



OPEN

DATA DESCRIPTOR

# An annotated wing interferential pattern dataset of dipteran insects of medical interest for deep learning


Arnaud Cannet<sup>1,7</sup>, Camille Simon-chane<sup>2,7</sup> , Aymeric Histace<sup>2,7</sup>, Mohammad Akhoundi<sup>3,7</sup>, Olivier Romain<sup>2,7</sup>, Marc Souchaud<sup>2,7</sup>, Pierre Jacob<sup>2,4</sup>, Darian Sereno<sup>5</sup>, Philippe Bousses<sup>6</sup> & Denis Sereno<sup>5,6,7</sup>  

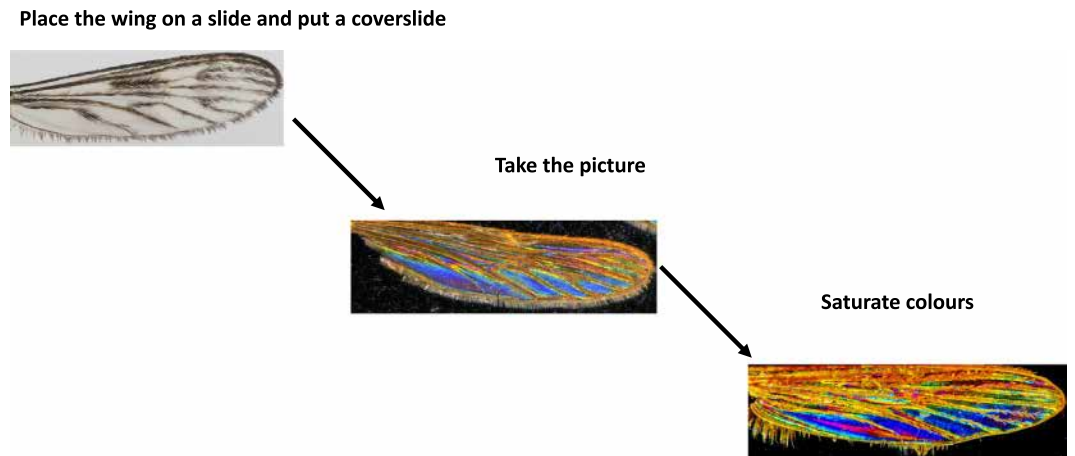
Several Diptera species are known to transmit pathogens of medical and veterinary interest. However, identifying these species using conventional methods can be time-consuming, labor-intensive, or expensive. A computer vision-based system that uses Wing interferential patterns (WIPs) to identify these insects could solve this problem. This study introduces a dataset for training and evaluating a recognition system for dipteran insects of medical and veterinary importance using WIPs. The dataset includes pictures of Culicidae, Calliphoridae, Muscidae, Tabanidae, Ceratopogonidae, and Psychodidae. The dataset is complemented by previously published datasets of Glossinidae and some Culicidae members. The new dataset contains 2,399 pictures of 18 genera, with each genus documented by a variable number of species and annotated as a class. The dataset covers species variation, with some genera having up to 300 samples.

## Background & Summary

Blood-sucking insects, such as mosquitoes, ticks, and sandflies, transmit viral, parasitic, or bacterial pathogens that cause severe diseases, including arboviruses, malaria, Lyme disease, and others. Climate fluctuations, global economic growth, migration, and increased trade are factors that influence the distribution of many organisms, not just insects. The expansion of the tiger mosquito, *Aedes albopictus*, into new climates, which it has recently done, is a concern as it is an established vector for Zika, chikungunya, and dengue viruses<sup>1</sup>. To address the threats of emerging vector-borne diseases, robust and rapid species identification is crucial. However, current global vector surveillance systems are unstandardized and facing a global shortage of entomologists. Identification is typically conducted by skilled personnel using quantitative and qualitative criteria, such as specimen size, shape, texture, or the presence or absence of certain key features. Nevertheless, when sympatric species have medical importance, these distinctions based on morphological characteristics may not always be discriminative. Additionally, older adult specimens may have missing or damaged body parts or characters essential for exact identification. These samples can also have damage in critical diagnostic character regions, making it challenging to separate vector species from closely related non-vector ones. Finally, the identification of dipteran species is complex and requires highly specialized expertise if diversity is to be fully addressed.

