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, as well as 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 species 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 used to discriminate between different objects on-the-ground, which includes intra-comparisons of trees or vegetation. Hyperspectral imagery (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 (2007), 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 (K-Nearest Neighbors, SVM) were tested on classifying tree species with SVM demonstrating the highest classification accuracy.

1. **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), and in our case has a total of 369 bands per pixel.

*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

**Figure 1** – Each plot represents the same 20 by 20-meter area in the MLBS site, only at a different spectral band. The color of the pixels corresponds to the reflectance values of the 1-m2 land area. The white square is the delineated ITC for the species *Acer pensylvanicum L.* found in that image.

1. **Methods:**

*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.

* 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.

I proceeded with pixel-level classification and used Principal Component Analysis (PCA) to reduce the dimensionality of the pixel (i.e., the number of bands). I extracted the 40 bands (features) with the highest explained variance (i.e., largest eigenvalues). The number of features was set to 40 as per the findings of Pal and Mather (2007). The data was then standardized as follows:

I split the data by applying 80 to 20 training to test ratio. The classes within the training dataset were imbalanced. For each of the 6 classes there are 4647, 8700, 4392, 2066, 2073, 6207 pixels. Training a Machine Learning (ML) model with imbalanced data would result in a classifier which is heavily biased towards the class with the largest number of observations. To address this, the data was both up- and down-scaled. More specifically, if a class has less than 6198 observations, the data in the class is resampled until there are 6198 observations. Conversely, for classes with higher than 6198 observations, a subsample of 6198 observations is randomly chosen. A fixed class size of 6198 pixels was chosen as 3 times the number of observations of the class with the least number of observations, the 3 multiplier was an arbitrary choice.

*Classifying the Data:*

Five different ML algorithms were compared on the test data. The algorithms were: Logistic Regression (LogReg), Random Forest, XGBoost (XGB), SVM, and K-Nearest Neighbors (KNN). While LogReg and XGBoost were not directly referenced in literature, they have demonstrated high efficacy in object classification from imagery and were therefore also experimented with. The standard settings of each of the five models were used. Different kernels were trialed for the SVM classifier, with the linear and RBF kernels performing best. In the final comparison only the linear kernel, SVM classifier was tested.

1. **Results:**

To replicate a real-world implementation of the model use, initially a 5-fold cross-validation (CV) on the training dataset is run. Afterwards, we can compare the results of the CV and the test data results to see if our model choice that would have resulted from CV performs the best.

For the CV, the computation time along with three different cross-validation metrics are recorded. The metrics chosen were accuracy score, precision score, and F1 score. Accuracy scores return the fraction of correctly classified samples. The precision score is a measure of the ratio of true positives to the total number of predicted positives. And the F1 score is the weighted average of the precision and recall of a classifier. In our multi-classification task, the F1 score is the average of the F1 score of each class. RF and XGB classifiers perform the best across all validation metrics in Table 1. Between the two, the RF model takes almost 1/1000th of the time to run than the XGB model, with only a 1% loss in accuracy. The model with the shortest compute time was KNN, however it is approximately 10% less accurate or precise than either the RF or XGB classifiers.

Table 2 outlines the test accuracy of the different trained models (from the CV process discussed in the last paragraph). The models with the top accuracy are still RF and XGB, but there is an approximately %10 drop in the accuracy. This overfitting in the CV process can be attributed to the up-sampling done in order to balance the number of images in each class. The model with the least change in accuracy is Logistic Regression along with SVM. Nonetheless, even with overfitting, RF and XGB produce the highest Test Accuracy and can be seen to perform the best on unseen data. Figure 2 gives the confusion matrix for each classifier, demonstrating how well the model performs on each of the 6 classes. Overall, the confusion matrices align with the average results in Table 1 and 2 and show that the RF and XGB models have the fewest instances of misclassifications or false positives (or false negatives).

The confusion matrices allow us to analyze how each of the models perform on each of the classes (i.e., species). All models performed exceptionally on the *Pinus palustris Mill.* species, with accuracy and precision scores ranging between 0.94 and 0.97. This is likely due to *Pinus palustris Mill.* having the highest number of training samples (8700). Contrastingly, other species faced widely different F1 scores between classifiers. For example, *Quercus laevis Walter* is classified with a 0.75 to 0.81 F1 score, except with the KNN for which the F1 score was 0.61, due to a very low precision score of 0.48. In general the species with the lowest number of training samples (prior to resampling), *Quercus laevis Walter* and *Quercus coccinea,* had the lowest precision score – or the highest instances of false positives.

**Table 1.** Mean statistics of 5-fold Cross-Validation on Training data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Time (s) | CV Accuracy | CV Precision, weighted | CV F-1, weighted |
| LogReg | 82.686 | 0.708 | 0.709 | 0.708 |
| RF | 11.539 | 0.814 | 0.813 | 0.814 |
| XGB | 1172.369 | 0.825 | 0.823 | 0.823 |
| KNN | 0.053 | 0.716 | 0.718 | 0.713 |
| SVM | 92.497 | 0.718 | 0.720 | 0.716 |

**Table 2.** Test results

|  |  |
| --- | --- |
| Model | Test Accuracy |
| LogReg | 0.697 |
| RF | 0.710 |
| XGB | 0.730 |
| KNN | 0.640 |
| SVM | 0.702 |

**A picture containing graphical user interface

Description automatically generated**

**Figure 2** – Confusion matrix results for each of the five ML algorithms.

1. **Conclusion:**

The results of the study demonstrate that investing in the expansion of ITC-labelled hyperspectral imagery to a larger number of species would significantly improve the performance of ML models on the data. The study also emphasized Random Forest and XGBoost classifiers as highly effective at tree-species classification, especially for the *Pinus palustris Mill* and *Quercus rubra L.* species. Corroborating supporting literature which claims hyperspectral images can substitute time and labor-intensive field studies. Future work should experiment with including ancillary information about each tree species (e.g., canopy height or LiDAR imagery) and tuning RF and XGB hyperparameters.

**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: Becca Browder {bbrowder@mit.edu}; Björn Lütjens {lutjens@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.

*Literature References*

Ballanti, L., Blesius, L., Hines, E., & Kruse, B. (2016). Tree Species Classification Using Hyperspectral Imagery: A Comparison of Two Classifiers. *Remote Sensing 2016, Vol. 8, Page 445*, *8*(6), 445. https://doi.org/10.3390/RS8060445

Dalponte, M., Ørka, H. O., Ene, L. T., Gobakken, T., & Næsset, E. (2014). Tree crown delineation and tree species classification in boreal forests using hyperspectral and ALS data. *Remote Sensing of Environment*, *140*, 306–317. https://doi.org/10.1016/j.rse.2013.09.006

Nezami, S., Khoramshahi, E., Nevalainen, O., Pölönen, I., & Honkavaara, E. (2020). Tree species classification of drone hyperspectral and RGB imagery with deep learning convolutional neural networks. *Remote Sensing*, *12*(7). https://doi.org/10.3390/rs12071070

Pal, M., & Mather, P. M. (2007). Some issues in the classification of DAIS hyperspectral data. *27*(14), 2895–2916. https://doi.org/10.1080/01431160500185227

Raczko, E., & Zagajewski, B. (2017). Comparison of support vector machine, random forest and neural network classifiers for tree species classification on airborne hyperspectral APEX images. *European Journal of Remote Sensing*, *50*(1), 144–154. https://doi.org/10.1080/22797254.2017.1299557

Wu, Y., & Zhang, X. (2020). Object-Based tree species classification using airborne hyperspectral images and LiDAR data. *Forests*, *11*(1). https://doi.org/10.3390/f11010032