electronics

Article

Assessment of Tree Species Classification by Decision Tree Algorithm Using Multiwavelength Airborne Polarimetric LiDAR Data

Zhong Hu [1] **and Songxin Tan** [2,*]

[1] Department of Mechanical Engineering, J.J. Lohr College of Engineering, South Dakota State University, Brookings, SD 57007, USA; zhong.hu@sdstate.edu

[2] Department of Electrical Engineering and Computer Science, J.J. Lohr College of Engineering, South Dakota State University, Brookings, SD 57007, USA

* Correspondence: songxin.tan@sdstate.edu; Tel.: +1-605-688-4994

Abstract: Polarimetric measurement has been proven to be of great importance in various applications, including remote sensing in agriculture and forest. Polarimetric full waveform LiDAR is a relatively new yet valuable active remote sensing tool. This instrument offers the full waveform data and polarimetric information simultaneously. Current studies have primarily used commercial non-polarimetric LiDAR for tree species classification, either at the dominant species level or at the individual tree level. Many classification approaches combine multiple features, such as tree height, stand width, and crown shape, without utilizing polarimetric information. In this work, a customized Multiwavelength Airborne Polarimetric LiDAR (MAPL) system was developed for field tree measurements. The MAPL is a unique system with unparalleled capabilities in vegetation remote sensing. It features four receiving channels at dual wavelengths and dual polarization: near infrared (NIR) co-polarization, NIR cross-polarization, green (GN) co-polarization, and GN cross-polarization, respectively. Data were collected from several tree species, including coniferous trees (blue spruce, ponderosa pine, and Austrian pine) and deciduous trees (ash and maple). The goal was to improve the target identification ability and detection accuracy. A machine learning (ML) approach, specifically a decision tree, was developed to classify tree species based on the peak reflectance values of the MAPL waveforms. The results indicate a re-substitution error of 3.23% and a *k*-fold loss error of 5.03% for the 2106 tree samples used in this study. The decision tree method proved to be both accurate and effective, and the classification of new observation data can be performed using the previously trained decision tree, as suggested by both error values. Future research will focus on incorporating additional LiDAR data features, exploring more advanced ML methods, and expanding to other vegetation classification applications. Furthermore, the MAPL data can be fused with data from other sensors to provide augmented reality applications, such as Simultaneous Localization and Mapping (SLAM) and Bird's Eye View (BEV). Its polarimetric capability will enable target characterization beyond shape and distance.

Keywords: remote sensing; LiDAR; polarimetric LiDAR; full waveform; classification; machine learning; decision tree

check for updates

Citation: Hu, Z.; Tan, S. Assessment of Tree Species Classification by Decision Tree Algorithm Using Multiwavelength Airborne Polarimetric LiDAR Data. *Electronics* **2024**, *13*, 4534. https://doi.org/10.3390/electronics13224534

Academic Editor: Esteban Tlelo-Cuautle

Received: 21 October 2024
Revised: 13 November 2024
Accepted: 18 November 2024
Published: 19 November 2024

1. Introduction

Global climate change has become a pressing issue. The Sixth Assessment Report from the Intergovernmental Panel on Climate Change shows that the global average surface temperature in the past decade (2011–2020) was 1.09 °C higher than the average temperature between 1850 to 1900 [1,2]. The United Nations General Assembly approved the 2030 agenda, which urges the world to take urgent actions to address climate change and its impacts [3]. As a major source of carbon sink, vegetation and forest help reduce greenhouse gases and slow down climate change. Hence, mapping the forests on both the global and local scale is necessitated. Classifying and mapping vegetation is also needed in order to understand how climate change affects ecosystems and to predict future changes in vegetation distribution due to altered temperature and precipitation patterns. It allows scientists to monitor and study how plant life responds to climate change by determining which vegetation types are expanding, shrinking, or migrating to new areas [4]. In addition, species diversity is essential for the provision of environmental services and ecosystem health.

Not surprisingly, therefore, many metrics have been developed to accurately map the vegetation distribution. These metrics should be statistically rigorous, show monotonic relationships, be sensitive to species' appearance and disappearance, be comparable and invariant across different scales, be affordable, and, most importantly, be easy to understand [5]. There have been a variety of sensors and platforms developed to meet these requirements. Remote sensing is an effective option for vegetation monitoring over large areas. It has been used in many fields, including geophysics, geography, land surveying, and most earth science disciplines. Recent technological advances have provided vast amounts of remote sensing data to meet the ever-increasing demands [6].

Light detection and ranging (LiDAR) is a remote sensing technique that measures distance by using a pulsed or continuous laser wave to aim at an object and counting the time it takes for the light to return to the receiver [7]. LiDAR can be used on the ground, or in an airborne manner, or satellite-borne [8,9]. Based on the sampling strategy, pulsed LiDAR systems can be categorized as discrete return systems, full waveform systems, photon-counting systems, and synthetic aperture LiDAR systems. In vegetation research, capturing the true canopy profile can provide crucial information. The full waveform LiDAR has its advantage in measuring the canopy distribution [10,11]. In addition, polarimetric discrimination plays an important role in target classification. Therefore, a polarimetric LiDAR is warranted, and it has the ability to measure the polarization state of the backscattered laser beam from the target [12,13]. Polarimetric measurements have been used in a variety of fields, including astronomy, chemistry, and food and beverage production, in addition to the vegetation classification in agricultural and forest remote sensing [14–17].

After acquiring the LiDAR measurement data, data analysis is performed to make informed decisions to improve processes and gain competitive advantage. With the exponential growth in the availability of acquired data and the advancement in experimental and computational techniques, researchers can now obtain valuable insights from large and complex datasets. A variety of interconnected data analysis methods have become available, including statistical techniques, machine learning (ML), deep learning (DL), neural networks (NNs), data mining, and artificial intelligence (AI), each of which intersects and overlaps with the others and can be used to extract meaningful information from datasets [18,19]. ML-based classification is a predictive modeling process in which the model attempts to predict the correct label for input data. In classification, the model is trained using the training data, validated using validation data, evaluated using test data with an error function, and, finally, used to make predictions on new, unseen data. Due to the powerful capabilities of classification and clustering algorithms in ML, ML has been widely used in many fields, including vegetation classification, land cover classification, and yield prediction in agriculture and forestry [20–24].

In this study, a customized Multiwavelength Airborne Polarimetric LiDAR (MAPL) system was developed. Four different LiDAR channels at dual wavelengths and dual polarization—near infrared (NIR) co-polarization, NIR cross-polarization, Green (GN) co-polarization, and GN cross-polarization—were designed to improve the identification ability and measurement accuracy. The MAPL was designed primarily for vegetation remote sensing [25]. Field experiments were conducted using the MAPL to collect data from a variety of different trees. Then, a decision tree approach was developed to classify the different species of trees based on the peak intensity values from the MAPL waveforms. This approach has not previously been applied to the unique MAPL dataset. The effectiveness of the proposed method is evaluated to guide future classification efforts.

