MOTIVATION.

[question1] For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.
[answer1] 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.

[question2] Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?
[answer2] 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.

[question3] Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.
[answer3] 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).

[question4] Any other comments?
[answer4] No.

COMPOSITION.

[question5] What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.
[answer5] 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.

[question6] How many instances are there in total (of each type, if appropriate)?
[answer6] 9,295 wav audio files, 1 csv.

[question7] Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).
[answer7] 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.

[question8] What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.

[answer8] 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.

[question9] Is there a label or target associated with each instance? If so, please provide a description.

[answer9] Yes, every recording matches a label.

[question10] Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.

[answer10] Though every single sample has a label, some recordings have greater availability of metadata than others; see the metadata csv for details.

[question11] Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

[answer11]

[question12] Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

[answer12] Yes, see Table 7. 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. The two tasks that the data splits encouraged are defined as follows:

• Mosquito Event Detection (MED): distinguishing mosquitoes of any species from their background surroundings, such as other insects, speech, urban, and rural noise.

• Mosquito Species Classification (MSC): the classification of detected mosquitoes into their respective species.

Table 7: Key audio metadata and division into train/test for the tasks of MED: Mosquito Event Detection, and MSC: Mosquito Species Classification. '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 documented in Appendix C.

[question13] Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

[answer13] 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.

[question14] Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or

relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

[answer14] 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.

[question15] Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

[answer15] No, explicit permission was obtained where speech is present.

[question16] Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

[answer16] The audio recordings of mosquitoes may cause distress or discomfort to individuals with medical issues that pertain to mosquito sound.

If the dataset does not relate to people, you may skip the remaining questions in this section.

[SkipA] NO

[question17] Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

[answer17] 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.

[question18] Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.

[answer18] 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.

[question19] Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

[answer19]
[question20] Any other comments?
[answer20] No.

COLLECTION PROCESS.

[question21] How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.
[answer21] 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 11 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 11 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.

[question22] What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?
[answer22] A summary of equipment
is as follows:
• Smartphone (Itel, Alcatel, and others) audio recording with the MozzWear application.
Smartphone devices may have variable sample rates, as denoted by the sample rate column
of the metadata. The version of MozzWear used in the curation of this dataset recorded
audio in 8,000 Hz mono wave format.
• Telinga EM-23 field microphone, and Tascam, Olympus recording devices recording at
44,100 Hz. The Telinga is a very sensitive, low-noise microphone which was widely adopted
in bioacoustic studies.
• 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.
[question23] If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?
[answer23]
[question24] Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?
[answer24] Researchers from the locations previously mentioned, paid salary from their respective institutions, through the
grants disclosed previously.
[question25] Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.
[answer25] 2015 to 2020 (and ongoing).
[question26] Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.
[answer26] 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

If the dataset does not relate to people, you may skip the remaining questions in this section.
[SkipB] NO
[question27] Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?
[answer27]
[question28] Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.
[answer28]
[question29] Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.
[answer29]
[question30] If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).
[answer30]
[question31] Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.
[answer31]
[question32] Any other comments?
[answer32] No.

PREPROCESSING/CLEANING/LABELING.

[question33] Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.
[answer33] 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):

```
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.

[question34] Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

[answer34] 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.

[question35] Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.

[answer35] 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.

[question36] Any other comments?

[answer36] No.

USES.

[question37] Has the dataset been used for any tasks already? If so, please provide a description.

[answer37] 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 [Kiskin et al., 2021]. Code for the training and resulting predictive pipeline is available on https://github.com/HumBug-Mosquito/MozzBNN.

Other past use cases and publications can be found in related works from the link of the following section. We summarise these here as:

- Bioacoustic classification with wavelet-conditioned neural networks [Kiskin et al., 2017, 2018].
- Cost-sensitive mosquito detection [Li et al., 2017a]
- A case study of species classification with field recordings [Li et al., 2018]
- A release of a subset of this database for crowdsourcing (with baseline mosquito detector model) [Kiskin et al., 2019, 2020]

[question38] Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

[answer38] Yes, the Zenodo data directory https://zenodo.org/record/4904800 contains all the references to projects, papers and code which are associated with this dataset.

[question39] What (other) tasks could the dataset be used for?

[answer39] A list of use cases is not limited to, but may include:

1. Validating species classification models from the literature.
2. Frequency analysis. Identifying the fundamental and harmonic 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. For example, the effect on the sound of flight as a result of age, gender, or any field supported in the database.

We now expand upon each point:

1. Validating species classification models from the literature. As a result of procuring curated data with species meta-information of both wild and lab mosquitoes, this dataset serves as an ideal test-bed to verify the effectiveness of existing species classification approaches. We encourage researchers to validate their models by making use of these data to form their own test sets without re-training their models on any parts of this dataset. Strong species discrimination performance would signify a great opportunity to utilise acoustics as a wide-scale surveillance tool.

It would also be very useful to examine transfer learning approaches, where previous models are re-trained and tested on the suggested splits of the data for either task. If you encounter any issues, or require further information do not hesitate to contact the database maintainers (Appendix D.7).

Frequency analysis. Earlier works in the literature proposed more hand-crafted approaches to building detection or classification models. These may be especially useful in very lowpower embedded devices which require fast and efficient algorithms. These approaches were typically centered around specific harmonic inter-peak ratios (See Kiskin [2020, Sec. 3.2] for an overview of relevant prior work). Frequency analysis may be performed on any parts of this dataset, including on species which are under-represented. In particular, the CDC dataset contains a wide range of unique species which are sparsely labelled, however the labelled sections have very high signal-to-noise ratio. As with previous suggested use cases, we recommend trialling approaches on disjoint sets of experiments (or at the very least individual mosquito recording within an experimental set). Once again, there exists an excellent opportunity to validate models from the literature on their ability to distinguish species on this dataset.

3. Examining the effect of species variability on their flight tone. It is well known that mosquitoes exhibit significant variability in their physical (and therefore acoustic) properties within a species. These occur due to a multitude of factors including the age, wingspan, gender. Additionally, confounding factors such as the temperature, humidity, and potentially their fed status, can increase the difficulty in distinguishing individuals within and across species. As we maintain as much metadata as possible, this dataset provides the opportunity to examine such factors. In future releases, temperature and humidity will be added where possible, and this data is expected to be available in an update on the Tanzanian cup recordings which has already good metadata coverage including species, age, gender, fed, method. If you wish to have early access to additional metadata, please contact us and we will make the availability of such metadata a higher priority.

[question40] Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?

For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal

risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?
[answer40] 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.
[question41] Are there tasks for which the dataset should not be used? If so, please provide a description.
[answer41] No.
[question42] Any other comments?
[answer42] No.

DISTRIBUTION.

[question43] Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.
[answer43]
[question44] How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?
[answer44]
[question45] When will the dataset be distributed?
[answer45]
[question46] Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.
[answer46]
[question47] Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.
[answer47]
[question48] Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.
[answer48]
[question49] Any other comments?
[answer49] No.

MAINTENANCE.

[question50] Who will be supporting/hosting/maintaining the dataset?
[answer50] Please contact Dr. Ivan Kiskin who is maintaining the dataset. Alternative contacts include Professor Steve Roberts at the University of Oxford Machine Learning Research Group.

[question51] How can the owner/curator/manager of the dataset be contacted (for example, email address)?
[answer51] ivankiskin1@gmail.com, and sjrob@robots.ox.ac.uk

[question52] Is there an erratum? If so, please provide a link or other access point.
[answer52]

[question53] Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?
[answer53] The data will be updated as new data and/or metadata from new trials is obtained and curated. We expect the following updates:
1. Recordings of wild captured mosquitoes in DRC:
• Date: Q2/Q3 2021
• Summary: 15 species, at minimum 2000 wild captured individual mosquitoes
• Metadata: species (with PCR identification where appropriate), gender, fed status, temperature, humidity, collection method, recording device information, time of collection
2. Additional metadata for IHI Tanzanian cup recordings:
• Date: Q2/Q3 2021
• Summary: Additional metadata
• Metadata: Temperature, humidity, wing span images (and wing lengths)

[question54] If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.
[answer54]

[question55] Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.
[answer55] To ensure the documentation is up to date, any additional metadata and code will be documented in this supplement and the datasheet for datasets. The supplement is available on GitHub alongside the baseline code at https://github.com/HumBug-Mosquito/HumBugDB/tree/master/docs.
Database revisions will be incremented in the format of X.X.X (current version 0.0.1). The main url, https://zenodo.org/record/4904800, will always resolve to the latest version. If you intend to use a specific version you may select the version from the main page. Both the data and metadata are fully supported with versioning.
Updates will be communicated as follows:
• GitHub commits (mailing list), and releases.
• Posts on the HumBug official twitter account https://twitter.com/oxhumbug.

• Updates on our official website on https://humbug.ox.ac.uk/news/.
• Follow-up publications utilising additional data (arXiv and proceedings where appropriate).

[question56] 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. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.

[answer56] 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.

[question57] Any other comments?

[answer57] No.