Forest Ecology and Management 508 (2022) 120031

Contents lists available at ScienceDirect

Forest Ecology and Management

journal homepage: www.elsevier.com/locate/foreco

Deep learning models for improved reliability of tree aboveground biomass prediction in the tropical evergreen broadleaf forests

Bao Huy [a,b], Nguyen Quy Truong [c,1], Nguyen Quy Khiem [a], Krishna P. Poudel [d], Hailemariam Temesgen [e,*]

[a] Forest Resources and Environment Management Consultancy (FREM), No. 06 Nguyen Hong, Buon Ma Thuot, Dak Lak 630000 Viet Nam
[b] Visiting Scholar of Department of Forest Engineering, Resources and Management, Oregon State University (OSU), Corvallis, OR 97333, USA
[c] Technical Consultant, Snapmart Inc., Hochiminh City 700000, Viet Nam
[d] Department of Forestry, Mississippi State University, P.O. Box 9681, Mississippi State, MS 39762, USA
[e] Department of Forest Engineering, Resources and Management, Oregon State University (OSU), Corvallis, OR 97333, USA

ARTICLE INFO

Keywords:
Aboveground biomass
Artificial intelligence
Deep learning
Evergreen broadleaf forest
Forest carbon sequestration

ABSTRACT

Aboveground biomass (*AGB*) and carbon uptake of a forest are key ecological indicators for various technical and scientific applications and sustainable forest management. Deep Learning (DL) methods have been considered as alternative to regression techniques to increase the reliability of tree *AGB* prediction. The objectives were to develop DL models to predict *AGB* in the tropical evergreen broadleaf forests and compare DL models with traditional regression equations for their reliability in *AGB* prediction. A total of 968 individual trees were destructively sampled from fourteen 1-ha and twenty-six 0.2-ha plots distributed across five ecoregions of Viet Nam to get a dataset of tree predictors of diameter at breast height (*DBH*), tree height (*H*), wood density (*WD*) and the response variable of *AGB* along with forest stand factors of basal area (*BA*) and density (*N*); ecological and environmental variables such as *ecoregion, slope, altitude, soil type*, averaged annual temperature (*T*), averaged annual rainfall (*P*) and averaged *dry season length*. The DL models were developed using different combinations of variables selected by factor analysis for mixed data and compared with traditional regression equations by using cross-validation. Trees *AGB* in tropical rainforest predicted by DL models had significantly higher reliability than the conventional regression equations when both had the same input variables. Of the 16 developed DL models with 1 to 9 predictors, the model with 9 predictors (*DBH, H, Ecoregion, Altitude, P, T, Soil type, N* and *WD*) was the best DL model which reduced root mean square percent error (RMSPE) and mean absolute percent error (MAPE) by up to 7.7% and 6.1%, respectively, compared to traditional allometric equations. The DL models created in this study should be applied for measured tree data following factors of the forest stand, ecology, and environment in sampled plots to predict the tree *AGB* and total *AGB*, carbon on a large scale with variation in the value of these factors. Thus, we recommend that the DL models apply for the Measurement, Reporting, and Verification (MRV) system of the Reducing Emissions from Deforestation and forest Degradation (REDD+) program at a large regional level, national or territorial level scale.

1. Introduction

Forests play an important role in mitigating climate change through carbon sequestration. Therefore, carbon stored in the forest tree aboveground biomass (*AGB*) is a key ecological indicator (Bosela et al., 2021) for various technical and scientific applications ranging from regional carbon and bioenergy policies to sustainable forest management (Temesgen et al., 2015; Huy et al., 2016a,b,c, 2019; Zhang et al., 2019, 2020; Nguyen and Kappas, 2020).

To implement the program on Reducing Emissions from Deforestation and forest Degradation (REDD+), Measurement, Reporting, and Verification (MRV) is required for emissions and removals from forests. Estimates of their change over time are also needed with the most transparent and accurate approach possible (Pelletier et al., 2012;

* Corresponding author.
E-mail addresses: baohuy.frem@gmail.com (B. Huy), quytruong.ng@gmail.com (N.Q. Truong), quykhiem.frem@gmail.com (N.Q. Khiem), Krishna.Poudel@msstate.edu (K.P. Poudel), Temesgen.Hailemariam@oregonstate.edu (H. Temesgen).
[1] Joint first author.

https://doi.org/10.1016/j.foreco.2022.120031
Received 25 September 2021; Received in revised form 9 January 2022; Accepted 11 January 2022

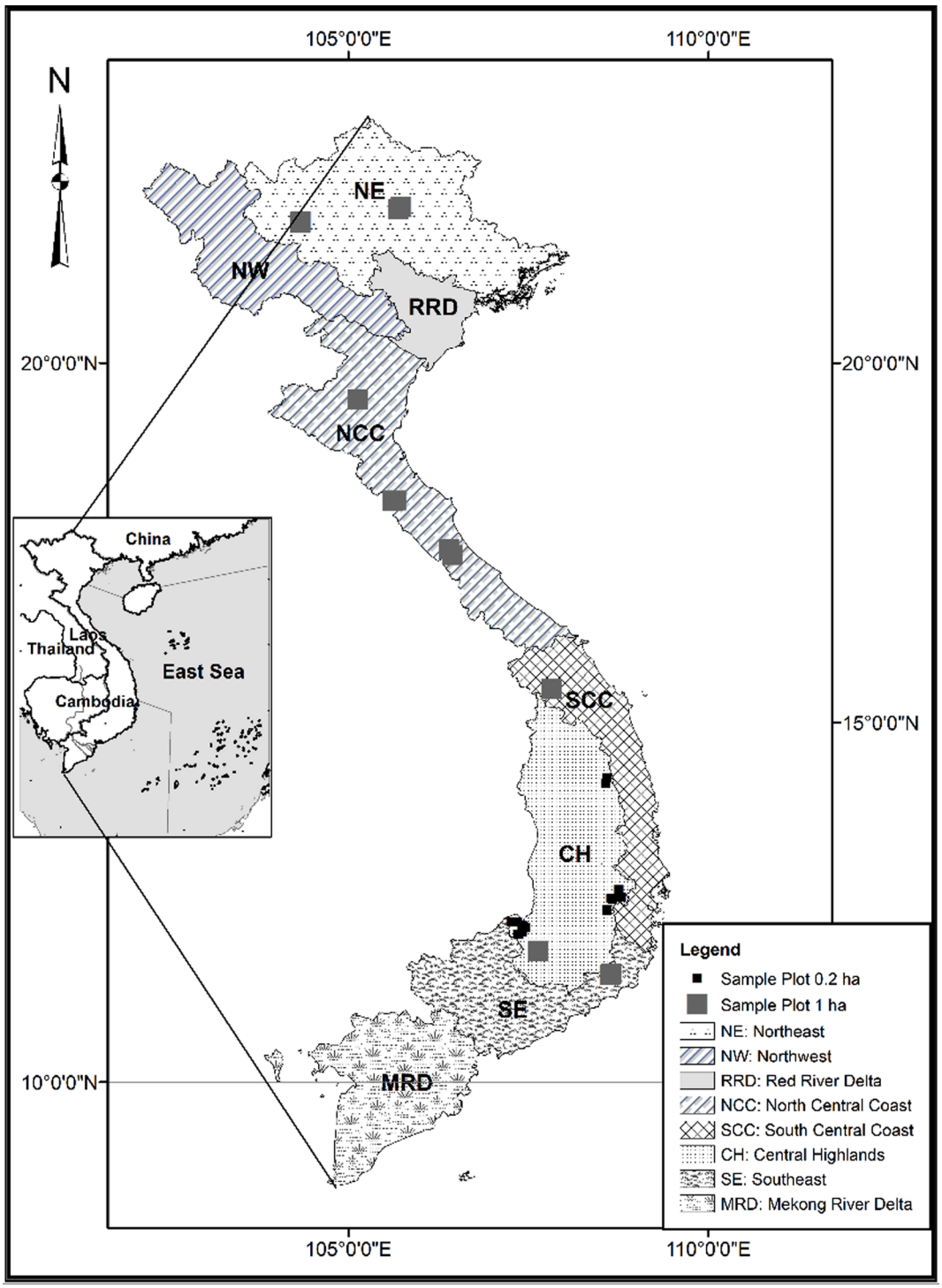


Fig. 1. Studied sample plots in five ecoregions in Viet Nam.

Montano et al., 2017). The participating developing countries must conduct complete MRV to receive results-based financing for RED-D+implementation (U.N., 2015).

Currently, the biomass and carbon accumulated in tropical forest trees are estimated mainly from regression models. These include equations developed either for pan-tropics (Brown, 1997; IPCC, 2003; Chave et al., 2005, 2014), for each sub-tropical ecoregion and forest type (Basuki et al., 2009; Huy et al., 2016b), and woody vegetation taxa such as species, genera, and family-specific biomass modeling systems (Basuki et al., 2009; Huy et al., 2016c, 2019). Mankou et al. (2021) compiled common allometric equations reported across the tropics to predict tree *AGB* from common dendrometric measurements such as diameter at breast height (*DBH*), tree height (*H*), and wood density (*WD*). The power function has often been used to estimate the *AGB* of tropical forest trees using one or more tree predictors. Brown (1997) and IPCC (2003) used a single variable, *DBH*; Basuki et al. (2009) used two variables *DBH* and *WD*; and Chave et al. (2005, 2014), Huy et al. (2016a, b,c, 2019) used three variables *DBH, H*, and *WD*.

The level of prediction error associated with *AGB* regression models has varied across studies, forest types, the generality of the models. For example, Huy et al. (2016a,b) reported that the cross-validation mean absolute percent errors (MAPE) of the *AGB* prediction equations for

evergreen broadleaf forest (EBLF) in Viet Nam fluctuated around from 19.5 up to 36.9%. Increasing the number of predictors or adding random effects of key ecological and environmental factors helped reduce error and improve the estimated *AGB* reliability for the EBLF type. In addition, models that are specific to each ecological region (Huy et al. 2016b), territory or country (Huy et al. 2016a), or forest type, ecological region (Huy et al., 2016c) provided high reliability compared to pan-tropical models.

The *AGB* power models are either fitted in a linear form after log-transformation or as non-linear fixed-effect models. These models have also been fit as weighted non-linear mixed effect models using maximum likelihood and incorporating random effects of environmental and ecological factors (Huy et al. 2016a). Kralicek et al. (2017) and Huy et al. (2019) used seemingly unrelated regression (SUR) for simultaneously estimating either tree *AGB* - belowground biomass *(BGB)* or *AGB* and its components. The weighted non-linear models set up for each ecological region, territory, with three predictors *DBH, H* and *WD*, and considering the random effects of environmental and ecological factors, the MAPE was still high of around 20% (Huy et al., 2016a,b). While applying weighted non–linear SUR models for each forest type and ecological region, based on taxon-specific model (family, genus, and dominant species), the error was decreased, but the MAPE was still high, ranging from 18.5 to 27.1% (Huy et al., 2019). Therefore, the search for methods to improve reliability and reduce the error of the *AGB* prediction of the EBLF is necessary and follows the IPCC requirements (2003, 2006).

