BIOACOUSTIC INSECT SPECIES CLASSIFICATION USING A SUB-SPECTROGRAM ARCHITECTURE

J.D. MULCKHUIJSE

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

2108203

COMMITTEE

dr. Dan Stowell
dr. Dimitar Shterionov

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

June 24th, 2024

WORD COUNT

7219

ACKNOWLEDGMENTS

I would like to thank my supervisor dr. Dan Stowell for inspiring me to pursue this topic and his guidance throughout the course of the thesis. Furthermore, I would like to thank my girlfriend, family, and family in law for their unconditional support on which I can always count.

CONTENTS

Abstract

This study researches the improvement of bioacoustic insect classification, focusing on Orthoptera and Cicadidae, through the use of sub-spectrogram CNN architectures. While in the process, evaluating both mel and linear spectrograms to determine their efficacy in representing high-frequency insect sounds. This research was motivated by the global decline in insect populations, which threatens ecosystem functions that are essential for human health and survival. Two state-of-the-art models using the InsectSet66 dataset were adapted to incorporate a sub-spectrogram approach. Results have shown that use of linear instead of mel spectrograms did not substantially change performance, indicating that the mel scale compression of higher frequencies is not a constraining factor in this problem setting. Even though these signals are generally in the high frequency spectrum. No universally optimal number of sub-spectrograms could be determined by the models; however, both demonstrated performance improvement. Model 1 showed a 2.1 percentage point increase over its full-spectrogram baseline, achieving a macro F_1 score of 90.3%. Model 2 showed a 8.2 percentage point increase over its full-spectrogram baseline, reaching a macro F_1 score of 77.6%. The findings suggest that sub-spectrogram architectures are a viable method for enhancing bioacoustic insect classification, potentially contributing to better monitoring and conservation efforts.

Keywords: Bioacoustics, Insect Classification, Convolutional Neural Networks, Sub-Spectrograms, Orthoptera, Cicadidae, Spectrogram Analysis, Deep Learning.

1 DATA SOURCE, ETHICS, CODE, AND TECHNOLOGY STATEMENT

This thesis project used the InsectSet66 dataset, which is publicly available online on Zenodo. The dataset is jointly owned by BioAcoustica, Xeno-canto, iNaturalist, and Baudewijn Odé. Reuse for academic purposes is allowed (Faiß, 2023).

The thesis project itself did not involve collecting data from human participants or animals. All figures and tables presented are the original creations of the author.

Code from previous research by Heily et al. (2023), Faiß & Stowell (2023), and Phaye et al. (2019) was used for the creation of Model 1 and Model 2. Reused code fragments are clearly referenced. Required Python packages for running the code can be found in Appendix A: Software Implementations.

Links to the InsectSet66 dataset, the original code of Heily et al. (2023), Faiß & Stowell (2023), Phaye et al. (2019), and code repository of this study can be found in Appendix B: Code and Data.

Considering writing, assistance of ChatGPT-4, Grammarly, Overleaf, and Microsoft Word was used for checking spelling and grammar.

2 PROBLEM STATEMENT AND RESEARCH STRATEGY

2.1 *Research Goal*

The goal of this research was to add to the field of bioacoustic insect monitoring by improving the performance of state-of-the-art bioacoustic Orthoptera and Cicadidae classification architectures. Specifically, with the problem statement focusing on evaluating whether sub-spectrogram convolutional neural network (CNN) architectures enhance classification performance.

2.2 *Problem Statement*

Global insect population decline is putting insect mediated ecosystem functions and services at risk, while these are essential for human health and human survival (van der Sluijs, 2020). This alarming trend has fuelled the desire for improved insect population monitoring. While insect monitoring can be done in various ways, the efficacy of techniques differs per insect grouping (Montgomery et al., 2021).

The research domain of bioacoustics studies animal sounds and is an important source of evidence for monitoring biodiversity. This field increasingly utilizes machine learning, and in recent years, particularly deep learning techniques for analyzing audio recordings of insect sounds (Stowell, 2022). Due to their loud species specific communication sounds, bioacoustic monitoring is an optimally suitable technique for the Orthoptera and Cicadoidea insect groupings. While, at the same time, bioacoustic monitoring is the only suitable method for this insect grouping (Montgomery et al., 2021). The reason being that most of the insects in this grouping produce species-specific sounds, while their morphological differences can be vague (Heller et al., 2021). This indicates the need for robust bioacoustic Orthoptera and Cicadoidea monitoring architectures. However, the signals produced by these groupings are different from highly studied signals from humans and birds (Faiß & Stowell, 2023). Additionally, research into bioacoustic insect classification has been minimal (Stowell, 2022), resulting in room for improvement of the state-of-the-art architectures.

State-of-the-art architectures typically use CNNs trained on spectrogram representations of audio recordings. Spectrograms visually represent the frequency spectrum of a sound signal over time, with time on the X-axis and frequency on the Y-axis (Figure 2). Historically, mel spectrograms are used because they try to mimic the human perception of sound as early acoustic deep learning tasks focused on music-based, speech, or

language recognition problems (Faiß & Stowell, 2023; Stowell, 2022). These mel spectrograms compress higher frequencies to give more emphasis to lower frequencies, mimicking human perception of sound (Faiß & Stowell, 2023; Stevens & Volkmann, 1940). However, insect sounds are generally much higher in frequency and cover broader frequency bands than most mammals and birds (Robinson & Hall, 2002). Compressing the higher frequencies could therefore obscure a valuable part of the data. Suggesting that linear spectrograms, which use a linearly scaled frequency axis, may be more suitable for bioacoustic insect classification.

Additionally, state-of-the-art architectures generally classify the entire spectrogram. This approach may be problematic as insects' sounds can exhibit slightly different characteristics across wide frequency ranges (Faiß & Stowell, 2023). Their sounds are often broadband and non-harmonic, leading to non-uniform acoustic properties that vary depending on the frequency (Faiß & Stowell, 2023). Sub-spectrogram multi-band architectures might address this issue. These architectures capture differentiating features by splitting spectrograms in frequency bands, creating sub-spectrograms. These train independent CNNs on the slices and merge these for a global classification (Figure 1). This sub-spectrogram multi-band architecture approach has shown high performance improvements in another acoustic problem setting (Phaye et al., 2019).

For these reasons, this research applied a sub-spectrogram multi-band architecture approach, trained on linear spectrograms, to improve bioacoustic Orthoptera and Cicadoidea classification.

2.3 *Societal and Scientific Impact*

The societal impact of this research is grounded in addressing and contributing to mitigating the global insect population decline and the associated risks. It is estimated that 40% of insect species are at risk of extinction (Francisco Sánchez-Bayo, 2019). This poses an international threat to vital ecosystem functions and services mediated by insects, such as freshwater and soil processes (including nutrient cycling, soil formation, decomposition, and water purification), biological pest control and, pollination. These functions are crucial for both ecosystem health and human survival. The repercussions could extend beyond the realm of insects, leading to loss of biodiversity higher up in the food-web, such as insect-eating birds, and to a reduction in ecosystem resilience (van der Sluijs, 2020).

The scientific impact of this research flows from adding to the minimal research that has been done into bioacoustic insect classification. It aims to

provide new insights into the effectiveness of using linear spectrogram data representation and sub-spectrogram architectures for this problem setting. Exploring these techniques might advance the bioacoustic research domain by potentially offering improvements in monitoring insect populations.

2.4 *Research Strategy*

To address the research goal, the following research question was formulated:

Research Question: *What is the impact of sub-spectrogram architectures on bioacoustic Orthoptera and Cicadidae sound classification compared to complete spectrograms?*

This research question was answered by answering two sub-questions:

Sub-question 1: To what extent does the use of linear spectrograms, instead of mel spectrograms, impact the performance of state-of-the-art models from the literature?

Two state-of-the-art models from literature were recreated to serve as baselines. Both were originally created using mel spectrograms. However, due to mel spectrograms compressing part of the informative high-frequency spectrum, they were recreated with the option to use linear spectrograms instead. To determine the effectiveness of this alternative data representation, both models were also trained using mel spectrograms. Comparing the results enabled answering this sub-question. Additionally, the usage of linear spectrograms allowed for a more isolated comparison with the sub-spectrogram models, which were trained on linear spectrograms. Essentially meaning that the linear trained models served as baselines for the sub-spectrogram models of sub-question 2.

