MASC AI: A Novel Method for Effective Mosquito Data Classification and Mapping

Student Researchers: Nikita Agrawal, Samhitha Duggirala, Elizabeth Gorman, William Hong, Joseph (Alex) Kim, Nithin Reddy, and Aesha Shah

Mentors: Dr. Di Yang, Bill Lam, Kellen Meymarian, and Matteo Kimura

***Abstract* -** Anopheles mosquitoes are densely populated in Africa but can be found in other regions and continents, including but not limited to North America, South America, Europe, and Asia (specifically the Indian subcontinent). A point of concern with Anopheles mosquitoes is their unique ability to carry and transmit malaria, a parasitic infection that attacks red blood cells. WHO studies report that this deadly disease infects more than 200 million and kills over 500,000 people annually. The severity of this disease illustrates the need to engineer a practical method to mitigate the spread. The GLOBE Mosquito Habitat Mappers (MHM) app allows users worldwide to photograph and identify mosquito larva they encounter or trap. The collected data is available in a global database, accessible to researchers for labeling and classifying, a key step in tracking the population growth of the Anopheles mosquitoes. However, the manual verification for the classification of this data is time-consuming and inefficient, limiting the expansion of mosquito research. The use of Machine Learning (ML), a subset of Artificial Intelligence (AI), has substantial benefits for practical implementation, serving to improve image classification of mosquito larva of the Anopheles genus. This project proposes a method of implementing logistic regression for developing a mosquito identification model, with an accuracy greater than 80%, and plotting detected locations on a global map. A total of 3,275 images were extracted from the GLOBE MHM application. Each image is classified based on the absence of a siphon, a distinctive feature of the Anopheles mosquito, allowing accurate identification of the genus. The publicly available machine learning model and precise mapping of detected mosquito locations on a comprehensive world map will help mosquito ecologists, governments, and public health organizations effectively track and mitigate the spread of mosquito-borne diseases.

Index Terms – Artificial Intelligence, Data Classification, Logistic Regression, Machine Learning.

1 INTRODUCTION

Mosquitoes are the deadliest animal to humans, causing over 700,000 deaths annually worldwide. A specific genus of mosquitoes, *Anopheles*, accounts for around 500,000 of said deaths due to their transmission of the disease malaria. This parasitic disease enters the bloodstream through the liver and targets red blood cells (CDC, 2021). Malaria's mortality rate is extremely high (especially for children under 5) with some complications at ~15-20% with treatment. Nearly 100% without, presenting a significant problem for countries that lack adequate medical resources for the disease (Dvorin, 2018, para. 3). According to the CDC, 95% of malaria deaths were recorded in African countries, due to a combination of their immense *Anopheles* mosquito population and an insufficient health-care system (CDC, 2021, para. 4).

Due to the lack of medical resources to combat the virus worldwide, scientists are documenting the population growth of *Anopheles* mosquitoes to moderate and control the spread of malaria. However, mosquitoes are spread worldwide, causing experts to struggle with gathering data across the world. The immense magnitude of the problem presents the need for citizen scientists to collect an adequately representative database. According to National Geographic, citizen science is "the practice of public participation and collaboration in scientific research to increase scientific knowledge" (Ullrich et al., 2022). The GLOBE Mosquito Habitat Mappers (MHM) tool in the GLOBE Observer app allows anyone worldwide with a smartphone to photograph and document mosquito larvae/habitats and add them to the global database. The tool lets users label and classify mosquito larvae while recording the precise location (up to ~1-2 feet accuracy) of the documentation, giving insight into mosquito population demographics around the globe (Leidner et al., 2022).

Despite the convenience of the GLOBE Observer's database, many of the classifications of larvae are either inaccurate or incomplete due to a lack of accountability and training of the remote citizen scientists. The main distinction between different genera of mosquito is *Anopheles* larvae's absence of a siphon (seen as a dark tube on larvae's rear allowing *non-Anopheles* larvae to breathe on the surface of water) (Skiff & Yee, 2014, para. 1). To an untrained eye, this distinction is often difficult to spot, which leads to the inconsistencies in the classification. Manually reclassifying thousands of photographs from the GLOBE MHM database to account for these inconsistencies is exceedingly tedious and time-consuming for scientists.

