Student Checklist (1A)

This form is required for ALL projects.

1. a. Student/Team Leader: Rithika Narayan Grade: 11

   Email: narayanrithika@gmail.com Phone: 631-601-5476

   b. Team Member: _____ c. Team Member: _____

2. Title of Project:

   Coral Grief: Machine Learning on Crowd-sourced Data to Highlight an Ecological Crisis

3. School: Elwood John Glenn High School School Phone: 631-266-5410

   School Address: 478 Elwood Rd, Elwood, NY 11731

4. Adult Sponsor: Anthony Pellicano Phone/Email: apellicano@angion.com

5. Does this project need SRC/IRB/IACUC or other pre-approval? ☐ Yes ☑ No Tentative start date: _____

6. Is this a continuation/progression from a previous year? ☐ Yes ☑ No

   If Yes:

   a. Attach the previous year's ☐ Abstract **and** ☐ Research Plan/Project Summary

   b. Explain how this project is new and different from previous years on ☐ Continuation/Research Progression Form (7)

7. This year's laboratory experiment/data collection:

   06/26/19 (Actual Start Date: (mm/dd/yy)) 11/19/19 (End Date: (mm/dd/yy))

8. Where will you conduct your experimentation? (check all that apply)

   ☑ Research Institution ☐ School ☐ Field ☑ Home ☐ Other: _____

9. List name and address of all non-home and non-school work site(s):

   Name: Angion Biomedica Corp.

   Address: 51 Charles LIndbergh Boulevard, Uniondale, NY 11553

   Phone/email: 516-326-1200

10. **Complete a Research Plan/Project Summary following the Research Plan/Project Summary instructions and attach to this form.**

11. **An abstract is required for all projects after experimentation.**

"Coral Grief: Machine Learning on Crowd-sourced Data to Highlight an Ecological Crisis"
Rithika Narayan
Earth and Environmental Sciences

a. Rationale

Research has proven time and time again that for the past few decades, coral reefs have been experiencing unprecedented rates of death and disease due to pollution, predation, and warming oceans; few are left in pristine health (Pandolfi et al., 2003). There is much cause for concern and a need for action: coral reefs are the most biodiverse marine environments, housing millions of organisms from thousands of species in just 1% of the Earth's surface, and protect coastal areas from tidal waves and erosion. In addition to providing vast ecological benefits, coral reefs economically support millions of people who rely on the organisms in the reef for food and the beauty of the environment to draw tourists. ("Corals,: n.d.) Machine learning (ML) has already been adapted to measure the area covered by coral in the benthic zone as a measure of reef health. Automating this process allows for the analysis of the millions of images of reefs taken each year in government and private surveys such as the XL Catlin Seaview Survey (Beijbom et al., 2012). Seeing that less than 5% of the images collected of coral reefs are able to be annotated manually, it is important to find efficient ways to thoroughly analyze these images (Beijbom et al., 2012). Benthic cover is not the only method of analyzing reef health. For instance, the amount of coverage says nothing about the state of the corals that are remaining. In the Caribbean, Arabian Sea, Great Barrier Reef, and elsewhere, many corals are affected by common diseases including bleaching, black-, yellow-, and white band disease, dark spot disease, white plague, and white pox ("Common Identified Coral Diseases," n.d.). Many of these diseases are direct results of human actions: bleaching is caused by rises in ocean temperature and white pox is linked to the pollution of the oceans with human fecal matter (National Oceanic and Atmospheric Administration, n.d.; Joyner et al., 2015). Although the causes of some of these diseases are known, researchers are still not able to predict the timing and location of the next disease outbreak or mass bleaching event. A Convolutional Neural Network capable of detecting and identifying disease in corals could be used to annotate the most recent and historical images that there simply is not enough manpower to put human eyes on. These annotations could be compiled to visualize trends in the spread and occurrence of coral disease in order to assess which areas need what resources and to predict the next outbreaks. Furthermore, the use of the algorithm does not have to be limited to academics. Private companies such as National Geographic could run the model through their archives of footage of coral reefs in order to expand the locations and times when data was collected. The public could be encouraged to join the efforts by downloading a mobile app with the model and using it on their vacation photos from scuba diving and such in order to be a part of conservation efforts. The annotations collected could be integrated into NOAA's Deep Sea Coral Data Portal (DSCDP), which houses images of and information about coral found around the world. This technology can also be used in citizen science projects such as the Great Reef Census, which asks volunteers to upload 10 photos of their diving site and its GPS coordinates to be able to track the condition of the reef at various unmonitored locations ("The Great Reef Census," n.d.).

b. Research Question

How can ML algorithms be applied to the detection and identification of coral diseases in images of reefs and individual corals? How can this algorithm be used in order to complement image analysis by human researchers?

Hypothesis/Engineering Goal

A Convolutional Neural Network (CNN) can be trained on images of corals with varying degrees of health in order to create a model that can identify these diseases. This model can then be made into an app where users can upload images of coral that they have taken and the image and its annotations can be compiled in a database to create a historical log of the conditions in a particular reef.

