Predicting Mosquito Abundance in Chicago Using Remote Sensing Climate Data and Machine Learning

Sheil Dharan[1], Daisy Li[2], Alan Monteiro[3], Giovanni Victorio[4]

[1]*Denmark High School, Alpharetta, GA, USA*
[2]*Alexander W. Dreyfoos School of the Arts, West Palm Beach, FL, USA*
[3]*Colégio ETAPA, São Paulo, Brazil*
[4]*Lamar High School, Houston, TX, USA*

Abstract **- In recent years, mosquito-borne diseases such as the Zika virus, West Nile virus, Chikungunya virus, Dengue, and Malaria have become more prevalent in urban areas due to various climate and anthropogenic factors. This led to a greater need for mosquito abundance prediction to improve the response to disease outbreaks, especially during the summer when mosquito season peaks and outdoor activities increase significantly. The objective of this study was to evaluate the accuracy of six machine learning models for classifying extreme mosquito abundance events based on climate data. Data sourced from the Mosquito Habitat Mappers challenge on GLOBE and a City of Chicago dataset were matched to area-averaged time-series climate data for Chicago from GIOVANNI, a NASA open access remote sensing database for Earth science. Data was cleaned and then aggregated to a single weekly time-series dataset consisting of mosquito abundance, and the past week's three climate variable averages. The models were trained and tested on climate data, namely surface humidity, precipitation, and daytime temperature. The mosquito and climate data were recorded from five Chicago summers. The results indicated that the best models for predicting mosquito abundance events were the ensemble learning methods of AdaBoost and Random Forest, respectively. Future avenues of research include using other, more-specific factors for prediction such as the chlorophyll from algal blooms (increasingly common due to direct and indirect anthropic activities, such as fertilizer runoff and warming waters due to climate change), more localized predictions, accounting for the microclimates of urban areas, and using regression models to predict precise mosquito numbers.**

Keywords: Mosquito Abundance, Machine Learning, Classification, Climate Variables, Remote Sensing

I. INTRODUCTION

Mosquito-borne diseases account for over 17% of all infectious diseases and cause more than 700,000 annual deaths according to (World Health Organization [WHO], 2020). As noted in (Petersen et al., 2019), in recent years, cases of vector-borne diseases in the United States have rapidly increased along with sporadic outbreaks of domestic and invasive mosquito-borne diseases. (Petersen et al., 2019) also notes that proven and scalable public health control measures do not exist and measures that do may not be effective, timely, or occur at all. In fact, (National Association of County & City Health Officials [NACCHO], 2017) found in a survey that 84% of surveyed vector-control operations are lacking in at least one out of the five core competencies of vector-control. More

specifically, urban areas are at a higher risk of mosquito-borne disease outbreak and experience very high transmission rates as shown in (Thang et al., 2019). Chicago is one city that experiences mosquito-borne diseases, particularly the West Nile Virus. Instances such as the Chicago West Nile Virus outbreak of 2002 as well regular cases of the virus occur in the City of Chicago. The virus affects the urban and suburban areas of Chicago and is a very serious illness. The prevalence of the West Nile Virus has been shown in (Tedesco et al., 2010). The West Nile Virus has no specific treatment or vaccine according to (Center for Disease Control [CDC], 2021). Summer is the optimal time for the highest mosquito frequency rates due to warm weather posing favorable conditions for mosquito reproduction and habitat creation. Compared to the numbers year-round, these sudden outbreaks can be difficult to control mosquito-borne diseases. It is also understood that these population peaks are influenced by climate and overall weather conditions, so along with many other studies, this study aims to predict their magnitude to avoid surprises to the city's health infrastructure. The influence of temperature was shown to be significant and mostly positive, augmenting growth rates of populations (Paz, 2008): warming of the mosquito environment boosted their rates of reproduction and number of blood meals, prolonged their breeding season, excluding the case of extreme temperatures exceeding mosquitos' survivability limits (Drakou et al., 2020). In this factor trifecta, rainfall had the least predictable relationship (but still has a correlation) with mosquito presence, as various studies have analyzed its influence and found that for each habitat and mosquito-specific situation, there were different lag times in between higher and lower precipitation periods along with increased or decreased mosquito numbers. Unlike rainfall, relative humidity plays a more general role. For instance, mosquitoes become inactive to maintain body fluids and reduce energy use in low humidity environments. As a result of the insufficient treatment methods and prevalence of mosquito-borne disease outbreak in urban areas, a method of predicting mosquito abundance is vital to preventing the spread of disease. Predicting mosquito abundance can help increase the efficacy of response efforts to a mosquito abundance event. Since correlations have been found in past research between climate variables and mosquito abundance, considering climate variables such as humidity, temperature, and precipitation can be used to predict mosquito abundance. Much of recent research in this area has applied machine learning to this task because some machine learning algorithms, particularly supervised machine learning algorithms, are efficient at modeling relationships between features, such as climate variables, and targets, such as abundance classifications. Many past studies have utilized machine learning to prevent the spread of deadly infectious diseases such as COVID-19. (Alfred & Obit, 2021) is an example of one such study which overviewed the use of machine learning in disease prevention. However, most past studies that utilized machine learning for disease prevention utilized only the Neural Network and Support Vector Machine (SVM) machine learning models according to (Schaefer et al., 2020). Machine learning has also been used to predict mosquito abundance based on socioeconomic and land cover data such as that of (Chen et al., 2019). Many of these machine learning studies have found moderate to high success rates in predicting diseases and mosquito abundance. This study aims to implement and compare the performance of six different machine learning classifiers (Random Forest, Neural Network, Naïve Bayes, Support Vector Machine, AdaBoost, and k-Nearest Neighbor) in predicting mosquito abundance based on three remote sensing climate variables (temperature, humidity, and precipitation) in the City of Chicago. The City of Chicago was chosen due to the availability of GLOBE citizen science data,

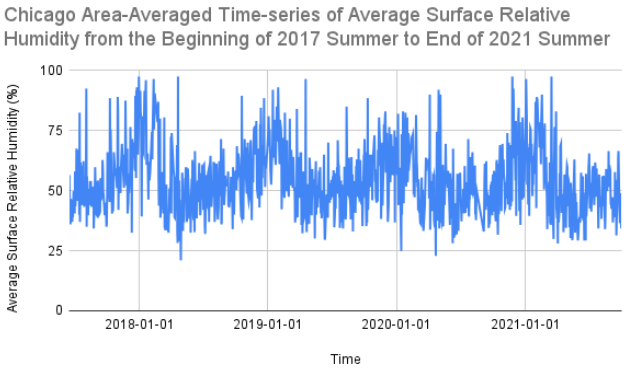
governmental mosquito data, and remote sensing climate data. In addition, Chicago has a high population density, is an urban area, and has ecological variability. The research questions answered by this study are: "Can supervised machine learning models be used to predict mosquito abundance in Chicago based on the remote sensing climate variables of temperature, humidity, and precipitation?" and "If so, which machine learning model will perform the best?". We (the authors) hypothesized that mosquito abundance could be predicted by machine learning models based on remote sensing climate data and that the AdaBoost model would perform the best due to its boosting nature and its low generalization error as stated in ref.

II. RESEARCH METHODS

There were four main stages to this research process: data acquisition, data preparation, machine learning, and evaluation. Data acquisition involved retrieving mosquito abundance data and climate data. Data preparation organized the data into a single, uniform time series comma-separated values file that could be analyzed. The machine learning process utilized six different models to predict a mosquito bloom event (defined as greater than 1386.743 mosquitos). The evaluation processes used standard machine learning evaluation metrics to compare the performance of the models.

A. Data Acquisition

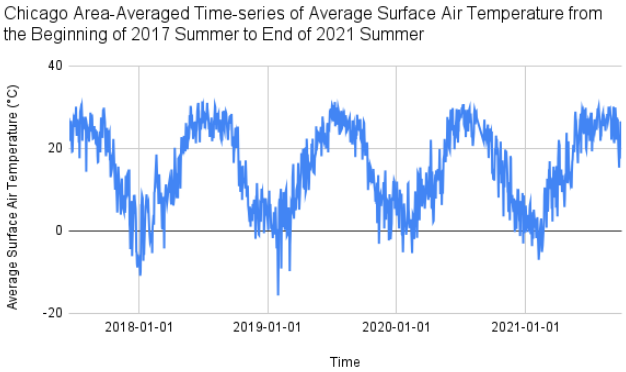
The mosquito abundance data was sourced from sample sizes of regular observations of mosquito traps in the City of Chicago's Data Portal as well as GLOBE Observer Mosquito Habitat Mapper observations in the Chicago area during the summer months from 2017 to 2021. The climate variables selected in this experiment were the average daily surface relative humidity (SRH) as shown below in Figure 1



,

Fig. 1. Chicago area-averaged time-series of the average surface relative humidity from the beginning of 2017 summer to the end of 2021 summer. Graph created by authors; data retrieved from GIOVANNI, (AIRS Science Team & Teixeira, 2013).

the average daily daytime surface air temperature (SAT) as shown below in Figure 2



,

Fig. 2. Chicago area-averaged time-series of the average surface air temperature from the beginning of 2017 summer to the end of 2021 summer. Graph created by authors; data retrieved from GIOVANNI, (AIRS Science Team & Teixeira, 2013).

and the daily precipitation as shown on the next page in Figure 3

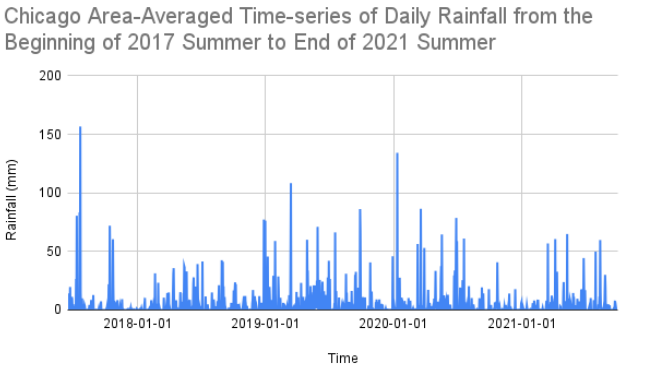


Fig. 3. Chicago area-averaged time-series of the daily precipitation from the beginning of 2017 summer to the end of 2021 summer. Graph created by authors; data retrieved from GIOVANNI, (Huffman et al., 2019).

These variables were chosen because mosquitos' abundance is based on their toleration of conditions such as humidity, precipitation, and temperature. The rapid change of these factors in Chicago provide a good base for machine learning to make predictions of mosquito outbreaks. Climate data was retrieved from NASA's *Geospatial Interactive Online Visualization and Analysis Infrastructure* (GIOVANNI), a web-based tool for visualizing, analyzing, and accessing Earth science remote sensing data. Each parameter was area-averaged for the Chicago area (Bounding Box Coordinates: -87.9110W, 41.60581N, -87.4606W, 42.0417N)

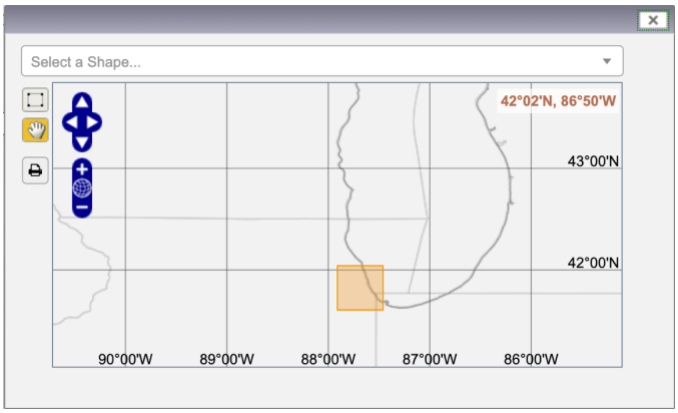


Fig. 4. Bounding box of Chicago used for mosquito habitats and area-averaged remote sensing climate data as shown in the GIOVANNI interface, ("Giovanni").

and downloaded as a time-series onto a comma-separated values (CSV) file. SRH and SAT data were sourced from the Atmospheric Infrared Sounder (AIRS) on NASA's Aqua satellite. The precipitation data was obtained from an international network of satellites that provide global observations of precipitation called the Global Precipitation Measurement (GPM).

