Fall 2022 Capstone Project Progress Report 1

Tree Species Detection

Cyndi (Shengdi) Chen (sc4928), Kenny (Anshuo) Wu (aw3395),

Yunshu Cai (yc4000), Yunze Pan (yp2599), Ziyan Liu(zl3098)

October 22, 2022

1 Introduction

1.1 Abstract

Being able to detect species of trees has several industrial applications, including tracking invasive species, understanding growth rate, above-ground biomass, and encroachment into utility power lines, railway power lines, et cetera. Utility companies and railways will use the output as a reference to remove unhealthy trees from their right of way. Companies can also use tree species data to understand growth patterns in their corridors and plan trimming schedules accordingly. Due to various tree species characteristics, tree species information is also essential in estimating the carbon content in a certain area, especially in understanding the year-over-year effectiveness of reforestation or sequestration efforts by various companies.

This capstone project is aimed to use remote-sensing data such as multi-spectral satellite images and labeled tree species data with geo-locations to determine tree species in two regions in Austin, Texas. The main purpose is to train and build classification models that detect tree species at any location in the satellite imagery. More specifically, we will classify over 2.5 million pixels in aerial imagery and use multiple ML algorithms and deep learning techniques to predict which tree species each pixel contains. Besides, we compute the Normalized Difference Vegetation Index (NDVI) and extract various texture-based features from the NAIP imagery to improve the accuracy of our tree species classification model.

1.2 Related Work

Research has been conducted on plant species detection, where researchers first delineate trees from LiDAR images, then classify tree species based on detected tree polygons and additional spectral information. In Heinzel et al.'s research, the method applied is to first transform the band information in red, green, blue, and near-infrared to hue (H), saturation (S), and intensity (I), and then classify species based on Gaussian-smoothed histograms of the spectral features since there can be found local maximum and minimum.[1] Deur et al. used satellite imagery and further introduced machine learning models- random forest and SVM- into the detection process. They also combined extra texture features of the image like GLCM in order to improve the model performance.[2] In the past decade, deep learning methods (e.g., 3D-CNN) are most commonly used and are proven to have higher accuracy in tree species classification tasks.[3] With the development of unmanned aerial vehicles (UAV), there is more and more research done with data collected by UAVs as well, since the cost is lower than satellite imagery. Natesan et al. used high-resolution imagery from UAVs to train a dense convolutional neural network and acquired 84% test accuracy, where they also confirmed that consecutive temporal data could also lead to better classification.[4]

2 Data Description

In this section, we will describe what kinds of data we collect. When dealing with spatial data, we will frequently work with two major data types: vector data and raster data. The main vector data types are points, lines, and polygons. In a shape file, a point must contain its coordinate pair (x, y). A line must consist of an ordered set of points and a polygon is represented as a set of closed polylines, meaning that the last coordinate pair must coincide with the first pair. A shape file can include not only the geometry of vector data but also additional attributes of points, lines, and polygons. In contrast to vector data, the geometry of raster data is not stored as pairs of coordinates. Instead, a coordinate system will be used to record where the raster is located in geographic space. A raster image file is commonly represented as a rectangular array of pixels. A

raster file can potentially have many bands such as red, green, and blue. For each band, each pixel in that band has a value ranging from 0-255.

2.1 2D multi-spectral NAIP imagery

Raster files used in this project are high-quality pixel files from the National Agriculture Imagery Program (NAIP), which acquires aerial imagery at a high resolution that combines the image characteristics of an aerial photograph with the georeferenced qualities of a map. Tag Image File (TIF) is the raster file type that is provided as the input image files of the project, which includes aerial images of four regions of Austin, Texas (southwest, southeast, northwest, northeast), with geometric information (x and y coordinates of each pixel) and color band information (Red, Green, Blue, and Near Infrared). One reason that we utilize NAIP images is that it is quite accessible for individuals or companies to acquire such a dataset for analyzing tree species in the future at a very low cost. Besides, those images obtained from Austin, Texas are collected during full leaf seasons. It is convenient for us to detect tree species without too much preprocessing work of satellite imagery.

2.2 Tree species and tree segmentation

The vector files used in the project are in the form of the delimited text file (.csv file) and shapefile (.shp file). The raw data is given in the form of a .csv file, which is later transformed into a .shapefile for following geospatial analysis. A shapefile is a common format used to store the vectors in QGIS, and the vectors can include data points and a shape with a list of vertices. The two types of shape files that we touch with are the tree species data points and the polygons for identifying the tree crown boundary. The tree species data set is composed of different types of tree species names, their associated geometry information (longitude and latitude), and their diameters. The polygons that are used to classify the tree species are determined by visual interpretation of WV-3 imagery and by using field data on tree species and tree locations. The accuracy of the interpretation that generates the polygons is under consideration for our further classification process.

3 Data Cleaning and Exploratory Data Analysis

3.1 Missing data

In our tree species shape file, there are 64 missing values in the species column, and we decided to drop those rows because we prioritize the completeness of the dataset. Also, the tree species column contains a few unrecognizable names, such as “unknown”, we remove them as well.

3.2 Similar species names and duplicates

The column contains many alias entries, for instance, “Southern Oak” and “Oak (Southern)”, which are different strings that point to the same species. To standardize the format, we first remove all the parentheses and convert all tree species' names to lowercase strings. Then, we create a dictionary that maps different names of the same species to one based on our data referring to Trees of Texas[5]. After relabeling them, we find that 204 rows are duplicated with other rows, so we delete them to prevent potential bias when finding the popular species in the region. Species marked as “vacant” are dropped as well.

3.3 Feature engineering

3.3.1 NDVI: We calculate and analyze the vegetation index, known as the NDVI index. It is measured by extracting the red and near-infrared spectral information from the raster file. That is, near-infrared radiation minus visible radiation divided by near-infrared radiation plus visible radiation. For a given pixel, the value of NDVI close to +1 indicates the highest possible density of green leaves because healthy vegetation absorbs most of the visible light that hits it, and reflects a large portion of the near-infrared light. We think the existence of the NDVI index will be a great feature source for our model to detect various tree species.

3.3.2 Min, Max, Mean, Std (of RGB, NIR, and NDVI): Besides the red, green, blue, and near-infrared band information, we will also take some transformations of these values. The idea is to expand the spectral characteristics without adding too many correlated feature values. For

any given pixel in the tiff imagery, we will take its 8 surrounding pixels, and calculate the average, minimum, maximum, and standard deviation of the total 9 pixels. We will use this kind of 3x3 feature window to slide through the entire image. These calculations help us understand the sharing characteristics or disparities among different regions in the image. After getting these feature statistics, we still need to examine the correlation between each pair of them. Having too many correlated features in our training dataset has no benefit to the accuracy of the final models, but will slow down the system performance and take up a lot of memory.

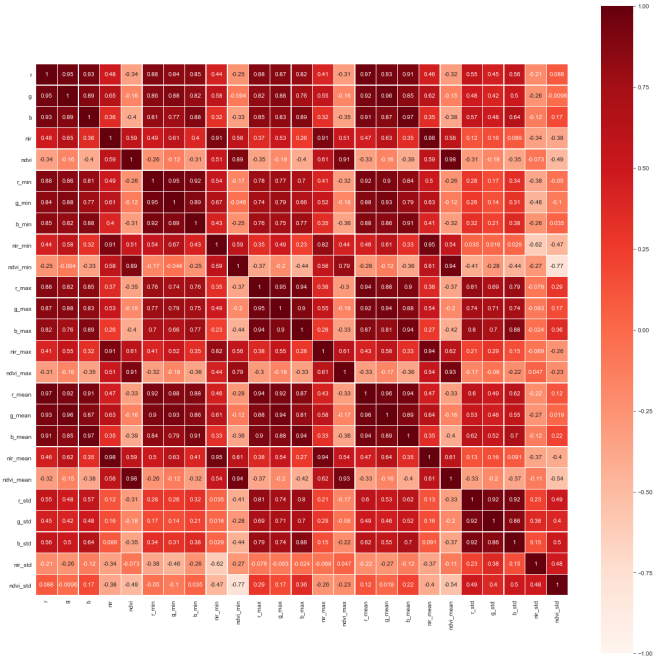

Figure 1. Correlation matrix showing correlations between each pair of 25 features

3.4 Tree species to be classified

As shown in the bar plot, we pick the top 7 most common species in the dataset as our final classifications: southern live oak, elm cedar, pecan, crape myrtle, escarpment live oak, hackberry and ashe juniper. The remaining species will be labeled as “others”.