2. MAPL System

The LiDAR system used in this study is the MAPL system. The MAPL has four receiving channels and is used for forest remote sensing [25]. It consists of three main subsystems: the laser source—an optical receiver assembly, and the data acquisition and processing hardware and software. The system employs an Nd: YAG laser that simultaneously emits radiation at two wavelengths: 1064 nm (NIR) and 532 nm (GN). Both laser beams are highly linearly polarized with an extinction ratio of 100:1. The beam divergence is approximately 4 mrad, which will produce a laser footprint of approximately 4 m in diameter at a distance of 1000 m. The laser source has a pulse repetition frequency of 10 Hz and a pulse width of 10 ns, yielding a range resolution of 1.5 m. The receiving aperture is 25 mm in diameter and the receiver optics are designed to be simple to reduce any possible modifications to the polarization state of the backscattered light. The maximum laser output power is 30 mJ per pulse and can be adjusted to meet the detection needs. The receiver has four channels, allowing dual-wavelength and dual-polarization detection, namely, the co-polarization and cross-polarization at the NIR and GN wavelengths, respectively. The laser pulses are backscattered by the vegetation or other targets and received by the four photomultiplier tube detectors. The precise timing capability of the digital delay generator is used to control the transmission and reception of the laser pulses, which improves the ranging accuracy and helps to obtain precise LiDAR waveforms that contain information about the vegetation canopy structure. The MAPL system, including the laser, the receivers, and data storage, is controlled by a self-developed LabVIEW program [7]. The data are stored in a hard drive and post-processed using MATLAB. The MAPL system is capable of performing both vegetation canopy structure studies and the characterization of vegetation polarimetric reflectance and depolarization. This configuration has proven to be able to improve target classification [16]. The block diagram of the MAPL is shown in Figure 1a, and Figure 1b shows a photograph of the receiver package with the laser head and four receivers.

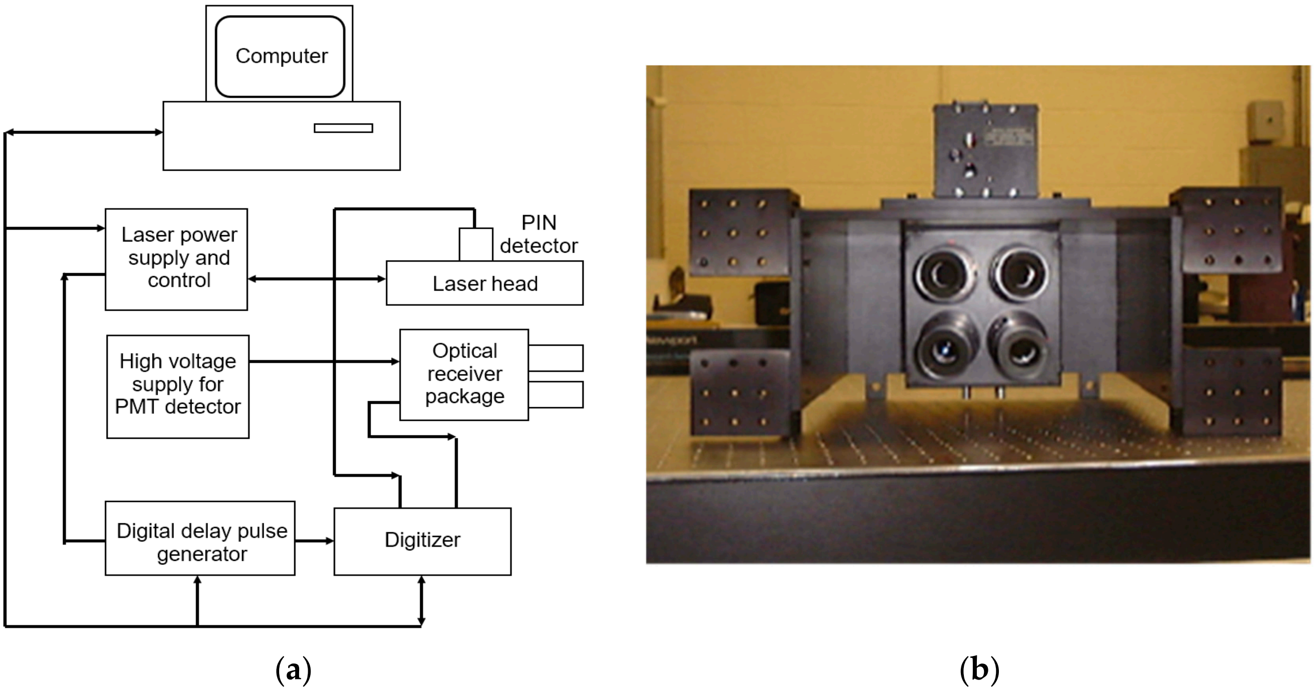
Computer
PIN detector
Laser power supply and control
Laser head
High voltage supply for PMT detector
Optical receiver package
Digital delay pulse generator
Digitizer
(a)
(b)

Figure 1. Schematic diagram of the MAPL system (**a**), and a photograph of the receiver package (**b**).

The LiDAR equation describing the relationship between the transmitted laser power and the received laser power is as follows [7]:

$$P_R = \frac{\pi P_T \rho_T D^2}{16R^2} \cdot {T_A}^2 \cdot \eta_T \cdot \eta_R, \tag{1}$$

where P_R is the LiDAR received power, P_T is the LiDAR transmitted power, R is the one-way distance from the LiDAR to the target, D is the optical detector aperture diameter, ρ_T is the target reflectivity, T_A is the one-way atmospheric transmission coefficient from the LiDAR to the target, η_T is the transmitter transmission efficiency, and η_R is the receiver transmission efficiency. As is seen, the received power is reversely proportional to the range squared for the extended target.

The received power is in two polarization directions, i.e., the co-polarization P_{RCO} and the cross-polarization P_{RX}. The cross-polarization ratio is defined as follows [25]:

$$\delta = \frac{P_{RX}}{P_{RCO}}. \tag{2}$$

The cross-polarization ratio is an important parameter used to quantify the target laser scattering property and has been used in various applications to characterize various targets [26].

3. Data Collection and Processing

The main objective of this study is to characterize the peak reflectance intensity from the trees captured by the MAPL system toward species classification. To achieve this objective, several steps need to be taken, starting with data collection and data preparation, followed by data analysis, and, finally, the interpretation of the findings. A flowchart is provided in Figure 2, detailing the steps of the proposed classification approach.

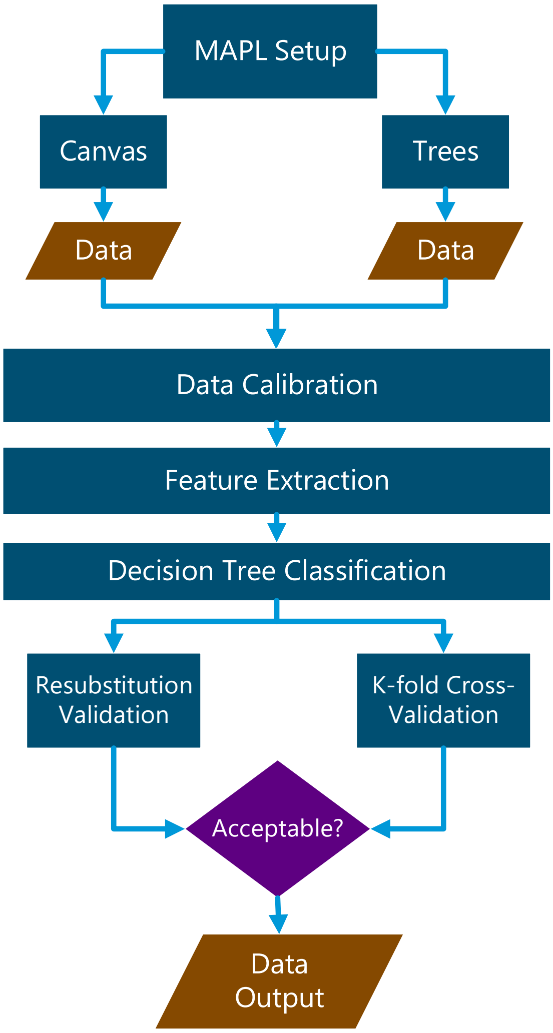

Figure 2. A flowchart of the proposed decision tree classification using MAPL data.

