

Article

Combining Multisource Data and Machine Learning Approaches for Multiscale Estimation of Forest Biomass

Yifeng Hong^{1,†}, Jiaming Xu^{1,†} , Chunyan Wu², Yong Pang³ , Shougong Zhang², Dongsheng Chen^{2,*} and Bo Yang⁴

¹ East China Academy of Inventory and Planning, National Forestry and Grassland Administration, Hangzhou 310019, China; yifeng_hong2022@126.com (Y.H.); xjmcugb@163.com (J.X.)

² State Key Laboratory of Efficient Production of Forest Resources, Key Laboratory of Tree Breeding and Cultivation of State Forestry and Grassland Administration, Research Institute of Forestry, Chinese Academy of Forestry, Beijing 100091, China; cywu@caf.ac.cn (C.W.); sgzhang@caf.ac.cn (S.Z.)

³ Research Institute of Forest Resources Information Technique, Chinese Academy of Forestry, Beijing 100091, China; pangy@ifrit.ac.cn

⁴ China Forestry Group Corporation, Beijing 100026, China; yangb@cfgc.cn

* Correspondence: chends@caf.ac.cn

† These authors contributed equally to this work.

Abstract: Forest biomass is an important indicator of forest ecosystem productivity, and it plays vital roles in the global carbon cycling, global climate change mitigating, and ecosystem researches. Multiscale, rapid, and accurate extraction of forest biomass information is always a research topic. In this study, comprehensive investigation of a larch (*Larix olgensis*) plantation was performed using remote sensing and field-based monitoring methods, in combination with LiDAR-based multisource data and machine learning methods. On this basis, a universal, multiscale (single tree, stand, management unit, and region), and unit-high-precision continuous monitoring method was proposed for forest biomass components. The results revealed the following. (1) Airborne LiDAR point cloud variables exhibited significant correlation with the aboveground components (except leaves) and the whole-plant biomass ($R_{adj}^2 > 0.91$), suitable for extraction or estimation of forest parameters such as biomass and stock volume. (2) In terms of biomass monitoring at forest stand and management unit scale, a random forest model performed well in fitting accuracy and generalization ability, whereas a multiple linear regression model produced clearer explanation regarding the biomass of each forest component. (3) Using seasonal phenological characteristics in the study area, larch distribution information was extracted effectively. The overall accuracy reached 90.0%, and the kappa coefficient reached 0.88. (4) A regional-scale forest biomass component estimation model was constructed using a long short-term memory model, which effectively reduced the probability of biomass underestimation while ensuring good estimation accuracy, with R^2 exceeding 0.6 for the biomass of the aboveground and whole-plant components. This research provides theoretical support for rapid and accurate acquisition of large-scale forest biomass information.

Keywords: forest; biomass; LiDAR; machine learning; remote sensing



Citation: Hong, Y.; Xu, J.; Wu, C.; Pang, Y.; Zhang, S.; Chen, D.; Yang, B. Combining Multisource Data and Machine Learning Approaches for Multiscale Estimation of Forest Biomass. *Forests* **2023**, *14*, 2248. <https://doi.org/10.3390/f14112248>

Academic Editor: Stefano Grigolato

Received: 28 September 2023

Revised: 22 October 2023

Accepted: 26 October 2023

Published: 15 November 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Forest biomass is a critical structural characteristic that serves as an indicator of a forest ecosystem's productivity and carbon storage. This measurement plays a pivotal role in understanding both local and regional carbon cycles, ultimately influencing terrestrial carbon sequestration [1–6]. Furthermore, precise forest biomass estimates are essential for effective forest management and the associated industries [7]. Consequently, the demand for a swift and precise approach to assess multiscale forest biomass information is on the rise.

Traditionally, forest biomass estimation relies on establishing specific relationships between structural traits, such as breast radius and tree heights, and biomass [8,9]. These traits are typically gathered through field investigations. By recording these structural traits for all or representative trees within designated plots, the results obtained through these conventional methods serve as benchmarks for other biomass estimation attempts. Thus far, more than 2000 models encompassing over 100 tree species have been developed for biomass estimation [10–12]. Despite the high accuracy and wide applicability of such models, long-term dynamic monitoring of biomass can be a time-consuming, labor-intensive, destructive, and often challenging endeavor [13]. Furthermore, when dealing with heterogeneous regions, the spatial representativeness of the investigation samples remains uncertain. As a more robust and intricate approach to biomass estimation, the utilization of dynamic global vegetation models (DGVMs) is considered to possess greater potential for predicting forest dynamics under the influence of climate change [14]. However, this method requires the calibration of numerous type-specific parameters based on diverse field experiments. This suggests that DGVMs cannot be swiftly adapted to new forest categories and lack the ability to diagnose biomass status in current forests with high spatial resolution.

The rapid advancement of remote sensing (RS) technology has enabled the swift assessment of the spatial distribution of forest biomass across a broad spectrum. Leveraging fitted relationships between biomass and RS-derived predictors, it is possible to generate comprehensive biomass maps that cover entire regions [15]. The selection of input features typically depends on data availability and the research objectives. Conventional input features include optical RS images, given that a variety of RS indices has been devised to capture various forest characteristics, some of which are closely linked to biomass. Additionally, microwave RS observations are sensitive to canopy structure and water volume, making them valuable for biomass prediction [16]. Microwave observations with extended waveforms such as L-band or P-band have proven effective in global forest biomass retrieval [17]. However, it is worth noting that the spatial resolution of related products is not comparable to optical RS observations, typically at a 0.1- or 0.25-degree grid scale.

On a regional scale, numerous studies have combined optical and microwave RS data to construct biomass inversion models that have achieved a certain degree of accuracy, while challenges like saturation and instability hinder the assurance of sufficient accuracy in regional applications using such an approach [18–20]. Therefore, LiDAR (Light Detection And Ranging) technology, a novel non-imaging RS technology capable of measuring the absolute distance between a target (such as canopy tops in the current context) was introduced. This ability ensures accurate canopy height measurements, a crucial factor in biomass estimation [21]. LiDAR has demonstrated reasonable performance in estimating forest biomass on small to medium scales, including single trees, stands, and management units, due to its precise acquisition of forest vertical structure [22,23]. However, the discrete nature and high cost of airborne LiDAR limit its continuous application on a large scale [24]. Consequently, the question remains: Can an integrated strategy that leverages the strengths of various data and technologies, mitigates the limitations of individual methods, and establishes a multiscale, rapid, and accurate approach for acquiring forest biomass information using multisource data be realized?

To integrate multiple data sources effectively, a proficient algorithm is essential for learning or summarizing the relationship between input features and forest biomass. In comparison to traditional algorithms like linear regression, machine learning approaches such as LSTM, U-Net, and transformer approach exhibit superior capability and reliability in enhancing the accuracy of forest biomass estimation and conducting in-depth analysis of multisource data [25–27]. Machine learning's applications in remote sensing within the field of forestry primarily concentrate on tasks such as land-use/land-cover classification, vegetation succession prediction, tree species identification, forest pest and fire damage detection, and the estimation of parameters like forest leaf area index, stock volume, and

biomass [28–31]. Regarding the remote sensing estimation of forest biomass using machine learning, it can be categorized into image-based and object-oriented methods [32–36]. Nevertheless, many prior studies focused on establishing relationships solely between remote sensing image pixels or objects and forest biomass. Little attention has been given to exploring relationships between adjacent pixels (or objects) or between pixels with similar information but separated by a certain distance. Moreover, machine learning often necessitates numerous high-precision samples, which may be challenging to obtain through traditional methods. Therefore, research that combines machine learning with multi-source remote sensing data is also a promising avenue worth considering.

Larch (*Larix olgensis*) forests represent a forest category with notable economic potential. These forests are extensively distributed throughout the northeastern region of China, characterized by a temperate and cold temperate monsoon climate [14]. The ongoing global climate warming has had a marked impact on larch forests, which is particularly significant given larch's heightened sensitivity to climate variations. Furthermore, the northeastern part of China is experiencing a more pronounced warming trend [37]. In order to ensure the continued productivity of larch forests amidst climate change and to sustain economic benefits for local communities, it is imperative to enhance our understanding of larch growth and optimize the management of these forests [38]. This optimization process necessitates the accurate mapping of larch biomass.

In this study, our aim was to develop a methodology capable of addressing the challenges associated with extensive and large-scale ground-based surveys. To achieve this, we integrated LiDAR observations with ground-based sample survey data and applied it to estimate biomass in a larch plantation located in Heilongjiang Province, northeast China. Additionally, we harnessed the “memory” characteristics of the long short-term memory (LSTM) machine learning method for data analysis in two scenarios: (1) homogenous forest stands not adjacent in space and (2) forest stands (planted forests) with a stable status closely linked to a specific factor (age). The outcomes of this study aim to illustrate the suitability of a universal, multiscale (including single tree, forest stand, management unit, and region), and high-precision continuous monitoring approach for assessing forest biomass components by integrating data from multiple remote sensing sources.

2. Materials and Methods

2.1. Study Area

The study area was located in the Mengjiagang Forest Farm in Heilongjiang Province (46°20′–46°30′ N, 130°32′–130°52′ E). It is located in the western foothills of Wanda Mountain, dominated by low mountains and hills with elevation of 170–575 m (average elevation: 250 m) and gentle slope of 10°–20°. The soil is dominated by typical dark-brown earth. The forested area of the forest farm covers 13,671 ha. The forest coverage rate is 86.3%, and the total standing tree stock volume is 1.46 million m³. The main forest types are plantations of *Pinus sylvestris* L. var. *mongolica* Litv. and *Pinus koraiensis* Sieb. et Zucc., which occupy approximately two-thirds of the total area. Other natural secondary forests account for the remaining one-third. To objectively evaluate the generalization capability of the regional-scale biomass component extrapolation model, an independent sample test area (45°3′–45°58′ N, 129°42′–130°34′ E) was designated in the Linkou Forestry Bureau. The topography, landforms, climate, flora, and other factors of this test area are broadly consistent with those of the Mengjiagang Forest Farm. The forest area of the Linkou Forestry Bureau covers 213,600 ha, the standing tree stock volume is 14.1 million m³, and the forest coverage rate is 78.9%.

2.2. Data

2.2.1. Field Data

A field survey conducted in June 2016 divided the *L. olgensis* plantation into young stands (age: ≤20 years), middle-aged stands (21–30 years), near-mature stands (31–40 years), and mature stands (≥41 years) and established 40 sample plots (each with area of 0.06 ha)

(Figure 1). The diameter at breast height (DBH, in cm) of a single tree was measured using a measuring tape. The height of a tree and the height (H, in m) of its branches were measured using an ultrasonic altimeter (Vertex III, HAGLOF Company, Långsele, Sweden). A summary of the data details of the sample plots is presented in Table 1.

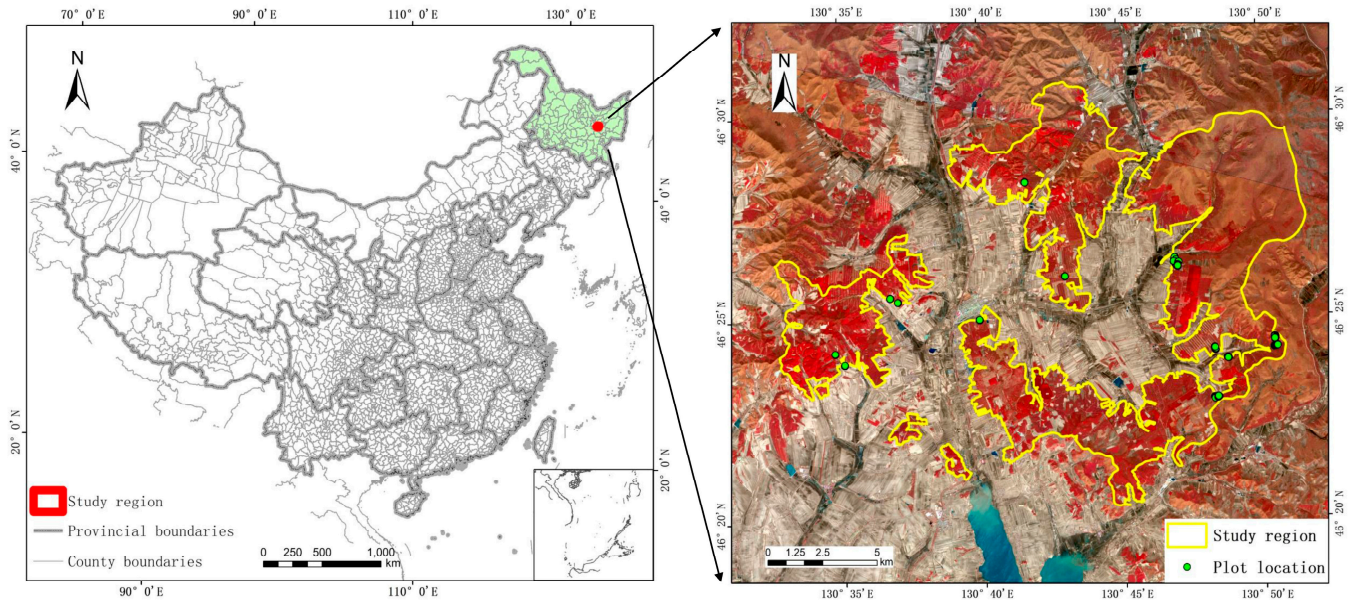


Figure 1. Location of the study area (Mengjiagang Forest Farm).

Table 1. Summary statistics of the sampled plots.

Ages	DBH (cm)		H (m)		Basal Area (m ² ·ha)		Density (Trees·ha)	
	Mean	Range	Mean	Range	Mean	Range	Mean	Range
Young stand	8.6	5.6~11.0	18.51	12.94~26.16	21.6	16.8~24.7	1825	1700~1950
Middle-aged stand	12.7	11.8~13.3	15.81	13.55~24.86	25.2	23.2~28.7	1589	1283~1750
Near-mature stand	17.7	16.6~19.5	20.27	19.28~21.16	33.6	27.3~49.9	1233	1050~1467
Mature stand	22.8	16.9~31.5	26.82	23.85~29.41	26.4	16.8~33.4	672	417~1133

Notes: DBH, diameter at breast height; H, height of a tree.

On the basis of the sample plot investigation, four standard plots were selected in areas of the four forest ages. Each standard plot comprised one superior tree, one inferior tree, and two standard trees [39]. Thus, 64 sample trees were selected (Table 2). The aboveground survey adopted the Monsic layered cutting method [40–42] to determine the weight of fresh stem, bark, branch, and leaf of each sample tree. The underground survey used the whole-root excavation method to determine the weight of fresh small roots (≤ 2 cm), thick roots (2–5 cm), and rhizomes (≥ 5 cm). After each element was sampled separately, it was processed for 30 min in an oven at 105 °C. Then, the temperature of the oven was adjusted to 80 °C to dry the sample to constant mass, and the moisture content of each component sample was measured. The biomass of each component was calculated by multiplying the moisture content of the sample by the total fresh mass.

Following similar procedures, 10 sample plots were designated in the independent sample inspection area, and individual trees were surveyed (Table 3). All data collected and survey standards adopted were consistent with those used in the study area to ensure uniformity between the samples from the inspection area and those from the study area.

Table 2. Summary statistics of the sampled trees.

Ages	Numbers of Trees	DBH/cm	Height/m	Stem/kg	Bark/kg	Branch/kg	Leaf/kg	Root/kg
Young stand	16	3.4~9.2	3.7~13.8	1.7~15.9	0.3~2.9	0.7~4.8	0.6~2.0	0.8~5.2
Middle-aged stand	16	9.2~13.1	6.5~17.1	10.1~48.6	2.1~7.1	3.7~9.0	1.5~3.1	4.7~11.9
Near-mature stand	16	12.9~16.7	11.3~18.4	26.9~106.4	4.0~12.3	5.4~12.3	1.8~3.5	7.4~20.1
Mature stand	16	17.2~23.0	11.7~20.8	64.0~185.7	7.9~19.2	8.7~19.4	2.5~5.6	15.8~49.9

Note: Stem, bark, branch, leaf, and root represent the biomass of different tree parts.

Table 3. Sampled plot statistics of the test area.

Sample NO.	DBH (cm)		Basal Area (m ² ·ha)	Density (Trees·ha)
	Mean	Range		
1	8.85	5.3~15.4	18.46	2850
2	10.63	5.5~19.7	16.78	1783
3	10.29	4.8~16.2	22.13	2517
4	12.53	7.0~18.6	15.85	1233
5	19.19	6.5~26.4	16.44	550
6	16.87	9.5~23.4	16.25	700
7	17.61	10.1~25.5	18.86	750
8	16.64	6.5~22.2	21.08	933
9	17.03	5.6~23.5	16.38	683
10	17.80	9.9~26.3	15.64	600

Notes: DBH, diameter at breast height.

2.2.2. Remotely Sensed Data

(1) LiDAR Data

In June 2016, an observational flight was conducted simultaneously with the ground-based sample site survey. The LiCHy (LiDAR, CCD, and Hyperspectral) system was used to collect LiDAR point clouds and high-definition images of the study area. The detailed parameters of the LiCHy system are available in the literature [43]. A “Yun-5” aircraft was used as the flight platform. The observations were acquired as the aircraft flew at an average height of 2500 m with relative ground speed of approximately 200 km·h⁻¹. A total area of 300 km² was obtained. The average scan width of each flight strip was approximately 1000 m, and the overlap rate of adjacent flight strip was 60%. The coverage of the flight path strips is shown in Figure 2. The average point cloud density within the focusing region is 3.6 pts/m². An improved progressive encryption triangulation filtering algorithm was used to classify the point cloud into ground points and non-ground points [44]. Using the elevation threshold method (elevation: >2 m) in combination with manual editing with ArcGIS 10.8.2 software for several days, the non-ground points were divided into vegetation points and non-vegetation point cloud types. The classified ground points were applied to the triangulated irregular network interpolation method to generate a digital elevation model (DEM), and the difference between the vegetation point and the digital elevation model was derived to obtain the canopy height model (CHM).

The height and density variable groups of the LiDAR point cloud data (3.6 pts/m²) extracted from the sample plot statistical unit were taken as the input variables of the *L. olgensis* biomass estimation model [45]. The altitude variable groups included 15 altitude percentiles ($H_1, H_5, H_{10}, \dots, H_{95}, H_{99}$), 15 cumulative altitude percentiles ($AIH_1, AIH_5, AIH_{10}, \dots, AIH_{95}, AIH_{99}$), maximum height (H_{max}), minimum height (H_{min}), average height (H_{mean}), height median (H_{median}), height percentile interquartile range ($H_{interval}$), height kurtosis ($H_{kurtosis}$), mean deviation (H_{mab}), coefficient of variation (H_{cv}), standard deviation (H_{stddev}), and so on. The 10 density variables group was used to divide the point cloud (>2 m) into 10 equal height layers from low to high, and the ratio of the number of echoes in each layer was taken as the corresponding density variable ($D_5, D_{10}, D_{20}, \dots, D_{80}, D_{90}$). Figure 3 shows the height and density variables extracted from LiDAR point cloud data.

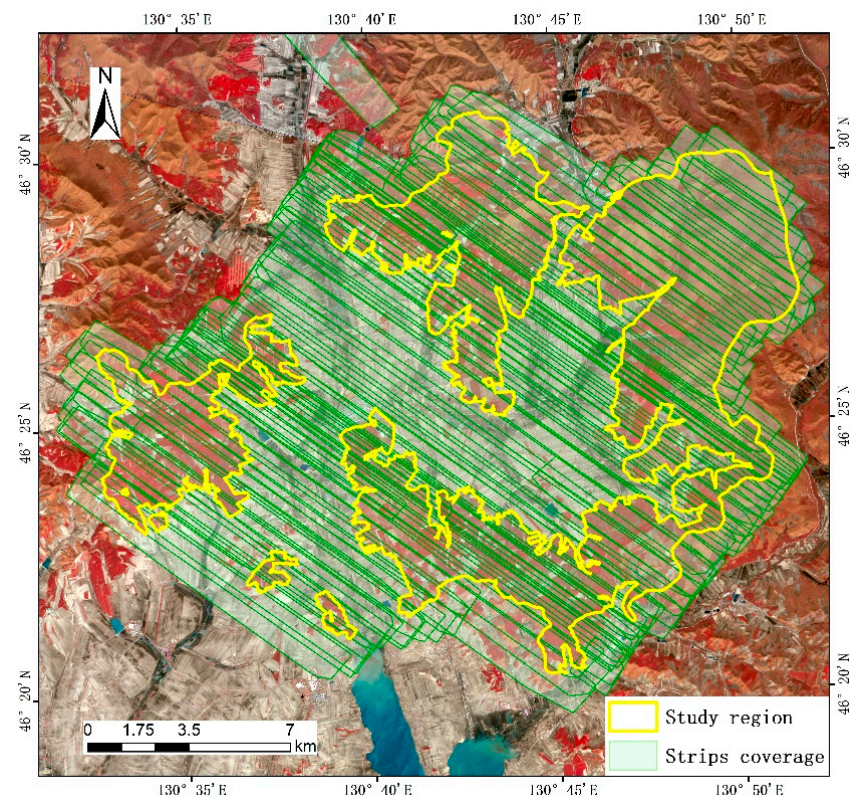


Figure 2. Coverage of experimental strips.

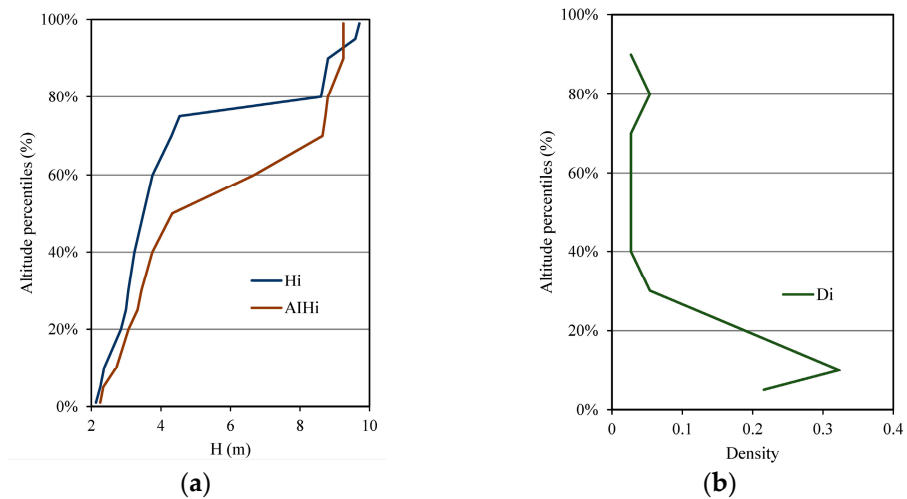


Figure 3. Display of height and density variables extracted from LiDAR point cloud data. (a) Altitude variable. (b) Density variable.

(2) Optical Data

Two scenes acquired by the GF-1 satellite PMS1 camera on 26 March and 6 July 2016 (orbit number: 26 March, E 130.7/N 46.3, E 130.8/N 46.6. 6 July, E 130.7/N 46.6, E 130.6/N 46.3) were selected for extraction of *L. olgensis* distribution information and the construction of the biomass estimation extrapolation model. Additionally, one scene acquired by the GF-1 satellite PMS1 camera on 18 September 2018 (orbit number: E 129.9/N 45.5) was selected for evaluation of the biomass estimation extrapolation model. Three scenarios of the selected images were displayed in Figure 4. The detailed information of the data is shown in Table 4. The data were provided by the China Center for Resources Satellite Data and Application (<https://www.cresda.com/zgzywxyyzx/index.html>, accessed on

20 October 2023) at a size of 2 m × 2 m for multispectral images and 8 m × 8 m for panchromatic images.



Figure 4. Multispectral image of GF-1 satellite PMS1 camera on 26 March 2016. (a–c) Different scenarios.

Table 4. Detailed Information of optical data.

Satellite/Sensor	Resolution	Date	Orbit Number	Use
GF-1/PMS1	2 m/8 m	26 March 16	E 130.7/N 46.3, E 130.8/N 46.6;	Extraction of <i>L. olgensis</i> distribution information and the construction 171 of the biomass estimation model
		6 July 16	E 130.7/N 46.6, E 130.6/N 46.3	
		18 September18	E 129.9/ N45.5	Evaluation of the biomass estimation model

The GF-1 data should be preprocessed before further application, which included radiometric calibration, atmospheric correction, geometric correction, and image fusion; the spectral features, vegetation indexes, and texture feature indicators were extracted. Spectral features included the band information (Band 1, Band 2, Band 3, and Band 4), mean (\bar{x}), variance (σ^2), standard deviation (σ), and entropy ($h(x)$). The vegetation indexes included the normalized difference vegetation index, vegetation index ratio, difference vegetation index, red and green vegetation index, and soil-adjusted vegetation index. The eight most commonly used statistics in the gray level co-occurrence matrix were selected for texture features: mean, variance, homogeneity, contrast, dissimilarity, entropy, correlation, and second moment. The parameters were set as follows: the grayscale compression parameter had 64 levels, the sliding window was 9×9 , the step size was 1, and the direction was 135° (Table 5).

Table 5. Processed set-up for gray-level co-occurrence matrix.

Grayscale Compression Parameters	Sliding Window	Step Size	Direction
64 levels	9×9	1	135°

2.3. Methods

We deployed a deep learning method to create a regional biomass extrapolation model grounded in GF-1 data. We meticulously evaluated the extrapolation efficacy of this model using independent sample datasets. The schematic representation of our comprehensive technical workflow is illustrated in Figure 5.

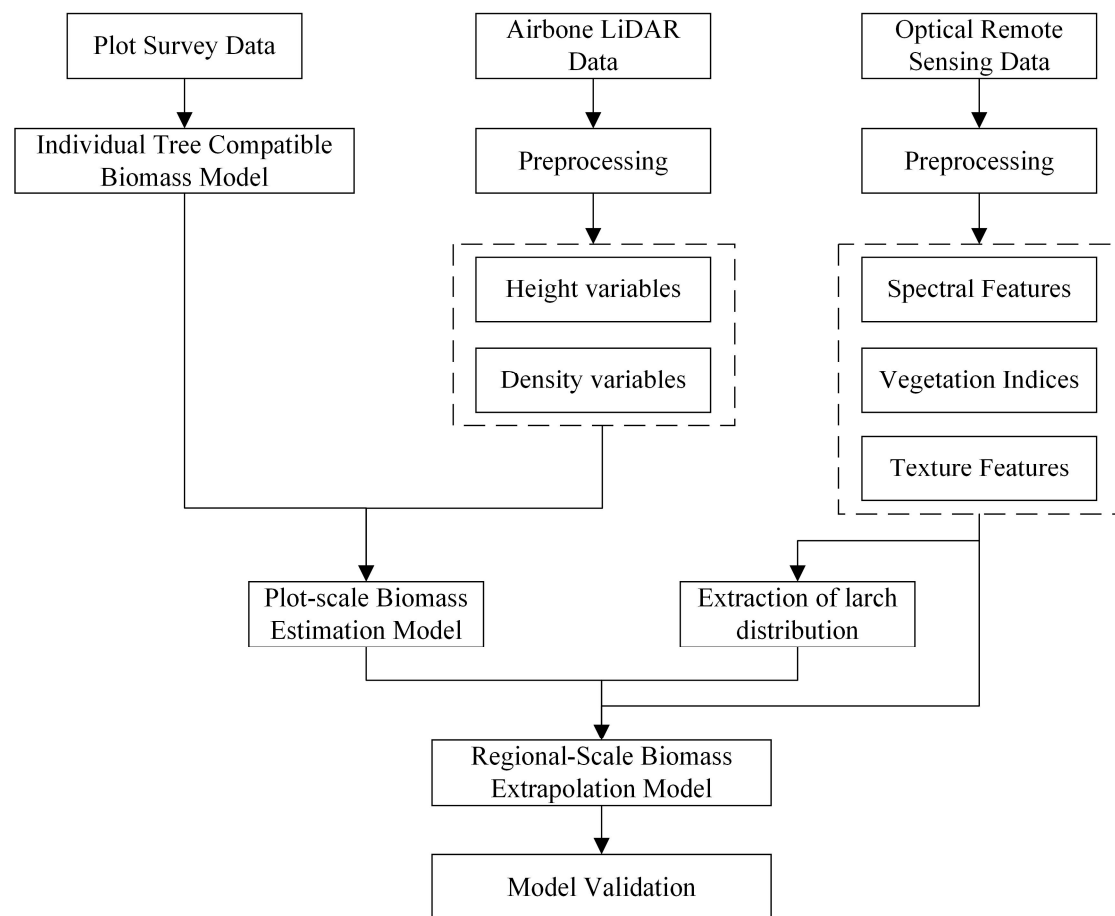


Figure 5. Flowchart of multiscale estimation of forest biomass.

2.3.1. Build a Basic Model of Larch Biomass Compatibility

This study was based on analytical data from 64 sampled trees and the allometric growth model, and a compatible biomass model for larch plantations in the study area was constructed using the non-linear, seemingly unrelated regressions [46–48]. Subsequently, the established single-tree-level biomass model was used to calculate the total biomass and the biomass of each component in each plot. The total biomass should be the sum of each component (trunk, bark, branches, leaves and roots). In this paper, the biomass modeling was adopted to consider the compatibility of total biomass and each component biomass. Details of the specific process are available in the literature [49]. Furthermore, 400 samples were taken to training model and the remaining 100 samples were used as verification data.

2.3.2. Build a Plot-Scale Biomass Component Estimation Model on the Basis of Airborne LiDAR Data

In this study, multiple linear regression (MLR) and random forest (RF) methods were used to establish a plot-scale biomass component estimation model on the basis of airborne LiDAR data. The RF approach is a type of machine learning method that is widely applied to data mining and provides the benefits of good fitting effects and generalization ability [50]. The dependent variable was the biomass, and the independent variable was all the extracted parameters extracted in the second part.

2.3.3. Extraction of Larch Distribution Information on the Basis of Vegetation Phenology Characteristics

Extraction of larch distribution information was the prerequisite for achieving regional-scale biomass estimation. The approach made full use of the phenological characteristics of the deciduous winter leaves of larch among the coniferous species in the study area and

used the GF-1 data of the two phases of winter (March) and summer (June) to improve the accuracy of larch information extraction. The support vector machine (SVM) approach was used to extract the distribution information of larch in the study area.

Land-cover types in the study area were divided into two types: forest land and non-forest land. Forest land included three secondary types: larch, other conifers, and broadleaved trees. Non-forest land types included three secondary types: cultivated land and unused land, construction land, and water area. Using GF-1 data acquired in the two periods, the spectral features, vegetation indexes, and texture feature variable groups were extracted as SVM input variables. In accordance with the established classification system, an SVM-supervised classification and interpretation symbol library was established using the field survey data and analysis of the GF-1 remote sensing images, and a total of 3000 GPS interpretation symbols were collected according to the six secondary types, i.e., 500 samples of each type (Figure 6). For each type, 400 samples were taken for the training model, and the remaining 100 samples were used as verification data.

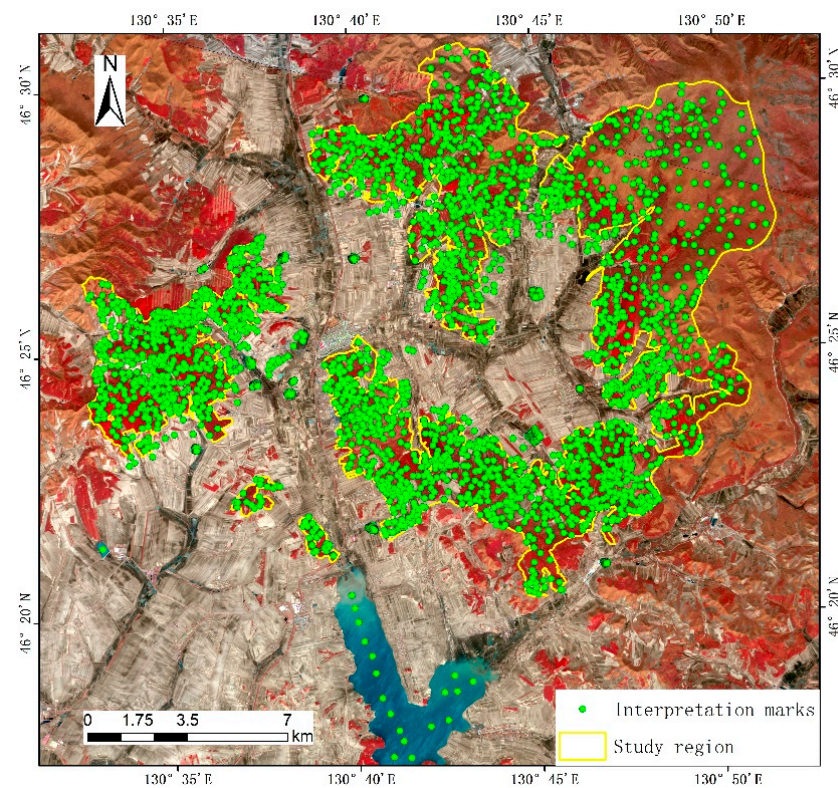


Figure 6. Spatial distribution of field samples.

When using SVM to solve a classification problem, the core difficulty is to determine the appropriate choice of the kernel function. This study randomly selected 200 test samples to pretrain the most commonly used kernel functions (i.e., linear kernel, quadratic polynomial kernel, third polynomial kernel, fine Gaussian kernel, medium Gaussian kernel, and coarse Gaussian kernel) to classify the overall accuracy and prediction speed. The evaluation results of indicators such as classification accuracy and training time were used as the basis for selecting the optimal kernel function. The results showed that the quadratic polynomial kernel had the highest classification accuracy (overall accuracy: 88.3%), shortest training time (14.154 s), and fastest prediction speed (18,000 items/s). Therefore, it was chosen for use as the classifier in this study. The SVM classification process was implemented in MATLAB R2023b.

The evaluation of the accuracy of the results of information extraction used a combination of qualitative and quantitative analyses. The qualitative analysis comprised comparison with the results of a forest resource census performed in the study area to

establish whether there were too many points, missing points, or wrong points. Quantitative analysis was achieved by constructing a confusion matrix. The overall accuracy, user accuracy, mapping accuracy, kappa coefficient (Equation (1)), and other indicators were calculated to test the degree of agreement between the classification results and the verification sample while taking into account the correctly classified diagonal samples and noncorresponding samples that were missed and misclassified at the corners [51–53]:

$$\text{Kappa} = \frac{N \sum_{i=1}^m x_{ii} - \sum_{i=1}^m (x_{i+} x_{+i})}{N^2 - \sum_{i=1}^m (x_{i+} x_{+i})}, \quad (1)$$

where m is the total number of columns in the error matrix, which is also the total number of classifications. x_{ji} is the number of pixels in row i and column i of the confusion matrix, representing the number of samples classified correctly, and x_{i+} and x_{+i} are the total numbers of pixels in the row and column, respectively. N is the total number of pixels used for the accuracy evaluation.

2.3.4. Construction of a Model for Extrapolation of Biomass Components at the Regional Scale

A 20×30 m sample plot was systematically arranged within the distribution range of larch extracted in Section 2.3.3, and the biomass components of each plot were retrieved in accordance with the RF method outlined in Section 2.3.2. Overall, 53,348 initial samples were obtained. To enable the accuracy and objectivity of the fitted model, only 38,276 samples were finally selected through a boxplot method. The sample data were divided into training data (26,793 samples), verification data (3828 samples), and prediction data (7655 samples) in a ratio of 7:1:2. In the 10 sample plots established in the independent sample inspection area, the method of Section 2.3.1 was used to calculate both the total biomass and the biomass of each component of each sample plots, which were used as observational values to test the generalization capability of the biomass extrapolation model.

In this section, the RF and LSTM neural networks were used to construct a regional-scale biomass component extrapolation model. Recurrent neural network (RNN) belongs to the deep learning algorithms in machine learning, which is suitable for complex nonlinear problems and can mine the relationship between remote sensing and forest biomass. However, trapped by the problems of gradient explosion and gradient disappearance, simple RNN may not be the best model for predicting time-series data with long-term dependence [54,55]. LSTM, as a special recurrent neural network, inserts a storage unit to the hidden layer of the RNN, which controls the flow of information through each unit and the neural network using input gates, forget gates, and output gates, thereby avoiding the common problems of recurrent neural network [56–58].

The input is x_t , the hidden layer output is h_t , the previous output is h_{t-1} , the unit input state is \tilde{C}_t , the unit output state is C_t , the previous state is C_{t-1} , and the input gate, forget gate, and output gate states are i_t ($i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$), f_t ($f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$), and o_t ($o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$), respectively. The structure of the LSTM unit indicates that C_t and h_t were transmitted to the next neural network in the recurrent neural network. To calculate C_t and h_t , it is necessary to first calculate the state of each of the three gates and the unit input state ($\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c)$), where W_f , W_i , W_o , and W_c (representing h_{t-1}) and x_t connect the weights on the forget gate, input gate, output gate, and unit state input, respectively, and b_f , b_i , b_o , and b_c represent their corresponding offset items, respectively.

To predict the whole-plant biomass and the biomass of its components, a three-layer neural network comprising an input layer, hidden layer, and output layer was designed. The input layer contained 44 input neurons, i.e., 44 factor variables extracted from the GF-1 remote sensing images. The hidden layer comprises two LSTM layers, and each layer has 100 neurons. To avoid overfitting of the LSTM model, the dropout parameter value (which was a deactivation probability value between 0 and 1) was added after each LSTM layer.

The output layer was a fully connected neural network that included seven output neurons, i.e., the seven biomass factors to be predicted. The mean systematic error (MSE) was chosen as the loss function in the LSTM biomass prediction model. The smaller the MSE, the closer the predicted value was to the actual value, the higher the accuracy of the model, and the stronger the prediction capability. Adam was used as the optimizer of the adaptive learning rate, and the initial value of the learning rate was set to 0.001. The deactivation probability value of dropout was set to 0.2, the activation function of the output layer fully connected output layer of the network was set to the function, the batch size of each feed to the neural network was set to 50, and the number of iterations was set to 1000 [59].

2.3.5. Model Evaluation

The model results in this study were evaluated using several indicators, including the decision coefficient (R^2), root mean square error (RMSE), relative root mean square error (rRMSE), adjusted decision coefficient (R_{adj}^2), standard deviation of the estimated value (SEE), total relative error (TRE), mean percentage error (MPE), mean squared error (MSE), and mean percentage standard error (MPSE). To make full use of the effective information of the samples, the 10-fold cross-validation method was adopted in the modeling, and the average value of the cross-validation results was taken as the final model evaluation index. Equations for the formulation of each of the indicators can be expressed as follows:

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y}_i)^2}, \quad (2)$$

$$RMSE = \sqrt{\frac{1}{n} \sum (y_i - \hat{y}_i)^2}, \quad (3)$$

$$rRMSE = \frac{\sqrt{\frac{1}{n} \sum (y_i - \hat{y}_i)^2}}{\sum y_i}, \quad (4)$$

$$R_{adj}^2 = 1 - \frac{(n-1) \sum (y_i - \hat{y}_i)^2}{(n-p) \sum (y_i - \bar{y}_i)^2}, \quad (5)$$

$$SEE = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{(n-p)}}, \quad (6)$$

$$TRE = \frac{\sum (y_i - \hat{y}_i)}{\sum \hat{y}_i} \times 100, \quad (7)$$

$$MPE = t_{\alpha} \cdot (SEE / \bar{y}) / \sqrt{n} \times 100, \quad (8)$$

$$MSE = \frac{\sum (y_i - \hat{y}_i)}{\sum \hat{y}_i / n} \times 100, \quad (9)$$

$$MPSE = \frac{\sum |y_i - \hat{y}_i|}{\sum \hat{y}_i} / n \times 100 \quad (10)$$

where n is the total number of training samples, y_i is the observed value, \hat{y}_i is the estimated value, \bar{y}_i is the average value of the samples, t_{α} is the t value at the confidence interval of α , and p is the number of parameters.

3. Results

3.1. Basic Model of Larch Biomass Compatibility

The results showed that the R_{adj}^2 values of the biomass models of the whole plant and each component (except leaves) were >0.91 , and those of the aboveground and whole-plant biomass models were >0.95 , indicating that DBH explains more than 95% of the changes in the standing tree biomass (Table 6). The small values of SEE, TRE, and MSE indicate that the two models have good fitting effects. The MPE for the whole plant and for each component was approximately 4%–7%, indicating that the average prediction accuracy

of the model was >93%. In summary, the model has good fitting effect and prediction accuracy that meet the requirements of application.

Table 6. Parameter estimates and evaluation statistics of the biomass model.

Components	Parameter Estimates				Evaluation Statistics			
	a	b	R^2_{adj}	SEE (t)	TRE (%)	MPE (%)	MSE (%)	MPSE (%)
Stem	0.06	2.51	0.97	14.59	9.03	7.03	6.31	18.47
Bark	0.03	1.99	0.95	1.78	7.70	7.21	6.52	17.34
Branch	0.16	1.46	0.92	1.69	5.91	5.40	5.98	16.81
Leaf	0.24	0.91	0.84	0.49	5.69	4.73	5.69	16.21
Root	0.04	2.10	0.96	4.20	10.85	8.01	7.50	17.15
Aboveground			0.96	18.01	8.41	6.58	6.38	16.74
Total			0.96	21.02	8.80	6.44	6.53	16.38

Note: $B_{Above} = B_{Stem} + B_{Bark} + B_{Branch} + B_{Leaf}$. $B_{Total} = B_{Above} + B_{Root}$. The biomass model of each component is $\ln B_i = \ln(a_i) + b_i \ln D_i$, where a and b represent the fitting parameters in the biomass model. i is components. R^2 , SEE, TRE, MPE, MSE, and MPSE represent evaluation indicators referenced in Section 2.3.5.

3.2. Plot-Scale Biomass Component Estimation Model Based on Airborne LiDAR Data

Tables 7 and 8 show that LiDAR variables had strong correlation with biomass, and the correlations between Hinterval, H_{80} , D_{10} , and D_{20} and the biomass of each component obtained by stepwise regression were generally significant ($p < 0.05$) or extremely significant ($p < 0.01$). Except for branches and leaves, the biomass models constructed using the MLR method have $R^2 > 0.82$, and the R^2 for stem, ground, and whole-plant models were >0.90 . The biomass models constructed using the RF method have $R^2 > 0.91$. Moreover, they have smaller rRMSE and TRE values, indicating that the models constructed using the two methods have better fitting effects and that the explainable variation accounts for a higher proportion of the total.

Table 7. Parameter estimates and evaluation statistics of the MLR model.

Components	Parameter Estimates				Evaluation Statistics			
	a_0	b_1	b_2	b_3	R^2	RMSE (t)	rRMSE	TRE (%)
Stem	−2.76 **	−0.63 **	1.70 **	11.85 **	0.91	0.41	0.08	0.60
Bark	−3.12 **	−0.88 **	1.19 **	10.78 **	0.83	0.05	0.08	0.66
Branch	−0.87 *	−0.79 **	0.44 **	7.35 *	0.54	0.07	0.10	0.91
Leaf	0.11	−1.10 **	−0.167	5.86	0.634	0.15	0.68	0.96
Root	−2.57 **	−0.60 **	1.145 **	8.38 **	0.87	0.08	0.067	0.45
Aboveground	−1.59 **	−0.65 **	1.40 **	10.58 **	0.91	0.46	0.067	0.45
Total	−1.30 **	−0.64 **	1.36 **	10.22 **	0.91	0.534	0.07	0.45

Note: *, significant correlation ($p < 0.05$); **, extremely significant correlation ($p < 0.01$). The model form of branch composition is $\ln B_{Branch} = a_0 + b_1 \ln H_{interval} + b_2 \ln H_{90} + b_3 \ln D_{20}$. The form of other component models is $\ln B_{components} = a_0 + b_1 \ln H_{interval} + b_2 \ln H_{90} + b_3 \ln D_{10}$, where a_0 , b_1 , b_2 , and b_3 represent the fitting parameters in the biomass model. R^2 , RMSE, rRMSE, and TRE represent evaluation indicators referenced in Section 2.3.5.

Comparison of the two methods revealed that the RF model had better fitting effect. In terms of the B_{stem} , B_{above} , and B_{total} models, the R^2 was 0.97 for RF and 0.91 for MLR, and the values of rRMSE and TRE were similar. In terms of the branch and leaf models, the R^2 of the MLR model was only 0.5–0.6, while that of the RF model was >0.9 , indicating that the RF model had the better fitting effect. Additionally, irrespective of model type (MLR or RF), underestimation of the interval of biomass $B_{total} > 8$ ton increased substantially (Figure 7). The high-precision RF method was used to invert the biomass of larch in the study area (Figure 8), and the results show satisfactory agreement with the observed values.

Table 8. Evaluation statistics of the RF model.

Components	R^2	RMSE (t)	rRMSE	TRE (%)
Stem	0.97	0.54	0.11	1.067
Bark	0.96	0.05	0.09	0.73
Branch	0.92	0.07	0.10	0.94
Leaf	0.91	0.03	0.14	1.94
Root	0.96	0.10	0.09	0.71
Aboveground	0.97	0.62	0.10	0.85
Total	0.97	0.72	0.09	0.81

Note: R^2 , RMSE, rRMSE, and TRE represent evaluation indicators referenced in Section 2.3.5.