Sub-question 2: What number of sub-spectrograms achieves the highest performance and how does this compare to the baselines?

The two state-of-the-art models from literature recreated for sub-question 1 were adapted to enable sub-spectrogram training. The sub-spectrogram approach involved one critical hyperparameter: The number of sub-spectrograms. This specifies in how many equally sized slices the spectrogram will be divided. Slices are continuous, i.e., without gaps or overlap. Figure 1 illustrates this concept with an example of a four sub-spectrogram model. Performance of different numbers was then compared to enable answering sub-question 2. Afterwards, the best performing models were compared to the baseline models, i.e., the linear full-spectrogram model of sub-question 1, that they were based on.

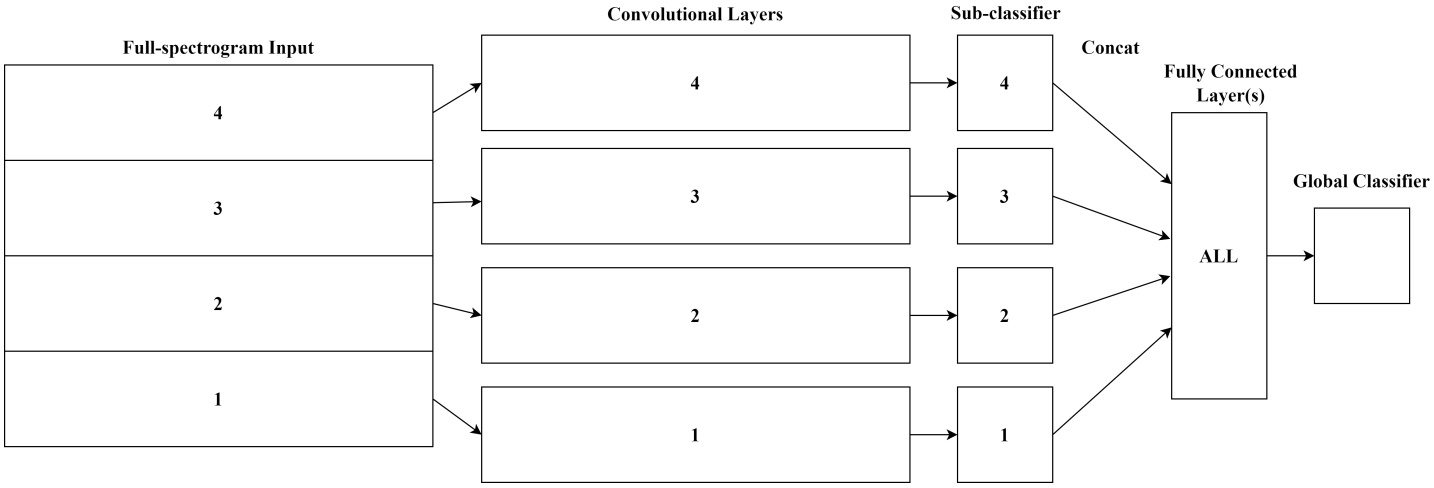

Figure 1: Visual representation of the sub-spectrogram architecture.

2.5 *Main Findings*

The research has shown that the use of linear spectrograms instead of mel spectrograms did not substantially change performance of the two state-of-the-art models. Contrary to what has been hypothesised, mel spectrogram models achieved slightly worse results in Model 1 but slightly better performance in Model 2. Model 1 trained on linear spectrograms achieved a macro F_1 score of 0.882, while the same model trained on mel spectrograms scored 0.867. Conversely, Model 2 performed slightly better with mel spectrograms, achieving a macro F_1 score of 0.713 compared to 0.694 with linear spectrograms. Visual and acoustic analysis of species performing substantially different in the linear spectrogram representation compared to the mel spectrogram representation showed no abnormal or unexpected signal characteristics, e.g., extraordinary low or high frequency signal or sparse activity. This indicates that the mel scale compression of higher frequency is not a constraining factor in this problem setting. Even though these signals are generally in the high frequency spectrum.

The usage of the sub-spectrogram architecture has shown that no universally optimal number of sub-spectrograms could be determined for the Orthoptera and Cicadidae classification problem setting. This due to the sub-spectrogram approach impacting both state-of-the-art models differently, resulting in no pattern being observed. The two sub-spectrogram approach of Model 1 performed best and showed a 2.1 percentage point increase over its baseline, achieving a 90.3% macro F_1 score. For Model 2, the four sub-spectrogram approach performed best and exhibited a 8.2 percentage point increase, reaching a 77.6% macro F1 score. Nevertheless,

both Model 1 and Model 2 showed that the sub-spectrogram architectures have the potential to improve classification performance over the current state-of-the-art models in the Orthoptera and Cicadidae sound classification problem setting.

A more extensive analysis and discussion of the results can be found in Chapter 5: Results and Chapter 6: Discussion.

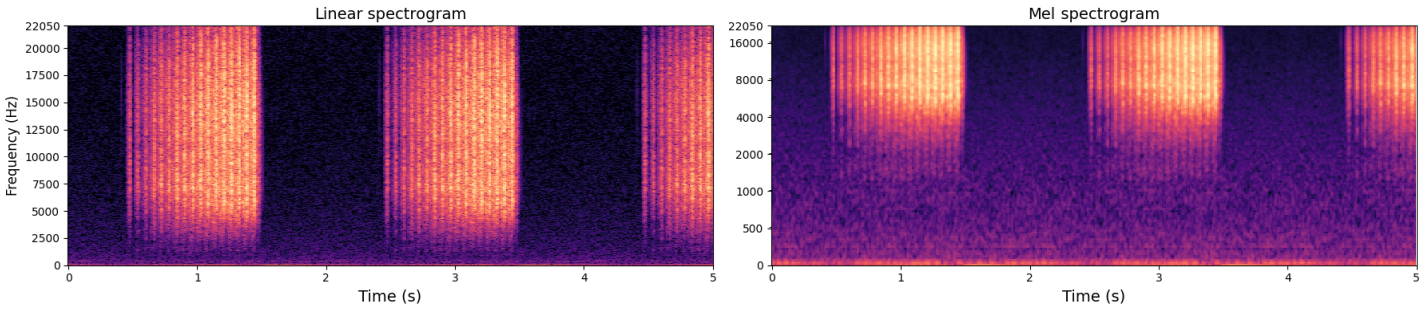

Figure 2: A linear spectrogram compared to a mel spectrogram of Chrysochraon Dispar (Orthoptera) communication. The color map indicates the magnitude in decibels (dB) with values ranging from 0 to 80 dB.

3 RELATED WORK

3.1 *Bioacoustic Monitoring and Insect Bioacoustics*

In recent years, the field of bioacoustic monitoring has seen rapid growth by adaptation of deep learning methods for the detection and classification of species. These methods have substantially reduced the time and effort required for manually processing audio data (Digby et al., 2013; Sharma et al., 2023). The lion's share of this research has been done on birds, (marine) mammals, and amphibians, with insects play a minimal role (Sharma et al., 2023; Stowell, 2022).

Focusing on bioacoustic insect classification, most research has gone to classification and detection of Orthopterans and Cicadoidea as it being the most suitable method for this insect grouping (Montgomery et al., 2021). This is because, most of the insects in this grouping produce species-specific sounds, while their morphological differences can be vague (Heller et al., 2021).

3.2 *State-of-the-art*

Generally, these deep learning methods use a convolutional neural network (CNN) architecture trained on spectrogram representations of audio recordings. Spectrograms are visual representations of the frequency spectrum of a sound signal over time, displaying time on the X-axis and frequency on the Y-axis (Figure 2). These are created by applying the Short-Time Fourier Transform (STFT) with a sliding Hann-window on an audio signal, displaying the magnitude. These magnitudes are then transformed to a

decibel scale, resulting in a linear spectrogram (Anastasia Natsiou, 2022; Faiß & Stowell, 2023; Huzaifah, 2017).