The identification of specimens at the species/subspecies level is crucial during proactive surveys to address health risks associated with their introduction or presence. However, morphological criteria are inadequate when samples are damaged or for extensive geographic surveys, and identification methods based on heavy biological protocols (DNA and mass spectrometry) are expensive, incompatible with in-field analyses, and

<sup>1</sup>Direction des affaires sanitaires et sociales de la Nouvelle-Calédonie, Nouméa, France. <sup>2</sup>ETIS UMR 8051, Cergy Paris University, ENSEA, CNRS, F-95000, Cergy, France. <sup>3</sup>Parasitology-Mycology, Hopital Avicenne, AP-HP, Bobigny, France. <sup>4</sup>Univ. Bordeaux, CNRS, Bordeaux INP, LaBRI, UMR 5800, F-33400, Talence, France. <sup>5</sup>InterTryp, Univ Montpellier, IRD-CIRAD, Infectiology, Entomology and One Health Research Group, Montpellier, France. <sup>6</sup>MIVEGEC, Univ Montpellier, CNRS, IRD, Montpellier, France. <sup>7</sup>These authors contributed equally: Arnaud Cannet, Camille Simon-chane, Aymeric Histace, Mohammad Akhoundi, Olivier Romain, Marc Souchaud, Denis Sereno.  e-mail: [denis.sereno@ird.fr](mailto:denis.sereno@ird.fr)



**Fig. 1** Schematic representation of the image acquisition and processing, example of *Culiseta* samples.

destructive to samples, including pathogens. Therefore, fine-grain, non-destructive entomological surveillance methods that allow for later pathogen identification with high efficiency, accuracy, and reduced costs are needed. Guidelines for mosquito surveillance are publicly available<sup>2</sup>.

Thin-film interference generated on the transparent wings with a thin membrane allows the formation of a colored pattern. With incoming external light wings in light-absorbing and dark environments, WIPs are displayed on the wing membranes. These WIPs vary significantly among species but faintly between specimens of the same species or between sexes. Since the 2010s, WIPs (Wing Interference Patterns) have received significant attention for their potential as a method for species identification<sup>3–5</sup>. The role of WIPs in sexual selection of *Drosophila melanogaster* is such that males with more vivid wings are more attractive to females than to males with dull wings<sup>6</sup>. This enhances the visual aspect of the mating tool array of *Drosophila*. Unlike iridescence, which depends on the angle of a flat film, wing structures act as diopters, making WIPs appear non-iridescent<sup>4</sup>. The Newton color series observed on wings resembles that of a soap bubble and is proportional to the wing membrane thickness at any given point, which helps in species identification<sup>7,8</sup>. Collecting colored patterns is relatively easy, and deep learning-trained descriptors extracted from pictures have demonstrated exceptional accuracy in identifying insect species<sup>9–13</sup>. The image dataset, raw or processed, combined with already publicly available ones on Glossinidae, Some Culicidae members can serve as an authenticated dataset for recognizing seven families or twenty-one genera, including those with medical or veterinary interests, and can be utilized by users such as machine learning engineers, app developers, data scientists, taxonomists, and medical and veterinary entomologists.

## Methods


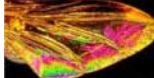
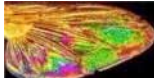
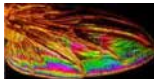

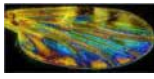


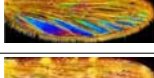
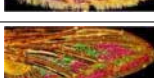

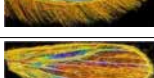




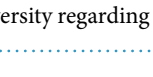

This method has been previously used, on Glossinidae and some Culicidae members, and results of the identification process were published<sup>11–13</sup>. Here, we provide dataset that complete other previously published to expand it and on which the procedure can be applied to dipteran insects belonging to 7 families (Culicidae, Calliphoridae, Muscidae, Glossinidae, Tabanidae, Ceratopogonidae, Psychodidae) and 21 genera. The method consists of selecting insects from 7 families; whenever possible, at least ten specimens, including male and female ones, were chosen for being included in the database. Then, wings are dissected, and WIPs are captured to fill the database. The automatic classification was performed as previously described<sup>9,11–13</sup> using a larger dataset of Dipteran insects collected during this study<sup>14</sup>.

