`

Identifying *Anopheles* Larvae Using a Convolutional Neural Network

Spencer Burke (spencerburke22@gmail.com)

Juan Durante (juan.durante2005@gmail.com)

Christopher Grizzaffi (grizzaffichristopher@gmail.com)

Amyn Macknojia (amyn.macknojia@gmail.com)

1. **Abstract**

As the prominence of citizen science grows and an increasing number of people are able to quickly and efficiently collect mosquito-related data, a method of classification that is accurate, fast, and simple to use becomes a growing necessity in order to be able to mitigate the spread of mosquito-borne diseases like malaria. This data is made publicly accessible through citizen science applications like the GLOBE Observer app. By using the mosquito larvae imagery stored on the GLOBE platform combined with images retrieved from public sources through the use of search engines, our study seeks to train and implement a binary image classifier that can determine, given an image of a mosquito larva, whether or not the photographed larva belongs to the genus *Anopheles*. Approximately 20 percent of the available GLOBE images were deemed to possess sufficient quality to form part of the classifier’s training set and therefore, our model lacks the needed accuracy and training to be available for public implementation. We hope to continue this study by building a larger training set that will not only create a more accurate model, but also contribute to other larvae identification efforts by providing a usable dataset for training any future larvae image identifiers and classifiers, and by refining the parameters and architecture of our classifier to improve accuracy.

**Keywords-** citizen science, deep learning, image classification, *Anopheles*, convolutional neural network

1. **Research Questions**
* Can deep learning models accurately identify *Anopheles* mosquito larvae?
* Can a deep learning model be used to aid malaria initiatives like the President’s Malaria Initiative?

These questions can have profound impacts on ecological and environmental initiatives, specifically those related to vector-borne disease prevention, while also testing the capability and applicability of deep learning in the field. The *Anopheles* mosquito, being the main transmitter of malaria, has a significant ecological impact and a classifier able to rapidly identify its larvae has the potential to be a powerful tool in limiting malaria transmission. In addition, the implementation of such a model also presents a way to take advantage of citizen science, as untrained citizens would have a simple tool to aid them in identification of larvae. This in turn would increase the documentation, and thus the mitigation, of malaria outbreaks.

1. **Introduction and Literature Review**

Many researchers consider mosquitoes to be the most dangerous animals in the world because of the sheer destruction and death they cause; they infect approximately 700 million people with mosquito borne diseases each year and cause over one million deaths annually.In 2019, there were approximately 229 million malaria cases and 409,000 malaria deaths (World Health Organization, 2021), and only *Anopheles* mosquitoes infect people with malaria (Talpako *et al.,* 2019). Consequently, the ability to identify *Anopheles* mosquitoes and prepare for possible epidemics is essential, especially considering the recent prominence of lethal diseases.

The current literature about identifying mosquito larvae using machine learning is limited; there are few research papers and studies pertaining to this field. As a result, much of the research is not purely focused on machine learning combined mosquito identification, but instead on machine learning and mosquito identification as two separate but interdisciplinary fields which stem off one another.

The current literature suggests that *Anopheles* mosquito larvae have noticeable characteristic differences from other mosquito larvae, especially in regard to larvae of the *Culex* and *Aedes* genuses. Most notably, it can be seen even without the use of a microscope that *Anopheles* mosquitoes tend to lie parallel to the water surface, whereas *Culex* and *Aedes* mosquitoes tend to lie at an angle. However, when looking at *Anopheles* mosquitoes through microscopes, the differences become more apparent--while most types of mosquitoes have a siphon, including the *Culex* and *Aedes* mosquitoes, *Anopheles* mosquitoes do not have a siphon. This tends to be the greatest indicator when determining the genus of mosquito present, however, more subtle distinctions are also present, such as the head length (longer in *Anopheles* and *Uranotaenia*) and the arrangement shape and color of comb scales (Burkett-Cadena).

Furthermore, recent innovations in the fields of citizen science and machine learning have led to more accurate models. Developments in machine learning and image processing models have allowed for a greater accuracy in classification for the ability to construct and analyze models in a deeper regard. Therefore, as a result of greater processing power availability and a significant improvement in classifier architecture, modern deep learning models have the potential to be substantially more beneficial than prior generations of models (Sarker *et al.,* 2021).

To summarize, research teams have recently created a machine learning algorithm that can detect certain species. The team used a Convolutional Neural Network and used only 310 images and were able to originally get an accuracy of 70%, and later were able to achieve an accuracy of 96.8% with 200 epochs (Sanchez-Ortiz *et al.,* 2017).



1. **Research Methods**

(i) Data Collection:

Our study used two main sources of data: the GLOBE Observer application and various search engine results. From the GLOBE Observer application, we used the data from their Mosquito Habitat Mapper protocol, which consisted of pictures of various mosquito larvae taken by NASA SEES interns during the months of June and July of 2021. Once we had the data, we then sifted through it to select the pictures deemed acceptable for training. This consisted of manual checks to ensure that key features of interest such as the larva’s siphon were visible without excessive debris, coverage, or noise. After removing all inadequate images, we had a training set of 69 images, consisting of 14 *Anopheles* photos and 55 non-*Anopheles*  photos. However, due to the size of datasets required to train deep learning software, we also utilized the search engines Google, Yahoo, Bing, Ecosia, and DuckDuckGo to help supplement our training set with additional images. After browsing the internet for *Anopheles* mosquito larvae, as well as some non-*Anopheles* larvae imagery, 38 new *Anopheles* photos and 17 non-*Anopheles* photos were pulled from search engines and added to the dataset, for a total 124 images.

(ii) Analysis:

Once the quality images were selected, we separated them into pictures showing an *Anopheles* larva and pictures showing any other genus of larva. The images were all form fitted to 720 by 480 pixels and made into a dataset used to train a convolutional neural network (CNN) named AlexNet. AlexNet has a deep architecture, and so it can be used to better describe and classify the images (Figure 2). In total, we had a dataset composed of 124 images, with 52 of them being of the *Anopheles* genus and the other 72 being of a variety of mosquito genuses, but the model was only trained with a 104-picture dataset (52 *Anopheles*, 52 non-*Anopheles*) in order to minimize overfitting.. Another important aspect of the training of the neural network was the model’s utilization of the ReLU and sigmoid functions. Mathematically, they can be defined as y =max(0,max) (relu) and S(x) = 1/(1+e-x) (sigmoid). These functions determine which values to pass as output and what not to pass. They influence the way the neural network interacts and learns, ultimately affecting the final outputs and predictions of the program.

1. **Results**



**Fig. 3.** The same classifier trained for 20 epochs (left) and 15 epochs (right) and tested on the same picture of an *Anopheles* larva, which was taken from the training set. A prediction of 0 indicates that the model classifies the image as an *Anopheles* larva, and a prediction of 1 indicates a larva that is not an *Anopheles*. The 15-epoch model correctly identifies the picture as an *Anopheles* larva with more certainty than the 20-epoch model.



**Fig. 4.** The training accuracy for models trained for 15 (left) and 20 (right) epochs

1. **Discussion**

As shown inFig. 3, the model that had only been trained for 15 epochs had a more accurate reading than the model trained for 20, since it gave the test image a higher probability of being of the *Anopheles* genus than the 20-epoch model. We infer that these differing results stem from overfitting in the models. This potential overfitting is also indicated in how the model’s training accuracy is much more stabilized by epoch 20, while in the model only trained for 15 epochs, the training accuracy has not yet completely stabilized and the graph has not fully flattened.

We identify lack of sufficient data as another potential source for error. Due to the limited availability of intern-provided imagery that was of acceptable quality and the similar absence of photos available through search engines, our model could not trained to reach higher levels of accuracy, and as demonstrated, increasing the number of epochs only increased the amount of overfitting, lowering the overall accuracy.

Another source of error is the source images themselves. While some images are taken by NASA professionals, some images were taken by ameteur scientists. Consequently, this creates an inconsistency in the quality of pictures and data, providing an additional “noise” in the images which we believe contributed to our ultimately inconsistent testing results.

Finally, we argue image noise as a cause of error. Lack of processing in the pictures, such as cropping to only show the critical details, was something that we believe may have caused errors in the model. Due to there simply being too much noise in the images, the classifier struggled with identifying key details.

