A Licenses

A.1 Code license

A.2 Database license

CC-BY-4.0, https://creativecommons.org/licenses/by/4.0/

B Code use

B.1 Code access and structure

- The audio recordings and metadata csv are hosted on Zenodo http://doi.org/10.5281/zenodo.4904800. under a CC-BY-4.0 license.
- Code (and the metadata csv for completeness) is hosted on https://github.com/HumBug-Mosquito/HumBugDB under the MIT license.

The GitHub data directory structure is as follows:

```
HumBugDB
 ├─README.md
 ├─*requirements.txt
 ├─notebooks
 │  ├─main.ipynb
 │  └─supplement.ipynb
 ├─data
 │  ├─metadata
 │  │  └─*.csv
 │  └─audio
 │     └─*.wav
 ├─lib
 │  ├─config.py
 │  ├─config_keras.py
 │  ├─config_pytorch.py
 │  ├─feat_util.py
 │  ├─runKeras.py
 │  ├─runTorch.py
 │  ├─ResNetDropoutSource.py
 │  ├─evaluate.py
 │  └─write_audio.py
 └─outputs
    ├─models
    │  ├─keras
    │  └─pytorch
    ├─features
    └─plots
```

A README and several requirements are included for installing Keras, PyTorch, and dependencies for the code. The metadata is located in /data/metadata/ as a csv file.

Extract the audio from Zenodo to the folder /data/audio/ and launch the Jupyter notebook main.ipynb to perform train-test splitting, feature extraction, model training, and evaluation. The notebook imports from lib the necessary files depending on the choice of kernel and PyTorch or Keras.

B.2 Code manual

Top-level notebook main.ipynb performs data partitioning, feature extraction and segmentation in get_train_test_from_df(), model training in train_model(), and model evaluation in get_results(). The code is configured with config.py, where data directories are specified for the data, metadata and outputs, and feature transformation parameters are supplied. Model hyperparameters are given in config_keras.py or config_pytorch.py. The notebook supports both Keras [Chollet et al., 2015] and PyTorch [Paszke et al., 2019] with a common interface for convenience. In more detail, each top-level function is described as follows:

- get_train_test_from_df(df_train, df_test_A, df_test_B) extracts, reshapes, strides, and normalises librosa features for use as tensors, and saves them to config.dir_out, if features with that particular configuration do not exist already. The

Table 3: Feature transformation parameters, in samples. Audio processed with `librosa` at 8,000 Hz. The size of 1 frame in w is equal to `hop_length`. For our parameterisation this is 64 ms, resulting in an input feature slice of $w = 1.92$ s duration and $h = 128$ height.

`NFFT`	`win_size`	`hop_length`	h (`n_mels`)	w	Stride
2048	2048	512	128	30	512

data is split into train and test based on the matches of experiment ID to the audio tracks from the metadata given in `df_train, df_test_A, df_test_B`. It is important that no test recordings from these experiments are seen during training in advance, as otherwise model performance is overestimated. Appendix B.3, Table 4 shows the result of feature extraction with baseline feature parameters.

- `train_model(X_train, y_train, X_val=None, Y_val=None)` trains the BNNs on the data supplied (with validation data optional). The assumed input shape is that of the features produced by `get_train_test_from_df()`. The model architecture and training strategies may be changed in `runKeras.py` or `runTorch.py`.
- `get_results(model, X, y, n_samples=1)` evaluates the model object on test data `{X, y}` with the number of MC dropout samples as `n_samples`. If using deterministic networks, leaving the input argument blank will default to a single evaluation.

Supplementary notebook `supplement.ipynb` is used to reproduce the plots of species distribution in this paper (Figure 6) and contains utilities that were used for debugging and visualising the data, should they be helpful for researchers using their own functions.

B.3 Feature parameters

We first need to define the number of feature windows that are used to represent a sample, $\mathbf{X}_i \in \mathbb{R}^{h \times w}$, where h is the height of the two-dimensional matrix, and w is the width. The longer the window, w, the better potential the network has of learning appropriate dynamics, but the smaller the resulting dataset in number of samples. It may also be more difficult to learn the salient parts of the sample that are responsible for the signal, resulting in a weak labelling problem [Kiskin et al., 2019]. Early mosquito detection efforts have used small windows due to a restriction in dataset size. For example, Fanioudakis et al. [2018] supplies a rich database of audio, however the samples are limited to just under a second. However, despite the mosquito's simple harmonic structure, its characteristic sound also derives from the temporal variations, as is visible from spectrograms. We suspect this flight behaviour tone is better captured over longer windows, since we achieved more robust results with $w = 30, h = 128$, corresponding to 30 frames per window, each of 64 ms duration for a total audio slice of 1.92 seconds per sample. Nevertheless, we encourage researchers to make use of any data they wish to augment their model, for example by padding with noise to match the window size of this architecture, or by choosing a smaller window to extract features from.

To create an augmented dataset, we stride the input signal feature window with a step of 5 feature windows (a duration of 320 ms) Note that the training data is segmented by using overlapping strides specified with `config.step_size=5`, whereas the test data is created with no overlap. Samples that do not divide evenly into the window size are discarded (this is a very small number when using such a small step, and we prefer this option over padding with zeros or noise, though alternate solutions are welcome).

B.4 Baseline models

Keras We give the full model structure in Figure 3. Lambda layers are dropout layers which are placed to perform MC dropout at test-time. This structure bares similarity to VGGish[4], which uses 0.96 second log-mel spectrogram patches as inputs, and 11 weight layers (primarily convolutional layers and max-pool layers).

[4] https://github.com/tensorflow/models/tree/master/research/audioset/vggish

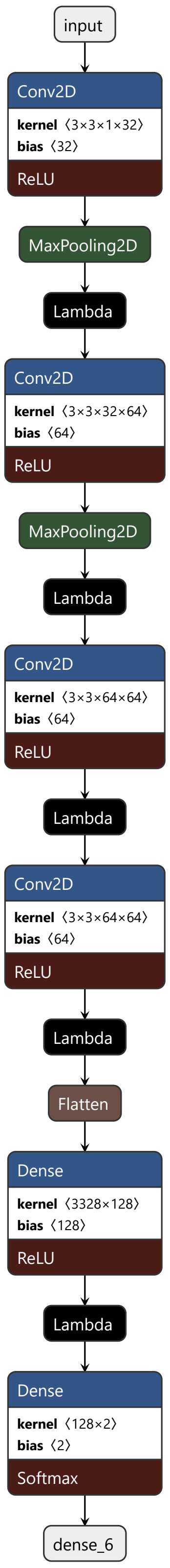


Figure 3: BCNN Keras model. Log-mel spectrograms are input with $w = 40, h = 128$, and passed through the above model. Lambda layers are dropout layers with probability 0.2. Made with `https://github.com/lutzroeder/netron`.

Table 4: Statistics of samples passed to models. Shown in number of data samples **X** resulting from log-mel feature transformation with a window size of 1.92 s, a step size of 0.32 s over the data of Table 5. All the variables affecting the size of the dataset created are found in `config.py`. and documented in `main.ipynb`.

Data	Mosquito	Negative	Total
Train	164,677	138,784	303,461
Test A: Tanzania (bednet)	1,714	2,068	3,782
Test B: Oxford Zoology (caged)	430	1,015	1,445

PyTorch ResNet-X We modify the final layers for compatibility with our data (in `runTorch.py`). Furthermore, we have augmented the construction blocks `BasicBlock()` and `Bottleneck()`, as well as the overall model construction, to feature dropout layers to act as an approximation for the model posterior at test-time. Dropout is implemented implicitly in `ResNetSource.py`, to not interfere with the behaviour of `model.eval()`, which by default disables dropout layers at test-time, removing the necessary stochastic component. Select which version of ResNet to use by modifying the class `ResnetDropoutFull` in `runTorch.py`. For ResNet-18, 34 the final `self.fc1` layer is of size `[512,1]`, whereas ResNet-50 the size is `[2048,1]`. A quick way to check this is to print `x.shape()` before the creation of the `fc1` layer.

Model training To select the loss that is used to define the best performing model, edit `runTorch.py` to make use of `train_acc` (or any other metric as desired) by replacing line 126. Similarly, amend the training epoch loop starting at line 85 to change other metrics or properties during training. In `runKeras.py`, supply arguments and any other desired callbacks and model checkpointing strategies to `model.fit()` in line 105.

Memory optimisation Note that the default settings require at least 16 GB RAM to load into memory for ResNet-50 processing, as channels are replicated 3 times to match the pre-trained weights model. To reduce the strain on memory, increase the `step_size` parameter in `config.py` to reduce the number of windows created by feature extraction. This reduces the overlap between samples.

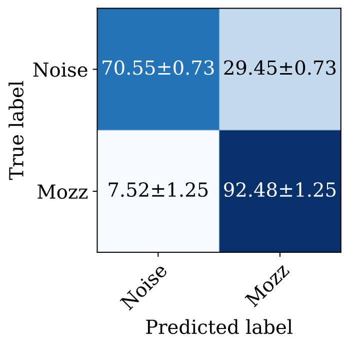
Alternatively, it is possible to use a non-pretrained architecture and change the tensor creation code in `build_dataloader()` from `runTorch.py` to remove `.repeat(1,3,1,1)` as there will be no need to copy over identical data over three channels.

Note that once the tensors have been created, VRAM is not an issue due to the batching over the dataloaders (this code has been run on a GTX970 with 3.5 GB useable VRAM).

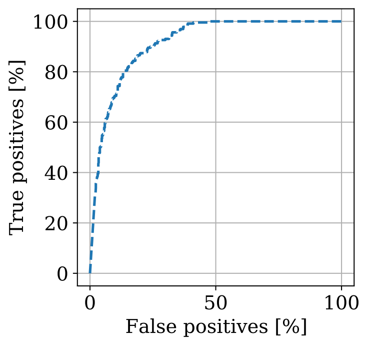
Hyperparameters Configure the hyperparameters in `config_pytorch.py` and `config_keras.py`. The number of epochs was set by observing the learning rate of the network. Within a few epochs, the models began to strongly overfit, with the training accuracy failing to improve validation accuracy. For this reason, both models are set to a low epoch number, and have a fairly low `max_overrun` counter, which determines the maximum number of steps taken for which the target metric fails to improve. The dropout rate and batch size were set to 0.2 and 32, values which are generally risk-free. We note here that the point at which we stop training the model made a fairly significant difference to the balance between true positive and true negative errors (despite a similar overall ROC AUC score). In this respect, the optimisation procedure for the models could be improved with more careful thought about the metrics used for training. If error types are important, consider using loss-calibrated approaches such as that of Cobb et al. [2018].

B.5 Test performance

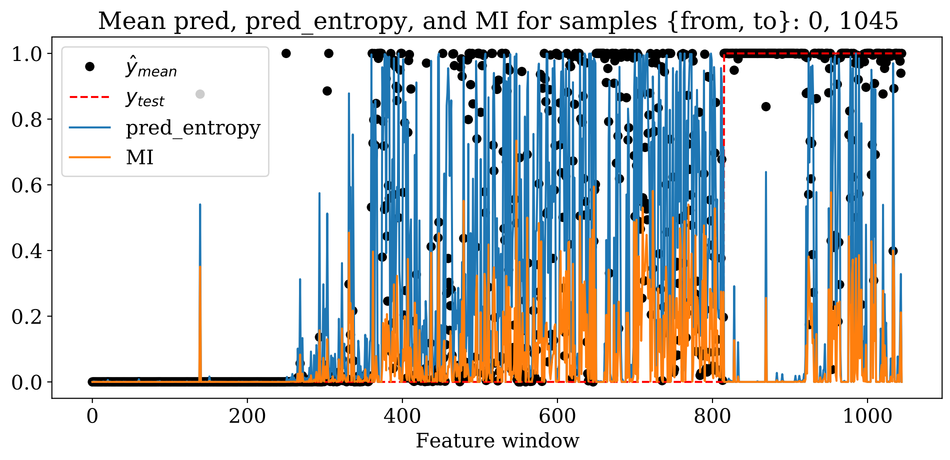
Verifying data integrity of Test B To support the validity of this data, we train the Keras model on half of Test B and test on the other half (with recordings held out), with the settings: `epochs = 7, tau = 1.0, dropout = 0.2, validation_split = None, batch_size = 32, lengthscale = 0.01` to achieve the results of Figure 4.



(a) Confusion matrix



(b) ROC AUC: 0.915 ± 0.003



(c) Mean prediction, predictive entropy and mutual information for feature windows

Figure 4: Performance on half of a hold-out test set constructed from Test B. Confusion matrices given in normalised percentage, and ROC in the form of mean $\pm$ standard deviation, across $N = 10$ MC dropout samples.

C PostgreSQL Database

C.1 Database metadata

The data presented in this paper are regularly maintained in a PostgreSQL database. For completeness, we include the full schema in Figure 5. We note that since data upload is a constant work in progress, some fields have not yet been populated sufficiently to be useful upon data extraction. We thus restrict the metadata to the fields that have been verified, and are most likely to be of greatest use. The command we use to extract all the metadata for this paper is as follows:

```
\copy (SELECT label.id, fine_end_time-fine_start_time, name,
sample_rate, record_datetime, sound_type, species, gender, fed, plurality,
age, method, mic_type, device_type, country, district, province, place,
location_type
FROM label
LEFT JOIN mosquito ON (label.mosquito_id = mosquito.id)
RIGHT JOIN audio ON (label.audio_id = audio.id)
RIGHT JOIN device ON (audio.dev_id = device.id)
WHERE type = 'Fine'
AND fine_start_time IS NOT NULL AND sound_type in
('mosquito', 'background', 'audio', 'wasp','fly') AND
(path LIKE '%Kenya%'
OR path LIKE '%Thai%'
OR path LIKE '%Tanzania%'
OR path LIKE '%LSTMH%'
OR path LIKE '%CDC%'
OR path LIKE '%Culex%')
ORDER BY path) to '/data/export/neurips_2021_zenodo_0_0_1.csv' csv header;
```

We will now break down each metadata field in the data release by the table it originated from and its column heading.

Label:

- `label.id` selects the column `id` from the table `label`, which is joined to `audio` on `label.audio_id = audio.id`. This allows us to now extract a labelled section of audio as indicated by the start and end times of the label.
- `fine_start_time`, `fine_end_time` are the tags for start and end of the audio label, with reference to the original audio recording. Once audio is extracted, we assign the labelled section the filename set to the `label.id`, and define a column `length` which takes the value of `fine_start_time - fine_end_time` for each new label.

Audio:

- `name`: The original filename of the recording (including file extension).
- `sample_rate`: The sample rate of the recording.
- `record_datetime` The time of recording, as SQL `DATETIME` object (easy to parse with either `pandas` or built-in `datetime.datetime`). For newer data, this timestamp is exact, however data collected prior may only be correct to the month.

Mosquito:

- `species` is the species of the mosquito, either the species complex, or more specifically the species if available (e.g. *An. arabiensis* of the complex *An. gambiae s.s.*). If no species information is available, this field is blank (or `NaN` when imported by `pandas` with default settings). A full breakdown of the available species per experimental group is given in Figure 6.
- `gender`: Gender of mosquitoes (`M` or `F`) or blank if not known.
- `fed`: Whether mosquito has been fed (`t` or `f`) or blank.

- `plurality`: The quantity of mosquitoes recorded at one instant: `single`, `plural` or blank if unknown.
- `age`: The age of mosquito in days.
- `sound_type`: denotes whether the label corresponds to a mosquito event if `mosquito`, but can take the value of `background` for corresponding background, `audio` for sections of dense audio events not containing mosquito or `wasp` and `fly`. When parsing data, a binary distinction between `mosquito` and `NOT mosquito` can be made safely.

Device:

- `method`: The method of capture of mosquitoes, taking values `HBN`, `LT`, `ABN`, `LC`, `HLC` or none if not known (or applicable). Human-baited nets (`HBN`) are a form of mosquito intervention where humans are surrounded by a mosquito net. As part of the HumBug project, adapted bednets were used where an additional canopy to hold smartphones for recording was sewn on (from 2020 onwards) [Sinka et al., 2021].

  Animal-baited nets follow the same concept but involve an animal as the main attractant for mosquitoes.

  CDC light traps (`LT`) use several attractants to lure mosquitoes into the collection chamber. Light is the primary source, but bottled CO_2, gas or dry ice can also be used.

  Larval collections, where the eggs of young mosquitoes are collected, are denoted `LC`.

  Human landing catches, where mosquitoes that landed on humans are caught, are denoted `HLC`.

  For mosquitoes raised from culture and not released into the wild and/or near any nets, this field is blank.
- `mic_type`: The microphone used. Takes values `telinga`, `phone` to denote the microphone type. Use this field to filter audio by the type of sound produced, if you wish to check for bias arising from recording device. Further refine the search with the phone model as specified in `device_type`.
- `device_type`: the device to which the microphone was connected. E.g. the field microphone (Telinga) was connected to a `Tascam` or `Olympus` recorder. If a smartphone was used, the device is the phone model (e.g. `itel A16` or `Alcatel 4015X`).

Location:

- `location_type`: The environment in which the mosquitoes were recorded in, taking values `cup` for mosquitoes recorded in sample cups, `field` for mosquitoes recorded free-flying in the field (applicable to Tanzania 2020 bednet recordings), or `culture` for mosquitoes recorded in culture cages.
- `country`, `district`, `province`, `place`: The country, district, province, and name of the recording site (e.g. `USA`, `Georgia`, `Atlanta`, `CDC insect culture, Atlanta`). Use these values combined with `location_type` to filter data by recording experiment.

C.2 Miscellaneous commands

To generate the metadata of Table 5, we include a list of commands used to generate one row for completeness.

Count total length of labelled audio for a certain path and sound type:

```sql
SELECT SUM(fine_end_time-fine_start_time)
FROM label
LEFT JOIN mosquito ON (label.mosquito_id = mosquito.id)
RIGHT JOIN audio ON (label.audio_id = audio.id)
RIGHT JOIN location ON (audio.loc_id = location.id)
WHERE path LIKE '%Thai%' and sound_type='mosquito';
```

Count number of audio files for a certain path and sound type:

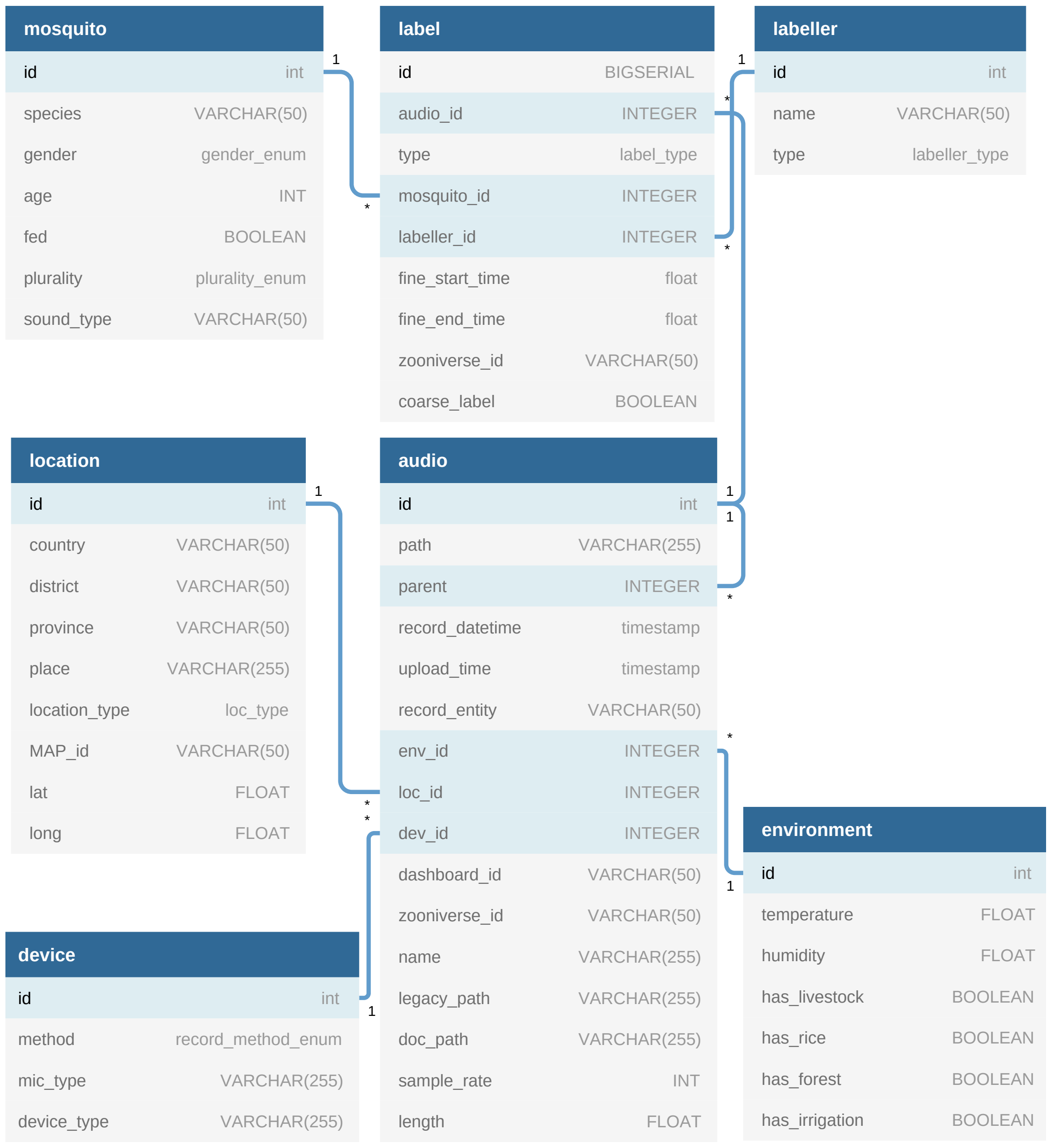


Figure 5: Relational tables of the full PostgreSQL database which was used to generate the data for this paper. The structured nature of the database enforces a standard in label format, ensuring we can efficiently mix and match data from a wide range of experiments with differing protocols. For example, if we wish to investigate the effect of mosquito gender or microphone type on the ability to detect mosquitoes, we may sub-select data with the appropriate metadata with one query. Database schema generated with dbdiagram.io from with `pg_dump -s`.

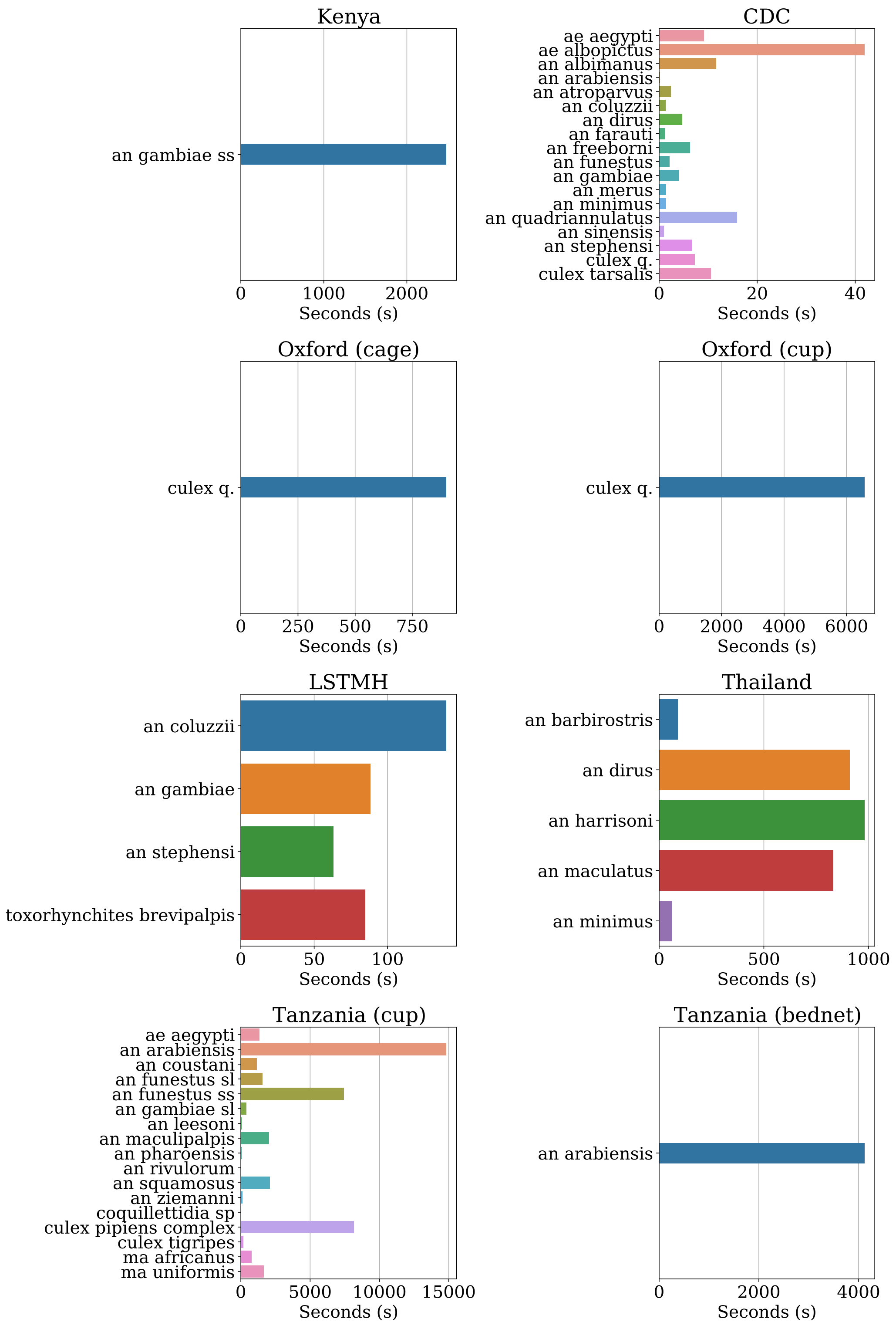


Figure 6: Species distribution per experiment corresponding to Table 5.

```
SELECT COUNT (DISTINCT path)
FROM label
LEFT JOIN mosquito ON (label.mosquito_id = mosquito.id)
RIGHT JOIN audio ON (label.audio_id = audio.id)
RIGHT JOIN location ON (audio.loc_id = location.id)
WHERE path LIKE '%Thai%' and sound_type='mosquito';
```

Return location, device types, and recording methods for dataset:

```
SELECT DISTINCT country, location_type, method, mic_type, device_type
FROM audio
RIGHT JOIN location ON (audio.loc_id = location.id)
RIGHT JOIN device ON (audio.dev_id = device.id)
WHERE path LIKE '%Tanzania%';
```

D Datasheet for dataset

We follow the structure outlined in Datasheets for Datasets by Gebru et al. [2018]. A data nutrition project label [Holland et al., 2018] is also work in progress, but has a longer timeline of delivery, and will be ready for the camera ready version.

D.1 Motivation

For what purpose was the dataset created? This dataset was created for academic research, and applications of machine learning for global health. One such application is the monitoring of deadly mosquito species from their acoustic signature, for which quality training data is required to capture the variation that species may exhibit.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)? This dataset was curated by the Machine Learning Research Group of the University of Oxford. Data was collected by the Department of Zoology, University of Oxford, the Centers for Disease Control and Prevention, Atlanta, the United States Army Medical Research Unit in Kenya (USAMRU-K), at the London School of Tropical Medicine and Hygiene, the Dept of Entomology, Kasesart University, Bangkok, and by the Ifakara Health Institute in Tanzania.

Who funded the creation of the dataset? A Google Impact Challenge Award 2014, The Bill and Melinda Gates Foundation (2019–present), available on `https://www.gatesfoundation.org/about/committed-grants/2019/07/opp1209888` (last accessed: June 2021).

D.2 Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? This dataset is a collection of acoustic recordings in `wav` PCM format. We also supply all the metadata, generated in PostgreSQL to a `csv` file.

How many instances are there in total (of each type, if appropriate)? 9,295 `wav` audio files, 1 `csv`.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? The audio files are a sub-sample of complete audio recordings, with the recordings corresponding to one complete label defined with a label ID, extracted from the original audio with the markers `start_time, end_time`. We are unable to release the full unlabelled audio due to potential issues with privacy and personally identifiable information. The metadata is a curated sub-sample of all available metadata, where fields which were not sufficiently populated or unverified are excluded.

What data does each instance consist of? Each instance corresponds to a labelled section of audio with the event times originally tagged in the original recording with a `start_time, end_time`, either manually by human domain experts, or by machine learning models. The label type is supplied in the metadata.

Is there a label or target associated with each instance? Yes, every recording matches a label.

Is any information missing from individual instances? Though every single sample has a label, some recordings have greater availability of metadata than others; see the metadata `csv` for details.

Are there recommended data splits (e.g., training, development/validation, testing)? Yes, see Table 5. The splits are carried out to increase the chance of generalisation to recordings conducted in varying conditions. The validation split is part of the challenge of this benchmark, left to the discretion of the users. The test data is automatically split in the supplied code.

Table 5: Key audio metadata and train-test partition. *'Wild'* mosquitoes captured and placed into paper *'cups'* or attracted by bait surrounded by *'bednets'*. *'Culture'* mosquitoes bred specifically for research. Total length (in seconds) of mosquito recordings per group given, with the availability of species meta-information in parentheses. Total length of corresponding non-mosquito recordings, with matching environments, given as *'Negative'*. Full metadata is given in Appendix C.

Data (mosquitoes)	Site (country)	Recorded in	Device (sample rate)	Mosquito (s) (with species)	Negative (s)
Train (wild)	Kasetsart (Thailand)	cup (2018)	Telinga (44.1 kHz)	9,306 (2,869)	7,896
Train (wild)	IHI (Tanzania)	cup (2020)	Telinga (44.1 kHz)	45,998 (45,998)	5,600
Train (culture)	Zoology (Oxford, UK)	cup (2017)	Telinga (44.1 kHz)	6,573 (6,573)	1,817
Train (culture)	LSTMH (UK)	cup (2018)	Telinga (44.1 kHz)	376 (376)	147
Train (culture)	CDC (USA)	cage (2016)	phone (8 kHz)	133 (127)	1,121
Train (culture)	USAMRU (Kenya)	cage (2016)	phone (8 kHz)	2,475 (2,475)	31,930
Test A (culture)	IHI (Tanzania)	bednet (2020)	phone 8 kHz	4,118 (4,118)	3,979
Test B (culture)	Zoology (Oxford, UK)	cage (2016)	phone (8 kHz)	737 (737)	2,307
All	All	All	All	71,286 (64,843)	53,227

Are there any errors, sources of noise, or redundancies in the dataset? To our knowledge there are no redundancies, duplicate files, corrupt files or unintended bugs. Despite comprehensive manual checks, label errors due to human entry and ambiguity in sound type may remain.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? The data is self-contained, generated from a PostgreSQL database which is hosted on University of Oxford servers. The data itself is hosted on Zenodo, and the code is accessible on GitHub.

Does the dataset contain data that might be considered confidential? No, explicit permission was obtained where speech is present.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? The audio recordings of mosquitoes may cause distress or discomfort to individuals with medical issues that pertain to mosquito sound.

Does the dataset identify any subpopulations (e.g., by age, gender)? The metadata identifies subpopulations of species complexes by species, and further by gender, age and plurality type (for example, if there was more than one mosquito recorded at a label). Further discriminating factors are described in Appendix C.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? Yes, the speakers may announce the recording ID at the start of a recording, however explicit consent was obtained. It may be possible to trace to the person conducting the experiment indirectly.

D.3 Collection Process

How was the data associated with each instance acquired? The data was collected globally at numerous research facilities. We summarise the data collection efforts below:

- **UK, Kenya, USA**: Recordings from laboratory cultures at the London School of Tropical Medicine and Hygiene (LSTMH), the United States Army Medical Research Unit-Kenya (USAMRU-K); Center for Diseases Control and Prevention (CDC), Atlanta as well as with mosquitoes raised from eggs at the Department of Zoology, University of Oxford. Mosquitoes were recorded by placing a recording device into the culture cages where one or multiple mosquitoes were flying, or by placing individual mosquitoes into large sample cups and holding these close to the recording devices.
- **Tanzania i)**: Mosquitoes recorded at Ifakara Health Institute's semi-field facility (*'Mosquito City'*) at Kining'ina. The facility houses six chambers containing purpose-built experimental huts, built using traditional methods and representing local housing constructions, with grass roofs, open eaves and brick walls. Four different configurations of the HumBug Net, each with a volunteer sleeping under the net, were set up in four chambers. Budget smartphones were placed in each of the four corners of the HumBug Net. Each night of the study, 200 laboratory cultured *An. arabiensis* were released into each of the four huts and the MozzWear app began recording.
- **Tanzania ii)** A collection and recording project in the Kilombero Valley, Tanzania. HBNs, larval collections and CDC-LTs were used to sample wild mosquitoes and record them in sample cups in the laboratory. *Anopheles gambiae* and *An. funestus* (another highly dangerous mosquito found across sub-Saharan Africa), are also siblings within their respective species complexes. Thus, standard PCR identification techniques [Scott et al., 1993] were used to fully identify mosquitoes from these groups. The Tanzanian sampling has collected 17 different species including: *An. arabiensis* (a member of the *gambiae* complex), *An. coluzzii*, *An. funestus*, *An. pharoensis* (see Appendix C, Figure 6 for a full breakdown).
- **Thailand**: Mosquitoes were sampled using ABNs, HBNs and larval collections over a period of two months during peak mosquito season (May to October 2018). Sampling was conducted in Pu Teuy Village (Sai Yok District, Kanchanaburi Province, Thailand) at a vector monitoring station owned by the Kasetsart University, Bangkok. The mosquito fauna at this site include a number of dominant vector species, including *An. dirus* and *An. minimus* alongside their siblings (*An. baimaii* and *An. harrisoni*) respectively (Appendix C, Figure 6 gives a species histogram for this dataset). Sampling ran from 6 pm to 6 am, as most anopheline vectors prefer to bite during the night. Mosquitoes were collected at night, carefully placed into large sample cups and recorded the following day using the high-spec Telinga field microphone and a budget smartphone.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? A summary of equipment is as follows:

- Smartphone (Itel, Alcatel, and others) audio recording with the MozzWear application.
- Telinga, Tascam, Olympus field microphone and recording devices.
- Human labelling with Excel.
- Human labelling with Audacity (GNU GPLv2 license).
- Labels produced by a Bayesian convolutional neural network (our own, MIT license, included in paper).
- Voice activity detection and removal with WebRTC (BSD license).
- Python (BSD-style license), MongoDB (Server Side Public License), Django (BSD license), Apache (GPLv3 license), PostgreSQL (BSD/MIT-like license), Unix for databases, HTML dashboards, and post-processing.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)? Researchers from the locations previously mentioned, paid salary from their respective institutions, through the grants disclosed previously.

Over what timeframe was the data collected? 2015 to 2020 (and ongoing).

Were any ethical review processes conducted (e.g., by an institutional review board)? We have obtained the ethics approval from the following committees:

- Oxford Tropical Research Ethics Committee (OxTREC Ref. 548-19) – University of Oxford (UK).
- Ifakara Health Institute (IHI)-IRB – Tanzania
- National Institute for Medical Research – Tanzania
- School of Public Health at the University of Kinshasa (KSPH) – DRC

D.4 Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? The data underwent rigorous curation, from manual adjustment to labels supplied in text files, to commands in the database to deal with incorrectly entered label times resulting in missing data. To encourage reproducibility and compatibility for future data release, all the label and audio quality control is performed before uploading to the database, and within the dataset itself.

Example of quality control code to check that the label end does not exceed the length (which happens frequently when labels are entered by hand into Audacity with end times longer than the recording and then exported to a text file):

```sql
SELECT path, fine_start_time, fine_end_time, sound_type, length
FROM label
LEFT JOIN mosquito ON (label.mosquito_id = mosquito.id)
RIGHT JOIN audio ON (label.audio_id = audio.id)
RIGHT JOIN location ON (audio.loc_id = location.id)
WHERE fine_end_time > length;
```

Sources with low estimated label quality were either removed or manually re-labelled and amended in the database.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? Yes, all data that may have future utility (and has not been yet used for that purpose) has been released. Unprocessed, and currently unlabelled data is also all stored on the database server, but requires further curation and data entry to the specific data tables before release. We plan to periodically update the database as more data becomes available.

Is the software used to preprocess/clean/label the instances available? The software to do so included Audacity, PostgreSQL, Python, Excel, and is available and well-maintained. We will make use of it in future for future data curation.

D.5 Uses

Has the dataset been used for any tasks already? If so, please provide a description. A subset of this data (recorded in Thailand, Kenya, UK, USA) has been used to train a machine learning model to distinguish and detect a mosquito from its acoustic signature. The model was a 4-layer Bayesian convolutional neural network implemented in Keras. The predictive entropy and mutual information were used to screen predictions over thousands of hours of data. Hand labels were added to correct predictions, and the labels were fed back into the database [Anonymous et al., 2021].

Is there a repository that links to any or all papers or systems that use the dataset? Yes, the Zenodo data directory contains all the references to papers and code.

What (other) tasks could the dataset be used for? A list of use cases is not limited to, but may include:

1. Training a machine learning model to distinguish between various species (or other properties such as age, gender, or any field supported in the database).
2. Identifying the fundamental frequencies of flight tone for a particular species, to improve upon the understanding of bioacoustics literature, and entomological research.
3. Examining inter-species (or similar) variability.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? No, the dataset is specifically organised in PostgreSQL in a way to be consistent with future release. However, in future more metadata may become available for legacy datasets, and larger subsets may become available upon addition of labels.

Are there tasks for which the dataset should not be used? No.

D.6 Database maintenance

Who is supporting/hosting/maintaining the dataset? Please contact Dr. Ivan Kiskin at `ivankiskin1@gmail.com`, who is maintaining the dataset. Alternative contacts include Professor Steve Roberts at `sjrob@robots.ox.ac.uk` at the University of Oxford Machine Learning Research Group.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? The data will be updated as new data from new trials is obtained and curated. We expect updates to occur every few months in 2021. Updates will be communicated by commits and pushes to the GitHub repository. Please also follow the link on Zenodo for versioning details, where older versions will continue to be hosted.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. If you would like to contribute to this data, please contact the database host and supervising professor. We would be happy to curate data and provide requirements which would help qualify a dataset for hosting. All contributions will be credited appropriately in future work.

References

H. Ali. Real-time Communication Using WebRTC. Technical report, Georgia Institute of Technology, 2018.

Anonymous et al. Automatic acoustic mosquito tagging with Bayesian neural networks. *Under review*, X(X):X, 2021.

R. J. Bomphrey, T. Nakata, N. Phillips, and S. M. Walker. Smart wing rotation and trailing-edge vortices enable high frequency mosquito flight. *Nature*, 544(7648):92–95, 2017.

F. Chollet et al. Keras, 2015. URL https://keras.io. Accessed: 2018-06-07.

A. D. Cobb, S. J. Roberts, and Y. Gal. Loss-calibrated approximate inference in Bayesian neural networks. *arXiv preprint arXiv:1805.03901*, 2018.

E. Fanioudakis, M. Geismar, and I. Potamitis. Mosquito wingbeat analysis and classification using deep learning. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 2410–2414, 2018.

Y. Gal and Z. Ghahramani. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059, 2016.

T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé III, and K. Crawford. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*, 2018.

R. Harbach. Mosquito taxonomic inventory, 2013. URL http://mosquito-taxonomic-inventory.info/. Last accessed: 2021-06-07.

S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, et al. CNN architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 131–135. IEEE, 2017.

S. Holland, A. Hosny, S. Newman, J. Joseph, and K. Chmielinski. The dataset nutrition label: A framework to drive higher data quality standards. *arXiv preprint arXiv:1805.03677*, 2018.

N. Houlsby, F. Huszár, Z. Ghahramani, and M. Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

HumBug. The HumBug Project, 2021. URL https://humbug.ox.ac.uk/. Accessed: 2021-06-21.

S. Jakhete, S. Allan, and R. Mankin. Wingbeat frequency-sweep and visual stimuli for trapping male Aedes aegypti (Diptera: Culicidae). *Journal of medical entomology*, 54(5):1415–1419, 2017.

B. J. Johnson and S. A. Ritchie. The siren's song: exploitation of female flight tones to passively capture male Aedes aegypti (Diptera: Culicidae). *Journal of medical entomology*, 53(1):245–248, 2016.

R. Karrer. Google WebRTC Voice Activity Detection module, 2020. URL https://github.com/rafaelkarrer/mex-webrtcvad/releases/tag/v0.1. Accessed: 2021-06-05.

I. Kiskin, B. P. Orozco, T. Windebank, D. Zilli, M. Sinka, K. Willis, and S. Roberts. Mosquito detection with neural networks: the buzz of deep learning. *arXiv preprint arXiv:1705.05180*, 2017.

I. Kiskin, D. Zilli, Y. Li, M. Sinka, K. Willis, and S. Roberts. Bioacoustic detection with wavelet-conditioned convolutional neural networks. *Neural Computing and Applications: Special Issue on Deep Learning for Music and Audio*, Aug 2018. ISSN 1433-3058.

I. Kiskin, U. Meepegama, and S. Roberts. Super-resolution of time-series labels for bootstrapped event detection. *Time-series Workshop at the International Conference on Machine Learning*, 2019.

I. Kiskin, L. Wang, A. Cobb, et al. Humbug Zooniverse: a crowd-sourced acoustic mosquito dataset. *International Conference on Acoustics, Speech, and Signal Processing 2020, NeurIPS Machine Learning for the Developing World Workshop 2019*, 2019, 2020.

Y. Li, D. Zilli, H. Chan, I. Kiskin, M. Sinka, S. Roberts, and K. Willis. Mosquito detection with low-cost smartphones: data acquisition for malaria research. *NeurIPS Workshop on Machine Learning for the Developing World*, 2017.

Y. Li, I. Kiskin, M. Sinka, D. Zilli, H. Chan, E. Herreros-Moya, T. Chareonviriyaphap, R. Tisgratog, K. Willis, and S. Roberts. Fast mosquito acoustic detection with field cup recordings: an initial investigation. *Detection and Classification of Acoustic Scenes and Events*, 2018.

T. Marinos, S. Lin, D. Zilli, and H. Chan. MozzWear, 2021. URL https://github.com/HumBug-Mosquito/MozzWear. Pending update on Google Play store, GitHub private, accessed: 2021-06-05.

MongoDB Inc. Mongodb, 2021. URL https://www.mongodb.com/. Accessed: 2021-06-05.

H. Mukundarajan, F. J. H. Hol, E. A. Castillo, C. Newby, and M. Prakash. Using mobile phones as acoustic sensors for high-throughput mosquito surveillance. *eLife*, 6:e27854, Oct 2017. ISSN 2050-084X.

K. Palanisamy, D. Singhania, and A. Yao. Rethinking CNN models for audio classification. *arXiv preprint arXiv:2007.11154*, 2020.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

V. P. Perevozkin and S. S. Bondarchuk. Species specificity of acoustic signals of malarial mosquitoes of anopheles maculipennis complex. *International Journal of Mosquito Research*, 2(3):150–155, 2015.

J. Pons and X. Serra. musicnn: Pre-trained convolutional neural networks for music audio tagging. *arXiv preprint arXiv:1909.06654*, 2019.

J. Pons, O. Nieto, M. Prockup, E. Schmidt, A. Ehmann, and X. Serra. End-to-end learning for music audio tagging at scale. *arXiv preprint arXiv:1711.02520*, 2017.

PostgreSQL Global Development Group. PostgreSQL, 2021. URL https://www.postgresql.org/docs/9.3/app-psql.html. Accessed: 2021-06-05.

J. Ramirez, J. M. Górriz, and J. C. Segura. Voice activity detection. fundamentals and speech recognition system robustness. *Robust speech recognition and understanding*, 6(9):1–22, 2007.

A. Sahoo. Voice activity detection for low-resource settings. *Department of Electrical Engineering, Stanford University*, 2020.

J. A. Scott, W. G. Brogdon, and F. H. Collins. Identification of single specimens of the anopheles gambiae complex by the polymerase chain reaction. *The American journal of tropical medicine and hygiene*, 49(4):520–529, 1993.

K. Shimada, N. Takahashi, S. Takahashi, and Y. Mitsufuji. Sound event localization and detection using activity-coupled cartesian doa vector and rd3net. Technical report, DCASE2020 Challenge, July 2020.

P. M. Simões, R. A. Ingham, G. Gibson, and I. J. Russell. A role for acoustic distortion in novel rapid frequency modulation behaviour in free-flying male mosquitoes. *Journal of Experimental Biology*, 219(13):2039–2047, 2016.

M. Sinka, D. Zilli, I. Kiskin, Y. Li, D. Kirkham, W. Rafique, H. Chan, B. Gutteridge, E. Herreros-Moya, H. Portwood, S. J. Roberts, and K. J. Willis. HumBug – an acoustic mosquito monitoring tool for use on budget smartphones. *Methods in Ecology and Evolution*, 2021.

M. E. Sinka, M. J. Bangs, S. Manguin, Y. Rubio-Palis, T. Chareonviriyaphap, M. Coetzee, C. M. Mbogo, J. Hemingway, A. P. Patil, W. H. Temperley, et al. A global map of dominant malaria vectors. *Parasites & vectors*, 5(1):1–11, 2012.

D. Vasconcelos, N. J. Nunes, and J. Gomes. An annotated dataset of bioacoustic sensing and features of mosquitoes. *Scientific Data*, 7(1):1–8, 2020.

World Health Organization. Fact Sheet, 2020. URL `https://www.who.int/news-room/fact-sheets/detail/vector-borne-diseases`. Accessed: 2020-01-26.