Improving Remote Monitoring of Carbon Stock in Tropical Forests Using Machine Learning: A Case Study in Indonesian Borneo

Andrew J. Chamberlin * , Zac Yung-Chun Liu , Christopher G.L. Cross , Julie Pourtois , Iskandar Zulkarnaen Siregar , Dodik Ridho Nurrochmat , Yudi Setiawan , Kinari Webb , Skylar Hopkins , Susanne H. Sokolow , Giulio A. De Leo

Posted Date: 11 June 2024

doi: 10.20944/preprints202406.0671.v1

Keywords: Carbon Stock Estimation; Deep Learning; Deforestation

 doi:10.20944/preprints202406.0671.v1

Article

Improving Remote Monitoring of Carbon Stock in Tropical Forests Using Machine Learning: A Case Study in Indonesian Borneo

Andrew J. Chamberlin [1,*], Zac Yung-Chun Liu [1], Christopher G.L. Cross [2], Julie Pourtois [1], Iskandar Zulkarnaen Siregar [3], Dodik Ridho Nurrochmat [3], Yudi Setiawan [3,4],Kinari Webb [5], Skylar Hopkins [6], Susanne H. Sokolow [1,7] and Giulio A. De Leo [1]

[1] Hopkins Marine Station, Department of Oceans and of Earth System Science, Stanford University, Pacific Grove, CA 93950, USA
[3] Department of Computer Sciences, Stanford University, Stanford 94305, CA, USA
[3] Faculty of Forestry and Environment, IPB University, Bogor, Indonesia
[4] Environmental Research Center, International Research Institute for Environment and Climate Change, IPB University, Bogor, Indonesia
[5] Health in Harmony (HIH), Portland, OR, USA
[6] Department of Applied Ecology, North Carolina State University, Raleigh 27606, NC, USA
[7] Department of Ecology, Evolution, and Marine Biology and Marine Science Institute, University of California, Santa Barbara, CA 93106, USA
* Correspondence: achamb@stanford.edu

Abstract: Deforestation through land-use conversion, illegal logging, and timber trafficking is believed to cause ~10% of the annual human carbon dioxide emissions at the global level. Given the large contribution, local and national policies have been set in place in the effort to reduce deforestation and support reforestation. However, accurate assessment of forest carbon stock is expensive and challenging in remote areas and on large geographical scales. To improve carbon stock monitoring and evaluation of fine-scale forest loss, we developed a rapid, automatic, scalable, and cost-efficient generalized deep learning framework that uses diverse remote sensing data and satellite imagery to derive aboveground carbon density from accurate estimates of tree canopy heights at fine-grained resolution (30x30 meters) in remote tropical rainforests. The remote sensing data is composed of Landsat-8, Sentinel-1, land cover, digital elevation model, and NASA CMS airborne LiDAR, as well as vegetation indices, texture metrics, and climatic data. Data sources were compiled into a data pipeline which produced >300 features and 2 million observations over forests in Indonesian Borneo. Using LiDAR swath data on canopy forest height for ~100,000 hectares in Indonesian Borneo to create a training and validation datasets, our neural network model produced aboveground carbon density estimates with R^2 of 0.82, which is a significant improvement from comparable works by using Random Forest (R^2 of 0.3-0.5). This deep learning framework can be used to facilitate further carbon stock modeling in other forest regions (e.g., Brazil) as well as for the general purpose of climate change mitigation.

Keywords: carbon stock estimation; deep learning; deforestation

1. Introduction

1.1. Biomonitoring of Carbon in Tropical Forests

Tropical countries primarily contribute to carbon emissions through forest degradation and deforestation, and this accounts for approximately 10 percent of total global annual carbon emissions [1]. Local, national, and international initiatives (e.g., REDD+2) are underway to reduce carbon emissions. To assess whether these initiatives are delivering the expected results, it is crucial to

 doi:10.20944/preprints202406.0671.v1

develop precise, high-resolution, and cost-effective methods to track aboveground changes in carbon stocks.

Monitoring carbon stock in tropical rainforests can be expensive and challenging [2]. Traditional fieldwork to estimate aboveground carbon is a labor-intensive and time-consuming process, requiring at least 5-10 technicians working for weeks to measure a wide range of structural traits of trees (e.g., wood density of different tree species, density of individual trees, trunk diameters, canopy height) over a relatively small area (few hectares) to provide a localized estimate of carbon stock [2]. By replicating the same measures over a network of field plots, it is possible to derive an allometric relationship that estimates aboveground carbon density (ACD) based on these variables [3]. Whereas field methods provide accurate localized estimates, ACD extrapolation beyond the inventory plot networks remains challenging, as it would require deriving information on the same set of forest/tree structural traits at a fine scale over wide geographical areas [4]. Remote sensing methods are cheaper and are feasible at larger scales.

One useful remote sensing platform for monitoring and managing forest cover is Hansen Global Forest Watch (GFW) [5]. GFW data are available for free on the GFW Open Data Portal, which provides global coverage at several spatial and temporal resolutions. However, while useful for estimating forest area and changes on a yearly basis, Hansen GWF currently does not offer estimates of aboveground carbon density, nor can it be used for quantifying weekly or monthly changes in forest cover [5]. Therefore, novel datasets and pipelines are required in order to produce high-resolution, accurate time series data on ACD changes [6].

By providing high-definition, accurate three-dimensional data on forest canopy height and structure, airborne Light Detection and Ranging (LiDAR) optical sensing technology can be used to estimate ACD beyond the field plot inventory used to parametrize the allometric relationships between canopy height and ACD [6,7]. LiDAR-derived canopy height has been used to map biomass and estimate ACD in tropical forests including Central America [7], South America [8], Asia [9], Africa [10–12], and the oceanic islands [13]. However, high-definition airborne LiDAR data is also geographically limited in extent, a result of costs and logistical hindrances associated with the use of aircraft to carry the optical sensing technology [13,14]. Therefore, LiDAR is often paired with other data, such as optical remote sensing data of a range of spatial and spectral properties, to model LiDAR-derived properties across the larger geographic scale afforded by satellites [1,8,15,16].

Coupling airborne LiDAR along with satellite images and other geospatial datasets has become a more frequent approach to map tropical forest ACD, more so across geographic regions that do not have enough LiDAR measurements [17,18]. For instance, Asner et al. [19] coupled Landsat-derived parameters along with Shuttle Radar Topography Mission elevation metrics to regulate an ACD model that used airborne LiDAR for Peru. Baccini and Asner [20] used MODIS, Moderate Resolution Imaging Spectroradiometer, image data along with airborne LiDAR, to create pantropical high-resolution ACD maps. Notably, Mascaro et al. [6], for the first time, not only incorporated LiDAR data, but also used machine learning algorithms to estimate ACD. Today, evaluating forest properties, such as tree canopy heights (TCH) and ACD, from optical images can be accomplished rapidly via machine learning (ML) regression modeling [21–25].

1.4. Machine Learning (ML) to Estimate Forest Properties

Random Forest (RF) is the ML approach of choice, as it has proven to be superior to conventional ML techniques and tools for tropical forest carbon mapping applications [6]. This is in part due to RF being non-parametric and robust even with skewed data and a high number of input variables [26]. RF accuracy can be greatly enhanced by computing two widely used textural measures, namely Fourier transform textural ordination (FOTO) [27,28] and gray-level co-occurrence matrices (GLCM) [29–31]. Although textural features have been applied to imagery acquired with a variety of sensors, including WorldView-2 and -4 [32], Cartosat-1a [33], SPOT-5 [34], and IKONOS-2 [35], these texture features have yet to be tested on other gradient boosting decision tree models for large scale ACD mapping, such as XGBoost [36] and LightGBM [37], or with deep learning (DL) methods (e.g., neural network models).

 doi:10.20944/preprints202406.0671.v1

In this study, we present an open-source, rapid, automated, fine-scaled, cost-efficient, and generalizable framework to accurately estimate ACD. By stacking multiple sources of coarse-resolution satellite imagery and remotely-sensed earth data atop the high-resolution canopy tree height LiDAR data, we have trained multiple ML algorithms, including tree-based models (e.g., RF, XGBoost, LightGBM) and DL models (i.e., neural networks) to estimate TCH and ACD values over areas that lack high-quality LiDAR data. Our work demonstrates a solution for increased accessibility of carbon estimation of forests from freely-available and open-source remote sensing data.