Historically, mel spectrograms are used because they try to mimic the human perception of sound, as early acoustic deep learning tasks focused on music-based, speech, or language recognition problems (Faiß & Stowell, 2023; Stowell, 2022). Humans perceive sound on a non-linear scale, making changes in the low-frequency spectrum much more noticeable than similar changes in the high-frequency spectrum (Stevens & Volkmann, 1940). This has led to the widespread usage of the mel spectrogram, which uses a mel scale that compresses the higher frequencies to give more emphasis to the lower frequencies (Faiß & Stowell, 2023). mel spectrograms are created by applying mel-spaced filter banks to the output of the STFT, by summing the energies of the STFT bins according to the mel scale filter shapes, which are triangular filters spaced according to the mel scale (Anastasia Natsiou, 2022; Faiß & Stowell, 2023; Huzaifah, 2017).

However, insect sounds are generally much higher in frequency and cover wider frequency bands than most mammals and birds (Robinson & Hall, 2002). Compressing the higher frequencies would therefore obscure a valuable part of the data, as shown in Figure 2, possibly resulting in worse classification results due to a less informative representation of the data. Literature addressing this feature representation design choice on environmental and urban sounds datasets showed that linear and mel spectrograms performed comparably, with mel spectrograms performing slightly better (Huzaifah, 2017). Nevertheless, these datasets contain many classes that are generally lower in the frequency spectrum and therefore might benefit from the mel scale. The study does not provide a confusion matrix to analyse performance per class, making it difficult to assess the performance of high- and low-frequency signals. Thus, the better feature representation method for Orthopterans and Cicadoidea remains unclear.

One other potential bottleneck of Orthopterans and Cicadoidea classification is that CNN architectures generally classify the whole spectrogram. This approach may be problematic as these insects' sounds can exhibit slightly different characteristics across wide frequency ranges (Faiß & Stowell, 2023). Their sounds are often broadband and non-harmonic, leading to non-uniform acoustic properties that vary depending on the frequency (Faiß & Stowell, 2023). Sub-spectrogram multi-band architectures might help address this bottleneck. These architectures capture differentiating features by splitting spectrograms in frequency bands, creating sub-spectrograms, and use these to train CNN architectures. The original version of this architecture was created by Phaye et al. (2019) and is known as SubSpectralNet. The fundamentals are splitting spectrograms in

horizontal slices to capture frequency bands, train independent CNNs on these slices and merging these for a global classification (Phaye et al., 2019). This has shown to yield a 14% accuracy increase over the baseline model of the DCASE 2018 acoustic sound classification challenge, classifying scenes i.e., soundscapes or environments, with best performance using three, eighteen, and nineteen sub-spectrograms (Phaye et al., 2019). Figure 1 shows a simplified visualization of the sub-spectrogram architecture.

Considering sub-spectrograms, research is very limited. A sub-spectrogram acoustic bird classification model returned no substantial improvement compared to other CNN architectures. The study only experimented with a three sub-spectrogram approach, not comparing other numbers, leaving the best performing number unclear for this problem setting (Xie et al., 2019). Another study used a sub-spectrogram convolutional recurrent neural network (CRNN) approach on the ESC-50 dataset, containing environmental sound recordings of animals, natural soundscapes human non-speech sounds, domestic sounds, and urban noises. This study achieved a 9.1% accuracy increase over its baseline with a four sub-spectrogram approach. Three to six sub-spectrogram approaches scores comparable, although slightly less (Qiao et al., 2019).

However, due to the limited research and this being conducted on a different problem setting it, remains unclear if sub-spectrogram CNN models will improve performance of Orthopterans and Cicadoidea classification.

3.3 *Relevant Datasets and State-of-the-art*

Relevant datasets used in the state-of-the-art models for Orthopterans and Cicadoidea classification include the InsectSet32, InsectSet47, and InsectSet66 datasets, consisting of the corresponding amount of Orthoptera and Cicadidae (sub-family of Cicadoidea) species (Faiß, 2022, 2023), and the European Orthoptera dataset consisting of 9 species (Gandini, 2022).

Classification performance varies with approximately 90% accuracy scores for the European Orthoptera dataset and a custom 7-species dataset (Gandini, 2022; Hibino et al., 2021). In general, research on the InsectSet datasets has been limited. Approximately 70% - 90% macro F_1 scores have been achieved for these datasets (Faiß & Stowell, 2023; Heily et al., 2023). However, it is important to mention that performance of different studies cannot be directly compared, as some classify individual 5-second segments of an audio recording, while others combine all segments of a recording to make a global classification for the complete recording.

This research used the publicly available InsectSet66 dataset. The rationale for using this dataset is it having most species of any research grade Orthoptera and Cicadoidea audio set available. Previous research achieved approximately 70% - 90% macro F_1 scores for this dataset (Faiß & Stowell, 2023; Heily et al., 2023). This shows relevance for improvement.

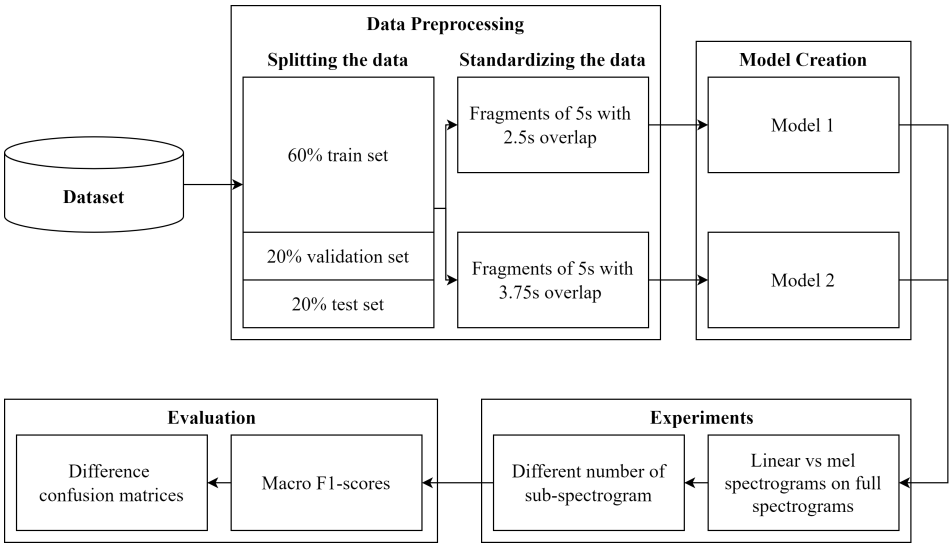

Figure 3: A visual representation of the methodology, illustrating the sequential steps and the relationships between different components.

4 METHOD

4.1 *Methodology*

Following what has already been discussed in the research strategy, two state-of-the-art models from literature were recreated to serve as baselines. Both models were adapted to also enable linear spectrogram data representation and sub-spectrogram training. The methodology explained in the following sections is visualized by Figure 3. Adding this aims to provide a clear, visual representation of the methodology, illustrating the sequential steps and the relationships between different components, thereby aiming to improve understandability and reproducibility.

4.1.1 *Preprocessing*

For enabling the models to run, some preprocessing operations needed to be conducted. Firstly, the dataset was split in the predefined train, validation, and test sets made by the authors of the dataset. This was done by utilizing the attached annotations CSV file (Table 1). Secondly, the duration of the spectrograms had to be fixed for being able to run CNN models. As recording lengths of the data varies, the recordings were be split into segments of five seconds. This is because the majority of sounds is either brief and rhythmic or long and monotone. Sequences longer than five seconds that repeated are rare in the dataset, suggesting that a five-second segment would likely retain the unique rhythmic traits specific

to each species' calls. Recordings shorter than five seconds were looped to extend them to the required five-second length (Faiß & Stowell, 2023). Longer files were divided into five-second segments with a 2.5-second overlap for Model 1 and 3.75-second overlap for Model 2. If the end of a file was reached, the start of the recording was appended to complete the five-second segment, provided the remaining portion was at least 1.25 seconds long. The same considerations have been made by the original authors of the models that were recreated.

4.1.2 *Model 1*

Model 1 is a recreation of the I&D Austria model from the Capgemini Global Data Science Challenge 2023, using their publicly available code from GitHub (Heily et al., 2023). It was created with a transfer learning approach, using the EfficientNetV2-S architecture. This architecture contains approximately 20 million trained parameters, making it smaller than most state-of-the-art transfer learning models, resulting in faster computation. While, at the same time, maintaining state-of-the-art performance (Heily et al., 2023; Tan & Le, 2021). For Model 1, the class imbalance of the data was mitigated by adding custom class weights to the categorical cross entropy loss function, weighing minority classes more and majority classes less. These were calculated with the following formula:

$$1 - (n\_files\_per\_class / n\_total\_files)$$

For this calculation, files were already standardized to five-second fragments. These custom class weights made the model penalize misclassifications of the minority classes more than those of the majority classes, increasing the model sensitivity to the minority classes. Additionally, label smoothing was applied to the loss function. This regularization technique adjusts the target labels, making them slightly less confident, improving generalization and preventing overfitting the majority class (Szegedy et al., 2015).

It can be considered standard practise to apply data augmentation methods to enhance the diversity of the data and improve robustness (Alomar et al., 2023; Perez & Wang, 2017; Stowell, 2022). This was conducted by randomly applying Noise Injection, Gaussian Noise, or Pink Noise, and Impulse Response to 10 - 15% of the training data.

4.1.3 *Model 2*

To provide a more comprehensive understanding of the effects of the linear spectrogram approach and the sub-spectrogram architecture, a second model was created. By including a second model, the study aimed to

ensure that any observed performance improvements were not limited to a single model's specifics but are broadly applicable, providing a more robust and generalizable understanding of the sub-spectrogram architecture's impact.

Model 2 is based on the model used by Faiß & Stowell (2023), which scored approximately 70% macro F_1 scores on individual 5-second fragments classification and has close to 30,000 parameters. The training data was augmented by applying Colored Noise and Impulse Response to varying frequencies distributions. With frequency power decay randomized between -2 and 1.5 and variable frequency distribution to 90% of the data and impulse response to 70% of the data. This increased the diversity of the data, improving the generalizability and preventing overfitting (Alomar et al., 2023; Perez & Wang, 2017). However, this did not mitigate the class imbalance as no classes' samples increased. No additional measures were taken to reduce the class imbalance (Faiß & Stowell, 2023).

Spectrograms were created using different hyperparameters compared to Model 1. This approach was chosen to closely replicate the state-of-the-art model of Faiß & Stowell (2023). This difference did not compromise the comparability between Model 1 and 2, as the performance was solely compared to the other data representation approach and sub-spectrogram architecture within the model itself. Just the patterns in performance of both models were compared, e.g., both Model 1 and 2 perform better on low number sub-spectrograms.

4.1.4 *Sub-spectrogram Architecture*

The multi-band sub-spectrogram approach was introduced in 2019 for an acoustic scene classification problem. It introduced the SubSpectralNet, that is a modified version of the DCASE 2018 baseline model, trained on mel spectrograms, for acoustic scene recognition (Phaye et al., 2019).

For this research, Model 1 and 2 were adapted to incorporate the core principles and technical framework of the SubSpectralNet, while matching their original model specifics. Meaning that the models utilized the same layers, hyperparameters, optimizer, and data representation parameters as its corresponding original version. This while retaining the sub-spectrogram approach of splitting the full-spectrogram, the sub-spectrogram classification, and the global classification. This meant that using one sub-spectrogram was equivalent to full-spectrogram classification, and was therefore used to serve as baselines. With also the mel spectrogram experiment being conducted with an one sub-spectrogram model.

Both Model 1 and 2 ran for a maximum of 50 epochs. An early stopping mechanism was applied, terminating training after 10 epochs without improving validation macro F_1. The model from the epoch with the highest validation macro F_1 was used for final prediction on the test set. Monitoring the validation F_1 score can be more effective than validation loss when dealing with imbalanced classification tasks, because it gives a balanced view of the model's performance by considering both precision and recall. This stimulates that the model performs well across all classes, not just the majority class (Jeni et al., 2013; Müller & Guido, 2016; Saito & Rehmsmeier, 2015). Final prediction is done for complete recordings, not individual 5-second fragments. This by predicting all preprocessed fragments of a recording and averaging these for a global prediction.

To keep the scope of the experiments in line with the time-constraints of the study, comparison of linear and mel spectrograms was only conducted for full-spectrogram classification and eight number of sub-spectrograms were tested for Model 1 and 2. For the same reason, only equally sized continuous sub-spectrogram were experimented with for each test.

Afterwards, the best performing sub-spectrogram models were compared to their corresponding baseline models that used linear full-spectrogram classification. This enabled answering the research question: *"What is the impact of sub-spectrogram architectures on bioacoustic Orthoptera and Cicadidae sound classification compared to complete spectrograms?"*.

4.1.5 *Evaluation*

The generalizability of the models was validated by using the holdout approach. Train, validation, and test set were created according to the predefined split. This split is stratified, ensuring approximately equal distribution of audio files and material of species for each subset.

Models were trained by minimizing the training loss and monitoring the validation macro F_1 score. The models from the epochs with the highest validation macro F_1 score were used to make the final predictions on the hold-out set, i.e., the test set. This ensured that the models can generalize beyond the validation set. K-fold Cross-validation was not used due to time constraints, as this would increase the already high computational cost by K.

Results were evaluated by comparing macro F_1 scores on the test set. This metric computes the binary F_1 score for each class, considering that class as the positive class while treating the other classes as negative. The per class F_1 scores are then averaged without weight, giving equal importance to all classes regardless of their support (Müller & Guido,

2016). This metric was used for multiple reasons. Firstly, because there is no reason to emphasize minimizing false positives or false negatives, as both types of errors are equally important in this context. This makes the precision and recall classification evaluation metrics less relevant (Müller & Guido, 2016). Secondly, accuracy was deemed inappropriate due to severe class-imbalance in the dataset, which would result in misleading and uninformative high accuracy by predicting the majority classes (Müller & Guido, 2016). Thirdly, species are considered to be of equal importance. Therefore, micro F_1 scores and weighted F_1 scores are inappropriate, as both incorporate the support in their calculation (Müller & Guido, 2016).

The errors were further analysed using difference confusion matrices. These matrices, illustrate the differences between the confusion matrices of the full-spectrogram models trained with linear spectrograms and mel spectrograms. The values from the linear spectrogram confusion matrices were subtracted from those of the mel spectrogram confusion matrices. The resulting values were then transformed into percentages of mistakes relative to their support.

This transformation results in the relative performance difference. For example, a positive value X on the diagonal means that the mel spectrogram model was better at classifying X% of that species compared to the linear spectrogram model. Conversely, the species with which the linear spectrogram model was confused can be found in the row corresponding to that species, with the sum of the negative values adding up to the positive value on the main diagonal.

The black column on the right shows the support per species. Red rows indicate species with a substantial performance difference of at least 50% between the two approaches.

4.2 *Experimental Setup*

4.2.1 *Dataset Description*

This research used the publicly available InsectSet66 dataset, containing 1554 manually inspected recordings from 66 Orthoptera and Cicadidae species. The total length is over 24 hours, and each species has minimally 10 recordings. These are standardized to 44.1 kHz mono WAV files. Length ranges from less than one second to multiple minutes. The dataset is highly class imbalanced. The number of recordings varies from 10 to 152 per species, while total recording time varies from 0:01:20 to 1:37:39. Recordings were removed if strong noise inference or filtering was found, and when it contained multiple species. A stratified train, validation, and

	File Name	Species	Subset
1	Roeselianaroeselii_XC751814-dat028-019_edit1.wav	Roeselianaroeselii	train
2	Roeselianaroeselii_XC752367-dat006-010.wav	Roeselianaroeselii	train
3	Yoyettacelis_GBIF2465208563_IN36000894_50988.wav	Yoyettacelis	train
4	Gomphocerippusrufus_XC752285-dat001-045.wav	Gomphocerippusrufus	train
5	Atrapsaltacorticina_GBIF2901504947_IN62966536_143690.wav	Atrapsaltacorticina	validation
	Unique File	**Link**	**Contributor**
1	Roeselianaroeselii_XC751814-dat028-019	https://xeno-canto.org/751814	Baudewijn Odé
2	Roeselianaroeselii_XC752367-dat006-010	https://xeno-canto.org/752367	Baudewijn Odé
3	Yoyettacelis_GBIF2465208563_IN36000894_50988	https://www.inaturalist.org/observations/36000894	Christie
4	Gomphocerippusrufus_XC752285-dat001-045	https://xeno-canto.org/752285	Baudewijn Odé
5	Atrapsaltacorticina_GBIF2901504947_IN62966536_143690	https://www.inaturalist.org/observations/62966536	Nathan Emery

Table 1: The InsectSet66 dataset comes with an annotations CSV file specifying the file name, species, subset, the unique source file, link to the source, and the contributor. This table displays the first five rows.

test split has already been made by the publishers to ensure approximately equal distribution of audio files and material of species for each subset. The data is split 60/20/20 on file amount and 64/19.5/16.5 by length. Recordings including silent periods of more than 5 seconds were split but included in the same split to prevent data leakage. The dataset was created by combining recordings from BioAcoustica, xeno-canto, Baudewijn Odé and iNaturalist (Faiß, 2023).

