1. **Introduction:**

As the rate of global deforestation by human activity increases, as do its negative effects on natural ecosystems and the climate. Efforts to mitigate these negative effects, conserve individual tree species, and maintain biodiversity all hinge on researchers’ ability to quantify and monitor the density and distribution of individual tree species in forests. However, collecting data on individual trees in the field is expensive, time consuming, and laborious. As a result, the frequency and scale at which this crucial data can be collected is increasingly limited. Remotely sensed imagery from satellites, airplanes, and drones circumvent these limitations by enabling real-time observation of ecosystems at high spatial resolution. As such, the identification of tree species with remote sensing datasets offers a cost-effective way to inventory, protect and manage forest resources. The goal of this project is to investigate the use of hyperspectral, remote sensing, imagery, for the classification of individual trees to their taxonomic species. The project is inspired by the 2020 data science competition: IDTReeS, [Integrating Data science with Trees and Remote Sensing](https://idtrees.org/competition/).

1. **Related Work:**

Remotely sensed images detect the unique reflectance values (or spectral signatures) of different surface objects. The unique values are then used to discriminate between different objects on-the-ground, this includes intra-comparisons of trees or vegetation. Hyperspectral images (HSI), which detect reflectance values at a wide spectrum (typically between 64 and 256 wavelengths), provide significant levels of detail and allow for the distinction of fine spectral variations among tree species (Ballanti et al., 2016). This has resulted in the extensive use of hyperspectral imagery for tree species classification (Ballanti et al., 2016; Dalponte et al., 2014; Nezami et al., 2020; Raczko & Zagajewski, 2017; Wu & Zhang, 2020).

The Ballanti (2016) study compares the use of two non-parametric classifiers on HSI: Support-Vector-Machine (SVMs) and Random-Forest (RF) algorithms. The SVM classifier outperformed the RF. Ballanti (2016) apply a forward minimum noise fraction (MNF) transform to the mosaicked image and identify the most meaningful spectral bands to reduce the high data dimensionality and redundancy of HIS. Raczko and Zagajewski (2017) conduct a similar study, except choose to compare the performance of SVM, RF, and Artificial Neural Network (ANN) classifiers; the ANN performed best followed by the SVM and RF. Their study also makes use of the work by Pal and Mather (2006), which showed that a 40-band dataset was optimal for reducing processing times of HSI, while preserving enough data to obtain satisfactory results. Raczko and Zagajewski (2017) apply Principal Component Analysis (PCA) for selecting the 40 most important bands. Wu and Zhang (2020) combined hyperspectral data with simultaneously acquired LiDAR data to extract multiple tree-crown features, including canopy height, texture feature, and spectral indices. Different feature combinations and classifiers (KNN, SVM) were tested on classifying tree species with SVM demonstrating the highest classification accuracy.

**Explaining the Data:**

*On Remote Sensing Data:*

The competition provides three primary data sources: remote sensing, field data, and individual tree crowns. The following project will only consider remote sensing geospatial datasets, specifically passive sensing systems. Passive systems measure the amount of reflectance at different wavelengths for ground-detected objects. The remote sensing datasets are generated by the NEON Airborne Observation Platform (AOP), and distributed in RGB and Hyperspectral formats at 100 cm2 and 1 m2 spatial resolutions, respectively. Data is stored as raster files, which means an image or array of pixels, whereby each pixel is stored as a vector of numbers. An RGB image is stored as a 3-band raster (3-element vector pixels). Each band represents the reflectance at different points in the electromagnetic spectrum corresponding to red, green, and blue wavelengths, respectively. Hyperspectral data consist of reflectance information from a much wider electromagnetic spectrum (380-2510 nanometers). Our data has a total of 369 bands.

*On Individual Tree Crown (ITC) Delineations:*

Individual tree crown (ITC) delineates are generated by IDTReeS research group. Each delineation is a 2-D rectangular bounding box defining the maximum tree crown extent in an image and is provided in vector format as ESRI shape files.

*Location of the Tree Crown Data*

The data consist of three NEON ecoclimatic sites in Eastern United States. In other words, each site is characterized by distinctive environmental, geographic, and vegetative properties. The sites are:

* **Ordway-Swisher Biological Station, Florida (OSBS):** The region contains mixed forests of hardwood and conifers, mostly dominated by pine trees.
* **Talladega National Forest, Alabama (TALL):** Forests made of mixed hardwood and conifers (mostly pine) in the Ozarks complex.
* **Mountain Lake Biological Station, Virginia (MLBS):** The region is mainly made of hardwood forests in the Appalachians and Cumberland Plateau.

