Classifying FIA Forest Type from a Fusion of Hyperspectral and LiDAR Data

Caileigh Shoot

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington

2018

Committee:

L. Monika Moskal, Co-Chair

Hans-Erik Andersen, Co-Chair

David Butman

Program Authorized to Offer Degree:

School of Environmental and Forest Sciences

University of Washington

Abstract

Classifying FIA Forest Type from a Fusion of Hyperspectral and LiDAR Data

Caileigh Shoot

Co-Chairs of the Supervisory Committee:
L. Monika Moskal, Associate Professor
Hans-Erik Andersen, Affiliate Assistant Professor
School of Environmental and Forest Sciences

In this study, we develop a methodology for classifying FIA defined forest type across the Tanana Inventory Unit (TIU) using a fusion of hyperspectral and LiDAR data. The hyperspectral and LiDAR data used in this study were collected as part of the 2014 acquisition with the NASA Goddard's LiDAR, Hyperspectral & Thermal Imager (G-LiHT). In order to determine the best classification method, we tested 5 classification algorithms: Naive Bayes Classifier, K-Nearest Neighbor, Multinomial Logistic Regression, Support Vector Machine, and Random Forests.

Each model was trained and validated using the forest type corresponding to each FIA subplot, alongside raw hyperspectral data (114 spectral bands in total), hyperspectral vegetation indices, and selected LiDAR-derived canopy height and topography metrics. Six different combinations of this input data were tested to determine the most accurate classification algorithm and model inputs. A 3-fold cross validation was performed in order to ensure that all data was included in

both training and validation, but never within the same model. Of the five models and six model input combinations tested, we found Random Forest with hyperspectral vegetation indices as well as topography and canopy height metrics as model inputs had the highest accuracy at 77.53% overall. With the completion of this work, we hope to use this "best" model to classify forest types across the Tanana Inventory Unit in central inland Alaska where there is G-LiHT coverage.

There are three primary sections of this thesis document. The first is an informal introduction to the study wherein the study, the people involved, and the lessons learned along the way are all introduced. The second section is an overview the research performed. This is the publication-ready document which details the work performed in this study. It provides an overview of the relevant literature, methods and materials used, and discusses the implications of the findings from this study. The third and final section is an extended discussion of future work. It gives detailed descriptions of future research that can be undertaken following this study, and gives open and detailed critique of the methods and materials used in this study in order to inform future research. It also provides a detailed description of the bootstrap aggregation process that was attempted to improve accuracies.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

This work would not have been possible without the dedicated guidance of my committee co-chairs, **L. Monika Moskal** and **Hans-Erik Andersen**, and committee member **David Butman**. I would also like to thank the funding sources that provided me with the tools necessary to complete this project: The UW School of Environmental and Forest Sciences for a yearlong fellowship which funded my education and work during my first year in the program; The **USDA Forest Service** and **NASA** for providing funding and all necessary data; And the **Precision Forestry Cooperative** for housing me and funding many of my conference presentations.

I received a great deal of help and guidance from many people over the past 2 years while I worked to obtain my masters. Thank you to all my friends, family, coworkers, UW staff members, and others who helped me achieve this lifelong goal. I would not have been able to do this without you:

Anthony Shoot
Bernease Herman
Bob McGaughey
Bruce Cook
Bridget Daly
Chad Babcock
Christy Heaton
David Campbell
Douglas Morton
Gabriel Torres
Jeff Richardson
Jonathan Kane
John LeFevre
Justin McAllister
KC Deterling
Leta Kao
Megan O'Shea
Meghan Halabisky
Michelle Trudeau
Miles LeFevre
Sean Jeronimo
Tami Shoot
Travis Axe
Van Kane

DEDICATION

This thesis is dedicated to my Grandfather, Henry Renz.

Chapter 1. INTRODUCTION

Alaska is the largest state in the United States at 665,384 square miles in total area [1]. It is also the most northern state, with almost a third of its land area lying within the Arctic Circle. In 2017, Alaska was said to have an estimated population of 739,795, making it the 4th least populated U.S. state [2]. According to the 2011 National Land Cover Database, Alaska is made up of 20.26% forest and 39.67% shrubland [3]. The federal government owns the majority of the land in Alaska at about 346,875 square miles, or about 60% of the total land area [4]. Of that, the United States Department of Agriculture (USDA) Forest Service manages 21,956,250 acres of land, including wilderness, wetlands, roads, streams, trails, and much more [5].

The USDA Forest Service Forest Inventory and Analysis (FIA) program has provided vital information for assessing the status of forests in the United States of America since 1930 [6]. The modern Enhanced FIA Program implements a 3-phase sampling design. The first phase uses ancillary data, such as satellite or aerial imaging, to stratify the land area by cover. The second phase is the installation and measurement of permanent ground plots, so long as any portion of the plots contain forest land use. The plots are typically established along a hexagonal grid, wherein there is one ground plot in each 6000-acre hexagon. The third and final phase is conducted on 1/16ths of all established FIA plots, or approximately one plot per 96,000 acres. These plots are also included in the Phase 2 plots; thus, all the same measurements are made, but they also measure biotic and abiotic features associated with forest and ecosystem health [7].

In Alaska, the first inventory was conducted in the late 1950s and early 1960s. In interior Alaska, this inventory had 3 levels of sampling: large scale aerial photography (37,177 photos in total), an air check of 10% (3,774) of photos to verify photo-interpreted information, and field

visits and measurement of a small subsample (355) of the air-checked plots. This inventory allowed for estimation (with high levels of error) of a variety forest characteristics across the entire region, including forest area, size class, species composition, and volume. In the 1980's, a four-phase, multilevel inventory design (the Alaska Integrated Resource Inventory System (AIRIS)) was implemented in the Tanana Valley of Interior Alaska. This used satellite and aerial imagery to interpret vegetation class, volume, stand size class, foliar cover by vegetation class, understory component, tree crown diameters, tree and (or) shrub heights, and land use. Ground plots were also implemented to collect standard timber inventory variables [8]. More recently, plots were established in southeast and south-central Alaska between 1995 and 2003 [9]–[11]. The annual coastal inventory started in 2004, with a total of 2,227 plots collected between 2004 and 2013 [12]. This inventory only covered a small portion of Alaska's forests, leaving out approximately 112 million acres of interior Alaska, or 15% of forested land in the United States. This area was not included in surveying due to remoteness, size, lack of transportation infrastructure, complex logistics, and cost [13]. In 2014, a pilot project was established to begin the work of surveying interior Alaska, starting with the Tanana Valley of Interior Alaska. This area included the Tanana Valley State Forest and Tetlin National Wildlife Refuge which were sampled at a 1:4 intensity (or 1 plot per 24,000 acres) on a hexagonal grid [14].

At the same time that these new FIA plots were established, G-LiHT (Goddard-Lidar/Hyperspectral/Thermal) was simultaneously collected in order to augment the sparse-sample design. This new suite of sensors developed by NASA Goddard Space Flight Center simultaneously maps the composition, structure, and function of terrestrial ecosystems at high (~1 m^2) spatial resolution using LiDAR, hyperspectral, and thermal imaging [15]. This allows for improved assessment of forest conditions beyond FIA plots.

One key part of FIA plot protocol is the mapping of condition classes within each plot. The condition class attributes recorded for each plot describe the forest structure, composition, and disturbance history. This information allows for the FIA program to estimate and account for changes in forest land. A basic requirement of this estimation and accounting is to report the current status of forest lands by forest type [16]. Forest type is defined as the code associated with the dominant stocking of live trees that are not overtopped and is recorded for each subplot [7]. This essentially records the dominant tree species present at a given plot. If there are no trees present or if the trees do not meet a specific stocking threshold, then the plot is marked as non-forest [16]. This data allows for the understanding of the distribution of different species throughout the forests. In combination with the G-LiHT data collected, this can be used to classify forest type throughout that Tanana valley of interior Alaska.

In this study, we develop a methodology for classifying FIA defined forest type across the Tanana Inventory Unit (TIU) of interior Alaska using a fusion of hyperspectral and LiDAR data collected as part of the 2014 G-LiHT mission. This study consists of two main parts. In the first part, we aimed to determine the best classification algorithm for classifying forest type within our study area. We tested 5 classification algorithms: Naive Bayes Classifier, K-Nearest Neighbor, Multinomial Logistic Regression, Support Vector Machine, and Random Forests. Each model was trained and validated using the forest type corresponding to each FIA subplot.

The second part of this study aimed to determine which combination of predictor variables resulted in the highest classification accuracies. Six different combinations of predictor input data were tested. These predictor inputs were different combinations of the available data for each plot: raw hyperspectral bands, hyperspectral vegetation indices, DTM metrics, and CHM metrics. We found that the Random Forests classification method with hyperspectral vegetation indices, 5 raw

hyperspectral bands (red, green, blue, near infrared, and red edge), and DTM and CHM metrics resulted in the highest overall and kappa accuracy.

1.1 PURPOSE

Within the rural communities of interior Alaska, the availability of fuel sources is an extremely important issue in many households. With this, interior Alaska's communities are interested to know how much biomass is available in Alaska's forests for use in homes, biofuel production, and timber sales [17]. In addition, understanding biomass in Alaska helps us to better quantify and measure changes in boreal forests under climate change.

In order to accurately predict biomass, it is important that we understand the likelihood of a vegetation species being present in a given location. The 3D structural and spectral data from G-LiHT and the field data from the newly installed FIA plots, provide the information necessary to predict forest type within the TIU. The best classification algorithm and model inputs for this process is unknown. This study tests 5 commonly used algorithms and 6 model inputs derived from LiDAR and Hyperspectral data.

1.2 ABOUT OUR TEAM

This study was performed by Caileigh Shoot, a graduate student in the University of Washington (UW) Remote Sensing and Geospatial Analysis Laboratory (RSGAL). The UW RSGAL is the remote sensing and geospatial research partner of the Precision Forestry Cooperative (PFC) in the College of the Environment, School of Environmental and Forest Sciences (SEFS) at the University of Washington.

The PFC was created to conduct pioneering research in forest management and assessment at a new scale of resolution and accuracy with the goal of producing economic and environmental

benefits. The PFC's overarching mission is to develop advanced technology solutions to improve the quality and reliability of information needed for planning, implementation, and monitoring of natural resource management, to ensure sustainable forest management. The PFC frequently works with collaborators throughout the Pacific Northwest and the world in order to achieve this mission. For this work, the PFC worked with Hans-Erik Andersen of the USDA Forest Service Pacific Northwest Research Station. Hans provided all the necessary data to perform these analyses, and acted as co-chair on Caileigh Shoot's committee, alongside PFC Director L. Monika Moskal.

Chapter 2. CLASSIFYING FIA FOREST TYPE FROM A FUSION OF HYPERSPECTRAL AND LIDAR DATA

2.1 ABSTRACT

In this study, we develop a methodology for classifying FIA defined forest type across the Tanana Inventory Unit (TIU) using a fusion of hyperspectral and LiDAR data. The hyperspectral and LiDAR data used in this study were collected as part of the 2014 acquisition with the NASA Goddard's LiDAR, Hyperspectral & Thermal Imager (G-LiHT). In order to determine the best classification method, we tested 5 classification algorithms: Naive Bayes Classifier, K-Nearest Neighbor, Multinomial Logistic Regression, Support Vector Machine, and Random Forests.

Each model was trained and validated using the forest type corresponding to each FIA subplot, alongside raw hyperspectral data (114 spectral bands in total), hyperspectral vegetation indices, and selected LiDAR-derived canopy height and topography metrics. Six different combinations of this input data were tested to determine the most accurate classification algorithm and model inputs. A 3-fold cross validation was performed in order to ensure that all data was included in both training and validation, but never within the same model. Of the five models and six model input combinations tested, we found Random Forest with hyperspectral vegetation indices as well as topography and canopy height metrics as model inputs had the highest accuracy at 77.53% overall. With the completion of this work, we hope to use this “best” model to classify forest types across the Tanana Inventory Unit in central inland Alaska where there is G-LiHT coverage.

2.2 INTRODUCTION

The United States Forest Service (USFS) Forest Inventory and Analysis (FIA) program is designed to give the USFS the information necessary to monitor and assess our nation's forests.

Since 1930 the USFS has continually monitored much of our nation's forests [18]. In Alaska, the first statewide inventory was conducted in the late 1950s and early 1960s. The interior Alaska component of this inventory had 3 levels of sampling: large scale aerial photography (37,177 photos in total), an air check of 10% (3,774) of photos to verify photo-interpreted information, and field visits and measurement of a small subsample (355) of the air-checked plots. This inventory allowed for estimation (with high levels of error) of a variety forest characteristics across the entire region, including forest area, size class, species composition, and volume.

In the 1980's, a four-phase, multilevel inventory design (the Alaska Integrated Resource Inventory System (AIRIS)) was implemented in the Tanana Valley of Interior Alaska. This used satellite and aerial imagery to interpret vegetation class, volume, stand size class, foliar cover by vegetation class, understory component, tree crown diameters, tree and (or) shrub heights, and land use. Ground plots were also implemented to collect standard timber inventory variables [8].