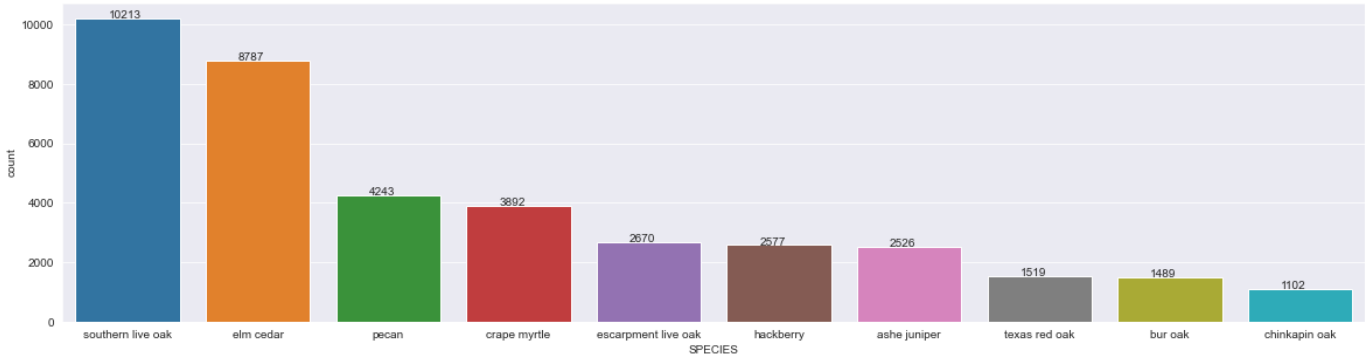

Figure 2. Bar plot showing the top 10 species in the tree species csv

3.5 Data visualization/manipulation using QGIS

In order to gain a better understanding of the satellite imagery, in particular the composition of multispectral band distribution(Red, Green, Blue, Near-Infrared), we use an open GIS tool called QGIS, which visualizes text and numerical file inputs by transforming them into graphic layers. We import the satellite images in TIFF format and build statistics on the band distribution. A sample tree species classification map is shown in Figure 3. The left and right plot below shows the distribution of the top 7 most common species in the southeast and southwest region, respectively.

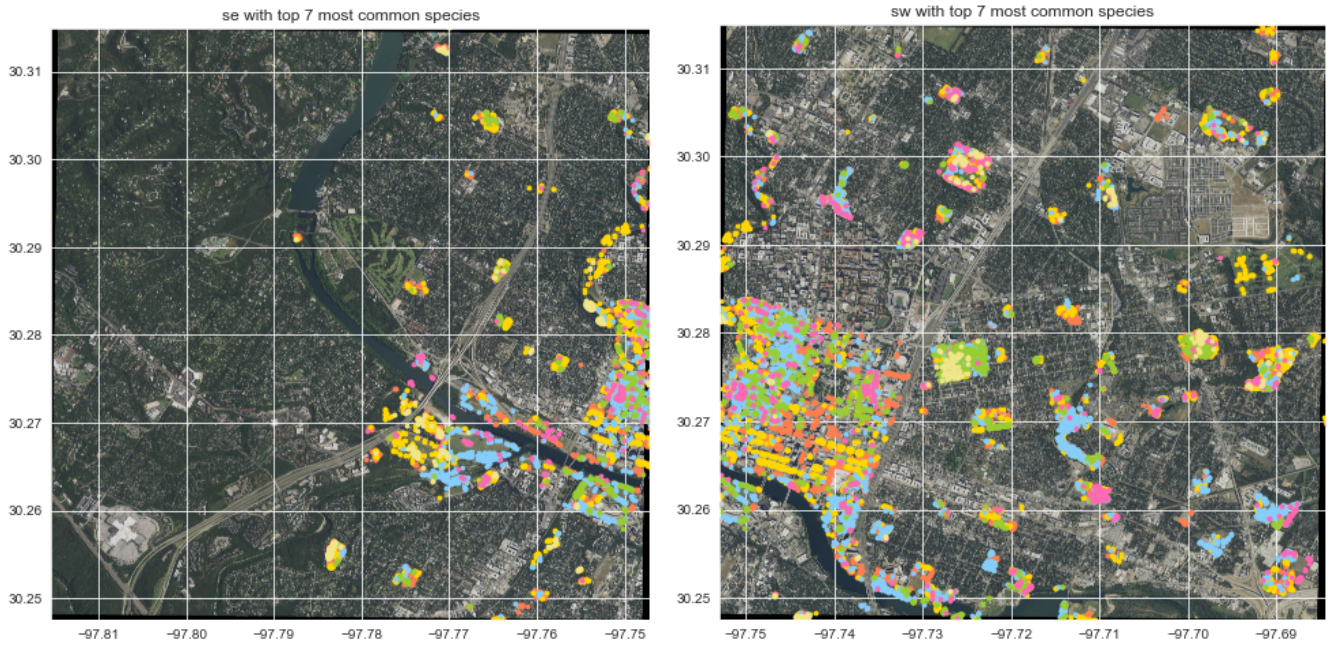

Figure 3. Tree species classification map for seven tree species in Austin, Texas.

Moreover, a map of NDVI is depicted in Figure 4. for the southeast and southwest regions of Austin. Figure 4 reflects the coverage of vegetation. We can easily distinguish vegetation areas from other classes such as rivers, roads, and buildings. The black dots in the graph are represented as our labeled trees.

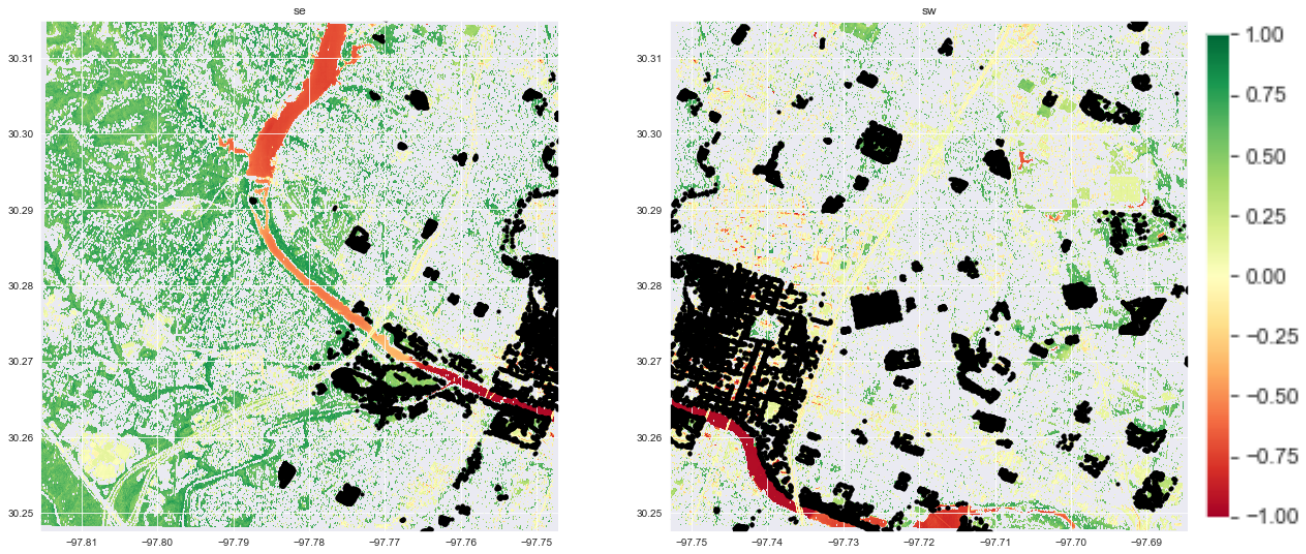

Figure 4. NDVI map in Austin, Texas.

4 Methodology

4.1 Data Association and Labeling

The tree species dataset is composed of 62274 rows and 6 columns. Each row is represented by one species chosen from 482 distinct species types. The six columns are the geometry location of the species, the species name, the diameter of the species, the latitude, longitude, and the new georeferenced location.

One of our main goals is to associate all pixels within polygons with identified species. In order to get our final species pixel shapefile, first we matched the coordinate reference system(CRS) of species and satellite image: we reprojected the NAIP images of the Southwest and Southeast region of Austin from EPSG:26914 to EPSG:4326. Secondly, we matched the polygons shapefile with the species information by choosing the majority or random species inside each polygon. Lastly, we associated every pixel within polygons with the species information. Besides that, we want to associate every pixel in the polygons with feature values, so we created 25 tiff files showing individual feature values for each pixel. Then we cropped the tiff files using the polygons that contain the species and used QGIS to convert raster pixels to point shapefile. In the

end, we did spatial join for 25 points shapefiles and eventually gathered them into species pixel shapefile.

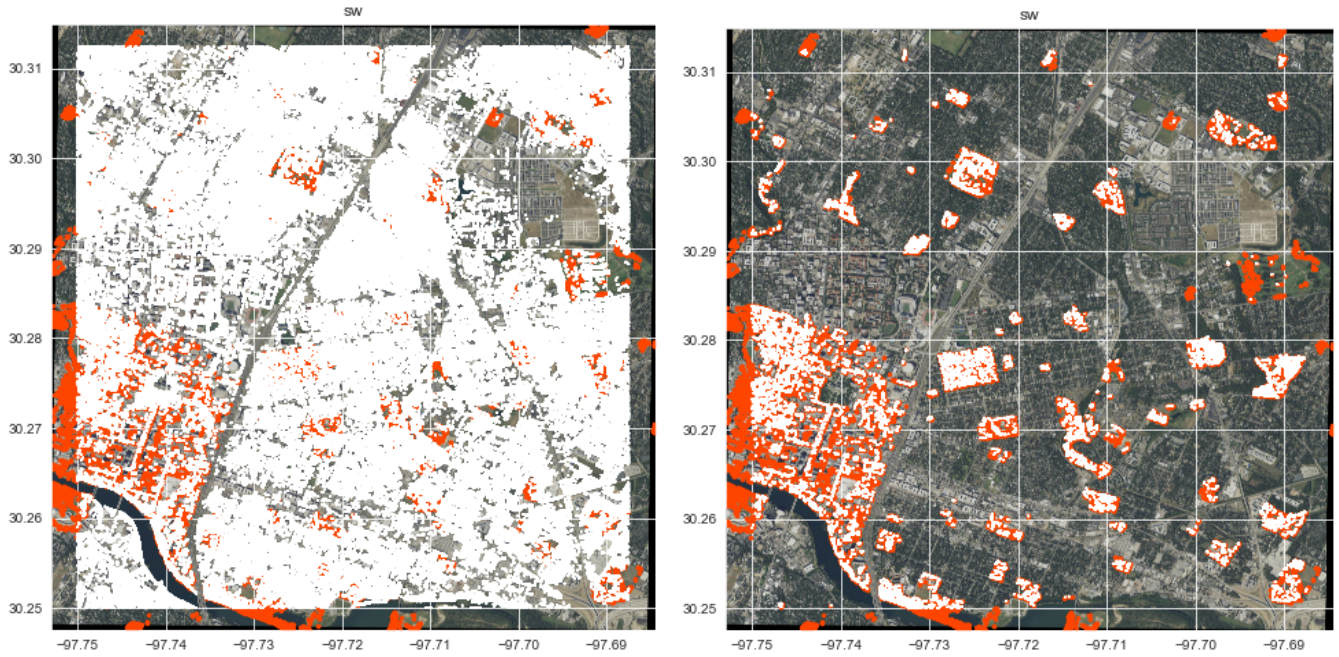

Figure 5. Cropped raster files left with polygons with species in the southwest region

After our data processing stage, we get our final training data, which consists of 1845082 pixels in southwest polygon regions and 739690 pixels in southeast polygon regions. Each row is composed of pixel locations, associated feature values, and designated species in the area. The task is to use the feature information to classify the top 7 tree species in the two regions. In this project, several machine learning models are deployed for classification: random forest(RF), supervised vector machines(SVM), and Neural Networks.

4.2 Train-test split

We split our training dataset into 80% of the training set and 20% test set. The details of the dataset are shown as follows. The first index in the brackets is the number of rows of the dataset, and the second index shows the number of columns.

	x_train(feature)	y_train(target)	x_test(feature)	y_test(target)

Southwest	(1476065, 25)	(1476065,1)	(369017, 25)	(369017, 1)
Southeast	(591752, 25)	(591752, 1)	(147938, 25)	(147938, 1)

4.3 Target variables

The target variables for two sets of training data (southwest and southeast) are divided into 8 class labels, where the seven of the class labels are the top species in each polygon region, and the last class ("others") is a mix of the rest of tree species.

Southwest	pecan, southern live oak, elm cedar, escarpment live oak, crape myrtle, hackberry, ashe juniper, others
Southeast	elm cedar, escarpment live oak, pecan, southern live oak, ashe juniper, hackberry, crape myrtle, others

5 Model Review

All the models we used are subject to changes. They are recorded as baselines for our models. Some of the train/test accuracies below are not finalized. We will proceed with our modeling and tuning procedure during the remaining semester, with the main focus on building more accurate neural network models.