Deep Learning (DL) is a subset of machine learning (ML) which is an important branch of Artificial Intelligence (AI) (Ganatra and Patel, 2018). There has been a revolution in ML applications because of the introduction and advancement of DL (Kumar and Garg, 2018; Kriegeskorte and Golan, 2019). ML is widely used in satellite image interpretation combined with ground data to estimate forest stand variables and forest biomass (Dang et al., 2019; Zhang et al., 2019, 2020; Nguyen and Kappas, 2020). DL provides a more adaptive way of using deep neural networks (DNNs) to learn a function from a given input and allow the machine to make decisions. Inputs can be of any kind, structured or unstructured. DL models can produce consistent results without human intervention, making them promising for solving real-time problems (Ganatra and Patel, 2018). Therefore, DL has been successfully applied in many major fields. ML and its DL methods have recently been applied for estimating tree volume and have proven to have better accuracy than regression (Mushar et al., 2020).

The ML and its DL methods are increasingly being applied in forest ecological sciences such as species distribution model, carbon cycle assessment, and climate and environmental change prediction on forest ecosystems (Liu et al. 2018). As it is known, the ecological relationships among the components of the tropical forest ecosystem are complex. Therefore, if only traditional regression models are applied, it is difficult to detect the complex relationships of the ecosystem and biological processes. Future applications of ML and its DL in forest ecology will become increasingly attractive techniques for ecologists (Liu et al. 2018).

Traditional modeling approaches have a great capacity to quantify and predict carbon cycles and can be upscaled from local to regional or global scales. However, the adaptability of these models is typically unsatisfactory, which generally leads to uncertain predictions if spatial and temporal scales change. ML and its DL techniques can be used to address forestry problems where climate and environmental conditions are diverse and complex (Liu et al., 2018). Therefore, DL needs to be considered an alternative to regression techniques (Montano et al., 2017) to improve the prediction of *AGB*. Nevertheless, up to now, there have been very few publications applying ML, specifically DL to biometric science for estimating tree volume, biomass in tropical rain forests. For example, Mushar et al. (2020) used the ML technique to estimate tree volume and found that it produced a better precision and accuracy than the regression method. Ogana and Ercanli (2021) showed that using equations to model the relationships between tree height and diameter in complex rainforest ecosystems remains a challenge, while DL algorithm models that were used to predict tree height as a predictor for estimating tree *AGB*, overperformed other classical regression techniques.

Table 1
Summary statistics of variables used to develop tree aboveground biomass models in this study.

ID	Variables	Min	Mean	Max	Std.
1.	*AGB* (kg tree^{-1})	2.9	553.7	8633.0	917.5
2.	*DBH* (cm)	4.7	25.0	87.7	17.2
3.	*H* (m)	3.9	17.4	41.4	7.2
4.	*WD* (g cm^{-3})	0.166	0.547	0.964	0.139
5.	*BA* (m^2 ha^{-1})	9.0	30.9	49.0	10.1
6.	*N* (tree ha^{-1})	370.0	938.9	3330.0	416.3
7.	*Ecoregion* (code)	1.0	2.5	5.0	1.3
8.	*Slope* (degree)	0.0	18.8	40.0	11.3
9.	*Altitude* (m)	154.0	547.4	1335.0	280.7
10.	Soil type (code)	1.0	2.2	3.0	0.9
11.	*T* (^{0}C averaged)	16.9	22.8	25.0	2.1
12.	*P* (mm year^{-1} averaged)	1055	1962	2500	421.2
13.	*Dry season length* (month averaged)	1.0	3.6	5.0	1.2

Note: Summary statistics based on a dataset of n = 968 destructive sampled trees in 40 sample plots located in 5 ecological regions of EBLFs distribution.

In this study, we hypothesize that the DL approach to predict tree *AGB* of tropical rain forests based on multivariate data consisting of tree-level predictors, forest stand factors, and forest ecological and environmental variables would provide significantly higher reliability than traditional regression equations on a limited number of tree-level predictors. The objectives of this study were to 1) develop DL models to predict *AGB* in the tropical EBLFs and 2) compare DL models with traditional regression equations for their reliability in *AGB* prediction. The contribution of this study is to document DL techniques and their use to improve the reliability of *AGB* predictions in complex tropical forests.

2. Materials and methods

2.1. Study sites

This study used the dataset published by Huy et al. (2016a) and added data on forest stand attributes and ecological and environmental factors. The dataset was collected in five out of eight agro-ecological regions of Viet Nam, containing most of the country's forest cover: northeast (NE), north-central coastal (NCC), central highlands (CH), south-central coastal (SCC), and southeast (SE). These ecoregions span a range of ecological, climatic, and structural site characteristics and are the main sites of the EBLFs (Fig. 1).

2.2. Sampling design and data collection

Fourteen 1-ha (100 × 100 m) and twenty-six 0.2-ha (20 × 100 m) sample plots were established in the five ecoregions CH, NCC, NE, SCC and SE, where the majority of the country's EBLFs are distributed. Within a plot, species and *DBH* (cm) were recorded for all trees larger than 5 cm (1.3 m above ground). Sample trees were selected from each plot and harvested for tree *AGB* (kg tree^{-1}) measurement. Sample tree selection focused on the main dominant species and the number of trees sampled according to the proportion of the diameter distribution. In total, 968 trees ranging from 4.7 to 87.7 cm *DBH* and 3.9 to 41.4 m *H* were destructively sampled. Detailed sampling and measurement procedure was given in Huy et al. (2016a). Table 1 shows summary statistics for the three predictors of *DBH* (cm), *H* (m), *WD* (g cm^{-3}), and the response variable *AGB* of the destructive sample trees.

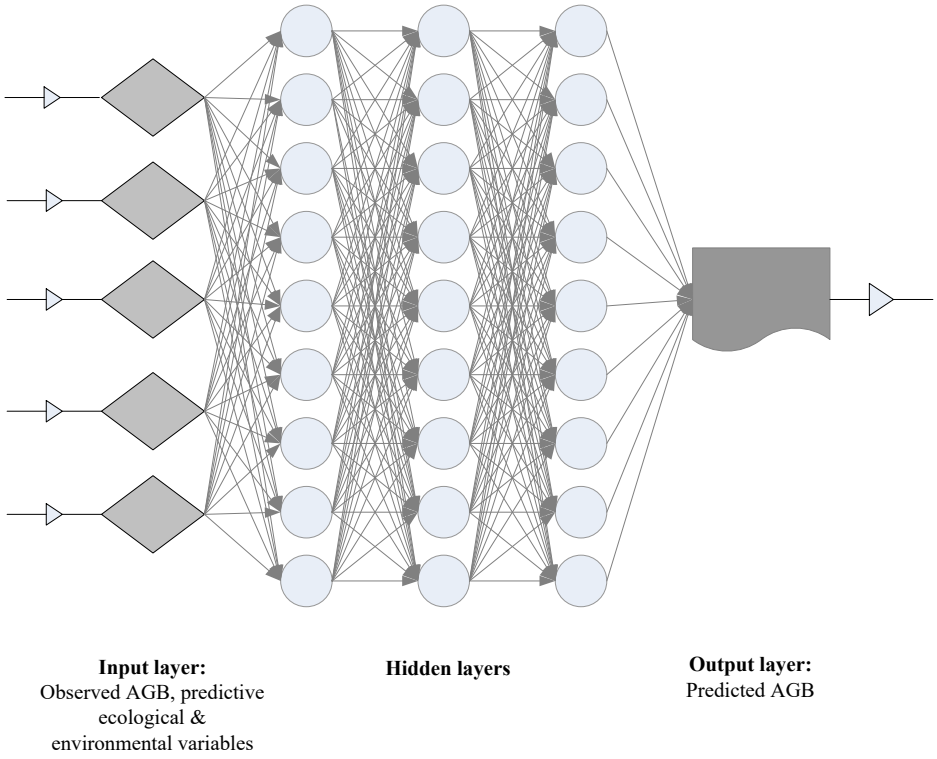


Fig. 2. Deep Neural Network for predicting tree aboveground biomass (*AGB*) as output with tree, stand, ecological, and environmental variables as input predictors of *AGB*.

2.3. Input dataset and variable selection

In this study, the input data layers consisted of observed *AGB* and multiple predictor variables. Individual tree measurements were *DBH, H,* and *WD,* and stand-level variables were basal area (*BA*, m^2 ha^{-1}) and density of trees with *DBH* $\geq$ 5 cm (*N*, tree ha^{-1}). Ecological and environmental variables included ecoregion, slope, altitude, soil type, averaged annual temperature (*T*, °C), averaged annual rainfall (*P*, mm $year^{-1}$), and averaged dry season length (month).

Forest stand variables of *BA* and *N* were calculated from data of sample plots. Ecoregions were encoded 1 to 5 corresponding to NE, NCC, CH, SCC, and SE, respectively. Slope in degree and altitude in meter were also recorded at the sampled plots. Soil types, including crystalline shists, igneous rocks, and sedimentary rocks (Fischer et al., 2008) and were encoded as 1, 2 and 3, respectively. Data on *T, P,* and dry season length were determined from the coordinates of the sample plots based on Hijmans et al., (2005) as well as Fick and Hijmans (2017). Table 1 presents the summary statistics of 13 input variables that were used in the study.

The dataset included both numerical (11 variables as observed *AGB, DBH, H, WD, BA, N, Slope, Altitude, P, T,* and *Dry season length*) and categorical variables (two variables as *Ecoregion* and *Soil type*). To identify the factors that account for the highest variability in the *AGB* among the 12 predictive variables (3 tree-level predictors, 2 stand factors, and 7 ecological and environmental variables), we used factor analysis for mixed data (FAMD) using the FAMD package in R version 4.0.5 (R Core Team, 2021). FAMD was applied to reduce predictive variables and increase the interpretability of multiple factors affecting forest tree *AGB*. FAMD is a principal component method and is appropriate when dealing with both numerical and categorical variables. Used to examine the association between both numerical and categorical variables, FAMD is a hybrid approach between principal component analysis (PCA) and multiple correspondence analysis (MCA) (STHDA, 2021; Pages, 2004; R Core Team, 2021). The variables were normalized during the analysis, the numerical variables were scaled to unit variance, and the categorical variables were transformed into a crisp coding and then scaled using MCA. This ensures to balance of the influence of both numerical and categorical variables in the analysis (R Core Team, 2021).