4.2.2 *Model 1*

The mel spectrograms were created using 128 mel bins, a 2048-point Fast Fourier Transform (FFT), a window size of 2048, a hop length of 1024, a power of 2, a minimum frequency of 400 Hz, and a maximum frequency of 30,000 Hz. For the linear spectrograms, a 2048-point FFT, a window size of 2048, a hop length of 1024, and a power of 2 were used. Both spectrogram types were normalized by dividing by the sum of the window function values. Subsequently, the amplitude values were converted to the decibel (dB) scale with a max dB value of 80.

To ensure proper comparability and to prevent computational constrains, the linear spectrogram feature maps were resized to match the size of the mel spectrogram feature maps, specifically 128 x 216 pixels. The downscaling was performed using PyTorch's bicubic interpolation with anti-aliasing, as this method best preserved feature map quality (Gonzalez & Woods, 2018).

The model ran for 50 epochs with a batch size of 32, label smoothing of 0.1, an AdamW optimizer, and a learning rate of 0.0017.

4.2.3 *Model 2*

This model used a CNN architecture containing five 2D convolutional layers utilizing a ReLU activation function and batch normalization. Afterwards, the feature maps were pooled, flattened, and put into a linear layer. Dropout was implemented on the final linear as well as L2 regularization of the weights (Faiß & Stowell, 2023).

The mel spectrograms used 64 mel bins, a 1000-point Fast Fourier Transform (FFT), a window size of 294, a hop length of 147, and a maximum frequency of 22,050 Hz. For the linear spectrograms, a 1000-point FFT, a window size of 294, and a hop length of 147 were used. Both spectrogram amplitude values were converted to the decibel (dB) scale with a max dB value of 80. To ensure proper comparability and to prevent computational constrains, the linear spectrogram feature maps were resized to match the size of the mel spectrogram feature maps, specifically 64 x 1501 pixels. The downscaling was performed using PyTorch's bicubic interpolation with anti-aliasing, as this method best preserved feature map quality (Gonzalez & Woods, 2018).

The model a maximum of 50 epochs with a batch size of 14, an Adam optimizer, and a learning rate of 0.001.

4.2.4 *Software Implementations*

The architectures were made in Python using PyTorch for training the deep learning models. For running the models, Google Colab Pro was used as it provided the computational requirements to run the deep learning models. An overview of the required Python packages can be found in Appendix A: Software Implementations.

Model 1	F_1 **score**
Mel spectrogram	0.867
Linear spectrogram	**0.882**
Model 2	
Mel spectrogram	**0.713**
Linear spectrogram	0.694

Table 2: Mel versus linear full-spectrogram macro F_1 scores.

5 RESULTS

5.1 *Linear Versus Mel Spectrograms*

This section covers and compares the classification results of the mel and linear full-spectrogram trained Model 1 and Model 2.

Table 2 shows that models trained on mel spectrograms do not substantially underperform compared to those trained on linear spectrograms, contrary to what has been hypothesised. The mel spectrogram approach achieved slightly worse performance in Model 1 but slightly better performance in Model 2. These results are contradictory, however, with the performance of Model 1 differing by only 1.5 percentage points and Model 2 differing by just 1.9 percentage points, the difference can be considered very small. Because of this, the results seem to indicates that the mel scale compression of higher frequency is not a constraining factor in this problem setting, as this would have likely resulted in substantial improvements of the linear spectrogram trained models. Even though these signals in this problem setting are generally in the high frequency spectrum.

Further analysis of the results has been conducted with difference confusion matrices. These can be found in Appendix C. These figures illustrate the differences between the confusion matrices of the full-spectrogram models trained with linear spectrograms and mel spectrograms. The values from the linear spectrogram confusion matrices were subtracted from those of the mel spectrogram confusion matrices. The resulting values were then transformed into percentages of mistakes relative to the support per species. This transformation results in the relative performance difference. For example, a positive value X on the diagonal means that the mel spectrogram model was better at classifying X% of that species compared to the linear spectrogram model. Conversely, the species with which the linear spectrogram model was confused can be found in the row corresponding to that species, with the sum of the negative values adding

	Number of sub-spectrograms							
Model 1	1 (baseline)	2	3	4	5	6	7	8
F_1 score	0.882	**0.903**	0.867	0.858	0.874	0.857	0.871	0.864
+/- Baseline	0.0%	**+2.1%**	-1.5%	-2.4%	-0.8%	-2.5%	-1.1%	-1.8%
Model 2	1 (baseline)	2	3	4	5	6	7	8
F_1 score	0.694	0.735	0.741	**0.776**	0.738	0.739	0.769	0.698
+/- Baseline	0.0%	+4.1%	+4.7%	**+8.2%**	+4.4%	+4.5%	+7.5%	+0.4%

Table 3: Macro F_1 scores of linear spectrogram trained models with different numbers of sub-spectrograms, including percentage point difference to the corresponding baseline.

up to the positive value on the main diagonal. Red rows indicate species with a substantial performance difference of at least 50% between the two approaches. The black column on the right shows the support per species.

Spectrograms of species reaching the 50% threshold were visually and acoustically analyzed to infer a possible cause of the difference in error patterns. For Model 1 this concerns the 'Stenobothrus stigmaticus', 'Atrapsalta collina', 'Gomphocerippus rufus', and 'Ephippiger diurnus'. For Model 2 the 'Atrapsalta collina', 'Cyclochila australasiae', 'Yoyetta celis', 'Omocestus rufipes', 'Gryllus campestris', 'Eupholidoptera schmidti', and 'Tettigonia viridissima' passed the threshold. Only the 'Atrapsalta collina' passed it in both models. Most interesting being species with a medium to high support, as the threshold can otherwise be easily crossed. This being the 'Gomphocerippus rufus', 'Stenobothrus stigmaticus', 'Gryllus campestris', and 'Tettigonia viridissima'. However, no abnormal or unexpected signal characteristics, e.g., extraordinary low or high frequency signal or sparse activity, were observed.

5.2 *Sub-spectrogram Models*

This section covers and compares the classification results of the linear sub-spectrogram trained Model 1 and Model 2.

The results displayed in Table 3 show that the sub-spectrogram approach impacts the two models differently. For Model 1, employing a two sub-spectrogram approach yields a modest improvement of 2.1 percentage point over its corresponding baseline. However, with larger number of sub-spectrograms, the performance decreases, resulting in outcomes that are slightly worse than the baseline full-spectrogram classification.

Conversely, Model 2 shows a substantial performance increase using the sub-spectrogram approach. All number of sub-spectrograms show improved performance, with an average increase of 4.8 percentage points. The model utilizing four sub-spectrograms achieves the highest performance, showing a 8.2 percentage point increase in performance over its corresponding baseline. However, the eight sub-spectrograms model only improves performance with 0.4 percentage point. Therefore, it can not be concluded that a higher or lower number of sub-spectrograms consistently performs better in this model. Overall, the results indicate a clear advantage of the sub-spectrogram approach for Model 2.

The contradictory nature of these results prevents to identify any clear performance patterns between the two models.

6 DISCUSSION

The goal of this research was to add to the field of bioacoustic insect monitoring by improving the performance of state-of-the-art bioacoustic Orthoptera and Cicadidae classification architectures. Specifically, with the problem statement focusing on evaluating whether sub-spectrogram convolutional neural network architectures enhance classification performance, and in the process, evaluate if a linear spectrogram representation of the data performs better in this problem setting.

6.1 *Sub-question 1*

To what extent does the use of linear spectrograms, instead of mel spectrograms, impact the performance of state-of-the-art models from the literature?

Comparing linear and mel spectrogram trained models showed that mel trained models do not substantially underperform compared to those trained on linear spectrograms, contrary to what has been hypothesised. Mel spectrogram models achieved slightly worse results in Model 1 but slightly better performance in Model 2. However, the performance of Model 1 differing by only 1.5 percentage points and Model 2 differing by just 1.9 percentage points, the difference can be considered very small. Because of this, the results seem to indicates that the mel scale compression of higher frequency is not a constraining factor in this problem setting, as this would have likely resulted in substantial improvements of the linear spectrogram trained models. Even though these signals in this problem setting are generally in the high frequency spectrum.

When comparing this result with the limited literature addressing this problem, the findings seem to support earlier observations that showed that linear and mel spectrograms performed comparably. Although mel spectrograms performed marginally better (Huzaifah, 2017). However, the datasets used in this research contained many classes that were generally lower in the frequency spectrum and therefore might have benefited from the mel scale. It could therefore be that in low frequency spectrum problem settings, the mel spectrogram trained models actually perform slightly better than the linear spectrogram trained models.

A possible explanation to why both approaches also performed comparably in Orthoptera and Cicadidae classification could be the generally broadband nature of these insects sounds. This possibly ensures that, even when the higher frequencies are compressed, there is still enough representation of the signal to adequately learn it. An additional reason

might be that the hyperparameters used by the model, that were optimized by the original authors of the models, were tuned on the mel spectrogram representation of the data. Separate hyperparameter tuning on the linear spectrogram representation of the data might result in a different set of optimal hyperparameters and possibly better performance.

Considering scientific impact, the result adds to the evidence that mel spectrograms are a superior representation method for acoustic data, as these seem to perform better on low frequency problems and comparable on high frequency problems. However, it should again be mentioned that the literature is limited. Future research on narrow-band high frequency problems and additional hyperparameter tuning should be conducted to be able establish a more conclusive answer to if linear spectrograms improve performance in some problem settings. However, until that is conducted, no performance increase is observed by using linear spectrograms in the Orthoptera and Cicadidae classification problem setting.

6.2 *Sub-question 2*

What number of sub-spectrograms achieves the highest performance and how does this compare to the baselines?

The results showed that the sub-spectrogram approach impacts the two models differently. With Model 1 showing best performance with a two sub-spectrogram approach, indicating that dividing the spectrogram into two frequency bands best captured relevant features for this model. Model 2 performed best with the four sub-spectrograms approach, suggesting that splitting the spectrogram into four frequency bands captured more relevant features for classification in this model. These differing results indicate that the optimal number of sub-spectrograms appears to be model-dependent, varying based on the architecture and specific characteristics of each model. Additionally, these differing results prevent this research from determining an universally optimal number of sub-spectrograms for the Orthoptera and Cicadidae classification problem setting.

Considering the limited research into sub-spectrograms, results vary based on the studies and models used. The SubSpectralNet, created by the original developer of the sub-spectrogram architecture, showed the best performance using three, eighteen, and nineteen sub-spectrograms (Phaye et al., 2019). This suggests that the number of sub-spectrograms can greatly impact performance. The sub-spectrogram acoustic bird classification model only experimented with a three sub-spectrogram approach, not comparing other numbers (Xie et al., 2019). This limitation means that

the potential benefits of using different numbers of sub-spectrograms was not explored, leaving the optimal number unclear for this problem setting. Additionally, the environmental sound recordings sub-spectrogram model achieved its best results using a four sub-spectrogram convolutional recurrent neural network approach, with similarly good performance between three and six sub-spectrograms (Qiao et al., 2019). The varying optimal number of sub-spectrograms further illustrates the variability and lack of a clear pattern in determining the optimal number. From this literature, no pattern in the superior number of sub-spectrograms can be determined as the research is limited, is partially done with different neural network architectures, and is done on different problem settings. This variability shows need for more extensive and systematic research to identify the factors that determine the optimal number of sub-spectrograms for different classification tasks.

Finding a possible explanation for why Model 1 and 2 show different characteristics when using different numbers of sub-spectrograms is challenging. Several factors could contribute to these differences. Firstly, Model 1 and Model 2 have different neural network architectures, meaning that they process and interpret the data differently. The layers, activation functions, and overall structure of each model can substantially influence how the bioacoustic signals are learned. Secondly, both models have been tuned with different hyperparameters setting, such as learning rates, batch sizes, and regularization techniques. This can also affect the ability of the model to learn and generalize. Thirdly, parameters used to generate the spectrograms, such as window size, hop length, and number of Fast Fourier Transforms, can impact the features captured by the spectrogram. Fourthly, the different augmentation techniques applied to the training data can affect how well the models learn from the data. Lastly, the slight differences in creating five second fragments might differ how well the data is learned. Despite these possible explanations, the black box nature of neural networks prevents giving a conclusive reason.

Scientific impact flows from the results contributing to the growing body of literature on bioacoustic classification, specifically in the use of sub-spectrogram architectures. The study suggests that there is no universally superior number of sub-spectrograms for Orthoptera and Cicadidae classification, pointing to the complexity and model-dependence of this approach.

6.3 *Research Question*

What is the impact of sub-spectrogram architectures on bioacoustic Orthoptera and Cicadidae sound classification compared to complete spectrograms?

The results have shown that the impact of sub-spectrogram architectures on bioacoustic Orthoptera and Cicadidae sound classification compared to complete spectrograms varies. For Model 1, employing a two sub-spectrogram approach yielded a modest improvement of 2.1 percentage point over its corresponding baseline, with larger number of sub-spectrograms, resulting in outcomes that were slightly worse than its baseline full-spectrogram classification. Conversely, Model 2 experienced a substantial performance boost with all configurations showing enhanced performance, averaging an increase of 4.8 percentage points, with a maximum increase of 8.2 percentage points over its corresponding baseline. This discrepancy suggests that the effectiveness of sub-spectrograms is highly dependent on the model specifics i.e., model layers, hyperparameters, optimizer, and data representation parameters of the models being used. Therefore, these findings emphasize the necessity of empirical testing with different sub-spectrogram configurations to determine the optimal setup for each specific model and dataset. Nevertheless, both Model 1 and Model 2 showed that sub-spectrogram architectures have the potential to improve classification performance over the current state-of-the-art models in the Orthoptera and Cicadidae sound classification problem setting.

The limited literature on sub-spectrograms shows that performance varies based on the studies and models used. The SubSpectralNet, developed by the original creator of the sub-spectrogram architecture, showed a 14% accuracy increase over its baseline model (Phaye et al., 2019). The environmental sound recordings sub-spectrogram model achieved a 9.1% accuracy increase over its baseline (Qiao et al., 2019). Contradicting these improvements, the sub-spectrogram acoustic bird classification model returned no improvement compared to other convolutional neural network (CNN) architectures, performing on par with a common mel spectrogram CNN (Xie et al., 2019). This result shows that the sub-spectrogram approach does not universally outperform traditional methods across all problem setting. It shows the importance of specific data characteristics and the classification task at hand when choosing an appropriate spectrogram processing method. From this literature the sub-spectrogram approach's strength can still be doubted due to the contradictory results.

Finding a possible explanation for why the impact differs between both models is difficult. Firstly, Model 1 and Model 2 have different neural network architectures, meaning that they process and interpret the data differently. The layers, activation functions, and overall structure of each model can substantially influence how the bioacoustic signals are learned. Especially when using pretrained EfficientNetV2-S architecture, that is already a very optimized model, limiting the potential improvement that

can be found in other models. Secondly, both models have been tuned with different hyperparameters setting, such as learning rates, batch sizes, and regularization techniques. This can also affect the ability of the model to learn and generalize. Thirdly, parameters used to generate the spectrograms, such as window size, hop length, and number of Fast Fourier Transforms, can impact the features captured by the spectrogram. Fourthly, the different augmentation techniques applied to the training data can affect how well the models learn from the data. Lastly, the slight differences in creating fragments might differ how well the data is learned. Despite these possible explanations, the black-box nature of neural networks prevents giving a conclusive reason.