B. Data Preparation

Regular mosquito traps reportings from the GLOBE Mosquito Habitat Mappers database as well as the City of Chicago data portal (which includes mosquito trap numbers from a dataset tracking the West Nile Virus) were aggregated into a single weekly time-series dataset of the mosquito count in each of the five Chicago summers from 2017 to 2021. The City of Chicago mosquito observations can be seen below in Figure 5.

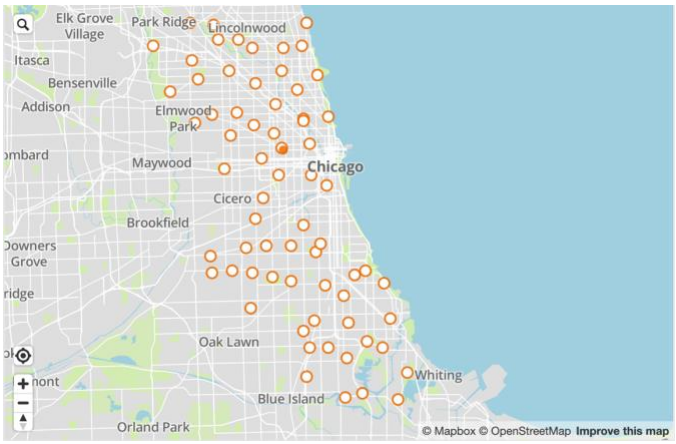


Fig. 5. City of Chicago Data Portal of mosquito observation locations for the five Chicago summers from 2017 to 2021, (City of Chicago, 2022).

The GLOBE Mosquito Habitat protocol data availability is shown on the next page in Figure 6.

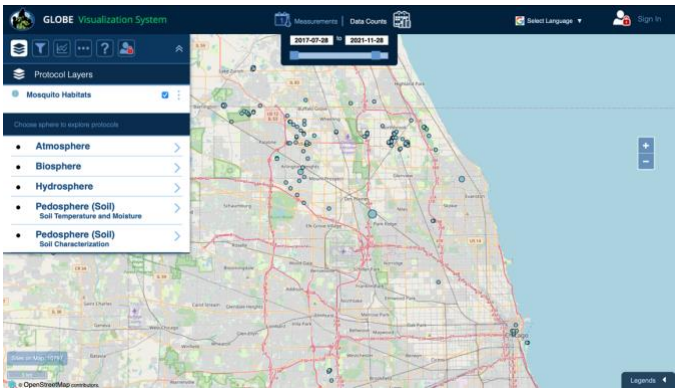


Fig. 6. GLOBE Data Visualization of the Mosquito Habitat protocol locations for the five Chicago summers from 2017 to 2021, ("GLOBE Program").

From the GLOBE Data, we utilized the location (ensuring it was in Chicago), the date, the location type (only using regular traps), and the mosquito count as shown in Figure 7 below.

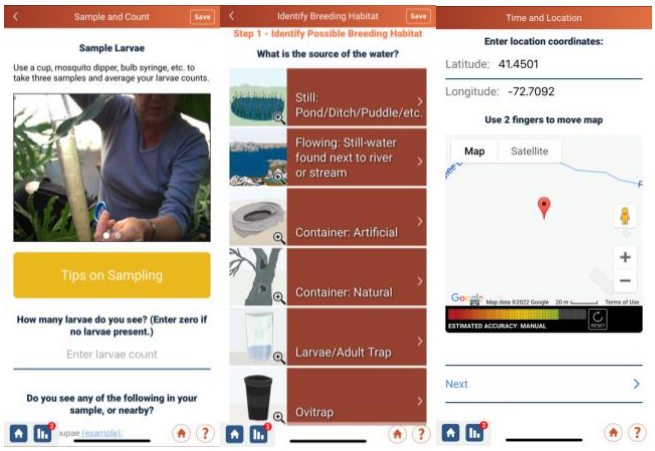


Fig. 7. GLOBE Observer Mosquito Habitat Mapper Data Entry screenshots. These three images include the data that we used from the GLOBE Mosquito Habitat Mapper protocol to create our mosquito count dataset, ("GLOBE Program")

Climate data from GIOVANNI was downloaded for the same period as the mosquito data and matched with each other. Since the data was daily, it was averaged on a weekly basis according to dates of mosquito observance in the mosquito dataset. There were null and missing values in the climate data, so the AVERAGEIF() function was used to skip days with no clean data. Next, the mosquito data was matched to the preceding week's climate averages for each of the three factors. Finally, a new column was added to the data, a Boolean true or false: according to whether the mosquito number was above (true) or below (false) the average mosquito frequency of 1386.743 excluding outliers, assuming mosquito frequency was a Poisson distribution (a type of distribution regularly used for time-series data as described in (Heinen, 2003)).

C. Machine Learning

The final CSV file from the data preparation process was uploaded into the Orange Data Mining software (Demsar et al., 2013), an open-source machine learning and data visualization toolbox. The data were then split into a training category comprised of seventy percent of the data and a testing category comprised of the remaining thirty percent of the data. The data was then put through the machine learning pipeline for six different models and evaluated under different measures of effectiveness. Using the pre-processing feature of Orange, the data were normalized using the following equation:

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}}.$$

Fig. 8. Normalization equation. Image generated using (Codecogs).

