Identifying *Anopheles*/Non-*Anopheles* Larvae with AI Implications

Vishnu Rajasekhar, Jocelyn Browning, Ashley Hume, Kellen Meymarian, Robyn Ogombe, and Hannah Slocum

NASA SEES Mosquito Mapper Virtual Internship Science Fair

Table of Contents

Abstract

Mosquitoes are vector organisms that spread diseases to thousands of people worldwide, contributing to the considerable number of deaths due to vector-borne illness. This study serves to identify larvae of the malaria-spreading *Anopheles* genus of mosquitoes in North America while fueling methods to refine the NASA GLOBE Observer data set and promoting Citizen Science.

Citizen Science data are considered highly inaccurate in the scientific community, with the reasoning that almost anyone, trained expert or not, can contribute to these data sets with varying levels of precision. To improve the accuracy and credibility of such Citizen Science data

sets--in this case, the GLOBE Observer database--a sample of 155 unique mosquito observations were pulled from the database. Using this sample, trained classifiers and mosquito experts reclassified each reported observation to check Citizen Scientist's accuracy when identifying *Anopheles* larvae. As a result of this reclassification the majority of Citizen Science data, in this aspect, is proven to be accurate, but perhaps not at the desired level of accuracy. For this reason, this refined sample of data can be used as a baseline for an AI recognition system to automatically classify images taken by Citizen Scientists as either *Anopheles* or non-*Anopheles*, in further development of this study.

Purpose + Literature Review

The purpose of this study was to manually identify *Anopheles* larvae vs. other species of mosquito larva. This data set was collected through NASA's GLOBE Observer application. The manual verification data will be used by an artificial intelligence (AI) team to create an AI algorithm for computerized larvae identification. The data submitted through the GLOBE Observer was collected by Citizen Scientists, and a sample was reviewed in this study to refine submission data before use in artificial intelligence contexts.

This project relies on Citizen Scientists, so it's important to understand what they do. Oxford languages define Citizen Science as "the collection and analysis of data relating to the natural world by members of the general public, typically as part of a collaborative project with professional scientists." Citizen Science is crucial to the scientific research industry, substantially lowering fieldwork costs and providing more varied data to researchers.

The goal of this data analysis study was to analyze multiple citizen science submissions from the GLOBE Observer app and verify *Anopheles* and Non-*Anopheles* mosquito larvae observations. Citizen Science is a form of research that is being used more and more frequently

in the modern scientific community. It allows researchers to gain information from different people in areas all across the globe. Through the use of websites and even mobile apps, the common citizen can now participate in essential research with scientific professionals. This lowers the cost of fieldwork for researchers, while also educating the general public through hands-on participation in research (Murindahabi et al., 2018).

Despite the many benefits, some claim that Citizen Science is not a viable method of gathering data because it uses data from both trained and untrained individuals. Citizen Science is also typically performed remotely, so it can be difficult to train citizens well enough that they can perform research on the same level as a highly trained scientist. This data quality project went through several different mosquito photographs and flagged any entries that were incorrectly labeled as "*Anopheles*" or "Non-*Anopheles*" mosquitoes. This data verification research is essential to creating an artificial intelligence algorithm that will analyze different entries submitted by future Citizen Scientists and correct any obvious errors.

The protocol used for the manual identification of *Anopheles* larvae is as follows. The two most notable differences that were used in the identification of *Anopheles* larvae are the lack of a breathing siphon on the larvae's tail end and the presence of palmate hairs on the larva's abdominal sections (Classification and Identification…, n.d.). Since *Anopheles* larvae exclusively breathe through openings on their back at the surface of the water, they lack a breathing siphon and lay parallel to the surface (Rios & Roxanne, 2018). *Anopheles* larvae also tend to have a darkly colored, elongated head, which greatly helps with identification (Burkett-Cadena, n.d.). This predetermined protocol for larvae identification keeps each classifier's classifications constant throughout the entire sheet.

Methods

To classify each specimen as accurately as possible, each classifier underwent ten weeks of training, learning about different types of mosquitoes, their larvae, and their habitats. These weeks of training give each classifier an advantage in being able to identify if a mosquito larva is an *Anopheles* or not in comparison to the common Citizen Scientist. The classifications analyzing the sample from the NASA GLOBE Observer data set will allow an AI to properly analyze future GLOBE Observer app entries with greater accuracy, due to the minimization of human error.

The identification system's Google Sheets spreadsheet workbook starts with a reminder of the thorough protocol, with supporting sample *Anopheles* images (Image Reference, Figure 1) and diagrams (Image Reference, Figure 2), explaining the process of identifying each specimen. This protocol primarily consists of written descriptors--mentioned in the above section--to aid the classifiers in identifying each specimen. Each classifier has a Sheet, private to themself, to make unbiased classifications of each data entry.

Each classifier's Sheets are identical to start before the classifier begins their analysis of each specimen. Every data entry includes a unique Mosquito Habitat Mapper ID, the ID of the GLOBE application user who made the observation, the application user's classification of the specimen, an exact location tag, and the date of the observation. Using this information from the GLOBE application's record, the classifier must tag the entry with an identifier, indicating whether the specimen is an *Anopheles* or not, or another extraneous classification.

The tagging system consists of four options to choose from for each specimen. The tag options are "y", "n", "ind", "non", and answer the question "Is this larval specimen an *Anopheles*?". The "y" tag indicates that the specimen in the photo has been determined by the

classifier to be an *Anopheles* larva. The "n" tag indicates that the specimen in the photo has been determined by the classifiers to **not** be an *Anopheles* larva. The "ind" tag means that the specimen in the photo was indistinguishable by the classifier for a multitude of reasons. This ranges from the photo being too blurry, too dark, too small, too low of a resolution, etc. If this tag is used, the classifier places a comment noting their reason for its use. The "non" tag indicates that the specimen in the photo has been determined by the classifier to not be a mosquito larva. Instead, it could be plant debris, sand, or another organism altogether. If it is clear to the classifier that the image does not show a mosquito larva, this tag would be used, with a comment explaining the reason for its use. Additionally, if the classifier was not certain in their classification, a comment was left with their confidence level in their classification (ie. If a classifier placed a "y" tag on a photo of a specimen, but they are not completely sure, they would put a "confidence - *x*/5 " in the comment section for that specimen, noting that they were confident to *x* level in their classification.

Using this classification system, a master sheet was created to compile the results of all the classifiers' manual classifications. As the classifier team contained both trained individuals and mosquito experts, the trained individual's classification counted as one vote for a classification tag, while the expert's classification counted as two votes. This adds credibility to the study as it gives expert classifiers more weight in the final decision, and their classifications are bound to be more accurate.

Results

Following the classification methodology discussed in the previous section, the results of the accuracy of Citizen Scientist classifications were identified. The data set studied in this project contained a total of 155 entries obtained from the GLOBE Mosquito Habitat Mapper

database by Citizen Scientists throughout North America. The classifiers verifying the classifications of the mosquito data entries were well trained on the distinguishing characteristics of *Anopheles* mosquitoes versus Non-*Anopheles* mosquitoes.

The classifiers concluded that of the 155 entries, 12.9% were indistinguishable and 4.5% were not mosquito larva. As previously mentioned, a classification of indistinguishable was assigned to entries that required more data to draw a conclusion, were too blurry, or were low resolution (Image Reference, Figure 3). Some entries that were deemed "indistinguishable" were identified as mosquito larvae, but classifiers could not identify whether or not the mosquito larva was *Anopheles* or not, due to the lack of visibility of identifying features. Additionally, the 4.5% of entries that were identified as "not mosquito" contained images that did not resemble any mosquito genus. These could include plant debris or other organisms (Image Reference, Figure 4).