1. **Conclusion**

Despite the lack of a large dataset, these results reinforce the idea that a deep learning model can be used for image processing, given enough data. Although we were unable to make our model as accurate as we had originally set out for, we found that the model, albeit limited, supports our hypothesis and seems to show possible applications for deep learning image classifiers in citizen science environments. The model was able to differentiate between *Anopheles* and Non-*Anopheles* photos with some level of accuracy, providing evidence for the model’s future potential as a publicly implemented larvae classifier.

We believe this research to be extremely applicable to citizen science fields. The use of a rapid and accessible *Anopheles* identifier could greatly aid in the tracking, monitoring, and prediction of malaria outbreaks. Additionally, this image classifier could be quickly adapted to identify other mosquito species or genuses, aiding in the overall identification of mosquitoes.

Future studies can improve several aspects of our project. Currently, the largest issue our study faces is simply a lack of quality and consistent data. Simply due to the nature of the data needed, it is difficult to find enough imagery to train the model to reach a significant accuracy and the data found comes from differing sources, creating additional noise in the pictures. Future studies can address these issues by creating their own datasets through their own collection methods. This would cut down on noise and ensure that enough imagery is present to use.

To conclude, we state that this study would not have been possible without the aid of our project mentors. They aided in data collection and interpretation when assembling our dataset and were critical in helping us find new directions with our project when faced with issues. Finally, they suggested various ways to present and interpret results, such as varying certain aspects of the models and also identifying possible discussion points in our project.

1. **Acknowledgements**

We would like to thank the Earth System Explorers/Mosquito Mappers mentors (Cassie Soeffing, Rusty Low, Peder Nelson) and peer mentors (Matteo Kimura, Vishnu Rajasekhar).

This project included; Spencer Burke, Juan Durante, Chris Grizzaffi, Amyn Macknojia

**Spencer Burke** Managed the type of deep learning model being used, helped refine the model to better suit the data, found additional data to improve model accuracy, sorted data, and researched the critical differences between each larva. Although Spencer is familiar with JavaScript, he had to learn how to code in Python and to use the various types and classifications of machine learning models. To do this, he had to extensively research types of machine learning models, read many research papers, and watch many videos.

**Juan Durante** Managed the implementation, adaptation, training, and testing of the AlexNet convolutional neural network and assisted in maintaining a good overall pace of work throughout the project. From this study, Juan has learned to use Python for the first time and has gained experience coding in interactive development environments like Virtual Studio Code and Google Colaboratory.

**Chris Grizzaffi** Researched the characteristics, structure, and anatomy of *Anopheles* larvae. As well as contributing to the first models of the code on Google Colabatory and understanding of the deep learning model like the algorithms associated with code, the different types of model, and structure of it. Finally, he contributed to the written aspect of the project and research paper. Chris learned how to use Python, write and structure a research paper, and use different coding techniques and environments.

**Amyn Macknojia** Contributed and organized the deep learning model being used in the machine learning algorithm we developed on Google Colaboratory, along with implementing the learnings and new knowledge from the literature as a gateway to find variables and preemptive hypotheses to further delve into in later stages of the project.

1. **References**

Burkett-Cadena, N. (n.d.). Morphology of Adult and Larval Mosquitoes. University of Florida.

Sanchez-Ortiz, A. & Fierro, Atoany & Arista-Jalife, A. & Cedillo-Hernandez, M. Nakano-Miyatake, M. & Robles-Camarillo, Daniel & Cuatepotzo-Jimenez, V.. (2017). Mosquito larva classification method based on convolutional neural networks. 1-6. 10.1109/CONIELECOMP.2017.7891835.

Sarker, I.H. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN COMPUT. SCI.* 2, 160 (2021). <https://doi.org/10.1007/s42979-021-00592-x>

Talapko, J., Škrlec, I., Alebić, T., Jukić, M., & Včev, A. (2019). Malaria: The Past and the Present. *Microorganisms*, *7*(6), 179. https://doi.org/10.3390/microorganisms7060179

World Health Organization. (2021, April 21). *Fact sheet about malaria*. World Health Organization. <https://www.who.int/news-room/fact-sheets/detail/malaria>.