Expected Outcome

The expected outcome is an app containing a model that can be used to analyze images of coral reefs and add the image and the assessment of its health into a database that human researchers can use to identify trends in health of different reefs. Based on these trends, resources for treatment of the diseases can be allocated, protection zones can be set up, and local awareness can be raised to lower risk factors.

c. Procedure

1) Select an ML algorithm to deploy for the model. Literature indicates that a CNN has a high Mathew's Correlation Coefficient (MCC), which is a performance parameter that assesses classifiers, as compared to other algorithms that were applied to coral images (Brown & Dharma, 2018). The Mask RCNN algorithm is an open source CNN that has been successfully used in the analysis of nuclei in microscope images, detection of sports fields in satellite images, and other object detection and segmentation projects. The name of the algorithm refers to the masks that are generated above the classified regions of interest when the model gives its output.
2) Modify the algorithm's source code to best suit the specific qualities of the dataset. This includes adding and changing the classes and their names. Select an initial learning rate that will be updated as training progresses to make the process more effective and tailored to the data.
3) Collect visual data to train and test the algorithm on. Ensure the accuracy of the labeling of the health of the coral in these images by collecting them from peer-reviewed journal articles, government or university databases, and research projects such as the XL Catlin Seaview Survey. There should be hundreds of images with hundreds of instances of each type of disease represented.
4) Select a platform on which to annotate these images. Labelbox is one of many such platforms including LabelMe and the VGG Image Annotator which allow users to upload datasets, create classes and segmentation types with which to label images. For this purpose, the segmentation polygon tool would be best to create masks of the coral that follow the outline of the coral itself as well as the disease lesions. These masks are what the computer will learn from and them emulate.
5) Annotate the images in two groups, one group for training and one for testing. The setup for both groups should be identical in terms of classes and annotation type, but the data split between the two should be 80-20 training to validation, which literature

indicates is an ideal split for midsize data sets ("Pareto principle," n.d.). Use the platform to generate a JSON (Java Script Object Notation) format that contains image paths and the annotations for each image. JSONs are readable by humans and computers alike, making them a viable method for providing data to a model. This JSON will be used by the Mask R-CNN algorithm to train and validate.

6) Start an Amazon Web Services (AWS) EC2 instance to train the model in. These instances are remote computers dedicated to only the customer's purpose and therefore are able to carry out heavy training in a timely manner. Use one that has TensorFlow GPU capabilities in order to increase computing power. GPU stands for the graphic processing unit and when used in conjunction with a computer's CPU, or central processing unit, can make training more efficient. Transfer the training and validation images, the JSON files, and the code to the instance and begin training within a powershell.
7) Save the logs created during the training. Each subsequent training session will begin from the last log such that the model does not have to relearn and instead builds on what it has learned.
8) Monitor the progress of the training using Tensorboard (see data analysis) and the most recent logs. Once the box and class losses for training and validation near zero and accuracy approaches one, stop the training. Avoid overfitting by using image augmentation in the code and by not overtraining; monitor the losses for signs of rapid descent to near zero loss.
9) Once training has been completed, create an interface for the use of the model using the Raspberry Pi 4.

Risk and Safety

There is no risk to any living organisms in this research. All work is done digitally.

Data Analysis

Tensorboard, a feature of Tensorflow, is used to visualize the progress of ML. By inputting the most recent log to Tensorboard, the platform will graph the accuracy and loss rates for the training and validation data sets over time for class loss as well as box loss. The goal for the losses is to be as close to zero without overfitting the model, which means that the program has learned to recognize the classes but only as the appear in the training images rather than being able to be deployed successfully on new images, and the goal for the accuracies to be as close to one as possible without overfitting. Class accuracy reflects how well the model is able to categorize what it sees into the provided classes based on the class annotations made by the researcher on the datasets. Box accuracy reflects how well the program is able to pinpoint the location and boundaries of the object in the image, in other words how well it can segment the images. Tensorboard can be used to decide when to stop training as well as if learning rates need to be changed. If the loss or accuracy rates are stagnant, it can indicate the program has been caught in the loss landscape and needs a new learning rate to ascend out of the loss landscape. Once the model is actually in use, the confidence level it displays can be a useful measure of how well the program is performing. Within the code, the researcher can set a minimum confidence threshold in order for the model to display its results; results below that threshold will be ignored. If confidence is consistently low, it can indicate that more training

needs to be done with a wider set of images, perhaps for a specific disease if that is what shows low confidence most often.

d. Bibliography

Beijbom, O., Edmunds, P. J., Kline, D. I., Mitchell, B. G., & Kriegman, D. (2012). Automated annotation of coral reef survey images. *2012 IEEE Conference on Computer Vision and Pattern Recognition*. doi: 10.1109/cvpr.2012.6247798

Common Identified Coral Diseases. (n.d.). Retrieved June 14, 2019, from http://www.artificialreefs.org/Corals/diseasesfiles/Common Identified Coral Diseases.htm

Corals. (n.d.). Retrieved June 16, 2019, from https://oceanservice.noaa.gov/education/kits/corals/coral07_importance.html

Joyner, J. L., Sutherland, K. P., Kemp, D. W., Berry, B., Griffin, A., Porter, J. W., … Lipp, E. K. (2015). Systematic Analysis of White Pox Disease in Acropora palmata of the Florida Keys and Role of Serratia marcescens. *Applied and Environmental Microbiology*, *81*(13), 4451–4457. doi: 10.1128/aem.00116-15

Mary, N. A. B., & Dharma, D. (2018). A novel framework for real-time diseased coral reef image classification. *Multimedia Tools and Applications*, *78*(9), 11387–11425. doi: 10.1007/s11042-018-6673-2

National Oceanic and Atmospheric Administration. (2010, March 15). What is coral bleaching? Retrieved June 15, 2019, from https://oceanservice.noaa.gov/facts/coral_bleach.html

Pandolfi, J. M., Bradbury, R. H., & Sala, E. (2003). Global Trajectories of the Long-Term Decline of Coral Reef Ecosystems. *Science*, *301*(5635), 955–958. doi: 10.1126/science.1085706

Pareto principle. (n.d.). Retrieved June 17, 2019, from

https://en.wikipedia.org/wiki/Pareto_principle#In_computing

The Great Reef Census. (n.d.). Retrieved June 17, 2019, from https://census.citizensgbr.org/

Addendum

No changes were made to this research plan.