Because the input dataset is composed of features of different units, ranges, magnitudes, and properties, normalization is a vital step to ensure maximum efficiency for algorithms that are sensitive to differences and make predictions based on differences between data points.

C (i). Random Forest

The Random Forest classifier is a supervised machine learning algorithm consisting of a "forest" of decision trees. A decision tree is comprised of nodes where "decisions" are made based on certain features of the input data. A random forest contains a large group of independent decision trees which make independent decisions on the classification of an instance based on the input data. The independent decision trees have different nodes and make differing decisions at nodes to output different classifications. The classification that is made most often becomes the random forest's "final decision". The key to this algorithm is the low correlation between the decision trees, which helps mitigate errors that a specific decision tree may be prone to. (Yiu, 2021)

C (ii). Neural Network

The Neural Network classifier used in this study was a multi-layer perceptron neural network. The input layer in this study consisted of three nodes, through which the three climate variables were input. The input signals feedforward through the network comprised of nodes or "neurons". In each layer of the neural network, the value of a node is multiplied by the weight of the connection to the node in the next layer which is added to a bias value and passed through an activation function. Activation functions are used to prevent linearity and transform the input to do more complex tasks. The neural network in this study used the rectified linear unit (ReLu) activation function. Additionally, the model used adaptive learning rate optimization (Adam). This optimizer decreases computation time and requires fewer parameters for tuning. The network eventually converges to an output layer of two nodes: one with a classification of mosquito abundance event, and one with the classification of no mosquito abundance event. The model is then re-run and neurons are adjusted to increase the classification accuracy of the final output. (Hardesty, 2017)

C (iii). Naïve Bayes

The Naive Bayes classifier is a probabilistic machine learning model that is used for classification. The classifier is modeled on the Bayes theorem:

$$P(y|X) = \frac{P(X|y) \cdot P(y)}{P(X)}.$$

Fig. 9. Bayes theorem. Image generated using (Codecogs).

The model is considered "naive" because it assumes that each input variable is independent. This is unrealistic for real-life data; however, the technique can be very effective on a large range of computing problems. The Bayes theorem is used to calculate the membership probability for an instance based on the features for each classification. The classification with the highest probability is the classification output by the classifier. (Gandhi, 2018)

C (iv). Support Vector Machine (SVM)

The Support Vector Machine (SVM) classifier observes each instance (in our study each date of mosquito recording) in an n-dimensional space, where n is the number of features in the dataset. Since the dataset in this study includes three climate variables, the classifier works in a three-dimensional space, although values of n that are more than three can exist. The data are first passed through a linear kernel function that converts the data into a separable format that will fit in the three-dimensional space. The classifier works to find the optimal hyperplane based on the number of output classifications. Since this study focuses on a binary classification, the hyper-plane is a one-dimensional line. This plane maximizes the distance between the data points that are known to be in separate classes. Thus, when a

testing data point is introduced, its location in the three-dimensional plane relative to the hyperplane determines the classification that is output by the SVM classifier. The data points closest to the hyperplane are called support vectors and play an important role in determine the orientation and position of the hyperplane. (Yadav, 2018).

C (v). k-Nearest Neighbors (k-NN)

The k-Nearest Neighbors classifier acts in a similar fashion to the SVM classifier. The classifier also attempts to segregate the data points in a multidimensional space. The k-NN algorithm assumes that objects of the same class will be near one another. Thus, the algorithm computes the distance between a new given point and the distance between each of the other points in the dataset using the formula

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Fig. 10. k-NN standard Euclidean distance formula. Image generated using (Codecogs).

where n is the number of variables The k-nearest data points will be considered by the algorithm. The classification of the new data point by the algorithm is the same classification of most of the neighboring points. A k-value of ten was used in this study as it was determined to be experimentally optimal for out dataset. (Kumar, 2021).

C (vi). AdaBoost

The AdaBoost classifier is a boosting technique that is used as an Ensemble learning method. AdaBoost builds on top of another classifier, in the case of this study: the Random Forest classifier. The algorithm utilizes multiple weak classifiers to build a single strong classifier. Weak classifiers are classifiers that perform better than random guessing, but still poor in general. This study utilized decision stumps, which are like the decision trees of a Random Forest but are not fully grown containing only one node and two leaves. Although these stumps are not a good way to make decisions on their own, using AdaBoost over the stumps can lead to a more accurate classifier. As per the process, initially the decision stumps are used for each variable to see how well each stump classifies sample to their correct target class. More weight is assigned to the incorrectly classified samples, so they are classified correctly by the next decision stump. Each classifier also receives a weight with higher weights being assigned for more accurate classifiers. This process is iterated until all the training data is classified correctly or the maximum iteration level has been reached. (Freund & Schapire, 1996).

D. Evaluation

Each of the models was evaluated under five standard machine learning classification metrics: Area under the receiver operating characteristic (ROC) curve (AUC), classification accuracy, harmonic mean between precision and recall (F1), precision, and recall ("Classification", 2022). All these metrics are based on the notion of a confusion matrix:

$$\begin{array}{cc} & \textit{Actual} \\ & \begin{array}{cc} \textit{Positive} & \textit{Negative} \end{array} \\ \textit{Predicted} \begin{array}{c} \textit{Positive} \\ \textit{Negative} \end{array} & \begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix} \end{array}.$$

Fig. 11. Confusion matrix. Image generated using (Codecogs).

True positives (TP) occur when the model predicts a mosquito abundance event for the week a mosquito

abundance event occurred. False positives (FP) occurred when the model predicts a mosquito abundance event for a week when a mosquito abundance event does not occur. False negatives (FN) occur when the model predicts no mosquito abundance event for a week where a mosquito abundance event occurred. True negatives (TN) occur when the model predicts no mosquito abundance event for a week where no mosquito abundance event occurred. The first letter (T or F) describes the correct or incorrect classification of the model, and the second letter (P or N) describes the actual classification. Thus, the TP and TN are correct classifications. These four values were then used to calculate the five-evaluation metrics. The

D (i). Area Under ROC Curve (AUC)

The AUC is used to measure the ability of each of the classifiers to distinguish between classes and is used as a summary score for the ROC curve. The ROC curve shows the performance of a classifier at different classification thresholds plotting the true positive rate (TPR) or recall:

$$TPR = \frac{TP}{TP + FN}$$

Fig. 12. True positive rate. Image generated using (Codecogs).

on the y-axis and the false positive rate (FPR):

$$FPR = \frac{FP}{TP + TN}$$

Fig. 13. False positive rate. Image generated using (Codecogs).

on the x-axis. Lowering the classification threshold classifies more positives so both the TPR and FPR increase. The area under the curve is found using integration to provide an aggregate measure of performance across every classification threshold. The closer to one the AUC score is, the better the model's performance.

D (ii). Classification Accuracy (CA)

Classification accuracy is considered the raw accuracy of a model. It is simply the percent of classifications that the model got correct and is calculated with the following formula:

$$CA = \frac{TP + TN}{TP + TN + FP + FN}.$$

Fig. 14. Classification accuracy formula. Image generated using (Codecogs).

D (iii). Precision

Precision is used to measure what proportion of positive identifications were correct. The higher the precision (and thus closer to 1), the more effective the classifier. The formula used to calculate precision is

$$Precision = \frac{TP}{TP + FP}.$$

Fig. 15. Precision formula. Image generated using (Codecogs).

D (iii). Recall

Recall is used to measure what proportion of the positives were identified correctly. The higher the recall (and thus closer to 1), the more effective the classifier. The formula used to calculate recall is

$$Recall = \frac{TP}{TP + FN}.$$

Fig. 16. Recall formula. Image generated using (Codecogs).

D (iv). F1 Score

The F1 score is the harmonic mean between precision and recall. The metric is a supposed improvement on the two simpler performance metrics of precision and recall. The F1 gives equal weight to precision and recall. Thus, a higher F1 score would indicate a higher performing classifier. The F1 score is calculated using the following formula:

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}.$$

Fig. 17. F1 score formula. Image generated using (Codecogs).

III. RESULTS

	Classification Metrics				
Model	*AUC*	*CA*	*Precision*	*Recall*	*F1*
Random Forest	**0.988**	**0.944**	**0.945**	**0.944**	**0.944**
Neural Network	0.871	0.748	0.748	0.748	0.748
Naïve Bayes	0.920	0.868	0.854	0.856	0.855
SVM	0.885	0.772	0.771	0.772	0.771
k-NN	0.822	0.748	0.746	0.748	0.747
AdaBoost	**0.996**	**0.973**	**0.963**	**0.962**	**0.962**

Table. 1. Comparison of the performance off six different classification machine learning models described in II-C, as measured by the five classification metrics as described in II-D. The highest performing models are AdaBoost, followed closely by Random Forest.

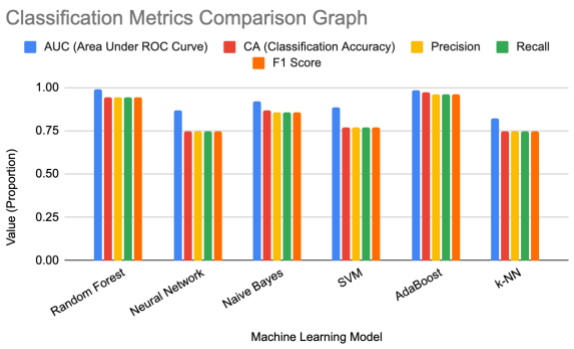


Fig. 18. Comparison of the six machine learning models under the five classification metrics in the form of a bar graph.

IV. DISCUSSION

All six models performed well with classification accuracies above 75%. This finding is likely due to the high quality and large datasets on mosquito counts and climate variables in the Chicago area. These results show that there is a high correlation between daily dynamic climate variables and mosquito abundance. Algorithms that performed the best were the ensemble learning algorithms of the Random Forest classifier (with a classification accuracy of 94.4%) and the AdaBoost classifier (with a classification accuracy of 99.6%), which built on top of the Random Forest classifier (see II-C(vi)). The high performance of these two algorithms can be explained by the algorithms' ability to consider the most important attributes of the climate variables and the massive number of iterations and decision trees used to derive a classification. In addition, AdaBoost is known to have low generalization error, which means that the algorithm is much less prone to overfitting and performs better classification on previously unseen (testing) data. This phenomenon was noted in (Vezhnevets & Vezhnevets, 2005). The Random Forest classifier may have also performed well because of its ability to balance data when the amount of data in one class outnumbers the amount of data in another class, as it it did in this study. This feature of the Random Forest algorithm is widely known and is stated in (More & Rana, 2017). Furthermore, the AdaBoost and Random Forest algorithms had AUCs, Precisions, Recalls, and F1 scores all above 90%. This shows the algorithms do not have a weakness in classifying a certain class or classifying the data at different classification thresholds. These evaluation metrics show that the algorithms are good all-around classifiers for the data from this study. The Neural Network, SVM, and k-NN performed the worst with classification accuracies of 74.8%, 77.2%, and 74.8% respectively. The Neural Network's performance could be limited

as this dataset involved dealing with both changes in patterns in the data as well as pattern recognition. This corroborates with (Oleinik, 2019), which states that because neural networks are good at identifying patterns in structured data, they are limited when they must combine pattern combination and recognition. The SVM's performance may be limited because the dataset is very noisy in terms of its target classes being classified as above or below a mosquito abundance value. In addition, SVM is less effective in low-dimensional spaces, so the use of only three climate variables may have limited the model's effectiveness. These limitations have been notes in the past in (Yadav, 2018). Finally, the k-NN classifier's performance may be limited to redundant features in the data as all features contribute similarly as noted in (Imandoust & Bolandraftar, 2013). The dataset used in this study included the climate variables of humidity, temperature, and precipitation, all of which are interrelated to each other. Although the models in this study performed well, there are some possible sources of error that could have affected the accuracies of the model in a positive way overexaggerating the effectiveness of the classifiers or a negative way, underexaggerating the effectiveness of the classifiers. These sources of error mainly lie in the data sources themselves. The GIOVANNI data sources did have missing data values which if the climate variables were extreme for that day, could have affected the weekly averages and caused a mosquito abundance event. The GIOVANNI data also did not include errors which is unusual for historical climate data. Adding on, the mosquito counts, although taken from regular traps, could have external factors inflating or deflating the mosquito counts. Since, Chicago is an urban area, many microclimates exist where extreme climates can exist as a result for factors not accounted for in this study. These factors range from river eutrophication to man-made factors such as cars or human activity could be influencing mosquito counts. In other words, since data for this study was not specifically collected for the purpose of this study, but just for citizen science in general, a host of confounding variables other than the climate variables collected from remote sensing could be acting upon mosquito counts. To improve this, future studies should be longitudinal and measure mosquito counts from controlled traps specifically aimed at collecting mosquito data for the purpose of the study. One similar study to this one, (Chen et al., 2019), also uses several machine learning models to predict mosquito abundance. However, the predictions were based on socioeconomic, and landscape (land cover) features rather than the climate features utilized in this study. In addition, Charlotte was chosen as the city of interest due to its vast socioeconomic inequality and landscape differences. Furthermore, the study only compared three classifiers: k-NN, SVM, and a Neural Network. F1 scores higher as the best classifiers in this study were only achieved when both the socioeconomic and landscape factors were combined on a continuous input. Another study in Chicago, (Gardner et al., 2013), used machine learning to characterize the relationship between terrestrial vegetation and aquatic chemistry and mosquito abundance. This study measured mosquitoes for the study itself and focused on the spatial and temporal variation in the mosquito vectors and their larval production in relation to the spread of the West Nile Virus. The results support the original hypothesis that machine learning classification could be applied to predict mosquito abundance and that the AdaBoost classifier would perform the best at predicting mosquito abundance in the Chicago area based on the three climate variables. This is shown through the AdaBoost's AUC, CA, Precision, Recall, and F1 score all over 95% as well as each of the classification metrics being higher than those of the other classifiers.

V. CONCLUSION

All six of the machine learning classifiers performed well in predicting mosquito abundance based on climate variables, as measured by the classification metrics. In particular, the AdaBoost and Random Forest models are particularly strong at predicting mosquito abundance. The findings from this study can be applied to current readily available remote sensing climate data to predict future mosquito abundance events. Predicting mosquito abundance events is important as preventative measures can be taken such as mosquito habitat eradication. Preventing mosquito abundance in urban areas is especially important because of the rise of mosquito-borne diseases such as the Zika virus, West Nile virus, Chikungunya virus, Dengue, and Malaria. These diseases spread rapidly in urban areas and can cause severe illness and in some cases even death. Preventing mosquitos' abundance and removing breeding sites are key to stopping the spread of these diseases. The link between mosquitoes and diseases is widely known and has been written about in articles such as (Tolle, 2009). In addition, machine learning methods like the one in this study are much more cost-effective and can work on a wide range of land cover without needing to record data in specific locations, as remote sensing data is used. Additionally, the findings show the importance for the creation of large and high-quality publicly available datasets for easily tracked variables such as mosquito habitats. More citizen science initiatives will lead to more available data for data analysis studies to reach findings that can advance scientific knowledge. An improvement to the methodology of this study could be doing fieldwork in addition to the machine learning data analysis. Collecting mosquito data for the purpose of this study could help make data more regular and control for external variables. In addition, sensor data could be collected for the locations to account for the vast microclimates in urban areas rather than using the area-averaged remote sensing climate data. Furthermore, regression analysis could be done to predict precise mosquito numbers as opposed to classifying mosquito abundance events. One final improvement to the methodology would be to use satellite imagery and a convolutional neural network to identify potential mosquito habitats, and then climate variables to help predict mosquito abundance. This could narrow down the prediction of mosquito abundance to specific locations where preventative measures could be taken. In the future, more variables could be used for the machine learning model to make a correlation with for mosquito abundance prediction. This could include more specific factors such as chlorophyll (which is readily available through remote sensing). Chlorophyll can be an indicator of algal blooms and eutrophication, both of which have and will become increasingly frequent in local Chicago water bodies as predicted in (Schelske & Stoermer, 1971). Eutrophication has been linked to mosquito survival and development as found in (Schrama et al., 2018). Furthermore, it might be beneficial to look at the link between mosquito abundance and socioeconomic factors such as housing prices and average household income as poorer zip codes may have infrastructure that is more prone for mosquito habitat development and abundance as studied in (Chen et al., 2019). Future GLOBE protocols that could be added include Land Cover for mosquito habitat location predictions; pedosphere protocols such as Soil Temperature, Soil Moisture, and other soil characterization protocols (such as Soil Density); and other hydrosphere protocols such as nitrates or pH. Working with project mentors throughout the research process really enhanced the efficiency of our research. They helped us narrow down our vast ideas to a specific project that could completed in a realistic timeframe. The mentors also introduced and taught us how to use many of the tools that are used heavily in the Earth science community such as

GIOVANNI and GLOBE. Furthermore, the mentors facilitated us through the novelties of the scientific research process including literature review and formatting research papers.

ACKNOWLEDGMENT

We would like to acknowledge our research mentors for their continued technical support throughout our research: Dr. Rusty Low, from the Institute of Global Environmental Strategies; Dr. Cassie Soeffing from the Institute for Global Environmental Strategies; Andrew Clark from the Institute for Global Environmental Strategies; Peder Nelson from Oregon State University; Dr. Erika Podest from NASA's Jet Propulsion Laboratory; and our peer mentor, Alessandro Greco, from Western Canada High School. In addition, we would like to thank everyone from the SEES STEM Enhancement in Earth Science 2022 cohort and especially the GLOBE Earth System Explorers 2022 cohort including all our mentors, administrative members, guest speakers, and peers.

AUTHOR CONTRIBUTION STATEMENT

S.D. cleaned the data, conceived the machine learning models, and obtained the results. D.L. obtained the mosquito and climate date. A.M. coordinated the research and obtained background information. G.V. reviewed literature and produced media. All authors drafted the final manuscript and other written products.

COMPETING INTERESTS

The authors declare that they have no competing interests.

REFERENCES

[1] Vector-borne diseases. (2020). *World Health Organization.* Retrieved from https://www.who.int/news-room/fact-sheets/detail/vector-borne-diseases

[2] Petersen, L. R., Beard, C. B., & Visser, S. N. (2019). Combatting the Increasing Threat of Vector-Borne Disease in the United States with a National Vector-Borne Disease Prevention and Control System, *The American Journal of Tropical Medicine and Hygiene*, 100(2), 242-245. Retrieved from https://www.ajtmh.org/view/journals/tpmd/100/2/article-p242.xml

[3] Mosquito Control Capabilities in the U.S. (2017). *National Association of County & City Health Officials*. Retrieved from https://www.naccho.org/uploads/downloadable-resources/Mosquito-control-in-the-U.S.-Report.pdf

[4] Thang Nguyen-Tien, Åke Lundkvist & Johanna Lindahl (2019) Urban transmission of mosquito-borne flaviviruses – a review of the risk for humans in Vietnam, *Infection Ecology & Epidemiology*, 9:1, https://doi.org/10.1080/20008686.2019.1660129

[5] Tedesco, C., Ruiz, M., & McLafferty, S. (2010). Mosquito politics: Local vector control policies and the spread of West Nile Virus in the Chicago region. *Health & Place* (Vol. 16, Issue 6, pp. 1188–1195). https://doi.org/10.1016/j.healthplace.2010.08.003

[6] Treatment & Prevention. (2021). *Center for Disease Control*. Retrieved from https://www.cdc.gov/westnile/healthcareproviders/healthCareProviders– TreatmentPrevention.html

[7] Paz, S., Albersheim, I. (2008). Influence of Warming Tendency on *Culex*

pipiens Population Abundance and on the Probability of West Nile Fever Outbreaks (Israeli Case Study: 2001–2005). *EcoHealth* **5,** 40–48 (2008). https://doi.org/10.1007/s10393-007-0150-0

[8] Drakou, K., Nikolaou, T., Vasquez, M., Petric, D., Michaelakis, A., Kapranas, A., Papatheodoulou, A., & Koliou, M. (2020). The Effect of Weather Variables on Mosquito Activity: A Snapshot of the Main Point of Entry of Cyprus. *International journal of environmental research and public health*, *17*(4), 1403. https://doi.org/10.3390/ijerph17041403

[9] Alfred, R., & Obit, J. H. (2021). The roles of machine learning methods in limiting the spread of deadly diseases: A systematic review. *Heliyon* (Vol. 7, Issue 6, p. e07371). https://doi.org/10.1016/j.heliyon.2021.e07371

[10] Schaefer, J., Lehne, M., Schepers, J., Prasser, F., & Thun, S. (2020). The use of machine learning in rare diseases: a scoping review. *Orphanet Journal of Rare Diseases* (Vol. 15, Issue 1). https://doi.org/10.1186/s13023-020-01424-6

[11] Chen, S., Whiteman, A., Li, A., Rapp, T., Delmelle, E., Chen, G., Brown, C. L., Robinson, P., Coffman, M. J., Janies, D., & Dulin, M. (2019). An operational machine learning approach to predict mosquito abundance based on socioeconomic and landscape patterns. *Landscape Ecology* (Vol. 34, Issue 6, pp. 1295–1311). https://doi.org/10.1007/s10980-019-00839-2

[12] AIRS Science Team/Joao Teixeira (2013), AIRS/Aqua L3 Monthly Standard Physical Retrieval (AIRS-only) 1 degree x 1 degree V006, Greenbelt, MD, USA, *Goddard Earth Sciences Data and Information Services Center (GES DISC),* https://doi.org/10.5067/Aqua/AIRS/DATA321

[13] Huffman, G.J., E.F. Stocker, D.T. Bolvin, E.J. Nelkin, Jackson Tan (2019), GPM IMERG Final Precipitation L3 1 day 0.1 degree x 0.1 degree V06, Edited by Andrey Savtchenko, Greenbelt, MD, *Goddard Earth Sciences Data and Information Services Center (GES DISC)*, https://doi.org/10.5067/GPM/IMERGDF/DAY/06

[14] Giovanni. (n.d.). The Bridge Between Data and Science. *NASA GIOVANNI.* Retrieved from https://giovanni.gsfc.nasa.gov/giovanni

[15] City of Chicago (2022). West Nile Virus (WNV) Mosquito Test Results: *City of Chicago: Data Portal*. Retrieved from https://data.cityofchicago.org/Health- Human- Services/West- Nile- Virus- WNV- Mosquito- Test- Results/jqe8-8r6s

[16] Global Learning and Observations to Benefit the Environment (GLOBE) Program, *Data Accessed: 2022, July 10*, Retrieved from globe.gov

[17] Heinen, A. (2003). Modelling time series count data: An autoregressive conditional Poisson model. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1117187

[18] Demsar J, Curk T, Erjavec A, Gorup C, Hocevar T, Milutinovic M, Mozina M, Polajnar M, Toplak M, Staric A, Stajdohar M, Umek L, Zagar L, Zbontar J, Zitnik M, Zupan B (2013) Orange: Data Mining Toolbox in Python, *Journal of Machine Learning Research* 14(Aug): 2349−2353.