2. Materials and Methods

2.1. Study Area and Data Pipeline

Borneo, a 750,000 km² island, contains some of the world's most biodiverse and carbon-dense tropical forests. However, Borneo has lost about 30% of its forests within the last 40 years [38]. In 2014, the NASA Carbon Monitoring System (CMS) surveyed 85 sites in Indonesian Borneo (Kalimantan) using plane-based LiDAR [39], creating canopy height digital surface models for 100,000 hectares of tropical forest TCH (Figure 1). We included this TCH data into our data pipeline and utilized it to calculate ACD at the pixel-level, as our ML regression target (see Section 2.2).

Figure 1. (Left) Location of Kalimantan in South-East Asia. (Right) NASA CMS LiDAR flight path locations, indicating LiDAR tracks (black color) in 85 sites in Indonesia Borneo.

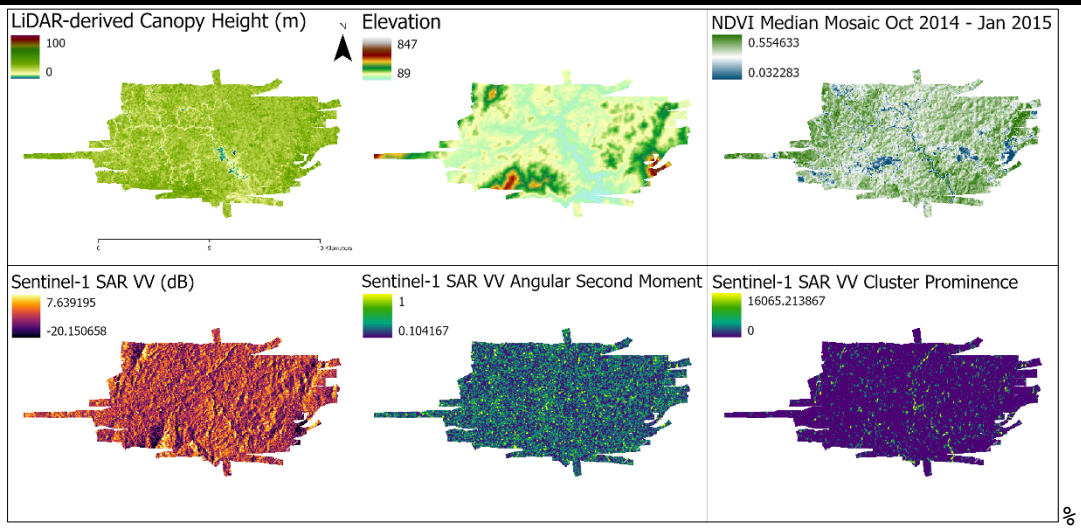
Like previous studies [6,40], we integrated several sources of common remote sensing data in our data pipeline, including Landsat-8 (8 spectral bands), Sentinel-1 (4 band Synthetic Aperture Radar), land cover fractional coverages from the Copernicus satellite mission, digital elevation model, and NASA CMS airborne LiDAR. Because of data loss due to cloud cover and other atmospheric effects, Landsat data was temporally sampled from August 2014 – January 2015 (the time period of the LiDAR data plus 2 months before and after). In addition to these data, we also integrated other data sources, such as indices commonly used to assess vegetation (e.g., NDVI, NDWI, NDII, EVI, for Landsat 8), Gray-Level Co-occurrence Matrix (GLCM) texture metrics [29] for all spectral data, and climatic data from WorldClim [41]. All these data sources (listed in Table 1) were compiled into a data processing pipeline which produced our dataset of 336 features for each LiDAR value (examples in Figure 2), resulting in about 2 million observations. The definition of each data feature is listed in Supplementary Material Table S1.

Table 1. Data sources in the GEE data pipelines.

Data	Type	Temporal Range	Spatial Resolution	Source
LANDSAT-8	OLI/TIRS sensors	Aug. 2014 – Jan. 2015	30 m	USGS

 doi:10.20944/preprints202406.0671.v1

Vegetation Indices - NDVI, NDWI, NDII, EVI, calculated for Landsat-8	Various measures associated with vegetation properties	Same as input	Same as input	Same as input
Sentinel-1	Synthetic Aperture Radar (SAR) instrument	Aug. 2014 – May 2015	10 m	ESA
Gray-Level Co-Occurrence Matrix (GLCM), derived from Landsat-8, Sentinel-1, and Landsat-8 vegetation indices	Textural image features derived from pixel spatial relationships	Same as input	Same as input	Same as input
Canopy Height Model (CHM)	Plane-mounted LiDAR	2014	1 m	NASA
NASA SRTM V3	Digital Elevation Model (elevation, slope, aspect)	2000	30 m	NASA
Bioclim	Climate	1970-2000	927.67 m	WorldClim
CopCover	Land Cover - Copernicus	2015	100 m	ESA
Land Use/Land Cover	Land Cover classification of Kalimantan - includes types of forests, other natural habitats, and developed lands	2011	30m	Indonesian Ministry of Forestry
OpenLandMap soil variables - soil bulk density (kg/m3), clay content (%), sand content (%), soil organic carbon (g/kg), soil pH in H20, soil water content (volumetric %), all at 0 cm depth.	Modeled soil properties from various global datasets of soil samples	1950-2018	250 m	OpenLandMap
Upstream drainage area (km2) and height above nearest drainage (m)	Hydrological flow dataset	1987-2017	90m	MERIT Hydro
Geomorphic layers - compound topographic index, terrain roughness index, vector ruggedness measure, roughness, topographic position index, and stream power index	Topographical relief characterization derived from MERIT Digital Elevation Model	1987-2017	90m	Geomorpho90m Geomorphometric Layers



 doi:10.20944/preprints202406.0671.v1

Figure 2. Examples of data enrichment for a selected LiDAR track. (Top Left) Canopy height model derived from LiDAR. (Top Center) Elevation. (Top Right) NDVI median mosaic, contemporaneous with LiDAR flights. (Bottom Left) Sentinel-1 SAR. (Bottom Center) GLCM Angular Second Moment texture feature from Sentinel-1. (Bottom Right) GLCM Cluster Prominence texture feature from Sentinel-1. The other features are described in Table S1.

All data was reprojected to the coordinate reference system of the corresponding Landsat imagery (30x30m pixel); LiDAR data was resampled using bicubic interpolation (Figure 3). Land Cover classification values were transformed via label encoding. This diverse remote sensing dataset was critical to providing the foundation on which we produced a robust, predictive ML framework. Here we utilized Google Earth Engine (GEE) [42] programmatically for data acquisition, processing, analysis, and modeling. The data processing pipelines we built in GEE can be further used to facilitate new data acquisition and processing for any future work in different forest regions and scope expansion.

