Method for Effective Mosquito Data Classification to Identify Potential Hosts of Malaria with AI Implications

Student Researchers: AJ Caesar, Walker Gaines, Rahul Gajendran, Tejas Ram, Aaron Lee, Obumneme Nwosu, Angelina Richter

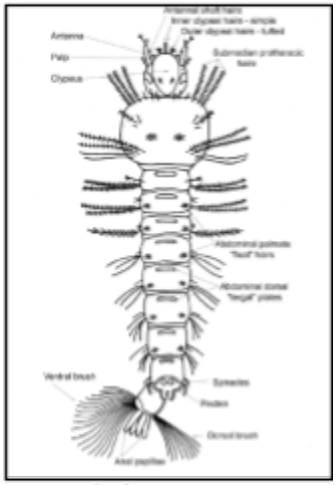
Mentors: Dr. Di Yang, Bill Lam, Kellen Meymarian, Matteo Kimura, Dr. Rusty Low, Cassie Soeffing

Table of Contents

Abstract

Most of Earth's mosquito-borne illnesses are transmitted by mosquitoes in one of three genera: *Anopheles, Aedes,* and *Culex.* Mosquitos of such genera are located in all continents but Antarctica and infect millions of humans with parasitic viruses yearly. However, a special concern is reserved for *Anopheles* mosquitoes for their unique ability to carry and transmit Malaria, a disease that, according to WHO, infects more than 200 million and kills over 500,000 humans annually (*Malaria*, 2022). While it is most prevalent in Africa, Southeast Asia, and Central America, Malaria could soon spread to northern and southern latitudes with a changing global climate. Therefore, it is crucial to track the extent of the *Anopheles* range and identify any changes that could have detrimental consequences on public health. One way this can be done is using the GLOBE Observer Mosquito Habitat Mapper (MHM) tool, which allows global users free access to photograph mosquito larvae, attempt to identify their genus, and upload said images to a worldwide database that records the location at which they were taken. While citizen science data is extremely helpful for mosquito research, it can be difficult for citizens with minimal training to classify the genus of their discovered larva correctly. A large portion of mosquito photos uploaded to the GLOBE MHM database is either unidentified or misidentified. Therefore, this research paper aims to devise and assess how the MHM database can be appropriately classified to create an accurate dataset with all *Anopheles* larvae photos classified by their proper genus. Besides being a vector of Malaria, another unique characteristic of the *Anopheles* mosquito is the absence of a siphon, so by scanning for this trait among MHM larvae photographs and noting positive matches, researchers

Image Reference #1



Anopheles Anatomy
(Tusting, 2014)

created a dataset of mosquito larvae that could become vectors of Malaria as adults (Image Reference #1). This data set could then be used to train AI models utilizing Convolutional Neural Networks (CNN) or Vision Transformers (ViT) to classify the MHM database autonomously in the near future.

Research Question

Question 1: How can scientists manually classify photos of *Anopheles* mosquito larva within the GLOBE MHM database to create a dataset with AI capabilities?

Devising such methods will allow us to create datasets of classified mosquitos which can be used as train data for future AI developers. It will also illuminate differences in the appearance of mosquito larva which can be taught to future citizen scientists to improve the accuracy of their field research.

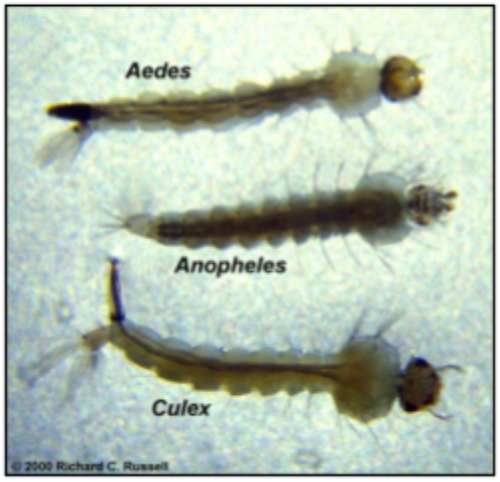
Question 2: What will our classification results show about the relative quantity of *Anopheles* mosquitoes in the Africa, Asia-Pacific, and Latin America regions within this data set?

