Deep learning methods for naturalness evaluation of forests based on canopy height model

Master's thesis in Complex Adaptive Systems

Andreas Bauner

DEPARTMENT OF MECHANICS AND MARITIME SCIENCES

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2024
www.chalmers.se

Master's Thesis in Complex Adaptive Systems

Deep learning methods for naturalness evaluation of forests based on canopy height model

Andreas Bauner

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Mechanics and Maritime Sciences
Division of Vehicle Engineering and Autonomous Systems
Chalmers University of Technology
Gothenburg, Sweden 2024

Deep learning methods for naturalness evaluation of forests based on canopy height model
Andreas Bauner

Examiner and Supervisor:
Marco L. Della Vedova, Dept. of Mechanics and Maritime Sciences

Master's Thesis 2024
Department of Mechanics and Maritime Sciences
Chalmers University of Technology
SE-412 96 Gothenburg
Sweden
Telephone +46 31 772 1000

Typeset in LaTeX
Gothenburg, Sweden 2024

Deep learning methods for naturalness evaluation of forests based on canopy height model

Andreas Bauner
Department of Mechanics and Maritime Sciences
Division of Vehicle Engineering and Autonomous Systems
Chalmers University of Technology

Abstract

Forest evaluation has historically been done through field surveys by experts from national forest agencies or from forestry companies. This is costly manual labor that consumes a lot of time. A solution could be to use remote sensing ecological data and automate the naturalness evaluation of forests with the use of computers. The aim of this thesis is to develop a machine learning model that could help automate naturalness evaluation of forests. The remote sensing data is in the form of a Canopy Height Model (CHM), that is height of trees obtained from airborne laser-scanning. Ground truth data for forest naturalness is given in the form of annotated, georeferenced polygons. The study area is limited to a $50 \times 50\,\mathrm{km}^2$ area north-east of the city of Jönköping in Sweden. After applying different processing steps on the data, it is then used for training a convolutional neural network, based on U-Nets, on this semantic segmentation task. The evaluation of the model shows good results, achieving an accuracy of 94.1% on the test set. This performance is competitive with currently used models for related tasks and shows the feasibility of using machine learning in the relatively new field of automated naturalness evaluation of forests.

Keywords: machine learning, artificial intelligence, U-Net, forests, canopy height model, remote sensing, semantic segmentation.

Acknowledgements

I would first like to thank my supervisor Marco L. Della Vedova for the many helpful supervisor meetings and all the time spent discussing this Master's thesis. I would also like to thank my family for all their love and support in my daily life and their encouragement.

Andreas Bauner, December 2024

The computations were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement n. 2022-06725.

Contents

List of Figures

List of Tables

1
Introduction

Forest ecosystems change dynamically over time, both as a result of natural processes but also due to human activities. In 2020, the Food and Agriculture Organization (FAO) of the United Nations (UN) published a report stating that forests cover 4.06×10^9 ha or 31% of the total land area on Earth [7]. Some examples of positive environmental effects of forests are on climate, landscape, hydrology, water and air quality, CO_2 sequestration and aesthetics [1]. These facts makes it important to monitor and assess how the naturalness and ecosystem functions of forests change over time. It is usually done through forest field inventories by experts in ecology.

This is costly due to requiring a high degree of manual labor and being time-consuming. The aim of this Master Thesis is to develop a machine learning model that can automate the task of naturalness evaluation of forests with high-accuracy predictions. It will make use of remote sensing data in the form of airborne laser scanning data of forests as input. In recent years, many similar studies have made use of the combination of remote sensing and machine learning, for example in forest detection [11] and assessing forest ecological function levels using multi source data [13]. However, there is limited research treating the topic of naturalness of forests specifically.

In particular, the UNet 3+ model will be used in this Master's Thesis, which is a semantic segmentation model that has shown good performance on various image processing tasks. It outperformed other models, such as PSPNet and DeepLabV3+, on the medical diagnosis task for which it was introduced [9]. If used in practice, this could potentially lower the cost of assessing forests for forest management agencies and other stakeholders.

There are several limitations that have been set in this work. I limited the scope by only using the model UNet 3+, a limitation set primarily due to time constraints. UNet 3+ is deemed a good model due to its good performance in the study it was first introduced, having an architecture that is relatively easy to understand and that similar models were used for related tasks when exploring the literature. No other models were used for comparison with UNet 3+. As a result, it was not possible to estimate what effect choosing this particular model had on performance. Focus is instead on measuring what effects on accuracy, that changing hyperparameters in the model can cause. In regards to data, it is limited to using only a subset of features. The data was annotated, georeferenced polygons of expert ecologists representing the ground truth labels, and the Canopy Height Model (CHM) data from Skogsstyrelsen.

1.1 Sustainability and society

In terms of sustainable development, questions on social as well as environmental issues can be raised regarding the use of a black-box model. UNet 3+ is a black-box model due to its large amount of parameters and its complex architecture. This means that it's

impossible for a human to interpret exactly how the model produced its predictions from the input. Such a property is clearly undesirable in high-stakes decision making, where the logic and reasoning is a relevant aspect of what makes the decision credible and legitimate. The prediction whether a forest area has low or high naturalness can be considered a high-stakes decision as it could affect decisions on which forest areas to harvest or preserve. It is therefore important to be critical to its predictions and try to find independent sources that support or refute it. Another issue about UNet 3+ is that training its weights requires a lot of computational operations, which consumes a significant amount of electrical power. High resource consumption tend to have a negative impact on the environment. This negative effect needs to be weighed against the positive environmental impact when using the model. If its predictions are used for preserving important forest areas, increasing biodiversity and natural carbon storage, then there could be a positive net effect on the environment.

2
Background

This chapter provides background on different approaches for evaluating naturalness of forests, both traditionally and by applying machine learning. In addition, it will briefly mention research in related tasks of forest evaluation were machine learning has been applied. Subsequently, details on the type of data and how it is used in this work will be described. Finally, the machine learning model chosen in this work for forest naturalness evaluation is explained.

2.1 Related work in forest evaluation

The evaluation of forest naturalness has traditionally relied on ecological experts conducting surveys through extensive fieldwork. While this approach remains the gold standard for accuracy, it is time-consuming, labor-intensive, and difficult to scale over large geographic areas. Despite the growing advancements in machine learning, its application to the specific task of assessing forest naturalness has been limited. However, significant progress has been made in applying machine learning models to related forest evaluation tasks, particularly those involving remote sensing data.
For instance, Grabska et al. [8] explored the use of three machine learning models—Random Forest (RF), Support Vector Machine (SVM), and Extreme Gradient Boosting (XGBoost)—to map forest stand species. Their research leveraged high-resolution remote sensing data from the Sentinel-2 satellite mission[1], demonstrating the potential of these models for handling complex spectral and spatial information inherent in satellite imagery.
Another relevant example is the work by Bragagnolo et al. [10], where the U-Net model was employed for semantic segmentation to detect forest cover changes in the Amazon rainforest. Using satellite imagery, this study highlighted U-Net's ability to achieve high accuracy in identifying and quantifying deforestation, showcasing its reliability for large-scale environmental monitoring. U-Net, a widely used deep learning architecture for image segmentation tasks, will be described in detail in Section 2.3.
Furthermore, machine learning models have been applied to the semantic segmentation of land cover types using satellite imagery. Singh et al. [14] proposed an enhanced model called Deep-U-Net, which builds on the original U-Net architecture by incorporating additional layers and techniques to improve segmentation performance. Their study demonstrated the effectiveness of Deep-U-Net in differentiating between various land cover types, suggesting its potential applicability in forest-related studies.
These examples illustrate that while machine learning has not yet been extensively applied to the evaluation of forest naturalness, it has proven highly effective in related domains, such as species mapping, deforestation detection, and land cover classification. This growing body of research highlights the potential for further exploration of machine learning

[1] `https://sentiwiki.copernicus.eu/web/s2-mission`

techniques in addressing the complex task of forest naturalness assessment.

2.2 Data

The forest areas in this study, are located in Southern Sweden, in a study area of a 50×50 km square. Its boundaries are parallel to the SWEREF99 coordinate system axes with $(515784, 6408954)$ as the coordinate for the corner with smallest coordinate values. This corresponds to a location a small distance northeast from the city of Jönköping in Sweden.
In this work, the Canopy Height Model (CHM) of forests in Southern Sweden, is used as data. The CHM consists of a set of rasters with information on the height of the trees at different georeferenced locations. More specifically, the CHM has a ground resolution of 1 m, such that each pixel in the raster corresponds to a $1 \times 1\,\mathrm{m}^2$ square on the ground. The canopy height is measured in units of 0.1 m and stored as non-negative integer values. In the rasters, there is also metadata on the date when the CHM was measured. The date of the measurements range from 2018 to 2022, and were collected using airborne laser scanning. The data is freely available as open data[1], provided by Skogsstyrelsen, the Swedish Forest Agency.
In addition to the CHM, there is also ground truth data containing labels on the naturalness of different areas in the shape of polygons. These annotated areas were obtained from ecological field surveys made by ecological experts. The labels are either low naturalness or high naturalness. As a result of the manual and costly method of annotation, the labels are sparsely distributed. Outside these polygons, the naturalness of forests are unlabeled. The data contains three classes, which are assigned class labels. Unlabeled areas, low naturalness areas and high naturalness areas are assigned class labels 0, 1 and 2 respectively. For high naturalness areas, three different data sources were used; *Naturvardsverket*, with habitat-classed areas within Natura 2000, a network of protected areas in Europe protected by EU legislation, *Storskogsbruket*, which makes an inventory of "key habitats" done by forestry companies and *SksBorealSyd*, which is an inventory of key habitats in the South Boreal region done by the Swedish Forest Agency. The process of extracting the high naturalness areas in these datasets was done in a previous related work by filtering on attributes of the datasets [15]. For low naturalness forest areas, three other data sources were used, *BestandEjNaturvarden*, which are forest stands with low naturalness evaluation according to forestry companies, *Hyggen1990-2000*, which are forests that were harvested between 1990 and 2000 and *Pskog30till80*, which are forest stands between 30 and 80 years old.
Once the CHM and labeled georeferenced polygons are obtained, the data is split into three disjoint subsets, a training set, a validation set and a test set. The algorithm used for this process was first developed by Della Vedova and Wahde [15]. The first step is to define a uniform 1.28×1.28 km grid from the study area. Next, squares with no polygons are discarded and the different subsets are obtained from random sampling. The ratio of selected squares is 64% for the training set, 16% for the validation set and 20% for the test set. These squares define regions for the three subsets. Polygons located inside a region belongs to that subset of the data. If a polygon is located inside two or more regions it is split between the subsets accordingly. In addition, polygons with an area less than $0.01\,\mathrm{km}^2$ are discarded as the naturalness of such small polygons cannot be reliably assessed.

[1] `https://www.skogsstyrelsen.se/sjalvservice/karttjanster/skogliga-grunddata/`

2.3 U-Net

The advancement in research on machine learning models has resulted in a great number of different models. One successful type of model in visual recognition tasks is the deep Convolutional Neural Network (CNN). The architecture of these models consists of many convolution layers connected together, with millions of trainable parameters. In order to train these networks i.e. tune their parameters, typically thousands or millions of training samples are required. An early example of CNNs performing well on image tasks is the Fully Convolutional Neural Network (FCN). It was used on the PASCAL VOC 2011 and 2012 datasets, on which it performed significantly better than the previous state-of-the-art method called SDS [3].
The development of CNNs with good performance showed that CNNs could be used in semantic segmentation. U-NET is a model inspired by and extending the FCN, and was used for semantic segmentation in a medical diagnosis task [4]. It consists of two parts, a contracting path and an expansive path. The contracting path transforms the input image to lower resolution feature maps. These feature maps are obtained by applying two 3×3 (unpadded) convolution layers followed by a 2×2 max-pooling layer with stride 2. Note that the convolution layers use Rectified Linear Unit (ReLU) as activation functions. The two convolutional layers reduce the resolution only slightly and are used for creating feature maps with larger scale features from finer detail feature maps. The max-pooling operation functions as a down-sampling step that halves the "height" and "width" of the feature maps. These three layers are then repeated but the number of filter kernels in the convolution layers are doubled every time. As a result of applying these layers repeatedly, a feature map with low resolution but many feature channels is obtained, which contains information on large scale features in the input.
In the expanding path, up-sampling steps are used instead, which increase the resolution again. They consist of an up-sampling operation, which is then followed by a 2×2 convolution layer that halves the number of feature channels, a concatenation with the corresponding cropped feature map from the contracting path and two 3×3 convolution layers with ReLU functions. The last layer is a 1×1 convolution layer, to map the feature vector to a segmentation map with predicted class labels for each pixel of the original image.
In Figure 2.1, the U-Net model architecture is visualized, showing its contracting and expanding path from left to right. It shows the different operations applied to the feature maps as color coded arrows. One of the fundamental principles behind U-NET is its use of skip-connections, connecting feature layers from the contracting path with their corresponding feature layers in the expanding path. As a result of combining features of different spatial resolutions in this way, the U-Net model has a greater ability to retain and make use of information at different spatial scales.

2.4 UNet 3+

UNet 3+ is a model that was constructed to improve on the semantic segmentation performance of U-Net [9]. It was inspired by the observation that U-Net fuses feature maps from the contracting path with feature maps in the expanding path, that are semantically dissimilar. As a result, some loss or degradation of semantic information occur during the fusion. This difference arise due to the feature maps containing semantic information of the input at different scales. Feature maps in the contracting path contain detailed

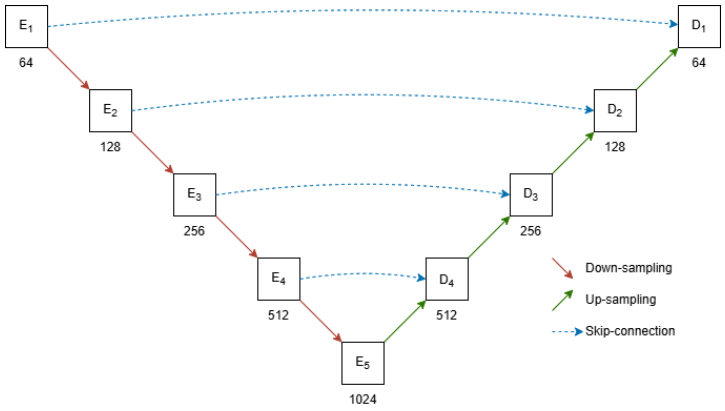

Figure 2.1: U-Net model architecture without its last 1×1 convolution layer, where the white boxes represent the feature maps after applying two 3×3 convolutions with ReLU activation functions. The numbers below the boxes are the number of feature channels after applying the convolutions. Red arrows represent down-sampling with 2×2 max-pooling and stride 2, and the blue dotted arrows are skip-connections, which copies and crops the feature map. The Green arrows represent up-sampling by 2×2 "up-convolution" with stride 2, doubling the spatial resolution but with half the number of feature channels. Feature maps from the skip-connections are concatenated with the feature maps from up-sampling [4].

features with local spatial information, while the feature maps in the expanding path contain high-level information such as the position of larger features. UNet 3+ bridges this semantic gap through the use of full-scale skip-connections. These skip-connections are different from those in U-Net and instead all feature maps from the contracting path with same or smaller scale feature maps are fused with the feature maps in the expanding path. Remember that the feature map in the expanding path before fusing, is obtained by up-sampling a previous feature map in the expanding path with large scale features. With this approach, semantic information at all scales are better preserved[9].

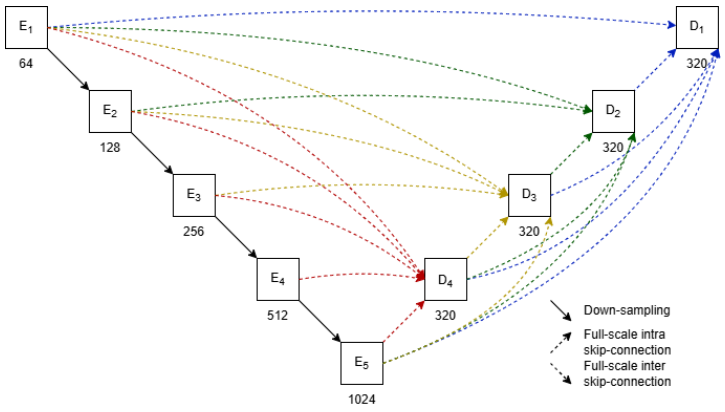

Figure 2.2: The network architecture of the UNet 3+ model, represented in a simplified graph showing intermediary feature maps as blocks. Below each block is the number of feature channels of that block. It shows how the UNet 3+ model has Full-scale skip connections that collect and merge features at all possible scales [9].

In Figure 2.2 the UNet 3+ model architecture is shown in a simple graph, providing some explanation on how its different parts are connected. The original UNet 3+ also made use of supervised learning in intermediate feature maps of the expanding path. It was called full-scale deep supervision but this capability was not used in this work. The transformations in the contracting path are the same as in the U-Net architecture. However, in UNet 3+ the full-scale skip-connections are different from those in U-Net. This difference is required since the full-scale skip-connections merge feature maps with different resolutions and number of feature channels. Down-sampling in the skip-connections between the two paths is achieved by using "non-overlapping max-pooling operations", to decrease the resolution of the feature maps from the contracting path with small scale features [9]. Later, these feature maps are merged with larger scale feature maps from earlier "stages" in the expanding path. The large scale feature maps in the expanding path are up-sampled using bilinear interpolation. In order to merge the feature maps with the same resolution but different number of feature channels, a 3×3 convolution with 64 filters were applied to obtain feature maps that all have 64 feature channels. Later, these were concatenated along the feature channel dimension resulting in a feature map with 320 feature channels, because the UNet 3+ has five "levels" as shown in Figure 2.2. Finally, the features are combined using 320 convolution filters with size 3×3, followed by batch normalization and ReLU as activation function. In the last block D_1 in Figure 2.2, batch normalization and ReLU is not applied, and instead a convolution layer with the number of predicted classes as number of filters[9]. In order to obtain a segmentation map, a Softmax activation function is applied such that the output can represent pixel-wise predictions of low and high naturalness, as model confidence scores.

3
Methods

The data used in this work has been described, as well as the architecture of UNet 3+, the deep learning model selected for evaluating forest naturalness. To effectively utilize this data within the machine learning framework, extensive preprocessing is required. The raw inputs consists of a Canopy Height Model (CHM) and georeferenced polygons that represents annotated naturalness classes. From these inputs, square rasters of fixed dimensions are generated. Each raster contains two key components: the CHM data, which provides the height profile of the forest canopy, and the annotated class labels derived from the polygons. These labels are then discretized to correspond to individual pixels within the raster, which guarantees correct spatial alignment between the input features and the target outputs.

However, to make the data compatible with the UNet 3+ model, additional preprocessing operations are required. First, the CHM data underwent normalization to standardize its values, ensuring consistency across all rasters and facilitating more efficient learning by the model. Normalization also helps to prevent numerical instability during training by scaling the CHM values to a range suitable for the neural network.

Next, the dataset is filtered to ensure that only high-quality samples are used for training and evaluation. This step involves removing rasters with a low fraction of naturalness class labels. Such rasters are produced as artifacts, when the CHM is split into fixed size squares. The reason is that the position or shape of the polygons do not align with the annotated ground truth naturalness labels.

In the following sections, each preprocessing step will be described in greater detail, including the specific techniques and thresholds applied to normalize and filter the data, as well as any challenges encountered during this process. These steps are crucial in preparing the data for input into the UNet 3+ model and to satisfy the criteria of robust and reliable evaluation of forest naturalness.

3.1 Pre-processing data to raster files

One of the key differences between applying UNet 3+ on the problem of evaluating forest naturalness and the domain of medical diagnosis is in the pre-processsing step. The UNet 3+ in medical diagnosis expects stacked gray scale images, corresponding to volumetric medical scans. These have three channels with integer values ranging from 0 to 255. In this application, the CHM has integer values representing canopy height in decimeter, with values outside of this range. In addition, the data is not nicely delimited in fixed image dimensions. Instead, it is specified by georeferenced, labeled polygons of varying sizes and shapes.

The given dataset contains cases where polygons with different class labels overlap, which leaves the class labels of these areas ambiguous. In order to account for this case, the

overlapping areas are removed. The algorithm then proceeds to generate square tiles with a fixed dimension in the SWEREF99 coordinate system. From the georeferenced polygons, ground truth labels are discretized to pixel locations in the square tile. It is then combined with the corresponding CHM at those pixel locations to obtain a raster of two bands. The first band has the CHM values and the second band has the class labels of each pixel (either $0, 1$ or 2). Note that the rasters also contains information on their spatial location in the SWEREF99 coordinate reference frame.

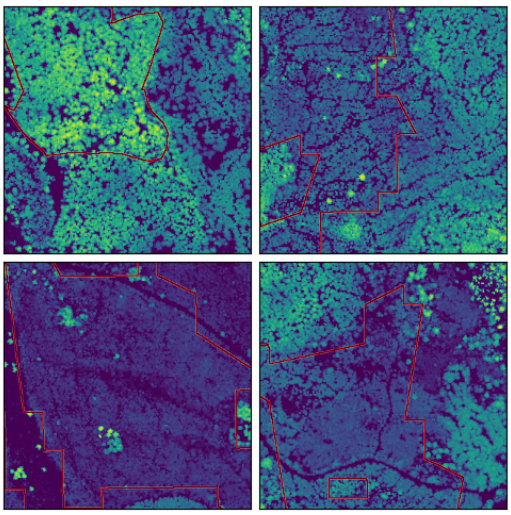

Figure 3.1: The CHM of 4 different forest areas evaluated as having low naturalness inside their red borders.

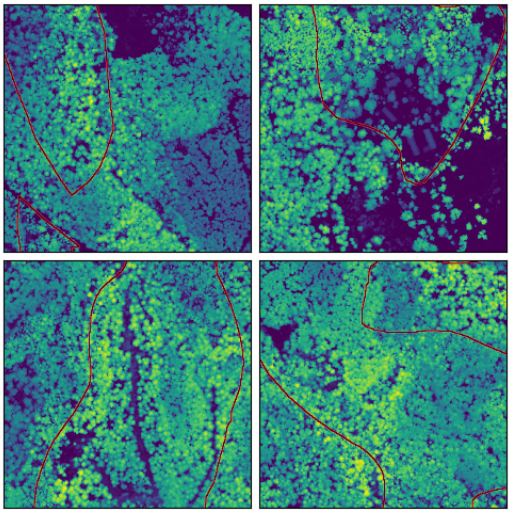

Figure 3.2: The CHM of 4 different forest areas evaluated as having high naturalness inside their red borders.

3.2 Normalization and scaling

The CHM for the annotated areas and their surroundings, are stored as files without modification in the previous step. As a result, the CHM of the rasters contain integer

values within an interval not known in advance. In order to constrain the data distribution of the input and make it less sensitive to measurement noise, a fixed max value is set. CHM values greater than the max value is set to the max value. The input to the model is then scaled with the inverse of the max value to real values in $[0, 1]$. This significantly lowers the mean value and variance in the input data which improves numeric stability during training of the model. Selecting the max value for the CHM is done by testing different max values and visually inspecting the corresponding histogram. As a result of folding in values greater than the max value, there was a spike in frequency at the max value. The aim is to set the max value such that the spike of frequency is not too big relative to neighboring bins. This approach ensures that not too much information is lost from the data with the greatest canopy height. At the same time, the max value should be set low enough to benefit from a lower variance in the input.

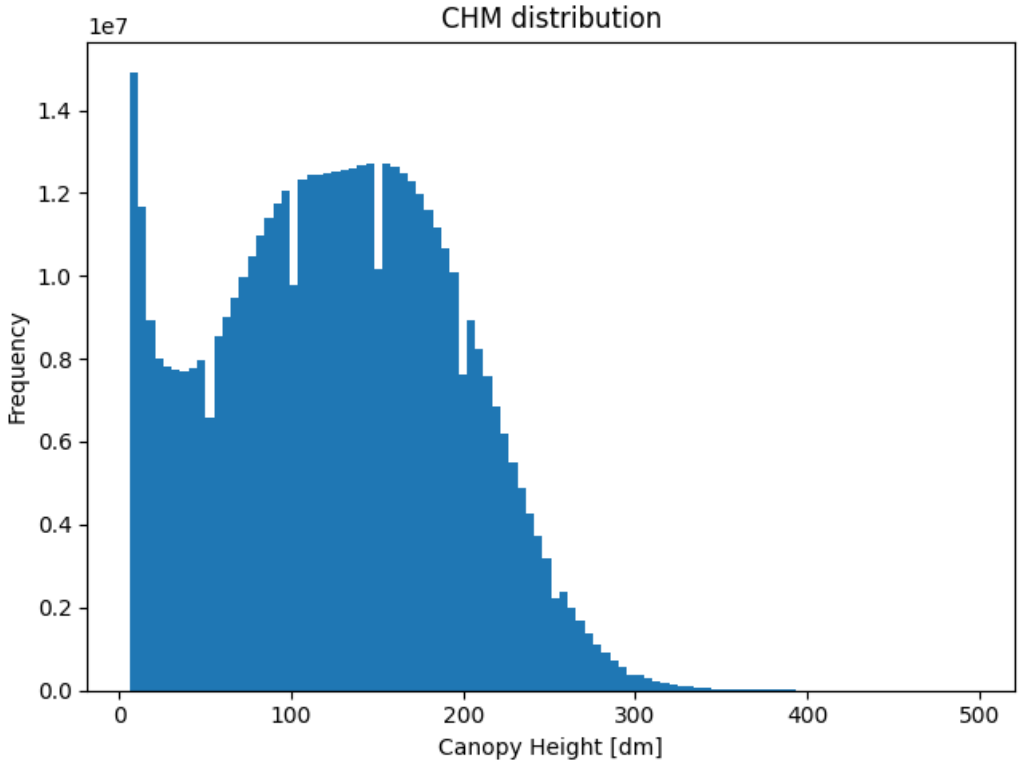

Figure 3.3: The histogram of the CHM for the 50 by 50 km study area, for all the labeled polygons and their surroundings. Used 100 uniformly spaced bins for this histogram. Note that bins with values less than 6 dm contain 31.5% of the data and creates a dominant peak. This peak has been omitted for a better visualization of other parts.

In Figure 3.3, the original data distribution of all canopy height data points is shown, for all labeled polygons in the study area. It shows that the distribution is heavily skewed towards 0 dm canopy height. In addition, it shows that the data distribution at large CHM values has few data points, a so called "tail".

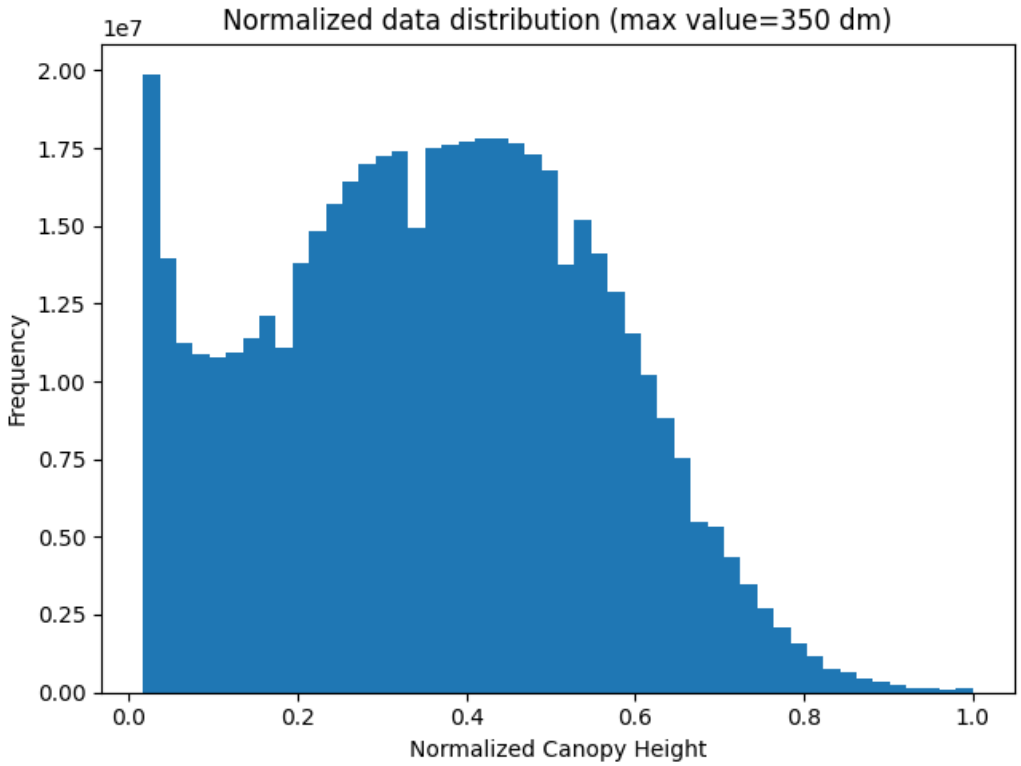

Figure 3.4: The normalized histogram of the CHM with max value 350 dm, for the 50 by 50 km study area, for all the labeled polygons and their surroundings. Used 50 uniformly spaced bins for this histogram. Note that bins with values less than 6 dm contain 31.5% of the data and creates a dominant peak. This peak has been omitted for a better visualization of other parts.

In Figure 3.4, the effects of normalizing the CHM from the pre-processing step, is shown. It shows that the tail is removed and the re-scaling to values in $[0, 1]$. For the normalization, the max value was set to 350 dm. This max value was deemed suitable with respect to the data used in this task but would probably not be suitable for the CHM of forests in other regions of the world. The max value would in that case need to be adjusted using a similar approach. It is also worth emphasizing that the lowest values (values below 6 dm) have been omitted from Figure 3.3 and Figure 3.4. These data points were predominant, constituting 31.5% of all data points, resulting in bad scaling of the histogram if not omitted.

3.3 Model training and loss functions

Following pre-processing and normalization of the rasters, the next step involves training the UNet 3+ model on the prepared data. The training process requires careful configuration of several hyperparameters to optimize model performance. To maintain consistency and adhere to the baseline implementation of UNet 3+, default hyperparameter values from the original architecture are utilized wherever possible. Specifically, the Adam optimizer is employed with learning rate $\eta = 3 \times 10^{-4}$, and batch size $m_s = 2$ iss used to accommodate memory limitations. The number of epochs, a critical factor influencing both training time and model convergence, is adjusted iteratively during experimenta-

tion. Ultimately, it was determined that 25 epochs provided sufficient training duration for the model to converge and achieve its optimal training accuracy without overfitting or excessive computational expense. The Adam optimizer performs optimization steps on the model weights such that a loss function is minimized. In this work the "hybrid loss function" from the original UNet 3+ paper is selected as a baseline [9]. This loss is defined by combining three loss functions: Focal loss, Intersection over Union (IoU) loss and Multi-Scale Structural Similarity Index (MS-SSIM) loss. In addition, UNet 3+ will be trained on each of the loss functions that can be additively combined from these three loss functions, including single loss functions. Afterwards, the model performance of each parameter instantiation of UNet 3+ will be compared, in order to analyze how the performance of UNet 3+ depends on the loss function.

3.3.1 Focal loss

Focal loss was first introduced in the context of dense object detection to enhance the performance of one-stage detectors. Object detection refers to the task of identifying and localizing objects in an image that belong to a predefined set of classes. Lin et al. observed that one-stage detectors face significant challenges due to class imbalance during training. These detectors generate a vast number of candidate object locations, but only a small fraction of these candidates correspond to actual objects. Specifically, the majority of the candidates belong to the negative class, corresponding to the object not being present [6]. This imbalance causes the model to disproportionately focus on the dominant negative class, impairing its ability to effectively learn to classify object locations. To address this issue, Lin et al. proposed focal loss, a novel modification of the standard cross-entropy (CE) loss designed to reduce the impact of easy-to-classify negative examples and prioritize harder, informative examples. The focal loss l_{fl} is defined as:

$$l_{fl}(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t), \tag{3.1}$$

where α_t is a parameter for scaling the loss, $\gamma > 0$ is called a focusing parameter and p_t is the probability or confidence score of the model predicting the correct ground truth class [6]. Note that focal loss is the cross-entropy loss modified with two factors α_t and $(1-p_t)^{\gamma}$. The second factor lowers the loss for easily predicted samples (p_t high), relative to those samples that are difficult to predict (p_t low). It effectively drives the model towards learning how to predict the most difficult inputs, which was the original motivation for defining this loss. The parameters were set to $\alpha_t = 4$ and $\gamma = 2$, which were the parameter values used in the original UNet 3+ paper.

3.3.2 Intersection over Union loss

The second loss function employed is the Intersection over Union (IoU) loss, denoted as l_{iou}, which is derived from the widely used IoU segmentation metric. The IoU metric measures the overlap between the predicted segmentation and the ground truth and takes values in the range $[0, 1]$, where values closer to 1 indicate better agreement. To transform this metric into a loss function suitable for optimization, IoU loss is defined as:

$$l_{iou} = 1 - \text{IoU}. \tag{3.2}$$

This formulation ensures that minimizing the IoU loss is equivalent to maximizing the IoU metric. The name Intersection over Union reflects the way this metric is computed:

it quantifies the ratio of the overlapping area (intersection) between the ground truth and predicted segmentation masks to the total combined area (union). This intuitive interpretation highlights its relevance for segmentation tasks, as it explicitly penalizes both oversegmentation and undersegmentation. Cheng et al. provided a general definition [12] of the IoU metric when applied to binary image segmentation:

$$\text{IoU} = \frac{|G \cap P|}{|G \cup P|}, \tag{3.3}$$

where G is the binary ground truth mask and P is the predicted binary mask. In this formula, the numerator $|G \cap P|$ represents the number of pixels correctly predicted as part of the object (i.e., the true positives). The denominator $|G \cup P|$ accounts for all pixels in either the predicted mask or the ground truth mask, effectively summing the true positives with the incorrectly classified pixels (false positives and false negatives). This property makes the IoU metric particularly robust for evaluating segmentation performance, as it balances precision and recall in a single unified measure.
By using IoU loss, the optimization process directly focuses on improving the overlap between predictions and the ground truth, making it an effective choice for segmentation tasks were pixel-wise accuracy alone is insufficient to capture the quality of the results.
As stated previously, the IoU metric in binary segmentation is usually defined to focus exclusively on the positive class. This is because, in many segmentation tasks, the negative class tends to represent the background rather than an object of interest. In this work, the "negative class" is the class for low naturalness, and the IoU metric needs to account for both classes. Instead, mean Intersection over Union (mIoU) is used for the IoU loss, which is calculated by first computing the IoU for the two classes and then averaging these values. For an alternative mathematic formulation of the mIoU metric refer to [5], which provides additional insights on the properties of mIoU.

3.3.3 Multi-scale structural similarity loss

The third and final loss function is the Multi-Scale Structural Similarity Index (MS-SSIM) loss, denoted $l_{ms-ssim}$. which is defined using a modified version of the metric MS-SSIM [2]. It is computed on pairs of small patches extracted from the prediction mask p and ground truth labels g, as described in the UNet 3+ paper. MS-SSIM loss compares spatial structure at multiple spatial scales and assigns higher loss for fuzzy boundaries. Consequently, the MS-SSIM loss encourages UNet 3+ to learn how to predict sharper, and more precise boundaries between segments [9]. Huang et al. define the MS-SSIM loss in the UNet 3+ paper as:

$$l_{ms-ssim} = 1 - \prod_{m=1}^{M} \left(\frac{2\mu_p\mu_g + C_1}{\mu_p^2 + \mu_g^2 + C_1} \right)^{\beta_m} \left(\frac{2\sigma_{pg} + C_2}{\sigma_p^2 + \sigma_g^2 + C_2} \right)^{\gamma_m}, \tag{3.4}$$

where M is the total number of scales, μ_p, μ_g are the mean of patches p and g respectively, while σ_p, σ_g and σ_{pg} are standard deviations and the covariance of the patches. MS-SSIM considers multiple scales and the parameters β_m and γ_m determine how much the two factors in Equation (3.4) contribute at each scale. The number of scales was set to 5 and the parameter values were set with values from [2], in agreement with the UNet 3+ paper [9]. Three loss functions have now been defined and the hybrid loss l_{seg} is then defined as:

$$l_{seg} = l_{fl} + l_{ms-ssim} + l_{iou}. \tag{3.5}$$

Lastly, note that the ground truth labels can assume three values, while the predictions are binary (low or high naturalness). This asymmetry was handled by not using the "unlabeled pixels" when computing the loss and metric. As a result, rasters with a large fraction of unlabeled pixels have a low loss independent of the model predictions. This could cause problems when training UNet 3+, because the loss function should reflect properties of the model predictions and not on the sparsity of ground truth labels.

3.3.4 Evaluation metric

In this work, accuracy is chosen as the primary evaluation metric because it appropriately treats incorrect predictions equally for both the low and high naturalness classes. This balanced treatment is particularly important in this context, as there is no inherent reason to prioritize one class over the other in terms of classification performance. Furthermore, accuracy was also used in the research paper on forest naturalness evaluation by [15], on which this thesis builds. Adopting the same evaluation metric facilitates direct comparison with their results, allowing for a clearer assessment of improvements or differences in performance.

It is worth noting that while accuracy is suitable for this task, other evaluation metrics could also be considered, provided they satisfy the requirement of equal treatment for the two classes. For instance, metrics such as balanced accuracy, F1 score, or mean Intersection over Union (mIoU) could be used if specific characteristics of the dataset or task warranted their application. However, for the purposes of this thesis, accuracy remains a logical and effective choice, given its straightforward interpretability and compatibility with prior work in the field.

3.4 Post-processing

In Section 3.3, it was discussed that unlabeled pixels were excluded when computing both the loss and accuracy of the UNet 3+ model. While this approach ensures that the model is evaluated only on meaningful regions of the data, it introduces a potential issue: rasters containing a high proportion of unlabeled pixels tend to have artificially low loss values. This occurs because, with fewer labeled pixels contributing to the loss calculation, the overall loss becomes disproportionately small, even if the model's predictions on the labeled regions are suboptimal.

If such rasters with many unlabeled pixels are prevalent in the training and validation datasets, the artificially low loss values may mislead the optimization process and the evaluation of the model. During training, the optimizer may mistakenly interpret these low-loss rasters as indicators of good performance, even though they are simply an artefact of insufficient labeled data. This can result in the model failing to improve on genuinely labeled regions, as the loss no longer accurately reflects prediction quality. Furthermore, during validation, the model's performance may appear better than it truly is, leading to an overestimation of its accuracy.

This issue highlights the importance of careful dataset preparation and loss weighting strategies when working with datasets that include significant proportions of unlabeled pixels. Techniques such as balancing the contribution of different rasters to the overall loss or introducing additional penalties for underperforming on labeled pixels could mitigate these effects and ensure that the model is evaluated and trained more reliably.

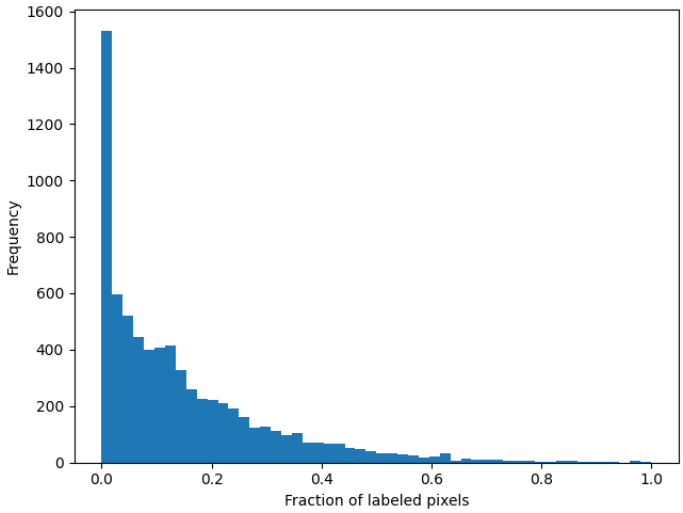

Figure 3.5: Histogram over the fraction of annotated labels computed for each raster. A fraction of 0 represents a raster with only unlabeled pixels (label 0). In contrast, a fraction of 1 represents that all labels are either 1 or 2 (low or high naturalness). All rasters from the training, validation and test set were combined.

In Figure 3.5, the distribution for the fraction of labeled pixels in each raster is shown, where the fraction could be called "label density". The distribution clearly shows that rasters with few labels are the most common in the data. In order to mitigate the negative effect this could have, rasters with a label density below 10% are removed from the training and validation set before training as well as in the test set during evaluation. This percentage was subjectively chosen based on the shape of the label density distribution in Figure 3.5. It has a dominant peak around 0% and bins with gradually lower frequency as the label density increases. The aim with filtering was to remove inputs with the lowest label density, and it was weighed against the adverse effect this had on model training due to less training data.

4

Results & Discussion

In the previous chapter, the methods for processing the CHM in order to better use it as input in the UNet 3+ model, was described. The steps to train the model was also described, were UNet 3+ used the CHM and the corresponding ground truth labels on forest naturalness, to adjust its parameters. The performance of UNet 3+ will be evaluated both by comparing ground truth labels with predicted labels visually, as well as its accuracy on the test set. In addition, the loss function will be varied to ascertain its effects on model performance.

4.1 Model training evaluation

It has been extensively explained how the UNet 3+ is applied to the CHM input data to give the model output and how it will be evaluated using accuracy as metric. In order to train UNet 3+ efficiently to obtain high performance, it is necessary to analyze and optimize the model training. Improving on model training is an iterative process, where visualizing the evolution of training loss with validation loss, as functions of number of epochs can provide insights on how well the model learns to generalize from the data.

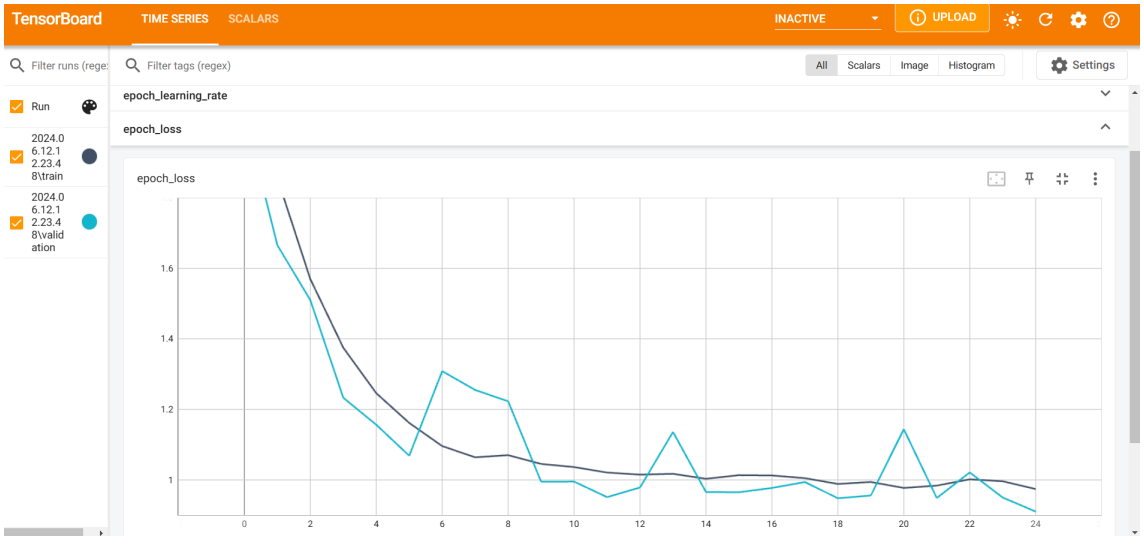

Figure 4.1: Tensorboard visualization of the training (dark blue) and validation loss (light blue) as functions of training epochs with the sum of Focal and IoU loss as total loss. Observe that the loss stops decreasing at the end of the training around epoch 25 and that training and validation loss are at a comparable level.

In Figure 4.1, the training and validation losses are plotted as they evolve over successive

epochs during the model training process. This example illustrates the case where the total loss is computed as the sum of the Focal Loss and IoU Loss. A common objective when training a machine learning model is to strike a balance between avoiding underfitting and overfitting, ensuring the model generalizes well to unseen data.
Underfitting occurs when the model performs poorly on both the training and validation sets because it fails to capture the underlying relationships in the data. This is often indicative of insufficient training time, a lack of model complexity, or an inadequate dataset. The simplest solution is to extend the training duration if the model architecture is sufficiently complex and the dataset contains enough representative examples. Alternatively, addressing underfitting might require increasing the size or diversity of the training dataset, adjusting hyperparameters, or switching to a more expressive model architecture better suited to the data's structure.
In this study, underfitting was rarely observed due to the use of the large and expressive UNet 3+ architecture combined with a sufficiently large dataset. Instead, the primary challenge was mitigating overfitting, a phenomenon where the model fits the training data too closely, capturing noise or spurious patterns that do not generalize to the validation set. Overfitting was managed primarily through careful selection of the number of training epochs.
The loss curves shown in Figure 4.1 provide insights into the training process. The nearly monotonically decreasing training loss, accompanied by a validation loss that steadily decreases to a comparable level, indicates that the model maintained good generalization without significant overfitting. This suggests that training the model for 25 epochs was an appropriate choice. If training had continued for substantially more epochs, there might have been a risk of overfitting, in which case it would show as a divergence between the training and validation loss.

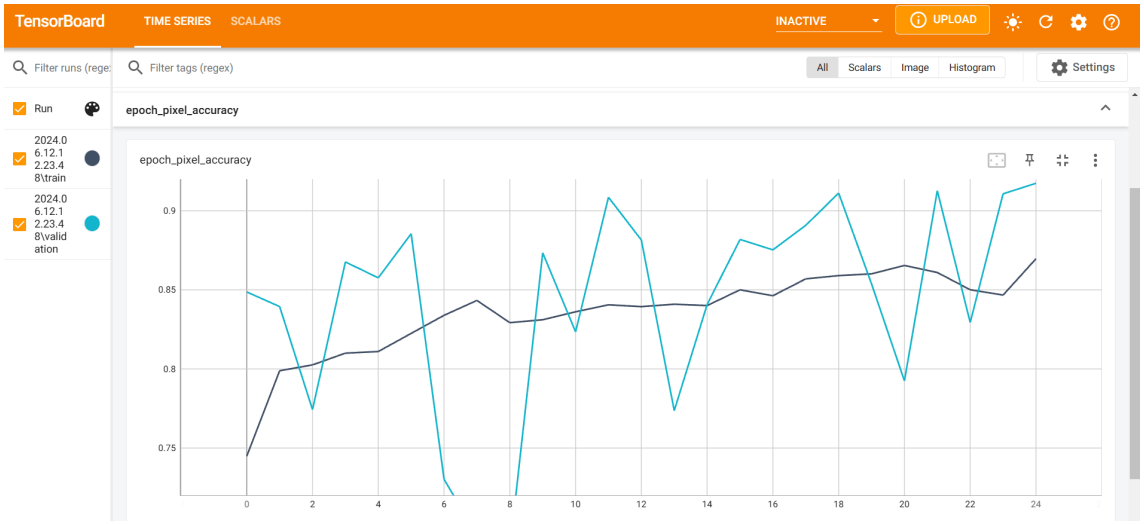

Figure 4.2: Tensorboard visualization of the training (dark blue) and validation accuracy (light blue) as functions of training epochs with focal_iou loss. Observe that the validation accuracy at the end of the training is approximately 92%, which is significantly higher than the training accuracy at approximately 87%.

In Figure 4.2 the training and validation accuracy of UNet 3+ is shown, which was trained with the sum of Focal and IoU loss as its total loss. Observe that the training accuracy increases steadily without much fluctuation the longer the model is trained and reaches its

maximum accuracy around 87%. In contrast, the validation accuracy increases in a volatile pattern of oscillations around the training accuracy as the training epochs increases. A surprising fact is that the validation accuracy is greater than the training accuracy at the end of training, with accuracy at roughly 92%. The volatility in validation accuracy suggests that the model performance on unseen data is sensitive to parameter updates at the end of model training.

4.2 Effect of loss function on model performance

In section 3.3 the loss function for UNet 3+ was defined as a hybrid loss function consisting of a sum of three loss functions. These loss functions are the focal loss, the IOU loss and the MS-SSIM loss. In machine learning, model training is done by minimizing the loss function by iteratively adjusting model weights. As a result, the choice of loss function could have a significant impact on how the model is trained and its performance. To determine the effect each one the three loss functions has on model performance, UNet 3+ is trained using seven different combinations of loss functions. The model trained using the hybrid loss function could be considered a baseline for comparison. For the other loss functions the model either uses a single loss function or a combination of two of the loss functions. Note that these loss functions are selected through a binary choice of including a loss function or not. This would correspond to scaling the loss functions in the hybrid loss with binary weights $w_i \in \{0, 1\}$, $i = 1, 2, 3$. Instead of this approach with binary choices, one could weigh the loss functions with positive, real valued weights. Optimizing these weights would become an optimization problem in a subset of $\mathbb{R}^3$ instead of comparing seven discrete cases. In this thesis, comparing seven discrete cases was done for the sake of simplicity.

Each UNet 3+ model is trained on the training and validation set after filtering as previously described, and with its own unique loss function. Next, each trained UNet 3+ with its unique model weights, has their accuracy evaluated on the test set after filtering. The filtering helps with avoiding the case of evaluating on samples with no labels, where evaluating accuracy becomes undefined. It also increase the robustness of the evaluation as when evaluating on samples with few labels, each mislabeled pixel prediction has a greater adverse effect on accuracy than with samples that have many ground truth labels. This would make such evaluations more sensitive to noise in the ground truth labels. The ground truth inadvertently has some incorrectly classified labels due to being represented as georeferenced polygons of low resolution. This makes it reasonable to exclude samples with less than 10% ground truth labels when evaluating the model. Evaluating UNet 3+ with the hybrid loss function results in an accuracy of approximately 93.6%.

The accuracy of UNet 3+ with different loss functions is presented in Table 4.1, with each trained model evaluated on the test set after filtering. It shows that the accuracy is almost constant regardless of the loss function used, with the exception of MS-SSIM loss. The accuracy of the six other loss functions belongs to the short interval $[0.933, 0.941]$. The largest difference in accuracy between these loss functions is only 0.7 percent units. Such a small difference in accuracy is not significant and can be attributed to other factors than the accurateness of the model. Consequently, the choice of loss function from this set of six loss functions does not have a great impact on the accuracy of the trained UNet 3+ model. The best accuracy was obtained with the UNet 3+ that was trained using the sum of focal loss and IoU loss as its loss function, with an accuracy of 94.1%.

When training UNet 3+ using only the MS-SSIM loss as the loss function, the model

Table 4.1: The accuracy of UNet 3+, when evaluated on the test set after 10% filtering, trained with different loss functions.

Loss function	Accuracy
unet3p_hybrid_loss	0.936
ms_ssim_loss	0.887
iou_ms_ssim	0.934
iou_loss	0.939
focal_ms_ssim	0.933
focal_iou	**0.941**
focal_loss	0.938

achieved an accuracy of 88.7%, representing a significant drop in performance compared to the results obtained using other loss functions. Specifically, the decrease in accuracy is approximately 5 percentage points, indicating that the MS-SSIM loss may not be well-suited for the task of predicting forest naturalness.
The MS-SSIM loss was originally introduced in the UNet 3+ paper to penalize model predictions with fuzzy boundaries, encouraging the model to produce sharper, well-defined boundaries between the positive class (e.g., an organ) and the negative class (e.g., the background). This property makes logical sense in the context of semantic segmentation for medical diagnosis, where boundaries are often inherently sharp, such as the edges of organs or lesions in imaging data.
However, for the task of forest naturalness evaluation, this focus on sharp boundaries is less appropriate. Naturalness is expected to transition more gradually across forest areas, as ecological changes rarely occur as discrete, abrupt shifts. For instance, a forest might gradually change from high to low naturalness due to factors like human intervention, tree age, or biodiversity loss. Enforcing sharp boundaries between these regions does not reflect the true nature of the problem and may introduce inaccuracies in model predictions.
A further complication arises from the ground truth data itself, which is derived from low-resolution polygons with artificially sharp edges that do not correspond to the actual, gradual boundaries between forest areas of differing naturalness. By relying on MS-SSIM loss, the model is encouraged to align its predictions with these unrealistic sharp edges, which misrepresents the underlying ecological transitions. Consequently, the use of MS-SSIM loss leads the model to produce outputs with unnaturally sharp boundaries, which is undesirable in this context.

4.3 Evaluation of prediction maps

The accuracy of the UNet 3+ model predictions was previously evaluated by applying it on the test set and comparing the predictions with the ground truth labels on the naturalness. This gave a good measurement on how well the model predicts naturalness for different loss functions. These model predictions will now be visualized for specific samples in the test set. This will hopefully provide some general insight on how the UNet 3+ model makes its predictions from the CHM. All figures will show predictions from the UNet 3+ model trained using the sum of focal and IOU loss, as it performed best in terms of accuracy. The model gives probabilities of high naturalness p, which are transformed to labels using the condition $p > 0.5$, to high or low naturalness respectively. This has the effect that

two model predictions p_1 and p_2, which are arbitrarily close but one is smaller and one is larger than 0.5, will give two different labels. The model can be therefore be said to be uncertain in its prediction if it's close to 0.5. In the visualization, the predicted labels will be represented with black pixels if the model prediction is uncertain. In our case, a model prediction p is said to be uncertain if $|p - 0.5| <= 0.15$, or equivalently, $p \in [0.35, 0.65]$. The threshold 0.15 was chosen such that the length of the probability interval for each label is roughly equal.

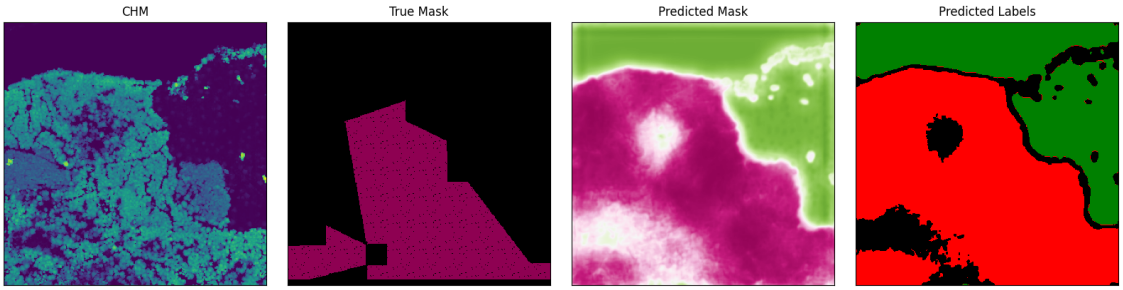

Figure 4.3: The CHM (left), the ground truth labels (left of center), predicted probability of high naturalness (right of center) and the predicted labels (right) for a 320 m × 320 m raster. Black pixels in the ground truth represent unlabeled pixels, while for the predicted labels they represent pixels where the model is uncertain. The red and purple pixels represent low naturalness, while pixels in the two shades of green represent high naturalness. It is sampled from the post-processed test set.

In Figure 4.3, the predicted labels are shown to agree with the ground truth on most pixels. The predicted labels also shows smooth and naturally drawn borders between different segments, which correspond to clearly visible segments in the CHM. In contrast, the border of the ground truth has sharp edges and looks to be drawn somewhat arbitrarily in relation to the CHM. Model predictions seem to become less certain when there's a rapid change in the CHM, which is visible at the border between non-forest (the dark blue area) and the forest. It is also uncertain when there's a change in a local spot of the low naturalness forest, which otherwise mostly has a similar spatial distribution. Note also how the model predicts high naturalness for the large dark blue area in the CHM that has no trees. This is an artefact produced because the model can only predict high or low naturalness and that the model recognizes that the non-forest area is not part of the forest with low naturalness.

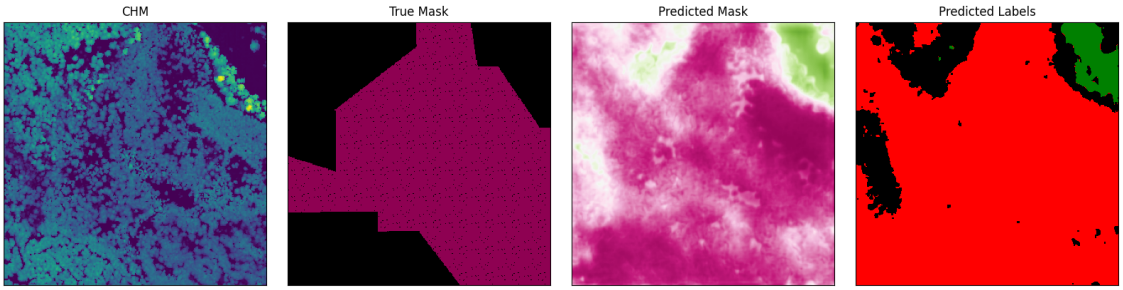

Figure 4.4: The CHM (left), the ground truth labels (left of center), predicted probability of high naturalness (right of center) and the predicted labels (right) for a 320 m × 320 m raster. Black pixels in the ground truth represent unlabeled pixels, while for the predicted labels they represent pixels were the model is uncertain. The red and purple pixels represent low naturalness, while pixels in the two shades of green represent high naturalness. It is sampled from the post-processed test set.

In Figure 4.4, the predicted and ground truth labels mostly agree and the UNet 3+ model also seems quite confident with its predictions. This suggests that forests with similar spatial distribution in their CHM, are typically forests with low naturalness.

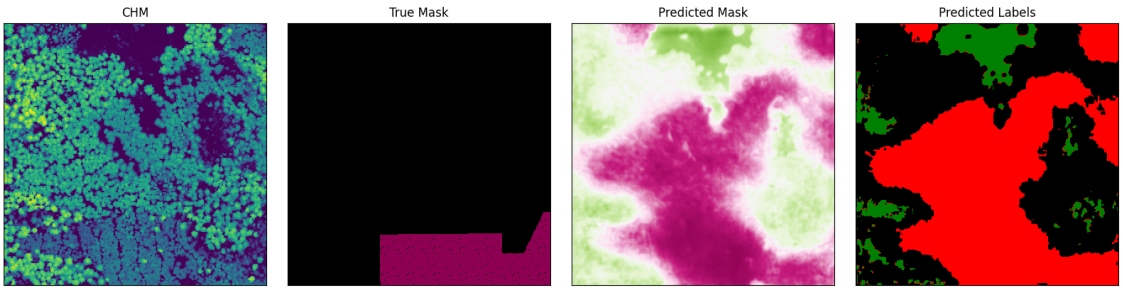

Figure 4.5: The CHM (left), the ground truth labels (left of center), predicted probability of high naturalness (right of center) and the predicted labels (right) for a 320 m × 320 m raster. Black pixels in the ground truth represent unlabeled pixels, while for the predicted labels they represent pixels were the model is uncertain. The red and purple pixels represent low naturalness, while pixels in the two shades of green represent high naturalness. It is sampled from the post-processed test set.

In Figure 4.5, the CHM includes a mix of forest areas with trees that have low and high canopy height. Although the UNet 3+ model predicts the ground truth labels correctly as a low naturalness forest, the substantial area of black pixels in its predicted labels suggests that overall, there's a significant amount of uncertainty in the model predictions. An explanation for the uncertainty could be that UNet 3+ find it difficult to obtain useful information from combining spatial features at different scales because the changes in CHM are too different at various scales. It could also be related to the fact that if the forest is fragmented into many small areas with different naturalness, then there's also many local borders between such areas. The naturalness of such small scale forest areas could be considered inherently uncertain as the concept of naturalness breaks down at small enough spatial scales.

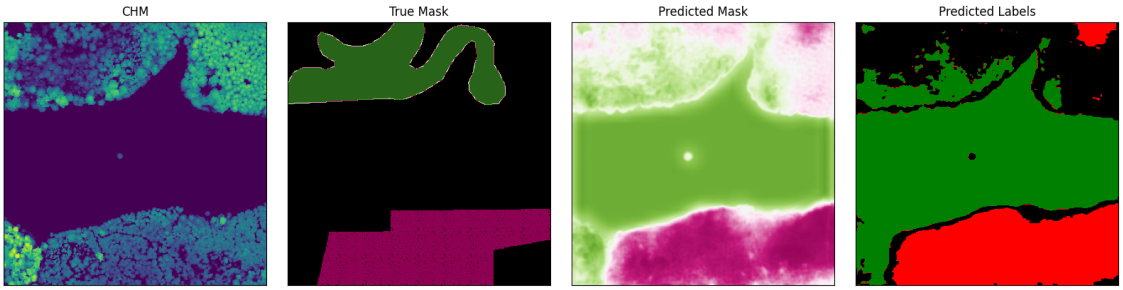

Figure 4.6: The CHM (left), the ground truth labels (left of center), predicted probability of high naturalness (right of center) and the predicted labels (right) for a 320 m × 320 m raster. Black pixels in the ground truth represent unlabeled pixels, while for the predicted labels they represent pixels were the model is uncertain. The red and purple pixels represent low naturalness, while pixels in the two shades of green represent high naturalness. It is sampled from the post-processed test set.

In Figure 4.6, both forest areas with low and high naturalness are present in the ground truth. From observing the CHM it seems like the areas are divided by a river or lake since the area in the middle is completely devoid of trees and vegetation. Comparing the ground truth with the predicted labels for the low naturalness forest at the bottom of the figure confirms they are in agreement. In contrast, the model predictions of the high naturalness area at the top of the figure can be considered uncertain. This is possibly a result of the irregular shape of the river or lake, which prevents the model from connecting forest areas of high naturalness at different sides of the "wedge".

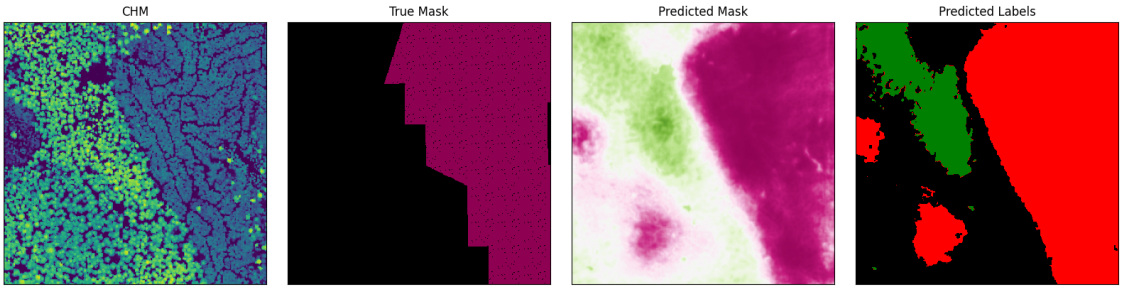

Figure 4.7: The CHM (left), the ground truth labels (left of center), predicted probability of high naturalness (right of center) and the predicted labels (right) for a 320 m × 320 m raster. Black pixels in the ground truth represent unlabeled pixels, while for the predicted labels they represent pixels were the model is uncertain. The red and purple pixels represent low naturalness, while pixels in the two shades of green represent high naturalness. It is sampled from the post-processed test set.

The CHM in Figure 4.7 shows two distinct regions separated by a smooth boundary. However, the ground truth labels instead display a boundary with many edges. This is a good example of how the low resolution of the ground truth polygons could introduce measurement noise when training the model, or in this case, when evaluating the model accuracy. Interestingly, the UNet 3+ model gives a prediction with the smooth boundary one would expect from the CHM.

5
Conclusion

This work set out to develop a machine learning model capable of accurately evaluating forest naturalness using data derived from a Canopy Height Model (CHM). The overarching goal was to take a step towards automating the evaluation of forest naturalness, reducing the reliance on labor-intensive and costly field studies. To achieve this, an input pipeline was designed to combine CHM data with ground truth labels extracted from georeferenced polygons, transforming these inputs into rasters, which was a more suitable format for machine learning.
A machine learning model called UNet 3+ with a strong track record in semantic segmentation tasks—particularly in the domain of medical diagnostics—was adapted and optimized for the specific challenges of forest naturalness evaluation. This model, modified to process raster data from forest environments, achieved an accuracy of 94.1% on the test set after filtering rasters with too few ground truth labels. This result demonstrates that the proposed approach is competitive with, and potentially superior to, existing methods for evaluating forest naturalness.
While the model and input pipeline developed in this thesis show strong performance, certain limitations remain. For instance, the level of automation achieved may not yet be sufficient for fully autonomous large-scale deployments, and aspects such as generalization to diverse forest types or other ecological datasets were not fully explored. Despite these challenges, the outcomes of this work demonstrate the viability of applying advanced machine learning techniques to this domain and set a strong foundation for further research.
The results presented here provide a valuable source of inspiration and a comparative benchmark for future studies aiming to refine and expand automated forest naturalness evaluation. By building on the methods and insights of this work, researchers can contribute to the broader goal of integrating advanced technologies into sustainable forest management practices.

5.1 Future Work

During the course of this work, certain limitations in scope had to be made, which raised several questions requiring further investigation. Addressing these questions could significantly enhance the accuracy, scalability, and practicality of forest naturalness prediction models.
One crucial avenue for future work is the development of a more automated pipeline for predicting forest naturalness from Canopy Height Models (CHMs). In its current form, the pipeline involves significant manual effort, particularly in preparing the data, such as generating labeled rasters from georeferenced polygons. Automating this process would greatly improve scalability, enabling the analysis of larger areas with reduced human intervention. One major obstacle to automation is the supervised nature of the UNet 3+ model,

which requires manually annotated labels for training. Future studies could explore alternative approaches, such as unsupervised learning or self-supervised learning techniques, which might alleviate the dependence on labeled data while still providing meaningful predictions.

Another promising area of research is the exploration of interpretable models for forest naturalness evaluation. While the UNet 3+ is a high-performing model, it functions as a "black-box" system, offering limited insight into how predictions are made. Investigating interpretable models could provide valuable explanations for predictions, helping forest managers and ecologists better understand the factors influencing naturalness assessments. Comparing the performance of interpretable models with black-box models like UNet 3+ would also shed light on whether transparency can be achieved without sacrificing accuracy.

This work also raised practical questions about how to best divide the study area into smaller fixed-size squares (rasters) to ensure efficient and accurate segmentation. The current approach led to many rasters containing large regions of unlabeled pixels, which affected the loss computation and training dynamics of the model. Developing improved strategies for dividing georeferenced polygons into squares—such as optimizing for minimal overlap and coverage efficiency—could significantly mitigate this problem, leading to cleaner and more representative datasets.

Extending the scope of this work to include multimodal data sources is another promising direction. While this study focused on CHMs, incorporating additional remote sensing data, such as multispectral or hyperspectral imagery, LiDAR-derived metrics, or ecological indicators (e.g., biodiversity indices), could improve model robustness and prediction accuracy. Such multimodal approaches might capture finer details of forest structure and composition that CHMs alone cannot provide.

To enhance the practicality of the model, temporal analysis could also be explored. Forest naturalness is not static—it evolves over time due to natural growth, disturbances, and management interventions. Future work could investigate the use of time-series data to track changes in naturalness and predict trends. Temporal models, such as recurrent neural networks or temporal convolutional networks, could be adapted for this purpose, offering a dynamic perspective on forest health.

Finally, this study provides an important foundation for real-world applications of automated forest evaluation. Future efforts could focus on deploying the model in collaboration with forest management agencies or conservation organizations, testing its performance in diverse environments and under practical conditions. This would also enable the integration of feedback to improve the model's usability and scalability.

By addressing these challenges and opportunities, future work can build on the findings of this thesis to develop more robust, scalable, and interpretable models for automated forest naturalness evaluation. This would ultimately contribute to more effective forest management practices and broader ecological sustainability efforts.

Bibliography

[1] Erwin Führer. "Forest functions, ecosystem stability and management". In: *Forest Ecology and Management* 132.1 (June 15, 2000), p. 30. ISSN: 0378-1127. DOI: 10.1016/S0378-1127(00)00377-7. URL: https://www.sciencedirect.com/science/article/pii/S0378112700003777 (visited on 06/02/2024).

[2] Z. Wang, E.P. Simoncelli, and A.C. Bovik. "Multiscale structural similarity for image quality assessment". In: *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003.* Thirty-Seventh Asilomar Conference on Signals, Systems and Computers 2003. Vol. 2. Pacific Grove, CA, USA: IEEE, 2003, 1398–1402 Vol.2. ISBN: 0-7803-8104-1. DOI: 10.1109/ACSSC.2003.1292216.

[3] Jonathan Long, Evan Shelhamer, and Trevor Darrell. "Fully Convolutional Networks for Semantic Segmentation". In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015, pp. 3431–3440. URL: https://openaccess.thecvf.com/content_cvpr_2015/html/Long_Fully_Convolutional_Networks_2015_CVPR_paper.html (visited on 05/05/2024).

[4] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015.* Ed. by Nassir Navab et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2015, pp. 234–241. ISBN: 978-3-319-24574-4. DOI: 10.1007/978-3-319-24574-4_28.

[5] Gellert Mattyus, Wenjie Luo, and Raquel Urtasun. "DeepRoadMapper: Extracting Road Topology From Aerial Images". In: Proceedings of the IEEE International Conference on Computer Vision. 2017, pp. 3438–3446. URL: https://openaccess.thecvf.com/content_iccv_2017/html/Mattyus_DeepRoadMapper_Extracting_Road_ICCV_2017_paper.html (visited on 11/30/2024).

[6] Tsung-Yi Lin et al. *Focal Loss for Dense Object Detection.* version: 2. Feb. 7, 2018. DOI: 10.48550/arXiv.1708.02002. arXiv: 1708.02002[cs]. URL: http://arxiv.org/abs/1708.02002 (visited on 06/06/2024).

[7] FAO. *Global Forest Resources Assessment 2020.* 1st ed. Rome, Italy: FAO ; 2020. ISBN: 978-92-5-132974-0. URL: https://openknowledge.fao.org/handle/20.500.14283/ca9825en (visited on 06/01/2024).

[8] Ewa Grabska, David Frantz, and Katarzyna Ostapowicz. "Evaluation of machine learning algorithms for forest stand species mapping using Sentinel-2 imagery and environmental data in the Polish Carpathians". In: *Remote Sensing of Environment* 251 (Dec. 15, 2020), p. 112103. ISSN: 0034-4257. DOI:

10.1016/j.rse.2020.112103. URL: https://www.sciencedirect.com/science/article/pii/S0034425720304764 (visited on 06/05/2024).

[9] Huimin Huang et al. "UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation". In: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). ISSN: 2379-190X. May 2020, pp. 1055–1059. DOI: 10.1109/ICASSP40776.2020.9053405. URL: https://ieeexplore.ieee.org/abstract/document/9053405 (visited on 02/15/2024).

[10] L. Bragagnolo, R. V. da Silva, and J. M. V. Grzybowski. "Amazon forest cover change mapping based on semantic segmentation by U-Nets". In: *Ecological Informatics* 62 (May 1, 2021), p. 101279. ISSN: 1574-9541. DOI: 10.1016/j.ecoinf.2021.101279. URL: https://www.sciencedirect.com/science/article/pii/S1574954121000704 (visited on 08/24/2024).

[11] Gabriel D. Caffaratti et al. "Improving forest detection with machine learning in remote sensing data". In: *Remote Sensing Applications: Society and Environment* 24 (Nov. 1, 2021), p. 100654. ISSN: 2352-9385. DOI: 10.1016/j.rsase.2021.100654. URL: https://www.sciencedirect.com/science/article/pii/S2352938521001907 (visited on 06/03/2024).

[12] Bowen Cheng et al. "Boundary IoU: Improving Object-Centric Image Segmentation Evaluation". In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021, pp. 15334–15342. URL: https://openaccess.thecvf.com/content/CVPR2021/html/Cheng_Boundary_IoU_Improving_Object-Centric_Image_Segmentation_Evaluation_CVPR_2021_paper.html (visited on 06/06/2024).

[13] Ning Fang et al. "Assessment of Forest Ecological Function Levels Based on Multi-Source Data and Machine Learning". In: *Forests* 14.8 (Aug. 2023). Number: 8 Publisher: Multidisciplinary Digital Publishing Institute, p. 1630. ISSN: 1999-4907. DOI: 10.3390/f14081630. URL: https://www.mdpi.com/1999-4907/14/8/1630 (visited on 06/03/2024).

[14] Ningthoujam Johny Singh and Kishorjit Nongmeikapam. "Semantic Segmentation of Satellite Images Using Deep-Unet". In: *Arabian Journal for Science and Engineering* 48.2 (Feb. 1, 2023), pp. 1193–1205. ISSN: 2191-4281. DOI: 10.1007/s13369-022-06734-4. URL: https://doi.org/10.1007/s13369-022-06734-4 (visited on 06/05/2024).

[15] Marco L. Della Vedova and Mattias Wahde. *Naturalness Indicators of Forests in Southern Sweden derived from the Canopy Height Model*. Oct. 14, 2024. DOI: 10.48550/arXiv.2410.10465. arXiv: 2410.10465. URL: http://arxiv.org/abs/2410.10465 (visited on 11/17/2024).

DEPARTMENT OF MECHANICS AND MARITIME SCIENCES
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2024
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY