Machine Learning for Tree Species Identification from LiDAR & Imagery

Trevor Hooper[1], Mike Parlow[2], Hazel Jeong[3]
D'Laine Robertson-Hooper[4], Geoff Lawless[5]

[1]Forsite Consultants Ltd, PO BOX 2079, Salmon Arm, BC, V1E 4R1
Email: thooper@forsite.ca

[2]Forsite Consultants Ltd, PO BOX 2079, Salmon Arm, BC, V1E 4R1
Email: mparlow@forsite.ca

3Forsite Consultants Ltd, PO BOX 2079, Salmon Arm, BC, V1E 4R1
Email: hjeong@forsite.ca

[4]Forsite Consultants Ltd, PO BOX 2079, Salmon Arm, BC, V1E 4R1
Email: drobertson-hooper@forsite.ca

[5]Forsite Consultants Ltd, PO BOX 2079, Salmon Arm, BC, V1E 4R1
Email: glawless@forsite.ca

1. Introduction

The investigators sought to explore and examine the impact of varied remote sensing inputs for species identification using machine learning. The foundation of the research was a multi-class support vector machine (SVM) learning adaptation originally developed in 2012. The latter used LiDAR as the exclusive input for species identification and leveraged spatial density, trunk & branch geometry, and the intensity attribute. In the new effort, the team attempted to adapt and incorporate spectral as well as land form information to the SVM descriptor list. 251 new descriptors were created and ranked alongside the existing 847 descriptors. The research team sought to determine the optimal combination of descriptors for species identification accuracy. The result was an increase in stem accuracy in a cross-fold validation test of between 8% and 13% for a mix of 13 conifer and deciduous species.

1.1 Background

The Tree Species Identifier (TSI) system uses a bottom-up approach where metrics are measured and predicted first at the individual tree level. The individual tree process captures height, canopy characteristics, and species from the LiDAR. From those inputs, the system can calculate estimates of the diameter at breast height (DBH) as well as volume. Once the analysis has been completed at the single tree level, the outputs can be rolled up to larger reporting units. The results individual tree inventory can also be used for statistical adjustments across the land base using and area-based enhanced forest inventory approach.
First developed in 2012, the process has successfully analysed over 2 billion trees and produced operational inventories derived from hundreds of terabytes of LiDAR across millions of forested hectares.

1.2 Tree Species Identifier Process

First the LAS is reviewed, cleaned, and prepared for analysis. Analysts review an array of factors including the consistency of point density, the intensity calibration, and any gaps in the coverage. Then TSI segments the individual trees from the point cloud and produces an area shapefile for each tree. The

Published in: Markus Hollaus, Norbert Pfeifer (Eds.): Proceedings of the SilviLaser Conference 2021, Vienna, Austria, 28–30 September 2021.
Technische Universität Wien, 2021. DOI: 10.34726/wim.1861
This paper was peer-reviewed. DOI of this paper: 10.34726/wim.2025

system calculates a number of attributes including height, slope, crown area, aspect, local density, and live crown percentage. Each tree also receives a unique ID at this point in the process.

The segmentation parameters used are selected based on a variety of stand characteristics using a blend of classic watershed techniques and point finding routines. Individual tree inventories from LIDAR tend to underestimate the number of stems as the software can only include what the sensor sees. Missed stems are typically smaller ones hiding under larger ones or those in tight clumps with a common height. Conversely, leaning trees can sometimes be segmented into multiple trees. As LiDAR point densities increase, say above 16-20 pts/m^2, the segmentation algorithm is able to adapt resulting in higher overall tree segmentation accuracy as well as understory segmentation.