This question will give insight into the changing concentrations of *Anopheles* mosquitoes across the globe which can help researchers model the future spread of Malaria. According to Climate.gov, Earth's average temperature has risen by about 1° Celsius since the pre-industrial era, and this global warming trend will continue (Lindsey & Dahlman, 2022). *Anopheles* mosquitoes have thrived in the hot, muggy environment around the equator for millions of years and may spread to northern and southern latitudes. Our research will establish a critical baseline to reference in the future and analyze any bias in the data set that could induce errors in an AI model that it trains.

Introduction & Review of Literature

The United Nations acknowledges the "complex relationship between Malaria and Climate Change" and admits that "gaps in knowledge still exist in the mechanisms of the linkage" (*Climate Change and Malaria - A Complex Relationship | United Nations*, n.d.). Therefore, it is safe to assume that new research methods will be necessary to assess this problem in the near future. By tracking its source, mosquitoes of the *Anopheles* genus, scientists can better understand how this vector-borne virus spreads across continents. To achieve a model of how and where mosquitoes inhabit new territories, thousands of mosquito larvae must be frequently analyzed to determine their genus, location and concentration, specifically *Anopheles* mosquitoes. Photographs of such larvae and their location can be found in the GLOBE MHM database. However, the majority are unclassified or misclassified and therefore difficult for scientists to use in a professional research setting. However, one helpful tool for identification of mosquito genus is that only larvae of the *Anopheles* genus present no siphon (Image Reference #2. The long-term goal of this project is to manually verify enough photos so that an AI can be trained to classify *Anopheles* larvae based on the absence of a siphon, as this will predict future vectors of Malaria. An AI can help scientists quickly analyze larvae and avoid manual labor.

Image Reference #2



Anopheles tail compared to Culex and Aedes (Terezinha, 2017)

Additionally, from this data set we aim to determine the relative concentration of photographed larva with siphons and those without in the GLOBE database. When used to train AI models for image classification, disproportionate datasets in which one category dominates in quantity can lead to bias in the final AI model (Shahinfar et al., 2020). This article uses the example of a camera that captures many photographs of birds but few of koalas. This creates bias in favor of birds when developers utilize the dataset to train

AI models that classify photographs between the two. In the same manner, an overwhelming quantity of larvae photos with or without siphons within this study could alter the accuracy of a future AI model. Therefore, it is important to note the relative frequency of the two categories contained in our final dataset - siphon, no siphon - to inform them properly. In total, the goal of this study to create a final dataset from GLOBE MHM data that contains 3000+ correctly classified (siphon, no siphon) mosquito larva photos with AI implications means that the dataset not only needs to be extremely accurate but also checked for disparities in quantities of each final classification.

Research Methods

The group of researchers was self-selected from the NASA GLOBE Earth System Explorers program, an eight-week summer internship for high schoolers across the globe.

From the publicly accessible MHM dataset researchers sourced 15,000 photos captured in Africa, Asia Pacific, and Latin America. With each photo researchers also obtained MHM ID, user-attempted classification (if applicable), Container Type, Userid, Longitude, Latitude, and Date. Then, the number of photos was divided up between researchers to examine and identify. Each researcher was also tasked with checking the previous classification work of another after the first round of classification was completed. All work was done using the platform Google Sheets.

The absence of a siphon is a unique trait of *Anopheles* mosquito larvae (*Identify Specimens*, n.d.) (Image Reference #2). Therefore, with the help of a classification training throughout the NASA GLOBE Earth System Explorers Internship, researchers were given the skills to identify Anopheles mosquito larvae within photos by identifying those without a siphon. To further ensure researchers in this project uniformly identified data, everyone referenced the first 100 images sorted. This gave a standard for all classifications so the data is proper and can be used elsewhere. Researchers reviewed photos and sorted them into three classifications: no siphon present, siphon present, and indeterminate. Indeterminate photos were not used, while no siphon present and siphon present photos were uploaded to a group document with their respective classification.

