Machine Learning and Multi-source Remote Sensing in Forest Aboveground Biomass Estimation: A Review

Autumn Nguyen
Mount Holyoke College
Computer Science
South Hadley, MA, United States
autumn.yngoc@gmail.com

Sulagna Saha
Mount Holyoke College
Computer Science
South Hadley, MA, United States
saha23s@mtholyoke.edu

Abstract

Quantifying forest aboveground biomass (AGB) is crucial for informing decisions and policies that will protect the planet. Machine learning (ML) and remote sensing (RS) techniques have been used to do this task more effectively, yet there lacks a systematic review on the most recent working combinations of ML methods and multiple RS sources, especially with the consideration of forest-specific variables. This study systematically analyzed 25 papers that met strict inclusion criteria from over 80 related studies, identifying all ML methods and combinations of RS data used. Random Forest had the most frequent appearance (88% of studies), while Extreme Gradient Boosting showed superior performance in 75% of the studies in which it was compared with other methods. Sentinel-1 emerged as the most utilized remote sensing source, with multi-sensor approaches (e.g., Sentinel-1, Sentinel-2, and LiDAR) proving especially effective. Our findings provide grounds for recommending which sensing sources, variables, and methods to consider using when integrating ML and RS for forest AGB estimation.

1 Introduction

The main driver of the increasing global deforestation is because forests are mainly valued in terms of their economic value, such as how much timber or area of land they can provide, rather than on how much they help regulate climate [27]. To quantify how keeping certain areas of forests can help the climate, we need to quantify their carbon stock. The best way to measure the amount of carbon sequestered in a forest is to do so with direct field measurements [5], but since manually collecting field measurements at large scale is too costly, RS has been utilized. RS can fly over large areas of forests to capture information, such as tree density, vegetation cover, and 3D structures, with minimal disturbance, and these data can be put into forestry allometric equations to calculate AGB. Combining multiple sources remote sensing is promising because it allows the capabilities of one source to compensate for the limitations of the other sources, as we summarized in Table 1. There have been studies aiming to summarize existing work in this area. Ouaknine et al. [21] provided the comprehensive list of open forest monitoring datasets. Sun and Liu [29] reviewed fundamental estimation methods, but only for studies in China, and they did not review any ML methods. Rolnick et al. [24] provided a broad overview of ML in climate change, but with limited focus on forest carbon. Hamedianfar et al. [10] detailed deep learning methods for forest inventory, but no common non-DL methods like RF; plus this highly technical approach may be inaccessible to a broad audience. Matiza et al. [18] reviewed ML and RS approaches for carbon storage, but they did not analyze the specific combinations of data sources or forest characteristics. Our study addresses those gaps by reviewing studies done around the world, focusing on forest AGB estimation with ML and combinations of RS sources (i.e. multi-source RS), and communicating the results in an accessible way to people who may not have deep expertise in those areas.

2 Methods

The papers in our review were drawn from these **search terms**: "(estimation OR estimating OR "machine learning") (multisource OR multi-source OR multisensor OR multi-sensor) forest carbon biomass map". The aim was to find papers that surely had a Machine Learning component, used a combination of different sources of remote sensing data, and for the purpose of mapping carbon stick or biomass of forests. We retrieved the papers from this search into a database using the public API from [1], specifically the google_scholar_internal function. The 25 papers that we drew quantitative results from were the first 25 that satisfied five of our **inclusion criteria**: (1) full paper accessible to us (so most are open access articles, since we are college students with almost no subscriptions to any journals), (2) used ML in the study, (3) used multiple sources of remote sensing data, (4) had the end goal being estimation forest carbon, whether it was AGB or BGB or soil carbon, and (5) was written in the recent 10 years (2014-2024).

The following **data** was collected for the **quantitative database**: 1) All the remote sensing sources that the study took data from. 2) All the ML methods used in the study. 3) If the study used multiple ML methods for comparison of performance on the same task, or for each method to be used on a different task. 4) If the study used multiple ML methods for comparison, which method(s) were found to have the best performance? Since there are many different ways to define "best", we just included the methods that were explicitly mentioned in the abstract or conclusion with a keyword "best", or "highest" for metrics like R^2 for accuracy, or "lowest" for metrics like RMSE or uncertainty. 5) Any limitations or future steps thoroughly explained. 6) The ultimate task, such as AGB map, BGB map, general biomass map, multi-scale biomass maps, uncertainty estimation, etc. 7) The location(s) of the studied forests. 8) The types or characteristics or dominant species of the forests. The forest type categories we used in our review are not at all mutually exclusive or deterministic — they are meant to facilitate readers in identifying the papers that work on forests of similar types to their interest. Since the words people used to describe their forests varied widely between papers, we did our best to identify the common terms used across papers, and refer to a few sources ([33, 2]) to determine the forest types of the papers that did not use exactly those common terms (so we related their terms to the common terms), and of the papers that did not have any terms about the type of their forest (for those, we used the geographical latitude and longitude of the area to determine the type based on the external sources cited above). 9) The scale of the study: region, country, or global. Python libraries, namely Pandas, Matplotlib, Seaborn, and NumPy, were used to manipulate and visualize data.

