Species Identification Of Cicadas And Orthopteran Through CNN – Based Sound Classification

Karan Sanjay Kajrolkar

Final Thesis Report

APRIL 2024

Table of Contents

ACKNOWLEDGEMENTS

I would like to thank my supervisor Dr. Akash Talwariya for all his help and advice with this thesis. I would also like to thank my parents; without whose support this would have not been possible. I also appreciate all the support I received from the other students in by batch. Lastly, I would like to thank the LJMU and Upgrad for the studentship that allowed me to conduct this thesis.

ABSTRACT

Insects play vital roles in ecosystems, affecting pollination, pest control, and nutrient cycling. However, the decline of certain insect populations has raised concerns about ecosystem health. By accurately classifying insect species through audio data, we can monitor and assess insect populations non-invasively. In this study we have demonstrated of insect acoustic classification by comparative investigation into the recognition capabilities of diverse deep learning models and proposing deep learning solution to automate the identification of insect species. In signal processing, features are extracted based on Learnable audio frontend LEAF. Data augmentation technique used to handle generalizability of model performance. Used orthopteran and cicada's species dataset published in xeno-canto publicly available dataset of 66 species. All files are mono audio with 44.1 kHz sample rate. However, orthopteran and cicadas know for ability of sound production. However, our proposed model has outperformed Baseline LEAF and Mel- Spectrogram frontend. We have achieved 92% accuracy over imbalanced dataset of insect audio.

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

LEAF…….. Learnable Audio Frontend

STFT..... Short-Term Fourier Transform

MFCC……. Mel Frequency Cepstral Coefficients

LFCC…… Linear Frequency Cepstral Coefficients

PCEN……. Per Channel Energy Normalization

CNN…… Convolution Neural Network

LSTM…….. Long Short-Term memory

ROC-AUC…… Receiver Operating Characteristic Area Under the Curve

TD Filterbank…… Time Domain Filterbank

AGC…… Automatic Gain Control

DRC……. Dynamic Range Compression

VGG……. Visual Geometry Group

sPCEN…… Smoothing Coefficient Per Channel Energy Normalization

LOGTBN……... Log1p (each frequency band) + Subtract Median filter + TBN (temporal batch normalization) compression function.

CHAPTER 1: INTRODUCTION

1.1 Background of the Study

Exploring animal vocalizations and the natural soundscape presents a captivating avenue for research, offering valuable insights into animal behaviors, populations, and ecosystems. The realm of acoustic insect recognition introduces a complementary dimension to conventional surveillance methods, such as camera-based approaches. Numerous significant insect groups exhibit distinct species-specific sounds, and the classification of these sounds holds the potential to revolutionize and expedite specimen identification, thereby enhancing the monitoring of biodiversity and species distribution. Given the vast number of insect species – reaching hundreds of thousands, the task of their identification is exceptionally intricate. To put this into context, in the Netherlands alone, 40% of total biodiversity is estimated to consist of insects, whereas mammals and plants are just 5% and 8%, respectively.

Insect population decreases have received considerable attention from both the scientific community and the general public, even though many of these papers either examine a limited number of species of interest or focus on a specific geographic region (Montgomery et al., 2019). Because of their size, camouflage, activity, or environment, certain species are difficult to view, but their noises can help identify them much more quickly. Additionally, this methodology is less complex and mostly non-invasive than other popular monitoring techniques (Riede, 2018). Due to their small size, ability to blend into their surroundings, and cryptic lives in sometimes inhospitable and challenging locations like tropical rainforests, insects are a particularly difficult group to identify using standard monitoring methods (Riede, 2018).

Even though there are many different kinds of insects, for this study we considered classifying two types, orthopteran and cicada. Cicadas are best known for their ability to produce loud sound (Moulds, 2009). The majority of the 3200 species in the Cicadidae family use speedily deforming tymbal membranes to make loud clicking noises that cause the tymbals to resonate (Young and Bennet-Clark, 1995). The sounds of orthopterans are often species-specific and serve a crucial role in the identification of species (Heller et al., 2021). However, the orthopteran order of insects includes the grasshoppers, crickets, and their relatives. The process by which orthopteran make sound, termed as stridulation, involves

rubbing one modified part of the body against another (Naskrecki, 2001) . A call's information may be stored by frequency modulation, where the pitch varies over time like it does in birds, time modulation where the pitch remains constant over the call's duration, but the temporal pattern is unique to the species, or a mixture of both types (Naskrecki, 2001).

Historically, when trying to identify Orthoptera based on their sounds, research has predominantly depended on manually extracting sound parameters such as carrier frequency and pulse rates (Riede, 1998). Before using these characteristics for automatic classification, their parameters must be manually chosen and made (Faiß and Stowell, 2023) .Because Orthoptera are poikilothermic species, the ambient temperature during the recording may have an impact on how frequently they sing. As a result, these parameters might not always work as intended (Dolbear, 1897).

1.2 Problem Statement

Deep learning has been investigated by the researcher as a flexible strategy for the problem of classifying and on input data achieving high-precision identification of acoustic signals with minimal manual pre-processing (Stowell, 2022; Faiß and Stowell, 2023). However, there are also challenges for acoustic monitoring such as background heavy rain, wind, loudness, distance of the sound source from the recorder etc.

Previous studies have combined audio features to increase performance by their features, such as the log mel and log gammatone (Chi et al., 2019) and the MFCC and LFCC (Noda et al., 2019). The researchers discovered that fusion features outperform cutting-edge methods.

Long-form field recordings can be classified when used in conjunction with sound event detection, eliminating the need for any manual feature extraction or relevant clip identification (Faiß and Stowell, 2023). Where (Noda et al., 2019) used (Härmä, 2003) procedure for segmentation of multiple calls of one insect species in single recording as result gives high performance except some insect not classified correctly due to bad segmentation of the calls due some unnecessary information low-amplitude noise gets captured because as formulation algorithm captures frequency if it is greater than threshold (max frequency - MINDB). As improvement on bad segmentation problem (Cicada et al., 2022) introduce improved Harma. However, (Cicada et al., 2022) removed low-amplitude noise helps the

model obtain the highest classification accuracy on dataset of species katydids, crickets and cicadas.

In the past, Mel filterbanks, an unchanging, manually-engineered representation of sound, were employed for audio classification tasks since strong features can result in high performance models (Zeghidour et al., 2021). Mel-filterbanks create a spectrogram first using the STFT's squared modulus. Then, to mimic the non-linear human perception of pitch, the spectrogram is run through a bank of triangular bandpass filters that are spaced on a logarithmic scale in mel-scale (Zeghidour et al., 2021). Finally, a logarithm compression is applied to the coefficients to mimic our non-linear sensitivity to loudness (Zeghidour et al., 2021).

As in recent study by (Faiß and Stowell, 2023) mentioned Insect sounds are produced utilizing stridulatory or tymbal mechanisms that produce a different structure of frequencies and overtones from source-filter systems used by mammals or birds. Many species of insects produce ultrasonic sounds, which are often significantly higher in frequency than the majority of mammal or bird sounds. Depending on how they go about it, the emphasis on high-frequency sounds that are occasionally completely and far outside of the human hearing range (between 20 Hz and 20 kHz) might affect how well audio categorization networks operate. The mel filter bank technique based on human perception is probably not optimal to identify and differentiate between slight variations in high frequencies for many insect noises, even though it works well enough for other sounds like bird song.

In recent work of deep learning audio classification introduce adaptive waveform-based methods such as LEAF (Zeghidour et al., 2021) represent a new generation of frontends implemented as layers in a neural network, which can be optimized alongside the model to better fit insect.

Additionally, most studies have compared the effectiveness of handcrafted and deep learning features (Siddhartha Varma et al., 2021; Anderson and Harte, 2022; Faiß and Stowell, 2023). The utilization of a SincNet based deep learning feature layer combined with a CNN layer yielded the highest accuracy, achieving 97% on insect database of cornell university (Siddhartha Varma et al., 2021). Where LEAF 83.7% performs worse than PCEN 89.9% (Wang et al., 2017) and TD 87.6% on BirdVox-DCASE-20k dataset combined with supplemented with 10s length, recorded at 44.1 KHz, and normalized to -2dBFS with

EfficentNet-B0 model, it has been determined that learnable filterbanks can increase performance (Anderson and Harte, 2022).

In their study, (Faiß and Stowell, 2023) contrasted conventional spectrogram-based audio representation with a novel neural network approach known as LEAF (Zeghidour et al., 2021) for analyzing insect sounds, particularly those beyond human hearing range. Using data from 66 insect species, they compared Mel-spectrogram and LEAF feature extraction methods and found that combining LEAF with a CNN led to an impressive 83% accuracy in classifying insect species. Notably, they revealed that individual LEAF component learnable filterbank shows higher contribution in overall LEAF performance. The study also suggested the potential of enhancing CNN performance through techniques like sound-event detection, thereby indicating avenues for further improvement in achieving higher classification accuracy.

Furthermore, mosquito classification in these studies have harnessed raw audio from mosquito wingbeats (Yin et al., 2023) applied a 1D-CNN + LSTM approach achieving a 93% accuracy in classifying species and sex based on wingbeats, exploring microphone factors and comparing Bayesian CNN techniques. (Bilal et al., 2023) employing variables including mel spectrogram, amplitude envelope, zero crossings, MFCCs, and spectral centroid, audio characteristics were manually extracted for CNN modeling. Considering the difficulty of class imbalance, their suggested model, a modified ResNet-50, achieved acceptable results with ROC-AUC, Precision Recall AUC with 0.921 and 0.885.

(Li et al., 2022) introduced a technique to classify diverse acoustic scenes by combining data augmentation methods and a lightweight ResNet model. Their approach included random time stretching and shifting for augmented training data, enhancing model robustness. They minimized complexity by using a compact ResNet variation and optimized it for low power devices with real time applications. The results of the experiments showed that the proposed method outperformed existing proposed techniques for acoustic scene classification on the bases of accuracy and efficiency.

(Zhang et al., 2023) , has demonstrated classification of urban sound scenes including insect, human and bird audio. where they achieved good performance on pre-trained model DenseNet, MobileNet and EfficientNet with the help hand-tunned mel-spectrogram. However, author described misclassification happened due complex sound patterns involving humans, birds, and insects has not fully learned during training.

In their work (Chu et al., 2023), a sound classification mechanism was proposed, which utilized a refined approach to hand-crafted MFCC features. The study addressed the limitation of classification efficiency due to data scarcity by focusing on Mel filters, designed to mirror the human auditory system's frequency response. This involves segmenting speech signals into distinct frequency bands using these filters, quantifying each band's strength logarithmically for use as acoustic features in classification. The identical signal produced somewhat different logarithmic energy representations depending on the different sizes of triangle bandpass filters used. The authors developed a method for data augmentation on the ESC-50 that outperformed the original dataset score of 63% by 97% using a particular set of 5 triangular bandpass filters. Classification accuracy in the UrbanSound8K dataset, where data volume is more balanced, attained 90%, marginally increasing to 92% following data augmentation. This augmentation proved instrumental in mitigating data imbalance and labeling challenges. Notably, the models were developed without accounting for real-world factors such as environmental noise, interference, and sound overlap, which significantly impact model accuracy during practical sound reception.

The researcher (Anderson and Harte, 2022) has discussed about how initialization of filterbank in learnable frontend LEAF can increase performance on 2 different frequency distribution domain like human and bird. However, using liner instead of mel used by LEAF perform better than another initialization. These changes can add valuable study in deep learning acoustic classification.

To explain key ideas and fill in knowledge gaps, a recent review by (Stowell, 2022), put a special emphasis on deep learning in computational bioacoustics. The review outlined a standard approach involving CNN architectures (such as ResNet, VGGish, MobileNet) pretrained on AudioSet, with spectrograms (linear, mel, log-frequency) as typical input data. Alternatives like wavelet representations, raw waveforms (e.g., Wavenet), or the LEAF frontend were also noted. Effective data augmentation techniques like noise mixing, time shifting, and mixing were highlighted, along with methods such as time warping and frequency adjustments to enhance dataset diversity. Given the often-imbalanced nature of bioacoustics datasets, Stowell suggested employing macro-averaging for evaluation, providing equal weight to each class by calculating and averaging class-specific performance (Mesaros et al., 2016).

Although the study found many researchers has focuses on how different feature extraction technique either hand crafted or learnable with standard CNN architecture can improve model

performance. As result they found learnable frontend outperform hand-crafted fronted (Siddhartha Varma et al., 2021; Anderson and Harte, 2022; Faiß and Stowell, 2023; Yin et al., 2023) It is necessary to further enhance findings when they propose classification, and it is possible to do so by comparing the model performances to determine which one is the most appropriate.

1.3 Aim and Objectives

The main aim of this research is to develop and evaluate an effective insect audio classification using CNN. Accurate classification model designed to identify distinct sounds, facilitating the monitoring and identification of groups of insect species in remote habitats.

The research objectives are formulated based on the aim of this study which are follows:

- To analyze the data augmentation and pre-process technique helps to handle data imbalance e.g., mixing, one of (Gaussian Noise, Pink Noise, Noise Injection) technique while using for models.

- To investigate the potential of transfer learning by fine-tuning a pre-trained CNN (e.g., from image classification tasks) on the insect audio dataset. Assess how this strategy improves convergence speed and classification accuracy.

- To analyze the learned features with the help of LEAF Frontend within the pre-trained CNN to gain insights into what audio characteristics contribute to accurate classification.

1.4 Research Questions

The following research questions are suggested for each of the research objective as highlighted as follows:

- Can deep learning techniques be employed to assess the capacity of various models in recognizing acoustic scenes associated with insects?

- Can data augmentation technique help to handle data imbalance?

1.5 Scope of the Study

The main purpose of the study to demonstrated comparative investigation into the recognition capabilities of diverse deep learning models and proposing deep learning solution to automate the identification of species of orthopteran and cicadas.

This study aims to focus on characteristic of learnable frontend feature for insect audio. This study covers various combination of data augmentation technique on raw waveform and spectrogram with different CNN model to increase generalization of model performance.

In this study we have worked on mono waveform type audio files not stereo. However, we are using chunk of 1-10s to handle dataset imbalance (Stowell, 2022; Faiß and Stowell, 2023). Furthermore, used PyTorch library includes pre-trained models for Acoustic insect classification. Dataset is imbalance due to some species have longer audio recording than other.

This study will be helpful to tracking population trends, assessing biodiversity, and identifying potential conservation threats, environmental monitoring, Agriculture and Pest Control.

For future studies need to focus on how insect syllable can extract from long audio know as Sound Event Detection. In above literature review (Anderson et al., 2023) has found using liner initialization instead Mel in LEAF can increase performance. However, need to address potential contributions of RNN (Recurrent Neural Network) and Hybrid-CNN models. By incorporating these models, we seek to gain valuable insights into their effectiveness in addressing the challenges posed by insect species identification through vocalization (Stowell, 2022).

1.6 Significance of the Study

For many species or texa vocalization are explored using Deep Learning Technique e.g. bird, bat, insect etc. However, they found deep learning can replace manual process to detect and analyze by using either images or audio signal. Image-based classification of insect species can be difficult, but it can supplement audio-based techniques and provide a visual perspective on the biodiversity and behavior of insects. However, difficulties are visual appearance, background noise, Differentiation in Appearance, Environmental Factor such are whether condition or lighting, Incorrect labeling due similar appearance insect.

This study designed to help identify acoustic insect through vocalization using deep learning. Author (Faiß and Stowell, 2023) proposed a methodology for comparison between Mel spectrogram and learnable frontend for feature extraction with own CNN model. However, they have not achieved promising performance. So instead of training model from scratch in this research we are applying some CNN models which pre-trained on large image database ImageNet which can help in reducing model size as well as performance.

This study specifically addresses the augmentation and pre-processing technique used for model generalization because dataset used in this study is imbalance. If this study is achieved good performance, then we can use this for filed study. Insects are fundamental components of ecosystems worldwide. This research carries global significance by addressing a universal challenge in biodiversity monitoring, with potential applications in ecosystems around the world.

By successfully applying deep learning to acoustic insect identification, this research opens doors to cross-species insights. Lessons learned from this study can potentially be transferred to similar endeavors in the identification of other species.

1.7 Structure of the Study

The structure of the study is as follows:

Chapter 1 – Introduction: This section serves as an introduction and provides the background context for the research endeavor.

Chapter 2 – Literature Review: This chapter outlines the existing works in the domains of classifying insect audio using CNN and feature extraction frontends.

Chapter 3 – Research Methodology: This chapter gives a detailed walkthrough of the methodology followed during the experimentation stage.

Chapter 4 – Analyze and Implementation: This chapter elaborates on the various experiments conducted for the task of insect audio classification, required augmentation and feature extraction technique.

Chapter 5 – Results and discussion: This chapter discusses the results of the experiments performed in Chapter 4.

Chapter 6 – Conclusion and recommendation: This chapter concludes the work done in the thesis and discusses future improvements.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

The growing body of evidence worldwide substantiates the looming risk of extensive insect population declines (Wagner, 2020). To enhance comprehension of insect declines amid data gaps and other obstacles, researchers require more systematic and prolonged monitoring of insect abundance and diversity. The primary bottleneck in insect monitoring lies in transitioning from trap data to accessible information. There is a critical need to expedite the time-consuming process of processing and identifying specimens to enhance efficiency in insect monitoring. Various machine learning techniques have been explored, and deep learning have emerged as a particularly popular approach for such tasks.

2.2 The Role of Deep Learning in Acoustic Classification

As outlined in recent literature by (Stowell, 2022) deep learning is an emerging field for conducting research on bioacoustics computation. Here, "Bioacoustics" refers to the study of animal sounds. Where deep learning are applied on various task as speech, music etc. Nevertheless, deep learning is employed for the analysis of vocalizations across various species or taxonomies, encompassing birds, cetaceans, other marine mammals, bats, anurans, insects, and fish.

Several of the mentioned taxa hold significant importance in biodiversity and conservation monitoring, such as birds, insects, and bats. Additionally, they play crucial roles in comparative linguistics and behavior studies, including rodents, cetaceans, primates, and songbirds (Stowell, 2022).So we mainly discuss the process adopted for bird and insects.

Distinct variations exist in the sound frequency ranges generated by birds, insects. However, majority of the noises that are used to identify insects are related to flight (Branding et al., 2023). Where insect flight sound much below than bird frequency range. However, birds lies between 3 kHz to 5 kHz and insect’s flight sound lies 60Hz to 200Hz (Branding et al., 2023). Nevertheless, dealing with acoustic insect sounds poses challenges, particularly in the

presence of environmental noise, which often falls within the lower frequency to 1/f noise. (Bennet-Clark, 1998).

Deep learning often used after typically precomputed spectrogram which are calculated using either CNN, RNN, Transformers (Stowell, 2022). Previous studies employed a multi-layer perceptron (MLP) (Koops et al., 2015) with manually crafted features such as syllable duration and peak frequency. However, these approaches have been surpassed in performance by CNN & RNN (Stowell, 2022). Deep learning not only assists in feature extraction but also in the identification of complex patterns directly from raw acoustic data. This eliminates the need for manual feature engineering, as learnable filterbanks, especially prominent in models like Convolutional Neural Networks (CNNs), play a crucial role in autonomously discerning and adapting filters to capture hierarchical and abstract representations from the input acoustic signals.

SincNet (Ravanelli and Bengio, 2019) adopts a unique approach by extracting rectangular bandpass filters for a more comprehensive understanding, diverging from the triangular filters. Conversely, Leaf employs Gabor 1D Convolution (Ravanelli and Bengio, 2019) , and the option for incorporating learnable time-domain filterbanks is available through Conv1D (Siddhartha Varma et al., 2021; Lostanlen et al., 2023).

Deep learning enables end-to-end learning, where the model learns to map raw input (acoustic signals) directly to output (class labels). This approach simplifies the overall pipeline and can lead to more efficient and effective classification systems. Deploying deep learning models for real-time acoustic classification becomes more feasible due to their adaptability, eliminating the necessity for domain-specific knowledge. This adaptability enhances their applicability across diverse real-world scenarios.