More recently, plots were established in southeast and south-central Alaska between 1995 and 2003 [9]–[11]. The annual coastal inventory started in 2004, with a total of 2,227 plots collected between 2004 and 2013 [12]. This inventory only covered a small portion of Alaska's forests, leaving out approximately 112 million acres of interior Alaska, or 15% of forested land in the United States. This area was not included in surveying due to remoteness, size, lack of transportation infrastructure, complex logistics, and cost [13]. In 2014, a pilot project was established to begin the work of surveying interior Alaska, starting with the Tanana Valley of Interior Alaska. This area included the Tanana Valley State Forest and Tetlin National Wildlife Refuge which were sampled at a 1:4 intensity (or 1 plot per 24,000 acres) on a hexagonal grid [14].

One key part of FIA plot protocol is the mapping of condition classes within each plot. The condition class attributes recorded for each plot describe the forest structure, composition, and

disturbance history. This information allows for the FIA program to estimate and account for changes in forest land. A basic requirement of this estimation and accounting is to report the current status of forest lands by forest type. Forest type is a measure of the dominant stocking of live trees that are not overtopped at a given subplot [16].

These forests play an important role in the global carbon cycle, thus monitoring them is critical to our understanding and quantification of the rate of climate change and its impacts on forest systems [19], [20]. In order to better quantify carbon stored in forests, it is important that we determine not only the volume of woody material found in forests, but also the species of that material. This allows for more accurate quantification of the carbon stored by forests [21].

In 2014, NASA flew the Goddard's LiDAR, Hyperspectral & Thermal Imager (G-LiHT) sensor in 250-meter-wide strips spaced 9.3 km apart over the Tanana valley of central Alaska, covering a variety of state and federal lands, including many recently installed FIA plots [14]. The G-LiHT unit simultaneously collects light detection and ranging (LiDAR), Hyperspectral, and Thermal data with three separate sensors integrated with GPS [22]. LiDAR has been shown to be an extremely useful tool for characterizing forest structure across landscapes [23], [24]. Hyperspectral imaging can be used to differentiate the spectra of different land cover types [25]–[27].

The goal of this study is to develop a methodology for classifying US Forest Service Forest Inventory and Analysis (FIA)-defined forest type using a combination of hyperspectral and LiDAR data over central Alaska's forests. A model will be trained on the available FIA subplot forest type data, alongside hyperspectral and LiDAR remotely sensed data. 632 subplots fall within the study area and intersect with the available hyperspectral and LiDAR data. All of these plots will be used for training and validating the models. Five different classification algorithms and 6 model input

combinations of LiDAR and Hyperspectral data were tested to compare their accuracies and determine the best classification algorithm and model inputs to use across USFS land where FIA forest type information is available. With this methodology, we were able to achieve 77.53% overall prediction accuracy.

2.3 MATERIALS

2.3.1 *Study Location*

This study was conducted within the Tanana Inventory Unit (TIU) located within the Tanana Valley of Interior Alaska (Figure 2.1). This region is 136,482 km^2 in area, dominated by black spruce (*Picea mariana*), white spruce (*Picea glauca*), tamarack (also referred to as larch, *Larix laricina*), Alaska paper birch (*Betulapapyrifera var. neoalaskana*), quaking aspen (*Populus tremuloides*), and balsam poplar (*Populus balsamifera*). These forests frequently have a significant non-tree vegetation component which includes shrubs such as dwarf birch (*Betula spp.*) and willow (*Salix spp.*) [28]–[32].

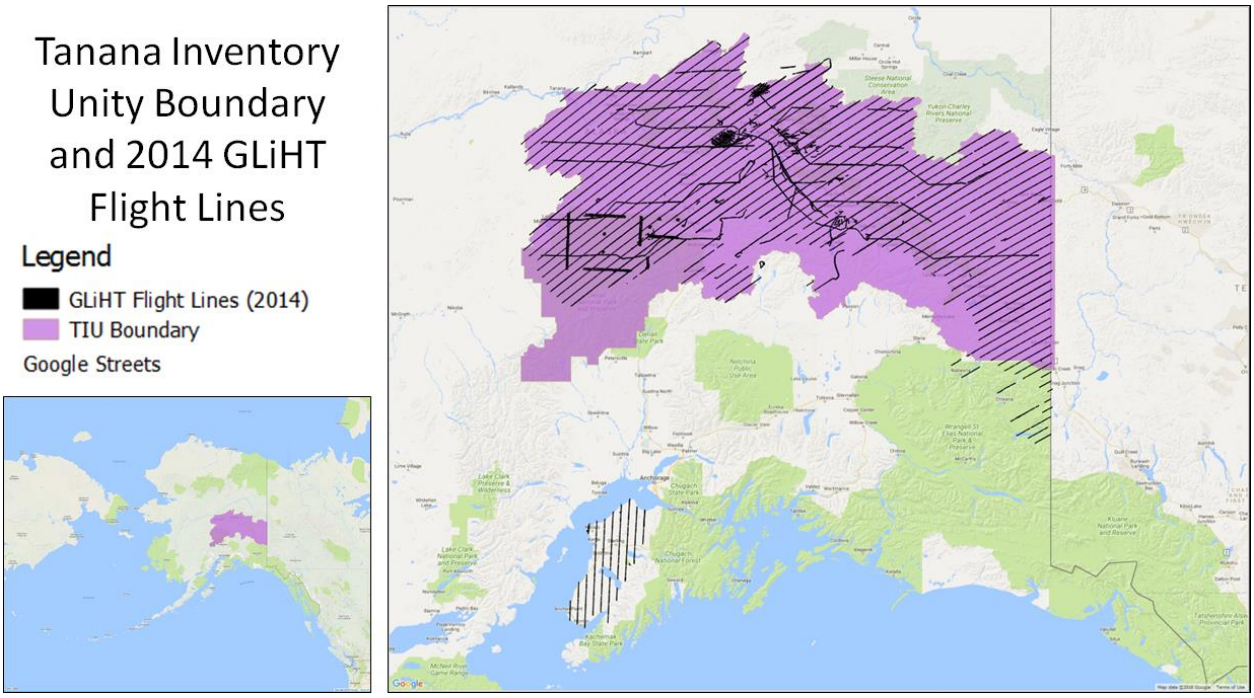

Figure 2.1. Tanana Inventory Unit Boundary and 2014 G-LiHT Flight Lines.

2.3.2 *FIA Field Data*

Much of Alaska's vast, inland forests have not been included in this census as they are extremely remote, with much of the area being accessible only by helicopter. In 2014 the US Forest Service began the long and arduous process of installing FIA plots throughout Alaska, starting with plots within the Tanana Inventory Unit (TIU) (Figure 2.1).

The modern Enhanced FIA Program implements a multi-phase sampling design. The first phase uses ancillary data, such as satellite or aerial imaging, is used to stratify the land area by cover. The second phase is the installation and measurement of permanent ground plots, so long as any portion of the plots contain forest land use. The plots are typically established along a hexagonal grid, wherein there is one ground plot in each 6000-acre hexagon. The plots are laid out such that there is one central subplot, and 3 outer subplots that are each 120 ft. from center-to-center (Figure 2.2). In the 2014 pilot project in the Tanana Valley of Interior Alaska the plots were

sampled along a 24,000-acre hexagon grid due to the cost and accessibility issues associated with these plots [14].

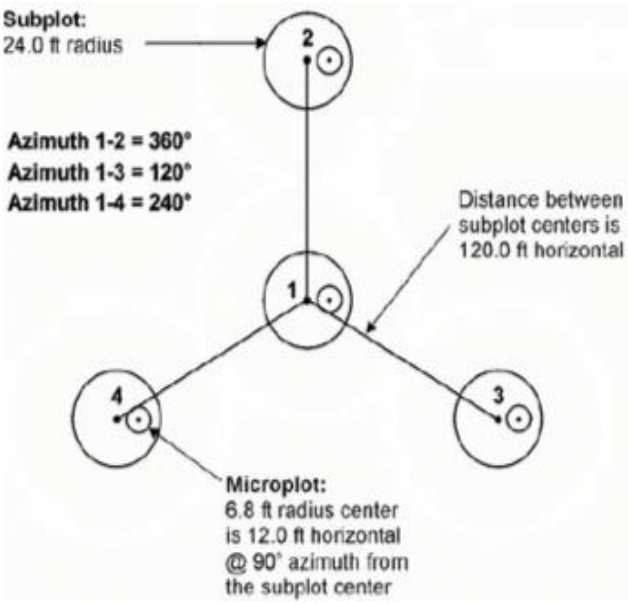

Figure 2.2. FIA Subplot and Micro plot layout (obtained from the Coastal Alaska FIA manual).

This study focuses on the Forest Type condition class information at each subplot. For each subplot, field crews recorded the forest type code (Table 2.1) associated with the dominant stocking of live trees that are not overtopped. This essentially records the dominant tree species present at a given plot. If there are no trees present or if the trees do not meet a specific stocking threshold, then the plot is marked as non-forest [16]. Additionally, survey grade GPS coordinates were collected at each FIA subplot center within the TIU, ensuring the data could be accurately aligned with remotely sensed data.

Table 2.1. Each forest type is indicated by a forest type code, the key for which can be seen here.

Code	Species Common Name	Species Scientific Name
122	White spruce	Picea glauca
125	Black spruce	Picea mariana
126	Tamarack	Larix laricina
264	Pacific silver fir	Abies amabilis
268	Subalpine fir	Abies lasiocarpa
270	Mountain hemlock	Tsuga mertensiana
271	Alaska-yellow-cedar	Cupressus nootkatensis
281	Lodgepole pine	Pinus contorta
301	Western hemlock	Tsuga heterophylla
304	Western redcedar	Thuja plicata
305	Sitka spruce	Picea sitchensis
703	Cottonwood	Populus L.
901	Aspen	Populus tremuloides
902	Paper birch	Betula papyrifera
904	Balsam poplar	Populus balsamifera
911	Red alder	Alnus rubra

2.3.3 *Data Pre-Processing*

In 2014, at the same time that field crews were on the ground installing FIA plots, NASA flew the Goddard's LiDAR, Hyperspectral and Thermal Imager (G-LiHT) sensor in strips over the TIU, covering most of the recently installed FIA plots. The G-LiHT unit simultaneously collects LiDAR, Hyperspectral, and Thermal data with three separate sensors integrated with a high precision positioning system [22].

This study focuses on the Hyperspectral and LiDAR data collected as part of this acquisition. The hyperspectral data was pre-processed by a team at NASA, resulting in 114 bands equally spaced with 4 nm distance between 418 and 918 nm at 1 m spatial resolution. This preprocessing was performed as a data and noise reduction step. Radiometric calibration of the data was also performed [22]. LiDAR point clouds were processed in FUSION to obtain digital terrain models (DTM) from the ground-classified points and canopy height models (CHM) from the highest point in each grid cell, each at 1 meter resolution [33].

2.4 METHODS

The DTM was processed to obtain topography metrics using the gdaldem function within the gdalUtils package in R [34]. The metrics obtained using this function were: slope, aspect, topographic roughness index (TRI), topographic position index (TPI), and roughness [35].

The CHM was processed to obtain canopy metrics using the GridSurfaceStats command line utility in FUSION [33]. This resulted in the metrics max height, potential volume, surface area ratio, surface volume, and surface volume ratio.

The hyperspectral data was processed in R to obtain a variety of hyperspectral vegetation indices, each of which describe a different and unique spectral characteristic of vegetation. In total 27 metrics were tested in order to ensure they describe all unique characteristics of each forest type. EVI, ISR, ISR (NDRE), NDRE, NDVI, SR, SR (RE), VARI, and VIgreen can be used to describe vegetation structure such as the amount of green biomass, and leaf area index. PSND (a and b), PSRI, PSSR, and SIPI can be used to describe biochemical pigments that make up vegetation. CARI, CI (RE), CSM, CSM (RE), Datt, and MCARI have been found to describe chlorophyll content in plants. ACI, ARI, and mARI describe anthrocyanin content. CRI (1 and 2) describe carotenoid content (Table 2.2).

Table 2.2. Hyperspectral vegetation indices used as data inputs for classification algorithms tested in this study. Adapted from table 14.1 found in Thenkabail & Lyon, 2016 [36].

Index	Equation	Reference
Raw Hyperspectral Bands		
Red (R_{Red})	R_{650}	
Green (R_{Green})	R_{550}	
Blue (R_{Blue})	R_{475}	
NIR (R_{NIR})	R_{710}	
Red Edge ($R_{Red\ Edge}$)	R_{800}	
Metrics Describing Structure (Green Biomass, Leaf Area Index, etc.)		
EVI	$\frac{2.5 \times (-R_{Red})}{(R_{NIR} + (6 \times R_{Red}) - (7.5 \times R_{Blue}) + 1)}$	[37]
ISR	$\frac{1 - NDVI}{1 + NDVI}$	[38], [39]
ISR $_{NDRE}$	$\frac{1 - NDRE}{1 + NDRE}$	[39]
NDRE	$\frac{R_{NIR} - R_{Red\ Edge}}{R_{NIR} + R_{Red\ Edge}}$	[40]
NDVI	$\frac{R_{NIR} - R_{Red}}{R_{NIR} + R_{Red}}$	[41]
SR	$\frac{R_{NIR}}{R_{Red}}$	[42]
SR $_{Red\ Edge}$	$\frac{R_{NIR}}{R_{Red\ Edge}}$	
VARI	$\frac{R_{Green} - R_{Red}}{R_{Green} + R_{Red} - R_{Blue}}$	[43]
VIgreen	$\frac{R_{Green} - R_{Red}}{R_{Green} + R_{Red}}$	[43]
Metrics Describing Biochemical Characteristics		
Pigments		
PSND $_a$	$\frac{R_{Red\ Edge} - R_{675}}{R_{Red\ Edge} + R_{675}}$	[44]
PSND $_b$	$\frac{R_{Red\ Edge} - R_{Red}}{R_{Red\ Edge} + R_{Red}}$	[44]
PSRI	$\frac{R_{680} - R_{500}}{R_{750}}$	[45]
PSSR $_a$	$\frac{R_{Red\ Edge}}{R_{675}}$	[46]
PSSR $_b$	$\frac{R_{Red\ Edge}}{R_{Red}}$	[46]
SIPI	$\frac{R_{Red\ Edge} - R_{445}}{R_{Red\ Edge} - R_{680}}$	[47]
Chlorophyll		
CARI	$(R_{700} - R_{670}) - 0.2 \times (R_{700} - R_{550})$	[48]

CI $_{\text{Red Edge}}$	$\frac{R_{NIR}}{R_{Red\ Edge}} - 1$	[49]
CSM	$\frac{R_{Red}}{R_{NIR}}$	[50]
CSM $_{\text{Red Edge}}$	$\frac{R_{Red\ Edge}}{R_{NIR}}$	[50]
Datt	$\frac{R_{NIR} - R_{Red\ Edge}}{R_{NIR} - R_{Red}}$	[51]
MCARI	$\left((R_{700} - R_{670}) - 0.2\ \times (R_{700} - R_{550})\right) \times \frac{R_{700}}{R_{670}}$	[52]
Anthrocyanins		
ACI	$\frac{R_{Green}}{R_{NIR}}$	[53]
ARI	$\frac{1}{R_{Green}} - \frac{1}{R_{Red\ Edge}}$	[54]
mARI	$(\frac{1}{R_{Green}} - \frac{1}{R_{Red\ Edge}}) \times\ R_{NIR}$	[49]
RGRI	$\frac{R_{Red}}{R_{Green}}$	[55]
Carotenoids		
CRI1	$\frac{1}{R_{510}} - \frac{1}{R_{550}}$	[56]
CRI2	$\frac{1}{R_{510}} - \frac{1}{R_{700}}$	[56]

2.4.1 *Averaging Data over Subplots*

All DTM and CHM metrics were averaged over each circular subplot (7.3-meter radius). When aggregating the hyperspectral indices, shadow pixels were eliminated by removing pixels that fell into the lower 10% quantile for the hyperspectral indices Normalized Difference Vegetation Index (NDVI), photochemical reflectance index (PRI), or Red Edge normalized difference vegetation index (RENDVI) [57], [58]. These pixels were marked and removed in all vegetation indices when averaging over each circular subplot. All of these resulted in average DTM and CHM metrics, and Hyperspectral indices for each FIA subplot (Figure 2.3).

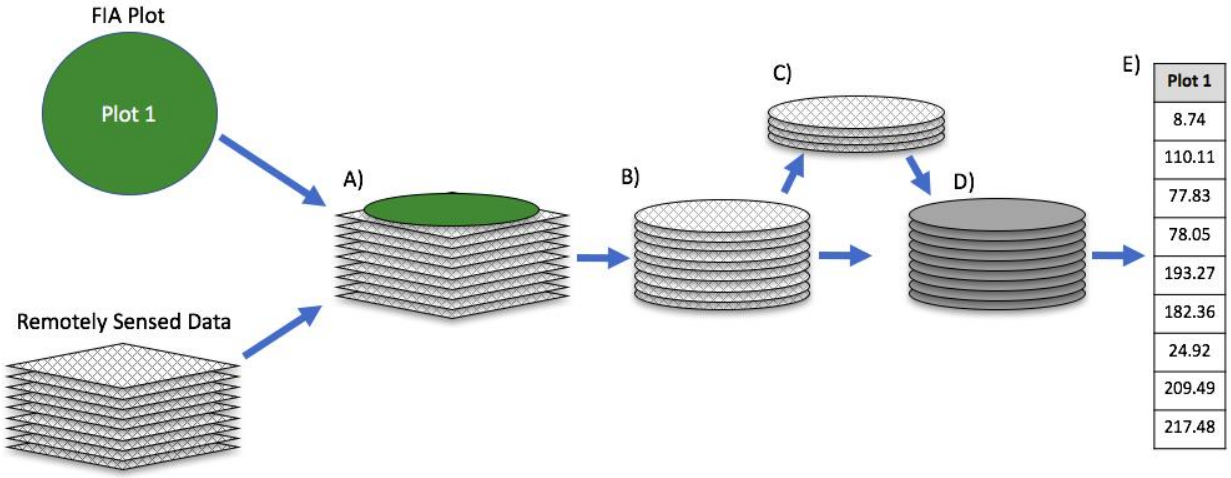

Figure 2.3. Aggregating the values of the LiDAR and Hyperspectral data contained within each subplot boundary started by overlaying the subplot boundary on top of the data (A). Then, the plot boundary was used to clip the data to only include the pixels whose cell center was within the boundary of the subplot (B). Then, shadow pixels were eliminated from the hyperspectral vegetation index layers by removing pixels that fell into the lower 10% quantile for the hyperspectral indices Normalized Difference Vegetation Index (NDVI), photochemical reflectance index (PRI), or Red Edge normalized difference vegetation index (RENDVI) (C). All included pixels from the hyperspectral and LiDAR data that fell within the subplot boundary were averaged (D) and inserted into a data frame (E).

2.4.2 *Data Subsetting*

In total, 632 FIA subplots are covered by the G-LiHT data used in this study. The nested subplot design of FIA plots leads to an issue of pseudo replication. Subplots that lie within the same plot are only 120 ft. from one another, and thus are more likely to be spectrally and texturally (both in the canopy and topography) similar to one another than plots in different parts of the forest. Thus, the models are more likely to classify the forest types of these subplots correctly, even if they trained on a different subplot within the same plot.

Most studies that focus on the FIA subplots overcome this issue by selecting only one subplot per plot, and eliminating the remaining plots from the analysis [59], [60]. In this study, we came to realize that when this was done, all forest types could not be represented in both the

training and validation data as there were insufficient subplots to ensure that both training and validation subsets had at least one subplot of each forest type (Table 2.3).

Table 2.3. Number of subplots per forest type.

Forest Type	Number of Subplots
Aspen	23
Balsam Poplar	7
Black Spruce	280
Cottonwood	12
Paper Birch	175
Tamarack	4
White Spruce	93
Non-forest	38

The way to overcome this would be to replicate these subplots so that the same subplots could be used in validation and training, to ensure that all species were represented. But, this direct replication of the data seemed even less desirable than the pseudo-replication created by the nested plot design. Thus, it was concluded that all subplots would be used in this analysis, and that it must be recognized that pseudo replication is present and thus these results will be an overestimation of the true accuracy of this methodology. Until there is a sufficient number of subplots of each forest type, this issue cannot be addressed further. Bootstrap aggregating, or bagging, was tested, but did not improve model performance [61].

2.4.3 *Models*

The 5 classification algorithms tested in this study were: Random Forest, Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Naive Bayes Classifier (NBC), and Multinomial Logistic Regression (Table 2.4). These five algorithms were chosen because they are commonly used in ecological classification studies, but have not been simultaneously compared in any other known ecological study. Random forest is a machine learning algorithm that uses an ensemble of decision trees that are split at each node using a random subset of input predictors. This is repeated

a set number of times and used to create models for prediction of classes [62], [63]. Support vector machines use hyperplanes in an n-dimensional space to separate and distinguish distinct classes in a dataset [64], [65]. K-Nearest Neighbor classifiers find the *k* closest training vectors in Euclidean space to the validation set, and determine the class based on a majority vote of the *k* training vectors [66], [67]. A naive Bayes classifier is a probability-based classifier that uses Bayes' theorem and assumes conditional independence and a Gaussian distribution in the predictor metrics. The model computes the conditional posterior probability of the response and uses that to perform classifications [68]–[71]. Multinomial logistic regressions use a logistic regression to predict the probability of different classes given a set of independent predictor variables [72]. In this study, all Multinomial Logistic Regression models were fit using neural networks [72].

Table 2.4. Classification algorithms used and their associated R functions and packages, as well as the run specifications that were used with that function when running the models.

Classification Algorithm	**R Function**	**R Package**	**Model Run Specifications**
Random Forest	randomForest	randomForest	ntrees = 5000
Support Vector Machine (both Tuned and Untuned)	svm	e1071	Tuning parameters were set to values between cost=$10^{(-1:2)}$ gamma= c(.5,1,2) The best parameters were selected and used in the final models.
K-Nearest Neighbor	knn	class	K values between 0 and 100 were tested, the value that resulted in the most accurate predictions was used in the final models.
Naive Bayes Classifier	naiveBayes	e1071	Default
Multinomial Logistic Regression	multinom	nnet	Default

These five classification algorithms were tested to determine the best algorithm for classifying forest type from hyperspectral and LiDAR data. The overall and kappa prediction accuracies for each model were compared in order to determine the best model for the data. Overall

prediction accuracy is the accuracy is defined as the rate at which the model correctly predicted a given class, or forest type in this case. It is calculated by taking the sum of the total number of correct predictions divided by the total number of predictions made. Kappa accuracy, or the kappa coefficient of agreement, is a commonly used metric in remote sensing that takes into account the expected error rate. Kappa is calculated by subtracting the expected accuracy (E) from the observed accuracy (O), and dividing that by 1 minus the expected accuracy ($\frac{O-E}{1-E}$) [73]. The kappa statistic is highly controversial in the field of remote sensing, which is why it was used in combination with overall accuracy in this study [74].

When necessary, parameters within the models were tuned in order to improve accuracy of the models. In the case of Random Forest models, 5000 trees were used instead of the default 500 in order to boost model performance and obtain a more consistent estimation of variable importance between different model runs [75]. When tuning the Support Vector Machine, a variety of tuning parameters were tested (cost values between 10^{-1} and 10^{2} and gamma values .5, 1, 2). The best parameters were selected from the tuning function and used in the final models with a radial kernel. In the K-Nearest Neighbor models, k values between 0 and 100 were tested in order to determine the k value that resulted in the highest prediction accuracy. This k value was individually selected and used for each model. In the case of the Naive Bayes Classifier and Multinomial Logistic Regression models, all default parameters were used.

All available subplot data was used in a 3-fold cross validation in order to ensure that all data was included in both validation and training. The 3-fold cross validation used ~33% of the data for validation and ~66% for training, and then was rotated so that different thirds of the data were used in 3 different models for each modeling algorithm tested. These models were used to make predictions over the validation data, and compared with the known values for each of the 3

models **(**Figure 2.4). This ensures that a prediction was made on every data point within the model, without overfitting the models by validation and training on the same data. In total, there were 15 separate models (3 models for each of the 5 classification algorithms) which were compared for each response input.

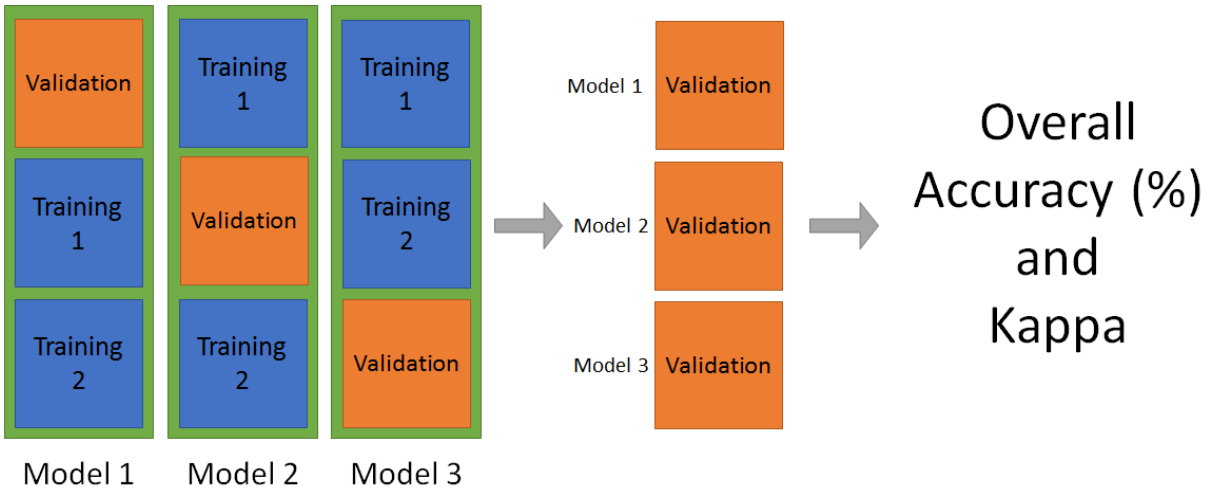

Figure 2.4. This figure illustrates the 3-fold cross validation performed as part of this study. The data was broken up into thirds such that ⅓ was used for validation and ⅔ were used for training. The validation and training always included different data, but was cycled in 3 models so that all parts of the data were included in some validation, allowing for unbiased prediction of forest type over the entire dataset.

2.4.4 *Predictor Inputs*

For each classification algorithm tested, 6 predictor data inputs were also tested, excluding the Multinomial Logistic Regression which had 5 inputs. These inputs were tested in order to determine which data or combination of data best predicted forest type.

