scientific data

Check for updates

OPEN

DATA DESCRIPTOR

An annotated image dataset for training mosquito species recognition system on human skin

Song-Quan Ong[1]✉ & Hamdan Ahmad[2]

This paper introduces a new mosquito images dataset that is suitable for training and evaluating a recognition system on mosquitoes in normal or smashed conditions. The images dataset served mainly for the development a machine learning model that can recognize the mosquito in the public community, which commonly found in the smashed/damaged form by human. Especially the images of mosquito in hashed condition, which to the best of our knowledge, a dataset that fulfilled such condition is not available. There are three mosquito species in the dataset, which are *Aedes aegypti*, *Aedes albopictus* and *Culex quinquefasciatus*, and the images were annotated until species level due to the specimen was purely bred in a WHO accredited breeding laboratory. The dataset consists of seven root files, six root files that composed of six classes (each species with either normal landing, or random damaged conditions) with a total of 1500 images, and one pre-processed file which consists of a train, test and prediction set, respectively for model construction.

Background & Summary

Mosquito surveillance programs are probably the most important components to prevent disease outbreaks. To monitor the population of mosquitoes, field sampling and data collection of adult mosquitoes are commonly conducted, and later, the species are classified and counted in a laboratory[1,2]. However, these standard procedures to obtain mosquito number data have key constraints, such as labor, time, and cost consumption. Since humans-as-bait traps are more effective than physical traps and the public commonly encounter mosquitoes on their own[3], the idea of engaging the public community in a mosquito surveillance program provides an excellent alternative to collect mosquito data with spatial-temporal information. To ensure that the idea of community-based mosquito surveillance is feasible, we need to support the community with the knowledge of mosquito pest recognition. The support can be achieved with a mobile application or recognition system that able to classify the mosquito at the household level.

To address these challenges, we need to define operationally of mosquitoes in the community, where two criteria need to be fulfilled: the type of mosquito species and their conditions. For the mosquito species, according to the WHO[4], the most prevalent viral infections are dengue, chikungunya fever, Zika virus fever, yellow fever, West Nile fever, and Japanese encephalitis, and the diseases are primarily transmitted by *Aedes aegypti* (L.), *Aedes albopictus* (L.), and *Culex quinquefasciatus*. The condition of the mosquito that is likely to be observed by the community is either landing alive or smashed or damaged on human skin. Therefore, we present an annotated dataset that able to be used for training a mosquito recognition system that able to distinguish the mosquito species in harsh condition.

In general, the mosquitoes were bred and grew to adult stage, 4–5 days old in a fully control laboratory, Vector Control Research Unit, Universiti Sains Malaysia, which is accredited by WHO for insecticides susceptibility test[5]. The data collection process was illustrated in Fig. 1.

The mosquito obtained from the mosquito breeding was transferred by a Polyethylene terephthalate (PET) container (diameter 12 cm, height 6 cm, Fig. 2a) to the net cage for image acquisition. The container and camera were placed in the cage for 30 minutes to allow the mosquito to adapt to the environment before images acquisition. The images were acquired by a digital single-lens reflex (DSLR) camera (Canon 7D, 18MP APS-C CMOS sensor, ISO 3200, auto white balance) with Tamron SP AF 90 mm f/2.8 Di Macro Lens. The images acquisition

[1]Institute for Tropical Biology and Conservation, Universiti Malaysia Sabah, Jalan UMS, 88400, Kota Kinabalu, Sabah, Malaysia. [2]Vector Control Research Unit, School of Biological Sciences, Universiti Sains Malaysia, 11800, Penang, Malaysia. ✉e-mail: songquan.ong@ums.edu.my

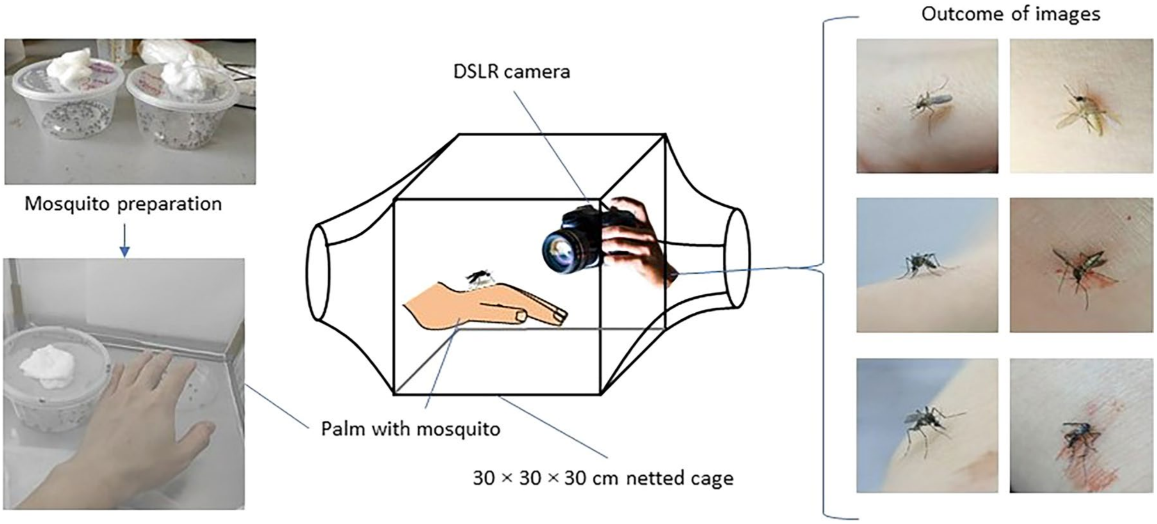

Fig. 1 Outline of mosquito preparation and image collection.

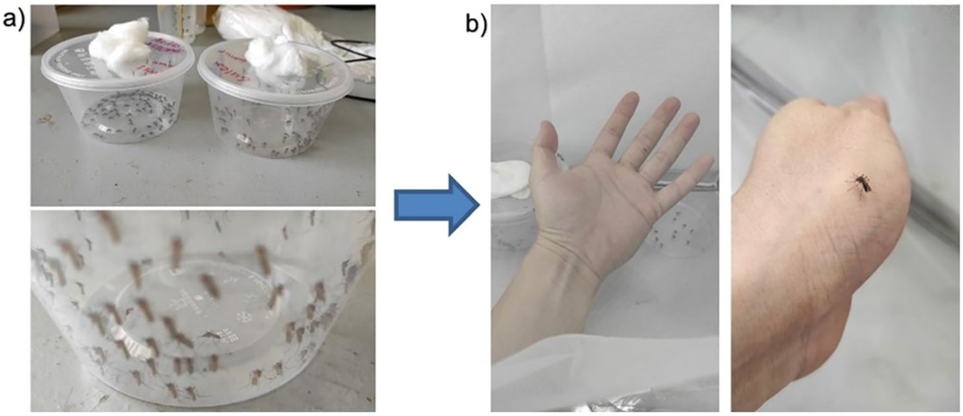

Fig. 2 (**a**) Mosquitos' colonies and culture from VCRU USM. Mosquito was released one by one for image acquisition, (**b**) The process Image acquisition is carried out within a 30 × 30 × 30 cm netted cage with 36 W LED Ring Light white colored illumination (5500 K).

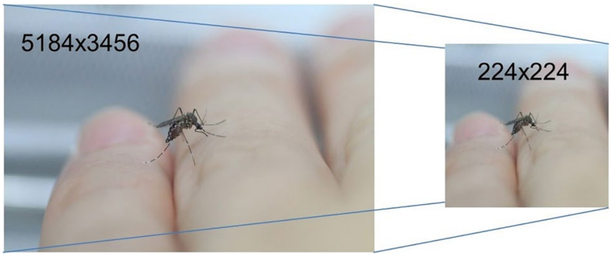

Fig. 3 All the images were resized into 224 × 224 pixels from the original dimension.

was performed on 4- to 5-day-old females' adult in a netted cage with 34 W white light illumination on top of the cage. The volunteer consists of three ethnicities – Malay, Chinese and India, which aim to reflect the diversity of human skin tone. The volunteer's palm is rest in the cage and different angles of the landed mosquitos' images were acquired. Smashed mosquitoes were generated by smashing the mosquito randomly by a human palm in a non-feeding, partial, or fully repletion situation (Fig. 2b). The images were saved in JPEG format in the folders according to their classes. Images were later resized from original dimension into 224 × 224 pixels, to lower the file size of the images (as lower the computational power) that required to initiate the machine learning model training pipeline (Fig. 3), which is a common input image dimension expected by most of the deep convolutional neural network such as AlexNet[6], ResNet and VGG-16[7]

Six root files of raw image data				
Sample of images	**Labels**	**Species**	**Conditions on human skin**	**Number of images**
	Aedes aegypti landing	*Aedes aegypti* L.	Normal landed	250
	Aedes aegypti smashed		Smashed or damaged	250
	Aedes albopictus landing	*Aedes albopictus* L.	Normal landed	250
	Aedes albopictus smashed		Smashed or damaged	250
	Culex quinquefasciatus landing	*Culex quinquefasciatus* Say	Normal landed	250
	Culex quinquefasciatus smashed		Smashed or damaged	250
One pre-processed data file*				
data_splitting	Train			4200
	Test			1800
	Prediction			3600

Table 1. Description, Labels, and Example of images for the dataset: Six root files that represent six classes of mosquitoes, and one pre-processed file. *Pre-processed the image data with augmentation and data splitting.

Methods

Mosquito.

The adult of the susceptible strain WHO/VCRU of *Ae. aegypti, Ae. albopictus* and *Cx quinquefasciatus* were obtained from the Vector Control Research Unit (VCRU), Universiti Sains Malaysia. The mosquitoes were cultured in insectarium for more than 20 years and used for the WHO insecticides susceptibility test; the colonies were maintained at 27 ± 1 °C and 75 ± 5% relative humidity in insectariums. The larvae were reared in dechlorinated water and fed with lab food (Dog biscuit: yeast: milk powder: beef liver powder at a 3:1:1:1 ratio). The pupae were transferred into a 30 × 30 × 30 cm netted cage for adult emergence. The adult mosquitoes were fed with 10% sucrose mixed with a Vitamin B complex as an energy supply. Four to five-day-old female adults were used for the images acquisition.

The dataset consists of three mosquito species - *Aedes aegypti* L., *Aedes albopictus* L., and *Culex quinquefasciatus* Say in normal landing (dorsal-ventral axis is roughly perpendicular to human skin) and smashed/damaged (lateral/dorsal of mosquito's thorax is touch on human skin) condition, respectively. Table 1 summarized the labels, descriptions, and examples of the images in the dataset.

Ethics statements.

Ethical approval for using participants palm and mosquito imaging was obtained from the ethics commission of the Universiti Malaysia Sabah (EM1012/2021). All authors confirm that we have complied with all relevant ethical regulations.

Data Records

The image dataset consists of six root files which are raw image data of three mosquito species with two conditions, respectively, and one data pre-processed file that could serve as an authenticated dataset in recognise three of the mosquitoes, and subsequently applied by potential user such as machine learning engineer, apps developer, data scientist, etc. The ultimate goal for the application can benefit in developing a more effective tools in recognise the mosquito species, which is crucial in mosquito surveillance. The dataset is publicly available in Mendeley Data, Identification number: https://doi.org/10.17632/zw4p9kj6nt.2[8].

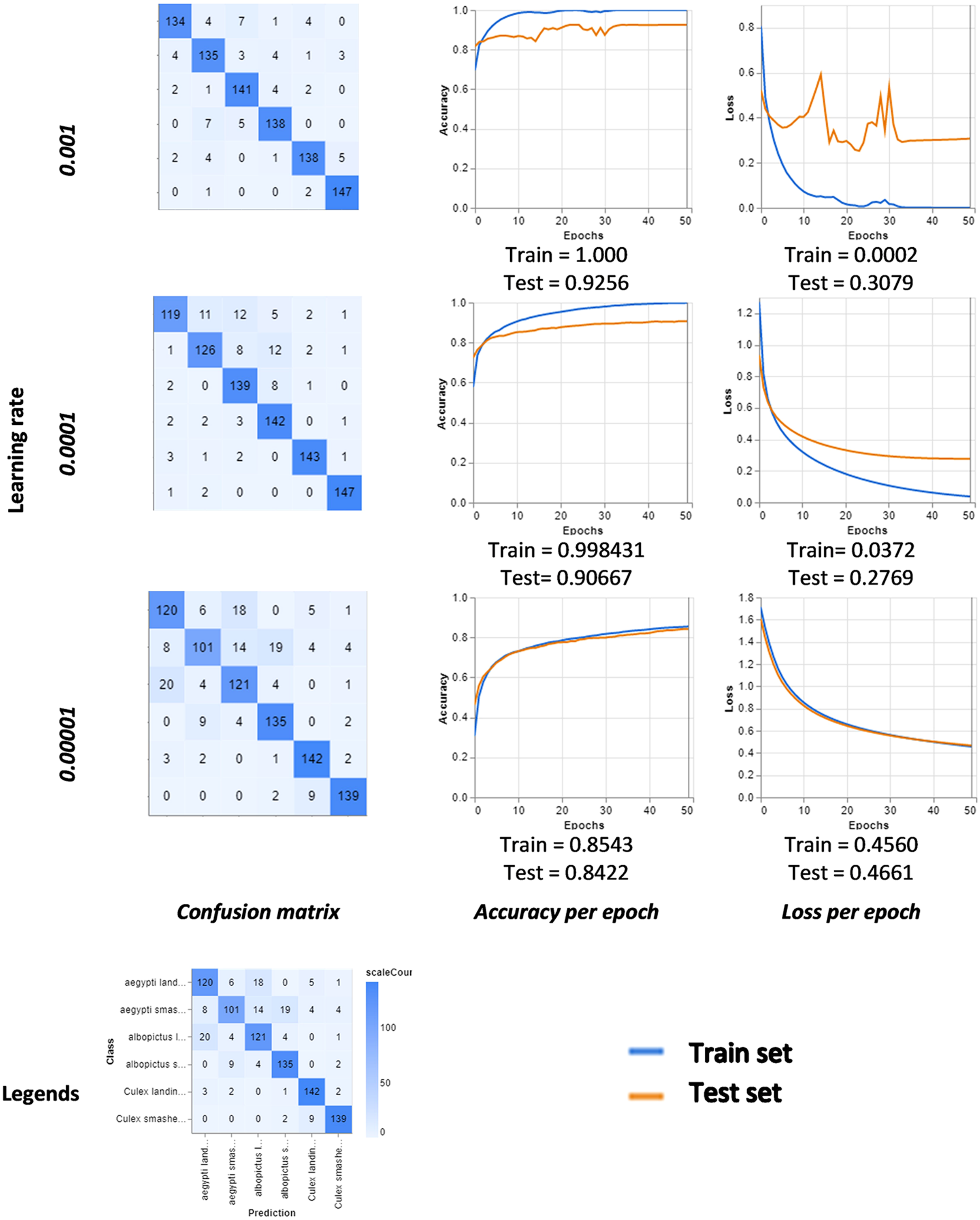

Fig. 4 Confusion matrix, accuracy, and error loss of the pilot test of a deep learning model by using the dataset at three learning rates.

Technical Validation

Sources of mosquito and annotation validation. The source of mosquito adults is the pure bred of the susceptible strain of *Ae. aegypti, Ae. albopictus* and *Cx quinquefasciatus* from Vector Control Research Unit (VCRU), Universiti Sains Malaysia. The mosquitoes were cultured in insectarium for more than 20 years and used for the WHO insecticides susceptibility test[9,10]. Furthermore, before and after the image acquisition, the taxonomy of the mosquito were validated by two medical entomologist

A pilot test with a basic model build-up. We conducted a pilot test on the datasets to validate the quality of the dataset in terms of the feasibility of deep convolutional neural networks (DCNN) model construction. We utilize a web-based tool from Google Creative Lab—Teachable Machine 2.0—that allow us to train a deep learning model with no coding required[11,12]. The data splitting and partitioning used for training and testing are: - training set (85%) and the prediction is carried out on a testing set (15%). The platform also allows us to fine-tune the model with hyperparameters, such as the learning rate, batch size, and epoch. We demonstrate the output of the models by using the datasets at three levels of learning rates - 0.01, 0.001, and 0.0001, which controls the rate of the change to the model during each step of the optimization process. Figure 4 summarises the result – confusion matrix, training, and testing accuracy and loss, respectively for validating the dataset for deep learning model construction.

Usage Notes

The dataset contains a data pre-processed file that has data that have been augmented with four degrees of rotation – 0°, 90°, 180°, 270°, and partitioned into a training and testing set, and one prediction set to evaluate the model performance. Therefore, the file directory in Mendeley can be used directly as a URL and imported into the programming environment. Nevertheless, the dataset posted some limitations as below:

1. Lack of human skin tone diversity. The volunteers that participated in this dataset were Asian, and therefore is not covering the skin tone background of American, African, European, and Australian
2. Image data were taken in a high-resolution camera and under standardized laboratory conditions. The images were acquired by using a DSLR camera and under a condition of enough light illumination. Therefore, images from a smartphone that have been internally processed to enhance the visualization of an image and images from the field may not be recognized by the model that constructed by this dataset.
3. The dataset consists of only three mosquito species. The dataset consists of two visually similar species – *Aedes aegypti* and *Aedes albopictus*; nevertheless, other visually similar mosquitoes such as *Armigeres* are not covered in the dataset.

Code availability

There is no customized code in generation or processing of datasets.

Received: 5 April 2022; Accepted: 8 July 2022;
Published online: 15 July 2022

References

1. Kweka, E. J. & Mahande, A. M. Comparative evaluation of four mosquitoes sampling methods in rice irrigation schemes of lower Moshi, northern Tanzania. *Malar. J.* **8**(1), 1–5 (2009).
2. Gao, Q. *et al.* Comparison of mosquito population composition and dynamics between human-baited landing and CO2-baited trapping monitoring methods. *Chin J Hyg Insect Equip.* **21**, 254–8 (2015).
3. Lima, J. B. P., Galardo, A. K. R., Bastos, L. S., Lima, A. W. D. S. & Rosa-Freitas, M. G. MosqTent: An individual portable protective double-chamber mosquito trap for anthropophilic mosquitoes. *PLoS Negl Trop Dis.* **11**(3), e0005245 (2017).
4. WHO-World Health Organization. "Vector-borne diseases," Who.int, https://www.who.int/news-room/fact-sheets/detail/vector-borne-diseases (Mar 2020).
5. Krizhevsky, A., Sutskever, I., & Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Adv Neural Inf Process Syst*, **25** (2012).
6. He, K., Zhang, X., Ren, S., & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conf. Comput. Vis. Pattern Recognit* (pp. 770–778) (2016).
7. Simonyan, K., & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
8. Ong, S.-Q. Mosquito-on-human-skin. *Mendeley Data* https://doi.org/10.17632/zw4p9kj6nt.2 (2022).
9. WHO-World Health Organization. Insecticide Resistance Monitoring in Disease Vectors Procedures and conditions for supply of test kits https://www.who.int/malaria/areas/vector_control/WHO_test_kit_catalogue_and_requisition_form_may2013.pdf?ua=1.
10. Urban and vector control. School of Biological Sciences, Universiti Sains Malaysia https://bio.usm.my/services/urban-and-vector-control/
11. Ong, S. Q., Ahmad, H. & Majid, A. H. A. Development of a deep learning model from breeding substrate images: a novel method for estimating the abundance of house fly (Musca domestica L.) larvae. *Pest Manag. Sci.* **77**(12), 5347–5355 (2021).
12. Ong, S. Q., Ahmad, H., Nair, G., Isawasan, P. & Majid, A. H. A. Implementation of a deep learning model for automated classification of Aedes aegypti (Linnaeus) and Aedes albopictus (Skuse) in real time. *Sci. Rep.* **11**(1), 1–12 (2021).

Acknowledgements

The work was partly supported by Ministry of Higher Education Fundamental Research Grant Scheme (FRGS) (FRGS/1/2021/STG03/KDUPG/02/1) SQO, Principal Investigator.

Author contributions

S.Q.O. compiled the data, created the first dataset version, and wrote the first version of the manuscript with inputs from H.A. and S.Q.O. All authors contributed substantially to providing data, checking the information on distribution and status of the species.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to S.-Q.O.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.