3 Results & Discussions

We created a interactive database[1] which everyone can filter by forest types, data sources, ML methods, or any keywords they want to find relevant papers that we reviewed. A summary table and the abbreviations of ML methods and data sources can be found in the Appendix.

Most of the ML methods had roughly similar appearance frequency (see Appendix), except for Random Forest, which was used in around 88% of the studies – as the model for the end task—AGB estimation and sometimes as the model for other intermediate tasks in the data processing pipeline. We put the methods into groups in Figure 1 to see the trend in a bigger picture. Random Forests are still the most commonly used methods, and most of the methods found to have the best performance fall into the three most frequently used groups: Random Forest (RF, QRF, RRF), Gradient Boosting (XGB, LGBM, CatBoost), and Neural Network (CNN, BayesResNN). 11 out of 25 studies compared multiple ML methods for the AGB estimation task, and found the model(s) that performed best. RF was part of all the studies that compared multiple MLs, but was only found to be best in 4, whereas XGB was only used in 4 studies, but was found to perform the best in 3. For instance, [17] found that XGB had best estimations of AGB in high and low range values, while XGB, RF, LR performed similarly in medium range, so XGB also improved the overestimation-underestimation issue.

The most frequently used data source was Sentinel-1, followed by Sentinel-2, ALOS-PALSAR, Landsat, MODIS, GLAS/ICESat LiDAR, and GEDI LiDAR. However, since GLAS/ICESat and GEDI are both spaceborne LiDAR, we can also say that spaceborne LiDARs were the most frequently

[1] `https://itsautumn.notion.site/10a0d405e6518047b073ddd00c71dc65?v=12b0d405e6518078be9b000cbf6ccde5&pvs=4`

used source. In 2022, Sentinel-1 only came at the 7th, ALOS-PALSAR the 13th, while Landsat sat at the top of the frequency rank of sensors in [18].

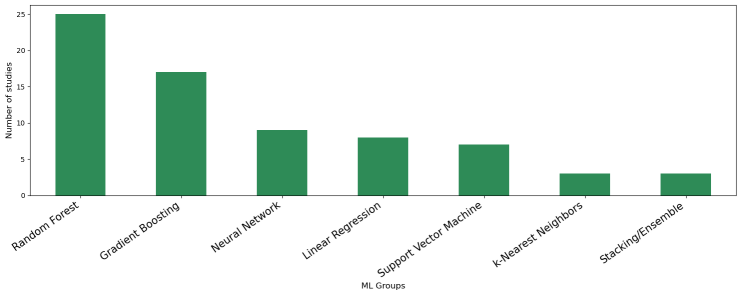

Figure 1: Frequency of ML methods by Groups

The heatmap 2 shows that Sentinel-1, Sentinel-2, and spaceborne LiDAR (GEDI) were most often used together. They made a combination of passive optical, active optical, and radar, complementing each other's strengths and limitations. LiDAR had very limited availability and high costs, so when it was not available, combinations of passive optical and radar also used well together quite often. For example, Landsat and Sentinel-1, or MODIS and Sentinel-2, or MODIS and PALSAR. Nonetheless, some of those studies that used only passive optical and radar sensors faced a common issue of saturation, and LiDAR was usually the recommended solution to that issue. Another observation was that when a study used Sentinel-2, they'd likely also include LiDAR or DEM. This is likely because Sentinel-2 is a passive optical sensor with no ability to infer canopy height, an important variable in estimating AGB, and LiDAR or DEM can provide that information.

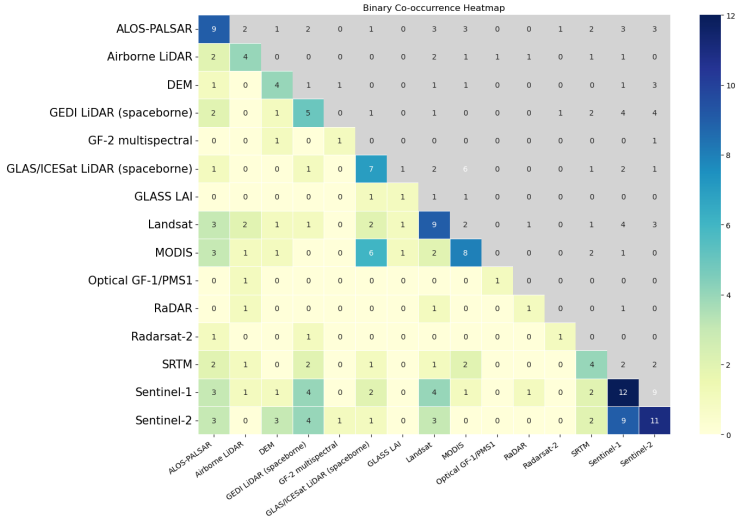

Figure 2: Heatmap: binary co-occurrence of two data sources

Feature selection was found to be a critical factor influencing the models' performances in many studies. A key finding from the literature is that a well-executed variable selection process can significantly enhance model accuracy. For instance, Li et al. [16] demonstrated that feature selection substantially improved the performance of all their models, with XGBoost showing the most significant gains. Similarly, Huang et al. [13] used the Least Absolute Shrinkage and Selection Operator (Lasso) to reduce over 30 initial numerical parameters to just seven. However, there is a contrasting perspective in the literature. Some studies, such as that by Li et al. [17], opted against feature selection. Their decision was based on two main arguments. First, their datasets had a limited number of variables, and they were concerned that multiple rounds of selection would eliminate crucial data, particularly from Sentinel sensor variables. Second, they argued that their chosen models, specifically

Random Forest (RF) and XGBoost, were inherently robust against noisy variables. They noted that RF is generally unaffected by noisy predictors, while XGBoost's regularization objective helps to restrain their influence. Since this was somewhat contrasting with how many studies using RF and XGB did perform feature selection and noted the improved performance, it would be an interesting topic to look into the relationship between feature selection and choice of ML models.

3D structural data from Shuttle Radar Topography Mission or other digital elevation models seems to be frequently used in the studies that didn't have LiDAR data, when it was used, it was usually one of the the most important predictor variables. [14] had Landsat OLI and Sentinel-2 as the two main remote sensors, with the addition of DEM data, and found that DEM was the most important variable. [13] also had optical sensors and radar sensor as Sentinel-2B, Sentinel-2A, Sentinel 1A, with the inclusion of DEM data.

There are also **variables not from remote sensors** that were found to be critical as well. **Phenological characteristics**, the seasonal patterns and timing of biological events of different forest types and different tree species, was found to be a valuable one. When [36] inputed phenological variables into their model, they achieved a higher R-squared result. Although their study area had a specific dominant forest type and dominant species, an implication of their work was that incorporating data about phenological characteristics and dominant species significantly improved the accuracy of AGB estimation. Phenological characteristics were also used to help extracting the distribution information of their study subject (larch trees) in [11]. In addition, the close relationship between phenological data and **time** were a major advantage for AGB estimations. Forests' carbon flux varies over seasons, but the commonly used spectral variables from optical sensors only reflected the state of the forests at one point in time. Therefore, the AGB estimation based solely on those variables couldn't be scaled temporally [36]. With phenological variables in play, we can create what [36] called time-consistent AGB models. In [11]'s study, their LSTM model did well in the last stage of their study pipeline, which was extrapolating biomass components at the regional scale. It is reasonable because LSTM is a type of recurrent neural networks that is specialized in working with time-series and sequential data, and time is an important indicator in the phenological data they used. When [11] compared LSTM with RF for this task, they found that the LSTM model was less prone to underestimation of biomass, and this characteristic became more obvious when the sample unit biomass was increased.

Another insight was that **different ML algorithms may be suitable for different stages of a forest carbon mapping pipeline** In mapping AGB in alpine regions of Yunnan, [36] used three different ML methods through their pipeline: logistic regression to extract phenological parameters from Landsat and work with MCCDC; SVM to take in in phenological parameters and classify forest dominant tree groups; and RF to take in forest dominant tree groups mapping and create AGB map for the region. [11] compared RF and MLR for creating Plot-Scale Biomass Component Estimation Model; used SVM for the extraction of Larch Distribution Information on the Basis of Vegetation Phenology Characteristics; and compared RF and LSTM for the extrapolation of Biomass Components at the Regional Scale. [32] used an optimized RF regressor to calculate early estimates of carbon storage at the canopy scale in the footprints of ICESat-2/ATLAS LiDAR data; and used a deep neural network to create regional-scale carbon storage maps from those early LiDAR estimates and Landsat data.

4 Conclusion & Future Steps

This review highlights the machine learning methods and the remote sensing sources and combinations with the highest usage frequency and performance. Our recommendation for future studies on estimating forest AGB is to, in terms of remote sensing data sources, combine multiple sources of remote sensing data, at least passive optical and radar optical, such as Sentinel-1 with Sentinel-2 or MODIS with ALOS-PALSAR, to address coverage and saturation limitations. In addition, including data that can indicate the forest's 3D structure, like from DEM or active optical sensors, can enhance accuracy and mitigate the overestimation-underestimation problem. In terms of ML methods, Random Forest is a great baseline method due to its long history of reliability, but it may also be worth trying other methods that had proven success recently, such as Extreme Gradient Boosting or CNN. We also emphasize the importance of feature selection and ensuring the spatial heterogeneity of sample plots to improve model performance. Additionally, rather than just using one ML method, different ML methods can be leveraged at various stages of the data processing pipeline. Future work should also take into account the types, phenological characteristics, and dominant species of the forests in building estimation models.

Acknowledgments and Disclosure of Funding

We are grateful for Professor **Alyx Burns**, who has met with us every week throughout Fall 2024 to advise us on the whole process, from reading papers to compiling the database to visualizing results. We are also deeply thankful for Dr. **Sreedath Panat**, Dr. **David Dao**, and Dr. **Björn Lütjens** who have given us not only advice and insights, but also great encouragement.

No funding was received for this work.

References

[1] Paul Ankit. Github repository, 2021. https://github.com/monk1337/resp.

[2] ArcGIS. Map of wwf ecoregions, 2024. `https://www.arcgis.com/apps/View/index.html?appid=d60ec415febb4874ac5e0960a6a2e448`.

[3] Polyanna da Conceição Bispo, Pedro Rodríguez-Veiga, Barbara Zimbres, Sabrina do Couto de Miranda, Cassio Henrique Giusti Cezare, Sam Fleming, Francesca Baldacchino, Valentin Louis, Dominik Rains, Mariano Garcia, et al. Woody aboveground biomass mapping of the brazilian savanna with a multi-sensor and machine learning approach. *Remote Sensing*, 12(17):2685, 2020.

[4] Lin Chen, Chunying Ren, Guangdao Bao, Bai Zhang, Zongming Wang, Mingyue Liu, Weidong Man, and Jiafu Liu. Improved object-based estimation of forest aboveground biomass by integrating lidar data from gedi and icesat-2 with multi-sensor images in a heterogeneous mountainous region. *Remote Sensing*, 14(12):2743, 2022.

[5] Keith W. Cunningham and Marci N. Montgomery. Remote sensing for the audit and assurance of the carbon market. In *2011 IEEE Global Humanitarian Technology Conference*, pages 114–116, 2011.

[6] Dekker Ehlers, Chao Wang, John Coulston, Yulong Zhang, Tamlin Pavelsky, Elizabeth Frankenberg, Curtis Woodcock, and Conghe Song. Mapping forest aboveground biomass using multisource remotely sensed data. *Remote Sensing*, 14(5):1115, 2022.

[7] Rakesh Fararoda, R Suraj Reddy, Gopalakrishnan Rajashekar, TR Kiran Chand, Chandra Shekhar Jha, and VK Dadhwal. Improving forest above ground biomass estimates over indian forests using multi source data sets with machine learning algorithm. *Ecological Informatics*, 65:101392, 2021.

[8] Sujit M Ghosh, Mukunda D Behera, Subham Kumar, Pulakesh Das, Ambadipudi J Prakash, Prasad K Bhaskaran, Parth S Roy, Saroj K Barik, Chockalingam Jeganathan, Prashant K Srivastava, et al. Predicting the forest canopy height from lidar and multi-sensor data using machine learning over india. *Remote Sensing*, 14(23):5968, 2022.

[9] Sujit Madhab Ghosh and Mukunda Dev Behera. Aboveground biomass estimation using multi-sensor data synergy and machine learning algorithms in a dense tropical forest. *Applied Geography*, 96:29–40, 2018.

[10] Alireza Hamedianfar, Cheikh Mohamedou, Annika Kangas, and Jari Vauhkonen. Deep learning for forest inventory and planning: a critical review on the remote sensing approaches so far and prospects for further applications. *Forestry*, 95(4):451–465, 2022.

[11] Yifeng Hong, Jiaming Xu, Chunyan Wu, Yong Pang, Shougong Zhang, Dongsheng Chen, and Bo Yang. Combining multisource data and machine learning approaches for multiscale estimation of forest biomass. *Forests*, 14(11):2248, 2023.

[12] Tianyu Hu, YingYing Zhang, Yanjun Su, Yi Zheng, Guanghui Lin, and Qinghua Guo. Mapping the global mangrove forest aboveground biomass using multisource remote sensing data. *Remote sensing*, 12(10):1690, 2020.

[13] Huajian Huang, Dasheng Wu, Luming Fang, and Xinyu Zheng. Comparison of multiple machine learning models for estimating the forest growing stock in large-scale forests using multi-source data. *Forests*, 13(9):1471, 2022.

[14] Tianbao Huang, Guanglong Ou, Yong Wu, Xiaoli Zhang, Zihao Liu, Hui Xu, Xiongwei Xu, Zhenghui Wang, and Can Xu. Estimating the aboveground biomass of various forest types with high heterogeneity at the provincial scale based on multi-source data. *Remote Sensing*, 15(14):3550, 2023.

[15] Xinyu Li, Meng Zhang, Jiangping Long, and Hui Lin. A novel method for estimating spatial distribution of forest above-ground biomass based on multispectral fusion data and ensemble learning algorithm. *Remote Sensing*, 13(19):3910, 2021.

[16] Yingchang Li, Chao Li, Mingyang Li, and Zhenzhen Liu. Influence of variable selection and forest type on forest aboveground biomass estimation using machine learning algorithms. *Forests*, 10(12):1073, 2019.

[17] Yingchang Li, Mingyang Li, Chao Li, and Zhenzhen Liu. Forest aboveground biomass estimation using landsat 8 and sentinel-1a data with machine learning algorithms. *Scientific reports*, 10(1):9952, 2020.

[18] Collins Matiza, Onisimo Mutanga, Kabir Peerbhay, John Odindi, and Romano Lottering. A systematic review of remote sensing and machine learning approaches for accurate carbon storage estimation in natural forests. *Southern Forests: a Journal of Forest Science*, 85(3-4):123–141, 2023.

[19] Mohamed Musthafa and Gulab Singh. Improving forest above-ground biomass retrieval using multi-sensor l-and c-band sar data and multi-temporal spaceborne lidar data. *Frontiers in Forests and Global Change*, 5:822704, 2022.

[20] Jean Pierre Ometto, Eric Bastos Gorgens, Francisca Rocha de Souza Pereira, Luciane Sato, Mauro Lúcio Rodrigures de Assis, Roberta Cantinho, Marcos Longo, Aline Daniele Jacon, and Michael Keller. A biomass map of the brazilian amazon from multisource remote sensing. *Scientific Data*, 10(1):668, 2023.

[21] Arthur Ouaknine, Teja Kattenborn, Etienne Laliberté, and David Rolnick. Openforest: A data catalogue for machine learning in forest monitoring. *arXiv preprint arXiv:2311.00277*, 2024.

[22] Florian Pötzschner, Matthias Baumann, Nestor Ignacio Gasparri, Georgina Conti, Dante Loto, María Piquer-Rodríguez, and Tobias Kuemmerle. Ecoregion-wide, multi-sensor biomass mapping highlights a major underestimation of dry forests carbon stocks. *Remote sensing of environment*, 269:112849, 2022.

[23] Asim Qadeer, Muhammad Shakir, Li Wang, and Syed Muhammad Talha. Evaluating machine learning approaches for aboveground biomass prediction in fragmented high-elevated forests using multi-sensor satellite data. *Remote Sensing Applications: Society and Environment*, 36:101291, 2024.

[24] David Rolnick, Priya L Donti, Lynn H Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, et al. Tackling climate change with machine learning. *ACM Computing Surveys (CSUR)*, 55(2):1–96, 2022.

[25] Ghasem Ronoud, Parviz Fatehi, Ali A Darvishsefat, Erkki Tomppo, Jaan Praks, and Michael E Schaepman. Multi-sensor aboveground biomass estimation in the broadleaved hyrcanian forest of iran. *Canadian journal of remote sensing*, 47(6):818–834, 2021.

[26] Faseela V Sainuddin, Guljar Malek, Ankur Rajwadi, Padamnabhi S Nagar, Smitha V Asok, and C Sudhakar Reddy. Estimating above-ground biomass of the regional forest landscape of northern western ghats using machine learning algorithms and multi-sensor remote sensing data. *Journal of the Indian Society of Remote Sensing*, pages 1–18, 2024.

[27] Frances Seymour and Jonah Busch. Forests for growth—invisible contributions. *Finance & Development*, 54(1), 2017.

[28] RK Singh, CM Biradar, Mukunda Dev Behera, A Jaya Prakash, P Das, MR Mohanta, G Krishna, A Dogra, SK Dhyani, and J Rizvi. Optimising carbon fixation through agroforestry: Estimation of aboveground biomass using multi-sensor data synergy and machine learning. *Ecological Informatics*, 79:102408, 2024.

[29] Wanlong Sun and Xuehua Liu. Review on carbon storage estimation of forest ecosystem and applications in china. *Forest Ecosystems*, 7:1–14, 2020.

[30] Zhi Tang, Xiaosheng Xia, Yonghua Huang, Yan Lu, and Zhongyang Guo. Estimation of national forest aboveground biomass from multi-source remotely sensed dataset with machine learning algorithms in china. *Remote Sensing*, 14(21):5487, 2022.

[31] Xiaoyi Wang, Caixia Liu, Guanting Lv, Jinfeng Xu, and Guishan Cui. Integrating multi-source remote sensing to assess forest aboveground biomass in the khingan mountains of north-eastern china using machine-learning algorithms. *Remote Sensing*, 14(4):1039, 2022.

[32] Lei Xi, Qingtai Shu, Yang Sun, Jinjun Huang, and Hanyue Song. Carbon storage estimation of mountain forests based on deep learning and multisource remote sensing data. *Journal of Applied Remote Sensing*, 17(1):014510–014510, 2023.

[33] Chen Xu, Xianliang Zhang, Rocío Hernandez-Clemente, Wei Lu, and Rubén D Manzanedo. Global forest types based on climatic and vegetation data. *Sustainability*, 14(2):634, 2022.

[34] Fanyi Zhang, Xin Tian, Haibo Zhang, and Mi Jiang. Estimation of aboveground carbon density of forests using deep learning and multisource remote sensing. *Remote Sensing*, 14(13):3022, 2022.

[35] Rong Zhang, Xuhui Zhou, Zutao Ouyang, Valerio Avitabile, Jiaguo Qi, Jiquan Chen, and Vincenzo Giannico. Estimating aboveground biomass in subtropical forests of china by integrating multisource remote sensing and ground data. *Remote Sensing of Environment*, 232:111341, 2019.

[36] Yali Zhang, Ni Wang, Yuliang Wang, and Mingshi Li. A new strategy for improving the accuracy of forest aboveground biomass estimates in an alpine region based on multi-source remote sensing. *GIScience & Remote Sensing*, 60(1):2163574, 2023.

[37] Yuzhen Zhang, Jun Ma, Shunlin Liang, Xisheng Li, and Manyao Li. An evaluation of eight machine learning regression algorithms for forest aboveground biomass estimation from multiple satellite data products. *Remote sensing*, 12(24):4015, 2020.

[38] Yan Zhu, Zhongke Feng, Jing Lu, and Jincheng Liu. Estimation of forest biomass in beijing (china) using multisource remote sensing and forest inventory data. *Forests*, 11(2):163, 2020.

5 Appendix

	Passive optical	Active optical (LiDAR)	Radar (SAR)
Sensors	Sentinel-2 MODIS Landsat	GLAS (ICESat)	Sentinel-1 ALOS-PALSAR
Capabilities	Widely and freely available High resolution	Able to measure 3D structural data Can work at nighttime High resolution	Can penetrate dense canopies and clouds Can work at nighttime
Limitations	Cannot work at nighttime Cannot penetrate dense canopies or clouds	Expensive Available in only certain small areas Cannot penetrate clouds	Lower resolution than optical Saturation issues

Table 1: Capabilities and limitations of most common categories of remote sensor modalities

5.1 Abbreviations of ML methods

- Random Forest (RF)
- Quantile Random Forest (QRF)
- Regularized Random Forest (RRF)
- Extremely Randomized Trees (ERT)
- Gradient Tree Boosting (GTB)
- Gradient-Boosted Regression Tree (GBRT)
- Boosted Regression Tree (BRT)
- Gradient Boosting Machine (GBM)
- Light Gradient Boosting Machine (LGBM)
- Stochastic Gradient Boosting (SGB)
- Extreme Gradient Boosting (XGB)
- Categorical Boosting (CatBoost)
- Linear Regression (LR)
- Multi-Linear Regression (MLR)
- Stepwise Linear Regression (StepwiseLR)
- Multivariate adaptive regression splines (MARS)
- Random Forest with Stacking Algorithm (RFStacking)
- Cubist Regression Tree Ensemble (CubistRTEns)
- Stacked Ensemble for RF and boosting algorithms
- Bayesian Regularization Neural Network (BayesRegNN)

5.2 Groupings of ML methods

The ML methods were grouped as follows:

- 'Random Forest': ['RF', 'RRF', 'QRF', 'ERT'],
- 'Gradient Boosting': ['GTB', 'GBM', 'GBRT', 'BRT', 'LGBM', 'SGB', 'XGB', 'CatBoost'],
- 'Linear Regression': ['MLR', 'LR', 'Stepwise LR', 'MARS'],

- 'Neural Networks': ['LSTM', 'QRNN', 'CNN', 'ANN', 'BayesRegNN', 'Keras'],
- 'Support Vector Machines': ['SVM', 'SVR'],
- 'Stacking/Ensembles': ['RFStacking', 'StackedEnsemble'],
- 'Cubist': ['CubistRTEns'],
- 'k-NN': ['kNN']

5.3 Summary table

Study	Data sources	ML methods used
[7]	ALOS-PALSAR, DEM, MODIS	RF, kNN
[38]	ALOS-PALSAR, Landsat	MLR, RF
[11]	Airborne LiDAR, Optical GF-1/PMS1	LSTM, MLR, RF, SVM
[31]	ALOS-PALSAR, GLAS/ICESat LiDAR (spaceborne), MODIS	LR, QRNN, RF, SVM, Stepwise LR
[35]	GLAS/ICESat LiDAR (spaceborne), Landsat, MODIS	CubistRegrTree
[34]	ALOS-PALSAR, Sentinel-1, Sentinel-2	CNN, Keras, MLR, RF, SVM
[30]	GLAS/ICESat LiDAR (spaceborne), MODIS, SRTM	CatBoost, GBM, LGBM, RF, XGB
[6]	Airborne LiDAR, Landsat, RaDAR, Sentinel-1	RF
[12]	GLAS/ICESat LiDAR (spaceborne), MODIS	RF
[14]	DEM, Landsat, Sentinel-2	BayesRegNN, GBM, QRF, RF, RRF, kNN
[20]	ALOS-PALSAR, Airborne LiDAR, MODIS, SRTM	RF
[9]	Sentinel-1, Sentinel-2	RF, SGB
[3]	ALOS-PALSAR, Airborne LiDAR, Landsat	RF
[22]	GLAS/ICESat LiDAR (spaceborne), MODIS, Sentinel-1	GBM
[28]	Sentinel-1, Sentinel-2	ANN, RF, SVM
[25]	Landsat, Sentinel-1, Sentinel-2	MLR, RF, SVR, kNN
[23]	DEM, GEDI LiDAR (spaceborne), Sentinel-1, Sentinel-2	CatBoost, GTB, LGBM, RF, XGB
[4]	GEDI LiDAR (spaceborne), GLAS/ICESat LiDAR (spaceborne), Sentinel-1, Sentinel-2	RF
[19]	ALOS-PALSAR, GEDI LiDAR (spaceborne), Radarsat-2	RF
[15]	DEM, GF-2 multispectral, Sentinel-2	RFStacking
[26]	GEDI LiDAR (spaceborne), SRTM, Sentinel-1, Sentinel-2	BRT, RF, XGB
[37]	GLAS/ICESat LiDAR (spaceborne), GLASS LAI, Landsat, MODIS	ANN, CatBoost, ERT, GBRT, MARS, RF, SGB, SVR
[8]	ALOS-PALSAR, GEDI LiDAR (spaceborne), Landsat, SRTM, Sentinel-1, Sentinel-2	RF
[17]	Landsat, Sentinel-1	LR, RF, XGB

Table 2: Data sources and ML methods used in selected studies.