2.2. ACD Estimation

In this study, we used the empirical relationship specific to Borneo, built by [9], to derive ACD based on TCH from airborne LiDAR. The empirical equations are the follows:

$$ACD = 0.567 \times TCH^{0.554} \times BA^{1.081} \times WD^{0.186}$$

$$BA = 1.112 \times TCH$$

$$WD = 0.385 \times TCH^{0.097}$$

where ACD is in *Mg C ha*$^{-1}$, TCH is in *m*, basal area (BA) is in m^2 ha^{-1}, and wood density (WD) is in g cm^{-3}. Jucker et al. [9] combined field data of tree height measurements with airborne laser scanning and hyperspectral imaging to characterize how topography shapes the vertical structure, wood density, diversity and ACD of nearly 15 km^2 of old-growth forest in Borneo. This general form of empirical relation was first developed by Asner et al. [8] and the allometric scaling coefficients in the equations are region-specific and are based on the field work that took place in that region. The error range of the ACD values calculated from the empirical equations is 39.3 *Mg C ha*$^{-1}$ [9].

 doi:10.20944/preprints202406.0671.v1

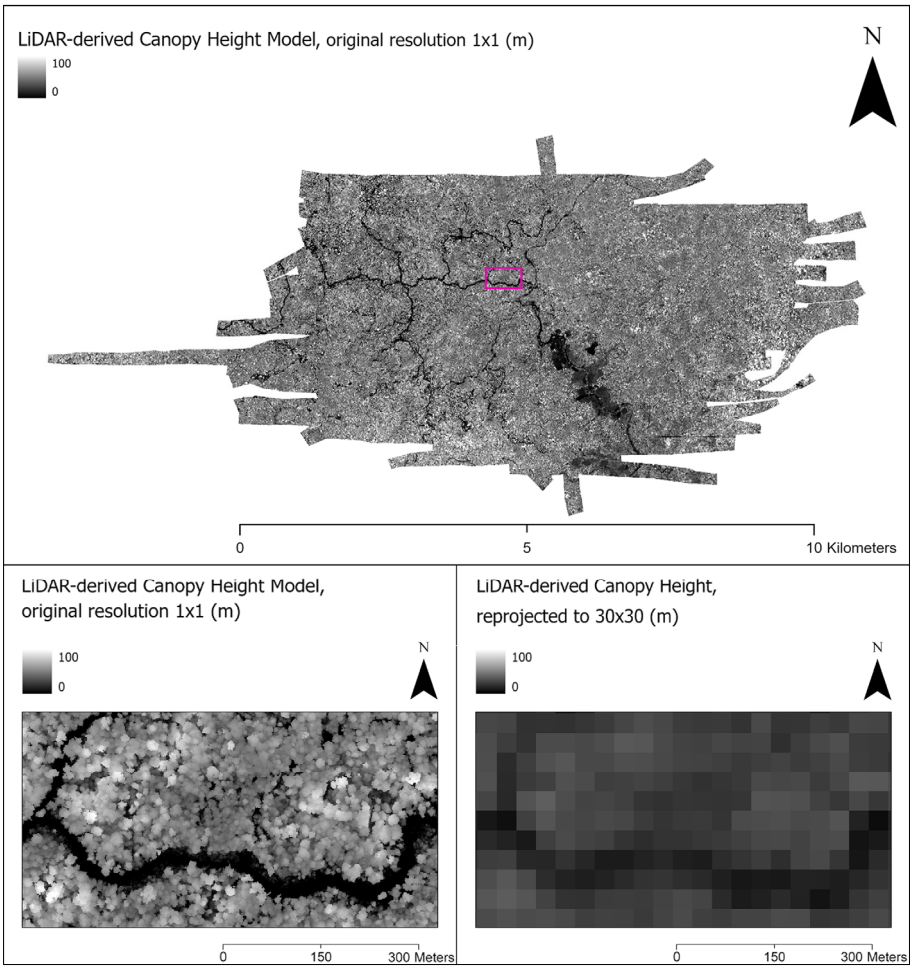


Figure 3. (Top) An original full extent, full resolution (1x1 m) LiDAR-derived Canopy Height Model (CHM) with a rectangle (pink) showing inset. (Left) Zoomed-in inset of CHM at original 1x1 m resolution, showing individual trees as light-colored circular shapes and a river/low ground in dark. (Right) CHM image after resampling to Landsat-8 scale (30x30 m).

2.3. ML Framework

In our framework of ACD estimation from ML modeling, we experimented with several ML and DL algorithms, including artificial neural network [43], random forest (RF) [44], XGBoost [36], and LightGBM [37], as well as the most recent AutoMLs [45,46].

From our data pipeline, we ran each ML algorithm with our dataset containing 336 features and 2 million TCH and ACD observations. The regression target for each ML model is the ACD value in pixel level. The goal was to develop a generalized ML model that is capable of predicting ACD solely based on open-source remote sensing data in the regions that LiDAR TCH is generally not available. The framework workflow is presented in Figure 4.

The workflow for ML model development consisted of two stages: (1) training and validation and (2) testing. The selected ML model takes the input data (GEE compiled remote sensing data) and ACD as the predicted output values. During the training and validation phase, the cost (or loss) function, the mean squared error (MSE), was used to evaluate the performance of the ML model by comparing the ACD estimation to the output values predicted by the ML model. Since this is a regression task, the evaluation metrics we used here is Coefficient of determination (R^2). We split the dataset into a 90% training subset, a 5% validation subset, and a 5% test set. The subsets were kept consistent for each round of experiment of the selected ML algorithms. The validation set was used to evaluate if the model is over-fitting to the training data, while the test set was treated as out-of-bag data that was not used in in the training phase.

 doi:10.20944/preprints202406.0671.v1

Note that in general, with large datasets (e.g., > 1 million data points), the artificial neural network (ANN) model is more generalizable and likely has higher predictive performance than the tree-based algorithms, such as RF. As RF has been tested rigorously in the remote sensing literatures in recent years, we treat the tree-based algorithms tested in this study as the baseline.

We used Python environment to establish our code base and ML framework. The ML computation, model training, and inference were done in Google Cloud Platform's Vertex AI.

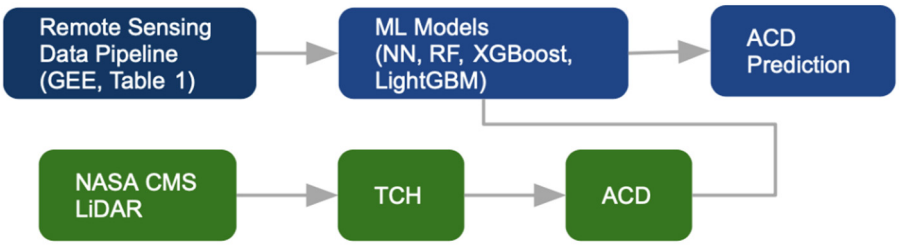


Figure 4. ML framework with GEE remote sensing data pipeline and GCP infrastructure with ML model development and deployment for ACD prediction.

2.3.1. Tree-Based Models

In addition to commonly used RF in the remote sensing community, we experimented with gradient boosted tree algorithms, such as XGBoost and LightGBM, as they have previously been effective in remote sensing image analysis and classifications of forests and land cover [47–49].

RF is a supervised ensemble algorithm that generates unpruned classification trees to predict a response [44]. RF implements bootstrapping samples of the data and randomized subsets of predictors. Final predictions are ensembled across a forest of classification trees (n = 100) based on an averaging of the probabilistic prediction of each classifier. Tree depth, number of leaf nodes, and number of features are tested with the dataset in each iteration in the training process.

XGBoost (eXtreme Gradient Boosting) is designed for both linear and tree-based supervised models [36]. XGBoost builds an ensemble of weak learners, where each weak learner is a decision tree. These decision trees are trained in a sequential manner, with each tree trying to correct the mistakes of the previous tree. The final output is a combination of all the individual trees, which results in a highly accurate model. XGBoost also includes a variety of regularization techniques, such as L1 and L2 regularization, and the ability to handle missing values and categorical variables.

Like XGBoost, LightGBM also builds an ensemble of weak learners, where each weak learner is a decision tree [37]. The key difference between LightGBM and XGBoost is the way they handle the decision tree learning process. LightGBM uses a technique called gradient-based one-side sampling to sample the data, which improves the speed of the training process. It also uses an algorithm called exclusive feature bundling to handle categorical variables, which reduces the number of splits required in the decision tree. Another difference between LightGBM and XGBoost is that LightGBM uses a leaf-wise tree growth algorithm, whereas XGBoost uses a level-wise tree growth algorithm. The leaf-wise algorithm tends to grow deeper trees than the level-wise algorithm, which results in better accuracy but can also lead to overfitting if not carefully controlled.

2.3.2. Neural Network Models

Deep learning algorithms (i.e., neural network models) have previously been effective for discovering underlying patterns in large datasets [43]. Therefore, we expected that neural network models could find highly predictive features from our diverse remote sensing data, which contained many data features (336) and data points (2 million). We specifically constructed the simple form of feedforward artificial neural network (ANN).

