Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

Improved Insect Sound Classification using Modulation Spectrograms and Late Integration

Ali Saudi[1] | Dragan Chobanov[2] | Hanna Lukashevich[1] | Jakob Abeßer[1]

[1]Semantic Music Technologies Group, Fraunhofer Institute for Digital Media Technologies (IDMT), Ilmenau, Germany

[2]Institute of Biodiversity and Ecosystem Research at the Bulgarian Academy of Sciences, Sofia, Bulgaria

Correspondence
Corresponding author: Jakob Abeßer.
Email: jakob.abesser@idmt.fraunhofer.de

Abstract

Acoustic classification of insect species is essential for biodiversity monitoring, agricultural protection, and ecological research. Identifying insect species through their distinctive sound patterns provides an efficient and non-invasive alternative to the traditional visual methods, especially in ecosystems where insects play critical roles in pollination, nutrient cycling, and population control. This research significantly improves insect sound classification by employing modulation spectrograms as input features, successfully capturing repetitive insect sound patterns. Additionally, a late integration strategy is proposed, combining output decisions from traditional features, such as Mel spectrograms and LEarnable Audio Front-end, with modulation spectrograms. Experimental evaluations on three public datasets demonstrate that the proposed multi-branch neural network architecture achieves up to 88% classification accuracy across 47 classes of insects, outperforming existing approaches with a 3% of relative improvement.

KEYWORDS

audio classification, insect sound, biodiversity monitoring, deep learning, convolutional neural networks

1 | INTRODUCTION

Insects are the most diverse group of animals on Earth, comprising around 75% to 80% of all known animal species Bánki et al. (2024). According to recent estimates, there might be ca. 8.7 million eukaryotic species on Earth, of which 2.6 Mora et al. (2011) or even 5.5 million Stork (2018) are insects. Despite various even larger estimates, only 1.1 to 1.7 million insect species have been described, leaving many species yet to be discovered Sankarganesh (2017), Cardoso et al. (2020). Their diversity is attributed to evolutionary traits such as exoskeleton, flight, diverse feeding and ecological specializations (including econiche split in larvae and adults of holometabolous insects). This diversity is closely linked to plant life through co-evolutionary processes like herbivory and pollination, which play essential roles in ecosystems Scudder (2017).

Insects are crucial for ecosystems, including human-modified ecosystems, providing vital functions such as pollination, pest control, and nutrient cycling. These functions are key to maintaining biodiversity and food security. Bees, butterflies, beetles, dipterans, and ants, as important pollinators, help to ensure plant reproduction. In addition, insects serve as bioindicators, providing valuable information on environmental health, soil quality, levels of pollution, and biodiversity Kalita and Das (2023). However, insect populations are rapidly declining due to climate change, habitat destruction, pollution, and pesticide use Hallmann et al. (2017), Leather (2017), Sánchez-Bayo and Wyckhuys (2019), Hailay Gebremariam (2024). This decline poses a significant threat to ecosystems, agriculture, and public health and highlights the urgent need for conservation efforts.

Bio-monitoring techniques for tracking insect populations, identifying species at risk, and supporting conservation efforts, are essential to address these challenges Van Klink et al. (2022). The monitoring of insect sounds offers a non-invasive method to monitor and study insect populations in various habitats and therefore to guide conservation actions and improve ecosystem health Faiß and Stowell (2023), Branding et al. (2024), He et al. (2024). Automatic bioacoustic monitoring systems, which have been successful in tracking vertebrate species, are now being adapted for insect research. This approach has demonstrated potential to advance ecological research, pest management, and bio-monitoring Varma et al. (2021), Branding et al. (2024). Machine learning techniques, particularly those designed to classify acoustic signals, have demonstrated remarkable effectiveness. This capability contributes significantly to bio-monitoring and conservation efforts

Abbreviations: CNN: Convolutional neural network; LI: Late integration; MFCC: Mel-frequency cepstral coefficients; LEAF: Learnable front-end for audio classification.

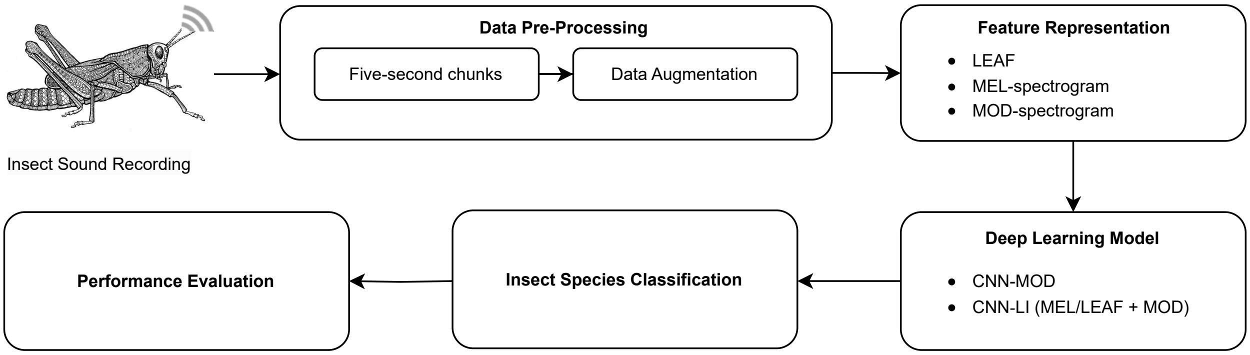

FIGURE 1 The general pipeline of the proposed deep-learning models for insect sound classification.

Faiß and Stowell (2023), He et al. (2024). Recent advances in deep learning, especially the development of convolutional neural networks (CNNs), have further refined the accuracy of insect sound classification, enabling high-precision identification of singing insect species Hibino et al. (2021).

Fusion or integration techniques involve combining multiple data sources, features, or models to enhance the performance and robustness of machine learning systems. These techniques can be broadly categorized into feature-level fusion, decision-level fusion, and model-level fusion Raghuwanshi and Kaushik (2024), He et al. (2024). Feature-level fusion or early integration (EI) combines different types of features (e.g., spectral and temporal) to create a richer representation of the data. Decision-level fusion or late integration (LI) aggregates the outputs from multiple models to make a final prediction, while model-level fusion combines the strengths of different architectures, such as CNNs and recurrent neural networks (RNNs), to improve overall accuracy. Fusion techniques have played an essential role in improving the precision of insect sound classification Travieso et al. (2021), Raghuwanshi and Kaushik (2024), He et al. (2024). For instance, integrating Mel Frequency Cepstral Coefficients (MFCC) with Linear Frequency Cepstral Coefficients (LFCC) has been shown to be beneficial in distinguishing insect species based on their unique acoustic signatures Noda et al. (2019). Furthermore, the combination of temporal and spectral features improves the classification accuracy of insect monitoring systems Raghuwanshi and Kaushik (2024), underscoring the importance of combining diverse feature sets for robust acoustic analysis.

The classification of insect species using acoustic signatures has advanced significantly, particularly through the incorporation of machine learning and deep learning methodologies. Early contributions to this field Phung et al. (2017) established a foundational framework by developing an automated system for insect detection. The system integrated low-level (LL) attributes with MFCC and employed various classifiers such as linear discriminant (LD), k-nearest neighbour (kNN), support vector machine (SVM), and Bagged Tree (BT). Their findings showed that ensemble methods like BT achieved the best detection accuracy, and combining LL and MFCC features improved classification, showcasing the benefits of integrating multiple feature sets. Building on these initial efforts, an improved insect sound recognition system was introduced that incorporates Contrast-Limited Adaptive Histogram Equalization (CLAHE) to improve the clarity of the spectrogram and reduce noise Dong et al. (2018). The system attained a significant classification accuracy of 97.87% across 47 insect species from the U.S. Department of Agriculture (USDA) library, utilizing a CNN. This study underscored the importance of advanced preprocessing techniques in the refinement of feature representations for more accurate insect sound classification. In 2019, a directed acyclic graph (DAG) approach integrated with hidden Markov models (HMMs) was proposed to classify insect species based on wingbeat sounds. This effectively captured the temporal dynamics of Mel-scaled spectrograms Ntalampiras (2019). Concurrently, a tool was developed to differentiate between crickets, katydids, and cicadas by integrating MFCC and LFCC early in the sound parametrization process Noda et al. (2019). Utilizing SVM and Random Forest (RF) classifiers, an impressive classification accuracy of 98.07% was achieved across 343 species of katydids, crickets, and cicadas, showcasing the effectiveness of combining diverse feature sets for enhanced species identification.

The integration of deep learning techniques continued to shape the field, as demonstrated in Varma et al. (2021). The authors introduced a hybrid approach that combines traditional signal processing with modern deep learning techniques for insect classification. By comparing handcrafted Mel-spectrogram features with automatic feature extraction from models like 1D-CNN and SincNet, they found that SincNet achieved the highest accuracy of 97%. In Hibino et al. (2021), the impact of various data augmentation techniques was investigated to increase the variability of the training dataset. The study demonstrated that by applying augmentations such as time shift, time stretch, and white noise to Mel-spectrograms, classification performance could be significantly improved.

A recent study Faiß and Stowell (2023) proposed a comprehensive methodology for classifying orthopteran and cicada species based on their distinctive acoustic features. The study compared traditional Mel-spectrograms as input features with Learnable Front-end for Audio Classification (LEAF), which directly processes audio waveforms. The experiments showed that LEAF outperformed Mel spectrograms by dynamically adjusting feature extraction parameters during training, resulting in more accurate species identification.

All of these studies illustrate the rapid evolution of the classification of insect species, driven by the integration of traditional signal processing techniques with state-of-the-art deep learning approaches. The continuous development of feature extraction methods, classifier models, and data augmentation strategies is crucial to advance the field, paving the way for more effective and scalable systems to classify insect species based on their acoustic signatures.

As the main contribution of this paper, we propose two improvements of the feature representation and the model architecture of an insect sound classification model. First, we introduce the modulation spectrograms as additional input features to better model species-specific recurring sound patterns. Second, we propose a late integration strategy to combine output decisions derived from Mel spectrograms or LEAF features with those derived from modulation spectrogram features. Figure1 illustrates the general processing pipeline of the proposed deep learning models for insect sound classification.

The organization of this paper is as follows. Section 2 provides an overview of the characteristics of insect sounds and their significance in classification tasks. Section 3 details the materials and methods, including datasets, data pre-processing techniques, feature representation strategies, and the proposed CNN model architectures. Section 4 presents experimental results, comparing the performance of the proposed models across multiple datasets and evaluating the contributions of different feature representations and integration strategies. Finally, Section 5 concludes the paper with a summary of the findings and a discussion of potential directions for future work.

2 | INSECT SOUNDS

Vibration and sound are inevitable products of the physiology and activity of organisms. Their ubiquitous presence in nature has led to the coupling of the evolution of the animal nervous system with the evolution of a wide array of organs to detect such stimuli. Hearing and vibration sensing are originally related to orientation within the environment, as well as toward food sources or away from predators. Using substrate- or fluid-borne vibrations to communicate is a next level of vibration utilization that evolved independently in various animal groups, such as bony fish, amphibians, birds, mammals, spiders, crustaceans and insects Greenfield (2002), Brumm (2013), Hedwig et al. (2014).