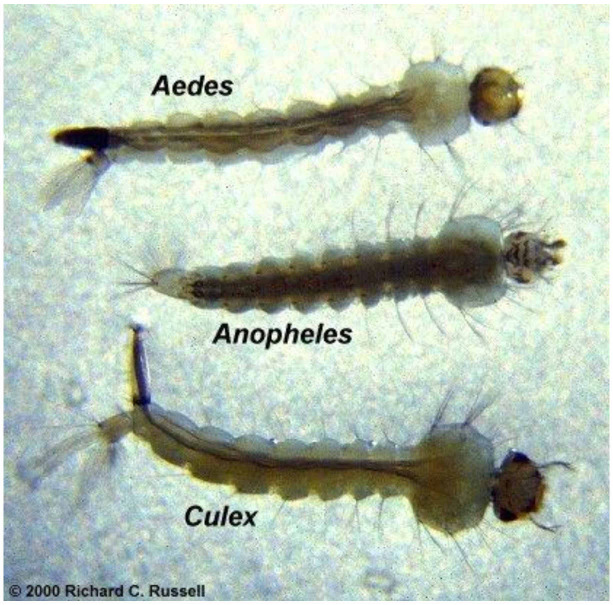

Figure 1: *Differences between mosquito larvae.* Credit: Richard C. Russell (2000), retrieved from *How to identify Culex, Anopheles and Aedes mosquitoes from their larvae.* Researchgate.net (2017)

MASC AI (Mosquito lArvae Siphon Classification Artificial Intelligence) presents an efficient solution to this problem by preventing the need for manual reclassification. MASC utilizes supervised machine learning (a subset of AI) to automatically classify images of mosquito larvae into categories of *Anopheles* and *Non-Anopheles*. Our team used a training set of 3,275 images, classified by another NASA SEES intern team. The training set would indicate whether the mosquito had a siphon, the species of mosquito, and the location where the image was taken.

From this training set, our team used Logistic Regression to develop MASC in order to have high accuracy while also efficiently identifying the mosquito larvae. Logistic Regression is a supervised ML algorithm used for binary classification, making it appropriate in our case since we only had two possible categories of *Anopheles* and *Non-Anopheles.* It uses a sigmoid function to estimate the probability of a data point being one of the two categories based upon the determined decision boundary (Boateng & Abaye, 2018). The exact specifics of our logistic regression process are gone into more depth in the methods section.

The main goal of this study was to utilize Logistic Regression to create an accurate (80%+) AI model for classifying mosquito larvae images.

2 RESEARCH QUESTIONS

To effectively prepare for our study and construct MASC, we chose to ask and research questions that would both directly affect portions of our study and have the potential for our MASC to affect in the future. The term "Research Question" is abbreviated as "RQ" and followed by a number that indicates the order in which these questions are posed in our research study.

RQ1: *What is the general understanding of Anopheles population demographics worldwide, and how does it compare to the GLOBE MHM app's database?*

According to the *Malaria ATLAS Project* conducted by the University of Oxford and WHO, *Anopheles* mosquitoes are found most often in tropical/subtropical areas and densely populated in East African countries such as Ethiopia, Sudan, and Kenya (Malaria *ATLAS* Staff, 2022). This was consistent with the GLOBE MHM data, with around 80% of the mosquito larvae initially classified as *Anopheles* taken from the Africa database.

RQ2: *What percentage of the GLOBE MHM database species classification was inaccurate before correction from the data verification team?*

After communicating with the data verification team, they found that over 90% of the data was either inaccurate or not even classified in the first place. The data they covered are representative of the entire GLOBE MHM database (over 3,000 images from 5 different continents), which shows how severely inaccurate the current classification system is.