3.1. Data Collection

The purpose of selecting individual trees was to ensure that trees with different biophysical characteristics from a variety of representative forest groups were included. The tree species selected for this study were grouped into coniferous trees (blue spruce, ponderosa pine, and Austrian pine) and deciduous trees (ash and maple) based on their leaf structure, crown shape, and crown size. Generally, coniferous species are a group of plants with needle-like leaves and cone-shaped crowns. Whereas deciduous species are broad-leaf trees that shed their leaves annually to save energy, and the crown is generally more spherical in shape. These species are commonly found in eastern South Dakota. There are large amounts of coniferous and deciduous trees and shrubs, including a range of species native to the Mid-Eastern United States. The study sites are home to many different tree species with varying canopy structures. The presence of individual trees detected at each location in the study makes them suitable field sites for studying forest parameters at the individual tree level. In general, canopy overlapping is not an issue in field data collections, as the trees have been managed and are well-separated from each other. Consequently, the collected MAPL waveform contains information on tree canopy polarimetric scattering property, tree morphological structure, tree canopy size, etc.

All field data collections were carried out under clear weather conditions close to midday. This is a data consistency measure in order to minimize such effects of tree diurnal change and solar radiation change on the collected data. In addition, data collections were carried out in July during the leaf-on season. Our interest is the tree canopy and how canopy data can be used to discover different tree characteristics, and, potentially, tree stress or health conditions. Further, meteorological data were collected using a tripod-mounted weather station (Kestrel 4000 Pocket Weather Tracker). Wind speeds at each study site were acquired from the same weather station. As wind is a potential factor in altering the LiDAR waveform, all data collection was carried out with a wind speed of less than 3 mph. Field data were collected at various locations around the Brookings area in eastern South Dakota. The distance between the trees and the MAPL system was maintained at about 500 m. Before each tree measurement, a canvas tarp was set up close to the tree as a calibration standard. The MAPL laser was first aimed at the canvas to collect the calibration data; subsequently, it was aimed at the tree canopy for tree data collection. The same procedures were carried out for all trees. In our case, the green laser is visible and the near-infrared laser is collocated with the green laser by design, ensuring that the correct target is measured.

3.2. Processing of LiDAR Waveform Data

A LabVIEW program was written to control the MAPL hardware and the data collection [7]. It also displays the real-time waveform on the monitor. The output data were stored and post-processed by MATLAB. The sampled LiDAR waveforms of an Austrian Pine tree captured by the MAPL system are shown in Figure 3.

Data processing starts with radiometric calibration, range calibration, and then outlier removal. Firstly, radiometric calibration is achieved by using the canvas calibration standard. As mentioned previously, the MAPL system measures the canvas calibration standard before collecting any tree data [7]. A previous in-lab measurement of the canvas revealed that the canvas has a reflectivity of 7.1% at GN and 11.8% at NIR, and a cross-polarization ratio of 55% at GN and 65% at NIR, respectively. Four calibration constants for the four MAPL channels are then calculated based on the field canvas data to make sure that the field data agree with the in-lab measurement. And these constants are then used to calibrate the subsequent tree data through rescaling. This radiometric calibration ensures the tree data are consistent with the canvas standard. Therefore, any effect on the data caused by noise such as laser output power variation can be removed. Secondly, as the trees are not always the same distance away from the MAPL, a longer range will yield a weaker signal return as shown in Equation (1). Therefore, range calibration is performed to remove the effect of the range difference. This is carried out by normalizing all LiDAR data with respect to the range. In this case, the LiDAR return data are multiplied by the range squared. Finally, during data collection, there may be many sources introducing random noise into the data. Some noisy data are too far away from the mean and are considered as outliers by using a three standard-deviation from the mean criterion for each individual tree data. Any waveform with the peak value falling outside the three standard-deviation range is removed from further data processing. In our case, the outlier does not happen very often. Typically, there would be one outlier in several hundred samples. After outlier removal, we have a total number of 2106 datasets from all trees, in Table 1, which translates to a total of 8424 LiDAR waveforms. Table 1 also lists the number of valid datasets for each tree species. As is revealed in Figure 3, the waveforms from different channels are different, and so are the peak reflectance values. Therefore, in this study, the peak reflectance values at dual wavelength and dual polarization were selected as the feature for the subsequent classification.

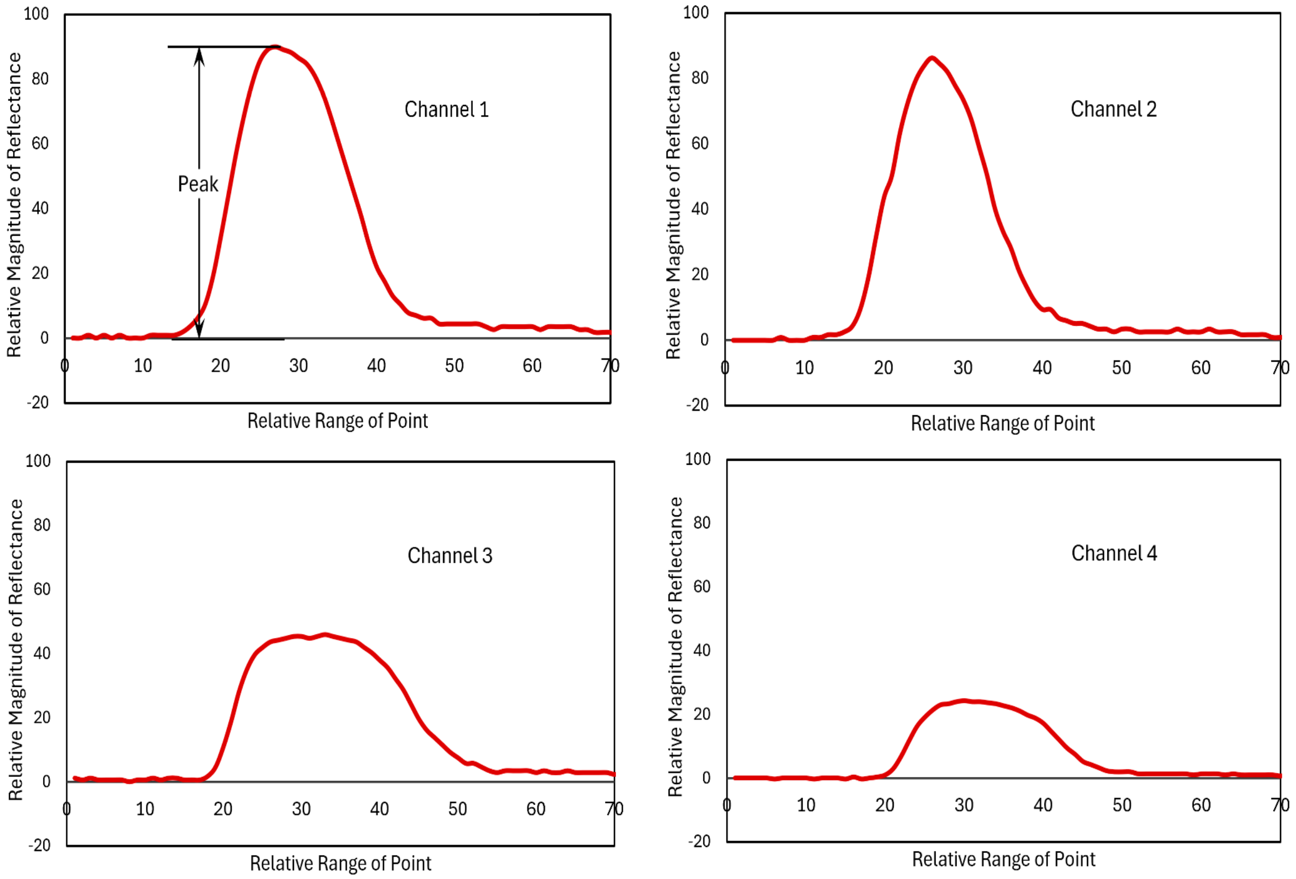

Figure 3. Sample LiDAR waveforms of an Austrian pine tree captured by four different polarimetric channels of the MAPL system, with *x*-axis representing the relative range in pixels and *y*-axis representing the relative reflectance in arbitrary unit (Ch1: NIR co-polarized, Ch2: NIR cross-polarized, Ch3: GN co-polarized, and Ch4: GN cross-polarized).

Table 1. The tree species and their corresponding number of datasets used in this study.

Tree Species	Dataset
Blue Spruce	244
Ash	277
Ponderosa Pine	795
Austrian Pine	318
Maple	472

4. Decision-Tree-Based ML Classification

Classification methods have wide applications in ML and data analysis. They provide effective tools to categorize data into predefined classes or groups based on specific features or attributes. Classification algorithms return predictions based on captured and processed data. A decision tree is a type of predictive modeling algorithm, which uses the divide-and-conquer strategy to obtain a hierarchical data structure. It is an effective nonparametric method for classification and regression [27,28]. The decision tree method has been used in LiDAR data classification. For instance, the LiDAR point cloud density and the standard deviation of the intensity/elevation have been used to classify water and nonwater at an Arctic region [29] or road boundaries for intelligent transportation [30]. Other efforts involve the use of deep learning to classify trees using the point cloud [31]. Comprehensive reviews on using ML for LiDAR data feature selection [32] and tree species classification are also available [33]. However, as the MAPL data are unique with polarization and full waveform information, the feature selection and classification are therefore different from these in the literature. Essentially, the polarimetric scattering property of the target is used for the classification in the case of MAPL data, instead of the commonly used LiDAR point cloud.

4.1. Decision Tree Classification

In this study, the peak values of the LiDAR waveforms, as shown in Figure 3, were first identified as the features/attributes. Then, a set of predefined tree class labels or classifiers, in this case, blue spruce, ash, ponderosa pine, Austrian pine, and maple, were assigned to each input data. Then, the decision tree algorithm assigns class labels using a tree-like structure, in which each internal node represents a "test" on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label. The paths from root to leaf represent classification rules which are used for the tree classification analysis. The approach was validated using re-substitution validation and *k*-fold cross-validation [34].

In order to directly view the relationships between different tree species and the tree feature distribution within their respective groups, a scatter plot matrix is grouped together to visualize the bi-variate relationships between the MAPL data. Figure 4 depicts the scatter plot matrix of the peak reflectance intensity captured by the MAPL four channels from five different tree species, where the blocks from left to right and from top to bottom represent Ch1 to Ch4, respectively. A preliminary analysis of Figure 4 shows that the maple data (green cross in the figure) are distinctly separated from other tree species. This facilitates the accurate and fast identification of the maple tree. In contrast, other species exhibit some degree of overlapping patterns in all plots and it will be more difficult to classify.

The scatter plot of Ch1 vs. Ch2 is further provided in Figure 5, where the distribution of the five tree species, each occupying a certain region within the plot to form a rough linear relationship, is revealed. However, for other pairs of variables, no obvious correlation is observed as the data are scattered throughout the plot area. To further analyze this initial visual observation, the cross-channel correlation coefficients were calculated and listed in Table 2. It is seen that the correlation coefficient of Ch1 and Ch2 is 0.9227 while the correlation coefficients between other channels are much lower. The high positive correlation between Ch1 and Ch2 is yet to be explained.

Table 2. Cross-channel correlation coefficients.

	Ch1	**Ch2**	**Ch3**	**Ch4**
Ch1	1.0000	0.9227	0.3258	−0.4050
Ch2	0.9227	1.0000	0.3537	−0.4343
Ch3	0.3258	0.3537	1.0000	−0.3435
Ch4	−0.4050	−0.4343	−0.3435	1.0000

