remote sensing

MDPI

Article

Deep Neural Networks with Transfer Learning for Forest Variable Estimation Using Sentinel-2 Imagery in Boreal Forest

Heikki Astola *, Lauri Seitsonen †, Eelis Halme †, Matthieu Molinier † and Anne Lönnqvist

VTT Technical Research Centre of Finland Ltd., P.O. Box 1000, FI-02044 Espoo, Finland; lauri.seitsonen@vtt.fi (L.S.); eelis.halme@vtt.fi (E.H.); matthieu.molinier@vtt.fi (M.M.); anne.lonnqvist@vtt.fi (A.L.)

* Correspondence: heikki.astola@vtt.fi

† These authors contributed equally to this work.

Abstract: Estimation of forest structural variables is essential to provide relevant insights for public and private stakeholders in forestry and environmental sectors. Airborne light detection and ranging (LiDAR) enables accurate forest inventory, but it is expensive for large area analyses. Continuously increasing volume of open Earth Observation (EO) imagery from high-resolution (<30 m) satellites together with modern machine learning algorithms provide new prospects for spaceborne large area forest inventory. In this study, we investigated the capability of Sentinel-2 (S2) image and metadata, topography data, and canopy height model (CHM), as well as their combinations, to predict growing stock volume with deep neural networks (DNN) in four forestry districts in Central Finland. We focused on investigating the relevance of different input features, the effect of DNN depth, the amount of training data, and the size of image data sampling window to model prediction performance. We also studied model transfer between different silvicultural districts in Finland, with the objective to minimize the amount of new field data needed. We used forest inventory data provided by the Finnish Forest Centre for model training and performance evaluation. Leaving out CHM features, the model using RGB and NIR bands, the imaging and sun angles, and topography features as additional predictive variables obtained the best plot level accuracy (RMSE% = 42.6%, $|BIAS\%|$ = 0.8%). We found 3×3 pixels to be the optimal size for the sampling window, and two to three hidden layer DNNs to produce the best results with relatively small improvement to single hidden layer networks. Including CHM features with S2 data and additional features led to reduced relative RMSE (RMSE% = 28.6–30.7%) but increased the absolute value of relative bias ($|BIAS\%|$ = 0.9–4.0%). Transfer learning was found to be beneficial mainly with training data sets containing less than 250 field plots. The performance differences of DNN and random forest models were marginal. Our results contribute to improved structural variable estimation performance in boreal forests with the proposed image sampling and input feature concept.

Keywords: deep learning; deep neural networks; transfer learning; sentinel-2; boreal forest; forest variables; stem volume

Citation: Astola , H.; Seitsonen, L.; Halme, E.; Molinier, M.; Lönnqvist, A. Deep Neural Networks with Transfer Learning for Forest Variable Estimation Using Sentinel-2 Imagery in Boreal Forest. *Remote Sens.* **2021**, *13*, 2392. https://doi.org/10.3390/rs13122392

Academic Editor: Brigitte Leblon

Received: 5 May 2021
Accepted: 14 June 2021
Published: 18 June 2021

1. Introduction

Forest structural variables, such as the growing stock volume (GSV), offer valuable information for forest conditions and long-term development trends [1,2]. This information can be utilized by various stakeholders, e.g., for sustainable forest management and commercial utilization, carbon balance assessment or predictions of ecosystem response to climate change [3,4]. In situ measurements provide transparent, comparable and consistent information on the availability of forest resources for the development of forest-associated policies and to support decision making [5]. Combined with remote sensing instruments, the sparse network of field plots can provide analyses for larger areas [6]. Airborne LiDAR measurements and aerial imagery enable accurate forest inventory and monitoring, but airborne acquisitions are expensive, limited in spatial and temporal coverage, and can

contain data gaps, thus restricting their utility for wall-to-wall mapping of large areas [7]. Consequently, open satellite imagery becomes appealing for cost-effective wide-area forest inventories. However, there is a constant need to improve the accuracy of spaceborne forest inventory for it to be more attractive for a wider selection of users.

The Sentinel-2 mission of the European Copernicus program offers free optical imagery at 10 to 60 m spatial resolution, with 13 spectral bands and a revisit frequency of five days at the equator [8], enabling cost-effective and frequent forest monitoring at a large scale. The volume of openly available Earth Observation (EO) data through, e.g., ESA and USGS is increasing exponentially [9], which induces the need for reliable, robust, and transferable machine learning techniques for large-scale information extraction [10]. In recent years, the remote sensing community has shown great interest in non-parametric methods, particularly kernel-based regression methods and neural networks. For large area mapping, non-parametric methods are considered more versatile than traditional parametric methods [11] because they only rely on the available data and do not assume any fixed functional form [12]. Therefore, they typically perform better in the presence of a nonlinear relationship between the explanatory and response variables [13]. In particular, deep learning methods have seen a great increase in popularity for optical satellite image analysis [14–16]. Typical applications of deep learning in remote sensing include image classification and segmentation tasks, object detection or hyperspectral image analysis [17,18], and the most commonly used deep learning models include convolutional neural network (CNN) and recurrent neural network (RNN) [15].

Despite promising results for classification tasks, deep learning has not yet been used extensively for regression in environmental studies [19]. A recent review on CNNs in vegetation remote sensing [20] included just ten regression studies, only two of which used satellite images as the primary source. Several regression studies focus on estimating continuous biophysical variables in the agricultural sector. Reference [21] used CNNs and RNNs to predict soybean yield in the USA, while [22] used CNNs to estimate crop yields in India. In Finland, reference [23] estimated wheat and barley yield using CNNs. Deep learning has also been tested in estimating strawberry yield [24] or predicting soil moisture [25].

Few studies have used deep learning for estimating forest aboveground biomass (AGB). For instance, [26] used stacked sparse autoencoders (SSAE) to estimate AGB in China, in mixed broadleaved and coniferous forests characterized by a subtropical monsoon climate. Their SSAE model outperformed traditional methods, such as RF, support vector machine (SVM), and k-nearest neighbor (k-NN). Their results were confirmed by [27] in the same area in China. Forest biomass has been estimated using deep learning also in the U.S. [28] and in Spain [29]. In both studies, the use of deep learning was found to be a promising approach for further investigation. Authors in [7] introduced a deep multi-input recurrent convolutional neural network called Chimera to classify forest and land cover to five classes, and to estimate forest structure metrics (including AGB) simultaneously using satellite and aerial imagery, and ancillary data as model inputs. The Chimera predictor outperformed RF and SVM in both tasks when using the same input data for all classification and regression tasks.

Other continuous forest variables that have been estimated using deep learning models recently are growing stock volume and vegetation height. Authors in [30] were able to predict GSV accurately (RMSE = 30.5 m^3/ha) in China using tree pixel information extracted from on-ground digital camera images. Authors in [31] trained a CNN to extract suitable textural and spectral features from Sentinel-2 images and used the model for spatial prediction of vegetation height in Switzerland and Gabon. They obtained RMSE values of 3.4 m and 5.6 m for tree height estimates in the two countries respectively, with a CNN model consisting of nearly 20 million trainable weights. Recently in Finland, boreal forest variables have been estimated with neural networks [32] and kernel-based regression methods [33,34]. Although deep learning has been used for estimating continuous forest

variables, studies focusing on the boreal forest biome are still lacking. The present study will contribute to closing this research gap.

Usually with machine learning algorithms, the assumption is that the distribution of the data used for model training is the same as the data that the model is applied to [35]. However, transfer learning [36,37] allows these distributions to be different. In short, transfer learning is a fine-tuning method that aims to improve learning in a new target data set utilizing information and knowledge learned in a data set from another domain. In remote sensing, transfer learning offers appealing means to refine models for new target distributions without the need to organize expensive and slow field data campaigns, or compromising on model accuracy. It may be used to refine an existing model for a new target data distribution, e.g., for a new geographical area, or to update a model outdated due to, e.g., forest growing or management. Transfer learning can also be used to transform an existing model for a new target variable, or with a new type of source data (e.g., another satellite). For deep learning models in remote sensing, transfer learning is a useful tool for training a comprehensive target network and avoiding over-fitting, even if the target domain data set would be smaller than the source data set [38].

Studies combining transfer learning and deep learning in remote sensing are still relatively rare, especially for regression. Reference [39] estimated soybean yield in Brazil utilizing transfer learning with an RNN model trained on Argentinian soybean harvests. Their transfer learning model outperformed all other used models, which were trained only on data from Brazil. Reference [40] found transfer learning to be extremely valuable in retrieving information on small-scale urban structures. They reported that transfer learning from one optical domain (QuickBird) to another optical domain (Sentinel-2) worked well and improved segmentation accuracy.

We applied deep neural networks for continuous forest variable estimation in the boreal forest biome. The motivation to use deep neural networks in our study was their ability to automatically extract useful features from complex data, which is beneficial when combining multiple data sources of different natures, and the simplicity to implement transfer learning. We investigated the benefit of combining the Sentinel-2 image data sampled from an extended spatial context of a target pixel with information of the satellite imaging and sun angles and with topography data. We assumed this to improve model performance when compared with the previously used methods such as simple neural networks with pixel-wise image sampling. Furthermore, as the potential of transfer learning in remote sensing has not been widely explored yet, especially in the forest sector, we investigated its applicability to refine models from one spatial region to another characterized with different data distribution, aiming to reduce the need for additional expensive field work. The special interest was to evaluate the amount of new training data needed to obtain models with accuracy close to models that are trained with extensive training data. An important goal of our study was also to obtain information on how to apply the developed deep learning concept and transfer learning in practice.

The main objective of this study was to investigate the performance and practical usability of deep neural networks for forest variable prediction in the boreal forest biome by combining predictive data from multiple sources. The experiments to achieve this objective were (1) to investigate the relevance of different input features, the effect of DNN depth, the amount of training data, and the size of image data sampling window to model prediction performance; (2) to examine the relation of available reference data (number of training samples) to forest variable prediction accuracy; (3) to examine the benefits of using transfer learning for a new geographic region with a limited set of reference data.

2. Materials

2.1. Study Site

Our study area (Figure 1) was located in Central Finland and covered an area of four forestry districts: Central Ostrobothnia (Keski-Pohjanmaa), Central Finland (Keski-Suomi),

North Savo (Pohjois-Savo), and North Karelia (Pohjois-Karjala). The total area covered by the four districts is 68,315 km^2, and it extended to southern and central boreal zones.

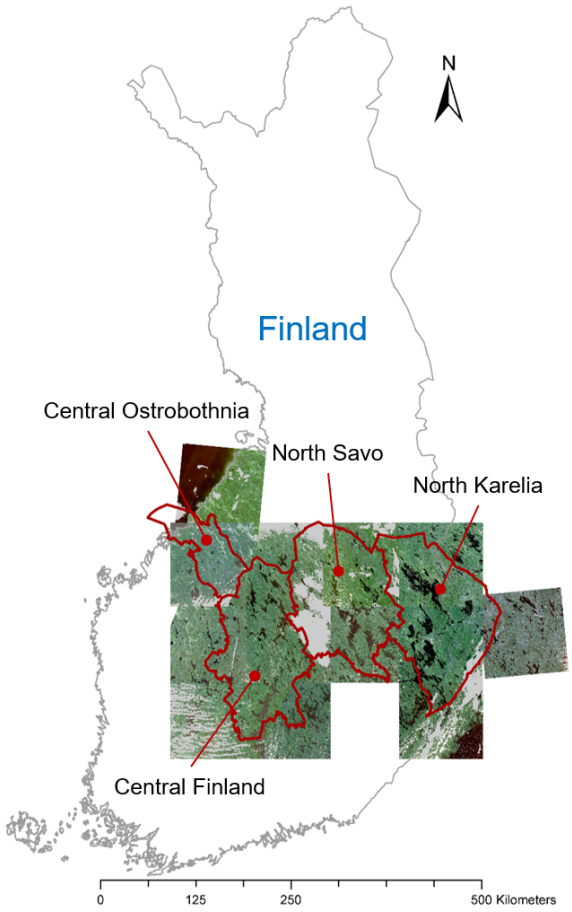


Figure 1. The study area in Central Finland with the 22 Sentinel-2 images overlaid, and the four forestry districts delineated in red. The total area of the four districts is 68,315 km^2.

The study sites are typical Finnish boreal forests. The topography is relatively flat with the changes in elevation increasing towards the east. The forest area is conifer-dominated with Scots pine (*Pinus sylvestris*) and Norway spruce (*Picea abies*) as the main overstory species. The deciduous trees usually occur as mixed-species stands. Birches are the most common deciduous species, especially Silver birch (*Betula pendula*) and Downy birch (*Betula pubescens*).

2.2. Field Reference Data

The Finnish Forest Centre provided the field plots that we used as reference data. The plots were distributed over the four forestry districts. The data included information on stem volume, mean diameter at breast height, mean height, basal area, age, stem number, and additional stand information, such as development class, dominant tree species, proportion on timber, and regeneration situation.

The reference data were acquired during 2016–2017 using three different plot radii: 9 m in young and advanced managed forests with a relatively high tree density, 12.62 m in a forest with a low stem density and typically high volume due to the mature development stage, and 5.64 m in seedling stands. A total of 42.7% field plots used in the study were on green heathland, 27.1% on dryish heathland, and 24.1% on grovy heathland, with the remaining 5.8% residing on dry heathland, groves, and barren heathlands. A total of 34.0% of the plots were located on grown up growing forests, 25.2% on young growing forests,

22.4% on recently planted forest stands, and 18.4% on mature forests. Table 1 shows the mean and standard deviation of major forest variables in the four forestry districts.

Table 1. Forest variable means and standard deviations (Stdev) for the four forestry districts. Species-specific values for spruce and broadleaved are indicated by "spr" and "bl", respectively. Numbers of used field plots for each forestry district are also shown.

	Central Ostrobothnia		Central Finland		North Savo		North Karelia	
Forest Variable	**Mean**	**Stdev**	**Mean**	**Stdev**	**Mean**	**Stdev**	**Mean**	**Stdev**
Total Stem Volume (m^3/ha)	100.8	84.6	130.3	104.4	165.7	122.4	140.1	117.0
Stem Volume—pine (m^3/ha)	69.4	67.0	69.1	86.4	64.1	91.4	48.4	75.5
Stem Volume—spr (m^3/ha)	14.1	43.6	36.7	68.1	64.3	113.2	59.9	93.0
Stem Volume—bl (m^3/ha)	17.5	33.7	24.3	41.3	37.2	61.1	31.8	45.9
Basal Area—G (m^2/ha)	14.0	8.7	16.6	10.2	19.4	11.1	17.4	10.8
Basal Area—pine (m^2/ha)	9.5	7.3	8.5	9.2	7.4	9.3	5.7	7.8
Basal Area—spr (m^2/ha)	1.9	4.8	4.8	7.7	7.3	11.3	7.4	9.3
Basal Area—bl (m^2/ha)	2.6	4.4	3.2	4.8	4.7	6.9	4.3	5.5
Stem Diameter—D (cm)	15.0	7.7	15.5	7.8	17.2	9.5	15.6	8.8
Stem Diameter—pine (cm)	15.0	8.8	13.1	10.4	11.5	11.9	10.7	11.2
Stem Diameter—spr (cm)	5.5	8.2	9.7	8.6	10.2	10.6	12.0	9.4
Stem Diameter—bl (cm)	7.4	7.2	8.9	8.1	9.7	9.1	9.9	8.2
Tree Height—H (m)	12.2	5.7	13.4	6.3	14.8	7.2	13.5	6.7
Tree Height—pine (m)	11.7	6.5	10.8	8.2	9.4	9.3	8.9	8.9
Tree Height—spr (m)	4.4	6.3	8.2	6.8	8.5	8.2	9.9	7.2
Tree Height—bl (m)	7.5	6.5	9.5	7.6	10.2	8.1	10.5	7.5
Age—T (years)	54.0	30.5	39.1	21.5	43.6	27.1	41.2	27.2
Age—pine (years)	53.4	34.1	32.9	27.0	30.9	33.1	28.8	32.3
Age—spr (years)	20.1	28.7	27.5	22.7	26.4	26.1	33.7	26.8
Age—bl (years)	27.2	24.9	22.6	19.3	25.3	23.0	25.7	21.8
Stem Number—N (count)	2196	3351	2287	3931	2625	4200	2685	4134
Stem Number—pine (count)	1049	1401	639	1043	630	1222	343	564
Stem Number—spr (count)	169	382	466	586	547	694	697	692
Stem Number—bl (count)	976	2683	1184	3594	1448	3568	1645	3852
Number of field plots	374		387		362		712	

We used four target variables (stem volume, basal area, (mean) stem diameter, and (mean) tree height) and their species-specific components (pine, spruce, and broadleaved) in our study. We focused on stem volume for most of our tests, but we also produced performance figures for all of them with our forest variable prediction concept.

2.3. Optical Satellite Imagery

We used the optical remote sensing data from the Sentinel-2 Multispectral Instrument (MSI). A set of 22 Level-1C products (https://sentinel.esa.int/web/sentinel/user-guides/sentinel-2-msi/product-types/level-1c; accessed 6 June 2021) over Southern/Central Finland were downloaded from the Copernicus Open Access API Hub (Figure 1), with the acquisition time window set in summer months (1 June 2017–6 September 2017).

To perform the necessary atmospheric correction, we could not rely on approaches requiring dark vegetation targets, such as the Dense Dark Vegetation (DDV) algorithm [41]. Dark objects, typically pure old spruce forest stands, were very rare in the target forestry districts mostly populated by managed pine forests on mineral soil. The selected Sentinel-2 images were also partly cloudy, which reduced the possibilities of finding the required reference dark target stands. Consequently, we computed the bottom-of-atmosphere (BOA) reflectances by applying the SMAC4 radiation transfer code [42] using a constant atmospheric optical density (AOD) parameter of 0.01, and masked out clouds in the image.

We used the S2 image channels at 10 m and 20 m spatial resolution, with the 20 m channels re-sampled to 10 m (Table 2). In addition, we included image metadata and auxiliary data as model input features. These data are listed in (Table 3).

Table 2. Spectral bands of Sentinel-2 MSI. The S2 bands with 10 m or 20 m spatial resolution were used in this study.

Band	Description	Spectral Range (nm)	Resolution (m)
1	Coastal aerosol	433–453	60
2	Blue	458–523	10
3	Green	543–578	10
4	Red	650–680	10
5	Vegetation Red Edge (RE1)	698–713	20
6	Vegetation Red Edge (RE2)	733–748	20
7	Vegetation Red Edge (RE3)	773–793	20
8	Near-Infrared (NIR)	785–900	10
9	Narrow NIR (nNir)	855–875	20
10	Water vapor	935–955	60
11	Shortwave infrared - Cirrus	1360–1390	60
12	Shortwave infrared (SWIR1)	1565–1655	20
13	Shortwave infrared (SWIR2)	2100–2280	20

Table 3. The different input data used as model predictor variables. Each data source was used as model inputs either separately (referred to by data set number) or in data groups (referred to by group code). The dimensions of the S2 image data varied depending on the image channels used as well as the image sampling window size (1×1, 3×3 or 5×5 pixels), which also affected the dimensions of terrain slope data. The CHM statistics (mean, median, stdev) were computed for 1×1 and 3×3 S2 window areas.

Data Set#/Group	Input Data	Data Dimension
1/I	S2 10 m and 20 m band data	4, 10, 36, 90, 100, or 250
2/A	S2 acquisition date	2
3/B	S2 target pixel location (Lat, Lon)	2
4/C	S2 angles (azimuth and elevation)	2
5/C	S2 sun angles (azimuth and elevation)	2
6/A	S2 acq. time difference to field measurement (days)	1
7/D	DEM data	1
8/E	CHM data	6
9/D	Terrain slope	2, 18, or 50

2.4. Elevation Models

2.4.1. Digital Elevation Model

A digital elevation model (DEM) was acquired from the National Land Survey of Finland. The DEM depicts the elevation of the ground surface of the whole of Finland in relation to sea level, with a grid size of 10 m $\times$ 10 m and accuracy of elevation data of 1.4 m. Slopes in north–south and east–west directions were computed by shifting the DEM 10 m along these directions then subtracting the shifted DEM from the original DEM.

2.4.2. Canopy Height Model

The canopy height model (CHM) computed from LiDAR measurements was acquired at 1 m pixel resolution by the Finnish Forest Centre. The mean, median, and standard deviation values of CHM data were computed for 1×1 and 3×3 Sentinel-2 pixel windows (i.e., 10 m $\times$ 10 m and 30 m $\times$ 30 m). These data were included in combination with the other data sources to evaluate the approximate level of accuracy that can be obtained when they are available for forest variable modeling (Table 3).

2.5. Training, Validation, and Test Sets

The 2970 reference field plots were divided into training, validation, and test sets using stratified random sampling with five strata of equal bin widths and relative set sizes of 50%, 30%, and 20% for training, validation, and test sets, respectively. The stratification variable was total stem volume, and the bin edges were 0, 160, 320, 480, 640, and 800 $m^3\,ha^{-1}$. The resulting numbers of field plots were 1488, 895, and 587 for training, validation, and test sets, respectively.

According to [43], the spatial auto-correlation effect in Finnish forests reduces to close to zero within 200–300 m. Adopting this information, we set the minimum distance requirement of field plots belonging to different data sets (i.e., training, validation, or test set) to 250 m, and we removed all field plots in training and validation sets that did not exceed this threshold. This reduced the number of field plots in the resulting training and validation sets to 1129 and 338, respectively. In order to keep enough field plots in the test set, we kept the set of test field plots intact. The resulting minimum distances between the three sets were 252.6 m (from training to validation), 251.1 m (from training to test), and 256.9 m (from validation to test).

Multiple Sentinel-2 images were included from the test area for the selected period, with the number of images for each field plot location varying from 1 to 22. After combining the image data with the field plot data, the resulting data set included in total 10,686 data vectors. The sizes for training, validation, and test sets were 5826, 1826, and 3034 data vectors, respectively.

Some of the Sentinel-2 images contained heavy cloud cover (up to 86% of image area), and the image pixels that were not masked out as clouds were often contaminated by haze or fog that could not be removed or compensated. To avoid the use of noisy data in model training, the maximum cloud cover percentage of the training and validation images was set to 70%. The same threshold was set to 10% for the test set data to use only the best quality data for computing performance measures.

With the maximum cloud percentage limitation, the number of data vectors in training, validation, and test set reduced to 5762, 1804, and 908, respectively. To obtain a unique test set, we randomly selected a single data vector from those field plot locations that included multiple data vectors, and thus the test data set size reduced from 908 to 368 vectors. The numbers of field plots in the resulting training, validation, and test sets were then 1129, 338, and 368, respectively.

This compound set of data that consisted of the field plots with the associated image data, image metadata, DEM, and CHM was used throughout the study. The data were combined into a feature matrix that was stored in text format. A summary of the reference data sets with statistics of main forest variables is shown in Table 4.

Table 4. Forest variable means and standard deviations (Stdev) for training, validation, and test data sets. N_{tr}, N_v, N_{te} = Number of field plots in training, validation, and test sets, respectively.

	Training Set (N_{tr} = 1129)		**Validation Set (N_v = 338)**		**Test Set (N_{te} = 368)**	
Forest Variable	**Mean**	**Stdev**	**Mean**	**Stdev**	**Mean**	**Stdev**
Stem Volume, V (m^3/ha)	136.1	113.8	130.3	107.7	136.4	108.3
Basal Area, G (m^2/ha)	17.1	10.8	16.3	9.9	17.1	10.0
Stem Diameter, D (cm)	15.6	8.6	15.8	8.6	16.2	8.7
Tree Height, H (m)	13.4	6.6	13.6	6.6	13.7	6.5
Age, T (years)	43.3	27.6	44.0	27.6	45.5	26.3
Stem Number, N (count)	2532	3923	2487	4355	2361	3691

2.5.1. Training Data Subsets

To investigate the relationship between the prediction performance and the amount of data used for model training, we split the training and validation reference data into smaller subsets. We ordered the field plot data within these two sets according to increasing

total stem volume and split the ordered data sets into 30 sub-groups (strata) with an equal number of field plots (37 or 11 for training and validation sets, respectively) in each subgroup. As the numbers of field plots in training and validation sets (1129 and 338, respectively) were not divisible by the number of strata, a 31st group included the remaining field plots with the highest stem volumes (19 plots for training, 8 for validation). Then we randomly assigned the subset labels (1 to 37) for training and (1 to 11) for validation data within the ordered subgroups. The resulting 37 training and 11 validation data subsets then contained 30 or 31 field plots, and their total stem volume distribution followed roughly the distribution of the entire field plot data set.

3. Methods

3.1. Study Concept

We defined a series of tests to investigate the effect of different factors on the model prediction performance. Such factors included the size of the Sentinel-2 image sampling window, the number of DNN hidden layers (DNN depth), the amount of reference data for model training, and the effect of auxiliary data (e.g., topography data, and imaging angles or acquisition dates). In addition, we used transfer learning to fine-tune DNN models to evaluate the applicability of the models in other geographical regions.

The input data for the forest variable models were a combination of data from Sentinel-2 satellite, topography data (digital elevation model (DEM) and terrain slope), and canopy height model (CHM). The data from Sentinel-2 included image pixel values from spectral bands with 10 and 20 m spatial resolution, acquisition date, time difference between acquisition and field measurement, imaging angle, sun angle, and target pixel location (Lat, Long). The spatial resolution of the model output was 10 m, i.e., the pixel size of Sentinel-2 RGBNIR bands.

The motivation for selecting group A auxiliary features was to compensate for seasonal effects (feature #2/A; Table 3) and to capture the effect of vegetation growth between image acquisition and field measurement dates (feature #6/A). For selecting group B or auxiliary feature #3 the idea was to capture areal variations in the landscape, mostly applying to the west–east direction in our study area. The reason to include groups C and D (features #4, 5, 7 and 9) was to compensate the differences in the pixel reflectance values due to variation in terrain elevation (effect of shadowing), if these data were combined with the windowed S2 image sampling.

The performance of the DNN models was compared with results produced with commonly used random forest (RF) models [44]. The performance of the transfer learning models was compared with the results obtained with models trained from scratch using the same training data. Total stem volume (V) was used as the target variable in most of the tests, although we produced DNN models also for a wider set of forest variables to assess the improvement obtained with the proposed data fusion concept when compared with a simple data model.

3.2. Deep Neural Network (DNN) Models

A fully connected deep neural net (DNN) implemented with Keras deep learning library [45] using Tensorflow [46] platform was used for the forest variable regression. We used rectified linear units (ReLU) as the activation functions for both hidden and output layer neurons.

We used the Adam algorithm [47] as the optimizer in the network training. The parameter values used for Adam were beta1 = 0.9 , beta2 = 0.999. To prevent the network from overfitting the training data, we used regularization and weight decay in combination with early stopping to interrupt training at an appropriate point. The L2 regularization method [48] was used, after checking that it produced roughly an equal regularization effect on network training as dropout [49]. The weight decay rate for the Adam algorithm was set to learning rate/number of epochs. As the early stopping criterion, the error

measure to monitor was the mean absolute error on the validation set. We defined 0.001 as its minimum improvement (min_delta) with the patience parameter set to three epochs.

3.2.1. Transfer Learning

In transfer learning, a model developed for a given source data set is used as the starting point for tuning a model on a target data set. Two strategies are commonly adopted: either the network weights learned in the source data set are used as initial weights to retrain (or fine-tune) the entire network on the target data set, or a part of the network weights are fixed and only the non-fixed weights are retrained. The non-fixed weights may also be initialized randomly instead of using the pre-trained weight values. The network structure may also be changed by adding or removing layers and/or outputs to the network.

We tested transfer learning using a DNN model trained with reference forest data from the Central Ostrobothnia district, which has a lower mean stem volume and different species composition than the other three districts (Table 1). The data used to fine-tune the pre-trained model consisted of the field plot data from the three other forestry districts: Central Finland, North Savo, and North Karelia. We used the field plot data subsets described in Section 2.5.1 to evaluate the effect of a limited amount of training data to transfer learning. We compared the transfer learning model with a model of equal structure, and that was trained starting from randomly initialized network weights ('training from scratch') with an identical training data set.

3.2.2. Hyper-Parameter Search

To find the best hyper-parameter values for model training, we used the empirical cost function presented by [32] that combined relative root-mean-square error, absolute value of relative bias, and coefficient of determination. We used a factor of 3 instead of 10 for the bias term in this study:

$$\text{cost} = \frac{RMSE\% + 3 \cdot abs(BIAS\%)}{0.5 + R^2} \quad (1)$$

The cost for each hyper-parameter combination was computed using the mean values of validation data set performance figures from 25 modeling iterations, and the hyper-parameter combination with minimum cost was selected for subsequent model training. A modeling iteration wrapper [50] was then run with the found optimal hyper-parameters for 100 times for each separate test setup (see Figure 2).

The learning rate was searched among values [0.00001, 0.0001, 0.0003, 0.0005, 0.001, 0.005]. For mini-batch size, two ranges of search values were used: [8, 16, 32, 64, 128] for the smallest sets of training data, and [32, 64, 128, 256] for others. The number of training epochs was set to 250, which was considered adequate for convergence as the early stopping criterion interrupted training before this limit was reached. We used an L2-regularization weight penalty value of L2 = 0.15 for each test run.

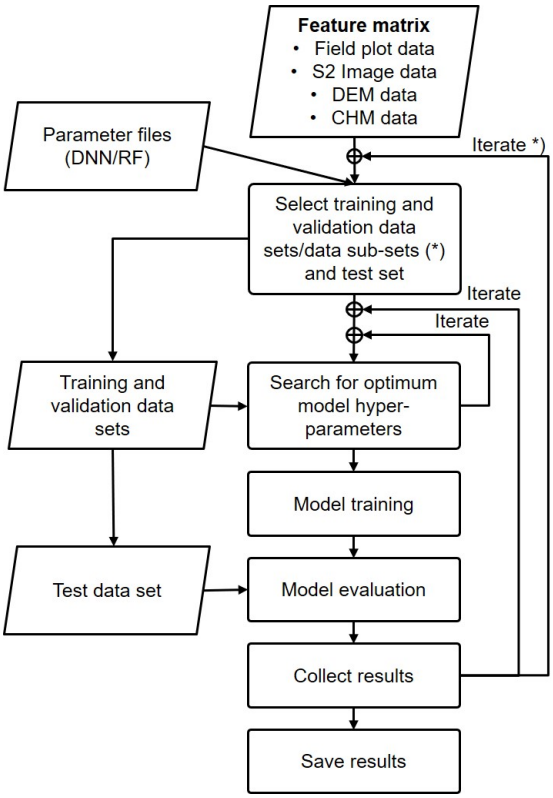


Figure 2. Flowchart of the iteration wrapper used in tests. (*) Refers to iteration including subsets of whole training/validation data (i.e., tests with limited training data and transfer learning tests).

3.3. Test Setups

We defined different test setups to evaluate the effect of various factors to the DNN model performance. The contribution of each factor to model performance was tested by individually changing the parameters related to the factor being investigated, while fixing other parameters to default values that were found close to optimal for each factor in preliminary tests (Table 5).

Table 5. Different factors studied for the performance of the deep neural network model. These factors are listed here together with the parameter default values. Refer to the corresponding subsection for more details on test setups.

Nbr	Factor	Default Value	Subsection
1	Different input data sources (features)—Table 3	All except CHM data	Section 3.3.1
2	Size of image sampling window	3 × 3 pixels	Section 3.3.2
3	Number of field plots	All (train + valid: 1467)	Section 3.3.3
4	Number of DNN hidden layers	3	Section 3.3.4

3.3.1. Different Input Data Sources

Input data of the estimation model, i.e., the features that were used in the study, included the data sources listed in Table 3. To investigate the individual significance of different input data, each of the sources was added one at a time to the used Sentinel-2 image channels. For the Sentinel-2 data we used two different combinations: spectral bands with 10 m spatial resolution (i.e., B2, B3, B4, and B8) or spectral bands with 10 and 20 m spatial resolution (i.e., B2, B3, B4, B5, B6, B7, B8, B8A, B11, and B12).

3.3.2. Size of Image Sampling Window

Three different sampling sizes were used: 1×1, 3×3, and 5×5 pixels. The model performance was assumed to improve with the windowed approach, as the effects of shadows in images would be taken into account in combination with DEM and slope information.

3.3.3. Minimal Amount of Reference Data

In this experiment, our goal was to find the smallest amount of reference data (field plots) for proper training of the models. To evaluate this, the training and validation data were drawn randomly from the 37 training and 11 validation data subsets described in Section 2.5.1, while aiming at a nominal number of field plots N = 60, 90, 120, 150, 210, 300, 510, 750, and 990 in the combined training and validation sets. The ratio of training and validation data varied from 50/50 to 67/33 with the different nominal amounts of field plots, with a more equal split for the smallest nominal values.

This random selection and the subsequent modeling iteration run was repeated 10 times for each nominal number of field plots N (i.e., 10 random combinations of N field plots). For each of these 10 iterations, one optimal hyper-parameter search loop was performed, followed by 25 DNN training iteration runs with the same hyper-parameters. The results for the training data sub-sample tests were computed as mean values from the 10×25 modeling runs, with the standard deviation of the means from the 10 different training data combinations.

3.3.4. Number of DNN Hidden Layers

A selection of fully connected networks with a varying number of hidden layers was tested, between 1 and 20 hidden layers (Table 6). In the tests, we used two sets of model inputs: one set with all ten Sentinel-2 (S2) image channels and all auxiliary input features, and another set with the 10 m spatial resolution channels (i.e., B2, B3, B4, and B8) and S2 imaging/sun angles and DEM/slope features. The 3×3 image sampling was used for both input feature groups. The numbers of model inputs were 118 and 59, respectively.

The numbers of neurons in hidden layers were empirically searched for stable convergence of the network, aiming to keep the number of network parameters low, while preserving the network ability to learn the desired prediction task without overfitting to the fairly limited set of reference data. The number of neurons in the first hidden layer was set to <*number of inputs*+3> as this was found to be close to optimal for the single hidden layer network. The number of neurons in subsequent hidden layers followed a decreasing trend from the first to the last layer, as in [28]. For instance, the structure of an eight hidden layer DNN was [59-**62-32-24-16-12-9-6-3**-1] for the input configuration **S2-4-CD** (the number of neurons in hidden layers in bold; for the model structure coding, see Section 4). Table 6 shows the number of hidden layers and the number of weights in the used networks.

Table 6. Different depths of neural networks tested in the study. Number of hidden layers shown with the number of trainable weights for two different sets of network inputs (**S2-4-CD w3** and **S2-10-ABCD w3**) that were used in the tests. In addition, the ratio of the number of training data vectors (*Nt* = training and validation sets combined) and the number of network weights (*Nw*) are shown. For the model structure coding, see Section 4.

Number of Hidden Layers	S2-4 [CD]		S2-10 [ABCD]	
	Nw	*Nt/Nw*	*Nw*	*Nt/Nw*
1	3783	2.00	14521	0.52
2	5257	1.44	17352	0.44
3	5545	1.36	17640	0.43
4	5617	1.35	17712	0.43
5	6913	1.09	19480	0.39
6	6931	1.09	19498	0.39
8	7334	1.03	19901	0.38
12	7857	0.96	20424	0.37
16	10439	0.72	23006	0.33
20	12332	0.61	25843	0.29

3.4. Random Forest (RF) Models

The Random forest algorithm [44] is an ensemble learning method that combines the bootstrap aggregation method ('bagging' [51]) with the random subspace method to train a set of decision trees that have a high variance and low bias [52]. The outputs of the trees are combined by mode (classification trees) or mean (regression trees) to produce the prediction output. Random forests avoid overfitting to the training set [51], provide fast processing [53], and means to rank the importance of the predictors [51,52]. The random forest models were implemented with scikit-learn [54], a frequently used machine learning library for Python programming language, using the function 'sklearn.ensemble. RandomForestRegressor'.

For the random forest models we tuned the hyper-parameters for tree maximum depth (search values = [20, 30, no limitation]), number of estimators ([80, 160, 200]), number of minimum samples for tree split ([2, 3]), and number of minimum samples for tree leaves ([1, 2, 5]). We also examined two alternative values for the RF parameter that controls the number of features to consider when looking for the best split. Setting this parameter value to the square root of the total number of inputs (predictor variables) speeded up modeling roughly by a factor of 10. However, this setting caused a slight increase in root-mean-square error and bias, when compared to the case of considering all features. Thus, we decided to use all features in node splitting, as speed was not an issue in testing. As the variance in validation set performance was very small for different iterations, the iteration count for random forest hyper-parameter search was set to 10.

3.5. Model Performance Evaluation

To evaluate model performance, we used the absolute and relative root-mean-square error (RMSE and RMSE%), bias (BIAS and BIAS%), and the coefficient of determination (R^2).

$$\text{RMSE} = \sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{n}} \tag{2}$$

$$\text{RMSE\%} = \frac{RMSE}{\bar{y}} \cdot 100\% \tag{3}$$

$$\text{BIAS} = \frac{\sum_i (\hat{y}_i - y_i)}{n} \tag{4}$$

$$\text{BIAS\%} = \frac{BIAS}{\bar{y}} \cdot 100\% \tag{5}$$

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y}_i)^2} \quad (6)$$

with y_i the observed values from reference forest inventory and $\bar{y}$ their mean, $\hat{y}_i$ the model estimates, n the total number of measurements, SS_{tot} the total sum of squares, and SS_{res} the residual sum of squares; all sums are for $i = 1 \ldots n$.

Due to the stochastic nature of the model training processes, the models performed slightly differently in separate training runs, even with the same model structure and hyper-parameters. Consequently, we repeated the modeling iteration wrapper 100 times for each separate DNN model test setup, or 25 times for tests with a limited number of field plots (Section 3.3.3). The arithmetic mean values of the performance measures and their lower and upper confidence levels were computed from the results of these 100 (or 25) iterations. In addition, we computed the number of failed modeling iteration runs of all trials until the amount of 100 (or 25) successful iterations was reached. As the bias may obtain both positive and negative values, we used absolute bias values in the mean computations.

4. Results

The results presented in this section were obtained using total stem volume (V) as the target variable in each test. Sections 4.1–4.4 present the sensitivity analysis of various factors (Table 5) to model performance, and the results of transfer learning are presented in Section 4.5. Section 4.6 presents the performance figures for a selection of other forest variables, namely basal area, stem diameter, and tree height.

In each following result, figure, and table, the used model input features, Sentinel-2 image sampling scheme, and DNN depth are indicated by a label following a template **S2-**** [****] **w*** ***L**, with:

- **S2-****: the used Sentinel-2 image channels: **S2-4** = [B2, B3, B4, B8] (i.e., RGB and NIR) or **S2-10** = [B2, B3, B4, B8, B5, B6, B7, B8A, B11, B12] (i.e., RGB, NIR, SWIR, and VRE bands);
- **[****]**: the auxiliary features or feature groups, labelled according to Table 3;
- **w* = w1, w3 or w5**: window size of the used Sentinel-2 image sampling scheme (1 × 1, 3 × 3 or 5 × 5 pixels, respectively);
- **[*L]**: the number of hidden layers in DNN (e.g., 3L = three hidden layers).

For instance, a label **S2-4 [CD] w3 3L** indicates that model inputs were four Sentinel-2 channels with auxiliary inputs from groups C and D (DEM and slope data), and that the model used 3 × 3 pixel image sampling window and a DNN structure with three hidden layers. For the DNN structure we use the notation i-$h_1 - h_2 - ...h_n$-o, where i = number of input nodes, $h_1 - h_n$ = number of hidden layer neurons, and o = number of output neurons. Table 7 summarizes the factors tested and parameters varied in the different test setups.

As CHM data may not always be available, they were included only in the tests considering different input features (Section 4.1). Comparisons and discussions of the results after Section 4.1 refer to the models without CHM, unless otherwise stated.

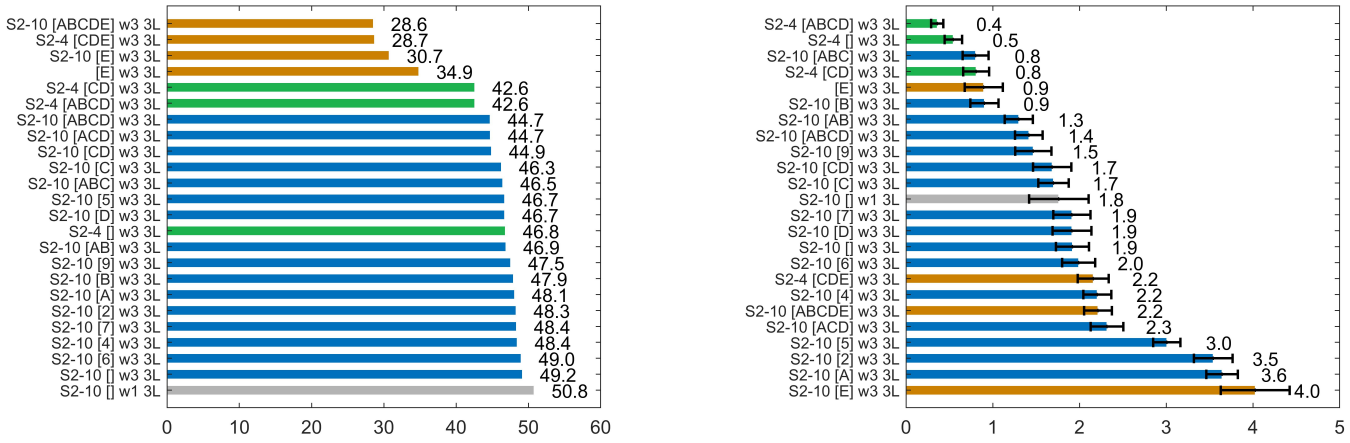
Table 7. Details of different test setups to study the influence of various factors on the performance of the deep neural network model. Input data sources are described in Table 3. Refer to the corresponding subsection for results of the different test setups. Coding of test setups: S2-4 = Sentinel-2 bands B2, B3, B4, and B8 (at 10 m resolution)); S2-10 = Sentinel-2 bands B2, B3, B4, B5, B6, B7, B8, B8A, B11, and B12 (20 m bands resampled to 10 m); **** = none, 2, 3 (= B), 4, 5, 6, 7, 8 (= E), 9, A, C, D, AB, CD, ABC, ACD, ABCD, ABCDE; ++++ = none, CD, CDE, ABCD.

Factor Tested/Parameter Varied	Parameter Values	Section
Effect of different input features /DNN input feature combination	24 different input combinations: (1) Ten S2 image bands, 1 × 1 sampling, 3 hidden layer DNN (S2-10 [] w1 3 L = baseline) (2) Ten S2 image bands + auxiliary features, 3 × 3 sampling, 3 hidden layer DNN (S2-10 [****] w3 3L) (3) Four S2 image bands + auxiliary features, 3 × 3 sampling, 3 hidden layer DNN (S2-4 [++++] w3 3L) (4) CHM features only (E)	Section 4.1
Effect of DNN depth /Number of DNN hidden layers	Two input variable combinations: (1) S2-4-CD w3 xL (2) S2-10-ABCD w3 xL with ten different numbers of hidden layers: x = 1–6, 8, 12, 16 and 20	Section 4.2
Effect of image sampling window size /S2 image sampling window size	1 × 1, 3 × 3 and 5 × 5 pixels windows	Section 4.3
Effect of the amount of training data /Nominal number of field plots, training + validation sets combined	60, 90, 120, 150, 210, 367 300, 510, 750, and 990	Section 4.4
Transfer learning tests		Section 4.5

The tests were run with an HP power laptop PC with Intel® Core™ i7-10610U CPU @ 1.80 GHz 2.30 GHz processor, with 24 GB RAM memory installed, and with 64-bit operating system. The runtimes for the different tests, and for a single model, were from 30 min to 6 h, depending on the DNN size, hyper-parameter tuning, and the number of iterations defined for the test. Typical time for training a single model (with hyper-parameter tuning, no performance statistics collection) varied roughly between 10 and 15 min (training time for S2-4 [CD] w3 2L: hyper-parameter tuning: 10 min; model Training: 2 min).

4.1. Effect of Different Input Features

When testing the effect of different input features as model predictors, we chose a DNN structure with three hidden layers for the models (**3L**) and the 3 × 3 window (**w3**) for S2 image sampling. As a baseline model, we used a three hidden layer DNN having only ten S2 image channels with pixel-wise (1 × 1) image sampling as inputs, i.e., **S2-10 [] w1 3L**. Figure 3 shows the relative root-mean-square error (RMSE%) and the average of relative bias absolute values (BIAS% abs.) with different combinations of input features.



(**a**) RMSE% (**b**) BIAS% abs.

Figure 3. Test set relative root-mean-square error (RMSE%) and average of relative bias absolute values (BIAS% abs.) for DNN models predicting total stem volume with different input feature combinations. The results are mean values from 100 iterations of the modeling wrapper with the same hyper-parameter values. The bars are ordered with decreasing performance (the best performing on top). The bar colors correspond to the same feature combinations in both figures: gray = the baseline model; blue = models with all ten Sentinel-2 image channels (**S2-10**) and various auxiliary features; green = models with four Sentinel-2 image channels (**S2-4**); orange = models including CHM data (auxiliary input group E). The 99% confidence intervals are shown for BIAS% abs. only, as they are not distinguishable for RMSE% at this scale.

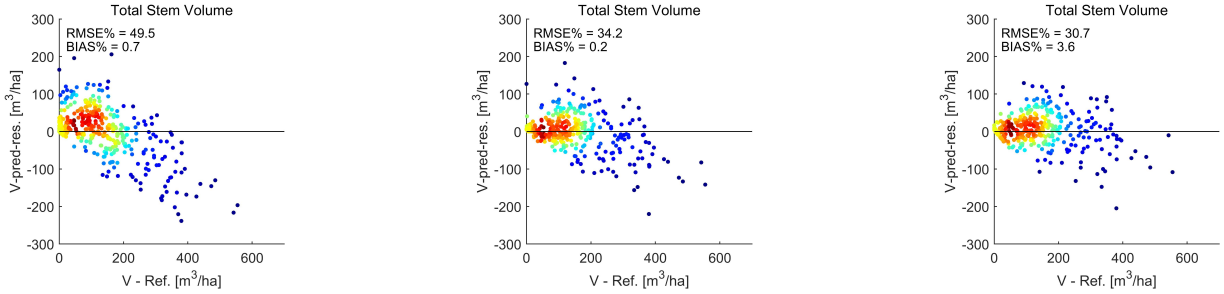
Using 3 × 3 image sampling and adding auxiliary features as model predictors improved the RMSE% accuracy in comparison to the baseline model. The best performance was obtained by adding LiDAR-based canopy height model (CHM) features to satellite data (models with orange bars), with 2022% (pp) improvement with respect to the baseline. The orange bar labeled with [E] shows the performance of the model using only the CHM features as predictive inputs (RMSE% = 34.9%, BIAS% abs. = 0.9%). When combining S2 data and auxiliary features with the CHM data the RMS error was 28.6–30.7% with different feature combinations.

In addition to the CHM features, almost all the auxiliary features caused a clear improvement in RMSE% accuracy, when added one by one as model inputs in addition to the ten S2 image channels (models **S2-10 [2–7]**). In general, the DNN models with a subset of four S2 image channels produced better RMSE% performance (and BIAS% as well) when compared to the models that included all ten S2 channels. The improvement with respect to the baseline RMSE% performance was 8.1% (pp) with models **S2-4 [CD]** and **S2-4 [ABCD]**.

The BIAS% (abs.) varied from 0.4% to 4.0% in the test set. The smallest bias was obtained with the model including four S2 image channels and all auxiliary data sources (**S2-4 [ABCD] w3 3L**). In addition, two other models using the reduced set of four S2 image channels as inputs (**S2-4 []** * and **S2-4 [CD]** *) were among the four best performing models in terms of the bias. Unlike with RMSE%, adding auxiliary features to model inputs did not systematically improve the bias when compared to models **S2-10 [] w3 3L** or **S2-10 [] w1 3L**. This was most surprising with the CHM features, which obtained the worst performance in terms of bias, when added alone to the ten S2 image channels.

An explanation for this was found by examining the residual error plots of the test set estimates, for which the example with CHM features is shown in Figure 4. The CHM inputs reduced the underestimates above 200 m^3/ha, but the slight overestimates at stem volumes lower than this caused an increased positive bias, as most of the test data volume is concentrated there. This was also the case with other auxiliary inputs that caused an increase in the bias when added to the S2 image data as model inputs. It must be noted, however, that the increase in BIAS% with respect to model **S2-10 [] w3 3L** was

significant only for the four models shown in the bottom of Figure 3b, as indicated by the non-overlapping 99% confidence intervals.

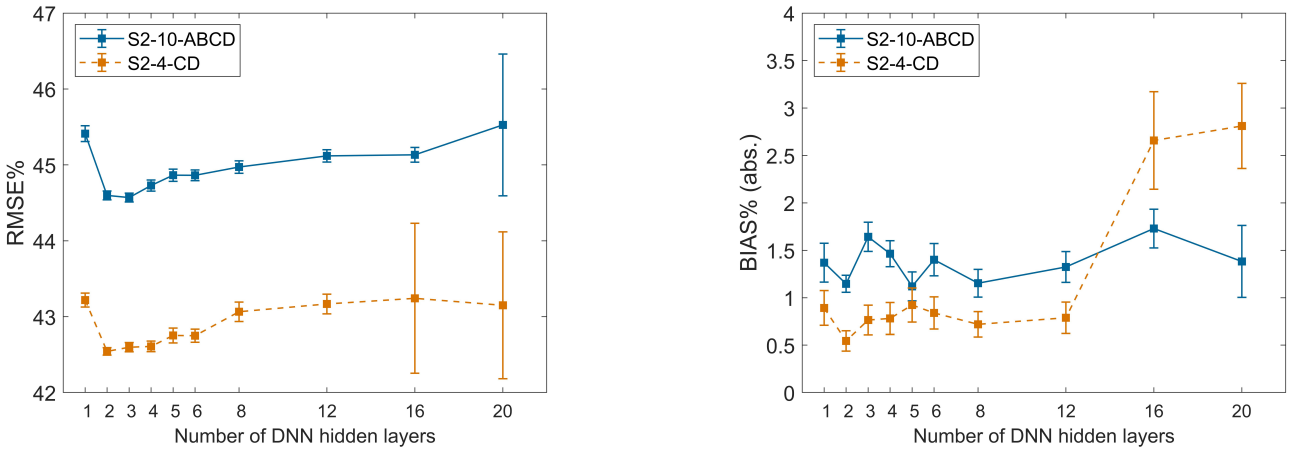


(**a**) Model S2-10 [] w3 3L (**b**) Model [E] w3 3L (**c**) Model S2-10 [E] w3 3L

Figure 4. Residual error plots of total stem volume test set estimates for models with (**a**) 10 S2 bands (S2-10-[]), (**b**) CHM features ([E]), and (**c**) combined 10 S2 bands and CHM features (S2-10 [E]) as predictive features. In (**a**) the S2 data model the underestimates above 200 m^3/ha and are compensated by the overestimates for lower stem volume resulting in low bias. This also applies to the model (**b**) with CHM features, although the errors are smaller. In combined data model (**c**) the underestimates above 200 m^3/ha have been reduced significantly (thanks to CHM data), but the overestimates under this stem volume cause increased positive bias due to the data distribution.

4.2. Effect of DNN Depth

The RMSE% and BIAS% performance figures for DNNs with different numbers of hidden layers are shown in Figure 5 for two different input feature combinations with a 3 × 3 window as S2 image sampling scheme (**S2-10 [ABCD] w3 *** and **S2-4 [CD] w3 ***). These were the two models with the lowest RMSE%, including all input features and a reduced set of S2 image channel data (excluding models with CHM features; Figure 3a).



(**a**) RMSE% (**b**) BIAS% abs.

Figure 5. Test set relative root-mean-square error (RMSE%) and average of relative bias absolute values (BIAS% abs.) for DNN models with different numbers of DNN hidden layers. Blue: model with ten S2 image channels and all auxiliary features. Orange: model with four S2 image channels, DEM, and slope information as input features. S2 image sampling with 3 × 3 pixels in each model. The whiskers show the 99% confidence intervals from 100 modeling iterations.

There was about a 1% (pp) decrease in RMSE% with both of the feature sets, when moving from one to two hidden layer networks. The minimum RMSE% values were obtained already with two or three hidden layer networks, after which the error increased slowly along with the number of hidden layers. The variance in the 100 model test results increased considerably for the models with 16 or 20 hidden layers.

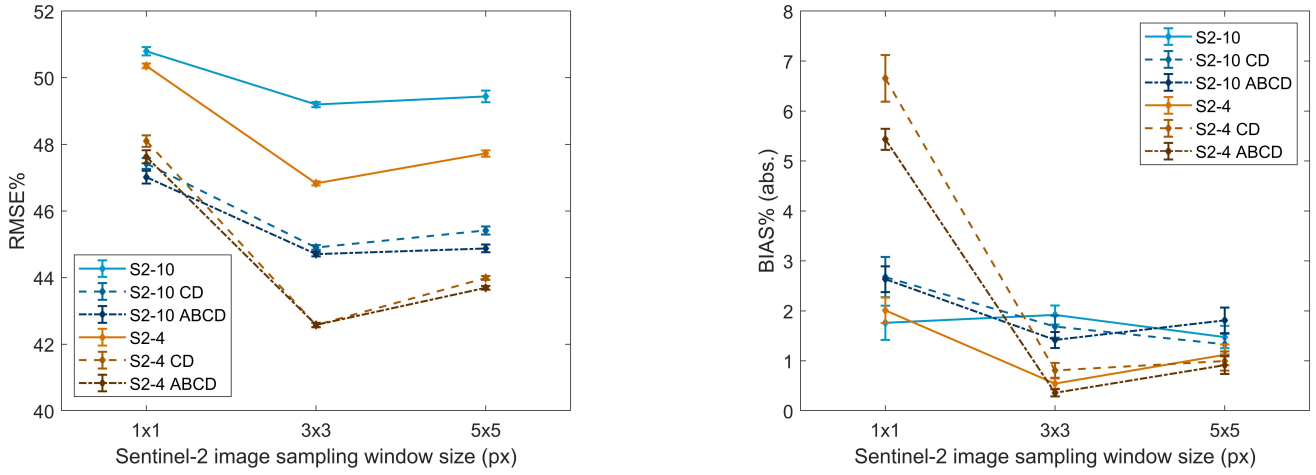
In bias values (BIAS% abs.) we did not get a clear indication of the optimum network depth, although there seemed to be a minimum at five hidden layers with the model having the input set of **S2-10 [ABCD]**, and at two hidden layers with the model having the input

set of **S2-4 [CD]**. As with the RMSE% values, the variance in model performance increased with the deepest models (16 or 20 hidden layers).

Training failed occasionally and more frequently with the deepest networks. The proportion of the failed runs was zero for the one- or two-hidden-layer networks, and was 1% for the three-layer network, and then increased gradually, exceeding 50% for the 20-layer network. In these training attempts the weights in the first two network layers were practically zero, and the network produced a zero output. Thus, the training suffered from vanishing gradients [55,56] along increasing depth of the DNN. We addressed this by replacing the ReLU activations with the Scaled Exponential Linear Unit (SELU) activation function, which reduced the proportion of failed training runs considerably. However, as the use of SELU activation caused an increase in RMSE% and bias variances, we used ReLU for the final results.

4.3. Effect of Image Sampling Window Size

The effect of the S2 image sampling window size on model performance is shown in Figure 6. We used six different sets of input features in these tests, combining two S2 image channel sets (S2-10 and S2-4) and three selections of auxiliary features [none, CD, and ABCD]. All DNN models had three hidden layers.



(**a**) RMSE% (**b**) BIAS% abs.

Figure 6. Test set relative root-mean-square error (RMSE%) and average of relative bias absolute values (BIAS% abs.) for DNN models with six different input feature combinations. All DNN networks consisted of three hidden layers. The whiskers show the 99% confidence intervals from 100 modeling iterations.

All six curves in Figure 6a obtained a minimum at 3 × 3 window sampling. The improvement of using 3 × 3 pixels rather than 1 × 1 was 1.5–5.5% (pp), being largest with the input feature set **S2-4 [CD]**. Using four of the S2 image channels instead of ten and adding auxiliary features (CD or ABCD) resulted in greater improvement. Extending the S2 image sampling to 5 × 5 pixels did not improve the results any more. On the contrary, there was a systematical decrease in RMSE% performance with all the tested input feature combinations between 3 × 3 and 5 × 5 sampling schemes.

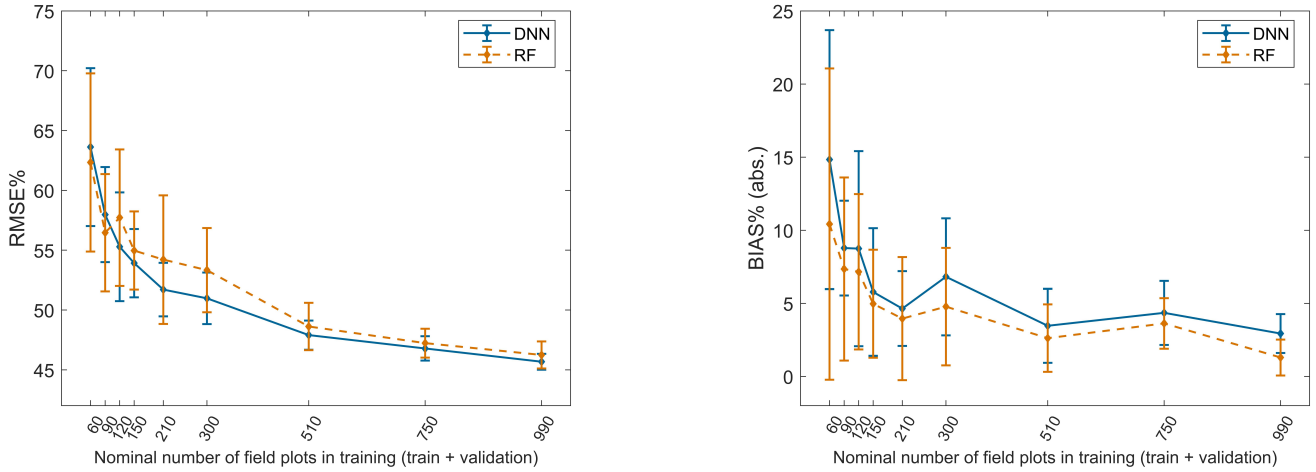
Increasing the S2 image sampling window size from 1 × 1 reduced the test set bias on average (Figure 6b). The bias reached the minimum with 3 × 3 sampling with four of the six input combinations (**S2-10 [ABCD]**, **S2-4**, **S2-4 [CD]**, **S2-4 [ABCD]**). The last two input combinations (**S2-10**, **S2-10 [CD]**) produced the smallest bias with the 5 × 5 sampling window.

Sampling the Sentinel-2 image data using 3 × 3 or 5 × 5 pixel windows thus clearly improved the model performance in comparison to pixel-wise sampling. As the usage of the 5 × 5 window did not bring any benefit to the model performance in comparison to 3 × 3 sampling, and on the other hand increased the model complexity along increased

number of trainable weights, our choice for optimum image sampling window size was 3×3 pixels.

4.4. Effect of the Amount of Training Data

The performances of the deep neural networks and random forest were compared by training total stem volume prediction models using limited numbers of field plots. Figure 7a,b show the results according to increasing nominal number of field plots used in training (Section 3.3.3).



(**a**) RMSE% (**b**) BIAS% abs.

Figure 7. Test set relative root-mean-square error (RMSE%) and average of relative bias absolute values (BIAS% abs.) for DNN and RF models with different nominal amounts of field plots used in training+validation sets. The whiskers show the standard deviation (1σ) of the error from the 10 iterations of same nominal number of field plots.

The RMSE% values of both model types increased about linearly when the number of field plots in training was reduced from 990 to 300 (including validation set). With fewer than 300 field plots the RMSE% increased rapidly, showing almost exponential growth under 210 field plots. The variation between the ten iteration run results also increased heavily below 510 field plots. In general, there was no large difference between the performance of DNN and RF methods. DNN produced slightly more accurate stem volume estimations between 120 and 510 training samples. On the other hand, the RF method yielded lower BIAS% values with all sample counts. However, the confidence levels of both RMSE% and BIAS% (abs.) overlapped heavily, so no preference between these two methods could be made.

Our results indicate that a practical minimum number of field plots for DNN (or RF) model training is somewhere between 200 and 300, and amounts larger than 500 plots are preferred.

4.5. Transfer Learning Tests

We tested transfer learning with a DNN model (fully connected 118-**121-24-12-6**-1 network) pre-trained on data from the Central Ostrobothnia forestry district. We chose a model with enough layers to separate the possible low level features in the earlier hidden layers, from the more specific features in the last layers. On the other hand, we selected a network structure with close to optimal RMSE% performance (Figure 5a).

To first investigate the best transfer learning configuration, we trained (refined) the pre-trained model by freezing the weights from 0 to 4 hidden layers using field plot data from the three other forestry districts (Table 1). We then computed the test set performance (BIAS% and BIAS% abs.) for the refined models (see Figure 8). The RMSE% increased significantly with 1 to 4 layer weights frozen. The bias did not show a monotonic increase, but the general trend was the same as with RMSE%. In general, we obtained the best

performance by letting all weights learn from new data; thus, we used this principle in the subsequent transfer learning tests.

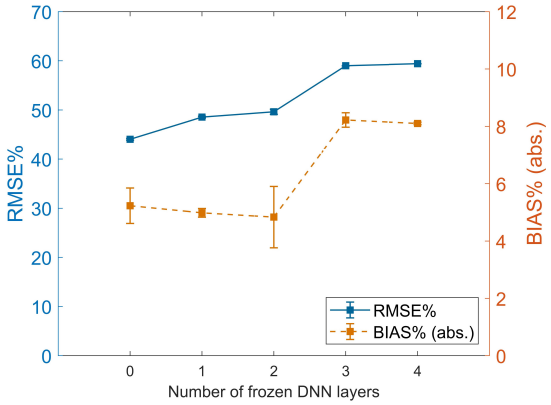
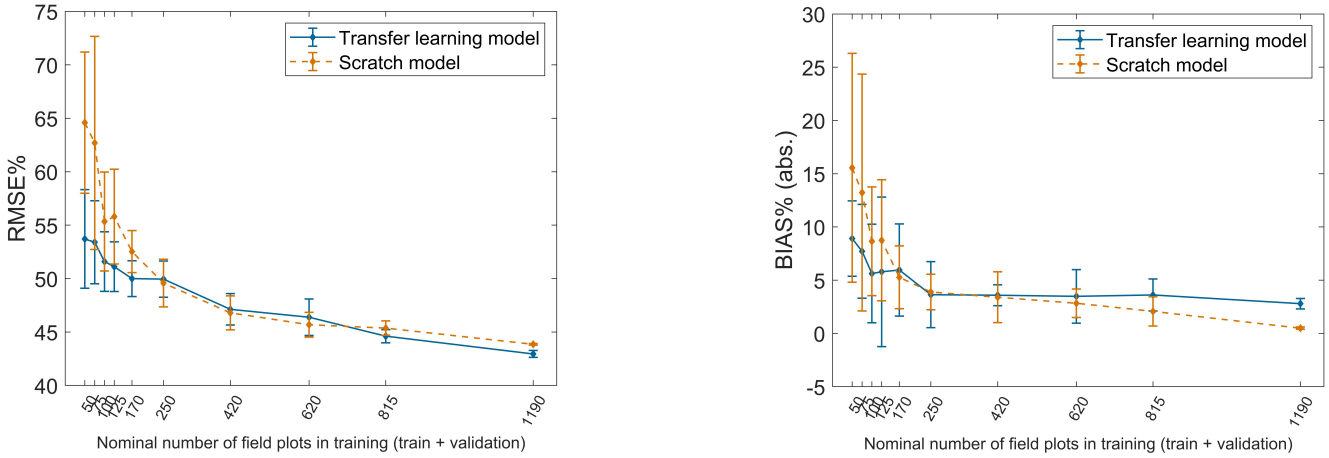


Figure 8. Test set average relative root-mean-squared error (RMSE%) and average of relative bias absolute values (BIAS% abs.) for transfer learning models with 0–4 hidden layer weights frozen. The whiskers show the 99% confidence intervals from 100 modeling iterations (not visible for RMSE%).

We refined the pre-trained DNN model with different numbers of field plots. The RMSE% and BIAS% performances of the refined models are compared in Figure 9 with models having equal structure, and that were trained from scratch with the same field plots as the transfer learning network. The RMSE% performance of the transfer learning models and models trained from scratch were practically equivalent when 250 field plots or more were used in training. The mean RMSE% performance of the transfer learning models was better than the models trained from scratch with fewer than 250 field plots, but the variances overlapped heavily (except with the smallest number of field plots).



(**a**) RMSE% (**b**) BIAS% abs.

Figure 9. Test set relative root-mean-squared error (RMSE%) and average of relative bias absolute values (BIAS% abs.) for the transfer learning model and for the model trained from scratch for different nominal numbers of field plots. The whiskers show the standard deviation (1σ) of the error from the 10 iterations of the same nominal number of field plots.

In addition, BIAS% performance with the transfer learning model was better with small numbers of training plots (<170). An interesting result was that with the largest number of nominal field plots (≥815), the RMSE% and BIAS% were higher for the transfer learning models than with the models trained from scratch, although the differences in RMSE% were small.

Our results indicated that transfer learning improved total stem volume prediction accuracy, if the number of training samples is limited (i.e., less than 250). With larger

amounts of training data transfer, learning did not bring any benefit in comparison to training the models from scratch, as can be seen from Figure 9.

4.6. The Performance for a Wider Set of Forest Variables

To get a broader perspective, we produced prediction models and the performance figures for a wider set of forest variables with the proposed concept. The proposed models consist of a two-hidden-layer DNN with 3 × 3 image sampling window trained with image data and all auxiliary features except CHM (**S2-4 [ABCD] w3 2L**). We chose the four S2 image channels and all auxiliary variables as the model predictors as they produced the best RMSE% and bias accuracies of the tested input features (Figure 3). On the other hand the two-layer DNN produced the best performance when the four S2 channels were used as model predictors (Figure 5).

We also produced predictions for the same set of variables with more simple models that represent a straightforward forest variable modeling using satellite image data without windowed image sampling, and any auxiliary features as predictive inputs. We used two hidden layers for the simple models as well (**S2-4 [] w1 2L**).

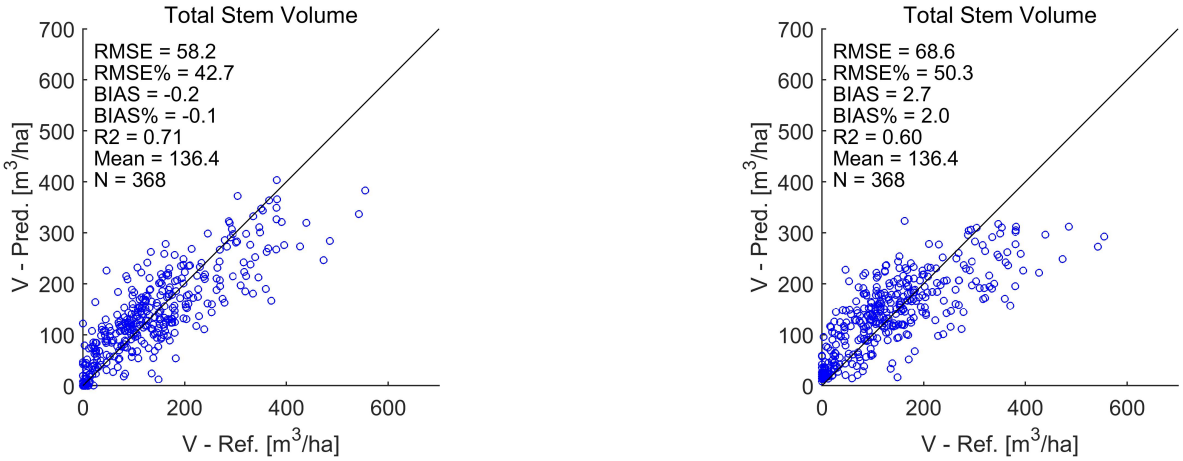
We produced models for four forest variables: stem volume (V), basal area (G), stem diameter (D), and tree height (H). We predicted both species-specific data and variable totals. We used (64-**67-24**-1) networks for the proposed models and (4-**7-3**-1) networks for the simple models. Relative root-mean-square error (RMSE%), bias (BIAS% abs.), and coefficient of determination (R^2) for the models are shown in Table 8.

Table 8. Test set relative root-mean-square error (RMSE%), average of relative bias absolute values (BIAS% abs.), and coefficient of determination (R^2) for forest variables stem volume (V), basal area (G), stem diameter (D), and tree height (H), and their species-specific components (suffixes -pine for pine, -spr for spruce, and -bl for broadleaved). The proposed DNN models with two hidden layers using [ABCD] input variables and 3 × 3 window image sampling compared to models using pixel-wise S2 image data and two hidden layers. The best performance values in bold.

	Proposed DNN Models S2-4 [ABCD] w3 2L			**Simple Models S2-4 [] w1 2L**		
Forest Variable	**RMSE%**	**BIAS% abs.**	R^2	**RMSE%**	**BIAS% abs.**	R^2
Stem Volume (V)	**42.4**	**1.1**	**0.71**	50.3	1.5	0.60
Stem Volume/pine (V-pine)	**95.6**	**1.8**	**0.41**	104.0	3.1	0.30
Stem Volume/spruce (V-spr)	**138.3**	**14.3**	**0.47**	153.1	16.6	0.35
Stem Volume/broadleaved (V-bl)	**125.8**	11.5	**0.46**	151.5	**6.1**	0.22
Basal Area (G)	**33.0**	**0.4**	**0.68**	36.9	1.5	0.60
Basal Area/pine (G-pine)	**77.9**	**3.8**	**0.50**	82.3	5.7	0.44
Basal Area/spruce (G-spr)	**96.3**	**5.0**	**0.64**	134.3	23.2	0.31
Basal Area/broadleaved (G-bl)	**107.3**	13.1	**0.50**	125.8	**9.9**	0.31
Stem Diameter (D)	**33.1**	4.8	**0.62**	35.4	**4.2**	0.57
Stem Diameter/pine (D-pine)	**36.6**	**4.5**	**0.51**	39.7	4.6	0.42
Stem Diameter/spruce (D-spr)	**41.6**	2.9	**0.60**	44.9	**1.8**	0.53
Stem Diameter/broadleaved (D-bl)	**47.3**	5.7	**0.36**	48.4	**5.2**	0.33
Tree Height (H)	**28.2**	2.9	**0.65**	30.7	2.9	0.59
Tree Height/pine (H-pine)	**30.3**	3	**0.63**	33.4	**2.8**	0.55
Tree Height/spruce (H-spr)	**34.1**	1.8	**0.68**	37.2	**1.0**	0.62
Tree Height/broadleaved (H-bl)	**35.5**	3.7	**0.54**	37.1	**3.6**	0.50

The proposed DNN models produced better RMSE% and R^2 results than the simple models for all forest variables. The simple models produced lower BIAS% than the proposed models for 8 out of 16 forest variables. The relative improvements with proposed concept with respect to the simple models were from 2.3% to 28.3%), with the largest decrease obtained with stem volume and basal area variables (G-spr: 28.3% and V-bl: 16.9%,

relative improvement). Figure 10 shows the total stem volume model response for the test set when using the proposed concept (a) and with the simple model (b). The proposed model reduced both the saturation at about 320 m^3/ha stem volumes, and the overestimates below 200 m^3/ha that were visible in the response of the simple model. Similar effects can be seen from the scatterplots of basal area, stem diameter, and tree height (Appendix A).

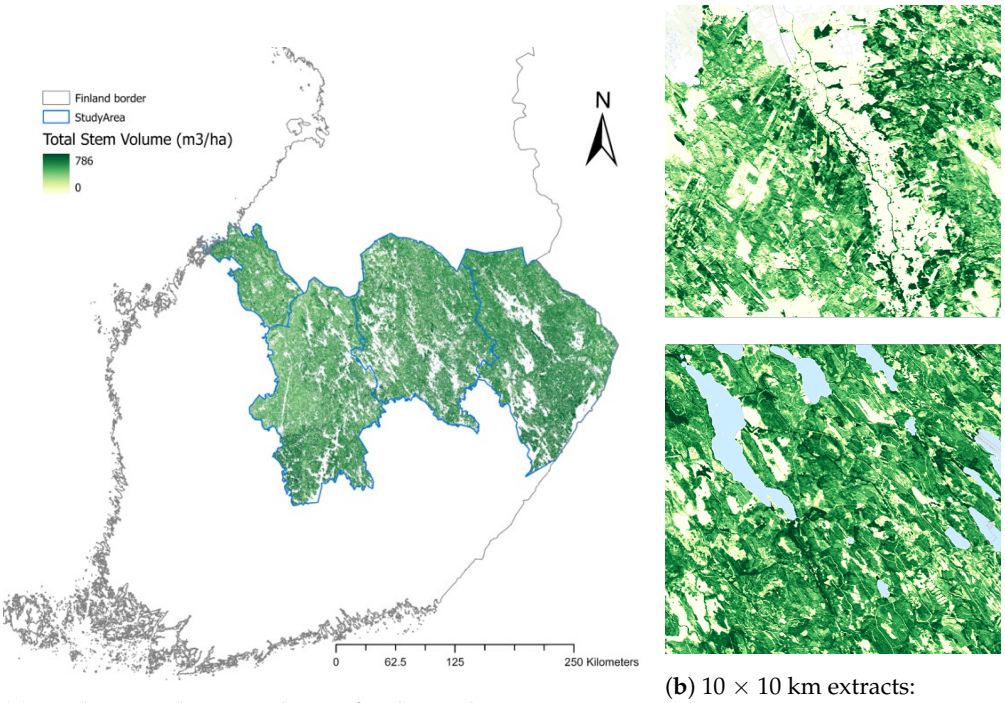


(**a**) Model: S2-4 [ABCD] w3 2L (**b**) Model: S2-4 [] w1 2L

Figure 10. Scatterplots of total stem volume test data set predictions with the proposed model concept (**a**) and with the simple model (**b**).

4.7. Total Stem Volume Estimates

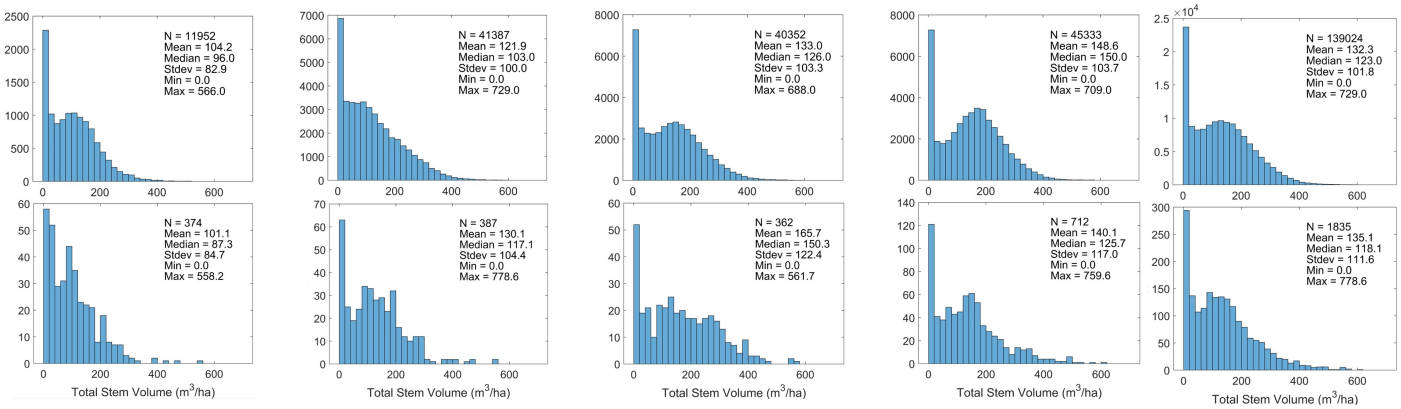
Figure 11a shows a stem volume estimate with spatial resolution of 10 m computed for the four forestry districts with the model: S2-4 [CD] w3 3L. Figure 11b shows 10 km × 10 km extracts from Central Ostrobothnia (top) and North Savo (bottom). The difference in the forest area, and the higher stem volumes in North Savo, is clearly visible in the zoomed figures. Figure 12 (top row) shows the predicted stem volume distributions for the entire study area and the four forestry districts. The predictions were sampled from random locations, with minimum mutual distance of 250 m. The bottom row shows the reference data (training, validation, and test sets combined) histograms from corresponding areas. The distributions of the reference data and predictions show relatively similar behavior in all forestry districts, with the biggest differences in North Savo area. The differences in the stem volumes between the four areas can be seen from the histograms and the statistics printed on the graphs.



(**a**) Total stem volume prediction for the study area

(**b**) 10 × 10 km extracts:
Central Ostrobothnia (top)
North Savo (bottom)

Figure 11. Total stem volume predictions of model S2-4 [CD] w3 3L (**a**) for the whole study area, and (**b**) for two 10 km × 10 km close-ups from Central Ostrobothnia (**top**) and North Savo (**bottom**).



(**a**) Central Ostrobothnia (**b**) Central Finland (**c**) North Savo (**d**) North Karelia (**e**) Entire study area

Figure 12. Total stem volume distributions for the four forestry districts and the entire study area with the histogram of the prediction (**top**) and the reference (**bottom**).

5. Discussion

5.1. Sensitivity Analysis of Training Parameters

Our test results showed that there was a clear benefit in combining the auxiliary features with S2 image data that were sampled with 3 × 3 or 5 × 5 window instead of using a pixel-wise approach. We obtained a substantial improvement in total stem volume prediction performance: the best models using 3 × 3 image sampling, topography data (DEM and slope; input group D), and the sun and imaging angles (input group C) obtained about 8% (pp) lower RMSE% values than the model with only pixel-wise image data. The improvement of R^2 value from about 0.6 to 0.7 was observed between these models as well. The effect of the image acquisition date and the time to field data collection (feature group A) had a relatively small improving effect to the model RMSE performance (see Figure 3). This was no surprise as there was not much seasonal variation in the green

vegetation during the imaging period (1 June 2017–6 September 2017), and as the time differences from imaging to field data collection were also small. Inclusion of the image pixel location (Lat, Lon; feature 3/B) caused slightly more improvement, which was also an expected result, as the forest characteristics and the topography vary significantly between the western and eastern parts of the study area.

The results also indicated that the S2 sun angles and terrain slope were the two sets of features that most improved the RMSE% when added alone to the S2 image channels (CHM excluded). When selecting the auxiliary predictive variables to the DNN models, we assumed that by including the topography data from the close spatial context of the target pixel, and combining it with S2 imaging and sun angles, we could at least partly compensate the effect of shadowing in the pixel reflectance values. The effect of these features can be seen in the saturation effect above 300 m^3/ha stem volumes that correspond to the lowest image reflectance values (on forest area) (Figure 10; the scatterplots of models with features S2-4 [CD] and S2-4 [ABCD] are almost identical). The windowed S2 image sampling combined with topography data (feature group D) and sun and imaging angles (feature group C) reduce this saturation, i.e., the DNN model is able to capture the correlations between sunny/shadowed terrain slopes and the image reflectance variations.

In our tests, the optimum Sentinel-2 image sampling window size was 3 × 3 pixels, and increasing to 5 × 5 pixels did not improve the total stem volume prediction accuracy. An obvious explanation is that the image sampling area of 3 × 3 pixels matches best the field plot area with a maximum radius of R = 12.6 m (mature stands), and the outer 16 pixels of the 5 × 5 pixel sampling window do not necessarily overlap the field plot area at all. In other words, the 5 × 5 px area contains data that may not correspond to the ground truth data, which is thus seen as noise in the model training process.

In addition, we observed the best performance using a subset of the ten S2 image channels, i.e., RGBNIR bands. We did not use any procedure for selecting the best predictive image features, but we used two different sets of S2 image channels: one with the 10 m resolution bands and the other with all the 10 m and 20 m bands. The systematic difference in model performance is clearly visible in Figure 5 with RMSE% about 2% (pp) lower for the models having fewer inputs. We hypothesize that the better results with fewer spectral bands are due to the better ratio of the amount of training data to the number of trainable weights in the network (see ratio (Nt/Nw in Table 6): the training algorithm has more weights to optimize with the models with more inputs that also bring additional noise components to the optimization task. The ratio (Nt/Nw may also partly explain the differences in the model accuracy between 3 × 3 px and 5 × 5 px image sampling windows. However, as the selection of the best predictive S2 bands vary for different forest variables, one should include a feature selection step for optimal performance, especially if the amount of good quality reference data for DNN model training is limited, as in this study.

The benefit of adding layers to the neural network was marginal, and we obtained the best RMSE% performance already with networks having two hidden layers (Figure 5a). Using more than two hidden layers did not improve RMSE% and seemed to increase the variance in bias (Figure 5b). On the other hand, the DNNs with 3 to 12 hidden layers kept performing almost as well as the two-hidden-layer network.

The probability of the gradients to go to zero during DNN training seemed to increase rapidly with the depth of the network. As seen in our study, the selection of SELU activation instead of ReLU improved the situation but with the drawback of increased uncertainty of the model quality. Choosing a different weight initialization might improve the situation. Another way would be to use, e.g., a ResNet-like architecture with skip connections [57]. However, as the skip connections require the number of network elements not to change in between the skipped layers, this solution tends to increase the number of weights in the network. As there was no benefit seen from using more than two or three hidden layers in the DNN, this was not studied further.

As seen from the results with a limited number of field plots, the practical minimum number of field plots is somewhere between 200 and 330 plots in our experimental setting. On the other hand, the stem volume prediction performance (RMSE%) improved by less than 1% (pp) when increasing the number of field plots from 990 to 1467. Based on this observation, one could expect the marginal utility of adding more training data (i.e., field plots) with the present predictive inputs would be relatively small. This means that to gain more accuracy, one must include additional features that contain relevant information about the variable to be predicted. Alternatively, one could choose a completely different approach, e.g., the use of time series analysis and different DNN architecture.

Typically, the amount of training data is very large in deep learning applications, which enables DNN architectures with up to several millions of trainable parameters. When building models for forest variable estimation this is seldom the case, due to the costs of fieldwork and the size of data sets available for the model training (often consisting of only 200 to 500 up-to-date field plots). A widely used technique to increase the size of training set is data augmentation [58], in which the available relevant training data are multiplied with different techniques (e.g., image mirroring or rotation). It has to be noted that when using these data augmentation techniques the consistency with other related input variables must be preserved. In our case, this would mean that, in addition to image data, the sun and imaging azimuth angles as well as the slope data must be rotated/mirrored accordingly. The analysis for the impact of data augmentation in the current context will be the subject of a future study.

5.2. Comparison with Similar Studies

A few recent remote sensing studies estimating stem volume in boreal forests have used similar inputs or methods as the present study [43,59–61]. A study concentrating on improving Finnish multi-source national forest inventory by 3D aerial imaging compared the accuracy and spatial characteristics of 2D satellite (Landsat 8) and aerial imagery as well as 3D airborne laser scanning (ALS) and photogrammetric remote sensing data in the estimation of forest inventory variables using the KNN method [43]. Volume estimations using Landsat 8 data or aerial imagery produced poor accuracies with RMSE% values of over 60%. Even their combination produced poor accuracy of approximately 58%. Estimations carried out in the present study using only 2D optical data produced clearly higher accuracies (Figure 3). This result was in line with earlier study by Astola et al. [32] that compared the use of Landsat 8 and Sentinel-2 data for forest variable estimation. The highest accuracy (RMSE = 27.80%) for the volume estimation in [43] was obtained using a combination of ALS and 2D aerial imagery, which had very similar accuracy to the best accuracy obtained in the present study (RMSE% = 28.6%, Figure 3). Thus, it can be noticed that our models with CHM features included as predictors in combination with S2 data achieved accuracies comparable to those reported by [43] using 3D ALS data and aerial imagery. This is encouraging to note and highlights the potential for using openly available Sentinel-2 data instead of aerial imagery for large area inventories, considering the increasing availability of ALS data.

In a study conducted in southern Finland the performance of multitemporal Sentinel-2 data was determined for estimating several forest variables using random forest [59]. In the study, the multitemporal results were compared to those of single date Sentinel-2 data, ALS, stereo-SAR, and high-resolution 3D satellite data. The study area was relatively small and consisted of only 74 plots (32×32 m^2). ALS produced the best estimation accuracy (RMSE%) for volume (14.17%), and the 3D photogrammetric data from the high-resolution satellite data produced the second highest accuracy (17.33%). The multitemporal Sentinel-2 gave higher accuracy for volume than the stereo data from SAR (27.20% and 30.22%, respectively). These results highlight the benefit of 3D information when estimating forest structure variables. Our results support this, since the addition of canopy height information increased the volume estimation accuracy (Figure 3). In Sweden, a couple of recent studies have obtained similar results to ours when estimating stem volume.

For instance, Sentinel-2 and TanDEM-X data were combined in a study to estimate stem volume over large areas in Sweden using the KNN method (k = 5) [60]. The Level-1C data and 10 m and 20 m spatial resolution bands were used together with windowed radar data (3 × 3 and 5 × 5 pixels, similarly to our study). The obtained stand level accuracy (RMSE%) using the combination of Sentinel-2 and TanDEM-X was 30.2%. Using only Sentinel-2 and TanDEM-X the accuracies were 37.9% and 32.0%, respectively. The use of interferometric phase height from TanDEM-X provided the basis for better volume estimations. Sentinel-2 data were considered important for tree species mapping. Another study that was conducted in mid- and south Sweden developed linear regression models using only 3D information from stereophotogrammetry of aerial images and ALS (i.e., no 2D optical data) [61]. The estimations were validated on stand level. In mid Sweden, where forest types are more similar to Finland than in southern Sweden, the accuracy (RMSE%) for stem volume estimation was approximately 22% when using point cloud metrics from aerial images and ALS as predictor variables. In addition, these studies carried out in Sweden agree with the fact that 3D information is important for volumetric predictions in boreal forests, as shown by our results.

Another study using rather similar inputs or methods compared with the present study was carried out in the southeast of Poland, outside a boreal forest. In the study, growing stock volumes of Scots pine stands were estimated using Sentinel-2 images and airborne image-derived point clouds [62]. Multiple linear regression and random forest were used for plot level (400 m^2) estimations. It was investigated whether the inclusion of Sentinel-2 data improved the accuracy of models based on the 3D point cloud data. The estimation accuracy (RMSE%) with random forest using only Sentinel-2 data was approximately 36%. The accuracy was notably improved when using only the point cloud data or when combining Sentinel-2 and point cloud data. Accuracy for both was approximately 20% using random forest. Even though the study did not find any major benefits from combining Sentinel-2 data with 3D information in plot level estimations, the difference in accuracy was notable when compared with a situation where only satellite data were used, similarly to our results, e.g., in Figure 3.

A recent study in different forest conditions, but with similar model inputs to our study, was conducted in southwestern Germany [63]. In the study, multi-temporal Sentinel-2 data and 3D photogrammetric point clouds were combined and used to estimate timber volume. In the study, Sentinel-2 data were used to model the percentage of broadleaf tree volume. This was also used as one the metrics in timber volume estimations in addition to data from CHM that were derived from the photogrammetric point cloud. A non-linear logistic regression model was used to estimate timber volume, and the achieved accuracy was RMSE% = 31.7%, which is very similar to the accuracy produced by our model with combined 10 Sentinel-2 bands and CHM features (S2-10 [E] w3 3L, Figures 3 and 4).

Other studies using deep neural networks in different conditions, but with similar findings to our study regarding, e.g., Sentinel-2 sampling window size and the importance of Sentinel-2 spectral bands or the optimal DNN depth, were reported by [28,31].

In this study, the focus was on applying DNN to satellite data, and we did not make full use of the 3D information provided by ALS data, when computing statistical features of the 1 m resolution CHM data. A more sophisticated way would be to include the CHM at its original 1 m resolution or to use the ALS metrics from point clouds directly. This would most likely require a change to the DNN architecture, e.g., to select CNNs instead of fully connected dense layers [64,65]. Furthermore, as the number of trainable parameters would increase along with the number of predictive inputs, a larger amount of training data would probably be needed.

5.3. Comparison between DNNs and RF

In our study, the deep neural network models and random forest models produced equivalent test set performances, and we could not find any preference for selecting between these two methods for this kind of application. The hyper-parameter tuning was

slightly more demanding for DNNs, but with models of the size used in our study, this is not an issue. One benefit of available DNN architectures is that they are able to extract meaningful features from image texture (e.g., CNN) or time series (e.g., LSTM), which must be produced separately in prediction systems based on other methods, as pointed out in [7].

Additionally, in some other studies, deep learning models have not been outstandingly better than RF models. For instance, [64] reported that, in general, their deep learning method for growing stock volume prediction produced the best results, when compared to other methods (e.g., RF or k-NN), but that the differences were small. Using the Landsat and ALS predictors, they reported RMSE% = 24.16%, BIAS% = −2.2%, and R^2 = 0.38 with deep learning, and RMSE% = 25.18%, BIAS% = −5.32%, and R^2 = 0.39 with RF model. Narine et al. [28] compared DNN to RF methods in three different scenarios: daytime, nighttime, and no noise. They reported equal performance in AGB prediction for both models, DNN being slightly more accurate for daytime and nighttime, and vice versa for the no noise scenario.

On the other hand, many other studies have reported results showing DNNs outperforming RF in remote sensing applications [7,26,27]. The Chimera model presented in [7] outperformed SVM and RF models with the same input data for all tested classification and regression tasks, including, e.g., basal area and quadratic mean diameter. In the study, data augmentation was used to train the Chimera RCNN (with more than 1.3 million parameters) with a relatively small amount of reference data (<10,000 field plots). The root-mean-square errors they obtained for basal area (7.1 m^2/ha) and quadratic mean diameter (9.9 cm) were of the same order as in our study: 5.8 m^2/ha for basal area and 5.4 cm for mean diameter at breast height. Shao et al. [26] estimated forest biomass in China. They obtained satisfactory estimation accuracies when using LiDAR-derived AGB as ground reference data. The best mapping accuracy was obtained using SSAE model that outperformed more traditional methods, such as KNN, SVM, and RF. Zhang et al. [27] also mapped AGB in China. Their study targeted at developing a novel approach for predicting AGB by integrating Landsat 8 imagery with LiDAR data through a deep learning-based workflow. SSAE was used as deep learning model, and it was found to be superior over the other prediction models (e.g., RF, KNN, SVM).

5.4. Transfer Learning Performance and Applicability

Transfer learning produced equal or better overall performance than the reference model when all the pre-trained model weights were refined with field plot data. Our tests showed that transfer learning is beneficial mainly with small training data sets (fewer than 250 field plots), not with larger amounts of training data.

In our tests, we observed the difficulty of the network to learn the characteristics of the new area, when even only the lowest layer weights of the pre-trained network were frozen, as seen also by [38]. However, their observation that even features transferred from a distant task would be better than random weights was not supported by our results: when enough training data were available (i.e., >250 field plots), the performance of the refined transfer learning model was not better than the model trained from scratch. However, in our case the model was much simpler; thus, the results are not necessarily comparable. Authors in [20] showed that training deep models from scratch on remote sensing data usually gave higher accuracy than transfer learning for tasks related to vegetation.

6. Conclusions

In this study we tested different combinations of model predictors and two predictive methods, i.e., deep neural network (DNN) and random forest (RF) algorithm, for forest variable estimation. The model inputs were selected from several data sources: Sentinel-2 image data and metadata, digital elevation model (DEM), terrain slope data, and canopy height model (CHM). We tested different factors influencing on the model predictive performance: the image sampling window size, the importance of different input features

as model predictors, the DNN depth, and the number of field plots used for training. In addition, we tested the utility of transfer learning in Finnish boreal forest conditions.

The concept of using Sentinel-2 metadata features and terrain elevation and slope with the image reflectance data in combination with windowed image sampling scheme proved successful, and we were able to gain significant improvement (8% pp. in RMSE%) in total stem volume prediction accuracy. Increasing the number of hidden layers in DNN models improved the performance only marginally, with two or three hidden layers being the optimum depth in our tests.

Both DNN and RF models were suitable for the forest variable estimation in the proposed modeling setup, with approximately equal performance. The tuning of hyper-parameters was slightly easier with RF models, which makes this method more appealing. When fewer than 300 field plots were available for model training, the errors with both methods increased rapidly along with decreasing amount of training data.

The advantage of deep neural network is the possibility to implement transfer learning easily. DNN transfer learning facilitates the refinement of an existing model to new areas and for new variables with a relatively small amount of new reference data. Thus, the major advantage of using transfer learning is that the expensive and time-consuming fieldwork may be reduced considerably, which may be found profitable in, e.g., regularly repeating forest inventories. The reduction in model performance with small amounts of reference data may also be compensated by using transfer learning.

As an outcome of the study we were able to improve the forest variable estimation performance with the proposed image sampling and input feature concept that is straightforward to implement with DNN or RF algorithms, and that has small computational impact. The developed deep learning concept with the transfer learning option will be implemented in Forestry Thematic Exploitation Platform Forestry-TEP (European Space Agency, https://f-tep.com/; accessed 17 June 2021) in the near future to be available for the remote sensing community.

Our research interests in the future include the potential improvement in model performance when using data augmentation techniques to multiply the amount of reference data, and to investigate the usage of multitask learning for forest variable regression. In addition, the utilization of ALS point clouds in combination with deep neural networks is an interesting topic to explore.

Author Contributions: Conceptualization, H.A., L.S., M.M. and A.L.; Data curation, H.A. and L.S.; Formal analysis, H.A.; Funding acquisition, H.A. and A.L.; Investigation, H.A.; Methodology, H.A. and L.S.; Project administration, A.L.; Resources, L.S.; Software, H.A. and L.S.; Supervision, A.L.; Validation, H.A. and L.S.; Writing—original draft, H.A. L.S., E.H. and M.M.; Writing—review & editing, H.A., E.H. and M.M. All authors have read and agreed to the published version of the manuscript.

Funding: This study was part of the joint Forestry One Stop Shop project (2018–2019) of Business Finland's New Space Economy program. Additional funding was provided by the Academy of Finland (grant 317387).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available upon reasonable request from the corresponding author. The data are not publicly available due to reference data license conditions.

Acknowledgments: We thank the Finnish Forest Centre for providing the reference field plots.

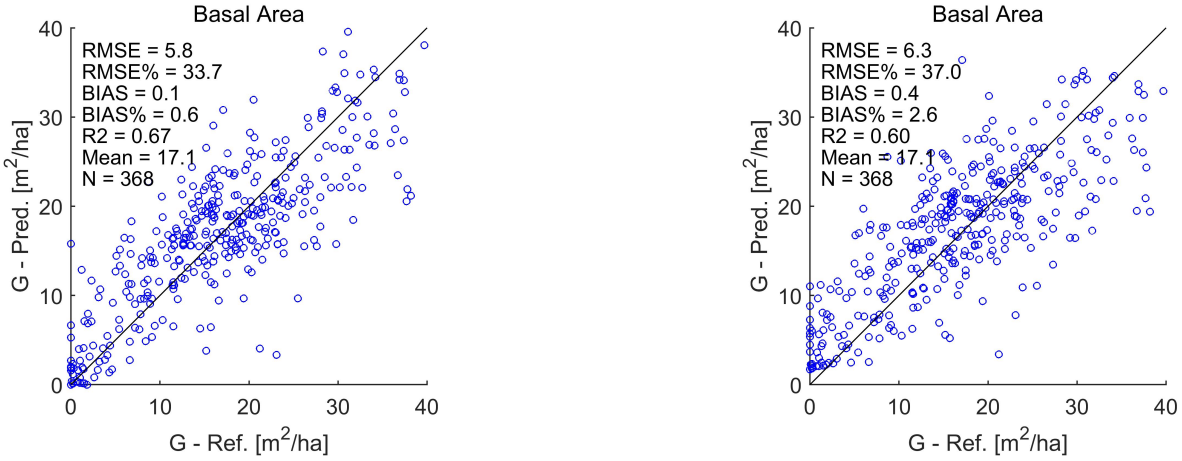
Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

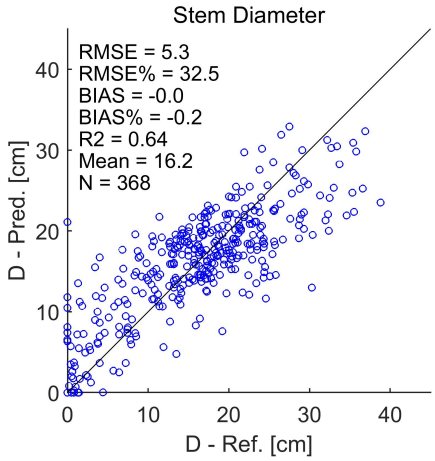
AGB	Aboveground Biomass
ALS	Airborne Laser Scanning
CHM	Canopy Height Model
CNN	Convolutional Neural Network
DBN	Deep Belief Network
DEM	Digital Elevation Model
DNN	Deep Neural Network
EO	Earth Observation
GSV	Growing Stock Volume
KNN	k-Nearest Neighbor
LiDAR	Light Detection And Ranging
NIR	Near Infra-Red
ReLU	Rectified Linear Unit
RF	Random Forest
RGB	Red Green Blue
RMSE	Root Mean-Squared Error
RNN	Recurrent Neural Network
S2	Sentinel-2
SELU	Scaled Exponential Linear Unit
Stdev	Standard Deviation
SVM	Support Vector Machines

Appendix A. Test Set Scatterplots for Basal Area (G), Stem Diameter (D), and Tree Height (H)

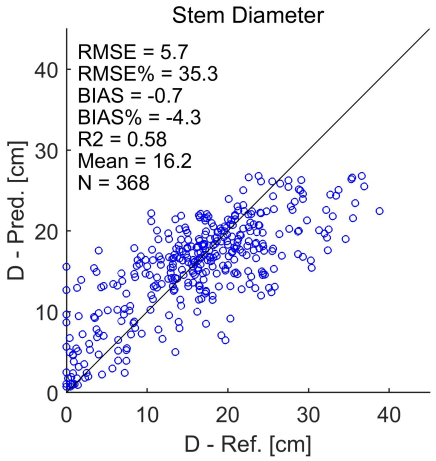


(**a**) Model: S2-4 [ABCD] w3 2L (**b**) Model: S2-4 [] w1 2L

Figure A1. Scatterplots of basal area (G) test data set predictions with the proposed model concept (**a**) and with the simple model (**b**).

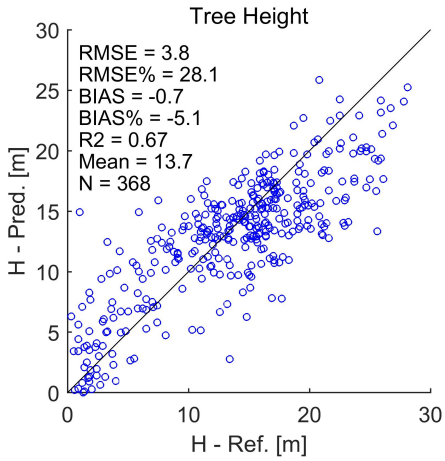


(**a**) Model: S2-4 [ABCD] w3 2L

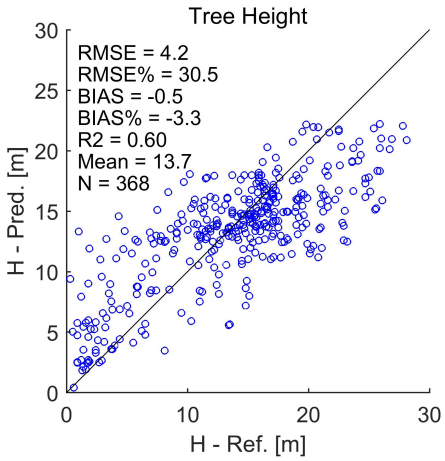


(**b**) Model: S2-4 [] w1 2L

Figure A2. Scatterplots of stem diameter (D) test data set predictions with the proposed model concept (**a**) and with the simple model (**b**).



(**a**) Model: S2-4 [ABCD] w3 2L



(**b**) Model: S2-4 [] w1 2L

Figure A3. Scatterplots of tree height (H) test data set predictions with the proposed model concept (**a**) and with the simple model (**b**).

References

1. Chrysafis, I.; Mallinis, G.; Siachalou, S.; Patias, P. Assessing the relationships between growing stock volume and Sentinel-2 imagery in a Mediterranean forest ecosystem. *Remote Sens. Lett.* **2017**, *8*, 508–517. [CrossRef]
2. Antropov, O.; Rauste, Y.; Tegel, K.; Baral, Y.; Junttila, V.; Kauranne, T.; Häme, T.; Praks, J. Tropical Forest Tree Height and Above Ground Biomass Mapping in Nepal Using Tandem-X and ALOS PALSAR Data. In Proceedings of the IGARSS 2018—2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; pp. 5334–5336. [CrossRef]
3. Keenan, R.; Reams, G.; Achard, F.; de Freitas, J.; Grainger, A.; Lindquist, E. Dynamics of global forest area: Results from the FAO Global Forest Resources Assessment 2015. *For. Ecol. Manag.* **2015**, *352*, 9–20. [CrossRef]
4. MäKelä, A.; Pulkkinen, M.; Kolari, P.; Lagergren, F.; Berbigier, P.; Lindroth, A.; Loustau, D.; Nikinmaa, E.; Vesala, T.; Hari, P. Developing an empirical model of stand GPP with the LUE approach: analysis of eddy covariance data at five contrasting conifer sites in Europe. *Glob. Chang. Biol.* **2008**, *14*, 92–108. [CrossRef]
5. Alberdi, I.; Bender, S.; Riedel, T.; Avitable, V.; Boriaud, O.; Bosela, M.; Camia, A.; Cañellas, I.; Castro Rego, F.; Fischer, C.; et al. Assessing forest availability for wood supply in Europe. *For. Policy Econ.* **2020**, *111*, 102032. [CrossRef]
6. Haakana, H. *Multi-Source Forest Inventory Data for Forest Production and Utilization Analyses at Different Levels*; Dissertationes Forestales; Finnish Society of Forest Science: Helsinki, Finland; Faculty of Agriculture and Forestry of the University of Helsinki: Helsinki, Finland; School of Forest Sciences of the University of Eastern Finland: Joensuu, Finland; August 2017; [CrossRef]
7. Chang, T.; Rasmussen, B.P.; Dickson, B.G.; Zachmann, L.J. Chimera: A multi-task recurrent convolutional neural network for forest classification and structural estimation. *Remote Sens.* **2019**, *11*, 768. [CrossRef]

8. Drusch, M.; Del Bello, U.; Carlier, S.; Colin, O.; Fernandez, V.; Gascon, F.; Hoersch, B.; Isola, C.; Laberinti, P.; Martimort, P.; et al. Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. *Remote Sens. Environ.* **2012**, *120*, 25–36. [CrossRef]
9. Soille, P.; Burger, A.; De Marchi, D.; Kempeneers, P.; Rodriguez, D.; Syrris, V.; Vasilev, V. A versatile data-intensive computing platform for information retrieval from big geospatial data. *Future Gener. Comput. Syst.* **2018**, *81*, 30–40. [CrossRef]
10. Li, J.; Huang, X.; Gong, J. Deep neural network for remote-sensing image interpretation: Status and perspectives. *Natl. Sci. Rev.* **2019**, *6*, 1082–1086. [CrossRef]
11. Verrelst, J.; Malenovský, Z.; Van der Tol, C.; Camps-Valls, G.; Gastellu-Etchegorry, J.P.; Lewis, P.; North, P.; Moreno, J. Quantifying Vegetation Biophysical Variables from Imaging Spectroscopy Data: A Review on Retrieval Methods. *Surv. Geophys.* **2019**, *40*, 589–629. [CrossRef]
12. Tuia, D.; Verrelst, J.; Alonso, L.; Perez-Cruz, F.; Camps-Valls, G. Multioutput support vector regression for remote sensing biophysical parameter estimation. *IEEE Geosci. Remote Sens. Lett.* **2011**, *8*, 804–808. [CrossRef]
13. Wu, C.; Tao, H.; Zhai, M.; Lin, Y.; Wang, K.; Deng, J.; Shen, A.; Gan, M.; Li, J.; Yang, H. Using nonparametric modeling approaches and remote sensing imagery to estimate ecological welfare forest biomass. *J. For. Res.* **2018**, *29*, 151–161. [CrossRef]
14. Zhu, X.; Tuia, D.; Mou, L.; Xia, G.S.; Zhang, L.; Xu, F.; Fraundorfer, F. Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 8–36. [CrossRef]
15. Ma, L.; Liu, Y.; Zhang, X.; Ye, Y.; Yin, G.; Johnson, B. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS J. Photogramm. Remote Sens.* **2019**, *152*, 166–177. [CrossRef]
16. Li, J.; Wu, Z.; Hu, Z.; Zhang, J.; Li, M.; Mo, L.; Molinier, M. Thin cloud removal in optical remote sensing images based on generative adversarial networks and physical model of cloud distortion. *ISPRS J. Photogramm. Remote Sens.* **2020**, *166*, 373–389. [CrossRef]
17. Ball, J.; Anderson, D.; Chan, C. Comprehensive survey of deep learning in remote sensing: Theories, tools, and challenges for the community. *J. Appl. Remote Sens.* **2017**, *11*, 042609. [CrossRef]
18. Mountrakis, G.; Li, J.; Lu, X.; Hellwich, O. Deep learning for remotely sensed data. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 1 – 2. [CrossRef]
19. Yuan, Q.; Shen, H.; Li, T.; Li, Z.; Li, S.; Jiang, Y.; Xu, H.; Tan, W.; Yang, Q.; Wang, J.; et al. Deep learning in environmental remote sensing: Achievements and challenges. *Remote Sens. Environ.* **2020**, *241*, 111716. [CrossRef]
20. Kattenborn, T.; Leitloff, J.; Schiefer, F.; Hinz, S. Review on Convolutional Neural Networks (CNN) in vegetation remote sensing. *ISPRS J. Photogramm. Remote Sens.* **2021**, *173*, 24 – 49. [CrossRef]
21. You, J.; Li, X.; Low, M.; Lobell, D.; Ermon, S. Deep Gaussian process for crop yield prediction based on remote sensing data. In Proceedings of the 31st AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; pp. 4559–4565.
22. Wolanin, A.; Mateo-García, G.; Camps-Valls, G.; Gómez-Chova, L.; Meroni, M.; Duveiller, G.; Liangzhi, Y.; Guanter, L. Estimating and understanding crop yields with explainable deep learning in the Indian Wheat Belt. *Environ. Res. Lett.* **2020**, *15*, 024019. [CrossRef]
23. Nevavuori, P.; Narra, N.; Lipping, T. Crop yield prediction with deep convolutional neural networks. *Comput. Electron. Agric.* **2019**, *163*, 104859. [CrossRef]
24. Chen, Y.; Lee, W.; Gan, H.; Peres, N.; Fraisse, C.; Zhang, Y.; He, Y. Strawberry yield prediction based on a deep neural network using high-resolution aerial orthoimages. *Remote Sens.* **2019**, *11*, 1584. [CrossRef]
25. Song, X.; Zhang, G.; Liu, F.; Li, D.; Zhao, Y.; Yang, J. Modeling spatio-temporal distribution of soil moisture by deep learning-based cellular automata model. *J. Arid Land* **2016**, *8*, 734–748. [CrossRef]
26. Shao, Z.; Zhang, L.; Wang, L. Stacked Sparse Autoencoder Modeling Using the Synergy of Airborne LiDAR and Satellite Optical and SAR Data to Map Forest Above-Ground Biomass. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2017**, *10*, 5569–5582. [CrossRef]
27. Zhang, L.; Shao, Z.; Liu, J.; Cheng, Q. Deep learning based retrieval of forest aboveground biomass from combined LiDAR and Landsat 8 data. *Remote Sens.* **2019**, *11*, 1459. [CrossRef]
28. Narine, L.; Popescu, S.; Malambo, L. Synergy of ICESat-2 and Landsat for mapping forest aboveground biomass with deep learning. *Remote Sens.* **2019**, *11*, 1503. [CrossRef]
29. García-Gutiérrez, J.; González-Ferreiro, E.; Mateos-García, D.; Riquelme-Santos, J.C. A Preliminary Study of the Suitability of Deep Learning to Improve LiDAR-Derived Biomass Estimation. In *Hybrid Artificial Intelligent Systems*; Martínez-Álvarez, F., Troncoso, A., Quintián, H., Corchado, E., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 588–596.
30. Liu, J.; Wang, X.; Wang, T. Classification of tree species and stock volume estimation in ground forest images using Deep Learning. *Comput. Electron. Agric.* **2019**, *166*, 105012. [CrossRef]
31. Lang, N.; Schindler, K.; Wegner, J. Country-wide high-resolution vegetation height mapping with Sentinel-2. *Remote Sens. Environ.* **2019**, *233*, 111347. [CrossRef]
32. Astola, H.; Häme, T.; Sirro, L.; Molinier, M.; Kilpi, J. Comparison of Sentinel-2 and Landsat 8 imagery for forest variable prediction in boreal region. *Remote Sens. Environ.* **2019**, *223*, 257–273. [CrossRef]
33. Halme, E.; Pellikka, P.; Mõttus, M. Utility of hyperspectral compared to multispectral remote sensing data in estimating forest biomass and structure variables in Finnish boreal forest. *Int. J. Appl. Earth Obs. Geoinf.* **2019**, *83*, 101942. [CrossRef]

34. Mutanen, T.; Sirro, L.; Rauste, Y. Tree height estimates in boreal forest using Gaussian process regression. In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–15 July 2016; pp. 1757–1760. [CrossRef]
35. Pan, S.; Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **2010**, *22*, 1345–1359. [CrossRef]
36. Bengio, Y. Deep Learning of Representations for Unsupervised and Transfer Learning. In *Proceedings of Machine Learning Research, ICML Workshop on Unsupervised and Transfer Learning*; Guyon, I., Dror, G., Lemaire, V., Taylor, G., Silver, D., Eds.; PMLR: Bellevue, DC, USA, 2012; Volume 27, pp. 17–36.
37. Pratt, L.Y. Discriminability-Based Transfer between Neural Networks. *Adv. Neural Inf. Process. Syst.* **1993**. [CrossRef]
38. Yosinski, J.; Clune, J.; Bengio, Y.; Lipson, H. How Transferable Are Features in Deep Neural Networks? In *Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS'14)—Volume 2*; MIT Press: Cambridge, MA, USA, 2014; pp. 3320–3328. [CrossRef]
39. Wang, A.X.; Tran, C.; Desai, N.; Lobell, D.; Ermon, S. Deep Transfer Learning for Crop Yield Prediction with Remote Sensing Data. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS '18)*; Association for Computing Machinery: New York, NY, USA, 2018; pp. 1–5. [CrossRef]
40. Wurm, M.; Stark, T.; Zhu, X.; Weigand, M.; Taubenböck, H. Semantic segmentation of slums in satellite images using transfer learning on fully convolutional neural networks. *ISPRS J. Photogramm. Remote Sens.* **2019**, *150*, 59–69. [CrossRef]
41. Kaufman, Y.J.; Sendra, C. Algorithm for automatic atmospheric corrections to visible and near-IR satellite imagery. *Int. J. Remote Sens.* **1988**, *9*, 1357–1381. [CrossRef]
42. Rahman, H.; Dedieu, G. SMAC: A simplified method for the atmospheric correction of satellite measurements in the solar spectrum. *Int. J. Remote Sens.* **1994**, *15*, 123–143. [CrossRef]
43. Tuominen, S.; Pitkänen, T.; Balazs, A.; Kangas, A. Improving Finnish Multi-Source National Forest Inventory by 3D Aerial Imaging. *Silva Fennica* **2017**, *51*, 7743. Available online: https://www.silvafennica.fi/article/7743/related/7743 (accessed on 17 June 2021). [CrossRef]
44. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
45. Chollet, F. Keras. 2015. Available online: https://keras.io (accessed on 17 June 2021).
46. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; Devin, M.; et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. 2015. Available online: https://www.tensorflow.org (accessed on 17 June 2021).
47. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:cs.LG/1412.6980.
48. Hinton, G.E.; Srivastava, N.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R.R. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv* **2012**, arXiv:cs.NE/1207.0580.
49. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
50. Kohavi, R.; John, G.H. Wrappers for feature subset selection. *Artif. Intell.* **1997**, *97*, 273–324. [CrossRef]
51. Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123–140. [CrossRef]
52. Belgiu, M.; Dragut, L. Random forest in remote sensing: A review of applications and future directions. *ISPRS J. Photogramm. Remote Sens.* **2016**, *114*, 24–31. [CrossRef]
53. Rodriguez-Galiano, V.; Ghimire, B.; Rogan, J.; Chica-Olmo, M.; Rigol-Sanchez, J. An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS J. Photogramm. Remote Sens.* **2012**, *67*, 93–104. [CrossRef]
54. Pedregosa, F. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
55. Bengio, Y.; Simard, P.; Frasconi, P. Learning long-term dependencies with gradient descent is difficult. *IEEE Trans. Neural Netw.* **1994**, *5*, 157–166. [CrossRef]
56. Pascanu, R.; Mikolov, T.; Bengio, Y. Understanding the exploding gradient problem. *arXiv* **2012**, arXiv:1211.5063.
57. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp, 770–778. [CrossRef]
58. Shorten, C.; Khoshgoftaar, T.M. A survey on Image Data Augmentation for Deep Learning. *J. Big Data* **2019**, *6*, 1–48. [CrossRef]
59. Wittke, S.; Yu, X.; Karjalainen, M.; Hyyppä, J.; Puttonen, E. Comparison of two-dimensional multitemporal Sentinel-2 data with three-dimensional remote sensing data sources for forest inventory parameter estimation over a boreal forest. *Int. J. Appl. Earth Obs. Geoinf.* **2019**, *76*, 167–178. [CrossRef]
60. Persson, H.J.; Jonzén, J.; Nilsson, M. Combining TanDEM-X and Sentinel-2 for large-area species-wise prediction of forest biomass and volume. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *96*, 102275. [CrossRef]
61. Bohlin, J.; Bohlin, I.; Jonzén, J.; Nilsson, M. Mapping forest attributes using data from stereophotogrammetry of aerial images and field data from the national forest inventory. *Silva Fenn.* **2017**, *51*, 18. [CrossRef]
62. Hawryło, P.; Wezyk, P. Predicting growing stock volume of Scots pine stands using Sentinel-2 satellite imagery and airborne image-derived point clouds. *Forests* **2018**, *9*, 274. [CrossRef]
63. Schumacher, J.; Rattay, M.; Kirchhöfer, M.; Adler, P.; Kändler, G. Combination of multi-temporal Sentinel 2 images and aerial image based canopy height models for timber volume modelling. *Forests* **2019**, *10*, 746. [CrossRef]

64. Hawryło, P.; Francini, S.; Chirici, G.; Giannetti, F.; Parkitna, K.; Krok, G.; Mitelsztedt, K.; Lisańczuk, M.; Stereńczak, K.; Ciesielski, M.; et al. The use of remotely sensed data and Polish NFI plots for prediction of growing stock volume using different predictive methods. *Remote Sens.* **2020**, *12*, 3331. [CrossRef]
65. Ayrey, E.; Hayes, D.J.; Kilbride, J.B.; Fraver, S.; Kershaw, J.A.; Cook, B.D.; Weiskittel, A.R. Synthesizing Disparate LiDAR and Satellite Datasets through Deep Learning to Generate Wall-to-Wall Regional Forest Inventories. *bioRxiv* **2019**. [CrossRef]