sensors

MDPI

Article

MOSQUITO EDGE: An Edge-Intelligent Real-Time Mosquito Threat Prediction Using an IoT-Enabled Hardware System

Shyam Polineni [1], Om Shastri [1], Avi Bagchi [1], Govind Gnanakumar [1], Sujay Rasamsetti [1] and Prabha Sundaravadivel [2,*]

[1] STEM Enhancement in Earth Sciences, NASA Center for Space Research, Austin, TX 78723, USA; polineni.shyam@gmail.com (S.P.); omshastri@gmail.com (O.S.); avibagchi32@gmail.com (A.B.); govind.gnanakumar@gmail.com (G.G.); mosquitoedge@gmail.com (S.R.)
[2] Department of Electrical Engineering, The University of Texas at Tyler, Tyler, TX 75702, USA
* Correspondence: psundaravadivel@uttyler.edu; Tel.: +1-903-566-6118

Abstract: Species distribution models (SDMs) that use climate variables to make binary predictions are effective tools for niche prediction in current and future climate scenarios. In this study, a Hutchinson hypervolume is defined with temperature, humidity, air pressure, precipitation, and cloud cover climate vectors collected from the National Oceanic and Atmospheric Administration (NOAA) that were matched to mosquito presence and absence points extracted from NASA's citizen science platform called GLOBE Observer and the National Ecological Observatory Network. An 86% accurate Random Forest model that operates on binary classification was created to predict mosquito threat. Given a location and date input, the model produces a threat level based on the number of decision trees that vote for a presence label. The feature importance chart and regression show a positive, linear correlation between humidity and mosquito threat and between temperature and mosquito threat below a threshold of 28 °C. In accordance with the statistical analysis and ecological wisdom, high threat clusters in warm, humid regions and low threat clusters in cold, dry regions were found. With the model running on the cloud and within ArcGIS Dashboard, accurate and granular real-time threat level predictions can be made at any latitude and longitude. A device leveraging Global Positioning System (GPS) smartphone technology and the Internet of Things (IoT) to collect and analyze data on the edge was developed. The data from the edge device along with its respective date and location collected are automatically inputted into the aforementioned Random Forest model to provide users with a real-time threat level prediction. This inexpensive hardware can be used in developing countries that are threatened by vector-borne diseases or in remote areas without cloud connectivity. Such devices can be linked with citizen science mosquito data platforms to build training datasets for machine learning based SDMs.

Keywords: mosquito prediction; edge sensing frameworks; citizen science

Citation: Polineni, S.; Shastri, O.; Bagchi, A.; Gnanakumar, G.; Rasamsetti, S.; Sundaravadivel, P. MOSQUITO EDGE: An Edge-Intelligent Real-Time Mosquito Threat Prediction Using an IoT-Enabled Hardware System. *Sensors* **2022**, *22*, 695. https://doi.org/10.3390/s22020695

Academic Editors: Catia Prandi, Pietro Manzoni and Ruidong Li

Received: 22 November 2021
Accepted: 7 January 2022
Published: 17 January 2022

1. Introduction

Mosquitoes are one of the world's most dangerous organisms, spreading deadly diseases like malaria, Dengue, and Zika. They have spread to nearly every continent and are further increasing their range due to the extreme weather conditions caused by climate change [1]. The ability to identify mosquito hotspots (areas of high probability of mosquito presence) can be especially valuable in preventing the spread of mosquitoes and the diseases they carry [2].

Due to the information revolution, we are now capable of custom manufacturing circuit boards (PCBs), i.e., more specialized boards akin to Raspberry Pi and Arduino that are well suited for certain endeavors over others. Compared to their satellite counterparts, such edge devices are significantly less expensive both to build and deploy. This makes them ideal for use in areas in which there is minimal existing infrastructure. Rather than

deploying a network of satellites, we simply deploy a network of edge devices, collecting massive troves of data on their localized climates using external sensors. With such large amounts of data, we are capable of making real time predictions of greater accuracy.

The notion that mosquitoes pose a significant threat to humans across the world is the primary basis from which the underlying question of this study arises: How can climate and citizen science mosquito data be used to develop a machine learning algorithm that can predict mosquito hotspots? This question aims to shed light on a new perspective towards understanding and predicting potential mosquito hotspots by analyzing climate variables (such as temperature, climate cover, precipitation, humidity, and water vapor) as key determinants of species' habitats. This technique combines older work done with species distribution models with Random Forest methods to increase accuracy. If an accurate mosquito threat prediction model is created and deployed on the edge at a low cost using completely autonomous technology, underdeveloped areas can be more easily prepared for mosquito-borne disease outbreaks. Our device would be able to connect with other devices (when possible), to form a highly linked system with a small energy footprint and high predictive capability. More importantly, this end-to-end process is built around consideration for the user and provides a reliable way for any individual to understand the mosquito threat near them, as well as the sensor data used to determine that threat. Not only are users capable of accessing that sensor data, but they are also able to modify and manually verify it as necessary. Figure 1 provides the overview of the proposed methodology.

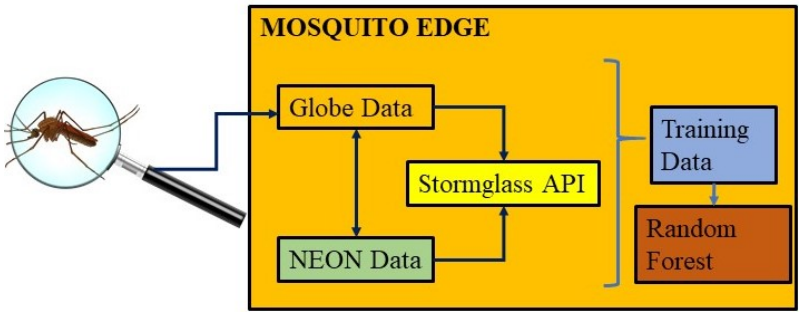

Figure 1. Overview of the proposed method in MOSQUITO EDGE framework.