Feedforward ANNs consist of inter-connected computational units referred to as neurons that are arranged in layers. Input data is passed in one direction through an input layer and subsequently propagated through a defined number of hidden layers whereby the sum of the products of the

 doi:10.20944/preprints202406.0671.v1

connection weights from each layer approximate a function to estimate the observed output values [43]. Under repeated iteration and adjustment of the connection weights, a function between inputs and output is as closely estimated as possible given the parameter space available in the network [50,51]. Detailed development of model architecture lies in the selection of node functions (i.e., activation functions), layer sizes (number of hidden layers and numbers of nodes in each layer), learning rate, regularization parameters, and parameter dropout, which is discussed in Section 2.3.3.

As described above, the neural network used the input parameters as the first layer of neurons, and the last layer of neurons represents the predicted output values. Stochastic gradient descent, a common optimization method for ANN, was then used to iteratively adjust the weights and biases for each neuron to allow the ANN to best approximate the training data output. At each iteration, a backpropagation algorithm estimated the partial derivatives of the cost function (MSE) with respect to incremental changes of all the weights and biases, to determine the gradient descent directions for the next iteration. Note that in our model, the neurons of each hidden layer were composed of Rectified Linear Units (i.e., a ReLU activation function), to avoid the vanishing gradients [52]. Validation data were not used in the optimization or backpropagation algorithms. Instead, the cost function was evaluated over the validation data and served as an independent tuning metric of the performance of the ANN.

2.3.3. AutoML (Automated Machine Learning)

AutoML (Automated Machine Learning) is a technique that automates the process of building machine learning models, from data preparation to model selection and hyperparameter tuning [45]. The goal of AutoML is to make the process of building models more accessible and efficient for ML practitioners [46]. In this study, we primarily utilized AutoML for model selection and fine-tuning, as manually constructing the architecture of ANN model and fine tune hyperparameters, such as number of hidden layers, numbers of nodes in each layer, learning rate, and dropout rate, can be a time-consuming and complex process. We implemented a Python library, Auto-Keras [53], to automate the search process for ANN architecture that yields the best model prediction R^2. To evaluate the robustness of the ANN model, we compared it to the best performing tree-based model as our baseline. We implemented a Python library, TPOT, a tree-based pipeline optimization tool for automating machine learning [54], to automate the search process for tree-based models that yields the best model prediction R^2.

3. Results

In our experiments, with 336 data features and 2 million data points on ACD compiled from our GEE processing, the tree-based ML models achieved good R^2 performance on the test set. Our tree-based baseline ML models all have similar and consistent predictive performance: $R^2 = 0.572$ for the RF model, 0.583 for the XGBoost model, and 0.581 for the LightGBM model. The 3 most important features of each model are presented in Table 2. TPOT, the AutoML approach, found the best tree-based model to be the decision tree with fine-tuned parameters (maximum depth = 9, minimum sample leaf = 2, minimum samples split = 15) and the best R^2 was 0.593. These results are better than the comparable studies [6,55], here used as benchmark (R^2 of 0.3-0.5), due to the large number of features used in our study. For instance, the RF model in Mascaro et al. [6] was trained with data that contained only multispectral satellite channels and land cover features, which had a total of 11 data features and 80,000 data points.

 doi:10.20944/preprints202406.0671.v1

Table 2. Tree-Based Models Top 3 Features.

Model	Band Name/Feature	Importance Rank
Random Forest	Sentinel-1 VH Cluster Shade	1
Random Forest	Landsat-8 Thermal Infrared Information Measure of Correlation 2	2
Random Forest	Landsat-8 Thermal Infrared Inertia	3
XGBoost	Sentinel-1 VH Cluster Shade	1
XGBoost	Landsat-8 Red Difference Entropy	2
XGBoost	Landsat-8 Red Information Measure of Correlation 2	3
LightGBM	Sentinel-1 VH Cluster Shade	1
LightGBM	Landsat-8 Shortwave Infrared 1 Sum Entropy	2
LightGBM	Landsat-8 Blue Difference Entropy	3

For the DL experiments, the AutoML approach via Auto-Keras found the best neural network model to be 3 dense layers that applied multi-category encoding and normalization; each hidden layer had 32 neurons. ReLU activation function layers are implemented twice in this best-fit model, which allows for the neural network to learn hierarchical non-linear representations of the input data. The model architecture summary is presented in Table 3. This neural network had 13,090 total parameters in which 11,296 parameters were trainable. This custom feed-forward ANN achieved R^2 = 0.821 on the test set, which is higher than our tree-based baseline models and outperformed comparable studies [6,55]. After obtaining the best neural network model, we used it to produce a prediction ACD map over a section of Borneo where high-resolution LiDAR does not have coverage (Figure 5).

Table 3. Neural network model summary.

Layer (type)	Output shape	Param #
Input Layer	(None, 336)	0
Multi-category Encoding	(None, 336)	0
Normalization	(None, 336)	705
Dense	(None, 32)	11,296
ReLu	(None, 32)	0
Dense	(None, 32)	1,056
ReLu	(None, 32)	0
Dense (regression head)	(None, 1)	33

Total params: 13,090; Trainable params: 12,385; non-trainable params: 705.

 doi:10.20944/preprints202406.0671.v1

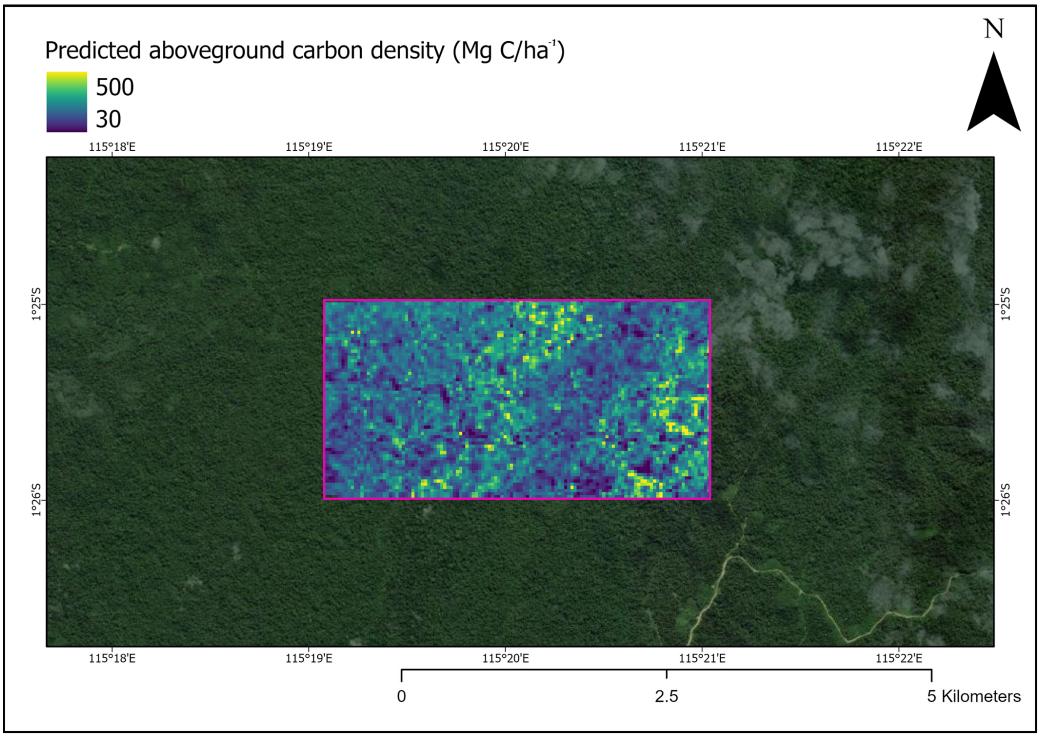


Figure 5. Section of forest in Indonesia Borneo with estimated aboveground carbon density visualized. The same trained model can be applied to any forested area in Indonesian Borneo that does not have LiDAR data.

4. Discussion

Here we show that large remote sensing datasets produce tree-based ML models with high predictive accuracy. Our three tree-based ML models trained with 336 data features and 2 million data points all had R^2 values greater than 0.57, which is substantially higher than prior studies with fewer features and data points that had R^2 values in the 0.3 – 0.5 range [6,55]. This resonates with the recent data-centric AI approach, which emphasizes the importance of improving data quality and quantity, instead of focusing on improving ML algorithms [56–58].

We also found that our deep learning neural network model yielded substantially higher R^2 than the tree-based models ($R^2 > 0.8$). This is consistent with previous studies that showed that neural networks are more generalizable and capable of extracting underlying patterns from a large dataset, resulting in a higher predictive performance [59]. This is a large benefit, but there is also a cost, because neural network models are more difficult to interpret than tree-based models. While it is mathematically possible to determine which nodes of a deep neural network were activated, it is difficult to assess how each of the neuron layers of neurons performed collectively. This is less straightforward than interpreting conventional tree-based models, which provide crisp rules on feature selections and encode for feature importance, making it easier to interpret the reasoning behind the model. In this study, we focus more on model predictive performance than interpretability, so we considered the DL based neural network modeling to be the best approach for our ADC estimation application.

As our models were not built for interpretability, but rather to maximize predictive power by including a large number of features, it is not sensical to select a single feature and comment on its significance. However, we did calculate the feature importance from the RF, XGBoost, and LightGBM models. In general, the top features that have the most predictive power for the ACD estimation were Sentinel-1 VH Cluster Shade (a GLCM texture feature), Landsat-8 thermal and infrared GLCM features, and Landsat-8 Red and Blue GLCM features. GLCM features comprised the top 5 features of every model run, which indicates the significance of texture analysis of medium-scale satellite imagery in the analysis of forest canopy structure.

 doi:10.20944/preprints202406.0671.v1

Note that the Landsat imagery used in our data pipeline is 30 x 30 m, so the sub-meter LiDAR data was resampled to derive mean height in each 30 x 30 m plot. The new generation of high temporal and spatial resolution satellite data, e.g., PlanetScope, which began observations over Borneo in 2018, consists of pixel cell sizes at ~3-5 m^2 and would thus reduce the amount of data resampling, potentially providing a better estimation of TCH and, in turn, of ACD [60]. PlanetScope sensors capture eight spectral bands, the same number as from Landsat, but don't have a thermal infrared sensor. Nevertheless, in our models, texture features derived from red, blue, and infrared spectra consistently were ranked the most important, suggesting that PlanetScope sensors would still capture the most useful information, especially as the amount of textural information would increase greatly with the finer resolution. The use of PlanetScope imagery is particularly appealing as a future avenue of research because PlanetScope satellites now cover the entire globe daily, allowing high temporal frequency imagery of the area of interest in Borneo. This would greatly help with dealing with cloud cover, which is a common challenge in rainforests like the ones we considered in Borneo.

5. Conclusions

In this study, we developed an improved method for evaluating fine-scale carbon stock by integrating LiDAR with satellite data and using a range of ML and DL algorithms. We found that Neural Network models trained on a large dataset of observations and features outperformed tree-based ML algorithms and allowed to achieve better prediction accuracy than any previous studies. Future efforts will employ Google Cloud to host and deploy this trained model on GEE. Our pipeline can be modified to estimate ACD in regions other than Borneo where location-specific equations exist to relate TCH to ACD and will be especially accurate in regions lacking large LiDAR coverage. This will provide insights for the necessary actions required to mitigate the effects of climate change.

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org, Table S1: Definition for each feature in the dataset.

Author Contributions: Conceptualization, A.J.C., Z.Y.-C.L., S.H.S. and G.A.D.L.; Data curation, A.J.C.; Funding acquisition, G.A.D.L.; Investigation, A.J.C., Z.Y.-C.L., C.G.L.C., J.P., S.H. and S.H.S.; Methodology, A.J.C. and Z.Y.-C.L.; Project administration, G.A.D.L.; Supervision, S.H.S. and G.A.D.L.; Validation, A.J.C. and Z.Y.-C.L.; Visualization, A.J.C. and Z.Y.-C.L.; Writing – original draft, A.J.C. and Z.Y.-C.L.; Writing – review & editing, A.J.C., Z.Y.-C.L., C.G.L.C., J.P., I.Z.S., D.R.N., Y.S., K.W., S.H., S.H.S. and G.A.D.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by a grant from the David and Lucile Packard Foundation (#2020-70906) and the National Geographic Society (#NGS-91292R-21), and Belmont Collaborative Forum on Climate, Environment and Health (NSF-ICER-2024383).

Data Availability Statement: Open-source dataset developed in this study can be accessed via this link (https://github.com/deleo-lab/carbon-remote-sensing-ml).

Acknowledgments: The authors would like to thank our institutions for providing facilities, resources, and administrative and technical support.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Baccini, A.; Goetz, S. J.; Walker, W. S.; Laporte, N. T.; Sun, M.; Sulla-Menashe, D.; Hackler, J.; Beck, P. S. A.; Dubayah, R.; Friedl, M. A.; Samanta, S.; Houghton, R. A. Estimated Carbon Dioxide Emissions from Tropical Deforestation Improved by Carbon-Density Maps. *Nature Clim Change* **2012**, *2* (3), 182–185. https://doi.org/10.1038/nclimate1354.
2. Mitchell, L. E.; Lin, J. C.; Bowling, D. R.; Pataki, D. E.; Strong, C.; Schauer, A. J.; Bares, R.; Bush, S. E.; Stephens, B. B.; Mendoza, D.; Mallia, D.; Holland, L.; Gurney, K. R.; Ehleringer, J. R. Long-Term Urban Carbon Dioxide Observations Reveal Spatial and Temporal Dynamics Related to Urban Characteristics and Growth. *Proceedings of the National Academy of Sciences* **2018**, *115* (12), 2912–2917. https://doi.org/10.1073/pnas.1702393115.
3. Chave, J.; Andalo, C.; Brown, S.; Cairns, M. A.; Chambers, J. Q.; Eamus, D.; Fölster, H.; Fromard, F.; Higuchi, N.; Kira, T.; Lescure, J.-P.; Nelson, B. W.; Ogawa, H.; Puig, H.; Riéra, B.; Yamakura, T. Tree Allometry and

 doi:10.20944/preprints202406.0671.v1

Improved Estimation of Carbon Stocks and Balance in Tropical Forests. *Oecologia* **2005**, *145* (1), 87–99. https://doi.org/10.1007/s00442-005-0100-x.

4. Brearley, F. Q.; Adinugroho, W. C.; Cámara-Leret, R.; Krisnawati, H.; Ledo, A.; Qie, L.; Smith, T. E. L.; Aini, F.; Garnier, F.; Lestari, N. S.; Mansur, M.; Murdjoko, A.; Oktarita, S.; Soraya, E.; Tata, H. L.; Tiryana, T.; Trethowan, L. A.; Wheeler, C. E.; Abdullah, M.; Aswandi; Buckley, B. J. W.; Cantarello, E.; Dunggio, I.; Gunawan, H.; Heatubun, C. D.; Arini, D. I. D.; Istomo; Komar, T. E.; Kuswandi, R.; Mutaqien, Z.; Pangala, S. R.; Ramadhanil; Prayoto; Puspanti, A.; Qirom, M. A.; Rozak, A. H.; Sadili, A.; Samsoedin, I.; Sulistyawati, E.; Sundari, S.; Sutomo; Tampubolon, A. P.; Webb, C. O. Opportunities and Challenges for an Indonesian Forest Monitoring Network. *Annals of Forest Science* **2019**, *76* (2), 54. https://doi.org/10.1007/s13595-019-0840-0.
5. Hansen, M. C.; Potapov, P. V.; Moore, R.; Hancher, M.; Turubanova, S. A.; Tyukavina, A.; Thau, D.; Stehman, S. V.; Goetz, S. J.; Loveland, T. R.; Kommareddy, A.; Egorov, A.; Chini, L.; Justice, C. O.; Townshend, J. R. G. High-Resolution Global Maps of 21st-Century Forest Cover Change. *Science* **2013**, *342* (6160), 850–853. https://doi.org/10.1126/science.1244693.
6. Mascaro, J.; Asner, G. P.; Knapp, D. E.; Kennedy-Bowdoin, T.; Martin, R. E.; Anderson, C.; Higgins, M.; Chadwick, K. D. A Tale of Two "Forests": Random Forest Machine Learning Aids Tropical Forest Carbon Mapping. *PLOS ONE* **2014**, *9* (1), e85993. https://doi.org/10.1371/journal.pone.0085993.
7. Mascaro, J.; Asner, G. P.; Muller-Landau, H. C.; van Breugel, M.; Hall, J.; Dahlin, K. Controls over Aboveground Forest Carbon Density on Barro Colorado Island, Panama. *Biogeosciences* **2011**, *8* (6), 1615–1629. https://doi.org/10.5194/bg-8-1615-2011.
8. Asner, G. P.; Mascaro, J. Mapping Tropical Forest Carbon: Calibrating Plot Estimates to a Simple LiDAR Metric. *Remote Sensing of Environment* **2014**, *140*, 614–624. https://doi.org/10.1016/j.rse.2013.09.023.
9. Jucker, T.; Bongalov, B.; Burslem, D. F. R. P.; Nilus, R.; Dalponte, M.; Lewis, S. L.; Phillips, O. L.; Qie, L.; Coomes, D. A. Topography Shapes the Structure, Composition and Function of Tropical Forest Landscapes. *Ecology Letters* **2018**, *21* (7), 989–1000. https://doi.org/10.1111/ele.12964.
10. Asner, G. P.; Mascaro, J.; Muller-Landau, H. C.; Vieilledent, G.; Vaudry, R.; Rasamoelina, M.; Hall, J. S.; van Breugel, M. A Universal Airborne LiDAR Approach for Tropical Forest Carbon Mapping. *Oecologia* **2012**, *168* (4), 1147–1160. https://doi.org/10.1007/s00442-011-2165-z.
11. Bouvet, A.; Mermoz, S.; Le Toan, T.; Villard, L.; Mathieu, R.; Naidoo, L.; Asner, G. P. An Above-Ground Biomass Map of African Savannahs and Woodlands at 25m Resolution Derived from ALOS PALSAR. *Remote Sensing of Environment* **2018**, *206*, 156–173. https://doi.org/10.1016/j.rse.2017.12.030.
12. Vaglio Laurin, G.; Chen, Q.; Lindsell, J.; Coomes, D.; Del Frate, F.; Guerriero, L.; Pirotti, F.; Valentini, R. Above Ground Biomass Estimation in an African Tropical Forest with Lidar and Hyperspectral Data. *ISPRS Journal of Photogrammetry and Remote Sensing* **2014**, *89*, 49–58. https://doi.org/10.1016/j.isprsjprs.2014.01.001.
13. Hughes, R. F.; Asner, G. P.; Baldwin, J. A.; Mascaro, J.; Bufil, L. K. K.; Knapp, D. E. Estimating Aboveground Carbon Density across Forest Landscapes of Hawaii: Combining FIA Plot-Derived Estimates and Airborne LiDAR. *Forest Ecology and Management* **2018**, *424*, 323–337. https://doi.org/10.1016/j.foreco.2018.04.053.
14. Asner, G. P.; Clark, J. K.; Mascaro, J.; Galindo García, G. A.; Chadwick, K. D.; Navarrete Encinales, D. A.; Paez-Acosta, G.; Cabrera Montenegro, E.; Kennedy-Bowdoin, T.; Duque, Á.; Balaji, A.; von Hildebrand, P.; Maatoug, L.; Phillips Bernal, J. F.; Yepes Quintero, A. P.; Knapp, D. E.; García Dávila, M. C.; Jacobson, J.; Ordóñez, M. F. High-Resolution Mapping of Forest Carbon Stocks in the Colombian Amazon. *Biogeosciences* **2012**, *9* (7), 2683–2696. https://doi.org/10.5194/bg-9-2683-2012.
15. Saatchi, S. S.; Harris, N. L.; Brown, S.; Lefsky, M.; Mitchard, E. T. A.; Salas, W.; Zutta, B. R.; Buermann, W.; Lewis, S. L.; Hagen, S.; Petrova, S.; White, L.; Silman, M.; Morel, A. Benchmark Map of Forest Carbon Stocks in Tropical Regions across Three Continents. *Proceedings of the National Academy of Sciences* **2011**, *108* (24), 9899–9904. https://doi.org/10.1073/pnas.1019576108.
16. Yang, Y.; Saatchi, S. S.; Xu, L.; Yu, Y.; Choi, S.; Phillips, N.; Kennedy, R.; Keller, M.; Knyazikhin, Y.; Myneni, R. B. Post-Drought Decline of the Amazon Carbon Sink. *Nat Commun* **2018**, *9* (1), 3172. https://doi.org/10.1038/s41467-018-05668-6.
17. Asner, G. P.; Rudel, T. K.; Aide, T. M.; Defries, R.; Emerson, R. A Contemporary Assessment of Change in Humid Tropical Forests. *Conserv Biol* **2009**, *23* (6), 1386–1395. https://doi.org/10.1111/j.1523-1739.2009.01333.x.
18. Asner, G. P.; Brodrick, P. G.; Philipson, C.; Vaughn, N. R.; Martin, R. E.; Knapp, D. E.; Heckler, J.; Evans, L. J.; Jucker, T.; Goossens, B.; Stark, D. J.; Reynolds, G.; Ong, R.; Renneboog, N.; Kugan, F.; Coomes, D. A. Mapped Aboveground Carbon Stocks to Advance Forest Conservation and Recovery in Malaysian Borneo. *Biological Conservation* **2018**, *217*, 289–310. https://doi.org/10.1016/j.biocon.2017.10.020.
19. Asner, G. P.; Knapp, D. E.; Martin, R. E.; Tupayachi, R.; Anderson, C. B.; Mascaro, J.; Sinca, F.; Chadwick, K. D.; Higgins, M.; Farfan, W.; Llactayo, W.; Silman, M. R. Targeted Carbon Conservation at National Scales with High-Resolution Monitoring. *Proceedings of the National Academy of Sciences* **2014**, *111* (47), E5016–E5022. https://doi.org/10.1073/pnas.1419550111.

 doi:10.20944/preprints202406.0671.v1

20. Baccini, A.; Asner, G. P. Improving Pantropical Forest Carbon Maps with Airborne LiDAR Sampling. *Carbon Management* **2013**, *4* (6), 591–600. https://doi.org/10.4155/cmt.13.66.
21. Ali, A. M.; Darvishzadeh, R.; Skidmore, A. K.; Duren, I. van. Effects of Canopy Structural Variables on Retrieval of Leaf Dry Matter Content and Specific Leaf Area From Remotely Sensed Data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2016**, *9* (2), 898–909. https://doi.org/10.1109/JSTARS.2015.2450762.
22. Baccini, A.; Walker, W.; Carvalho, L.; Farina, M.; Sulla-Menashe, D.; Houghton, R. A. Tropical Forests Are a Net Carbon Source Based on Aboveground Measurements of Gain and Loss. *Science* **2017**, *358* (6360), 230–234. https://doi.org/10.1126/science.aam5962.
23. De'ath, G.; Fabricius, K. E. Classification and Regression Trees: A Powerful yet Simple Technique for Ecological Data Analysis. *Ecology* **2000**, *81* (11), 3178–3192. https://doi.org/10.1890/0012-9658(2000)081[3178:CARTAP]2.0.CO;2.
24. Gleason, C. J.; Im, J. Forest Biomass Estimation from Airborne LiDAR Data Using Machine Learning Approaches. *Remote Sensing of Environment* **2012**, *125*, 80–91. https://doi.org/10.1016/j.rse.2012.07.006.
25. Prasad, A. M.; Iverson, L. R.; Liaw, A. Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction. *Ecosystems* **2006**, *9* (2), 181–199. https://doi.org/10.1007/s10021-005-0054-1.
26. Evans, J. S.; Murphy, M. A.; Holden, Z. A.; Cushman, S. A. Modeling Species Distribution and Change Using Random Forest [Chapter 8]. *In: Drew, A. C.; Wiersma, Y.; Huettmann, F., eds. Predictive Species and Habitat Modeling in Landscape Ecology. New York, NY: Springer. p.139-159.* **2011**, 139–159. https://doi.org/10.1007/978-1-4419-7390-0_8.
27. Couteron, P.; Barbier, N.; Gautier, D. Textural Ordination Based on Fourier Spectral Decomposition: A Method to Analyze and Compare Landscape Patterns. *Landscape Ecology* **2006**, *21*, 555–567. https://doi.org/10.1007/s10980-005-2166-6.
28. Ploton, P.; Pélissier, R.; Proisy, C.; Flavenot, T.; Barbier, N.; Rai, S. N.; Couteron, P. Assessing Aboveground Tropical Forest Biomass Using Google Earth Canopy Images. *Ecological Applications* **2012**, *22* (3), 993–1003. https://doi.org/10.1890/11-1606.1.
29. Haralick, R. M.; Shanmugam, K.; Dinstein, I. Textural Features for Image Classification. *IEEE Transactions on Systems, Man, and Cybernetics* **1973**, *SMC-3* (6), 610–621. https://doi.org/10.1109/TSMC.1973.4309314.
30. Meng, S.; Pang, Y.; Zhang, Z.; Jia, W.; Li, Z. Mapping Aboveground Biomass Using Texture Indices from Aerial Photos in a Temperate Forest of Northeastern China. *Remote Sensing* **2016**, *8* (3), 230. https://doi.org/10.3390/rs8030230.
31. Wood, E. M.; Pidgeon, A. M.; Radeloff, V. C.; Keuler, N. S. Image Texture as a Remotely Sensed Measure of Vegetation Structure. *Remote Sensing of Environment* **2012**, *121*, 516–526. https://doi.org/10.1016/j.rse.2012.01.003.
32. Wen, D.; Huang, X.; Bovolo, F.; Li, J.; Ke, X.; Zhang, A.; Benediktsson, J. A. Change Detection From Very-High-Spatial-Resolution Optical Remote Sensing Images: Methods, Applications, and Future Directions. *IEEE Geoscience and Remote Sensing Magazine* **2021**, *9* (4), 68–101. https://doi.org/10.1109/MGRS.2021.3063465.
33. Pargal, S.; Fararoda, R.; Rajashekar, G.; Balachandran, N.; Réjou-Méchain, M.; Barbier, N.; Jha, C. S.; Pélissier, R.; Dadhwal, V. K.; Couteron, P. Inverting Aboveground Biomass–Canopy Texture Relationships in a Landscape of Forest Mosaic in the Western Ghats of India Using Very High Resolution Cartosat Imagery. *Remote Sensing* **2017**, *9* (3), 228. https://doi.org/10.3390/rs9030228.
34. Zhang, H.; Li, Q.; Liu, J.; Du, X.; Dong, T.; McNairn, H.; Champagne, C.; Liu, M.; Shang, J. Object-Based Crop Classification Using Multi-Temporal SPOT-5 Imagery and Textural Features with a Random Forest Classifier. *Geocarto International* **2018**, *33* (10), 1017–1035. https://doi.org/10.1080/10106049.2017.1333533.
35. Kayitakire, F.; Hamel, C.; Defourny, P. Retrieving Forest Structure Variables Based on Image Texture Analysis and IKONOS-2 Imagery. *Remote Sensing of Environment* **2006**, *102* (3), 390–401. https://doi.org/10.1016/j.rse.2006.02.022.
36. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; KDD '16; Association for Computing Machinery: New York, NY, USA, 2016; pp 785–794. https://doi.org/10.1145/2939672.2939785.
37. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc., 2017; Vol. 30.
38. Gaveau, D. L. A.; Sloan, S.; Molidena, E.; Yaen, H.; Sheil, D.; Abram, N. K.; Ancrenaz, M.; Nasi, R.; Quinones, M.; Wielaard, N.; Meijaard, E. Four Decades of Forest Persistence, Clearance and Logging on Borneo. *PLOS ONE* **2014**, *9* (7), e101654. https://doi.org/10.1371/journal.pone.0101654.
39. Melendy, L.; Hagen, S.; Sullivan, F. B.; Pearson, T.; Walker, S. M.; Ellis, P.; KUSTIYO; Sambodo, K. A.; Roswintiarti, O.; Hanson, M.; Klassen, A. W.; Palace, M. W.; Braswell, B. H.; Delgado, G. M.; Saatchi, S. S.;

 doi:10.20944/preprints202406.0671.v1

Ferraz, A. CMS: LiDAR-Derived Canopy Height, Elevation for Sites in Kalimantan, Indonesia, 2014. *ORNL DAAC* **2017**. https://doi.org/10.3334/ORNLDAAC/1540.

40. Vafaei, S.; Soosani, J.; Adeli, K.; Fadaei, H.; Naghavi, H.; Pham, T. D.; Tien Bui, D. Improving Accuracy Estimation of Forest Aboveground Biomass Based on Incorporation of ALOS-2 PALSAR-2 and Sentinel-2A Imagery and Machine Learning: A Case Study of the Hyrcanian Forest Area (Iran). *Remote Sensing* **2018**, *10* (2), 172. https://doi.org/10.3390/rs10020172.
41. Fick, S. E.; Hijmans, R. J. WorldClim 2: New 1-Km Spatial Resolution Climate Surfaces for Global Land Areas. *International Journal of Climatology* **2017**, *37* (12), 4302–4315. https://doi.org/10.1002/joc.5086.
42. Gorelick, N.; Hancher, M.; Dixon, M.; Ilyushchenko, S.; Thau, D.; Moore, R. Google Earth Engine: Planetary-Scale Geospatial Analysis for Everyone. *Remote Sensing of Environment* **2017**, *202*, 18–27. https://doi.org/10.1016/j.rse.2017.06.031.
43. LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521* (7553), 436–444. https://doi.org/10.1038/nature14539.
44. Breiman, L. Random Forests. *Machine Learning* **2001**, *45* (1), 5–32. https://doi.org/10.1023/A:1010933404324.
45. He, X.; Zhao, K.; Chu, X. AutoML: A Survey of the State-of-the-Art. *Knowledge-Based Systems* **2021**, *212*, 106622. https://doi.org/10.1016/j.knosys.2020.106622.
46. Karmaker ("Santu"), S. K.; Hassan, Md. M.; Smith, M. J.; Xu, L.; Zhai, C.; Veeramachaneni, K. AutoML to Date and Beyond: Challenges and Opportunities. *ACM Comput. Surv.* **2021**, *54* (8), 175:1-175:36. https://doi.org/10.1145/3470918.
47. Bhagwat, R. U.; Uma Shankar, B. A Novel Multilabel Classification of Remote Sensing Images Using XGBoost. In *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)*; 2019; pp 1–5. https://doi.org/10.1109/I2CT45611.2019.9033768.
48. Samat, A.; Li, E.; Wang, W.; Liu, S.; Lin, C.; Abuduwaili, J. Meta-XGBoost for Hyperspectral Image Classification Using Extended MSER-Guided Morphological Profiles. *Remote Sensing* **2020**, *12* (12), 1973. https://doi.org/10.3390/rs12121973.
49. Łoś, H.; Mendes, G. S.; Cordeiro, D.; Grosso, N.; Costa, H.; Benevides, P.; Caetano, M. Evaluation of Xgboost and Lgbm Performance in Tree Species Classification with Sentinel-2 Data. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*; 2021; pp 5803–5806. https://doi.org/10.1109/IGARSS47720.2021.9553031.
50. Nielsen, M. *Neural Networks and Deep Learning*; Determination Press, 2015. http://neuralnetworksanddeeplearning.com/
51. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press, 2016.
52. Nair, V.; Hinton, G. Rectified Linear Units Improve Restricted Boltzmann Machines. In *Proceedings of the 27th International Conference on Machine Learning*; 2010; Vol. 27, p 814.
53. Jin, H.; Song, Q.; Hu, X. Auto-Keras: An Efficient Neural Architecture Search System. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*; ACM: Anchorage AK USA, 2019; pp 1946–1956. https://doi.org/10.1145/3292500.3330648.
54. Olson, R. S.; Moore, J. H. TPOT: A Tree-Based Pipeline Optimization Tool for Automating Machine Learning. In *Automated Machine Learning*; Hutter, F., Kotthoff, L., Vanschoren, J., Eds.; The Springer Series on Challenges in Machine Learning; Springer International Publishing: Cham, 2019; pp 151–160. https://doi.org/10.1007/978-3-030-05318-5_8.
55. Pham, T. D.; Yoshino, K.; Le, N. N.; Bui, D. T. Estimating Aboveground Biomass of a Mangrove Plantation on the Northern Coast of Vietnam Using Machine Learning Techniques with an Integration of ALOS-2 PALSAR-2 and Sentinel-2A Data. *International Journal of Remote Sensing* **2018**, *39* (22), 7761–7788. https://doi.org/10.1080/01431161.2018.1471544.
56. Ng, D. T. K.; Leung, J. K. L.; Chu, S. K. W.; Qiao, M. S. Conceptualizing AI Literacy: An Exploratory Review. *Computers and Education: Artificial Intelligence* **2021**, *2*, 100041. https://doi.org/10.1016/j.caeai.2021.100041.
57. Jarrahi, M. H.; Lutz, C.; Newlands, G. Artificial Intelligence, Human Intelligence and Hybrid Intelligence Based on Mutual Augmentation. *Big Data & Society* **2022**, *9* (2), 20539517221142824. https://doi.org/10.1177/20539517221142824.
58. Whang, S. E.; Roh, Y.; Song, H.; Lee, J.-G. Data Collection and Quality Challenges in Deep Learning: A Data-Centric AI Perspective. *The VLDB Journal* **2023**, *32* (4), 791–813. https://doi.org/10.1007/s00778-022-00775-9.
59. Liu, T.; Yao, L.; Qin, J.; Lu, J.; Lu, N.; Zhou, C. A Deep Neural Network for the Estimation of Tree Density Based on High-Spatial Resolution Image. *IEEE Transactions on Geoscience and Remote Sensing* **2022**, *60*, 1–11. https://doi.org/10.1109/TGRS.2021.3101056.
60. Csillik, O.; Kumar, P.; Mascaro, J.; O'Shea, T.; Asner, G. P. Monitoring Tropical Forest Carbon Stocks and Emissions Using Planet Satellite Data. *Sci Rep* **2019**, *9* (1), 17831. https://doi.org/10.1038/s41598-019-54386-6.

