



OPEN

DATA DESCRIPTOR

# An annotated image dataset of medically and forensically important flies for deep learning model training

Song-Quan Ong<sup>1</sup>✉ & Hamdan Ahmad<sup>2</sup>

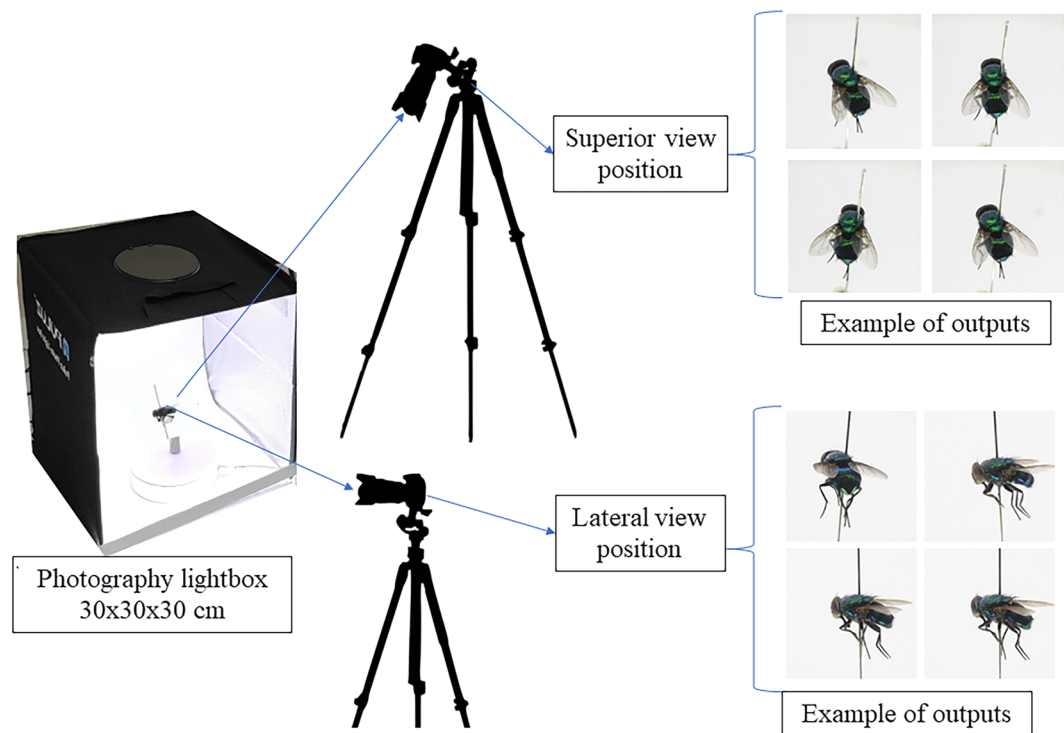
Conventional methods to study insect taxonomy especially forensic and medical dipterous flies are often tedious, time-consuming, labor-intensive, and expensive. An automated recognition system with image processing and computer vision provides an excellent solution to assist the process of insect identification. However, to the best of our knowledge, an image dataset that describes these dipterous flies is not available. Therefore, this paper introduces a new image dataset that is suitable for training and evaluation of a recognition system involved in identifying the forensic and medical importance of dipterous flies. The dataset consists of a total of 2876 images, in the input dimension ( $224 \times 224$  pixels) or as an embedded image model ( $96 \times 96$  pixels) for microcontrollers. There are three families (Calliphoridae, Sarcophagidae, Rhiniidae) and five genera (Chrysomya, Lucilia, Sarcophaga, Rhiniinae, Stomorhina), and each class of genus contained five different variants (same species) of fly to cover the variation of a species.