The predictor inputs tested were different combinations of the available data for each plot: raw hyperspectral bands, hyperspectral vegetation indices, DTM metrics, and CHM metrics. In the first predictor input group, all available data for the subplots were input into the model. This group was not tested in the Multinomial Logistic Regression models as there were too many predictor inputs, causing an error when attempting to run the model. In the second predictor group, most of

the raw hyperspectral bands (excluding one band each for red, green, blue, near infrared, and red edge) were removed and included only the hyperspectral indices and DTM and CHM metrics. The other four predictor groups consisted of the individual predictor data groups: raw hyperspectral bands, hyperspectral vegetation indices and 5 raw bands (one band each for red, green, blue, near infrared, and red edge; all are listed in Table 2.2), DTM metrics, and CHM metrics (Figure 2.5 and Figure 2.6).

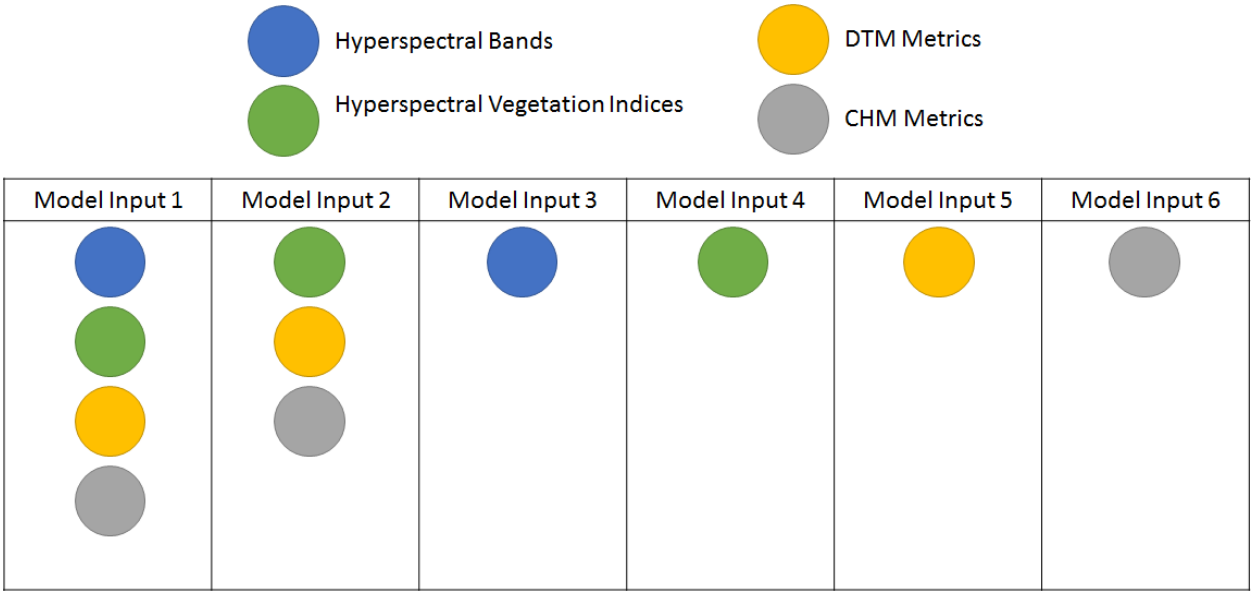

Figure 2.5. Six model predictor data inputs were tested for each modeling algorithm tested. These predictor inputs were different combinations of the available data for each plot: raw hyperspectral bands, hyperspectral vegetation indices, DTM metrics, and CHM metrics. In the first predictor input group, we tested all available data for the subplots. This group was not tested in the Multinomial Logistic Regression models as there was too many predictor inputs, causing an error when attempting to run the model. In the second predictor group, we removed most of the raw hyperspectral bands (excluding one band each for red, green, blue, near infrared, and red edge; all are listed in Table 2.2) and included only the hyperspectral indices and DTM and CHM metrics. The other four predictor groups consisted of the individual predictor data groups: raw hyperspectral bands, hyperspectral vegetation indices, DTM metrics, and CHM metrics.

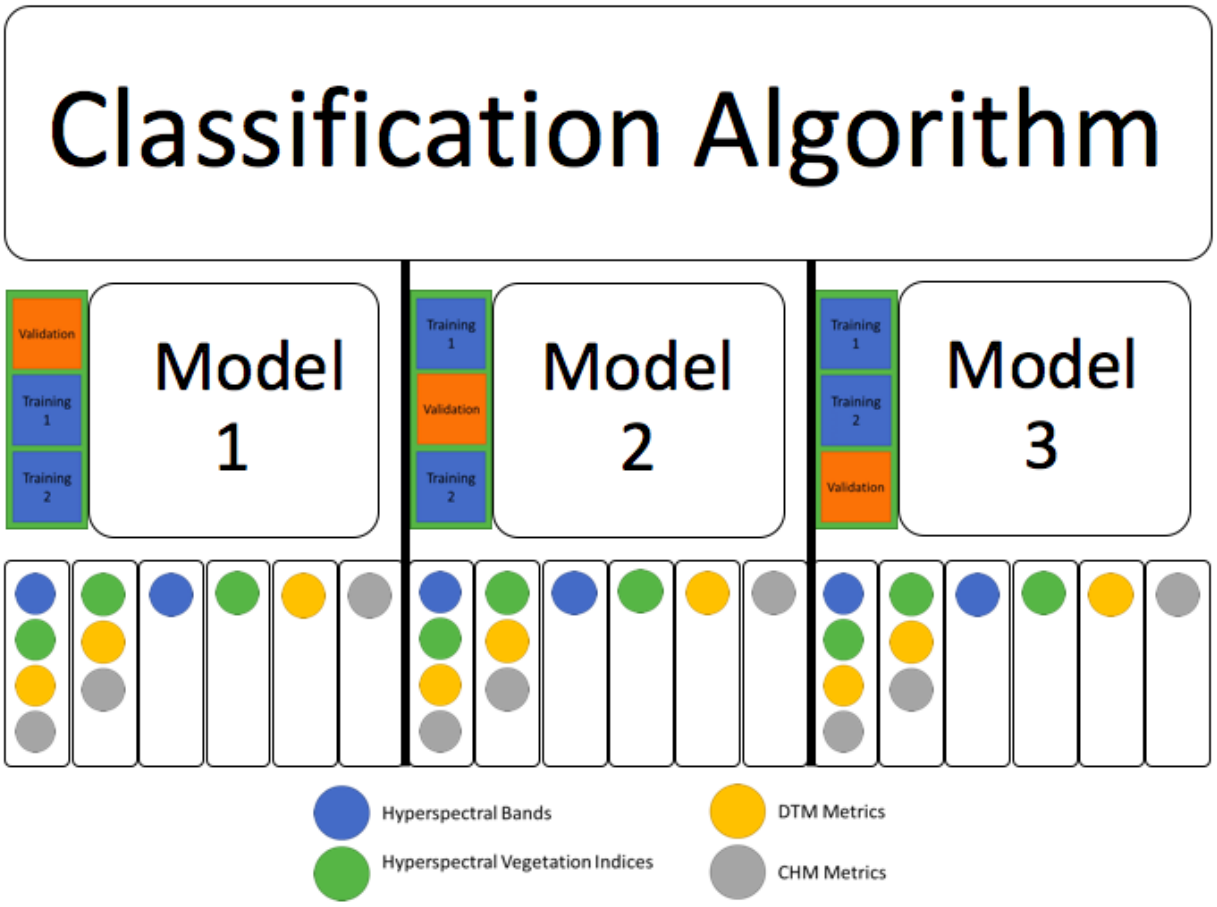

Figure 2.6. This shows all the models inputs and 3-fold validation process undertaken for each classification algorithm tested. For each algorithm tested, a 3-fold cross validation was performed. Within that 3-fold cross validation, each model tested 6 predictor data inputs, giving a total of 18 models for each classification algorithm tested.

2.4.5 *Important Predictor Variables from Random Forest Model*

Random forest estimates the importance of a predictor variable by determining how much out of bag (OOB) error increases when data for that variable is permuted while others are left unchanged [62]. There are multiple ways that one can measure variable importance; this study used mean decrease in accuracy to measure variable importance rather than mean decrease in Gini [75].

In order to assess overall variable importance, the predictor input group which resulted in the highest accuracy and response data for all subplots were input into a single Random Forest

model. The three-fold cross validation was not necessary as this model was not being used for assessing prediction accuracy, rather it was only used for assessing variable importance. In this final model, 15,000 trees were used in order to ensure extremely stable estimates of variable importance [62], [75].

2.5 RESULTS

In comparing model performance for each model with each predictor input group, the overall prediction accuracy and kappa value were used to evaluate the model performance. Model accuracy was calculated by comparing the predicted forest types to the actual forest types over the validation datasets. Kappa was used because it not only compares the agreement between actual and predicted forest types, but also accounts for the chance that the agreement was a product of random chance [76].

2.5.1 *Random Forest*

Random forest with hyperspectral vegetation indices and DTM and CHM metrics as model inputs had the highest accuracy and kappa of all the classification algorithms and model inputs tested, at an overall accuracy of 77.53% and a kappa of 0.66 (Table 2.5). The kappa value of 0.66 indicates substantial agreement between the known and predicted forest types for each subplot (Table 2.6) [77].

Table 2.5. Random Forest Model accuracy and kappa values.

Model Inputs	Accuracy	Kappa
Hyperspectral Bands and Vegetation Indices, with DTM and CHM Metrics	75.16%	0.62
Hyperspectral Vegetation Indices, with DTM and CHM Metrics	**77.53%**	**0.66**
CHM Metrics	69.30%	0.54
Hyperspectral Vegetation Indices	67.09%	0.50
Hyperspectral Bands	44.78%	0.23

DTM Metrics	65.03%	0.47

Table 2.6. Interpretation of kappa values from Landis, & Koch, 1977.

Kappa Statistic	Strength of Agreement
< 0.00	Poor
0.00-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Almost Perfect

When all available data were included in the model (Hyperspectral Bands and Vegetation Indices, with DTM and CHM Metrics), the results were also promising with accuracy reaching 75.16%. In this model, the kappa value, 0.62, also indicated substantial agreement between known and predicted forest type. The other model inputs tested were the individual model inputs: CHM Metrics, Hyperspectral Vegetation Indices, Hyperspectral Bands, and DTM Metrics. These models all resulted in moderate accuracies, with CHM metrics resulting in the highest accuracy at 69.30% overall and a kappa value of 0.54, indicating just moderate agreement between known and predicted forest type. The lowest accuracy and kappa was obtained when inputting only hyperspectral bands, with an accuracy of 44.78% and a kappa of just 0.23. The raw hyperspectral band data was significantly less predictive than the band-derived hyperspectral vegetation indices.

2.5.2 *Support Vector Machine (Both Tuned and Untuned)*

The Support Vector Machine (SVM) models were tested as both tuned and untuned models. The untuned model performed poorly, with the highest accuracy at just 50.32% and the associated kappa value at just 0.32, indicating fair agreement (Table 2.7). Similar to what was found in the Random Forest model, the highest accuracy model included Hyperspectral Vegetation Indices, with DTM and CHM Metrics.

Table 2.7. Untuned Support Vector Machine model accuracy and kappa results.

Model Inputs	Accuracy	Kappa

Hyperspectral Bands and Vegetation Indices, with DTM and CHM Metrics	46.84%	0.27
Hyperspectral Vegetation Indices, with DTM and CHM Metrics	50.32%	0.32
CHM Metrics	45.73%	0.20
Hyperspectral Vegetation Indices	43.83%	0.21
Hyperspectral Bands	43.04%	0.20
DTM Metrics	45.73%	0.20

In contrast, the tuned SVM model had a relatively high accuracy in the "best" model with 73.89% accuracy and a kappa of 0.60 (Table 2.8). Once again, the highest accuracy model was the one which included the model inputs Hyperspectral Vegetation Indices, with DTM and CHM Metrics. The second most accurate SVM model was the model with Hyperspectral Vegetation Indices as inputs at 64.72%, with a kappa value indicating moderate agreement.

Table 2.8. Tuned Support Vector Machine model accuracy and kappa results.

Model Inputs	Accuracy	Kappa
Hyperspectral Bands and Vegetation Indices, with DTM and CHM Metrics	45.57%	0.10
Hyperspectral Vegetation Indices, with DTM and CHM Metrics	**73.89%**	**0.60**
CHM Metrics	47.94%	0.26
Hyperspectral Vegetation Indices	64.72%	0.47
Hyperspectral Bands	43.83%	0.20
DTM Metrics	48.10%	0.27

2.5.3 *K-Nearest Neighbor*

The K-Nearest Neighbor (KNN) models had moderate overall accuracy at ~64% overall with almost all model inputs (Table 2.9). DTM metrics had the lowest accuracy of any KNN model, at just 48.10% accuracy. The most accurate model included only hyperspectral vegetation indices, but there not a significant difference in accuracy between this model input and others (excluding DTM metrics).

Table 2.9. K-Nearest Neighbor model accuracy and kappa results.