Furthermore, 71.6% of the citizen scientist's Anopheles vs. Non-Anopheles mosquito classifications were correct. Correctly identified entries were those where the Citizen Scientist's classification corresponded with the classification made by the classifiers. On the contrary, there were discrepancies between expert researcher classification and citizen scientist classification on 11% of the 155 mosquito data entries. Furthermore, when considering entries that were definitely mosquitos (excluding indistinguishable and non-mosquito entries), 13.3% of the entries were incorrectly identified by Citizen Scientists (Image Reference, Figure 5).

These data points, particularly the 13.3% of identification discrepancies, support the need for this study. By fueling the back-end of research needed to create an algorithm that can verify or correct Citizen Scientist mosquito classifications, incorrect classifications can be limited to much fewer than they currently are. Using an algorithm will maintain the positive aspect of

Citizen Science that yields increased and more accessible data while removing the negative aspect of incorrect classifications and will minimize the human interaction needed to manually verify classifications.

Conclusion & Further Discussion

This study's goal was to cross-validate classifications of the type of mosquitos from photos sent in through the Globe Observer application by citizen scientists, with a focus on the *Anopheles* genus due to its malaria-spreading tendencies. This ensures that the mosquitos were correctly identified so the data would be valid and credible for future research.

On average, the trained classifiers and experts were 97.81% sure in their classifications--a quantitative score reached by averaging all the final classification's confidence intervals. The classifiers were also, on average, over 80% accurate in their individual classifications (Image Reference, Figure 6). To allow GLOBE Observer application users to report observations with the same level of confidence and accuracy as the classifiers, there are multiple training protocols and GLOBE application changes that could be used. Such a training protocol could include a training course built into the GLOBE application that better trains individuals to identify different larvae types. An application improvement that could improve the accuracy of data is user verification or the concept of a "superuser." This would give users who are known to provide highly accurate observations the opportunity to be verified manually, making their data more reputable for use in scientific research contexts.

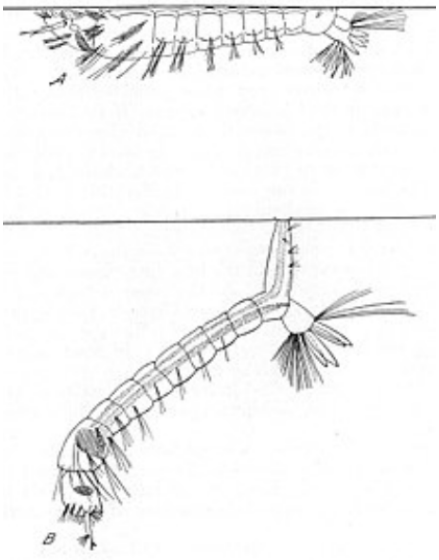
Due to this manual identification done by trained classifiers and mosquito experts, the data and conclusions from this study are being used to create an artificial intelligence-based solution to the Citizen Scientist classification error problem. Such a solution would eliminate the

majority of the human error in this process, refining the quality of Citizen Science data and giving way to a more technologically advanced and scientifically accurate future.

Image Reference

Figure 1

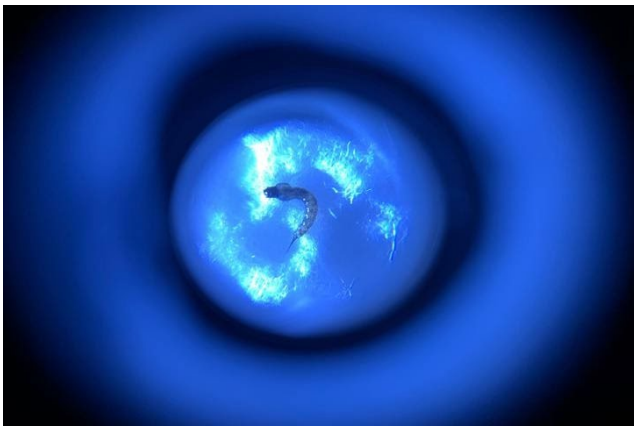
Anopheles (top) vs Non-*Anopheles* (bottom) Larvae Diagram



This diagram shows the difference in feeding position between the two mosquito classifications--parallel to the surface in *Anopheles* vs. “hanging” in Non-*Anopheles*.