## Background & Summary

Flies have strongly associated with various microorganisms such as bacteria, viruses, protozoa, fungi, and helminth parasites. Some species are serious medical pests due to their capability as mechanical vectors that carry pathogens or parasitize livestock or humans, causing myiasis<sup>1–3</sup>. Other species called carrion flies are forensically important and considered important scavengers due to their necrophagous feeding behaviors<sup>4</sup>. In terms of forensic entomology, some species provide an alternative way to estimate the minimum post-mortem interval (PMI) of a victim in forensic investigations<sup>5</sup>. Flies are important in many different fields and show great diversity in morphology, behavior, and ecology. Conventional taxonomy and systematic identification of the flies especially those involved in forensic and medical still rely heavily on human observation with or without the aid of microscopic tools. However, these methods are often tedious, time-consuming, labor-intensive, and expensive. Therefore, computer vision and deep learning could provide an excellent alternative for these global challenges, and a suitable dataset is a key to an accurate and reliable machine learning model. This paper provides an image dataset that has three common families of flies that are crucial in medical and forensic entomology. The images were formatted in a JPEG file, processed into  $224 \times 224$  pixels for machine learning or convolutional neural network (CNN) training, and  $96 \times 96$  pixels with smaller file size, as an embedded image model for the microcontroller. Both dimensions consist of 96 dpi resolution and 24-bit depth, and images were annotated into taxonomy levels of genus.

The image dataset is a raw data that could serve as an authenticated dataset in recognise three families or five genera of medical and forensically important flies. Subsequently, the dataset could be used by potential user such as machine learning engineer, apps developer, data scientist, taxonomist, medical and forensic entomology etc. Figure 1 illustrate the general workflow to record the dataset and organised into the labelled classes, and Table 1 summarizes the structure and labels of the dataset.

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



**Fig. 1** General workflow to record the dataset and organised into the labelled classes.

Annotation (number of images)	
Family	Genera
• Calliphoridae (1318)	• Chrysomya (731)
• Sarcophagidae (570)	• Lucilia (587)
• Rhiniidae (988)	• Sarcophaga (570)
	• Rhiniinae (488)
	• Stomorphina (500)

**Table 1.** Summary of the image annotations.

Methods

**Resources of insect specimen.** The insect specimens were obtained from the Insect Collection Room of Borneensis, Institute for Tropical Biology and Conservation (ITBC), Universiti Malaysia Sabah (UMS). The Insect Collection Room kept more than 200,000 insect specimens that have been preserved and stored in a compactor at temperature of 18 °C and humidity of 40 ± 5%. The taxonomy of insects has been identified and validated until the taxonomy of genus-level by two taxonomists. The adult stage of the insects were used for image acquisition, and the annotation of the dataset was set on the genus level.

**Data collection.** The insects’ images were acquired by a digital single-lens reflex (DSLR) camera (Canon EOS 50D, 15.0 MP APS-C CMOS Sensor) with Tamron 90 mm f/2.8 Di Macro. The image acquisitions process was conducted in a photography lightbox 30x30x30 cm (Fig. 2) with 34 W white light illumination. The insect specimen was placed in a pin on an electronic motorized rotating plate (the 30 s per resolution) and the camera acquired the images with three-frames per second. The images of the insect were acquired at two levels of position – superior view and lateral view of the insect (Fig. 2). The quality of the images (sharpness, brightness etc.) was checked after the acquisition, poor quality images were removed and caused the genera having different total number of images (Table 1). Table 2 shows the description and example of annotated classes of flies.

**Ethics Statements.** All authors confirm that we have complied with all relevant ethical regulations.

Data Records

The dataset is publicly available in figshare, with direct URL to data: <https://doi.org/10.6084/m9.figshare.19607193.v2><sup>6</sup>. Figure 1 illustrated the general workflow to record the dataset and organised it into the labelled classes. In general, after the images were acquired from the museum specimen, they were formatted into 224×224 for DCNN model training or 96×96 for embedded image model training. Users can train the model based on the label of genus – five classes. The image of a genus consists of 5 variants of specimens and consists of

