My Research Background

My research journey began in my sophomore year. In the Summer of my freshman into my sophomore year of high school, I became intrigued with the idea of Artificial Intelligence and, more specifically, the concept of machine learning. During that summer, out of pure curiosity, I read countless scientific research papers on various models of neural networks, most of which were landmark studies. The Summer of 2021 was my formal introduction to research. Although I've conducted "research"-like projects in school and have written scientific reports in science classes, the meticulousness, formality, and rigorousness of scientific research was nothing like I've seen before. Through learning about neural networks, specifically convolutional neural networks (CNN), I discovered that CNNs function just as a human brain does—a form of biomimicry. I decided then that CNNs would be my focus point.

My first scientific research project was a machine learning-assisted cancer diagnosis project the following winter in 2022. The title of the project is *Determining Critical Lung Cancer Subtypes from gigapixel multi-scale whole slide H&E stain images*. A project on using machine learning to assist in the diagnosis of two lung cancer subtypes: lung adenocarcinoma and squamous cell carcinoma. I used four different CNN models (VGG16, ResNet50, DenseNet121and Inceptionv3) accompanied by two different learning methods: transfer learning and fine-tuning. My research procedure in this project involved downloading and processing the cancer image data, training the four models, and applying the two learning methods to each model. My results section comprised a detailed analysis of the performance of my models. This project was the first hands-on experience of machine learning and research. Aside from the novelty of using a gaming PC with 32GB RAM with a GeForce RTX graphic processing unit, I was struck by two discoveries: 1) the staggeringly accurate prediction results and 2) the ability to

which the models learned. In my head, I likened them to a child's mind: as they grow, certain knowledge is taught to them (such as the difference between a pear and an apple), and over time, they would be able to make decisions based on their previous knowledge. However, these models, unlike children, can "grow up" and learn at such alarming rates that they would be able to discern the difference between two lung cancer subtypes on H&E-stained cell image slides in only a few hours. The accuracy of my models in my experiment was between 97-99%. Although my research was vastly primitive in the scope of clinical research or even application, the potential that these models held was obvious. Suppose the machine learning-assisted cancer diagnosis becomes successful, the time it takes to train a doctor or a model would be significantly decreased. Unlike the previous studies on state-of-art models that I had read, I was determined to use CNNs as a tool for application. I participated in the Massachusetts Science and Engineering Fair with this project in March 2022 and then published the paper on ieee.org.

The Tree Project & Procedure

In the following year, my junior year, I decided to start another project using CNNs and machine learning. My inspiration for my project came from an entrepreneurship project from Summer to October of 2022 with a few teammates for a competition. We created a device to catch the spotted lanternflies as they are a deleterious invasive species prevalent in the Mid-Atlantics in the U.S. I was greatly inspired by that project. Since I have some machine learning background from my research project in the previous year, I wanted to initiate a project on monitoring invasive species, specifically the spotted lanternflies. During my initial research stage, I came across a study on using satellite imagery and cancer survey data to predict cancer prevalence in the United States by producing geographical heatmaps, and I was inspired to use satellite imagery to remotely monitor spotted lanternflies and create geographical heatmaps but with

densities of invasive species instead. I quickly discovered during my research preparation that it is very difficult to monitor spotted lanternflies as they move and travel much faster than other species, such as plants. Instead of monitoring the flies themselves, I realized I could monitor something they relied on, such as environmental factors, etc. After more literature review, I discovered that the spotted lanternflies primarily relied on a type of plant named *Ailanthus Altissima* as the host plants, also known as the Tree of Heaven. These trees are also invasive species, and I decided to monitor the spread of the Tree of Heaven using remote sensing combined with satellite imagery.

At that point, I was certain that my research involved using satellite imagery combined with surveys of the Tree of Heaven locations; I planned to use machine learning, more specifically CNN (convolutional neural network), to train the data. I wasn't sure at the beginning of my planning stage how I would proceed with the experiment, specifically with details such as the exact machine learning models, data sources, and how I would apply my trained model to generate geographical density maps. I reached out to my current mentor, Dr. Zhao, and he provided me with guidance on the framework of my experiment. He suggested that I incorporate a novel computer vision model, Vision Transformer (ViT), in my experiment and implement inference testing by using the trained models to perform prediction on areas in the United States. After my mentor provided some guidance in the direction of my study, I conducted the experiment from 01/2023 to 03/2023. I wrote all code used for data processing, image segmentation, data preparation, model training, validation, and inference testing. All equipment used in my research was Python and its various libraries through Jupyter Notebook, which was installed on my personal computer and conducted at home. Examples of libraries I used were Pandas, Numpy, Matplotlib, PyTorch, GeoPandas, etc. When I encountered problems with my

code or procedure, I primarily used the internet to search for solutions, and my mentor occasionally provided me with website links and documentation links as well.