Model Inputs	Accuracy	Kappa

Hyperspectral Bands and Vegetation Indices, with DTM and CHM Metrics	64.87%	0.49
Hyperspectral Vegetation Indices, with DTM and CHM Metrics	63.61%	0.45
CHM Metrics	64.08%	0.47
Hyperspectral Vegetation Indices	65.98%	0.49
Hyperspectral Bands	64.56%	0.47
DTM Metrics	48.10%	0.28

2.5.4 *Naive Bayes Classifier*

Although maximum likelihood models are standard in the classification-focused studies, in this case the Naive Bayes Classifier (NBC) model performed the worst overall. Of all NBC model inputs tested, the model that included Hyperspectral Vegetation Indices, with DTM and CHM Metrics was the most accurate at 53.80% overall accuracy and a kappa of 0.38 (Table 2.10). The worst performing model included only hyperspectral bands and had an overall accuracy of just 19.62% and a kappa of 0.08.

Table 2.10. Naive Bayes Classifier model accuracy and kappa results.

Model Inputs	Accuracy	Kappa
Hyperspectral Bands and Vegetation Indices, with DTM and CHM Metrics	44.78%	0.28
Hyperspectral Vegetation Indices, with DTM and CHM Metrics	53.80%	0.38
CHM Metrics	21.52%	0.13
Hyperspectral Vegetation Indices	48.89%	0.31
Hyperspectral Bands	19.62%	0.08
DTM Metrics	37.66%	0.18

2.5.5 *Multinomial Logistic Regression*

The Multinomial Logistic Regression model (MLR) came in third place for highest overall accuracy at 71.20% (Table 2.11). The highest accuracy models were obtained with Hyperspectral Vegetation Indices, with DTM and CHM Metrics as model inputs. This is consistent with most other classification algorithms tested in this study. The second most accurate model input was

hyperspectral vegetation indices, which was also commonly observed as the second most accurate input in other models. Overall, the kappa values indicate a moderate to substantial agreement.

Table 2.11. Multinomial Logistic Regression model accuracy and kappa results.

Model Inputs	Accuracy	Kappa
Hyperspectral Vegetation Indices, with DTM and CHM Metrics	71.20%	0.58
CHM Metrics	48.26%	0.27
Hyperspectral Vegetation Indices	68.04%	0.52
Hyperspectral Bands	45.41%	0.25
DTM Metrics	57.59%	0.32

2.5.6 *Most Important Predictor Variables from Random Forest Model*

Random forest with hyperspectral vegetation indices and DTM and CHM metrics as model inputs had the highest overall and kappa accuracy of all the classification algorithms and model inputs tested. Thus, a single model was produced using all subplots in order to assess variable importance. Figure 2.7 shows the variable importance plot produced from this model. This plot shows that many of the predictor variables are highly correlated, due to their similarity in variable importance.

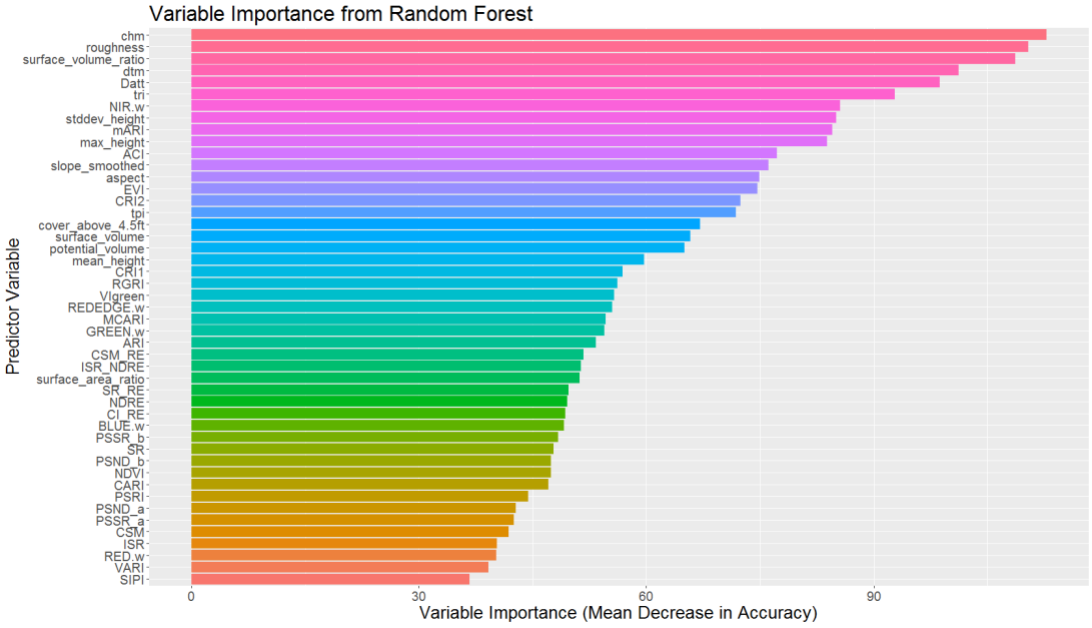

Figure 2.7. Variable Importance Plot for model with all data as input.

The canopy height model (CHM) values were shown to be the most important predictor, with roughness, surface volume ratio, and elevation (DTM) following close behind. These were followed by the Datt hyperspectral vegetation index [51] and topographic roughness index (TRI). It should be noted that many of these predictor variables are highly correlated, thus this may not be truly representative of which predictors are most important.

2.5.7 *Individual Class Prediction Accuracies*

The three most accurate models were the Random Forest, SVM, and MLR models which included hyperspectral vegetation indices and DTM and CHM metrics. For random forest, the model confusion matrix (Table 2.12) revealed that the model predicts black spruce forest type most accurately, and balsam poplar least accurately. It also frequently confuses white spruce forest type subplots with black spruce and paper birch. Paper birch was also frequently predicted accurately at 83.43% overall accuracy. Black Spruce and Paper Birch are the most common forest types, with white spruce following close behind.

Table 2.12. Confusion Matrix for Random Forest model with Hyperspectral Vegetation Indices, with DTM and CHM Metrics.

		Reference							
		Non-Forest	White Spruce	Black Spruce	Tamarack	Cottonwood	Aspen	Paper Birch	Balsam Poplar
PREDICTION	Non-Forest	**52.63%**	2.15%	0.71%	25.00%	0.00%	0.00%	0.57%	14.29%
	White Spruce	2.63%	**51.61%**	1.43%	0.00%	25.00%	13.04%	3.43%	28.57%
	Black Spruce	23.68%	20.43%	**93.57%**	50.00%	8.33%	34.78%	12.00%	14.29%
	Tamarack	0.00%	1.08%	0.00%	**25.00%**	0.00%	0.00%	0.00%	0.00%
	Cottonwood	0.00%	0.00%	0.00%	0.00%	**58.33%**	0.00%	0.00%	0.00%
	Aspen	0.00%	0.00%	0.00%	0.00%	0.00%	**26.09%**	0.57%	0.00%
	Paper Birch	18.42%	24.73%	4.29%	0.00%	8.33%	26.09%	**83.43%**	42.86%

	Balsam Poplar	2.63%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	**0.00%**

In the case of the SVM model, black spruce was accurately predicted at 93.93% accuracy (Table 2.13). Paper birch also had high prediction accuracy at 87.43%. In contrast to the 26.09% accuracy for predicting aspect forest type with the Random Forest model, the SVM model was able to predict Aspen forest type with 52.17% accuracy. But, the SVM model had poor prediction accuracy for cottonwood at just 16.67%, whereas Random Forest achieved 58.33% accuracy.

Table 2.13. Confusion Matrix for tuned Support Vector Machine Model with Hyperspectral Vegetation Indices, with DTM and CHM Metrics.

		Reference							
		Non-Forest	White Spruce	Black Spruce	Tamarack	Cottonwood	Aspen	Paper Birch	Balsam Poplar
PREDICTION	Non-Forest	**13.16%**	0.00%	0.36%	0.00%	0.00%	0.00%	0.00%	0.00%
	White Spruce	2.63%	**33.33%**	1.07%	0.00%	0.00%	4.35%	2.29%	0.00%
	Black Spruce	23.68%	22.58%	**93.93%**	50.00%	50.00%	21.74%	10.29%	14.29%
	Tamarack	0.00%	0.00%	0.00%	**25.00%**	0.00%	0.00%	0.00%	0.00%
	Cottonwood	0.00%	0.00%	0.36%	0.00%	**16.67%**	0.00%	0.00%	0.00%
	Aspen	0.00%	0.00%	0.00%	0.00%	0.00%	**52.17%**	0.00%	0.00%
	Paper Birch	60.53%	44.09%	4.29%	25.00%	33.33%	21.74%	**87.43%**	85.71%
	Balsam Poplar	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	**0.00%**

Both Random Forest and SVM achieved high accuracies in the two most common classes, Paper Birch and Black Spruce, but had lackluster results in the less common classes, Tamarack, Cottonwood, Aspen, and Balsam Poplar. The MLR model had very different results - it excelled in predicting the least common species and achieved moderate accuracies in the more common species. The MLR model achieved 84% accuracy for Tamarack, 75% for Cottonwood, 83% for Aspen, and 70% for Balsam Poplar (Table 2.14). This contrasts the 0% accuracy achieved by Random Forest and SVM when predicting Balsam Poplar, and 25% accuracy when predicting

Tamarack (Table 2.15). Overall, MLR had more consistently accurate performance when predicting forest type, despite having a slightly lower overall accuracy when compared to the equivalent Random Forest and SVM models.

Table 2.14. Confusion Matrix from Multinomial Logistic Regression Model with Hyperspectral Vegetation Indices, with DTM and CHM Metrics.

		Reference							
		Non-Forest	White Spruce	Black Spruce	Tamarack	Cottonwood	Aspen	Paper Birch	Balsam Poplar
PREDICTION	Non-Forest	**51.79%**	4.29%	7.50%	0.00%	6.79%	0.00%	2.86%	11.43%
	White Spruce	10.71%	**50.00%**	10.36%	0.00%	0.00%	8.21%	8.57%	0.00%
	Black Spruce	13.21%	19.29%	**66.79%**	0.00%	0.00%	4.29%	5.71%	0.00%
	Tamarack	4.29%	3.21%	4.29%	**84.29%**	0.00%	0.00%	0.36%	0.00%
	Cottonwood	7.86%	2.50%	0.36%	0.00%	**75.71%**	0.00%	1.43%	0.00%
	Aspen	5.36%	6.43%	6.07%	15.71%	0.00%	**83.21%**	8.57%	0.00%
	Paper Birch	3.21%	6.07%	4.64%	0.00%	17.50%	4.29%	**67.14%**	17.86%
	Balsam Poplar	3.57%	8.21%	0.00%	0.00%	0.00%	0.00%	5.36%	**70.71%**

Table 2.15. Comparison of class and overall accuracies for Random Forest, SVM, and MLR.

Forest Type	**Random Forest**	**Support Vector Machine**	**Multinomial Logistic Regression**	**Total # of Plots**
Non-Forest	52%	13%	51%	38
White Spruce	51%	33%	50%	93
Black Spruce	**93%**	**93%**	66%	280
Tamarack	25%	25%	**84%**	4
Cottonwood	58%	16%	**75%**	12
Aspen	26%	52%	**83%**	23
Paper Birch	**83%**	**87%**	67%	175
Balsam Poplar	0%	0%	**70%**	7
Overall Accuracy	**77.53%**	73%	71.20%	

2.6 DISCUSSION

Previous studies have worked to answer similar questions and achieved promising results. One study compared the accuracies of Support Vector Machines (SVM), K-Nearest Neighbor, and Gaussian maximum likelihood with leave-one-out-covariance algorithms for the classifications of tree species using both hyperspectral and LiDAR data [78]. They found that the SVM classifier was the most accurate, yielding just over 89% kappa accuracy.

In this study, SVM was also found to yield better results than the K-Nearest Neighbor classifier, but the Gaussian maximum likelihood (GML) with leave-one-out-covariance algorithm was not tested. Instead, a naive Bayes classifier which is based on GML assumptions was tested and found to perform the worst of any model. This contradicts the findings with the GML model in [78] which found their GML algorithm to perform better than their K-Nearest Neighbor algorithm. This difference is likely due to differences in GML and KNN algorithm and model parameters used in each study.

A second study compared the results of classifying 7 tree species and a non-forest class in the Southern Alps of Italy using hyperspectral and LiDAR data with Support Vector Machines and Random Forest [79]. They found that SVM consistently outperforms Random Forest, achieving 95% overall accuracy.

This differs from our findings which revealed that Random Forest outperformed SVM models, no matter the model input, achieving 77.53% overall accuracy with the best Random Forest model, and 73.89% with the best SVM model. This is likely due to differences in data pre-processing, especially in the case of hyperspectral data wherein [79] performed a minimum noise fraction transformation in order to reduce the data.

A minimum noise fraction transformation is a common first step in working with hyperspectral data, but was not performed in this study as it could not be performed at a large scale [80]. This same rationale was used when deciding to not perform a principal components analysis (PCA) in order to reduce the hyperspectral data. The model which was found to have the best performance in this study will be used to predict forest type over the entire TIU where there is G-LiHT data coverage. This is a massive area, making it impossible to perform a PCA or any other data reduction over the entire dataset, in order to create the same predictor variables in the data swaths used for predicting forest type. All predictor variables used in this study were easy to create and can be scaled over large areas, which is why they were used here.

