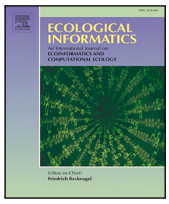
Ecological Informatics 86 (2025) 103010

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf



Transformer Models improve the acoustic recognition of buzz-pollinating bee species

Alef Iury Siqueira Ferreira [a], Nádia Felix Felipe da Silva [a], Fernanda Neiva Mesquita [a], Thierson Couto Rosa [a], Stephen L. Buchmann [b], José Neiva Mesquita-Neto [c,*]

[a] Universidade Federal de Goiás, Instituto de Informática, Goiânia, 74690-900, Goiás, Brazil
[b] Departments of Entomology and Ecology and Evolutionary Biology, University of Arizona, Tucson, 85721, AZ, USA
[c] Laboratorio Ecología de Abejas, Departamento de Biología y Química, Facultad de Ciencias Básicas, Universidad Católica del Maule, Talca, 3480112, Maule, Chile

ARTICLE INFO

Dataset link: https://doi.org/10.5281/zenodo.14611302, https://github.com/alefiury/Transformers-Bee-Species-Acoustic-Recognition

Keywords:
Buzz-pollinated crops
Ecosystem services
Crop pollination
Deep learning

ABSTRACT

Buzz-pollinated crops, such as tomatoes, potatoes, kiwifruit, and blueberries, are among the highest-yielding agricultural products. The flowers of these cultivated plants are characterized by having a specialized flower morphology with poricidal anthers that require vibration to achieve a full seed set. At least 446 bee species, in 82 genera, use floral sonication (buzz pollination) to collect pollen grains as food. Identifying and classifying these diverse often look-alike bee species poses a challenge for taxonomists. Automated classification systems, based upon audible bee floral buzzes, have been investigated to meet this need. Recently, convolutional neural network (CNN) models have demonstrated superior performance in recognizing and distinguishing bee-buzzing sounds compared to classical Machine-Learning (ML) classifiers. Nonetheless, the performance of CNNs remains unsatisfactory and can be improved. Therefore, we applied a novel transformer-based neural network architecture for the task of acoustic recognition of blueberry-pollinating bee species. We further compared the performance of the Audio Spectrogram Transformer (AST) model and its variants, including Self-Supervised AST (SSAST) and Masked Autoencoding AST (MAE-AST), to that of strong baseline CNN models based on previous work, at the task of bee species recognition. We also employed data augmentation techniques and evaluated these models with a data set of bee sounds recorded during visits to blueberry flowers in Chile (518 audio samples of 15 bee species). Our results revealed that Transformer-based Neural Networks combined with pre-training and data augmentation outperformed CNN models (maximum F1-score: 64.5% ± 2; Accuracy: 82.2% ± 0.8). These innovative attention-based neural network architectures have demonstrated exceptional performance in assigning bee buzzing sounds to their respective taxonomic categories, outperforming prior deep learning models. However, transformer approaches face challenges related to small dataset size and class imbalance, similar to CNNs and classical ML algorithms. Combining pre-training with data augmentation is crucial to increase the diversity and robustness of training data sets for the acoustic recognition of bee species. We document the potential of transformer architectures to improve the performance of audible bee species identification, offering promising new avenues for bioacoustic research and pollination ecology.

1. Introduction

Insect pollinators are key providers of essential ecosystem services including agricultural crop pollination and for flowering plants in adjacent wildlands. The production of about 35% of the global food supply humans eat depends on animal pollination, primarily by social and solitary bees (Potts et al., 2016). Approximately, 75% of the world's 1440 crops benefit from animal pollination (Klein et al., 2007). However, pollination can fail or be insufficient, affecting agricultural productivity (Lippert et al., 2021; Wilcock and Neiland, 2002). Managed pollinators are often used in agriculture plantings to supplement native pollinators and increase the yield and quality of crop production (Velthuis and Van Doorn, 2006; Rucker et al., 2012). However, some crops have a specialized flower morphology with poricidal anthers that require vibration to achieve a full seed set, called "buzz-pollinated crops" (Cooley and Vallejo-Marín, 2021), such as tomatoes, potatoes, kiwifruit, cranberries, and blueberries, which are among the highest-yielding and most valuable agricultural products.

* Correspondence to: Avenida San Miguel 3696, Talca, Región del Maule, Chile.
E-mail address: jmesquita@ucm.cl (J.N. Mesquita-Neto).

https://doi.org/10.1016/j.ecoinf.2025.103010
Received 14 June 2024; Received in revised form 8 January 2025; Accepted 9 January 2025
Available online 22 January 2025

To efficiently extract pollen from flowers of buzz-pollinated plants and pollinate them, bees vibrate these tubular anthers using their indirect flight muscles, which shakes the pollen inside the hollow anthers causing the pollen grains to be expelled through the apical pores of the anthers (Buchmann et al., 1983) and striking the bees. These vibrations produce audible sounds "bee buzzes" that give the name to this phenomenon, known as buzz pollination or floral sonication (Buchmann et al., 1983; De Luca and Vallejo-Marin, 2013). Numerous studies have shown that bee visits, especially by those capable of buzz pollination, can improve fruit yield and quality and achieve a full seed set in several buzz-pollinated crops (tomato: Banda and Paxton, 1990 kiwifruit: Pomeroy and Fisher, 2002; Kim et al., 2005 eggplant: Hikawa, 2004 blueberry and cranberry: Stubbs and Drummond, 1996; Javorek et al., 2002).

Blueberries, for example, can be highly dependent upon buzzing bees to set fruit or increase fruit size. These buzzing visits can deposit up to five times more pollen and produce fruits 1.8 times heavier than those visited solely by honeybees or other non-buzzing insects (Cortés-Rivas et al., 2023a). However, the quality of pollination (i.e. pollen delivery) provided varies among bee species, even among those capable of vibrating flowers (Cortés-Rivas et al., 2023a). Differences in body size (relative to flower size) and foraging behavior of visiting bees (number of flowers visited per unit time) have been proposed to explain the different pollination efficiencies among bee visitors (Solís-Montero and Vallejo-Marín, 2017; Mesquita-Neto et al., 2021). Consequently, the species of the local bee pollinator guild differ in their ability to pollinate, and the use of a bee species that is not suitable for a particular crop reduces its pollination services (Greenleaf and Kremen, 2006; Macias-Macias et al., 2009; Benjamin and Winfree, 2014) Therefore, it is important to correctly identify crop-visiting bees, which are the true and most efficient crop pollinators for the crop plant and localities under consideration.

Nonetheless, the considerable diversity of bee species and other insects that visit flowers poses a challenge to taxonomists (Troudet et al., 2017). Only 446 species of bees have been directly observed to buzz pollinate flowers with porose anthers. Since these species are scattered across 82 bee genera, and there are at least 20,507 bee species in 7 families, there are likely additional bee species that use floral sonication to collect pollen (Ascher and Pickering, 2020; Orr et al., 2020). Furthermore, traditional taxonomic recognition and classification of bees and other insects is not a trivial activity, as species can be almost identical in appearance (Gradišek et al., 2017). Taxonomists rely primarily upon morphological, genetic or behavioral characteristics to identify bees and other insects. These methods, however, tends to be time-consuming and often error-prone, since they are generally dependent on human expertise and experience. The lack of suitably trained taxonomists exacerbates this problem (Francoy et al., 2012; Santana et al., 2014). Due to the limitations of traditional taxonomy, it is advisable to develop and implement new and affordable technologies that also meet the required taxonomic rigor (Gaston and O'Neill, 2004; Zapponi et al., 2017; Neuenschwander et al., 2010).

To address this requirement, automated classification systems for plants and animals have recently been developed and evaluated (Schroder et al., 2002; Santana et al., 2014; Yanikoglu et al., 2014; Valliammal and Geethalakshmi, 2011; Gao et al., 2024), among which, acoustic classification systems stand out. Sound samples are relatively easy to record in the field and can also be recorded remotely and continuously over long periods in a scalable and minimally invasive manner (Gradišek et al., 2017). However, capturing bee sounds has its unique challenges, which are further complicated by the quality of the recordings, the presence of background noise (generally much louder than the buzzing sounds), and the simultaneous occurrence of multiple often overlapping sound events from different species (Ferreira et al., 2023). In addition, species-specific sound event detection requires the identification, classification, and quantification of individual acoustic events (You et al., 2023), which often require hundreds of hours of manual effort to properly label (e.g., determine time markers, species IDs, and sound types) (Oswald et al., 2022).

On the other hand, Deep Learning (DL) has become the standard for the automatic recognition and classification of bee buzzing sound signals with efficiency. One of the major challenges is that acoustic spectrograms, in contrast to common images or audio spectrograms, often lack sufficient textural features due to multivariate influences in the environment such as ambient noise interference (e.g., native and domesticated animals, traffic or wind). As such, it provides only certain low-level correlations between time and frequency axes, which implies a need for models with strong robustness. DL-based models, specifically convolutional neural networks (CNNs), and their variants have added certain improvements for bee species audio recognition (Ferreira et al., 2023). Indeed, CNN models can outperform classical ML classifiers in the recognition of bee-buzzing sounds (Truong et al., 2023). However, CNNs rely heavily on pre-training with large acoustic data sets and data augmentation to outperform classical Machine Learning (ML) classifiers. Despite these enhancements, the performance of CNNs in bee buzzing recognition remains unsatisfactory compared to ML standards, achieving a maximum F1 score of only 58% (Ferreira et al., 2023). In addition, both DL and classical ML models typically demand substantial amounts of training set data to adequately capture the inherent variability present in the data being modeled (O'Mahony et al., 2020). Hence, while CNNs remain valuable models, there is still room for improvement, particularly with the emergence of novel attention-based neural network architectures such as "transformers/perceivers" (Elliott et al., 2021; Wolters et al., 2021).

Indeed, transformers seem to achieve state-of-the-art results and have previously demonstrated considerable potential for various audio classification benchmarks. In particular, their notable success in audio processing further enhances the feasibility of bioacoustic research, such as the Audio Spectrogram Transformer (AST) model (Gong et al., 2021c). Compared with CNNs for the bee sound recognition task, transformer architecture is expected to perceive both global and local information from acoustic spectrograms. Therefore, we use the transformer as the backbone, instead of CNNs, for the acoustic recognition of bee species. To our knowledge, researchers have not widely applied the attention-based mechanism in the field of bioacoustics (Stowell, 2022; Fundel et al., 2023). Therefore, we applied the novel transformer-based neural network architectures to the task of acoustic recognition of blueberry-pollinating bee species. Due to the strong representational capability and demonstrated ability in the acoustic domain of transformers, we expected that this architecture would outperform CNNs and enhance the acoustic recognition of bee species (Hypothesis 1). However, the success of audio spectrogram transformers relies on supervised pre-training, which requires a large amount of labeled data for training compared to CNNs (Dosovitskiy et al., 2021). Conversely, collecting bee audio data in the field typically demands domain expertise and entails extensive, time-consuming sampling and labeling efforts, often resulting in small and unbalanced datasets (e.g., Ribeiro et al. (2021), Ferreira et al. (2023) and Fundel et al. (2023)). To reduce the need for large amounts of labeled data for the AST model, we expect that adding regularization techniques to the fine-tuning stage, such as data augmentation techniques, would greatly improve the performance of the AST model in the acoustic recognition of bee species (Hypothesis 2).

2. Materials and methods

We used the data set of Ferreira et al. (2023), who also trained DL models for the acoustic detection of bee species visiting blueberry flowers in Chile. In contrast to our study, their analyses relied on DL models utilizing multi-layer artificial neural networks, specifically Convolutional Neural Networks (CNNs). The data set comprises 518 audio samples (totaling 3595 buzzing-sound segments, with 1728 being sonication events) from 15 bee species during their visits to highbush

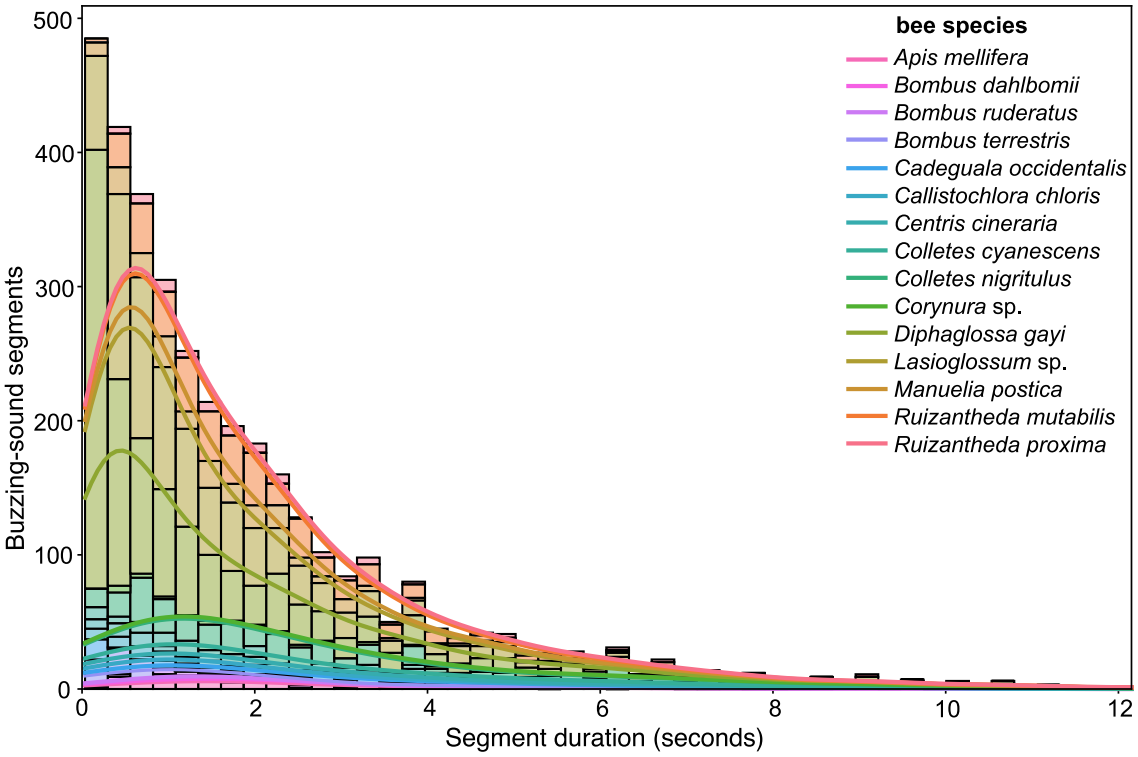


Fig. 1. Histogram showing the distribution of the duration of sampled buzzing-sound segments across bee species visiting blueberry flowers. The curves show the proportion of values in each range of segment duration (in seconds) per bee species.

blueberry (*Vaccinium corymbosum*) flowers in five orchards situated in southern Chile (Maule and Los Ríos regions), spanning September to November of 2020 and 2021. While most bee species are native to Chile (12 species), *Bombus terrestris*, *B. ruderatus*, and *Apis mellifera* are exotics. The duration of audio samples ranges from 5 s to over one minute, recorded at a sampling rate of 44.1 kHz. The distribution of samples per bee species exhibited significant imbalance, ranging from just eight samples (*Corynura chloris*) to 108 (*Cadeguala occidentalis*; also see Fig. 1; Table S4).

2.1. Acoustic pre-processing

The data set of Ferreira et al. (2023) was already acoustically pre-processed. We used it here without any changes, meaning we kept the bee buzzes that had already been detected and selected. Ferreira et al. (2023) conducted data pre-processing before training DL models to enhance their performance. The original sound file recordings (in .wav format) were manually categorized, and segments featuring bee-buzzing sounds were selected (see examples in Fig. 2). These segments were labeled as being of one of two categories: (1) sonication, which includes floral buzzing sounds produced by bees vibrating blueberry flowers, or (2) flight, encompassing wingbeat sounds from bees flying between flowers. In particular, Ribeiro et al. (2021) found that sonication sounds contributed more to the performance of ML models than did flight sounds for the acoustic recognition bee species. However, a dataset containing both sonication and flight sound segments contributed as much to the training of a classifier as the same dataset containing only sonication sounds (Ribeiro et al., 2021). Therefore, we used both categories of sounds together in all experiments, since flight and sonication together yielded a larger number of audio segments and included bee species not capable of sonication. The analysis was carried out using Raven Lite software (Cornell Laboratory of Ornithology in Ithaca, New York).

Background noise was common in the data set (e.g., traffic, birds, crickets, people talking, etc.). We kept portions of field recordings without bee sounds for later analysis (see Fig. 2). We chose to input the original audio without removing or attenuating the ambient noise because noise is almost unavoidable in real-world situations (Ribeiro et al., 2021). Training our neural network on noisy data means it would generalize similarly to noisy test data.

2.2. Spectrogram generation

Audio feature extraction techniques transform raw audio data generated by acoustic pre-processing into features that explicitly represent properties of the data that may be relevant for later ML classification. We used two main functions to convert waveform information into time–frequency representations of audio signals: the Log Mel Spectrogram and the Mel Filterbank. We incorporated these image-analog inputs to facilitate complex data interpretation by exploiting their visual similarities, following the protocols proposed by Dosovitskiy et al. (2021), Gong et al. (2021a) and Baade et al. (2022a).

We used the Log Mel Spectrogram within the PANNs (CNN14) model, as recommended by Kong et al. (2020a) and it is also used in EffNet V2 Small. The Log Mel Spectrogram is a sophisticated transformation of audio signals that greatly improves the representation of sound for analysis and signal processing. We first decomposed the audio signals into their component frequencies over time to create a spectrogram using the Short-Time Fourier Transform (STFT) algorithm. We then filtered this spectrogram through a Mel filter bank, which distorts the frequency scale to match the Mel scale. The Mel scale is a conversion of frequency in Hertz (f) to the Mel scale, as shown in Eq. (1). Finally, the Log Mel Spectrogram and the logarithm of the amplitude of the frequencies were generated to convert the intensity scale to one that reflects the logarithmic perception of loudness by the human auditory system. The specific parameters used to extract the Log Mel Spectrogram can be seen in Table 1

$$Mel = 2.595 \log_{10}(1 + \frac{f}{700}) \quad (1)$$

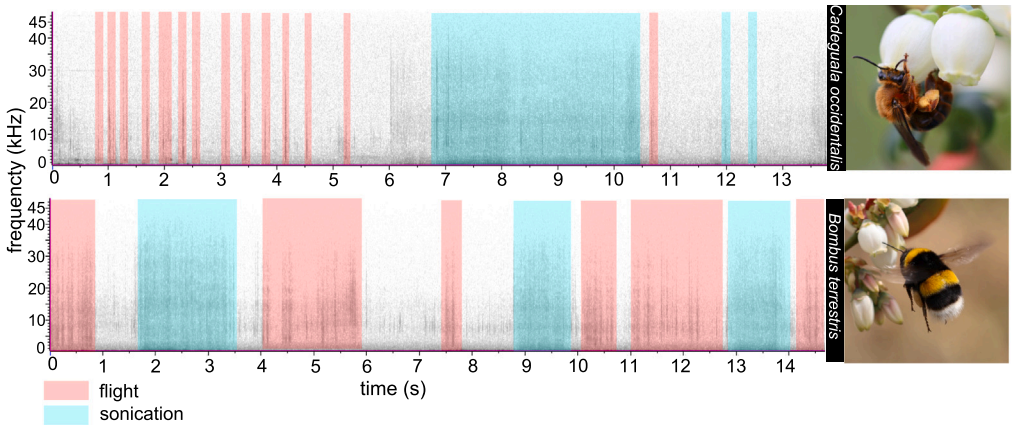


Fig. 2. Examples of spectrograms of the two categories of buzzing sounds (sonication and flight) produced by two species of bees visiting blueberry flowers (*Cadeguala occidentalis* and *Bombus terrestris*). Note that the duration, amplitude, and frequency of the buzzing sounds vary between species and between types of bee sounds (sonication and flight).

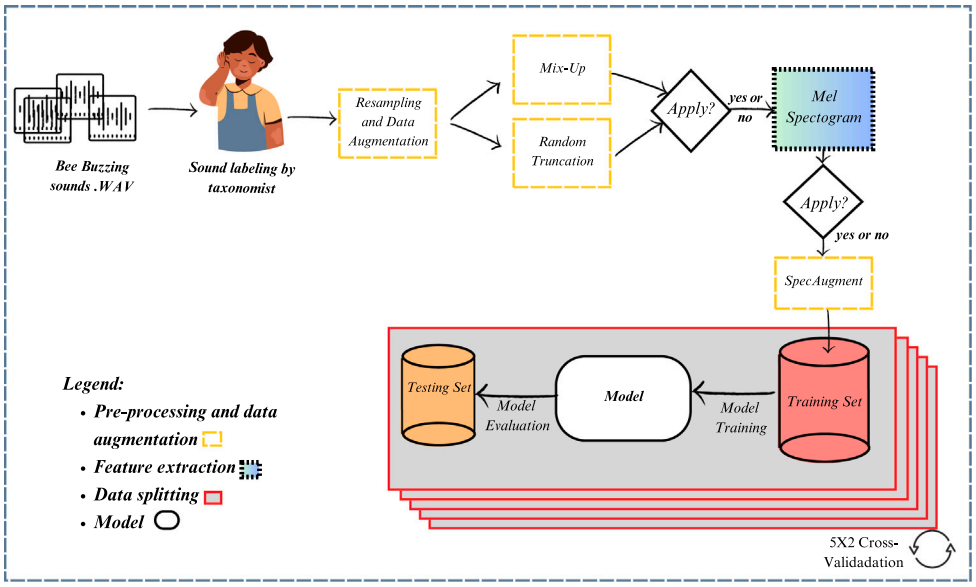


Fig. 3. Overview of the approach adopted for the acoustic classification of bee buzzing sounds and the Deep Learning pipeline enhanced by data augmentation techniques during the pre-processing phase. The original audio files (.wav format) containing recordings of bee buzzing sounds during visits to blueberry flowers were manually classified into sonication or flight (wingbeat sound) segments. This was followed by resampling and the application of data augmentation methods directly to the waveforms (mix-up and random truncation). Additionally, Mel spectrograms were extracted from the waveforms and another augmentation technique was applied (SpecAugment). After the pre-processing stage, the resulting data set was split into the training/development set and the test data set. The role of the test set was to evaluate the effectiveness of the model classifiers in accurately assigning the buzzing sound to the respective bee taxa.

We used the Mel filterbank as input for the models PSLA, AST, SSAST, and MAE-AST, following the guidelines described in their original papers. For all Transformer-based models, we first divide the spectrogram features into overlapping 16 × 16 pixel patches with a 6-pixel overlap in both time and frequency dimensions. Each patch is then processed through a learned linear projection layer, transforming it into a 1D patch embedding vector. To enable classification and maintain consistency with the Transformer architecture, we prepend a learnable [CLS] token to the sequence and add learnable positional embeddings to preserve spatial information. The CNN-based models employ different input processing approaches. EfficientNet, originally designed and pre-trained for RGB image processing, requires multiple replications of the single-channel log Mel spectrogram to create a three-channel input. In contrast, PANNs and PSLA were specifically designed for audio spectrograms and directly processing single-channel 2D inputs through specialized convolutional layers. The specific parameters used can be seen in Table 1. The Mel filterbank used in the feature extraction process of the Kaldi toolkit was developed as part of an open-source framework dedicated to human speech recognition research (see Povey et al. (2011)). While it is aligned with the conventional methodology used to generate log mel spectrograms, the resulting feature it yields may exhibit differences. This divergence is due to Kaldi's comprehensive suite of feature extraction tools, which are optimized for speech recognition applications. Such specialized adjustments may result in nuances that distinguish the features extracted by Kaldi from those obtained by a standard Log Mel Spectrogram approach.

We did not use feature engineering or selection because deep learning methods (CNNs and transformers) can do this automatically (Goodfellow et al., 2016).

2.2.1. Data splitting

To facilitate cross-validation, we subjected the audio sample data set to a rigorous partitioning process. Initially, we divided the dataset into two equal-sized subsets for training and testing purposes in each replication (Alpaydm, 1999). However, unlike previous work (see Ribeiro et al. (2021) and Ferreira et al. (2023)), we further divided the training set into two portions: 80% designated for training and 20% for validation. In the end, each replication followed this data distribution: 40%

for training, 10% for validation, and 50% for testing, resulting in a total of 10 runs (see Fig. 3). It is important to note that, due to the use of distinct seeds for each replication, the data distribution varied across runs.

To assess the effectiveness of supervised classification learning algorithms, we utilized the Combined 5 × 2 Cross-validated F-Test (Alpaydm, 1999), which is considered a more reliable alternative to the 5 × 2 cross-validated t-test (Dietterich, 1998). The Combined 5 × 2 Cross-validated F-test addresses the shortcomings of the cross-validated t-test and offers improved statistical power (Fig. 3). This approach involved five replications of two-fold cross-validation to ensure robust and reliable results.

All splits were stratified based on bee species. This approach kept the original class distribution across all folds, ensuring any existing class imbalance was consistently represented in each split. We also used fixed seeds when randomly choosing the samples to generate the splits to ensure data reproducibility and to guarantee that all methods were trained and tested with the same splits. The implementation details of the splitting are available in our repository: https://github.com/alefiury/Transformers-Bee-Species-Acoustic-Recognition

2.3. Data augmentation

Data augmentation seeks to improve the performance of ML algorithms by generating additional data for the training set of the model (Chlap et al., 2021; Kumar et al., 2024). It is particularly useful when the training set is small and/or imbalanced (Abayomi-Alli et al., 2022), which was the case of our buzzing bee dataset. Data augmentation has also improved the performance of CNNs for acoustic recognition of bee species (see Ferreira et al. (2023)). Thus, we used data augmentation to increase the variety and robustness of our training data for both CNN and transformer models.

To augment our bee audio dataset, we applied the following data augmentation techniques that have proven to be useful tools (Fig. 3): SpecAugment, Random Truncation (RT), and Mixup. SpecAugment applies time warping, time masking, and frequency masking to log mel spectrograms, making it an effective method for enriching the training data set (Park et al., 2019). Although we did not apply time warping in our experiments, we adopted the following SpecAugment parameters: the maximum width of frequency masks (F), the maximum width of time masks (T), the number of frequency masks (m_F) and the number of time masks (m_T) applied. The values used for these parameters were taken directly from the literature, as they have shown consistent effectiveness across different audio tasks.

Random Truncation involves sampling segments of audio samples for each forward pass of a DL model. This approach contrasts with fixed-segment approaches and contributes to improving learning (Ferreira et al., 2023). This technique introduces temporal variability in audio segments, thereby simulating natural conditions and improving the robustness of the recognition model. The initial step was to load the audio recording that will be utilized as the input. The parameters that control the randomness of segments and intervals were then defined. The lengths of the segments and the intervals between them were generated using a uniform distribution. Subsequently, the segments were extracted from the original recording by the generated start times and lengths, and then combined to form a new augmented recording. The Random Truncation augmentation technique is formalized in Algorithm S5, which delineates the procedural steps for implementing this method.

Mixup uses the convex combination of two different features and labels of audio samples to increase the variability of the training data. It blends samples from different bee species using the specified mixing parameter λ drawn from a Beta(α, α) distribution (Zhang et al., 2017). In our implementation, we used α (alpha) = 0.5 for the Beta distribution, which we empirically found worked best for our specific dataset and task, despite values around 0.3–0.4 being often used in the literature. The α parameter controls the shape of the Beta distribution from which λ is drawn, with higher α values leading to λ values closer to 0.5, resulting in more balanced mixing between samples.

To ensure the integrity of the data augmentation techniques, we adhered closely to the parameters originally utilized in the original works for the models. Table 1 illustrates the specific parameters utilized in each method. Data augmentation was conducted dynamically throughout the training phase, with a 100% augmentation probability. Each sample is augmented distinctly in each epoch due to the stochastic nature of the augmentation methods employed.

2.4. Configurations of transformers

Audio data often have complex and subtle patterns that can be difficult to capture using traditional analytical methods. Ideally, the attention mechanism helps to identify and focus on these intricate patterns. Thus, we selected attention-based Transformer models with demonstrated potential for processing acoustic signals, in particular the Audio Spectrogram Transformer (AST), the Self-Supervised Audio Spectrogram Transformer (SSAST), and the Masked Autoencoding Audio Spectrogram Transformer (MAE-AST), as described in the sections above. Unlike traditional Convolutional Neural Networks (CNNs), which map audio spectrograms directly to labels, these transformer architectures are built entirely around attention mechanisms, which helps the model to prioritize and give more weight to significant parts of the data during processing.

2.4.1. Audio Spectrogram Transformer (AST)

We first applied the Audio Spectrogram Transformer (AST) to improve the task of the acoustic recognition of bee species. The Audio Spectrogram Transformer (AST) leverages attention-based mechanisms to capture complex audio patterns (Gong et al., 2021a), making it more effective at detecting subtle nuances in audio data compared to CNNs (Islam et al., 2023). This is primarily due to the ability of AST to simultaneously process the entire audio sequence, enabling it to understand long-range dependencies often critical within audio data samples.

Initially, we used models that had been pre-trained and made publicly accessible via the official repository of the authors (see Gong et al. (2021b)). We then tested the AST extensions as described in the following subsections.

2.4.2. Self-Supervised Audio Spectrogram Transformer (SSAST)

The Self-Supervised Audio Spectrogram Transformer (SSAST) is a groundbreaking framework that combines discriminative and generative self-supervised learning, setting a new benchmark for self-supervised learning within the context of the Audio Spectrogram Transformer (AST) Atito et al. (2021). SSAST significantly enhances AST performance across various downstream tasks (Li et al., 2023). On average, it improves performance by 60.9%, often matching or surpassing the results achieved by models pre-trained with supervised data. One of the key advantages of SSAST is that it can perform exceptionally well without needing labeled/annotated data. In addition, SSAST supports different patch sizes and shapes, providing greater flexibility in its applications compared to supervised ImageNet pre-training, which restricts patches to squares. As a result of these key features, SSAST has outperformed previous models in many tasks, particularly in speech-related challenges (Bayraktar et al., 2023). For our experiments, we used the pre-trained SSAST models available from the authors' official repository (see Gong et al. (2022)).

Table 1
Hyperparameter configuration used for all tested CNNs (EffNet V2 Small, PANNs, PSLA) and transformer models (AST, SSAST, MAE-AST) combined with the best hyperparameter configuration strategies. Also included are model parameters (e.g., number of trainable parameters, batch size, optimizer settings), data preprocessing parameters (e.g., feature type, sampling rate, window size, hop size, mel bins), and data augmentation parameters (e.g., Mixup α, SpecAugment settings). Mixup α controls the strength of interpolation between two random training examples and their labels. SpecAugment parameters include m_T (number of time masks), T (maximum width of time masks), m_F (number of frequency masks), and F (maximum width of frequency masks). Min. Frequency and Max. Frequency refers to the frequency range used in mel spectrogram computation.

Hyperparameter	EffNet V2 Small	PANNs (CNN14)	PSLA	AST	SSAST	MAE-AST
Model						
Number of trainable parameters	22.18M	81.87M	7.87M	87.74M	87.27M	99.84M
Batch size	32	32	16	16	16	32
Number of epochs	120	120	120	120	120	120
Optimizer	Adam	Adam	Adam	Adam	Adam	Adam
Learning rate	1e−3	1e−3	1e−3	1e−4	1e−4	1e−4
Adam ϵ	1e−08	1e−08	1e−08	1e−08	1e−08	1e−08
Adam β	[0.9, 0.999]	[0.9, 0.999]	[0.9, 0.999]	[0.9, 0.999]	[0.95, 0.999]	[0.9, 0.999]
Weight decay	0.0	0.0	0.0	0.0	5e−7	0.0
Learning rate schedule	LinearWarmup	LinearWarmup	LinearWarmup	LinearWarmup	LinearWarmup	LinearWarmup
Data pre-processing						
Feature	Log Mel Spectrogram	Log Mel Spectrogram	Mel Filterbank	Mel Filterbank	Mel Filterbank	Mel Filterbank
Sampling rate	16 khz	32 khz	32 khz	16 khz	32 khz	32 khz
Window size	1024	1024	800	400	800	800
Hop size	320	320	320	160	320	320
Mel bins	64	64	128	128	128	128
Min. Frequency	0	0	20	20	20	20
Max. Frequency	8000	16 000	16 000	8000	16 000	16 000
Data augmentation						
Mixup α	0.5	0.5	0.5	0.5	0.5	0.5
SpecAugment m_F	2	2	1	1	1	1
SpecAugment T	64	64	96	96	96	96
SpecAugment m_F	2	2	1	1	1	1
SpecAugment F	8	8	48	24	48	48

2.4.3. Masked Autoencoding Audio Spectrogram Transformer (MAE-AST)

The Masked Autoencoding Audio Spectrogram Transformer (MAE-AST) combines the architecture of Masked Autoencoders, such as Scalable Vision Learners (MAE) (Baade et al., 2022), with SSAST. MAE-AST offers notable efficiency advantages, requiring only one-third of the time and half the memory compared to SSAST, despite sharing a similar model architecture. Remarkably, MAE-AST consistently outperforms SSAST in various downstream tasks when they share the same encoder depth, with all other factors being constant. We used a pre-trained MAE-AST model that was obtained directly from the authors' official repository (see Baade et al. (2022a)).

2.5. Hyperparameter setting

For the sake of replicability, continuity, maintainability, and advancement of this research, all hyperparameters were detailed per model in Table 1. The selected value ranges were based on the parameters suggested in the original studies. We employed the optimization algorithm known as Adam (Kingma and Ba, 2014) for all our experiments. Adam utilizes adaptive estimates of lower-order moments for first-order gradient-based optimization of stochastic objective functions. Thus, it played a pivotal role in our optimization process.

All experiments were performed on a single RTX 5000 GPU equipped with 16 GB of VRAM. The source code is available at the following link: https://github.com/alefiury/Transformers-Bee-Species-Acoustic-Recognition

2.6. Evaluation methods

In scenarios with class imbalances, as is the case with our data set, it is crucial to select appropriate evaluation metrics. Therefore, we primarily evaluated our models based on their F1-scores. The F1-score (MacF1) is a metric that balances precision and recall, making it particularly useful when the class distribution is imbalanced.

To evaluate the performance of our CNNs and transformer classifiers, we relied on Accuracy (Acc) and Macro-F1 (MacF1), which are commonly used metrics derived from the confusion matrix. However, we based the performance of our models primarily on the F1-score because our dataset is highly imbalanced and Accuracy tends to underestimate classes with fewer samples (Steiniger et al., 2020). For more details on these metrics, see Ferreira et al. (2023).

2.6.1. Training and fine-tuning

To evaluate the importance of using pre-trained models for both standard CNNs and transformers, we compared the performance of our models with and without pre-training. All the methods used (CNNs and transformers) are available in their code repositories in two versions: a pre-trained version and a non-pre-trained version. The pre-trained versions were trained by their authors mainly on large-scale image datasets like ImageNet (Deng et al., 2009). Some were further pre-trained on sound datasets, like AudioSet (Gemmeke et al., 2017). For detailed information about the datasets used in the original pre-training of these models, we refer the reader to the following publications: Kong et al. (2020b), Gong et al. (2021c,b, 2022) and Baade et al. (2022a).

The parameter values of pre-trained models are learned during the pre-training process using large-scale datasets, and they are made available to be used for various tasks. When applying a pre-trained model to a specific task (such as classifying bee species, in our case), these models undergo further training on a dataset specific to the task at hand (e.g. our training sets of bee species sounds). This additional training is usually referred to as *fine-tuning* the pre-trained model refined for a specific task. Essentially, pre-training gives the model a head start by leveraging what it has learned from a related task before fine-tuning it for the specific task (Yu et al., 2021). This process is particularly useful for datasets like ours, where the amount of labeled data is limited and the classes are imbalanced.

We opted to use pre-trained models for two reasons: firstly, they have achieved better performance in audio classification tasks in recent years (Gwardys and Grzywczak, 2014; Müller et al., 2020; Palanisamy et al., 2020; Zhong et al., 2020; Gong et al., 2021a). Secondly, we expected that the accumulated "previous knowledge" acquired by the pre-training models in their parameters would be better than training the models without pre-training, given the small size and the unbalanced class distribution of our training data.

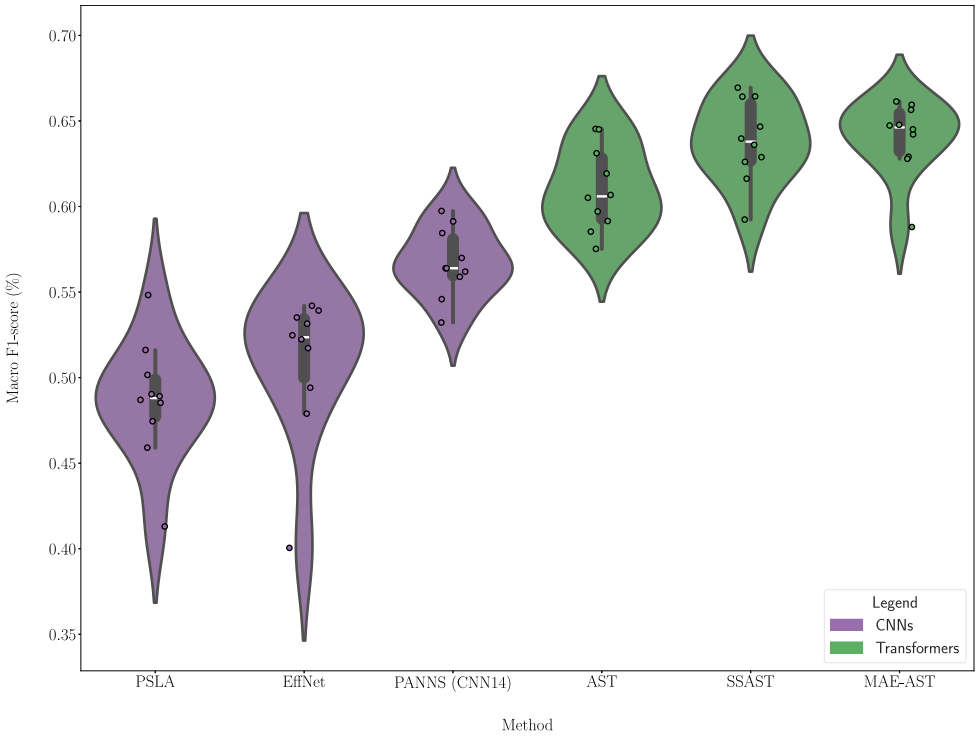


Fig. 4. Violin plots of the best models (higher macro F1-score) of CNNs (EffNet and PANNs, purple plots) and transformer models (AST, SSAST, and MAE-AST, green plots) with the pre-processing techniques (sound feature extraction, pre-training, and/or data augmentation) for the acoustic detection of bee species based on their buzzing sounds produced during visits to blueberry fields in southern Chile. While the black box plots show common summary statistics (with the data medians as white lines), the surrounding violin plots show the probability density of the data at different values. The width of each curve corresponds to the approximate frequency of the data points in each region. Each point on the plot represents the F1-score obtained by an independent model run (10 runs per model, 120 epochs). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The non-pre-trained versions had their parameters initialized with random values before the training process in our experiments. During the training process, these parameters were optimized on our training data only. We used the same protocol to fine-tune the pre-trained versions and to train the non-pre-trained versions. In our approach, we used a combined 5 × 2 cross-validated F-test, with audio files longer than 2 s trimmed to 2 s, and shorter audios padded with silence to achieve a uniform length. For each model, we replaced the original classification head used in the pre-training stage with a classification head designed for 15 classes, representing the 15 bee species buzz or flight sounds being classified. Additionally, we added a sigmoid activation function to the last layer of each model to store the probabilities. This configuration allowed us to take advantage of confounding and training the entire dataset using the same configuration outlined in the hyperparameters, as given in Table 1.

2.6.2. Baseline convolutional neural networks

To compare transformer-based models with state-of-the-art models, we selected the CNN classifiers that demonstrated the highest performance in previous studies on acoustic recognition of bee species (Ferreira et al., 2023): EfficientNet V2, Pre-trained Audio Neural Networks (PANNs), and an additional robust baseline: Pretraining, Sampling, Labeling, and Aggregation (PSLA). We tested both CNN classifiers individually and in combination with pre-training. Additionally, we used a data augmentation ensemble previously established as optimal in prior studies (Ferreira et al., 2023). Classical ML classifiers were not included as baselines in the current study because their performance was extensively evaluated against CNNs in Ferreira et al. (2023), where CNNs were found to outperform classical models consistently on this dataset using identical data splits.

3. Results

3.1. Performance of standard CNNs and transformer classifiers

The performance of the CNN models without any pre-training or data augmentation technique was somewhat low, with Macro F1-scores ranging from 22.6% to 43.6%. This makes sense, as the original training dataset is rather small. However, these scores demonstrated significant improvements when CNNs were supplemented with pre-training and/or audio data augmentation techniques (see Table S1).

The CNN models varied in recognition performance among bee species, as evidenced by Fig. 6. Conversely, the top-performing CNN model struggled to correctly identify the majority of audio samples from minority classes (with less than 50% of the audio samples being incorrectly predicted), namely *Corynura chloris* 28%, *Ruizantheda mutabilis* 24%, *Colletes nigritulus* 23%, and *Manuelia postica* 19% (Fig. 6).

Without any data pre-processing (audio augmentation, sampling, or pre-training), the transformers did not show significantly higher performance (based on Macro F1-score; $p > 0.05$, combined $5\times2cv$ F-test) compared to the best CNNs (Fig. 4). Single transformer models (without any pre-training and data augmentation techniques) performed only slightly better or even had lower Macro F1-scores than single CNNs, although these comparisons were not statistically significant (Table 2). The performance of transformer models became significantly higher than that of CNNs when they were combined with various pre-processing techniques, such as sound feature extraction, data augmentation, and pre-training (Table S1).

However, the transformer models (AST, SSAST, and MAE-AST) boosted by the best combination with pre-processing techniques (pre-training and data augmentation), reached better performance at the acoustic recognition of bee species visiting blueberry crops than the

Table 2
The average predictive performance of the top-performing CNNs (EffNet V2 Small, PSLA, PANNs) and transformer models (AST, SSAST, MAE-AST), was evaluated for recognizing bee species based on their buzzing sounds during visits to blueberry cultivar flowers in southern Chile. These models were enhanced with pre-training and various data augmentation techniques. Model performance was assessed using the average Macro F1-score and Macro Accuracy (± standard deviation). Different superscript letters denote significant differences in F1-score among algorithms (based on MacF1 score; $p \leq 0.05$, $5 \times 2cv$ combined F-test).

Algorithm	Data augmentation	With pre-training		Without pre-training	
		Macro F1 (%)	Accuracy (%)	Macro F1 (%)	Accuracy (%)
CNN (EffNet V2 Small)	RT + Mixup	50.8% (±4)^{c,d}	73.2% (±3)	38.7% (±4)^{d}	62.1% (±3)
PANNs (CNN14)	SpecAugment + Mixup	56.7% (±2)^{b,c}	79.1% (±1)	44.9% (±3)^{d}	75.0% (±1)
PSLA	RT + Mixup	48.6% (±3)^{c,d}	73.8% (±2)	48.2% (±3)^{d}	73.6% (±1)
AST (Transformer)	SpecAugment + RT + Mixup	61.0% (±2)^{a,b}	80.9% (±1)	42.9% (±2)^{d}	67.8% (±1)
SSAST (Transformer)	SpecAugment + RT + Mixup	**63.8%** (±2)^{a}	**81.9%** (±1)	42.3% (±2)^{d}	68.7% (±1)
MAE-AST (Transformer)	SpecAugment + RT + Mixup	**64.5%** (±2)^{a}	**82.2%** (±1)	26.5% (±4)^{e}	61.0% (±3)

best CNNs (see Table 2; Fig. S3). The Macro F1-score of MAE-AST combined with SpecAugment, RT, Mixup, and pre-training was 7.8% higher than that of PANNs with Mixup and SpecAugment (see Table 2; Fig. 4).

3.1.1. Models complexity

Learning curves were employed to identify instances of underfitting and overfitting. This is achieved by plotting the training and validation performance, in terms of Macro F1-score on the *y*-axis, against the training time in terms of epochs on the *x*-axis. The blue curve is the training performance, while the red curve is the performance of the models on the validation set (Fig. 5). A large gap between the training and validation lines indicates overfitting. Consequently, the learning curves indicate that CNN models, both with and without pre-training, tended to overfit during high-intensity training, but showed reduced performance during cross-validation (Fig. 5). The best CNN model (combined with pre-training, Mixup and SpecAugument) was still overfitting, but less so than the single model or with only pre-training. The CNN model can usually perfectly predict bee species identity for the training set, especially when combined with pre-training, but did not generalize well when predicting results for new test samples (validation), characterizing the overfitting.

As with CNNs, all transformer models were overfitted to the training set (Fig. 5). However, the best transformer model (combined with pre-training, Mixup, SpecAugument, and RT) overfitted much less, especially compared to the single model or pre-training alone. The transformer model combined with data pre-training achieved the highest overfitting, characterized by a nearly perfect prediction of bee species identity for the training set, but did not generalize well when predicting results for new test samples (validation). Even the single transformer model overfitted less than when combined with pre-training.

3.1.2. Classifier performance per bee species

Although the per-species performance of the transformers was generally better than that for CNNs, the performance of MAE-AST was nonuniform among classes, varying from 18% (*Ruizantheda mutabilis*) to 96% (*Centris cineraria*). On one hand, the transformers failed to recognize most of the minority classes (less than 50% of the samples were incorrectly predicted): *Colletes cyanescens* 34%, *Colletes nigritulus* 32%, *Manuelia postica* 29%, and *Ruizantheda mutabilis* 18% (see Fig. 6; Fig. S2). On the other hand, the best transformer model (MAE-AST combined with pre-training, SpecAugment, RT, and Mixup) performed well by discriminating among the most represented bee species based on their buzzing sounds (Fig. S2). The models achieved higher hits recognizing the bee species with the highest number of audio samples, in descending order: *Centris cineraria* (96% correctly predicted, N = 208 buzzing audio segments), *Cadeguala occidentalis* (92%, N = 762 audio segments), *Bombus terrestris* (91%, N = 589 buzzing audio segments), and *Bombus dahlbomii* (82%, N = 589 audio segments).

4. Discussion

Recently, CNN models have demonstrated superior performance to classical ML classifiers in the species recognition of bee buzzing sounds (Ferreira et al., 2023). However, when compared to ML standards, their performance remained unsatisfactory (Ferreira et al., 2023). Our results showed that neural networks based on transformers outperformed CNN models, when combined with pre-training and data augmentation techniques for the acoustic recognition of bee species, as predicted by Stowell (2022). These innovative attention-based neural network architectures exhibited superior performance when assigning bee buzzes to their respective taxonomic categories (Genus and species) compared to traditional DL models, achieving an approximately 14.6% improvement in macro F1-score. Nevertheless, transformers still face challenges related to overfitting and require more data sets to achieve results equivalent to other architectures (Dosovitskiy et al., 2021). Combining pre-training with data augmentation becomes critical to increasing the diversity and robustness of training data for the acoustic recognition of bee species.

4.1. Transformers showed a greater dependence on data pre-training to outperform CNNs

Our results showed that standard CNN and transformer models are prone to overfitting, a phenomenon in which a model excels on the training data but struggles to generalize to new evaluation data (Guo et al., 2016). Overfitting can occur for several reasons, including lack of training data (when the training data set is too small), noisy data (the presence of irrelevant information in the training data), overtraining (excessive focus on the training data rather than learning the underlying patterns), and high model complexity (which learns the noise in the training data; see Srivastava et al. (2014)). However, with pre-training, we observed a substantial absolute improvement of approximately 10% by the transformers over the current state-of-the-art model (standard CNNs). Interestingly, in some cases, transformer models even outperformed standard CNNs when pre-training was applied without task-specific methods. This highlights the critical role of pre-training in evaluating future methods for robustness and uncertainty tasks Hendrycks et al. (2019). Therefore, pre-training and data augmentation were used to address the relatively small dataset size and class imbalance, subsequently increasing the diversity and robustness of the training data for acoustic bee species recognition. These techniques emerge as critical factors in achieving improved performance for bee species detection in blueberry crops, whether CNN or transformer models are used.

While both CNNs and Transformers rely on pre-training, CNNs tend to overfit less than transformer models without pre-training. Thus, the performance gap between single transformer models and those combined with pre-training is wider than that of CNNs. Only when combined with pre-training were the Transformers capable of outperforming the best-performing CNN model (PANNs combined with SpecAugment, Mixup, and pre-training) in the task of acoustic recognition of bee species (Hypothesis 1). This implies that transformers may

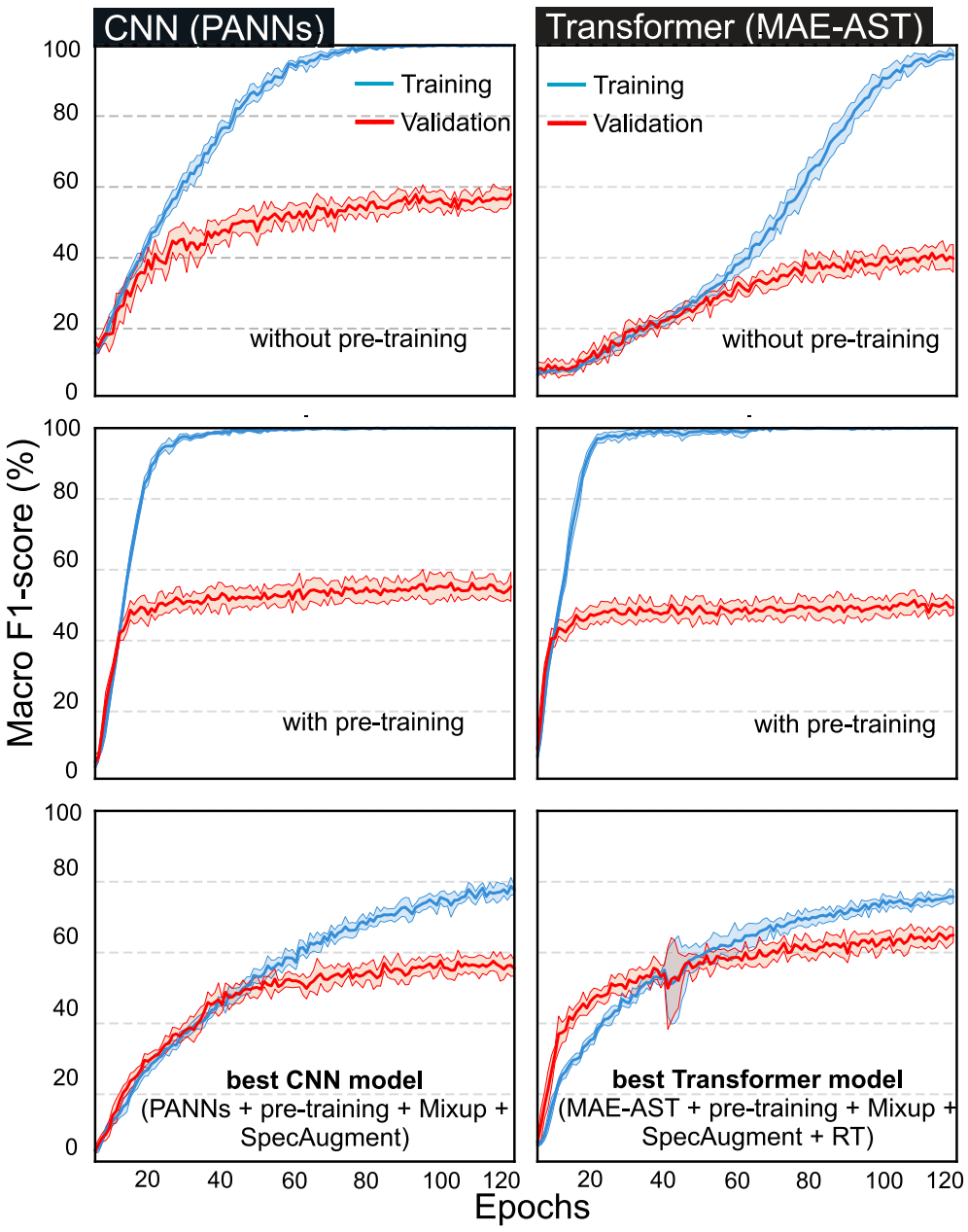


Fig. 5. Learning curves showing the performances of CNN and transformer models performances in training and cross-validation sets. The blue lines represent the performance of the models on the training set, while the red curves represent the performance of the models on the validation set. A large gap between the training and validation lines indicates overfitting. The shaded areas of the training Macro F1-score (expressed in blue) and validation Macro F1-score (expressed in red) indicate the variance of the estimates. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

have a greater capacity to capture complex patterns in bee buzzing sounds; however, their performance benefits significantly from appropriate pre-training. The application of pre-training significantly improved the performance of transformers, confirming their superiority over CNNs in the acoustic recognition of bee species, in line with our expectations outlined in Hypothesis 2.

4.2. Retaining environmental noise does not necessarily improve model performance, but there are advantages for real-world applications

It is well established that the efficacy of signal denoising directly correlates with the quality of the output from subsequent processes and, ultimately, classification performance Xie et al. (2021). The absence of denoising techniques may impede the extraction of meaningful information from raw field-collected audio data, particularly in the context of ML-based methods. However, while the removal of noise may enhance efficiency by reducing the overall volume of data, excessive removal (cleaning or filtering) may result in the loss of crucial information from the original signals (Napier et al., 2024). Furthermore, this is a less significant concern in the context of DL-based methods, which tend to be more resilient to noise (Brown et al., 2017). Therefore, rather than filtering out environmental noises during the acoustic preprocessing stage, we recommend saving them for further analysis, as we have done. The potential benefits of integrating environmental noise into bioacoustic analysis with DL models lie in the development of more robust models that are closer to real-world conditions, even if it means a reduction in performance. Therefore, CNNs and especially transformer-based models seem to have the potential to achieve reasonable accuracy in detecting and classifying bee buzzing amidst other environmental

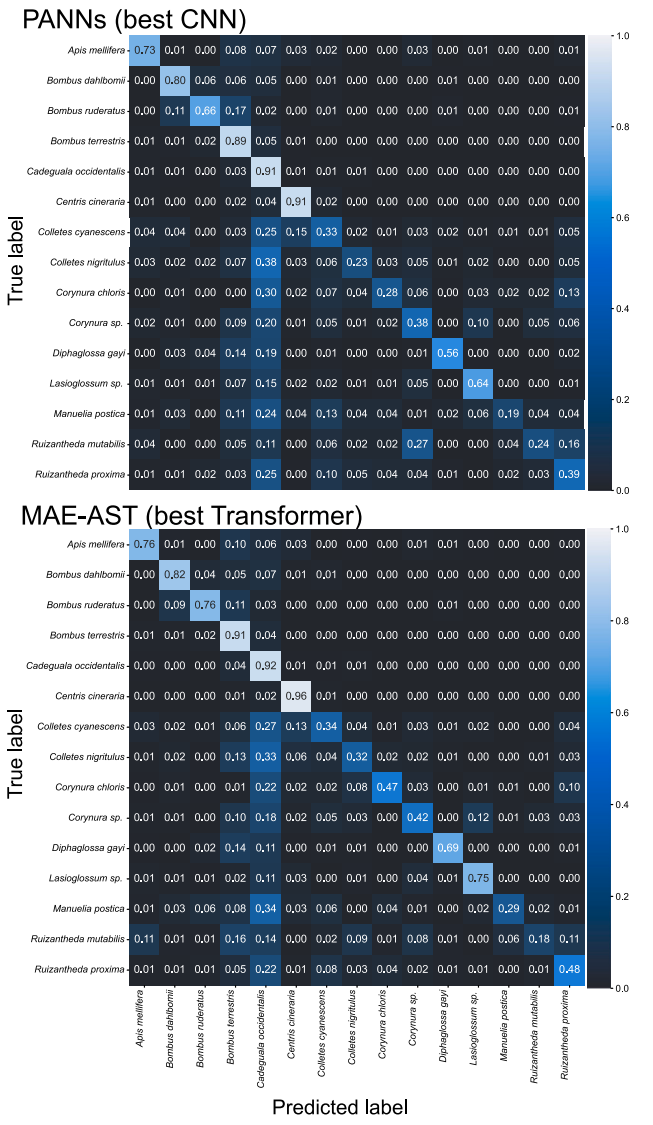


Fig. 6. The confusion matrices depict the number of audio segments correctly assigned to each bee identity (diagonal elements) versus those incorrectly assigned (non-diagonal elements) by the best transformer model (MAE-AST combined with SpecAugment, RT, Mixup, and pre-training) and the best CNN model (PANNs combined with SpecAugment, Mixup, and pre-training). Cell color signifies the corresponding count (log-transformed) of predicted audio segments, ranging from black (indicating zero predicted audio segments) to lightest blue (representing all audio segments correctly predicted). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

noise sources.

A recent alternative for dealing with excessive environmental noise is automated noise reduction using DL models, initially tested on birds (Zhang et al., 2024). With these models, it is possible to achieve a significantly shorter average noise reduction time for the test set compared to traditional bioacoustic noise reduction methods (Priyadarshani et al., 2016; Brown et al., 2017). However, automated noise reduction with DL models relies heavily on a large number of labeled, clean recordings for model training (Zhang et al., 2024). As far as we know, there is a lack of audio datasets of noise-free bee buzzes. Therefore, further studies must first address the lack of clean datasets before applying DL models for automated noise reduction.

4.3. Not only data imbalance but especially small data sets are the main challenges limiting the performance of transformer models

In general, the performance of the best transformer model at recognizing bee species based on acoustic cues from their buzzing sounds is reasonable, with accuracies exceeding 80%. However, this success is tempered by a persistent problem in our audio data set, namely data imbalance, which mainly affects the underrepresented bee taxonomic classes. Data imbalance is a significant factor that affects the stability and effectiveness of not only Transformers, but all ML algorithms (Chawla, 2010; Fundel et al., 2023). Imbalanced data can significantly skew the performance of classifiers, leading to a prediction bias in favor of the majority class (Wang et al., 2016). Nevertheless, class imbalance is a common occurrence in real-world data sets, especially when it comes to species abundance, which naturally varies among species, locations, and seasons (Fundel et al., 2023). Thus, this bias is inherent within our system since bees spontaneously visit the flowers at different frequencies, and the local availability of individuals per species also naturally varies. On the other hand, imbalanced classifications pose a unique challenge for predictive modeling, as most ML algorithms for classification assume an equal number of examples for each class. This often results in models with poorer predictive performance, especially for the minority classes. This is a critical issue because in some cases the minority classes are very important, making the problem more sensitive to classification errors within that class. Despite the difficulty posed by our inherently imbalanced dataset, we preserved the diversity and balance we found in the field to better reflect the actual species diversity, even if it meant sacrificing performance in our ML models.

In our specific case, the majority of classes are those of bee species commonly found in blueberry flowers and most relevant to crop pollination (Cortés-Rivas et al., 2023a,b). While these bees are frequent visitors to the flowers, they can contribute differently and even adversely to blueberry pollination (Cortés-Rivas et al., 2023a,b). For example, although honeybees, bumblebees, and some wild solitary bees are the majority classes here, bumblebees and some wild solitary bees using floral sonication behavior contributed more to blueberry pollination (i.e. pollen grain transfer to floral stigmas) than honeybees and other non-sonicating wild species (Cortés-Rivas et al., 2023a,b). Therefore, the effect of class imbalance is somewhat reduced in practice because the most frequent flower visitors are also the most relevant agents for both functional roles, whether they are effective pollinators or pollen/nectar thieves. As a result, while Transformer models may not excel at classifying all bee species, they show better performance at discriminating among the set of most frequent bee species for blueberry pollination.

A particular problem with transformers is that they require large amounts of training data to achieve higher performance than standard CNNs (Dosovitskiy et al., 2021). Indeed, many transformer models experience a significant drop in performance when working with insufficient training data, like the one we used. In other words, detection transformers require a large amount of data, characterizing them as data-hungry models (Wang et al., 2022). However, acquiring and labeling a training data set of audio recordings for multiple bee species can be a long and tedious task (Christin et al., 2019). Labeled training datasets rely on human judgment, as we still rely on traditional taxonomic recognition of bees to validate our model classifications. In addition, audio data collection is limited to the intrinsic availability of bee species, thus limiting the number of samples for less common species (see also Ribeiro et al. (2021) and Ferreira et al. (2023)). To alleviate the need for data-intensive training examples, several solutions have emerged in recent years and are readily available, the most popular being importing sounds from public datasets, crowdsourcing, transfer learning, and data augmentation (Christin et al., 2019). Since labeled training datasets depend on the traditional taxonomic recognition of bees, which relies on human judgments to recognize morphological features at the microscopic level (Gradišek et al., 2017),

and specialized public datasets are scarce, the use of data augmentation seems to be the most likely option to mitigate this data-intensive problem. Indeed, data augmentation played an important role in increasing the diversity and robustness of our dataset. Among all the pre-training we used, data augmentation regularization (SpecAugment, Random Truncation, and Mixup) may contribute the most to fully exploit the potential of transformers for the acoustic recognition of bee species. While excessive data augmentation does not necessarily enhance the performance of CNNs, transformers exhibited improved performance when all data augmentation methods were applied, compared to using the algorithm alone. This may be because the data augmentation techniques work together to alleviate the intrinsic need for few-shot of transformer models (Kumar et al., 2019, 2020; Ghani et al., 2023). Thus, our results underscore the greater importance of data augmentation techniques to alleviate the data-intensive nature of Transformer compared to CNN models.

While overparameterization can lead to overfitting, data hunger reflects a model's need for substantial training data to achieve strong performance. Insufficient data relative to model complexity often results in overfitting. Data hunger emphasizes the necessity for large datasets to train models effectively. In contrast, overparameterization refers to a model containing more parameters (weights) than are essential for the task at hand (Wang et al., 2022). Our Transformer models exhibit data hunger, as we discussed before, but overparameterization is less of a concern. Recent studies on scaling laws indicate that while parameter count is crucial, model performance depends on a finely tuned-relationship between parameters, data, and computational resources (Kaplan et al., 2020; Rosenfeld, 2021; Bahri et al., 2024). Our Transformer-based models, with 87–99 million parameters, were similar in size to our best-performing CNN baseline (PANNs), which contained 81 million parameters (see Table 1). This alignment suggests that our Transformer models performance gains were due to architectural innovation, not just over-parameterization. This finding is supported by results from the Vision Transformer (ViT; Dosovitskiy et al., 2021), which underlies our Transformer models. Architectural refinements, paired with increased data usage, have enabled Transformers to surpass CNNs on various computer vision tasks (Dosovitskiy et al., 2021). While model performance predictably scales with model size, it requires careful balancing of model complexity, dataset size, and computational resources (Thompson et al., 2020). Given our limited and imbalanced dataset, we employ preprocessing techniques and data augmentation to facilitate efficient learning. This approach demonstrates that Transformers can perform well even with limited data and resources if an appropriate training strategy is applied. We mitigated overfitting by balancing model capacity with dataset size and enhancing sampling efficiency through data augmentation. The comparable parameter counts across architectures, combined with our preprocessing efforts, suggest that the performance gains were due to the transformer's inductive biases, which are particularly well suited for acoustic bee detection, rather than parameter counts alone (see also Dosovitskiy et al. (2021)).

4.4. The field of bee sound processing evolves in tandem with cutting-edge research

Recently developed ML models, such as DL, have not yet fully realized their potential to automate the acoustic recognition of bee species. Transformers have achieved a maximum Macro F1-score and accuracy of 64% and 82%, respectively. However, these cutting-edge models combined with sophisticated data pre-training and the availability of larger data sets outperformed CNNs, the previous best model for automatic bee species recognition. CNNs, on the other hand, had recently outperformed classical ML models (Ferreira et al., 2023). Thus, the success of various DL techniques has been demonstrated in the development and implementation of speech recognition systems for bioacoustic applications worldwide (Rodrigues et al., 2021). Recent research suggests that task-specific model designs and training approaches for audio event recognition can achieve performance levels comparable to complex architectures used in other domains (Stowell, 2022; Ghani et al., 2023). Nevertheless, as important as it is to implement models from other domains, it will be necessary to develop specialized models and training schemes that use ambient sounds, including real-world noise, to improve the accuracy and reliability of bioacoustic analysis in more specific domains. There is no one-size-fits-all model for pattern recognition in DL, and the choice of the optimal approach depends on the specific problem and the characteristics of a particular dataset. As a future direction for new research, the possibility of domain-specific pre-training for bioacoustics is likely to be explored.

Furthermore, while we relied on pre-training techniques, such as data augmentation, to address the challenges of bioacoustic classification of bee species, there is an opportunity to further improve the performance of neural networks by pre-training them using public repositories containing recorded data of bee buzzing sounds. To our knowledge, there has been no research on pre-training Transformers using publicly available repositories tailored to the unique characteristics and challenges of bioacoustic data. Therefore, computational bioacoustics would benefit greatly, especially those based on extensive in-field human audio data collection. Despite these potential benefits, online public data platforms must meet minimum standards for reliable data curation and sharing. Data curation plays a critical role in ensuring this reliability by organizing, cleaning, and maintaining the data to improve its accuracy, consistency, and usability. Proper curation helps mitigate errors, reduce bias, and ensure that the dataset remains relevant and trustworthy over time (Freitas and Curry, 2016). With these precautions, this avenue of investigation could lead to improved models with a better understanding of the acoustic patterns specific to the field of bioacoustic classification of bee species, ultimately improving the accuracy and effectiveness of bioacoustic classification tasks.

By exploring these avenues, future work could improve the applicability and performance of, for example, few-shot learning strategies. The use of few-shot learning strategies for bee pattern recognition could be a valuable approach, especially when dealing with classes that have a limited number of samples. Few-shot learning techniques are designed to train models on tasks with a small number of examples, making them suitable for scenarios where the availability of training data is low (Ghani et al., 2023). In the context of bee species recognition, where classes are highly imbalanced, such as *Cadeguala albopilosa* with only five samples (see Table 2), the use of few-shot learning methods can have potential benefits, such as improving generalization, adaptability, data efficiency, and transfer learning. However, it is important to consider potential challenges and limitations, such as the risk of overfitting due to limited data and the need for careful selection of few-shot learning methods.

Deep learning has revolutionized the field of artificial intelligence by providing sophisticated models for a wide range of applications within bioacoustics (Stowell, 2022). However, DL models are typically black box models where the reason for the predictions is unknown (Hassija et al., 2024). Consequently, model reliability is questionable in many circumstances (Qamar and Bawany, 2023). Nevertheless, bioacoustic studies have explored which acoustic features may be relevant for taxonomic recognition of bee taxa, from which we can speculate about those that may also be important for DL model learning. A key acoustic feature is the particular buzzing frequencies, which tend to vary among different bee genera and species, with bee species differing in the frequency of floral vibrations even when visiting the same plant species (Rosi-Denadai et al., 2018; De Luca et al., 2014; Vallejo-Marín, 2019). Buzz pollination vibrations contain a fundamental frequency (typically 100–400 Hz) and often many higher frequency harmonics of rapidly decreasing magnitude (Szyszkowski and King, 1993). The frequencies generated during buzzing vary less than their duration, mainly because the frequency depends on inherent physical and physiological properties of the vibration-generating bee flight muscles and

their transmitting mechanisms, i.e. bee size and flight frequency, staminal resonance (Burkart et al., 2011). Other energy-related acoustic parameters of sound may be relevant to the task of acoustic identification of bee species, such as sound amplitude. However, amplitude depends upon the measurement procedures (e.g., recorder model and configuration, distance from the focal object) and does not necessarily correspond to vibration amplitude (Gradišek et al., 2017; De Luca et al., 2018). Therefore, since bee species tend to have different floral buzz and wingbeat frequency patterns, we would expect that fundamental frequency and harmonics can be the main features that help both CNN and transformer models to assign a buzzing sound to a given bee species. This assumption needs to be further tested.

Finally, one might ask whether these models have some practical value. For applications of AI-based approaches in fields such as medicine, high accuracy thresholds are essential because of the potential impact of a single error (Begoli et al., 2019; Chua et al., 2023). However, the context is different for species recognition in ecological studies, such as our work on blueberry pollinators. There is generally more uncertainty in this domain, and the tendency is to capture patterns and insights within complex, variable systems rather than to make precise predictions on a case-by-case and context-dependent basis (Catford et al., 2022). This uncertainty may be due to large amounts of unexplained variance resulting from unmeasured complex interactions between abiotic and biotic factors, or large amounts of randomness and noise in the data (Møller and Jennions, 2002; Fischer, 2019). Our Transformer model may not reach the ML-level accuracy standards found in more deterministic fields, but it still provides valuable predictions about pollinator identity. Given the challenges of data availability and the inherent variability of the ecological systems in which species are inserted (Fischer, 2019), the performance of these Transformer models represents a significant step forward. Thus, while the performance of the algorithms may not meet medical standards, they are practical and useful in the context of ecological interactions, helping to address important questions about pollinator contributions to crops such as blueberries, where absolute precision is less important than identifying general trends and interactions.

In summary, we compared the performance of the newly discovered transformer models with the current best models for automatic recognition of blueberry-pollinating bees by their buzzing sounds. Transformers demonstrate a superior ability to capture complex patterns in bee buzzing sounds, but their performance benefits greatly from appropriate pre-training and data augmentation techniques. Our findings indicate that neural networks using transformers, powered by a combination of pre-training techniques and robust data augmentation, outperformed the conventional CNNs in automated taxonomic recognition of bee species visiting flowers of cultivated blueberry in Chile. Nonetheless, there is still potential for greater enhancement in the performance of transformers. Further studies combining pre-training with data augmentation will be crucial to increase the diversity and robustness of training data for acoustic bee species recognition. Finally, we would like to emphasize that our pipeline is not limited to the acoustic recognition of bee species and could be applied to other domains. One particularly promising avenue for its use is in the sound classification of other flying insects, where a robust model like ours may offer advantages over other approaches that are constrained by the limitations of other modeling approaches for in acoustically distinguishing species.

CRediT authorship contribution statement

Alef Iury Siqueira Ferreira: Writing – original draft, Visualization, Software, Methodology, Formal analysis, Conceptualization. **Nádia Felix Felipe da Silva:** Writing – review & editing, Visualization, Supervision, Methodology, Investigation, Conceptualization. **Fernanda Neiva Mesquita:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Conceptualization. **Thierson Couto Rosa:** Writing – review & editing, Supervision, Methodology, Investigation, Conceptualization. **Stephen L. Buchmann:** Writing – review & editing, Resources, Investigation, Conceptualization. **José Neiva Mesquita-Neto:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.

Acknowledgments

The authors thank the two anonymous reviewers who provided constructive feedback to improve this paper and the ANID/CONICYT FONDECYT Regular for funding this work under grant No. 1231212. Powered@NLHPC: This research was partially supported by the supercomputing infrastructure of the NLHPC (ECM-02). The authors also thank the Artificial Intelligence Lab at Recod.ai, the Institute of Computing, University of Campinas (Unicamp), and the Center of Excellence in Artificial Intelligence (CEIA) at the Federal University of Goiás (UFG) for providing the necessary computational resources to run our experiments. Furthermore, the authors thank the Artificial Intelligence and Cognitive Architectures Hub (H.IAAC) for their invaluable support.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ecoinf.2025.103010.

Data availability

The data set used for this study is available in the Zenodo repository (https://doi.org/10.5281/zenodo.14611302); the code is available via GitHub (https://github.com/alefiury/Transformers-Bee-Species-Acoustic-Recognition).

References

Abayomi-Alli, O.O., Damaševičius, R., Qazi, A., Adedoyin-Olowe, M., Misra, S., 2022. Data augmentation and deep learning methods in sound classification: A systematic review. Electronics 11 (22), http://dx.doi.org/10.3390/electronics11223795, URL https://www.mdpi.com/2079-9292/11/22/3795.

Alpaydm, E., 1999. Combined 5× 2 cv f test for comparing supervised classification learning algorithms. Neural Comput. 11 (8), 1885–1892.

Ascher, J., Pickering, J., 2020. Discover life bee species guide and world checklist (Hymenoptera: Apoidea: Anthophila).

Atito, S., Awais, M., Kittler, J., 2021. Sit: Self-supervised vision transformer. arXiv preprint arXiv:2104.03602.

Baade, A., Peng, P., Harwath, D., 2022. Mae-ast: Masked autoencoding audio spectrogram transformer. arXiv preprint arXiv:2203.16691.

Baade, A., Peng, P., Harwath, D., 2022a. MAE-AST: Masked Autoencoding Audio Spectrogram Transformer - GitHub Repository. https://github.com/AlanBaade/MAE-AST-Public. (Accessed 10 May 2023).

Bahri, Y., Dyer, E., Kaplan, J., Lee, J., Sharma, U., 2024. Explaining neural scaling laws. Proc. Natl. Acad. Sci. 121 (27), e2311878121.

Banda, H., Paxton, R., 1990. Pollination of greenhouse tomatoes by bees. In: VI International Symposium on Pollination 288. pp. 194–198.

Bayraktar, U., Kilimci, H., Kilinc, H.H., Kilimci, Z.H., 2023. Assessing audio-based transformer models for speech emotion recognition. In: 2023 7th International Symposium on Innovative Approaches in Smart Technologies. ISAS, IEEE, pp. 1–7.

Begoli, E., Bhattacharya, T., Kusnezov, D., 2019. The need for uncertainty quantification in machine-assisted medical decision making. Nat. Mach. Intell. 1 (1), 20–23.

Benjamin, F.E., Winfree, R., 2014. Lack of pollinators limits fruit production in commercial blueberry (Vaccinium corymbosum). Environ. Entomol. 43 (6), 1574–1583.

Brown, A., Garg, S., Montgomery, J., 2017. Automatic and efficient denoising of bioacoustics recordings using mmse stsa. IEEE Access 6, 5010–5022.

Buchmann, S.L., et al., 1983. Buzz pollination in angiosperms. Buzz Pollinat. Angiosperms 28 (1), 73–113.

Burkart, A., Lunau, K., Schlindwein, C., 2011. Comparative bioacoustical studies on flight and buzzing of neotropical bees. J. Pollinat. Ecol. 6 (2), 491–596.

Catford, J.A., Wilson, J.R., Pyšek, P., Hulme, P.E., Duncan, R.P., 2022. Addressing context dependence in ecology. Trends Ecol. Evolut. 37 (2), 158–170.

Chawla, N.V., 2010. Data mining for imbalanced datasets: An overview. Data Min. Knowl. Discov. Handb. 875–886.

Chlap, P., Min, H., Vandenberg, N., Dowling, J., Holloway, L., Haworth, A., 2021. A review of medical image data augmentation techniques for deep learning applications. J. Med. Imaging Radiat. Oncol. 65 (5), 545–563.
Christin, S., Hervet, É., Lecomte, N., 2019. Applications for deep learning in ecology. Methods Ecol. Evol. 10 (10), 1632–1644.
Chua, M., Kim, D., Choi, J., Lee, N.G., Deshpande, V., Schwab, J., Lev, M.H., Gonzalez, R.G., Gee, M.S., Do, S., 2023. Tackling prediction uncertainty in machine learning for healthcare. Nat. Biomed. Eng. 7 (6), 711–718.
Cooley, H., Vallejo-Marín, M., 2021. Buzz-pollinated crops: A global review and meta-analysis of the effects of supplemental Bee Pollination in Tomato. J. Econ. Entomol. 14 (1), 179–213.
Cortés-Rivas, B., Monzón, V.H., Rego, J.O., Mesquita-Neto, J.N., 2023a. Pollination by native bees achieves high fruit quantity and quality of highbush blueberry: a sustainable alternative to managed pollinators. Front. Sustain. Food Syst. 7, 114–262.
Cortés-Rivas, B., Smith-Ramirez, C., Monzón, V.H., Mesquita-Neto, J.N., 2023b. Native bees with floral sonication behaviour can achieve high-performance pollination of highbush blueberry in Chile. Agric. for Entomol. 25 (1), 91–102.
De Luca, P.A., Cox, D.A., Vallejo-Marín, M., 2014. Comparison of pollination and defensive buzzes in bumblebees indicates species-specific and context-dependent vibrations. Naturwissenschaften 101 (4), 331–338.
De Luca, P.A., Giebink, N., Mason, A.C., Papaj, D., Buchmann, S.L., 2018. How well do acoustic recordings characterize properties of bee (Anthophila) floral sonication vibrations? Bioacoustics 29 (1), 1–14.
De Luca, P.A., Vallejo-Marin, M., 2013. What's the 'buzz'about? The ecology and evolutionary significance of buzz-pollination. Curr. Opin. Plant Biol. 16 (4), 429–435.
Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255. http://dx.doi.org/10.1109/CVPR.2009.5206848.
Dietterich, T.G., 1998. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. Neural Comput. 10 (7), 1895–1923. http://dx.doi.org/10.1162/089976698300017197.
Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations. pp. 1–22.
Elliott, D., Otero, C.E., Wyatt, S., Martino, E., 2021. Tiny transformers for environmental sound classification at the edge. arXiv preprint arXiv:2103.12157.
Ferreira, A.I.S., da Silva, N.F.F., Mesquita, F.N., Rosa, T.C., Monzón, V.H., Mesquita-Neto, J.N., 2023. Automatic acoustic recognition of pollinating bee species can be highly improved by Deep-Learning models accompanied by pre-training and strong data augmentation. Front. Plant Sci. 14, 11–51.
Fischer, M.M., 2019. Quantifying the uncertainty of variance partitioning estimates of ecological datasets. Environ. Ecol. Stat. 26 (4), 351–366.
Francoy, T.M., de Faria Franco, F., Roubik, D.W., 2012. Integrated landmark and outline-based morphometric methods efficiently distinguish species of Euglossa (Hymenoptera, Apidae, Euglossini). Apidologie 43, 609–617.
Freitas, A., Curry, E., 2016. Big data curation. In: New Horizons for a Data-Driven Economy: a Roadmap for Usage and Exploitation of Big Data in Europe. pp. 87–118.
Fundel, F., Braun, D.A., Gottwald, S., 2023. Automatic bat call classification using transformer networks. Ecol. Inform. 78, 102288.
Gao, Y., Xue, X., Qin, G., Li, K., Liu, J., Zhang, Y., Li, X., 2024. Application of machine learning in automatic image identification of insects-a review. Ecol. Inform. 102539.
Gaston, K.J., O'Neill, M.A., 2004. Automated species identification: why not? Philos. Trans. R. Soc. London [Biol] 359 (1444), 655–667.
Gemmeke, J.F., Ellis, D.P.W., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M., 2017. Audio Set: An ontology and human-labeled dataset for audio events. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 776–780. http://dx.doi.org/10.1109/ICASSP.2017.7952261.
Ghani, B., Denton, T., Kahl, S., Klinck, H., 2023. Global birdsong embeddings enable superior transfer learning for bioacoustic classification. Sci. Rep. 13 (1), 22–76.
Gong, Y., Chung, Y.-A., Glass, J., 2021a. Ast: Audio spectrogram transformer. arXiv preprint arXiv:2104.01778.
Gong, Y., Chung, Y.-A., Glass, J., 2021b. AST: Audio Spectrogram Transformer - GitHub Repository. https://github.com/YuanGongND/ast. (Accessed 10 May 2023).
Gong, Y., Chung, Y.A., Glass, J., 2021c. PSLA: Improving audio tagging with pretraining, sampling, labeling, and aggregation. IEEE/ ACM Trans. Audio Speech Lang. Proc. 29, 3292–3306. http://dx.doi.org/10.1109/TASLP.2021.3120633.
Gong, Y., Lai, C.-I., Chung, Y.-A., Glass, J., 2022. SSAST: Self-Supervised Audio Spectrogram Transformer - GitHub Repository. https://github.com/YuanGongND/ssast. (Accessed 10 May 2023).
Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press, http://www.deeplearningbook.org.
Gradišek, A., Slapničar, G., Šorn, J., Luštrek, M., Gams, M., Grad, J., 2017. Predicting species identity of bumblebees through analysis of flight buzzing sounds. Bioacoustics 26 (1), 63–76.
Greenleaf, S.S., Kremen, C., 2006. Wild bees enhance honey bees' pollination of hybrid sunflower. Proc. Natl. Acad. Sci. 103 (37), 13890–13895.
Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., Lew, M.S., 2016. Deep learning for visual understanding: A review. Neurocomputing 187, 27–48.
Gwardys, G., Grzywczak, D., 2014. Deep image features in music information retrieval. Int. J. Electron. Telecommun. 60, 321–326.
Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M., Hussain, A., 2024. Interpreting black-box models: a review on explainable artificial intelligence. Cogn. Comput. 16 (1), 45–74.
Hendrycks, D., Lee, K., Mazeika, M., 2019. Using pre-training can improve model robustness and uncertainty. In: International Conference on Machine Learning. PMLR, pp. 2712–2721.
Hikawa, M., 2004. Effects of pollination by honeybees on yield and the rate of unmarketable fruits in forcing eggplant [Solanum melongena] cultures. Hortic. Res. (Jpn.).
Islam, S., Haque, M.M., Sadat, A.J.M., 2023. Capturing spectral and long-term contextual information for speech emotion recognition using deep learning techniques. arXiv preprint arXiv:2308.04517.
Javorek, S., Mackenzie, K., Vander Kloet, S., 2002. Comparative pollination effectiveness among bees (Hymenoptera: Apoidea) on lowbush blueberry (Ericaceae: Vaccinium angustifolium). Ann. Entomol. Soc. Am. 95 (3), 345–351.
Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D., 2020. Scaling laws for neural language models. arXiv preprint arXiv:2001.08361.
Kim, Y., Jo, Y., Lee, S., Lee, M., Yoon, H., Lee, M., Nam, S., 2005. The comparison of pollinating effects between honeybees (Apis mellifera) and bumblebee (Bombus terrestris) on the Kiwifruit raised in greenhouse. Korean J. Apic..
Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.
Klein, A.M., Vaissière, B.E., Cane, J.H., Steffan-Dewenter, I., Cunningham, S.A., Kremen, C., Tscharntke, T., 2007. Importance of pollinators in changing landscapes for world crops. Proc. R. Soc. B: Biol. Sci. 274 (1608), 303–313.
Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., Plumbley, M.D., 2020a. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. IEEE/ ACM Trans. Audio Speech Lang. Process. 28, 2880–2894.
Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., Plumbley, M.D., 2020b. PANNs: Large-scale pretrained audio neural networks for audio pattern recognition. IEEE/ ACM Trans. Audio Speech Lang. Process. 28, 2880–2894. http://dx.doi.org/10.1109/TASLP.2020.3030497.
Kumar, V., Choudhary, A., Cho, E., 2020. Data augmentation using pre-trained transformer models. In: Campbell, W.M., Waibel, A., Hakkani-Tur, D., Hazen, T.J., Kilgour, K., Cho, E., Kumar, V., Glaude, H. (Eds.), Proceedings of the 2nd Workshop on Life-Long Learning for Spoken Language Systems. Association for Computational Linguistics, Suzhou, China, pp. 18–26.
Kumar, V., Glaude, H., de Lichy, C., Campbell, W., 2019. A closer look at feature space data augmentation for few-shot intent classification. In: Cherry, C., Durrett, G., Foster, G., Haffari, R., Khadivi, S., Peng, N., Ren, X., Swayamdipta, S. (Eds.), Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019). Association for Computational Linguistics, Hong Kong, China, pp. 1–10. http://dx.doi.org/10.18653/v1/D19-6101.
Kumar, A.S., Schlosser, T., Kahl, S., Kowerko, D., 2024. Improving learning-based birdsong classification by utilizing combined audio augmentation strategies. Ecol. Inform. 82, 102699. http://dx.doi.org/10.1016/j.ecoinf.2024.102699, URL https://www.sciencedirect.com/science/article/pii/S1574954124002413.
Li, X., Shao, N., Li, X., 2023. Self-supervised audio teacher-student transformer for both clip-level and frame-level tasks. arXiv preprint arXiv:2306.04186.
Lippert, C., Feuerbacher, A., Narjes, M., 2021. Revisiting the economic valuation of agricultural losses due to large-scale changes in pollinator populations. Ecol. Econom. 180, 106–860.
Macias-Macias, O., Chuc, J., Ancona-Xiu, P., Cauich, O., Quezada-Euán, J., 2009. Contribution of native bees and africanized honey bees (Hymenoptera: Apoidea) to Solanaceae crop pollination in tropical México. J. Appl. Entomol. 133 (6), 456–465.
Mesquita-Neto, J.N., Vieira, A.L.C., Schlindwein, C., 2021. Minimum size threshold of visiting bees of a buzz-pollinated plant species: consequences for pollination efficiency. Am. J. Bot..
Møller, A., Jennions, M.D., 2002. How much variance can be explained by ecologists and evolutionary biologists? Oecologia 132, 492–500.
Müller, R., Ritz, F., Illium, S., Linnhoff-Popien, C., 2020. Acoustic anomaly detection for machine sounds based on image transfer learning. CoRR abs/2006.03429 arXiv:2006.03429 URL https://arxiv.org/abs/2006.03429.
Napier, T., Ahn, E., Allen-Ankins, S., Schwarzkopf, L., Lee, I., 2024. Advancements in preprocessing, detection and classification techniques for ecoacoustic data: A comprehensive review for large-scale passive acoustic monitoring. Expert Syst. Appl. 124220.
Neuenschwander, A., Zabaleta, C., Francisco, J., et al., 2010. El cambio climático en el sector silvoagropecuario de Chile. Fundactión Innov. Agrar. 12 (6), 1091–1106.
O'Mahony, N., Campbell, S., Carvalho, A., Harapanahalli, S., Hernandez, G.V., Krpalkova, L., Riordan, D., Walsh, J., 2020. Deep learning vs. traditional computer vision. In: Advances in Computer Vision: Proceedings of the 2019 Computer Vision Conference (CVC), Volume 1 1. Springer, pp. 128–144.

Orr, M.C., Hughes, A.C., Chesters, D., Pickering, J., Zhu, C.-D., Ascher, J.S., 2020. Global patterns and drivers of bee distribution. Curr. Biol. 50 (3), 53–78.
Oswald, J.N., Erbe, C., Gannon, W.L., Madhusudhana, S., Thomas, J.A., 2022. Detection and classification methods for animal sounds. Explor. Anim. Behav. Through Sound 1, 269–317.
Palanisamy, K., Singhania, D., Yao, A., 2020. Rethinking CNN models for audio classification. http://dx.doi.org/10.48550/ARXIV.2007.11154, URL https://arxiv.org/abs/2007.11154.
Park, D.S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E.D., Le, Q.V., 2019. Specaugment: A simple data augmentation method for automatic speech recognition. arXiv preprint arXiv:1904.08779.
Pomeroy, N., Fisher, R.M., 2002. Pollination of kiwifruit (actinidia deliciosa) by bumble bees (Bombus terrestris): effects of bee density and patterns of flower visitation. N. Z. Èntomol. 25 (1), 41–49.
Potts, S.G., Imperatriz Fonseca, V., Ngo, H.T., Biesmeijer, J.C., Breeze, T.D., Dicks, L., Garibaldi, L.A., Hill, R., Settele, J., Vanbergen, A.J., et al., 2016. Summary for policymakers of the assessment report of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services on pollinators, pollination and food production. Intergov. Sci. Platf. Biodivers. Ecosyst. Serv..
Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al., 2011. The Kaldi speech recognition toolkit. In: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society, pp. 1–4.
Priyadarshani, N., Marsland, S., Castro, I., Punchihewa, A., 2016. Birdsong denoising using wavelets. PLoS One 11 (1), e0146790.
Qamar, T., Bawany, N.Z., 2023. Understanding the black-box: towards interpretable and reliable deep learning models. PeerJ Comput. Sci. 9, e1629.
Ribeiro, A.P., da Silva, N.F.F., Mesquita, F.N., Araújo, P.d.C.S., Rosa, T.C., Mesquita-Neto, J.N., 2021. Machine learning approach for automatic recognition of tomato-pollinating bees based on their buzzing-sounds. PLoS Comput. Biol. 17 (9), e1009426.
Rodrigues, J.F., Florea, L., de Oliveira, M.C., Diamond, D., Oliveira, O.N., 2021. Big data and machine learning for materials science. Discov. Mater. 1, 1–27.
Rosenfeld, J.S., 2021. Scaling laws for deep learning. arXiv preprint arXiv:2108.07686.
Rosi-Denadai, C.A., Araújo, P.C.S., Campos, L.A.d., Cosme Jr., L., Guedes, R.N.C., 2018. Buzz-pollination in Neotropical bees: genus-dependent frequencies and lack of optimal frequency for pollen release. Insect Sci. 27 (1), 133–142.
Rucker, R.R., Thurman, W.N., Burgett, M., 2012. Honey bee pollination markets and the internalization of reciprocal benefits. Am. J. Agric. Econ. 94 (4), 956–977.
Santana, F.S., Costa, A.H.R., Truzzi, F.S., Silva, F.L., Santos, S.L., Francoy, T.M., Saraiva, A.M., 2014. A reference process for automating bee species identification based on wing images and digital image processing. Ecol. Inform. 24, 248–260.
Schroder, S., Wittmann, D., Drescher, W., Roth, V., Steinhage, V., Cremers, A.B., 2002. The new key to bees: automated identification by image analysis of wings. Pollinating Bees– Conserv. Link Between Agric. Nat. Minist. Environ. Bras. 209–218.
Solís-Montero, L., Vallejo-Marín, M., 2017. Does the morphological fit between flowers and pollinators affect pollen deposition? An experimental test in a buzz-pollinated species with anther dimorphism. Ecol. Evol. 7 (8), 2706–2715.
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15 (1), 1929–1958.
Steiniger, Y., Stoppe, J., Meisen, T., Kraus, D., 2020. Dealing with highly unbalanced sidescan sonar image datasets for deep learning classification tasks. In: Global Oceans 2020: Singapore–US Gulf Coast. IEEE, pp. 1–7.
Stowell, D., 2022. Computational bioacoustics with deep learning: a review and roadmap. PeerJ 10, e13152.
Stubbs, C., Drummond, F., 1996. Blueberry and cranberry (Vaccinium spp.) pollination: a comparison of managed and native bee foraging behavior. In: VII International Symposium on Pollination 437. pp. 341–344.
Szyszkowski, W., King, J., 1993. Optimality criterion for maximum fundamental frequency of free vibrations of frames including axial and bending effects. Struct. Optim. 5, 250–255.
Thompson, N.C., Greenewald, K., Lee, K., Manso, G.F., 2020. The computational limits of deep learning. arXiv preprint arXiv:2007.05558, 10.
Troudet, J., Grandcolas, P., Blin, A., Vignes-Lebbe, R., Legendre, F., 2017. Taxonomic bias in biodiversity data and societal preferences. Sci. Rep. 7 (1), 91–132.
Truong, T.H., Du Nguyen, H., Mai, T.Q.A., Nguyen, H.L., Dang, T.N.M., Phan, T.T.H., 2023. A deep learning-based approach for bee sound identification. Ecol. Inform. 78, 102274.
Vallejo-Marín, M., 2019. Buzz pollination: studying bee vibrations on flowers. New Phytol. 224 (3), 1068–1074.
Valliammal, N., Geethalakshmi, S., 2011. Automatic recognition system using preferential image segmentation for leaf and flower images. Comput. Sci. Eng. 1 (4), 13.
Velthuis, H.H., Van Doorn, A., 2006. A century of advances in bumblebee domestication and the economic and environmental aspects of its commercialization for pollination. Apidologie 37 (4), 421–451.
Wang, S., Liu, W., Wu, J., Cao, L., Meng, Q., Kennedy, P.J., 2016. Training deep neural networks on imbalanced data sets. In: 2016 International Joint Conference on Neural Networks. IJCNN, IEEE, pp. 4368–4374.
Wang, W., Zhang, J., Cao, Y., Shen, Y., Tao, D., 2022. Towards data-efficient detection transformers. In: European Conference on Computer Vision. Springer, pp. 88–105.
Wilcock, C., Neiland, R., 2002. Pollination failure in plants: why it happens and when it matters. Trends Plant Sci. 7 (6), 270–277.
Wolters, P., Daw, C., Hutchinson, B., Phillips, L., 2021. Proposal-based few-shot sound event detection for speech and environmental sounds with perceivers. arXiv preprint arXiv:2107.13616.
Xie, J., Colonna, J.G., Zhang, J., 2021. Bioacoustic signal denoising: a review. Artif. Intell. Rev. 54, 3575–3597.
Yanikoglu, B., Aptoula, E., Tirkaz, C., 2014. Automatic plant identification from photographs. Mach. Vis. Appl. 25 (6), 1369–1383.
You, L., Coyotl, E.P., Gunturu, S., Van Segbroeck, M., 2023. Transformer-based bioacoustic sound event detection on few-shot learning tasks. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 1–5.
Yu, W., Zhu, C., Fang, Y., Yu, D., Wang, S., Xu, Y., Zeng, M., Jiang, M., 2021. Dict-bert: Enhancing language model pre-training with dictionary. arXiv preprint arXiv:2110.06490.
Zapponi, L., Mazza, G., Farina, A., Fedrigoli, L., Mazzocchi, F., Roversi, P.F., Pevierieri, G.S., Mason, F., 2017. The role of monumental trees for the preservation of saproxylic biodiversity: re-thinking their management in cultural landscapes. Nat. Conserv. 19, 231–243.
Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D., 2017. Mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412.
Zhang, C., He, K., Gao, X., Guo, Y., 2024. Automatic bioacoustics noise reduction method based on a deep feature loss network. Ecol. Inform. 80, 102517.
Zhong, M., LeBien, J., Campos-Cerqueira, M., Dodhia, R., Lavista Ferres, J., Velev, J.P., Aide, T.M., 2020. Multispecies bioacoustic classification using transfer learning of deep convolutional neural networks with pseudo-labeling. Appl. Acoust. 166, 107375. http://dx.doi.org/10.1016/j.apacoust.2020.107375, URL https://www.sciencedirect.com/science/article/pii/S0003682X20304795.