The specific models used in my experiment were three CNN models: ResNet50, VGG16, EfficientNetv2, and one ViT model. I fine-tuned the four models using Google Earth satellite images combined with a set of data on the Tree of Heaven distribution in Italy, a survey result I found on the internet. My study was unique in that not only did I employ a variety of models, but I also utilized a wide range of methods to both train my models and validate my inference testing result in the final stages of my experiment.

I conducted all procedures relating to data collection and preparation. I obtained my data from a literature review and finding available open-source data. I also emailed local and government environmental agencies for data, such as the Massachusetts Department of Agricultural Resources and the New York State Integrated Pest Management Program. Through more literature review, I eventually discovered a study conducted in Alta Murgia National Park located in Italy on the Tree of Heaven. The dataset in Italy included a detailed latitude and longitude location of each Tree of Heaven found within the area, which allowed me to combine Google satellite images with the Tree of Heaven survey points and generate my model training dataset.

For each satellite imagery, I also performed image processing by segmenting each image into smaller grids for CNN training. For each image, I assigned a label of either “1” or “0” for binary classification or the number of trees present in the image for multiclass classification in the later stages of validating our final result. I also downloaded satellite images of New Jersey for the inference testing stage, where I eventually produced geographical heat maps of the density of the Tree of Heaven.

\ The datasets(satellite imagery and survey locations) were prepared and packaged into an HDF5 file and split into training, validation, and testing datasets for the models. The data analysis was conducted through a variety of methods, including binary classification, multiclass classification, and feature extraction with logistic regression. The performance of each model and method was quantified with confusion matrices and ROC curves. Subsequently, I selected the best-performing model to perform inference testing in the United States. The previously trained models were used to predict the density of the Tree of Heaven in each satellite image of New Jersey. The prediction, or inference testing, results were used to graph geological heat maps of the Tree of Heaven density. In order to validate the results of our inference testing prediction, we utilized a different method. By using multiclass classification, we can compare the resulting graphs qualitatively to our previous testing results using binary classification. If significant plotting overlaps could be observed, it could then be inferred that our predictions are validated. I formulated all aspects of the discussions and conclusions for this study. Since there were no validation datasets in the United States to evaluate the density prediction geological heatmaps, I found other datasets, such as New Jersey population density for each municipality and the tree cover map, to attempt to construct a correlation between the distribution of the Tree of Heaven with other environmental factors. By following the procedure and predicting the spread of A. altissima in the U.S.A. using satellite imagery, I verified that the multiclass prediction results in a prediction pattern similar to that of binary classification. Using the same dataset but a different method, I was able to verify the consistency of the models. Although the distribution ranges were different, the correlated densities of the two maps were nearly identical. This finding demonstrates that the models can associate features unique to each satellite image with the absence or presence of trees. I discovered through the second stage of validation that the Tree of

Heaven distribution closely matches areas with high population densities and tree cover in New Jersey.

Advice

I remember in the Zoom meeting in January 2024, after the Regeneron STS Scholar was announced, the director of the program said that as the new generation of young scientists stepped up, the challenges faced by the old generation still persisted. In scientific research, a number of unforeseen challenges and setbacks could surface from any and all aspects of your research. Specific to my project, I was faced with constant code and software malfunctions, a lack of accessible data, etc. And in many cases, doubts and pressures, both internal and external, may serve as an impediment. The director told us that the reason we were present in that Zoom meeting was because we persisted in our belief, and in which case was the purpose of our projects. For all the younger scientists, really believe in your project, believe that what you're doing is right, and the rest will come.