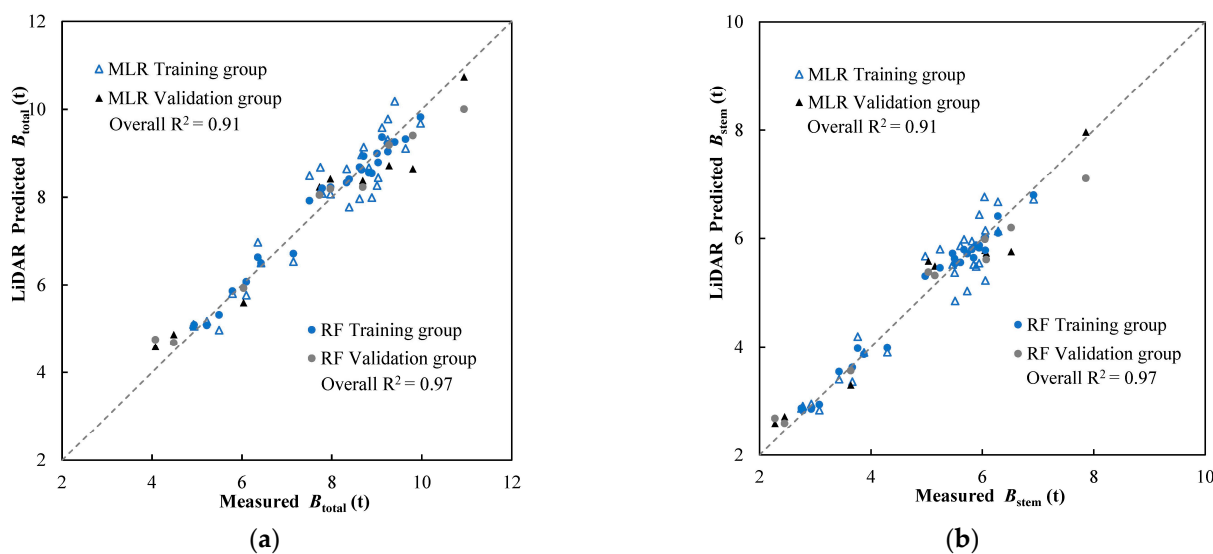


Figure 7. Comparison of biomass from field-measured and airborne LiDAR data. (a) Biomass of whole plant. (b) Biomass region of stem. The unit of biomass is ton. MLR, multiple linear regression; RF, random forest.

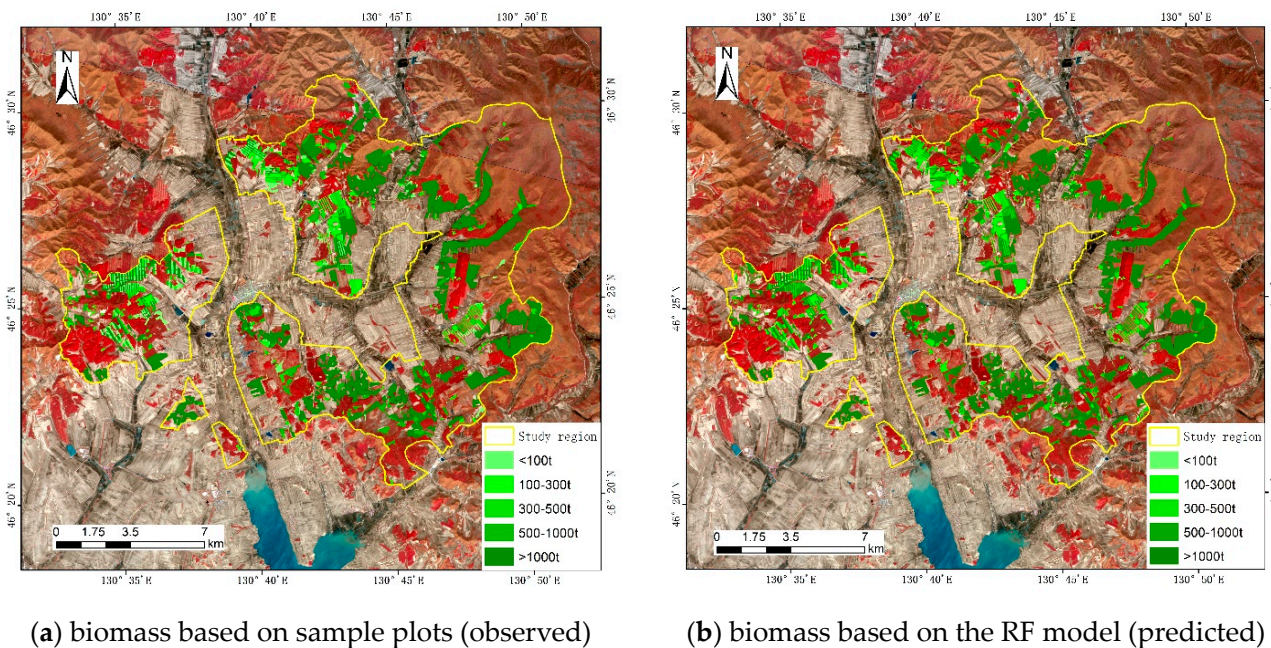


Figure 8. Comparison between observed and predicted values of total biomass in the study area.

3.3. Extraction of Larch Distribution Information Based on Vegetation Phenology Characteristics

The classification results were showed in Figure 9. The overall accuracy of the classification results was 90.0%, and the kappa coefficient was 0.88, indicating that the correct matching rate between the classification results and the actual land cover types was >90.0% (Table 9), and the numbers on the diagonal in the table represent the number of correctly classified pixels. The overall results were considered satisfactory and meet the requirements of subsequent work. In terms of user accuracy, the classification accuracy of water was highest, reaching 93.0%, indicating that this land-cover type had the highest probability of being classified correctly. The classification accuracy of conifers was the same as that of cultivated land and unused land, i.e., 92.0%, and the classification accuracy of larch was 91.0%. The classification accuracy of broadleaved trees and construction land was slightly lower, i.e., only 88.0% and 84.0%, respectively. In terms of mapping accuracy, the accuracy of the water area was the highest, reaching 96.9%, indicating that water areas within the study area have the highest probability of being classified correctly. The classification accuracy of larch and broadleaved trees was 95.8% and 95.7%, respectively, and that of other conifers was 92.9%. The classification accuracy of construction land was 86.6%, while that of cultivated land and unused land was only 76.0%.

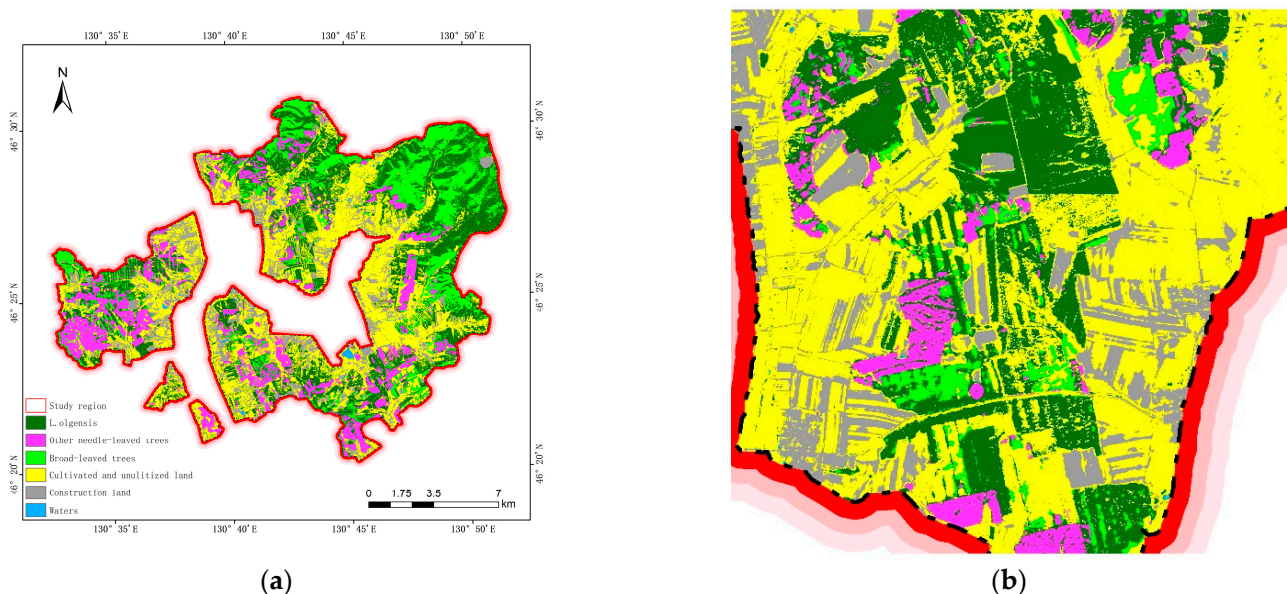


Figure 9. Classification results based on support vector machine. (a) The study area. (b) Enlarged view of part of the area.

Table 9. Confusion matrix and precision evaluation of *L. olgensis* support vector machine classification.

Classification Types	<i>L. olgensis</i>	Other Coniferous	Broad-Leaved Trees	Cultivated and Unutilized Land	Construction Land	Waters	User Accuracy(%)
<i>L. olgensis</i>	91	3	1	5			91.0
Other coniferous	3	92	2	2	1		92.0
Broad-leaved trees	1	3	88	6	2		88.0
Cultivated and unutilized land		1		92	7		92.0
Construction land			1	12	84	3	84.0
Waters				4	3	93	93.0
Graphic accuracy (%)	95.8	92.9	95.7	76.0	86.6	96.9	

Overall accuracy = 90.0%; kappa coefficient = 0.88.

3.4. Construction of a Regional-Scale Biomass Component Extrapolation Model Based on Multisource Data Fusion

The RF and LSTM methods demonstrated strong capability in fitting and predicting the biomass of larch, and the prediction accuracy of the whole-plant biomass and the biomass of its components was broadly > 70% (Table 10). In terms of the whole-plant, aboveground, and tree root biomass, the statistical results presented in Table 10 confirm that the fitting effect of the LSTM model was slightly better than that of the RF model, with higher R^2 and lower $rRMSE$ and TRE . In terms of the bark, branch, and leaf component models, the RF model performs slightly better than the LSTM model. For the stem model, the indicators reflecting the fitting effect and prediction accuracy of the two models were broadly consistent.

Table 10. Evaluation statistics of comparison of biomass extrapolation using the LSTM and RF models.

Components	Evaluation Statistics							
	R^2		RMSE (t)		$rRMSE$		TRE (%)	
	RF	LSTM	RF	LSTM	RF	LSTM	RF	LSTM
Stem	0.70	0.70	0.68	0.62	0.14	0.14	1.92	1.93
Bark	0.69	0.66	0.04	0.04	0.08	0.08	0.64	0.69
Branch	0.35	0.26	0.04	0.04	0.06	0.06	0.32	0.39
Leaf	0.50	0.44	0.02	0.02	0.10	0.11	0.92	1.11
Root	0.71	0.71	0.10	0.10	0.09	0.08	0.76	0.75
Aboveground	0.71	0.72	0.62	0.61	0.11	0.10	1.09	1.06
Total	0.71	0.71	0.74	0.72	0.11	0.10	1.08	1.04

Note: R^2 , $RMSE$, $rRMSE$ and TRE represent evaluation indicators referenced in Section 2.3.5.

From Table 11, among the 7655 predicted samples, 3897 were underestimated when using the LSTM method, accounting for 50.9% of the predicted samples, i.e., broadly the same as the number of overestimated samples. Under the RF method, 4066 samples were underestimated, accounting for 53.1% of the predicted samples, i.e., an increase of 2.2 percentage points in comparison with the LSTM model. When $B_{total} > 8$ t for the 2336 samples, under the LSTM method, 1824 samples were underestimated, accounting for 78.1% of the predicted samples. Under the RF method, 1966 samples were underestimated, accounting for 84.2% of the predicted samples, i.e., an increase of 6.1% in comparison with the LSTM model. Additionally, the average biomass based on the LiDAR estimation model was 7.0104 t (i.e., the true value of the extrapolated model), the estimated average biomass of the LSTM extrapolated model was 6.9993 t, and the average estimated biomass of the RF extrapolated model was 6.9450 t. In summary, in comparison with the RF model, the LSTM model was less prone to underestimation of biomass, and this characteristic becomes more obvious with increase in the sample unit biomass.

Table 11. Comparison of the LSTM and RF models in terms of estimated biomass.

Models		All Samples		Samples of Biomass > 8 ton	
		Overestimated	Underestimated	Overestimated	Underestimated
LSTM	Number of samples	3758	3897	512	1824
	Proportion	49.1%	50.9%	21.9%	78.1%
RF	Number of samples	3589	4066	370	1966
	Proportion	46.9%	53.1%	15.8%	84.2%

It can be seen from Table 12 that the biomass of the aboveground components and whole-plant biomass of the LSTM model have the highest R^2 , reaching 0.65 and 0.63, respectively, and $rRMSE$ and TRE were only 0.07 and 0.43%, respectively, indicating that the model had high prediction accuracy. The R^2 of the stem, root, and bark biomass models were 0.53, 0.50, and 0.45, respectively, with smaller $rRMSE$ and TRE . The prediction effect

of branches and leaves was poor, with R^2 of only 0.08 and 0.20, respectively. Overall, the generalization capability of the extrapolation model of larch component biomass based on the LSTM method was considered satisfactory, and the prediction accuracy of the model can reach >50%.

Table 12. Evaluation statistics of the LSTM model in the test area.

Components	Evaluation Statistics			
	R^2	RMSE (t)	rRMSE	TRE (%)
Stem	0.53	0.31	0.11	1.01
Bark	0.45	0.04	0.09	0.71
Branch	0.08	0.11	0.19	3.31
Leaf	0.20	0.06	0.28	7.06
Root	0.50	0.06	0.08	0.58
Aboveground	0.65	0.29	0.07	0.43
Total	0.63	0.35	0.07	0.43

Note: R^2 , RMSE, rRMSE and TRE represent evaluation indicators referenced in Section 2.3.5.

4. Discussion

4.1. Performance of Regression-Based Method and Machine Learning Algorithms

In this study, we employed two different modeling approaches to construct plot-scale biomass models for various components using LiDAR data: the conventional regression-based method, MLR, and a machine learning algorithm, RF. These choices were made for specific reasons. Firstly, MLR, representing regression methods, has a long history of use in establishing growth functions for diverse forest species, making it a well-established choice. Secondly, we opted for the RF algorithm due to its inherent capability to capture non-linear relationships between input features and outcomes. The results demonstrated that it is very effective to use the height and the density information from LiDAR data to retrieve forest biomass [23]. Our machine-learning-based approach for larch aboveground biomass (AGB) estimation surpasses traditional methods such as MLR. In a prior study [60], utilizing only optical RS observations resulted in an R^2 value of about 0.36 for larch AGB estimation. However, by incorporating stereo features, we were able to enhance the accuracy to 0.64. Reducing the number of input variables and the spatial resolution of RS observations had an adverse effect on larch AGB estimation accuracy. For instance, when using greenness RS indices with varying spatial resolutions, the R^2 increased from 0.29 to 0.49, while employing NDVI increased the R^2 from 0.26 to 0.79 [61]. Conversely, including unrelated input variables did not enhance larch AGB estimation, as the R^2 value remained around 0.3 despite the incorporation of numerous input variables [62]. Furthermore, the effective AGB estimation performance can, in part, be attributed to the fact that the study area under focus consists of a single tree-height layer [63]. When comparing the two modeling techniques, we observed that the RF model demonstrated superior fitting effects and generalization capabilities. This enhanced performance can be attributed to the RF approach's strong learning capacity and stable generalization accuracy. Furthermore, it is also linked to the complex, nonlinear relationship between forest biomass and remote sensing data. As stand ages increase, the growth efficiency decreases, and RF is better equipped to capture these intricate dynamics. The variables H_{interval} , H_{80} , D_{10} , and D_{20} selected by MLR were found generally related to the biomass of each component at significant ($p < 0.05$) or extremely significant ($p < 0.01$) levels. This signifies that LiDAR variables possess excellent explanatory power concerning biomass variation. Integrating these significant input features into the RF model can simplify the model's complexity, thereby enhancing its efficiency.

4.2. Extraction of Larch Distribution Information

The classification accuracy of larch reached 91.0%, and the mapping accuracy reached 95.8%. Our research follows a conventional forest classification workflow, integrating optical RS reflectance and related indices as inputs for a machine-learning-based classifier.

In comparison to prior large-scale forest classification endeavors, our findings showcase a notably enhanced level of accuracy. For instance, when we compare our results to a similar workflow employing similar input datasets, Yang and Huang reported producer accuracies ranging from 64% to 79% and user accuracies ranging from 69% to 87% for various forest products [64]. This heightened accuracy can be attributed to our strategic focus on a limited area, one relatively free from complicating land-cover categories such as shrublands and plantations. However, it is important to acknowledge that the task of tree species classification has long presented a formidable challenge. The classification accuracy may potentially diminish in future applications within natural forests where there is a more intricate combination of tree species. To mitigate this, employing hyperspectral RS observations proves to be a valuable strategy for bolstering the overall robustness of the classification method [65]. However, there were still misclassifications and omissions between larch and other coniferous species, young larch forests, uncultivated and cultivated land, and unused land. This might be the reason that the extracted optical remote sensing features were introduced into the model to participate in training without being screened, although multitemporal image features can significantly improve the classification effect [66–68]. Under conditions with complex surface covering, the advantages are more significant. However, the resulting optimal feature selection has become a crucial issue, and the optimal temporal features should be selected according to different vegetation biological phenology information, and redundant features should be removed to obtain the optimal classification effect [69]. It has been suggested that an appropriate number of multiple time-series images should be selected to fully reflect the characteristic differences of various vegetation types in different regions. Therefore, in the next step, we considered encrypting additional time-phase images. According to the evaluation results of the classification accuracy of deciduous pine trees, the optimal time phase for extracting larch distribution information was discussed. Analysis of misclassifications and omissions related to extraction of larch distribution information revealed that the probability of misclassification of larch into cultivated land and unused land was the highest (reaching 5.0%). The main reason was that the spatial distribution of larch was broadly the same as that of cultivated land and unused land. Moreover, the larch in the study area were mostly distributed in strips with obvious gaps between forest belts. Therefore, young larch trees and uncultivated land could easily be divided into cultivated land and unused land owing to the large areas of bare land. Furthermore, the phenomenon of mixed classification of larch and other conifers was also more obvious. Each type was misclassified into the category of the other at a rate of 3.0%. The possible main reasons for this were as follows: (1) The image characteristics of spruce and other young coniferous forests and immature forests were very similar to those of the middle-aged and young larch forests, which might cause misclassification, and (2) in a mixed forest of larch and other coniferous species, the image characteristics might be complicated and mixed owing to the variation of tree species composition.

4.3. Extrapolation Model of Biomass Components

In the previous section, we revealed a complex nonlinear relationship between remote sensing variables and biomass of components, and it would be difficult for traditional methods such as MLR to achieve better estimation results. The LSTM and RF methods adopted in this study achieved satisfactory results. Only in terms of the evaluation of the fitting effect and estimation accuracy was the advantage of the LSTM model not obvious. This might be related to the lack of information in the GF-1 band. Although this study supplemented the vegetation indexes and texture index to increase the sample characteristics, these indexes were all dependent on the calculation of band information, which does not substantially increase the feature dimension. The advantages of deep learning such as LSTM reflect its powerful capability in solving complex problems through its deeper net structure. When the complexity of the data is insufficient and the feature dimensions not high, it is difficult to express the capability of the LSTM approach effectively [70–72]. Figure 10 shows the data distribution structure of the predicted values and observations of

different remote sensing models in comparison with the RF model. It can be seen that the median values and the upper and lower quartiles of the predicted data set were closer to the observations for high-value areas. However, the data prediction in the low-value areas was more accurate, and the structure of the prediction result was more consistent with the observed value. To a certain extent, the phenomena of underestimation of high-value areas and overestimation of low-value areas were reduced.

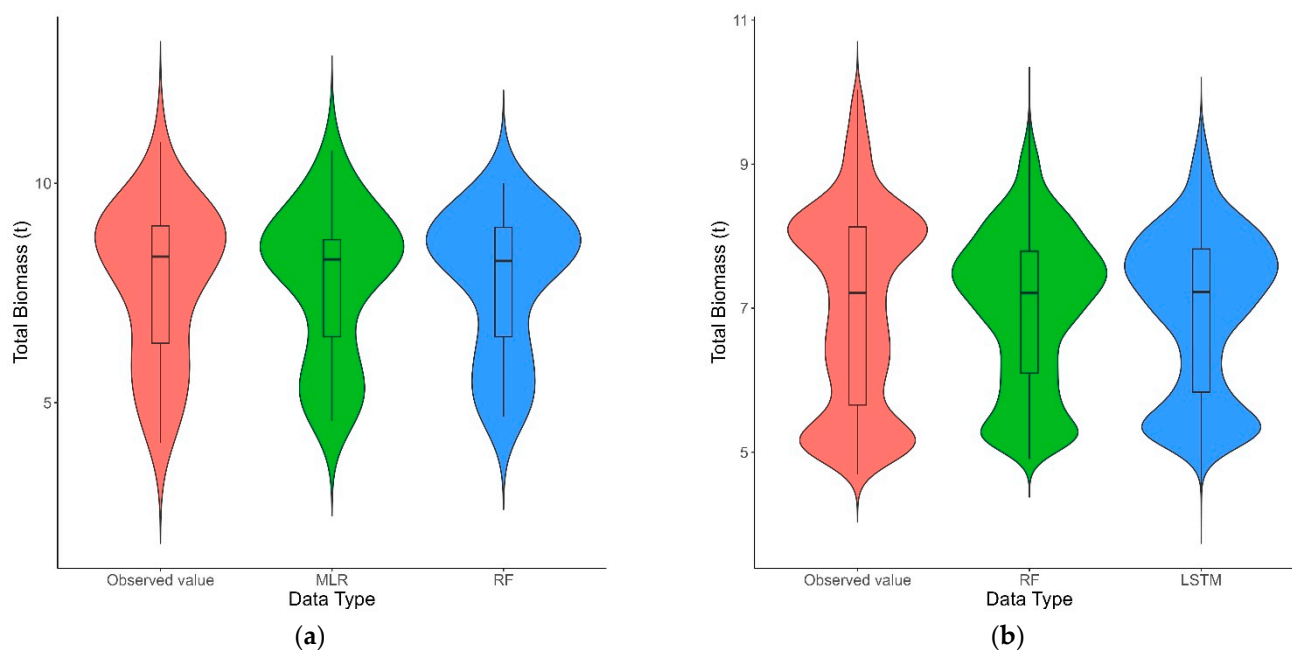


Figure 10. Schematic of each type of data structure. (a) Description of what is contained in the first panel. (b) Description of what is contained in the second panel.

Using independent test samples to evaluate the effect of LSTM model extrapolation, it was found that the R^2 of the whole plant and the biomass model of each component decreased to varying degrees. The time phase of remote sensing images in the study area was July 2016, and the time phase of remote sensing images in the inspection area was September 2018. Because optical remote sensing was very susceptible to environmental factors such as light and atmosphere, the relationship between ground objects and their reflection spectra in different regions and at different time phases was not stable, although the data underwent careful image screening, preprocessing, and normalization in this study. Therefore, it remains difficult to ensure complete consistency in the spectral information of the same features. Controversy remains regarding whether the use of remote sensing images to construct a regional-scale extrapolation model requires the introduction of geographic environmental factors [73,74]. The distribution of larch in the study area has small undulations, gentle slopes, and insignificant differences in site conditions such as soil, which were factors that are difficult to incorporate in model training as variable characteristics. Additionally, the selection of the texture window, the determination of the extrapolation scale, and the corresponding error measurement should be studied further in future work to enhance the generalization capability of the proposed extrapolation model.

4.4. Trade-Off between the Accuracy and Cost in Biomass Estimation

In the course of this research endeavor, we undertook an extensive data collection effort, drawing from diverse sources rather than relying on pre-existing datasets. Our data acquisition process encompassed field observations, airborne LiDAR surveys, and the utilization of high-resolution optical remote sensing imagery, along with derived indices. Throughout the data collection phase, we encountered a noteworthy observation: Traditional biomass investigation proved to be exceedingly time-intensive. It necessitated

approximately one full day to collect data from a sampled plot. It is worth noting that these field-derived data are indispensable for training and evaluating the efficacy of methods grounded in remote sensing imagery. Subsequently, by employing these trained machine learning models in conjunction with remote sensing imagery, we were able to generate a comprehensive, wall-to-wall larch biomass map covering the entire study area. This strategic approach, when juxtaposed with traditional field investigations, not only significantly reduces the time and manual labor required but also minimizes the need for extensive laboratory work. Furthermore, our study demonstrated that machine-learning-based methods, with their inherent computational efficiency, offer a practical alternative. They achieve substantial time savings without compromising on the overall accuracy of forest biomass estimation.

Nonetheless, it is imperative to address several critical considerations in regional biomass estimation studies. Foremost among these concerns is the quality of LiDAR observations. When the density of the LiDAR point cloud is sufficiently high, and the canopy waveform signal is inadequate, there is a non-negligible likelihood of missing the tree tops and canopy edges. This can result in bias in AGB estimation. When the density of the LiDAR point cloud is not high, the canopy echo signal is inadequate, and analysis of the tree tops and canopy edges is likely to miss certain features, resulting in underestimation of the biological amount [45]. On the other hand, it could be related to depletion of the branch and leaf biomass in the field sample collection [75].

5. Conclusions

The *L. olgensis* plantation in Heilongjiang Province considered in this study was surveyed using airborne LiDAR data, ground-based monitoring, and optical remote sensing techniques, and a set of methods, namely rapid, universal, multiscale (single tree, stand, management unit, and region), and unit-high-precision continuous monitoring methods, was proposed for forest biomass components. The analysis indicated that the variables extracted from the airborne LiDAR point cloud data had significant correlation with the biomass of each component. The correlation with the biomass of each component obtained through MLR screening was generally significant ($p < 0.05$) or extremely significant ($p < 0.01$) and thus very suitable for the extraction or estimation of biomass and other indicators.

The incorporation of phenological features resulted in a satisfactory model performance, demonstrating their utility for SVM and other machine learning algorithms. Moreover, the phenological data within remote sensing images can be effectively harnessed through time-series algorithms, such as the memory mechanism offered by LSTM. This enhancement can significantly improve biomass estimation accuracy, particularly when applied at a large scale.

Author Contributions: Conceptualization, Y.H. and D.C.; methodology, C.W.; software, J.X.; validation, Y.P. and S.Z.; formal analysis, Y.H.; investigation, C.W. and B.Y.; resources, D.C.; data curation, J.X.; writing—original draft preparation, Y.H.; writing—review and editing, D.C.; visualization, C.W. and B.Y.; supervision, S.Z.; project administration, Y.P.; funding acquisition, Y.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by The National Key Research and Development Program of China (2022YFD2201001, 2022YFE0112700), Scientific Research Program of Baishanzu National Park: Research on Regional Carbon Neutrality Pathway Facilitated by Baishanzu National Park (2022JBGS08), Project of Industry-University-Research cooperation between Tsinghua University and China Forestry Group Corporation on Forestry carbon sink development (ZLJT-THU2022110101), General Program of National Natural Science Foundation of China (31971652), and National Natural Science Foundation of China (32171787).

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: Author Bo Yang was employed by the company China Forestry Group Corporation. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Glossary

Abbreviation	Definition	Units
LSTM	Long short-term memory	
DBH	Diameter at breast height	cm
H	Height of a tree	m
H_i	Altitude percentiles extracted from LiDAR points	
AIH_i	Cumulative altitude percentiles extracted from LiDAR points	
D_i	Density variable extracted from LiDAR points	
GF-1	Gaofen-1 satellite	
PMS1	One of high-resolution cameras on the GF-1	
MLR	Multiple linear regression	
RF	Random forest	
SVM	Support vector machine	
Kappa	Kappa coefficient, an indicator to evaluate the accuracy of classification	
RNN	Recurrent neural network	
R^2	Coefficient of determination	
R^2_{adj}	Adjusting coefficient of determination	
RMSE	Root mean square error	
rRMSE	Relative root mean square error	
SEE	Standard deviation of the estimated value	
TRE	Total relative error	
MPE	Mean estimation error	
MSE	Mean systematic error	
MPSE	Mean percentage standard error	
a	The fitting parameter in biomass model	
b	The fitting parameter in biomass model	
B	Biomass	ton
B_{Stem}	Biomass of stem	ton
B_{Bark}	Biomass of bark	ton
B_{Branch}	Biomass of branch	ton
B_{Leaf}	Biomass of leaf	ton
B_{Root}	Biomass of root	ton
B_{Above}	Aboveground biomass	ton
B_{Total}	Total biomass	ton
t	Ton	ton

References

- Zeng, W.; Xiao, Q.; Hu, J.; Liao, Z. Establishment of single-tree biomass equations for *Pinus massoniana* in southern China. *J. Cent. South Univ. Technol.* **2010**, *30*, 50–56.
- Myneni, R.B.; Dong, J.; Tucker, C.J.; Kaufmann, R.K.; Kauppi, P.E.; Liski, J.; Zhou, L.; Alexeyev, V.; Hughes, M. A large carbon sink in the woody biomass of Northern forests. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 14784–14789. [[CrossRef](#)]
- Yu, G.; Fang, H.; Fu, Y.; Wang, Q. Research on carbon budget and carbon cycle of terrestrial ecosystems in regional scale: A review. *Shengtai Xuebao/Acta Ecol. Sin.* **2011**, *31*, 5449–5459.
- Awad, M.M. FlexibleNet: A New Lightweight Convolutional Neural Network Model for Estimating Carbon Sequestration Qualitatively Using Remote Sensing. *Remote Sens.* **2023**, *15*, 272. [[CrossRef](#)]
- Nishizono, T.; Iehara, T.; Kuboyama, H.; Fukuda, M. A forest biomass yield table based on an empirical model. *J. For. Res.* **2005**, *10*, 211–220. [[CrossRef](#)]
- Yu, Q.; Wang, Y.; Van Le, Q.; Yang, H.; Hosseinzadeh-Bandbafha, H.; Yang, Y.; Sonne, C.; Tabatabaei, M.; Lam, S.S.; Peng, W. An overview on the conversion of forest biomass into bioenergy. *Front. Energy Res.* **2021**, *9*, 684234. [[CrossRef](#)]
- Herold, M.; Carter, S.; Avitabile, V.; Espejo, A.B.; Jonckheere, I.; Lucas, R.; McRoberts, R.E.; Næsset, E.; Nightingale, J.; Petersen, R. The role and need for space-based forest biomass-related measurements in environmental management and policy. *Surv. Geophys.* **2019**, *40*, 757–778. [[CrossRef](#)]

8. Tang, X.; Zhao, X.; Bai, Y.; Tang, Z.; Wang, W.; Zhao, Y.; Wan, H.; Xie, Z.; Shi, X.; Wu, B. Carbon pools in China's terrestrial ecosystems: New estimates based on an intensive field survey. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 4021–4026. [[CrossRef](#)]
9. Zhou, G.; Meng, C.; Jiang, P.; Xu, Q. Review of carbon fixation in bamboo forests in China. *Bot. Rev.* **2011**, *77*, 262–270. [[CrossRef](#)]
10. Lambert, M.; Ung, C.; Raulier, F. Canadian national tree aboveground biomass equations. *Can. J. For. Res.* **2005**, *35*, 1996–2018. [[CrossRef](#)]
11. Shen, Y.; Sun, X.; Zhang, J.; Ma, J. Study on the individual tree biomass of *Larix kaempferi* plantation in Xiaolong Mountain, Gansu Province. *For. Res.* **2011**, *24*, 517–522.
12. Fu, L.; Zeng, W.; Tang, S.; Sharma, R.; Li, H. Using linear mixed model and dummy variable model approaches to construct compatible single-tree biomass equations at different scales-A case study for Masson pine in Southern China. *J. For. Sci.* **2012**, *58*, 101–115. [[CrossRef](#)]
13. Haara, A.; Leskinen, P. The assessment of the uncertainty of updated stand-level inventory data. *Silva Fenn.* **2009**, *43*, 87–112. [[CrossRef](#)]
14. Lei, X.; Yu, L.; Hong, L. Climate-sensitive integrated stand growth model (CS-ISGM) of Changbai larch (*Larix olgensis*) plantations. *For. Ecol. Manag.* **2016**, *376*, 265–275. [[CrossRef](#)]
15. Zhang, Y.; Liang, S.; Yang, L. A review of regional and global gridded forest biomass datasets. *Remote Sens.* **2019**, *11*, 2744. [[CrossRef](#)]
16. Wigneron, J.-P.; Li, X.; Frappart, F.; Fan, L.; Al-Yaari, A.; De Lannoy, G.; Liu, X.; Wang, M.; Le Masson, E.; Moisy, C. SMOS-IC data record of soil moisture and L-VOD: Historical development, applications and perspectives. *Remote Sens. Environ.* **2021**, *254*, 112238. [[CrossRef](#)]
17. Tong, X.; Brandt, M.; Yue, Y.; Ciais, P.; Rudbeck Jepsen, M.; Penuelas, J.; Wigneron, J.-P.; Xiao, X.; Song, X.-P.; Horion, S. Forest management in southern China generates short term extensive carbon sequestration. *Nat. Commun.* **2020**, *11*, 129. [[CrossRef](#)]
18. Fayad, I.; Baghdadi, N.; Guitet, S.; Bailly, J.-S.; Hérault, B.; Gond, V.; El Hajj, M.; Minh, D.H.T. Aboveground biomass mapping in French Guiana by combining remote sensing, forest inventories and environmental data. *Int. J. Appl. Earth Obs. Geoinf.* **2016**, *52*, 502–514. [[CrossRef](#)]
19. Cao, Q.; Xu, D.; Ju, H. Biomass estimation of five kinds of mangrove community with the KNN method based on the spectral information and textural features of TM images. *For. Res.* **2011**, *24*, 144–150.
20. El Hajj, M.; Baghdadi, N.; Labrière, N.; Bailly, J.-S.; Villard, L. Mapping of aboveground biomass in Gabon. *Comptes Rendus Géosci.* **2019**, *351*, 321–331. [[CrossRef](#)]
21. Guo, Q.; Su, Y.; Hu, T.; Guan, H.; Jin, S.; Zhang, J.; Zhao, X.; Xu, K.; Wei, D.; Kelly, M. Lidar boosts 3D ecological observations and modelings: A review and perspective. *IEEE Geosci. Remote Sens. Mag.* **2020**, *9*, 232–257. [[CrossRef](#)]
22. Liu, Q.; Li, Z.; Chen, E.; Pang, Y.; Tian, X.; Cao, C. Estimating biomass of individual trees using point cloud data of airborne LIDAR. *High Technol. Lett.* **2010**, *20*, 765–770.
23. Næsset, E.; Gobakken, T. Estimation of above- and below-ground biomass across regions of the boreal forest zone using airborne laser. *Remote Sens. Environ.* **2008**, *112*, 3079–3090. [[CrossRef](#)]
24. Mougin, E.; Proisy, C.; Marty, G.; Fromard, F.; Puig, H.; Betoulle, J.; Rudant, J.-P. Multifrequency and multipolarization radar backscattering from mangrove forests. *IEEE Trans. Geosci. Remote Sens.* **1999**, *37*, 94–102. [[CrossRef](#)]
25. Qiu, S.; Xing, Y.; Xu, W. Estimation of regional forest aboveground biomass combining spaceborne large footprint LiDAR and HJ-1A hyperspectral images. *Acta Ecol. Sin.* **2016**, *36*, 7401–7411.
26. Tolan, J.; Yang, H.-I.; Nosarzewski, B.; Couairon, G.; Vo, H.; Brandt, J.; Spore, J.; Majumdar, S.; Haziza, D.; Vamaraju, J. Sub-meter resolution canopy height maps using self-supervised learning and a vision transformer trained on Aerial and GEDI Lidar. *arXiv* **2023**, arXiv:2304.07213.
27. Wagner, F.H.; Roberts, S.; Ritz, A.L.; Carter, G.; Dalagnol, R.; Favrichon, S.; Hirye, M.; Brandt, M.; Ciais, P.; Saatchi, S. Sub-Meter Tree Height Mapping of California using Aerial Images and LiDAR-Informed U-Net Model. *arXiv* **2023**, arXiv:2306.01936.
28. Yang, S.; Zhang, K.; Shao, Y. Classification of Airborne LiDAR Point Cloud Data Based on Multiscale Adaptive Features. *Acta Opt. Sin.* **2019**, *39*, 0228001. [[CrossRef](#)]
29. Fan, S.; Zhang, A.; Hu, S.; Sun, W. A method of classification for airborne full waveform LiDAR data based on random forest. *Chin. J. Lasers* **2013**, *40*, 0914001.
30. Shu, Z.; Sun, K.; Qiu, K.; Ding, K. Pairwise-Svm for On-Board Urban Road LIDAR Classification. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2016**, *41*, 109. [[CrossRef](#)]
31. Ma, H.; Gao, X.; Gu, X. Random forest classification of Landsat 8 imagery for the complex terrain area based on the combination of spectral, topographic and texture information. *J. Geo-Inf. Sci.* **2019**, *21*, 359–371.
32. Dube, T.; Mutanga, O. Evaluating the utility of the medium-spatial resolution Landsat 8 multispectral sensor in quantifying aboveground biomass in uMgeni catchment, South Africa. *ISPRS J. Photogramm. Remote Sens.* **2015**, *101*, 36–46. [[CrossRef](#)]
33. Kajisa, T.; Murakami, T.; Mizoue, N.; Top, N.; Yoshida, S. Object-based forest biomass estimation using Landsat ETM+ in Kampong Thom Province, Cambodia. *J. For. Res.* **2009**, *14*, 203–211. [[CrossRef](#)]
34. Foody, G.M.; Boyd, D.S.; Cutler, M.E. Predictive relations of tropical forest biomass from Landsat TM data and their transferability between regions. *Remote Sens. Environ.* **2003**, *85*, 463–474. [[CrossRef](#)]
35. Wang, L.-H.; Xing, Y.-Q. Remote sensing estimation of natural forest biomass based on an artificial neural network. *Ying Yong Sheng Tai Xue Bao J. Appl. Ecol.* **2008**, *19*, 261–266.

36. Mountrakis, G.; Im, J.; Ogole, C. Support vector machines in remote sensing: A review. *ISPRS J. Photogramm. Remote Sens.* **2011**, *66*, 247–259. [[CrossRef](#)]
37. Sun, F.; Yuan, J.; Lu, S. The change and test of climate in Northeast China over the last 100 years. *Clim. Environ. Res.* **2006**, *11*, 101–108.
38. Wang, X.-Y.; Zhao, C.-Y.; Jia, Q.-Y. Impacts of climate change on forest ecosystems in Northeast China. *Adv. Clim. Chang. Res.* **2013**, *4*, 230–241.
39. Husch, B.; Beers, T.W.; Kershaw, J.A., Jr. *Forest Mensuration*; John Wiley & Sons: Hoboken, NJ, USA, 2002.
40. Xuan, Z.; Zhang, Q.; Ge, L.; He, H.; Xu, M.; Xu, W. Biomass structure and distribution of Korean Larch Plantations. *For. Resour. Manag.* **2013**, *1*, 53.
41. Mo, D.; Wu, Q.; Lin, N.; Zhuo, Y. Carbon and nitrogen storage and their allocation pattern in *Cryptomeria fortunei* plantations in southeastern Guangxi of South China. *Chin. J. Ecol.* **2012**, *1*, 794–799.
42. Yan, S.; Gao, R.; Chen, G.; Zhang, R. Carbon Stock and Carbon Sequestration of Successive Planting Chinese Fir in Different Rotations. *J.-Northeast. For. Univ. Chin. Ed.* **2006**, *34*, 42.
43. Pang, Y.; Li, Z.; Ju, H.; Lu, H.; Jia, W.; Si, L.; Guo, Y.; Liu, Q.; Li, S.; Liu, L. LiCHy: The CAF's LiDAR, CCD and hyperspectral integrated airborne observation system. *Remote Sens.* **2016**, *8*, 398. [[CrossRef](#)]
44. Zhao, X.; Guo, Q.; Su, Y.; Xue, B. Improved progressive TIN densification filtering algorithm for airborne LiDAR data in forested areas. *ISPRS J. Photogramm. Remote Sens.* **2016**, *117*, 79–91. [[CrossRef](#)]
45. Li, Z.; Liu, Q.; Pang, Y. Review on forest parameters inversion using LiDAR. *J. Remote Sens.* **2016**, *20*, 1138–1150.
46. Dong, L.; Li, F. Additive stand-level biomass models for natural larch forest in the east of Daxing'an mountains. *Sci. Silvae Sin.* **2016**, *52*, 13–20.
47. Fu, L.; Lei, Y.; Sun, W.; Tang, S.; Zeng, W. Development of compatible biomass models for trees from different stand origin. *Acta Ecol. Sin.* **2014**, *34*, 1461–1470.
48. Parresol, B.R. Additivity of nonlinear biomass equations. *Can. J. For. Res.* **2001**, *31*, 865–878. [[CrossRef](#)]
49. Hong, Y.; Chen, D.; Shen, J.; Sun, X.; Zhang, S. Compatible biomass models for *Larix olgensis* plantation based on tree-level and stand-level. *For. Res.* **2019**, *32*, 33–40.
50. Hong, Y.; Zhang, S.; Chen, W.; Chen, D.; Xiang, W.; Pang, Y. Inversion of Biomass Components for *Larix olgensis* Plantation Using Airborne LiDAR. *For. Res.* **2019**, *32*, 83–90.
51. Li, M.; Kang, X.; Fan, W. Burned area extraction in Huzhong forests based on remote sensing and the spatial analysis of the burned severity. *Sci. Silvae Sin.* **2017**, *53*, 163–174.
52. Wang, X.; Chen, E.; Li, Z.; Yao, W.; Zhao, L. Multi-temporal and dual-polarization SAR for forest land type classification. *Sci. Silvae Sin.* **2014**, *50*, 83–91.
53. Xing, Z.; Li, Y.; Deng, R.; Zhu, H.; Fu, B. Extracting farmland shelterbelt automatically based on ZY-3 remote sensing images. *Sci. Silvae Sin.* **2016**, *52*, 11–20.
54. Giles, C.L.; Kuhn, G.M.; Williams, R.J. Dynamic recurrent neural networks: Theory and applications. *IEEE Trans. Neural Netw.* **1994**, *5*, 153–156. [[CrossRef](#)]
55. Schuster, M.; Paliwal, K.K. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* **1997**, *45*, 2673–2681. [[CrossRef](#)]
56. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [[CrossRef](#)]
57. Schumann, O.; Wöhler, C.; Hahn, M.; Dickmann, J. Comparison of random forest and long short-term memory network performances in classification tasks using radar. In Proceedings of the 2017 Sensor Data Fusion: Trends, Solutions, Applications (SDF), Bonn, Germany, 10–12 October 2017; pp. 1–6.
58. Van Houdt, G.; Mosquera, C.; Nápoles, G. A review on the long short-term memory model. *Artif. Intell. Rev.* **2020**, *53*, 5929–5955. [[CrossRef](#)]
59. Kinga, D.A. A method for stochastic optimization. In Proceedings of the International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015.
60. Li, G.; Xie, Z.; Jiang, X.; Lu, D.; Chen, E. Integration of ZiYuan-3 multispectral and stereo data for modeling aboveground biomass of larch plantations in North China. *Remote Sens.* **2019**, *11*, 2328. [[CrossRef](#)]
61. Talucci, A.C.; Forbath, E.; Kropp, H.; Alexander, H.D.; DeMarco, J.; Paulson, A.K.; Zimov, N.S.; Zimov, S.; Loranty, M.M. Evaluating post-fire vegetation recovery in Cajander Larch Forests in Northeastern Siberia using UAV derived vegetation indices. *Remote Sens.* **2020**, *12*, 2970. [[CrossRef](#)]
62. Naik, P.; Dalponte, M.; Bruzzone, L. Prediction of forest aboveground biomass using multitemporal multispectral remote sensing data. *Remote Sens.* **2021**, *13*, 1282. [[CrossRef](#)]
63. Pang, Y.; Li, Z.-Y. Inversion of biomass components of the temperate forest using airborne Lidar technology in Xiaoxing'an Mountains, Northeastern of China. *Chin. J. Plant Ecol.* **2012**, *36*, 1095. [[CrossRef](#)]
64. Yang, J.; Huang, X. The 30 m annual land cover dataset and its dynamics in China from 1990 to 2019. *Earth Syst. Sci. Data* **2021**, *13*, 3907–3925. [[CrossRef](#)]
65. Zhao, Y.; Zeng, Y.; Zheng, Z.; Dong, W.; Zhao, D.; Wu, B.; Zhao, Q. Forest species diversity mapping using airborne LiDAR and hyperspectral data in a subtropical forest in China. *Remote Sens. Environ.* **2018**, *213*, 104–114. [[CrossRef](#)]
66. Badhwar, G. Classification of corn and soybeans using multitemporal thematic mapper data. *Remote Sens. Environ.* **1984**, *16*, 175–181. [[CrossRef](#)]

67. Conese, C.; Maselli, F. Use of multitemporal information to improve classification performance of TM scenes in complex terrain. *ISPRS J. Photogramm. Remote Sens.* **1991**, *46*, 187–197. [[CrossRef](#)]
68. Tian, J.; Xing, Y.; Yao, S.; Zeng, X.; Jiao, Y. Comparison of Landsat-TM image forest type classification based on cellular automata and BP neural network algorithm. *Sci. Silvae Sin.* **2017**, *53*, 26–34.
69. Lo, T.C.; Scarpace, F.L.; Lillesand, T.M. Use of multitemporal spectral profiles in agricultural land-cover classification. *Photogramm. Eng. Remote Sens.* **1986**, *52*, 535–544.
70. Cao, L.; Li, H.; Han, Y.; Yu, F.; Gu, H. Application of convolutional neural networks in classification of high resolution remote sensing imagery. *Sci. Surv. Mapp.* **2016**, *41*, 170–175.
71. Liu, D.; Han, L.; Han, X. High spatial resolution remote sensing image classification based on deep learning. *Acta Opt. Sin.* **2016**, *36*, 306–314.
72. Fu, W.; Zou, W. Review of Remote Sensing Image Classification Based on Deep Learning. *Appl. Res. Comput.* **2018**, *35*, 3521–3525.
73. Zeng, J.; Zhang, X. Laoshan forest biomass estimation based on GF-1 images with inversion algorithm. *J. Cent. South Univ. For. Technol.* **2016**, *36*, 46–51.
74. Liu, F.; Feng, Z.; Zhao, F.; Song, Y. Biomass inversion study of ZY-3 remote sensing satellite imagery. *J. Northwest For. Univ.* **2015**, *30*, 175–181.
75. Huang, X.; Sun, X.; Zhang, S.; Chen, D. Compatible biomass models for *Larix kaempferi* in mountainous area of eastern Liaoning. *For. Res.* **2014**, *27*, 142–148.

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