Vibrational communication is widespread in insects, known in at least 80% of the insect families Cocroft and Rodríguez (2005), possibly due to their small size. Although vibrational communication enables small animals, such as insects, to transmit low-frequency signals efficiently, their communication range is limited by one- or two-dimensional dispersal, short-range transmission (on plant tissues, typically a few centimetres to two meters), and vibrations are inevitably attenuated when the medium is altered Cocroft and Rodríguez (2005), Greenfield (2016). However, vibrations emitted in air or water have the potential to disperse equally in all directions and to great distances, depending on their wavelength and energy, or factors like the movements or density layering of the medium, the shape of the emitter, or the presence of physical barriers. Airborne sounds are thus an excellent way to perform long-range signalling. To utilize airborne sound, some insects evolved the so-called tympanal hearing organs. These organs composed of chordotonal organs (stretch receptors), a tympanum (a thin-cuticle membrane surrounded by a thick-cuticle frame), and a tracheal cavity (allowing the tympanum to resonate). Examples include Mantodea, Orthoptera, Hemiptera (Corixidae and Cicadidae), Neuroptera (Chrysopidae), Diptera (Sarcophagidae and Tachinidae), Coleoptera (Cicindellidae and Scarabaeidae), and Lepidoptera (eight superfamilies).

Intraspecific far field communication in air is further restricted in insects and is widely distributed form of signalling only in orthopterans (order Orthoptera; grasshoppers, crickets and bush-crickets) and cicadas (order Hemiptera, family Cicadidae). Acoustic signalling in long-horned orthopterans (Ensifera; currently present in crickets and bush-crickets) and cicadas likely evolved in the context of sexual communication Greenfield (2016), Song et al. (2020). In grasshoppers (Caelifera; currently present in Acrididae and Pneumoridae) tympanal hearing preceded sound production and sexual communication evolved secondarily as a result of 'sensory bias' Song et al. (2020).

Song of grasshoppers, crickets, and bush-crickets are very diverse. Grasshoppers of the family Acrididae usually sing at daytime and utilize femoro-alar (or vice versa; rubbing the femur of the hind leg against the folded forewing) stridulation with the sound resonator usually being a certain area of the forewing, with a few exceptions involving wing crepitation (rubbing the hindwings against each other). In grasshoppers of the Pneumoridae family, fundamentally different hearing and sound producing mechanisms developed, the latter involving femoro-abdominal stridulation with the resonator being the whole abdomen. Grasshopper sounds usually characterize with maximum energy at frequencies within the audible frequency range (5-10 kHz) and their hearing is tuned to specific peaks within the same range (e.g. Miller (1977)). The loudest songs are characteristic for male Pneumoridae, with their sound travelling up to 2 km distance at certain air properties Van Staaden and Römer (1997).

Bush-crickets (Tettigoniidae), being the most diverse group with intraspecific acoustic communication, show diverse song characteristics and circadian song-production activity. Most species produce non-resonant broadband songs, rarely resonant, tonal songs. Most large species have their maximum spectra in the audio range, while others show frequency bands of over 100 kHz with maxima well within the ultrasonic range. Some of the highest-frequency and loudest insect songs are produced by the neotropical bush-crickets of the genus *Supersonus*, with components of the male ultrasonic calling songs reaching 150 kHz and sound intensity exceeding 110 dB SPL at 15 cm Sarria-S et al. (2014).

Gryllids (Grylloidea) show yet another type of specific song, usually resonant with bandwidths between 2 and 8 kHz, thus, perceived as melodic by the human ear. Being small animals (up to a few centimetres in diameter) with dipole sound sources that produce long sound waves (4-20 cm), crickets must overcome certain mechanical limitations to spread the sound produced due to the effect of acoustic 'short circuit'. Male mole crickets (Gryllotalpidae) dig specific acoustic chambers in the soil, using them as resonators, while tree crickets (Oecanthidae) position their forewings in the same plane as tree leaves (or even dig a hole in the middle of a leave from where they sing), thus extending the surface of sound-producing structures. In these ways, these insects receive a gain in sound pressure of the radiated sound of 10-24 dB Bennet-Clark (1987), Pollack (2017).

Cicadas, the other group of far-field insect signallers, use a distinct type of sound production – tymbalation. Tymbals are structures of rib-like formed sclerotized cuticle located on each side of the first abdominal segment. The sound is produced when powerful muscles in the abdomen buckle the tymbals, while an air sac in the abdomen serves as a resonator. Cicadas are the loudest insect singers that produce audio sounds. The sound pressure within the air sac reaches 158 dB, while the radiated sound at 1 m of the insect can have levels of 100 dB Pollack (2017).

Insect sounds are fundamentally different from those of mammals and birds, primarily due to the distinct mechanisms insects use for sound production. Insects produce a wide range of sounds, each serving specific ecological functions, such as attracting mates, defending territories, or communicating with others of their species Bennet-Clark (1999), Rothenberg (2013), Neil and Holderied (2021). These sounds are generated through various mechanisms, each unique to a certain species in terms of frequency, rhythm, and complexity. These mechanisms create unique spectral and temporal sound structures.

With well more than 16,000 species Song et al. (2020) that communicate acoustically, and with approximately 500 new species described each year Eades et al. (2010), Cigliano et al. (2024), orthopterans outnumber all higher taxa (the classes Amphibians, Reptiles, Birds, Mammals) that communicate acoustically, as well as sound communicating fishes and other invertebrates. Cicadas add to this diversity with more than 3000 singing species. Orthopterans and cicadas, together with birds, frogs, and mammals, shape the acoustic environment in most habitats on Earth with differing seasonal and circadian contribution depending on the climate and habitat specifics. In certain environments and hours, these insects may represent the main acoustic source and thus are excellent umbrella-object for applying acoustic monitoring for studying overall insect population trends and habitat health. However, the specifics of the insect songs make their automated study challenging for us. Traditional feature extraction techniques, such as MFCCs, are often inadequate for insect sound classification due to their focus on frequencies relevant to human perception Rasmussen et al. (2024). Therefore, in order to make the large-scale monitoring of natural and anthropogenic habitats possible, we need to strive to improve the analytical tools for automatic sound recognition of far-field communicating insects.

3 | MATERIALS AND METHODS

This section provides a detailed overview of the methods and materials used in this study, focusing on the design and development of the proposed insect classification models. The section explores the datasets and data pre-processing techniques, which include the handling of varying audio file durations and the application of data augmentation techniques to increase data variability and model robustness. In addition, different feature extraction approaches are discussed, covering the LEAF, Mel spectrogram, and modulation spectrogram. Finally, several CNN architectures are introduced, focussing on the proposed CNN-MOD and CNN with Late Integration (CNN-LI) models.

3.1 | Datasets

The three datasets used in this study, namely InsectSet32 Faiß (2022), InsectSet47, and InsectSet66 Faiß (2023), represent progressive expansions in size, diversity, and sources. InsectSet32, as shown in Table 1, comprises 335 audio recordings from 32 species, totalling 57 minutes, recorded in real-world environments. The dataset combines private collections of Orthoptera recordings contributed by Baudewijn Odé and Cicadidae recordings provided by Ed Baker. Approximately 147 recordings correspond to nine Orthoptera species, while 188 are from 23 Cicadidae species. The InsectSet47 dataset, presented in Table 1, is an extension of InsectSet32, incorporating 2,266 audio files from 47 species, totalling 22 hours of recordings. These recordings were sourced from the Xeno-Canto platform Vellinga et al. (2017), collected by experts and citizen scientists from all over the world, along with two private collections: Orthoptera recordings contributed by Baudewijn Odé and Cicadidae recordings provided by Ed Baker.

InsectSet66 dataset, presented in Table 1, further expands on InsectSet47 by including recordings from five different source datasets: Orthoptera and Cicadidae datasets from iNaturalist Boone and Basille (2019), Orthoptera data from xeno-canto, and additional data from Baudewijn Odé and Ed Baker. This dataset contains 2,887 recordings from 66 species, totalling more than 24 hours of audio, with at least ten files per species. Like InsectSet32 and InsectSet47, InsectSet66 comes with a pre-defined split into training, validation, and test sets, with an equal distribution of files maintained across all subsets. These standardized splits enable reproducibility of previous experiments and facilitate direct comparability across studies. The annotations in InsectSet66 include similar details to those in InsectSet32 and InsectSet47, such as the file name, species name, class ID, a unique identifier, data subset (training, validation, or testing), observation link, and contributor name, ensuring reproducibility and direct comparability of experimental results. To maintain consistency for further neural network training, the audio files, which had varying sampling rates, were resampled to a 44.1 kHz mono WAV format.

TABLE 1 Insect Datasets—Species, class_id (cid), number of audio recordings (N), and number of five-second chunks (N_c).

(a) InsectSet32

Species	cid	N	N_c	Species	cid	N	N_c	Species	cid	N	N_c
Azanicada zulensis	0	4	28	Nemobius sylvestris	11	18	387	Platyleura plumosa	22	19	133
Brevisiana brevis	1	5	35	Oecanthus pellucens	12	14	209	Platyleura sp04	23	8	56
Chorthippus biguttulus	2	20	154	Pholidoptera griseoptera	13	15	73	Platyleura sp10	24	16	100
Chorthippus brunneus	3	13	100	Platyleura capensis	14	6	42	Platyleura sp11 cfritpennis	25	4	28
Gryllus campestris	4	22	158	Platyleura cfcatena	15	22	148	Platyleura sp12 cfritpennis	26	10	70
Kikihia muta	5	6	42	Platyleura chalyaea	16	7	47	Platyleura sp13	27	12	84
Myopsalta leona	6	7	49	Platyleura deusta	17	9	57	Pseudochorthippus paralleus	28	17	71
Myopsalta longicauda	7	4	28	Platyleura divisa	18	6	42	Pycna semiclara	29	9	63
Myopsalta mackinlayi	8	7	48	Platyleura haglundi	19	5	35	Roeseliana roeselii	30	12	37
Myopsalta melanobasis	9	5	29	Platyleura hirtipennis	20	6	36	Tettigonia viridissima	31	16	63
Myopsalta xerogracidia	10	6	42	Platyleura intercapedis	21	5	35				

(b) InsectSet47

Species	cid	N	N_c	Species	cid	N	N_c	Species	cid	N	N_c
Acheta domesticus	0	23	2647	Gomphocerus sibiricus	16	14	1234	Pholidoptera littoralis	32	13	127
Barbitistes yersini	1	14	900	Gryllus bimaculatus	17	17	1309	Platycleis albopunctata	33	15	1181
Chorthippus albomarginatus	2	11	1928	Gryllus campestris	18	38	4492	Platypleura cf catenata	34	22	831
Chorthippus apricarius	3	20	1339	Leptophyes punctatissima	19	13	1271	Platypleura plumosa	35	19	686
Chorthippus biguttulus	4	52	1355	Melanogryllus desertus	20	11	1191	Platypleura sp10	36	17	852
Chorthippus brunneus	5	32	945	Metrioptera brachyptera	21	13	973	Platypleura sp12 cf hirtipennis	37	10	361
Chorthippus mollis	6	38	1275	Myrmeleotettix maculatus	22	20	2594	Platypleura sp13	38	12	328
Chorthippus vagans	7	10	524	Nemobius sylvestris	23	28	1798	Pseudochorthippus montanus	39	12	478
Chrysochraon dispar	8	17	708	Oecanthus pellucens	24	22	1352	Pseudochorthippus parallelus	40	33	1105
Conocephalus dorsalis	9	18	1093	Omocestus petraeus	25	10	401	Roeseliana roeselii	41	33	1574
Conocephalus fuscus	10	34	2468	Omocestus rufipes	26	21	746	Stenobothrus lineatus	42	18	1550
Decticus verrucivorus	11	31	3411	Omocestus viridulus	27	25	2119	Stenobothrus stigmaticus	43	39	170
Ephippiger diurnus	12	29	1876	Phaneroptera falcata	28	20	1331	Tettigonia cantans	44	32	2687
Eupholidoptera schmidti	13	11	443	Phaneroptera nana	29	16	1414	Tettigonia viridissima	45	24	1212
Gampsocleis glabra	14	26	2625	Pholidoptera aptera	30	13	447	Tylopsis lilifolia	46	11	115
Gomphocerippus rufus	15	28	1387	Pholidoptera griseoaptera	31	21	545				

(c) InsectSet66

Species	cid	N	N_c	Species	cid	N	N_c	Species	cid	N	N_c
Acheta domesticus	0	24	2702	Eumodicogryllus bordigalensis	22	16	507	Pholidoptera littoralis	44	13	127
Aleeta curvicosta	1	23	184	Eupholidoptera schmidti	23	11	443	Platycleis albopunctata	45	15	1181
Atrapsalta collina	2	10	52	Galanga labeculata	24	43	269	Platypleura cfcatenata	46	22	831
Atrapsalta corticina	3	15	97	Gampsocleis glabra	25	27	2637	Platypleura plumosa	47	19	686
Atrapsalta encaustica	4	15	206	Gomphocerippus rufus	26	28	1387	Platypleura sp10a	48	17	852
Barbitistes yersini	5	14	900	Gomphocerus sibiricus	27	14	1234	Platypleura sp12cfhirtipennis	49	10	361
Bicolorana bicolor	6	10	421	Gryllus bimaculatus	28	20	1364	Platypleura sp13	50	12	328
Chorthippus albomarginatus	7	11	1928	Gryllus campestris	29	57	4630	Popplepsalta aeroides	51	10	79
Chorthippus apricarius	8	21	1345	Leptophyes punctatissima	30	13	1271	Popplepsalta notialis	52	14	133
Chorthippus biguttulus	9	53	1383	Melanogryllus desertus	31	11	1191	Psaltoda plaga	53	14	199
Chorthippus brunneus	10	35	985	Metrioptera brachyptera	32	14	993	Pseudochorthippus montanus	54	13	482
Chorthippus mollis	11	39	1285	Myrmeleotettix maculatus	33	21	3098	Pseudochorthippus parallelus	55	37	1123
Chorthippus vagans	12	10	524	Nemobius sylvestris	34	30	1823	Roeseliana roeselii	56	37	1612
Chrysochraon dispar	13	17	708	Neotibicen pruinosus	35	15	216	Ruspolia nitidula	57	11	599
Cicada orni	14	21	317	Oecanthus pellucens	36	29	1437	Stauroderus scalaris	58	10	978
Clinopsalta autumna	15	19	185	Omocestus petraeus	37	10	401	Stenobothrus lineatus	59	19	1633
Conocephalus dorsalis	16	18	1093	Omocestus rufipes	38	22	749	Stenobothrus stigmaticus	60	39	170
Conocephalus fuscus	17	36	2485	Omocestus viridulus	39	27	2136	Tettigonia cantans	61	37	2729
Cyclochila australasiae	18	13	78	Phaneroptera falcata	40	20	1331	Tettigonia viridissima	62	33	1300
Decticus verrucivorus	19	34	3581	Phaneroptera nana	41	18	1458	Tylopsis lilifolia	63	11	115
Diceroprocta eugraphica	20	11	239	Pholidoptera aptera	42	16	465	Yoyetta celis	64	152	346
Ephippiger diurnus	21	31	1189	Pholidoptera griseoaptera	43	27	651	Yoyetta repetens	65	40	223

3.2 | Data Pre-processing

In order to create a feature extraction method capable of generating meaningful data for identification and classification tasks, a preprocessing phase was implemented.

3.2.1 | Handling Different Audio Lengths

All audio recordings were standardized into five-second chunks to retain species-specific rhythmic patterns in insect calls. This duration was chosen for two key reasons: first, it aligns with the typical repetition

TABLE 2 Datasets Summary

Dataset	No. of species	No. of audio recordings (N)	No. of five-second chunks (N_c)			
			Train	Validation	Test	Total
InsectSet32	32	335	1,548	379	602	2,529
InsectSet47	47	2,266	39,600	11,937	9,861	61,398
InsectSet66	66	2,887	43,801	13,369	11,195	68,365

intervals of insect vocalizations (e.g., pulse rates, chirp sequences), ensuring that critical acoustic features are captured within a single chunk; second, it balances practicality, as most recordings met or exceeded this length, while enabling shorter recordings to be extended through looping without disrupting temporal patterns. Shorter recordings were looped until they reached the required length, while longer recordings were split into overlapping five-second chunks with a 3.75-second overlap. If the remaining portion of a recording at the end of a splitting window was at least 1.25 seconds, the chunk was completed by wrapping around to the beginning of the file. This pre-processing technique was uniformly applied to all three datasets. The number of five-second chunks of each species is represented by N_c as shown in Table 1. The total number of five-second chunks generated for each dataset is detailed in Table 2. The assignment of chunks to training, validation, or test sets follows the pre-defined annotations of the original files (i.e., before splitting into chunks). For example, if a file was annotated for training, all resulting chunks from that file are allocated to the training set. The distribution of chunks across training, validation, and test sets is also summarized in Table 2.

3.2.2 | Data Augmentation

Insect sounds vary significantly according to factors such as species, environment, behaviour, and recording conditions Romer and Lewald (1992). The challenges of limited available samples, environmental noise, and inconsistencies in recording equipment further complicate the modeling of acoustic data. To address these issues, data augmentation approaches that generate modified versions of the original audio samples are essential. Such approaches, for instance, simulate diverse acoustic recording environments while preserving the core features of insect sound recordings.

After standardizing the audio data to five-second segments, two augmentation techniques were applied as implemented in the Audiomentations Python library[‡]: *AddColoredNoise* and *ApplyImpulseResponse*. These techniques introduce noise and simulate room-acoustic variations, enabling the training of more robust models by exposing them to a wider range of acoustic conditions. Noise was added to the original audio to enhance the model's ability to handle interference. Subsequently, impulse responses (IRs) from the Open Acoustic Impulse Response (OpenAIR) dataset Shelley and Murphy (2010) were applied to simulate real-world acoustic effects. The IRs, which include high-sample-rate recordings captured in various outdoor environments, were carefully selected to ensure the model could generalize across diverse acoustic conditions. The IRs were chosen containing high-sample-rate recordings made in various locations Shelley and Murphy (2010). Specifically, 12 IRs from three different outdoor environments (two forest areas and one campus) were randomly applied during the enhancement with a probability of 70%. The IR-processed audio was then mixed with the original files in random ratios, creating additional variability in the impact of these effects.

3.3 | Features Representation

In this paper, the performance of the insect sound classification system is evaluated using three distinct feature representations: LEAF, Mel spectrogram, and modulation spectrogram. These front-end representations serve as input features for CNN models, capturing various spectral and temporal characteristics of insect sounds and will be detailed in the following sections.

3.3.1 | LEarnable Audio Front-End

LEAF is a neural network-based method for extracting features from audio, designed to replace traditional handcrafted audio feature representations. LEAF builds on the conventional steps to extract features using Mel filter-banks but adds trainable components to improve representation for specific tasks Zeghidour et al. (2021). The traditional Mel filter-bank process involves three key steps. First, the audio is split into short fixed-duration segments to capture its time-variability. Next, these segments pass through a bank of fixed-frequency filters that imitate human hearing, focusing more on low frequencies where humans are more sensitive to changes in pitch and loudness. Finally, the signal is compressed to reflect the human ear's logarithmic sensitivity to loudness, meaning the power of a sound must double to be perceived as 3 decibels louder.

LEAF follows this basic structure but replaces the fixed steps (filtering, windowing, and compression) with trainable components. In LEAF, the filtering layer is initialized with Gabor filters, the windowing function is set for optimization, and the compression step is adapted during training. This allows LEAF to produce a time-frequency representation similar to Mel filter-banks, but with parameters that can be trained

[‡] https://github.com/iver56/audiomentations

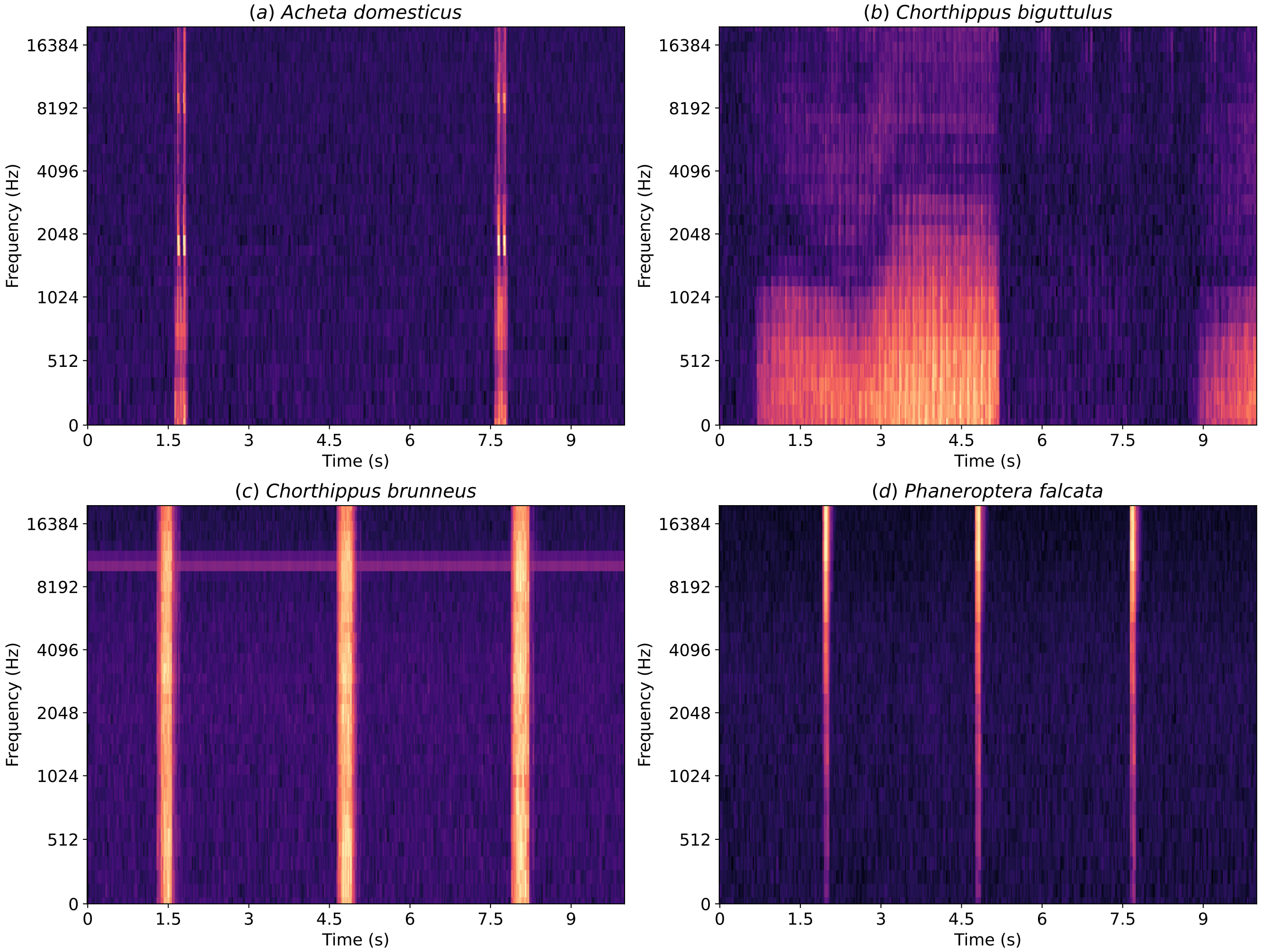

FIGURE 2 The Mel spectrograms of four different types of insect species: (a) *Acheta domesticus*, (b) *Chorthippus biguttulus*, (c) *Chorthippus brunneus*, and (d) *Phaneroptera falcata*.

to suit the specific classification task. Unlike Mel filter-banks, which employ a predefined, fixed frequency scale (e.g., the Mel scale), LEAF dynamically learns the optimal frequency scale for each task based on the data, adjusting the filter-bank parameters (e.g., Center frequencies and bandwidths) to better match the acoustic characteristics of the target domain.

LEAF has shown strong results in various audio classification tasks such as speech recognition, speaker identification, music classification, acoustic scene analysis, and bird song detection Zeghidour et al. (2021), Faiß and Stowell (2023), Song et al. (2024). Its trainable nature makes it particularly useful for tasks where traditional methods underperform, such as insect sound classification Faiß and Stowell (2023). In these cases, LEAF's ability to adjust filter frequencies, bandwidths, normalize data, and pool features during training helps it identify unique patterns in insect sounds, which differ from audio recordings of human-produced audio, such as speech or music. For example, insect sounds tend to be higher-pitched and have a different structure than human sounds, requiring a method that focuses more on high-frequency details and less on low frequencies. LEAF achieves this by adapting the filter settings and pooling techniques to focus on the most important frequency bands. This makes LEAF a useful tool for analyzing audio data where fixed spectral methods might miss relevant details.

Although LEAF has been tested on human-related tasks like speech and music recognition, its potential for bioacoustic tasks like insect classification is still largely unexplored. Insect sounds have distinct spectral characteristics that differ from human-focused audio features. By allowing LEAF to learn specific features for each task, it can provide better spectral resolution for important frequencies, improve feature extraction, and enhance classification performance when working with insect sound datasets.

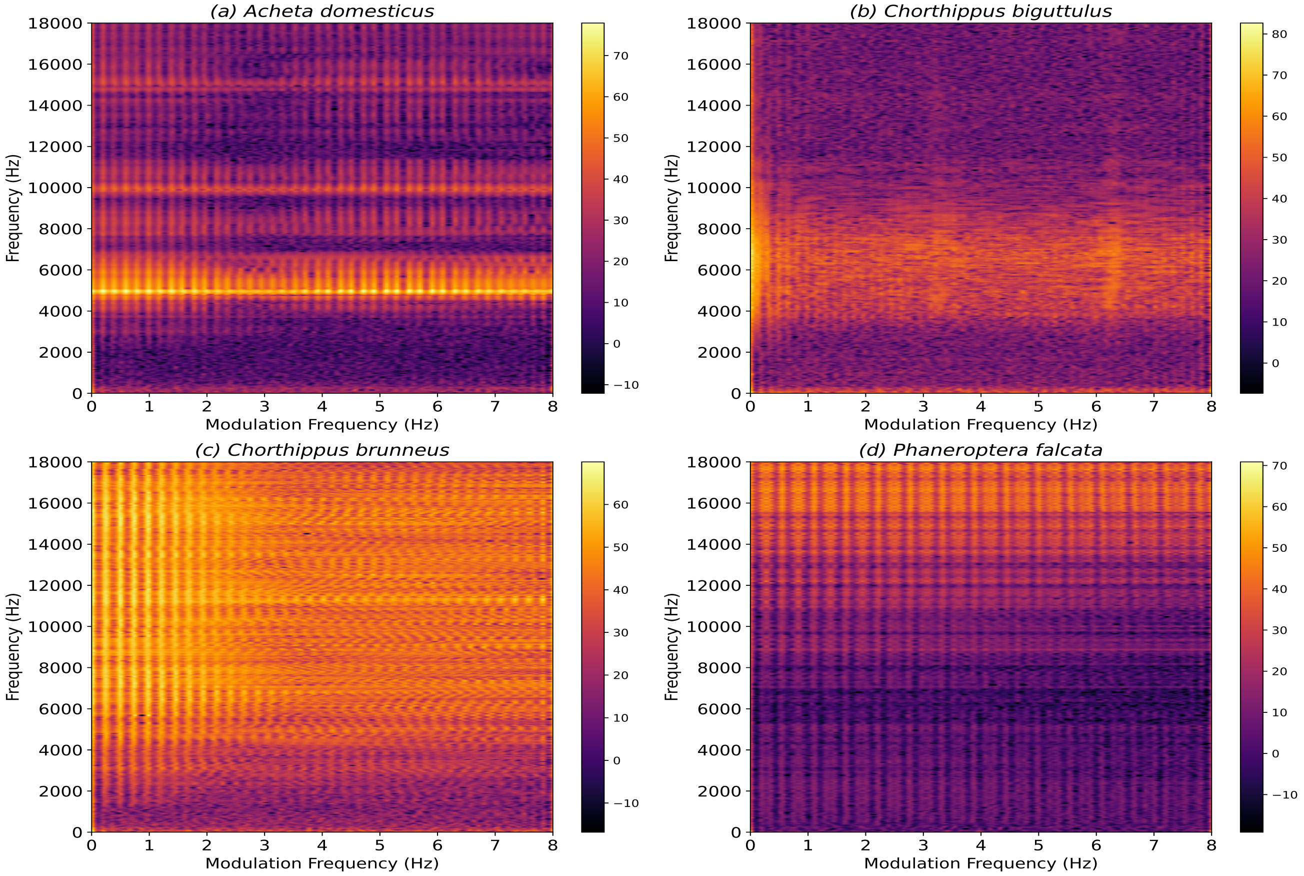

FIGURE 3 The modulation spectrograms of four different types of insect species: (a) *Acheta domesticus*, (b) *Chorthippus biguttulus*, (c) *Chorthippus brunneus*, and (d) *Phaneroptera falcata*.

3.3.2 | Mel Spectrogram

Insect sounds often span a wide range of frequencies and can vary greatly due to environmental conditions, behaviour, and recording equipment. To effectively represent these signals for classification, the Mel scale is used as a non-linear frequency scale. It transforms the physical frequency (*f*) in Hertz to Mel frequency (*Mel*) He et al. (2024). This scale is particularly useful for emphasizing lower frequencies while compressing the higher-frequency range, which often contains less significant variations.

To map spectral information onto the Mel scale, a set of Mel filters is applied. These filters are evenly spaced on the Mel scale. The Mel spectrum is generated by computing the energy within each filter's frequency band. First, the input signal is converted into the frequency domain using the short-time Fourier transform (STFT) He et al. (2024). This transformation decomposes the signal into frequency components in short time windows, allowing the extraction of temporal and spectral characteristics Noda et al. (2019).

The resulting Mel spectrogram is a visual representation of the frequency and time characteristics of the sound. By applying a logarithmic transformation, we generate a logarithmic Mel spectrogram, which compresses the dynamic range and highlights essential features of the sound. This makes the data more suitable for training deep learning models Raghuwanshi and Kaushik (2024), Albouy et al. (2024), Faiß and Stowell (2023). The Mel spectrogram effectively captures both temporal and spectral features of the audio. It provides a compact and meaningful representation of the sound, making it easier for machine learning models to distinguish between different insect species Gandini (2022).

Figure 2 presents the mel spectrograms of four insect species: (a) *Acheta domesticus*, (b) *Chorthippus biguttulus*, (c) *Chorthippus brumeus*, and (d) *Phaneroptera falcata*. The vertical axis represents acoustic frequency (Hz), and the horizontal axis denotes time (s). A general observation is that the distribution of acoustic energy varies across species. For instance, some species exhibit energy concentrated in lower frequency bands with steady temporal patterns like *Acheta domesticus*, while others display broader frequency coverage, extending to higher ranges and more intermittent or fluctuating energy over time such as *Phaneroptera falcata*. These differences may reflect distinct acoustic behaviours, with species like *Acheta domesticus* relying on sustained low-frequency signals for communication, whereas *Phaneroptera falcata* employs dynamic, high-frequency calls that vary temporally.

3.3.3 | Modulation Spectrogram

Insects produce signals with rapid temporal changes and fine spectral details, particularly in high-frequency ranges. Modulation spectrograms provide a more nuanced representation by capturing spectro-temporal modulations in the acoustic signal Albouy et al. (2024), Bakhtyari et al. (2022). This is particularly important for distinguishing between species with overlapping frequency ranges or subtle differences in their calls. Modulation spectrograms can highlight the harmonic structures, amplitude modulations, and frequency sweeps common to insect sounds, offering a robust feature set for training machine learning models.

A modulation spectrogram represents the time-varying spectral content of a signal, capturing the dynamics of both amplitude and frequency modulations Elliott and Theunissen (2009). It mimics the functionality of the human ear, capturing repetitive patterns in audio signals across different frequency bands. By applying an FFT across time for each frequency band of the Mel spectrogram, we obtain the modulation spectrogram.

Unlike traditional spectrograms that display power spectral density over time, modulation spectrograms depict how spectral features evolve, emphasizing patterns and structures in the signal modulation domain. This approach is particularly valuable for the analysis of complex auditory signals, such as speech, music, and animal vocalizations Albouy et al. (2024).

Modulation spectrograms have been extensively used in the field of speech processing. Elliott and Theunissen (2009) demonstrated the criticality of low modulation frequencies in both time and frequency for speech comprehension. Their study highlighted that temporal modulations from 1 to 7 Hz and spectral modulations below 1 cycle/kHz are vital for intelligibility, underscoring the importance of modulation spectrograms in preserving essential speech features during compression and noise reduction processes.

The study presented by Albouy et al. (2024) demonstrates that spectro-temporal modulation (STM) features extracted from the modulation spectrogram provide a robust basis for differentiating between speech and song across a diverse range of cultures and languages. The findings show that speech typically exhibits higher temporal modulation rates, align with the fast syllable rates and varied articulation necessary for conveying detailed linguistic information. In contrast, song features are characterized by slower temporal modulations but richer spectral variations, reflecting the sustained pitches and harmonic complexity essential for musical expression. Using these differences, the modulation spectrogram provides a robust framework for accurately classifying vocalizations, transcending cultural and linguistic boundaries, and aligning with the neural processing mechanisms that humans naturally employ.

The study presented by Gallardo-Antolín and Montero (2021) explored the development and assessment of various machine learning models to classify the intelligibility of audio recordings. The research implemented several systems, including those based on SVM and Long Short-Term Memory (LSTM) networks, utilizing log-Mel spectrograms and modulation spectrograms features. The results indicated that the LSTM-based systems significantly outperform the SVM-based systems, emphasizing the importance of accurately modelling the temporal dynamics of the acoustic characteristics. The fusion of log-Mel and modulation spectrogram features further enhances performance, with attention mechanisms proving especially beneficial. These findings suggest that modulation spectrograms carry vital complementary information for classification tasks, thus improving the overall accuracy and robustness of the system.

To compute the modulation spectrogram features, as presented in Gallardo-Antolín and Montero (2021), Singh et al. (2023), the process begins computing the Mel spectrogramS_{mel} from the input waveform x by first computing its Short-Time Fourier Transform (STFT) and then applying a triangular filter-bank with centre frequencies along the Mel frequency scale. The Mel spectrogram is further compressed to decibel scale as

$$S_{\text{db}} = 20 \cdot \log_{10}(S_{\text{mel}}) \tag{1}$$

The modulation spectrogram S_{mod} is computed by means of a Fast Fourier Transform (FFT) across time to detect periodic (modulating) components in S_{db} as

$$S_{\text{mod}}(\tau, k) = \sum_{n=-\infty}^{\infty} S_{\text{db}}(n, k) e^{-j\frac{2\pi}{N}\tau n} \tag{2}$$

with τ denoting the modulation frequency index, k denoting the frequency index in S_{db}, n denoting the frame index in S_{db}, and N denoting the total number of frames in S_{db}.

Figure 3 presents the modulation spectrograms of four insect species: (a) *Acheta domesticus*, (b) *Chorthippus biguttulus*, (c) *Chorthippus brunneus*, and (d) *Phaneroptera falcata*. In all cases, the horizontal and vertical axes represent, respectively, the modulation frequency (Hz) and acoustic frequency (Hz). It can be observed that the modulation spectrograms of these species exhibit distinct patterns, reflecting differences in their acoustic communication signals. For instance, *Acheta domesticus* shows a broad distribution of modulation frequencies over a wide range of acoustic frequencies, suggesting a complex temporal structure. In contrast, *Chorthippus biguttulus* and *Chorthippus brunneus* display more concentrated modulation frequencies, particularly at lower acoustic frequencies, indicating a more rhythmic and repetitive signal pattern. Finally, *Phaneroptera falcata* exhibits a unique pattern with distinct peaks at specific modulation frequencies, suggesting a more structured and periodic signal, possibly used for species-specific communication.

3.4 | CNN Model Architectures

The CNN architectures presented in this study consist of two primary components: a front-end for audio feature representation and a back-end for classification. The design of the front-end varies depending on the input representation (Mel spectrogram, modulation spectrogram, or LEAF features), while the back-end maintains a consistent convolutional structure adapted to process these representations. The following sub-subsections detail the configuration of these components, their integration, and the training methodology.

3.4.1 | Network Front-End

Three neural network front-ends are compared in our experiments: the conventional Mel spectrogram, the modulation spectrogram, and the adaptive waveform-based LEAF front-end. The Mel spectrograms and modulation spectrograms were generated through transformations applied to the audio signal before being used as input for the convolutional network. The Mel spectrograms are extracted using the *MelSpectrogram* module from the *torchaudio* Python package then converted to a decibel scale. In contrast, the modulation spectrogram features were subsequently computed through additional processing. The FFT size was set to match the number of frequency bins in the Mel spectrogram, while the hop length was set to one-third of the Mel spectrogram's own hop length parameter to increase temporal resolution, with the window length set to twice the hop length to maintain consistency. A Hann window was applied during STFT computation to reduce spectral leakage. The modulation spectrogram was then obtained by applying FFT across the time axis of the decibel-scaled Mel spectrogram as detailed in Equation (2).

In contrast, when the LEAF front-end was used, the full waveforms were directly processed in the network's front-end. This approach allows parameters such as filter frequency and bandwidth, per-channel compression and normalization, and low-pass pooling to be trainable by leveraging gradient descent learning. To ensure a fair comparison, the initialization parameters for all front-ends were aligned as closely as possible. The audio files were sampled at 44.1 kHz, with a mono-channel input shape of [1, 220500] (one channel, five seconds), which the front-ends transformed into a representation shape of [1, 64, 1500]. This representation included 64 filter bands along the frequency axis and 1500 time steps. The hop length for the Mel spectrograms was set at approximately 3.335 ms, with a corresponding window size of 6.67 ms, aligning with the LEAF settings. The LEAF filter bank was initialized on a scale similar to the Mel front-end, spanning the frequency range of 0 to 22.05 kHz. The inputs were processed in batches of 14 and passed through the network for training and evaluation. This pipeline ensured that the extracted features were consistent and allowed for a robust evaluation of the model's performance using different front-ends.

3.4.2 | Network Back-End

The network's back-end was adapted from a CNN architecture proposed by Faiß and Stowell (2023). It consists of four convolutional layers, with the option to extend to five or six layers, depending on the configuration. Each convolutional layer is followed by a rectified linear unit (ReLU) activation function and batch normalization. To ensure a fixed output size regardless of the input dimensions, adaptive average pooling was applied to the feature maps after the convolutional layers. Unlike regular average pooling, which uses a fixed-size pooling window and produces output sizes dependent on the input dimensions, adaptive average pooling dynamically adjusts the pooling window size to achieve a user-specified output size. This flexibility is particularly advantageous when dealing with variable input sizes, as it ensures consistent feature map dimensions for downstream layers.

The pooled output is flattened and fed into a linear layer to generate logits for each class in the dataset. A softmax activation function is then applied to these logits during inference to convert them into class probabilities. The final predicted class for each training example is determined by selecting the highest prediction value. To prevent over-fitting on the limited training dataset, dropout is applied to the final linear layer (initially set at a rate of 0.4), alongside L2 regularization of the weights (with a weight decay of 0.001). For InsectSet47 and InsectSet66, the dropout rate is subsequently lowered to 0.23 to address initial under-fitting caused by the heightened complexity of the data. For additional experimentation, a fifth or sixth convolutional layer is optionally added to the model to test its impact on performance. In general, the main four-layer model contains 28,319 trainable parameters that are adjusted during the training phase, with the inclusion of the LEAF front-end.

3.4.3 | Training Procedure

During training, early stopping was used to monitor the network's performance by evaluating it on the validation set after each epoch. The validation loss was tracked to predict how well the model would perform in the validation set and training was stopped if no improvement was observed for eight consecutive epochs. The dataset was split into training, validation, and test sets using the same partitioning strategy as described in the work by Faiß and Stowell (2023). Performance was assessed using metrics such as accuracy, F1 score, precision, and recall. To ensure consistency, each model was run three times in the InsectSet32, InsectSet47, and InsectSet66 datasets. Due to randomness in training, results varied between runs, but the best performing runs of InsectSet47 and InsectSet66 were further tested by incorporating an additional fifth or sixth convolutional layer to examine how increased model complexity affects classification performance.

3.4.4 | CNN Model Variants

This study introduces two convolutional neural network architectures for audio classification: the CNN-MOD and the CNN with Late Integration (CNN-LI) (MEL/LEAF + MOD) models, as illustrated in Figure 4.

CNN-MOD Model: This model is a CNN specifically designed to utilize modulation spectrogram (MOD) features for audio classification. This architecture is particularly advantageous for tasks requiring detailed temporal modulation analysis, as MOD features emphasize subtle spectral and temporal variations in audio signals. By focusing exclusively on these features, the CNN-MOD model effectively captures the fine-grained temporal patterns necessary for accurate classification.

CNN-LI (MEL/LEAF + MOD) Model: This model employs a dual-branch convolutional architecture with late integration, leveraging the complementary strengths of different audio features. It processes two types of input features: Mel spectrogram or LEAF features in one branch (first front-end), and modulation spectrogram (MOD) features in the other (second front-end). Each branch follows the general architectural framework described in this section. During the late integration phase, the outputs of these branches are concatenated along the channel dimension to form a unified feature representation. A final dense layer processes the concatenated features to predict class probabilities, with dropout applied at this stage to enhance generalization.

This architecture is uniquely suited for tasks where combining complementary feature representations yields better performance. The MEL or LEAF features provide a comprehensive spectral overview, capturing overall frequency distributions and energy patterns. In contrast, MOD features focus on fine-grained temporal modulations, highlighting dynamic variations critical for distinguishing similar classes. This model offers several advantages:

- **Complementary fusion**: Late integration combines the strengths of spectral (MEL) and temporal (MOD) features. By allowing each branch to independently learn optimized feature representations.
- **Enhanced robustness**: Using two different feature sets, the model becomes more resilient to variability in the data set. For example, when noise obscures one type of feature, the other branch can provide compensatory information.
- **Task-specific flexibility**: The modular architecture allows for straightforward replacement or tuning of one front-end without disrupting the other. For instance, the trainable LEAF front-end can be fine-tuned for specific tasks, while the MOD branch maintains consistent performance.
- **Improved generalization**: Late integration provides a richer feature representation, which enhances the model's ability to generalize to unseen data He et al. (2024). This is particularly beneficial for datasets with significant inter-class variability or classes with similar characteristics.
- **Scalability**: The architecture can be easily extended to incorporate additional front-ends or modified for different audio classification tasks by adjusting the branches or the final integration layer.

4 | EXPERIMENTAL RESULTS

This section reports the experimental results of the classification performance of the CNN-MOD and CNN-LI (MEL/LEAF + MOD) models across datasets comprising 32, 47, and 66 insect species.

4.1 | CNN-MOD Model Results

The CNN-MOD model employs the modulation spectrogram (MOD) as its front-end and is evaluated with configurations of four to six convolutional layers on both the test and validation sets. The experimental results are presented in Table 3.

Using six convolutional layers, the CNN-MOD model achieved its best classification performance in all datasets. For InsectSet32, the highest classification accuracy was 61%. For InsectSet47, the model achieved an accuracy of 76%, while for InsectSet66, the best accuracy reached was 77%. The results demonstrate that the model generalizes better with larger datasets and benefits from deeper architectures. Increasing the number of convolutional layers from four to six consistently improved performance across all datasets, improving classification accuracy.

The confusion matrix shown in Figure 5 illustrates the classification results of the CNN-MOD model on a dataset consisting of 47 insect classes (labeled 0 to 46). A notable trend is the misclassification of all test samples from class_id 37 (*Platypleura sp12 cf hirtipennis*) and most samples from class_id 38 (*Platypleura sp13*) as class_id 35 (*Platypleura plumosa*). This can be attributed to the shared physiological features, as they all belong to the genus *Platypleura* and share similar acoustic characteristics. Additionally, most test samples of class_id 32 (*Pholidoptera littoralis*) are misclassified as class_id 44 (*Tettigonia cantans*). This misclassification likely results from insufficient training data for class_id 32, which challenges the model's ability to distinguish between these two classes effectively.

4.2 | CNN-LI (MEL/LEAF + MOD) Model Results

The CNN-LI (MEL/LEAF + MOD) model employs a dual front-end approach, with the Mel spectrogram (MEL) as first front-end and the modulation spectrogram (MOD) as second front-end, and its performance was assessed using four to six convolutional layers. The experimental results are presented in Table 4. The results demonstrated that the model consistently achieved its best classification performance when learning from both MEL and MOD spectrograms, highlighting the effectiveness of combining these complementary feature representations.

For the InsectSet47 dataset, the model achieved its highest classification accuracy of 88% when using five convolutional layers and MEL-MOD features. The performance superiority in this dataset underscores the advantages of integrating spectral and temporal features extracted from MEL and MOD, respectively. The robustness of the proposed model was further validated by comparing with the results reported by Faiß and Stowell (2023), which used the MEL or LEAF characteristics under identical conditions. In all data sets and configurations, the proposed model outperformed the reference system Faiß and Stowell (2023). For instance, on the InsectSet47 dataset with 5 convolutional layers, the proposed model achieved an accuracy of 88% with MEL-MOD features, compared to 85% and 86% achieved by Faiß and Stowell (2023) using MEL and LEAF features, respectively. Similarly, on the InsectSet66 dataset, the proposed model achieved an accuracy of 84% using the MEL-MOD features, outperforming the results of 82% (MEL) and 83% (LEAF) of the Faiß and Stowell (2023) system.

The superior performance of the proposed model can be attributed to its ability to effectively learn from the complementary information

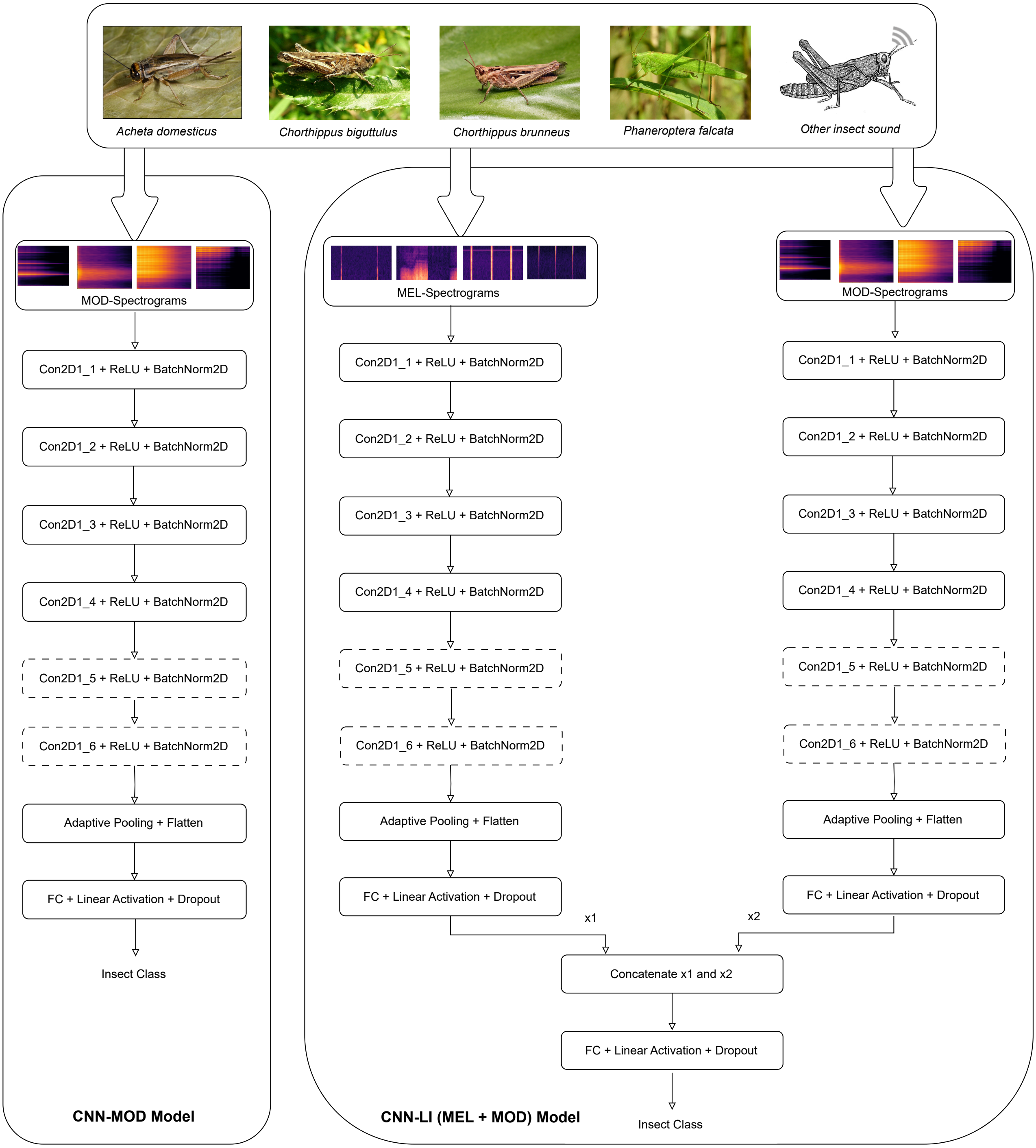
Acheta domesticus
Chorthippus biguttulus
Chorthippus brunneus
Phaneroptera falcata
Other insect sound
MOD-Spectrograms
Con2D1_1 + ReLU + BatchNorm2D
Con2D1_2 + ReLU + BatchNorm2D
Con2D1_3 + ReLU + BatchNorm2D
Con2D1_4 + ReLU + BatchNorm2D
Con2D1_5 + ReLU + BatchNorm2D
Con2D1_6 + ReLU + BatchNorm2D
Adaptive Pooling + Flatten
FC + Linear Activation + Dropout
Insect Class
CNN-MOD Model
MEL-Spectrograms
Con2D1_1 + ReLU + BatchNorm2D
Con2D1_2 + ReLU + BatchNorm2D
Con2D1_3 + ReLU + BatchNorm2D
Con2D1_4 + ReLU + BatchNorm2D
Con2D1_5 + ReLU + BatchNorm2D
Con2D1_6 + ReLU + BatchNorm2D
Adaptive Pooling + Flatten
FC + Linear Activation + Dropout
x1
MOD-Spectrograms
Con2D1_1 + ReLU + BatchNorm2D
Con2D1_2 + ReLU + BatchNorm2D
Con2D1_3 + ReLU + BatchNorm2D
Con2D1_4 + ReLU + BatchNorm2D
Con2D1_5 + ReLU + BatchNorm2D
Con2D1_6 + ReLU + BatchNorm2D
Adaptive Pooling + Flatten
FC + Linear Activation + Dropout
x2
Concatenate x1 and x2
FC + Linear Activation + Dropout
Insect Class
CNN-LI (MEL + MOD) Model

FIGURE 4 The proposed CNN-MOD and CNN-LI (MEL + MOD) models pipeline.

provided by MEL and MOD spectrograms. By integrating these two feature representations, the model gains a more comprehensive understanding of the acoustic data, enabling it to better distinguish between insect species with overlapping acoustic characteristics.

TABLE 3 CNN-MOD Model Results.

Dataset	Layers	Test				Validation	
		Acc.	F1 Score	Recall	Precision	Acc.	Loss
InsectSet32	4	0.55	0.44	0.43	0.48	0.53	1.51
	5	0.57	0.46	0.47	0.50	0.54	1.43
	6	0.61	0.45	0.50	0.53	0.54	2.03
InsectSet47	4	0.70	0.54	0.55	0.62	0.66	1.37
	5	0.72	0.61	0.62	0.67	0.70	1.10
	6	**0.76**	**0.67**	**0.67**	**0.73**	**0.72**	**1.13**
InsectSet66	4	0.69	0.59	0.60	0.64	0.67	1.32
	5	0.70	0.64	0.65	0.67	0.69	1.22
	6	**0.77**	**0.67**	**0.66**	**0.75**	**0.71**	**1.20**

In conclusion, the results clearly demonstrate that the proposed CNN-LI model with MEL-MOD features and deeper architectures, specifically with 5 convolutional layers, consistently outperforms existing methods, including the Faiß and Stowell (2023) system. The combination of spectral and temporal features allows the model to achieve superior accuracy, recall, and precision in classifying insect species, establishing its effectiveness as a robust tool for bioacoustic classification tasks.

The confusion matrix illustrated in Figure 6 shows the performance of the proposed CNN-LI (MEL/LEAF + MOD) model and reflects the overall performance since it shows a clearer diagonal for accurate classifications, with less incorrect classifications around it. A notable observation is the consistent misclassification of all test samples annotated as Platypleurasp13 and Platypleuraplumosa. Faiß and Stowell (2023) discuss that the LEAF front-end they tested 'did not fine-tune its parameters to distinguish between specific sound characteristics of closely related species'. This is particularly true for the largest dataset of samples from the genus Platypleura, which also produced misclassification outputs in our analyses. The songs of Platypleura in the data set show high similarities in frequency spectra (peaks between 7 and 10 kHz) and the amplitude-temporal pattern. Moreover, some samples have not been identified at the species level, which in addition to song similarity brings the possibility that some samples belong to forms or populations of the same species (see also Faiß and Stowell (2023)). In fact, published data reveal very similar patterns in the song of sympatric species of Platypleura. Although differences in the peak frequency or pulse repetition rate have been reported Sanborn (2003), Tatsuta et al. (2017), Matsuo (2023), these differences are small enough to be masked by the frequency bands of the songs, distance to the recorded object, or by temperature differences of the animal body reflecting the pulse repetition rate. Such similarities probably caused the model to confuse the species-specific characteristics of the sampled sound, leading to systematic errors. Such genus-level overlap in acoustic patterns may require additional distinctive features or refined pre-processing to improve differentiation.

5 | CONCLUSION

This study presents notable advances in insect sound classification using modulation spectrograms and a late integration strategy. Initially, we used modulation spectrograms as input features, which effectively captured repetitive patterns inherent in insect sounds, resulting in a classification accuracy of 77%. This demonstrated the strong predictive power of the CNN model when utilizing these features. Subsequently, we explored the late integration strategy, which merges output decisions from traditional features, such as Mel spectrograms and LEAF, with modulation spectrograms. The experimental results revealed that the model achieved its highest classification accuracy of 88% when trained with MEL and MOD spectrograms, outperforming existing models in the literature and demonstrating superior robustness and performance across the same datasets and in the same experimental settings.

Our findings were validated on multiple datasets, including InsectSet32, InsectSet47, and InsectSet66, confirming the generalizability of our proposed method with minimal dependency on specific datasets. Notably, the InsectSet47 and InsectSet66 datasets, which included a greater number of species and audio samples, showed improved classification performance compared to InsectSet32. This enhancement can be attributed to the increased number and length of audio samples, which allowed the models to generalize more effectively on previously unseen data.

Looking ahead, we anticipate that further performance improvements can be achieved by modifying the model architecture, such as adding additional convolutional layers. In addition, future research should explore multimodal studies that integrate both audio and visual data with biological theories and ecological concepts. This approach has the potential to provide deeper insight into animal behaviours and ecosystem interactions, advancing our understanding of these complex systems.

AUTHOR CONTRIBUTIONS

Funding acquisition: Hanna Lukashevich, Jakob Abeßer, Dragan Chobanov

Conceptualization: Ali Saudi, Jakob Abeßer, Dragan Chobanov

TABLE 4 Comparison of the Proposed CNN-LI (MEL/LEAF + MOD) Model and Faiß and Stowell (2023) Model Results.

Dataset	Proposed Model Results						Faiß and Stowell (2023) Model Results					
	Features	Layers	Test				Features	Layers	Test			
			Acc.	F1 Score	Recall	Precision			Acc.	F1 Score	Recall	Precision
InsectSet32	MEL-MOD	4	0.65	0.55	0.55	0.53	MEL	4	0.62	0.52	0.53	0.61
	LEAF-MOD	4	0.77	0.69	0.69	0.71	LEAF	4	0.76	0.66	0.68	0.70
InsectSet47	MEL-MOD	4	0.83	0.70	0.70	0.78	MEL	4	0.77	0.66	0.66	0.69
	LEAF-MOD	4	0.82	0.76	0.77	0.75	LEAF	4	0.81	0.71	0.72	0.77
	MEL-MOD	**5**	**0.88**	**0.80**	**0.80**	**0.82**	**MEL**	**5**	**0.85**	**0.78**	**0.79**	**0.81**
	LEAF-MOD	5	0.86	0.82	0.82	0.78	LEAF	5	0.86	0.81	0.81	0.85
	MEL-MOD	6	0.87	0.79	0.79	0.81						
	LEAF-MOD	6	0.86	0.85	0.85	0.80						
InsectSet66	MEL-MOD	4	0.80	0.70	0.71	0.76	MEL	4	0.78	0.66	0.66	0.73
	LEAF-MOD	4	0.81	0.75	0.75	0.74	MEL	4	0.80	0.68	0.68	0.77
	MEL-MOD	5	0.84	0.75	0.74	0.81	MEL	5	0.82	0.74	0.74	0.80
	LEAF-MOD	5	0.83	0.79	0.78	0.80	LEAF	5	0.83	0.76	0.77	0.81
	MEL-MOD	6	0.85	0.78	0.77	0.82						
	LEAF-MOD	6	0.84	0.80	0.79	0.81						

Project administration: Hanna Lukashevich

Supervision: Jakob Abeßer

Writing: Ali Saudi, Dragan Chobanov, Jakob Abeßer, Hanna Lukashevich (**review & editing**).

All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

This work was supported by funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101081964.

REFERENCES

Albouy, P., Mehr, S.A., Hoyer, R.S., Ginzburg, J., Du, Y. & Zatorre, R.J. (2024) Spectro-temporal acoustical markers differentiate speech from song across cultures. *Nature Communications*, 15(1), 4835.

Bakhtyari, M., Davoudi, S. & Mirzaei, S. Evaluating various feature extraction methods and classification algorithms for music genres classification. In: *2022 27th international computer conference, computer society of Iran (CSICC). IEEE, 2022*, pp. 1–6.

Belluco, S., Bertola, M., Montarsi, F., Di Martino, G., Granato, A., Stella, R. et al. (2023) Insects and public health: an overview. *Insects*, 14(3), 240.

Bennet-Clark, H. (1987) The tuned singing burrow of mole crickets. *Journal of Experimental Biology*, 128(1), 383–409.

Bennet-Clark, H.C. (1998) How cicadas make their noise. *Scientific American*, 278(5), 58–61.

Bennet-Clark, H.C. (1999) Resonators in insect sound production: how insects produce loud pure-tone songs. *Journal of experimental Biology*, 202(23), 3347–3357.

Boone, M.E. & Basille, M. (2019) Using inaturalist to contribute your nature observations to science: Wec413/uw458, 6/2019. *Edis*, 2019(4), 5–5.

Branding, J., von Hörsten, D., Böckmann, E., Wegener, J.K. & Hartung, E. (2024) Insectsound1000 an insect sound dataset for deep learning based acoustic insect recognition. *Scientific Data*, 11(1), 475.

Brumm, H. (2013) *Animal communication and noise*. Vol. 2. : Springer.

Bánki, O., Roskov, Y., Döring, M., Ower, G., Hernández Robles, D., Plata Corredor, C. et al. (2024) *Catalogue of life*. URL https://www.checklistbank.org/dataset/299029

Cardoso, P., Barton, P.S., Birkhofer, K., Chichorro, F., Deacon, C., Fartmann, T. et al. (2020) Scientists' warning to humanity on insect extinctions. *Biological conservation*, 242, 108426.

Cigliano, M.M., Braun, H., Eades, D.C. & Otte, D. (2024) Orthoptera species file,.

Cocroft, R.B. & Rodríguez, R.L. (2005) The behavioral ecology of insect vibrational communication. *Bioscience*, 55(4), 323–334.

Dong, X., Yan, N. & Wei, Y. Insect sound recognition based on convolutional neural network. In: *2018 IEEE 3rd International Conference on Image, Vision and Computing (ICIVC). IEEE, 2018*, pp. 855–859.

Eades, D.C., Otte, D., Cigliano, M. & Braun, H. (2010) Orthoptera species file online. *Version*, 2(3.5), 13.

Elliott, T.M. & Theunissen, F.E. (2009) The modulation transfer function for speech intelligibility. *PLoS computational biology*, 5(3), e1000302.

Ewing, A.W. (1989) *Arthropod Bioacoustics: Neurobiology and Behaviour*. Ithaca, New York: Comstock Publishing.

Faiß, M. (2022) Insectset32: Dataset for automatic acoustic identification of insects (orthoptera and cicadidae). *Zenodo*,.

Faiß, M. (2023) Insectset47 & insectset66: Expanded datasets for automatic acoustic identification of insects (orthoptera and cicadidae). *Zenodo*,.

Faiß, M. & Stowell, D. (2023) Adaptive representations of sound for automatic insect recognition. *PLOS Computational Biology*, 19(10), e1011541.

Flynn, M. & Bagnall, A. Classifying flies based on reconstructed audio signals. In: *Intelligent Data Engineering and Automated Learning–IDEAL 2019: 20th International Conference, Manchester, UK, November 14–16, 2019, Proceedings, Part II 20. Springer, 2019*, pp. 249–258.

Gallardo-Antolín, A. & Montero, J.M. (2021) On combining acoustic and modulation spectrograms in an attention lstm-based system for speech intelligibility level classification. *Neurocomputing*, 456, 49–60.

Gandini, D. (2022) Insect species sound classification using deep learning with small data. Ph.D. thesis, tilburg university.

Greenfield, M.D. (1997) Acoustic communication in orthoptera. *The bionomics of grasshoppers, katydids and their kin*„ 197–230.

Greenfield, M.D. (2002) *Signalers and receivers: mechanisms and evolution of arthropod communication*. : Oxford University Press.

Greenfield, M.D. (2016) Evolution of acoustic communication in insects. *Insect hearing*„ 17–47.

Hailay Gebremariam, G. (2024) A systematic review of insect decline and discovery: Trends, drivers, and conservation strategies over the past two decades. *Psyche: A Journal of Entomology*, 2024(1), 5998962.

Hallmann, C.A., Sorg, M., Jongejans, E., Siepel, H., Hofland, N., Schwan, H. et al. (2017) More than 75 percent decline over 27 years in total flying insect biomass in protected areas. *PloS one*, 12(10), e0185809.

He, H., Chen, J., Chen, H., Zeng, B., Huang, Y., Zhaopeng, Y. et al. (2024) Enhancing insect sound classification using dual-tower network: A fusion of temporal and spectral feature perception. *Applied Sciences*, 14(7), 3116.

Hedwig, B. et al. (2014) *Insect hearing and acoustic communication*. Vol. 1. : Springer.

Hibino, S., Suzuki, C. & Nishino, T. (2021) Classification of singing insect sounds with convolutional neural network. *Acoustical Science and Technology*, 42(6), 354–356.

Kalita, H. & Das, K. (2023) Chapter-4 exploring the ecological role of insects in biodiversity and ecosystems. *Advances in Entomology*„ 63.

Kohlberg, A.B., Myers, C.R. & Figueroa, L.L. (2024) From buzzes to bytes: A systematic review of automated bioacoustics models used to detect, classify and monitor insects. *Journal of Applied Ecology*,.

Laith, A.E., Alnimri, M., Ali, H., Alkhawaldeh, M. & Mihyar, A. (2024) Mosquito-borne diseases: assessing risk and strategies to control their spread in the middle east. *Journal of Biosafety and Biosecurity*,.

Leather, S.R. (2017) "ecological armageddon"-more evidence for the drastic decline in insect numbers. *Annals of Applied Biology*, 172(1), 1–3.

Low, M.L., Naranjo, M. & Yack, J.E. (2021) Survival sounds in insects: diversity, function, and evolution. *Frontiers in Ecology and Evolution*, 9, 641740.

Luo, C., Wei, C. & Nansen, C. (2015) How do "mute" cicadas produce their calling songs? *PLoS One*, 10(2), e0118554.

Mankin, R., Hagstrum, D., Guo, M., Eliopoulos, P. & Njoroge, A. (2021) Automated applications of acoustics for stored product insect detection, monitoring, and management. *Insects*, 12(3), 259.

Matsuo, I. (2023) Extraction of acoustic features of calls of five platypleura species using the field recordings. *The Journal of the Acoustical Society of America*, 154(4_supplement), A216–A216.

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

Miller, L.A. (1977) Directional hearing in the locust schistocerca gregaria forskål (acrididae, orthophera). *Journal of comparative physiology*, 119(1), 85–98.

Montealegre-z, F. (2009) Scale effects and constraints for sound production in katydids (orthoptera: Tettigoniidae): correlated evolution between morphology and signal parameters. *Journal of Evolutionary Biology*, 22(2), 355–366.

Mora, C., Tittensor, D.P., Adl, S., Simpson, A.G. & Worm, B. (2011) How many species are there on earth and in the ocean? *PLoS biology*, 9(8), e1001127.

Neil, T.R. & Holderied, M.W. (2021) Sound production and hearing in insects. In: *Advances in Insect Physiology*. Vol. 61Elsevier, pp. 101–139.

Noda, J.J., Travieso-González, C.M., Sanchez-Rodriguez, D. & Alonso-Hernández, J.B. (2019) Acoustic classification of singing insects based on mfcc/lfcc fusion. *Applied Sciences*, 9(19), 4097.

Ntalampiras, S. (2019) Automatic acoustic classification of insect species based on directed acyclic graphs. *The Journal of the Acoustical Society of America*, 145(6), EL541–EL546.

Phung, Q.V., Ahmad, I., Habibi, D. & Hinckley, S. (2017) Automated insect detection using acoustic features based on sound generated from insect activities. *Acoustics Australia*, 45, 445–451.

Pollack, G.S. (2017) Insect bioacoustics. *Acoustics Today*, 13(2), 26–34.

Potamitis, I. (2014) Classifying insects on the fly. *Ecological informatics*, 21, 40–49.

Raghuwanshi, P. & Kaushik, R. Insect classification using mel-cstft: A fusion of mel spectrogram and chroma stft features. In: *2024 First International Conference on Electronics, Communication and Signal Processing (ICECSP). IEEE, 2024*, pp. 1–5.

Rasmussen, J.H., Stowell, D. & Briefer, E.F. (2024) Sound evidence for biodiversity monitoring. *Science*, 385(6705), 138–140.

Robillard, T. & Desutter-Grandcolas, L. (2011) The complex stridulatory behavior of the cricket eneoptera guyanensis chopard (orthoptera: Grylloidea: Eneopterinae). *Journal of Insect Physiology*, 57(6), 694–703.

Romer, H. & Lewald, J. (1992) High-frequency sound transmission in natural habitats: implications for the evolution of insect acoustic communication. *Behavioral Ecology and Sociobiology*, 29, 437–444.

Rothenberg, D. (2013) *Bug music: How insects gave us rhythm and noise*. : St. Martin's Press.

Sanborn, A.F. (2003) Analysis of the calling songs of platypleura hirtipennis (germar, 1834) and p. plumosa (germar, 1834)(hemiptera: Cicadidae). *African entomology*, 11(2), 291–296.

Sánchez-Bayo, F. & Wyckhuys, K.A. (2019) Worldwide decline of the entomofauna: A review of its drivers. *Biological conservation*, 232, 8–27.

Sankarganesh, E. (2017) Insect biodiversity: The teeming millions-a review. *Bull Environ Pharmacol Life Sci*, 6, 101–5.

Sarria-S, F.A., Morris, G.K., Windmill, J.F., Jackson, J. & Montealegre-Z, F. (2014) Shrinking wings for ultrasonic pitch production: hyperintense ultra-short-wavelength calls in a new genus of neotropical katydids (orthoptera: Tettigoniidae). *PLoS One*, 9(6), e98708.

Savary, S., Willocquet, L., Pethybridge, S.J., Esker, P., McRoberts, N. & Nelson, A. (2019) The global burden of pathogens and pests on major food crops. *Nature ecology & evolution*, 3(3), 430–439.

Scudder, G.G. (2017) The importance of insects. *Insect biodiversity: science and society*„ 9–43.

Shelley, S. & Murphy, D.T. Openair: An interactive auralization web resource and database. In: *129th Audio Engineering Society Convention 2010, 2010*, pp. 1270–1278.

Singh, P., Sahidullah, M. & Saha, G. (2023) Modulation spectral features for speech emotion recognition using deep neural networks. *Speech Communication*, 146, 53–69.

Skendžić, S., Zovko, M., Živković, I.P., Lešić, V. & Lemić, D. (2021) The impact of climate change on agricultural insect pests. *Insects*, 12(5), 440.

Song, H., Béthoux, O., Shin, S., Donath, A., Letsch, H., Liu, S. et al. (2020) Phylogenomic analysis sheds light on the evolutionary pathways towards acoustic communication in orthoptera. *Nature communications*, 11(1), 4939.

Song, Z., Wu, J., Zhang, M., Shou, M.Z. & Li, H. Spiking-leaf: A learnable auditory front-end for spiking neural networks. In: *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2024*, pp. 226–230.

Stephen, R. & Hartley, J. (1995) Sound production in crickets. *Journal of Experimental Biology*, 198(10), 2139–2152.

Stork, N.E. (2018) How many species of insects and other terrestrial arthropods are there on earth? *Annual review of entomology*, 63(1), 31–45.

Tatsuta, H., Yokokura, R. & Saski, T. Preliminary studies of acoustic discrimination between an endangered cicadinae species, platypleura albivannata, and a closely related species platypleura yayeyamana. In: *Biological Shape Analysis: Proceedings of the 4th International Symposium*. *World Scientific*, *2017*, pp. 80–94.

Travieso, C.M., Noda, J.J. & Sánchez-Rodríguez, D. (2021) Acoustic identification of insects based on cepstral data fusion and hidden markov models. In: *Neuroendocrine Regulation of Animal Vocalization*Elsevier, pp. 31–37.

Van Klink, R., August, T., Bas, Y., Bodesheim, P., Bonn, A., Fossøy, F. et al. (2022) Emerging technologies revolutionise insect ecology and monitoring. *Trends in ecology & evolution*, 37(10), 872–885.

Van Staaden, M.J. & Römer, H. (1997) Sexual signalling in bladder grasshoppers: tactical design for maximizing calling range. *Journal of Experimental Biology*, 200(20), 2597–2608.

Varma, A.L.S., Bateshwar, V., Rathi, A. & Singh, A. Acoustic classification of insects using signal processing and deep learning approaches. In: *2021 8th International Conference on Signal Processing and Integrated Networks (SPIN)*. *IEEE*, *2021*, pp. 1048–1052.

Vellinga, W.P., Planqué, B., Pieterse, S. & Jongsma, J. (2017) www. xeno-canto. org: A decade on. *Neotropical Birding*, 21, 40–47.

Young, D. & Bennet-Clark, H. (1995) The role of the tymbal in cicada sound production. *Journal of Experimental Biology*, 198(4), 1001–1020.

Zeghidour, N., Teboul, O., Quitry, F.D.C. & Tagliasacchi, M. (2021) Leaf: A learnable frontend for audio classification. *arXiv preprint arXiv:2101.08596*,.

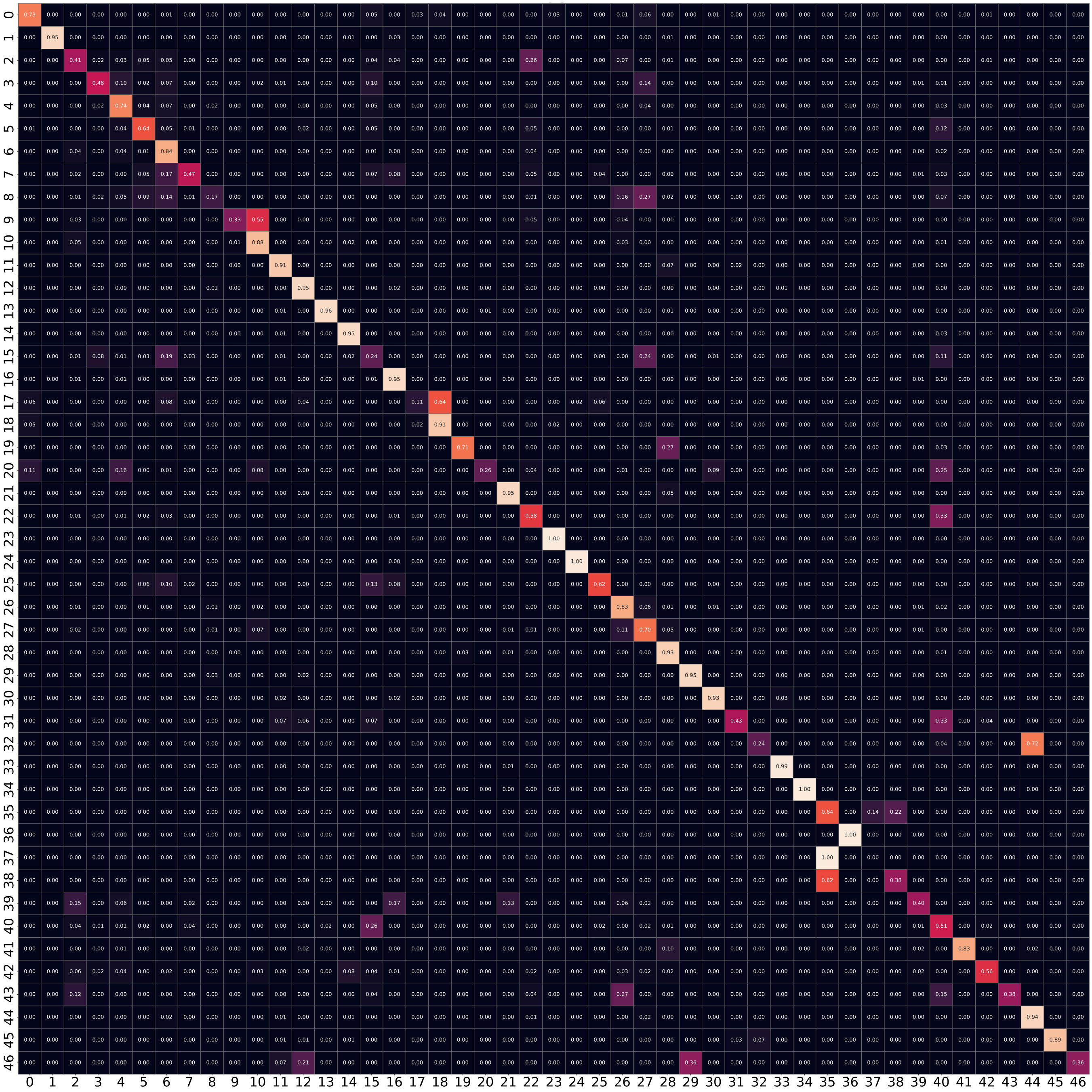

FIGURE 5 Confusion matrix for 47 species in the test set, using the Modulation Spectrogram ("MOD") as the front-end. The model consists of six convolutional layers and achieves 76% classification accuracy. The vertical axis represents the actual labels of the audio files, while the horizontal axis represents the labels predicted by the model, with numerical values 0–46 corresponding to the insect species detailed in Table 1.

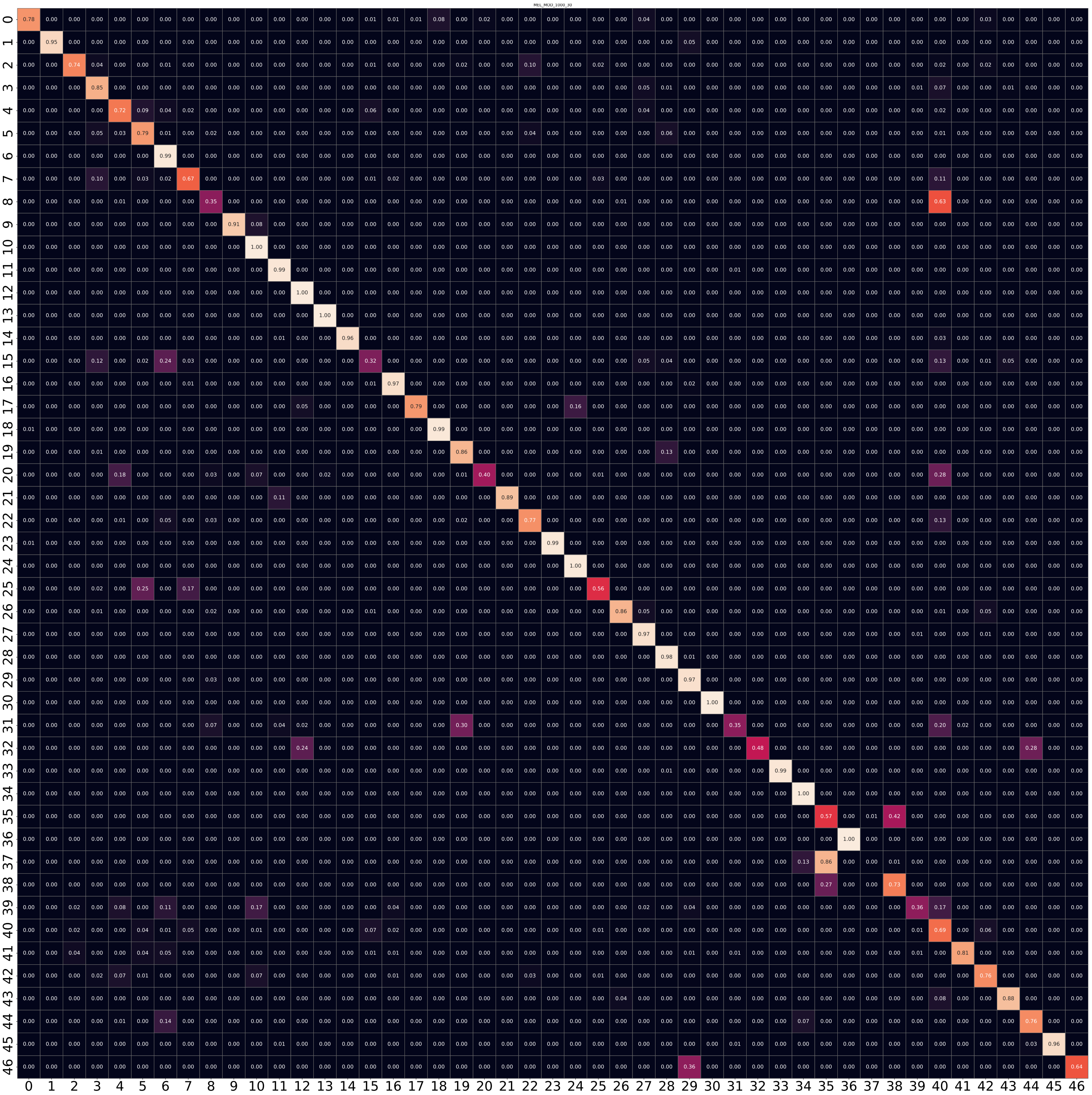

FIGURE 6 Confusion matrix for 47 species in the test set, using the Mel spectrogram (”MEL”) as Front-end 1 and the Modulation Spectrogram (”MOD”) as Front-end 2. The model consists of five convolutional layers in each branch and achieves 88% classification accuracy. The vertical axis represents the actual labels of the audio files, while the horizontal axis represents the labels predicted by the model, with numerical values 0–46 corresponding to the insect species detailed in Table 1.