5.4 Visualizations of Quantitative Results

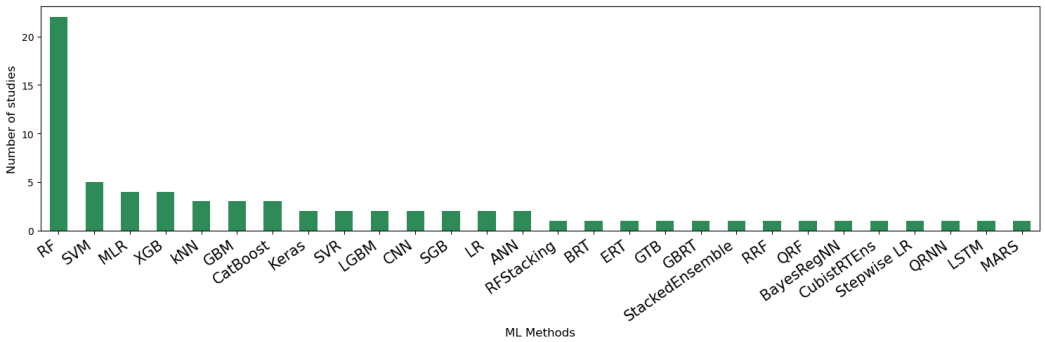

Figure 3: Frequency of ML methods used

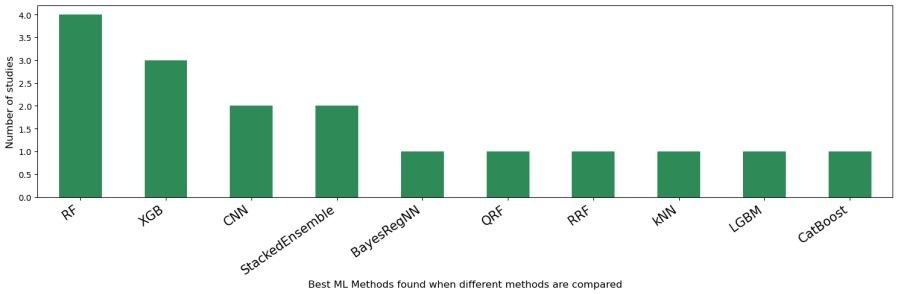

Figure 4: Best ML methods found in studies that compared multiple methods

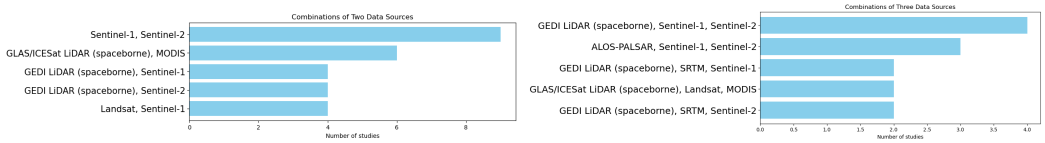

Figure 5: Most common combinations of data sources

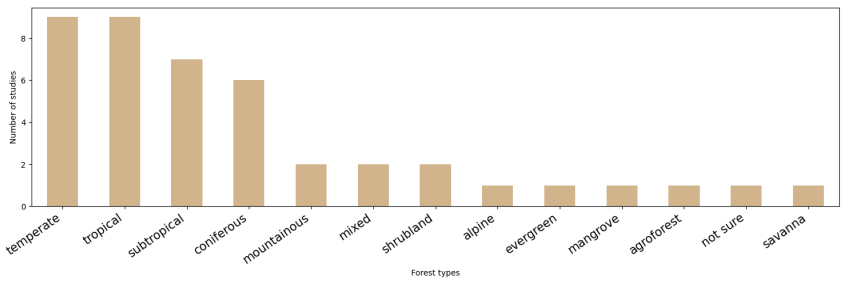

Figure 6: Frequency of forest types studied

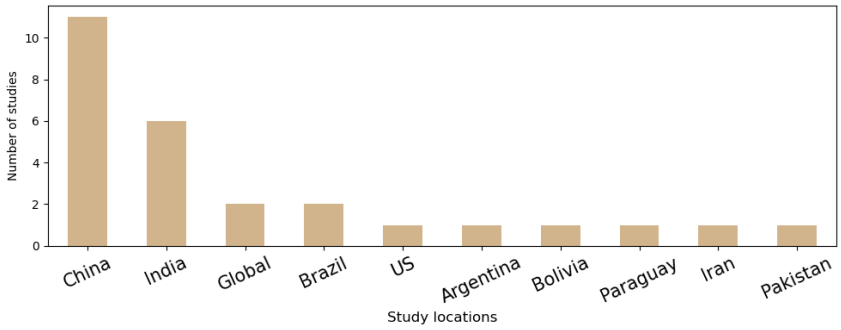

Figure 7: Frequency of geographical locations of the studied forests

China made up around half of the locations of the studies, making it the most frequently appeared country in Figure 7. Since none of the terms were China-related, and all papers were selected from the first-appeared results on Google Scholar rather than from related papers, this may point to some interesting geographical trend of research in the field.