RQ3: *To what accuracy can the MASC AI classify images of mosquitoes between three genera that are the leading cause of vector-borne diseases by mosquitoes?*

Upon completing the development of the Logistic Regression Model, it returned a train accuracy of 88.24% and a test accuracy of 86.11%, which is higher than originally expected. Compared to other alternatives for *Anopheles* verification, such as through molecular protein profiling, we find that the accuracy metrics from our AI model continue to outperform, indicating that MASC AI serves as a plausible solution for reclassification (Müller et al., 2013).

3 METHODS

3.1. Data Pre-Processing

The data of mosquito images was obtained from the GLOBE Observer App via the Mosquito Habitat Mapper selection, which was then compiled into a dataset sorted by region. This dataset contained whether or not the mosquito had a siphon, the mosquito classification type (Aedes, Anopheles, Culex), the geographical coordinates, and the image in URL format. The images were retrieved from the URLs using the Python package, skimage.io. These images were scaled to 64 x 64 pixels using the Python package, cv2, to remove potential bias and ensure that the machine learning model received all inputs of the same size. Finally, the images were converted into an array with the numerical RGB values using the Python package, NumPy, and the array values were divided by 255 to standardize the dataset. A separate array was created with the value "1" stored as the mosquito having a siphon and "0" stored as the mosquito not having a siphon.

3.2 Logistic Regression Model

This project implements a logistic regression model to binarily classify whether or not the mosquito presented in the image had a siphon. In order to develop this model, a Python notebook is used to write a script that defines the variables and functions required. The sigmoid function, $\sigma(z) = \frac{1}{1+e^{-z}}$, is used as the activation function for this model because of its ability to map a real value to another value between 0 and 1 as shown in Figure 2.

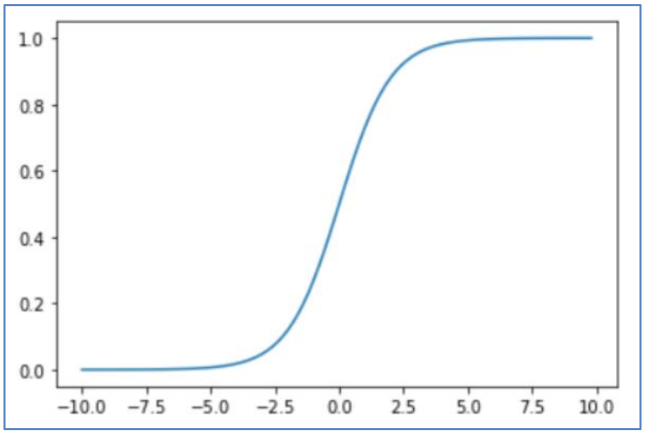

Figure 2: *Sigmoid Function.*
Credit: Nikita Agrawal

Forward propagation is used to determine the cost and the output. The cost function is given by

$$J = -\frac{1}{m}\sum_{i=1}^{m}\left(y^{(i)}\log(\mathrm{a}^{(i)}) + \left(1 - y^{(i)}\right)\log(1 - \mathrm{a}^{(i)})\right)$$

where m represents the number of features, y

represents the actual output, and a represents the predicted output. This equation is used to compute the cost as opposed to other common metrics such as Mean Squared Error because logistic regression is a nonlinear function. Back propagation is used to determine the gradient of loss by calculating the derivatives of the image parameter values, w. The derivatives of the parameters are given by the equation $\frac{\mathrm{d}J}{\mathrm{d}w} = \frac{1}{m}X(A - Y)$, where m represents the number of features, X represents the input, Yrepresents the actual output, and A represents the predicted output. When the derivatives stabilize, the weights of the parameters would be stored.

The next step is to optimize the model by minimizing the cost function and "learning" the parameter weights through forward and back propagation. This optimization function will run through 2000 iterations and the parameters will be updated using gradient descent. The equation used to update the parameters for a parameter θ is $\theta = \theta - \alpha d\theta$, where α represents the learning rate of the gradient descent update.

The final step is to use the parameter weights to predict the labels for the dataset using the sigmoid function. Since the sigmoid function outputs a number between 0 and 1 which corresponds to the probability of the mosquito having a siphon, all outputs less than 0.5 will be classified as images of mosquitoes that do not have a siphon and all outputs greater than or equal to 0.5 will be classified as images of mosquitoes that have a siphon.