5.1 Random Forest

Accuracy Report (SW)	Accuracy Report (SE)
Accuracy on train: 0.9999986450461192 Accuracy on test: 0.5006950899281063 **Best Params**: {'n_estimators': 1788, 'min_samples_split': 2, 'min_samples_leaf': 4, 'max_features': 'sqrt', 'max_depth': 90, 'bootstrap': True} (Use RandomSearch with 3 fold CV)	Accuracy on train: 1.0 Accuracy on test: 0.5065635604104423 **Best Params**: {'n_estimators': 311, 'min_samples_split': 2, 'min_samples_leaf': 1, 'max_features': 'sqrt', 'max_depth': None, 'bootstrap': False} (Use RandomSearch with 3 fold CV)

5.2 Linear/Kernel SVM

Accuracy Report (SW)	Accuracy Report (SE)
Accuracy on train: 0.3164 Accuracy on test: 0.3012 **Best Params**: (to be determined)	Accuracy on train: 0.2766 Accuracy on test: 0.267 **Best Params**: (to be determined)

5.3 XGBoost

Accuracy Report (SW)	Accuracy Report (SE)
Accuracy on train: 0.9958 Accuracy on test: 0.3494 **Best Params**: (to be determined)	Accuracy on train: 0.9986 Accuracy on test: 0.3308 **Best Params**: (to be determined)

5.4 Neural Network

Accuracy Report (SW)	Accuracy Report (SE)
Accuracy on train: 0.4812 Accuracy on test: 0.390 **Best Params**: (to be determined)	Accuracy on train: 0.4730 Accuracy on test:0.373 **Best Params**: (to be determined)

6 Conclusion and Future Work

We are currently at the 50% mark of our project. We have cleaned our datasets and associated different components, including all features and tree species information. Throughout this data analysis and labeling process, we worked with various multispectral satellite imagery data and geo data frames. After preparing the training data, we checked the correlation of the features and selected the essential ones, and eventually built some basic models.

Next, we will proceed from these aspects:

- **Improve feature selection**: we could include more features like texture, and height info from Lidar imagery (the ratio of height and diameter is different among different tree species).
- **Fine tune different parameters and find the best model:** resolve the overfitting problem in our current models; our goal is to reach 60% - 70% test accuracy.
- **Test and evaluate our final model**:
    - Increase or decrease the number of species we classify: run our final model on new labels to get accuracy and see if there are significant differences
    - Apply our final model to surrounding regions: currently our data comes from 2 regions satellite imagery from Austin. We can get more raster data from NAIP and test our model accuracy.

Our ultimate goal is to provide the model with high accuracy to help different companies understand tree species information. For most companies, our model should take some basic features (R, G, B, NIR) as input and generate a good outcome. For some companies that require higher accuracy, we could build another model that takes more complex features as input, like height, as a stretch goal. Those features come from Lidar data, which is expensive to get.

References

[1] Heinzel, J. N., Koch, B., & Weinacker, H. (2008, September). *Full automatic detection of tree species based on delineated single tree crowns - a data fusion approach for airborne laser scanning data and aerial photographs*. SilviLaser. Retrieved October 23, 2022, from https://www.researchgate.net/profile/Barbara-Koch-2/publication/228719924_Full_automatic_detection_of_tree_species_based_on_delineated_single_tree_crowns-a_data_fusion_approach_for_airborne_laser_scanning_data_and_aerial/links/0c960528b0fbc4fff6000000/Full-automatic-detection-of-tree-species-based-on-delineated-single-tree-crowns-a-data-fusion-approach-for-airborne-laser-scanning-data-and-aerial.pdf

[2] Deur, M., Gašparović, M., & Balenović, I. (2020). Tree species classification in mixed deciduous forests using very high spatial resolution satellite imagery and machine learning methods. *Remote Sensing*, *12*(23), 3926. https://doi.org/10.3390/rs12233926

[3] Pu, R. (2021). Mapping tree species using advanced remote sensing technologies: A state-of-the-art review and perspective. *Journal of Remote Sensing*, *2021*, 1–26. https://doi.org/10.34133/2021/9812624

[4] Natesan, S., Armenakis, C., & Vepakomma, U. (2020). Individual tree species identification using dense convolutional network (DenseNet) on multitemporal RGB images from UAV. *Journal of Unmanned Vehicle Systems*, *8*(4), 310–333. https://doi.org/10.1139/juvs-2020-0014

[5] Texas A&M Forest Service - Trees of Texas - list of trees. (n.d.). Retrieved October 22, 2022, from http://texastreeid.tamu.edu/content/listOfTrees/index.aspx