The scientific impact of this research flows from adding to the evidence that sub-spectrogram architectures can substantially improve performance and can do so in varying problem settings. Future research testing more sub-spectrogram architectures in in the Orthoptera and Cicadidae sound classification problem could provide a more conclusive answer to determine the impact of sub-spectrogram architectures on bioacoustic Orthoptera and Cicadidae sound classification compared to complete spectrograms.

7 CONCLUSION

This study advances Orthoptera and Cicadidae species classification by researching sub-spectrogram convolutional neural network architectures and spectrogram data representation approaches, addressing the urgent issue of global insect population decline. Findings contribute to the field of bioacoustics by exploring innovative methods to enhance the performance and efficiency of insect sound classification, which is vital for effective monitoring and conservation strategies.

Experiments revealed that models trained on linear and mel spectrograms performed comparably, indicating that the mel scale compression of higher frequencies is not a constraining factor in this problem setting. Even though these signals are generally in the high frequency spectrum. Additionally, experiments with Model 1 and Model 2 could not determine a generally optimal number of sub-spectrograms for the Orthoptera and Cicadidae classification problem setting. However, it was demonstrated that both models have the potential to improve classification performance over the current state-of-the-art models in the Orthoptera and Cicadidae sound classification problem setting, albeit with varying degrees of performance increase.

The findings suggest sub-spectrogram architectures are a viable method for improving bioacoustic insect classification, potentially aiding in better monitoring and conservation efforts.

REFERENCES

Alomar, K., Aysel, H. I., & Cai, X. (2023). Data augmentation in classification and segmentation: A survey and new strategies. https://doi.org/10.3390/jimaging9020046

Anastasia Natsiou, S. O. (2022). Audio representations for deep learning in sound synthesis: A review. https://doi.org/10.48550/arXiv.2201.02490

Digby, A., Towsey, M., Bell, B. D., & Teal, P. D. (2013). A practical comparison of manual and autonomous methods for acoustic monitoring. https://doi.org/10.1111/2041-210X.12060

Faiß, M. (2022). Insectset32: Dataset for automatic acoustic identification of insects (orthoptera and cicadidae). https://doi.org/10.5281/zenodo.7072196

Faiß, M. (2023). Insectset47 insectset66: Expanded datasets for automatic acoustic identification of insects (orthoptera and cicadidae). https://doi.org/10.5281/zenodo.8252141

Faiß, M., & Stowell, D. (2023). Adaptive representations of sound for automatic insect recognition. *Plos Computational Biology*. https://doi.org/10.1371/journal.pcbi.1011541

Francisco Sánchez-Bayo, K. A. W. (2019). Worldwide decline of the entomofauna: A review of its drivers. https://doi.org/10.1016/j.biocon.2019.01.020

Gandini, D. (2022). Insect species sound classification using deep learning with small data. https://arno.uvt.nl/show.cgi?fid=160119

Gonzalez, R. C., & Woods, R. E. (2018). *Digital image processing*. Pearson.

Heily, R., Kemetinger, L., Lemm, D., & Unterberger, L. (2023). Capgemini global data science challenge. https://github.com/Dom1L/GDSC23?tab=readme-ov-file#model-evaluation

Heller, K.-G., Baker, E., Ingrisch, S., Korsunovskaya, O., Liu, C.-X., Riede, K., & Warchałowska-Śliwa, E. (2021). Bioacoustics and systematics of mecopoda (and related forms) from south east asia and adjacent areas (orthoptera, tettigonioidea, mecopodinae) including some chromosome data. https://doi.org/10.11646/ZOOTAXA.5005.2.1

Hibino, S., Suzuki, C., & Nishino, T. (2021). Classification of singing insect sounds with convolutional neural network. https://doi.org/10.1250/ast.42.354

Huzaifah, M. H. (2017). Comparison of time-frequency representations for environmental sound classification using convolutional neural networks. https://doi.org/10.48550/arXiv.1706.07156

Jeni, L. A., Cohn, J. F., & Torre, F. D. L. (2013). Facing imbalanced data–recommendations for the use of performance metrics. https://doi.org/10.1109/acii.2013.4

Montgomery, G. A., Belitz, M. W., Guralnick, R. P., & Tingley, M. W. (2021). Standards and best practices for monitoring and benchmarking insects. *Frontiers in Ecology and Evolution*. https://doi.org/10.3389/fevo.2020.579193

Müller, A. C., & Guido, S. (2016). *Introduction to machine learning with python*. O'Reilly Media, Inc.

Perez, L., & Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. https://doi.org/10.48550/arXiv.1712.04621

Phaye, S. S. R., Benetos, E., & Wang, Y. (2019). Subspectralnet - using sub-spectrogram based convolutional neural networks for acoustic scene classification. https://doi.org/10.1109/ICASSP.2019.8683288

Qiao, T., Zhang, S., Zhang, Z., Cao, S., & Xu, S. (2019). Sub-spectrogram segmentation for environmental sound classification via convolutional recurrent neural network and score level fusion. https://doi.org/10.1109/SiPS47522.2019.9020418

Robinson, D. J., & Hall, M. J. (2002). Sound signalling in orthoptera. https://doi.org/10.1016/S0065-2806(02)29003-7

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. https://doi.org/10.1371/journal.pone.0118432

Sharma, S., Sato, K., & Gautam, B. P. (2023). A methodological literature review of acoustic wildlife monitoring using artificial intelligence tools and techniques. https://doi.org/10.3390/su15097128

Stevens, S. S., & Volkmann, J. (1940). The relation of pitch to frequency: A revised scale. https://doi.org/10.2307/1417526

Stowell, D. (2022). Computational bioacoustics with deep learning: A review and roadmap. https://doi.org/10.7717/peerj.13152

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2015). Rethinking the inception architecture for computer vision. https://doi.org/https://doi.org/10.48550/arXiv.1512.00567

Tan, M., & Le, Q. V. (2021). Efficientnetv2: Smaller models and faster training. https://doi.org/10.48550/arXiv.2104.00298

van der Sluijs, J. P. (2020). Insect decline, an emerging global environmental risk. *Current Opinion in Environmental Sustainability*. https://doi.org/10.1016/J.COSUST.2020.08.012

Xie, J., Hu, K., Zhu, M., Yu, J., & Zhu, Q. (2019). Investigation of different cnn-based models for improved bird sound classification. https://doi.org/10.1109/ACCESS.2019.2957572

8 APPENDICES

8.1 *Appendix A - Software Implementations*

Required Python packages for running the author's code:

- Pandas - 2.0.3
- Numpy - 1.25.2
- Tqdm - 4.66.4
- Librosa - 0.10.2
- Soundfile - 0.12.1
- Dcase_util - 0.2.11
- Torch - 2.3.0
- Torchaudio - 2.3.0
- Torchvision - 0.18.0
- Timm - 1.0.7
- Sklearn - 1.2.2

8.2 *Appendix B - Code and Data*

The code of Model 1 and 2 can be found using the link below to a GitHub repository. This code is a combination and adaptation of the original codes from Heily et al. (2023), Faiß & Stowell (2023), and Phaye et al. (2019). Therefore, also the links to these GitHub repositories are provided, as well as the link to the InsectSet66 dataset.

- Code created for this project
- Original code of Heily et al. (Model 1)
- Original code of Faiß & Stowell (Model 2)
- Original code of Phaye et al. (SubSpectralNet)
- InsectSet66

8.3 *Appendix C - Difference Confusion Matrices*

The figures in this appendix illustrate the differences between the confusion matrices of the full-spectrogram models trained with linear spectrograms and mel spectrograms. The values from the linear spectrogram confusion matrices were subtracted from those of the mel spectrogram confusion matrices. The resulting values were then transformed into percentages of mistakes relative to the support per species.