3.3 Mapping

To ensure effective understanding and use of this model for tracking and visual identification, there was a need to create a visual with the longitude-latitude locations for the mosquitoes of the Anopheles genus. Unique latitude and longitude points were plotted (as shown in red). This map was then superimposed with the longitude-latitude locations for the mosquitoes of the Aedes and Culex genera (as shown in yellow and blue, respectively). In order to create this visual depiction, a Python notebook was used to write a script that identifies the latitude and longitude points to correlate a specific location on a comprehensive world map, created by Google. The gmplot library, a matplotlib-like interface that generates the necessary HTML and javascript to render data on a Google Map, was used and the data was processed to separate the entries of Anopheles, Aedes, and Culex to plot and overlay each instance of a corresponding latitude and longitude. Upon sorting the latitude-longitude locations of each mosquito by genus, the data was added to a list for easier processing and split into latitude and longitude. In order to create a map that covers all latitude and longitude points in the entire dataset, its maximum and minimum latitude points and longitude points were identified and used in an equation which determines the scale and endpoints of the map.

$$\left(minimum\ latitude + \frac{maximum\ latitude - minimum\ latitude}{2},\right.$$

$$\left. minimum\ longitude + \frac{maximum\ longitude + minimum\ longitude}{2}\right)$$

Credit: Aesha Shah

This equation was derived by the researchers to determine the necessary range for the latitude and longitude without omitting any points that are in the dataset. The scale for the map was set to 10 to ensure visibility of the plotted locations.

4 RESULTS

Analyzing the results of the Logistic Regression model, we observe that our model achieves a training accuracy of 88.24% and a test accuracy of 86.11%. This was conducted with a learning rate of 0.01. As such, these results outperform our original expectations of reaching an accuracy rate of between 70% to 80%. Compared to other alternatives for Anopheles verification, such as through molecular protein profiling, our accuracy metrics from our AI model continue to outperform (Müller et al., 2013).

As we tested our model on the dataset we found its optimal learning rate to be 0.01. The rate was high enough for our model to adapt to identifying the less common Anopheles samples but also low enough to avoid skewing the model toward producing more false positives. This is seen in Figure 3 as the cost is minimized with more iterations.

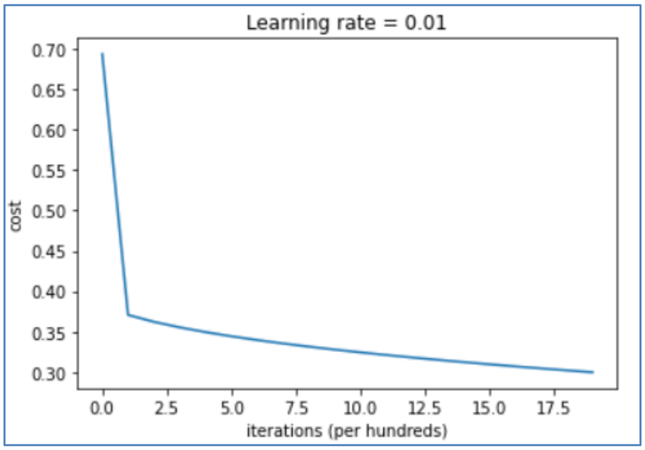

Figure 3: *Learning Rate.*
Credit: Nikita Agrawal

Figure 4 shows the confusion matrix output for our logistic regression model. We find that our model is doing an accurate job at correctly identifying true negatives (non-Anopheles). Still, we also see an increase in the number of false negatives. This is most likely due to the skewness of our original dataset, which had significantly more non-Anopheles than Anopheles mosquitoes and thus resulted in the model being more likely to label an image as negative rather than positive.