Once all classification was completed and reviewed, researchers inputted their final results into an additional Google Sheet where the quantities of photos presenting a siphon (non-*Anopheles)* and not showing a siphon (*Anopheles*) are totaled and the relative quantity within each region (Africa, Asia Pacific, Latin America) is determined using summation, fractional, and percentage models.

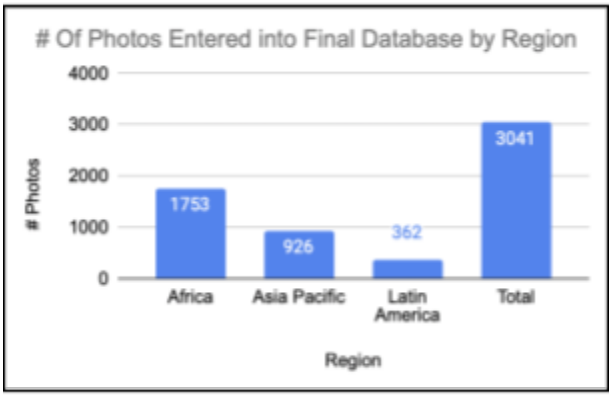
Results

After uploading a total of 3041 classified and rechecked mosquito larva photos to the final dataset (Image Reference #3) researchers found a number of significant trends within the dataset. Primarily, over 10,000 GLOBE MHM photos from these regions were discarded because they either did not show the tail section, presented ambiguous evidence of tail presence or absence, had poor image quality, or were duplicate images. The GLOBE MHM manual asks for two high-quality photos of the larvae tail section whenever a citizen scientist makes an observation. However, our quantitative observations show that this informal rule is either not followed or satisfied with poor image quality for a majority of uploads.

The relative quantity of No-Siphon vs. Siphon Present photos uploaded from each region is tightly correlated to the total quantity of photos uploaded from that region. In order from greatest to least, the number of total photographs, No-Siphon photographs, and Siphon Present photographs, went in order

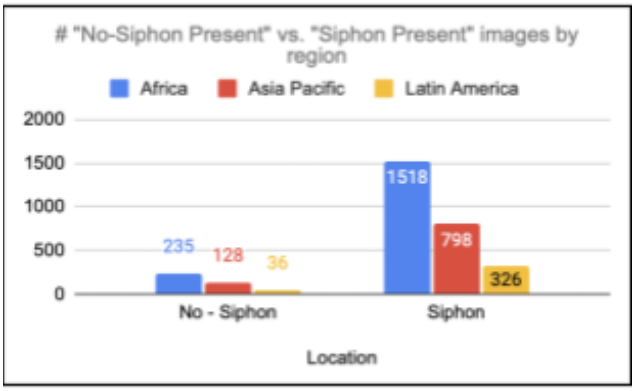
- Africa, Asia Pacific, Latin America - for all three categories (Image Reference #4). This strict correlation can be seen in the percentage model, as the percentage of "No-Siphon Present"

Image Reference #3



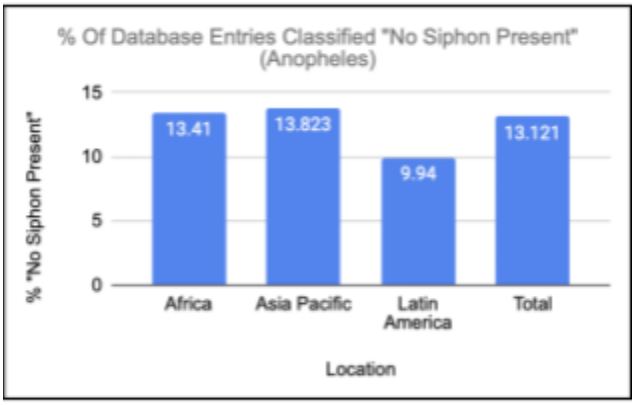
Total number of photos entered into final database by region

Image Reference #4



Siphon vs. No Siphon Breakdown by Region

Image Reference #5



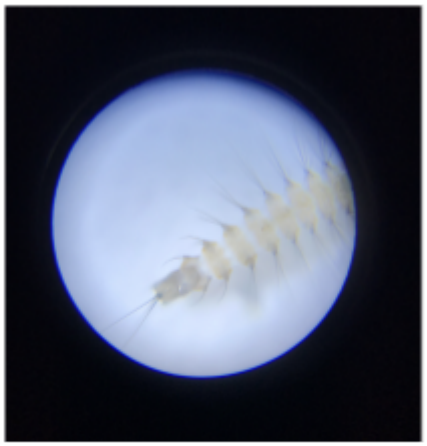
Percentage of No-Siphon (Anopheles) images within the final database by region

photographs in terms of total photographs uploaded by region was very similar across all regions. The Africa and Asia Pacific regions showed a slightly higher percentage of mosquito larvae presenting with no-siphon (*Anopheles)* than Latin America (Image Reference #5). However, this slight difference falls within the margin of error of classification by researchers. The overall breakdown of the final dataset (13.121% No-Siphon Present, 86.879% Siphon Present) (Image Reference #5) demonstrates a significant bias towards "Siphon Present" mosquito larvae (Aedes & Culex) in comparison to No-Siphon Present (*Anopheles*) within the final dataset that future AI developers must consider.

Discussion

In the first of the two research questions, qualitative observations must be made to determine with what ease researcher's could produce an accurate final dataset that classified thousands of GLOBE MHM photographs into the two categories of "No-Siphon Present" and "Siphon Present". The large majority of photos sourced from the original GLOBE MHM dataset that had to be discarded signals some difficulty within this endeavor. Researcher's found a number of concerns that led to this large quantity not being classified and uploaded to the final dataset. Only photos with clear signs of siphon or no siphon (Image References #6 & #7) could be included in the dataset. However, numerous other categories of photographs were included in the originally sourced GLOBE MHM data. Primarily, images that did not show the tail of the larvae (Image Reference #8) had to be excluded from the final dataset, in addition to those with sufficiently poor image quality (Image Reference #9) that the presence or absence of a siphon could not be determined despite the tail being photographed.

Image Reference #6



A clear "No-Siphon Present" mosquito larva tail
GLOBE MHM Database

Image Reference #7

A clear "Siphon Present" Mosquito Larvae
GLOBE MHM Database

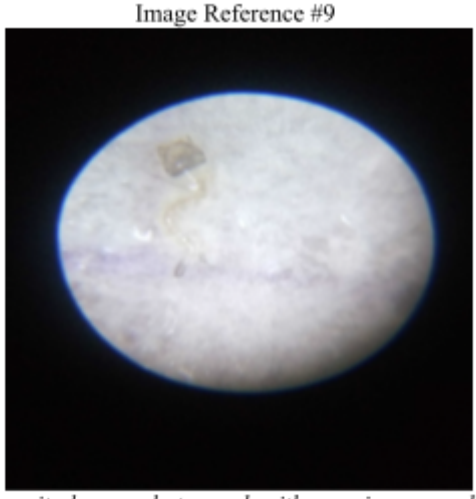
Image Reference #8

Mosquito larvae without tail section displayed
GLOBE MHM Database

Another point of contention in data classification regarded the presence of pupae images within GLOBE MHM data. Pupae is the phase of the mosquito life cycle following the larvae stage (Image Reference #11) and one in which the tail greatly resembles that of an *Anopheles* larva (Image Reference #10). Rather than breathing through a siphon, mosquito pupae of the three genera - Aedes, Culex, *Anopheles* - breathe through "horns" on the back end of their

head (*Mosquito Life Cycle | US EPA*, 2022). Therefore, the tails of pupae of all three major genera lack a siphon. Being that the goal of this study was to identify potential hosts of Malaria (*Anopheles* mosquitoes) using the absence of a siphon in larvae as a defining characteristic, this presented a great challenge in data classification. It was therefore within the discretion of the research team to check and determine if "No Siphon Present" images were in fact pupae rather than larvae of the *Anopheles* genus. Those that were determined to be pupae were eliminated, though some contentious entries were included in the final database under the "No Siphon Present" classification which may be a source of experimental error.

Image Reference #9



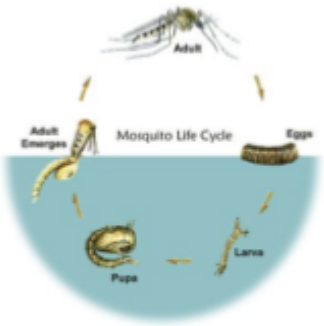
Mosquito larvae photograph with poor image quality
GLOBE MHM Database

Additionally, to answer the second research question, researchers would like to discuss the future capabilities of the final database (3041 Total Larvae Photos, 399 No Siphon, 2642 Yes Siphon) to be used in AI development. The large bias in the overall breakdown of the final database (13.121% No-Siphon Present, 86.879% Siphon Present) presents some concern. As mentioned, the closer to an even breakdown, 50% and 50%, that an AI train data set containing two categories of data can reach, the more accurate the final AI model will be at image recognition (Shahinfar et al., 2020). The percentage breakdown by region suggests that GLOBE MHM photographs from the Africa and Asia Pacific regions may be most effective

Image Reference #10

Mosquito Pupae
GLOBE MHM Database

Image Reference #11



Mosquito Life Cycle
(*Mosquito Life Cycle | US EPA*, 2022)

at creating an accurate database since the percentage breakdown is closest to even (as compared to Latin America), though they still presents large bias (Africa: 13.41% No-Siphon Present, 86.59% Siphon Present, Asia Pacific: 13.823% No Siphon Present, 86.177% Siphon Present). Therefore, in addition to sourcing their data from the Africa and Asia Pacific regions, an AI development team targeting the greatest precision possible should aim to selectively incorporate only certain photographs into their train data so that there is a relatively equal quantity of siphon presence and absence classifications.

Conclusion

In response to the first research question, it can be concluded that the MHM database can be classified into the three categories of "No Siphon Present", "Siphon Present" and "Unusable" in order to create a database that accurately predicts future hosts of Malaria. This is because only *Anopheles* possess the two unique traits of siphon absence and potential to host Malaria upon maturity. However, challenges involved in this process, including but not limited to human error and pupae identification, bring about the need for manual accuracy verification by at least one other trained classifier apart from the original.

Additionally, due the significant bias in the dataset towards "Siphon Present" Larvae photographs, the database as a complete entity cannot accurately train an AI model in image classification of mosquito larvae siphon presence or absence. However, by selecting specific photographs within the database and creating their data set that has a breakdown of classification much closer to even between the No Siphon Present and Siphon Present categories, AI developers can still effectively use this database for Image classification model development using Convolutional Neural Networks (CNNs) or Vision Transformers (ViTs).

From Image Reference #5, one might assume that Africa and Asia Pacific contain a higher percentage of Malaria hosting mosquitos than Latin America due to the higher percentage of "No Siphon Present" photographs in the final database. While this conclusion might be true in the real world, this conclusion cannot be drawn from our dataset as GLOBE MHM data is idealistic data. There are too many uncontrollable variables associated with the number of photos of each mosquito genera uploaded to the GLOBE MHM database by citizen scientists. Conclusions regarding the relative quantity of potential Malaria hosting mosquitos in each of these three regions requires further research to better understand what factors contribute to the relative quantity uploaded to the MHM database in order to understand how accurately it reflects the real world percentage breakdown.

Dataset

https://docs.google.com/spreadsheets/d/1NNwolG0aP6IJtp2DA2RUo5YVPtx4b4UN6e7Qc5H7z1I/edit?usp=sharing

Badges

I Am A Data Scientist

Utilizing GLOBE MHM data, which included photographed larvae, longitude, and latitude data researchers were able to analyze 3000+ entries and complete a final dataset which they hope will be utilized by future researchers in AI development.

I Am A STEM Professional

Researchers worked closely with STEM Professionals including Dr. Cassie Low to gain a better understanding of how we could identify potential vectors of Malaria by investigating the tail section of larvae and noting the presence or absence of a siphon. Researchers also utilized help from Matteo Kimura to obtain images and their location, so that we could use this data to classify and create their final dataset.

I Am A Collaborator

Researchers adequately divided work - classification, writing, video editing - between members to ensure that all contributed and had equal say in the paper's conclusions and delivery. Researchers also ensured effective communication using platforms including Discord and Messages to complete this project in a timely manner even while team members were in completely opposite time zones across the world.

Bibliography

Climate Change and Malaria - A Complex Relationship | United Nations. (n.d.). the United Nations. Retrieved July 21, 2022, from https://www.un.org/en/chronicle/article/climate-change-and-malaria-complex-relationship

Identify specimens. (n.d.). GLOBE Observer. Retrieved July 21, 2022, from https://observer.globe.gov/documents/19589576/8c9e4a31-ac81-48ff-82aa-8d07a86babc2

Lindsey, R., & Dahlman, L. (2022, June 28). *Climate Change: Global Temperature*. NOAA Climate.gov. Retrieved July 21, 2022, from https://www.climate.gov/news-features/understanding-climate/climate-change-global-temperature

Malaria. (2022, April 6). WHO | World Health Organization. Retrieved July 21, 2022, from https://www.who.int/news-room/fact-sheets/detail/malaria

Mosquito Life Cycle | US EPA. (2022, March 8). Environmental Protection Agency. Retrieved July 24, 2022, from https://www.epa.gov/mosquitocontrol/mosquito-life-cycle

Reiter, P. (n.d.). *Climate change and mosquito-borne disease. - PMC*. NCBI. Retrieved July 21, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1240549/

Shahinfar, S., Meek, P., & Falzon, G. (2020, May). "How many images do I need?" Understanding how sample size per class affects deep learning model performance metrics for balanced designs in autonomous wildlife monitoring. *Science Direct*, *57*. https://www.sciencedirect.com/science/article/abs/pii/S1574954120300352

Terezinha, S. (2017). *How to identify Culex, Anopheles and Aedes mosquitoes and their larvae?* Researchgate. Retrieved July 23, 2022, from https://www.researchgate.net/post/How_to_identify_Culex_Anopheles_and_Aedes_mosquitoes_and_their_larvae/58a41eedf7b67e73463c4a89

Tusting, L. (2014, 2 1). Larval Source Management: A supplementary Measure for Malaria Control. *Outlooks on Pest Management*, *25*.

https://www.researchgate.net/publication/263420029_Larval_Source_Management_A_Supplementary_Measure_for_Malaria_Control

Acknowledgements

SEES Earth System Explorer mentors; Dr. Rusanne Low, Ms. Cassie Soeffing, Mr. Peder Nelson, Dr. Erika Podest, Andrew Clark, Matteo Kimura, Kellen Meymarian.

The material contained in this poster is based upon work supported by the National Aeronautics and Space Administration (NASA) cooperative agreements NNX16AE28A to the Institute for Global Environmental Strategies (IGES) for the NASA Earth Science Education Collaborative (NESEC) and NNX16AB89A to the University of Texas Austin for the STEM Enhancement in Earth Science (SEES). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NASA.