A second factor that may have influenced the difference in results between both [79], [78], and this study are the model parameters used. In both SVM and KNN models, users must supply parameters in order to build models. In this study, an optimization script that tried 100 different k-values for the KNN model and chose the one that yielded the highest accuracy for each model input tested was used. For the SVM, multiple cost and gamma values were tested. A script was used to choose the values that resulted in the highest accuracy.

Additionally, this study used a radial kernel for the SVM, while Dalponte's studies used different kernels (a Gaussian kernel was used in the 2008 paper, and the kernel used in the 2012 paper was not specified). All of these differences in parameters used makes for difference in final results.

One additional issue with the training data is the difference between "forest type" as defined by the FIA protocol and the true vegetation species present at a given location. Each FIA subplot collected classifies a subplot as a given forest type based on the dominant stocking present (or not present in the case of non-forest plots), rather than giving a percentage of each species

present in a plot. This means that in any given subplot, the dominant stocking species present could theoretically be present on anywhere between 51-100% of the plot, while the other 0-49% could be any other tree species or vegetation type. A subplot that is 51% white spruce is spectrally very different from a plot that is 100% white spruce, but we are treating both of these situations as being the same class. This is not ideal due the difference in spectral profile, and likely contributed to our error rate.

Unlike the potential loss of accuracy due to the spectral mixing within subplots, the nested plot design in this study likely inflated overall accuracy. The FIA protocol is designed such that for each plot, there are four subplots on which forest type data is collected. There is just 120 ft. between the center subplot and the surrounding 3 subplots, leading to an issue of pseudo replication and spatial-autocorrelation. These subplots are so close together that they are likely to include similar vegetation and thus have similar hyperspectral and LiDAR data values. But in this study, each subplot was treated as if it is a unique and independent of others. This likely lead to an inflation in accuracy which cannot be overcome until a more balanced dataset with more equal representation of all forest types within the TIU is available.

Imbalanced datasets are common in ecological studies due to the nature of natural datasets. There are multiple ways that one can attempt to deal with imbalanced datasets. One method is to under sample the dataset so that there is an equal number of samples in each class that is equal to or less than the size of the smallest class, in order to ensure that the number of samples of each sample is the same. This was not feasible in this study as there were some classes that had less than 10 samples (Tamarack and Balsam poplar), with the smallest class having just 4 samples. Providing only 4 samples per class would not lead to good results, so under sampling was not performed.

A second method of overcoming this issue is to oversample the dataset so that each class is sampled with replacement until it reaches a quantity that is typically greater than or equal to the largest class in the dataset. This same methodology can be performed in an ensemble called bootstrap aggregating, or bagging. Bagging is the process by which one samples a dataset *D* of length *n* with replacement, in order to generate a new dataset *D1* of length *m* wherein each class is equally represented [61].

In this case, bagging was implemented in all models in order to determine how this would impact overall and kappa accuracy. Each class was sampled with replacement until we reached a set equal in length to the largest class within the dataset (280 data points). In most cases, bagging did not improve overall or kappa accuracy. In the cases where accuracy was improved, most models were already performing so poorly that the improved accuracy still did not make the algorithm or model inputs a contender for the best algorithm and inputs. Thus, bagging was not included as a part of this study.

2.6.1 *Future Research*

Pseudo replication and spatial autocorrelation caused by the nested plot design used in this study likely influenced the results in this study. This problem was unavoidable due to the limited availability of data for certain forest types. In order to better overcome this, this study suggest that future research collect more field validation data and ensure that there is a balance of all forest types present in the study area. This study also suggests that if a nested plot design is used, just one subplot should be selected from each plot to represent the entire plot. Additionally, training and validation datasets should be selected such that they are in spatially distinct regions, rather than performing a random selection.

Future studies may also consider exploring variable importance that takes into account highly correlated variables. Many of the predictor variables included in our final model were highly correlated, making it challenging to say which variables were truly the most important. It is also important to note that the variables used in this study are not the only LiDAR and hyperspectral data-derived metrics that can be used to describe forest composition. There are a multitude of other metrics, from LiDAR grid metrics to other hyperspectral vegetation indices that have been used in previous studies and may increase model performance [81], [82]. Additionally, when aggregating the hyperspectral and LiDAR metrics over each subplot, using a standard deviation of the metrics instead of or in addition to the average of the metrics may also improve model performance.

2.7 CONCLUSIONS

Overall, the Random Forest classification algorithm resulted in the highest overall and kappa accuracy. There are many factors that make ecological data complex and difficult to model [83], [84]. Many studies use Random Forests to overcome these complexities and still yield high prediction accuracies [85], [86]. This was likely also the case in this study, resulting in the highest overall accuracy coming from Random Forest.

The MLR model with hyperspectral vegetation indices, DTM metrics, and CHM metrics had the best performance when predicting individual classes, especially with the classes that had less representation in the dataset. This may indicate that the MLR model would be a better model in cases where the objective is to accurately predict rare classes. It also indicates that it may be more suitable to use multiple models when predicting forest type. One could use Random Forest for predicting the more common forest types, and MLR for predicting the less common forest types.

This study also tested a multitude of model inputs and found that including hyperspectral vegetation indices, DTM metrics, and CHM metrics as model inputs yielded the highest overall and kappa accuracy with almost all classification algorithms. These model inputs each described different ecological factors that are key in distinguishing forest type groups. Topography plays an important role in the distribution of vegetation across the landscape, and the topographic variables that can be derived from a DTM can in predicting vegetation distribution [87]–[89]. Unsurprisingly, when these model inputs were tested with all subplots in one Random Forest model, roughness, topographic roughness index, and elevation were found to be extremely important in predicting forest type.

The canopy height model describes the over story structure and can be used to describe species composition when used alongside optical sensors [90]–[93]. In this study, we used both the raw canopy height model and a multitude of CHM-derived metrics to describe the over story. Canopy height was found to be the most important predictor for predicting forest type, with surface volume ratio following close behind.

When it comes to hyperspectral model inputs, including the raw hyperspectral band data with the hyperspectral vegetation indices and CHM and DTM metrics in the Random Forest model actually decreased accuracy. While it is unclear exactly why this is, we hypothesize that the fact that the 114 hyperspectral bands are so highly correlated, that they had a negative impact on prediction accuracies. Additionally, transforming the raw hyperspectral data into vegetation indices allows for the data to be relativized, which can boost model performance in many cases.

In summary, this study concluded that:

1. Of all classification algorithms tested, Random Forest resulted in the highest overall and kappa accuracy.

2. Canopy Height and Digital Terrain model metrics with Hyperspectral vegetation indices and 5 raw hyperspectral bands resulted in models with the highest overall accuracy.

Chapter 3. FUTURE WORK

3.1.1 *Forest Type Maps for TIU*

One of the key deliverables of this study is a model that allows for the classification of forest type across the TIU. This data product will be a 13 meter^2 resolution map of forest type over the entire area covered by both LiDAR and Hyperspectral data collected by the G-LiHT mission. Although both G-LiHT data products are available at 1 m^2 resolution, it was decided that a 13 m^2 resolution would be used in order to compensate for the 1 to 3 meter misalignment between the hyperspectral and LiDAR data. Aggregating the data to a lower resolution allows for this misalignment issue to have less of an impact on the classification accuracy. Additionally, the 13 m^2 resolution allows for much faster classification of the data when compared to the 1 m^2 resolution.

If there was not a 1 to 3-meter misalignment between the hyperspectral and LiDAR data, this study's methodology for classifying forest type would likely be drastically different. One methodology that was considered, had the data been better aligned, was a classification based on the individual tree or tree top [94]. The LiDAR data can be used to identify individual trees and obtain metrics for those trees [95]–[99]. Then, the hyperspectral data for those trees can be aggregated within the tree or tree-top area and used to classify species alongside the LiDAR metrics for that tree. This methodology is difficult to undertake in many studies due to the common problem of data misalignment and a lack of suitable field validation data.

3.1.2 *R Code*

A second product that will be published as part of this study is the R scripts that were wrote to perform almost every step of the process, from data pre-processing to model building. This

study hopes that this code could be used to repeat this study when more field data is available and/or in other regions with different datasets. Future work could consider adding additional hyperspectral vegetation indices and LiDAR-derived metrics. One could also consider testing different model parameters for the various classification algorithms tested.

All code will be published and available through githib.com, but will be edited to remove any classified information on FIA plot locations. The field datasets used in this study will not be made publicly available due to the classified nature of FIA plot locations.

3.1.3 *Bootstrap Aggregating*

There is an extreme imbalance between forest types represented in the response dataset. For example, there are 280 black spruce subplots while there are just 4 tamarack subplots. One way that many other studies deal with imbalanced response datasets such as the one seen here is to use bootstrap aggregating, or bagging. Bagging is the process by which one samples a dataset *D* of length *n* with replacement, in order to generate a new dataset *D1* of length *m* wherein each class is equally represented [100].

In this study, bagging was tested with each classification and model input in order to determine if it would improve model accuracies. Each class was sampled with replacement until a set equal in length to the largest class within the dataset was reached (Table 3.16).

Table 3.16. Number of subplots per forest type before and after bagging.

Forest Type	Number of Subplots	Number of Plots after Bagging
Aspen	23	280
Balsam Poplar	7	280
Black Spruce	280	280
Cottonwood	12	280
Paper Birch	175	280
Tamarack	4	280
White Spruce	93	280
Non-forest	38	280

Overall, bagging improved performance in cases where the model was already performing poorly, but did not improve accuracies when models were performing well. In the case of Random Forest, bagging decreased overall and kappa accuracy with all model input groups tested (Table 3.17).

Table 3.17. Random Forest Model Accuracy and Kappa values.

Model Inputs	Accuracy	Kappa	Bagging Accuracy	Bagging Kappa
Hyperspectral Bands and Vegetation Indices, with DTM and CHM Metrics	75.16%	0.62	59.69%	0.54
Hyperspectral Vegetation Indices, with DTM and CHM Metrics	**77.53%**	**0.66**	63.17%	0.58
CHM Metrics	69.30%	0.54	56.79%	0.51
Hyperspectral Vegetation Indices	67.09%	0.50	40.36%	0.32
Hyperspectral Bands	44.78%	0.23	32.68%	0.23
DTM Metrics	65.03%	0.47	37.05%	0.28

With the untuned Support Vector Machine, overall and kappa accuracies were extremely low, bagging improved the kappa accuracy in every case, and overall accuracy in all cases except in the case of raw hyperspectral bands (Table 3.18). Bagging made the untuned Support Vector Machine model more comparable to the results of the tuned Support Vector Machine, but accuracies were still lower.

Table 3.18. Untuned Support Vector Machine model accuracy and kappa results.

Model Inputs	Accuracy	Kappa	Bagging Accuracy	Bagging Kappa
Hyperspectral Bands and Vegetation Indices, with DTM and CHM Metrics	46.84%	0.27	62.99%	0.58

Hyperspectral Vegetation Indices, with DTM and CHM Metrics	50.32%	0.32	66.16%	0.61
CHM Metrics	45.73%	0.20	56.74%	0.51
Hyperspectral Vegetation Indices	43.83%	0.21	46.47%	0.39
Hyperspectral Bands	43.04%	0.20	37.50%	0.29
DTM Metrics	45.73%	0.20	47.81%	0.40

In the case of the tuned Support Vector Machine, bagging decreased overall accuracy in all cases, and decreased kappa accuracy in most cases, excluding CHM and DTM metrics where kappa accuracy increased slightly (Table 3.19).

Table 3.19. Tuned Support Vector Machine model accuracy and kappa results.

Model Inputs	Accuracy	Kappa	Bagging Accuracy	Bagging Kappa
Hyperspectral Bands and Vegetation Indices, with DTM and CHM Metrics	45.57%	0.10	14.46%	0.02
Hyperspectral Vegetation Indices, with DTM and CHM Metrics	**73.89%**	**0.60**	39.87%	0.31
CHM Metrics	47.94%	0.26	42.37%	0.34
Hyperspectral Vegetation Indices	64.72%	0.47	15.49%	0.03
Hyperspectral Bands	43.83%	0.20	15.27%	0.03
DTM Metrics	48.10%	0.27	38.57%	0.30

For the K-Nearest Neighbor Model, initial overall and kappa values were relatively similar in all cases, except in the case of DTM metrics where overall and kappa accuracy were much lower. When bagging was performed, overall accuracy for all model inputs changed to ~51%, and kappa accuracy went to ~49%. This was a decrease in overall and kappa accuracy in all cases except with DTM metrics (Table 3.20).

Table 3.20. K-Nearest Neighbor model accuracy and kappa results.

Model Inputs	Accuracy	Kappa	Bagging Accuracy	Bagging Kappa
Hyperspectral Bands and Vegetation Indices, with DTM and CHM Metrics	64.87%	0.49	51.29%	0.44
Hyperspectral Vegetation Indices, with DTM and CHM Metrics	63.61%	0.45	52.59%	0.46
CHM Metrics	64.08%	0.47	52.68%	0.46
Hyperspectral Vegetation Indices	65.98%	0.49	52.41%	0.46
Hyperspectral Bands	64.56%	0.47	51.47%	0.45
DTM Metrics	48.10%	0.28	51.12%	0.44

With the Naive Bayes classifier, overall and kappa accuracies were low prior to bagging. Bagging increased accuracies with some model inputs, but decreased it with others. The highest overall accuracy with the naive Bayes classifier was ~53%; this value did not significantly change with bagging (Table 3.21).