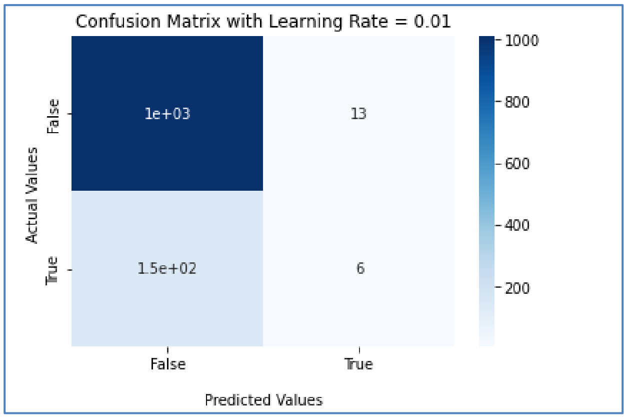

Figure 4: *Confusion Matrix.*
Credit: Nikita Agrawal

Our model successfully mapped the locations of each mosquito image that was inputted using coordinate information from the crowdsourced data. From the map in Figure 7, we see that specifically within the United States, Anopheles mosquitoes are most predominant in the eastern half of the nation while the west coast mainly houses Aedes mosquitoes.

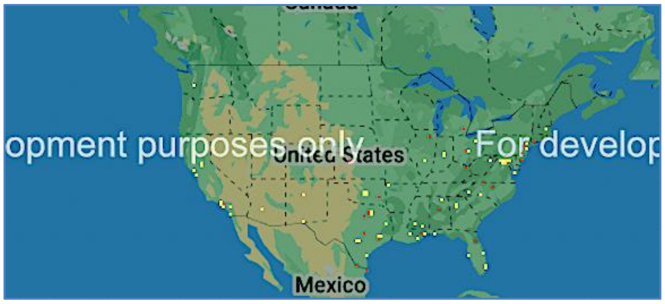

Figure 5: *United States Map.*
Credit: Aesha Shah

Additionally, a greater abundance of Anopheles mosquitoes, compared to Aedes mosquitos, can clearly be seen in the northern and eastern regions of South America. With its tropical climate, Thailand can also be seen as a location with a large mosquito population, containing an equal balance of both Anopheles and Aedes.

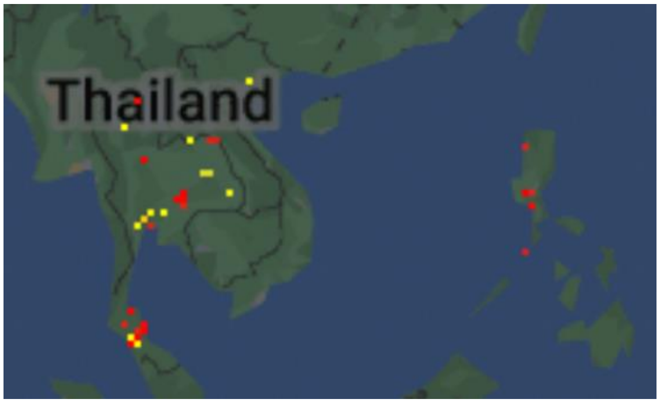

Figure 6: *Thailand Map.*
Credit: Aesha Shah

This directly reflects how Anopheles mosquitoes thrive in more tropical climates while Aedes can thrive in tropical, subtropical, and temperate climates (Potential Range of Aedes Aegypti and Aedes Albopictus in the United States, 2017 |

2017). Already showing an accurate representation of the distribution, our model will be able to provide a map that can be used to identify the likelihood of certain genera of mosquitoes being found in specific regions with increasing accuracy as even more data is inputted.

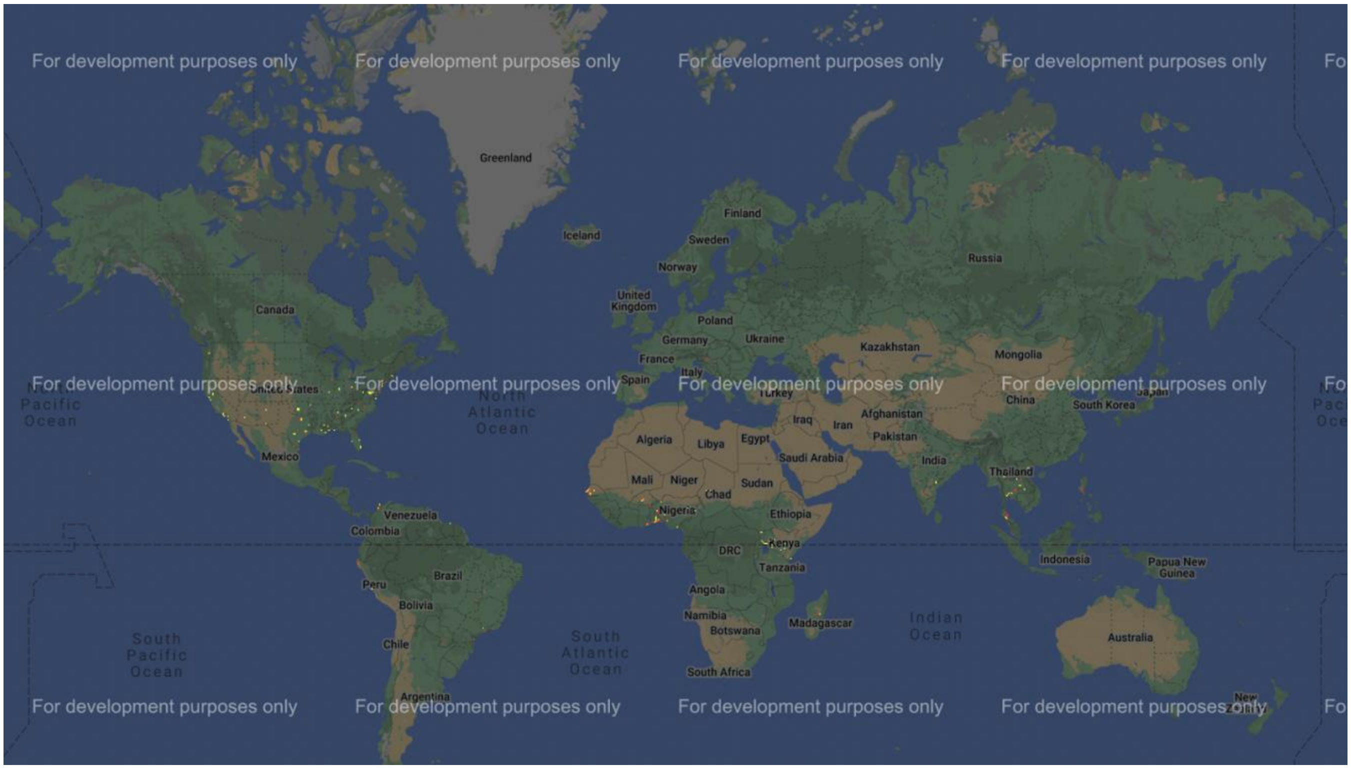

Figure 7: *Map of Aedes, Anopheles, and Culex mosquitoes derived from crowdsourced data: The red dots represent Anopheles mosquitoes, the yellow dots represent Aedes mosquitos, and the blue dots represent Culex mosquitoes.*

Credit: Aesha Shah

5 DISCUSSION

The importance of this model and its effectiveness are vital to identifying and classifying *Anopheles* larvae because of the genus' ability to transmit Malaria. The WHO has identified surveillance as a crucial part of mitigating Malaria outbreaks by pinpointing populations of *Anopheles* mosquitoes (*Malaria*, 2022). Automated, accurate, and reliable identification of Malaria-carrying mosquito larvae, as our team has generated with MASC AI, could be a very effective strategy to identify areas of concern. The further mapping of the location data by the model gives a visual representation of areas prone to exposure to *Anopheles* mosquitoes and, therefore, Malaria. Although vaccines that protect against malaria are recommended for children in areas most affected by the virus, the WHO recommends the vaccine in addition to current preventative measures like surveillance, vector control, and preventive medical treatment (*WHO Recommends Groundbreaking Malaria Vaccine for Children at Risk*, 2021). Malaria is common in sub-tropical areas of the world, and Sub-Saharan Africa has been identified as an area of high risk of Malaria transmission. Although parts of Asia, South America, and Central America have also been identified as areas where Malaria transmission can occur. The WHO African Region represents 95% of deaths from the virus (*Malaria - Malaria Worldwide - Impact of Malaria*, 2021). Early identification has shown to be helpful in both understanding transmission, and mitigating outbreaks.