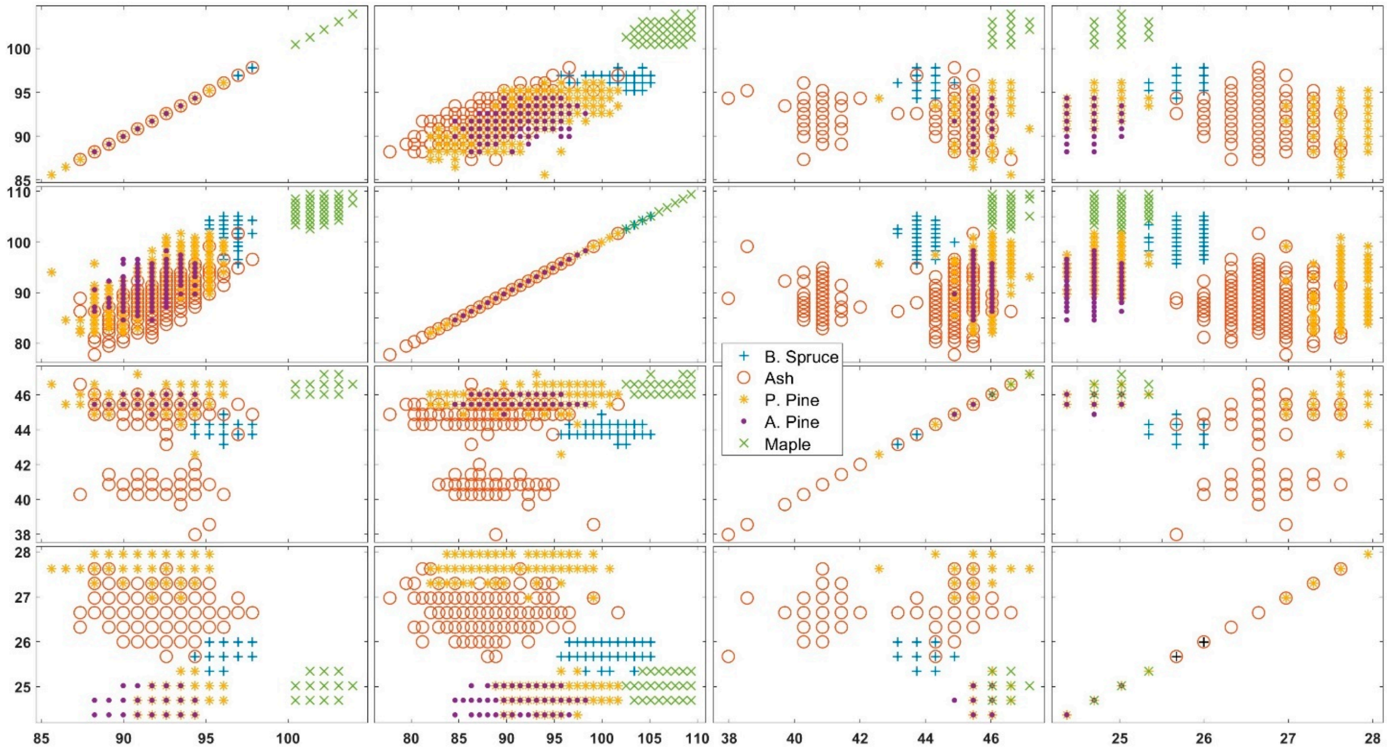

Figure 4. Matrix graphs of scatter plots grouped by different tree species. The horizontal axes from left to right blocks and vertical axes from top to bottom blocks represent Ch1, Ch2, Ch3, and Ch4, respectively. Both the x axes and y axes are peak intensity with arbitrary units.

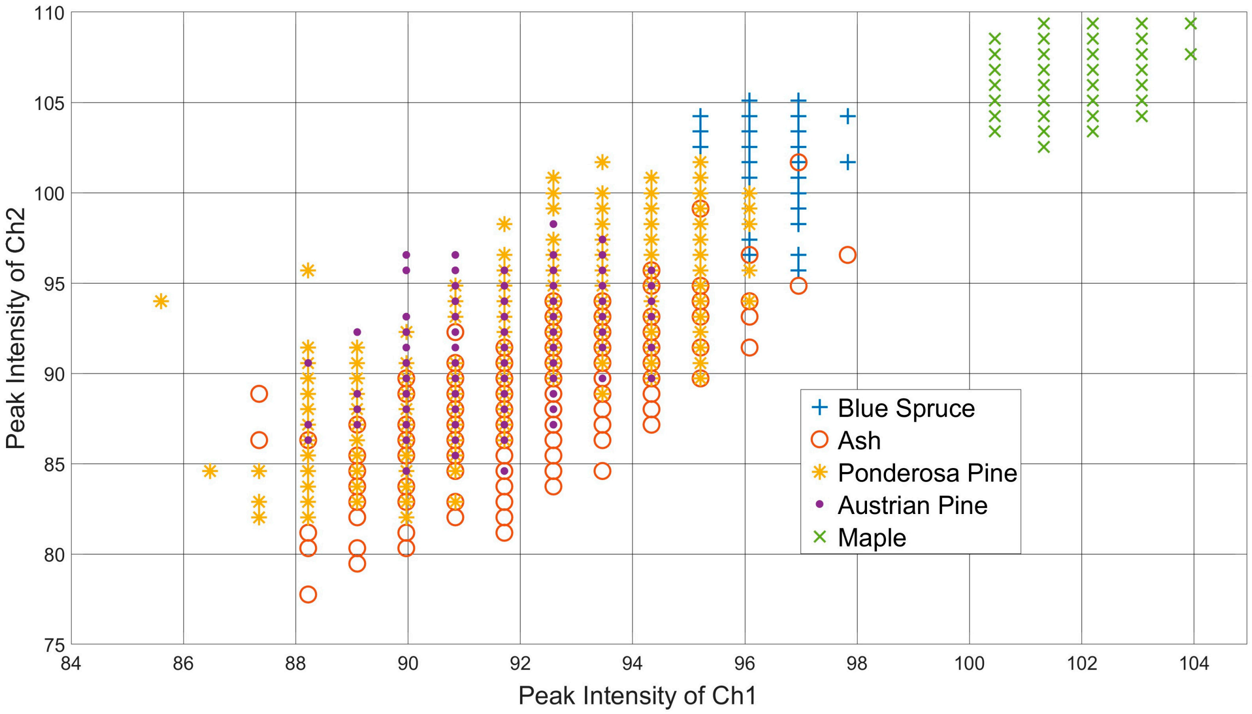

Figure 5. Scatter plot of Ch1 vs. Ch2 grouped by different tree species.

After running the decision tree program, Figure 6 graphically demonstrates the species decision-making process, where the nodes are numbered from 1 to 71. Moreover, $x1$ to $x4$ represent peak intensity values from Ch1 to Ch4, respectively. As is seen, for example, when $x1 \geq 99.14$, the decision-making process for the maple species is determined in a first step. The classification processes for other species are much more complex due to their overlapping as demonstrated in Figures 4 and 5.

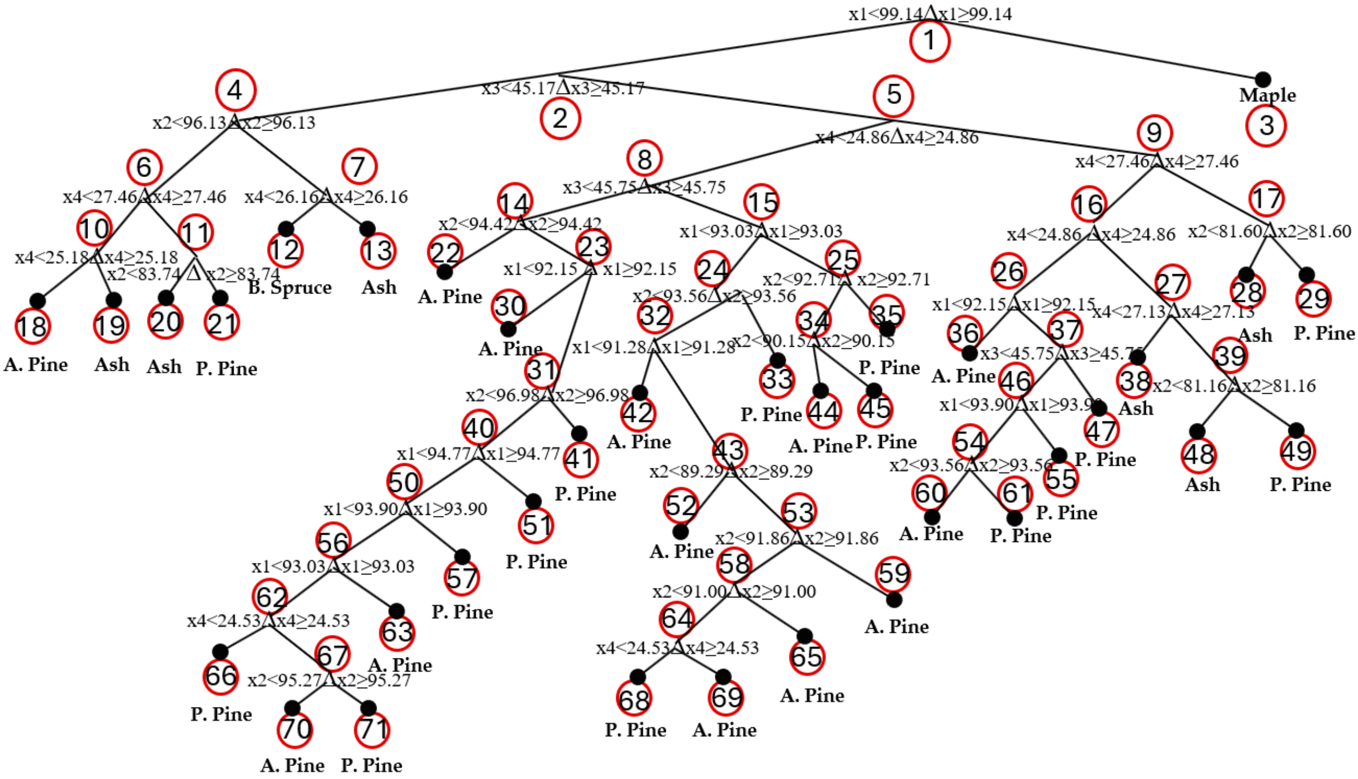

Figure 6. Graphical description of the decision-making process.

The decision tree text description structure helps convey the same information revealed in Figure 6 with enhanced readability by emphasizing the specific branches that lead to the decision or evaluation, as is listed below in node order:

1. if x1 < 99.14 then node 2 elseif x1 ≥ 99.14 then node 3
2. if x3 < 45.17 then node 4 elseif x3 ≥ 45.17 then node 5
3. **class = Maple**
4. if x2 < 96.13 then node 6 elseif x2 ≥ 96.13 then node 7
5. if x4 < 24.86 then node 8 elseif x4 ≥ 24.86 then node 9
6. if x4 < 27.46 then node 10 elseif x4 ≥ 27.46 then node 11
7. if x4 < 26.16 then node 12 elseif x4 ≥ 26.16 then node 13
8. if x3 < 45.75 then node 14 elseif x3 ≥ 45.75 then node 15
9. if x4 < 27.46 then node 16 elseif x4 ≥ 27.46 then node 17
10. if x4 < 25.18 then node 18 elseif x4 ≥ 25.18 then node 19
11. if x2 < 83.74 then node 20 elseif x2 ≥ 83.74 then node 21
12. **class = Blue Spruce**
13. **class = Ash**
14. if x2 < 94.42 then node 22 elseif x2 ≥ 94.42 then node 23
15. if x1 < 93.03 then node 24 elseif x1 ≥ 93.03 then node 25
16. if x4 < 25.83 then node 26 elseif x4 ≥ 25.83 then node 27
17. if x2 < 81.60 then node 28 elseif x2 ≥ 81.60 then node 29
18. **class = Austrian Pine**
19. **class = Ash**

20. **class = Ash**
21. **class = Ponderosa Pine**
22. **class = Austrian Pine**
23. if x1 < 92.15 then node 30 elseif x1 ≥ 92.15 then node 31
24. if x2 < 93.56 then node 32 elseif x2 ≥ 93.56 then node 33
25. if x2 < 92.71 then node 34 elseif x2 ≥ 92.71 then node 35
26. if x1 < 92.15 then node 36 elseif x1 ≥ 92.15 then node 37
27. if x4 < 27.13 then node 38 elseif x4 ≥ 27.13 then node 39
28. **class = Ash**
29. **class = Ponderosa Pine**
30. **class = Austrian Pine**
31. if x2 < 96.98 then node 40 elseif x2 ≥ 96.98 then node 41
32. if x1 < 91.28 then node 42 elseif x1 ≥ 91.28 then node 43
33. class = Ponderosa Pine
34. if x2 < 90.15 then node 44 elseif x2 ≥ 90.15 then node 45
35. **class = Ponderosa Pine**
36. **class = Austrian Pine**
37. if x3 < 45.75 then node 46 elseif x3 ≥ 45.75 then node 47
38. **class = Ash**
39. if x2 < 81.18 then node 48 elseif x2 ≥ 81.18 then node 49
40. if x1 < 94.77 then node 50 elseif x1 ≥ 94.77 then node 51
41. **class = Ponderosa Pine**
42. **class = Austrian Pine**
43. if x2 < 89.29 then node 52 elseif x2 ≥ 89.29 then node 53
44. **class = Austrian Pine**
45. **class = Ponderosa Pine**
46. if x1 < 93.90 then node 54 elseif x1 ≥ 93.90 then node 55
47. **class = Ponderosa Pine**
48. **class = Ash**
49. **class = Ponderosa Pine**
50. if x1 < 93.90 then node 56 elseif x1 ≥ 93.90 then node 57
51. **class = Ponderosa Pine**
52. **class = Austrian Pine**
53. if x2 < 91.86 then node 58 elseif x2 ≥ 91.86 then node 59
54. if x2 < 93.56 then node 60 elseif x2 ≥ 93.56 then node 61
55. **class = Ponderosa Pine**
56. if x1 < 93.03 then node 62 elseif x1 ≥ 93.03 then node 63
57. **class = Ponderosa Pine**
58. if x2 < 91.00 then node 64 elseif x2 ≥ 91.00 then node 65
59. **class = Austrian Pine**
60. **class = Austrian Pine**
61. **class = Ponderosa Pine**
62. if x4 < 24.53 then node 66 elseif x4 ≥ 24.53 then node 67
63. **class = Austrian Pine**
64. if x4 < 24.53 then node 68 elseif x4 ≥ 24.53 then node 69
65. **class = Austrian Pine**
66. **class = Ponderosa Pine**
67. if x2 < 95.27 then node 70 elseif x2 ≥ 95.27 then node 71
68. **class = Ponderosa Pine**
69. **class = Austrian Pine**
70. **class = Austrian Pine**
71. **class = Ponderosa Pine**

4.2. Model Performance Evaluation and Validation

After building a classification tree, it is necessary to assess the capability of the model in predicting newly observed data. A common approach is to calculate the re-substitution error rate, which is the difference between the predicted species classes and the actual species classes in the training dataset. This can be used as an initial estimate of the model's performance. To calculate the re-substitution error rate, the steps are as follows:

i. Fit the Decision Tree Model: Train the decision tree model using the training samples (in this case, a total of 2106 × 4 = 8424 waveforms), which include features and corresponding labels.
ii. Make Predictions: Use the trained decision tree model to make predictions on the training datasets. Each dataset in the training dataset will be classified into a specific class by the decision tree model.
iii. Compare Predictions with Actual Labels: The misclassification rate is the rate of incorrectly classified instances in the training dataset.
iv. Calculate the Re-substitution Error Rate [34]:

$$\text{Re\_substitution Error Rate} = \frac{\text{Number of Misclassified instances}}{\text{Total Number of Instances}} \times 100\%. \tag{3}$$

In this study, the re-substitution error of the model was calculated to be 0.0323 (i.e., 3.23%). A low re-substitution error value indicates that the classification tree accurately classifies the data.

