Naturalness indicators of forests in Southern Sweden derived from the canopy height model

European Journal of Remote Sensing

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/tejr20

Naturalness indicators of forests in Southern Sweden derived from the canopy height model

Marco L. Della Vedova & Mattias Wahde

To cite this article: Marco L. Della Vedova & Mattias Wahde (2025) Naturalness indicators of forests in Southern Sweden derived from the canopy height model, European Journal of Remote Sensing, 58:1, 2441834, DOI: 10.1080/22797254.2024.2441834

To link to this article: https://doi.org/10.1080/22797254.2024.2441834

Published online: 23 Dec 2024.

EUROPEAN JOURNAL OF REMOTE SENSING
2025, VOL. 58, NO. 1, 2441834
https://doi.org/10.1080/22797254.2024.2441834

Naturalness indicators of forests in Southern Sweden derived from the canopy height model

Marco L. Della Vedova and Mattias Wahde

Dept. of Mechanics and Maritime Sciences, Chalmers University of Technology, Gothenburg, Sweden

ABSTRACT

Forest canopies embody a dynamic set of ecological factors, acting as a pivotal interface between the Earth and its atmosphere. They are not only the result of an ecosystem's ability to maintain its inherent ecological processes, structures, and functions but also a reflection of human disturbance. This study introduces a methodology for extracting a comprehensive and human-interpretable set of features from the Canopy Height Model (CHM) with a resolution of 1 meter. These features are then analyzed to identify reliable indicators of the degree of naturalness of forests in Southern Sweden. Using these features, machine learning models – specifically, the perceptron, logistic regression, and decision trees – are trained with examples of forests exhibiting known high and low degrees of naturalness. These models achieve prediction accuracies ranging from 89% to 95% on unseen data, depending on the area of the region of interest. The predictions of the proposed method are easy to interpret, making them particularly valuable to various stakeholders involved in forest management and conservation.

ARTICLE HISTORY
Received 11 July 2024
Revised 19 November 2024
Accepted 25 November 2024

KEYWORDS
Machine learning; interpretability; forests; canopy height model; remote sensing

Introduction

Forests, as complex and dynamic ecosystems, play a pivotal role in maintaining ecological balance, biodiversity, and global climate regulation. In the context of forest ecosystems, *naturalness* refers to the degree to which an ecosystem retains its inherent ecological processes, structures, and functions in the absence of significant human intervention (McRoberts et al., 2012). The assessment of the naturalness of forests has emerged as an important endeavor both for defining nature protection areas, considering escalating anthropogenic pressures and loss of biodiversity, and in terms of forest management by national forest agencies worldwide.

Traditionally, the assessment of forest naturalness is carried out by experts with field inventories that are collected in national forest inventories. Various methods have been suggested for evaluating naturalness, yet there is ongoing debate surrounding the merits and drawbacks of these various approaches. Nevertheless, the typical traits of naturalness are considered to be: diversity in trees (species composition, stand structure), presence of dead wood, landscape age structure, and wildlife (fauna) composition (Barrette et al., 2020). Very similar traits have been used in the literature to define the concepts of *ecological integrity* of forests and *forest integrity* (Frego, 2007; Tierney et al., 2009), which are closely related to the definition of naturalness used in this paper. In any case, assessing these characteristics traditionally requires field work, which is costly and time-consuming.

On the other hand, more and more digital data about lands and their usage are collected by remote sensing with satellites and aircraft, both manned and UAVs (Lechner et al., 2020). There is therefore an opportunity to apply techniques from artificial intelligence (AI) to assess forest naturalness with automated and semi-automated methods. In fact, these techniques have been extensively used for similar processes (L. Zhang & Zhang, 2022), such as land use classification (C. Zhang et al., 2018), leaf-area index estimation (Chen et al., 2015), flooding prediction (Saravi et al., 2019; Z. Zhang et al., 2022), etc. Specifically regarding forests, AI methods have been proposed for stand delineation (Olofsson & Holmgren, 2014) and segmentation (Dechesne et al., 2016; Mustonen et al., 2008), prediction of tree species richness (Brugere et al., 2023), drone applications (Buchelt et al., 2024), and biomass estimation (Hong et al., 2023).However, interestingly, not many studies address the problem of automatic assessment of forest naturalness. The assessment of ecological function levels, which is related to naturalness, has been proposed by (Fang et al., 2023), where an overall accuracy of 0.82 was obtained using a random forest classifier (an ensemble of decision trees) applied to a multi-source dataset.

This paper introduces a novel automated method to assess forest naturalness from a set of features extracted from Canopy height model (CHM) with 1 m resolution as single source of data. The CHM is

CONTACT Marco L. Della Vedova marco.dellavedova@chalmers.se Chalmers University of Technology, Hörsalsvägen 7A, Gothenburg, SE-41296, Sweden

a georeferenced raster in which each pixel represents the height of the trees' canopy relative to the ground. CHM data can be obtained from airborne laser scanning surveys, which are becoming increasingly common (Moudrý et al., 2023; Stoker, 2020). Note that with a 1 m resolution CHM it is possible to distinguish individual trees, as also demonstrated by Ozdemir and Donoghue (2013). Among other applications, CHM has been exploited to classify the structure of forest stands (Torresan et al., 2014), to assess natural disturbances (Senf et al., 2020), and to estimate aboveground biomass (Chirici et al., 2016); but not for estimating forest naturalness in the way presented in this paper.

Figure 1 shows an overview of the study: The model takes as input the CHM and the area of interest, defined by a georeferenced polygon representing the boundary of the forest to analyze, and it produces as output the predicted probability of a forest to be of high naturalness. Internally, the method is composed of two components: The feature extraction and the machine learning (ML) classifier. The classifier is trained on forests with known levels of naturalness, categorized as either high or low (i.e. positive and negative examples of highly natural forests) based on data from national forest inventories (see Section 2). After the training phase, the model carries out inference on new, unseen data (details on the data split into training, validation, and test sets are provided in Section 2.3). Model evaluation focuses on several key aspects: Feature importance, comparisons across different classifier types, the impact of polygon area, and the confidence of predictions.

The main benefits of the proposed method are that (1) it is based on a single source of data (the CHM) at inference time, thus simplifying the requirements on data acquisition and processing; (2) it returns a probability of high naturalness, which can be used to measure naturalness on a continuous scale; and (3) it is interpretable, meaning that it is easy to understand its outcomes (glass box model).

There are multiple advantages with having an interpretable method. The distinctive characteristic of such methods is that their inner workings are human-understandable. This is in contrast to the currently popular black box approaches based on deep neural networks (DNNs), with millions (or

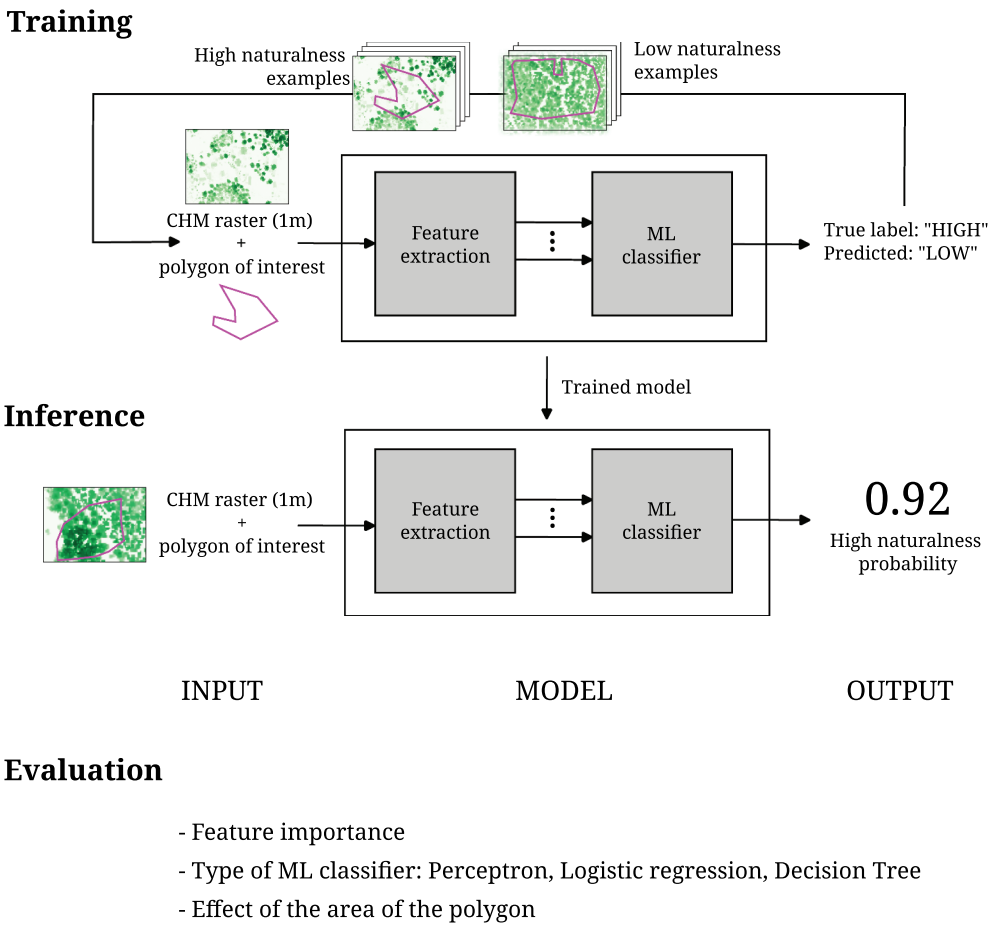

Figure 1. The workflow of this study centers on two core components of the proposed model: feature extraction and machine learning classification. Eight numerical features are derived from the canopy height model (CHM) and the polygon of interest, serving as input to the machine learning (ML) classifier. These features capture key characteristics relevant to forest naturalness, supporting the classifier in distinguishing between areas of high and low naturalness.

even billions) of non-linearly interconnected computational elements (artificial neurons), making it near-impossible for a human to follow their reasoning (Rudin, 2019). We believe that, in the context of assessing forest naturalness, understanding the reasons *why* a forest has been classified as having a certain degree of naturalness is of great value for many stakeholders, such as environmental experts, policy makers, forestry companies, and engineers that develop and maintain the classification method (see also Jiang et al. (2024)). With a human-understandable method, environmental experts can identify key features of forests that are associated with naturalness. Policy makers can enhance transparency and accountability of forest monitoring and assessment programs, and develop more targeted conservation strategies. The transparency of the system is important also from the perspective of forestry companies to understand why conservation strategies are in place, and why they can or cannot cut a particular forest. Last, understanding the inner workings of the system is important for the engineers who develop, debug, maintain, and extend it.

In summary the research questions we address in this paper are the following: (i) Is it possible to identify highly natural forests from the CHM with an interpretable method? (ii) Which of the features of a forest, extracted from the CHM, are indicators of naturalness?

The remainder of the paper is organized as follows: In Section 2 we describe the data used here, and we also provide information about how the data set has been split into training, validation, and test sets. Then, in Section 3, the method is described, starting with the definition of the features used, followed by a presentation of the different machine learning models applied here. Next, in Section 4 we present the results of the performance analysis of the models. A discussion is given in Section 5, and then the conclusions in Section 6.

Materials

The study area is located in Southern Sweden (56.17° – 60.78° N, 10.99° – 19.20° E); see also Figure 2. Forests of this region are predominantly coniferous, with Norway spruce (Picea abies) and Scots pine (Pinus sylvestris) accounting for around 75% of the total standing tree volume. However, there is also a significant presence of broadleaved trees, particularly in southernmost Sweden. These include birch (Betula), European beech (Fagus sylvatica), oak (Quercus spp.), ash (Fraxinus excelsior), and alder (Alnus spp.) (Drössler, 2010).

The data used in this work consists of CHM combined with polygons that delimit the forest areas under consideration and are also associated with a class label (high or low naturalness). Below, we first describe the CHM in Section 2.1 and the polygons in Section 2.2. The process used for dividing the data in the usual training, validation, and test splits is described in Section 2.3.

Canopy height model (CHM)

The main source of data is the CHM. This data set is derived from a large set of georeferenced raster images containing information about the height of the trees. The ground resolution is 1 meter and the measurement precision is 0.1. In other words, each pixel corresponds to a 1×1 m square on the ground and its (integer) value is the height of the upper canopy of the trees relative to the ground, in decimeters. In addition, information about the date of the survey is available as metadata. The data set was collected through airborne laser scanning from 2018 to 2022 and it is provided by Skogsstyrelsen, the Swedish Forest Agency, as open data.[1] Details about the airborne laser scanning are listed in Table 1. Data quality of the CHM, in terms of the root mean square difference in the canopy height as measured by the laserscan and by 2,786 dedicated field surveys, has been estimated to be about 8%.[2]

Labels – ground truth

To collect positive and negative examples of highly natural forests, we use data from different sources. The data contain georeferenced geometries (polygons and multi-polygons) defined by environmental experts in field campaigns. The coverage of these data is limited to selected regions in Sweden, due to the expensive (manual) detection process.

In detail, for positive examples, i.e. forests with high naturalness, we consider the following sources: (1) *Naturvardsverket*, habitat-classed areas within Natura 2000, a network of protected areas throughout the EU[3]; (2) *Storskogsbruket*, inventory of key habitats conducted by forestry companies; and (3) *SksBorealSyd*, inventory of key habitats in the South Boreal region conducted by the Swedish Forest Agency.

Natura 2000 forests are manually annotated and include details about the inventory method employed, along with the assigned naturalness classification for the forest stand. Among all the available forests, we limited our scope by filtering on attributes: We

[1]Available at https://www.skogsstyrelsen.se/sjalvservice/karttjanster/skogliga-grunddata/, *trädhöjd* layer.
[2]Cf. https://www.skogsstyrelsen.se/globalassets/om-oss/regeringsuppdrag/uppdatering-av-skogliga-grundata/kvalitetsbeskrivning-skogliga-skattningar-laserdata-20240115.pdf
[3]See https://ec.europa.eu/environment/nature/natura2000/index_en.htm

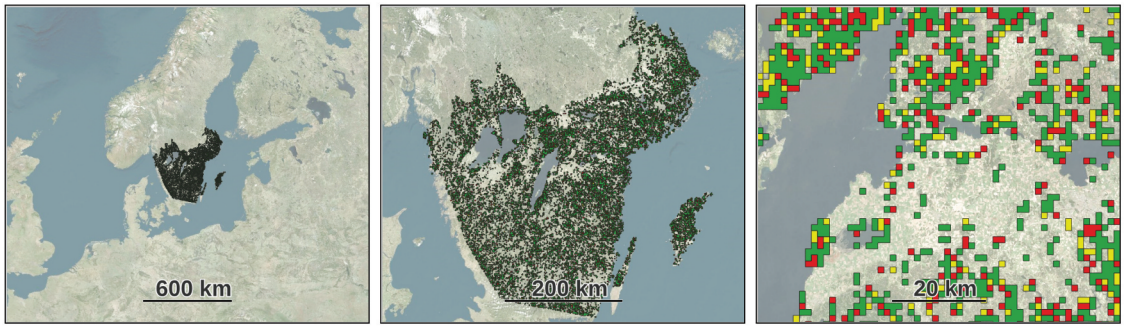

Figure 2. The study area in southern Sweden: continental view (left), overview (center), and zoom (right). Training, validation, and test regions are shown in (bright) green, yellow, and red, respectively. All the three images are centered in 58.4881 N, 15.1000 E; EPGS:3006 projection.

Table 1. Lidar data acquisition parameters. Cf. https://www.lantmateriet.se/globalassets/geodata/geodataprodukter/hojddata/quality_description_lidar.pdf for more details.

Point density	1–2 points per square meter
Flying altitude	ca. 3000 above ground
Scanning angle	maximum $\pm$ 20º
Side overlap	at least 20%
Footprint on the ground	$\leq$ 0.75m depending on flying altitude

considered only those annotated in the field or reviewed at the desk,[4] which were assigned a high natural value.[5] 40% of the polygons contain taiga (habitat code 9010) which could be both coniferous, deciduous, and mixed forest. The rest of the habitat is mostly coniferous-dominated forest. In *Storskogsbruket* approximately 14% of the polygons are dominated by deciduous forest and the remainder is dominated by coniferous forest. In *SksBorealSyd* approximately 43% of the polygons are predominantly deciduous forest and the rest is mainly coniferous forest.

For negative examples, i.e. forests with low naturalness, we consider the following sources: (1) *Hyggen1990–2000*, forests harvested between 1990 and 2000; (2) *Pskog30till80*, forest stands between 30 and 80 years old; and (3) *BestandEjNaturvarden*, older forest stands that have been assessed with low natural value by forestry companies. These areas represent a range of forests with different disturbance levels that compromise the naturalness value. Areas defined in *Hyggen1990–2000* contain young forests, mostly coniferous-dominated, detected by a yearly change detection analysis from satellite images conducted by the Swedish Forestry Agency.

Training-validation-test split

We divided the data in training, validation, and test sets with a process based on geography. First, we defined a uniform grid of 1.28 $\times$ 1.28 km squares in the study area. Then, we discarded squares without any labeled polygon. Finally, we randomly selected 64% of these squares to build the training set, 16% to the validation set, and 20% to the test set. In terms of size, the total surface areas are 29,409.28 km^2, 7353.14 km^2, and 9265.15 km^2 for the training, validation, and test sets, respectively. These regions are shown in Figure 2. Note that whenever a polygon belongs to two or more regions, we split the polygon accordingly. So, for example, if part of a polygon lies in the training region and the other part lies in the validation region then we split it into two polygons: one for the training set and the other for the validation set. Polygons with an area less than 0.01 km^2 have been discarded because they are too small for their naturalness to be reliably assessed. In addition, we discarded polygons where the CHM is not completely available and those with no trees, i.e. where all the values in the CHM are less than 4 m (see Sect. 3.1 for an explanation of the choice of this value).

The main characteristics of the three sets thus obtained are listed in Table 2. Note that, since the data sources are not perfect and surveys have been made independently of each other, there are overlaps between polygons in the same set and even some (rare) areas that belong both to low and high naturalness polygons. However, due to the splitting procedure, there are no overlaps among data sets. In other words, there are some geographical areas that belong to more than one polygon in the training set, for example, but there are no areas that belong both to a polygon in the training set and a polygon in the validation (or test) set.

Figure 3 shows the distribution of areas of the polygons in each split, overlaid by the naturalness (high or low) of the forest. As can be seen in the figure, the vast majority of areas are less than 0.1 km^2 (10 hectares).

Method

The proposed method is based on the characterization of the forests using human-understandable features

[4]Attribute *Kartering* (="mapping") $\in$ [2, 3, 4]
[5]Attribute *Naturtypss* (="natural type") $\in$ [1, 2]

Table 2. Summary statistics of the training, validation, and test sets.

	Number of polygons			Total area (km^2)		
Split	Total	High nat.	Low nat.	Total*	High nat.	Low nat.
Training	49,405	18,303	31,102	1,911	875	1,041
Validation	13,239	5,020	8,219	496	239	258
Test	16,516	6,416	10,100	592	276	317

*The total area slightly differs from the sum of the high and low naturalness areas because of overlaps.

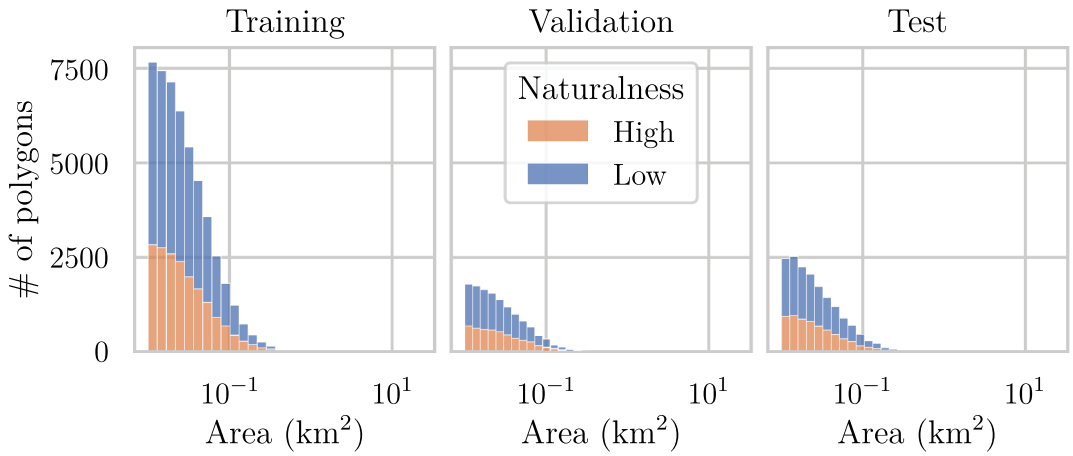

Figure 3. Areas of the forests considered in training, validation, and test sets.

related to tree heights and the spatial distributions of trees. In other words, given a region of the CHM defined by a polygon, we compute a set of values that are descriptive of the forest. Then, from this set of values (i.e. the features) we predict the probability of the region to be a highly natural forest, by applying our classifier. The classifier is trained using the dataset described in Section 2, containing positive and negative examples of high natural forests.

Feature extraction

For any given polygon (see Section 2) several features are computed for the pixels in the CHM raster that fall within the polygon, which defines the region of interest. In this process, we define a pixel as a tree only if its value is equal to or greater than a threshold h_{min}. Here, we have set this threshold as $h_{\text{min}} = 4$ m. Note that for the type of forests in the study area, trees lower than 4 m are very young. See Figure 4 for an example of tree-labeled pixels.

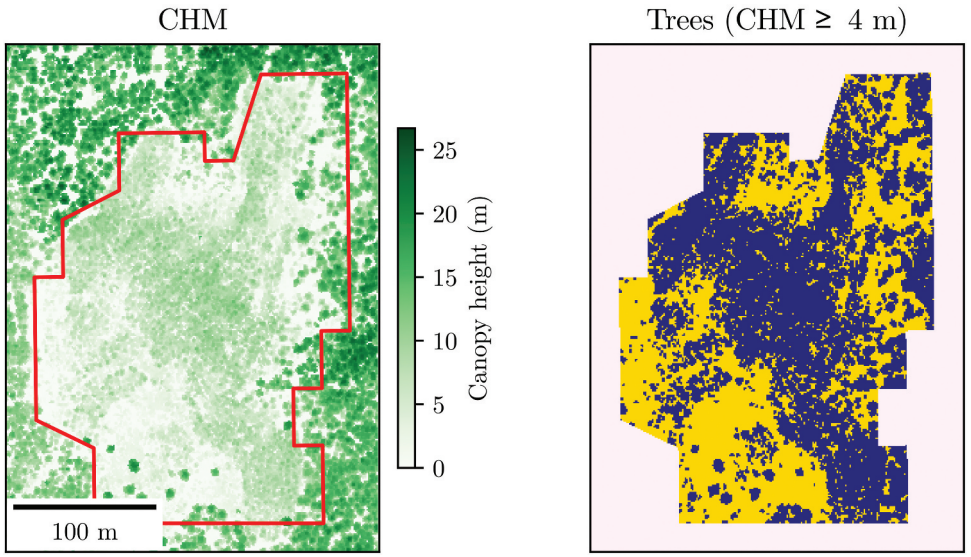

Figure 4. Example of tree density (TD = 0.574). The image on the left shows the CHM (shades of green) and the bounds of the region of interest (in red). The image on the right shows the area covered by trees (dark blue) and the remaining area (yellow) inside the region of interest, as well as the area outside the region of interest (light pink).

In addition, we label some pixels as *treetops*, so that we can distinguish single trees in the CHM. Different methods have been proposed in the literature to identify single trees and treetops (Ahmadi et al., 2022; Franceschi et al., 2018). Here, we identify treetops as a subset of the local maxima in the CHM raster. In particular, to compute the treetops, we apply a maximum filter with a window size of w_s to the CHM image and we consider the pixels where all the following conditions are met: (i) value of the pixel in the filtered image is equal to the original; (ii) the value is equal to or greater than $h_{\min}$; (iii) the pixels are inside the region of interest. Then, to avoid counting contiguous pixels with the same height as multiple treetops, we keep a single pixel for each region with contiguous maxima (considering 8-connectivity): the chosen pixel is the top-left-most. Note that w_s can be set as a parameter and it can be interpreted as $w_s = 2d_{\min} + 1$, where $d_{\min}$ corresponds to the minimal allowed distance separating treetops. Here, we have set $d_{\min} = 2m$. This value has been chosen by visual inspection of the resulting treetops on some areas considering the CHM and orthophotos. Note that the resolution of the CHM is 1 m, so it is not possible to select a lower value than 1 m for $d_{\min}$. See Figure 5 for an example of treetop identification. The features can then be computed, as will be described next.

Tree density (TD)

This feature measures the proportion of a polygon that is covered by trees. It is computed as

$$\mathrm{TD} = \frac{|H'|}{|H|}, \tag{1}$$

where H is the set of values of the region of interest of the CHM, H' is the set of values fulfilling the minimum height condition, i.e.

$$H' = \{h \in H \,|\, h \geq h_{min}\}, \tag{2}$$

and the $|\cdot|$ notation denotes the cardinality of a set, i.e. the number of elements. An example is shown in Figure 4.

Tree height mean (THM)

This feature measures the average value of the CHM, considering only the pixels that represent trees, i.e. the average value over the elements of the set H'. Over the entire training set (all polygons) the mean of THM is equal to 12.4 m.

Tree height variation (THV)

As its name implies, this feature measures the variation in canopy height over the region of interest. It is computed as the relative standard deviation (or coefficient of variation) of the value of the pixels in the CHM, again considering only pixels with a value greater than h_{min}. So, given H' as defined in Equation 2, we define

$$\mathrm{THV} = \frac{\sigma(H')}{\mathrm{THM}} \tag{3}$$

where $\sigma(H')$ denotes the standard deviation over the elements of the set H'.

Treetop density (TTD)

The treetop density measures the density of individual trees. It is computed from the treetop list, dividing the number of treetops inside the region of interest by its area:

$$\mathrm{TTD} = \frac{|T|}{A}, \tag{4}$$

where T is the set of treetop heights inside the region of interest and A is its area. For this feature, we measure the area in hectares, as is standard practice in forestry.

Treetop height mean (TTHM)

This feature measures the average value of the treetop heights in the region of interest, i.e. the average value over the set T. Over the entire training set (all polygons) the mean of TTHM is equal to 14.9 m.

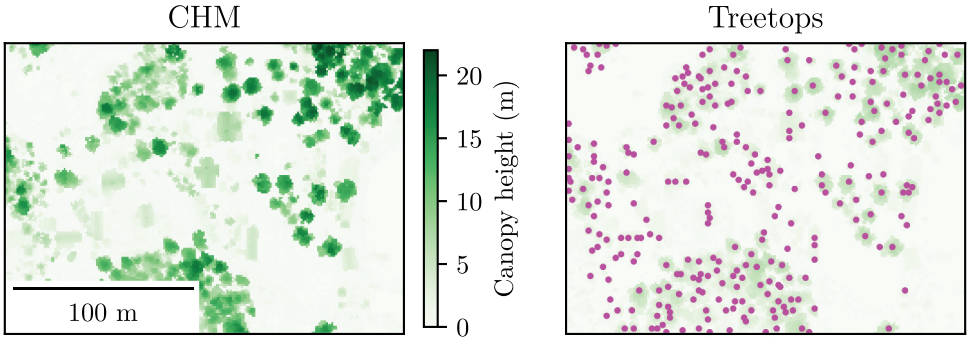

Figure 5. Example of treetop identification. The image on the left shows the CHM. The image on the right shows the treetops (in magenta) identified in the same area.

Treetop height variation (TTHV)

This feature measures the variability of the treetop heights over the region of interest. It is computed as the relative standard deviation of the treetop heights for trees that lie inside the region of interest, i.e. as

$$\mathrm{TTHV} = \frac{\sigma(T)}{\mathrm{TTHM}}. \tag{5}$$

Edge-like pixels (ELP)

This feature considers the texture of the CHD and measures the proportion of pixels, within the region of interest, that are edge-like. The intuition is that the human intervention in low naturalness forests results in more regular patterns in the CHD. So, the proportion of edge-like pixels is expected to be smaller in low naturalness forests because regular pattern has shorter edges than irregular patterns. To identify edge-like pixels we use the local binary patterns (LBP) technique, which is a well-known method in computer vision (Ojala et al., 1996). In summary, the LBP method assigns to each pixel a value that depends on the difference with its neighbour pixels. Two parameters can be set for defining the neighbor pixels, which lay in a circle centered in the considered pixel: the radius r of the circle and the number p of (circularly symmetric) points. In our experiments we set $p = 24$ and $r = 3$. The value of the ELP feature is then computed as the proportion of edge-like pixels, i.e. pixels with an LBP value in the range $p/2 \pm (r-1)$, which corresponds to the range [10, 14] with our choice of parameters. Figure 6 visualizes the LBP method for the computation of the ELP feature in two examples.

Treetop spatial distribution (TTSD)

This feature measures the regularity of the spatial distribution of treetops. The intuition is similar to the previous feature: Human intervention results in a more regular spatial distribution of the trees w.r.t. natural grown forests. In order to measure the regularity, we project the treetop locations onto binned lines (in various directions α) and we measure the fraction f_w of bins that contain a (projected) treetop. We consider a variable number of bins with a fixed width equal to w, in the range of treetop projections. We use n lines with different directions D_n, equally spaced in the range $0 - 180$. The TTSD feature is then defined as:

$$\mathrm{TTSD} = \min_{\alpha \in D_n} f_w(\alpha) \tag{6}$$

where the minimum f_w is found at the direction that maximizes the regular spacing of the trees. Note that the width w of the bins and the number of lines n can be seen as tunable parameters, for which we have here used the values $w = 1$ and $n = 100$. Figure 7 illustrates the intuition behind this feature.

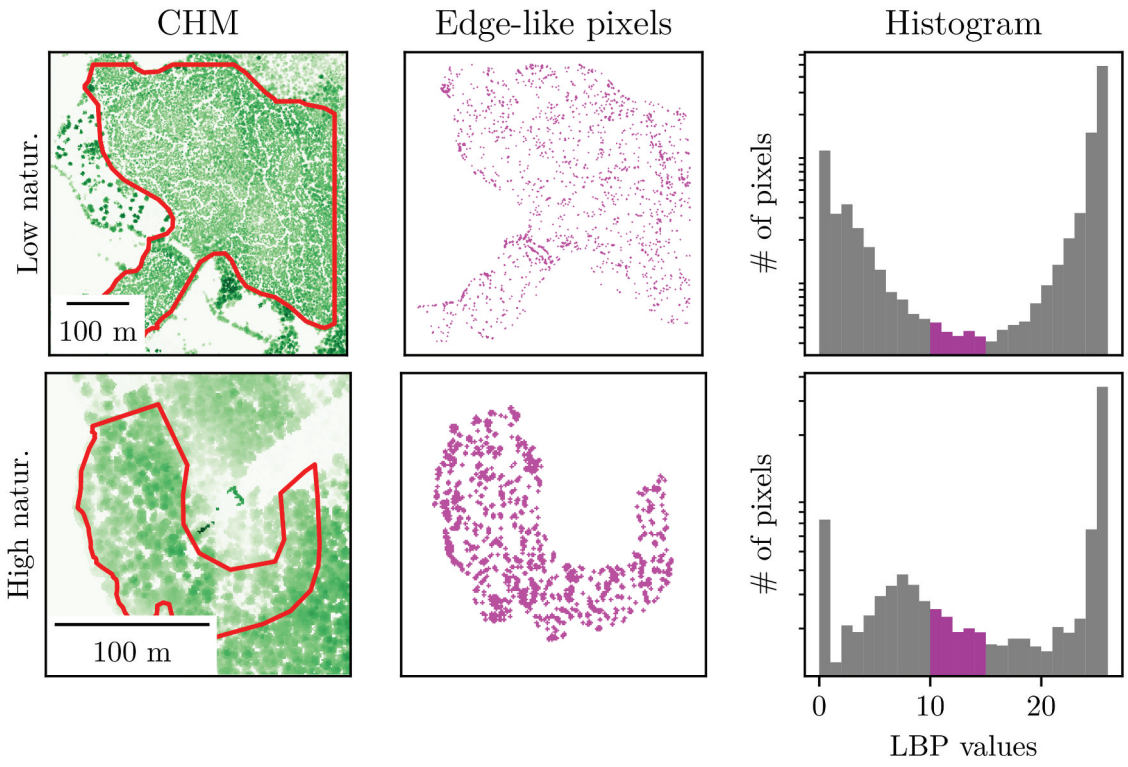

Figure 6. Examples of the ELP feature for a low naturalness forest and a high naturalness one, shown in the top and in the bottom row, respectively. The edge-like pixels are shown in the middle column and their counts are highlighted in magenta in the histograms in the right column.

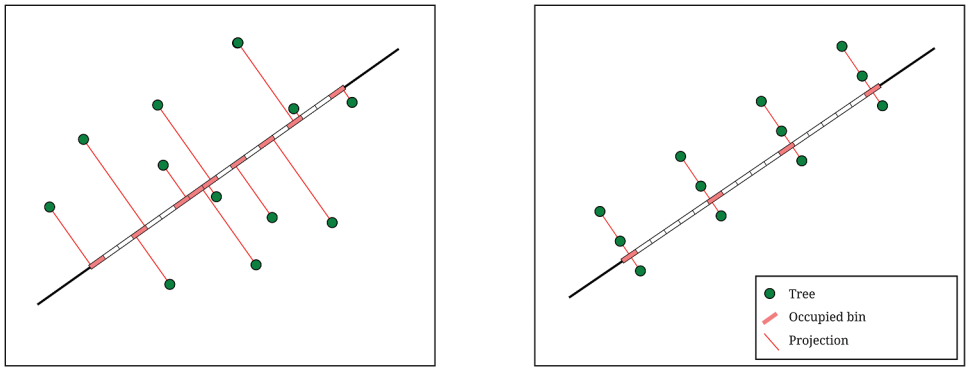

Figure 7. Visualization of the TTSD feature. On the left side treetops are distributed randomly and their projection on the line results in 9 of 18 bins being occupied. On the right side, the spatial distribution of the treetops is regular and it results in only 4 occupied bins.

Machine learning models

Considering the annotated polygons from field inventories (see Section 2.2) and their featurization as described above, we obtain a tabular data set in which forests (rows) are described by a set of features (columns) plus a binary label for naturalness, i.e. *low* and *high*, encoded as 0 and 1, respectively.

This results in a classification problem with tabular data, and many existing ML methods can be applied. Since we are primarily interested in interpretable models, we considered perceptrons, logistic regression, and decision trees.

Perceptron

A perceptron is a linear, binary classifier, in which a single hyperplane separates the two classes (McCulloch & Pitts, 1943). Thus, the output from the perceptron can be expressed as

$$y = \Theta\left(\sum_{i=0}^{n} w_i x_i\right), \tag{7}$$

where Θ denotes the Heaviside step function, w_i are the tunable weights, x_i, $i = 1, \ldots n$ are the values of the n features, and $x_0 \equiv 1$ (that multiplies the bias weight w_0), is introduced for convenience of notation.

The perceptron is trained using an iterative algorithm that is guaranteed to converge if the two classes are fully separable. If the classes are *not* fully separable, as is normally the case in applied problems, the algorithm does not formally converge but, provided that the number of data points is finite, the weights will remain bounded.

The training algorithm starts by setting the weights to zero (or random, small values). Next, for each training epoch, the algorithm runs through all m training examples, after first placing them in random order. For each training example k ($k = 1, \ldots, m$) the weights are then updated as follows

$$w_i \leftarrow w_i + \eta(d_k - y_k)x_{i,k}, \;\; i = 1, \ldots n \tag{8}$$

where d_k and y_k are the desired and actual outputs (class labels, in this case) for the k^{th} training example, η is the learning rate, and $x_{i,k}$ are the feature values for the same training example. The procedure is then repeated, epoch by epoch, until a termination criterion is reached, for example a certain number of epochs.

Logistic regression

Logistic regression is a linear classification method that is widely used in various fields. It is a statistical model that predicts the probability of a binary outcome, in this case high and low naturalness. The logistic regression model is characterized by its use of the logistic function, which is a sigmoid function that transforms a linear combination of input variables into a probability value between 0 and 1. This function is used to model the relationship between the input variables (the features) and the probability of the binary outcome, formally:

$$y = \Theta\big(\sigma(\beta^{\mathrm{T}} x) - 0.5\big) \tag{9}$$

where x is the input vector, Θ denotes the Heaviside step function, β is the vector of coefficients, and $\sigma(t) = (1 + e^{-t})^{-1}$ is the logistic function. Note that we rescale the feature values with a min-max normalization before applying the logistic regression, so that all components of x have the same range. In our experiments we use the scikit-learn[6] default implementation.

Decision trees

Decision trees are a non-parametric machine learning method widely used for classification problems.

[6]See https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html version 1.3.2.

The model predicts the value of a target variable, the naturalness in our case, by learning simple decision rules inferred from the data features. In our experiments we use the scikit-learn[7] default implementation, which relies on the CART algorithm (Breiman et al., 1984).

Prediction of the probability of high naturalness

The three ML models can provide a probability associated to the classification task, indicating the model's uncertainty about the predicted class. Hence, the probability of being in the high naturalness class can be used to measure forest naturalness on a continuous scale, even if the training of the model considers only binary labels, i.e. positive and negative examples. Moreover, this information can be valuable for decision-making, as it allows users to consider the uncertainty associated with the model's predictions.

In the case of logistic regression, the output of the logistic function is a value between zero and one, it can be interpreted as the probability p_1 of the observation to belong to Class 1 (high naturalness). Note that, since our classification problem is a binary one, the probability of the observation to belong to Class 0 (low naturalness) is simply $p_0 = 1 - p_1$.

In the case of the perceptron, the probability can be obtained by applying a sigmoid function (e.g. Gauss error function) to the results of the weighted sum (see Equation (7)).

In the case of the decision trees, the probability can be obtained by limiting the depth of the tree and considering the fraction of training samples of the class in each leaf.

Performance metrics

We evaluate the proposed approach using holdout validation: We use the training and validation set to build the models, and the test set for the evaluations.

We consider standard metrics commonly used for binary classification problems, namely accuracy, precision, recall, and the F1 score. Note that our data set is slightly unbalanced, since the low naturalness class represents more than half (about 62% in all the splits) of the data, see Table 2. For this reason, we also considered the *balanced accuracy*.

Using these metrics, in what follows, we evaluate the models' performance by examining the impact of varying the ML classifier, the effect of polygon area, the influence of selecting different feature subsets, and the confidence of the predictions.

Results

Feature importance

We first analyse the discriminatory power of each feature, individually. For each feature, we compute the *optimal discrimination threshold* such that if we classify all the polygons below that threshold as low (or high) naturalness and vice-versa, the accuracy over the training set is maximized. Results (see Table 3) show that features can be clearly divided into two categories: TTD, ELP, THM, and TTHM have a strong discriminatory power (accuracy in the range 84–85%); while TTSD, TD, THV, and TTHV have less discriminatory power (accuracy less than 67%).

Note that, compared to low-naturalness forests, those with high naturalness are characterized by larger values of tree and treetop height means, edge-like pixels, and randomness of treetop spatial distribution (THM, TTHM, ELP, TTSD). By contrast, low-naturalness forests are characterized by larger values of treetop density (TTD). High- and low-naturalness forests have similar values for tree and treetop height variations, and tree density (THV, TTHV, TD).

The value of the threshold is shown in the third column. The threshold is computed considering the accuracy over training set (fourth column). The corresponding accuracy over the validation set is also shown (fifth column).

Figure 8 shows a pair plot of the four most important features, individually and in pairs. Note that for

Table 3. Optimal discrimination threshold for each feature.

Feature descr.	Feature	Threshold	Acc. (train.)	Acc. (valid.)
Treetop density	TTD	214.2	0.856	0.857
Edge-like pixels	ELP	0.061	0.846	0.847
Tree height mean	THM	128.3	0.845	0.843
Treetop height mean	TTHM	155.8	0.845	0.844
Treetop spatial distr.	TTSD	0.099	0.666	0.666
Tree density	TD	0.165	0.640	0.634
Treetop height variation	TTHV	0.057	0.631	0.622
Tree height variation	THV	0.752	0.630	0.621

[7]See https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html version 1.3.2.

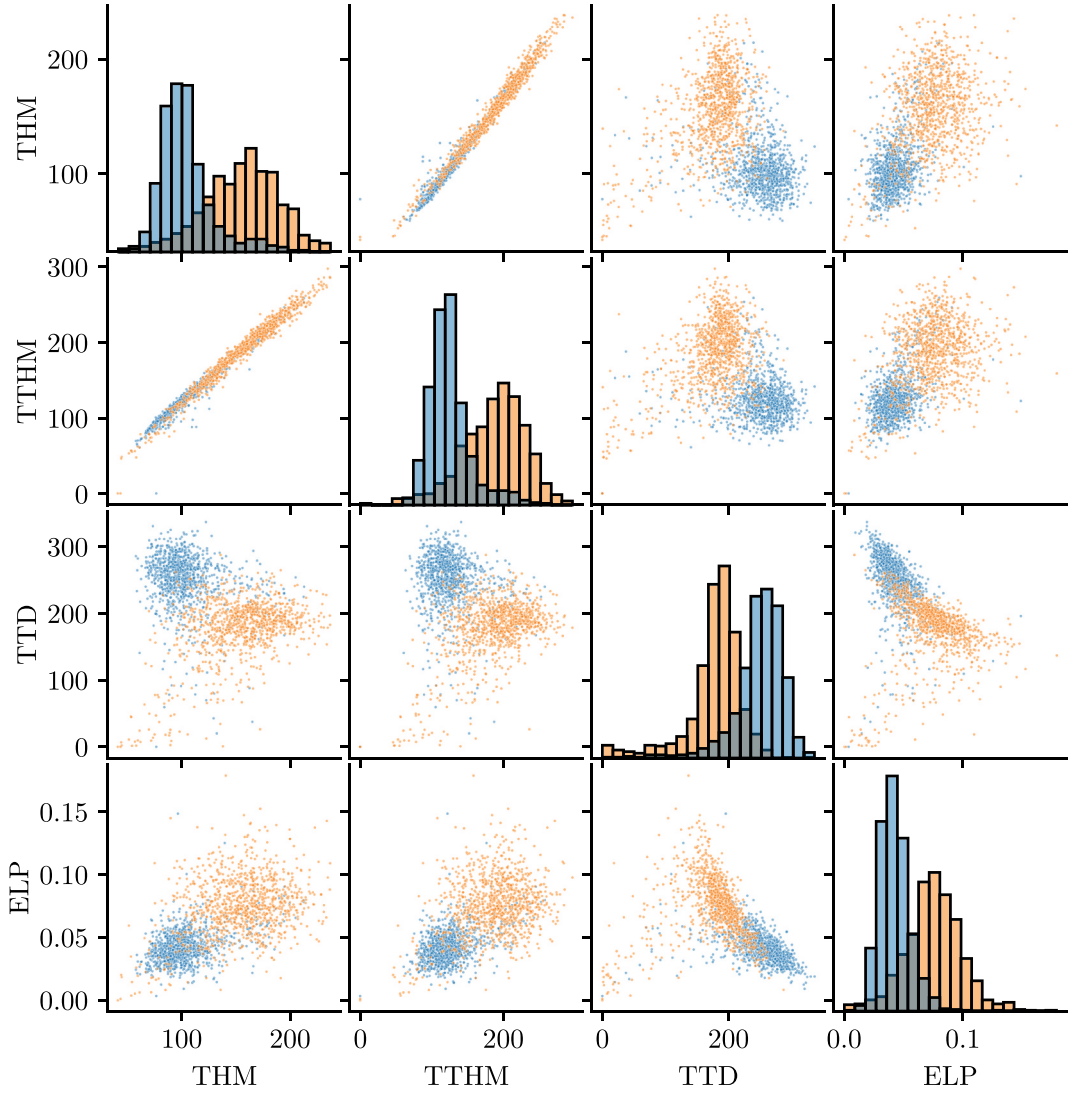

Figure 8. Pair plot of the four most important features (see Table 3). Histograms for the individual features are shown on the diagonal, while the scatter plots for each pair of features are shown on the other charts. Orange points and bars represent high naturalness, while blue points and bars represent low naturalness.

visual clarity, a random balanced subset (2,000 data points) of the training set has been considered when generating the figure. It is interesting also to note the good separation between the data points (for each class) for some of the pairs, especially in the case of the TTHM-TTD pair.

Model comparison

Overall, the three models have similar performances: about 0.9% accuracy on the test set for all the considered models. In detail, the perceptron scores 0.897%, logistic regression scores 0.898%, and the decision trees score 0.901%. In what follows, we investigate how the area of the polygon and the tree density affect the performance of the models. The results obtained for the different metrics (accuracy, precision, recall, F1, and balanced accuracy) are similar and follow the same pattern. Thus, for the sake of clarity, we present only those related to accuracy. Note that all the accuracy scores reported in what follows refer to the test set.

Accuracy vs. area of the polygon

The full data set used here involves a set of forested areas that differ widely in size, ranging from 0.01 km^2 (1 hectare) to 22.6 km^2, even though the vast majority of areas are under 0.1 km^2 in size (cf. Figure 3). It could be expected that larger areas would be easier to classify reliably than smaller ones, just because they contain more data. In order to investigate this hypothesis, the areas were sorted in increasing order, after which a given percentile range, based on area, was extracted from the training set. Next, the smallest and largest areas ($a_{\min}$ and $a_{\max}$) found in the extracted training subset were logged. Then, all areas falling in that range were extracted from the validation and test sets. Thus, this procedure resulted in three subsets, with the same range of areas. Optimization runs where then carried out, using the results over the extracted validation

subset to determine when to stop the runs, after which the results over the corresponding test subset were measured. Those results are shown in Table 4. As can be seen in the table, the accuracy increases as the size of the areas grows, reaching a value of 0.9497 for the largest 10%. Note that the extent of the area of interest is an exogenous variable, meaning that it is not incorporated as an input to the model.

In each case, the results pertain to the model with highest validation accuracy. The second column shows the area of the smallest polygon ($a_{\min}$) included in the training subset. The upper limit is 22.61 km^2, i.e. the maximum size, in all cases.

Accuracy vs. feature set

For the sake of the interpretability of the system, we are interested in reducing the complexity of the model, while not losing too much in accuracy. For this reason we report here the accuracy of the three machine learning models, considering subsets of the available features. In particular, we consider the first six features, which are arguably the most easy to understand since they involve standard operations (average and coefficient of variation), and are easily visualizable, making them comprehensible to most observers, in our assessment. We then consider the four most important features in terms of discriminatory power (see Section 4.1), namely THM, TTD, TTHM, and ELP. Finally, we consider only two features among those, discarding the less interpretable ELP as well as THM that is highly correlated with TTHM. Results are show in Table 5.

The second column shows the number (m) of features in the subset.

Accuracy vs. prediction confidence

The predictions of the proposed ML models are associated with a probability (see Section 3.3). The *confidence score* of the classification is defined as the probability of predicted class. Note that in binary classification the confidence score can take values from 0.5 to 1: values close to 0.5 when the outcome of the model is very uncertain; values close to 1 when there is no doubt. We naturally expect the higher the confidence, the higher the accuracy. Table 6 shows the accuracy of the complete (all eight features) logistic regression model as a function of the confidence score. Note that the model is confident about the prediction most of the times: In fact, for about 67% of the observations in the test set, the confidence score is greater than 0.9. In these cases, the performance of the model is the highest, namely 96.7%. Notably the accuracy matches the range of the confidence score, as expected.

The Support columns report the number of observations in absolute and relative terms.

Discussion

Our study presents an interpretable AI method for assessing the naturalness of forests from the CHM. The method is composed of a feature extraction process and an ML model. In our view, for high-stakes decision-making, a classifier should ideally be transparent and human-interpretable (as are our classifiers), in order to offer valuable insight, beyond merely a high probability of correct classification which is all that black box models can offer.

Our results show that the performance of the three considered ML models are similar: An accuracy of 0.89–0.90 over the test set. What is really crucial in the proposed method is the feature extraction process. This process involves selecting and transforming the raster input data into a set of values (i.e. the eight features) that is suitable for the ML models. In our

Table 4. Accuracy over the test set obtained when training over different subsets (percentile ranges based on area) of the training set, using all eight features.

Percentile range	$a_{\min}$ (km^2)	Perceptron	Log. regr.	D.T.
0 – 100	0.010	0.897	0.898	0.901
10 – 100	0.012	0.901	0.902	0.903
20 – 100	0.014	0.904	0.906	0.912
30 – 100	0.017	0.909	0.911	0.921
40 – 100	0.020	0.914	0.916	0.919
50 – 100	0.024	0.922	0.921	0.927
60 – 100	0.029	0.924	0.926	0.936
70 – 100	0.037	0.932	0.932	0.937
80 – 100	0.050	0.941	0.937	0.942
90 – 100	0.077	0.950	0.944	0.944

Table 5. Accuracy over the test set considering selected subsets of the available features.

Feature set	m	Perceptron	Log. regr.	D.T.
All	8	0.897	0.898	0.901
TD, THM, THV, TTD, TTHM, TTHV	6	0.895	0.897	0.895
THM, TTD, TTHM, ELP	4	0.889	0.889	0.871
TTD, TTHM	2	0.888	0.885	0.852

Table 6. Accuracy over the test set considering the confidence of the prediction for the logistic regression model with all the features.

Confidence score range	Support	Support %	Accuracy
0.5 – 0.6	913	5.5%	0.560
0.6 – 0.7	983	6.0%	0.650
0.7 – 0.8	1,374	8.3%	0.782
0.8 – 0.9	2,202	13.3%	0.873
0.9 – 1.0	11,044	66.9%	0.967

study, we used a combination of features extracted from the CHM, considering a region of interest delimited by a labeled polygon. The features include metrics such as tree height, canopy cover, and density. We found particularly important the procedure that identifies the treetops, which is the basis for two of the features with the most discriminatory power, namely Treetop density (TTD) and Treetop height mean (TTHM). This procedure leverages the high resolution of the CHM (i.e. 1), where a single tree typically spans multiple pixels. Through visual inspection, we observed that the accuracy of treetop detection varies with forest density. In dense forests, multiple trees are sometimes mistakenly identified as a single treetop, leading to a computed treetop density that is lower than the actual density. However, this discrepancy does not substantially affect model performance, as the computed treetop density remains higher than that of sparse forests.

It should be noted that with the proposed methodology we cannot claim that it is possible to assess the naturalness of every forest. What our model really does is to distinguish between the classes (low or high naturalness) of the data source, as we have defined them. Nevertheless, we have considered data sources for labeling that include a range of forests with different levels of human interference that can compromise the naturalness value. So, we believe that our dataset is representative of most types of forest in the region of study, i.e. Southern Sweden. To apply this method in other regions of the world with different types of forest, it will be necessary to re-train the model with new, meaningful examples of forests typical of those regions.

An additional strength of this model is its flexibility to incorporate new features, such as those derived from ALS point clouds – volume, ground area, ground area-weighted average height, and ground area-weighted average diameter, for instance. The model architecture readily accommodates these additions by integrating them directly into the machine learning classifiers. The classifiers we employ – perceptron, logistic regression, and decision trees – are well-suited for handling high-dimensional feature spaces, allowing us to experiment with a broader range of data inputs without compromising performance. While expanding the feature set is part of our planned future work, in this study, we intentionally limited the input to the CHM to ensure that the model remains applicable in situations where only CHM data are available. This is exemplified by a recent data set produced by Meta (Tolan et al., 2024), which provides a global high-resolution CHM derived from RGB imagery. Although this dataset has accuracy limitations (Moudrý et al., 2024), its global availability could significantly enhance the applicability and adoption of our model. By relying solely on CHM data, our model can leverage these widely accessible resources, offering a scalable solution for forest assessments even in regions lacking ALS data.

Regarding the interpretability of the system, it should be noted that the degree of interpretability is, to a great extent, a subjective one. In the case considered here, two factors are involved, namely the interpretability of the individual features *and* the interpretability of the model that uses them. Starting with the latter, and in contrast to black box models such as DNNs, the models presented here are fully transparent, generating their output as a linear combination of, or a set of rules based on, a small set of features. As for the features, the first six (TD, THM, THV, TTD, TTHM, and TTHV) involve standard operations (means and variances), and are easily visualizable, making them transparent to most observers, in our assessment. The remaining two features, ELP and TTSD, are a bit more complex, yet within reach of interpretation for an observer with some knowledge of image interpretation. Importantly, using only the six first features, one obtains a test set accuracy in the range 0.895 to 0.897 (depending on the model). Moreover, by further simplifying the model and considering just two of those features (namely, TTD and TTHM), the performance decreases only slightly to an accuracy of 0.88 for the linear models, which corresponds to a 1% loss (see Table 5).

In situations where, say, a forest owner is prevented by a forest authority from exploiting a particular area, it is a large advantage for every stakeholder to have a transparent model of the kind considered here. In fact, one can identify the exact contribution of each feature to the final decision. Note that this is different from the post-hoc explanations methods for black box models, whose explanations are typically partial, often contradictory in cases where several explanation methods are considered (Krishna et al., 2022), and sometimes simply unreliable (Slack et al., 2020).

Even though the explanation of a linear model is exact and readily understandable by a mathematically inclined observer, a lay person may require a *verbal* explanation. Thus, in future work, we aim to combine our classifier with an equally transparent dialogue system (DAISY) (Wahde & Virgolin, 2023). This combination will result in a system that, when prompted, can provide a verbal explanation, an ability that is an integral part of the DAISY dialogue manager.

Conclusion

This paper presents an interpretable methodology for automated assessment of forest naturalness. The method uses eight features extracted from Canopy height model data and applies several different interpretable machine learning classifiers, namely perceptrons, logistic regression, and decision trees. The high-resolution data make it possible to identify individual trees and treetops, facilitating comprehensive feature extraction.

We find that the accuracy of the method is largely independent of the specific machine learning model employed, but is influenced by the procedures used to extract the features from the CHM (treetop identification, in particular), the feature selection, the area of the region of interest, and the prediction confidence score. In the study area encompassing coniferous and deciduous forests in Southern Sweden, the proposed method achieved an overall accuracy of approximately 90%.

The interpretability of the method is a key feature that enhances transparency and facilitates understanding by various stakeholders, including environmental experts, policy makers, forestry companies, and engineers involved in method development and maintenance. Moreover, the interpretable nature of our approach also allows for deeper comprehension of the factors contributing to a forest's naturalness classification, empowering stakeholders to make informed decisions related to forest management and conservation efforts.

Acknowledgments

We thank our colleagues from Skogsstyrelsen, especially Alice Högström, Liselott Nilsson, and Johan Häggmark (Combitech AB), who provided insight and expertise that greatly assisted the research, although the authors take full responsibility for all interpretations and conclusions of this paper.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research was partially supported by VINNOVA, project 2022-01702.

Data availability statement

The Canopy Height Model is provided by Skogsstyrelsen and it is publicly available at https://www.skogsstyrelsen.se/sjalvservice/karttjanster/skogliga-grunddata/, *trädhöjd* layer.

References

Ahmadi, S. A., Ghorbanian, A., Golparvar, F., Mohammadzadeh, A., & Jamali, S. (2022, December). Individual tree detection from unmanned aerial vehicle (UAV) derived point cloud data in a mixed broadleaf forest using hierarchical graph approach. *European Journal of Remote Sensing*, *55*(1), 520–539.

Barrette, M., Dumais, D., Auger, I., Boucher, Y., Bouchard, M., & Bouliane, J. (2020, December). Naturalness assessment performed using forestry maps to validate forest management sustainability. *Ecological Indicators*, *119*, 106832. https://doi.org/10.1016/j.ecolind.2020.106832

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Wadsworth & Brooks/Cole Advanced Books & Software.

Brugere, L., Kwon, Y., Frazier, A. E., & Kedron, P. (2023, July). Improved prediction of tree species richness and interpretability of environmental drivers using a machine learning approach. *Forest Ecology & Management*, *539*, 120972. https://doi.org/10.1016/j.foreco.2023.120972

Buchelt, A., Adrowitzer, A., Kieseberg, P., Gollob, C., Nothdurft, A., Eresheim, S., Tschiatschek S., Stampfer K., & Holzinger, A. (2024, January). Exploring artificial intelligence for applications of drones in forest ecology and management. *Forest Ecology & Management*, *551*, 121530. https://doi.org/10.1016/j.foreco.2023.121530

Chen, B., Wu, Z., Wang, J., Dong, J., Guan, L., Chen, J., Yang K., & Xie, G. (2015, April). Spatio-temporal prediction of leaf area index of rubber plantation using HJ-1A/1B CCD images and recurrent neural network. *Isprs Journal of Photogrammetry & Remote Sensing*, *102*, 148–160. https://doi.org/10.1016/j.isprsjprs.2014.12.011

Chirici, G., McRoberts, R. E., Fattorini, L., Mura, M., & Marchetti, M. (2016, March). Comparing echo-based and canopy height model-based metrics for enhancing estimation of forest aboveground biomass in a model-assisted framework. *Remote Sensing of Environment*, *174*, 1–9. https://doi.org/10.1016/j.rse.2015.11.010

Dechesne, C., Mallet, C., Le Bris, A., Gouet, V., & Hervieu, A. (2016, June). Forest stand segmentation using airborne lidar data and very high resolution multispectral imagery. *ISPRS - International Archives of the Photogrammetry Remote Sensing and Spatial Information Sciences*, *XLI-B3*, 207–214. https://doi.org/10.5194/isprsarchives-XLI-B3-207-2016

Drössler, L. (2010, October). Tree species mixtures - a common feature of southern Swedish forests. *Forestry: An International Journal of Forest Research*, *83*(4), 433–441.

Fang, N., Yao, L., Wu, D., Zheng, X., & Luo, S. (2023, August). Assessment of forest ecological function levels based on multi-source data and machine learning. *Forests*, *14*(8), 1630. https://doi.org/10.3390/f14081630

Franceschi, S., Antonello, A., Floreancig, V., Gianelle, D., Comiti, F., & Tonon, G. (2018, January). Identifying treetops from aerial laser scanning data with particle swarming optimization. *European Journal of Remote Sensing*, *51*(1), 945–964.

Frego, K. A. (2007, April). Bryophytes as potential indicators of forest integrity. *Forest Ecology & Management*, *242*(1), 65–75. https://doi.org/10.1016/j.foreco.2007.01.030

Hong, Y., Xu, J., Wu, C., Pang, Y., Zhang, S., Chen, D., & Yang, B. (2023, November). Combining multisource data and machine learning approaches for multiscale

estimation of forest biomass. *Forests*, *14*(11), 2248. https://doi.org/10.3390/f14112248

Jiang, S., Sweet, L.-B., Blougouras, G., Brenning, A., Li, W., Reichstein, M., Denzler J., Shangguan W., Yu G., Huang F., & Zscheischler, J. (2024). How interpretable machine learning can benefit process understanding in the geosciences. *Earth's Future*, *12*(7), e2024EF004540. https://doi.org/10.1029/2024EF004540

Krishna, S., Han, T., Gu, A., Pombra, J., Jabbari, S., Wu, S., & Lakkaraju, H. (2022). The disagreement problem in explainable machine learning: A practitioner's perspective. https://doi.org/10.48550/arXiv.2202.01602. *arXiv preprint arXiv: 2202.01602*.

Lechner, A. M., Foody, G. M., & Boyd, D. S. (2020, May). Applications in remote sensing to forest ecology and management. *One Earth*, *2*(5), 405–412. https://doi.org/10.1016/j.oneear.2020.05.001

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5*(4), 115–133. https://doi.org/10.1007/BF02478259

McRoberts, R. E., Winter, S., Chirici, G., & LaPoint, E. (2012, June). Assessing forest naturalness. *Forest Science*, *58*(3), 294–309.

Moudrý, V., Cord, A. F., Gábor, L., Laurin, G. V., Barták, V., Gdulová, K., Gdulová, K., Malavasi, M., Rocchini, D., Stereńczak, K., Prošek, J., Klápště, P., & Wild, J. (2023). Vegetation structure derived from airborne laser scanning to assess species distribution and habitat suitability: The way forward. *Diversity & Distributions*, *29*(1), 39–50. https://doi.org/10.1111/ddi.13644

Moudrý, V., Gábor, L., Marselis, S., Pracná, P., Barták, V., Prošek, J., Navrátilová, B., Novotný, J., Potůčková, M., Gdulová, K., Crespo-Peremarch, P., & Wild, J. (2024). Comparison of three global canopy height maps and their applicability to biodiversity modeling: Accuracy issues revealed. *Ecosphere*, *15*(10), e70026.

Mustonen, J., Packalén, P., & Kangas, A. (2008, December). Automatic segmentation of forest stands using a canopy height model and aerial photography. *Scandinavian Journal of Forest Research*, *23*(6), 534–545.

Ojala, T., Pietikäinen, M., & Harwood, D. (1996, January). A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, *29*(1), 51–59.

Olofsson, K., & Holmgren, J. (2014, March). Forest stand delineation from lidar point-clouds using local maxima of the crown height model and region merging of the corresponding voronoi cells. *Remote Sensing Letters*, *5*(3), 268–276. https://doi.org/10.1080/2150704X.2014.900203

Ozdemir, I., & Donoghue, D. N. M. (2013, May). Modelling tree size diversity from airborne laser scanning using canopy height models with image texture measures. *Forest Ecology & Management*, *295*, 28–37. https://doi.org/10.1016/j.foreco.2012.12.044

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Saravi, S., Kalawsky, R., Joannou, D., Rivas Casado, M., Fu, G., & Meng, F. (2019, May). Use of artificial intelligence to improve resilience and preparedness against adverse flood events. *Water*, *11*(5), 973. https://doi.org/10.3390/w11050973

Senf, C., Mori, A. S., Müller, J., & Seidl, R. (2020, September). The response of canopy height diversity to natural disturbances in two temperate forest landscapes. *Landscape Ecology*, *35*(9), 2101–2112. https://doi.org/10.1007/s10980-020-01085-7

Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling lime and shap: Adversarial attacks on post hoc explanation methods. *Proceedings of the aaai/acm conference on ai, ethics, and society* (pp. 180–186).

Stoker, J. M. (2020). *Defining technology operational readiness for the 3D elevation program- a plan for investment, incubation, and adoption. Tech. Rep. No. 2020-1015*. US Geological Survey.

Tierney, G. L., Faber-Langendoen, D., Mitchell, B. R., Shriver, W. G., & Gibbs, J. P. (2009). Monitoring and evaluating the ecological integrity of forest ecosystems. *Frontiers in Ecology and the Environment*, *7*(6), 308–316. https://doi.org/10.1890/070176

Tolan, J., Yang, H.-I., Nosarzewski, B., Couairon, G., Vo, H. V., Brandt, J., Spore, J., Majumdar, S., Haziza, D., Vamaraju, J., Moutakanni, T., & Couprie, C. (2024, January). Very high resolution canopy height maps from RGB imagery using self-supervised vision transformer and convolutional decoder trained on aerial lidar. *Remote Sensing of Environment*, *300*, 113888. https://doi.org/10.1016/j.rse.2023.113888

Torresan, C., Strunk, J., Zald, H. S., Zhiqiang, Y., & Cohen, W. B. (2014, January). Comparing statistical techniques to classify the structure of mountain forest stands using chm-derived metrics in Trento province (Italy). *European Journal of Remote Sensing*, *47*(1), 75–94.

Wahde, M., & Virgolin, M. (2023). Daisy: An implementation of five core principles for transparent and accountable conversational AI. *International Journal of Human–Computer Interaction*, *39*(9), 1856–1873. https://doi.org/10.1080/10447318.2022.2081762

Zhang, C., Sargent, I., Pan, X., Li, H., Gardiner, A., Hare, J., & Atkinson, P. M. (2018, October). An object-based convolutional neural network (OCNN) for urban land use classification. *Remote Sensing of Environment*, *216*, 57–70. https://doi.org/10.1016/j.rse.2018.06.034

Zhang, L., & Zhang, L. (2022, June). Artificial intelligence for remote sensing data analysis: A review of challenges and opportunities. *IEEE Geoscience and Remote Sensing Magazine*, *10*(2), 270–294. https://doi.org/10.1109/MGRS.2022.3145854

Zhang, Z., Angelov, P., Soares, E., Longepe, N., & Mathieu, P. P. (2022, October). An interpretable deep semantic segmentation method for earth observation. *2022 IEEE 11th International Conference on Intelligent Systems (IS)* (pp. 1–8).