2.4. Deep learning for predicting tree AGB

Artificial Neural Network (ANN) is a DL technique that has recently attracted great attention. ANN has been used to model the relationship between independent input variables and dependent output variables simulating the learning of the biological neural system. Various types of ANN can be applied in different problem domains. The feed-forward structure of ANN has been widely used in many applications (Kumar and Garg, 2018). A DNN is a variant of multilayer feed-forward ANN. Almost all current DL applications are built upon DNNs (Zhou and Feng, 2019). It has more than one hidden layer between the input and output layers (Kumar and Garg, 2018; Chollet, 2018; Kriegeskorte and Golan, 2019).

DL is a mathematical framework for learning representations from data (Chollet, 2018), it is a computational process involving multiple layers to learn how to represent data with multiple levels of abstraction. DL explores complex structures in large data sets using back-propagation algorithms to change the internal parameters used to compute the representation in each layer compared to the representation in the previous layer (LeCun et al., 2015). DL models are multi-layered DNNs with hidden layers and hundreds to thousands of neurons. The DNNs represent a more complex structure similar to the human brain than those of ANNs (Ogana and Ercanli, 2021). The DL algorithm passes input data across several layers; each layer can extract features progressively and pass them on to the next layer. This process enables very complex relations functions among input variables that can be learned to form a complete output layer representation (Mathew et al., 2021; LeCun et al., 2015). The specification of what a layer does with its input data is stored in the weights of the layer. The transformation is performed by a layer parameterized by its weights. In this process, learning means finding a

set of values for the weights of all layers in a network. To control the output of a DNN, a loss function of the network is used. The loss function takes the predictions of the network and compares them with the real outputs and calculates the differences to obtain how well the network has performed (Chollet, 2018).

The architecture of the DNN used in this study is presented in Fig. 2. The DL models were trained by input data layers (Ogana and Ercanli, 2021) consisting of observed *AGB* and different combinations of the tree-level variables of tree predictors, forest stand, and ecological, environmental factors identified by the FAMD method, and the output data layer was *AGB* predicted.

DL algorithms work with numerical and categorical variables. However, the categorical variables need to be encoded for DNN (Dahouda and Joe 2021). In this study, the dataset for the DL process included both numerical variables (Observed *AGB, DBH, H, Altitude, P, T, N,* and *WD*) and categorical variables (*Ecoregion* and *Soil type*). We used One- Hot Encoding technique to apply these two categorical variables as inputs to deep learning algorithms (Dahouda and Joe 2021; Hayashi, 2020; Potdar et al., 2017; Hancock and Khoshgoftaar, 2020). All numerical input variables were scaled (Guo and Berkhahn 2016) by dividing by their maximum values and were within the range [0, 1]. This ensures to balance of the influence of all input variables in the DL process.

The learning process of the DNN is described as follows (Kumar and Garg, 2018):

$$y_{product1} = \sum_{i=1}^{n} I_i \times w_i + \varepsilon_1 \tag{1}$$

$$Y_{product} = \left[y_{product1}, y_{product2,} \cdots\cdots, y_{productm}\right] \tag{2}$$

$$y_{output1} = f\left(\sum_{i=1}^{n} I_i \times w_i + \varepsilon_1\right) \tag{3}$$

$$Y_{output} = \left[y_{output1}, y_{output2,} \cdots\cdots, y_{outputm}\right] \tag{4}$$

where $y_{product\ m}$ is defined as the cross product of the input vector $I = [I_1, I_2, \ldots I_i .., I_n]$, n is the number of input variables, and w_i is the weight on interconnection along with I_i with and ε is the bias value, m is number of neurons of the network, $Y_{product}$ is the product vector, f is the activation function used at the neuron, and $y_{output\ m}$ is the output of the neuron and Y_{output} is the output vector.

Keras library – deep learning Application Programming Interfaces (API) was used in free open-source Python (2021) for developing and evaluating deep learning models (Kumar and Garg, 2018; Chollet, 2018), and TensorFlow backend was applied for most deep learning needs (Chollet, 2018; Ganatra and Patel, 2018).

In this study, DL was performed with three hidden layers (layers of nodes between the input and output layers), including 512 neurons in each hidden layer. The number of epochs is a hyperparameter on which the learning algorithm works through the entire training dataset and is set to 5000 times. The batch size is a hyperparameter, which is the number of samples to work through before updating the internal model parameters was defined to be 64 samples. Thus, for 775 training samples (80% random dataset), an epoch comprised 775/64 = 12 batches. The DL used a ReLU activation function for the hidden layers and a linear activation function for the output and applied Adam's optimization algorithm (Kingma and Ba, 2015; Jais et al., 2019; Zaheer and Shaziya, 2019) to select the best fit between predicted and actual outputs. The DL process used 193 validation samples (20% random dataset) and loss function as the MAPE to select the best DL model. In addition, an early stop function with patience set at 1000 was used to stop the training when validation loss did not decrease further before the model was overfitted.

2.5. A non-linear fixed and mixed model with separate variable random effect

Huy et al. (2016a) used *DBH, WD*, and *H* as the covariates of the power model for estimating *AGB*. The Furnival index (Furnival, 1961) was used to compare the performance of log-linear and weighted non-linear models. As a result of that comparison, non-linear models were selected. Weighted non-linear models allow flexibility in model forms and can account for the heterogeneity of errors (Davidian and Giltinan, 1995; Picard et al., 2012; Huy et al., 2016b, 2019).

The models of Huy et al. (2016a) were fit based on the Maximum Likelihood procedure in R statistical software using the nlme package (Bates, 2010; Picard et al., 2012; Pinheiro et al., 2014), and model diagnostics were conducted using the ggplot2 package (Wickham and Chang, 2013). The general form of the *AGB* model was (Huy et al., 2016a):

$$Y_{ij} = \left(a + \alpha_j\right) \times X_{ij}^{\left(b+\beta_j\right)} + \varepsilon_{ij} \tag{5}$$

$$\varepsilon_{ij} \sim \ iid\mathcal{N}\left(0, \sigma^2\right) \tag{6}$$

where Y_{ij} was the *AGB* (kg tree^{-1}) for the i^{th} tree from the j^{th} class of a variable/factor; and a and b were the fixed effect parameters of the model; α_j and β_j were parameters associated with the j^{th} class of a variable; X_{ij} was the covariate *DBH* (cm), *H* (m), *WD* (g/cm^3), DBH^2H (m^3), or DBH^2HWD (kg) for the i^{th} tree in the j^{th} class of a variable; and ε_{ij} was the random error associated with the i^{th} tree from the j^{th} class of a variable. For example, the independent combined variables $DBH^2H = (DBH/100)^2 \times H$ and $DBH^2HWD = DBH^2H \times WD \times 1000$ were approximations of volume and *AGB*, respectively.

The variance function was as follows (Huy et al., 2016a):

$$Var(\varepsilon_{ij}) = \widehat{\sigma^2}\left(\nu_{ij}\right)^{2\delta} \tag{7}$$

where $\widehat{\sigma^2}$ was the estimated error sum of squares;ν_{ij} was the weighting variable (*DBH, DBH*^{2}*H,* or DBH^2HWD) associated with the i^{th} tree from the j^{th} class of the random effect; and δ was the variance function coefficient to be estimated.

Huy et al. (2016a) mainly used random effects of ecoregion on model parameters to test and evaluate their influence in the allometric relationship. Ecoregion at five levels (NE, NCC, CH, SCC, SE) represented the influence of ecological and climatic factors on *AGB*.

2.6. Non-linear fixed model with a combination of factors

Mixed-effects *AGB* model with random effects mentioned above set up a single model for each environmental and climatic factor (Huy et al., 2016a). Meanwhile, these factors interact and have synergistic effects on *AGB*. Therefore, the fixed-effects model with a combination of ecological and environmental variables was examined and compared with DL model performance in *AGB* prediction.

In this study, the form of the *AGB* model consisted of two components, an average *AGB* model and a modifier (Lessard et al. 2001; Huy et al., 2020) as follows:

$$\mathrm{AGB} = \mathrm{AVERAGE} \times \mathrm{MODIFIER} \tag{8}$$

where AVERAGE was the best equation selected by Huy et al. (2016a):

$$\mathrm{AVERAGE} = \mathrm{a} \times \left(\mathrm{DBH}^2 \times \mathrm{H} \times \mathrm{WD}\right)^{\mathrm{b}} \tag{9}$$

$$MODIFIER = \prod_{j=1}^{n} exp\left(factor_j - average \quad value \quad of \quad factor_j\right) \tag{10}$$

The modifier is an exponential function involving forest stand, ecological, and environmental factors as additional covariates. The modifier adjusts *AGB* based on the combined effects of these factors. In

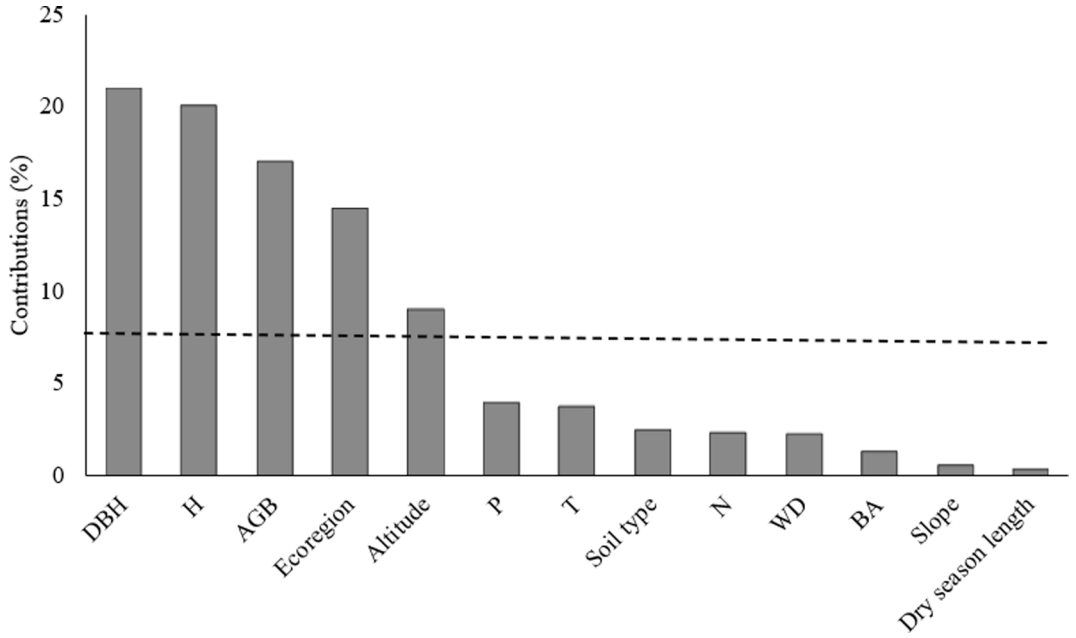


Fig. 3. Factor analysis for mixed data (FAMD): Contribution of mixed numerical and categorical variables to principal dimension 2. The dashed line indicates the expected average value.

this study, n factors consisted of variables that significantly affect *AGB* and were selected by the FAMD method. Average values of the variables presented in Table 1 were incorporated into the modifier. The model in (8) was fitted as weighted non-linear fixed-effects models with weighting variable *DBH* fit by the maximum likelihood (Bates, 2010; Pinheiro et al., 2014) using nlme package in R version 4.0.5 (R Core Team, 2021).

2.7. Cross-validation

The dataset was randomly split into two parts, with 80% for training and 20% for validation. The cross-validation process was repeated 10 times and the model performance was averaged over 10 realizations. The goodness-of-fit statistic used to validate, compare, and select models were R^2 or Fit Index (FI) (Parresol, 1999). The closer the FI is to 1, the better the model.

$$FI = \frac{1}{R}\sum_{r=1}^{R}\left(1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}\right) \tag{11}$$

Along with FI, bias (%), root mean square error (RMSE, kg), root mean square percent error (RMSPE, %), and mean absolute percent error (MAPE, %) were calculated. Smaller values for indicators are preferred.

$$Bias(\%) = \frac{1}{R}\sum_{r=1}^{R}\frac{100}{n}\sum_{i=1}^{n}\frac{y_i - \widehat{y}_i}{y_i} \tag{12}$$

$$RMSE(kg) = \frac{1}{R}\sum_{r=1}^{R}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2} \tag{13}$$

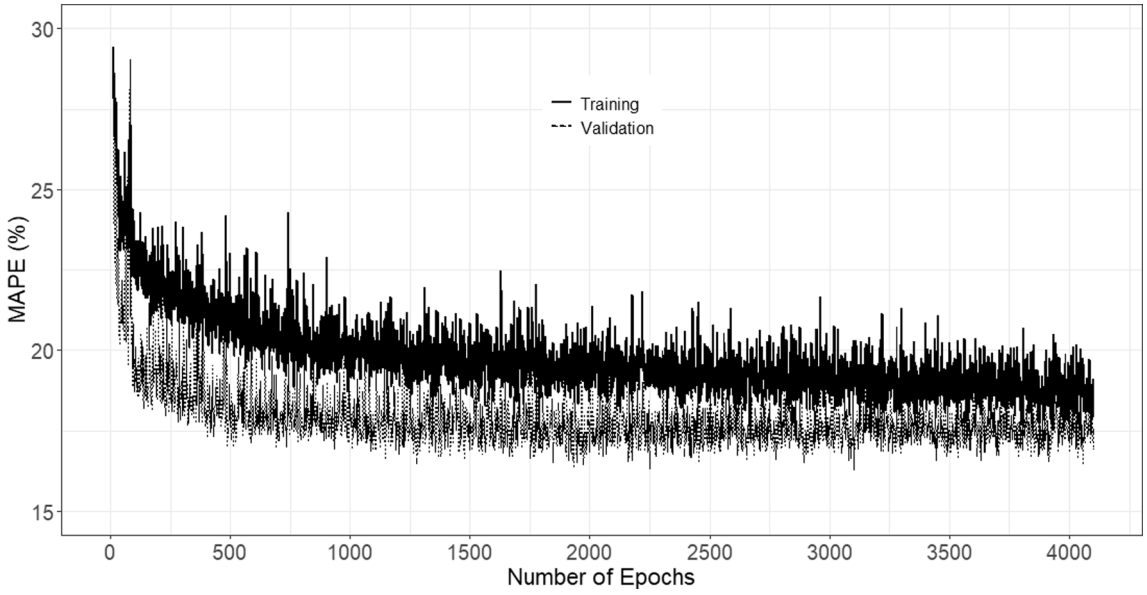


Fig. 4. Mean absolute percent error (MAPE, %) of training and validation vs. the number of epochs for deep learning model with all selected 10 input variables: Observed *AGB, DBH, H, Ecoregion, Altitude, P, T, Soil type, N,* and *WD*.

Table 2
Deep learning models – Cross-validation statistics in predicting *AGB* from the best and mean results based on different combinations of input variables.

ID	Input variables	FI	RMSE (kg)	Bias (%)	RMSPE (%)	MAPE (%)
1.	Input 10 variables: *Observed AGB, DBH, H, Ecoregion, Altitude, P, T, Soil type, N and WD*					
	Best result	0.957	170.2	4.6	19.1	15.0
	Mean result	0.934	229.6	0.7	24.0	17.4
2.	Input 9 variables: *Observed AGB, DBH, H, Ecoregion, Altitude, P, T, Soil type, and N*					
	Best result	0.941	253.8	2.6	27.8	21.3
	Mean result	0.910	271.0	3.1	30.3	22.9
3.	Input 8 variables: *Observed AGB, DBH, H, WD, Ecoregion, Altitude, P, and Soil type*					
	Best result	0.956	142.1	−0.6	23.7	15.9
	Mean result	0.918	254.7	1.1	25.8	18.2
4.	Input 8 variables: *Observed AGB, DBH, H, Ecoregion, Altitude, P, T and Soil type*					
	Best result	0.931	253.2	3.1	29.9	22.0
	Mean result	0.896	282.2	3.9	30.1	23.1
5.	Input 7 variables: *Observed AGB, DBH, H, Ecoregion, Altitude, P and Soil type*					
	Best result	0.899	247.8	2.7	29.2	21.6
	Mean result	0.905	255.3	3.7	30.3	22.9
6.	Input 7 variables: *Observed AGB, DBH, H, Ecoregion, Altitude, P and T*					
	Best result	0.920	223.3	2.7	29.1	21.7
	Mean result	0.895	278.7	3.9	29.9	23.2
7.	Input 7 variables: *Observed AGB, DBH, H, Ecoregion, Altitude, P and WD*					
	Best result	0.939	225.8	3.0	20.1	15.4
	Mean result	0.928	234.7	2.4	23.7	17.4
8.	Input 6 variables: *Observed AGB, DBH, H, Ecoregion, T and WD*					
	Best result	0.863	366.2	3.4	21.1	16.6
	Mean result	0.903	288.1	2.6	24.6	17.7
9.	Input 5 variables: *Observed AGB, DBH, H, Ecoregion and Altitude*					
	Best result	0.895	278.5	6.0	26.5	22.4
	Mean result	0.865	322.2	4.4	30.8	23.7
10.	Input 5 variables: *Observed AGB, DBH, H, WD and Ecoregion.*					
	Best result	0.918	214.4	1.2	20.8	15.7
	Mean result	0.918	252.4	1.7	24.6	17.6
11.	Input 4 variables: *Observed AGB, DBH, H, WD.*					
	Best result	0.910	262.9	4.6	22.2	17.0
	Mean result	0.914	252.7	3.2	24.3	18.1
12.	Input 4 variables: *Observed AGB, DBH, H, and Ecoregion.*					
	Best result	0.863	309.2	3.9	28.3	22.4
	Mean result	0.885	304.8	4.8	31.0	24.3
13.	Input 4 variables: *Observed AGB, DBH, WD, Ecoregion.*					
	Best result	0.941	244.6	2.8	24.2	17.7
	Mean result	0.908	256.9	2.2	27.4	19.8
14.	Input 3 variables: *Observed AGB, D, WD.*					
	Best result	0.915	213.0	3.5	24.2	18.8
	Mean result	0.890	314.4	2.8	28.1	20.7
15.	Input 3 variables: *Observed AGB, DBH, H.*					
	Best result	0.885	338.3	5.0	27.9	21.8
	Mean result	0.876	318.4	4.6	30.8	24.3
16.	Input 2 variables: *Observed AGB, DBH.*					
	Best result	0.884	277.0	2.4	30.9	24.5
	Mean result	0.840	369.7	6.0	33.7	26.6

Note: Cross-validation with 10 realizations, each repeating the dataset of 968 samples was split randomly into 80% for training and 20% for validation; statistics, errors in mean result were averaged over 10 times, and the best result was selected out of 10 validation results.

$$RMSPE(\%) = \frac{1}{R}\sum_{r=1}^{R} 100\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \widehat{y}_i}{y_i}\right)^2} \quad (14)$$

$$MAPE(\%) = \frac{1}{R}\sum_{r=1}^{R} \frac{100}{n}\sum_{i=1}^{n}\frac{|y_i - \widehat{y}_i|}{y_i} \quad (15)$$

where R was the number of realization (10); n was the number of sampled trees of validation dataset; and y_i, $\widehat{y}_i$ and $\overline{y}$ were observed, predicted, and averaged *AGB* (kg tree^{-1}) for the i^{th} sampled tree in realization R, respectively.

The diagnostic plots of the trend of fitted vs. observed *AGB* data and residuals vs. fitted *AGB* were also used to assess the model performance. Fitting models with the entire dataset obtained final parameter estimates for all the selected fixed effect modeling systems. Meanwhile, the DL model with the smallest error was saved in Python code. Then, the best DL model was used to predict tree *AGB* with a selected multivariate variable with code written in Python language.

3. Results

3.1. Multiple variables affect tree AGB

Based on the results of FAMD, five principal dimensions explained 76.7% of the variability in the original data. The second principal dimension was selected because the contribution of the *AGB* variable was the highest. Fig. 3 shows the contribution of variables to principal dimension 2, in which there were five variables namely *DBH, H, AGB, Ecoregion,* and *Altitude* that contribute the most. Meanwhile, three variables *BA, Slope,* and *Dry season length* showed the lowest effects, so these three factors were excluded, and the remaining 10 factors of observed *AGB, DBH, H, Ecoregion, Altitude, P, T, Soil type, N,* and *WD* were considered in *AGB* modeling and DL process.

3.2. Deep learning models for predicting tree AGB

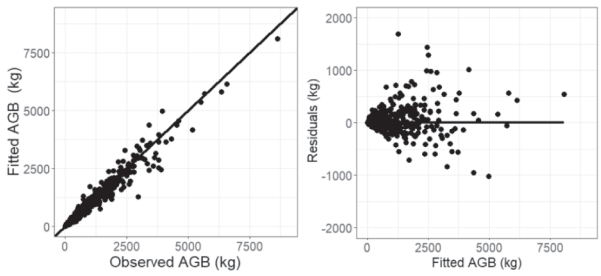
The MAPE for the DL model with 10 input variables selected through FAMD was nearly identical and saturated (Fig. 4) for training and validation datasets when the epoch numbers reached over 4000 times. We developed 16 DL models for *AGB* prediction based on different combinations of 2 to 10 input variables through FAMD results. The DL models were first established with two input variables: observed *AGB* and *DBH* and gradually increased to include all 10 input variables (Table 2).

The cross-validation results for the best and average model of 10 realizations are presented in Table 2. The errors such as RMSPE and MAPE of the DL model with 10 input variables selected by FAMD, including observed *AGB, DBH, H, Ecoregion, Altitude, P, T, Soil type, N,* and *WD* were almost the lowest. When forest stand, ecological and environmental variables were gradually eliminated from the models, the errors are increased. The plots of fitted vs. observed *AGB* and residuals vs. fitted *AGB* are shown in Fig. 5. The DL model that had 10 optimal input variables mentioned provided the best prediction of the tree *AGB* with FI = 0.957, Bias = 4.6%, RMSPE = 19.1% and MAPE = 15.0%.

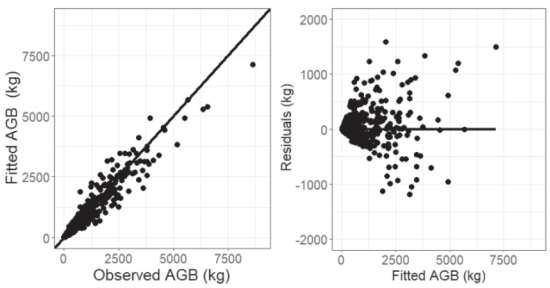
3.3. Comparison of deep learning and regression methods

Results from comparing bias, RMSPE, and MAPE in predicting *AGB* based on one to three tree predictors such as *DBH, H,* and *WD* and mixed-effects models with the random effects of the ecoregion Huy et al. (2016a) with the DL models are presented in Tables 3 and 4, respectively. Variables *T* and *N* were insignificant parameters with P values greater than 0.05 when the regression model was set up with the combination of 9 independent variables including *DBH, H, WD, Ecoregion, Altitude, P, T, Soil type,* and *N* affecting *AGB* as determined by FAMD method (Table 5). Therefore, the equation was developed with eight optimal input variables, including observed *AGB, DBH, H, WD, Ecoregion, Altitude, P,* and *Soil type* (Table 5).

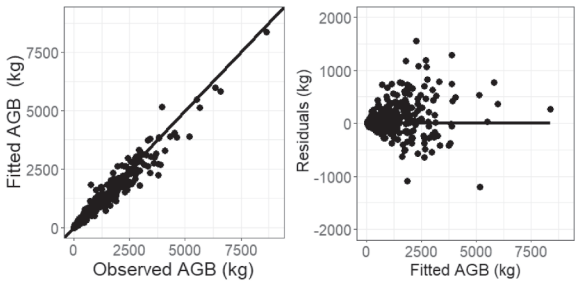
To compare the reliability of using the DL model and allometric equation (Table 5 and Fig. 6), we used the FI index and errors from cross-validation. From there, to predict tree AGB through regression equation, use the model with optimal predictors, presented in Table 5 and Fig. 6. Table 6 shows parameter values of regression with optimal predictors and its equation, which are presented as follow:



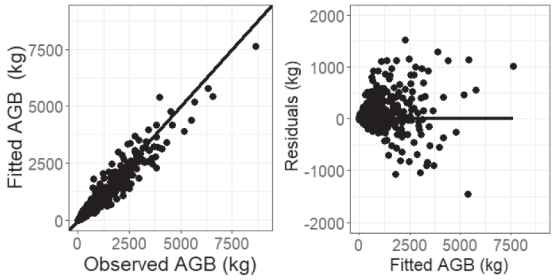
Input 10 variables: *Observed AGB, DBH, H, Ecoregion, Altitude, P, T, Soil type, N and WD.*



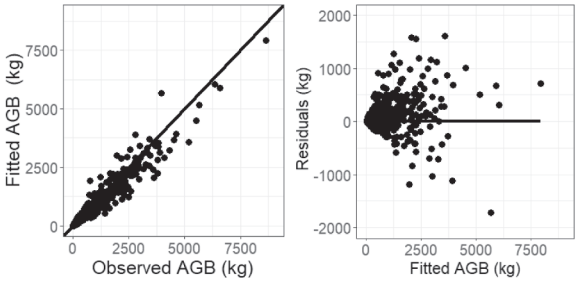
Input 9 variables: *Observed AGB, DBH, H, Ecoregion, Altitude, P, T, Soil type, and N*



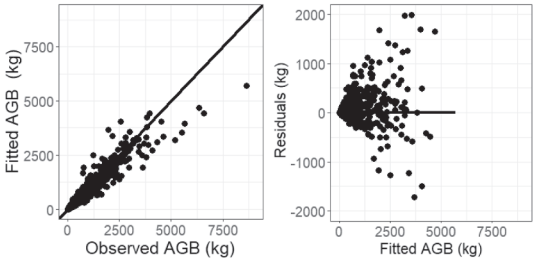
Input 8 variables: *Observed AGB, DBH, H, WD, Ecoregion, Altitude, P, and Soil type*



Input 8 variables: *Observed AGB, DBH, H, Ecoregion, Altitude, P, T and Soil type*



Input 7 variables: *Observed AGB, DBH, H, Ecoregion, Altitude, P and Soil type*



Input 7 variables: *Observed AGB, DBH, H, Ecoregion, Altitude, P and T*

Fig. 5. Plots of deep learning models for a different combination of input variables: Fitted vs. Observed *AGB* (left); Residuals vs. Fitted *AGB* (right) based on the entire dataset.

$$AGB = 0.079835 \times \left(DBH^2 \times H \times WD\right)^{0.938173} \times exp(0.015205 \times (Ecoregion - 2.5) + 0.000170 \times (Altitude - 547) + 0.000182 \times (P - 1962) + 0.016620 \times (Soil\ \ type - 2.2)) \quad (16)$$

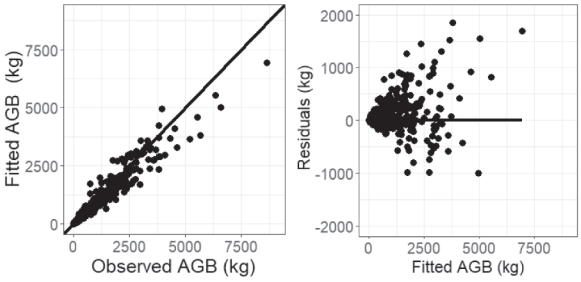
4. Discussion

4.1. Deep learning vs. regression for predicting AGB of tropical forest trees

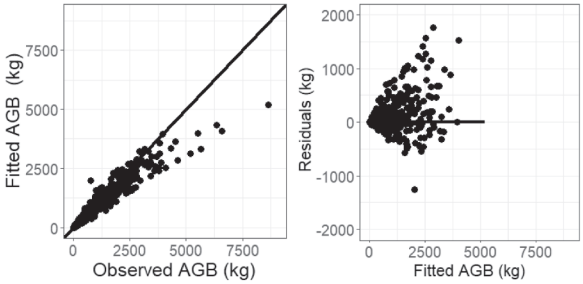
DL has been scarcely applied in biometric science to predict forest tree biomass and carbon. This study showed that DL models significantly improved the predictive reliability of tree *AGB* in tropical EBLFs compared to traditional allometric equations (Tables 3, 4, 5).

The DL models provided significantly better reliability than allometric equations for *AGB* predictions based on one to three tree predictors (*DBH*, *H*, and *WD*). DL models reduced the MAPE between 2.6% and 6.1% compared with allometric equations (Table 3). While the regression equations and DL models had the same predictors, including the random effect of the ecoregion, the DL methods reduced the MAPE between 3.3% and 5.0% (Table 4). In other words, DL models reduced MAPE by up to 6.1% compared with the regression equations.

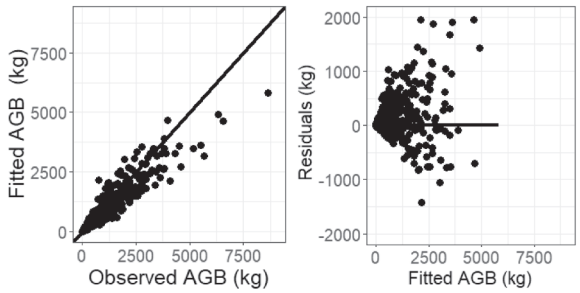
FI index for the DL model with ten optimal input variables was approximately equal and greater than 0.95. The RMSPE and MAPE were up to 7.7 % and 4.2 % less than that produced by the regression models with eight optimal input variables, respectively (Table 5). In addition, Bland- Altman plot (Bland and Altman, 1999) demonstrated many values of differences between *AGB* predicted by DL and regression were outside an interval within which 95% of differences between predictions by the two methods are expected to lie, so this indicated the differences significantly between predicted *AGB* by DL model and regression equation with the optimal input variables (Fig. 7). This result confirms



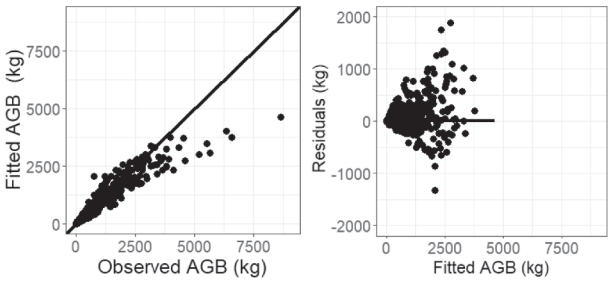
Input 7 variables: *Observed AGB, DBH, H, Ecoregion, Altitude, P and WD*



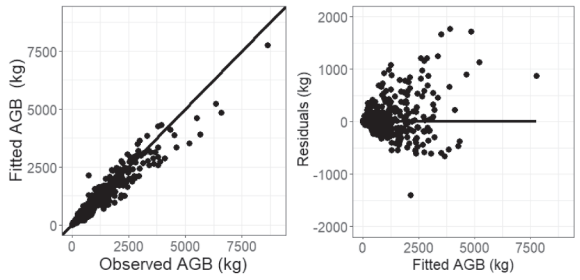
Input 6 variables: *Observed AGB, DBH, H, Ecoregion, T and WD*



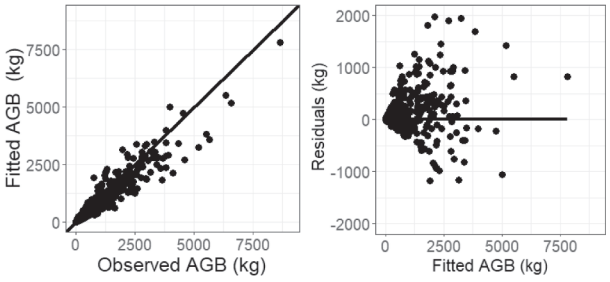
Input 5 variables: *Observed AGB, DBH, H, Ecoregion and Altitude*



Input 5 variables: *Observed AGB, DBH, H, WD and Ecoregion*



Input 4 variables: *Observed AGB, DBH, H, WD*



Input 4 variables: *Observed AGB, DBH, H, and Ecoregion*

Fig. 5. (*continued*).

the optimality when using the DL model under AI to predict *AGB* compared to the traditional regression method.

The result of this study is consistent with the comments of Montano et al. (2017) that the AI models have superior capabilities of estimating and predicting the biomass of tropical forest trees compared to the regression-based allometric equations. In addition, heterogeneity errors and deviations from normality are common issues in regression for biomass estimation. The AI models are an effective alternative to the regression technique, especially tropical forest tree biomass. In this study, the DL method increased the predictive reliability of the tree *AGB* of the tropical forests by allowing many kinds of input complex numerical and categorical variables, including tree predictors, forest stand factors, ecological and environmental variables.

4.2. Variables that affect the prediction of AGB in tropical EBLF

Most pantropical tree-level *AGB* equations are based on tree predictors such as *DBH, H*, and *WD* (e.g., Brown, 1997; IPCC, 2003; Chave et al., 2005, 2014; Basuki et al., 2009). Huy et al. (2016a) had the addition of the random effect of the ecoregion. Huy et al. (2016c, 2019) and Basuki et al. (2009) fitted the taxon-specific levels *AGB* equations.

In tropical forests, many environmental and ecological factors affect allometric relationships (Cysneiros et al., 2021) of biomass accumulation that have not been fully considered in the regression functions. This study indicated 9 independent variables/factors as *DBH, H, Ecoregion, Altitude, P, T, Soil type, N,* and *WD* influencing *AGB*. This result is consistent with Kassa's (2015) findings that the carbon pools in aboveground exhibited distinct patterns along environmental gradients (altitude, slope gradient, and aspect); or included climatic factors represented by *P* and *T* variables, and *Ecoregion, Soil type* variables that allow DL model to address differences of unique regions; this is consistent with Goslee et al. (2015).

The taxon-specific *AGB* equations, such as models for dominant family, genus significantly improved the reliability of the *AGB* estimates (Basuki et al., 2009; Huy et al., 2016c, 2019; Mankou et al., 2021). However, in this study, the family and genus factors were eliminated. This is explained by the fact that the variable *WD* can represent these two taxon factors in the DL process.

Trees in a higher mean altitude of 547 m had greater carbon uptake capacity of the tree *AGB* in the tropical EBLFs (Eq. (16)). Under the trend of tropical climate change (U.N., 2015), the *P* and *T* factors are changing and affect carbon sequestration in the tropical rain forests as demonstrated by the Eq. (16) and optimal DL model.

4.3. Application of deep learning models to predict tree AGB of the tropical EBLFs

This study developed and evaluated 16 best DL models to predict tree

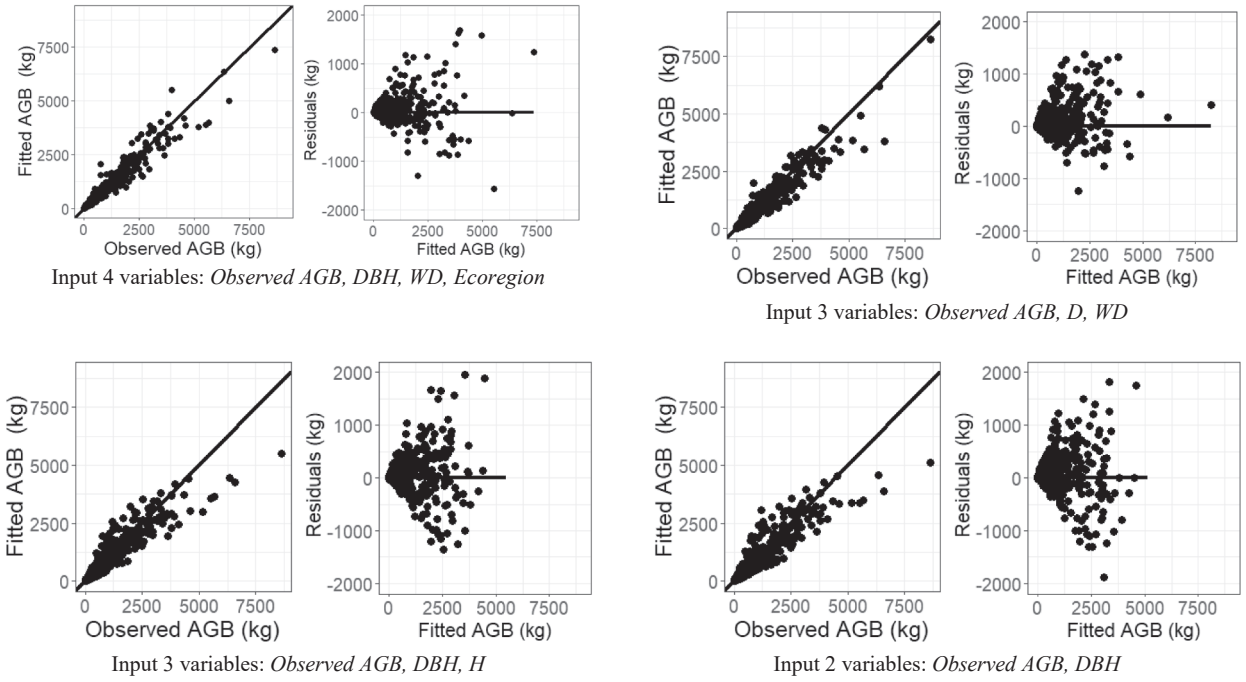


Fig. 5. (*continued*).

Table 3
Comparisons of deep learning vs. weighted non-linear fixed effect model with combinations of tree predictors. Results were based on cross validation of models in predicting *AGB*.

ID	Predictor/s	Method	Model	Bias (%)	RMSPE (%)	MAPE (%)	Source
1	*DBH*	Regression	$AGB = 0.128 \times DBH^{2.409}$	−12.2	42.1	30.6	Huy et al., 2016a
		Deep learning	*The best model*	2.4	30.9	24.5	This study, 2021
2	*DBH, H*	Regression	$AGB = 263.998 \times (DBH^2H)^{0.936}$	−8.6	36.6	27.4	Huy et al., 2016a
		Deep learning	*The best model*	5.0	27.9	21.8	This study, 2021
3	*DBH, WD*	Regression	$AGB = 0.248 \times DBH^{2.386} \times WD$	−4.5	30.0	21.4	Huy et al., 2016a
		Deep learning	*The best model*	3.5	24.2	18.8	This study, 2021
4	*DBH, H, WD*	Regression	$AGB = 0.806 \times (DBH^2HWD)^{0.920}$	−2.1	26.7	19.6	Huy et al., 2016a
		Deep learning	*The best model*	4.6	22.2	17.0	This study, 2021

Note: Cross validation for deep learning (DL) with 10 realizations, each repeating the dataset of 968 samples was split randomly into 80% for training and 20% for validation; the best DL model was selected out of 10 validation results. DBH^2H (m^3)= (DBH (cm)/100)^2 × H (m); DBH^2HWD (kg) = DBH^2H × WD (g cm^{-3}) × 1000.

Table 4
Comparisons of deep learning vs. weighted non-linear mixed effect model with random effects of the separate ecological variable. Results were based on cross validation of models in predicting *AGB*.

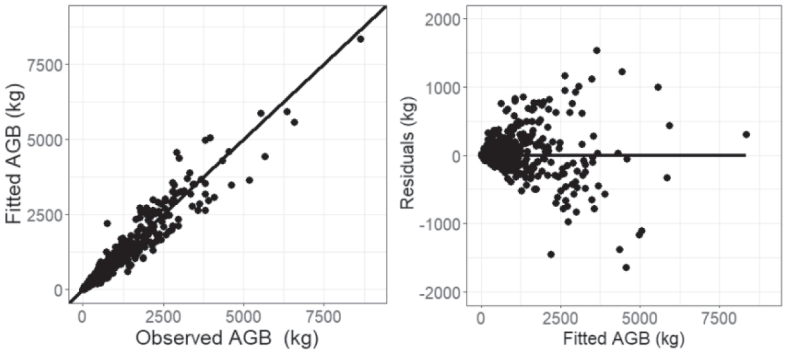
ID	Input variables	Method	Model	Bias (%)	RMSPE (%)	MAPE (%)	Source
1	*Observed AGB, DBH, H, Ecoregion*	Regression	$AGB = a_i \times (DBH^2H)^{0.951}$ with a_i *for ecoregions: CH: 304.167, NCC: 253.245, NE: 256.713, SCC: 272.080, SE: 236.586*	−10.4	37.6	27.4	Huy et al., 2016a
		Deep learning	*The best model*	3.9	28.3	22.4	This study, 2021
2	*Observed AGB, DBH, WD, Ecoregion*	Regression	$AGB = 0.229 \times DBH^{bi} \times WD$ with b_i *for ecoregions: CH: 2.461, NCC: 2.402, NE: 2.400, SCC: 2.410, SE: 2.391*	−4.8	29.9	21.0	Huy et al., 2016a
		Deep learning	*The best model*	2.8	24.2	17.7	This study, 2021
3	*Observed AGB, DBH, H, WD Ecoregion*	Regression	$AGB = a_i \times (DBH^2HWD)^{bi}$ with a_i and b_i *for ecoregions respectively: CH: 0.798 and 0.966, NCC: 0.681 and 0.938, NE: 0.680 and 0.938, SCC: 0.685 and 0.940, SE: 0.647 and 0.931*	−5.9	28.0	19.5	Huy et al., 2016a
		Deep learning	*The best model*	1.2	20.8	15.7	This study, 2021

Note: Ecoregion: CH: Central Highlands, NCC: North Central Coastal, NE: Northeast, SCC: South Central Coastal, SE: Southeast. Cross-validation for deep learning (DL) with 10 realizations, each repeating the dataset of 968 samples was split randomly into 80% for training and 20% for validation; the best DL model was selected out of 10 validation results. DBH^2H (m^3)= (DBH (cm)/100)^2 × H (m); DBH^2HWD (kg) = DBH^2H × WD (g cm^{-3}) × 1000.

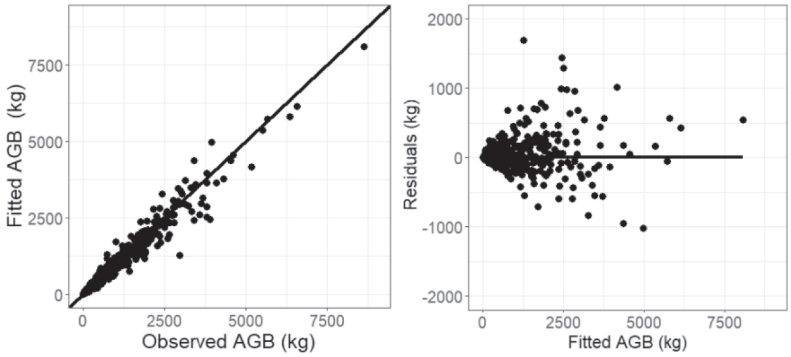
Table 5
Comparisons of the two methods of deep learning vs. regression of weighted non-linear fixed-effect model combined with optimal multivariate predictors and ecological, environmental, variables. Cross-validation results of predicting *AGB*.

ID	Method	Input variables	Equation/Model	FI	Bias (%)	RMSE (kg)	RMSPE (%)	MAPE (%)
1	Regression	Input 10 variables	$AGB = a\times(DBH^2 \times H \times WD)^b \times \exp(c_1\times (Ecoregion - 2.5) + c_2\times (Altitude - 547) + c_3\times(P - 1962) + c_4{}^*\times (T - 22.8) + c_5\times (Soil - 2.2) + c_6{}^*\times (N - 939))$	0.938	−2.7	217.8	26.4	19.3
2		Input 8 optimal variables	$AGB = a\times(DBH^2 \times H \times WD)^b \times \exp(c_1\times (Ecoregion - 2.5) + c_2\times (Altitude - 547) + c_3\times(P - 1962) + c_5\times (Soil\ type - 2.2))$	0.937	−3.0	203.4	26.8	19.2
3	Deep learning	Input 10 optimal variables: Observed *AGB, DBH, H, Ecoregion, Altitude, P, T, Soil type, N* and *WD.*	*The best model*	0.957	4.6	170.2	19.1	15.0

Note: Cross-validation for both deep learning (DL) and regression with 10 realizations, each repeating the dataset of 968 samples was split randomly into 80% for training and 20% for validation; statistics, errors in mean result were averaged over 10 times, and the best model of DL was selected out of 10 validation results. *: Parameter with $P_{value} > 0.05$.



Multivariate regression: Input 8 optimal input variables: *Observed AGB, DBH, H, WD, Ecoregion, Altitude, P,* and *Soil type.* Weighted fixed effect model with a combination of multi predictors of tree predictors, ecological, environmental variables fit by Maximum Likelihood.



Deep learning model: Input 10 optimal variables: *Observed AGB, DBH, H, Ecoregion, Altitude, P, T, Soil type, N* and *WD.*

Fig. 6. Plots compare performances of multivariate regression vs. deep learning models with input optimal variables. Fitted vs. Observed *AGB* (left) and Residuals vs. Fitted *AGB* based on the entire dataset.

AGB for tropical EBLFs, including at least 1 predictive variable and up to 9 predictors such as *DBH, H, Ecoregion, Altitude, P, T, Soil type, N,* and *WD*. One of the 16 best created DL models can be chosen to use depending on the actual ability to collect the predictive variables and their variation in the application region. The 9 predictive variables DL model had the highest reliability for predicting tree *AGB*.

The application of DL models to predict tree *AGB* requires input data collected at tree level in the sample plots. The measured and collected variables should be selected among nine predictive variables *DBH, H, Ecoregion, Altitude, P, T, Soil type, N,* and *WD*. Set up a measured data file formatted in *.csv to import the predictive variables data from all the measured trees following the collected, measured factors in sample plots. Each predictive variable is a column/field. Use a compiled Python code to read the saved best DL models together with the measured data file, thereby predicting the *AGB* for each tree in all sample plots. On that basis, calculate the total *AGB* of forest stand per hectare and the whole survey forest region.

In addition, the results of this study show that DL models can be

Table 6
Parameters of optimal multi predictors, ecological and environmental variables fixed effect model: $AGB = a \times (DBH^2 \times H \times WD)^b \times \exp(c_1 \times (Ecoregion - 2.5) + c_2 \times (Altitude - 547) + c_3 \times (P - 1962) + c_5 \times (Soil\ type - 2.2))$.

Parameters	Values	Std. Error
a	0.079835	0.002953
b	0.938173	0.004352
c_1	0.015205	0.006501
c_2	0.000170	0.000028
c_3	0.000182	0.000020
c_5	0.016620	0.008773

Note: Parameters were estimated by using the entire dataset of 968 samples.

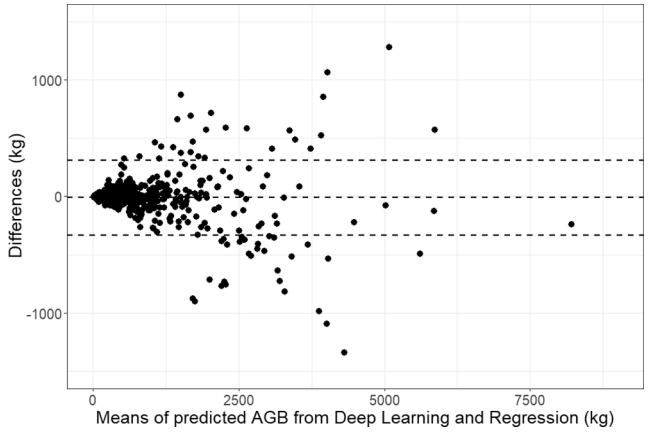


Fig. 7. Bland- Altman plot: Differences vs. Means of *AGB* predicted by the Deep Learning model and the Regression equation with their optimal input variables. The two dashed lines are the 95% limits of agreement estimated by mean difference ± 1.96 standard deviation of the differences.

applied more widely to simulate biological processes, biometrics and simulate complex ecological relationships of tropical forests for sustainable management. This is expected to be possible because DL does not require normality of the response variable, can handle heteroscedasticity (Montano et al., 2017), for a large number and different kinds of complex variables and samples, and does not need to examine for appropriate equation forms like traditional regression, and thanks to DL techniques through DNNs (Fig. 2) that can detect functions of complex biometric, environmental and ecological relationships (Ogana and Ercanli, 2021).

5. Conclusion

Trees *AGB* in tropical rain forests predicted by Deep Learning models had significantly higher reliability than the regression equations when both had the same input variables. In addition, DL models reduced RMSPE and MAPE by up to 7.7% and 6.1%, respectively, compared to traditional allometric equations. Sixteen best DL models were set up and stored with predictors from 1 to 9 variables. The DL model with 9 predictive variables *DBH, H, Ecoregion, Altitude, P, T, Soil type, N,* and *WD* was the best for predicting tree *AGB* in tropical EBLF.

The best DL models created in this study should be applied for measured tree data in accordance with factors of the forest stand, ecology, and environment variables in sampled plots to predict more accurately the tree *AGB* and total *AGB*, carbon on a large scale. This is because DL models involved many predictive variables such as forest stand, environmental, ecological factors. Therefore, when applied over a large area with variation in the value of these factors, it will provide higher predictive reliability of tree *AGB* than the allometric equations, which only have predictive variables of tree predictors as tradition. Thus, the DL models are recommended to apply in the MRV system of the REDD+program at a large regional or national, or territorial scale.

Credit authorship contribution statement

All authors do not have any financial and personal relationships with other people or organizations that could inappropriately influence (bias) their work. **Bao Huy:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing - original draft, Writing - review & editing. **Nguyen Quy Truong:** Conceptualization, Data curation, Formal analysis, Methodology, Software, Validation, Visualization, Writing - original draft. **Nguyen Quy Khiem:** Data curation, Formal analysis, Investigation, Project administration, Resources, Writing - original draft. **Krishna P. Poudel:** Methodology, Validation, Writing - review & editing. **Hailemariam Temesgen:** Methodology, Supervision, Validation, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The dataset used in this article was collected and compiled by Huy et al., 2016a. In terms of origin, part of the dataset came from the UN-REDD program in Viet Nam, and another part came from research funded by the Vietnamese Ministry of Education and Training.

References

Basuki, T.M., Van Lake, P.E., Skidmore, A.K., Hussin, Y.A., 2009. Allometric equations for estimating the aboveground biomass in the tropical lowland Dipterocarp forests. For. Ecol. Manag. 257 (2009), 1684–1694.
Bates, D.M., 2010. lme4: Mixed-effects modeling with R. Springer, p. 131.
Bland, J.M., Altman, D.G., 1999. Measuring agreement in method comparison studies. Stat. Methods Med. Res. 8 (2), 135–160.
Bosela, M., Stefancik, I., Marcis, P., Rubio-Cuadrado, A., Lukac, M., 2021. Thinning decreases aboveground biomass increment in central European beech forests but does not change individual tree resistance to climate events. Agric. For. Meteorol. 306 (2021), 108441 https://doi.org/10.1016/j.agrformet.2021.108441.
Brown, S., 1997. Estimating biomass and biomass change of tropical forests: A Primer. FAO Forestry paper – 134. ISBN 92-5-103955-0. Available on web site: http://www.fao.org/docrep/W4095E/w4095e00.htm#Contents.
Chave, J., Andalo, C., Brown, S., Cairns, M.A., Chambers, J.Q., Eamus, D., Folster, H., Fromard, F., Higuchi, N., Kira, T., Lescure, J.P., Nelson, B.W., Ogawa, H., Puig, H., Rier, B., Yamakura, T., 2005. Tree allometry and improved estimation of carbon stocks and balance in tropical forests. Oecologia 145 (2005), 87–99. https://doi.org/10.1007/s00442-005-0100-x.
Chave, J., Réjou-Méchain, M., Búrquez, A., Chidumayo, E., Colgan, M.S., Delitti, W.B.C., Duque, A., Eid, T., Fearnside, P.M., Goodman, R.C., Henry, M., Martínez-Yrízar, A., Mugasha, W.A., Muller-Landau, H.C., Mencuccini, M., Nelson, B.W., Ngomanda, A., Nogueira, E.M., Ortiz-Malavassi, E., Pélissier, R., Ploton, P., Ryan, C.M., Saldarriaga, J.G., Vieilledent, G., 2014. Improved allometric models to estimate the aboveground biomass of tropical trees. Glob. Change Biol. 20 (10), 3177–3190. https://doi.org/10.1111/gcb.12629.
Chollet, F., 2018. Deep Learning with Python. Manning, Shelter Island, NY, USA, p. 386.
Cysneiros, V.C., Souza, F.C.D., Gaui, T.D., Pelissari, A.L., Orso, G.A., Machado, S.D.A., Carvalho, D.C.D., Silveira-Filho, T.B., 2021. Integrating climate, soil and stand structure into allometric models: An approach of site-effects on tree allometry in Atlantic Forest. Ecol. Ind. 127 (2021), 107794 https://doi.org/10.1016/j.ecolind.2021.107794.
Dahouda, M.K., Joe, I., 2021. A Deep-Learned Embedding Technique for Categorical Features Encoding. Dig. Object Ident. 9, 114381–114391. https://doi.org/10.1109/ACCESS.2021.3104357.
Dang, A.T.N., Nandy, S., Srinet, R., Luong, N.V., Ghoshd, S., Kumara, A.S., 2019. Forest aboveground biomass estimation using machine learning regression algorithm in Yok Don National Park, Vietnam. Ecol. Inform. 50 (2019), 24–32.
Davidian, M., Giltinan, D.M., 1995. Nonlinear Mixed Effects Models for Repeated Measurement Data. Chapman and Hall, p. 356.
Fick, S.E., Hijmans, R.J., 2017. WorldClim 2: new 1km spatial resolution climate surfaces for global land areas. Int. J. Climatol. 37 (12), 4302–4315.
Fischer, G., Nachtergaele, F.O., Prieler, S., Teixeira, E., Toth, G., van Velthuizen, H., Verelst, L., Wiberg, D., 2008. Global Agro-ecological Zones Assessment for Agriculture (GAEZ 2008). IIASA, Laxenburg, Austria and FAO, Rome, Italy.

Furnival, G.M., 1961. An index for comparing equations used in constructing volume tables. For. Sci. 7, 337–341.
Ganatra, N., Patel, A., 2018. A Comprehensive Study of Deep Learning Architectures, Applications and Tools. Int. J. Comput. Sci. Eng. 6 (12), 701–705.
Goslee, K.M., Brown, S., Walker, S.M., Murray, L., Tepe, T., 2015. Review of aboveground biomass estimation techniques. Winrock International, Arkansas, USA, p. 31.
Guo, C., Berkhahn, F., 2016. Entity Embeddings of Categorical Variables. http://www.kaggle.com/c/rossmann-store-sales.
Hancock, J.T., Khoshgoftaar, T.M., 2020. Survey on categorical data for neural networks. J. Big Data 7 (1). https://doi.org/10.1186/s40537-020-00305-w.
Hayashi, Y., 2020. Does Deep Learning Work Well for Categorical Datasets with Mainly Nominal Attributes? Electronics 9, 1966. https://doi.org/10.3390/electronics9111966.
Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., Jarvis, A., 2005. Very high resolution interpolated climate surfaces for global land areas. Int. J. Climatol. 25 (15), 1965–1978.
Huy, B., Kralicek, K., Poudel, K.P., Phương, V.T., Khoa, P.V., Hung, N.D., Temesgen, H., 2016a. Allometric Equations for Estimating Tree Aboveground Biomass in Evergreen Broadleaf Forests of Viet Nam. For. Ecol. Manage. 382 (2016), 193–205. https://doi.org/10.1016/j.foreco.2016.10.021.
Huy, B., Nam, L.C., Poudel, K.P., Temesgen, H., 2021. Individual tree diameter growth modeling system for Dalat pine (*Pinus dalatensis* Ferré) of the upland mixed tropical forests, 118612: 1–15 For. Ecol. Manage. 480. https://doi.org/10.1016/j.foreco.2020.118612.
Huy, B., Poudel, K.P., Temesgen, H., 2016b. Aboveground biomass equations for evergreen broadleaf forests in South Central Coastal ecoregion of Viet Nam: Selection of eco-regional or pantropical models. For. Ecol. Mgmt. 376, 276–283.
Huy, B., Poudel, K.P., Kralicek, K., Hung, N.D., Khoa, P.V., Phuong, V.T., Temesgen, H., 2016c. Allometric Equations for Estimating Tree Aboveground Biomass in Tropical Dipterocarp Forests of Viet Nam. Forests 7 (180), 1–19. http://www.mdpi.com/1999-4907/7/8/180. https://doi.org/10.3390/f7080180.
Huy, B., Tinh, N.T., Poudel, K.P., Frank, B.M., Temesgen, H., 2019. Taxon-specific modeling systems for improving reliability of tree aboveground biomass and its components estimates in tropical dry dipterocarp forests. For. Ecol. Manage. 437 (2019), 156–174.
Ipcc, 2003. Good Practice Guidance for Land Use, Land-Use Change and Forestry. IPCC National Greenhouse Gas Inventories Programme, Hayama, Japan, p. 295.
IPCC, 2006. IPCC Guidelines for National Greenhouse Gas Inventories. In: Eggleston, H. S., Buendia, L., Miwa, K., Ngara, T., Tanabe, K. (Eds.), Prepared by the National Greenhouse Gas Inventories Programme. IGES, Japan.
Jais, I.K.M., Ismail, A.R., Nisa, S.Q., 2019. Adam Optimization Algorithm for Wide and Deep Neural Network. Knowl. Eng. Data Sci. (KEDS) 2 (1), 41–46.
Kassa, G.A., 2015. Forest Carbon Stock and Variations along Environmental Gradients in Yeka Forest and its Implication for Climate Change Mitigation. MSc. Thesis of Addis Ababa University, Ethiopia, p. 103.
Kingma, D.P., Ba, J.L., 2015. Adam: A Method for Stochastic Optimization. In: A Conference Paper at ICLR 2015.
Kralicek, K., Huy, B., Poudel, K.P., Temesgen, H., Salas, C., 2017. Simultaneous estimation of above- and belowground biomass in tropical forests of Viet Nam. ISSN: 0378-1127 For. Ecol. Manage. 390, 147–156. http://www.sciencedirect.com/science/article/pii/S0378112716307411.
Kriegeskorte, N., Golan, T., 2019. Neural network models and deep learning. Curr. Biol. 29 (7), R231–R236.
Kumar, V., Garg, M.L., 2018. Deep Learning as Frontier of Machine Learning: A Review. Int. J. Comput. Appl. 1 (182), 22–30.
LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (2015), 436–444. https://doi.org/10.1038/nature14539.
Lessard, V.C., McRoberts, R.E., Holdaway, M.R., 2001. Diameter Growth Models Using Minnesota Forest Inventory and Analysis Data. For. Sci. 47 (3), 301–310.
Liu, Z., Peng, C., Work, T., Candau, J.-N., DesRochers, A., Kneeshaw, D., 2018. Application of machine learning methods in forest ecology: recent progress and future challenges. Environ. Rev. 26 (4), 339–350. https://doi.org/10.1139/er-2018-0034.
Mankou, G.S., Ligot, G., Panzou, G.J.L., Boyemba, F., Loumeto, F.J., Ngomanda, A., Obiang, D., Rossi, V., Sonke, B., Yongo, O.D., Fayolle, A., 2021. Tropical tree allometry and crown allocation, and their relationship with species traits in central Africa. For. Ecol. Manage. 493 (2021), 119262 https://doi.org/10.1016/j.foreco.2021.119262.
Mathew, A., Amudha, P., Sivakumari, S., 2021. Deep Learning Techniques: An Overview. In: book: Advanced Machine Learning Technologies and Applications, pp. 599–608. http://dx.doi.10.1007/978-981-15-3383-9_54.
Montano, R.A.N.R., Sanquetta, C.R., Wojciechowski, J., Mattar, E., Corte, A.P.D., Todt, E., 2017. Artificial Intelligence Models to Estimate Biomass of Tropical Forest Trees. Polibits 56 (2017), 29–37.
Mushar, S.M.M, Sakinah Syed Ahmad, S., Kasmin, F., Hajar Zamah Shari, N., 2020. Machine learning approach for estimating tree volume. J. Phys. Conf. Ser. 1502 (1), 012039. https://doi.org/10.1088/1742-6596/1502/1/012039.
Nguyen, T.D., Kappas, M., 2020. Estimating the Aboveground Biomass of an Evergreen Broadleaf Forest in Xuan Lien Nature Reserve, Thanh Hoa, Vietnam, Using SPOT-6 Data and the Random Forest Algorithm. Int. J. For. Res. 2020, 1–13. https://doi.org/10.1155/2020/4216160.
Ogana, F.N., Ercanli, I., 2021. Modelling height-diameter relationships in complex tropical rain forest ecosystems using deep learning algorithm. J. For. Res. https://doi.org/10.1007/s11676-021-01373-1.
Pages, J., 2004. Analyse factorielle de donnees mixtes. Revue Statistique Appliquee LII (4), 93–111.
Parresol, B.R., 1999. Assessing Tree and Stand Biomass: A Review with Examples and Critical Comparisons. For. Sci. 45 (4), 573–593.
Pelletier, J., Codjia, C., Potvin, C., 2012. Traditional shifting agriculture: tracking forest carbon stock and biodiversity through time in western Panama. Glob. Change Biol. 2012 (18), 3581–3595. https://doi.org/10.1111/j.1365-2486.2012.02788.x.
Picard, N., Saint-André, L., Henry, M., 2012. Manual for building tree volume and biomass allometric equations: from field measurement to prediction. Food and Agricultural Organization of the United Nations, Rome, and Centre de Coopération Internationale en Recherche Agronomique pour le Développement, Montpellier, pp. 215.
Pinheiro, J., Bates, D., Debroy, S., Sarkar, D., Team, R.C., 2014. nlme: Linear and non-linear mixed effects models. R package version 3.1-117.
Potdar, K., S., T., D., C., 2017. A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers. Int. J. Comput. Appl. 175 (4), 7–9.
Python, 2021. Python Packaging User Guide. https://packaging.python.org/.
R Core Team, 2021. A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.r-project.org/index.html.
Temesgen, H., Affleck, D., Poudel, K., Gray, A., Sessions, J., 2015. A review of the challenges and opportunities in estimating above ground forest biomass using tree-level models. Scand. J. For. Res. 30 (4), 326–335. https://doi.org/10.1080/02827581.2015.1012114.
UN - United Nations, 2015. Paris Agreement, pp. 25.
Wickham, H., Chang, W., 2013. Package 'ggplot2': an implementation of the Grammar of Graphics.
Zaheer, R., Shaziya, H., 2019. A Study of the Optimization Algorithms in Deep Learning. In: International Conference on Inventive Systems and Control (ICISC 2019). IEEE Xplore Part Number: CFP19J06-ART, pp. 536–539 (ISBN: 978-1-5386-3950-4).
Zhang, L., Shao, Z., Liu, J., Cheng, Q., 2019. Deep Learning Based Retrieval of Forest Aboveground Biomass from Combined LiDAR and Landsat 8 Data. Remote Sens. 11, 1459. https://doi.org/10.3390/rs11121459.
Zhang, Y., Ma, J., Liang, S., Li, X., Li, M., 2020. An Evaluation of Eight Machine Learning Regression Algorithms for Forest Aboveground Biomass Estimation from Multiple Satellite Data Products. Remote Sens. 12, 4015. https://doi.org/10.3390/rs12244015.
Zhou, Z.H., Feng, J., 2019. Deep forest. Natl. Sci. Rev. 6, 74–86. https://doi.org/10.1093/nsr/nwy108.