In order to further assess the prediction accuracy of the decision tree, a k-fold cross-validation algorithm is adopted. The process partitions the datasets into k equal-sized folds, where $k - 1$ folds are used for training and the remaining one-fold is used for testing. The process will iterate k times; therefore, all data are tested. In this study, k = 10 is selected to partition the 8424 waveforms. Usually, k value should be $\geq$5, ensuring that the training process has enough samples. However, if the k value is too large, it will lead to less variance across the training set. Consequently, k = 10 is a good trade-off in our case to achieve high accuracy and better generalization and predictability. The model was trained and evaluated in k iterations, with each iteration using a different fold (e.g., the ith fold, where $i = 1, 2, \ldots, k$) as the validation set while using the remaining folds ($k - 1$ folds) as the training set. The results of each iteration were then averaged to produce a comprehensive performance estimation. The average performance metric was used to compare and select the best model among different algorithms or hyperparameter configurations. The most common way of calculating the k-fold cross-validation error is to use the mean squared error (MSE) [34]

$$MSE = \frac{1}{n}\sum (y_i - f(x_i))^2, \tag{4}$$

where n is the total number of the datasets, y_i is the actual labeled response value of the ith observation, and $f(x_i)$ is the prediction value of the ith observation.

Unlike the re-substitution error, this approach can reliably estimate the predictive accuracy of the resulting tree species because it tests the new trees with new and unknown data. In this study, the calculated cross-validation loss error is 0.0503 (i.e., 5.03%). The low error value suggests that the model performs well.

5. Discussions

Many factors affect the overall accuracy of the MAPL system. Radiometric calibration is a key factor as it rescales data from all four channels. The canvas tarp proves to be a reasonably good calibration standard. Although care was taken to make sure that the canvas surface remains intact during field campaigns, the folding of the canvas, dirt and dust getting on the canvas, aging of the tarp, etc. are all factors which may potentially affecting measurement accuracy [35]. There are many noise sources that introduce random noise into the signal during field measurement. For example, it was discovered that the field power source, a portable gas generator (Honda EG2200 series with a power output of 2.2 kW), was one source of noise. Cleaner power tends to produce better signals, and, hence, may improve classification accuracy.

On the other hand, exploring other ML approaches needs to be considered for further vegetation classification research, especially unsupervised learning approaches. For instance, clustering, one of the unsupervised ML methods, involves identifying groups of data based on the proximity between elements/measurements. Proximity can refer to either similarity or dissimilarity. Therefore, the classification of data groups depends on how similarity or dissimilarity is defined. Clustering algorithms group elements based on the mutual distances between them, where membership in a particular set is determined by the proximity of an element to other members of the same set. Recent developments on DL and image-based NNs provide alternative options for classification efforts.

Upon careful inspection of the subplots in Figure 4, it is difficult to identify the tree species category with high accuracy and easy pathways, because the profiles and responses of several tree species are very similar, resulting in overlapping responses. Instead, it may be easier to cluster the tree species into coniferous or deciduous tree types. Otherwise, in future studies, more tree features (such as adding the width of the captured signal pulse, or even the entire signal pulse profile/image) can be selected for classification or clustering to improve accuracy and robustness.

In summary, future research should prioritize enhancing radiometric calibration by utilizing a more reliable calibration standard rather than a commercial canvas tarp. Additionally, a more efficient power regulator may reduce power noise. It is also helpful to collect more tree data and investigate additional features to improve our understanding of the classification process. Most importantly, a comprehensive theoretical framework for laser polarimetric scattering in tree canopies has yet to be developed. Establishing such a theory would significantly advance our understanding of LiDAR signal generation and facilitate the design of more effective classification strategies. Furthermore, other ML methods such as clustering, DL, and image-based NNs should be explored. These new methods may provide fresh insights into the MAPL data and the classification process as well. Finally, the MAPL system, when fused with other imaging sensor data, can deliver augmented reality such as Simultaneous Localization and Mapping (SLAM) and Bird's Eye View (BEV) [25,36]. Recently, much LiDAR-based multi-sensor fusion SLAM research has emerged to make it more stable and accurate. Latest advancements include more complex systems integrated with multiple sensors, more sophisticated fusion algorithms using optimization, and more robust error reduction using tightly coupled complete graphical models. In addition, the unique polarimetric capability of the MAPL provides extra information and enables target characterization beyond shape and distance. For instance, if a cylindrical object is detected, polarimetric data can be used to identify whether the object is made of wood, stone, or metal.

6. Conclusions

The MAPL system provides a unique opportunity for vegetation and forest remote sensing. With its dual-wavelength, dual-polarization, and full waveform capabilities, it is still the only such system available for vegetation research. In this study, tree species classification was successfully performed using a decision-tree-based ML method. The approach is accurate and effective. The following key conclusions have been drawn:

- Polarimetric measurement has been proven to be an effective method for target detection. Polarimetric diversity enhances measurement and provides more information on target characterization.
- The MAPL peak reflectance intensity data, at dual wavelength and dual polarization, is an effective and simple feature for classification purposes.
- The decision tree algorithm proves to be effective in this case as suggested by the re-substitution error and the *k*-fold cross-validation loss error.

- The method developed in this study can be extended to new data and other vegetation classification applications.

Author Contributions: Conceptualization, S.T. and Z.H.; Data preparation and feature extraction, S.T.; decision tree analysis, Z.H.; writing—original draft preparation, Z.H.; writing—review and editing, S.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author, ST, upon reasonable request.

Acknowledgments: The authors would like to thank former graduate students Ali Haider and Nabeen Thakur for assisting in field data collections.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Masson-Delmotte, V.; Zhai, P.M.; Pirani, A.; Connors, S.L.; Péan, C.; Berger, S.; Huang, M.T.; Yelekçi, O.; Yu, R.; Zhou, B.Q. Climate Change 2021: The Physical Science Basis. In *Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*; IPCC: Geneva, Switzerland, 2021; Volume 2.
2. Ma, Z.Y.; Ren, J.X.; Chen, H.T.; Jiang, H.; Gao, Q.X.; Liu, S.L.; Yan, W.; Li, Z.M. Analysis and Recommendations of the IPCC Working Group I Assessment Report. *Resour. Environ. Sci.* **2022**, *35*, 2550–2558.
3. United Nations, Department of Economic and Social Affairs. *Envision2030: 17 Goals to Transform the World for Persons With Disabilities [Internet]*; United Nations, Department of Economic and Social Affairs: New York, NY, USA, 2024. Available online: https://social.desa.un.org/issues/disability/envision-2030/17goals-pwds (accessed on 28 October 2024).
4. Afuye, G.A.; Kalumba, A.M.; Orimoloye, I.R. Characterisation of vegetation response to climate change: A review. *Sustainability* **2021**, *13*, 7265. [CrossRef]
5. Canto-Sansores, W.G.; López-Martínez, J.O.; González, E.J.; Meave, J.A.; Hernández-Stefanoni, J.L.; Macario-Mendoza, P.A. The importance of spatial scale and vegetation complexity in woody species diversity and its relationship with remotely sensed variables. *ISPRS J. Photogramm. Remote Sens.* **2024**, *216*, 142–253. [CrossRef]
6. Rapinel, S.; Hubert-Moy, L. One-class classification of natural vegetation using remote sensing: A review. *Remote Sens.* **2021**, *13*, 1892. [CrossRef]
7. Tan, S. Development of a Multiwavelength Airborne Polarimetric Lidar for Vegetation Remote Sensing. Ph.D. Dissertation, University of Nebraska-Lincoln, Lincoln, NE, USA, 2003.
8. Kalshoven, J.E.; Tierney, M.R.; Daughtry, C.S.T.; McMurtrey, J.E. Remote sensing of crop parameters with a polarized, frequency-doubled Nd:YAG laser. *Appl. Opt.* **1995**, *34*, 2745–2749. [CrossRef]
9. Lefsky, M.A.; Harding, D.; Cohen, W.B.; Parker, G.; Shugart, H.H. Surface lidar remote sensing of basal area and biomass in deciduous forests of eastern Maryland, USA. *Remote Sens. Environ.* **1999**, *67*, 83–98. [CrossRef]
10. Kogut, T.; Bakuła, K. Improvement of Full Waveform Airborne Laser Bathymetry Data Processing based on Waves of Neighborhood Points. *Remote Sens.* **2019**, *11*, 1255. [CrossRef]
11. Haider, A.; Tan, S. Improvement of lidar data classification algorithm using machine learning technique. In Proceedings of the SPIE Polarization Science and Remote Sensing IX, San Diego, CA, USA, 14–15 August 2019; Volume 11132.
12. Tan, S.; Khan, A. Water stress detection of lilac leaves using a polarized laser. In Proceedings of the SPIE, Remote Sensing and Modeling of Ecosystems for Sustainability XII, San Diego, CA, USA, 1–12 August 2015; Volume 9610.
13. Tan, S.; Stoker, J.; Greenlee, S. Detection of foliage-obscured vehicle using a multiwavelength polarimetric lidar. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, Barcelona, Spain, 23–28 July 2007; Volume 4423, pp. 2503–2506.
14. Saarela, S.; Holm, S.; Healey, S.P.; Andersen, H.-E.; Petersson, H.; Prentius, W.; Patterson, P.L.; Næsset, E.; Gregoire, T.G.; Ståhl, G. Generalized Hierarchical Model-Based Estimation for Aboveground Biomass Assessment Using GEDI and Landsat Data. *Remote Sens.* **2018**, *10*, 1832. [CrossRef]
15. Wołk, K.; Tatara, M.S. A Review of Semantic Segmentation and Instance Segmentation Techniques in Forestry Using LiDAR and Imagery Data. *Electronics* **2024**, *13*, 4139. [CrossRef]
16. Tan, S.; Haider, A. A comparative study of polarimetric and non-polarimetric lidar in deciduous—Coniferous tree classification. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, Honolulu, HI, USA, 25–30 July 2010; pp. 1178–1181.
17. Tan, S.; Narayanan, R.; Helder, D. Polarimetric reflectance and depolarization ratio from several tree species using a multiwavelength polarimetric lidar. In Proceedings of the SPIE, Polarization Science and Remote Sensing II, San Diego, CA, USA, 2–4 August 2005; Volume 5888.
18. Hu, Z.; Zhou, R. Review on some important research progresses in biodegradable plastics/polymers. *Recent Prog. Mater.* **2024**, *6*, 015. [CrossRef]

19. Samuel, A. Some studies in machine learning using the game of checkers. *IBM J. Res. Dev.* **1959**, *3*, 210–229. [CrossRef]
20. Wang, T. Improved random forest classification model combined with C5.0 algorithm for vegetation feature analysis in non-agricultural environments. *Sci. Rep.* **2024**, *14*, 10367. [CrossRef] [PubMed]
21. Li, X.; Liu, Y.; Wang, L. Change in fractional vegetation cover and its prediction during the growing season based on machine learning in Southwest China. *Remote Sens.* **2024**, *16*, 3623. [CrossRef]
22. Piaser, E.; Villa, P. evaluating capabilities of machine learning algorithms for aquatic vegetation classification in temperate wetlands using multi-temporal Sentinel-2 data. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *117*, 103202. [CrossRef]
23. Drobnjak, S.; Stojanović, M.; Djordjević, D.; Bakrač, S.; Jovanović, J.; Djordjević, A. Testing a new ensemble vegetation classification method based on deep learning and machine learning methods using aerial photogrammetric images. *Front. Environ. Sci.* **2022**, *10*, 896158. [CrossRef]
24. Campos-Tabener, M.; García-Haro, F.J.; Martínez, B.; Izquierdo-Verdiguier, E.; Atzberger, C.; Camps-Valls, G.; Gilabert, M.A. Understanding deep learning in land use classification based on Sentinel-2 time series. *Sci. Rep.* **2020**, *10*, 17188.
25. Tan, S.; Narayanan, R. Design and performance of a multiwavelength airborne polarimetric lidar (MAPL) for vegetation remote sensing. *Appl. Opt.* **2004**, *43*, 2360–2368. [CrossRef]
26. Tan, S.; Narayanan, R. A multiwavelength airborne polarimetric lidar for vegetation remote sensing: Instrumentation and preliminary test results. *IEEE Geosci. Remote Sens. Symp.* **2002**, *5*, 2675–2677.
27. Breiman, L.; Friedman, J.H.; Olshen, R.A.; Stone, C.J. *Classification and Regression Trees*; Chapman & Hall: Boca Raton, FL, USA, 1984.
28. Quinlan, J.R. *C4.5: Programs for Machine Learning*; Morgan Kaufmann: San Mateo, CA, USA, 1993.
29. Crasto, N.; Hopkinson, C.; Forbes, D.L.; Lesack, L.; Marsh, P.; Spooner, I.; van der Sanden, J.J. A LiDAR-based decision-tree classification of open water surfaces in an Arctic delta. *Remote Sens. Environ.* **2015**, *164*, 90–102. [CrossRef]
30. Zheng, J.; Yang, S.; Wang, X.; Xia, X.; Xiao, Y.; Li, T. A Decision Tree Based Road Recognition Approach Using Roadside Fixed 3D LiDAR Sensors. *IEEE Access* **2019**, *7*, 53878–53890. [CrossRef]
31. Zou, X.; Cheng, M.; Wang, C.; Xia, Y.; Li, J. Tree Classification in Complex Forest Point Clouds Based on Deep Learning. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 2360–2364. [CrossRef]
32. Gharineiat, Z.; Tarsha Kurdi, F.; Campbell, G. 2022. Review of automatic processing of topography and surface feature identification LiDAR data using machine learning techniques. *Remote Sens.* **2022**, *14*, 4685. [CrossRef]
33. Michałowska, M.; Rapiński, J. A Review of tree species classification based on airborne LiDAR data and applied classifiers. *Remote Sens.* **2021**, *13*, 353. [CrossRef]
34. Ciaburro, G. *MATLAB for Machine Learning—Unlock the Power of Deep Leaning for Swift and Enhanced Results*, 2nd ed.; Packt Publishing: Birmingham, UK, 2024.
35. Tan, S.; Johnson, S.; Gu, Z. Laser depolarization ratio measurement of corn leaves from the biochar and non-biochar applied plots. *Opt. Express* **2018**, *26*, 14295–14306. [CrossRef]
36. Xu, X.; Zhang, L.; Yang, J.; Cao, C.; Wang, W.; Ran, Y.; Tan, Z.; Luo, M. A Review of Multi-Sensor Fusion SLAM Systems Based on 3D LIDAR. *Remote Sens.* **2022**, *14*, 2835. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.