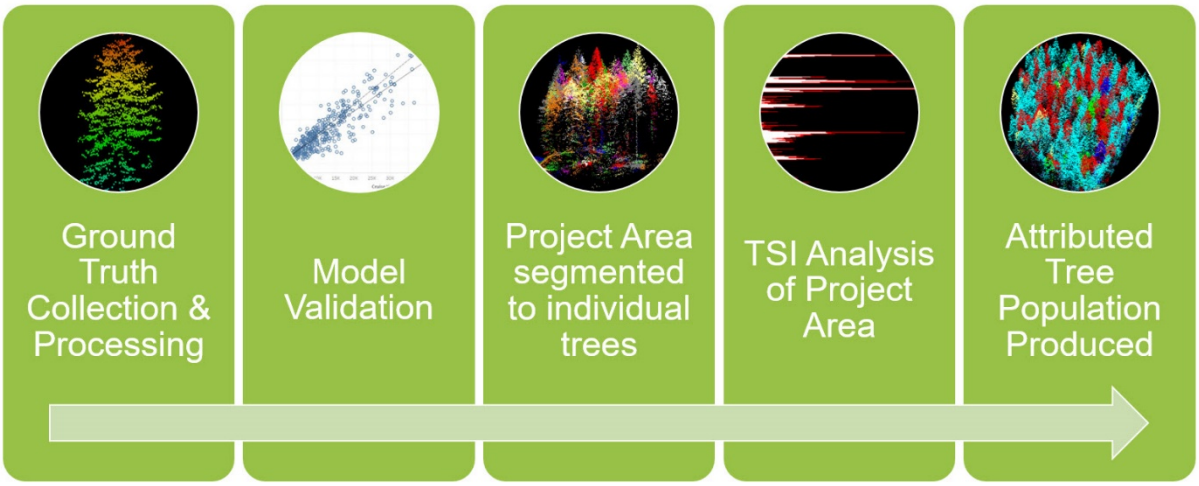


Figure 1: Tree Species Identifier Process

The next step in the species identification process is the collection of ground truth trees to be used in the TSI species prediction model. Trees are collected based on species, height, and location within the project area of interest by field crews and/or photo-interpretation. The goal is to acquire 100 to 300 samples of each species in the project area. The required number of samples per species varies project to project based on the complexity of the species mix and the size of the area under consideration. Using methods refined over 9 years, a trained 2-person field crew can collect 800 trees over 5 days.

Once the trees are captured, analysts attempt to match up the field or photo-interpretation collects with the correct tree in the LiDAR point cloud. The quality control success rates vary depending primarily on canopy density and GPS signal strength. Next the tree samples that have passed quality control are translated into machine-learning numeric "descriptors" in TSI. The software translates the information in each tree's point cloud into numeric descriptors based on geometry, density, and reflectivity. This step is at the heart of the TSI capability and the focus of the research. The model validation proceeds with the derived descriptors and the resulting model is used to perform a discrete analysis of each segmented tree.

Diameter at breast height (DBH) is derived from the tree height and species using established biometric models. The DBH is then used to calculate gross and merchantable volume for each tree.

2.0 Descriptor Research

The research team collected remote sensing inputs over forest areas in the Canadian province of British Columbia. The imagery data included 30 cm resolution 4-band RGB-NIR ortho-photo as well as 10 m resolution satellite Sentinel-2 multispectral. Terrain data included 2m resolution wetness and sunlight maps derived from the LiDAR. The Provincial Forestry Ministry also provides ecosite information and predictive ecosystem mapping, land base metrics that provide broad soil and moisture information, both of which were incorporated. The LiDAR was flown with a 10-12 pts/m^2 point density.

6 descriptor test combinations were tested:

1. Baseline: 10-12 ppm LiDAR only
2. Baseline plus LiDAR-derived intensity images from two channels
3. Baseline plus 4-band RGB-NIR
4. Baseline plus terrain characteristics
5. Baseline plus 4-band RGB-NIR & terrain characteristics
6. Baseline plus 4-band RGB-NIR & terrain characteristics & LiDAR-derived intensity images

The descriptors were run against a ground truth set of 4,395 trees including 13 species.