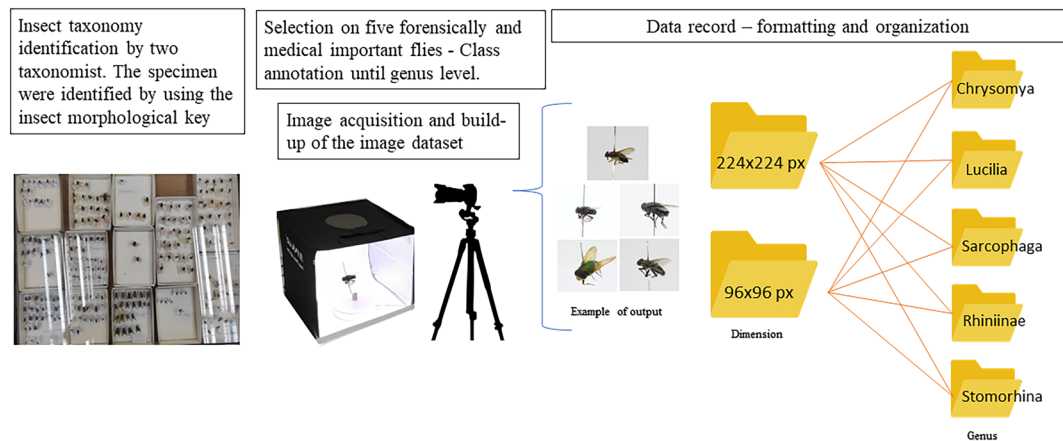


Fig. 2 Data collection process.


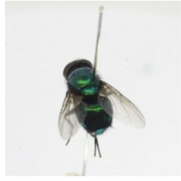



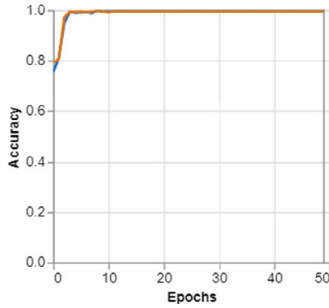
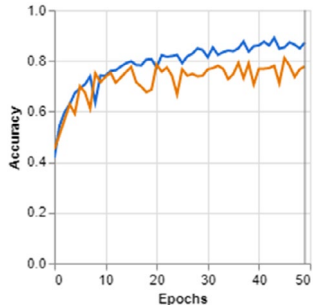
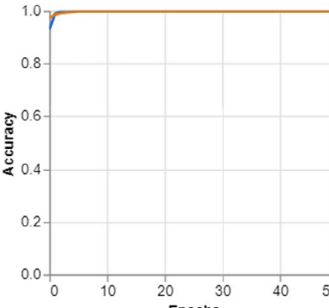
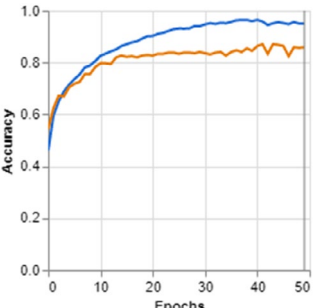
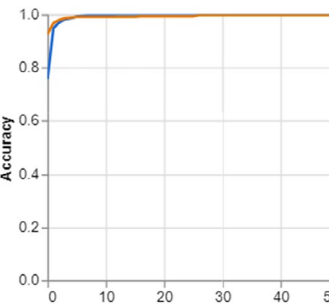
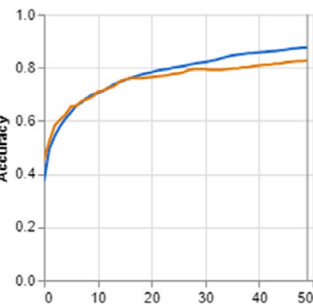
Order	Family	Genus	Examples
Diptera	Calliphoridae	Chrysomya	
		Lucilia	
	Sarcophagidae	Sarcophaga	
	Rhiniidae	Rhiniinae	
		Stomorhina	

Table 2. Description and example of annotated classes of flies.

360° view of a specimen. Therefore, for further species level identification by other user (to build a species level recognition system), we re-organized the images according to the individual specimen, and supply as a folder in this dataset.

Technical Validation

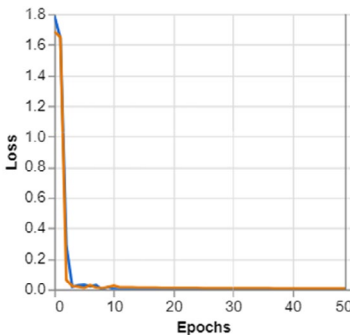
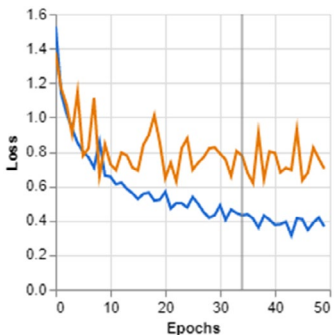
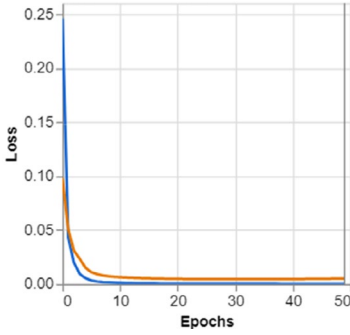
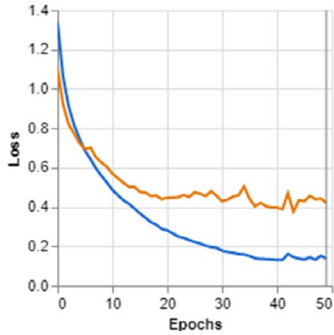
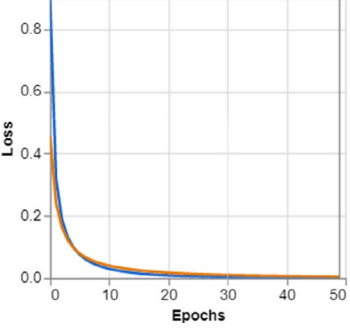
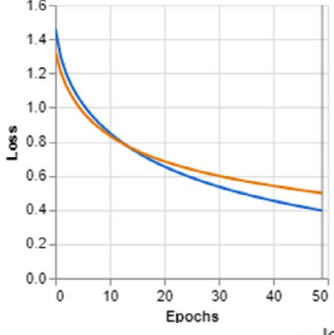
**Taxonomy.** The taxonomy of insects has been identified by two taxonomists based on the morphological practical keys to families, subfamilies, and genera as described by<sup>4,7–9</sup>.

		Dimension	
		224x224	96x96
Learning rate	0.01	Train: 1.00000 Test: 0.99538 	Train: 0.87224 Test: 0.77829 
	0.001	Train: 1.00000 Test: 0.99769 	Train: 0.95000 Test: 0.85910 
	0.0001	Train: 1.00000 Test: 0.99769 	Train: 0.87633 Test: 0.82679 

**Table 3.** Pilot test result: Training and testing accuracy of the deep learning model by using two different dimensions of dataset at three learning rates; blue line is representing training accuracy; orange line is representing testing accuracy.

**A pilot test with a model build-up.** We conducted a pilot test on the datasets to validate the quality in the terms of the development of a deep convolutional neural networks (DCNN) model. We utilize a web-based tool from Google Creative Lab—Teachable Machine 2.0—that is able to train a computational model with no coding required<sup>10</sup>.

The data splitting conducted on this dataset that used for training and testing are: - training (85%) and the prediction is carried out on the testing split (15%), which the images were randomly selected and not repeated with the train split. The platform also allows us to fine-tune the model with hyperparameters, such as the learning rate, batch size, and epoch. For the purpose of dataset quality validation but not presenting a new interpretations of deep learning model construction, this pilot test standardized the batch size to 16 and epoch to 50, and we only fine-tune with three levels of learning rate – 0.0001, 0.001, 0.01 to demonstrate the output of models by using the datasets Table 3 shows the result of the accuracy for the train and test split of the dataset, respectively. The learning curve consists of accuracy on the y-axis, which is the evaluation metric of the probability of

		Dimension	
		224x224	96x96
Learning rate	0.01	Train: 0.00000 Test: 0.0072 	Train: 0.36564 Test: 0.701945 
	0.001	Train: 0.00000 Test: 0.005053 	Train: 0.13772 Test: 0.4204 
	0.0001	Train: 0.00000 Test: 0.0045862 	Train: 0.39571 Test: 0.49825 

**Table 4.** Pilot test result: Training and testing loss of a deep learning model by using two different dimensions of dataset at three learning rates; blue line is representing training function loss, orange line is representing testing function loss.

accurate prediction against the epoch on the x-axis, which is the number of passes of the entire training dataset the deep learning algorithm has completed)<sup>11</sup>. Table 4 shows the loss for the train and test split of the dataset, respectively. The function loss curve consists of a loss function on the y-axis, which is a measurement of the differences between predicted and true values against the epoch on the x-axis. Table 5 shows the confusion matrix from the prediction on the test split (based on 433 images), which is a summary of prediction results that consists of correct and incorrect predictions (Prediction against the true value)<sup>11</sup>, or more machine learning model evaluation metrics such as precision and recall can be obtained from the confusion matrix as described in<sup>12</sup>.

		Dimension	
		224x224	96x96
Learning rate	0.01		
	0.001		
	0.0001		

**Table 5.** Confusion matrix of the deep learning model by using two different dimensions of dataset at three learning rates; the blue intensities indicate the frequency counts, the darker the blue colour the higher the frequency. Chy- Chrysomya; Luc- Lucilia; Sto- Stomorphina; Sar- Sarcophagidae; Rhi- Rhiniinae Number of images used for pilot test training and testing [class (train: test)]: Chy (621:110); Lucilia (499:88); Sto (425:75); Sar (484:86); Rhi (414:74).

### Usage Notes

The dataset posted some limitations as below:

1. Annotation of the specimen until genus level. The specimen was identified until genus level due to the restriction of the morphology key provided by<sup>7–9</sup>, and therefore able to be reused and identified until species level, and subsequently a recognition system until species level.
2. The dataset consists of imbalanced classes of images for the genus. This was due to the removal of blurry and poor-quality images during the process of image acquisition.

### Code availability

The original images were resized into  $224 \times 224$  and  $96 \times 96$  by using the web-based tools – <https://teachablemachine.withgoogle.com> by choosing a new image project with standard image model or embedded image model, respectively. There is no customized code in generation or processing of datasets.

Received: 25 April 2022; Accepted: 8 August 2022;

Published online: 20 August 2022

## References

1. Kano, R. & Shinonaga, S. Calliphoridae (Insecta: Diptera). (Biogeographical Society of Japan, National Science Museum, 1968).
2. Sawabe, K. *et al.* Detection and isolation of highly pathogenic H5N1 avian influenza A viruses from blow flies collected in the vicinity of an infected poultry farm in Kyoto, Japan, 2004. *Am. J. Trop. Med.* **75**(2), 327–332 (2006).
3. Tumrasvin, W., Kurahashi, H. & Kano, R. Studies on medically important flies in Thailand VII. Report on 42 species of calliphorid flies, including the taxonomic keys (Diptera: Calliphoridae). *Bull. Tokyo Dent. Coll.* **26**, 243–272 (1979).
4. Singh, K. I., Kurahashi, H. & Kano, R. A preliminary key to the common calliphorid flies of Peninsular Malaysia (Insecta: Diptera). *Bull. Tokyo Dent. Coll.* **26**(1), 5–24 (1979).
5. Catts, E. P. & Goff, M. L. Forensic entomology in criminal investigations. *Annu. Rev. Entomol.* **37**(1), 253–272, <https://doi.org/10.1146/annurev.en.37.010192.001345> (1992).
6. Ong, S. Q. Medical and forensically important flies. *Figshare* <https://doi.org/10.6084/m9.figshare.19607193.v2> (2022).
7. Kurahashi, H., Benjaphong, N. & Omar, B. Blow flies (Insecta: Diptera: Calliphoridae) of Malaysia and Singapore. *Raffles Bulletin of Zoology, School of Biological Sciences, University of Singapore, Singapore*, 1–88 (1997).
8. Nazni, W. A., Jeffrey, J., Heo, C. C., Chew, W. K. & Lee, H. L. Illustrated keys to adult flies of forensic importance in Malaysia. (Institute for Medical Research, 2011).
9. Yang, S. T., Kurahashi, H. & Shiao, S. F. Keys to the blow flies of Taiwan, with a checklist of recorded species and the description of a new species of *Paradichosia* Senior-White (Diptera, Calliphoridae). *ZooKeys*, **434**, 57 (2014).
10. 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).
11. Goodfellow, I., Bengio, Y. & Courville, A. Deep learning. (MIT press, 2016).
12. Markoulidakis, I. *et al.* Multiclass Confusion Matrix Reduction Method and Its Application on Net Promoter Score Classification Problem. *Technologies* **9**(4), 81 (2021).

## 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](http://www.nature.com/reprints).

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



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022