[19] Codecogs. (n.d.). Code Cogs Equation Editor. *CODECOGS.* Retrieved from https://latex.codecogs.com/eqneditor/editor.php

[20] Yiu, T. (2021). Understanding Random Forest. *Towards Data Science.* Retrieved from https://towardsdatascience.com/understanding- random- forest- 58381e0602d2

[21] Larry Hardesty. (2017). Explained: Neural networks. *MIT News.* Retrieved from https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414

[22] Gandhi, R. (2018). Naive Bayes Classifier. *Towards Data Science*, Retrieved from https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c

[23] Yadav, A. (2018). Support Vector Machines (SVM). *Towards Data Science*. Retrieved from https://towardsdatascience.com/support-vector-machines-svm-c9ef22815589

[24] Kumar, A. (2021). KNN Algorithm: When? Why? How? *Towards Data Science. Retrieved from* https://towardsdatascience.com/knn-algorithm-what-when-why-how-41405c16c36f

[25] Freund, Y., & Schapire, R.E. (1996). Experiments with a New Boosting Algorithm. *ICML*. Retrieved from https://www.semanticscholar.org/paper/Experiments-with-a-New-Boosting-Algorithm-Freund-Schapire/68c1bfe375dde46777fe1ac8f3636fb651e3f0f8#paper-header

[26] Classification | Machine Learning |. (2022). *Google Developers*. Retrieved from https://developers.google.com/machine-learning/crash-course/classification/video-lecture

[27] Vezhnevets, Alexander & Vezhnevets, Vladimir. (2005). 'Modest AdaBoost' - Teaching AdaBoost to Generalize Better. *Graphicon*. Retrieved from https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.64.2346&rep=rep1&type=pdf

[28] More, A.S., & Rana, D.P. (2017). Review of random forest classification techniques to resolve data imbalance. *2017 1st International Conference on Intelligent Systems and Information Management (ICISIM)*, 72– 78. https://doi.org/10.1109/ICISIM.2017.8122151

[29] Oleinik, A. (2019). What are neural networks not good at? On artificial creativity. *Big Data & Society*. https://doi.org/10.1177/2053951719839433

[30] Imandoust, S.B., & Bolandraftar, M. (2013). Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: *Theoretical Background*. Retrieved from https://www.ijera.com/papers/Vol3_issue5/DI35605610.pdf

[31] Gardner, A. M., Anderson, T. K., Hamer, G. L., Johnson, D. E., Varela, K. E., Walker, E. D., & Ruiz, M. O. (2013). Terrestrial vegetation and aquatic chemistry influence larval mosquito abundance in catch basins, Chicago, USA. *Parasites & Vectors* (Vol. 6, Issue 1). https://doi.org/10.1186/1756-3305-6-9

[32] Schelske, C. L., & Stoermer, E. F. (1971). Eutrophication, Silica Depletion, and Predicted Changes in Algal Quality in Lake Michigan. In Science (Vol. 173, Issue 3995, pp. 423–424). *American Association for the Advancement of Science (AAAS)*. https://doi.org/10.1126/science.173.3995.423

[33] Schrama, M., Gorsich, E. E., Hunting, E. R., Barmentlo, S. H., Beechler, B., & van Bodegom, P. M. (2018). Eutrophication and predator presence overrule the effects of temperature on mosquito survival and development. *P. Mireji (Ed.), PLOS Neglected Tropical Diseases* (Vol. 12, Issue 3, p. e0006354). https://doi.org/10.1371/journal.pntd.0006354

IVSS BADGES

A. I am a Data Scientist

The machine learning models in this study utilized a dataset created by us (the authors) which was compiled from a variety of data sources including the City of Chicago Data Portal, GLOBE, and GIOVANNI. We discuss the limitations of the data in our discussion (IV) as the GIOVANNI data was missing values and did not include errors and mosquito measurements might be influenced by confounding variables that are not controlled for as a result of using public data instead of collecting our own data. The data was also limited in terms of availability and reliability. Many data options were incompatible with our study such as climate data being recorded every 8-days instead of 7-days, which led us to averaging the daily values of a climate variable each week. In addition, this study uses the data with machine learning models to make inferences (predictions) about mosquito abundance events in the future. These inferences are made at a high accuracy and various standard classification metrics and statistical concepts are used to evaluate each model's performance in predicting mosquito abundance in Chicago. Our data analysis aimed to solve the problem of mosquito-borne disease outbreaks in urban areas as predicting mosquito abundance and thus enacting preventative measures can stop the spread of mosquito-borne diseases.

B. I am a Collaborator

As we (the authors) come from completely different backgrounds and parts of the world spanning three different time zones, we each brought our own skills which were vital to completing this project. This project required the integration of many skills including machine learning, mosquitoes, data analysis, literature review, scientific writing, and climatology. S.D. has a background in computer science and machine learning, so he used his knowledge to clean the data into a usable format and develop the machine learning models and their pipelines. D.L. has a background in Earth science data, so she was able to obtain and clean the mosquito data as well as climate data. A.M. has a background in scientific writing and was able to research the background and aide in the formatting of the report. G.V. has a background in graphic design and created graphics for the report and designed the presentation. All authors had experience with writing in general and wrote and gave feedback on the writing of the report. Without collaborative effort, this study would not have been possible as no one person had the background to complete this entire project. Furthermore, working as a team allowed us to get crucial feedback to improve all aspects of the study and report. With diverse backgrounds from schools across the Americas, as a team, we were able to develop creative solutions that we encountered throughout the research process.

C. I Make an Impact

All of us (the authors) come from humid climates near urban areas. This means that mosquito-borne diseases are a local issue in all four of our communities. There is a need for prediction of mosquito abundance to enact preventative measures to inhibit the spread of disease. However, not all communities have the resources nor the infrastructure to enact enhanced methods of mosquito prevention without prediction. Our results utilize readily available remote sensing climate data to predict mosquito abundance. This can act as a cost-effective way to predict mosquito abundance and precent disease spread for communities who cannot afford them. Though our results focused on Chicago, our study could be applied to our local communities once mosquito data is available. We plan to share our research with communities who can utilize them and take effective preventative measures to stop the spread of mosquito-borne diseases.