 doi:10.20944/preprints202406.0671.v1

61. McFEETERS, S. K. The Use of the Normalized Difference Water Index (NDWI) in the Delineation of Open Water Features. *International Journal of Remote Sensing* **1996**, *17* (7), 1425–1432. https://doi.org/10.1080/01431169608948714.
62. Kriegler, F., Malila, W., Nalepka, R., & Richardson, W. Preprocessing transformations and their effect on multispectral recognition. Proceedings of the 6th International Symposium on Remote Sensing of Environment, Ann Arbor, MI, 1969.
63. Hunt, E. R.; Rock, B. N. Detection of Changes in Leaf Water Content Using Near- and Middle-Infrared Reflectances. *Remote Sensing of Environment* **1989**, *30* (1), 43–54. https://doi.org/10.1016/0034-4257(89)90046-1.
64. Huete, A.; Didan, K.; Miura, T.; Rodriguez, E. P.; Gao, X.; Ferreira, L. G. Overview of the Radiometric and Biophysical Performance of the MODIS Vegetation Indices. *Remote Sensing of Environment* **2002**, *83* (1), 195–213. https://doi.org/10.1016/S0034-4257(02)00096-2.
65. Hijmans, R. J.; Cameron, S. E.; Parra, J. L.; Jones, P. G.; Jarvis, A. Very High Resolution Interpolated Climate Surfaces for Global Land Areas. *International Journal of Climatology* **2005**, *25* (15), 1965–1978. https://doi.org/10.1002/joc.1276.
66. Farr, T. G.; Rosen, P. A.; Caro, E.; Crippen, R.; Duren, R.; Hensley, S.; Kobrick, M.; Paller, M.; Rodriguez, E.; Roth, L.; Seal, D.; Shaffer, S.; Shimada, J.; Umland, J.; Werner, M.; Oskin, M.; Burbank, D.; Alsdorf, D. The Shuttle Radar Topography Mission. *Reviews of Geophysics* **2007**, *45* (2). https://doi.org/10.1029/2005RG000183.
67. Copernicus Global Land Service: Land Cover 100m: Collection 3: Epoch 2015: Globe. Available online: https://doi.org/10.5281/zenodo.3939038.
68. Bisong, E. Google Colaboratory. In *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners*; Bisong, E., Ed.; Apress: Berkeley, CA, 2019; pp 59–64. https://doi.org/10.1007/978-1-4842-4470-8_7.
69. Ferraz, A.; Saatchi, S. S.; Xu, L.; Hagen, S.; Chave, J.; Yu, Y.; Meyer, V.; Garcia, M.; Silva, C.; Roswintiarti, O.; Samboko, A.; Sist, P.; Walker, S. M.; Pearson, T.; Wijaya, A.; Sullivan, F. B.; Rutishauser, E.; Hoekman, D.; Ganguly, S. Aboveground Biomass, Landcover, and Degradation, Kalimantan Forests, Indonesia, 2014. *ORNL DAAC* **2019**. https://doi.org/10.3334/ORNLDAAC/1645.
70. Clay Content in % (Kg / Kg) at 6 Standard Depths (0, 10, 30, 60, 100 and 200 Cm) at 250 m Resolution. Available online: https://doi.org/10.5281/zenodo.2525663 (accessed on 30 November 2022).
71. Sand Content in % (Kg / Kg) at 6 Standard Depths (0, 10, 30, 60, 100 and 200 Cm) at 250 m Resolution. Available online: https://doi.org/10.5281/zenodo.2525662 (accessed on 30 November 2022).
72. Soil Water Content (Volumetric %) for 33kPa and 1500kPa Suctions Predicted at 6 Standard Depths (0, 10, 30, 60, 100 and 200 Cm) at 250 m Resolution. Available online: https://doi.org/10.5281/zenodo.2784001 (accessed on 30 November 2022).
73. Soil Organic Carbon Content in x 5 g / Kg at 6 Standard Depths (0, 10, 30, 60, 100 and 200 Cm) at 250 m Resolution. Available online: https://doi.org/10.5281/zenodo.2525553 (accessed on 30 November 2022).
74. Soil pH in H2O at 6 Standard Depths (0, 10, 30, 60, 100 and 200 Cm) at 250 m Resolution. Available online: https://doi.org/10.5281/zenodo.2525664 (accessed on 30 November 2022).
75. Soil Bulk Density (Fine Earth) 10 x Kg / m-Cubic at 6 Standard Depths (0, 10, 30, 60, 100 and 200 Cm) at 250 m Resolution. Available online: https://doi.org/10.5281/zenodo.2525665 (accessed on 30 November 2022).
76. Yamazaki, D.; Ikeshima, D.; Sosa, J.; Bates, P. D.; Allen, G. H.; Pavelsky, T. M. MERIT Hydro: A High-Resolution Global Hydrography Map Based on Latest Topography Dataset. *Water Resources Research* **2019**, *55* (6), 5053–5073. https://doi.org/10.1029/2019WR024873.
77. Amatulli, G.; McInerney, D.; Sethi, T.; Strobl, P.; Domisch, S. Geomorpho90m, Empirical Evaluation and Accuracy Assessment of Global High-Resolution Geomorphometric Layers. *Sci Data* **2020**, *7* (1), 162. https://doi.org/10.1038/s41597-020-0479-6.