2. Literature Review

2.1. Species Distribution Models

Species distribution models (SDMs) use climate variables and species observations to project the niche of an organism [3]. In the onset of the centralization of ecological data, species data are often used as a binary "presence" and "absence" variable for efficiency and ease of analysis [4]. This has popularized the use of AI-powered binary classification models in niche prediction. Popular models include CART, logistic regression, neural networks, Naive Bayes, and Maxent [5]. However, an approach with the label as a categorical variable does present challenges, namely class imbalance [6]. Without a large amount of high quality absence data collected in the field, pseudo-absence points generated in the ecological background often outnumber the observed species presence points [7]. In classification models, this imbalance can skew predictions in favor of the most populous label [8]. In this study, the problem of class imbalance was avoided in two key ways. Firstly, quality absence data were allocated in an amount equal to the amount of presence data. Secondly, an edge computing device whose role is to identify and collect species data can balance observed presences with an equal number of pseudo-absences or even true absence observations taken at regular intervals. With the preliminary issue of class imbalance resolved, the second conundrum becomes choosing the ideal model. The two primary contenders are Maxent and Random Forests as deep learning is an area that has not been proven to be successful in niche projects [9]. Maxent is a maximum-entropy based model that is advantageous in that it only needs presence data and can adjust to continuous

labels [10]. It generates pseudo-absences in a customized environmental background, allowing for efficient and often accurate projections free of class imbalance [11]. Random Forests, on the other hand, is sensitive to class imbalance [12] given its ensemble methods, but due to the aforementioned resolution of the class imbalance issue, this drawback need not be considered. Past literature has indicated that Random Forests is the most accurate in niche modeling because this ensemble approach enhances prediction power [13]. Given Random Forest's compatibility with edge computing in the form of TensorFlow Lite, and in consideration of Maxent's inability to be easily harnessed on external hardware, a Random Forests species distribution model was used.

To make use of these species distribution models, newer computing paradigms can be used for data collection and the running of the models. This way, niche prediction can be performed with low latency and high scalability through extending cloud computing abilities to allow computation to occur beyond the central network [14]. These computing paradigms in addition are stored through a diverse set of devices allowing us to use an Internet of Things (IoT) device capable of running SDMs. Citizen science has played a significant role in climate data collection and remote sensing by increasing the amount of locations and sources in which data are collected [15]. The crowdsensing of this collection has been made possible through the IoT service of mobile crowdsensing (MCS), which uses cloud centered architectures [16]. As this produces copious amounts of network traffic, mobile edge computing (MEC) has been suggested as a replacement that solves this problem by functioning on the network edge [16]. A form of crowdsensing combining elements of MEC and MCS is most beneficial for larger scale crowdsensing, such as global users inputting data regarding climate variables.

2.2. Edge Computing and Its Significance

As a result of a general increase in artificial intelligence services in use today, edge computing has been identified as a promising solution that shifts AI model training and prediction to the network edge [17]. Edge intelligence increases the operational efficiency of predictive decision-making and achieves low-latency data processing through shifting the computing capability to the network edge [17]. If a Hutchinson approach is to be used, a large database of climate variable values is required. Using edge intelligence over traditional data collection from a cloud data center can consume less bandwidth and make niche prediction perform at a higher level. In addition, energy consumed by the IoT devices is reduced by compressing sensed data and sending it to the edge [18].

2.3. AI for Climate Prediction

The use of climate variables and data in artificial intelligence can be extremely beneficial for a multitude of applications. From machine learning models that can perform disease prediction to habitat detection, several deep learning techniques have been utilized to advance the understanding of certain climate phenomena. For example, when detecting drought stress, computer vision has been found to be a useful machine learning tool [19]. Time-lapse image sequences of crops are combined with a transfer learning classification technique in order to perform image detection and classification [19]. When determining the climate variables to be used in a machine learning model, many factors must be taken into account. Historical data often show periods of disruption in the form of sudden extreme weather or natural disasters that could invalidate the relationship between these variables [20]. Therefore, certain care must be taken to accurately sample historical data. In order to account for drastic climate changes, there have been certain popular climate modeling tools for time series data. The auto-regressive integrated moving average model in particular has been extremely beneficial for analyzing time series data, with a common application being a forecasting model that utilizes several climate variables [21]. Along with this, researchers have used several other machine learning techniques such as Random Forest Regression, Support Vector Machines, and Hidden Markov models in order to predict the climate variables themselves [21]. This could be used in our study to identify

the most important features for detection of mosquito presence. The Random Forest algorithm in particular has been used in several climate-based applications. This supervised learning makes use of a large amount of decision trees to perform both classification and regression [22]. In addition to predicting certain environments and performing forecast analysis, the use of climate variables in artificial intelligence can also be vital in predicting trends of diseases based on how they are transmitted. For example, research conducted in Malaysia utilized a k-means clustering algorithm in order to determine relationships between temperature and humidity and the spread of dengue [23].

To make niche predictions using data from several different geographical locations, data collection from a network of IoT devices must be optimized. IoT data collection methods can be reviewed based on the trade-offs between data accuracy and frequency of measurement requests [24]. Beyond the traditional concurrent data collection, trees are used in an updated single-hop network structure that efficiently determines how many time-slots are required for concurrent data collection [25]. This novel network framework works under the assumption that several IoT devices are "neighboring" one another and thus can be used to scale up research involving several geographic locations.

2.4. Internet of Things for Citizen Science

Utilizing Internet of Things (IoT) and edge computing presents a promising avenue to collect data effectively. Moreover, in utilizing edge computing over the cloud, information collected from citizen scientists can be processed and transferred far more efficiently. Furthermore, research done on utilizing citizen science data with an edge device to map information proved that such networks "have the potential to transform the roles of citizens" [26]. By utilizing citizen science data in tandem with edge computing, there are new avenues that can be opened for the mapping and tracking of data real time. For example, crowdsourcing data can be effectively transferred over time with the use of edge computing with faster relay speeds [27]. On a similar note, the developments within IoT are being used to help shape citizens to be more digitally apt, resulting in a smarter city overall. IoT has "proven to better sustain the sensing and connectivity, becoming thus an efficient tool to address a broader range of qualitative public services. Moreover, the innate versatility that citizen scientists provide can allow "for the field assessment of any events which makes citizens a more reliable source of information" [28]. Beyond urban cities and smart citizen implementation, IoT and citizen science data can be implemented in rural areas, where mosquitoes are most prevalent. Edge computing is a cost effective and user-friendly solution that can allow individuals experiencing the digital divide to interact with complex datasets [29]. Although citizen science and IoT has its benefits, it is important to consider "data protocol procedures" regarding privacy and analyze "user inputs to ensure data quality" [30]. In short, fully harnessing the potential of citizen scientists is a vital step towards the successful implementation of broader edge computing and IoT solutions.

2.5. Deep Learning Models in Species Classification

Remote sensing IoT technology will help predict mosquito outbreaks and prevent some of the million cases of mosquito-borne disease every year. An existing computer vision-enabled IoT device films mosquitoes and uses convolutional neural networks to identify Aedes and Culex mosquitoes [31]. While it exhibits high accuracy, it is limited to classifying mosquitoes between two species and does not account for overlapping object detection [31]. This feature can be improved upon in future work to include better edge computing and classification and more features can be developed to better predict areas of high mosquito density.

3. Methodology

3.1. Data Extraction

To construct the training dataset for an empirical species distribution model, both mosquito and climate data are needed. GLOBE Observer (an app under the oversight

of the National Aeronautics and Space Administration) provides useful mosquito larvae abundance data obtained from mosquito traps built by citizen scientists. Location, date, and larvae abundance were extracted into a Pandas DataFrame through the GLOBE Observer API for Python [32]. Other variables concerning the mosquito trap itself were not considered in this study. Since our model is rooted on binary classification, the larvae abundance variable was converted into a binary, categorical variable. Sites with larvae abundances exceeding 25 larvae were classified as "infested" and assigned the location a label of "1". Locations with larvae abundances below 25 were assigned "0". About 2000 mosquito observations (latitude, longitude, date, and larvae abundance per observation) were exported.

To reinforce citizen science data with robust scientific data, mosquito observations sampled from carbon dioxide traps were appended to the GLOBE Observer data [33]. Technicians from the National Science Foundation's National Ecological Observatory Network (NEON) collect data from these traps, providing us with reliable and abundant data. Once again, latitude, longitude, and date were recorded. NEON's mosquito data were already binary (presence or absence), so no threshold was needed to record these mosquito observations. With the mosquito data stored, climate data at the locations and dates with mosquito observations were acquired.

The climate attributes used to develop the model were air temperature, humidity, cloud cover, pressure, and precipitation. Initially, an attempt was made to extract raw climate data from the SentinelHub API [34], which provides near real-time data from various satellites, e.g., Sentinels, Landsats. However, the low-level nature of the API made extracting information from raw pixel data time-consuming and error prone. Thus, another weather API—Storm Glass [35]—which provided both real time and historical data for a given point and time from various meteorological services (National Oceanic and Atmospheric Administration, Deutscher Wetterdienst, Météo-France, etc.) was identified. The API for the associated climate attributes at every point and time that was extracted earlier from the GLOBE and NEON datasets was queried. Storm Glass contained missing data for some required attributes at certain points and times—such points from our training set were removed. After deleting records with data gaps, the training set was composed of 15,838 mosquito observations from NEON and GLOBE Observer observations with associated climate data.

3.2. Machine Learning Approach

A decision tree is a technique used in supervised machine learning in which a label is generated based on decisions made at branches extending from nodes. Random Forest builds on this basic technique using the "Wisdom of the Crowd", the idea that an aggregated prediction tends to be more accurate than an individual prediction. The use of several rather than individual predictors is called Ensemble Learning with an ensemble of decision trees being a Random Forest [36].

The process begins with a training dataset—the data that will be used to build the model. Then, in a process called bagging or bootstrap aggregation, the training dataset is randomly sampled with replacement and distributed to different predictors. In a Random Forest model, these predictors are decision trees. Each decision tree will be trained on their respective subset of the training data. Once trained, labels will be associated with certain combinations of decisions. Therefore, when new data arrives at the decision trees (usually from a user input), the model has already learned what label is associated with what combination of decisions, and a predicted label can be outputted for the new data based on what decisions are made for that data. Since Random Forest is based on a "majority vote" system, whatever value the majority of decision trees outputs is the value that is outputted by the model. In this study, binary classification will be used since the model will attempt to classify the data into two groups: presence and absence. For example, if absence is defined by 0 and presence by 1, and 11 out of 20 decision trees output a 0, the

model output would be 0. A mathematical representation of this process of class prediction can be seen in Equation (1):

$$f_{i_i} = \frac{\sum j : node\ j\ splits\ on\ feature\ i\ n_j}{\sum k \in\ all\ nodes\ ni_k} \tag{1}$$

The model learns the splits (threshold values for features that break a node into two branches) that yield a homogeneous (in terms of the label) subset of the training data. The dominant label at the end node will be the output when these combinations of decisions are made again following a user input. A model must also be able to split the data and determine the "quality" of that split. In binary classification, the optimum split would exactly divide one class from the other. To determine the quality of a split, Gini impurity is typically used in Random Forest.

Gini impurity can be defined more simply as the probability of incorrectly classifying an element. Gini values range from 0 to 0.5 and are calculated as shown in Equation (2). For example, a Gini impurity value of 0.5 says that the probability of incorrectly classifying an element would be 50%—the model would simply be "guessing" between 0 and 1. To avoid "guessing" and to maximize the assurance that the model's prediction is correct, a decision tree in a Random Forest will continue to split the data until Gini impurity is as close as possible to 0. These procedures allow feature (e.g., climate variable) importance to be calculated:

$$\begin{aligned} G &= 1 - \sum_{i=1}^{c} p(i)^2 \\ C &= \text{Number of classes}, \\ p(i) &= \text{Probability of selecting class } i, \\ G &= \text{Gini impurity} \end{aligned} \tag{2}$$

By determining how much each feature reduced Gini impurity on average, the most influential feature can be found [37]. Feature importance is calculated as shown in Equation (3) to determine which climate variable had the most influence on a mosquito presence or absence prediction:

$$F(x) = \frac{1}{J}\sum_{j=1}^{J} c_{j_{full}} + \sum_{k=1}^{K}\left(\frac{1}{J}\sum_{j=1}^{J} \text{contribution}_{\,j}(x,k)\right) \tag{3}$$

This is based on Hutchinson's niche concept which states that environmental factors determine the space in which a species can live perpetually [38].

Matrix A in Equation (4) represents a set of linearly independent row vectors (an axis) that spans a space in which the ecological niche is defined. Another matrix will define the space where the species cannot survive. The Random Forest model attempts to determine the thresholds that divide the spaces defined by the matrix and the "absence matrix." Based on these thresholds, the model will be able to predict whether a mosquito will be present or absent given certain climate inputs. However, for a more convenient user experience, the user does not have to provide the climate variable values. Instead, the user enters latitude and longitude coordinates. Corresponding Storm Glass data are then fed into the model and a prediction generated.

A row vector was assigned a 0 if it represented an absence point and 1 if it represented a presence point. The row vectors are also organized by latitude and longitude, but since location was not a variable in the model, it is excluded from the matrix above. The Random Forest algorithm itself was written in Python using the scikit-learn library. One thousand decision trees were generated with 90% of data to be used for training and 10% for testing. The model was trained to output a 0 or a 1 (two classes). All other parameters were set to their default value (the default parameter is Gini impurity rather than entropy for this library):

$$A = \begin{bmatrix} a_1 & b_1 & c_1 & d_1 & e_1 \\ a_2 & b_2 & c_2 & d_2 & e_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_n & b_n & c_n & d_n & e_n \end{bmatrix} \overrightarrow{Y_1} = \{a_1, b_1, c_1, d_1, e_1\} \text{ defines one mosquito observation}$$

$$\begin{aligned} \overrightarrow{C_1} &= \{a_1, a_2, a_3 \cdots a_n\} \text{ defines temperature} \\ \overrightarrow{C_2} &= \{b_1, b_2, b_3 \cdots b_n\} \text{ defines humidity} \\ \overrightarrow{C_3} &= \{c_1, c_2, c_3 \cdots c_n\} \text{ defines air pressure} \\ \overrightarrow{C_4} &= \{d_1, d_2, d_3 \cdots d_n\} \text{ defines precipitation} \\ \overrightarrow{C_5} &= \{e_1, e_2, e_3 \cdots e_n\} \text{ defines cloud cover} \end{aligned} \quad (4)$$

3.3. Edge Computing

An IoT device capable of collecting localized sensor data and outputting a threat level entirely on the edge was configured as a potential solution for areas with little cloud connectivity. GPS data from a smartphone were used to extract latitude and longitude data, which was then sent to the microcomputer via Bluetooth/USB C. To analyze and compute on the edge in a cost-effective manner, a relatively high-performance Raspberry Pi 4 Model B was chosen.

To collect sensor data, the prebuilt Grove Smart Agriculture Kit from the agriculture industry was leveraged as using existing tools allows for greater flexibility and integration with preexisting frameworks. The kit provided access to several external Grove sensors that were assembled onto the Raspberry Pi 4 Model B as seen in Figure 2. These sensors collect air temperature, humidity, soil moisture, soil temperature, UV light, IR light, and visible light data.

This climate data as well as the exact date and GPS coordinates of its collection are recorded by the device. The data are then used as input to run a TensorFlow Lite model (a low footprint format) on the device, removing the need for a PC entirely and truly moving our device to the edge. This localization of data collection and analysis allows for increased speed and removes the need for Wi-Fi access.

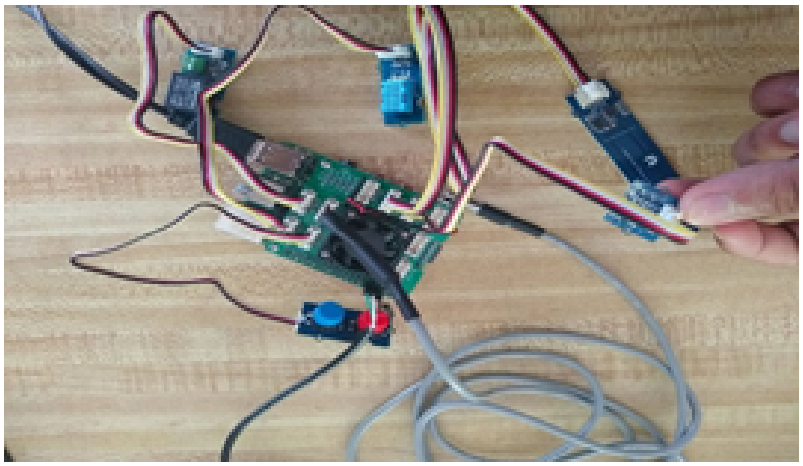

Figure 2. Mosquito edge framework configured with sensors and PC.

4. Results

The Hutchinson matrix that defines the ecological niche of a mosquito was implemented as the training data for the model. This is shown in Table 1.

Table 1. First 5 rows of training data.

	Temp (°C)	Rel. Humidity (%)	Pressure (hPa)	Precipitation (kg/m^3)	Cloud Cover (%)	Presence
0	15.67	87.87	1023.32	0.48	96.67	1
1	15.67	87.87	1023.32	0.48	96.67	1
2	15.67	87.87	1023.32	0.48	96.67	1
3	16.19	83.03	1024.08	0.24	97.33	1
4	16.19	83.03	1024.08	0.24	97.33	1

Then, the training dataset was loaded onto the Raspberry Pi 4 as a CSV file. Sensors that were used include a temperature and humidity sensor, and GPS coordinates were sent from a smartphone via USB C. When the model was run on the edge in Los Gatos, CA, with the inputs being the data collected from the climate sensors attached to the device, a threat level of 42.5% was successfully outputted with an accuracy was 86% as shown in Table 2. This is very accurate considering the vast number of variables that cannot be accounted for in an ecological niche model. The confusion matrix supports this accuracy rating—there is a high number of true positives (772) and true negatives (588) while there is a relatively small number of false positives (108) and false negatives (116).

Table 2. Random Forest model performance.

	Precision	Recall	F1-Score	Support
0	0.87	0.88	0.87	888
1	0.84	0.84	0.84	784
acccuracy			0.86	1584
macro avg	0.86	0.86	0.86	1584
weighted avg	0.86	0.86	0.86	1584

The relative feature importance was calculated for all five climate variables and can be seen in Table 3. The data suggest that temperature has the strongest association with threat level followed by humidity, pressure, cloud cover, and precipitation in order of weakening association.

Table 3. Random Forest model feature importance.

Feature	Score
Temperature	0.33783
Humidity	0.21056
Pressure	0.18738
Precipitation	0.12785
Cloud Cover	0.13718

Another objective of the Random Forest model was to produce a "threat level" or the probability of a mosquito presence point. This number can be attained by calculating the number of decision trees out of the 1000 in the forest that voted for a presence ("1") based on the user location input and its corresponding climate variables. This model can be run for any year after 2017 (limited by a lack of historical data in the Storm Glass API, which will soon be remedied) up until the present and at any location on Earth. After running the model hundreds of times, aggregate threat outputs were correlated with the climate variables.

In Figure 3, temperature vs. threat level (probability of a presence point) is plotted for the 2021 data. The relationship between these two variables is nonlinear. As temperature increases, threat level increases as well. However, past a threshold value of about 28 degrees Celsius, the threat rapidly decreases. This is most likely due to mosquitoes' aversion to very hot, dry climates such as the Sahara region in favor of fairly hot, humid regions. Once again, the similarities between the relationships seen in our model and ecological phenomena is a testament to the model's high accuracy.

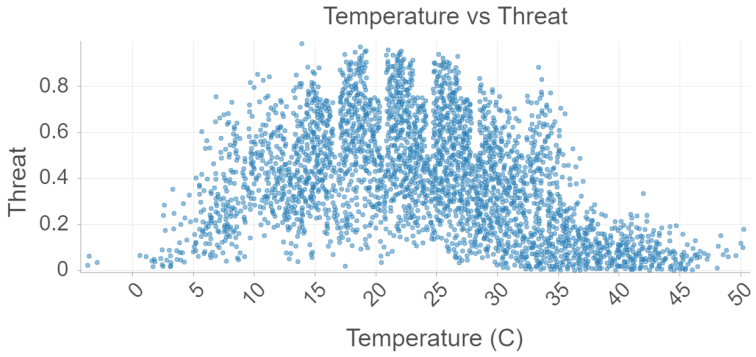

Figure 3. Temperature and Threat analysis using Random Forest.

Figure 4, humidity vs. threat level is plotted for the 2021 data. There is a moderate, linear association between humidity and threat. This is consistent with Table 3 which suggests a moderate association and the ecological literature which suggest that mosquitoes prefer humid climates.

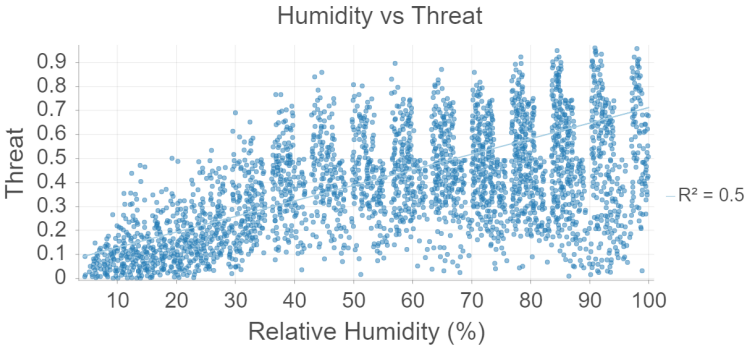

Figure 4. Humidity and Threat analysis using Random Forest.

Figure 5, pressure vs. threat level is plotted for the 2021 data. There is a weak, linear association between pressure and threat. However, this finding may be perhaps insignificant since there is no ecological support for why mosquitoes would prefer regions with high air pressure.

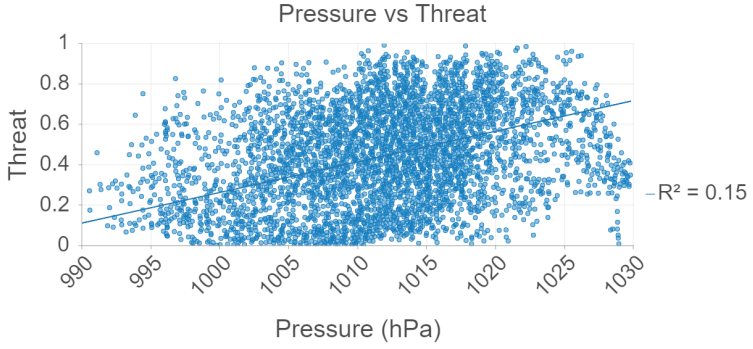

Figure 5. Pressure and Threat analysis using Random Forest.

There is no association between precipitation and threat as seen in Figure 6. This aligns with Table 3, which suggests a weak correlation. This is also further support to the belief that mosquitoes are attracted to standing water rather than precipitation itself. Future studies should examine the relationship between precipitation and threat days after the precipitation has already occurred.

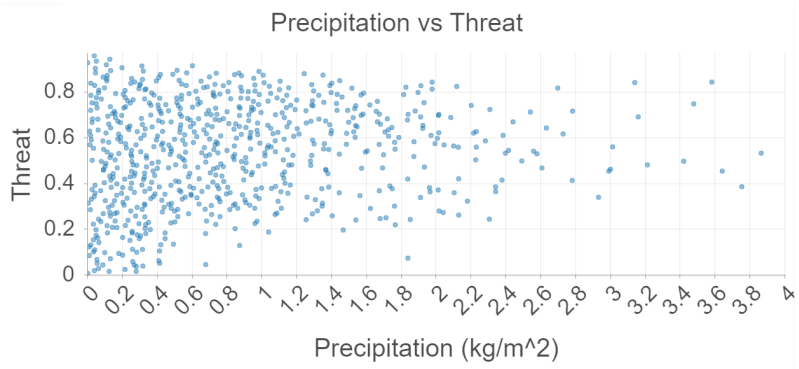

Figure 6. Precipitation and Threat analysis using Random Forest.

Cloud cover is somewhat correlated to temperature and precipitation, but there seems to be no relationship between cloud cover and threat level according to Figure 7.

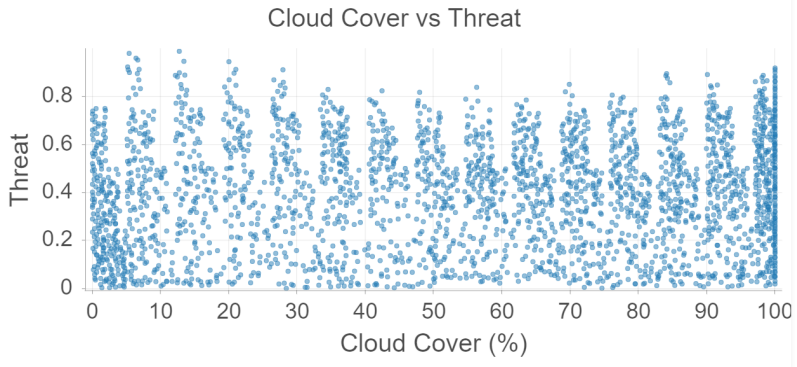

Figure 7. Cloud Cover and Threat analysis using Random Forest.

5. Discussion

A neural network, linear regression, and logistic regression were run on the aforementioned data. The neural network model has an accuracy score of 74% while the logistic regression performs at 67%. A comparable accuracy score cannot be calculated for a continuous target variable, but the linear regression has a coefficient of determination of 0.16, suggesting that the linear model is not accurate nor is it suitable for classification. All three algorithms performed at a lower accuracy than Random Forest. This reinforces past literature, which has suggested that niche prediction is best modeled by an ensemble approach.

Temperature clearly has the most influence on the model output—a result that is consistent with ecological wisdom that suggests mosquito presence and absence is strongly correlated with high and low temperatures respectively. Precipitation has a relatively low impact on the model, which is consistent with the ecological literature that suggests that mosquitoes are attracted to standing water rather than precipitation itself. Mosquitoes prefer humid environments—this is consistent with the relative importance results. We are able to perform this analysis in real time through the running of the edge device model and collecting of local data from the device.

As with any research, our model might have been affected by errors. For one, citizen science data are not as accurate as data collected by trained scientists. As our mosquito dataset did not contain information about specific mosquito species, it is possibly biased

towards the more common mosquito species (e.g., Aedes, Culex, and Anopheles), potentially leading to excessive generalization and low accuracy for predictions in areas with less common mosquitoes with different environmental niches.Furthermore, resource intensive AI implementations on the edge on low power IoT devices can lead to high latency and low accuracy.

While our model may use an ensemble approach similar to that of other mosquito tracking software, our model differs through its use in variables and scale. Specifically, other software focuses singularly on a specific genus (and often species) of a mosquito that is concentrated in a particular area of the world—our software by contrast focuses on all mosquitoes in all parts of the world. Additionally, our software uses precipitation, cloud cover, atmospheric pressure, temperature, and humidity to predict mosquito hotspots, while other software largely uses land cover data coupled with temperature and precipitation. This in part is what leads to a lower accuracy of 86%, whereas other software has a significantly higher accuracy.

Other research papers have acknowledged different variables that can affect certain mosquito populations such as *Culiseta annulata*, *Anopheles claviger* and *Ochlerotatus punctor* [14]. Species of mosquitoes are affected by different combinations and variables' importance to mosquito populations have changed, suggesting that not all mosquito's species are affected equally by changes in certain climate variables. Ultimately, researchers were able to achieve a 99% accuracy model using the random forest framework to predict the spatial distribution of *Anopheles Claviger*.

6. Conclusions

A combination of temperature, humidity, atmospheric pressure, precipitation, and cloud cover data were used to create a model with 86% accuracy. There was a direct correlation between moderate, linear association between humidity and threat, but a weak, linear association between pressure and threat.

Our findings reinforce the current theories that mosquitoes are heavily influenced by their environment, and although there are approximately 3500 species of mosquitoes, all desire a certain range of climate conditions. Our study has taken thousands of data points from NEON, Storm Glass API, and the GLOBE Observer, and coupled with our accuracy, our model makes statistically significant findings.

The accuracy of the model can be increased by adding more environmental features, such as normalized vegetation difference index and human population density. Our model could also become more accurate by tailoring it to specific mosquito species rather than treating all mosquitoes as a single species with the same environmental preferences. This would allow scientists to predict the threat level for a particular species with more confidence, especially for less common species. We are also working on integrating our model with the Find Your Invasive framework.

For the future, there are plans to pair the edge device with a mosquito trap and directly capture mosquito data with computer vision techniques. This would be an autonomous system that can be setup in remote areas without manual oversight, increasing efficiency dramatically.

Author Contributions: Conceptualization, O.S., A.B. and G.G.; methodology, O.S., A.B. and G.G.; edge-computing sensors, O.S.; model development, A.B.; data curation, S.P. (Shyam Polineni), G.G. and S.R.; project administration, S.P. (Shyam Polineni), writing—original draft preparation, S.P. (Shyam Polineni), O.S., A.B., G.G., S.R. and S.P. (Prabha Sundaravadivel); writing—review and editing, S.P. (Prabha Sundaravadivel); supervision, S.P. (Prabha Sundaravadivel). All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported in parts by the National Science Foundation under Grant No. OAC-1924117. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: The authors would like to thank the Earth System Explorers/Mosquito Mappers mentors Peder Nelson, Rusty Low, Cassie Soeffing, Erika Podest, and Becky Boger for their invaluable feedback. The authors would also like to thank NASA for facilitating the SEES program. Furthermore, we would like to thank Prabha Sundaravadivel for her invaluable guidance and mentorship.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ryan, S.J.; Carlson, C.J.; Mordecai, E.A.; Johnson, L.R. Global Expansion and Redistribution of Aedes-Borne Virus Transmission Risk with Climate Change. *PLoS Negl. Trop. Dis.* **2019**, *13*, e0007213. [CrossRef] [PubMed]
2. Davis, J.K.; Vincent, G.; Hildreth, M.B.; Kightlinger, L.; Carlson, C.; Wimberly, M.C. Integrating Environmental Monitoring and Mosquito Surveillance to Predict Vector-Borne Disease: Prospective Forecasts of a West Nile Virus Outbreak. *PLoS Curr.* **2017**, *9*. [CrossRef] [PubMed]
3. Jeschke, J.; Strayer, D. Usefulness of Bioclimatic Models for Studying Climate Change and Invasive Species. *Ann. N. Y. Acad. Sci.* **2008**, *1134*, 1–24. [CrossRef] [PubMed]
4. Liu, C.; White, M.; Newell, G. Measuring and Comparing the Accuracy of Species Distribution Models with Presence-Absence Data. *Ecography* **2010**, *34*, 232–243. [CrossRef]
5. Zhang, J.; Li, S. A Review of Machine Learning Based Species' Distribution Modelling. In Proceedings of the 2017 International Conference on Industrial Informatics-Computing Technology, Intelligent Technology, Industrial Information Integration (ICIICII), Wuhan, China, 2–3 December 2017. [CrossRef]
6. Johnson, R.A.; Chawla, N.V.; Hellmann, J.J. Species Distribution Modeling and Prediction: A Class Imbalance Problem. In Proceedings of the 2012 Conference on Intelligent Data Understanding, Boulder, CO, USA, 24–26 October 2012.
7. Van Der Wal, J.; Shoo, L.P.; Graham, C.; Williams, S.E. Selecting Pseudo-Absence Data for Presence-Only Distribution Modeling: How Far Should You Stray from What You Know? *Ecol. Model.* **2009**, *220*, 589–594. Available online: https://www.sciencedirect.com/science/article/abs/pii/S0304380008005486 (accessed on 15 July 2021). [CrossRef]
8. Lusa, L.; Blagus, R. The Class-Imbalance Problem for High-Dimensional Class Prediction. In Proceedings of the 2012 11th International Conference on Machine Learning and Applications, Boca Raton, FL, USA, 12–15 December 2012. [CrossRef]
9. Kaky, E.; Nolan, V.; Alatawi, A.; Gilbert, F. A Comparison between Ensemble and Maxent Species Distribution Modelling Approaches for Conservation: A Case Study with Egyptian Medicinal Plants. *Ecol. Inform.* **2020**, *60*, 101150. [CrossRef]
10. Kramer-Schadt, S.; Niedballa, J.; Pilgrim, J.D.; Schröder, B.; Lindenborn, J.; Reinfelder, V.; Stillfried, M.; Heckmann, I.; Scharf, A.K.; Augeri, D.M.; et al. The Importance of Correcting for Sampling Bias in Maxent Species Distribution Models. *Divers. Distrib.* **2013**, *19*, 1366–1379. [CrossRef]
11. Azari, A.; Namayanja, J.M.; Kaur, N.; Misal, V.; Shukla, S. Imbalanced Learning in Massive Phishing Datasets. In Proceedings of the 2020 IEEE 6th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing, (HPSC) and IEEE International Conference on Intelligent Data and Security (IDS), Baltimore, MD, USA, 25–27 May 2020. [CrossRef]
12. O'Brien, R.; Ishwaran, H. A Random Forests Quantile Classifier for Class Imbalanced Data. *Pattern Recognit.* **2019**, *90*, 232–249. [CrossRef] [PubMed]
13. Magness, D.R.; Huettmann, F.; Morton, J.M. Using Random Forests to Provide Predicted Species Distribution Maps as a Metric for Ecological Inventory & Monitoring Programs. In *Applications of Computational Intelligence in Biology. Studies in Computational Intelligence*; Smolinski, T.G., Milanova, M.G., Hassanien, A.E., Eds.; Springer: Berlin/Heidelberg, Germany, 2008; Volume 122. [CrossRef]
14. Kalluri, S.; Gilruth, P.; Rogers, D.; Szczur, M. Surveillance of Arthropod Vector-Borne infectious diseases using remote sensing techniques: A review. *PLoS Pathog.* **2007**, *3*, e116. [CrossRef] [PubMed]
15. Bugoro, H.; Hii, J.; Russell, T.L.; Cooper, R.D.; Chan, B.K.; Iro'ofa, C.; Butafa, C.; Apairamo, A.; Bobogare, A.; Chen, C.C. Influence of environmental factors on the abundance of Anopheles farauti larvae in large brackish water streams in Northern Guadalcanal, Solomon Islands. *Malar. J.* **2011**, *10*, 262. [CrossRef] [PubMed]
16. Marjanović, M.; Antonić, A.; Žarko, I.P. Edge Computing Architecture for Mobile Crowdsensing. *IEEE Access* **2018**, *6*, 10662–10674. [CrossRef]
17. Zhang, K.; Leng, S.; He, Y.; Maharjan, S.; Zhang, Y. Mobile Edge Computing and Networking for Green and Low-Latency Internet of Things. *IEEE Commun. Mag.* **2018**, *56*, 39–45. [CrossRef]
18. Sodhro, A.H.; Pirbhulal, S.; de Albuquerque, V.H.C. Artificial Intelligence-Driven Mechanism for Edge Computing-Based Industrial Applications. *IEEE Trans. Ind. Inform.* **2019**, *15*, 4235–4243. [CrossRef]
19. Ramos-Giraldo, P.; Reberg-Horton, C.; Locke, A.M.; Mirsky, S.; Lobaton, E. Drought Stress Detection Using Low-Cost Computer Vision Systems and Machine Learning Techniques. *IT Prof.* **2020**, *22*, 27–29. [CrossRef]

20. Szczupak, J.; Sica, D.; Silva, D.; Pinto, L.; Macedo, L.; Savi, F. AI identification of new Hydro-Climate models. In Proceedings of the 2009 52nd IEEE International Midwest Symposium on Circuits and Systems, Cancun, Mexico, 2–5 August 2009; pp. 901–904. [CrossRef]
21. Teggi, P.P.; Natarajan, S.; Malakreddy, B. Intelligent FORecasting Model for Climate Variations (InFORM): An Urban Climate Case Study. In Proceedings of the 2020 7th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 12–14 March 2020; pp. 130–137. [CrossRef]
22. Geetha, V.; Punitha, A.; Abarna, M.; Akshaya, M.; Illakiya, S.; Janani, A.P. An Effective Crop Prediction Using Random Forest Algorithm. In Proceedings of the 2020 International Conference on System, Computation, Automation and Networking (ICSCAN), Pondicherry, India, 3–4 July 2020; pp. 1–5. [CrossRef]
23. Mathur, N.; Asirvadam, V.S.; Dass, S.C.; Gill, B.S. Visualization of dengue incidences for vulnerability using K-means. In Proceedings of the 2015 IEEE International Conference on Signal and Image Processing Applications (ICSIPA), Kuala Lumpur, Malaysia, 19–21 October 2015; pp. 569–573. [CrossRef]
24. Li, Y.; Chen, Y.; Lan, T.; Venkataramani, G. MobiQoR: Pushing the Envelope of Mobile Edge Computing Via Quality-of-Result Optimization. In Proceedings of the 2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS), Atlanta, GA, USA, 5–8 June 2017; pp. 1261–1270. [CrossRef]
25. Incel, O.D.; Ghosh, A.; Krishnamachari, B.; Chintalapudi, K. Fast Data Collection in Tree-Based Wireless Sensor Networks. *IEEE Trans. Mob. Comput.* **2012**, *11*, 86–99. [CrossRef]
26. Liu, Y.; Piyawongwisal, P.; Handa, S.; Yu, L.; Xu, Y.; Samuel, A. Going Beyond Citizen Data Collection with Mapster: A Mobile+Cloud Real-Time Citizen Science Experiment. In Proceedings of the 2011 IEEE Seventh International Conference on e-Science Workshops, Washington, DC, USA, 5–8 December 2011; pp. 1–6. [CrossRef]
27. Disney, J.; Bailey, D.; Farrell, A.; Taylor, A.; McGreavy, B. Anecdata.org: An online citizen science platform for Building Climate Resilient Communities. In Proceedings of the OCEANS 2018 MTS/IEEE Charleston, Charleston, SC, USA, 22–25 October 2018; pp. 1–4. [CrossRef]
28. Alexandru, A.; Ianculescu, M.; Marinescu, I.A.; Popescu, T.D. Shaping the Digital Citizen into a Smart Citizen on the Basis of IoT Capabilities. In Proceedings of the 2019 22nd International Conference on Control Systems and Computer Science (CSCS), Bucharest, Romania, 28–30 May 2019; pp. 707–714. [CrossRef]
29. Lopez-Novoa, U.; Morgan, J.; Jones, K.; Rana, O.; Edwards, T.; Grigoletto, F. Enabling Citizen Science in Rural Environments with IoT and Mobile Technologies. Available online: https://orca.cardiff.ac.uk/126324/1/CPSS2019_paper_12.pdf (accessed on 5 August 2021).
30. Scheibner, J.; Jobin, A.; Vayena, E. Ethical Issues with Using Internet of Things Devices in Citizen Science Research: A Scoping Review. 14 July 2020. Available online: https://ssrn.com/abstract=3651447 (accessed on 17 July 2021).
31. Casey, A. A Convolutional Neural Network Model for Species Classification of Camera Trap Images. Mathematics Undergraduate Theses. 2018. Available online: https://scholarworks.boisestate.edu/math_undergraduate_theses/8 (accessed on 5 August 2021).
32. goutilsAP I. (n.d.). Available online: https://iges-geospatial.github.io/globe-observer-utils-docs/goutils.html (accessed on 4 July 2021).
33. NEON (National Ecological Observatory Network). Mosquitoes Sampled from CO_2 Traps, RELEASE-2021 (DP1.10043.001). Available online: https://data.neonscience.org/data-products/DP1.10043.001/RELEASE-2021 (accessed on 16 July 2021). [CrossRef]
34. Sentinel Hub. (n.d.). Available online: https://www.sentinel-hub.com/ (accessed on 16 July 2021).
35. Global Weather API. Storm Glass. (n.d.). Available online: https://storm-glass.io/ (accessed on 16 July 2021).
36. Géron, A. *Hands-On Machine Learning with Scikit-Learn and Tensorflow Concepts, Tools, and Techniques to Build Intelligent Systems*; O'Reilly: Springfield, MO, USA, 2019.
37. Lewinson, E. Explaining Feature Importance by Example of a Random Forest. Medium. (17 April 2020). Available online: https://towardsdatascience.com/explaining-feature-importance-by-example-of-a-random-forest-d9166011959e (accessed on 5 August 2021).
38. Hutchinson, E. Concluding Remarks-University of Lausanne. 1957. Retrieved 27 September 2020. Available online: https://www2.unil.ch/biomapper/Download/Hutchinson-CSHSymQunBio-1957.pdf (accessed on 4 July 2021).