FD	Douglas Fir	CW	Western Red Cedar
LW	Western Larch	AC	Black Cottonwood
BL	Balsam Fir	AT	Trembling Aspen
PY	Ponderosa Pine	EP	Paper Birch
PL	Lodgepole Pine	DP	Lodgepole Pine (Dead)
SX	Spruce (hybrid)	SN	Snag
RP	Lodgepole Pine (Red - Dying)		

Table 1: Species included in the test

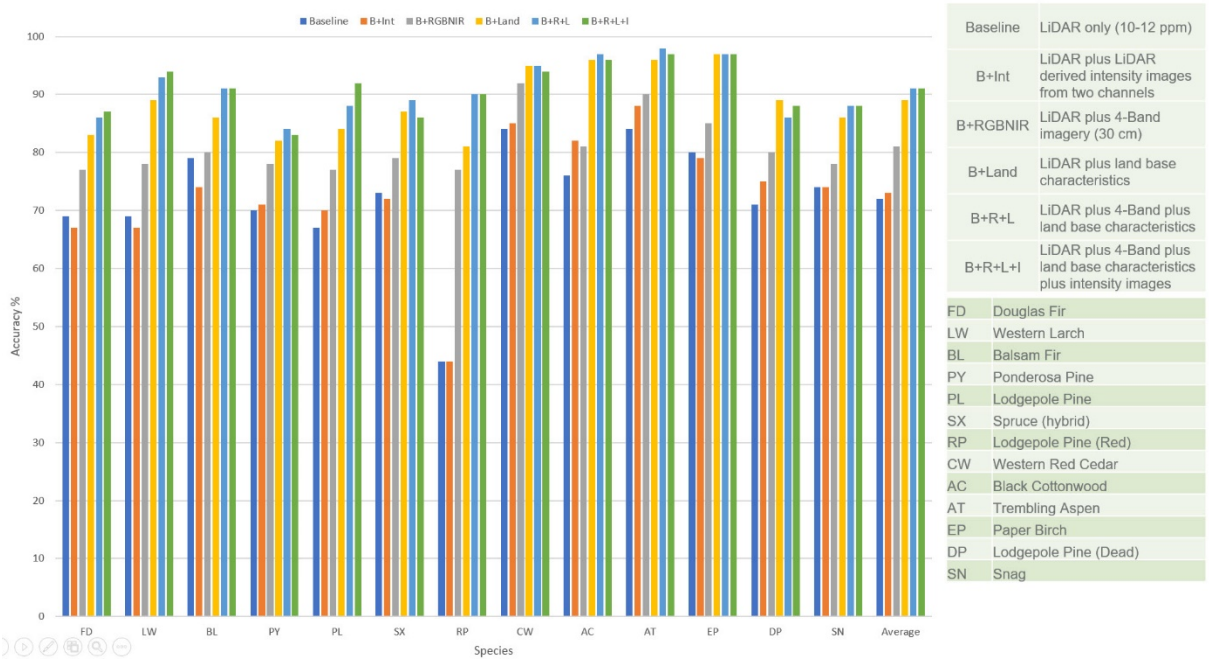
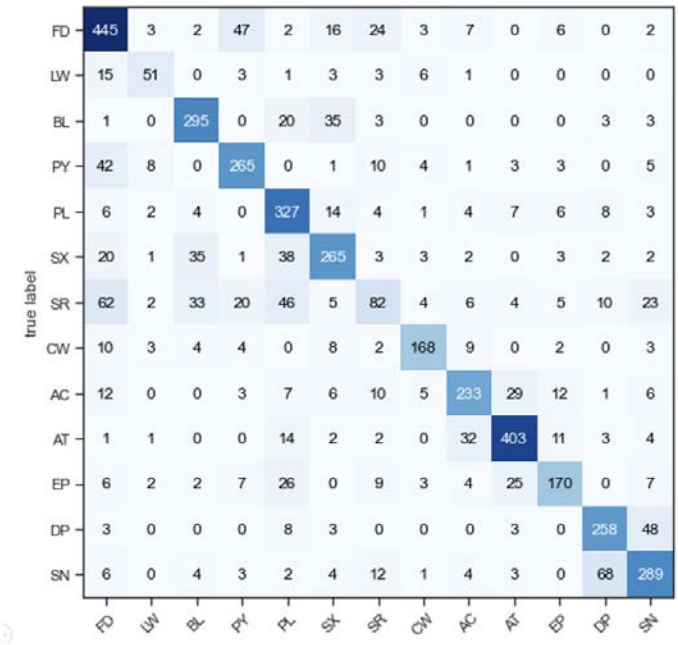


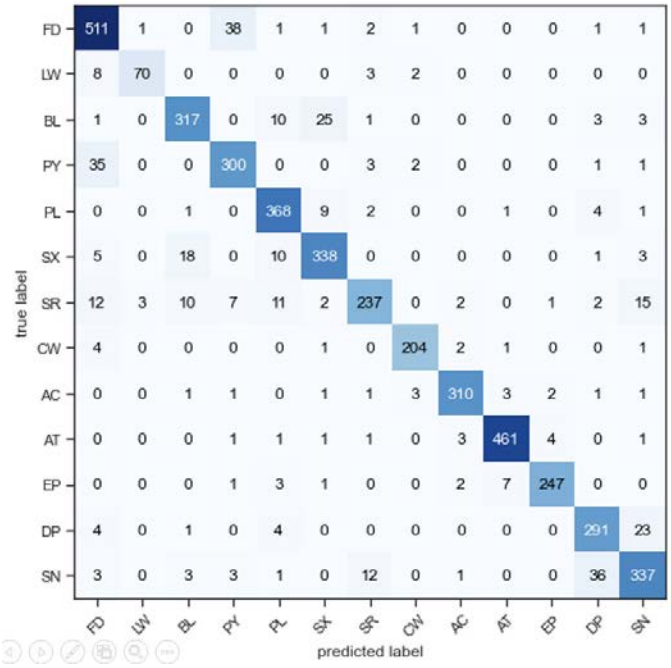
Figure 2: Species Results by descriptor set

The general trend evidenced by the trials was that as the descriptor sets added additional remote sensing and terrain input information, species accuracy improved. This trend held for conifer and deciduous classes broadly as well as dead or stressed trees. The largest improvement was seen for dying lodgepole pine (red) while the highest accuracies were recorded for the deciduous species.



	precision	recall	f1-score	support
FD	0.71	0.80	0.75	557
LW	0.70	0.61	0.65	83
BL	0.78	0.82	0.80	360
PY	0.75	0.77	0.76	342
PL	0.67	0.85	0.75	386
SX	0.73	0.71	0.72	375
SR	0.50	0.27	0.35	302
CW	0.85	0.79	0.82	213
AC	0.77	0.72	0.74	324
AT	0.84	0.85	0.85	473
EP	0.78	0.65	0.71	261
DP	0.73	0.80	0.76	323
SN	0.73	0.73	0.73	396
accuracy			0.74	4395
macro avg	0.73	0.72	0.72	4395
weighted avg	0.74	0.74	0.73	4395

Table 2: Baseline Species Accuracy Confusion Matrix



	precision	recall	f1-score	support
FD	0.88	0.92	0.90	557
LW	0.95	0.84	0.89	83
BL	0.90	0.88	0.89	360
PY	0.85	0.88	0.87	342
PL	0.90	0.95	0.93	386
SX	0.89	0.90	0.90	375
SR	0.90	0.78	0.84	302
CW	0.96	0.96	0.96	213
AC	0.97	0.96	0.96	324
AT	0.97	0.97	0.97	473
EP	0.97	0.95	0.96	261
DP	0.86	0.90	0.88	323
SN	0.87	0.85	0.86	396
accuracy			0.91	4395
macro avg	0.91	0.90	0.91	4395
weighted avg	0.91	0.91	0.91	4395

Table 3: Final Species Accuracy Confusion matrix

The most significant improvement was in dying lodgepole pine (SR). A working hypothesis was that SR would improve as a result of the addition of spectral descriptors. While true, SR accuracy also improved with the addition of terrain characteristics in the absence of spectral data. Douglas fir and ponderosa pine confusion was reduced from 89 direct errors to 73 direct errors. The latter improvement fell short of expectations as the two species have distinct spectral signatures. Those distinct signatures enable rigorous accuracies in photo-interpretation and so more improvement was expected. One possible explanation is that the geometry and density descriptors derived from the point cloud may have some embedded biases that need to be addressed in future research. Another possible explanation is the introduction of noise by the imagery due to parallax offsets that degraded the predictive power of the descriptor.