Table 3.21. Naive Bayes Classifier model accuracy and kappa results.

Model Inputs	Accuracy	Kappa	Bagging Accuracy	Bagging Kappa
Hyperspectral Bands and Vegetation Indices, with DTM and CHM Metrics	44.78%	0.28	40.63%	0.32
Hyperspectral Vegetation Indices, with DTM and CHM Metrics	53.80%	0.38	53.04%	0.46
CHM Metrics	21.52%	0.13	48.93%	0.42
Hyperspectral Vegetation Indices	48.89%	0.31	30.85%	0.21
Hyperspectral Bands	19.62%	0.08	21.70%	0.11
DTM Metrics	37.66%	0.18	39.96%	0.31

Finally, the Multinomial Logistic Regression model had the highest accuracy of any model after bagging at 68.71% overall and a kappa of 0.64. But, the overall accuracies decreased in all cases with bagging in the Multinomial Logistic Regression model. Kappa accuracy increased slightly in a few cases, but not significantly or to the point that it surpassed the results obtained with other classifiers without bagging (Table 3.22). Overall, this study found that bagging can increase model performance when a model is already not performing well, but it typically does not help when a model has generally high accuracies.

Table 3.22. Multinomial Logistic Regression model accuracy and kappa results.

Model Inputs	Accuracy	Kappa	Bagging Accuracy	Bagging Kappa
Hyperspectral Vegetation Indices, with DTM and CHM Metrics	71.20%	0.58	**68.71%**	**0.64**
CHM Metrics	48.26%	0.27	45.13%	0.37
Hyperspectral Vegetation Indices	68.04%	0.52	52.63%	0.46
Hyperspectral Bands	45.41%	0.25	40.13%	0.32
DTM Metrics	57.59%	0.32	46.34%	0.39

BIBLIOGRAPHY

[1] M. M. Miller and D. Lynch, "Alaska," Encyclopedia Britannica. Encyclopedia Britannica, inc., 02-May-2018 [Online]. Available: https://www.britannica.com/place/Alaska

[2] U. C. Bureau, "Census. gov," http://www. census. gov, vol. 612, p. 613, 2017.

[3] "MRLC NLCD 2011 Statistics." [Online]. Available: https://www.mrlc.gov/nlcd11_stat.php. [Accessed: 05-May-2018]

[4] Alaska Department of Natural Resources, "Land Ownership in Alaska," Alaska Department of Natural Resource, 05-2000. [Online]. Available: http://dnr.alaska.gov/mlw/factsht/land_fs/land_own.pdf. [Accessed: 05-May-2018]

[5] "Region 10 - About the Region." [Online]. Available: https://www.fs.usda.gov/main/r10/about-region. [Accessed: 06-May-2018]

[6] W. B. S. J.T. Vogt, "Forest Inventory and Analysis: Fiscal Year 2016 Business Report," United States Department of Agriculture, Aug. 2017 [Online]. Available: https://www.fs.fed.us/sites/default/files/fs_media/fs_document/publication-15817-usda-forest-service-fia-annual-report-508.pdf

[7] W. A. Bechtold, P. L. Patterson, and Others, The enhanced forest inventory and analysis program: national sampling design and estimation procedures, vol. 80. US Department of Agriculture Forest Service, Southern Research Station Asheville, North Carolina, 2005.

[8] H.-E. Andersen, J. Strunk, H. Temesgen, D. Atwood, and K. Winterberger, "Using multilevel remote sensing and ground data to estimate forest biomass resources in remote regions: a case study in the boreal forests of interior Alaska," Can. J. Remote Sens., vol. 37, no. 6, pp. 596–611, Jan. 2012.

[9] G. Reams, "Implementing FIA For All of Alaska: FIA & Partners," USDA Forest Service, Feb. 2014 [Online]. Available: https://www.fs.fed.us/pnw/workshops/interior-alaska-inventory/presentations/1_GregReams_Intro_AK_Meeting_Feb_2014.pdf

[10] "FIELD PROCEDURES FOR THE COASTAL ALASKA INVENTORY," USDA FOREST SERVICE PNW STATION - FORESTRY SCIENCES LAB & REGION 10 - ALASKA. 2003 [Online]. Available: https://www.fs.fed.us/pnw/rma/fia-topics/documentation/field-manuals/documents/Periodic/2003_coak_field_manual.pdf

[11] "FIELD PROCEDURES FOR THE SOUTHEAST ALASKA INVENTORY," USDA FOREST SERVICE PNW STATION - FORESTRY SCIENCES LAB & REGION 10 - ALASKA. 2000 [Online]. Available: https://www.fs.fed.us/pnw/rma/fia-topics/documentation/field-manuals/documents/Periodic/2000_seak_field_manual.pdf

[12] "Story Map Journal." [Online]. Available: https://usfs.maps.arcgis.com/apps/MapJournal/index.html?appid=d0464406188740fb81e2e4c3d1b48915. [Accessed: 15-May-2018]

[13] B. Mueller and D. Irvine, "Collaborating for success: implementation of the interior Alaska inventory," In: Stanton, Sharon M.; Christensen, Glenn A., comps. 2015. Pushing boundaries: new directions in inventory techniques and applications: Forest Inventory and Analysis (FIA) symposium 2015. 2015 December 8–10; Portland, Oregon. Gen. Tech. Rep. PNW-GTR-931. Portland, OR: U.S. Department of Agriculture, Forest Service, Pacific Northwest Research Station. p. 197., vol. 931, 2015 [Online]. Available: https://www.fs.usda.gov/treesearch/pubs/download/50361.pdf. [Accessed: 24-Apr-2018]

[14] H.-E. Andersen, C. Babcock, R. Pattison, B. Cook, D. Morton, and A. Finley, "The 2014 tanana inventory pilot: A USFS-NASA partnership to leverage advanced remote sensing

technologies for forest inventory," In: Stanton, Sharon M.; Christensen, Glenn A., comps. 2015. Pushing boundaries: new directions in inventory techniques and applications: Forest Inventory and Analysis (FIA) symposium 2015. 2015 December 8–10; Portland, Oregon. Gen. Tech. Rep. PNW-GTR-931. Portland, OR: U.S. Department of Agriculture, Forest Service, Pacific Northwest Research Station: 40-41., vol. 931, pp. 40–41, 2015.

[15] B. D. Cook et al., "NASA Goddard's LiDAR, Hyperspectral and Thermal (G-LiHT) Airborne Imager," Remote Sensing, vol. 5, no. 8, pp. 4045–4066, Aug. 2013.

[16] "FIELD INSTRUCTIONS FOR THE ANNUAL INVENTORY OF ALASKA," FOREST INVENTORY AND ANALYSIS RESOURCE MONITORING AND ASSESSMENT PROGRAM PACIFIC NORTHWEST RESEARCH STATION USDA FOREST SERVICE, 2018 [Online]. Available: https://www.fs.fed.us/pnw/rma/fia-topics/documentation/field-manuals/documents/Annual/2018_FIA_Interior_Alaska_Supplement.pdf

[17] H.-E. Andersen, J. Strunk, and H. Temesgen, "Using airborne light detection and ranging as a sampling tool for estimating forest biomass resources in the Upper Tanana Valley of Interior Alaska," West. J. Appl. For., vol. 26, no. 4, pp. 157–164, 2011.

[18] B. Ruefenacht et al., "Conterminous U.S. and Alaska Forest Type Mapping Using Forest Inventory and Analysis Data," Photogrammetric Engineering & Remote Sensing, vol. 74, no. 11, pp. 1379–1388, 2008.

[19] H.-E. Andersen, J. Strunk, H. Temesgen, D. Atwood, and K. Winterberger, "Using multilevel remote sensing and ground data to estimate forest biomass resources in remote regions: a case study in the boreal forests of interior Alaska," Can. J. Remote Sens., vol. 37, no. 6, pp. 596–611, 2012.

[20] C. Le Quéré et al., "Trends in the sources and sinks of carbon dioxide," Nat. Geosci., vol. 2, no. 12, pp. 831–836, 2009.

[21] K. R. Kirby and C. Potvin, "Variation in carbon storage among tree species: Implications for the management of a small-scale carbon sink project," For. Ecol. Manage., vol. 246, no. 2–3, pp. 208–221, 2007.

[22] B. Cook et al., "NASA Goddard's LiDAR, Hyperspectral and Thermal (G-LiHT) Airborne Imager," Remote Sensing, vol. 5, no. 8, pp. 4045–4066, 2013.

[23] G. P. Asner et al., "A universal airborne LiDAR approach for tropical forest carbon mapping," Oecologia, vol. 168, no. 4, pp. 1147–1160, Apr. 2012.

[24] K. T. Vierling, L. A. Vierling, W. A. Gould, S. Martinuzzi, and R. M. Clawges, "Lidar: shedding new light on habitat characterization and modeling," Front. Ecol. Environ., vol. 6, no. 2, pp. 90–98, 2008.

[25] Fuan Tsai, F. Tsai, and W. D. Philpot, "A derivative-aided hyperspectral image analysis system for land-cover classification," IEEE Trans. Geosci. Remote Sens., vol. 40, no. 2, pp. 416–425, 2002.

[26] M. Pal, "Support vector machine-based feature selection for land cover classification: a case study with DAIS hyperspectral data," Int. J. Remote Sens., vol. 27, no. 14, pp. 2877–2894, 2006.

[27] R. J. Zomer, A. Trabucco, and S. L. Ustin, "Building spectral libraries for wetlands land cover classification and hyperspectral remote sensing," J. Environ. Manage., vol. 90, no. 7, pp. 2170–2177, May 2009.

[28] R. M. Burns and B. H. Honkala, "Silvics of North America, Vol. 2--Hardwoods, Agricultural Handbook 654," US Department of Agriculture, Washington, DC, 1990.

[29] R. M. Burns, B. H. Honkala, and Others, “Silvics of North America. Volume 1. Conifers,” Agric. Handb., no. 654, 1990 [Online]. Available: https://www.cabdirect.org/cabdirect/abstract/19910653295
[30] L. A. Viereck, C. T. Dyrness, A. R. Batten, K. J. Wenzlick, and Others, “The Alaska vegetation classification,” 1992 [Online]. Available: https://www.frames.gov/files/5514/3352/7836/Viereck-et-al.-1992-Alaska-Vegetation-Classification.pdf
[31] C. A. Roland, J. H. Schmidt, and E. F. Nicklen, “Landscape-scale patterns in tree occupancy and abundance in subarctic Alaska,” Ecol. Monogr., vol. 83, no. 1, pp. 19–48, 2013.
[32] L. T. Ene et al., “Large-area hybrid estimation of aboveground biomass in interior Alaska using airborne laser scanning data,” Remote Sens. Environ., vol. 204, pp. 741–755, Jan. 2018.
[33] R. J. McGaughey and Others, “FUSION/LDV: Software for LIDAR data analysis and visualization,” US Department of Agriculture, Forest Service, Pacific Northwest Research Station: Seattle, WA, USA, vol. 123, no. 2, 2009.
[34] J. A. Greenberg and M. Mattiuzzi, “gdalUtils: wrappers for the geospatial data abstraction library (GDAL) utilities,” R Package Version, vol. 2, no. 1.7, 2015.
[35] M. F. J. Wilson, B. O’Connell, C. Brown, J. C. Guinan, and A. J. Grehan, “Multiscale Terrain Analysis of Multibeam Bathymetry Data for Habitat Mapping on the Continental Slope,” Mar. Geod., vol. 30, no. 1–2, pp. 3–35, May 2007.
[36] P. S. Thenkabail and J. G. Lyon, Hyperspectral Remote Sensing of Vegetation. CRC Press, 2016.
[37] A. R. Heute, H. Q. Liu, K. Batchily, and W. Van Leeuwen, “A comparison of vegetation indices over a global set of TM images for EOS-MODIS,” REMOTE SENSING OF ENVIRONMENT-NEW YORK-, vol. 59, pp. 440–451, 1997.
[38] P. Gong, R. Pu, G. S. Biging, and M. R. Larrieu, “Estimation of forest leaf area index using vegetation indices derived from Hyperion hyperspectral data,” IEEE Trans. Geosci. Remote Sens., vol. 41, no. 6, pp. 1355–1362, 2003.
[39] M. S. Torino, B. V. Ortiz, J. P. Fulton, and K. Balkcom, “EVALUATION OF DIFFERENCES IN CORN BIOMASS AND NITROGEN UPTAKE AT VARIOUS GROWTH STAGES USING SPECTRAL VEGETATION INDICES,” ispag.org [Online]. Available: https://www.ispag.org/abstract_papers/papers/abstract_1251.pdf
[40] A. Gitelson and M. N. Merzlyak, “Spectral Reflectance Changes Associated with Autumn Senescence of Aesculus hippocastanum L. and Acer platanoides L. Leaves. Spectral Features and Relation to Chlorophyll Estimation,” J. Plant Physiol., vol. 143, no. 3, pp. 286–292, Mar. 1994.
[41] J. W. Rouse Jr, “HAAS. RH, SCHELL, JA and DEHRINO. DW. 1973. Monitoring vegetation systems in the Great Plains with ERTS,” Proceedings of the 3rd Earth Resources Technology Satelhte-l Sytnposium, Wa. fhington. DC. USA, 10-15 December 1973, pp. 309–317, 1973.
[42] C. F. Jordan, “Derivation of leaf-area index from quality of light on the forest floor,” Ecology, vol. 50, no. 4, pp. 663–666, 1969.
[43] A. A. Gitelson, Y. J. Kaufman, R. Stark, and D. Rundquist, “Novel algorithms for remote estimation of vegetation fraction,” Remote Sens. Environ., vol. 80, no. 1, pp. 76–87, Apr. 2002.