**Resources of specimen.** The insect specimens were gained from ARIM collection belonging to IRD (Institut de Recherche pour le Développement) (<https://arim.ird.fr/>) from well-established laboratory-reared or field-caught specimens. The ARIM collection kept more than 100,000 preserved and stored insect specimens.

**Data collection.** Insect wings were dissected and deposited on a glass slide. Samples preserved in 70° ethanol were layered overnight at room temperature on a glass slide before being processed. A cover slide is deposited on the sample before image acquisition. The pictures are taken with a Keyence™ VHX 1000 microscope, with the VH-Z20r camera and a VH K20 adapter, an illumination incidence of 10°. Image acquisition was performed using the High Dynamic Range (HDR) function. Magnification was adjusted to ensure constant-size pictures; a schematic representation of the process and output is given in Fig. 1.

The numerical parameters settled were as follows:

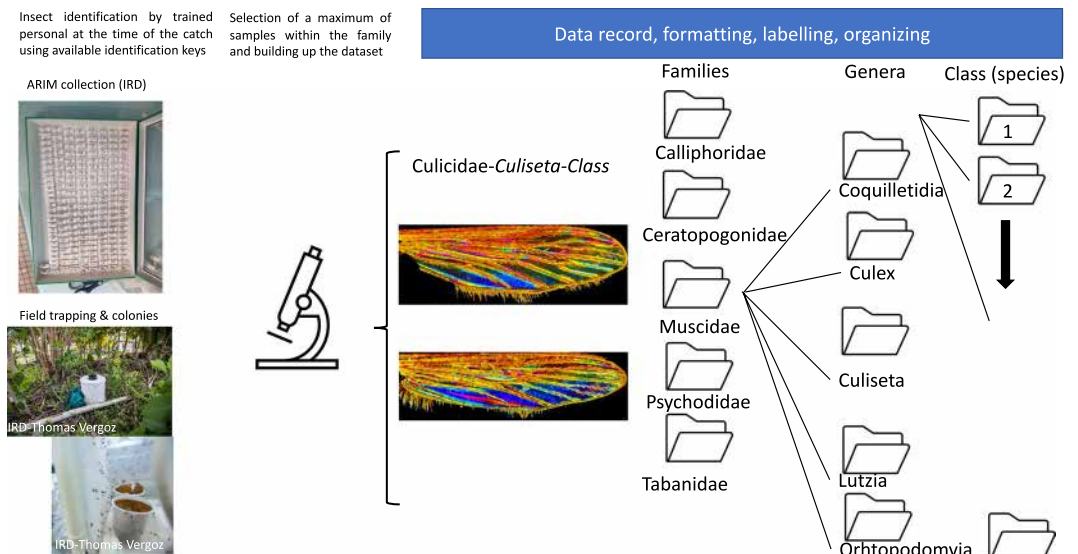
- Camera White Balance: 3200 K
- Shutter Speed: preset 1/15(sec)
- Gain: 0 dB
- Frame rate 15 F/s

Label		
Family	Genera	
Callyphoridae	<i>Auchmeromyia</i>	
	<i>Chrysomyia</i>	
	<i>Hemipyrelia</i>	
	<i>Lucilia</i>	
	<i>Tryciclea</i>	
Ceratopogonidae	<i>Culicoides</i>	
Culicidae	<i>Coquillettidia</i>	
	<i>Culex</i>	
	<i>Culiseta</i>	
	<i>Lutzia</i>	
	<i>Orthopodomyia</i>	
Muscidae	<i>Haematobia</i>	
Psychodidae	<i>Lutzomyia</i>	
	<i>Phlebotomus</i>	
	<i>Sergentomyia</i>	
Tabanidae	<i>Atylotus</i>	
	<i>Chrysops</i>	
	<i>Tabanus</i>	

**Table 1.** Illustration of the current dataset diversity regarding families, genera, and specimens.

HDR function:

- Brightness: 15%
- Texture: 15%
- Contrast: 45%
- Color: 100%



**Fig. 2** Schematic representation of the image acquisition and labeling workflow.

Next, the luminosity, contrast, shadow, reflection, and saturation were settled at 80, 100, 0, 0, and 100%, respectively, using Window 7 familial edition. Table 1 show examples of processed image for dataset implementation.

### Data Records

The dataset is publicly available in Figshare<sup>14</sup>. Figure 2 illustrates the workflow to record and organize it. Specimens belonging to the Culicidae (*Culiseta annulata*) family were used as examples to demonstrate the process, all samples being processed according to the same workflow. Only specimens displaying wing integrity >60% (arbitrarily set) with a distinguishable Wing Interferential Pattern are filled in the database. The sole exception is the Tabanus specimen, which doesn't display a distinguishable WIP.

The origin of samples is presented in Fig. 3; note that the geographic origin of specimens from laboratory-reared colonies is not representative of the original one.

The geographic distribution depicts that most samples originate from Africa, Madagascar, and La Reunion Island. Most specimens having a European origin are colony-reared ones.

For further species-level identification (species recognition system), the images were organized in individual specimens in the genus folder of the dataset. Spreadsheets are organized as follows: numeration of the picture, Order, family, Genus, and Class. Each class corresponds to an individual species see Fig. 2. In the dataset, the unique image of Tabanus wing filled doesn't display WIPs, and efforts must be engaged to gather them.

### Technical Validation

**Taxonomy.** The identification of insects at the genus, species, or subspecies level was performed by trained entomologists using the available keys at the time of their catch. Only the adult stage was used for WIPS image acquisition.

**A pilot test.** The method has been validated in our previous work<sup>9,11–13</sup>

### Usage Note

#### Usage of the dataset.

1. Entomologists can use the dataset gathering 2,399 pictures of 18 genera, for training for taxonomic and/or machine learning engineers.
2. Combining the dataset repository provided in this study<sup>14</sup> with the previously published dataset<sup>15</sup> allow to extend the diversity to 5516 pictures of 7 families (Culicidae, Calliphoridae, Muscidae, Glossinidae, Tabanidae, Ceratopogonidae Psychodidae) and 21 genera. See the Table 2 for Family, Genus and picture number filled in each dataset.

#### Limitations of the dataset:

1. The dataset consists of imbalanced classes (species) of images due to difficulties in gathering enough specimens because we cannot gather them in the ARIM database, we do not get financial resources to collect them in natura, or there are no colonies available.
2. The dataset does not represent the whole family/Genera/species diversity of dipteran insects of medical and veterinary interest.
3. Be aware that images in the dataset were resized, computed, and processed in terms of luminosity, contrast,



**Fig. 3** Geographic distribution of samples using Google Looker Studio (<https://lookerstudio.google.com/overview>).

Label (Number of pictures)		
Family	Genera	Repository reference
Calliphoridae (35)	<i>Auchmeromyia</i> (3)	14
	<i>Chrysomyia</i> (7)	14
	<i>Hemipyrelia</i> (5)	14
	<i>Lucilia</i> (6)	14
	<i>Tryciclea</i> (14)	14
Ceratopogonidae (28)	<i>Culicoides</i> (28)	14
Culicidae (1992)	<i>Aedes</i> (502)	15
	<i>Anopheles</i> (849)	15
	<i>Coquillettidia</i> (5)	14
	<i>Culex</i> (591)	14
	<i>Culiseta</i> (13)	14
	<i>Lutzia</i> (14)	14
	<i>Orhtopodomyia</i> (18)	14
Glossinidae (1766)	<i>Glossina</i> (1766)	15
Muscidae (4)	<i>Haematobia</i> (4)	14
Psychodidae (1673)	<i>Lutzomyia</i> (294)	14
	<i>Phlebotomus</i> (1272)	14
	<i>Sergentomyia</i> (107)	14
Tabanidae (18)	<i>Atylotus</i> (5)	14
	<i>Chrysops</i> (12)	14
	<i>Tabanus</i> (1)	14

**Table 2.** Families and genera of WIPs pictures included in the datasets.

shadow, reflection, and saturation, which is limiting for applications requiring wing thickness measurement deduced from Newton color series

- The eligibility criteria for data inclusion in the dataset are not restrictive; damaged samples were included that might be limiting for some application

## Code availability

The source code is publicly available in GitHub, with a direct URL: <https://github.com/marcensea/diptera-wips.git>.