This transformation results in the relative performance difference. For example, a positive value X on the diagonal means that the mel spectrogram model was better at classifying X% of that species compared to the linear spectrogram model. Conversely, the species with which the linear spectrogram model was confused can be found in the row corresponding to that species, with the sum of the negative values adding up to the positive value on the main diagonal.

Red rows indicate species with a substantial performance difference of at least 50% between the two approaches. The black column on the right shows the support per species.

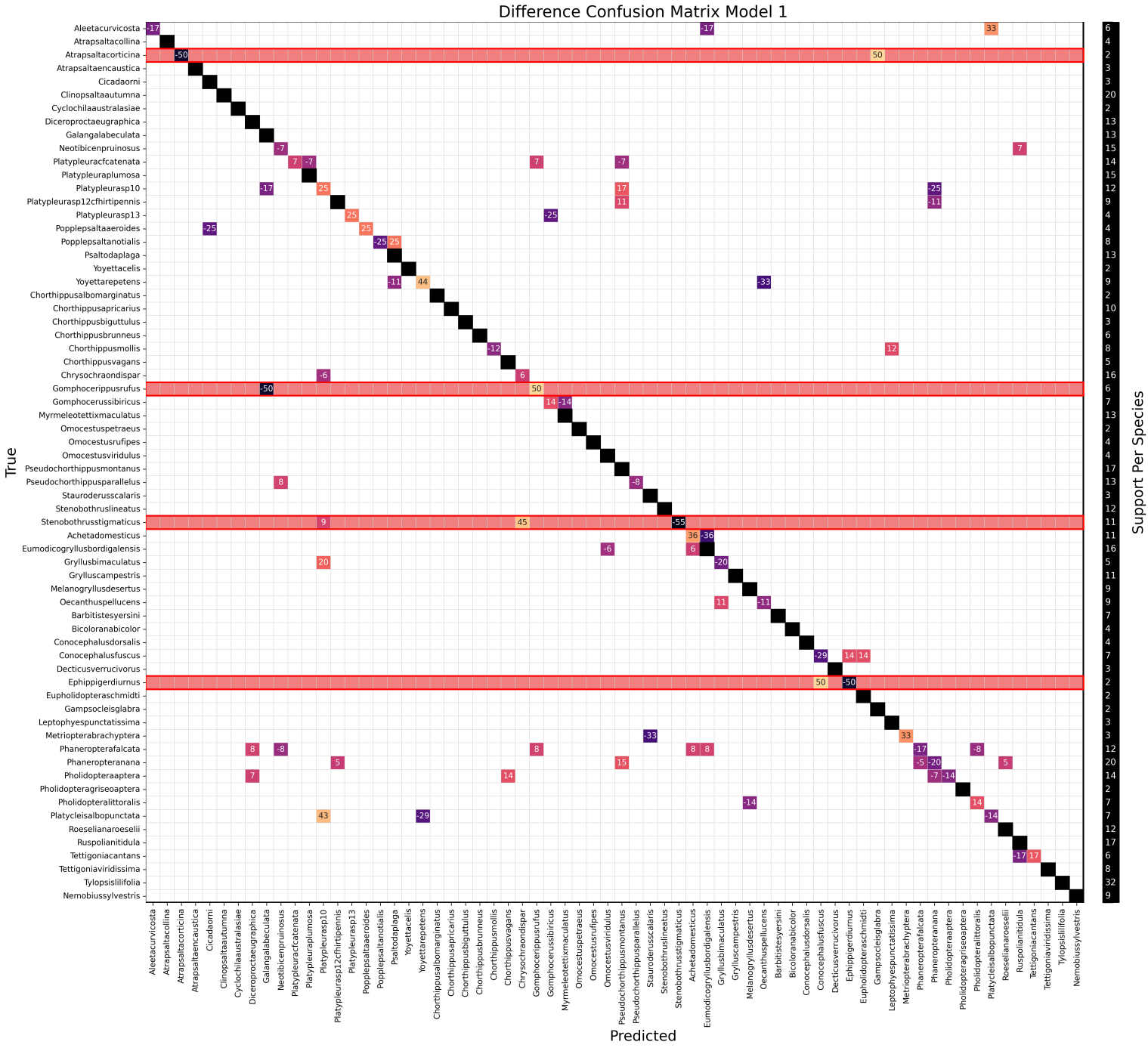

Figure 4: Difference confusion matrix Model 1.

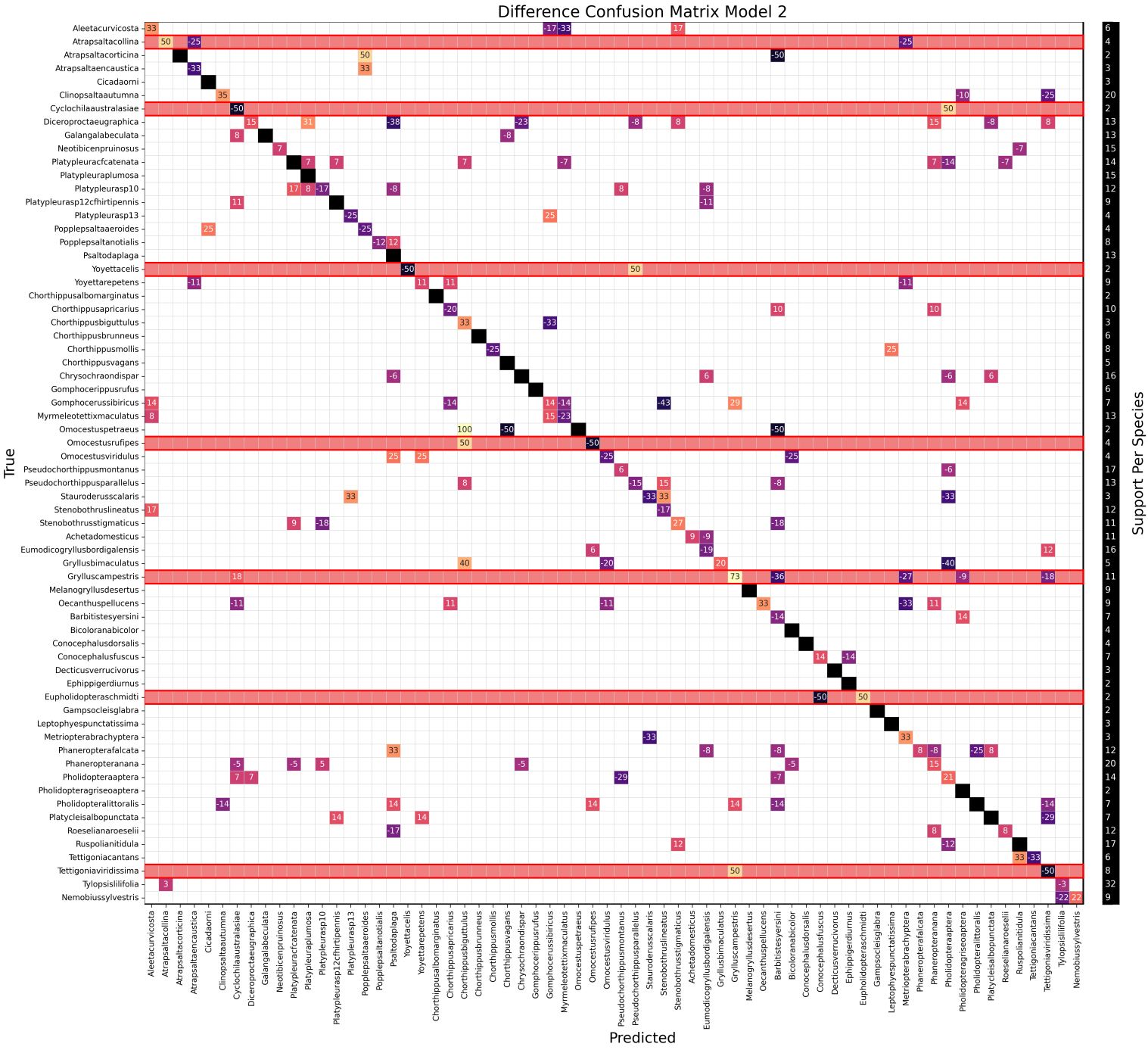

Figure 5: Difference confusion matrix Model 2.