[44] G. A. Blackburn, “Quantifying Chlorophylls and Caroteniods at Leaf and Canopy Scales: An Evaluation of Some Hyperspectral Approaches,” Remote Sens. Environ., vol. 66, no. 3, pp. 273–285, Dec. 1998.
[45] M. N. Merzlyak, A. A. Gitelson, O. B. Chivkunova, and V. Y. Rakitin, “Non-destructive optical detection of pigment changes during leaf senescence and fruit ripening,” Physiol. Plant., vol. 106, no. 1, pp. 135–141, 1999.
[46] G. A. Blackburn, “Spectral indices for estimating photosynthetic pigment concentrations: A test using senescent tree leaves,” Int. J. Remote Sens., vol. 19, no. 4, pp. 657–675, Jan. 1998.
[47] J. Penuelas, F. Baret, and I. Filella, “Semi-empirical indices to assess carotenoids/chlorophyll a ratio from leaf spectral reflectance,” Photosynthetica, vol. 31, no. 2, pp. 221–230, 1995.
[48] M. S. Kim, “The Use of Narrow Spectral Bands for Improving Remote Sensing Estimations of Fractionally Absorbed Photosynthetically Active Radiation (fapar),” University of Maryland at College Park, 1994.
[49] A. A. Gitelson, G. P. Keydan, and M. N. Merzlyak, “Three-band model for noninvasive estimation of chlorophyll, carotenoids, and anthocyanin contents in higher plant leaves,” Geophys. Res. Lett., vol. 33, no. 11, 2006 [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1029/2006GL026457/full
[50] G. A. Carter and R. L. Miller, “Early detection of plant stress by digital imaging within narrow stress-sensitive wavebands,” Remote Sens. Environ., vol. 50, no. 3, pp. 295–302, Dec. 1994.
[51] B. Datt, “Visible/near infrared reflectance and chlorophyll content in Eucalyptus leaves,” Int. J. Remote Sens., vol. 20, no. 14, pp. 2741–2759, Jan. 1999.
[52] C. S. T. Daughtry, C. L. Walthall, M. S. Kim, E. B. de Colstoun, and J. E. McMurtrey, “Estimating Corn Leaf Chlorophyll Concentration from Leaf and Canopy Reflectance,” Remote Sens. Environ., vol. 74, no. 2, pp. 229–239, Nov. 2000.
[53] A. K. van den Berg and T. D. Perkins, “Nondestructive Estimation of Anthocyanin Content in Autumn Sugar Maple Leaves,” HortScience, vol. 40, no. 3, pp. 685–686, Jun. 2005.
[54] A. A. Gitelson, M. N. Merzlyak, and O. B. Chivkunova, “Optical properties and nondestructive estimation of anthocyanin content in plant leaves,” Photochem. Photobiol., vol. 74, no. 1, pp. 38–45, Jul. 2001.
[55] J. A. Gamon and J. S. Surfus, “Assessing leaf pigment content and activity with a reflectometer,” New Phytol., vol. 143, no. 1, pp. 105–117, Jul. 1999.
[56] A. A. Gitelson, Y. Zur, O. B. Chivkunova, and M. N. Merzlyak, “Assessing carotenoid content in plant leaves with reflectance spectroscopy,” Photochem. Photobiol., vol. 75, no. 3, pp. 272–281, Mar. 2002.
[57] X. Liu, Z. Hou, Z. Shi, Y. Bo, and J. Cheng, “A shadow identification method using vegetation indices derived from hyperspectral data,” Int. J. Remote Sens., vol. 38, no. 19, pp. 5357–5373, 2017.
[58] S. C. Azevedo, E. A. Silva, and M. M. Pedrosa, “Shadow detection improvement using spectral indices and morphological operators in high resolution images from urban areas,” ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XL–7/W3, pp. 587–592, 2015.
[59] S. H. Hurlbert, “Pseudoreplication and the Design of Ecological Field Experiments,” Ecol. Monogr., vol. 54, no. 2, pp. 187–211, Feb. 1984.

[60] P. Weishampel, R. Kolka, and J. Y. King, “Carbon pools and productivity in a 1-km2 heterogeneous forest and peatland mosaic in Minnesota, USA,” For. Ecol. Manage., vol. 257, no. 2, pp. 747–754, Jan. 2009.
[61] L. Breiman, “Bagging predictors,” Mach. Learn., vol. 24, no. 2, pp. 123–140, 1996.
[62] A. Liaw, M. Wiener, and Others, “Classification and regression by randomForest,” R news, vol. 2, no. 3, pp. 18–22, 2002.
[63] L. Breiman, “Random Forests,” Mach. Learn., vol. 45, no. 1, pp. 5–32, Oct. 2001.
[64] C.-C. Chang and C.-J. Lin, “LIBSVM: A Library for Support Vector Machines,” ACM Trans. Intell. Syst. Technol., vol. 2, no. 3, pp. 27:1–27:27, May 2011.
[65] M. M. Adankon and M. Cheriet, “Support Vector Machine,” in Encyclopedia of Biometrics, S. Z. Li and A. Jain, Eds. Boston, MA: Springer US, 2009, pp. 1303–1308.
[66] B. D. Ripley, Pattern Recognition and Neural Networks. Cambridge University Press, 2007.
[67] L. E. Peterson, “K-nearest neighbor,” Scholarpedia J., vol. 4, no. 2, p. 1883, 2009.
[68] D. J. Hand and K. Yu, “Idiot’s Bayes—not so stupid after all?,” Int. Stat. Rev., vol. 69, no. 3, pp. 385–398, 2001.
[69] H. Zhang, “The optimality of naive Bayes,” Archit. Aujourdhui., vol. 1, no. 2, p. 3, 2004.
[70] K. M. Leung, “Naive bayesian classifier,” University Department of Computer Science/Finance …, 2007 [Online]. Available: http://cis.poly.edu/~mleung/FRE7851/f07/naiveBayesianClassifier.pdf
[71] K. P. Murphy, “Naive bayes classifiers,” Univ. B. C. Law Rev., vol. 18, 2006 [Online]. Available: https://datajobsboard.com/wp-content/uploads/2017/01/Naive-Bayes-Kevin-Murphy.pdf
[72] C. Kwak and A. Clayton-Matthews, “Multinomial logistic regression,” Nurs. Res., vol. 51, no. 6, pp. 404–410, Nov. 2002.
[73] M. Kuhn, “Predictive Modeling with R and the caret Package,” 2013 [Online]. Available: https://www.r-project.org/conferences/useR-2013/Tutorials/kuhn/user_caret_2up.pdf
[74] R. G. Pontius and M. Millones, “Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment,” Int. J. Remote Sens., vol. 32, no. 15, pp. 4407–4429, Aug. 2011.
[75] L. Breiman, “Manual on setting up, using, and understanding random forests v3. 1,” Statistics Department University of California Berkeley, CA, USA, vol. 1, 2002.
[76] J. Carletta, “Assessing Agreement on Classification Tasks: The Kappa Statistic,” Comput. Linguist., vol. 22, no. 2, pp. 249–254, Jun. 1996.
[77] J. R. Landis and G. G. Koch, “The measurement of observer agreement for categorical data,” Biometrics, vol. 33, no. 1, pp. 159–174, Mar. 1977.
[78] M. Dalponte, L. Bruzzone, and D. Gianelle, “Fusion of Hyperspectral and LIDAR Remote Sensing Data for Classification of Complex Forest Areas,” IEEE Trans. Geosci. Remote Sens., vol. 46, no. 5, pp. 1416–1427, 2008.
[79] M. Dalponte, L. Bruzzone, and D. Gianelle, “Tree species classification in the Southern Alps based on the fusion of very high geometrical resolution multispectral/hyperspectral images and LiDAR data,” Remote Sens. Environ., vol. 123, pp. 258–270, 2012.
[80] Z. Zhang, A. Kazakova, L. M. Moskal, and D. M. Styers, “Object-Based Tree Species Classification in Urban Ecosystems Using LiDAR and Hyperspectral Data,” For. Trees Livelihoods, vol. 7, no. 6, p. 122, Jun. 2016.

[81] Q. Cao et al., “Non-destructive estimation of rice plant nitrogen status with Crop Circle multispectral active canopy sensor,” Field Crops Res., vol. 154, pp. 133–144, Dec. 2013.
[82] E. R. Hunt, P. C. Doraiswamy, J. E. McMurtrey, C. S. T. Daughtry, E. M. Perry, and B. Akhmedov, “A visible band index for remote sensing leaf chlorophyll content at the canopy scale,” Int. J. Appl. Earth Obs. Geoinf., vol. 21, pp. 103–112, Apr. 2013.
[83] L. Breiman, “Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author),” Stat. Sci., vol. 16, no. 3, pp. 199–231, Aug. 2001.
[84] J. d. Olden, J. j. Lawler, and N. L. Poff, “Machine Learning Methods Without Tears: A Primer for Ecologists,” Q. Rev. Biol., vol. 83, no. 2, pp. 171–193, 2008.
[85] V. F. Rodriguez-Galiano, B. Ghimire, J. Rogan, M. Chica-Olmo, and J. P. Rigol-Sanchez, “An assessment of the effectiveness of a random forest classifier for land-cover classification,” ISPRS J. Photogramm. Remote Sens., vol. 67, pp. 93–104, Jan. 2012.
[86] D. R. Cutler et al., “Random forests for classification in ecology,” Ecology, vol. 88, no. 11, pp. 2783–2792, Nov. 2007.
[87] I. V. Florinsky and G. A. Kuryakova, “Influence of topography on some vegetation cover properties,” Catena, vol. 27, no. 2, pp. 123–141, Aug. 1996.
[88] D. G. Brown, “Predicting vegetation types at treeline using topography and biophysical disturbance variables,” J. Veg. Sci., vol. 5, no. 5, pp. 641–656, Oct. 1994.
[89] B. Wang, G. Zhang, and J. Duan, “Relationship between topography and the distribution of understory vegetation in a Pinus massoniana forest in Southern China,” International Soil and Water Conservation Research, vol. 3, no. 4, pp. 291–304, Dec. 2015.
[90] R. Dubayah, R. Knox, M. Hofton, and J. B. Blair, “Land surface characterization using lidar remote sensing,” Spatial information for, 2000 [Online]. Available: https://books.google.com/books?hl=en&lr=&id=B1hORdx0Gr0C&oi=fnd&pg=PA25&dq=DUBAYAH,+R.O.,+R.G.+KNOX,+M.A.+HOFTON,+J.B.+BLAIR,+and+J.B.+DRAKE.+2000.+Land+surface+characterization+using+lidar+remote+sensing.&ots=8L8DJqHKbK&sig=ondD3zsoa6ph7bcsZjlOkiIV_4Y
[91] R. Dubayah et al., “The vegetation canopy lidar mission,” Land satellite information in the next decade II: sources and applications, 1997.
[92] R. O. Dubayah and J. B. Drake, “Lidar Remote Sensing for Forestry,” J. For., vol. 98, no. 6, pp. 44–46, Jun. 2000.
[93] A. T. Hudak et al., “Mapping Forest Structure and Composition from Low-Density LiDAR for Informed Forest, Fuel, and Fire Management at Eglin Air Force Base, Florida, USA,” Can. J. Remote Sens., vol. 42, no. 5, pp. 411–427, Sep. 2016.
[94] C. Zhang and F. Qiu, “Mapping Individual Tree Species in an Urban Forest Using Airborne Lidar Data and Hyperspectral Imagery,” Photogrammetric Engineering & Remote Sensing, vol. 78, no. 10, pp. 1079–1087, Oct. 2012.
[95] A. Persson, J. Holmgren, and U. Soderman, “Detecting and measuring individual trees using an airborne laser scanner,” Photogramm. Eng. Remote Sens., vol. 68, no. 9, pp. 925–932, 2002.
[96] X. Yu, J. Hyyppä, M. Vastaranta, M. Holopainen, and R. Viitala, “Predicting individual tree attributes from airborne laser point clouds based on the random forests technique,” ISPRS J. Photogramm. Remote Sens., vol. 66, no. 1, pp. 28–37, Jan. 2011.
[97] J. Pitkänen, M. Maltamo, J. Hyyppä, and X. Yu, “Adaptive methods for individual tree detection on airborne laser based canopy height model,” International Archives of

Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. 36, no. 8, pp. 187–191, 2004.

[98] H. Kaartinen et al., “An International Comparison of Individual Tree Detection and Extraction Using Airborne Laser Scanning,” Remote Sensing, vol. 4, no. 4, pp. 950–974, Mar. 2012.

[99] B. Koch, U. Heyder, and H. Weinacker, “Detection of Individual Tree Crowns in Airborne Lidar Data,” Photogrammetric Engineering & Remote Sensing, vol. 72, no. 4, pp. 357–363, Apr. 2006.

[100] L. Breiman, “Bagging predictors,” Mach. Learn., vol. 24, no. 2, pp. 123–140, 1996.