Calendar

Description automatically generated

**Method:**

*Pre-Processing the Data:*

The training data consists of a total of 85 hyperspectral images are provided across all three sites. Each image represents the geographic extent of a single 20 x 20-meter plot, with array dimensions (20, 20, 369). Within these images are 1,165 delineated tree crowns with corresponding tree class labels (taxonomic species names).

To prepare the data, each of the 85 hyperspectral images were clipped into separate images corresponding to just the bounding boxes of each labelled ITC. This produced 1,165 images of varying heights and widths, with 369 spectral bands. The labels and extents of bounding boxes per image is given in the guideline file ‘data\_train\_mitree.csv’. There are a total of 33 tree species types delineated across all three NEON sites. Most of these species were underrepresented in the data, with some species corresponding to only 1 ITC. Given how limited the HSI data is, it is unlikely that a single classifier can demonstrate high prediction accuracies for all 33 species. As a result, the classification task is simplified to only the top 6, most frequently encountered, tree species and a 7th class labelled as ‘Other’, which encompasses the remaining 27 tree species in the NEON imagery.

* Object-Level Classification: is each clipped image, read-in as a tensor of shape , and is the associated tree label for that tensor. There would be a total of 1,165 tensors.
* Pixel-Level Classification: represents a single pixel, read-in as a 1-D array with shape and is the associated tree label for that pixel. There would be a total of 35,488 pixels.

*Dimensionality Reduction:*

*Standardize Data:*

The training data is then standardized by de-meaning and standardizing each

*Random Sample*

The classes within the dataset are imbalanced. For each of the 6 classes there are 2066, 2073, 4392, 4647, 6207, 8700. Which leads to a large class imbalance. Training a ML model with such imbalance will lead the classifier to heavily bias the class with the highest number of observations. In order to deal with this imbalance, a mixture of up-scaling and down-scaling is incorporated. More specifically, for each class, if the class has less than 6198 observations, the data is resampled from the class until there are 6198 observations. Similarly, for classes with higher than 6198 observations, a subsample of 6198 observations is randomly chosen. The choice of 6198 was chosen as 3 times the number of observations for the class with the least number of observations - 3 in itself was an arbitrary choice.

This random sampling/subsampling is done after the data has been split into training and test dataset. The ratio used for the split is 80/20 training/test.

*Classifiers Compared on Test Data*

For the main results, 5 different ML algorithms for getting accuracy results for the test data. These algorithms are Logistic Regression, Random Forest, XGBoost, SVM, and KNN classification.

*Hyperparameter Tuning*

Hyperparameter tuning is mostly done informally. Initially, mostly standard settings are used to get results. SVM is trialed with different kernels and the linear and RBF kernels perform the best in some initial tuning.

*Testing Data:* Tree clown delineations without tree class labels. Test our model by the classification of species on bounding boxes of unknown species identity.

**Results:**

*Performance Metrics:* F1 Scores (sklearn.metrics.f1\_score); Average Cross-Entropy Loss (sklearn.metrics.log-loss); Confusion Matrix (sklearn.metrics.confusion\_matrix).

Table 1. Mean statistics of Test data after 5 fold Cross-Validation

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Time (s) | Test accuracy | Test precision, weighted | Test F-1, weighted |
| KNN | 0.053 | 0.716 | 0.718 | 0.713 |
| LogReg | 82.686 | 0.708 | 0.709 | 0.708 |
| RF | 11.539 | 0.814 | 0.813 | 0.814 |
| SVM | 92.497 | 0.718 | 0.72 | 0.716 |
| XGB | 1172.369 | 0.825 | 0.823 | 0.823 |

**Conclusion:**

**References:**

*Collaborators*

As this project is an extension of an attempt at a Competition Submission, competition team members played a role in preliminary project scoping and data processing.

Team Members include: Björn Lütjens {lutjens@mit.edu}; Becca Browder {bbrowder@mit.edu}

*Neon Data References*

National Ecological Observatory Network. 2020. Data Product DP1.30010.001, High-resolution orthorectified camera imagery. Provisional data downloaded from http://data.neonscience.org on March 4, 2020. Battelle, Boulder, CO, USA NEON. 2020.

National Ecological Observatory Network. 2020. Data Product DP1.30003.001, Discrete return LiDAR point cloud. Provisional data downloaded from http://data.neonscience.org on March 4, 2020. Battelle, Boulder, CO, USA NEON. 2020.

National Ecological Observatory Network. 2020. Data Product DP1.10098.001, Woody plant vegetation structure. Provisional data downloaded from http://data.neonscience.org on March 4, 2020. Battelle, Boulder, CO, USA NEON. 2020.

National Ecological Observatory Network. 2020. Data Product DP3.30015.001, Ecosystem structure. Provisional data downloaded from http://data.neonscience.org on March 4, 2020. Battelle, Boulder, CO, USA NEON. 2020.