2.3 Implementing Learnable Filterbank

For an acoustic classification task, researchers used spatial and temporal features as inputs for deep learning models. A filterbank is a signal processing concept that involves the use of a bank of filters to analyze or process a signal. When dealing with audio, a filterbank can be employed to partition the incoming sound signal into distinct frequency bands. This enables a more thorough examination of the spectral content. Filterbanks play a crucial role in extracting pertinent features from audio signals, and the term "learnable" indicates that the parameters of these filters are optimized through the training process. Furthermore, resercher

adept either using frequency or time domain filterbank with passing coefficent as directly or produced by a parameterized function (Schlüter and Gutenbrunner, 2022).

2.3.1 Frequency – Domain Filterbank

Those techniques used to analyze and extract relevant information from frequency components of signals. A frequency domain filterbank comprises filters crafted to partition the input signal into distinct frequency bands. Generally non-parametric, these filters can be uncomplicated bandpass filters or more intricate ones such as Mel filters. The output of each filter signifies the energy or amplitude within the corresponding frequency range.

Furthermore, many researchers have chosen to utilize frequency domain filterbanks like mel-spectrograms, MFCC, and LFCC features as integral inputs to their models. Where author (Chi et al., 2019) adopt log Mel and log gammatone filterbank with logarithmic transformation is applied to better align with the perceptual characteristics of human hearing and (Noda et al., 2019) MFCC and LFCC in research. (Noda et al., 2019) proposed feature extraction method were used combined LFCC used for linear filterbank which has better in high resolution region with MFCC based on Mel-scale which mimic human auditory system. Similarly, (Siddhartha Varma et al., 2021) used filterbank initialized by Mel filterbank on insect Orthopteran (cricket and katydid) recognition achieved 95% accuracy on test dataset.

Furthermore, (Sainath et al., 2015) analysis is conducted to evaluate the difference between hand-crafted pseudo-filterbanks and learned filterbanks. These analyses demonstrate a correlation in the center frequencies of Mel-scale filterbanks and learned filterbanks. Similarly, (Seki et al., 2017) has adopted feature extraction technique with frequency domain parametric filterbank. He illustrated that learning the frequencies of Gaussian and triangular filters contributes to a reduction in parameters and enhances filter interpretability. However, it is important to note that this approach still relies on a manually adjusted Short-Time Fourier Transform (STFT).

2.3.2 Time- Domain Filterbank

Time- Domain representations of a signal directly derived from the amplitude values of the signal as a function of time. These features serve to elucidate various aspects of the signal's behavior within the time domain, thereby offering valuable insights into its temporal characteristics. The attempts to extract raw audio data feature so far different models have been proposed, which try to mimic the steps of generating a feature in the first layers of the model architecture by using convolution and pooling layers (Siddhartha Varma et al., 2021; Anderson and Harte, 2022; Branding et al., 2023; Faiß and Stowell, 2023)

Parametric time-domain convolutions play a crucial role in diminishing the number of trainable parameters and incorporating inductive biases, potentially enhancing the training process, especially when dealing with limited datasets (Schlüter and Gutenbrunner, 2022).

2.4 Using CNN for Insect Acoustic Classification

A significant distinction between human speech and insect sound signals lies in the requisite temporal context essential for accurate classification (Branding et al., 2023). However, the simplicity of insect sounds permits the utilization of considerably shorter signals. In addressing human speech tasks, variations of long-short-term-memory models (LSTM) are commonly employed (Branding et al., 2023). Conversely, the recognition of bird or insect audio can be effectively accomplished using simpler convolutional neural network (CNN) models (Noda et al., 2019; Cicada et al., 2022; Faiß and Stowell, 2023).

However, we can use CNN for feature extraction to identify complex pattern from raw audio, provides end-to-end solution and improve performance which lead to outperform traditional machine learning approaches. As feature extraction technique used by researcher are either using mel-filterbank which initialized by Mel scale having all parameter are fixed (Noda et al., 2019; Cicada et al., 2022; Faiß and Stowell, 2023) or by learnable filterbank where parameter learns while training. However, they have compared performance of multiple feature extraction technique.

In the existing literature, research has been conducted on utilizing acoustic characteristics of mosquito wingbeats to classify mosquito species. employed two (Yin et al., 2023) deep

learning approaches, namely 1D CNN and 1D CNN + LSTM, taking raw input signals for analysis. The study used a dataset comprising 6042 audio samples with a duration of 0.3 seconds and 0.15 seconds of overlap, encompassing four mosquito species found in Thailand, including Aedes aegypti and Aedes albopictus (vectors for dengue), Anopheles dirus (a primary malaria vector), and Culex quinquefasciatus (a common non-disease vector). The data were collected using an ECM8000 measurement condenser microphone and a low-cost microphone (Primo EM172), with mosquitoes placed in cylindrical containers.

The authors applied a simple auto-correlation function and Bayesian convolutional neural network (BCNN) to detect portions of audio signals containing mosquito wingbeat sounds. However, the experiment demonstrated poor performance in classifying the primary vectors for dengue and malaria, achieving F1-scores of 0.64 for 1D-CNN and 0.62 for 1D-CNN + LSTM.

When working with acoustic data, we face difficulties associated with gathering audio data in real environmental noises. In their study, (Bilal et al., 2023) employed the HumBerg mosquito dataset, which includes audio recordings from 36 species in diverse signal-to-noise ratios (SNR) and background environments spanning Thailand, Kenya, Tanzania, the USA, and the UK. The researchers manually extracted audio features for CNN modeling, specifically utilizing Mel spectrograms from the audio data. Addressing the challenge of an imbalanced dataset, they implemented oversampling of the minority class and down sampling of the majority class, alongside assigning class weights to mitigate imbalance. Remarkably, the utilization of a pre-trained ResNet-50 model on ImageNet demonstrated effective performance in classifying mosquito wingbeat sounds in their results.

In filled of multi-classification acoustic studies are done on urban forest including insect, human , bird or multiple cities scenes such as airport, shopping mall, metro station etc. (Zhang et al., 2023) conducted a comparative investigation into the recognition capabilities of diverse deep learning models for classifying biological acoustic scenes. They collected a range of audio categories, including human, insect (Cicadas), bird, silence using Song Meter SM4 acoustic recorders in an urban forest. Resampling audio to 22,050 Hz, 16-bit sampling, and durations of 3-5 seconds were standardized across all 1000 samples. Employing hand-crafted feature extraction through Mel spectrograms, the authors augmented data using techniques like noise addition, amplitude change, time shifting, frequency masking, and time masking with librosa. Several deep learning models were experimented including EfficentNet, two depth model for ResNet, DenseNet with bottleneck compression, edge device models like

MobileNet, Notably, the DenseNet_BC_34 model exhibited superior generalizability, yielding 93.81% score on overall validation dataset. This disparity possibly arises by test dataset encompassing novel sound patterns not fully learned during training, leading to occasional misclassifications, particularly among complex sound patterns involving humans, birds, and insects.

In prior studies focusing on deep learning, spectrogram representation emerged as the predominant input feature for acoustic classification (Stowell, 2022). However, CNN used spectro-temporal pattern from spectrogram representation. The key factors influencing the spectrogram representation of audio include the window size (which dictates the trade-off between time and frequency), scaling along the frequency axis, and amplitude compression (Anderson and Harte, 2022). Where insect sound are much higher frequency than birds 800-8000Hz (Anderson and Harte, 2022) are 150 kHz (Young and Bennet-Clark, 1995; Robinson and Hall, 2002). This high frequency far away from human hearing range. The Mel-filter bank approach, grounded in human perception, may not be optimal for discerning subtle differences in high frequencies present in various insect sounds. Nevertheless, numerous past endeavors to classify Orthoptera or Cicada based on their sounds have relied on different iterations of Mel spectrograms, occasionally in conjunction with other manually crafted features or spectrogram adjustments (Siddhartha Varma et al., 2021; Cicada et al., 2022; Faiß and Stowell, 2023; Zhang et al., 2023).

Moreover, SincNet (Ravanelli and Bengio, 2019) a learnable filterbank, demonstrates superior accuracy compared to manually crafted fixed filterbanks. In the context of Figure 1, SincNet and 1D-CNN share similarities. However, a notable distinction lies in the initial layer of SincNet, characterized by pre-defined filters. These filters consist of rectangular filterbanks, contributing to rapid convergence and a reduced number of parameters during training.

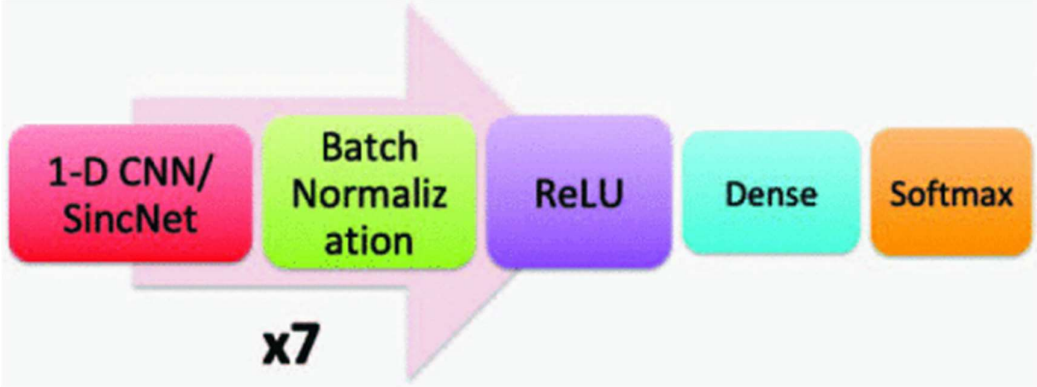


Figure 2. 1 : Architecture of 1-D Convolution Model (Siddhartha Varma et al., 2021)

By employing deep learning, the SincNet model was applied to an insect audio dataset belonging to the Orthopteran family, encompassing crickets and katydids. This approach yielded an impressive overall performance of 97%, surpassing the results reported in existing literature (Siddhartha Varma et al., 2021). Notably, each signal was individually analyzed, and samples of noise were eliminated from the entire dataset using the open-source speech processing tool, Audacity.

In their recent work (Zeghidour et al., 2021) , the authors introduced LEAF, a learnable filter that demonstrated promising outcomes when compared to fixed Mel spectrogram, SincNet, and Time-Domain filterbank. The evaluation encompassed diverse datasets such as speech commands, acoustic sounds, music, and birdsong. LEAF comprises a sequence of three integral components: Gabour Filterbank (Gabor, 1946) , Gaussian low-pass pooling designed to down sample the signal for reduced temporal resolution, and PCAN (Wang et al., 2017). Notably, all parameters, including smoothing, in PCAN are learnable. Where they use EfficientNet-B0, CNN as resulted both model with LEAF has performed well than other filterbanks. While training author used 1s window sample from audios.

After these proposed learnable filterbank feature many authors has worked upon how this better than fixed filterbanks. Where they have experiment on ASR (Automatic Speech Recognization) (Sainath et al., 2015; Seki et al., 2017), Insect (Branding et al., 2023; Faiß and Stowell, 2023) and bird (Zeghidour et al., 2021; Anderson and Harte, 2022; Schlüter and Gutenbrunner, 2022) classification.

Leaf (Zeghidour et al., 2021) will have two convolution layers for Gabor filtering and low pass pooling. Following this, the results will be sequentially passed through PCEN for normalization. However, it is important to note that this process is two orders of magnitude slower than typical spectrograms. As improvement on high computation and sequential normalization of PCEN, Efficient LEAF has proposed (Schlüter and Gutenbrunner, 2022).

Furthermore, instead sequential PCAN Normalization used learnable logarithmic compression, temporal median subtraction and temporal batch normalization, all of which are parallelizable and faster. For experiment it used 5 datasets include speech, birdCLIF, Synthetic audio, Crema-D (6 emotion), VoxForge (6 language) and deployed EfficientNet-B0 model. However proposed method helps to speed up learning faster than LEAF. Furthermore, experiment on different audio excerpt (1, 8, 32) Sec. Author resulted that Efficient Leaf outperform Original LEAF by computation speed on long sequences without impacting accuracy. EffiecintLeaf has observed that learned center frequencies and bandwidths do not deviate much from their initial value. Due to unexplored part of learnable window and stride are still seem room for improvement. However, author achieved 42.9% accuracy over BirdCLEF 2021dataset.

Moreover, learnable frontends are adaptable to various domains, as demonstrated by (Faiß and Stowell, 2023) who utilized LEAF to extract features for insect audio. They examined 66 species of orthoptera and cicadas in the Netherlands, employing a 2-second expert with data augmentation techniques such as impulse response, Gaussian noise, and shift. The results indicated that LEAF outperformed the standard Mel-spectrogram. Furthermore, the study concluded that learnable filterbanks (Gabour, Low-pass normalization) played a more significant role than learnable PCEN. The author suggests that sound event detection may aid in accurately detecting insects in noisy environments.

Moreover, it was noted in the training process that learnable filters with a backend remain unchanged after their initialization (Schlüter and Gutenbrunner, 2022; Faiß and Stowell, 2023). To Identifying this observation (Anderson et al., 2023), conducted experiments using LEAF (Gabour filterbank) initialized with different scales. Typically, all frontend initializations follow the Mel scale to simulate human speech tasks. The authors employed Mel, linear, bark, and random initializations. In the experiment, they utilized Efficient LEAF (Schlüter and Gutenbrunner, 2022) which incorporates a dynamic filterbank and pooling approach to reduce computation, while retaining the original LEAF (Zeghidour et al., 2021), PECN for further exploration.

For Voice Activity Detection, the TIMIT corpus was employed, utilizing the Spectro-Temporal Attention Voice Activity Detection model known for its noise robustness and lightweight characteristics. Additionally, the Bird Species Audio data was used with the baseline EfficientNet-B0 (Zeghidour et al., 2021) model, and all experiments were conducted using 40 filters. Augmentation techniques involved incorporating extra silence and additive noise.

The findings revealed that fixed filterbank learnable frontends outperformed the log-Mel spectrogram baseline features. In the context of this study, fixed filterbank learnable frontends refer to using a filter bank to analyze a signal. Unlike traditional fixed filterbanks with predefined parameters, the parameters of this filterbank are learned from the data during the training process. Notably, the linear initialization demonstrated strong performance in voice activity detection and bird species identification.

Additionally, apart from LEAF (Branding et al., 2023), an alternative approach involved incorporating a NBF layer for feature extraction and classification using a Wave Net Classifier for handling multi-channel insect raw audio. This method effectively employed the NBF (neural beamforming) layer to filter out background noises by leveraging spatial information inherent in the multichannel audio signals. The presence of fungus gnats in the audio recordings led to misclassifications. Notably, the baseline model (Mel Spectrogram with VGG16) accurately detected loud-flight insects, while the raw model utilized spatial information from the audio to mitigate the impact of noise. This allowed for the identification of the sound produced by Aphidoletes aphidimyza insects, which is inaudible to the human ear but discernible through high-quality measurements hardware.

2.5 Summary

In Insect audio classification task mainly differentiate into collecting data, pre-process, feature extraction and classification. However, focus mainly technique are uses less parameter, interpretable, less computation cost, accurate. Furthermore, author experimented on hand-crafted filterbank Mel-spectrogram but this learning parameters and choices are outside the training loop, they might be chosen sufficiently well to achieve a working model, but will never be chosen perfectly (Branding et al., 2023). To overcome these challenges learnable filterbank help in which mimic the step using convolution layer and train parameter while training to get with end-to-end solution.

Hence, this study is built upon the work of (Faiß and Stowell, 2023) who demonstrated the enhancement of the unaltered version (LEAF) compared to the Mel-spectrogram. Our emphasis is on conducting experiments with learnable filterbank on 66 species of insect audio. We employed a pre-trained CNN, including the baseline model EfficientNet-B0 from (Zeghidour et al., 2021) to achieve higher accuracy compared to the baseline paper.

CHAPTER 3: RESEARCH METHODOLOGY

This section will outline the general approach employed, including the mathematical models and algorithms that were implemented.

3.1 Introduction

Insects may be small, but they play a crucial role in ecosystems worldwide, providing irreplaceable benefits such as balancing ecosystems, supporting other animals, enabling plant pollination and reproduction, aerating soil, and much more. Losing even a few species could be catastrophic to biodiversity. Bioacoustics monitoring and classification of animal communication signals has developed into a powerful tool for measuring and monitoring species diversity within complex communities and habitats (Riede, 2018). The development of accessible digital sound recording technology and significant advancements in informatics, including big data, signal processing, and machine learning, have advanced the field of bioacoustics over the past few decades (Stowell, 2022).

Numerous studies are done as far on acoustic sound classification using deep learning. However, orthopteran and cicada know for ability of sound producing insect where some species of orthopteran produce ultrasonic sound. Deep learning algorithms are more frequently used in acoustic scene ecology for species-specific identification and target sound recognition. Many studies done so far for detection, classification of insect (Noda et al., 2019; Cicada et al., 2022; Bilal et al., 2023; Faiß and Stowell, 2023; Yin et al., 2023; Zhang et al., 2023). They used either fixed feature frontend (e.g., Mel-spectrogram), learnable feature frontend (e.g., SincNet, TD-filterbank, LEAF) or compare both to investigate performance.

Where (Faiß and Stowell, 2023) proposed robust deep learning model with comparison of leaf and Mel-spectrogram on various set of datasets of species orthopteran and cicada. They used 66 orthopteran and cicada species to experiment as learnable frontend outperforms fixed frontend. Given this foundation, we employed various deep learning models to train on these acoustic scene samples and subsequently evaluated and compared their classification performance.

3.2 Research Methodology

3.2.1 Data Description

In this investigation, two types of auditory family were employed, including orthopterans and cicada. The process of reviewing and examining recordings obtained from BioAcoustica, xeno-canto, iNaturalist, and private collections curated by Baudewijn Odé was undertaken. However, recordings that exhibited pronounced noise interference, extensive filtering artifacts, or included sounds from multiple species were excluded in order to curate and assemble those datasets. Publicly available data for classification are standardization to 44.1 kHz mono WAV format, with durations spanning from less than one second to several minutes. To address extended periods without insect sounds, file have edited into multiple smaller segments, ensuring that silent intervals did not surpass for 5 seconds. In public insect audio dataset (Faiß, 2023), there are a total of 2887 recordings representing these 66 species.

Table 3. 1: Dataset Description of 66 species of insect with labels

Label	Species	n	hh:mm:ss	Label	Species	n	hh:mm:ss	Label	Species	n	hh:mm:ss
0	Achetadomesticus	35	0:56:48	22	Eumodicogryllusbordigalensis	30	0:10:55	44	Pholidopteralittoralis	60	0:04:00
1	Aleetacurvicosta	23	0:04:04	23	Eupholidopteraschmidti	34	0:09:39	45	Platycleisalbopunctata	17	0:24:44
2	Atrapsaltacollina	11	0:01:19	24	Galangalabeculata	43	0:06:16	46	Platypleuracfcatenata	32	0:17:46
3	Atrapsaltacorticina	15	0:02:15	25	Gampsocleisglabra	28	0:55:17	47	Platypleuraplumosa	30	0:14:41
4	Atrapsaltaencaustica	15	0:04:33	26	Gomphocerippusrufus	51	0:29:38	48	Platypleurasp10	19	0:17:55
5	Barbitistesyersini	59	0:19:59	27	Gomphocerussibiricus	31	0:26:04	49	Platypleurasp12cfhirtipennis	12	0:07:41
6	Bicoloranabicolor	23	0:09:19	28	Gryllusbimaculatus	26	0:28:44	50	Platypleurasp13	15	0:07:01
7	Chorthippusalbomarginatus	26	0:40:29	29	Grylluscampestris	81	1:37:38	51	Popplepsaltaaeroides	10	0:01:46
8	Chorthippusapricarius	52	0:28:35	30	Leptophyespunctatissima	35	0:26:47	52	Popplepsaltanotialis	14	0:02:58
9	Chorthippusbiguttulus	118	0:30:25	31	Melanogryllusdesertus	29	0:25:24	53	Psaltodaplaga	14	0:04:20
10	Chorthippusbrunneus	95	0:21:57	32	Metriopterabrachyptera	19	0:20:55	54	Pseudochorthippusmontanus	73	0:11:35
11	Chorthippusmollis	84	0:27:49	33	Myrmeleotettixmaculatus	79	1:05:37	55	Pseudochorthippusparallelus	75	0:25:07
12	Chorthippusvagans	55	0:11:43	34	Nemobiussylvestris	52	0:38:43	56	Roeselianaroeselii	63	0:34:34
13	Chrysochraondispar	43	0:15:35	35	Neotibicenpruinosus	15	0:04:40	57	Ruspolianitidula	12	0:12:35
14	Cicadaorni	21	0:06:50	36	Oecanthuspellucens	48	0:30:32	58	Stauroderusscalaris	26	0:20:43
15	Clinopsaltaautumna	22	0:04:15	37	Omocestuspetraeus	56	0:09:21	59	Stenobothruslineatus	39	0:34:27
16	Conocephalusdorsalis	26	0:23:07	38	Omocestusrufipes	66	0:16:33	60	Stenobothrusstigmaticus	71	0:05:31
17	Conocephalusfuscus	98	0:53:33	39	Omocestusviridulus	92	0:45:48	61	Tettigoniacantans	110	0:58:10
18	Cyclochilaaustralasiae	13	0:01:52	40	Phaneropterafalcata	41	0:28:29	62	Tettigoniaviridissima	33	0:27:26
19	Decticusverrucivorus	42	1:15:03	41	Phaneropteranana	35	0:30:50	63	Tylopsislilifolia	44	0:03:30
20	Diceroproctaeugraphica	11	0:05:06	42	Pholidopteraaptera	48	0:10:54	64	Yoyettacelis	160	0:11:16
21	Ephippigerdiurnus	43	0:39:50	43	Pholidopteragriseoaptera	48	0:14:06	65	Yoyettarepetens	41	0:05:23

In Table 3.1 the name of species, total length of audio file per species and their average duration over species are shown. The combined recordings amount to over 24 hours in length, with each species having a minimum of 10 files. Furthermore, In Figure 3.1 they have provide train, test and validation split folder according to distribution of recording have for per species.

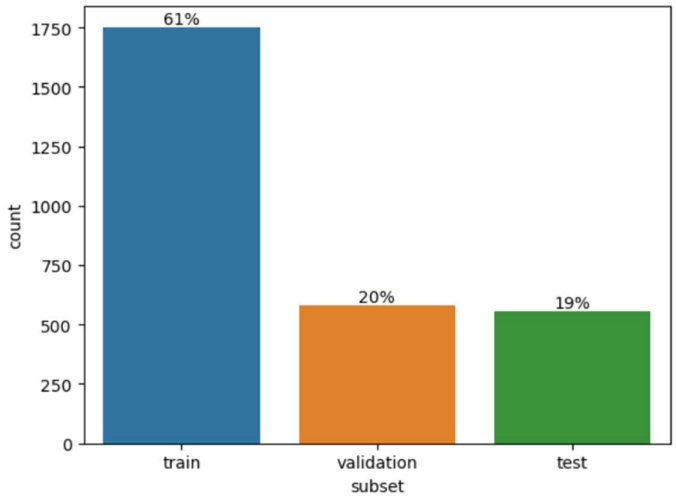


Figure 3. 1 : Dataset distribution for train, validation and test

3.2.2 Data Transformation

When handling audio data, various established methods exist for transforming an audio waveform into a feature representation. One of the most widely adopted approaches involves converting the audio into a time-frequency representation shown in Figure 3.2.

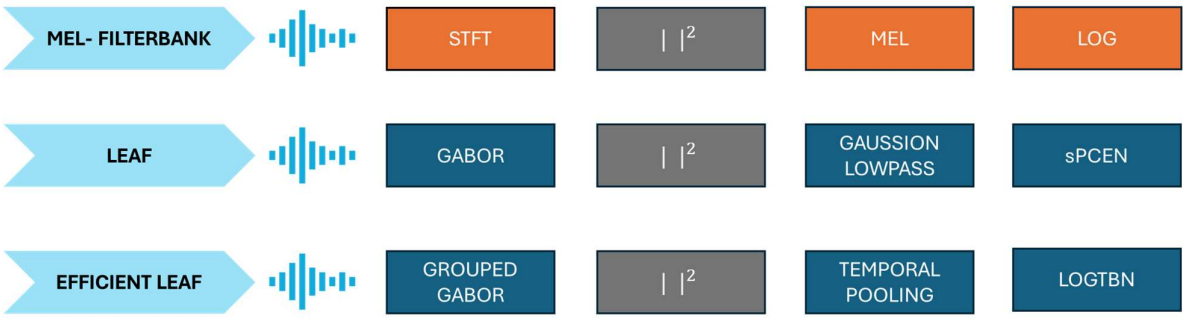


Figure 3. 2: Breakdown of computation of mel-filterbank, LEAF, EfficientLEAF. Orange boxes are fixed. While computation in blue boxes is learnable. Grey boxes represent activation functions.

In this thesis LEAF (Zeghidour et al., 2021) a Learnable frontend used which try to mimic process of extract handcrafted feature spectrogram using CNN layers. However, model can learn task-specific features directly from the raw audio data. This may be particularly beneficial if the task requires capturing intricate patterns or context-dependent features in the audio signals.

Where LEAF has divided into 3 component filtering, pooling, compression / normalization. The output filtering obtains Mel initialized gabour filter φ_n to compute bank of complex-valued filters as (1) where their frequency response is almost zero for negative frequencies and when combined with square modulus the resulting filterbank as a set of sub band Hilbert envelops, output following time-frequency representation. However, sub band Hilbert envelops allows for the analysis of amplitude modulation characteristics within different frequency components of a signal, which can be useful in tasks such as feature extraction.

$$\varphi_n(t) = \ e^{i2\pi\eta_n t} \frac{1}{\sqrt{2\pi\sigma_n}} e^{-\frac{t^2}{2\sigma_n^2}} \ , n = 1, \dots, N, t = \ -\frac{W}{2}, \dots, \frac{W}{2} \tag{1}$$

Where bandwidth and center frequencies are σ_n , η_n , impulse response as t, φ_n as bank of filters. The output of filtering component passed to second part pooling which down sample to lower sampling rate as shown in Equation (2). Which help in reduce its temporal resolution using Gaussian impulse response.

$$\Phi_n(t) = \ \frac{1}{\sqrt{2\pi\sigma_n}} e^{-\frac{t^2}{2\sigma_n^2}} \ , t = \ -\frac{W}{2}, \dots, \frac{W}{2} \tag{2}$$

The resultant time-frequency has passed to 3rd component for compression and normalization. Where replacement of log compression and mean-variance normalization to replicate non-human perception of loudness in mel-filterbank, LEAF used PCEN (Wang et al., 2017) as shown in equation (4). Where time- frequency $\mathcal{F}$ are normalized by exponential moving average of its past value as equation (3) where smoothing coefficient as s.

$$\mathcal{M}(t, n) = (1 - s)\mathcal{M}(t - 1, n) + s\mathcal{F}(t, n) \tag{3}$$

$$PCEN(\mathcal{F}(t, n)) = \left(\frac{\mathcal{F}(t,n)}{(\mathcal{E} + M(t,n))^{\alpha_n}} + \delta_n\right)^{r_n} - \ \delta_n^{r_n} \tag{4}$$

Where δn offset, r_n , α_n exponents and channel dependent s_n smoothing coefficient are trainable parameter in PCAN. Furthermore, resultant output passed to CNN classifier.

In deep learning, "generalization" signifies a neural network's ability to perform effectively on previously unseen data, as highlighted by (Zhu et al., 2020) . Data augmentation aims to enhance the training dataset's complexity and size, expanding the feature space for improved decision boundaries, thereby reducing false positives and false negatives and ultimately enhancing generalization on unfamiliar samples (Zhu et al., 2020). This section will provide a brief discussion of common augmentation techniques (Freer et al., 2020; Li et al., 2022; Chu et al., 2023) employed for audio classification tasks within the time domain.

However, to develop models with increased robustness against noise during the training process includes Noise Injection, Gaussian Noise and Pink Noise.

Audio augmentation employs noise injection by introducing controlled background sounds or distortions to training audio data. This enhances the model's adaptability to real-world audio variations, improving its overall performance and robustness. Injects Gaussian noise using a randomly chosen signal-to-noise ratio. Pink noise in audio augmentation is a balanced sound that mimics natural environments. It helps test and improve audio models by adding varied frequencies for a more realistic experience. One of these techniques has been implemented during the wave transformation, introducing a specified probability.

Impulse Response (IRs): Another method for dealing with complex noise effects is to include false negative noises (non-insect audios). However, by incorporating a small number of comparable types of noise samples into the training data, it is possible to successfully reduce this kind of false-positive prediction.

Random shift: The model can exhibit positional bias when trained with data anchored to a fixed reference time or with a narrow time shift around that reference point. In such cases, instead of learning broader patterns from the feature space, the neural network tends to memorize the specific anchor time, potentially limiting its ability to detect insect signals beyond the training shift window (Zhu et al., 2020) .

For applying transformations directly on the spectrogram of the audio signal, such as time or frequency masking by using Spec Augment (Park et al., 2019). After Augmentation require normalization techniques include scaling the values between 0 and 1 or standardizing them

with zero mean and unit variance. However, normalizing these values can improve convergence during training and prevent any feature from dominating the learning process.

3.2.3 Models

Deep learning is a subset of machine learning that utilizes multiple hidden nodes and nonlinear transformations to abstractly represent intricate data (Zhang et al., 2023). Convolutional neural networks (CNN) have gained popularity in image recognition, speech understanding, and various domains due to their ability to extract interconnected features from input data utilizing techniques like to those observed in the human brain (Zhang et al., 2023). For identification of insect, we are using convolutional models with head. However, for convolutional models we adapt transfer learning technique. A head is a single linear layer with exactly as many outputs as classes.

3.2.3.1 Transfer Learning

A model learned on one task or dataset is adjusted or improved for another, often related, task or dataset using the machine learning process known as transfer learning. Utilizing prior knowledge, it enhances performance on the new work while conserving time and resources. VGG, ResNet, DenceNet, Inception, Mobile Net, and EfficentNet are all CNN models used for the sound detection job. These models are all representative and have been used extensively in sound identification (Xiang et al., 2020; Zeghidour et al., 2021; Anderson and Harte, 2022; Anderson et al., 2023; Zhang et al., 2023).

However, ResNet solves the gradient disappearance issue brought on by increasing model depth by providing Skip Connection, allowing for the eventual building of models with more layers (He et al., 2016). ResNet can be used to process audio spectrograms for sound classification, and the residual connections support the maintenance and learning of key audio features over deep layers, resulting in enhanced performance (He et al., 2016). Similarly, an investigation of how to deepen such networks led to the development of the VGG, a standard convolutional neural network design of small 3 x 3 filters. Compared with ResNet, all layers are connected by DenseNet (Huang et al., 2017) more aggressive dense connection algorithm, and each layer takes input from all levels that came before it. In order to reduce the model

parameters as much as possible, the researchers added Bottleneck layers with Compression procedures to the model-building process to produce the DenseNet-BC model.

The Inception architecture proposed by (Szegedy et al., 2014), also known as GoogLeNet, employs multiple filters of different sizes and concatenates their outputs. This helps capture both local and global features effectively. Similarly, in sound classification, you can design filters to capture different frequency ranges or temporal patterns in audio signals, allowing Inception-like models to capture diverse sound characteristics. MobileNet is designed for efficiency and is well-suited for applications where computational resources are limited. In sound classification, MobileNet can be used to create lightweight models that can run on devices with lower processing power (Howard et al., 2017). However, audio data into spectrograms and leverage Mobile Net's depth wise separable convolutions to efficiently process them.

The effective design suggested by (Tan and Le, 2019). They discovered that depth, width, and resolution are the three key dimensions that have an impact on neural network accuracy. They used the neural architecture search (NAS) (Tan and Le, 2019) technique to obtain a model backbone called EfficientNet_b0, and then scaled the above three dimensions based on this backbone to obtain the models b1 to b7 with an amazing performance. EfficientNet-B0 is also used extensively to evaluate the original implementation of LEAF (Zeghidour et al., 2021).

Furthermore, the VGG, Inception, ResNet, DenseNet, MobileNet, and EfficientNet models were all initially pre-trained on the ImageNet dataset, which is a large dataset containing millions of labeled images from various categories. Pre-training on ImageNet allowed these models to learn general features from images, such as edges, textures, shapes, and higher-level object representations.

3.2.5 Evaluation

3.2.5.1 Accuracy

Accuracy measures the proportion of correctly classified insect sounds out of the total predictions. It provides an overall view of your model's performance.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

3.2.5.2 Precision & recall

Both metrics help you understand the model's ability to correctly classify insect sounds and avoid false positives. Precision is interpreted as the ratio of instances for accurately predicted positives to all instances of expected positives. The ratio of accurately predicted positive cases to actual positive instances is known as recall.

$$P = \frac{TP}{TP + FP} \qquad R = \frac{TP}{TP + FN}$$

3.2.5.3 Confusion Matrix

A confusion matrix provides a detailed breakdown of true positive, true negative, false positive, and false negative predictions. It helps you identify which classes are being misclassified.

3.2.5.4 F1-score

The F1-score is the harmonic mean of precision and recall. It balances the trade-off between precision and recall and is particularly useful when classes are imbalanced.

$$F1 = \frac{2.TP}{2.TP + FP + FN}$$

3.2.5.5 Macro-averaging

In bioacoustics datasets, addressing class imbalance is often achieved by employing macro-averaging, where individual performance metrics are computed for each class independently, and then their averages are taken to ensure equal consideration for each class. (Stowell, 2022).

3.3 Tools

3.3.1 Hardware

The hardware used for this project are listed below:

- Used Sage maker Notebook Instance
  - GPU (ml.g4dn.xlarge - Tesla-V4 GPU) having 15GB memory
  - Memory used 100 GB
- S3 Storage Bucket to store checkpoint and augmented datasets.

3.3.2 Software

The software used for this project are listed below:

- Python 3.9
- Lighting PyTorch for model building
- Used torch_audiomentations, librosa, torchaudio libraries for data augmentation.
- Matplotlib and seaborn to plot charts.

CHAPTER 4: ANALYSIS AND IMPLEMENTATION

4.1 Introduction

In this chapter, we delve into a comprehensive analysis of the data collected and the findings derived from our research study. This analysis serves as the backbone of our thesis, offering insights, interpretations, and implications drawn from the empirical evidence gathered. The analysis chapter not only showcases our ability to rigorously examine the research data but also highlights our capacity to critically evaluate and interpret the results in the context of the research objectives and theoretical framework.

4.2 Exploratory Data Analysis

While exploring the dataset, we gained deeper insights into the data. However, we discovered that some audio clips contained more than one insect audio clip. We utilized a total of 2887 audio clips corresponding to 66 species of cicadas and orthopteran. Additionally, distribution charts depicting the count of audio files with respect to subsets and the average length of audio for each species are presented in Figures 4.1 and 4.2, respectively.

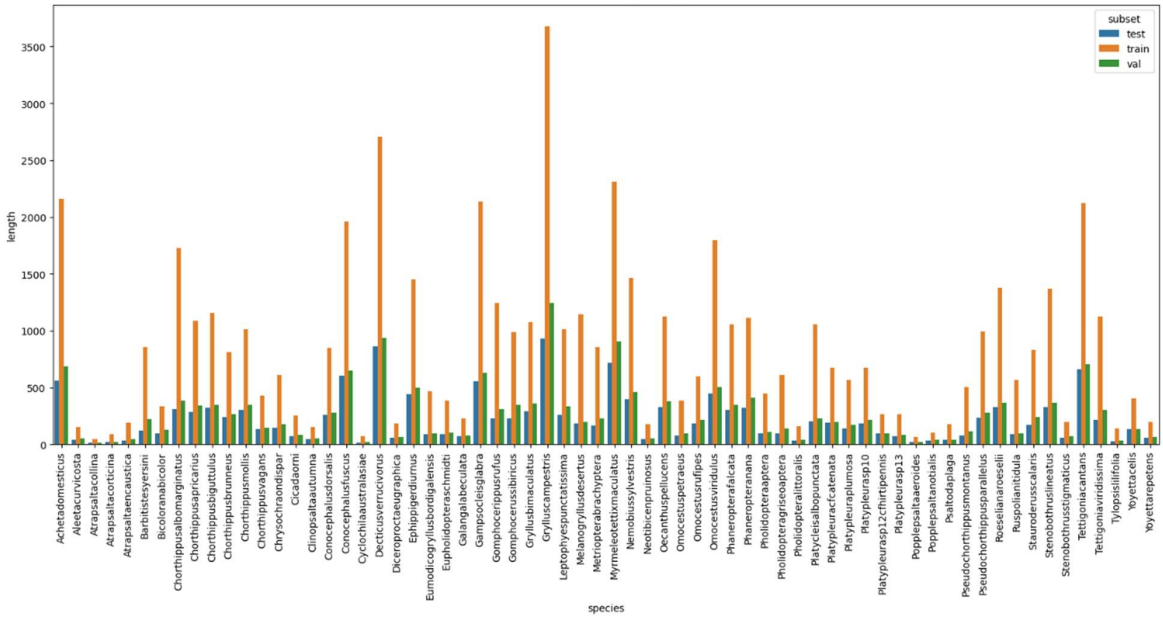


Figure 4. 1 : Average length of files per species over 3 subsets (train, val, test).

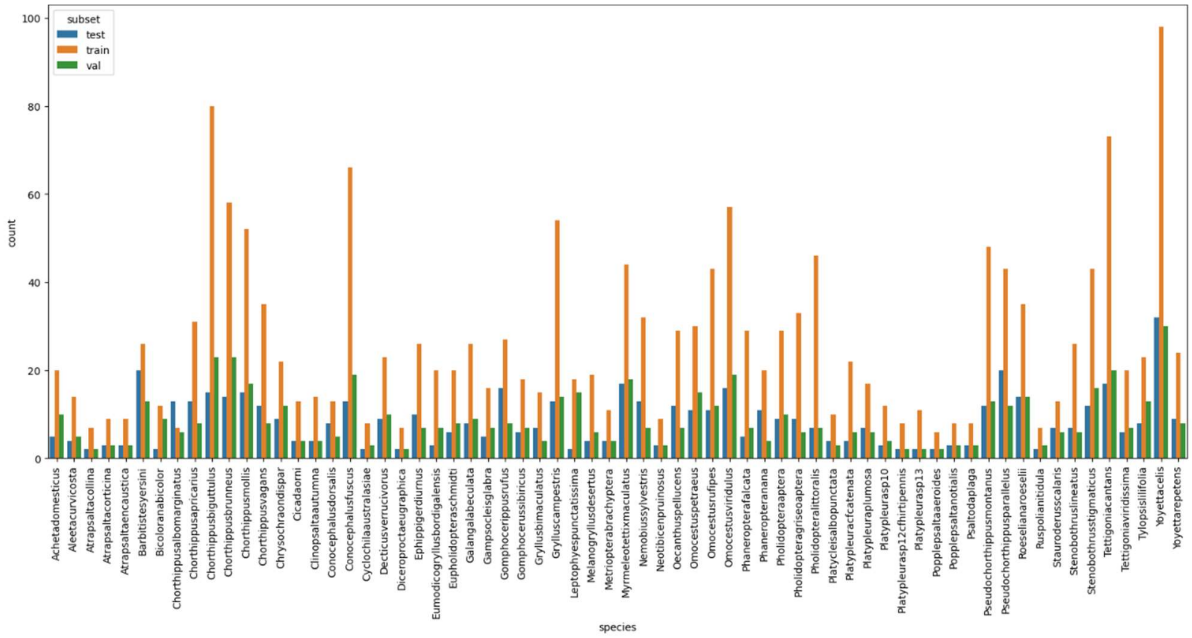


Figure 4. 2 : Number of files count per species over 3 subset (train, val, test)

We noticed that some audio excerpts had durations shorter than 5 seconds, yet they were valuable as they contained insect calling sounds, despite also containing human and environmental noises. Additionally, we observed that certain audio clips exhibited lower frequency spectra characteristic of insect species, exemplified by the spectrogram in Figure 4.3, which depicts the low-frequency sound of the Achetadomesticus species. Notably, the call of the Achetadomesticus species is evident from 150s to 250s in the spectrogram.

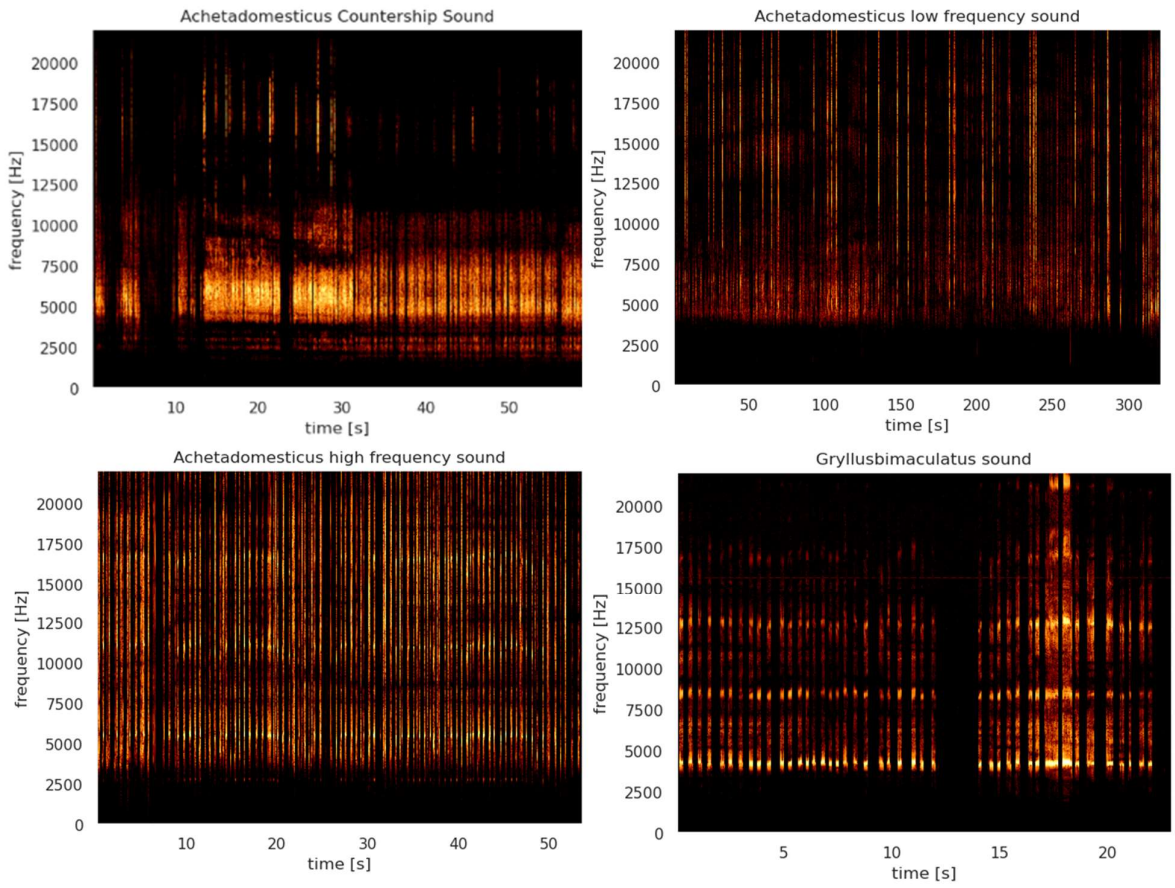


Figure 4.3 :Spectrogram representation of audio file of Achetadomesticus and Gryllusbimaculatus species.

4.3 Data Pre- Processing

Considering the different lengths of the recordings, it was necessary to partition them into segments of a consistent length suitable for input into the model. However, we have taken longer sequence of 5.5 sec with overlapping window of 2.75. As the splitting window reached the end of a file, the initial section of the recording was looped back to ensure the chunk extended to five seconds. Audio files shorter than number of samples per chunk are padded with zeros or looped to number of samples per chunk However it helps in handle imbalance dataset by increase count of files in subset of Train, Test, Val by 21060, 5415, 6438. To tackle imbalanced during training we have used cross_entropy function to give more attention to low probability species.

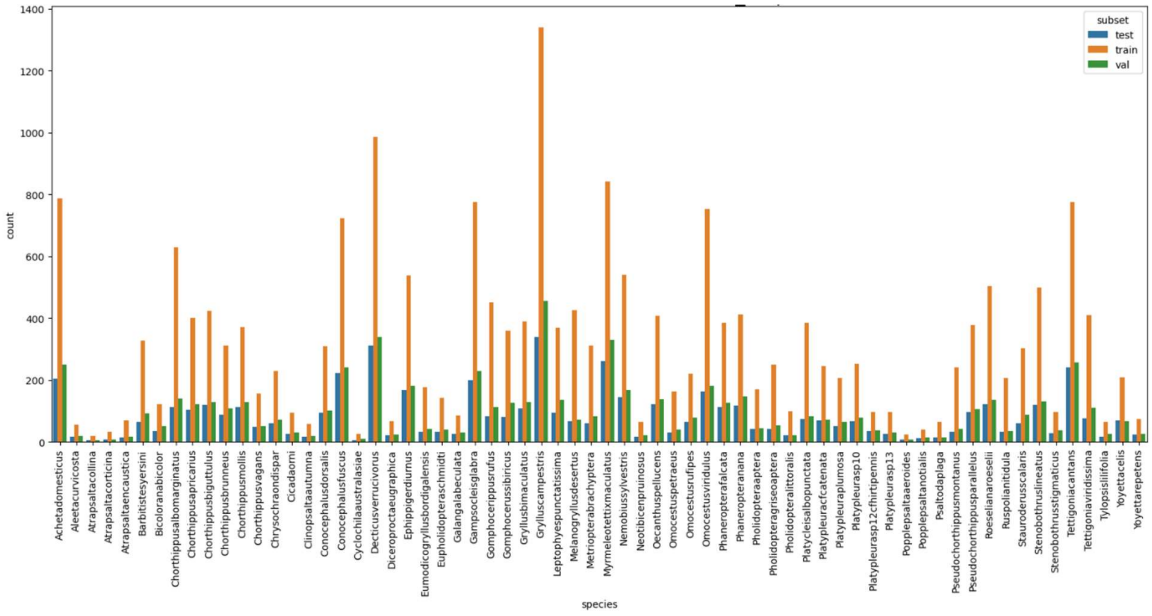


Figure 4. 4: Count of files in all subsets after 5.5sec Chunk

In deep learning, it's considered a best practice to enhance the generalizability of a dataset by applying data augmentation techniques. This aids the model in encountering a wider array of features during training. However, in our case, we employed the PyTorch Python library, torch audio, which offers a range of methods for augmenting audio data. Furthermore, we blended the original wave audio file with noise at a probability of 0.25. However, noise injection was applied with a probability of 1, ensuring a maximum noise level of 0.05. Gaussian Noise and Pink Noise were introduced at a single noise ratio ranging from 5dB to 20dB, each with a probability of 1. During the preprocessing step, we randomly selected one of these noises and applied it to the raw audio based on their probabilities over batches.

We incorporated mono audio impulse responses into our data using thes "ApplyImpulseResponse" function from the PyTorch Python library "torch_audiomentations". These impulse responses (IRs) originate from various sources such as buildings and spaces. Twenty-seven responses were randomly selected and applied with a probability of 0.25, thereby introducing additional variation in the severity of the effect.

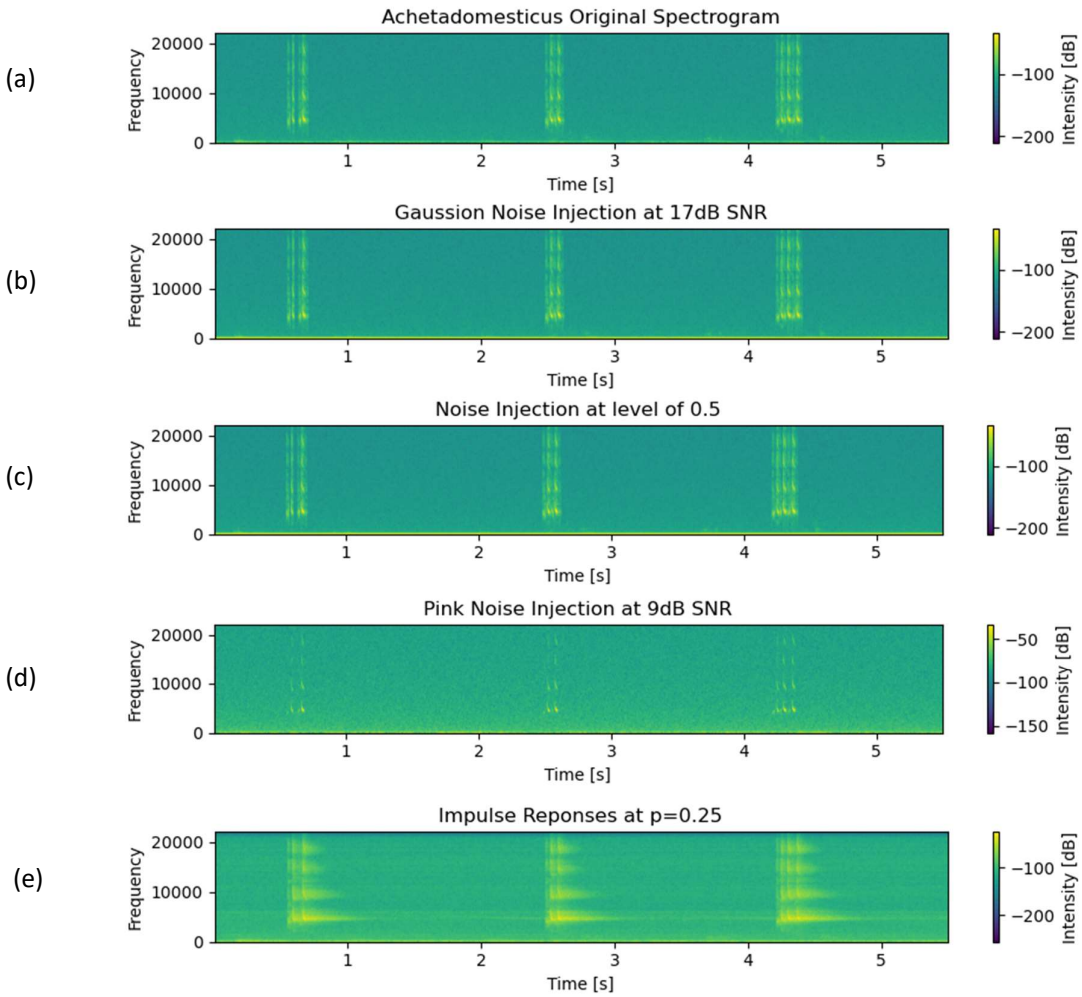


Figure 4. 5: Noise injection (b,c,d,e) data augmentation technique applied on waveform (a)

As shown in above figure 4.5(b,c), Noises are added at bottom of spectrogram shown by yellow line. However, on figure 4.5(d) the "colorednoise" Python library was employed to generate pink noise, where Gaussian distributed noise was applied with a power law spectrum featuring arbitrary exponents. Where Signal to noise ration are applied randomly between minimum 5dB to maximum 20dB, Further on figure 4.5(e) we can see impulse responses have applied one of effect over IRs mono files.

Additionally, we applied transformations to the spectrogram using the PyTorch library "torchaudio.transform". The transformations utilized were Oneof between MaskingTime and MaskingFrequency with a probability of 0.25 over each sample of batch and applied while training model. However, MaskingFrequency masks a random frequency range over each

individual sample, while MaskingTime masks a random time range over each individual sample within the batch. The parameter "time_mask_param" is utilized to define the maximum length of the mask, either along the time or frequency axis, within the range [0, time_mask_param), applying distinct masks to each sample in the batch. Moreover, we used a mixing procedure that blends two random samples in the batch with a mixing probability of 1. The impacts of masking on the original spectrogram are depicted in Figure 4.6.

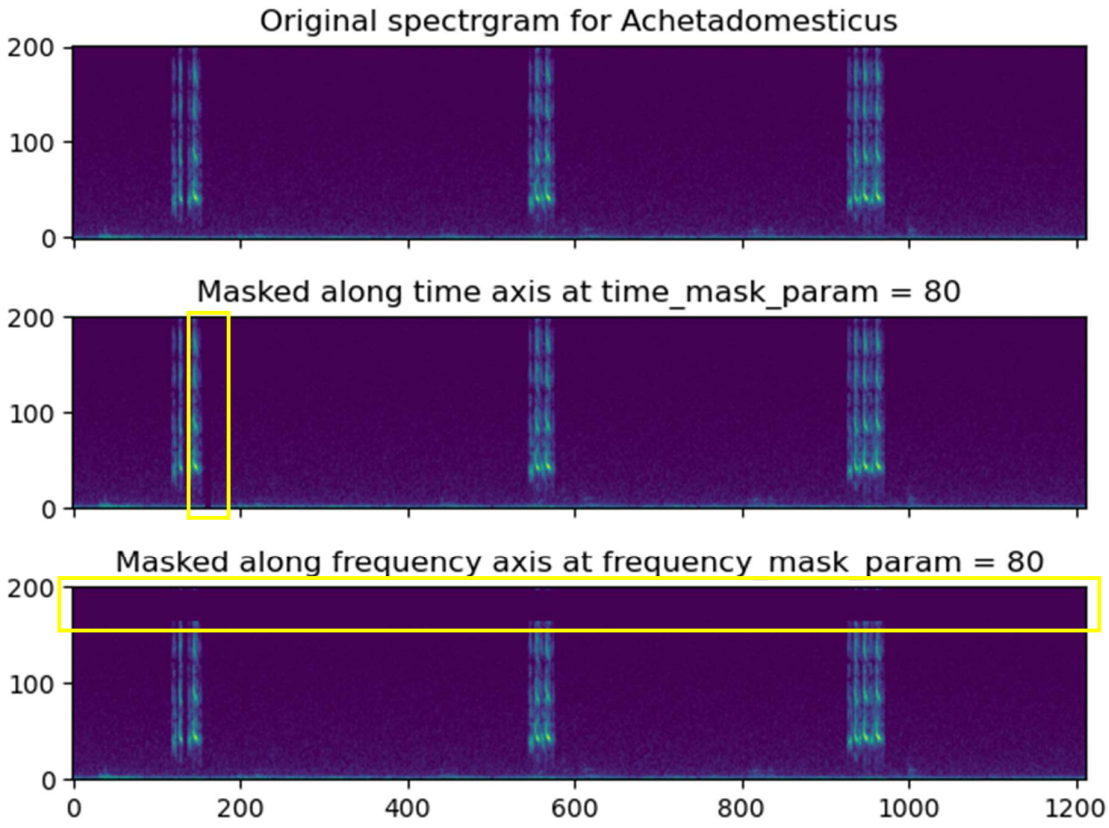


Figure 4. 6 : Data Transformation (FrequencyMasking and TimeMasking) over spectrogram

4.4 Model Building

In this experiment, we aim to demonstrate how utilizing a pre-trained model with a learnable frontend can enhance insect audio classification. Augmented audio files are fed through LEAF, a learnable frontend. LEAF's parameters, including filter frequency, bandwidth, per-channel compression and normalization, and low-pass pooling, are adjusted during training. These learned parameters enhance the network's capability to classify insect audio accurately. However, we have used pytorch implementation of leaf done by (Yadav, 2021; Faiß and Stowell, 2023). We employed default parameters, including the initialization of Mel in Gabor filterbank with 40 filters and a window stride of 3.335. All audio files were processed at a sample rate of 44100. Throughout training, we utilized PCEN for compression and normalization. The input is collected from leaf are fed into network.

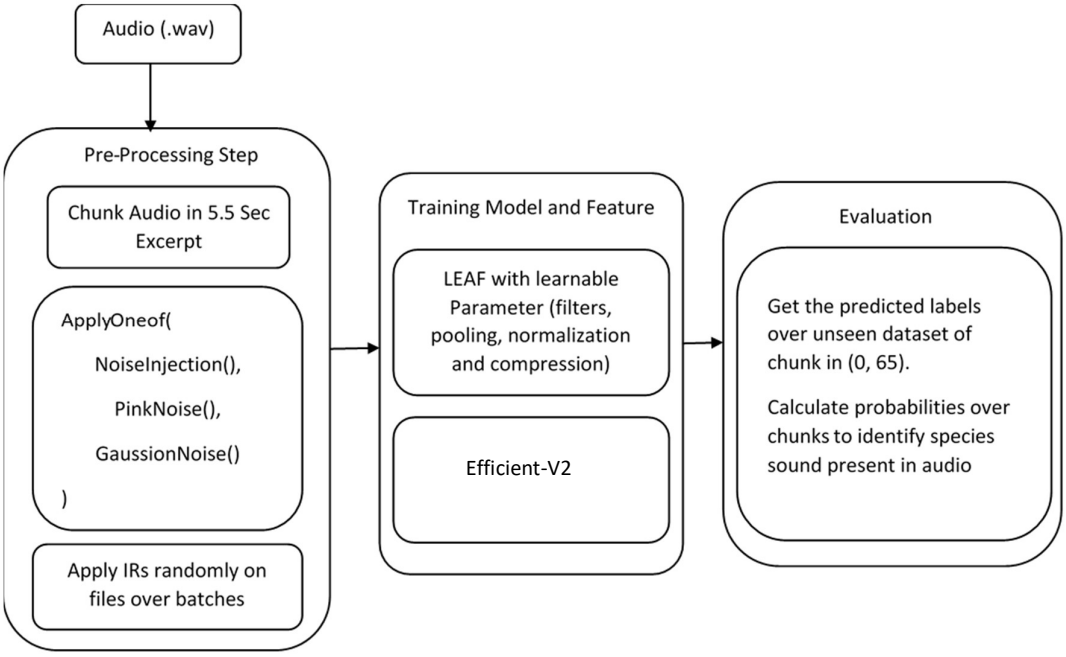


Figure 4. 7: Flow Architecture for Insect Audio Classification

Furthermore, we have also used Mel Spectrogram frontend with EffiecintNet-V2 model deployed by (Lemm, 2024) as baseline model for evaluation. However, Mel spectrogram frontend from pytorch library “torchaudio” where a spectrogram for raw audio created before passing to neural network. However, we have used 128 Mel bin filterbank with 2048 number of FFT. However resultant Mel spectrogram are normalized and used window size of 1024.

Also passed minimum frequency of 300Hz. Furthermore, these resultant tensors passed to pre-trained model.

In our backend network, we implemented PyTorch Lightning, leveraging a pre-trained baseline PyTorch EfficientNet-v2 model trained on a dataset containing 21,000 images. This model comprises 21 million parameters. Throughout our experiments, we utilized its pre-trained weights. We integrated TensorBoardLogger with PyTorch Lightning to enable the logging of various metrics and their visualization. During training, we continuously monitored the Learning Rate at logging intervals spanning epochs. Additionally, we implemented Early Stopping based on the validation F1-score, with a patience level set to 5 epochs. This strategy allowed us to halt training if there was no improvement in validation F1-score over the specified number of epochs.

Furthermore, we saved model checkpoints, retaining the model with the maximum validation F1-score, while also storing the latest checkpoint for continued training or analysis. This ensured that we could resume training from the most optimal point and track the progress of our model's performance throughout the training process. To address data imbalance during model training, we applied a weighted cross-entropy loss function over classes. This function assigns weights inversely proportional to the class frequency, calculated as 1 - ({class_total} / {total_files_of_classes}). Additionally, we incorporated label smoothing with a factor of 0.1 to regularize the predicted probabilities on cross entropy. This approach helped mitigate the effects of imbalanced class distributions, ensuring that the model learned from all classes effectively and generalized well to unseen data. Throughout the model construction phase, we employed the Adam optimizer with specific parameter settings: a weight decay and learning rate both set to 1e-4. This optimizer configuration facilitated efficient training by adjusting the learning rates of model parameters based on their gradients. We have used CosineAnnealingLR Schedular with Adam optimizer.

Subsequently, we conducted extensive training sessions lasting over 100 epochs. Notably, we have disabled the reloading of the data loader after each epoch. This optimization strategy minimized unnecessary overhead, ensuring a smoother and more efficient training procedure.

However, in experiment we have tried different batch sizes to train model accurately, as batch size is increasing the model size also increases. We have trained mel-spectrogram with EfficientNet-V2 on batch-size of 64 and LEAF with EfficientNet-V2 on batch size of 16,

EfficientLEAF with EfficientNet-V2 on batch size of 32 with however large batch size improves model performance.

CHAPTER 5: RESULTS AND DISCUSSIONS

5.1 Introduction

In this section we are discussing about result gain out of applying process over chapter 4.

5.2 Performance on Unseen Test Dataset

For assessment purposes, we employed macro-averaging metrics to assess the accuracy of predicting each class. Notably, we utilized two distinct approaches for calculating these metrics. The first approach involved passing the entire data batch to the model to predict multiple insect classes simultaneously. The second approach involved predicting each class by selecting k-random windows from a single file for scoring. Subsequently, final predictions were averaged over all k runs.

We assessed the model's performance on both full-length audio recordings and 5.5-second segments of the audio test dataset. However, some species within the full audio recordings were not accurately identified due to the presence of human voices, ambient noises, and other insect sounds, which obscured the actual insect sounds.

However, In Table 5.1 we clearly seen effect of data imbalance over evaluation matrix. However, if accuracy is lower compared to macro-average precision and recall, it could indicate that the model is performing relatively well on some classes but poorly on others, particularly in the case of imbalanced datasets where accuracy might be skewed by the dominant class. Therefore, macro-average precision and recall provide a more comprehensive evaluation of the model's performance across different classes.

Our approach using EfficientNet-v2 with pre-trained weights outperforms the base paper (Faiß and Stowell, 2023)on LEAF. We achieved a precision of 84%, surpassing the 81% reported in the base paper for LEAF. While our recall score remains similar at 77%, the F1-score reaches 78%, compared to the base paper's 77%, indicating an improvement in performance with our approach featuring learnable parameters. Despite its computational expense, LEAF remains slower compared to the Fixed mel-spectrogram frontend.

To address this, we adopted the EfficientLEAF Model, leveraging grouped Gabor filterbank at its conservative settings, which is 3 times faster (Schlüter and Gutenbrunner, 2022). Additionally, by substituting PCEN with log compression, median filtering, and temporal batch normalization, it achieved another 5% improvement in speed. EfficientLEAF with

default initial parameters demonstrates remarkable results on accuracy over unseen datasets compared to LEAF frontend, while maintaining superior computational efficiency. However, with default initial parameters and batch size of 32, EfficientLEAF has outperform the performance of the fixed-mel spectrogram over batch size of 64. However, it achieved an accuracy of over 92% across the total audio files, outperforming the mel-spectrogram's accuracy of 91%. Where Batch Sizes play crucial role to achieve good score. Furthermore, EfficientLeaf as achieved best performance on batch size 32 over mel-spectrogram batch size 64 as shown in Table 5.1.

Table 5. 1 : Evaluation table over all models with accuracy, precision, recall and f1-score.

Model		Over Each Species Audio	Macro- Averaging		
	Batch Size	Accuracy	Precision	Recall	F1-Score
Baseline Mel Spectrogram + EfficientNet-v2	64	0.91	0.92	0.89	0.89
Baseline LEAF + CNN Conv2D Model	-	0.83	0.81	0.77	0.76
Leaf (40 filters) + EfficientNet-v2 (pre-trained)	16	0.82	0.84	0.77	0.78
EfficientLEAF (40 filters) + EfficientNet-v2 (pre-trained)	32	0.92	0.89	0.88	0.87

5.3 Interpretation of Visualization

Upon scrutinizing misclassified instances in the test dataset, we discovered that the absence of Achetadomesticus counter ship audio in the training set led to misclassifications across all models. Notably, the genuine insect sounds within the full audio recordings were correctly identified over all models, as illustrated in Figure 5.1,5.3, 5.5 confusion plot for all models. However, Confusion plots shows count of audio files predicted correctly.

Additionally, audio signatures of insects Platypleurasp12cfhirtipennis and Platypleuraplumosa exhibit similarities. Among the 7 audio files of the Gryllus bimaculatus species, 4 files have been incorrectly classified as Achetadomesticus when using the mel-spectrogram and EfficientLEAF model. The macro-averaged precision, recall, and F1-score are depicted in Figures 5.2, 5.4, and 5.6, respectively, illustrating the performance across each species audio files.

However, the species Roeselianaroeselii exhibits misclassifications across 7 out of 14 files, with errors placing them within the insect families of Tettigoniidae (long-horned grasshoppers) and Acrididae (short-horned grasshoppers) within the Orthoptera order across all models. Additionally, the species Gryllusbimaculatus is misclassified in the mel and efficientLeaf models, being mistaken for species Achetadomesticus of the same cricket family Gryllidae. Some of the audio recordings of Achetadomesticus also contain noises resembling those of the Gryllusbimaculatus species. Interestingly, these misclassifications predominantly occur between species belonging to the same family, such as Gryllidae (crickets), Tettigoniidae (long-horned grasshoppers), Acrididae (short-horned grasshoppers), and Cicadidae.

An alternative method to ensure accurate classification involves segmenting the audio clips into 5.5-second chunks and performing predictions on each segment. Subsequently, grouping the predicted species across the chunks allows for the calculation of probabilities over the entire audio clip. However, some audio files may contain one or more insect sounds, with the actual insect audio present in small amounts throughout the entire clip, which can still be captured. For instance, considering one of the misclassified audio clips of the grasshopper species Roeselianaroeselii, as depicted in Table 5.2.

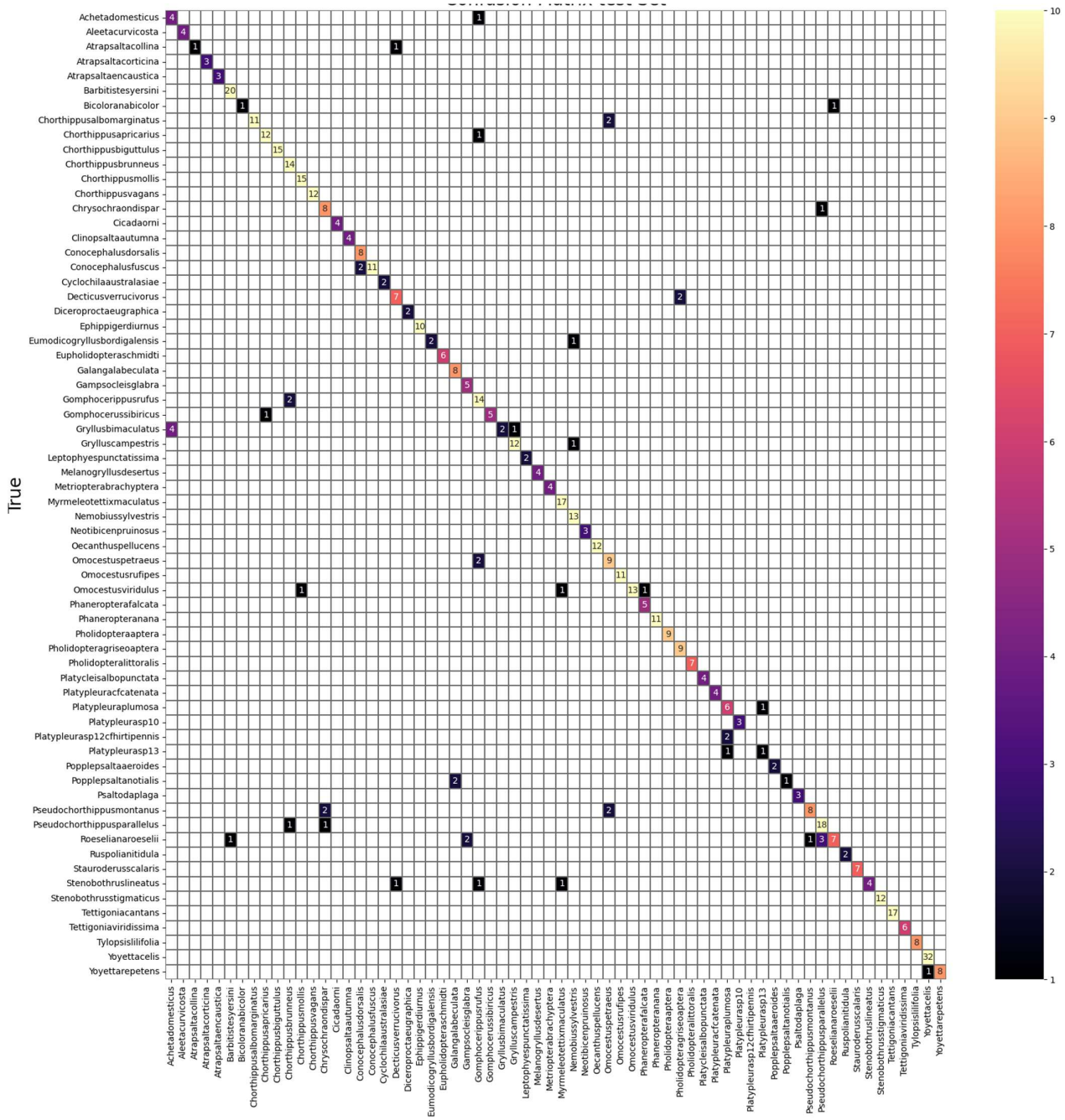


Figure 5. 1 : Confusion Matrix representation of Mel spectrogram + EfficientNet-v2

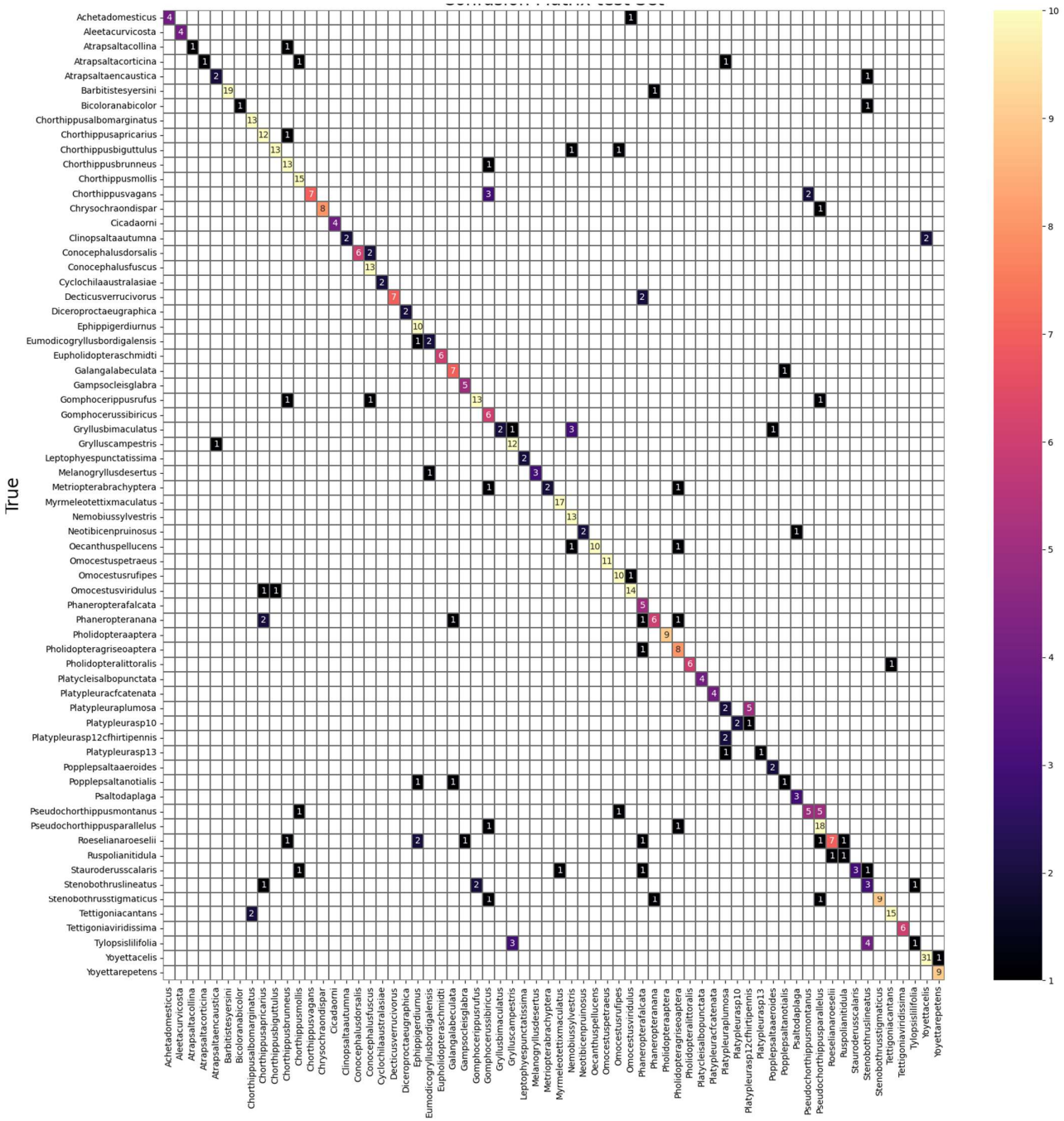


Figure 5. 2: Confusion Matrix representation of Leaf + EfficientNet-v2

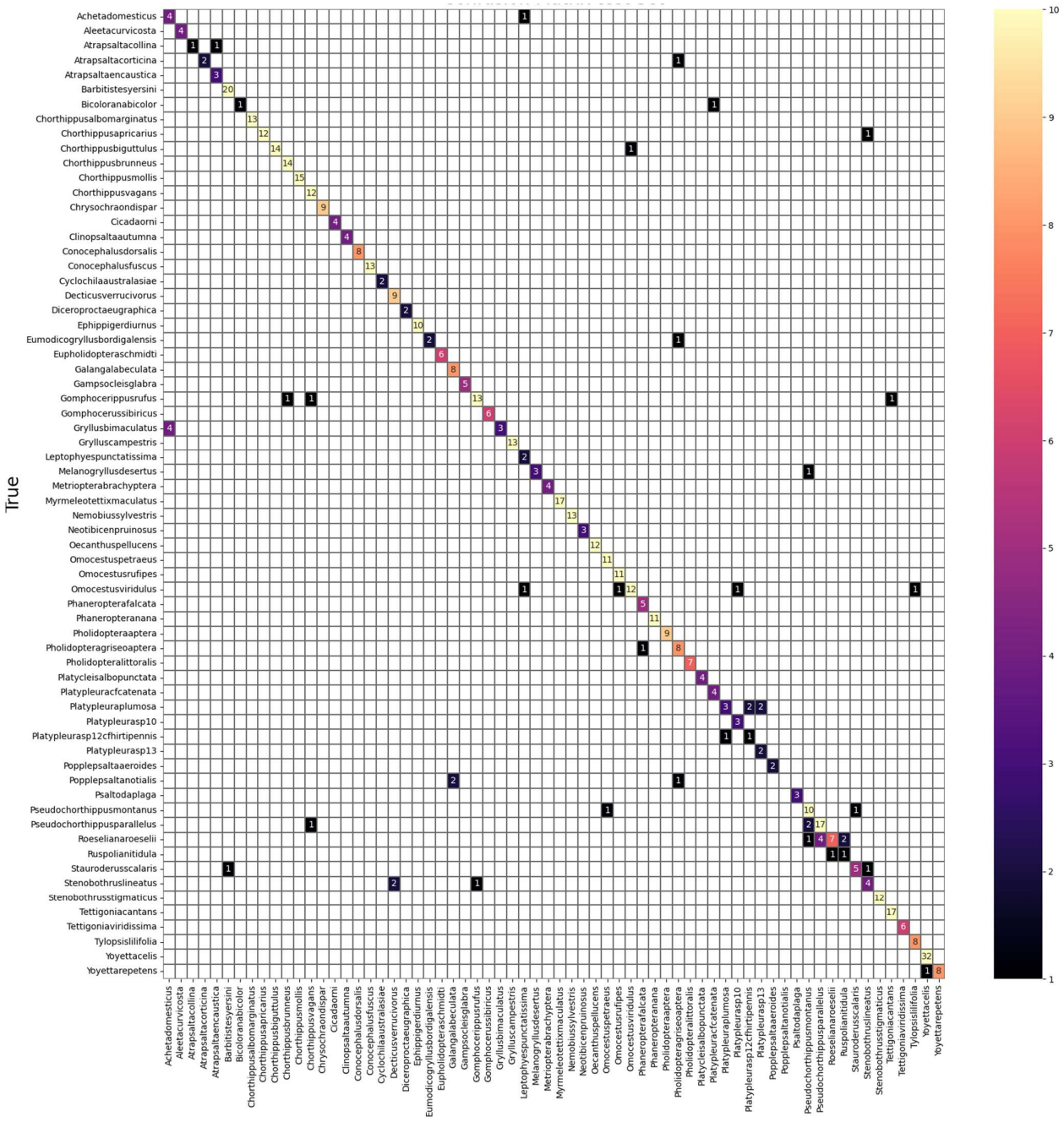


Figure 5. 3: Confusion Matrix representation of EfficientLEAF + EfficientNet-v2

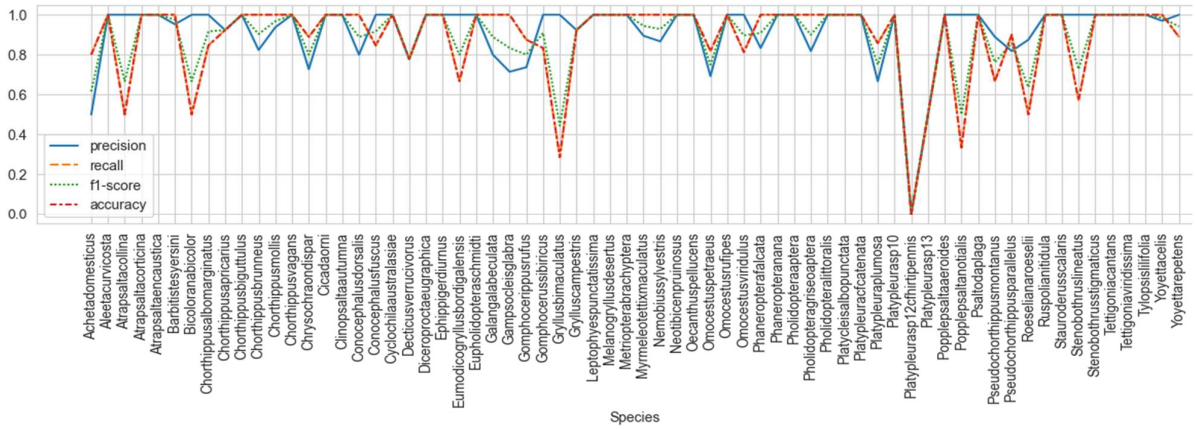


Figure 5. 4 : Evaluation Plot for Each Species (Baseline MelSpectrogram + EfficientNet v2)

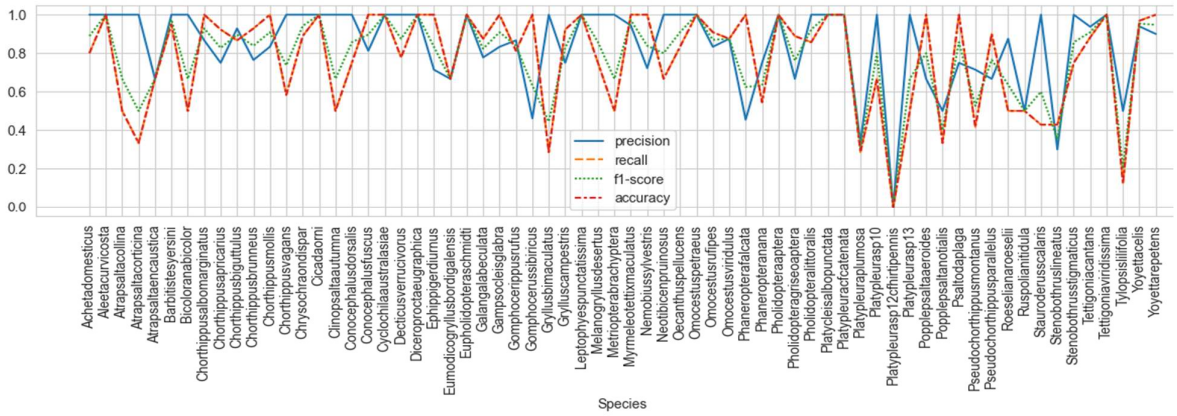


Figure 5. 5: Evaluation Plot for Each Species (Leaf + EfficientNet-v2)

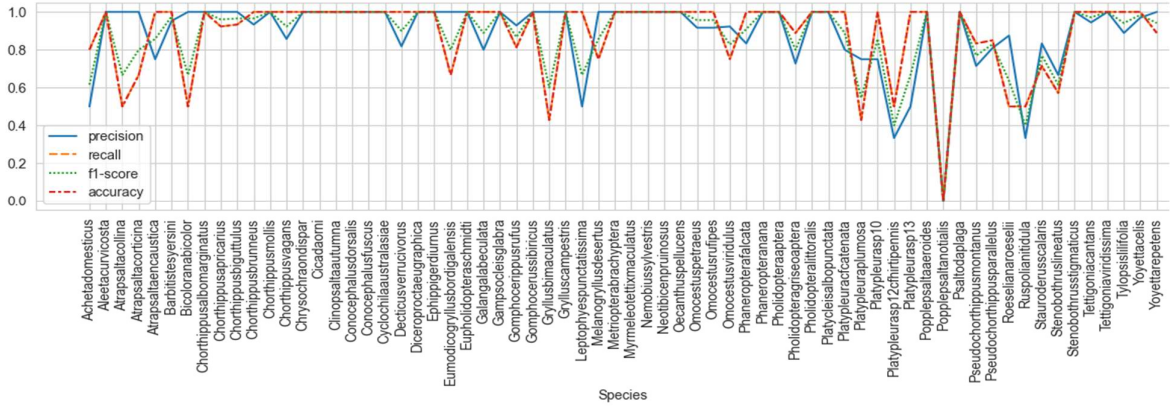


Figure 5. 6 : Evaluation Plot for Each Species (Efficient LEAF + EfficientNet-v2)

Table 5. 2: The table displays the probability of predicted species over 7 misclassified audio clip of Roeselianaroeselii.

File Name (From Unseen Test Dataset)	Predicted Species	Count of 5 Sec Chunk audio	probability
Roeselianaroeselii_GBIF2283054430_IN27820070_42340_edit1.wav	Roeselianaroeselii	1	0.25
Roeselianaroeselii_GBIF2283054430_IN27820070_42340_edit1.wav	Ruspolianitidula	3	0.75
Roeselianaroeselii_GBIF2283054430_IN27820070_42340_edit2.wav	Ruspolianitidula	5	1
Roeselianaroeselii_XC751671-dat015-003_edit2.wav	Pseudochorthippusmontanus	1	0.333333
Roeselianaroeselii_XC751671-dat015-003_edit2.wav	Pseudochorthippusparallelus	2	0.666667
Roeselianaroeselii_XC751671-dat015-003_edit3.wav	Pholidopteragriseoaptera	1	0.111111
Roeselianaroeselii_XC751671-dat015-003_edit3.wav	Pseudochorthippusparallelus	2	0.222222
Roeselianaroeselii_XC751671-dat015-003_edit3.wav	Roeselianaroeselii	6	0.666667
Roeselianaroeselii_XC751671-dat015-003_edit4.wav	Pholidopteragriseoaptera	1	0.25
Roeselianaroeselii_XC751671-dat015-003_edit4.wav	Pseudochorthippusparallelus	3	0.75
Roeselianaroeselii_XC751671-dat015-003_edit6.wav	Pseudochorthippusmontanus	2	1
Roeselianaroeselii_XC751671-dat015-003_edit7.wav	Pseudochorthippusparallelus	2	0.25
Roeselianaroeselii_XC751671-dat015-003_edit7.wav	Roeselianaroeselii	6	0.75

5.4 Summary

Upon reviewing the results of all models, it was observed that efficientLEAF achieved the best scores compared to LEAF Fronted. Moreover, it exhibited computational efficiency, outperforming LEAF in this regard. Additionally, the efficientLEAF model trained with a batch size of 32 surpassed the performance of the mel-spectrogram model trained with a batch size of 64.

CHAPTER 6: CONCLUSIONS AND RECOMMENDATIONS

6.1 Introduction

In this thesis, we embarked on a comprehensive exploration into the realm of audio classification, with a particular focus on the efficiency and effectiveness of various models in discerning between different species of insects. The primary objective of this research was to investigate an learnable frontend with data augmentation and pre trained models are capable of accurately identifying insect species based on audio recordings. With the increasing need for automated species recognition systems in ecological monitoring and biodiversity studies, the significance of this endeavor cannot be overstated. However, certain conclusions derived from this thesis are outlined below.

6.2 Discussion and Conclusion

Throughout this study, we delved into the intricacies of audio processing, model development, and evaluation methodologies. By leveraging advancements in machine learning techniques, we aimed to contribute to the advancement of automated species recognition systems, thereby facilitating more efficient and accurate biodiversity monitoring efforts.

However, our conclusion highlights that Efficient LEAF has surpassed both LEAF and Mel-spectrogram models in performance. Notably, batch sizes play a crucial role, with training Efficient LEAF on a batch size of 64 can improve species classification accuracy. However, it's important to note that larger batch sizes require more GPU memory for model training. Additionally, implementing data augmentation techniques, whether applied directly to raw waveforms or during training, proved instrumental in achieving high prediction scores.

These experiments, conducted over 100 epochs, have demonstrated the potential for achieving superior scores on these models trained specifically for each species. They provide a foundational understanding of utilizing a learnable frontend for feature extraction from raw audio. Furthermore, experimenting with different filters and convolution window strides within the learnable frontend framework can show better in achieving higher scores.

Ultimately, this thesis serves as a testament to our commitment to advancing the field of automated species recognition through innovative research and empirical analysis. By elucidating the nuances of audio classification and model development, we strive to pave the way for more robust and reliable species identification systems, with far-reaching implications for ecological research and conservation efforts.

6.3 Future Recommendations

Testing models over large batch sizes is imperative to understand their scalability and performance under different computational constraints. This aspect becomes particularly crucial when considering the deployment of models in real-world scenarios where efficiency and resource utilization are paramount. While the model sizes across all experiments hover around 220MB, exploring techniques such as pruning and quantization can significantly reduce these sizes. This reduction is essential for supporting deployment on low-memory services such as AWS Lambda and IoT devices, thereby extending the reach of automated species recognition systems to diverse environments and platforms.

Moreover, showcasing the capabilities of these models through web applications can provide researchers with valuable insights into the presence and probabilities of various insect species within audio recordings. Such applications serve as powerful tools for biodiversity monitoring efforts, enabling researchers to quickly and accurately assess the composition of insect populations in different ecosystems. By offering a user-friendly interface and real-time analysis, these applications empower researchers with the necessary information to make informed decisions and drive conservation efforts effectively.

REFERENCES

Anderson, M. and Harte, N., (2022) Learnable Acoustic Frontends in Bird Activity Detection. *International Workshop on Acoustic Signal Enhancement, IWAENC 2022 - Proceedings*.

Anderson, M., Kinnunen, T. and Harte, N., (2023) Learnable Frontends That Do Not Learn: Quantifying Sensitivity To Filterbank Initialisation. pp.1–5.

Bilal, M., Ata-Ur-Rehman and Razzaq, S., (2023) Mosquitoes Species Classification Using Acoustic Features of Wing Beats. *2023 International Conference on Communication Technologies (ComTech)*, [online] pp.28–33. Available at: https://ieeexplore.ieee.org/document/10165480/ [Accessed 21 Aug. 2023].

Branding, J., von Hörsten, D., Wegener, J.K., Böckmann, E. and Hartung, E., (2023) Towards noise robust acoustic insect detection: from the lab to the greenhouse. *KI - Kunstliche Intelligenz*, [online] 1, pp.1–17. Available at: https://link.springer.com/article/10.1007/s13218-023-00812-x [Accessed 10 Dec. 2023].

Chi, Z., Li, Y. and Chen, C., (2019) Deep Convolutional Neural Network Combined with Concatenated Spectrogram for Environmental Sound Classification. *Proceedings of IEEE 7th International Conference on Computer Science and Network Technology, ICCSNT 2019*, pp.251–254.

Chu, H.-C., Zhang, Y.-L. and Chiang, H.-C., (2023) A CNN Sound Classification Mechanism Using Data Augmentation. *Sensors (Basel, Switzerland)*, [online] 2315, p.6972. Available at: /pmc/articles/PMC10422379/ [Accessed 21 Aug. 2023].

Cicada, M., Singh, M., Singh, D., Werner, F., Teng Tey, W., Connie, T., Yeep Choo, K. and Kah Ong Goh, M., (2022) Cicada Species Recognition Based on Acoustic Signals. *Algorithms 2022, Vol. 15, Page 358*, [online] 1510, p.358. Available at: https://www.mdpi.com/1999-4893/15/10/358/htm [Accessed 21 Aug. 2023].

Dolbear, A.E., (1897) The Cricket as a Thermometer. *https://doi.org/10.1086/276739*, [online] 31371, pp.970–971. Available at: https://www.journals.uchicago.edu/doi/10.1086/276739 [Accessed 25 Aug. 2023].

Faiß, M., (2023) InsectSet47 & InsectSet66: Expanded datasets for automatic acoustic identification of insects (Orthoptera and Cicadidae). [online] Available at: https://zenodo.org/record/7828439 [Accessed 29 Aug. 2023].

Faiß, M. and Stowell, D., (2023) Adaptive Representations of Sound for Automatic Insect Recognition. [online] Available at: https://arxiv.org/abs/2304.12739v1 [Accessed 29 Aug. 2023].

Freer, D., Yang, G.-Z., Dai, G., Zhou, J., Huang, J., Wei, S., Zou, S., Liao, F. and Lang, W., (2020) A Comparison on Data Augmentation Methods Based on Deep Learning for Audio Classification. *Journal of Physics: Conference Series*, [online] 14531, p.012085. Available at: https://iopscience.iop.org/article/10.1088/1742-6596/1453/1/012085 [Accessed 28 Dec. 2023].

Gabor, D., (1946) Theory of communication. Part 1: The analysis of information. *Journal of the Institution of Electrical Engineers - Part III: Radio and Communication Engineering*, 9326, pp.429–441.

Härmä, A., (2003) Automatic identification of bird species based on sinusoidal modeling of syllables. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 5, pp.545–548.

He, K., Zhang, X., Ren, S. and Sun, J., (2016) Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-December, pp.770–778.

Heller, K.G., Baker, E., Ingrisch, S., Korsunovskaya, O., Liu, C.X., Riede, K. and Warchalowskaliwa, E., (2021) Bioacoustics and systematics of Mecopoda (and related forms) from South East Asia and adjacent areas (Orthoptera, Tettigonioidea, Mecopodinae) including some chromosome data. *Zootaxa*, 50052, pp.101–144.

Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M. and Adam, H., (2017) MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. [online] Available at: https://arxiv.org/abs/1704.04861v1 [Accessed 31 Aug. 2023].

Huang, G., Liu, Z., Van Der Maaten, L. and Weinberger, K.Q., (2017) Densely connected convolutional networks. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-January, pp.2261–2269.

Koops, H.V., Van Balen, J. and Wiering, F., (2015) Automatic segmentation and deep learning of bird sounds. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, [online] 9283, pp.261–267. Available at: https://link.springer.com/chapter/10.1007/978-3-319-24027-5_26 [Accessed 11 Dec. 2023].

Lemm, D., (2024) *Dom1L/GDSC23: Capgemini Global Data Science Challenge 2023 | Biodiversity Buzz | Solution of the team 'It is a bug'*. [online] Available at: https://github.com/Dom1L/GDSC23 [Accessed 14 Mar. 2024].

Li, Y., Cao, W., Xie, W., Huang, Q., Pang, W. and He, Q., (2022) Low-Complexity Acoustic Scene Classification Using Data Augmentation and Lightweight ResNet. *International Conference on Signal Processing Proceedings, ICSP*, 2022-October, pp.41–45.

Lostanlen, V., Haider, D., Han, H., Lagrange, M., Balazs, P. and Ehler, M., (2023) Fitting Auditory Filterbanks with Multiresolution Neural Networks. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2023-October.

Mesaros, A., Heittola, T. and Virtanen, T., (2016) Metrics for Polyphonic Sound Event Detection. *Applied Sciences 2016, Vol. 6, Page 162*, [online] 66, p.162. Available at: https://www.mdpi.com/2076-3417/6/6/162/htm [Accessed 25 Aug. 2023].

Moulds, M.S., (2009) Cicadas. *Encyclopedia of Insects*, pp.163–164.

Naskrecki, P., (2001) Grasshoppers and their Relatives. *Encyclopedia of Biodiversity*, pp.247–264.

Noda, J.J., Travieso-González, C.M., Sánchez-Rodríguez, D. and Alonso-Hernández, J.B., (2019) Acoustic Classification of Singing Insects Based on MFCC/LFCC Fusion. *Applied Sciences 2019, Vol. 9, Page 4097*, [online] 919, p.4097. Available at: https://www.mdpi.com/2076-3417/9/19/4097/htm [Accessed 21 Aug. 2023].

Park, D.S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E.D. and Le, Q. V., (2019) SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, [online] 2019-September, pp.2613–2617. Available at: http://arxiv.org/abs/1904.08779 [Accessed 31 Aug. 2023].

Ravanelli, M. and Bengio, Y., (2019) Speaker Recognition from Raw Waveform with SincNet. *2018 IEEE Spoken Language Technology Workshop, SLT 2018 - Proceedings*, pp.1021–1028.

Riede, K., (1998) Acoustic monitoring of Orthoptera and its potential for conservation. *Journal of Insect Conservation*, [online] 23–4, pp.217–223. Available at: https://link.springer.com/article/10.1023/A:1009695813606 [Accessed 25 Aug. 2023].

Riede, K., (2018) Acoustic profiling of Orthoptera: present state and future needs. *Journal of Orthoptera Research 27(2): 203-215*, [online] 272, pp.203–215. Available at: https://jor.pensoft.net/article/23700/ [Accessed 25 Aug. 2023].

Robinson, D.J. and Hall, M.J., (2002) Sound signalling in orthoptera. *Advances in Insect Physiology*, 29, pp.151–278.

Sainath, T.N., Weiss, R.J., Senior, A., Wilson, K.W. and Vinyals, O., (2015) Learning the speech front-end with raw waveform CLDNNs. *Interspeech*, 2015-January, pp.1–5.

Schlüter, J. and Gutenbrunner, G., (2022) EfficientLEAF: A Faster LEarnable Audio Frontend of Questionable Use. *European Signal Processing Conference*, 2022-August, pp.205–208.

Seki, H., Yamamoto, K. and Nakagawa, S., (2017) A deep neural network integrated with filterbank learning for speech recognition. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pp.5480–5484.

Siddhartha Varma, A.L.S.V., Bateshwar, V., Rathi, A. and Singh, A., (2021) Acoustic Classification of Insects using Signal Processing and Deep Learning Approaches. *Proceedings of the 8th International Conference on Signal Processing and Integrated Networks, SPIN 2021*, pp.1048–1052.

Stowell, D., (2022) Computational bioacoustics with deep learning: a review and roadmap. *PeerJ*, [online] 10. Available at: /pmc/articles/PMC8944344/ [Accessed 21 Aug. 2023].

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A., (2014) Going Deeper with Convolutions. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, [online] 07-12-June-2015, pp.1–9. Available at: https://arxiv.org/abs/1409.4842v1 [Accessed 31 Aug. 2023].

Tan, M. and Le, Q. V., (2019) EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *36th International Conference on Machine Learning, ICML 2019*, [online] 2019-June, pp.10691–10700. Available at: https://arxiv.org/abs/1905.11946v5 [Accessed 31 Aug. 2023].

Wagner, D.L., (2020) Insect Declines in the Anthropocene. *https://doi.org/10.1146/annurev-ento-011019-025151*, [online] 65, pp.457–480. Available at: https://www.annualreviews.org/doi/abs/10.1146/annurev-ento-011019-025151 [Accessed 6 Dec. 2023].

Wang, Y., Getreuer, P., Hughes, T., Lyon, R.F. and Saurous, R.A., (2017) Trainable frontend for robust and far-field keyword spotting. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pp.5670–5674.

Xiang, H., Liu, S., Zhuang, Z., -, al, Chu, Z., Liu, Z.J., Gu -, H.Y., Cao, X., Wei, Z., Gao, Y. and Huo, Y., (2020) Recognition of Common Insect in Field Based on Deep Learning. *Journal of Physics: Conference Series*, [online] 16341, p.012034. Available at: https://iopscience.iop.org/article/10.1088/1742-6596/1634/1/012034 [Accessed 31 Aug. 2023].

Yadav, S., (2021) *PyTorch implementation of the LEAF audio frontend*. [online] Available at: https://github.com/SarthakYadav/leaf-pytorch [Accessed 18 Mar. 2024].

Yin, M.S., Haddawy, P., Ziemer, T., Wetjen, F., Supratak, A., Chiamsakul, K., Siritanakorn, W., Chantanalertvilai, T., Sriwichai, P. and Sa-ngamuang, C., (2023) A deep learning-based pipeline for mosquito detection and classification from wingbeat sounds. *Multimedia Tools and Applications*, [online] 824, pp.5189–5205. Available at: https://link.springer.com/article/10.1007/s11042-022-13367-0 [Accessed 21 Aug. 2023].

Young, D. and Bennet-Clark, H.C., (1995) The Role of the Tymbal in Cicada Sound Production. *Journal of Experimental Biology*, [online] 1984, pp.1001–1020. Available at: https://dx.doi.org/10.1242/jeb.198.4.1001 [Accessed 30 Aug. 2023].

Zeghidour, N., Teboul, O., de Chaumont Quitry, F. and Tagliasacchi, M., (2021) LEAF: A Learnable Frontend for Audio Classification. *ICLR 2021 - 9th International Conference on Learning Representations*. [online] Available at: https://arxiv.org/abs/2101.08596v1 [Accessed 21 Aug. 2023].

Zhang, C., Zhan, H., Hao, Z. and Gao, X., (2023) Classification of Complicated Urban Forest Acoustic Scenes with Deep Learning Models. *Forests 2023, Vol. 14, Page 206*, [online] 142, p.206. Available at: https://www.mdpi.com/1999-4907/14/2/206/htm [Accessed 21 Aug. 2023].

Zhu, W., Mousavi, S.M. and Beroza, G.C., (2020) Seismic signal augmentation to improve generalization of deep neural networks. *Advances in Geophysics*, 61, pp.151–177.

APPENDIX A: RESEARCH PROPOSAL

SPECIES IDENTIFICATION OF CICADAS AND ORTHOPTERAN THROUGH CNN-BASED SOUND CLASSIFICATION

KARAN SANJAY KAJROLKAR

RESERCH PROPOSAL

AUGUST 2023

Abstract

Insects play vital roles in ecosystems, affecting pollination, pest control, and nutrient cycling. However, the decline of certain insect populations has raised concerns about ecosystem health. By accurately classifying insect species through audio data, we can monitor and assess insect populations non-invasively. In this study we have demonstrated of insect acoustic classification by comparative investigation into the recognition capabilities of diverse deep learning models and proposing deep learning solution to automate the identification of insect species. In signal processing, features are extracted based on Learnable audio frontend LEAF. Data augmentation technique used to handle generalizability of model performance. Used orthopteran and cicada's species dataset published in xeno-canto publicly available dataset of 66 species. All files are mono audio with 44.1 kHz sample rate. However, orthopteran and cicadas know for ability of sound production.

1. Background

Exploring animal vocalizations and the natural soundscape presents a captivating avenue for research, offering valuable insights into animal behaviors, populations, and ecosystems. The realm of acoustic insect recognition introduces a complementary dimension to conventional surveillance methods, such as camera-based approaches. Numerous significant insect groups exhibit distinct species-specific sounds, and the classification of these sounds holds the potential to revolutionize and expedite specimen identification, thereby enhancing the monitoring of biodiversity and species distribution. Given the vast number of insect species – reaching hundreds of thousands, the task of their identification is exceptionally intricate. To put this into context, in the Netherlands alone, 40% of total biodiversity is estimated to consist of insects, whereas mammals and plants are just 5% and 8%, respectively.

Insect population decreases have received considerable attention from both the scientific community and the general public, despite the fact that many of these papers either examine a limited number of species of interest or focus on a specific geographic region (Montgomery et al., 2019). Because of their size, camouflage, activity, or environment, certain species are difficult to view, but their noises can help identify them much more quickly. Additionally, this methodology is less complex and mostly non-invasive than other popular monitoring techniques (Riede, 2018). Due to their small size, ability to blend into their surroundings, and cryptic lives in sometimes inhospitable and challenging locations like tropical rainforests, insects are a particularly difficult group to identify using standard monitoring methods. (Riede, 2018).

Even though there are many different kinds of insects, for this study we considered classifying two types, orthopteran and cicada. Cicadas are best known for their ability to produce loud sound (Moulds, 2009). The majority of the 3200 species in the Cicadidae family use speedily deforming tymbal membranes to make loud clicking noises that cause the tymbals to resonate (Young and Bennet-Clark, 1995). The sounds of orthopterans are often species-specific and serve a crucial role in the identification of species (Heller et al., 2021). However, the orthopteran order of insects includes the grasshoppers, crickets, and their relatives. The process by which orthopteran make sound, termed as stridulation, involves rubbing one modified part of the body against another (Naskrecki, 2001). A call's information may be stored by frequency modulation, where the pitch varies over time like

it does in birds, time modulation where the pitch remains constant over the call's duration but the temporal pattern is unique to the species, or a mixture of both types. (Naskrecki, 2001).

Historically, when trying to identify Orthoptera based on their sounds, research has predominantly depended on manually extracting sound parameters such as carrier frequency and pulse rates. (Riede, 1998). Before using these characteristics for automatic classification, their parameters must be manually chosen and made (Faiß and Stowell, 2023). Because Orthoptera are poikilothermic species, the ambient temperature during the recording may have an impact on how frequently they sing. As a result, these parameters might not always work as intended (Dolbear, 1897).

2. Problem Statement OR Related Research OR Related Work

Deep learning has been investigated by the researcher as a flexible strategy for the problem of classifying and on input data achieving high-precision identification of acoustic signals with minimal manual pre-processing. (Stowell, 2022; Faiß and Stowell, 2023). However, there are also challenges for acoustic monitoring such as background heavy rain, wind, loudness, distance of the sound source from the recorder etc.

Previous studies have combined audio features to increase performance by their features, such as the log mel and log gammatone (Chi et al., 2019) and the MFCC and LFCC (Noda et al., 2019). The researchers discovered that fusion features outperform cutting-edge methods.

Long-form field recordings can be classified when used in conjunction with sound event detection, eliminating the need for any manual feature extraction or relevant clip identification (Faiß and Stowell, 2023). Where (Noda et al., 2019) used (Härmä, 2003) procedure for segmentation of multiple calls of one insect species in single recording as result gives high performance except some insect not classified correctly due to bad segmentation of the calls due some unnecessary information low-amplitude noise gets captured because as formulation algorithm captures frequency if it is greater than threshold (max frequency - MINDB). As improvement on bad segmentation problem (Cicada et al., 2022) introduce improved Harma. However, (Cicada et al., 2022) removed low-amplitude noise helps the model obtain the highest classification accuracy on dataset of species katydids, crickets and cicadas.

In the past, mel-filterbanks, an unchanging, manually-engineered representation of sound, were employed for audio classification tasks since strong features can result in high performance models (Zeghidour et al., 2021). Mel-filterbanks create a spectrogram first using the STFT's squared modulus. Then, to mimic the non-linear human perception of pitch, the spectrogram is run through a bank of triangular bandpass filters that are spaced on a logarithmic scale in mel-scale (Zeghidour et al., 2021). Finally, a logarithm compression is applied to the coefficients to mimic our non-linear sensitivity to loudness (Zeghidour et al., 2021).

As in recent study by (Faiß and Stowell, 2023) mentioned Insect sounds are produced utilizing stridulatory or tymbal mechanisms that produce a different structure of frequencies and overtones from source-filter systems used by mammals or birds. Many species of insects produce ultrasonic sounds, which are often significantly higher in frequency than the majority of mammal or bird sounds. Depending on how they go about it, the emphasis on high-frequency sounds that are occasionally completely and far outside of the human hearing range (between 20 Hz and 20 kHz) might affect how well audio categorization networks operate. The mel filter bank technique based on human perception is probably not optimal to identify and differentiate between slight variations in high frequencies for many insect noises, even though it works well enough for other sounds like bird song.

In recent work of deep learning audio classification introduce adaptive waveform-based methods such as LEAF (Zeghidour et al., 2021) represent a new generation of frontends implemented as layers in a neural network, which can be optimized alongside the model to better fit insect.

Additionally, most studies have compared the effectiveness of handcrafted and deep learning features (Siddhartha Varma et al., 2021; Anderson and Harte, 2022; Faiß and Stowell, 2023). The utilization of a SincNet based deep learning feature layer combined with a CNN layer yielded the highest accuracy, achieving 97% on insect database of Cornell University (Siddhartha Varma et al., 2021). Where LEAF 83.7% performs worse than PCEN 89.9% (Wang et al., 2017) and TD 87.6% on BirdVox-DCASE-20k dataset combined with supplemented with 10s length, recorded at 44.1 KHz, and normalized to -2dBFS with EfficentNet-B0 model, it has been determined that learnable filterbanks can increase performance (Anderson and Harte, 2022).

In their study, (Faiß and Stowell, 2023) contrasted conventional spectrogram-based audio representation with a novel neural network approach known as LEAF (Zeghidour et al., 2021) for analyzing insect sounds, particularly those beyond human hearing range. Using data from 66 insect species, they compared mel-spectrogram and LEAF feature extraction methods and found that combining LEAF with an CNN led to an impressive 83% accuracy in classifying insect species. Notably, they revealed that individual LEAF component learnable filterbank shows higher contribution in overall LEAF performance. The study also suggested the potential of enhancing CNN performance through techniques like sound-event detection, thereby indicating avenues for further improvement in achieving higher classification accuracy.

Furthermore, mosquito classification in these studies have harnessed raw audio from mosquito wingbeats. (Yin et al., 2023) applied a 1D-CNN + LSTM approach achieving a 93% accuracy in classifying species and sex based on wingbeats, exploring microphone factors and comparing Bayesian CNN techniques. (Bilal et al., 2023) employing variables including mel spectrogram, amplitude envelope, zero crossings, MFCCs, and spectral centroid, audio characteristics were manually extracted for CNN modeling. Considering the difficulty of class imbalance, their suggested model, a modified ResNet-50, achieved acceptable results with ROC-AUC, Precision Recall AUC with 0.921 and 0.885.

(Li et al., 2022) introduced a technique to classify diverse acoustic scenes by combining data augmentation methods and a lightweight ResNet model. Their approach included random time stretching and shifting for augmented training data, enhancing model robustness. They minimized complexity by using a compact ResNet variation and optimized it for low power devices with real time applications. The results of the experiments showed that the proposed method outperformed existing proposed techniques for acoustic scene classification on the bases of accuracy and efficiency.

(Zhang et al., 2023) conducted a comparative investigation into the recognition capabilities of diverse deep learning models for classifying biological acoustic scenes. They collected a range of audio categories, including human, insect (Cicadas), bird, silence using Song Meter SM4 acoustic recorders in an urban forest. Resampling audio to 22,050 Hz, 16-bit sampling, and durations of 3-5 seconds were standardized across all 1000 samples. Employing hand-crafted feature extraction through mel spectrograms, the authors augmented data using techniques like noise addition, amplitude change, time shifting,

frequency masking, and time masking with librosa. Several deep learning models were experimented including EfficentNet, two depth model for ResNet, DenseNet with bottleneck compression, edge device models like MobileNet, Notably, the DenseNet_BC_34 model exhibited superior generalizability, yielding 93.81% score on overall validation dataset. This disparity possibly arises by test dataset encompassing novel sound patterns not fully learned during training, leading to occasional misclassifications, particularly among complex sound patterns involving humans, birds, and insects.

In their work (Chu et al., 2023), a sound classification mechanism was proposed, which utilized a refined approach to hand-crafted MFCC features. The study addressed the limitation of classification efficiency due to data scarcity by focusing on Mel filters, designed to mirror the human auditory system's frequency response. This involves segmenting speech signals into distinct frequency bands using these filters, quantifying each band's strength logarithmically for use as acoustic features in classification. The identical signal produced somewhat different logarithmic energy representations depending on the different sizes of triangle bandpass filters used. The authors developed a method for data augmentation on the ESC-50 that outperformed the original dataset score of 63% by 97% using a particular set of 5 triangular bandpass filters. Classification accuracy in the UrbanSound8K dataset, where data volume is more balanced, attained 90%, marginally increasing to 92% following data augmentation. This augmentation proved instrumental in mitigating data imbalance and labeling challenges. Notably, the models were developed without accounting for real-world factors such as environmental noise, interference and sound overlap, which significantly impact model accuracy during practical sound reception.

The researcher (Anderson et al., 2023) has discussed about how initialization of filterbank in learnable frontend LEAF can increase performance on 2 different frequency distribution domain like human and bird. However, using liner instead of Mel used by LEAF perform better than another initialization. These changes can add valuable study in deep learning acoustic classification.

To explain key ideas and fill in knowledge gaps, a recent review by (Stowell, 2022), put a special emphasis on deep learning in computational bioacoustics. The review outlined a standard approach involving CNN architectures (such as ResNet, VGGish, MobileNet) pretrained on AudioSet, with spectrograms (linear, mel, log-frequency) as typical input

data. Alternatives like wavelet representations, raw waveforms (e.g., Wavenet), or the LEAF frontend were also noted. Effective data augmentation techniques like noise mixing, time shifting, and mixing were highlighted, along with methods such as time warping and frequency adjustments to enhance dataset diversity. Given the often-imbalanced nature of bioacoustics datasets, Stowell suggested employing macro-averaging for evaluation, providing equal weight to each class by calculating and averaging class-specific performance (Mesaros et al., 2016).

Although the study found many researchers has focuses on how different feature extraction technique either hand crafted or learnable with standard CNN architecture can improve model performance. As result they found learnable frontend outperform hand-crafted fronted (Siddhartha Varma et al., 2021; Anderson and Harte, 2022; Faiß and Stowell, 2023; Yin et al., 2023). It is necessary to further enhance findings when they propose classification, and it is possible to do so by comparing the model performances to determine which one is the most appropriate.

3. Research Questions

The following research questions are suggested for each of the research objective as highlighted as follows.

1. Can deep learning techniques be employed to assess the capacity of various models in recognizing acoustic scenes associated with insects?
2. Can data augmentation technique help to handle data imbalance?

4. Aim and Objectives

The main aim of this research is to develop and evaluate an effective insect audio classification using CNN. Accurate classification model designed to identify distinct sounds, facilitating the monitoring and identification of groups of insect species in remote habitats.

The research objectives are formulated based on the aim of this study which are as follows:

- To analyze the data augmentation and pre-process technique helps to handle data imbalance e.g., mixing, one of (Gaussian Noise, Pink Noise, Noise Injection) technique while using for models.
- To investigate the potential of transfer learning by fine-tuning a pre-trained CNN (e.g., from image classification tasks) on the insect audio dataset. Assess how this strategy improves convergence speed and classification accuracy.
- To analyze the learned features with the help of LEAF Frontend within the pre-trained CNN to gain insights into what audio characteristics contribute to accurate classification.

5. Significance of the Study

For many species or texa vocalization are explored using Deep Learning Technique e.g. bird, bat, insect etc. However, they found deep learning can replace manual process to detect and analyze by using either images or audio signal. Image-based classification of insect species can be difficult, but it can supplement audio-based techniques and provide a visual perspective on the biodiversity and behavior of insects. However, difficulties are visual appearance, background noise, Differentiation in Appearance, Environmental Factor such are whether condition or lighting, Incorrect labeling due similar appearance insect.

This study designed to help identify acoustic insect through vocalization using deep learning. Author (Faiß and Stowell, 2023) proposed a methodology for comparison between mel spectrogram and learnable frontend for feature extraction with own CNN model. However, they have not achieved promising performance. So instead of training model from scratch in this research we are applying some CNN models which pre-trained on large image database ImageNet which can help in reducing model size as well as performance.

This study specifically addresses the augmentation and pre-processing technique used for model generalization because dataset used in this study is imbalance. If this study is achieved good performance, then we can use this for filed study. Insects are fundamental components of ecosystems worldwide. This research carries global significance by addressing a universal challenge in biodiversity monitoring, with potential applications in ecosystems around the world.

By successfully applying deep learning to acoustic insect identification, this research opens doors to cross-species insights. Lessons learned from this study can potentially be transferred to similar endeavors in the identification of other species.

6. Scope of the Study

- The main purpose of the study to demonstrated comparative investigation into the recognition capabilities of diverse deep learning models and proposing deep learning solution to automate the identification of species of orthopteran and cicadas.
- This study aims to focus on characteristic of learnable frontend feature for insect audio. This study covers various combination of data augmentation technique on raw waveform and spectrogram with different CNN model to increase generalization of model performance.
- In this study we have worked on mono waveform type audio files not stereo. However we are using chunk of 1-10s to handle dataset imbalance (Stowell, 2022; Faiß and Stowell, 2023). Furthermore, used PyTorch library includes pre-trained models for Acoustic insect classification. Dataset is imbalance due to some species have longer audio recording than other.
- This study will be helpful to tracking population trends, assessing biodiversity, and identifying potential conservation threats, environmental monitoring, Agriculture and Pest Control.
- For future studies need to focus on how insect syllable can extract from long audio know as Sound Event Detection. In above literature review (Anderson et al., 2023) has found using liner initialization instead mel in LEAF can increase performance. However, need to address potential contributions of RNN (Recurrent Neural Network) and Hybrid- CNN models. By incorporating these models, we seek to gain valuable insights into their effectiveness in addressing the challenges posed by insect species identification through vocalization (Stowell, 2022).

7. Research Methodology

7.1 Introduction

Insects may be small, but they play a crucial role in ecosystems worldwide, providing irreplaceable benefits such as balancing ecosystems, supporting other animals, enabling plant pollination and reproduction, aerating soil, and much more. Losing even a few species could be catastrophic to biodiversity. Bioacoustics monitoring and classification of animal communication signals has developed into a powerful tool for measuring and monitoring species diversity within complex communities and habitats (Riede, 2018). The development of accessible digital sound recording technology and significant advancements in informatics, including big data, signal processing, and machine learning, have advanced the field of bioacoustics over the past few decades (Stowell, 2022).

Numerous studies are done as far on acoustic sound classification using deep learning. However, orthopteran and cicada know for ability of sound producing insect where some species of orthopteran produce ultrasonic sound. Deep learning algorithms are more frequently used in acoustic scene ecology for species-specific identification and target sound recognition. Many studies done so far for detection, classification of insect (Noda et al., 2019; Cicada et al., 2022; Bilal et al., 2023; Faiß and Stowell, 2023; Yin et al., 2023; Zhang et al., 2023). They used either fixed feature frontend (e.g., mel-spectrogram), learnable feature frontend (e.g. SincNet, TD-filterbank, LEAF) or compare both to investigate performance.

Where (Faiß and Stowell, 2023) proposed robust deep learning model with comparison of leaf and mel-spectrogram on various set of datasets of species orthopteran and cicada. They used 66 orthopteran and cicada species to experiment as learnable frontend outperforms fixed frontend. Given this foundation, we employed various deep learning models to train on these acoustic scene samples and subsequently evaluated and compared their classification performance.

7.2 Dataset Description

In this investigation, two types of auditory species were employed, including orthopterans and cicada. From Baudewijn Odé's private collections, recordings from BioAcoustica, xeno-canto, and iNaturalist were gathered recordings used in this study are publicly available. Files containing long periods without insect sounds were edited into multiple smaller files with silent periods no longer than 5 seconds. The files were standardized to 44.1 kHz mono WAV where minimum of ten files per species. Dataset is highly imbalance need augmentation technique to handle them.

Species	n	h:min:s	Species	n	h:min:s	Species	n	h:min:s
Yoyetta celis	152	0:11:16	Aleeta curvicosta	23	0:04:04	Gomphocerus sibiricus	14	0:26:05
Gryllus campestris	57	1:37:39	Platypleura cfcatenata	22	0:17:47	Barbitistes yersini	14	0:19:59
Chorthippus biguttulus	53	0:30:25	Omocestus rufipes	22	0:16:34	Psaltoda plaga	14	0:04:21
Galanga labeculata	43	0:06:16	Chorthippus apricarius	21	0:28:35	Popplepsalta notialis	14	0:02:58
Yoyetta repetens	40	0:05:23	Myrmeleotettix macula-tus	21	1:05:37	Pholidoptera littoralis	13	0:04:00
Chorthippus mollis	39	0:27:50	Cicada orni	21	0:06:50	Pseudochorthippus mon-tanus	13	0:11:36
Stenobothrus stigmati-cus	39	0:05:31	Phaneroptera falcata	20	0:28:30	Leptophyes punctatis-sima	13	0:26:48
Pseudochorthippus parallelus	37	0:25:08	Gryllus bimaculatus	20	0:28:44	Cyclochila australasiae	13	0:01:53
Roeseliana roeselii	37	0:34:34	Platypleura plumosa	19	0:14:42	Platypleura sp13	12	0:07:01
Tettigonia cantans	37	0:58:10	Stenobothrus lineatus	19	0:34:27	Chorthippus albomar-ginatus	11	0:40:29
Conocephalus fuscus	36	0:53:34	Clinopsalta autumna	19	0:04:16	Eupholidoptera schmidti	11	0:09:40
Chorthippus brunneus	35	0:21:57	Phaneroptera nana	18	0:30:50	Melanogryllus desertus	11	0:25:24
Decticus verrucivorus	34	1:15:04	Conocephalus dorsalis	18	0:23:07	Tylopsis lilifolia	11	0:03:30
Tettigonia viridissima	33	0:27:26	Platypleura sp10	17	0:17:55	Ruspolia nitidula	11	0:12:35
Ephippiger diurnus	31	0:39:51	Chrysochraon dispar	17	0:15:36	Diceroprocta eugraphica	11	0:05:07
Nemobius sylvestris	30	0:38:44	Pholidoptera aptera	16	0:10:55	Platypleura sp12cfhirtipennis	10	0:07:42
Oecanthus pellucens	29	0:30:32	Eumodicogryllus bor-digalensis	16	0:10:56	Omocestus petraeus	10	0:09:22
Gomphocerippus rufus	28	0:29:38	Platycleis albopunctata	15	0:24:45	Stauroderus scalaris	10	0:20:43
Pholidoptera griseoap-tera	27	0:14:07	Atrapsalta corticina	15	0:02:15	Chorthippus vagans	10	0:11:43
Omocestus viridulus	27	0:45:48	Neotibicen pruinosus	15	0:04:41	Bicolorana bicolor	10	0:09:19
Gampsocleis glabra	27	0:55:18	Atrapsalta encaustica	15	0:04:33	Popplepsalta aeroides	10	0:01:46
Acheta domesticus	24	0:56:48	Metrioptera brachyptera	14	0:20:56	Atrapsalta collina	10	0:01:20

Table 1: A total of 1,554 audio files representing 66 different species were curated from five distinct source datasets. These files collectively encompassed a recording duration of 24 hours and 32 minutes. The distribution includes the number of files per species (n) and the total recording time per species (h: min: s)

7.3 Data Pre-processing

In deep learning, "generalization" signifies a neural network's ability to perform effectively on previously unseen data, as highlighted by (Zhu et al., 2020). Data augmentation aims to enhance the training dataset's complexity and size, expanding the feature space for improved decision boundaries, thereby reducing false positives and false negatives and ultimately enhancing generalization on unfamiliar samples (Zhu et al., 2020).

7.3.1 Removing low pass and high pass noise

For denoising audio signal by using Butterworth, However, low pass filter applies to cut-off points 10 kHz and high pass filter applies to cut-off point 1 kHz. After denoising, the noises with frequencies above 10 kHz and below 1 kHz were mostly removed from the audio recording.

7.3.2 Data Augmentation

Furthermore, augmentations that are applied on waveforms are:

Noise Injection:

Audio augmentation employs noise injection by introducing controlled background sounds or distortions to training audio data. This enhances the model's adaptability to real-world audio variations, improving its overall performance and robustness.

Gaussian Noise:

Injects Gaussian noise using a randomly chosen signal-to-noise ratio.

Pink Noise:

Pink noise in audio augmentation is a balanced sound that mimics natural environments. It helps test and improve audio models by adding varied frequencies for a more realistic experience.

Impulse Response (IRs):

Another method for dealing with complex noise effects is to include false negative noises (non- insect audios). However, by incorporating a small number of comparable types of noise samples into the training data, it is possible to successfully reduce this kind of false-positive prediction.

Random shift:

The model can exhibit positional bias when trained with data anchored to a fixed reference time or with a narrow time shift around that reference point. In such cases, instead of learning broader patterns from the feature space, the neural network tends to memorize the specific anchor time, potentially limiting its ability to detect insect signals beyond the training shift window (Zhu et al., 2020).

For applying transformations directly on the spectrogram of the audio signal, such as time or frequency masking by using SpecAugment (Park et al., 2019). After Augmentation require normalization techniques include scaling the values between 0 and 1 or standardizing them with zero mean and unit variance. However, normalizing these values can improve convergence during training and prevent any feature from dominating the learning process.

7.4 Feature Extraction

We are using LEAF "LEarnable Audio Frontend" (Zeghidour et al., 2021) as feature extraction technique. LEAF is a special frontend that can be learned completely in all of its functions and is managed by just a few hundred parameters. These frontends are implemented as neural network, which can be optimized alongside of the model (Anderson and Harte, 2022).

In LEAF frontend have learnable parameter such as one-dimensional Gabor filters, low-pass smoothing and PCAN for normalization and compression with fixed squared modulus whose help in calculating time sequence to avoid frequency response does not have any negative frequency.

At the beginning of training, the Gabor filters are initialized to the mel scale, and the system makes an effort to learn the frequency bands of interest. (Eq. 1). Resulting time sequence is square-modulated to prevent negative frequency values.

The low pass filter used by LEAF has a Gaussian impulse response (Eq. 2).

$$\phi_n(t) = e^{2\pi j \eta_n t} \frac{1}{\sqrt{2\pi}\sigma_{n_{bw}}} e^{-\frac{t^2}{2\sigma^2_{n_{bw}}}} \quad (1)$$

$$\Phi_n(t) = \frac{1}{\sqrt{2\pi}\sigma_{n_{lp}}} e^{-\frac{t^2 \cdot n_{lp}}{2\sigma_n^2}} \quad (2)$$

The last stage of PCAN uses a learnable combination of AGC and DRC to handle normalization and noise, with AGC calculated using a learnable smoother. (Eq .3). However, the AGC is used before the DRC and produces output by (Eq .4). Where PCAN generate spectrogram-like output we can use input for our models.

$$M(t,f) = (1-s)M(t-1,f) + sE(t,f) \quad (3)$$

$$PCEN(\mathcal{F}(t,n)) = \left(\frac{\mathcal{F}(t,n)}{(\mathcal{E}+M(t,n))^{\alpha_n}} + \delta_n\right)^{r_n} - \delta_n^{r_n} \quad (4)$$

7.5 Models

Deep learning is a subset of machine learning that utilizes multiple hidden nodes and nonlinear transformations to abstractly represent intricate data (Zhang et al., 2023). Convolutional neural networks (CNN) have gained popularity in image recognition, speech understanding, and various domains due to their ability to extract interconnected features from input data utilizing techniques like to those observed in the human brain.(Zhang et al., 2023). For identification of insect, we are using convolutional models with head. However, for convolutional models we adapt transfer learning technique. A head is a single linear layer with exactly as many outputs as classes.

Transfer Learning Technique:

A model learned on one task or dataset is adjusted or improved for another, often related, task or dataset using the machine learning process known as transfer learning. Utilizing prior knowledge, it enhances performance on the new work while conserving time and resources. VGG, ResNet, DenceNet, Inception, MobileNet, and EfficentNet are all CNN models used for the sound detection job. These models are all representative and have been used extensively in sound identification (Xiang et al., 2020; Zeghidour et al., 2021; Anderson and Harte, 2022; Anderson et al., 2023; Zhang et al., 2023).

However, ResNet solves the gradient disappearance issue brought on by increasing model depth by providing Skip Connection, allowing for the eventual building of models with

more layers (He et al., 2016). ResNet can be used to process audio spectrograms for sound classification, and the residual connections support the maintenance and learning of key audio features over deep layers, resulting in enhanced performance (He et al., 2016). Similarly, an investigation of how to deepen such networks led to the development of the VGG, a standard convolutional neural network design of small 3 x 3 filters. Compared with ResNet, all layers are connected by DenseNet's (Huang et al., 2017) more aggressive dense connection algorithm, and each layer takes input from all levels that came before it. In order to reduce the model parameters as much as possible, the researchers added Bottleneck layers with Compression procedures to the model- building process to produce the DenseNet-BC model.

The Inception architecture proposed by (Szegedy et al., 2014), also known as GoogLeNet, employs multiple filters of different sizes and concatenates their outputs. This helps capture both local and global features effectively. Similarly, in sound classification, you can design filters to capture different frequency ranges or temporal patterns in audio signals, allowing Inception-like models to capture diverse sound characteristics. MobileNet is designed for efficiency and is well-suited for applications where computational resources are limited. In sound classification, MobileNet can be used to create lightweight models that can run on devices with lower processing power (Howard et al., 2017). However, audio data into spectrograms and leverage MobileNet's depth wise separable convolutions to efficiently process them.

The effective design suggested by (Tan and Le, 2019). They discovered that depth, width, and resolution are the three key dimensions that have an impact on neural network accuracy. They used the neural architecture search (NAS) (Tan and Le, 2019) technique to obtain a model backbone called EfficientNet_b0, and then scaled the above three dimensions based on this backbone to obtain the models b1 to b7 with an amazing performance. EfficientNet-B0 is also used extensively to evaluate the original implementation of LEAF (Zeghidour et al., 2021).

Furthermore, the VGG, Inception, ResNet, DenseNet, MobileNet and EfficientNet models were all initially pre-trained on the ImageNet dataset, which is a large dataset containing millions of labeled images from various categories. Pre-training on ImageNet allowed these models to learn general features from images, such as edges, textures, shapes, and higher-level object representations.

7.6 Evaluation Matrix

7.6.1 Accuracy

Accuracy measures the proportion of correctly classified insect sounds out of the total predictions. It provides an overall view of your model's performance

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

7.6.2 Precision & recall:

Both metrics help you understand the model's ability to correctly classify insect sounds and avoid false positives. Precision is interpreted as the ratio of instances for accurately predicted positives to all instances of expected positives. The ratio of accurately predicted positive cases to actual positive instances is known as recall.

$$P = \frac{TP}{TP + FP} \qquad R = \frac{TP}{TP + FN}$$

7.6.3 Confusion Matrix

A confusion matrix provides a detailed breakdown of true positive, true negative, false positive, and false negative predictions. It helps you identify which classes are being misclassified and in what way.

7.6.4 F1-score

The F1-score is the harmonic mean of precision and recall. It balances the trade-off between precision and recall and is particularly useful when classes are imbalanced.

$$F1 = \frac{2.TP}{2.TP + FP + FN}$$

7.6.5 macro-averaging

In bioacoustics datasets, addressing class imbalance is often achieved by employing macro-averaging, where individual performance metrics are computed for each class independently, and then their averages are taken to ensure equal consideration for each class. (Stowell, 2022).

8. Requirements Resources

Python will be used for this study implementation.

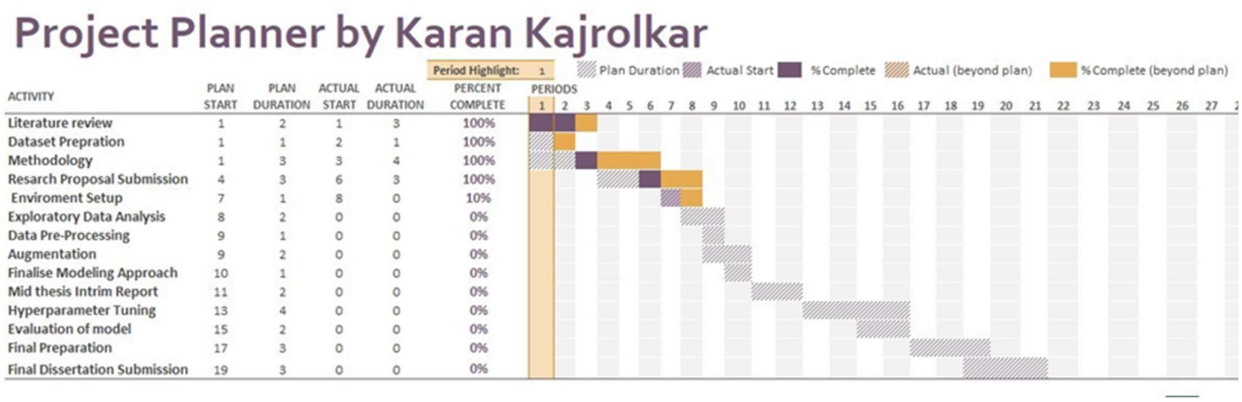
Software Requirement

- PyTorch libraries for data augmentation and implementing pre-trained CNN models.
- Using Google Colab and Jupyter Notebook Environment.

Hardware Requirement:

- Windows OS with Intel(R) Core(TM) i5-8365U CPU @ 1.60GHz, 1.90 GHz
- Required GPU for modelling LEAF frontend with CNN.

9. Research Plan



ACTIVITY	PLAN START	PLAN DURATION	ACTUAL START	ACTUAL DURATION	PERCENT COMPLETE
Literature review	1	2	1	3	100%
Dataset Prepration	1	1	2	1	100%
Methodology	1	3	3	4	100%
Resarch Proposal Submission	4	3	6	3	100%
Enviroment Setup	7	1	8	0	10%
Exploratory Data Analysis	8	2	0	0	0%
Data Pre-Processing	9	1	0	0	0%
Augmentation	9	2	0	0	0%
Finalise Modeling Approach	10	1	0	0	0%
Mid thesis Intrim Report	11	2	0	0	0%
Hyperparameter Tuning	13	4	0	0	0%
Evaluation of model	15	2	0	0	0%
Final Preparation	17	3	0	0	0%
Final Dissertation Submission	19	3	0	0	0%

References:

Anderson, M. and Harte, N., (2022) Learnable Acoustic Frontends in Bird Activity Detection. International Workshop on Acoustic Signal Enhancement, IWAENC 2022 - Proceedings.

Anderson, M., Kinnunen, T. and Harte, N., (2023) Learnable Frontends That Do Not Learn: Quantifying Sensitivity To Filterbank Initialisation. pp.1–5.

Bilal, M., Ata-Ur-Rehman and Razzaq, S., (2023) Mosquitoes Species Classification Using Acoustic Features of Wing Beats. 2023 International Conference on Communication Technologies (ComTech), [online] pp.28–33. Available at: https://ieeexplore.ieee.org/document/10165480/ [Accessed 21 Aug. 2023].

Cicada, M., Singh, M., Singh, D., Werner, F., Teng Tey, W., Connie, T., Yeep Choo, K. and Kah Ong Goh, M., (2022) Cicada Species Recognition Based on Acoustic Signals. Algorithms 2022, Vol. 15, Page 358, [online] 1510, p.358. Available at: https://www.mdpi.com/1999-4893/15/10/358/htm [Accessed 21 Aug. 2023].

Faiß, M., (2023) InsectSet47 & InsectSet66: Expanded datasets for automatic acoustic identification of insects (Orthoptera and Cicadidae). [online] Available at: https://zenodo.org/record/7828439 [Accessed 29 Aug. 2023].

Faiß, M. and Stowell, D., (2023) Adaptive Representations of Sound for Automatic Insect Recognition. [Online] Available at: https://arxiv.org/abs/2304.12739v1 [Accessed 29 Aug. 2023].

He, K., Zhang, X., Ren, S. and Sun, J., (2016) Deep residual learning for image recognition. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016-December, pp.770–778.

Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M. and Adam, H., (2017) MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. [Online] Available at: https://arxiv.org/abs/1704.04861v1 [Accessed 31 Aug. 2023].

Huang, G., Liu, Z., Van Der Maaten, L. and Weinberger, K.Q., (2017) Densely connected convolutional networks. Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, 2017-January, pp.2261–2269.

Moulds, M.S., (2009) Cicadas. Encyclopedia of Insects, pp.163–164.

Naskrecki, P., (2001) Grasshoppers and their Relatives. Encyclopedia of Biodiversity, pp.247–264.

Noda, J.J., Travieso-González, C.M., Sánchez-Rodríguez, D. and Alonso-Hernández, J.B., (2019) Acoustic Classification of Singing Insects Based on MFCC/LFCC Fusion. Applied Sciences 2019, Vol. 9, Page 4097, [online] 919, p.4097. Available at:

https://www.mdpi.com/2076-3417/9/19/4097/htm [Accessed 21 Aug. 2023].

Park, D.S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E.D. and Le, Q. V., (2019) SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, [online] 2019-September, pp.2613–2617. Available at: http://arxiv.org/abs/1904.08779 [Accessed 31 Aug. 2023].

Riede, K., (2018) Acoustic profiling of Orthoptera: present state and future needs. Journal of Orthoptera Research 27(2): 203-215, [online] 272, pp.203–215. Available at:https://jor.pensoft.net/article/23700/ [Accessed 25 Aug. 2023].

Siddhartha Varma, A.L.S.V., Bateshwar, V., Rathi, A. and Singh, A., (2021) Acoustic Classification of Insects using Signal Processing and Deep Learning Approaches. Proceedings of the 8th International Conference on Signal Processing and Integrated Networks, SPIN 2021, pp.1048–1052.

Stowell, D., (2022) Computational bioacoustics with deep learning: a review and roadmap. PeerJ, [online] 10. Available at: /pmc/articles/PMC8944344/ [Accessed 21 Aug. 2023].

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A., (2014) Going Deeper with Convolutions. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, [online] 07-12-June-2015, pp.1–9. Available at: https://arxiv.org/abs/1409.4842v1 [Accessed 31 Aug. 2023].

Tan, M., Chen, B., Pang, R., Vasudevan, V., Sandler, M., Howard, A. and Le, Q. V., (2018) MnasNet: Platform-Aware Neural Architecture Search for Mobile. Proceedings of the IEEE

Computer Society Conference on Computer Vision and Pattern Recognition, [online] 2019-June, pp.2815–2823. Available at: https://arxiv.org/abs/1807.11626v3 [Accessed 31 Aug. 2023].

Tan, M. and Le, Q. V., (2019) EfficientNet: Rethinking Model Scaling for Convolutional

Neural Networks. 36th International Conference on Machine Learning, ICML 2019, [online] 2019-June, pp.10691–10700. Available at: https://arxiv.org/abs/1905.11946v5 [Accessed 31 Aug. 2023].

Xiang, H., Liu, S., Zhuang, Z., -, al, Chu, Z., Liu, Z.J., Gu -, H.Y., Cao, X., Wei, Z., Gao, Y. and Huo, Y., (2020) Recognition of Common Insect in Field Based on Deep Learning. Journal of Physics: Conference Series, [online] 16341, p.012034. Available at: https://iopscience.iop.org/article/10.1088/1742-6596/1634/1/012034 [Accessed 31 Aug. 2023].

Yin, M.S., Haddawy, P., Ziemer, T., Wetjen, F., Supratak, A., Chiamsakul, K., Siritanakorn, W., Chantanalertvilai, T., Sriwichai, P. and Sa-ngamuang, C., (2023) A deep learning-based pipeline for mosquito detection and classification from wingbeat sounds. Multimedia Tools and Applications, [online] 824, pp.5189–5205. Available at: https://link.springer.com/article/10.1007/s11042-022-13367-0 [Accessed 21 Aug. 2023].

Young, D. and Bennet-Clark, H.C., (1995) The Role of the Tymbal in Cicada Sound Production. Journal of Experimental Biology, [online] 1984, pp.1001–1020. Available at: https://dx.doi.org/10.1242/jeb.198.4.1001 [Accessed 30 Aug. 2023].

Zeghidour, N., Teboul, O., de Chaumont Quitry, F. and Tagliasacchi, M., (2021) LEAF: A Learnable Frontend for Audio Classification. ICLR 2021 - 9th International Conference on Learning Representations. [online] Available at: https://arxiv.org/abs/2101.08596v1 [Accessed 21 Aug. 2023].

Zhang, C., Zhan, H., Hao, Z. and Gao, X., (2023) Classification of Complicated Urban Forest Acoustic Scenes with Deep Learning Models. Forests 2023, Vol. 14, Page 206, [online] 142,p.206. Available at: https://www.mdpi.com/1999-4907/14/2/206/htm [Accessed 21 Aug. 2023].

Zhu, W., Mousavi, S.M. and Beroza, G.C., (2020) Seismic signal augmentation to improve generalization of deep neural networks. Advances in Geophysics, 61, pp.151–177.

List of Abbreviations:

1. LEAF: Learnable Audio Frontend
2. STFT: Short-Term Fourier Transform
3. MFCC: Mel Frequency Cepstral Coefficients
4. LFCC: Linear Frequency Cepstral Coefficients
5. PCEN: Per Channel Energy Normalization
6. CNN: Convolution Neural Network
7. LSTM: Long Short Term memory
8. ROC-AUC: Receiver Operating Characteristic Area Under the Curve
9. TD Filterbank: Time Domain Filterbank
10. AGC: Automatic Gain Control
11. DRC: Dynamic Range Compression
12. VGG: Visual Geometry Group