The model created to help mitigate this global issue was trained with a dataset of 3,275 images. About 53% of those images were taken on the continent of Africa. Of those, about 13% were identified by the verification team as being *Anopheles*. It should be noted that the percentage from our dataset is not representative of the *Anopheles* mosquito population in Africa or the globe. However, the team selected Africa as the main data source because of the relative abundance of *Anopheles* larvae in the GLOBE Mosquito Habitat Mapper database. With the low percentage of verified images of Anopheles larvae in the dataset, the model was only trained with 432 images to represent the *Anopheles* genus. More images of *Anopheles* larvae would have benefited our team's AI model's training by giving it more data to work with in the initial training stage through by balancing out the data groups and in turn increasing the test accuracy of the model.

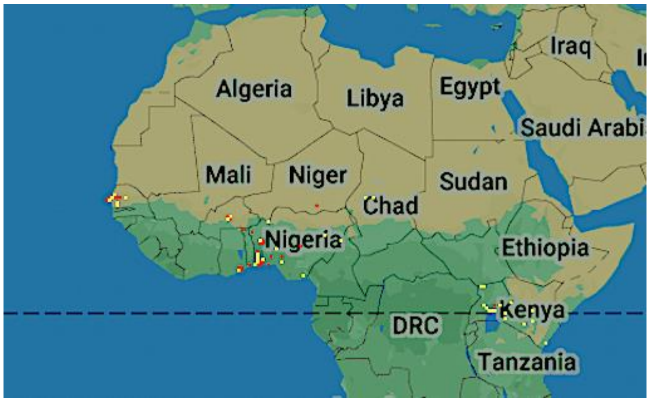

Figure 8: *Africa Map.*

Credit: Aesha Shah

The created model, with a learning rate of 0.01 and an accuracy rate of about 86%, was able to determine whether the mosquito larvae were of the *Anopheles* genus with higher accuracy than the general public who can add photos to the database. The data verification team, who created the dataset on which the model was trained, by determining the genus of all the images, determined that upwards of 90% of the photos in the database had incorrect determinations of larvae's genus. Usage of the MASC AI could improve the accuracy of genus classification for the entire GLOBE MHM database.

6 CONCLUSION

Adopting a more precise method of classifying data using an AI model like MASC, where the training was done using data with proper classification of mosquitoes, could help increase the number of classifications. Our group hopes to add our model as a tool for the GLOBE Observer app to allow people to get the results of their classifications after submitting them. This would let the users gain practice and help them classify with confidence in the future. To improve the way MASC works in correspondence with the Mosquito Habitat Mapper on the Globe Observer app, it would be to request users to retake the images of samples as soon as they submit their observations in case the users discard the samples too soon. One possible room for error with the way MASC can classify the genus of a mosquito is that the images are either blurry or the image does not consist of the region where the siphon might be present, which would lead to the image being classified as *Anopheles*. This model can be used not only for the GLOBE program but also for future researchers as a way to predict future outbreaks by tracking where populations of *Anopheles* reside. This could also be extended to *Culex* and *Aedes* genus mosquitoes as the main difference in their siphon would be that one is smaller and the other longer.

ACKNOWLEDGEMENTS

This project was successful thanks to the guidance of SEES Earth System Explorer mentors, Dr. Rusanne Low, Ms. Cassie Soeffing, Mr. Peder

Nelson, Dr. Erika Podest, Andrew Clark, Dr. Di Yang and peer mentors, Bill Lam, Kellen Meymarian, and Matteo Kimura.

The material contained in this paper 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, judgements or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NASA.

CONTRIBUTION STATEMENT

Nikita Agrawal contributed to the methods section of the paper and worked on data pre-processing, created the logistic regression model, and generated the learning curve graphs and confusion matrices. Samhitha Duggirala contributed to the results and discussions sections of the paper as well as worked on the geographical mapping of mosquito images. Elizabeth Gorman contributed to the discussions section as well as overall revising and formatting of the paper. William Hong contributed to the results and discussions section of the paper as well as worked on the preliminary model development. Joseph (Alex) Kim contributed to the introduction, literature review, and research questions sections of the paper, and worked on finalizing the dataset before model training. Nithin Reddy contributed to the conclusion section and overall revising and formatting of the paper as well as worked on data organization and finalization of the dataset before model training. Aesha Shah contributed to the abstract and methods sections of the paper, as well as worked on data organization, the logistic regression model, and the geographical mapping of mosquito images.

REFERENCES

[1] Albawi, S., Mohammed, T. A., & Al-Zawi, S. (2017, August). *Understanding of a convolutional neural network.* IEEE. Retrieved July, 2022, from https://ieeexplore.ieee.org/abstract/document/8308186?casa_token=INHvV_ahCXsAAAAA:GcEF9V2LHocTP5vIHnMRMTVOAnf1ksAVmJyGf5SKY4H2hwuPz2Y3-FaWLO95A9jAhnsHDCpN5g

[2] Boateng, E. Y., & Abaye, D. A. (2019, November). *A Review of the Logistic Regression Model with Emphasis on Medical Research. Scientific Research Publishing.* Retrieved July, 2022, from https://www.scirp.org/journal/paperinformation.aspx?paperid=95655

[3] Dvorin, J. D. (2017, November 8). *Getting Your Head around Cerebral Malaria - PMC. NCBI.* Retrieved July, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6077980/

[4] *Explorer - Malaria Atlas Project.* (n.d.). the Malaria Atlas Project. Retrieved July, 2022, from https://malariaatlas.org/explorer/#/

[5] Karsten, J. (2022, May 19). *citizen science.* National Geographic Society. Retrieved July, 2022, from https://education.nationalgeographic.org/resource/citizen-science

[6] *Malaria*. (2022, July 26). WHO | World Health Organization. Retrieved July, 2022, from https://www.who.int/news-room/fact-sheets/detail/malaria

[7] *Malaria - Malaria Worldwide - Impact of Malaria*. (2020). CDC. Retrieved July, 2022, from https://www.cdc.gov/malaria/malaria_worldwide/impact.html

[8] *Mosquito Habitats Toolkit - GLOBE Observer.* (n.d.). GLOBE Observer. Retrieved July, 2022, from https://observer.globe.gov/toolkit/mosquito-habitat-mapper-toolkit

[9] Müller, P., Pflüger, V., Wittwer, M., Ziegler, D., Chandre, F., Simard, F., & Lengeler, C. (2013, February 28). *Identification of Cryptic Anopheles Mosquito Species by Molecular Protein Profiling*. NCBI. Retrieved July, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3585343/

[10*] Potential Range of Aedes aegypti and Aedes albopictus in the United States, 2017* | Mosquitoes. (2017). CDC. Retrieved July, 2022, from https://www.cdc.gov/mosquitoes/mosquito-control/professionals/range.html

[11] Russell, R. C. (2000). *How to identify Culex, Anopheles and Aedes mosquitoes and their larvae?* ResearchGate. Retrieved July, 2022, from https://www.researchgate.net/post/How_to_identify_Culex_Anopheles_and_Aedes_mosquitoes_and_their_larvae

[12] Skiff, J., & Yee, D. A. (2014, March). *Behavioral Differences Among Four Co-occurring Species of Container Mosquito Larvae: Effects of Depth and Resource Environments.* NCBI. Retrieved July, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4011075/

[13] *WHO recommends groundbreaking malaria vaccine for children at risk.* (2021, October 6). WHO | World Health Organization. Retrieved July, 2022, from https://www.who.int/news/item/06-10-2021-who-recommends-groundbreaking-malaria-vaccine-for-children-at-risk