Received: 19 June 2023; Accepted: 11 December 2023;

Published online: 02 January 2024

## References

1. Kraemer, M. U. G. *et al.* Past and future spread of the arbovirus vectors *Aedes aegypti* and *Aedes albopictus*. *Nat Microbiol* **4**, 854–863, <https://doi.org/10.1038/s41564-019-0376-y> (2019).
2. Schaffner, F. *et al.* Development of guidelines for the surveillance of invasive mosquitoes in Europe. *Parasites vect* **6**, 209 (2013).
3. Buffington, L. M. & Sandler, J. R. The occurrence and phylogenetic implications of wing interference patterns in Cynipoidea (Insecta: Hymenoptera). *Invertebr Syst* **25**, 586–597 (2012).
4. Shevtsova, E., Hansson, C., Janzen, D. H. & Kjærandsen, J. Stable structural color patterns displayed on transparent insect wings. *Proc Natl Acad Sci USA* **108**, 668–673, <https://doi.org/10.1073/pnas.1017393108> (2011).
5. Simon, E. Preliminary study of wing interference patterns (WIPs) in some species of soft scale (Hemiptera, Sternorrhyncha, Coccoidea, Coccidae). *Zookeys*, 269–281, <https://doi.org/10.3897/zookeys.319.4219> (2013).
6. Katayama, N., Abbott, J. K., Kjærandsen, J., Takahashi, Y. & Svensson, E. I. Sexual selection on wing interference patterns in *Drosophila melanogaster*. *Proc Natl Acad Sci USA* **111**, 15144–15148, <https://doi.org/10.1073/pnas.1407595111> (2014).
7. Li, M. *et al.* Feasibility of Insect Identification Based on Spectral Fringes Produced by Clear Wings. *IEEE J Sel Top Quantum Electron* **29**, 1–8, <https://doi.org/10.1109/JSTQE.2022.3218218> (2023).
8. Müller, L. *et al.* Remote Nanoscopy with Infrared Elastic Hyperspectral Lidar. *Advanced Science* **10**, 2207110, <https://doi.org/10.1002/adv.202207110> (2023).
9. Souchaud, M. *et al.* Mobile Phones Hematophagous Diptera Surveillance in the field using Deep Learning and Wing Interference Patterns. *2018 IFIP/IEEE International Conference on Very Large Scale Integration (VLSI-SoC)*, 159–162 (2018).
10. Sereno, D., Cannet, A., Akhouni, M., Romain, O. & Histace, A. Système et procédé d'identification automatisée de diptères hématophages. France PCT/FR15/000229. patent (2015).
11. Cannet, A. *et al.* Wing Interferential Patterns (WIPs) and machine learning for the classification of some *Aedes* species of medical interest. *Sci Rep* **13**, 17628, <https://doi.org/10.1038/s41598-023-44945-3> (2023).
12. Cannet, A. *et al.* Deep learning and wing interferential patterns identify *Anopheles* species and discriminate amongst *Gambiae* complex species. *Sci Rep* **13**, 13895, <https://doi.org/10.1038/s41598-023-41114-4> (2023).
13. Cannet, A. *et al.* Wing interferential patterns (WIPs) and machine learning, a step toward automatized tsetse (*Glossina* spp.) identification. *Sci Rep* **12**, 20086, <https://doi.org/10.1038/s41598-022-24522-w> (2022).
14. Sereno, D. An exhaustive dataset of Diptera Wing Interferential Patterns (WIPs). *figshare* <https://doi.org/10.6084/m9.figshare.24444937.v2> (2023).
15. Sereno, D. *et al.* Listing and pictures of Diptera WIPs. *figshare* <https://doi.org/10.6084/m9.figshare.22083050.v4> (2023).

## Acknowledgements

We thank Pr. P. Marty and P. Delaunay (CHU Nice) for gaining access to the microscopic facility of the CHU de l'Archet, Nice. Dr. D. Fontenille (UMR MIVEGEC, Montpellier, France) for his support and fruitfully scientific discussions on medical entomology aspects. Mr JP Commes, former CEO of 2CSI, for his enthusiasm for the digital aspect of the project.

## Author contributions

Conceptualisation De.S., A.C., M.A., A.H., C.S.C., O.R. Data acquisition De.S., A.C., M.S., A.H., Da.S. Database construction De.S., Da.S., A.H., P.J., M.S., O.R. Sample collection and arthropod management P.B., De.S. Code management: M.S., A.H., C.S.C., O.R., P.J. Project management De.S., A.H., C.S.C. Writing first draft A.H., De.S., A.C. Writing and editing De.S. A.C., M.A., C.S.C., P.D., A.H.

## Competing interests

The authors declare no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to D.S.

**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 licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence 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 licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024