Figure 2

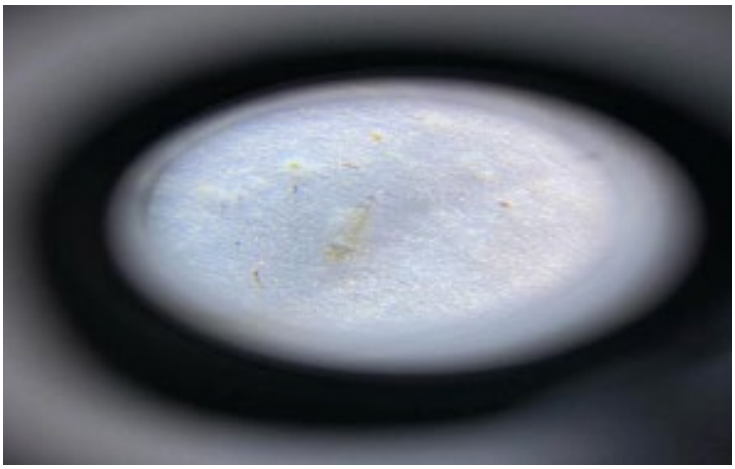
Anopheles Larva Example Image



This image clearly shows an *Anopheles* larva--a good baseline for classifiers to make their observations.

Figure 3

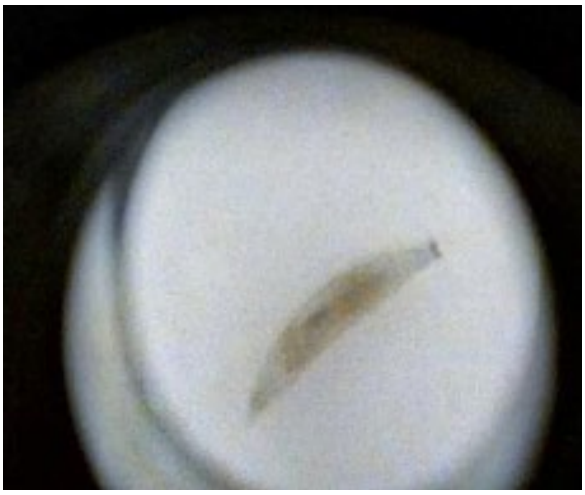
Indistinguishable Specimen



This image is an example of an entry that was classified as “indistinguishable” due to the low resolution and small image size. Classifiers could not see the distinguishing characteristics needed to accurately classify the specimen as either *Anopheles* or Non-*Anopheles*.

Figure 4

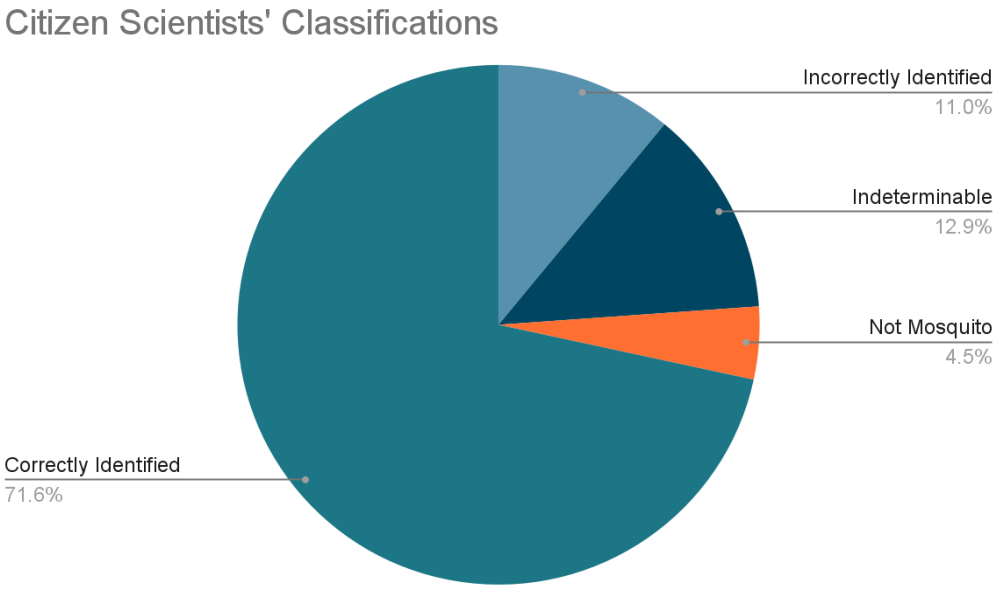
Non-mosquito Specimen



This image is an example of an entry that was classified as “non-mosquito.” While the contents of this image could be organic debris or another organism, it is certainly not a mosquito larva.

Figure 5

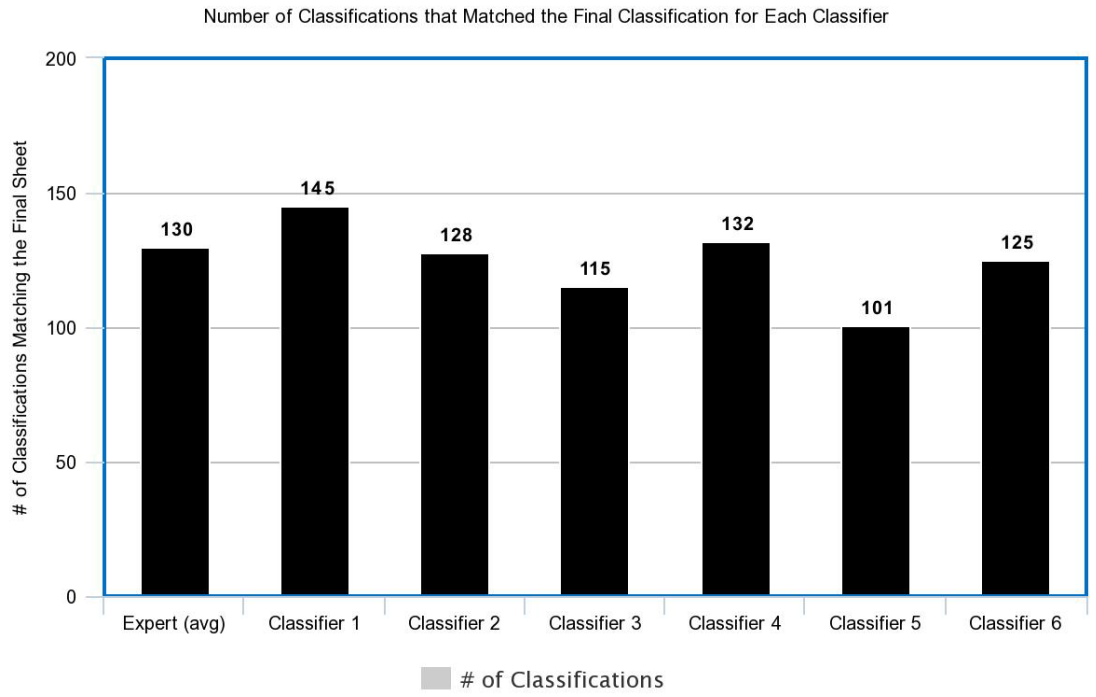
Citizen Scientists’ Classification Accuracy



This chart shows the accuracy of Citizen Scientists’ classifications of the overall sample data set. The majority of entries were classified correctly, but a large proportion of entries were also incorrectly identified or did not meet data validation standards.

Figure 6

Comparison of Classifiers' Classification Accuracy



This graph compares the number of classifications from each classifier that matched the final classification made for each specimen. This demonstrates the level of accuracy of each classifier, which on average has over an 80% individual accuracy rate.

Reference List

Murindahabi, M. M., Asingizwe, D., Poortvliet, M., van Vliet, A. J.H., Hakizimana, E., Mutesa, L., Takken, W., & Koenraadt, C. J.M. (2018, August 18). *A citizen science approach for malaria mosquito surveillance and control in Rwanda*. ScienceDirect. Retrieved July 26, 2021, from https://www.sciencedirect.com/science/article/pii/S1573521418301507

Rios, L., & Roxanne, C. (2018). Common malaria mosquito - anopheles quadrimaculatus say. https://entnemdept.ufl.edu/creatures/aquatic/Anopheles_quadrimaculatus.htm.

New Mexico Department of Health. (n.d.). *CLASSIFICATION AND IDENTIFICATION OF MOSQUITOES OF NEW MEXICO*. New Mexico Department of Health. https://www.nmhealth.org/publication/view/guide/986/.

Burkett-Cadena, N. (n.d.). *Morphology of Adult and Larval Mosquitoes*. University of Florida: Florida Medical Entomological Laboratory. https://fmel.ifas.ufl.edu/media/fmelifasufledu/workshop/Mosquito_Morphology.pdf.

King, W. V., Bradley, G. H., & McNeel, T. E. (1939, June). *USDA Miscellaneous Publication No. 336: The mosquitoes of the southeastern States*. USDA National Agriculture Library. https://naldc-legacy.nal.usda.gov/naldc/download.xhtml?id=CAT10307175&content=PDF.

Low, R., Boger, B., Nelson, P. and Kimura, M. (2021). GLOBE Observer Mosquito Habitat Mapper Citizen Science Data 2017-2020, v1.0. https://observer.globe.gov/get-data/mosquito-habitat-data

Acknowledgments

Thank you to the Earth System Explorers/Mosquito Mappers mentors, Dr. Becky Boger, Dr. Rusty Low, Peder Nelson, Dr. Erika Podest, Dr. Cassie Soeffing, and peer mentors, Vishnu Rajasekhar, Pratham Babaria, Faguni Gupta, Kavita Kar, and Matteo Kimura.

This project included Dr. Rusty Low, Dr. Becky Boger, Peer Mentor Vishnu Rajasekhar, Jocelyn Browning, Ashley Hume, Kellen Meymarian, Robyn Ogombe, and Hannah Slocum.

Dr. Rusty Low and *Dr. Becky Boger* were the expert validators for this study. As the mosquito experts, they also took part in classifying the sample of 155 observations pulled from the NASA GLOBE Observer database. Their work ensures that an expert opinion is present in each observation, increasing the credibility of this study.

Vishnu Rajasekhar wrote the abstract and set up the Google Sheets for classifiers to make classifications. Vishnu is credited with keeping the team together and on task, leading the intern group to complete a successful research project. Vishnu learned more about the benefits of Citizen Science, the dangers associated with *Anopheles* mosquitoes, data validation techniques, and project management.

Jocelyn Browning, *Ashley Hume*, *Kellen Meymarian*, *Robyn Ogombe*, and *Hannah Slocum* were all trained classifiers in this study. They analyzed a sample of 155 data entries of mosquito larvae from the NASA GLOBE Observer database and classified them into four different categories. They learned more about Citizen Science, data validation techniques, and the dangers associated with *Anopheles* mosquitoes.

Jocelyn Browning and *Kellen Meymarian* are credited with writing the Purpose & Literature Review section. *Kellen Meymarian* is also credited with editing and putting together the video presentation of our project. *Robyn Ogombe* is credited with writing the Methods section. *Ashley Hume* is credited with writing the Results section. *Hannah Slocum* is credited with writing the Conclusion & Further Discussion section.

Supplementary Links

Anopheles/Non-*Anopheles* Classification - Google Sheet - NA Anopheles MHM Images

Anopheles/Non-*Anopheles* Final Presentation - Google Slides - NA Anopheles Presentation

SUBMISSION FOR INTERNATIONAL VIRTUAL SCIENCE SYMPOSIUM

Identifying Anopheles/ Non-Anopheles Larvae with AI Implications

By Kellen Meymarian, Vishnu Rajasekhar, Jocelyn Browning, Ashley Hume, Robyn Ogombe, and Hannah Slocum

I AM A COLLABORATOR

During the summer 2021 intern program with the NASA and the University of Texas's STEM Enhancement in Earth and Space Sciences (SEES) program, our team conducted geo mapping, research and data analysis on disease vectors through field investigations. Despite the challenges of working 100% remotely with complete strangers who were scattered all over the country, the group successfully managed the team's efforts to create a video presentation for the 2021 NASA SEES Science Symposium detailing the team's accomplishments in building an artificial intelligence program to identify disease vectors through photographical analysis. The group video presentation can be found at https://youtu.be/5gKkjhP97Gw. Each member of the group was assigned a specific part of the project, and it subsequently resulted in a published paper and powerpoint in the Earth and Space Science Open Archive at https://www.essoar.org/doi/10.1002/essoar.10508741.1. The team members were responsibility for the following duties:

- Vishnu Rajasekhar: Peer Mentor, Data Analysis, and Lead Author on the report, and introduced our research project;
- Kellen Meymarian: Video editor, Supporting Author, Data Analysist, submitted paper for publication with ESSOAr, and presented why our group chose to study Anopheles mosquito larvae;
- Jocelyn Browning: Supporting Author, Data Analysist, and presented the impact of our research;
- Ashley Hume: Supporting Author, Data Analysist, and presented the results of our research;
- Robyn Ogombe: Supporting Author, Data Analysist, and presented where we collected our data from and how we classified each specimen; and
- Hannah Slocum: Supporting Author, Data Analysist, and presented the conclusions and overall findings of our research.

The team's patience and persistence is what allowed the team to finish the project on time and produce a quality project.

I AM A DATA SCIENTIST

The report examined the validity of data identification related to larvae of the malaria-spreading Anopheles genus of mosquitoes in North America. Additionally, the team examined if it was possible to refine or "clean" the NASA GLOBE Observer data set which is produced by collecting information from untrained citizen scientists. The GLOBE Observer app has facilitated mosquito research, allowing Citizen Scientists to report mosquito breeding grounds and the presence of larvae. In conjunction with other data, such as landcover or water, mosquito activity can be tracked and their effects can be mitigated. Citizen Science is often considered highly inaccurate, with the reasoning that almost anyone, trained expert or not, can contribute to data collection with varying levels of precision. To improve the accuracy and credibility of such Citizen Science data sets--in this case, the

GLOBE Observer database--a sample of 155 unique mosquito observations were pulled from the database. Using this sample, trained classifiers and mosquito experts reclassified each reported observation to gauge Citizen Scientists' accuracy in identifying Anopheles larvae. Using this reclassified data set, a convolutional neural network (CNN) was created as a machine learning (ML) solution to automatically identify a given larva photo as Anopheles or non-Anopheles. This model contains roughly 20% of the larval images in the GLOBE database, which were deemed usable for the training model.

The report concludes that

- Cross validating larvae identification (*Anopheles*/Not *Anopheles)*
- AI based identification eliminates human error in mosquito identification
- Easier to identify *Anopheles* mosquitoes due to reduction in manual work
- Tracking *Anopheles* to prevent malaria spreading
- Making citizen science more reliable and credible:
    - Introduce training section within the GLOBE Observer app to allow users to improve their knowledge on the types of observations they will be making and how to make them as accurately as possible
    - Introduce the concept of user verification or "super-users" who will have a special tag within the app and database to note their status; these users are verified manually and are recognized for having high accuracy rates due to completion of training and through experience by making more observations