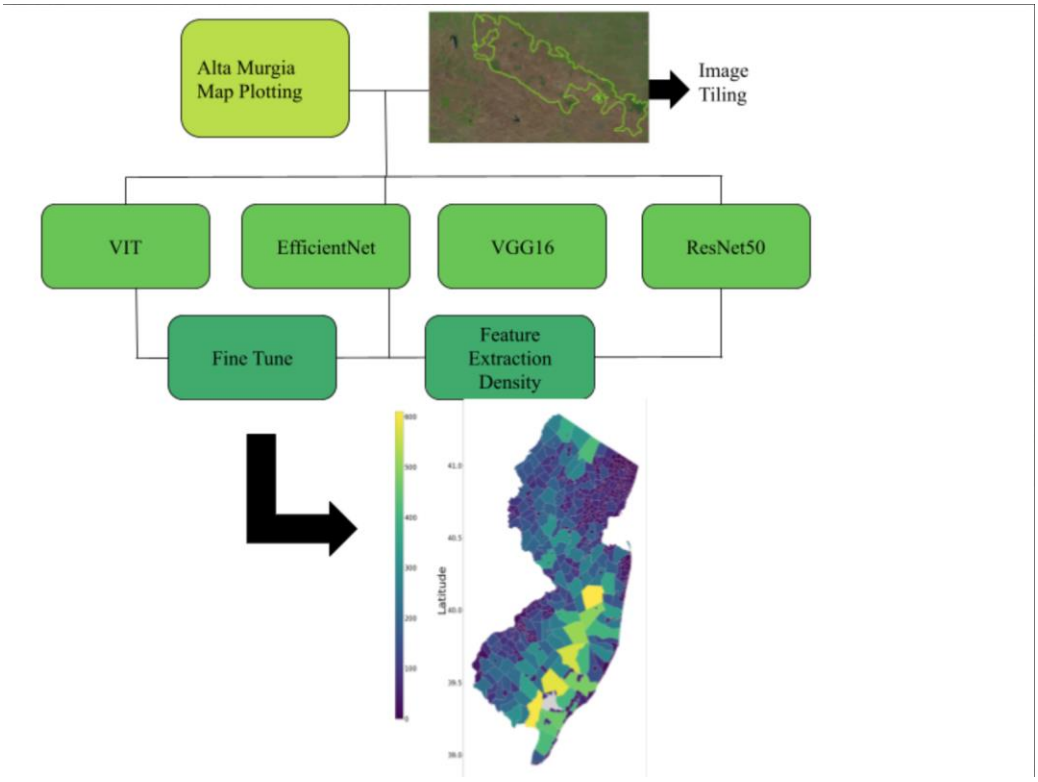
Lay Person's Summary

Ailanthus Altissima, also known as the Tree of Heaven, is an invasive species that causes ecological damage by secreting toxic chemicals into the soil and competing with native species. Aside from ecological damages, it also serves as a host plant for numerous pests and insects, including the spotted lanternfly. The challenge posed by the Tree of Heaven necessitates efforts in management. Through this research study, the Tree of Heaven is monitored remotely using satellite images and machine learning. Due to the absence of studies, surveys, and data on the Tree of Heaven in the United States, a survey dataset conducted in Italy was used to train machine learning models along with satellite imagery to predict the spread of the Tree of Heaven in the United States. Different models and methods were used in order to compare the

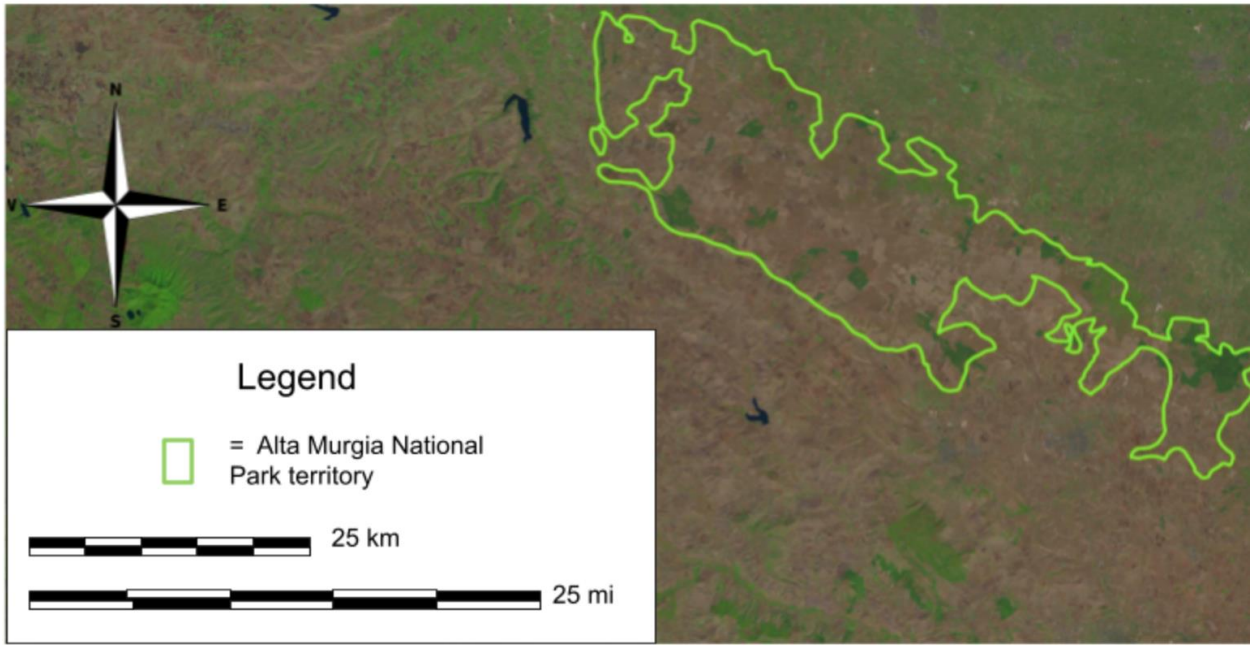
performance and accuracy. The most accurate model was used to create density maps of the Tree of Heaven in the United States. This study is important in the fields of invasive species management and remote sensing because it provides a comprehensive review of using different machine learning models as well as a density map to prevent further survey costs.

I've attached below a series of graphs that were important to my research with respect to each stage of my research:

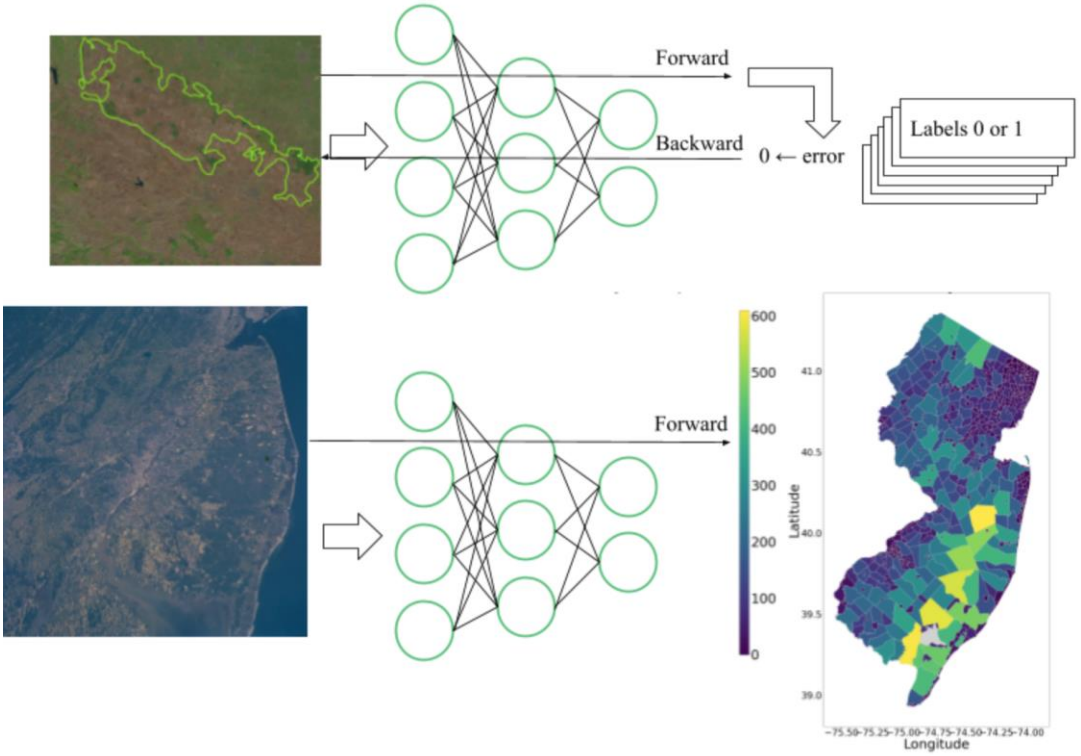
My overarching framework:



Satellite image gathering and tiling:



Research methodology:



Prediction result: