scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A dataset of insect sounds from 459 species for bioacoustic machine learning

Marius Faiß[1,2,3] ✉, Burooj Ghani[4] & Dan Stowell[4,5]

Automatic recognition of insect sound could help us understand changing biodiversity trends around the world—but insect sounds are challenging to recognize even for deep learning, due to the broad frequency ranges and limited amount of training data. We present a new dataset comprised of 26298 audio files (226.6 hours), from 459 species of Orthoptera (310 species) and Cicadidae (149 species). InsectSet459 is the first large-scale dataset of insect sound that is easily applicable for developing novel deep-learning methods. Its recordings were made with a variety of audio recorders using varying sample rates to capture the extremely broad range of frequencies that insects produce. We benchmark performance with two state-of-the-art deep learning classifiers, demonstrating good performance but also significant room for improvement in acoustic insect classification. This dataset can serve as a realistic test case for implementing insect monitoring workflows, and as a challenging basis for the development of audio representation methods that can handle highly variable frequencies and/or sample rates.

Background & Summary

Insects are crucial members of many ecosystems, and so the recent reports of insect population declines are a vital global topic[1]. Yet there is still much uncertainty about insect biodiversity and population changes, due to a poverty of monitoring information from a very limited number of species and geographic regions[2]. Automatic monitoring has a key role to play, including automatic recognition from sounds (and from images)[3–6].

There is a vast number of soniferous insect species including large clades whose sounds carry species-specific information, with two of the biggest and loudest being Orthoptera and Cicadidae[6]. To monitor population changes, distributions and rare species occurrence, automatic acoustic detection and classification methods can be applied in a wide range of environments while being minimally invasive[2] and automatable to a high degree[7,8]. The scale of this task seems daunting, due to the sheer number of insect species, but on the other hand, modern deep learning benefits from large-scale data, and certain tasks become tractable only when a certain scale is attained. For example, bird vocalization classifiers—trained on hundreds of species using deep learning—have now established themselves in the toolbox of bird monitoring techniques[9,10].

For soniferous insects, the data available for deep learning is not yet large-scale—neither in terms of audio duration, nor species coverage. In earlier work we introduced curated open datasets of insect sounds containing up to 66 species[11] of Orthopterans and Cicadas[12]. However, much larger collections are needed, in order for deep learning to become usefully deployable for acoustic insect monitoring. In the same work we evaluated a novel feature representation, based on trainable filter banks instead of spectrograms, which are commonly optimized for human-made sounds. Various other machine learning strategies may be useful for insect sound recognition, but it is not yet clear whether the best insect recognition will be based on waveforms, spectrograms, adaptive filters, or other new types of representation that suit the acoustic phenomena. There is a need for more comprehensive data, both for development of new methods and for validation of the technical readiness level of classifiers.

In 2024, the public animal sound database xeno-canto[13] (https://xeno-canto.org) witnessed a dramatic increase in insect sound recordings. This is due to the publication of several large collections of field and

[1]Department for the Ecology of Animal Societies, Max Planck Institute of Animal Behavior, Konstanz, Germany. [2]International Max Planck Research School for Quantitative Behavior, Ecology and Evolution, Konstanz, Germany. [3]Department of Biology, University of Konstanz, Konstanz, Germany. [4]Naturalis Biodiversity Centre, Leiden, The Netherlands. [5]Department of Cognitive Science and Artificial Intelligence, Tilburg University, Tilburg, The Netherlands. ✉e-mail: mfaiss@ab.mpg.de

laboratory recordings from insect sound experts, as well as increased adoption of citizen scientists uploading their insect sound observations to the website. At around the same time, the iNaturalist[14] project (https://www.inaturalist.org) has also witnessed large growth in the number of sound recordings collected, including sound from insects[15]. These two projects have in common the collection of species-labeled sound recordings, an open data philosophy (under varying licenses including multiple Creative Commons licenses), and the use of citizen science (community science) to bring together sounds from around the world. Their focus on wildlife, unlike general-purpose audio collections, encourages the community to work towards collecting representative audio clips and accurate species labeling.

We used this opportunity to compile the first large-scale dataset of insect sounds for training deep learning methods to detect and classify insect sounds in the wild. We note a recent parallel initiative to curate a general-purpose audio dataset from iNaturalist[15]. Unlike that initiative, we focused on preparing the data for the specific goal of enabling insect sounds to be better recognized through machine learning, since this is a current important gap.

In their systematic review, Kohlberg *et al.*[5] identified machine learning studies covering a total of 302 species for insect sound recognition. Our dataset thus represents an opportunity to substantially increase the coverage of bioacoustic machine learning in this domain by adding 406 species that have not been covered in machine learning studies according to the findings in Kohlberg *et al.*[5].

In this paper we introduce InsectSet459[16], an acoustic dataset of 459 insect species (26298 recordings), curated from open data sources and preprocessed to be of maximal usefulness to develop machine learning for insect sound. We describe the data curation process and the dataset characteristics. We also validate the usefulness of the dataset by applying deep learning algorithms for automatic species classification of insect sounds, demonstrating that the dataset enables high-quality acoustic recognition. We intend that others can use the dataset to create even stronger classifiers than those we demonstrate, so that intelligent acoustic monitoring can contribute to the urgent task of insect population monitoring.

Methods

Curation of InsectSet459. Our data curation was an expanded and revised version of the method of InsectSet66[12]. Recordings were downloaded and pooled from the following three sources:

- xeno-canto: Orthoptera[17]
- iNaturalist: Orthoptera[18] & Cicadidae[19]
- BioAcoustica: Cicadidae[20]

Initially, all files belonging to observations at the species-level were downloaded from xeno-canto and BioAcoustica. From iNaturalist, only research-grade observations were downloaded, meaning the species label was verified by other users. We exclude sound files with "no derivatives" license limitations to ensure re-usability of the dataset and to enable us to publish shortened versions of long audio files. All recordings in the dataset are licensed under creative commons licenses CC-BY-4.0 or CC0, which therefore means that users are permitted extensive rights to reuse, share and publish the resulting data.

Deduplication is an important curation step for machine learning, especially when multiple data sources are used. For iNaturalist observations with multiple attached audio files, only one file was downloaded. If users uploaded to both iNaturalist and xeno-canto, only the files from one of the platforms were used. Further, we performed deduplication based on sha256 checksums of the raw audio (python 3.9, hashlib.sha256()) to reduce the risk of duplicate files, especially considering that some files are in practice uploaded both to xeno-canto and to iNaturalist under different user names. We also found that some users had uploaded the same files multiple times to the same platform, sometimes even listed with differing species labels. This is likely done to indicate multiple individuals or species being present in the same audio file. This provides additional information for biodiversity monitoring purposes, but poses a problem for machine learning tasks by introducing data leakage, which is why we decided to remove duplicates. If recordings were uploaded more than once, but with the same species label, only one file was kept. Recordings that were uploaded under multiple different species names were completely excluded from the data pool, since we could not assign a focal species label. Another common source of redundancy is serial uploads from one location and time period split into separate observations (especially common on xeno-canto). Many of these likely include the same individual animals vocalizing over a period of time, but would show up as independent samples in a machine learning dataset. We addressed this problem by pooling all recordings by upload username, species label, geographic location, date and time, and then selecting only one recording from within a one-hour period. After these filtering steps, all files from species with at least 10 remaining sound examples were selected for the final dataset.

File duration and formatting. In our previous work, for InsectSet32[21] we used sources with fewer recordings available, but divided those recordings into multiple, overlapping sections. This is useful for creating enough training data for deep learning, but it does not yield a large diversity in the recording conditions, and thus it carries some risk of overfitting in training (a kind of "false replication" in experimental design). For the present work we restricted to 10+ separate recordings per species, to encourage that all of the classes that a classifier should learn are sufficiently representative and diverse. We give users of this dataset the choice of how to handle varying audio segment length and splitting strategies, by providing continuous files that are not pre-segmented.

Sound files from these platforms come in dramatically varying durations, from less than one second to many minutes. In order to maximize the acoustic diversity for a given size of dataset, we chose to trim each audio file down to a maximum duration (2 minutes). In a long audio file, we skip the first 2 minutes (since it may contain

Fold	Num files	Total duration (hours)
Train	15789	136.8
Validation	5290	46.1
Test	5219	43.7
Total	26298	226.6

Table 1. Sizes of the data folds in InsectSet459.

speech or handling noise) and take the next 2 minute segment. For those who wish to use as much audio data as possible, and thus more than 2 minutes per file, the original sources are available as links in the annotation file.

File formats were standardized to WAV and MP3, both of which are present in much of the source material, and also very widely compatible with audio tools. Less-common formats such as M4A were converted to high-quality MP3 to ensure wide compatibility and ease of use. All stereo files were converted to mono.

We chose not to reduce the sample rate down to a common value. This is a strategic difference from common practice in machine learning dataset work, and is motivated by the dramatic variation of bandwidths known in insect sound which can reach far outside of the human hearing range and can therefore not be represented at the most commonly used sample rates. We provide audio at the full sample rate given in the original source, which avoids information loss and allows for high-bandwidth and ultrasonic deep learning methods to be developed in the future.

Species labels and dataset splits. Species nomenclature was unified to the Orthoptera Species File taxonomy (COL24.4 2024-04-26 [294826]) using checklistbank[22]. Xeno-canto's metadata structure offers its users the possibility to label species that are audible in the background of a recording alongside the main species. However, since the majority of insect recordings at this time stem from batch uploads where this information was not given, we decided not to exclude these recordings from InsectSet459. In practice, field recordings of insects commonly contain sounds and chorusing from other species in the background and they are usually not labeled, especially on iNaturalist. Therefore, excluding recordings from xeno-canto where background species are indicated likely would not improve the overall quality of the dataset. We instead add this information to the metadata file in a separate column, which could potentially be useful for users of the dataset in some cases.

For machine-learning purposes, the dataset was split into the training, validation and test sets while ensuring a roughly equal distribution of audio files and audio material for every species in all three subsets. This resulted in a 60/20/20 split (train/validation/test) by file number and file length (Table 1). The final dataset is weakly labeled, with one species label per audio sample, and of variable file length and sample rate.

Data Record

The dataset is published under the CC 4.0 license on Zenodo[16]. The training, validation and testing datasets are stored in separate zip files. A csv file holds the metadata and species labels for all files in the dataset. An additional readme file gives further information about the metadata fields in the annotation file and an overview of all species in the dataset and their respective number of files and length of audio material. The metadata is given as follows:

- file_name: The file name used in this dataset, with the species name attached
- species_name: The name of the species in the recording
- group: Indicates whether the species belongs to Cicadidae or Orthoptera
- license: The Creative Commons Attribution license under which the original contributor or uploader published the source recording
- contributor: The contributor/recordist/uploader of the recording
- observation: Link to the observation record on xeno-canto or iNaturalist. Allows users to retrieve additional information associated with the observation from the source website.
- file: Direct link to the source file on xeno-canto or iNaturalist
- background:Lists of other species that are present in the recording. Only present for a subset of the xeno-canto observations
- original_name: Original species name, if it was changed to match nomenclature between datasets
- subset: The data-subset the file was included in for training, testing or validating machine learning techniques
- temperature: Temperature measured at the time of recording by the recordist in Celsius

Data Overview. This new dataset greatly increases the number of species compared against prior work, giving 459 unique species from the groups Orthoptera (310 species) and Cicadidae (149 species), while also strongly increasing the geographic coverage of recording locations (Fig. 1). The total duration of the dataset and number of sound examples is heavily expanded to a total of 26298 files containing 9.5 days of audio material (Table 1), with approximately equal amount of WAV and MP3 files. The sample rates range from 8 to 500 kHz, with roughly a third of the files containing ultrasonic frequencies (sample rates above 44 kHz; Table 2). Of the 459 species, many are in the "long tail" of the data distribution with very few recordings per species: the most prevalent have more than 500 recordings, but around half have fewer than 25 (Fig. 2). This level of class imbalance is a common aspect of datasets in ecology[23].

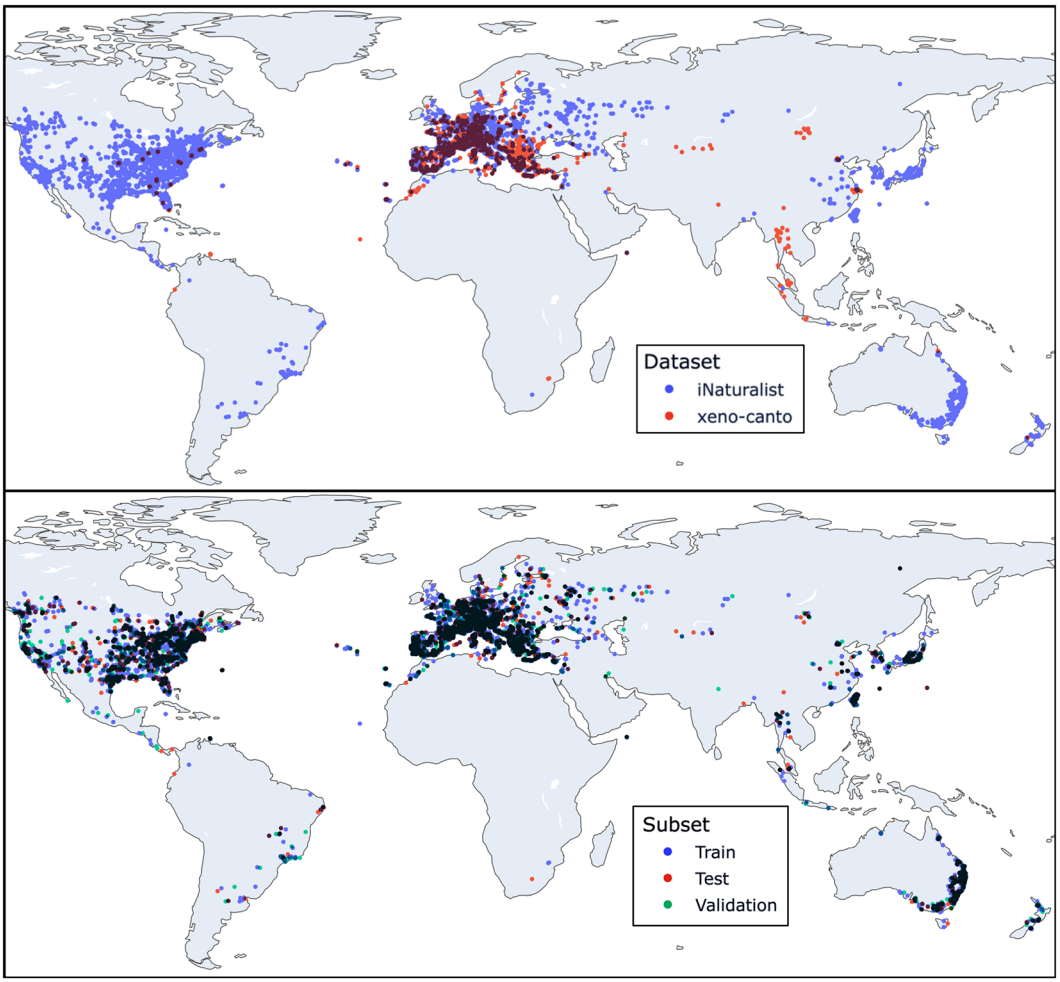

Fig. 1 The geographic locations of the files contained in InsectSet459. Out of 26298 files in the dataset, 305 do not have coordinates associated with them. (**a**) Upper plot: Recording locations split by source dataset. Blue dots indicate recordings sourced from iNaturalist, red dots show recordings from xeno-canto. Purple shows the overlap of both datasets. (**b**) Lower plot: Recording locations split into the training, validation and test sets. Blue dots indicate recordings in the training set, red dots show recordings in the testing subset and green dots show recordings in the validation subsets. Other colors show the overlap between the subsets.

Rate (kHz)	Num files	Rate (kHz)	Num files	Rate (kHz)	Num files
8	91	48	4596	192	861
11	3	49	2	200	38
16	85	50	3	250	193
22	58	63	11	256	639
24	102	64	111	300	3
26	3	88	28	313	297
32	930	96	2145	320	2
33	1	100	63	384	402
38	11	125	269	500	4
44	15344	152	3		

Table 2. Sample rates in InsectSet459, rounded to the nearest kHz.

The geographic coverage of this dataset is heavily biased towards Europe and Northern America (Fig. 1), likely due to the majority of contributors being located in these areas. Based on the recently published list of sound producing European grasshoppers[24], this dataset covers 29% of all European Orthoptera species that produce sounds meaningful for identification. In Central and Northern Europe 57% of species are represented, 47% in Iberia, 46% in Italy and Malta, 42% in the Balkan and 37% in Eastern Europe. This suggests that our dataset covers the most widely distributed species, while many endemic species are missing. According to the

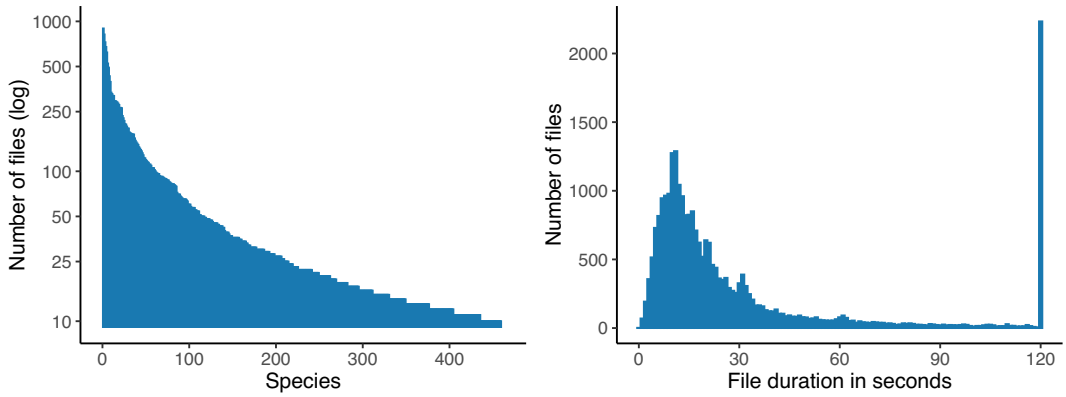

Fig. 2 (**a**) Left plot: The number of files for all 459 species in the dataset, sorted in descending rank order and scaled logarithmically. This illustrates the strongly imbalanced nature of the data, a common aspect of bioacoustic datasets. (**b**) Right plot: The distribution of file durations in the dataset. Most files are around 10 seconds long. The large peak at 120 seconds is the result of trimming longer files to a maximum of two minutes.

list "Singing Species of North America"[25], 11% of sound producing Orthoptera species in Northern America are covered in this dataset. Tropical regions are heavily underrepresented, especially considering their much higher biodiversity among insects. This could also be explained by the large number of species in this area that are not known to produce sounds, or whose calls have not been described[6].

Technical Validation

To provide a baseline of the recognition performance for insect sounds based on this dataset, following the current state of the art in acoustic machine learning, we trained two deep learning classifiers. Although our dataset is designed to encourage innovative approaches including the use of diverse sampling rates, for the present work we restricted our attention to the performance of algorithms which use the standard spectrogram approach of audio classification[26] i.e. converting the waveform data to a fixed non-ultrasonic sampling rate, transforming them into Mel spectrograms. The resulting classification problem can be addressed using standard pipelines very similar to those used for image classification. To handle the varying duration of audio files, spectrograms were divided into 5-second chunks (with 50% overlap) and treated as separate samples during training; classification decisions for a file are then averages over the chunks.

We tested the performance of algorithms from the two current leading families of deep learning algorithm: convolutional neural networks (CNNs) and transformers[26]. For the transformer, we used an algorithm named PaSST that Ghani *et al.*[27] reported as strongly-performing for birdsong. For the CNN, we used EfficientNetv2[28].

The EfficientNet algorithms are popular and robust CNNs. Since our particular use of EfficientNet has not been documented in the academic literature, we give here a brief summary of the implementation, which we will refer to as 'InsectEffNet'.

We selected EfficientNetv2 based on the outcome of a previous data contest (https://www.capgemini.com/in-en/news/inside-stories/catching-the-ai-bug/) which used the InsectSet66 dataset. We used the EfficientNetV2-S model, which has around 20 million parameters. We use version of the model publicly available in the PyTorch 'timm' library which had been pre-trained using ImageNet21k. This image-based pre-training is a common convenience which accelerates the convergence of training for the network. As input to the network we used mel spectrograms with 128 frequency bands (from 0.4 to 22 kHz), calculated from audio standardized to 44.1 kHz.

InsectEffNet and its training procedure were implemented using the PyTorch Lightning framework. The source code for InsectEffNet is available online under the open-source MIT license at (https://github.com/danstowell/insect_classifier_GDSC23_insecteffnet).

As a comparison, we also trained and evaluated the Transformer-based PaSST method. The method was used identically as described in Ghani *et al.*[27] which uses an upper sampling rate of 32 kHz for audio. We adapted the audio chunk size to the same (5 seconds) as InsectEffNet, and used the same number of 128 mel bands.

To counter class imbalance, for both classifiers we applied custom class weights during training. For each class (species), the weighting was chosen as

$$w_i = 1 - \frac{n_i}{\sum_{j=1}^{K} n_j}$$

where n_i is the number of files in a given class and K the number of classes. This down-weights the more common classes in the dataset, such that the training of machine learning should lead to more even per-class performance.

Data augmentation, which synthetically adds variety to the training data, is common in deep learning to improve generalization. The PaSST method uses additive noise and MixUp to the waveform data. For

Model	F1 score (%)	Accuracy (%)
InsectEffNet	56.8	72.2
PaSST	57.5	68.1

Table 3. Overall classification performance of classifiers trained on IS459, evaluated on the test set. F1 score is macro-averaged across classes; accuracy is averaged per-item.

InsectEffNet we used a similar procedure as in Faiß & Stowell, 2023[12], applying additive noise and/or impulse responses (environmental reverberation and sound absorption) to the waveform data.

We did not perform hyperparameter optimization (tuning) for either of these algorithms, because the focus of our current work is to provide baseline standard results; instead, we make use of the tuning that each algorithm had received prior to our main test.

To evaluate performance for our classifiers, we focus primarily on the 'F1 score' evaluated on the InsectSet459 test set[16]. The F1 score is the harmonic mean of the 'precision' and 'recall' measures. It is recommended as a more reliable measure than 'accuracy' for a dataset which is highly unbalanced in the number of examples per category, which is the usual case for wildlife observation data. We measured the F1 score in aggregate, and also for each of the 459 species separately, which enables us to inspect how performance varies from the commonly-represented to the less-common species.

Performance of classifiers. *InsectSet66.* First, the hyperparameters of InsectEffNet were trained and tuned using the previous, smaller dataset InsectSet66. That dataset also offers a starting point to inspect the classification performance of InsectEffNet, before considering our main results from the larger dataset. The performance was very strong, with an overall F1 score of 95.6% and accuracy of 97.7%. Although performance was best for the most commonly-represented species, it remained very high for most of the 66 species in this dataset (Supplementary Figure 1). InsectEffNet achieved stronger performance than the previous work trained on IS66; that work was not aimed at maximizing performance, but investigating feature representations using a simple CNN architecture[12]. These initial results indicate that EfficientNet is a very well-performing CNN architecture, on this task as it has been for many other tasks reported in the literature.

InsectSet459. We then performed a classification study using the new enlarged dataset IS459[16]. We retrained InsectEffNet on our full-size dataset for 459 species. We also trained and evaluated the transformer-based algorithm (named 'PaSST') used in Ghani *et al.*[27]. The deep learning classifiers attained F1 scores between 56% and 58%, with PaSST achieving the strongest results (Table 3). This level of performance is lower than achievable on IS66—as is to be expected for this substantially expanded task over 459 different species, including many species in the "long tail" of the data distribution with very few recordings per species (Fig. 2).

The choice of temporal pooling did not modify InsectEffNet outcomes, possibly indicating its predictions were relatively consistent over the duration of an audio file. For PaSST, max-pooling gave a slightly better F1 score (Table 3).

The long tail effect can be seen directly when the IS459 classification results are plotted per species, in decreasing order of species frequency (Fig. 3). The downward slope of the F1 curves indicates that the most common categories in IS459 can be classified by either algorithm with a performance above 80%, but this rapidly decreases as the less-frequent categories are included. For the infrequent categories, the per-species score has a very high variability (Supplementary Figure 2), which is to be expected when there are few examples available. Nevertheless, these infrequent categories are included so that the classifier can be exposed to a diverse variety of insect sounds. More advanced data augmentation techniques can be used to reduce the biases found in infrequent classes, but it is likely that in many cases more examples are needed to sufficiently capture the variation in insect sounds and background noise or recording device characteristics that may result in low classification performance. In some species, low performance could be caused by their frequency ranges lying outside the range of the spectrogram feature representations used in this work (InsectEffNet: up to 22 kHz, PaSST: up to 16 kHz), leaving little to potentially no species-specific information for the models to detect.

Discussion of the results. In our baseline classification experiment, we found that two very different deep learning classifiers attained a similar level of performance: around 57% F1-score overall (Table 3, rising to more than 80% for the most common classes, Fig. 3). As an indicative comparison: for a similar task with focal birdsong recordings, the BirdNet classifier was reported to show an overall accuracy score of 77.7% over 984 species[29], whereas in our case accuracy reaches 72.2%. Performance may thus already be good enough for some deployed use, especially for common species, although this should be undertaken with careful piloting.

There is a general decline in performance for the species for which we have a low number of examples. This decline was expected, given the dataset characteristics, since it is a common phenomenon in machine learning. For some data-poor categories, the F1 score per class was observed very low indeed, below 20% for both classifiers (Supplementary Fig. 2).

For the information given to the baseline classifiers, we limited attention to spectrograms in the audible frequency range (up to 22 kHz). This likely affected performance, and could be an explanation for the poor performance on certain species. Future work on developing and applying multi-sample-rate models to this data could thus improve performance especially for ultrasonic species.

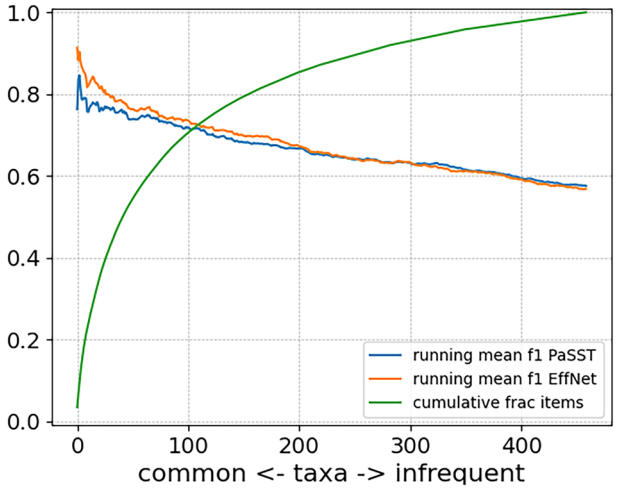

Fig. 3 Per-species classification performance of classifiers trained on IS459, evaluated on the test set. The species are ordered on the x-axis from most common to least common in the dataset, and for each one we plot a running mean of the F1 score—meaning the mean of the F1 score for all species that are equally or more common.

Regarding the machine learning algorithms, it is intriguing that the overall performance scores of the two models converge to very similar values, given that the underlying algorithms are very different from one another. It is unlikely that these different models extract the same features, yet the convergence implies they might both be extracting a very large fraction of the information that they can reasonably get from the audible-range mel-spectra we use. This convergence is not complete—as can be seen in the variation in per-class scores—leaving room for improvements.

Usage Notes

Our method was not designed to ensure complete species coverage for any given region; thus, using this dataset alone to train an automatic acoustic monitoring pipeline for a specific geographic region is not our recommended strategy, since all species from a given region are not necessarily represented. Instead, InsectSet459[16] could be useful for developing and testing machine learning methods specifically optimized for insect sounds and other datasets with highly variable sample-rates. Its total size and number of species could however simulate a comprehensive dataset that covers all sound-producing species of a large geographic region (especially in non-tropical areas). Therefore, methodologies that are developed and work well on this dataset are likely to function well when used to acoustically monitor soniferous insect species in a real environment. A further step towards that goal would be to use InsectSet459[16] to pre-train models and then fine-tune with a smaller, perhaps strongly labeled dataset that covers a geographic region, taxonomic group or a specific task on which a model should be deployed.

The annotation file for InsectSet459[16] contains weak labels, meaning that each audio sample is only associated with one species-label, without on- and offset times or durations annotated. Therefore, this dataset by itself is not suitable for training sound-event detection algorithms that predict precise on- and offsets in longer recordings. But pre-training with InsectSet459 and fine-tuning with a small, strongly-labeled dataset could be beneficial and reduce the amount of precise labeling work that would need to be done otherwise. Some recordings are annotated with additional labels that list species that are audible in the background. This could be useful for certain applications, but background labels are only available for a subset of the data and many recordings contain species in the background which are not labeled.

We also link each file in the dataset to its source observation record (on iNaturalist or xeno-canto) to allow users to retrieve additional information about the recording conditions if needed. For example, location data can be used to limit species predictions of a classifier to sensible geographic ranges. Ambient temperature can strongly influence insect songs, affecting frequency ranges and pulse rates[30,31], which could be useful additional information for classification. The annotation file lists the ambient temperatures at the time of recording for 19% of all files. Where this data is missing, approximate values can be gathered from historical weather data records using the time, date and location of the observations, since this information is much more commonly available.

The observation metadata could also be useful when combining IS459 with other datasets for training and testing models. Since many datasets source their data from the same public repositories as we do (e.g. the iNatSounds dataset[15]), there may be overlap in the recordings which could lead to data leakage between training and testing datasets. To avoid this, users can use the original observation records to filter out duplicate files when combining datasets.

Our curation of IS459[16] benefits from the growth in insect-focused contributions to community bioacoustics collections (xeno-canto and iNaturalist), which were also the source for our prior InsectSet66. IS459[16] is much larger in the number of insect species represented, thus having much better potential for complete species coverage in selected locations. IS459 is more than 9 times larger than IS66 in total audio duration (IS66: 24 hours;

IS459: 227 hours). This makes the dataset more useful for data-hungry applications such as deep learning. Regarding total data volume, note that we limited individual audio file durations to a maximum of 2 minutes. Without this limit, the dataset could be much larger still, but with a kind of false or unhelpful replication, with limited benefits. We chose instead to focus on increasing dataset size through diversity of audio samples.

We did not manually edit the audio files to trim them down to song bouts (echemes), as was done for IS66: instead we trimmed them to automatically-selected 2-minute segments. This is done to create a curation process that scales to large data sizes. It risks making some imperfect edits, which could be unhelpful for algorithm training. On the other hand it also corresponds more closely to automated passive acoustic monitoring, in which arbitrary sound segments are usually submitted for analysis.

An important goal in this work was also to preserve ultrasonic information where available (in around 25% of our data), since it can be an important acoustic component for various species and is a potential additional information source for recognition. This makes a distinctive characteristic of IS459 intended to improve automatic insect recognition in particular, by enabling the development of methods that could make use of the ultrasound component of many insect sounds.

Our dataset shares with all collections the limitation of incomplete coverage of insects and their acoustic behavior. Insect sounds may be simple compared against those of songbirds and other animals; however, they vary due to atmospheric conditions (especially temperature[30,31]), and are often heard in 'choruses', neither of which are addressed by the remit of this dataset. The work we have presented may help with the collection of more diverse insect sound data, as a step on the way to developing automatic systems that can distinguish the similar and dissimilar, the known and unknown, in insect sound.

Data availability

The dataset can be downloaded from Zenodo under the CC 4.0 license (https://doi.org/10.5281/zenodo.14056457).

Code availability

The code that was used to download the source data from iNaturalist and xeno-canto, as well as curating and processing the data as described in the methods section is published on Github (https://github.com/mariusfaiss/InsectSet459). All code for benchmarking the dataset using the InsectEffNet (https://github.com/danstowell/insect_classifier_GDSC23_insecteffnet) and the PaSST models (https://github.com/kkoutini/PaSST) is also available on Github.

Received: 16 September 2025; Accepted: 20 March 2026;
Published online: 27 March 2026

References

1. Wagner, D. L., Grames, E. M., Forister, M. L., Berenbaum, M. R. & Stopak, D. Insect decline in the anthropocene: Death by a thousand cuts. *Proceedings of the National Academy of Sciences*, **118**, https://doi.org/10.1073/pnas.2023989118 (2021).
2. Montgomery, G. A. *et al*. Is the insect apocalypse upon us? how to find out. *Biological Conservation* **241**, 108327, https://doi.org/10.1016/j.biocon.2019.108327 (2020).
3. van Klink, R. *et al*. Emerging technologies revolutionise insect ecology and monitoring. *Trends in Ecology & Evolution* https://doi.org/10.1016/j.tree.2022.06 (2022).
4. Van Klink, R. *et al*. Towards a toolkit for global insect biodiversity monitoring. *Philosophical Transactions of the Royal Society B* **379**, 20230101, https://doi.org/10.1098/rstb.2023.0101 (2024).
5. Kohlberg, A. B., Myers, C. R. & Figueroa, L. L. From buzzes to bytes: A systematic review of automated bioacoustics models used to detect, classify and monitor insects. *Journal of Applied Ecology* https://doi.org/10.1111/1365-2664.14630 (2024).
6. Riede, K. & Balakrishnan, R. Acoustic monitoring for tropical insect conservation. bioRxiv https://www.biorxiv.org/content/early/2024/07/05/2024.07.03.601657 (2024).
7. Riede, K. Acoustic profiling of Orthoptera: Present state and future needs. *Journal of Orthoptera Research* **27**, 203–215 (2018).
8. Bennett, D., Nissen, H., Maschke, M. A., Reck, H. & Diekötter, T. Recent technological developments allow for passive acoustic monitoring of Orthoptera (grasshoppers and crickets) in research and conservation across a broad range of temporal and spatial scales. *Basic and Applied Ecology* **84**, 147–157, https://doi.org/10.1016/j.baae.2025.03.004 (2025).
9. Pérez-Granados, C. BirdNet: applications, performance, pitfalls and future opportunities. *Ibis* (2023).
10. Sethi, S. S. *et al*. Large-scale avian vocalization detection delivers reliable global biodiversity insights. *Proceedings of the National Academy of Sciences* **121**, https://doi.org/10.1073/pnas.2315933121 (2024).
11. Faiß, M. InsectSet47 & 66: Expanded datasets for automatic acoustic identification of insects (Orthoptera and Cicadidae) https://zenodo.org/record/7828438 (2023).
12. Faiß, M. & Stowell, D. Adaptive Representations of Sound for Automatic Insect Recognition. *PLOS Computational Biology* arXiv:2304.12739, https://doi.org/10.1371/journal.pcbi.1011541 (2023).
13. Vellinga, W.-P. & Planque, R. The Xeno-canto collection and its relation to sound recognition and classification https://ceur-ws.org/Vol-1391/166-CR.pdf (2015).
14. iNaturalist. available from https://www.inaturalist.org accessed 09 august 2024.
15. Chasmai, M., Shepard, A., Maji, S. & Horn, G. V. The iNaturalist sounds dataset. In *Proceedings of the NeurIPS 2024 conference (Datasets and Benchmarks Track)* https://doi.org/10.52202/079017-4213 (2024).
16. Faiß, M. & Stowell, D. Insectset459: A large dataset for automatic acoustic identification of insects (orthoptera and cicadidae). *Zenodo* https://doi.org/10.5281/zenodo.14056457 (2025).
17. Xeno-canto Foundation for Nature Sounds. Xeno-canto api query: Recordings of species in the "grasshoppers" group https://xeno-canto.org/api/2/recordings?query=grp:grasshoppers (2024).
18. GBIF.Org User. Occurrence download https://www.gbif.org/occurrence/download/0058185-240626123714530 (2024).
19. GBIF.Org User. Occurrence download https://www.gbif.org/occurrence/download/0058173-240626123714530 (2024).
20. Baker, E., Price, B. W., Rycroft, S. D., Hill, J. & Smith, V. S. BioAcoustica: a free and open repository and analysis platform for bioacoustics. *Database* **2015**, bav054–bav054, https://doi.org/10.1093/database/bav054 (2015).
21. Faiß, M. InsectSet32: Dataset for automatic acoustic identification of insects (Orthoptera and Cicadidae) https://zenodo.org/record/7072196 (2022).

22. Döring, M., Jeppesen, T. & Bánki, O. Introducing checklistbank: An index and repository for taxonomic data. *Biodiversity Information Science and Standards* https://doi.org/10.3897/biss.6.93938 (2022).
23. Borowiec, M. L. *et al*. Deep learning as a tool for ecology and evolution. *Methods in Ecology and Evolution* **13**, 1640–1660, https://doi.org/10.1111/2041-210X.13901 (2022).
24. Odé, B. *et al*. A list of sound producing European grasshoppers and the availability of their sounds in Xeno-canto 95–130, https://doi.org/10.62323/a100191 (2025).
25. SINA. Singing Insects of North America. https://orthsoc.org/sina/index.htm (2023).
26. Stowell, D. Computational bioacoustics with deep learning: a review and roadmap. *PeerJ* **10**, https://doi.org/10.48550/arXiv.2112.06725 (2022).
27. Ghani, B. *et al*. Generalization in birdsong classification: impact of transfer learning methods and dataset characteristics https://arxiv.org/abs/2409.15383 (2024).
28. Tan, M. & Le, Q. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, 10096–10106, https://arxiv.org/abs/2104.00298 (PMLR, 2021).
29. Kahl, S., Wood, C. M., Eibl, M. & Klinck, H. BirdNET: A deep learning solution for avian diversity monitoring. *Ecological Informatics*, **61**, 101236, https://doi.org/10.1016/j.ecoinf.2021.101236 (2021).
30. Dolbear, A. E. The Cricket as a Thermometer **31**, 970–971 https://www.journals.uchicago.edu/doi/10.1086/276739 (1897).
31. Greenfield, M. D. *Acoustic Communication in Orthoptera* (CAB INTERNATIONAL, 1997).

Acknowledgements

We thank the many contributors to the xeno-canto and iNaturalist collections, for sharing their recordings, without whom this work would not have been possible. We highlight the following users who each contributed over 500 recordings: Baudewijn Odé (also for project advice), Christie (iNaturalist username), Joel Poyitt, K.-G. Heller, S. Ingrisch, Cedric Mroczko. The InsectEffNet classifier was developed as part of a CapGemini "Global Data Science Challenge", supported by Amazon Web Services. The classifier code was contributed by a CapGemini team and authored by Raffaela Heily, Lukas Kemetinger, Dominik Lemm and Lucas Unterberger. It is used here with permission. MF was supported by the EU MSCA Doctoral Network Bioacoustic AI (BioacAI, 101071532). BG was supported by EU-funded Horizon projects MAMBO (101060639), GUARDEN (101060693) and TETTRIs (101081903).

Author contributions

M.F. and D.S. conceived the data selection process, wrote code for the data processing and figure creation, wrote and edited the draft; M.F. wrote code for downloading, processing and annotating the data; D.S. wrote the code for benchmarking the InsectEffNet model; B.G. developed the PaSST-based model; all authors reviewed the final draft.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-026-07123-4.

Correspondence and requests for materials should be addressed to M.F.