https://doi.org/10.5194/egusphere-2025-970
Preprint. Discussion started: 28 May 2025

Identifying alpine treeline species using high-resolution WorldView-3 multispectral imagery and convolutional neural networks

Laurel A. Sindewald[1], Ryan Lagerquist[2], Matthew D. Cross[3], Theodore A. Scambos[4], Peter J. Anthamatten[5], Diana F. Tomback[1]

[1] Department of Integrative Biology, University of Colorado Denver, Denver, 80204, USA
[2] Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, 80521, USA
[3] Department of Geography and the Environment, University of Denver, Denver, 80208, USA
[4] Earth Science and Observation Center, Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder, Boulder, 80309, USA
[5] Department of Geography and Environmental Sciences, University of Colorado Denver, Denver, 80204, USA

Correspondence to: Laurel A. Sindewald (laurel.sindewald@ucdenver.edu)

Abstract

Alpine treeline systems are remote and difficult to access, making them natural candidates for remote sensing applications. Remote sensing applications are needed at multiple scales to connect landscape-scale responses to climate warming to finer-scale spatial patterns, and finally to community processes. Reliable, high-resolution tree species identification over broad geographic areas is important for connecting patterns to underlying processes, which are driven in part by species' tolerances and interactions (e.g., facilitation). To our knowledge, we are the first to attempt tree species identification at treeline using satellite imagery. We used convolutional neural networks (CNNs) trained with high-resolution WorldView-3 multispectral and panchromatic imagery, to distinguish six tree and shrub species found at treeline in the southern Rocky Mountains: limber pine *(Pinus flexilis)*, Engelmann spruce *(Picea engelmannii)*, subalpine fir *(Abies lasiocarpa)*, quaking aspen *(Populus tremuloides)*, glandular birch *(Betula glandulosa)*, and willow *(Salix spp.)*. We delineated 615 polygons in the field with a Trimble geolocator, aiming to capture the high intra- and interspecies variation found at treeline. We adapted our CNN architecture to accommodate the higher-resolution panchromatic and lower-resolution multispectral imagery within the same architecture, using both datasets at their native spatial resolution. We trained four- and two-class models with aims to 1) discriminate conifers from each other and from deciduous species, and 2) to discriminate limber pine—a keystone species of conservation concern—from the other species. Our models performed moderately well, with overall accuracies of 44.1%, 46.7%, and 86.2% for the six-, four-, and two-class models, respectively (as compared to random models, which could achieve 28.0%, 35.1%, and 80.3%, respectively). In future work, our models may be easily adapted to perform object-based classification, which will improve these accuracies substantially and will lead to cost-effective, high-resolution tree species classification over a much wider geographic extent than can be achieved with uncrewed aerial systems (UAS), including regions that prohibit UAS, such as in National Parks in the U.S.

https://doi.org/10.5194/egusphere-2025-970
Preprint. Discussion started: 28 May 2025

1 Introduction

Alpine treeline—the altitudinal limit of tree growth in mountain ecosystems—is remote, rugged, and often difficult or dangerous to access. These factors compound data limitations already prevalent in the field of ecology. Treeline systems are notoriously heterogeneous, and factors that limit treeline elevation exist on many scales (Malanson et al., 2007). Remote sensing technologies potentially overcome access limitations and may enable treeline ecologists to investigate patterns connected to the underlying processes that drive treeline ecologies at multiple scales (Garbarino et al., 2023).

Aerial photography and satellite imagery have been used for decades to map treeline position across regions, which has contributed to the identification of the variables that control treeline position (Wei and Karger and Wilson, 2020; Leonelli and Masseroli and Pelfini, 2016; Brown et al., 1994; Allen and Walsh, 1996). More recently, research has focused on determining where treelines have advanced to higher elevations and/or densified (Garbarino et al., 2023; Feuillet et al., 2020), as they are predicted to do with increasing average global temperatures (Harsch et al., 2009; Körner, 1998; Holtmeier and Broll, 2005; Brodersen et al., 2019). So far, remote sensing studies at treeline have been primarily focused on patterns and the connection to process is often missing. For example, it is known that some treelines are advancing while others are not (Harsch et al., 2009) but the reasons for this remain unknown and may vary by treeline system (Feuillet et al., 2020).

Bader et al. (2021) presented a useful global framework to guide hypothesis formation about how community structures and spatial patterns are driven by underlying ecological processes, aiming to identify parallels or commonalities across geographic regions. They postulated that concerted efforts to discover and connect these patterns and processes are key to understanding treeline ecosystems, including complex treeline shifts or responses to climate change on multiple scales. In other words, we require remote sensing applications that can extrapolate from finer-scale community processes to spatial patterns within treeline communities to larger scale patterns of treeline community distribution (Garbarino et al., 2023; Bader et al., 2021).

Field research at treeline has demonstrated important how species-specific tolerances and facilitative interactions may influence the position of alpine treeline (Brodersen et al., 2019; Mcintire and Piper and Fajardo, 2016; Resler and Butler and Malanson, 2005). For example, treelines formed by *Nothofagus* species in New Zealand and Hawaii are 200-500 m lower than expected from global isotherms, and *Metrosideros* treelines in Hawaii are also several hundred meters lower than those dominated by *Picea abies*, likely due to species-specific tolerances (Körner and Paulsen, 2004). Certain conifer species, such as limber pine *(Pinus flexilis)* and whitebark pine *(Pinus albicaulis)*, both white pines in Family Pinaceae and Subgenus *Strobus*, are more drought- and stress-tolerant at the seedling stage than other conifers in the Rocky Mountains, which may confer an advantage for establishment under harsh treeline conditions (Ulrich et al., 2023; Hankin and Bisbing, 2021; McCune, 1988; Bansal and Reinhardt and Germino, 2011). In general, the seedling stage is particularly vulnerable to abiotic

https://doi.org/10.5194/egusphere-2025-970
Preprint. Discussion started: 28 May 2025

stressors such as high growing season temperatures and drought (Cui and Smith, 1991; Germino and Smith and Resor, 2002), and recruitment at treeline tends to occur in pulses associated with consecutive years of higher moisture and cooler temperatures (Millar et al., 2015; Batllori and Gutiérrez, 2008). Both limber and whitebark pine are able to establish as seedlings without facilitative aid, i.e., from nurse objects or from other conifers, with greater frequency than do other forest trees (Sindewald and Tomback and Neumeyer, 2020; Resler et al., 2014; Wagner et al., 2018; Tomback et al., 2016a).

Species-specific facilitative interactions are also important for treeline advance in climatically limited systems, and stress-tolerant species are more likely to act as facilitators (Callaway, 1998; Resler and Butler and Malanson, 2005; Pyatt et al., 2016; Batllori et al., 2009). For example, whitebark pine is known to serve as a tree island initiator, facilitating the leeward establishment of other conifers and so conferring greater growth and survival for leeward trees and seedlings (Tomback et al., 2016a; Tomback et al., 2016b; Pyatt et al., 2016). Clearly, species identification is an important link between pattern and process in treeline systems where multiple species are present, and remote identification of species would be, quite literally, instrumental.

Remote sensing applications for tree species identification have proliferated with the continuous improvement of spatial, spectral, and radiometric resolutions. These advances in remote sensing technology have led to an exponential increase in species identification studies since 1990 (Fassnacht et al., 2016; Pu, 2021). To date, one study has attempted species identification at treeline using uncrewed aerial systems (UAS) (Mishra et al., 2018; Garbarino et al., 2023). Mishra et al. (2018) succeeded in achieving 73% overall accuracy in identifying four tree species across a ~ 140 m x 80 m region of the Himalayas using multispectral UAS imagery. This success highlights the potential and effectiveness of high-resolution UAS data for treeline species identification using an object-based classification approach with image segmentation. This technique can generate vegetation surveys in the Himalayas in a fraction of the time compared to previous methods, and demonstrates the importance of continuing the development of UAS methods for community-level treeline studies. However, work with UAS requires days in the field and some degree of site access, and so may not be suitable for locations with extremely rugged terrain.

Airborne hyperspectral imagery and lidar have also enabled tree species identification (Dalponte and Bruzzone and Gianelle, 2012; Matsuki and Yokoya and Iwasaki, 2015; Liu et al., 2017; Shen and Cao, 2017; Voss and Sugumaran, 2008), in some cases with classification accuracies ranging from 76.5-93.2% (Dalponte and Bruzzone and Gianelle, 2012). However, the sensors and overflights are relatively costly. Traditionally, satellite remote sensing is less costly but often comes at the cost of lower spatial resolution. Recently, with the advent of better access and more affordable imagery from higher resolution imaging systems, the scientific community has more choices for tree species research. Airborne multispectral imagery also generally yields high classification accuracy (85.8%) (Dalponte and Bruzzone and Gianelle, 2012), suggesting that several

https://doi.org/10.5194/egusphere-2025-970
Preprint. Discussion started: 28 May 2025

approaches with current technology can support species-level tree identification and mapping using a high resolution imaging system.

Despite the lower spatial resolution of satellite imagery, Cross et al. (2019) accomplished high-accuracy (85.37%) discrimination among seven rainforest tree species within the La Selva Research Center in Costa Rica using high resolution WorldView-3 (WV-3) imagery (Cross et al., 2019b; Cross et al., 2019a). In this work, a field spectroradiometer was used to determine the foliage spectral reflectance curves (light reflectance) of individual tree species. The curves measured in the field were then compared with the spectral reflectance curves observed in the WV-3 imagery, after atmospheric correction, as a spectral groundtruth (Cross et al., 2019a). Two spectral vegetation indices specific to WV-3 bands were developed (Cross et al., 2019b) and used for object-based classification of a segmented image. Prior to this work, applications of WV-3 imagery for species identification had mixed success and the imagery was often used in combination with machine learning or airborne lidar (Immitzer and Atzberger and Koukal, 2012; Li et al., 2015; Majid and Latif and Adnan, 2016; Wang et al., 2016; Rahman and Robson and Bristow, 2018).

Here we describe what may be the first to attempt to remotely identify plant species in a treeline system using satellite imagery (Garbarino et al., 2023). We aimed to discriminate six alpine treeline tree and shrub species in the southern Rocky Mountains, using a pixel-based convolutional neural network (CNN) classification of high-resolution WV-3 satellite imagery. CNNs are a type of deep learning model commonly used for image recognition tasks. They are effective at detecting patterns in imagery (or other gridded data) at multiple scales (Goodfellow and Bengio and Courville, 2016d; Dubey and Jain, 2019). The focus of our work was to discriminate limber pine—a keystone species of conservation concern—from other species (Schoettle et al., 2019). Limber pine populations are threatened by the spread of white pine blister rust, a disease caused by the non-native, invasive fungal pathogen *Cronartium ribicola;* limber pine has already been listed as endangered in Alberta, Canada (Jones et al., 2014; Schoettle et al., 2022). Limber pine is expected to migrate to higher elevations as climate changes (Monahan et al., 2013), but its current treeline distribution is unknown.

2 Methods

2.1 Satellite Imagery Acquisition and Treeline Community Composition

We purchased WV-3 panchromatic and multispectral imagery collected in July 2020 from Maxar, covering two treeline study areas in Rocky Mountain National Park (RMNP), Colorado, USA (Figure 1). RMNP includes a broad geographic area of treeline with many trails allowing for reasonable treeline access. The imagery was collected on July 21, 2020, with 0% cloud cover and an off-nadir angle of 16.8 degrees. WV-3 data include a panchromatic (black-and-white) band with 31 cm spatial resolution and eight multispectral bands with 1.24 m resolution: coastal blue (400-450 nm), blue (450-510nm), green (510-580 nm), yellow (585-625 nm), red (630-690 nm), red edge (705-745 nm), near-infrared 1 (N-IR1, 770-895 nm), and

https://doi.org/10.5194/egusphere-2025-970
Preprint. Discussion started: 28 May 2025

130 near-infrared 2 (N-IR2, 860-1040 nm). The panchromatic band pools spectral information from across the visible and near-infrared regions (450-800 nm) to yield a black-and-white image with higher spatial resolution than the multispectral bands.

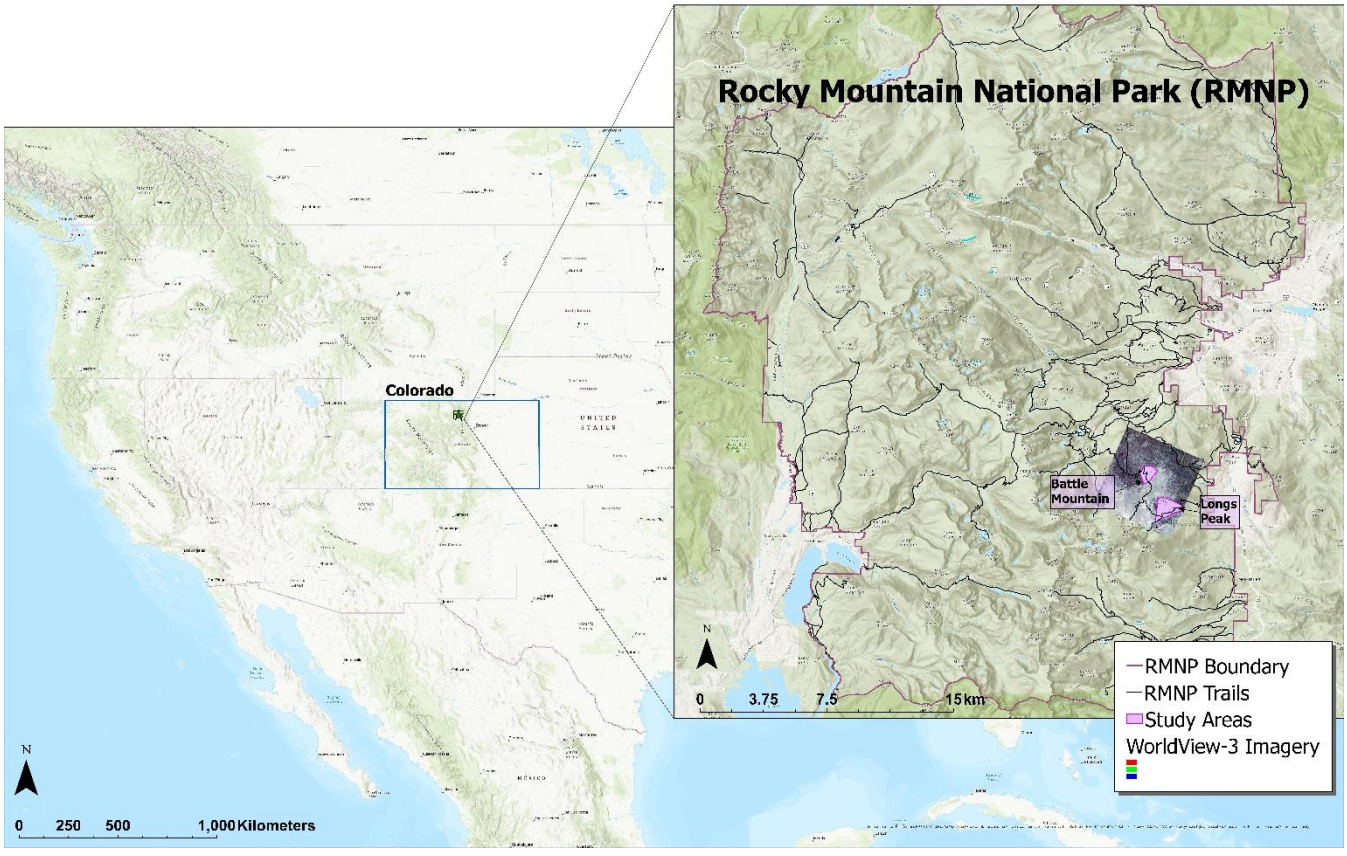


Figure 1. Locations of study areas and WV-3 imagery within RMNP and Colorado. The study areas are purple polygons visible over the WV-3 imagery extent, displayed in RGB (Red = red band, Green = green band, Blue = blue band). Basemap sources: Esri, 135 **TomTom, Garmin, FAO, NOAA, USGS © OpenStreetMap contributors, and the GIS User Community; topographic basemap sources: Esri, Airbus DS, USGS, NGA, NASA, CGIAR, N Robinson, NCEAS, NLS, OS, NMA, Geodatastyrelsen, Rijkswaterstaat, GSA, Geoland, FEMA, Intermap and the GIS user community.**

Limber pine is a dominant conifer at treeline in both study areas (Sindewald and Tomback and Neumeyer, 2020). The Longs Peak treeline study area communities include dense willow (*Salix glauca, Salix brachycarpa,* and hybrids), Engelmann 140 spruce *(Picea engelmannii),* subalpine fir *(Abies lasiocarpa),* glandular birch *(Betula glandulosa),* and quaking aspen *(Populus tremuloides).* The Battle Mountain treeline study area communities are predominately composed of limber pine with Engelmann spruce, subalpine fir, willow, and glandular birch as minor components.

2.2 Orthorectification and Atmospheric Correction

We collected ground control points (GCPs) at trail junctions and switchbacks, using a Trimble Geo7x Centimeter edition 145 geolocator with a Zephyr 2 antenna mounted on a 1-m carbon fiber pole. GCPs are high-accuracy (5-10 cm) positions of

https://doi.org/10.5194/egusphere-2025-970
Preprint. Discussion started: 28 May 2025

select landscape features visible in satellite imagery. We documented each GCP collection with photos from several angles, indicating the precise location of the point (see Figure A1), as well as image chips (see Figure A2) with the position marked from Google Earth Pro (Figure A2).

Maxar performed rigorous orthorectification to correct for distortions due to terrain; they did this using the nearest neighbor resampling method (preserving the original data values) with the GCP positions and documentation we provided. Cubic convolution resampling is commonly used because it results in a smoother image, but it alters the data values, effectively introducing noise into the data. "Pansharpening" is a similarly inappropriate technique for any analysis that relies on data precision. Maxar also applied the atmospheric compensation (ACOMP) correction to the imagery, which uses cloud, aerosol, water vapor, ice, and snow (CAVIS) band data collected at the same instance as the multispectral data to identify and correct for atmospheric influences (Pacifici, 2016).

2.3 Species Polygon Collections

In both study areas combined, we delineated 615 polygons of contiguous, single-species tree/shrub patches in the field that included a total of 165 limber pine, 129 Engelmann spruce, 84 subalpine fir, 71 willow, 141 glandular birch, and 25 aspen (a minor component of the treeline community) (Table 1). A large sample size is necessary to capture the intraspecies variation at treeline in plant condition caused by differences in frost desiccation, wind damage, or water availability, which influences the near-infrared wavelengths in particular (Curran, 1989; Campbell and Wynne, 2011). We walked the perimeter of each polygon with a Trimble GeoXT or a Geo7x, using differential correction enabled to obtain achieve sub-meter accuracy (typically 10-60 cm at treeline). A polygon may contain one or more individuals of a given species; apart from limber pine, all these species may spread clonally at treeline, making discrimination of individual trees or shrubs impossible without a genetic analysis. Limber pine has multi-stemmed growth forms at treeline, but multiple stems may comprise different individuals originating from a single Clark's nutcracker *(Nucifraga columbiana)* cache of limber pine seeds, presenting the same difficulties (Tomback and Linhart, 1990; Linhart and Tomback, 1985).

We imported the polygon data to ENVI (version 4.8, Exelis Visual Information Solutions, Boulder, CO) and selected WV-3 pixels that fell entirely within the bounds of each of the polygons. We examined both the panchromatic and multispectral imagery to identify systematic offsets between the polygons and vegetation patches in the imagery. The images from Maxar were orthorectified together; the pixels aligned across bands and sensors—4 x 4 panchromatic pixels align precisely with one multispectral pixel. We then exported the reflectance data from each region of interest as CSV files.

Table 1. Class frequencies (proportion of total pixels) for the CNN models. Frequencies are out of 615 polygons and 5,631 pixels, respectively. CNNs use a pixel-based classification approach (sample unit = the pixel, not the plant).

Species	Frequency (polygons)	Frequency (pixels)

https://doi.org/10.5194/egusphere-2025-970
Preprint. Discussion started: 28 May 2025

Subalpine fir—*Abies lasiocarpa* (ABLA)	0.137 (84)	0.280 (1,577)
Limber pine—*Pinus flexilis* (PIFL)	0.268 (165)	0.197 (1,109)
Engelmann spruce—*Picea engelmannii* (PIEN)	0.210 (129)	0.172 (969)
Willow—*Salix* sp. (*Salix*)	0.115 (71)	0.135 (760)
Quaking aspen—*Populus tremuloides* (POTR)	0.041 (25)	0.028 (158)
	Total polygons = 615	Total pixels = 5,631

2.4 Topographic Data

We obtained a 10 m digital elevation model (DEM) from the U.S. Geological Survey EROS Data Center. These raster data are generated by the National Mapping Program from cartographic information and are freely available from the National Map Data Delivery website (https://www.usgs.gov/the-national-map-data-delivery). We interpolated the DEM to the resolution of the multispectral data, using a cubic spline resampling method.

2.5 Convolutional Neural Network (CNN) Modelling Methods

In this section, we describe the methods we used to train the CNN models. Section 2.5.1 contains an overview of the CNN model architecture, including an adaptation to allow training on the higher resolution panchromatic imagery before concatenating those data with the multispectral imagery and DEM. Section 2.5.2 provides an overview of the hyperparameter experiment, used to determine the optimal combinations of fixed hyperparameters for the best model performance. In ML models, parameters are the trainable weights and biases within the models, and hyperparameters are values the user pre-defines that guide the training process and do not change during training. Section 2.5.3 describes the explainable artificial intelligence (XAI) methods used to determine which predictors were most important for model performance.

CNNs are a form of neural network that can be spatially aware, detecting patterns in gridded data on multiple spatial scales, and are often used for image recognition tasks (Goodfellow and Bengio and Courville, 2016d; Dubey and Jain, 2019). A CNN is structured as a series of convolutional blocks, each containing one or more convolutional layers and ending with a pooling layer (Dubey and Jain, 2019). Each convolutional layer contains many convolutional filters, or "kernels," containing learned parameters (weights and biases). During convolution, the input channels (or data layers, e.g., the multispectral WV-3 data contain eight channels) are multiplied by a 3-D convolutional filter, typically with dimensions of 3 pixels x 3 pixels x number of input channels. The products of this multiplication are summed to create one pixel in the output feature map (Goodfellow et al., 2016a). The filter slides over the spatial grid of the input imagery; at each position in this grid, it generates one pixel of the output feature map. To produce the desired number of feature maps (M), a convolutional layer contains M filters. Classic image-processing also uses convolutional filters, but the weights are pre-determined. For example, there are known, 3 x 3-pixel kernels that achieve blurring, sharpening, edge detection, etc. A strength of CNNs is that these

https://doi.org/10.5194/egusphere-2025-970
Preprint. Discussion started: 28 May 2025

weights are learned rather than pre-determined. The CNN uses multiple different filters to learn different patterns in the data (detecting edges, textures, parts of objects, whole objects, etc.), which generate multiple feature maps that are sent to the next convolutional layer. (The number of feature maps generated is a fixed hyperparameter set by the user.) The resulting feature maps (matrices containing the results of the convolutions) can be thought of as patterns found across all data channels. CNNs may be thought of as "a feature-detector (the convolution and pooling layers) attached to a traditional neural network," which learns from these detected features and transforms them into predictions (Lagerquist, 2020).

Parameters in a CNN consist of weights and biases in the convolutional filters. A single convolutional filter contains one bias and $K_h * K_w * C_{in}$ weights, where K_h is the height of the filter in pixels; K_w is the width in pixels; and C_{in} is the number of input channels. Thus, a single convolutional layer contains C_{out} biases and $K_h * K_w * C_{in} * C_{out}$ weights, where C_{out} is the number of filters or output feature maps. The model is trained in a series of epochs; in each epoch, the CNN learns from a series of many batches of training samples. In the first epoch, all the weights and biases start from a random initial seed. The first batch of data is input to the CNN, and the loss function is calculated. The loss function is an error metric used to "tell" the ML model how right or wrong it is in its predictions—a feedback mechanism. After the model learns from the training samples in an epoch, the weights and biases are adjusted through backpropagation (explained in detail in S1.1), using rules of gradient descent to minimize the loss function (the measure of model error).

Another batch of data is fed to the CNN, the loss function is calculated, and backpropagation is repeated. This process repeats for B training batches each containing N data samples, where B and N (both positive integers) are hyperparameters. At the end of the epoch, the validation loss is computed, using data in the validation fold. This in turn repeats for P epochs or stops early once the loss function for the validation fold (also known as out-of-bag error) has not decreased in the last Q epochs, where P and Q (both positive integers) are hyperparameters. The process of training a machine learning (ML) model with small batches (minibatches) of randomly selected training data is known as stochastic gradient descent, and is the most common algorithm used for training by contemporary ML developers to create deep learning models (Goodfellow and Bengio and Courville, 2016c, b; Li et al., 2014).

Most of the methods we use are standard in the ML literature, but we recognize our audience may include treeline ecologists for whom these methods are new. We provided further explanations of CNN methods—and their importance—for interested readers in Supplement S1, including loss functions (S1.1 and Appendix B), backpropagation (S1.1), ReLU activation (S1.2), batch normalization (S1.2), data augmentation (S1.3), and dropout (S1.3).

2.5.1 CNN Model Architecture

We used all eight multispectral WV-3 bands, the panchromatic band, and the interpolated DEM as the CNN inputs, yielding a total of 10 channels. A CNN model is a pixel-based classification approach, with a variable number of pixels within each

https://doi.org/10.5194/egusphere-2025-970
Preprint. Discussion started: 28 May 2025

tree or shrub polygon. Before modelling, the data were subset into 160 x 160 m image chips, each centered on one training pixel—one point within one tree or shrub polygon. This initial sub-setting process streamlined the dataset, reducing the RAM and time required to run the model.

The pixel dimensions of each image chip varied by the spatial resolution of the data: 513 x 513 pixels for the panchromatic imagery (31 cm resolution), and 129 x 129 pixels for the multispectral imagery and the interpolated DEM (1.24 m resolution). We adapted the model architecture to accommodate this discrepancy, to leverage the information in the higher-resolution panchromatic imagery, while ensuring that all 10 data channels could be processed together within the CNN. The first two convolutional blocks in the CNN processed only panchromatic imagery (at 31 cm resolution). Each convolutional block consisted of two convolutional layers (with a 3-by-3 convolutional filter), followed by 2-by-2 pooling, which reduces the spatial resolution by half. Hence, in two convolutional blocks, the 513 x 513 panchromatic data were downsampled to 257 x 257 (62 cm resolution) and then to 129 x 129 (1.24 m resolution), matching the resolution of the multispectral and interpolated DEM data (Figure 2). Max pooling helps with edge detection and allows the model to learn patterns on larger spatial scales. Using our classification problem as an example, the CNN could first learn size and shape patterns for species morphologies, associated with the spectral data, and, after pooling, could learn that certain species are found near other species or landscape features at coarser scales (such as rivers, or places of low or high prominence in the DEM). For a more detailed description of each convolutional block, please see Appendix C.

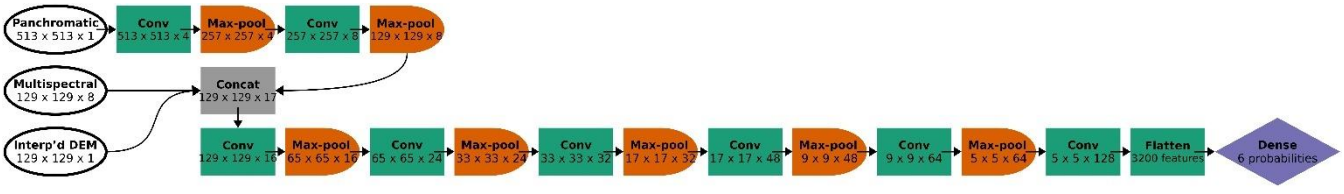


Figure 2. CNN architecture for the six-class model, incorporating panchromatic, multispectral, and DEM inputs at different spatial resolutions. Each convolutional block is represented by a green box labeled "Conv" (actually two series of convolution and activation, followed by batch normalization) and an orange half-oval labeled "Max-pool". After two convolutional blocks with only the panchromatic data, the multispectral and interpolated DEM data were concatenated with the feature maps from the convolutional blocks ("Concat"). At the end, the feature maps (64 of them, each with 5 x 5 pixels) were flattened (all of the values in all of the feature maps are appended to a 1-D vector, 3200 values long) and followed by one classic NN dense layer (or "fully connected layer"), yielding a vector of six probabilities, one for each class. At each step, the dimensions of the image patch are indicated in pixels, as well as the number of feature maps. For example, the first Conv step yielded four feature maps of 513 x 513 pixels.

After two convolutional blocks, once the panchromatic data were pooled down to the resolution of the multispectral imagery and the interpolated DEM, those channels were concatenated ("Concat" in Figure 2). Seventeen channels (one raw DEM channel, eight raw multispectral channels, and eight feature maps based on the one panchromatic channel) were fed into the next convolutional block. Then, after five more convolutional blocks (each halving the spatial resolution while increasing the number of feature maps), the last feature maps were flattened (re-organized into a 1-D vector) and fed into a dense layer, or

https://doi.org/10.5194/egusphere-2025-970
Preprint. Discussion started: 28 May 2025

"fully connected layer". A traditional (spatially agnostic) NN contains only dense layers, which are vectors of values connected to every adjacent value, including all the values in the next (and previous) dense layer(s).

Following the dense layer, we used the softmax activation function to force outputs to range from 0-1 with a sum of 1 (interpretable as probabilities representing all possible events/classes in the sample space) (Goodfellow and Bengio and Courville, 2016d). The non-terminal dense layers included a dropout rate of 0.650, and every convolutional and dense layer used an L_2 regularization strength of $10^{-6.5}$. Dropout and L_2 regularization are both regularization methods that help prevent the CNN from overfitting to the training data (see S1.3 for details). These were fixed hyperparameters, determined to be the best settings via the hyperparameter experiment described in section 2.5.2.

The total number of training samples was 5,631. We validated the models using five-fold cross-validation, so for each model, we used four-fifths of our data for training and the remaining fifth of the data for validation. We split the training and validation by individual—for a given organism, either 1) all of the pixels for that organism were in the training set or 2) all of the pixels were in the validation set. After creating the folds, we performed data augmentation (Goodfellow and Bengio and Courville, 2016a), turning each original training sample into eight augmented samples. To do this, we first normalized the data, transforming the data from physical values (x) to z-scores (z) based on the mean and standard deviation of each training fold. Data should be separated into training and validation folds *before* normalization, because otherwise we would be leaking information about the full data distribution from the validation data into the training data. To create an augmented data sample, we add Gaussian noise to the predictors in the original sample, with a mean of 0.0 and standard deviation of 0.2. (The parameters of the Gaussian distribution from which we drew the noise were fixed hyperparameters.) Thus, the final sample size for the training dataset was 45,048 image chips (5,631 x 8 augmented samples).

During each epoch, the model was trained with all 45,048 data samples in batches of 64 samples each (703 total batches). Each model was trained for 100 epochs, with a command to stop training early if the loss function did not improve for 15 epochs. We trained all CNN models with the Adam optimizer, using all the Keras defaults, including an initial learning rate of 0.001. The Adam optimizer updates the learning rate after every epoch and adjusts learning rates separately for every model parameter (every weight or bias) (Goodfellow and Bengio and Courville, 2016b; Kingma and Ba, 2014). We determined the optimal loss function for our classification problem through the hyperparameter experiments (described in the next section, 2.5.2.).

We trained three CNN models, each with a different number of classes. Classification becomes more difficult with each additional class. We pooled classes with fewer data to see whether that would improve model performance, particularly for distinguishing limber pine from the other species with higher accuracy. In addition to the six-class model, we trained a four-class model to separate the three conifers (limber pine, Engelmann spruce, and subalpine fir) from each other and from the deciduous plants (glandular birch, willow, and aspen), the latter grouped together in one class as "Other". Lastly, because

https://doi.org/10.5194/egusphere-2025-970
Preprint. Discussion started: 28 May 2025

managers may find it useful to discriminate limber pine from other treeline species, we trained a two-class model with limber pine and the other species grouped as “Other”. The CNN architecture for the four- and two-class models was the same as the six-class model, except for the output layer. The output layer produces one probability per class, so its output vector varied in length from two to six (Figures C1 and C2).

2.5.2 Hyperparameter Experiments and Model Selection

We ran hyperparameter experiments separately for the six-, four-, and two-class models (see Supplement S1.5 for detailed information on what each hyperparameter does within a CNN). Hyperparameters have a strong influence on model performance, so it was important to test a subset of hyperparameter combinations to identify the best versions of the six-, four-, and two-class models. The same four experimental hyperparameters were used for all three experiments: the number of dense layers at the end of the model, the dropout rate (used for non-terminal dense layers, i.e., all dense layers except the output layer that provides the final probabilistic predictions), the L_2 regularization strength (used for all convolutional and dense layers), and the loss function. Table 2 summarizes the values tested for each hyperparameter for each of the three models. The candidate values tested in the experiment for each hyperparameter were chosen for their usefulness in training skillful CNN models in past work (Lagerquist and McGovern and Gagne Ii, 2019; Lagerquist et al., 2020; Lagerquist et al., 2021).

We tested the models with both the Gerrity score and cross-entropy loss functions. Cross-entropy is widely used for classification problems; it comes from the field of information theory and describes the bits required to distinguish two distributions (i.e., the distribution of predictions from the distribution of observations) (Lagerquist, 2020). The class frequencies in our dataset were quite unbalanced (Table 1), so we also tested the Gerrity score, which rewards "risky" predictions. That is, in a problem with unbalanced classes, the Gerrity score rewards correct predictions of a lower-frequency class (which are harder) more strongly than correct predictions of a higher-frequency class (which are easier) (Gerrity, 1992). Equations for these metrics can be found in Appendix B and in other published works (Lagerquist and McGovern and Gagne Ii, 2019; Lagerquist, 2020).

The Gerrity score depends on how classes are ordered numerically. For example, in a six-class problem, the Gerrity score rewards correct predictions of "class 1" more strongly than correct predictions of "class 6," even if both classes have equal frequency. Thus, by default, we ordered classes from least to most frequent (using the pixel-based class frequencies in Table 1). We also tested the Gerrity score with two modifications. One was the class-weighted Gerrity score, where each data sample in the loss function was weighted by $\ln\frac{1}{f}$ or $\ln 50$, whichever was lower, where f is the frequency of the correct class. Class-weighting makes the Gerrity score reward risky predictions even more. The second modification was the limber-pine-first (PIFL-first) Gerrity score, where the aforementioned list was reordered to make limber pine "class 1". The PIFL-

https://doi.org/10.5194/egusphere-2025-970
Preprint. Discussion started: 28 May 2025

first Gerrity score gives higher rewards for correct predictions of limber pine. The Gerrity score varies between -1 and 1. A higher score is better, and as long as it is above 0, the model is performing better than a random model (Lagerquist and McGovern and Gagne Ii, 2019).

Table 2. Experimental hyperparameter values tested for the six-, four-, and two-class models.

Hyperparameter	Six-class model	Four-class model	Two-class model
Number of dense layers	1,2,3,4	1,2,3,4	1,2,3,4
Dropout rate	0.575, 0.650	0.650	0.650
L_2 regularization strength	$10^{-6.5}$, 10^{-6}	$10^{-6.5}$	$10^{-6.5}$
Loss function (all six loss functions were tried for all three models)	(1) Gerrity score with class-weighting and PIFL first, (2) Gerrity score with no class-weighting and PIFL first, (3) Gerrity score with class weighting and default order, (4) Gerrity score with no class weighting and default order, (5) cross-entropy, and (6) class-weighted cross-entropy		

In the six-class experiment, the total number of hyperparameter combinations was 96, representing all possible combinations of four numbers of dense layers, two dropout rates, two L_2 regularization strengths, and six functions (Table 2). For every hyperparameter combination (hyperparameter model) we performed five-fold cross-validation, thus training five sub-models; we refer to each set of five cross-validated sub-models as one model (Figure 3). This yielded a total of 480 sub-models for the six-class experiment (96 x 5) and 120 sub-models for each of the smaller experiments (24 x 5). To evaluate the performance of each hyperparameter model, we used predictions only on out-of-bag samples. For each model, every data sample was out-of-bag, appearing in the validation fold rather than one of the training folds only once. Given that our dataset contained 5,631 samples, the results for each hyperparameter model were therefore based on 5,631 out-of-bag predictions.

https://doi.org/10.5194/egusphere-2025-970
Preprint. Discussion started: 28 May 2025

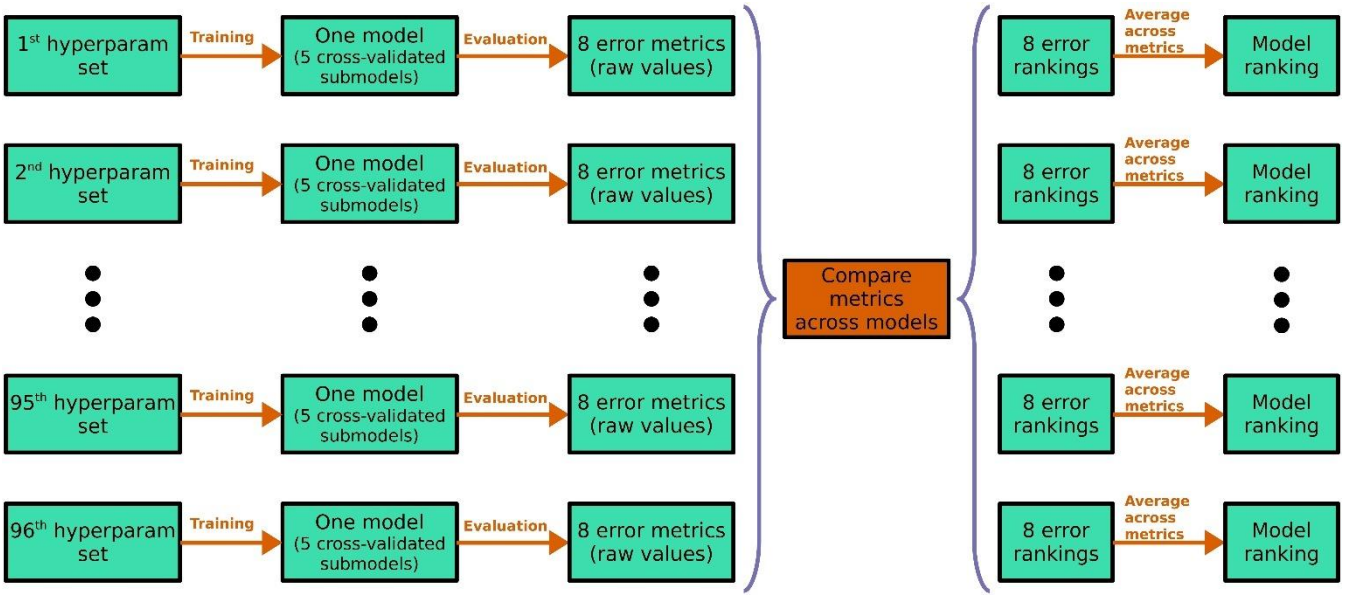


Figure 3. Schematic for the six-class hyperparameter experiment. The two- and four-class experiments follow the same methodology, except with 24 hyperparameter sets (hence, 24 models) instead of 96. Each row corresponds to one model; the black ellipses represent the 3rd through 95th models; each green box represents an object (hyperparameter set, model, set of error metrics or rankings); and each orange arrow/box represents a procedure. The purple braces indicate that models are being compared with each other—*i.e.*, they indicate a procedure that cannot be done independently for each model. The last procedure—implied but not shown—is choosing the model with the best "Model ranking," averaged over all eight error metrics.

We considered eight evaluation metrics: 1) top-1 accuracy, 2) top-2 accuracy, 3) top-3 accuracy, 4) cross-entropy, 5) Heidke score, 6) Peirce score, 7) Gerrity score, and 8) PIFL-first Gerrity score. Top-k accuracy is the fraction of data samples for which the correct class is in the k highest probabilities output by the model; for example, top-2 accuracy is the fraction of data samples for which the correct class is one of the two classes predicted with the highest probability. Top-1 accuracy is usually just called "accuracy," i.e., the fraction of samples for which the correct class receives the highest probability from the model. Further discussion of all metrics, including mathematical and conceptual definitions, can be found in Appendix B.

We ranked each model in terms of each evaluation metric out of the total number of models attempted in the experiment. For the six-class experiment, these rankings were out of 96; for the smaller experiments, these rankings were out of 24. Then, for each model, we averaged its rankings over all eight metrics. The model with the best average ranking was chosen as the best model.

2.5.3 Permutation Tests

Permutation tests are a form of XAI (McGovern et al., 2019), methods that enable interpretation of ML model results in terms of the predictors used. The permutation test comes in four varieties: the single-pass forward, multi-pass forward, single-pass backward, and multi-pass backward. The permutation tests successively permute (shuffle) or de-permute (clean) the values of predictor variables and quantifying the impact that has on model performance. For data with correlated

https://doi.org/10.5194/egusphere-2025-970
Preprint. Discussion started: 28 May 2025

predictors, each variety of the test gives different results (McGovern et al., 2019). Our image data contain both spatial autocorrelation, where nearby pixels are correlated with each other, and spectral correlation, where nearby wavelength bands in the multispectral imagery are correlated with each other. These problems always arise in image data, meaning that methods assuming mutual independence cannot be used. It is important to examine all four varieties of the test and determine which results are consistent to draw robust conclusions.

In the permutation test, permuting a predictor variable refers to shuffling the values of that variable across all data samples (in our case, across all out-of-bag samples), breaking the relationship between the predictor and the target variable (in our case, the species class). In the single-pass forward test, only one predictor variable is permuted at a time, leaving the other variables unchanged. After permuting one predictor variable, the model performance on the clean dataset is compared to performance on the dataset with that variable permuted. If model performance drops significantly with the predictor permuted, then that predictor is considered very important. In the multi-pass forward test, the single-pass forward test is carried out iteratively. Step 1 begins with a clean dataset and determines the most important variable, x_{1st}, which is then permuted forever. Step 2 begins with the output from step 1 (a dataset with x_{1st} permuted and all other variables unchanged) and determines the second-most important variable, x_{2nd}, which is then permuted forever. This continues until all predictor variables are permuted. In the single-pass backward test, we begin with a completely randomized dataset, where all predictor variables are permuted. Then only one predictor variable is cleaned up (restored to the correct order) at a time, leaving the other variables permuted. After cleaning up one predictor variable, we measure how much this improves model performance compared to the completely randomized dataset. If the model performance improves significantly when the predictor is cleaned up, then that predictor is considered very important. In the multi-pass backward test, the single-pass backward test is carried out iteratively. Step 1 begins with a completely randomized dataset and determines the most important variable, x_{1st}, which is then cleaned up forever. Step 2 begins with the output from step 1 (a dataset with x^*_{1st} cleaned up and all other variables still permuted) and determines the second-most important variable, x_{2nd}, which is then cleaned up forever. This continues until all the predictor variables have been cleaned.

3 Results

3.1 Hyperparameter Experiment Results

The hyperparameters for each of the selected models (six-class, four-class, and two-class) are shown in Table 3. Figure 4 shows the top-1 accuracy for all models in the six-class experiment. The selected model (circle) had accuracies of 0.44, 0.70, and 0.83 (top-1, top-2, and top-3, respectively), whereas the model with the highest top-1 accuracy (star) had accuracies of 0.45, 0.67, and 0.81. The selected model performed much better on the Gerrity score (0.38 vs. 0.30), comparably on the Heidke and Peirce scores (0.31 and 0.32 vs. 0.32 and 0.32), and only marginally worse on the PIFL-first Gerrity score (0.45 vs. 0.49). The selected model was the best model out of 96 models based on the average of all eight metrics.

https://doi.org/10.5194/egusphere-2025-970
Preprint. Discussion started: 28 May 2025

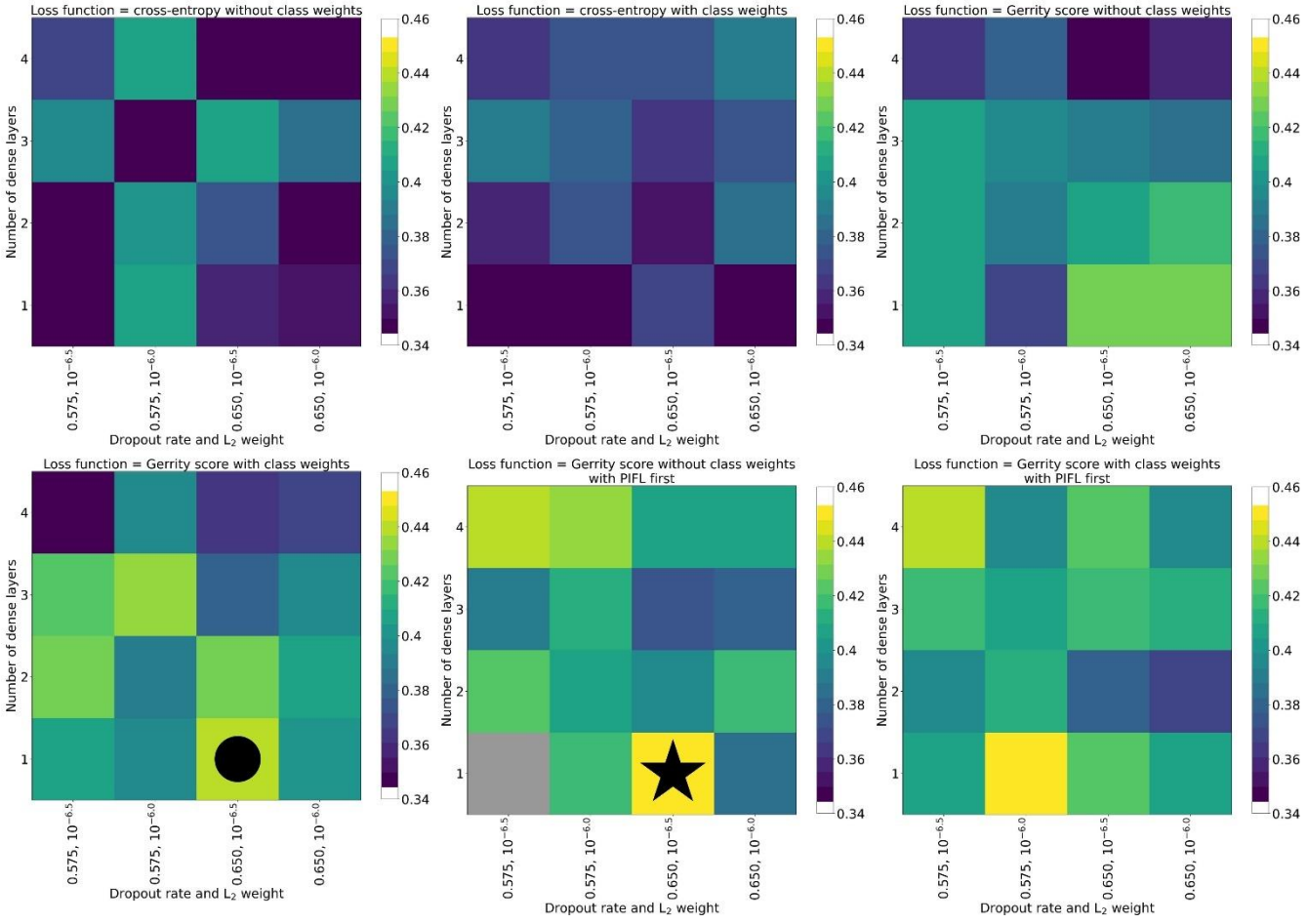


Figure 4. Hyperparameter experiment results for the six-class classification, evaluated by top-1 accuracy. Each grid cell of the figure shows results from a single CNN model (one set of hyperparameters), including results from all five sub-models. In the case of every data sample, the prediction came from a CNN that was not trained on that data sample. The color indicates the top-1 accuracy of the model. The models are organized by their hyperparameter values. The circle indicates the model that was selected from the set. The star indicates the model with the highest top-1 accuracy.

The selected four-class model was the best out of 24 evaluated, with the second-highest ranking for each top-k accuracy, as well as high rankings for the remaining metrics (Table 3). The selected two-class model was again the best balance out of 24 evaluated; it was the top-performing model based on the Gerrity, PIFL-weighted Gerrity, Heidke, and Peirce scores, and was the 3rd best model based on top-1 accuracy (Table 3). Strangely, across the board, models that performed well based on the cross-entropy metric performed poorly based on the other 7 metrics. See Supplement S2 for the remaining hyperparameter experiment result figures (S1-S21).

https://doi.org/10.5194/egusphere-2025-970
Preprint. Discussion started: 28 May 2025

Table 3. Hyperparameters and model performance metrics for the top-performing models. The rank of the selected model with respect to each performance metric is indicated in parentheses. Ranks are out of 96 for the six-class model and out of 24 for the four- and two-class models. All metrics are positively oriented (higher is better) except for cross-entropy, which is negatively oriented.

Hyperparameter	Six-class model	Four-class model	Two-class model
Number of dense layers	1	1	1
Dropout rate	0.65	0.65	0.65
L2 weight	$10^{-6.5}$	$10^{-6.5}$	$10^{-6.5}$
Loss function	Class-weighted Gerrity score	Default Gerrity score	Default Gerrity score
Performance Metric	**Six-class model**	**Four-class model**	**Two-class model**
Top-1 accuracy (rank)	0.441 (3rd)	0.467 (2nd)	0.862 (3rd)
Top-2 accuracy (rank)	0.700 (3rd)	0.767 (2nd)	NA
Top-3 accuracy (rank)	0.831 (4th)	0.922 (2nd)	NA
Gerrity score (rank)	0.379 (4th)	0.254 (3rd)	0.591 (1st)
PIFL-weighted Gerrity score (rank)	0.450 (7th)	0.318 (4th)	0.591 (1st)
Heidke score (rank)	0.314 (3rd)	0.247 (3rd)	0.576 (1st)
Pierce score (rank)	0.320 (2nd)	0.240 (5th)	0.591 (1st)
Cross-entropy (rank)	4.264 (86th)	5.749 (22nd)	2.980 (21st)

3.2 Six-Class Model Results

The six-class model performance was good considering the complexity of the task and the spatial and spectral resolution of the data, with a top-1 (overall) accuracy of 44.1% and a top-2 accuracy of 70.0% (Table 3). By comparison, a trivial model would have a top-1 accuracy of 28.0% and a top-2 accuracy of 47.6%. For a trivial model, the predicted probability of class *k* is always the frequency of class *k* in the data. In other words, a trivial model's predictions are the same for every data sample. For the six-class problem, a trivial model would always predict 28.0% probability of subalpine fir, 18.9% probability of glandular bitch, etc. (Table 3).

Using the class-weighted Gerrity score (Eq. B2) was effective, leading to higher overall model performance and for the correct identification of minority classes, though the model performed best at distinguishing the two highest-frequency classes: subalpine fir and limber pine (Figure 5). Figure 5 shows the classes which the model correctly identified with the greatest frequency, as well as the species it most often tended to confuse.

https://doi.org/10.5194/egusphere-2025-970
Preprint. Discussion started: 28 May 2025

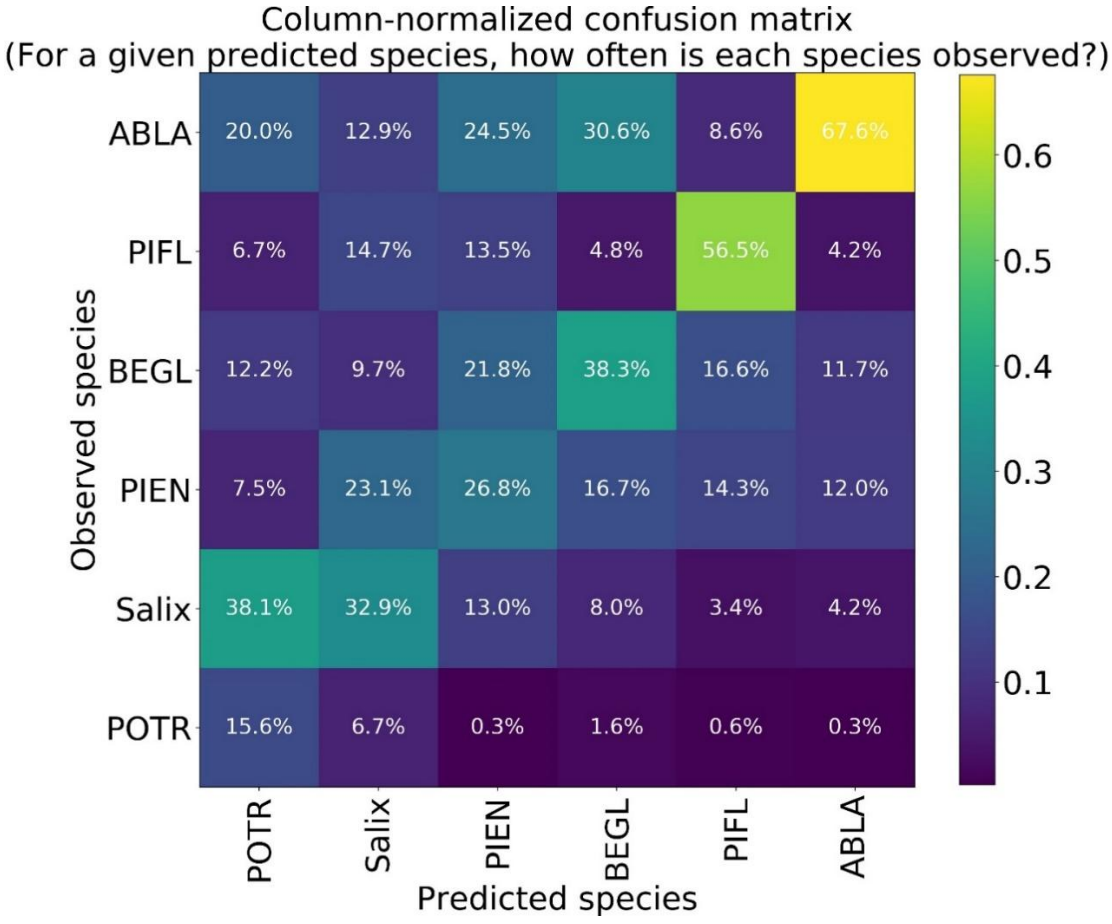


Figure 5. Column-normalized confusion matrix for the six-class model. Classes include subalpine fir (ABLA), glandular birch (BEGL), Engelmann spruce (PIEN), limber pine (PIFL), aspen (POTR), and willow (*Salix*).

Limber pine was most often confused with willow and Engelmann spruce in the six-class model (Figure 5), likely because of the model's reliance on the panchromatic band (see Figures 8b, 8d, and S39). While limber pine does not grow as true krummholz at treeline (Holtmeier, 2009), it is stunted and flagged and resembles willow or other shrubs in the panchromatic imagery. Figure 6 shows an example of a "best hit" identification of limber pine from the six-class model, which correctly predicted limber pine with 100% probability. The example is representative of the dispersed pattern and small size of limber pine at treeline. See Supplement S5.1 for other best hits.

https://doi.org/10.5194/egusphere-2025-970
Preprint. Discussion started: 28 May 2025

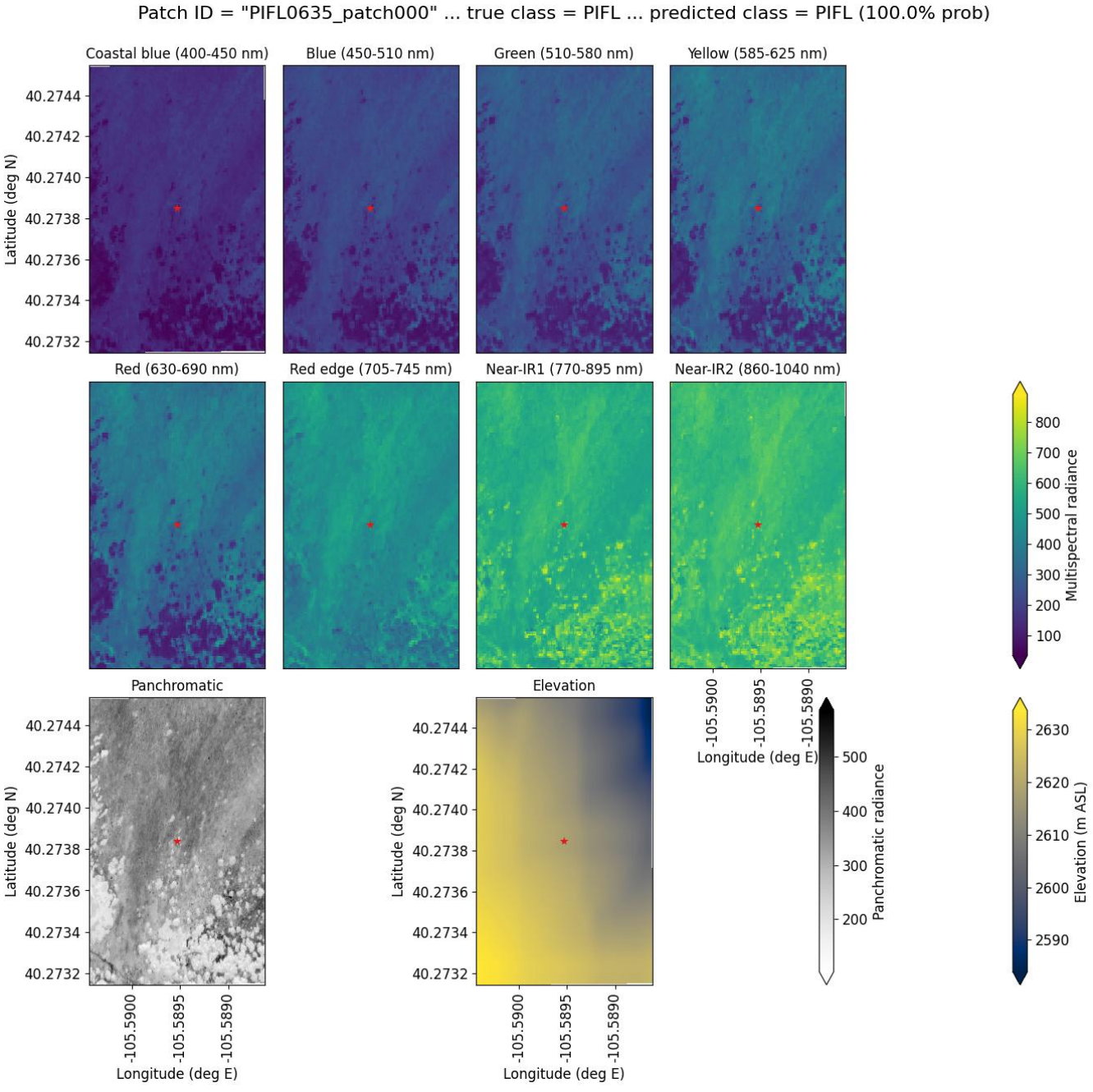


Figure 6. An example of a "best hit" classification of PIFL from the six-class model, where the model correctly predicted PIFL with 100% probability. This example (image chip or patch) is from the Battle Mountain study site. All eight multispectral bands, the panchromatic band, and the DEM are shown. The red star in the center of each image patch is the pixel being classified. Units of radiance are W m^{-2} sr^{-1} µm^{-1}.

https://doi.org/10.5194/egusphere-2025-970
Preprint. Discussion started: 28 May 2025

Subalpine fir was most often confused with glandular birch, once again likely due to reliance on textural information in the panchromatic band (Figures 8b, 8d, and S40 in Supplement S4); both species form large patches on the landscape that appear similar without additional structural information, such as height. Engelmann spruce was the least distinguishable after aspen, which was a lower-frequency class and has a similar growth pattern (in the panchromatic imagery) to both glandular birch and subalpine fir. Engelman spruce and subalpine fir tend to co-occur (with the same elevational distribution) (Sindewald et al., 2020; see also Figures S41, and S42). Their mean spectral radiance curves appear to be statistically distinguishable in the coastal blue, blue, green, yellow, red, and red edge bands (Figure S24), but the CNN relied less on these bands (Figure 8). An example of a worst-case confusion between Engelmann spruce and subalpine fir can be seen in Figure 7. Although the panchromatic band was very important for model performance (Figure 8b and 8d), likely due to its high spatial resolution, it was not enough to distinguish the treeline forms of these species by their morphology or spatial distribution on the landscape.

https://doi.org/10.5194/egusphere-2025-970
Preprint. Discussion started: 28 May 2025

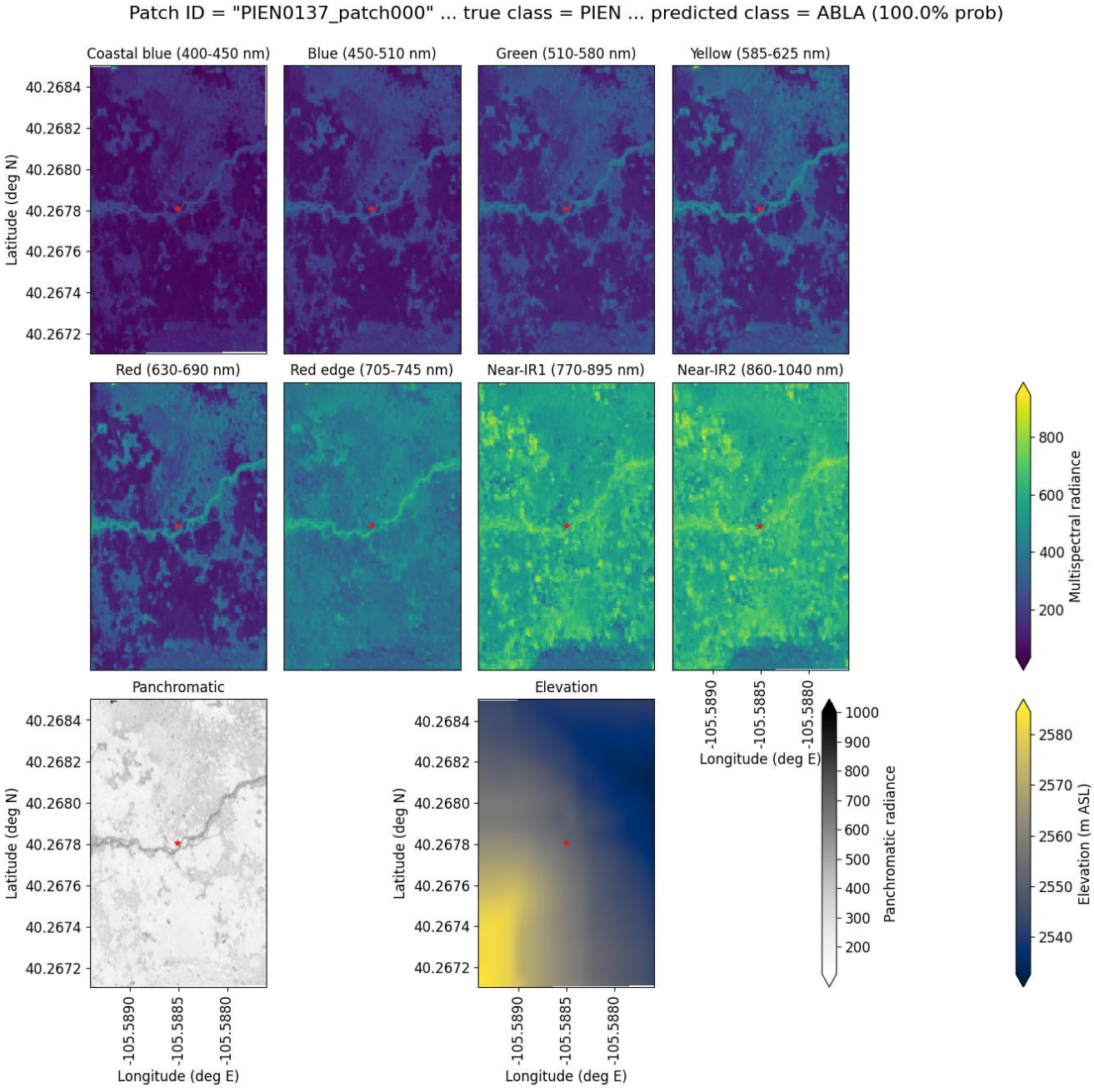


Figure 7. An example of a "worst confusion" classification of PIEN as ABLA from the six-class model, where the model incorrectly predicted PIEN as ABLA with 100% probability. This example (image chip or patch) is from the Longs Peak study site. All eight multispectral bands, the panchromatic band, and the DEM are shown.

Based on the results of the permutation tests, elevation was the most important predictor for model performance (Figure 8a, 8b, and 8d), which makes sense biologically. Willow species thrive in riparian areas, and at the Longs Peak site they are

https://doi.org/10.5194/egusphere-2025-970
Preprint. Discussion started: 28 May 2025

typically found along creeks or in topographic depressions where snow gathers (Figure S44). Similarly, glandular birch, subalpine fir, and Engelmann spruce grow more abundantly in areas where late-lying snowpack provides moisture through the start of the growing season (Figures S40 and S41) (Burns and Honkala, 1990; Hessl and Baker, 1997; Gill and Campbell and Karlinsey, 2015). Limber pine, by contrast, is a drought- and stress-tolerant conifer that often occupies on convex sites or wind-swept ridges where snow is blown clear (McCune, 1988; Ulrich et al., 2023; Steele, 1990).

Several multispectral WV-3 bands were also important for model performance, as evidenced by the multi-pass forward and backward tests (Figure 8b and 8d), including red, green, yellow, coastal, blue, and near-IR2. The visible bands emerged as important predictors for distinguishing aspen from the other species in the six-class model (Figure S43). While the multispectral bands were clearly important, they did have as strong an influence on model performance as did elevation or the panchromatic band, indicating the six-class CNN model did not rely on those data as heavily (Figure 8).

https://doi.org/10.5194/egusphere-2025-970
Preprint. Discussion started: 28 May 2025

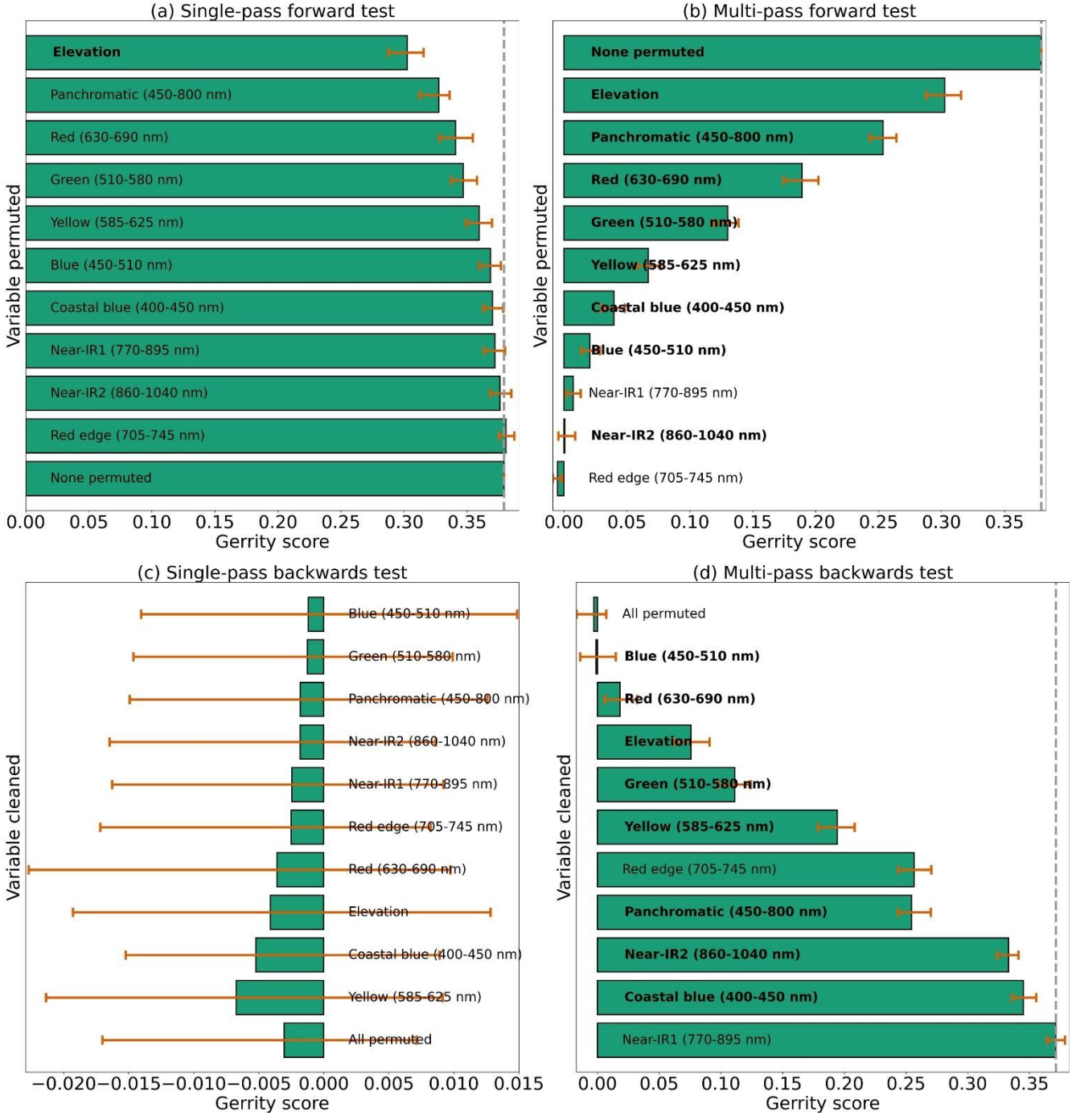


Figure 8: Results from each kind of permutation test to assess predictor importance for the six-class model. Predictors in bold have a significant effect on model performance when permuted, according to a 95% confidence interval over 100 random perturbations of the given predictor. Within each panel, predictor importance decreases from top to bottom, so the most important predictors are at the top. Our evaluation metric for this permutation test is the Gerrity score.

https://doi.org/10.5194/egusphere-2025-970
Preprint. Discussion started: 28 May 2025

3.3 Four-Class Model Results

The four-class model performed slightly better than the six-class model overall, with top-1 and top-2 accuracies of 46.7% and 76.7%, respectively (Table 3). A trivial model could at most have top-1 and top-2 accuracies of 35.1% and 63.1%, respectively. The model was best at correctly identifying subalpine fir, with the predicted species being correct 60.4% of the time (Figure 9). The model did least well at correctly identifying limber pine (39.3%), confusing it with the deciduous species pooled into the "Other" class. Engelmann spruce was the lowest frequency class in the four-class dataset (17.2% of the samples), and since the top model was trained with the default Gerrity loss and not the PIFL-first Gerrity loss, the model was most penalized for incorrect classification of Engelmann spruce. The result was that the four-class model was almost twice as effective at identifying Engelmann spruce as the six-class model (50% vs. 26.8%).

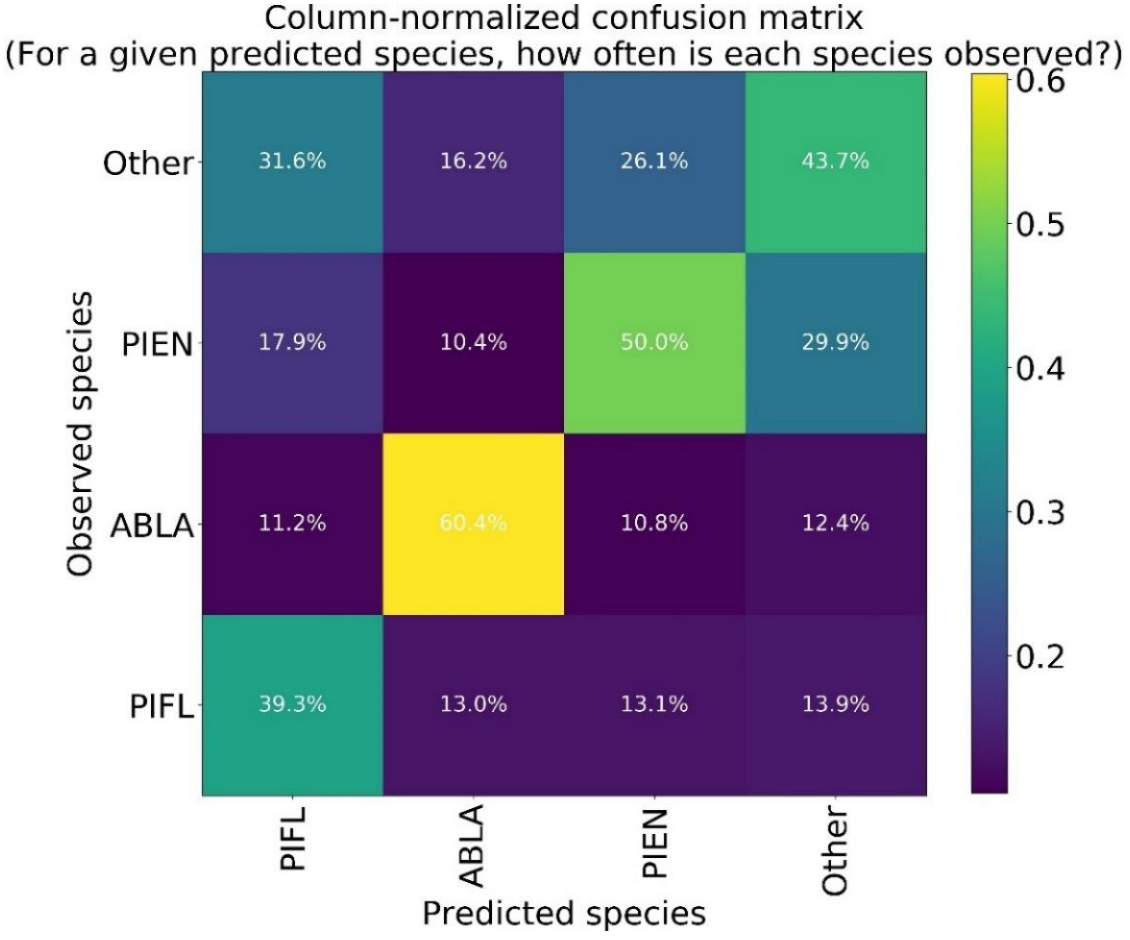


Figure 9: Column-normalized confusion matrix for the four-class model. Classes include subalpine fir (ABLA), Engelmann spruce (PIEN), limber pine (PIFL), and other classes pooled as "Other".

https://doi.org/10.5194/egusphere-2025-970
Preprint. Discussion started: 28 May 2025

The permutation tests were consistent in showing the importance of the panchromatic data to four-class model performance (Figures 10a, 10b, and 10d). As discussed above, the high spatial resolution of the panchromatic data likely allowed the CNN to distinguish growth forms and patterns of species occurrence and co-occurrence on the landscape. The panchromatic band emerged as the most important predictor for all four classes (Figures S45-S48). Multispectral bands red, green, blue, and yellow also emerged as significant predictors (Figures 10b and 10d), particularly for correct identification of ABLA, PIEN, and Other (Figures S45, S46, and S47).

https://doi.org/10.5194/egusphere-2025-970
Preprint. Discussion started: 28 May 2025

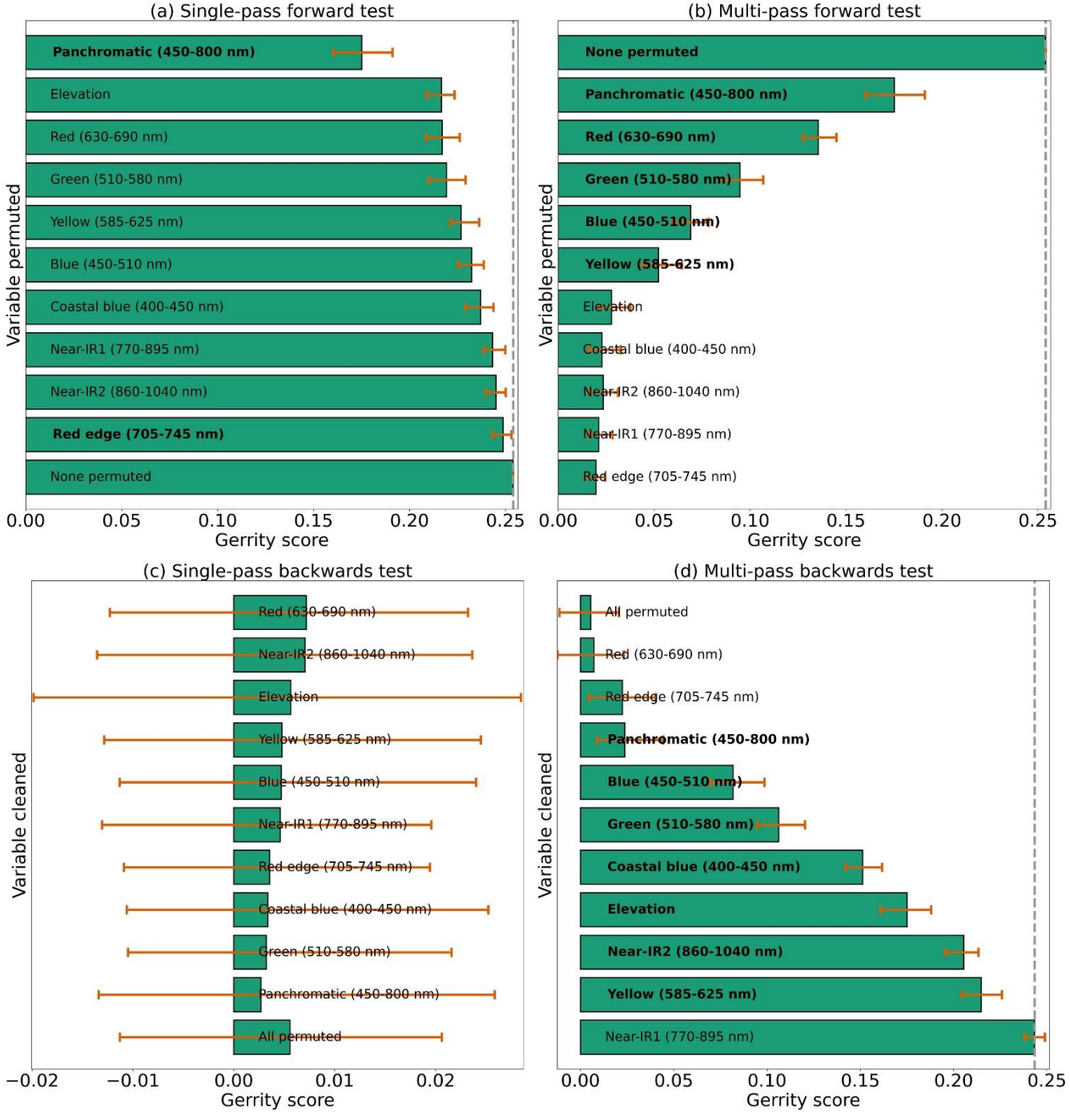


Figure 10: Results from each variety of the permutation test to assess predictor importance for the four-class model. Formatting is explained in the caption of Figure 9.

https://doi.org/10.5194/egusphere-2025-970
Preprint. Discussion started: 28 May 2025

3.4 Two-Class Model Results

The two-class model had the highest top-1 accuracy at 86.2%, though much of this skill is attributable to its correct prediction of the majority class ("Other"); a naïve or trivial model would have a top-1 accuracy of 80.3%, which is the frequency of Other. The model correctly identified limber pine 64.0% of the time (Figure 11), which is still an improvement over the success rates for the six-class (56.5%) and four-class (39.3%) models.

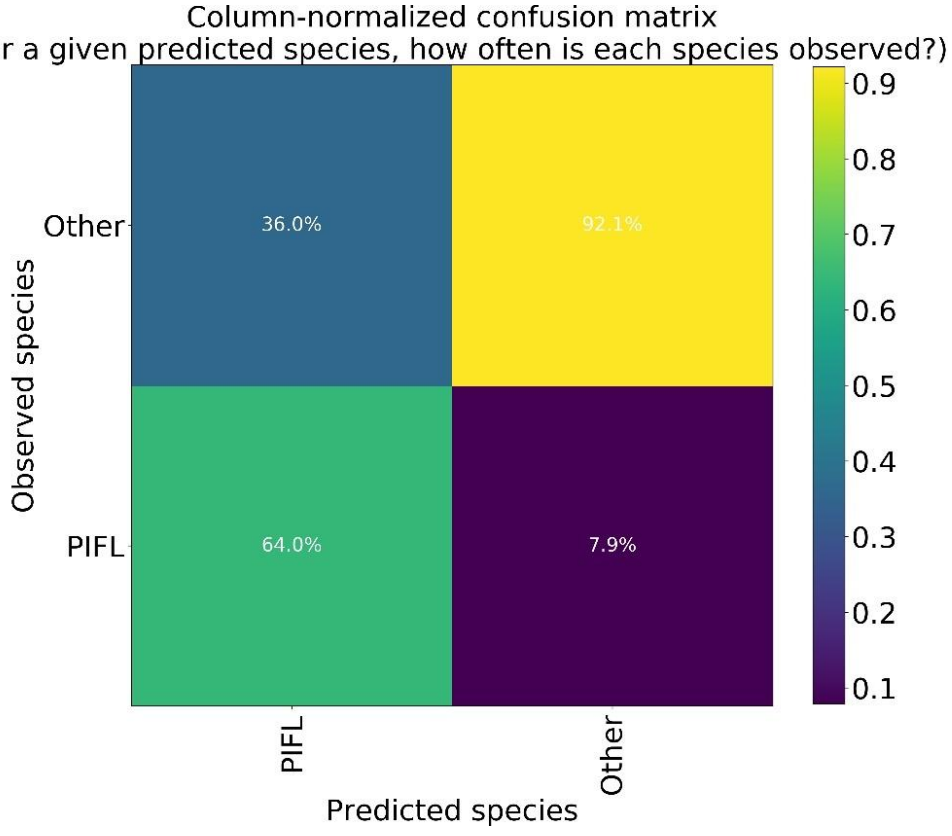
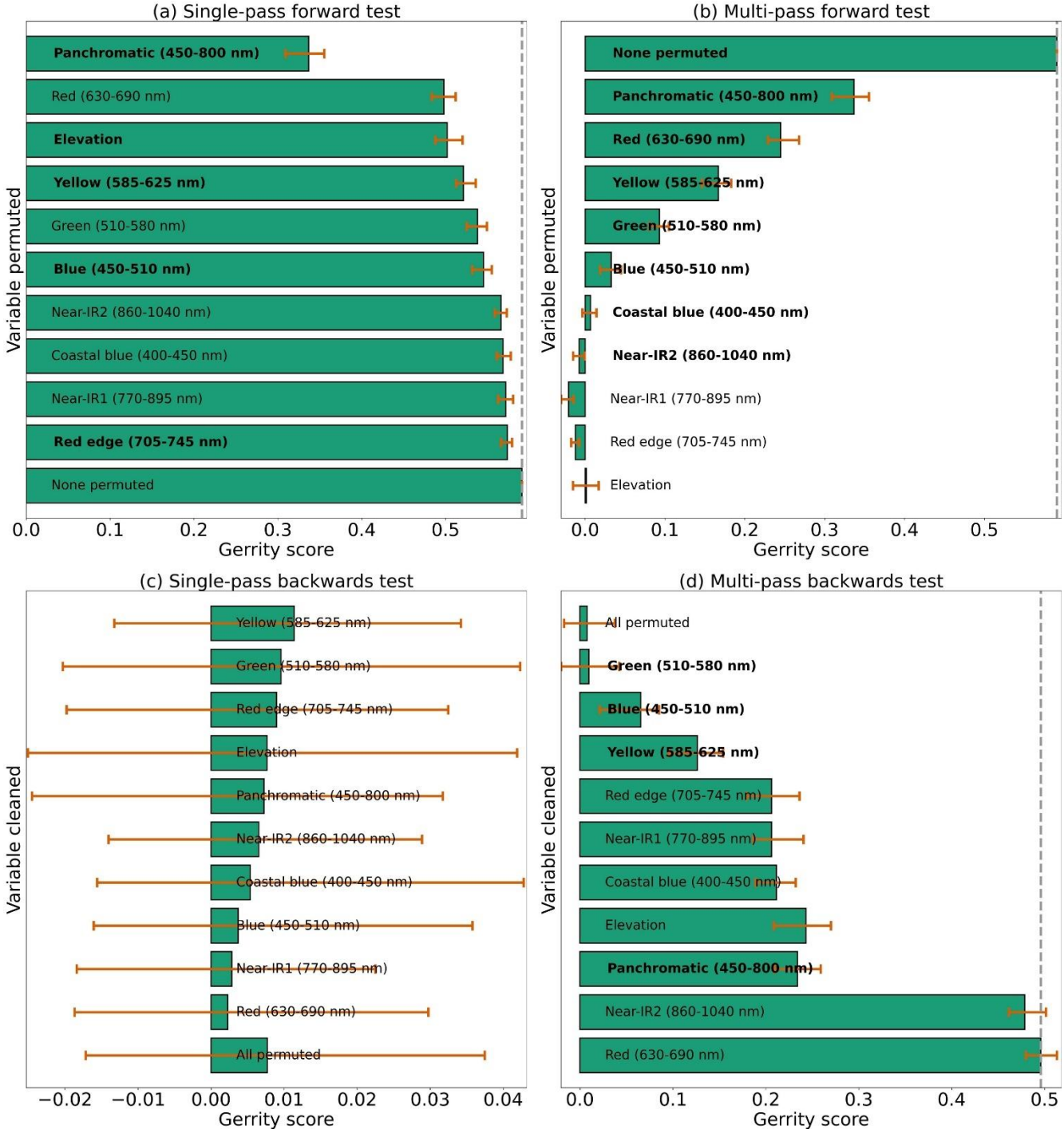


Figure 11: Column-normalized confusion matrix for the two-class model. Classes include limber pine (PIFL) and other classes pooled as "Other".

With the goal of distinguishing limber pine from other treeline species, the panchromatic band emerged as the most important predictor by far—nearly half the model skill depended on the higher spatial resolution panchromatic band (Figure 12a, 12b, and 12d). The red, yellow, green, and blue bands were also important for discriminating limber pine from other species. Limber pine tends to form low-density stands, with individuals spaced almost evenly across the landscape. Unlike subalpine fir and Engelmann spruce, limber pine also does not form large, sprawling krummholz mats or grow in large patches like glandular birch, willow, and aspen. The relatively lower spatial resolution of the multispectral WV-3 bands may

https://doi.org/10.5194/egusphere-2025-970
Preprint. Discussion started: 28 May 2025

have limited the usefulness of these data for distinguishing limber pine, although several bands were still significant predictors in the permutation tests for limber pine (Figure S49).



https://doi.org/10.5194/egusphere-2025-970
Preprint. Discussion started: 28 May 2025

Figure 12: Results from each variety of the permutation test to assess predictor importance for the two-class model. Predictors in bold have a significant effect on model performance when permuted, according to a 95% confidence interval over 100 random perturbations of the given predictor. Within each panel, predictor importance decreases from top to bottom, so the most important predictors are at the top. Our evaluation metric for this permutation test is the Gerrity score.

4 Discussion

This study may be the first attempt to use satellite imagery to identify woody plant species in a treeline ecosystem. Our goal was to discriminate six alpine treeline tree and shrub species in the Southern Rocky Mountains, using a pixel-based Convolutional Neural Network (CNN) classification of high-resolution Worldview-3 (WV-3) satellite imagery. We were particularly interested in identifying limber pine from other treeline species, especially since it is a species of conservation concern and its distribution at treeline in Rocky Mountain National Park is incompletely known.

4.1 Overview of CNN Model Performance

Overall, our results had mixed success. The six-class model performed reasonably well given the difficulty of the problem, with 44.1% top-1 and 70% top-2 accuracy—much better than a trivial model could achieve (28.0% and 47.6%, respectively). The four-class model did not improve classification accuracy as much as expected, with 46.7% top-1 and 76.7% top-2 accuracy (vs. 35.1% and 63.1%, respectively, for a trivial model). The simplified two-class model achieved a fairly high overall accuracy of 86.2% (vs. 80.3% for a trivial model). The two-class model distinguished limber pine from other trees and shrubs with 64.0% accuracy, which, on its own, may not be useful for more than identifying regions of treeline where high-probability limber pine pixels tend to cluster. However, the model did notably well (92.1%) at identifying pixels that are not limber pine.

4.2 Predictor Importance

The panchromatic band was the most important predictor based on the results of the XAI permutation tests, both for overall model performance in all three models, and for identifying limber pine specifically. This makes sense both biologically and in terms of the model structure. The models were trained on the panchromatic imagery for two convolutional blocks before the multispectral and elevation data were introduced to the model; this architectural design allowed us to make use of the higher resolution panchromatic data to detect fine-scale spatial patterns, and it likely also reinforced the mathematical influence of these data in the final model predictions. Treeline species often have distinctive physiological growth responses to stressors (e.g., high winds and heavy snowpack), and their krummholz growth forms vary in their spatial patterns. Limber pine frequently occurs as a solitary tree on the landscape, occupying very few pixels and surrounded by alpine tundra (Sindewald and Tomback and Neumeyer, 2020). Other species are more likely to form larger patches through vegetative layering and are also more likely to co-occur on the landscape. The CNN models still relied heavily on the panchromatic data for these species, but they also made greater use of the multispectral imagery for discrimination among these species.

https://doi.org/10.5194/egusphere-2025-970
Preprint. Discussion started: 28 May 2025

Elevation was the second most important predictor for six-class model performance after the panchromatic band. When discerning among six species classes, the CNN clearly picked up on topographic patterns in species distribution on the landscape. Willow is most abundant near creeks, with aspen not much farther away. Engelmann spruce and subalpine fir tend to be found further from creeks, but in topographic depressions with late-lying snowpack. Glandular birch and limber pine are more often found on slopes, and limber pine occupies windswept ridges with early snowmelt.

Multispectral bands were consistently identified as important predictors, suggesting that the CNN model utilized the multispectral imagery despite its lower spatial resolution, compared with the panchromatic imagery. The visible bands detect variation in concentrations of photosynthetic pigments in the leaves, which differ among species. The near-infrared (N-IR) region of the spectrum varies based on cellular structure, which changes with water content, making N-IR bands useful for assessing plant health (e.g., the normalized difference vegetation index or NDVI) (Curran, 1989; Campbell and Wynne, 2011). Trees vary widely in condition at treeline due to varying levels of frost desiccation, wind damage, or water availability. The relatively lower importance of the N-IR bands for discriminating species at treeline may mean that damage to trees and shrubs introduces variation across species that almost overshadows species-specific responses to these stressors.

However, the lower importance of the N-IR and red edge bands may also be because the CNN models prioritized the higher resolution spatial data and spatial patterns of species distributions in the panchromatic imagery over their spectral differences. Table S1 of section S3 of the Supplement summarizes the WV-3 bands where pairs of species may be statistically distinguished. Each pair of species differs in at least one band except for Engelmann spruce and willow, which overlapped across all eight bands. Interestingly, species pairs commonly differed significantly in the near-infrared bands. The fact that these bands were less important for CNN performance suggests that the models relied on panchromatic data *first* and only used the multispectral or DEM data *second*—whenever the panchromatic data were not informative. In those cases (e.g., cases where species growth forms were similar), the models made greater use of the DEM and visible bands. It is possible that the near-infrared bands are more correlated with the panchromatic band, and so the models obtained little additional information from the N-IR and red edge bands. This is plausible given that both the near-infrared bands and the panchromatic band are responsive to plant structure.

4.3 Model Generalizability

While our models have been validated through five-fold cross-validation, we need to test these models on the classification of geographically distinct treeline communities. Machine learning models tend to over-fit models, which is why we employed several regularization methods to reduce the risk of overfitting and to improve generalizability (see Supplement S1.4). However, the importance of the elevation data in species classifications, despite its low original resolution of 10 m, suggests that these models could have learned idiosyncratic topographic patterns at the two study sites (i.e., the “Clever

https://doi.org/10.5194/egusphere-2025-970
Preprint. Discussion started: 28 May 2025

Hans" problem, where machine learning models pick up on signals in the data that are sometimes collinear with more meaningful signals, much like the popular history horse, Clever Hans, who could respond to emotional anticipation in humans but could not, in fact, count). If the models learned that moisture-sensitive species tend to occur in topographic depressions, where snow persists into the spring, they could still generalize well to other treeline locations. On the other hand, if the models relied on proximity to a creek that ran through the study site, they may perform less well at treeline sites at comparable elevations but without a creek. We will test these models on a geographically independent dataset.

4.4 Trade-offs of Computation and Data Acquisition Costs and Increases in Model Accuracy

The use of combined hyperspectral and lidar data is increasingly considered the gold standard for identifying trees with remote sensing data, but our work opens the door for more cost-effective methods for researchers and managers. Ørka and Hauglin (2016) compared remote sensing data acquisition costs and found that high-resolution commercial satellite imagery is much less expensive than airborne aerial imagery (Ørka and Hauglin, 2016). Cost estimates do not include subsequent computational costs incurred through the analysis of such datasets, which can also represent a barrier for wider-scale implementation; not every manager and researcher has the training or money to make use of supercomputing resources.

WV-3 potentially represents a better balance of cost and classification accuracy. WV-3 imagery is cheaper than aerial imagery and provides comparable resolution. The addition of ACOMP atmospheric correction is important, because this method uses CAVIS data collected simultaneously with the multispectral and panchromatic data, yielding highly accurate corrections (Pacifici, 2016). However, as our study showed, even WV-3 data may have spatial resolution that is too low for classification of treeline vegetation species.

UAS data may yield better results than WV-3 given their very high spatial resolution (sub-meter). In fact, Onishi and Ise (2021) demonstrated that a classification accuracy of over 90% for seven species can be achieved using only red-green-blue (RGB) imagery, classified with a CNN (Onishi and Ise, 2021). These authors also used an XAI method, guided Grad-CAM, to determine that the CNN was focusing on canopy shapes for its classification. Our findings support theirs: spatial resolution is important for tree species classification problems. The UAS approach may also be best for classification problems requiring high temporal resolution (e.g., monthly). If data collection spans multiple years or multiple seasons, the UAS approach may be the most cost-effective. UAS, however, are prohibited on some public lands in the United States, particularly in national parks and congressionally designated wilderness areas, which include most treeline areas of conservation interest.

https://doi.org/10.5194/egusphere-2025-970
Preprint. Discussion started: 28 May 2025

5 Conclusions

To our knowledge, we are the first to use satellite imagery to distinguish tree and shrub species in alpine treeline ecosystems (Garbarino et al., 2023). Our study approach fills an important methodological gap, enabling researchers to connect landscape- and local-level treeline patterns and local treeline processes by leveraging field research on treeline species. Species-level maps of alpine treeline would enable researchers to better stratify field sampling efforts to understand interspecies dynamics, as well as how species tolerances influence treeline elevation advance (or lack thereof).

Our models have proven useful for identifying probability hotspots for limber pine occurrence at treeline, supporting ongoing research and management of this ecologically important conifer. Our methods are also more cost-effective than techniques relying on hyperspectral and lidar data collected with aircraft and are applicable to a broader geographic extent than UAS. Our work may also support the ongoing research and conservation of limber pine in the Rocky Mountains of the U.S. and Canada (Schoettle et al., 2019).

However, we believe that adapting these models to an object-based classification approach will improve classification accuracies and allow landscape-level species identification without the need to train additional models. We anticipate that outputs from our CNN models, in the form of pixel-level probability maps for classified species, may be easily processed into segmented images of tree and shrub objects. We will first apply an NDVI filter to isolate tree and shrub pixels, then run the CNN models on these pixels before segmenting the image into objects. The aggregate classifications of these tree and shrub objects (vegetation patches) are likely to have even higher accuracy, enabling landscape-level classification of tree species with high-resolution satellite imagery.

https://doi.org/10.5194/egusphere-2025-970
Preprint. Discussion started: 28 May 2025

Appendix A. Field Methods

Figure A1. Lucas Rudasill and Laurel Sindewald collecting a ground control point and descriptive metadata at the corner of a switchback next to a cairn in RMNP. This photo was included, with the GPS data, descriptive metadata, and image chips, to inform MAXAR orthorectification.

https://doi.org/10.5194/egusphere-2025-970
Preprint. Discussion started: 28 May 2025

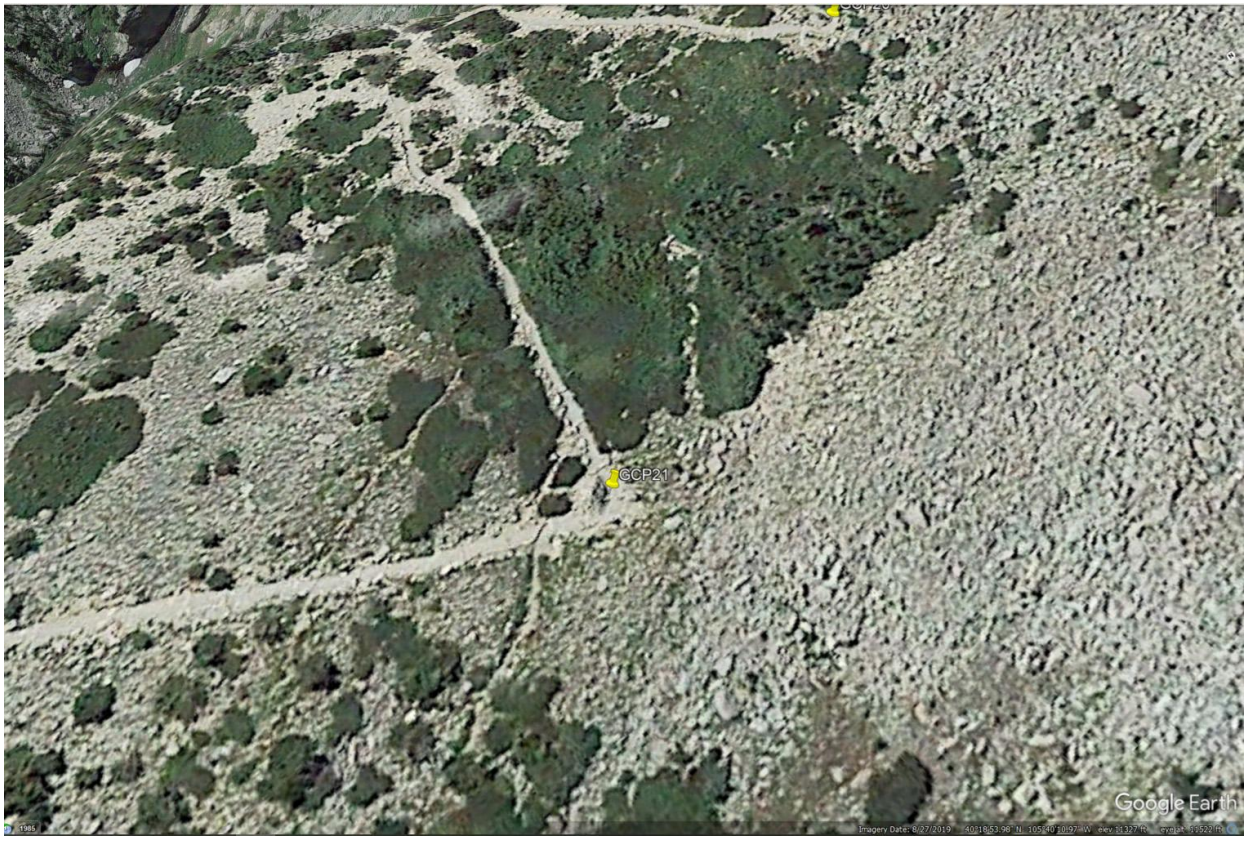


Figure A2. Example of an image chip from Google Earth Pro with the GCP position marked precisely on the image, corresponding to the photo documentation in Figure 26. This image chip was included, with the GPS data, descriptive metadata, and photos, to inform MAXAR orthorectification. © Google Earth.

https://doi.org/10.5194/egusphere-2025-970
Preprint. Discussion started: 28 May 2025

Appendix B. Loss Functions and Model Evaluation Metrics

We used eight metrics for model evaluation: top-1, top-2, and top-3 accuracies, the Gerrity score, the PIFL-first Gerrity score, the Heidke and Peirce scores, and cross-entropy. Note that evaluation metrics, used to assess the performance of an already-trained model, are not necessarily the same as the loss function used for training. Only the Gerrity score, PIFL-first Gerrity score, and cross-entropy were used as loss functions in this study. Table B1 summarizes characteristics of each metric.

Table B1. Characteristics of metrics used for model evaluation.

Evaluation Metric	Range of Possible Values	Optimal Value	Special Values
Top-k accuracies	0-1	1	NA
Default Gerrity score	[-1, +1]	1	0 indicates no skill (random model)
Class-weighted Gerrity score	[-1, +1]	1	0 indicates no skill (random model)
Heidke score	(-∞, 1]	1	0 indicates no skill and < 0 means worse than a random model
Peirce score	[-1, 1]	1	0 indicates no skill and < 0 means worse than a random model
Cross-entropy	[0,∞)	0	NA

Eq. B1 shows how the Gerrity score is calculated, without additional class-weighting (Gerrity, 1992).

$$\begin{cases} \text{Default GS} = \dfrac{1}{N}\sum_{i=1}^{K}\sum_{j=1}^{K} n_{ij}s_{ij}\,, \\ s_{ij} = s_{ji} = \dfrac{1}{K-1}\left[\sum_{k=1}^{i-1} a_k^{-1} - (j-i) + \sum_{k=j}^{K-1} a_k\right], \\ a_k = \dfrac{1-\frac{1}{N}\sum_{r=1}^{k} n(y_r)}{\frac{1}{N}\sum_{r=1}^{k} n(y_r)} \end{cases} \tag{B1}$$

where N is the total number of data samples, K is the number of classes, i and j index the predictions and observations, respectively, and n_ij is the number of data samples with the ith class predicted and jth class observed. When $i = j$ (when the prediction matches the observation), s is positive and higher to reward the correct prediction. When $i \neq j$, s is lower or even negative to penalize the incorrect prediction. s_{ij} is in turn determined by the second function in Equation B1, which determines the weights based on class frequencies to increase the reward when a low-frequency class is correctly predicted and decrease the penalty when a low-frequency class prediction is incorrect. In the second function, which determines the s_{ij}

https://doi.org/10.5194/egusphere-2025-970
Preprint. Discussion started: 28 May 2025

weights, a_k is the *cumulative observation frequency* of the first k classes and is defined in the third function. $n(y_r)$ is the number of data samples where the rth class is correct.

Table B1 is an example of an s-matrix (a matrix of s_{ij} weights) for the four-class model. The s_{ij} weights are larger for classes that have lower frequencies but also depend strongly on the order of the classes. For example, PIEN had a low class frequency (0.172) relative to Other (0.351), so the s_{ij} (PIEN, PIEN) weight was higher than any of the weights for Other. Because PIEN was the first class indexed for the Gerrity score, and because s_{ij} is calculated based on a_k, which is the cumulative observation frequency of the first k classes, the (PIEN, PIEN) weight was greater than the (PIFL, PIFL) weight, even though PIFL had the almost same class frequency as PIEN.

Table B2. The *s*-matrix for the default Gerrity score for the four-class model. The class frequences were 0.172 for PIEN, 0.197 for PIFL, 0.280 for ABLA, and 0.351 for Other.

	PIEN	PIFL	ABLA	Other
PIEN	2.35	0.42	-0.49	-1.00
PIFL	0.42	0.82	-0.08	-0.60
ABLA	-0.49	-0.08	0.44	-0.07
Other	-1.00	-0.60	-0.07	0.88

The reverse was true when PIFL was indexed first in the Gerrity score: PIFL had by far the highest weight, even though PIFL had only a slightly lower class frequency than PIEN (Table B3).

Table B3. The *s*-matrix for the PIFL-first Gerrity score for the four-class model. The class frequences were 0.172 for PIEN, 0.197 for PIFL, 0.280 for ABLA, and 0.351 for Other.

	PIFL	PIEN	ABLA	Other
PIFL	2.11	0.42	-0.49	-1.00
PIEN	0.42	0.83	-0.07	-0.59
ABLA	-0.49	-0.07	0.46	-0.06
Other	-1.00	-0.58	-0.06	0.89

Eq. B1 shows the default Gerrity score, which does not have additional class-weighting. We also added weights to the equation to prioritize the low-frequency classes even more heavily. Eq. B2 shows the class-weighted Gerrity score, which replaces the first function defined in Eq. B1. (The other two functions remain the same.)

https://doi.org/10.5194/egusphere-2025-970
Preprint. Discussion started: 28 May 2025

$$\text{Weighted GS} = \begin{cases} \dfrac{\sum_{i=1}^{K}\sum_{j=1}^{K} w_j n_{ij} s_{ij}}{\sum_{i=1}^{K}\sum_{j=1}^{K} w_j\, n_{ij}} \\ w_j = ln\left(min\left(\dfrac{1}{f_j}, 50\right)\right) \end{cases} \tag{B2}$$

where f_j is the observed frequency of the jth class and w_j is the resulting weight. The weighting function we devised limits the degree to which a class can be prioritized by capping the weight at the natural log of 50.

Cross-entropy is simpler than the Gerrity score. Cross-entropy quantifies the number of bits required to distinguish the distribution of model predictions from the distribution of observations. Cross-entropy is calculated using Eq. B3, and its characteristics are listed in Table B1.

$$\varepsilon = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik} log_2(p_{ik}) \tag{B3}$$

where N is the number of data samples, K is the number of classes, p_{ik} is the model-predicted probability that the i^{th} sample belongs to the k^{th} class, and y_{ik} is a binary indication of the correct class, which is 1 if the i^{th} example belongs to the k^{th} class and 0 otherwise.

The Heidke (Heidke, 1926) and Peirce (Peirce, 1884) scores are similar in that they both measure the proportion of correct predictions above and beyond those that would be expected from a random model (Lagerquist and McGovern and Gagne Ii, 2019). Their equations differ slightly. The Heidke score can be calculated with Eq. B4 and its domain is listed in Table B1.

$$Heidke\ score = \frac{\frac{1}{N}\sum_{k=1}^{K} n_{kk} - \frac{1}{N^2}\sum_{k=1}^{K} n(P_k)n(y_k)}{1 - \frac{1}{N^2}\sum_{k-1}^{K} n(P_k)n(y_k)} \tag{B4}$$

where N is the total number of data samples, K is the total number of classes, n_{kk} is a correct prediction of class k, $n(P_k)$ is the number of samples where the k^{th} class is predicted, and $n(y_k)$ is the number of samples where the k^{th} class is observed. The Peirce score can be calculated with Eq. B5 and shares these term definitions.

$$Peirce\ score = \frac{\frac{1}{N}\sum_{k=1}^{K} n_{kk} - \frac{1}{N^2}\sum_{k=1}^{K} n(P_k)n(y_k)}{1 - \frac{1}{N^2}\sum_{k=1}^{K} n(y_k)^2} \tag{B5}$$

https://doi.org/10.5194/egusphere-2025-970
Preprint. Discussion started: 28 May 2025

Appendix C. Convolutional Neural Network Architecture Diagrams (Four- and Two-Class Models)

The panchromatic data were input to the first convolutional block, labeled "Conv 513 x 513 x 4" and “Max pool 257 x 257 x 4” in Figure 1. All convolutional blocks in our CNN contained two convolutional layers, using 3 x 3-pixel convolutional filters or "kernels". The first convolutional layer in this block transformed the panchromatic data to two feature maps, via two learned convolutional filters. The second convolutional layer in the block then transformed the two feature maps to four feature maps, via four learned filters. Each convolutional layer was followed by an activation function, which is a pixel-by-pixel non-linear transformation. Activation functions are the key component enabling neural networks to learn non-linear relationships. Our specific activation function was the leaky rectified linear unit (ReLU) with a slope parameter of 0.2 (Nair and Hinton, 2010). Furthermore, each activation function was followed by batch normalization (Ioffe and Szegedy, 2015), which restores values in the maps to an approximately standard normal distribution. (The predictor variables followed a standard normal distribution, ensured by our z-score transform, but the operations carried out in a CNN can warp this distribution, which leads to slower training convergence.) After two convolutional layers with activation and batch normalization, we used max pooling to reduce the spatial resolution by half. With max pooling, the maximum value for each set of four pixels is retained. The full series of convolution > activation > batch normalization > convolution > activation > batch normalization > pooling made up one convolutional block (Figures 1, C1, and C2). CNN model architecture diagrams for the four-class and two-class models are shown in figures C1 and C2, respectively.

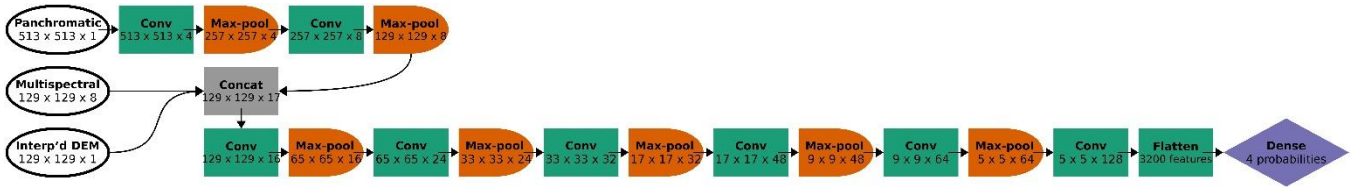


Figure C1. CNN architecture for the four-class model, incorporating panchromatic, multispectral, and DEM inputs at different spatial resolutions. The architecture is the same as the six- and two-class models, but with a different number of probabilities output by the final dense layer.

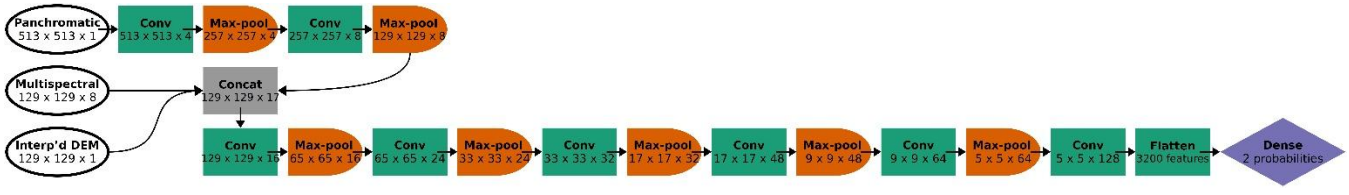


Figure C2. CNN architecture for the two-class model, incorporating panchromatic, multispectral, and DEM inputs at different spatial resolutions. The architecture is the same as the six- and four-class models, but with a different number of probabilities output by the final dense layer.

https://doi.org/10.5194/egusphere-2025-970
Preprint. Discussion started: 28 May 2025

Code availability

Code used to train the CNN models is available in a repository: https://doi.org/10.5281/zenodo.14946215.

Data availability

Data patches for all training samples and channels, as well as a CSV file with mean radiance values for each species region of interest polygon for each multispectral WV-3 band, are available in a repository: https://zenodo.org/records/14942410.

Author contributions

Field methods were developed by L.A.S, D.F.T, M.D.C, and T.A.S. Field data were collected by L.A.S. and R.L, with help from field assistants. WorldView-3 data were obtained by L.A.S. and M.D.C., with funding obtained by D.F.T., M.D.C., and P.J.A. CNN methods were developed and implemented by R.L. with domain-specific advisement from L.A.S. Additional analyses performed by L.A.S. Manuscript preparation and submission by L.A.S with reviews and editing provided by R.L., D.F.T., M.D.C., P.J.A., and T.A.S.

Competing Interests

The authors declare no competing interests.

Special Issue Statement

This article is part of the special issue, "Treeline ecotones under global change: linking spatial patterns to ecological processes". It is not associated with a conference.

Acknowledgements

The Office of Research Services, University of Colorado Denver, provided funding to DFT, MDC, and PJA for this research. We thank Maxar Technologies for an academic discount and for working with us on the imagery. We thank Frontier Precision for an academic discount on the Zephyr 2 antennae and Trimble Geo7x Centimeter Edition rental. The Continental Divide Research Learning Center in Rocky Mountain National Park provided lodging for our researchers and research assistants. We are grateful to the Center for Computational Mathematics, University of Colorado Denver, for providing computing resources and especially thank Jan Mandel for access to the Alderaan Cluster, which is supported by the National Science Foundation award OAC-2019089. Field assistants and volunteers—Lucas Rudisill, Nicole Hinostroza, Libby Pansing, and Aaron Wagner—provided invaluable help with data collection at treeline.

https://doi.org/10.5194/egusphere-2025-970
Preprint. Discussion started: 28 May 2025

References

Allen, T. R. and Walsh, S. J.: Spatial and compositional pattern of alpine treeline, Glacier National Park, Montana, Photogrammetric engineering and remote sensing., 62, 1261-1268, 1996.

Bader, M. Y., Llambí, L. D., Case, B. S., Buckley, H. L., Toivonen, J. M., Camarero, J. J., Cairns, D. M., Brown, C. D., Wiegand, T., and Resler, L. M.: A global framework for linking alpine-treeline ecotone patterns to underlying processes, Ecography, 44, 265-292, https://doi.org/10.1111/ecog.05285, 2021.

Bansal, S., Reinhardt, K., and Germino, M. J.: Linking carbon balance to establishment patterns: comparison of whitebark pine and Engelmann spruce seedlings along an herb cover exposure gradient at treeline, Plant Ecology, 212, 219-228, 10.1007/s11258-010-9816-8, 2011.

Batllori, E. and Gutiérrez, E.: Regional tree line dynamics in response to global change in the Pyrenees, Journal of Ecology, 96, 1275-1288, https://doi.org/10.1111/j.1365-2745.2008.01429.x, 2008.

Batllori, E., Camarero, J. J., Ninot, J. M., and Gutiérrez, E.: Seedling recruitment, survival and facilitation in alpine *Pinus uncinata* tree line ecotones. Implications and potential responses to climate warming, Global Ecology and Biogeography, 18, 460-472, 10.1111/j.1466-8238.2009.00464.x, 2009.

Brodersen, C. R., Germino, M. J., Johnson, D. M., Reinhardt, K., Smith, W. K., Resler, L. M., Bader, M. Y., Sala, A., Kueppers, L. M., Broll, G., Cairns, D. M., Holtmeier, F.-K., and Wieser, G.: Seedling survival at timberline is critical to conifer mountain forest elevation and extent, Frontiers in Forests and Global Change, 2, 10.3389/ffgc.2019.00009, 2019.

Brown, D. G., Cairns, D. M., Malanson, G. P., Walsh, S. J., and Butler, D. R.: Remote sensing and GIS techniques for spatial and biophysical analyses of alpine treeline through process and empirical models, Remote sensing and GIS techniques for spatial and biophysical analyses of alpine treeline through process and empirical models, Taylor & Francis Ltd., United Kingdom, 453-481 pp.1994.

Burns, R. M. and Honkala, B. H.: Volume 1: Conifers, in: Silvics of North America, Agriculture Handbook, 654, U.S. Department of Agriculture, Forest Service, Washington, DC, 1990.

Callaway, R. M.: Competition and facilitation on elevation gradients in subalpine forests of the Northern Rocky Mountains, USA, Oikos, 82, 561-573, 10.2307/3546376, 1998.

Campbell, J. B. and Wynne, R. H.: Plant Sciences, in: Introduction to Remote Sensing, The Guilford Press, New York, NY, 2011.

Cross, M., Scambos, T., Pacifici, F., Vargas-Ramirez, O., Moreno-Sanchez, R., and Marshall, W.: Classification of tropical forest tree species using meter-scale image data, Remote Sensing, 11, 1411, 2019a.

Cross, M. D., Scambos, T., Pacifici, F., and Marshall, W. E.: Determining effective meter-scale image data and spectral vegetation indices for tropical forest tree species differentiation, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 12, 2934-2943, 10.1109/JSTARS.2019.2918487, 2019b.

Cui, M. and Smith, W. K.: Photosynthesis, water relations and mortality in *Abies lasiocarpa* seedlings during natural establishment, Tree Physiol, 8, 37-46, 10.1093/treephys/8.1.37, 1991.

Curran, P. J.: Remote sensing of foliar chemistry, Remote Sensing of Environment, 30, 271-278, https://doi.org/10.1016/0034-4257(89)90069-2, 1989.

Dalponte, M., Bruzzone, L., and Gianelle, D.: Tree species classification in the Southern Alps based on the fusion of very high geometrical resolution multispectral/hyperspectral images and LiDAR data, Remote Sensing of Environment, 123, 258-270, https://doi.org/10.1016/j.rse.2012.03.013, 2012.

Dubey, A. K. and Jain, V.: Comparative study of convolution neural network's Relu and Leaky-Relu activation functions, Applications of Computing, Automation and Wireless Systems in Electrical Engineering, Singapore, 2019//, 873-880,

Fassnacht, F. E., Latifi, H., Stereńczak, K., Modzelewska, A., Lefsky, M., Waser, L. T., Straub, C., and Ghosh, A.: Review of studies on tree species classification from remotely sensed data, Remote Sensing of Environment, 186, 64-87, https://doi.org/10.1016/j.rse.2016.08.013, 2016.

Feuillet, T., Birre, D., Milian, J., Godard, V., Clauzel, C., and Serrano-Notivoli, R.: Spatial dynamics of alpine tree lines under global warming: What explains the mismatch between tree densification and elevational upward shifts at the tree line ecotone?, Journal of Biogeography, 47, 1056-1068, https://doi.org/10.1111/jbi.13779, 2020.

https://doi.org/10.5194/egusphere-2025-970
Preprint. Discussion started: 28 May 2025

Garbarino, M., Morresi, D., Anselmetto, N., and Weisberg, P. J.: Treeline remote sensing: from tracking treeline shifts to multi-dimensional monitoring of ecotonal change, Remote Sensing in Ecology and Conservation, 9, 729-742, https://doi.org/10.1002/rse2.351, 2023.
Germino, M. J., Smith, W. K., and Resor, A. C.: Conifer seedling distribution and survival in an alpine-treeline ecotone, Plant Ecology, 162, 157-168, 10.1023/a:1020385320738, 2002.
Gerrity, J. P.: A Note on Gandin and Murphy's Equitable Skill Score, Monthly Weather Review, 120, 2709-2712, https://doi.org/10.1175/1520-0493(1992)120<2709:ANOGAM>2.0.CO;2, 1992.
Gill, R. A., Campbell, C. S., and Karlinsey, S. M.: Soil moisture controls Engelmann spruce *(Picea engelmannii)* seedling carbon balance and survivorship at timberline in Utah, USA, Canadian Journal of Forest Research, 45, 1845-1852, 10.1139/cjfr-2015-0239, 2015.
Goodfellow, I., Bengio, Y., and Courville, A.: Regularization, in: Deep Learning, The MIT Press, 2016a.
Goodfellow, I., Bengio, Y., and Courville, A.: Optimization, in: Deep Learning, The MIT Press, 2016b.
Goodfellow, I., Bengio, Y., and Courville, A.: Machine Learning Basics, in: Deep Learning, The MIT Press, 2016c.
Goodfellow, I., Bengio, Y., and Courville, A.: Convolutional Networks, in: Deep Learning, The MIT Press, 2016d.
Hankin, L. E. and Bisbing, S. M.: Let it snow? Spring snowpack and microsite characterize the regeneration niche of high-elevation pines, Journal of Biogeography, 48, 2068-2084, https://doi.org/10.1111/jbi.14136, 2021.
Harsch, M. A., Hulme, P. E., McGlone, M. S., and Duncan, R. P.: Are treelines advancing? A global meta-analysis of treeline response to climate warming, Ecology Letters, 12, 1040-1049, 10.1111/j.1461-0248.2009.01355.x, 2009.
Heidke, P.: Berechnung Des Erfolges Und Der Güte Der Windstärkevorhersagen Im Sturmwarnungsdienst, Geografiska Annaler, 8, 301-349, 10.1080/20014422.1926.11881138, 1926.
Hessl, A. E. and Baker, W. L.: Spruce and fir regeneration and climate in the forest-tundra ecotone of Rocky Mountain National Park, Colorado, U.S.A, Arctic and Alpine Research, 29, 173-183, 10.1080/00040851.1997.12003230, 1997.
Holtmeier, F.-K. and Broll, G.: Sensitivity and response of Northern Hemisphere altitudinal and polar treelines to environmental change at landscape and local scales, Global Ecology and Biogeography, 14, 395-410, 2005.
Holtmeier, F. K.: Physiognomic and Ecological Differentiation of Mountain Timberline, in: Mountain Timberlines: Ecology, Patchiness, and Dynamics, 2 ed., Advances in Global Change Research, 36, Springer Netherlands, 2009.
Immitzer, M., Atzberger, C., and Koukal, T.: Tree species classification with random forest using very high spatial resolution 8-band WorldView-2 satellite data, Remote Sensing, 4, 2661-2693, 2012.
Ioffe, S. and Szegedy, C.: Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift, Proceedings from the International Conference on Machine Learning, 448-456,
Jones, B., Gutsell, R., Barnhardt, L., Gould, J., and Smith, C.: Alberta limber pine recovery plan 2014-2019, 2014.
Kingma, D. and Ba, J.: Adam: A Method for Stochastic Optimization, International Conference on Learning Representations, 2014.
Körner, C.: A re-assessment of high elevation treeline positions and their explanation, Oecologia, 115, 445-459, 10.1007/s004420050540, 1998.
Körner, C. and Paulsen, J.: A world-wide study of high altitude treeline temperatures, Journal of Biogeography, 31, 713-732, 10.1111/j.1365-2699.2003.01043.x, 2004.
Lagerquist, R.: Using deep learning to improve prediction and understanding of high-impact weather, School of Meteorology, University of Oklahoma, Norman, OK, 290 pp., 2020.
Lagerquist, R., McGovern, A., and Gagne Ii, D. J.: Deep learning for spatially explicit prediction of synoptic-scale fronts, Weather and Forecasting, 34, 1137-1160, https://doi.org/10.1175/WAF-D-18-0183.1, 2019.
Lagerquist, R., McGovern, A., Homeyer, C. R., Gagne II, D. J., and Smith, T.: Deep learning on three-dimensional multiscale data for next-hour tornado prediction, Monthly Weather Review, 148, 2837-2861, https://doi.org/10.1175/MWR-D-19-0372.1, 2020.
Lagerquist, R., Turner, D., Ebert-Uphoff, I., Stewart, J., and Hagerty, V.: Using deep learning to emulate and accelerate a radiative transfer model, Journal of Atmospheric and Oceanic Technology, 38, 1673-1696, 10.1175/JTECH-D-21-0007.1, 2021.
Leonelli, G., Masseroli, A., and Pelfini, M.: The influence of topographic variables on treeline trees under different environmental conditions, Physical Geography, 37, 56-72, 10.1080/02723646.2016.1153377, 2016.

https://doi.org/10.5194/egusphere-2025-970
Preprint. Discussion started: 28 May 2025

Li, D., Ke, Y., Gong, H., and Li, X.: Object-based urban tree species classification using bi-temporal WorldView-2 and WorldView-3 images, Remote Sensing, 7, 16917-16937, 2015.
Li, M., Zhang, T., Chen, Y., and Smola, A. J.: Efficient mini-batch training for stochastic optimization, Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, New York, New York, USA, 10.1145/2623330.2623612, 2014.
Linhart, Y. B. and Tomback, D. F.: Seed dispersal by nutcrackers causes multi-trunk growth form in pines, Oecologia, 67, 107-110, 10.1007/BF00378458, 1985.
Liu, L., Coops, N. C., Aven, N. W., and Pang, Y.: Mapping urban tree species using integrated airborne hyperspectral and LiDAR remote sensing data, Remote Sensing of Environment, 200, 170-182, https://doi.org/10.1016/j.rse.2017.08.010, 2017.
Majid, I. A., Latif, Z. A., and Adnan, N. A.: Tree species classification using WorldView-3 data, 2016 7th IEEE Control and System Graduate Research Colloquium (ICSGRC), 8-8 Aug. 2016, 73-76, 10.1109/ICSGRC.2016.7813304,
Malanson, G. P., Butler, D. R., Fagre, D. B., Walsh, S. J., Tomback, D. F., Daniels, L. D., Resler, L. M., Smith, W. K., Weiss, D. J., Peterson, D. L., Bunn, A. G., Hiemstra, C. A., Liptzin, D., Bourgeron, P. S., Shen, Z., and Millar, C. I.: Alpine treeline of western North America: linking organism-to-landscape dynamics, Physical Geography, 28, 378-396, 10.2747/0272-3646.28.5.378, 2007.
Matsuki, T., Yokoya, N., and Iwasaki, A.: Hyperspectral tree species classification of Japanese complex mixed forest with the aid of lidar data, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 8, 2177-2187, 10.1109/JSTARS.2015.2417859, 2015.
McCune, B.: Ecological diversity in North American pines, Am J Bot, 75, 353-368, 10.2307/2443983, 1988.
McGovern, A., Lagerquist, R., John Gagne, D., Jergensen, G. E., Elmore, K. L., Homeyer, C. R., and Smith, T.: Making the black box more transparent: Understanding the physical implications of machine learning, Bulletin of the American Meteorological Society, 100, 2175-2199, https://doi.org/10.1175/BAMS-D-18-0195.1, 2019.
McIntire, E. J. B., Piper, F. I., and Fajardo, A.: Wind exposure and light exposure, more than elevation-related temperature, limit tree line seedling abundance on three continents, Journal of Ecology, 104, 1379-1390, 10.1111/1365-2745.12599, 2016.
Millar, C. I., Westfall, R. D., Delany, D. L., Flint, A. L., and Flint, L. E.: Recruitment patterns and growth of high-elevation pines in response to climatic variability (1883–2013), in the western Great Basin, USA, Canadian Journal of Forest Research, 45, 1299-1312, 10.1139/cjfr-2015-0025, 2015.
Mishra, N. B., Mainali, K. P., Shrestha, B. B., Radenz, J., and Karki, D.: Species-level vegetation mapping in a Himalayan treeline ecotone using unmanned aerial system (UAS) imagery, ISPRS International Journal of Geo-Information, 7, 445, 2018.
Monahan, W. B., Cook, T., Melton, F., Connor, J., and Bobowski, B.: Forecasting distributional responses of limber pine to climate change at management-relevant scales in Rocky Mountain National Park, Plos One, 8, ARTN e83163 10.1371/journal.pone.0083163, 2013.
Nair, V. and Hinton, G. E.: Rectified Linear Units Improve Restricted Boltzmann Machines, Proceedings of the 27th International Conference on Machine Learning, Haifa, Israel,
Onishi, M. and Ise, T.: Explainable identification and mapping of trees using UAV RGB image and deep learning, Scientific Reports, 11, 903, 10.1038/s41598-020-79653-9, 2021.
Ørka, H. O. and Hauglin, M.: Use of remote sensing for mapping of non-native conifer species, Norwegian University of Life Sciences, INA fagapport 33, 2016.
Pacifici, F.: Validation of the DigitalGlobe surface reflectance product, 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 10-15 July 2016, 1973-1975, 10.1109/IGARSS.2016.7729508,
Peirce, C. S.: The Numerical Measure of the Success of Predictions, Science, ns-4, 453-454, doi:10.1126/science.ns-4.93.453.b, 1884.
Pu, R.: Mapping tree species using advanced remote sensing technologies: A state-of-the-art review and perspective, Journal of Remote Sensing, 10.34133/2021/9812624, 2021.
Pyatt, J. C., Tomback, D. F., Blakeslee, S. C., Wunder, M. B., Resler, L. M., Boggs, L. A., and Bevency, H. D.: The importance of conifers for facilitation at treeline: Comparing biophysical characteristics of leeward microsites in whitebark pine communities, Arctic, Antarctic, and Alpine Research, 48, 427-444, 10.1657/AAAR0015-055, 2016.

https://doi.org/10.5194/egusphere-2025-970
Preprint. Discussion started: 28 May 2025

Rahman, M. M., Robson, A., and Bristow, M.: Exploring the potential of high resolution WorldView-3 imagery for estimating yield of mango, Remote Sensing, 10, 1866, 2018.

Resler, L. M., Butler, D. R., and Malanson, G. P.: Topographic shelter and conifer establishment and mortality in an alpine environment, Glacier National Park, Montana, Physical Geography, 26, 112-125, 10.2747/0272-3646.26.2.112, 2005.

Resler, L. M., Shao, Y., Tomback, D. F., and Malanson, G. P.: Predicting functional role and occurrence of whitebark pine *(Pinus albicaulis)* at alpine treelines: Model accuracy and variable importance, Annals of the Association of American Geographers, 104, 703-722, 10.1080/00045608.2014.910072, 2014.

Schoettle, A. W., Burns, K. S., Cleaver, C. M., and Connor, J. J.: Proactive limber pine conservation strategy for the Greater Rocky Mountain National Park Area, 2019.

Schoettle, A. W., Burns, K. S., McKinney, S. T., Krakowski, J., Waring, K. M., Tomback, D. F., and Davenport, M.: Integrating forest health conditions and species adaptive capacities to infer future trajectories of the high elevation five-needle white pines, Forest Ecol Manag, 521, 120389, https://doi.org/10.1016/j.foreco.2022.120389, 2022.

Shen, X. and Cao, L.: Tree-species classification in subtropical forests using airborne hyperspectral and LiDAR data, Remote Sensing, 9, 1180, 2017.

Sindewald, L. A., Tomback, D. F., and Neumeyer, E. R.: Community structure and functional role of limber pine *(Pinus flexilis)* in treeline communities in Rocky Mountain National Park, Forests, 11, 838, doi: 10.3390/f11080838, 2020.

Steele, R.: *Pinus flexilis* James, in: Silvics of North America, Agricultural Handbook, No. 654, US Department of Agriculture, Washington, DC, 348-354, 1990.

Tomback, D., Resler, L., Keane, R., Pansing, E., Andrade, A., and Wagner, A.: Community structure, biodiversity, and ecosystem services in treeline whitebark pine communities: Potential impacts from a non-native pathogen, Forests, 7, 21, 2016a.

Tomback, D. F. and Linhart, Y. B.: The evolution of bird-dispersed pines, Ecological Economics, 4, 185-219, 1990.

Tomback, D. F., Blakeslee, S. C., Wagner, A. C., Wunder, M. B., Resler, L. M., Pyatt, J. C., and Diaz, S.: Whitebark pine facilitation at treeline: potential interactions for disruption by an invasive pathogen, Ecol Evol, 6, 5144-5157, 10.1002/ece3.2198, 2016b.

Ulrich, D. E. M., Wasteneys, C., Hoy-Skubik, S., and Alongi, F.: Functional traits underlie specialist-generalist strategies in whitebark pine and limber pine, Forest Ecol Manag, 542, 121113, https://doi.org/10.1016/j.foreco.2023.121113, 2023.

Voss, M. and Sugumaran, R.: Seasonal effect on tree species classification in an urban environment using hyperspectral data, LiDAR, and an object-oriented approach, Sensors, 8, 3020-3036, 2008.

Wagner, A. C., Tomback, D. F., Resler, L. M., and Pansing, E. R.: Whitebark pine prevalence and ecological function in treeline communities of the Greater Yellowstone Ecosystem, U.S.A.: Potential disruption by white pine blister rust, Forests, 9, 635, 2018.

Wang, T., Zhang, H., Lin, H., and Fang, C.: Textural–spectral feature-based species classification of mangroves in Mai Po Nature Reserve from WorldView-3 imagery, Remote Sensing, 8, 24, 2016.

Wei, C., Karger, D. N., and Wilson, A. M.: Spatial detection of alpine treeline ecotones in the Western United States, Remote Sensing of Environment, 240, 111672, https://doi.org/10.1016/j.rse.2020.111672, 2020.