



Dominant Expression of SAR Backscatter in Predicting Aboveground Biomass: Integrating Multi-Sensor Data and Machine Learning in Sikkim Himalaya

A. Jaya Prakash¹ · Sujoy Mudi¹ · Somnath Paramanik¹ · Mukunda Dev Behera¹ · Shanu Shekhar¹ · Narpati Sharma² · Bikash Ranjan Parida³

Received: 20 June 2023 / Accepted: 28 December 2023 / Published online: 3 February 2024
© Indian Society of Remote Sensing 2024

Abstract

Accurate assessment of aboveground biomass (AGB) is crucial for understanding carbon budgets, climate change impacts, and evaluating forest responses to environmental shifts. In this study, AGB was estimated in Sikkim State of India by leveraging the capabilities of machine learning (ML) and integrating multi-sensor satellite data. Specifically, the random forest (RF) and categorical boosting algorithm (CatBoost) models were utilised. Field estimated AGB ranges from 1.99 to 530.02 Mg/ha with an average of 252.58 Mg/ha, utilised for model prediction and validation. The RF model slightly outperformed the CatBoost model, with a coefficient of determination (R^2) of 0.71 and root mean square error (RMSE) of 72.98 Mg/ha, compared to the CatBoost model's R^2 of 0.67 and RMSE of 80.69 Mg/ha. The former showed a greater capacity to combat overfitting. Synthetic aperture radar variables have emerged as significant predictors because of their contribution to the structural properties of plants. This study acknowledges the limitations and challenges due to data availability, especially for ground truth measurements, which pose constraints on the accuracy and representativeness of AGB estimates. Uncertainties associated with AGB estimation, such as variations in vegetation structure and species composition, also affected model performance. Despite these limitations, this study emphasises the significance of multi-sensor data integration and ML models in AGB estimation and highlights their potential applications in forest management and climate change mitigation efforts in the Himalayan mountainous region.

Keywords Forest aboveground biomass · Random forest · Tropical forest · Sentinel-1 and 2 · PALSAR-2

Introduction

To address the growing threat of global warming, the Paris Agreement (Nations 2016) emphasised the urgent need for major reductions in global greenhouse gas emissions to

maintain global temperatures within 2 °C throughout the twenty-first century. This commitment was reinforced at the recent Conference of the Parties (COP) 27 meeting, which brought together world leaders and stakeholders to discuss climate change challenges and set ambitious emission reduction and sustainable development targets. These activities require accurate mapping of the aboveground biomass (AGB). They enable the development and execution of effective forest management plans, contributing to global efforts to reduce climate change. Evaluating forest carbon dynamics necessitates accurate estimation of AGB, especially in tropical forest ecosystems. Tropical forests sequester large amounts of atmospheric carbon dioxide (CO_2) through photosynthesis, recognised as essential carbon sinks (Xiao et al., 2019). The carbon stored in the aboveground components of forests contributes to global carbon cycles, making the accurate estimation of AGB crucial for accurate

✉ Mukunda Dev Behera
mdbehera@coral.iitkgp.ac.in

¹ Spatial Analysis and Modelling Lab, Centre for Ocean, River, Atmosphere and Land Sciences, Indian Institute of Technology Kharagpur, Kharagpur, West Bengal 721302, India

² Sikkim State Council of Science and Technology, Vigyan Bhawan, Department of Science and Technology, Gangtok 737102, Sikkim, India

³ Department of Geoinformatics, School of Natural Resource Management, Central University of Jharkhand, Ranchi 835222, India

carbon accounting and climate change mitigation strategies (Avitabile et al., 2016; Baccini et al., 2019).

In recent years, remote sensing (RS) techniques have emerged as significant tools for AGB estimation research. In particular, the employment of satellite data derived from Sentinel-2 (optical), Sentinel-1 (C-band), and PALSAR-2 (L-band) Synthetic Aperture Radar (SAR) data provides complementary information that improves AGB prediction accuracy in prior investigations (Ghosh & Behera, 2018; Prakash et al., 2022). Optical data, Sentinel-2, capture information about the spectral properties of vegetation and are utilised to estimate vegetation indices (VIs) associated with AGB (Lu et al., 2016; Wang et al., 2019). C-band SAR data, such as Sentinel-1 data, are sensitive to forest structure, including vegetation density and vertical structure (Forkuor et al., 2020). By capturing signals from the forest and lower vegetation layers, L-band SAR data, such as PALSAR-2, penetrate the tree canopy and offer extra signals for AGB estimation (Behera et al., 2016). Integrating multiple datasets enabled overcoming constraints associated with single data sources and enhancing the precision of machine learning (ML) models.

AGB estimation is typically achieved using various variables derived from different sensors. These sensors include LiDAR, optical, and radar sensors. Using these sensors allows the collection of data on forest structure, canopy cover, and other relevant parameters. The study by Laurin et al. (2014a) highlights the significance of optical spectral bands in characterising vegetation properties. The red-edge spectral band, in particular, is sensitive to chlorophyll content and canopy structure, which are known to affect variations in AGB. This finding emphasises the potential of optical data in accurately assessing AGB through its spectral bands. In recent years, researchers have established the potential of VIs derived from spectral bands to estimate AGB accurately (Mutanga et al., 2023). To accomplish this, numerous VIs have been employed as predictors. Among these indices, normalised difference vegetation index (NDVI), enhanced vegetation index (EVI), red-edge vegetation index (REVI), chlorophyll vegetation index (CVI), and soil-adjusted vegetation index (SAVI) are the most frequently employed (Forkuor et al., 2020; Ghosh & Behera, 2018; Jha et al., 2021). Various researchers have employed these indices, and it was found that the estimation of AGB can be improved by utilising a combination of spectral bands and VIs obtained from optical data (Fassnacht et al., 2016).

Similarly, microwave data provide useful information for estimating AGB in areas with dense vegetation and challenging weather conditions (Guerra-Hernández et al., 2022) where optical datasets are unavailable. The Sentinel-1 VV band measures the intensity of radar backscattering signal, providing information about vertical vegetation

structure and allowing for AGB variation estimation. The VH band records interactions between radar signals and vegetation canopy, supplementing the information in the VV band (Ghosh et al., 2018). PALSAR-2 HH and HV bands provide information about the forest canopy structure. The HH band measures backscatter intensity to canopy density and vertical structure, whereas the HV band measures vegetation scattering and penetration depth (Yu & Saatchi, 2016).

Numerous studies have explored AGB estimation in tropical forests using various data sources and methodologies (Singh et al., 2022; Vaglio Laurin et al., 2014a, 2014b). Over the past decade, ML algorithms have emerged as powerful tools for the accurate estimation of AGB. ML models, such as random forest (RF) and support vector machine (SVM), are widely employed owing to their ability to handle large and complex datasets. For instance, the RF model utilises an ensemble of decision trees to model the relationships between input features, such as spectral and synthetic aperture radar (SAR) data and targeted values (Ghosh et al., 2021; Prakash et al., 2022). In contrast, SVM aims to find the optimal hyperplane that separates different AGB classes in a high-dimensional feature space (Zhang et al., 2014). These ML algorithms leverage training data to identify underlying patterns and relationships, thereby enabling accurate AGB predictions for new observations (Singh et al., 2023). Furthermore, few studies have been conducted on the categorical boosting technique (CatBoost) for AGB estimation. This is due to its ability to handle categorical variables and its potential for improved performance (Luo et al., 2021). CatBoost employs a boosting technique to train a collection of weak models iteratively and combines their predictions to generate a robust final model.

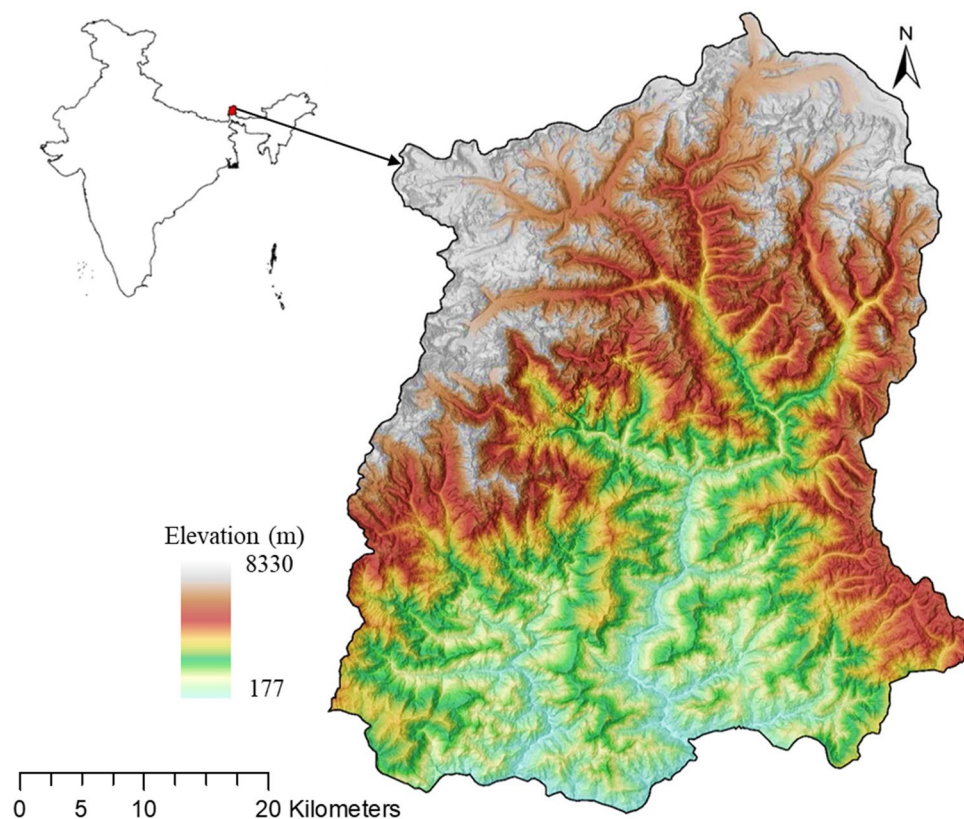
This study uses optical as well as microwave SAR data (C- and L-band) and ML algorithms to estimate AGB in Sikkim Himalaya. No study has employed the collaborative use of optical and SAR data for AGB prediction, particularly in the Sikkim Himalaya region, at a finer spatial resolution. This study aims to develop precise and reliable models for estimating AGB through data synergy and evaluate AGB's predictability using RF and CatBoost ML algorithms where the number of field samples is limited.

Material and Methods

Study Area Description

Sikkim is in the Himalayan range of north-eastern India (Fig. 1). Spanning an area of 7096 km², it is the second-smallest state in India. Sikkim is located between latitudes 27°5' and 28°9' north and longitudes 87°9' and 88°56' east. The elevation in the region varies from 300 to over 8586 m

Fig. 1 Study area depicts the geographical extent in Sikkim State with SRTM elevation map



above the mean sea level. It experiences mean precipitation between 2700 and 3200 mm, while the average yearly temperature fluctuates from 28 °C in the summer to sub-zero temperatures in the winter (FSI, 2019). Forests cover a significant portion of Sikkim, accounting for 47.08% of the total area, with very dense forests occupying 15.53% of the land. Open forests encompass 9.69%, and moderately dense forests cover 21.86% (FSI, 2021).

Field Inventory Data

Field sampling was conducted to ensure representation across different parts (north, south, east, west, and central) of Sikkim by accommodating 49 elementary sampling units (ESUs). In our study, 49 sample plots were established based on statistical considerations and research objectives. Of these, 60% (31 plots) were allocated for model training, whereas the remaining 40% (18 plots) were dedicated to model testing and validation. This partitioning strategy was chosen to ensure a robust assessment of the model's generalisation capabilities. To obtain accurate field inventory data, each ESU was measured at 0.04 hectares (20 m × 20 m). Within these plots, the circumference at breast height (CBH) of each tree species was measured using measuring tape, while individual tree height was determined using a Laser Range Finder. The sampling process employed stratified random sampling, with plots selected based

on criteria such as canopy composition, density, forest type, slope, and accessibility (Sharma et al., 2019). This approach allows for a comprehensive representation of forest ecosystems. Four distinct forest types were sampled, namely Tropical moist deciduous (TMD), Tropical semi-evergreen (TSE), Subtropical evergreen (STE), and Temperate evergreen (TE). By incorporating these field points or sample plots, this study aimed to capture the variability in AGB across different forest types. To compute tree volume within each plot, regional and species-specific volume equations (FSI, 1996) were used for *Shorea robusta*, *Tectona grandis*, *Schima wallichii*, and *Castanopsis indica*, while miscellaneous volume equations were employed for other species (Eqs. 1–5). Using wood-specific gravity values, the stand volumes were converted to AGB.

$$\text{Shorea robusta} = V/D^2 = -0.32546/D^2 + 9.78645 \quad (1)$$

$$\text{Tectona grandis} = V/D = 0.0341/D - 0.65623 + 7.881D \quad (2)$$

$$\text{Schima wallichii} = 0.28069 + 4.6198D - 1.65381 \log D \quad (3)$$

$$\text{Castanopsis indica} = V/D^2 = 0.001184 + 0.1812/D^2 - 0.02348/D \quad (4)$$

$$\text{Mixed species} = V = 0.3555 - 0.037D + 0.001259D^2 \quad (5)$$

where D = trunk diameter (in cm), V = volume (m^3) under bark, and ρ = density (in g/cm^3).

Satellite Data and Pre-Processing

The Sentinel-2 satellite, equipped with the advanced multispectral Instrument (MSI), facilitates the capture of high-resolution images worldwide. Operating within the visible to shortwave infrared range, it provides useful data for our research. To ensure the utmost accuracy, we exclusively utilised cloud-free Sentinel-2 Level-2A products, which offer meticulously atmospherically corrected surface reflectance bands. These bands are accessible at spatial resolutions of 10 and 20 m, enabling us to scrutinise intricate details with precision. Harnessing the potential of these images, we derived several vegetation indices (VIs), encompassing NDVI, SAVI, EVI, modified soil-adjusted vegetation index (MSAVI), difference vegetation index (DVI), ratio vegetation index (RVI), atmospherically resistant vegetation index (ARVI), and modified simple ratio (MSR).

Furthermore, the Sentinel-1 satellite complemented our study by providing dual-polarisation C-band SAR data,

which included vertical–vertical (VV) and vertical–horizontal (VH) polarisations, with a fine spatial resolution of 10 m. Leveraging the VH and VV images, we conducted mathematical operations to generate a ratio image and the square root of their product image. To ensure consistency, both the Sentinel-2 and Sentinel-1 datasets underwent processing in Google Earth Engine (GEE), incorporating resampling techniques to achieve a uniform 20×20 m resolution. We meticulously synchronised the acquisition dates of the images with the field observation time, ensuring accurate temporal alignment. These comprehensive datasets served as the foundation for subsequent processing and modelling of AGB, as illustrated in Fig. 2.

In addition, our study made use of the yearly mosaic ALOS-2/PALSAR-2 L-band data, which provided valuable insights. The dataset consisted of co-polarised horizontal–horizontal (HH) and cross-polarised horizontal–vertical (HV) waves at a spatial resolution of 25 m. Initially, the polarised data were stored as 16-bit digital numbers (DN). To enhance the precision and accuracy of our analysis, we converted the data to backscatter gamma-naught (γ^0) values expressed in decibels (dB). To perform this conversion, we leveraged Eq. (6) within the GEE platform, utilising the available HH and HV bands presented by Shimada et al. (2014). This transformation allowed us to derive more meaningful and interpretable information

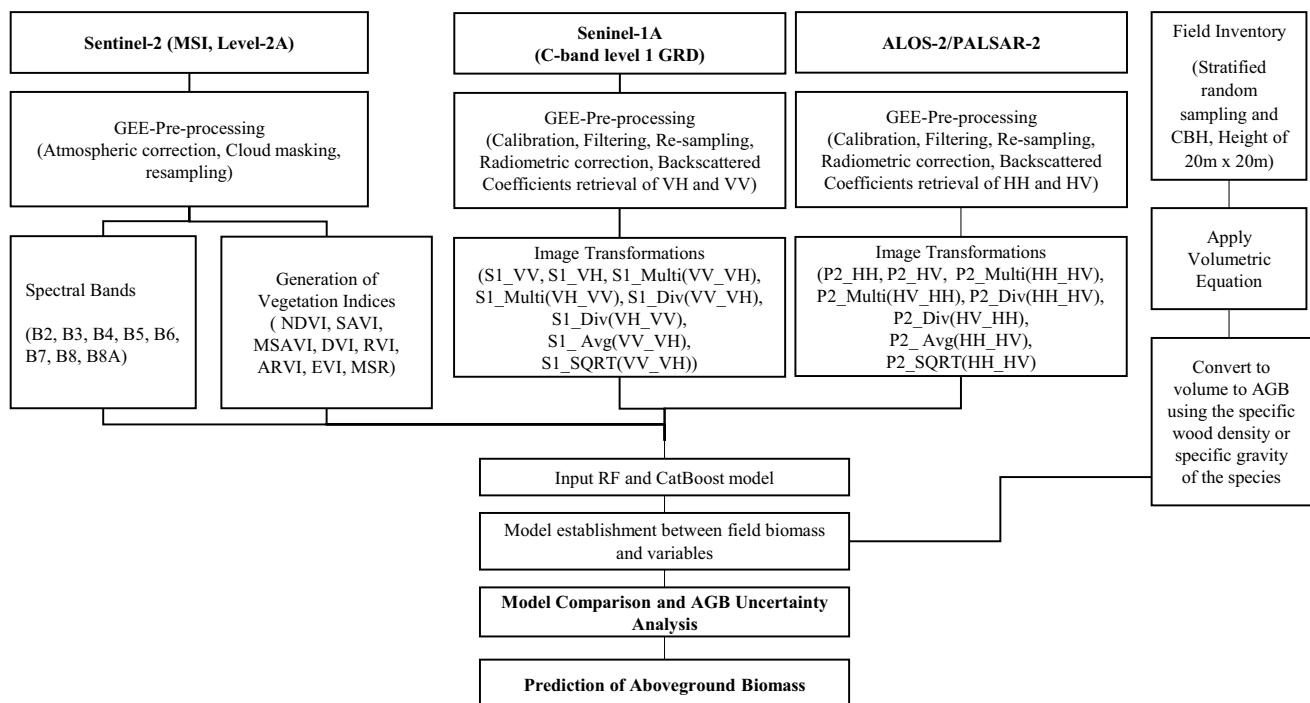


Fig. 2 The methodology overview highlights integrating Sentinel-2, Sentinel-1, and PALSAR-2 data with machine learning models (RF and CatBoost) for estimating aboveground biomass

from the PALSAR-2 data, enabling a comprehensive analysis of our study.

$$\sigma^0 = 10 * \log_{10} (DN^2) + CF \quad (6)$$

where σ^0 (sigma naught) is the backscattering coefficient in dB, DN is a digital number (raw pixel value), and CF is the calibration factor in dB (−83).

Machine Learning Models

In AGB estimation research, RF is most commonly and frequently used (Belgiu & Drăgu, 2016). This ensemble learning approach integrates numerous decision trees to generate a robust predictive model. Each tree in the ensemble was trained on a different subset of the data using a random selection of features. Subsequently, the final prediction is generated by adding the predictions of each tree (Breiman, 2001). The significance of RF lies in its capacity to handle large and complex datasets, capture nonlinear relationships, and accurately model variable interactions. It excels at processing high-dimensional RS data, such as spectral and SAR data, which are frequently used in AGB estimation. The intricate interdependencies between these variables and the target variable AGB can be effectively captured by RF, resulting in accurate predictions. Additionally, RF is less susceptible to overfitting than other ML algorithms. The ensemble nature of RF and the use of random subsets of data for each tree mitigate the risk of overfitting and improve the generalisation performance of the model.

CatBoost, boosting algorithm family, builds a robust predictive model by combining a group of weak learners, typically decision trees (Dorogush et al., 2018). In recent years, the CatBoost algorithm has emerged as a promising approach for gradient boosting in ML. A critical advantage of CatBoost is its ability to handle categorical variables without explicit data preprocessing. This feature makes CatBoost particularly well-suited for datasets that contain a mix of numerical and categorical features, which are commonly encountered in studies involving the estimation of AGB. Unlike traditional gradient boosting methods, CatBoost's ability to handle categorical variables allows for more efficient and accurate AGB modelling, a critical parameter in many ecological and environmental studies. The CatBoost algorithm was developed to address feature interactions and selection effectively, enabling it to capture intricate relationships between variables and the AGB target variable. This innovative algorithm was designed to automatically identify and select the most relevant features, thereby enhancing its ability to capture complex relationships between variables and the target variable. The CatBoost algorithm is an ML technique that utilises

advanced methods, including ordered boosting, to enhance the model's capacity to generalise effectively on novel data and mitigate overfitting. This study highlights the potential of CatBoost as a valuable ML tool. Specifically, the ability of the algorithm to effectively handle categorical variables, its feature interaction capabilities, and its robustness to overfitting were identified as key strengths (Hancock & Khoshgoftaar, 2020).

Selection of Predictor Variables

To assess the relationship between variables, we performed correlation analysis and feature importance assessment using the mean decrease impurity technique. For Sentinel-1 data, we examined variables such as backscatter intensity from the VV and VH bands and their mathematical transformations. Similarly, for Sentinel-2, we included predictor variables derived from spectral bands and Vis, such as NDVI, SAVI, MSAVI, DVI, RVI, ARVI, EVI, and MSR. Additionally, for PALSAR-2, we selected backscatter HH and HV bands and their transformations. Using correlation analysis, we identified the correlation between predictor variables and targeted AGB. Variables with higher absolute correlation coefficients were considered potentially more influential in estimating AGB (Table 1).

To further evaluate the importance of these variables, we employed the mean decrease impurity technique within an ML algorithm, such as RF and CatBoost. This technique quantifies the contribution of each predictor variable in reducing impurity during the construction of decision trees within the ensemble. The mean decrease in impurity score measures the extent to which the model's prediction accuracy decreases when a particular variable is randomly permuted. Higher mean decrease impurity scores indicate greater importance of a variable in predicting AGB. By calculating the mean decrease impurity scores for each predictor variable across multiple iterations, we obtained a ranking of the variables based on their importance in AGB estimation. This allowed us to identify the most influential variables that contribute significantly to the prediction of AGB using Sentinel-1, Sentinel-2, and PALSAR-2 data (Fig. 3a and b).

Evaluation of Models

We employed the train–test split method and K-fold cross-validation to evaluate the performance of the ML model. The dataset was divided into training and testing subsets, with the K-fold approach allowing us to overcome overfitting. Utilising the caret package in R (Kuhn, 2008), we conducted tenfold cross-validation, training on 60% of the data and validating on 40%. Performance metrics such as root mean square error (RMSE), coefficient of determination

Table 1 For AGB estimation models, a summary of predictor variables, including Sentinel-2 spectral, vegetation indices, Sentinel-1 backscatter, and PALSAR-2 backscatter

Variable type	Variable Name	Number of variables	Description
Sentinel-2 Spectral bands	B2, B3, B4, B5, B6, B7, B8, B8A	8	Blue, Green, Red, Red Edge 1, Red Edge 2, Red Edge 3, NIR, Red Edge 4
Sentinel-2 Vegetation indices	S2_NDVI S2_SAVI S2_DVI S2_RVI S2_ARVI S2_EVI S2_MSAVI S2_MSR	8	Normalised difference vegetation index (NDVI) Soil adjusted vegetation index (SAVI) Modified soil-adjusted vegetation index (MSAVI) Difference vegetation index (DVI) Ratio vegetation index (RVI) Atmospherically resistant vegetation index (ARVI) Enhanced vegetation index (EVI) Modified simple ratio (MSR)
Sentinel-1 Backscatter	S1_VV, S1_VH, S1_Multi(VV_VH), S1_Multi(VH_VV), S1_Div(VV_VH), S1_Avg(VV_VH), S1_SQRT(VV_VH)	8	Sentinel-1A dual polarization backscatter their mathematical manipulation like Multiplication, Average, Division, Square root
PALSAR-2 Backscatter	P2_HH, P2_HV, P2_Multi(HH_HV), P2_Multi(HV_HH), P2_Div(HH_HV), P2_Div(HV_HH), P2_Avg(HH_HV), P2_SQRT(HH_HV)	8	PALSAR-2 dual polarization backscatter and their mathematical manipulation like multiplication, average, division, square root

(R^2), mean absolute error (MAE), and bias were employed to assess the models. The RMSE measures the average difference between the predicted and observed AGB values, with lower values indicating better performance. R^2 measures the proportion of variance explained by the model, with higher values indicating a stronger relationship. MAE calculated the average absolute difference, and bias indicated the average difference between predicted and observed values. Through these evaluations, we obtained comprehensive insights into the effectiveness of the models without sacrificing their accuracy or clarity.

$$R^2 = 1 - \frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{\sum_{i=1}^n (\hat{y}_i - \bar{y}_i)^2} \quad (7)$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n}} \quad (8)$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |(\hat{y}_i - y_i)| \quad (9)$$

$$\text{Bias} = \frac{\sum_{i=1}^n (y_i - x_i)}{n} \quad (10)$$

Results

Field-Inventory Analysis

Field measurements and estimated AGB data ranging from 1.99 to 530.62 Mg/ha were employed for model construction and validation, with an average AGB of 224.58 Mg/ha (Fig. 4b). The diverse tree species included *Castanopsis indica*, *mixed species*, *Schima wallichii*, *Shorea robusta*, *Symplocos taurina*, and *Tectona grandis* (Fig. 4c). The diameter at breast height (DBH) characteristics of these species were determined through a comprehensive analysis. The maximum DBH recorded for *Castanopsis indica* was 148.01 cm, the minimum DBH was 27.06 cm, and the average DBH was 63.71 cm. Similarly, the mixed species had a maximum DBH of 157.56 cm, a minimum DBH of 5.41 cm, and an average DBH of 29.32 cm. The maximum DBH of *Schima wallichii* was 84.03 cm, the minimum was 5.09 cm, and the mean was 26.40 cm. The maximum DBH of *Shorea robusta* was 111.41 cm, the minimum was 5.67 cm, and the mean was 25.49 cm. The maximum DBH of *Symplocos taurina* was 21.33 cm, the minimum was 4.77 cm, and the mean was 10.49 cm. The average DBH of *Tectona grandis* was 30.03 cm (Fig. 4a).

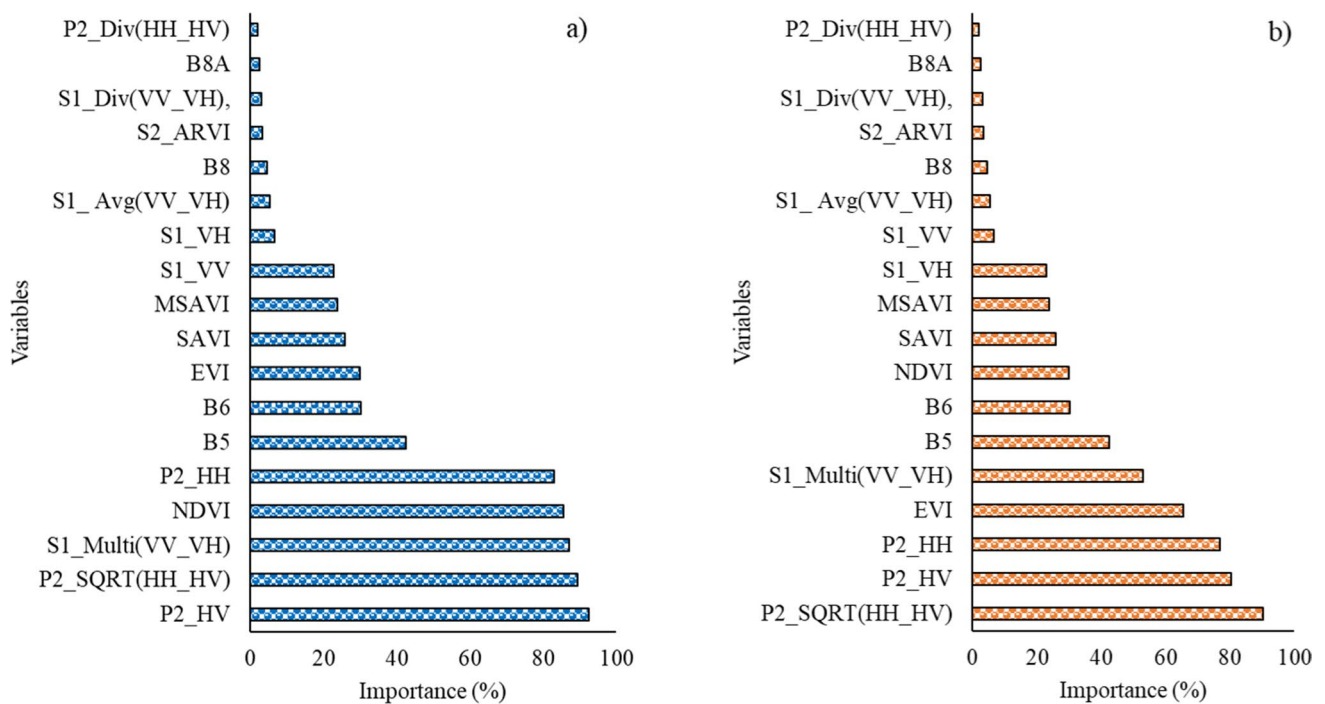


Fig. 3 Variable importance of **a** RF and **b** CatBoost ML models

Model-Based AGB Prediction

The ML models utilised satellite data from Sentinel-2, Sentinel-1, and PALSAR-2. Subsequently, performance metrics were employed to assess the accuracy of the models. The RF model was utilised on the integrated dataset, resulting in a cross-validation RMSE of 72.98 Mg/ha. This value represents the average discrepancy between the predicted and observed AGB values. The R^2 for the model was 0.71, suggesting that the predictors employed in the analysis were able to account for approximately 71% of the observed variability in AGB. The MAE was calculated to be 46.70 Mg/ha, which signifies the average absolute discrepancy between the predicted and observed AGB values, as illustrated in Fig. 5a. The RF model exhibited a bias of 4.32 Mg/ha, suggesting a marginal inclination towards either overestimating or underestimating AGB.

Similarly, the CatBoost model applied to the integrated dataset showed a cross-validation RMSE of 80.69 Mg/ha, reflecting the overall model performance. The R^2 value for the CatBoost model was 0.67, indicating that the predictors could explain approximately 67% of the variability in AGB. The MAE of the CatBoost model was calculated as 52.30 Mg/ha, representing the average absolute difference between predictor variables and targeted variable. The bias of the CatBoost model was 5.63 Mg/ha, indicating the presence of a slight systematic deviation in the AGB estimation (Fig. 5b). The RF model exhibited slightly better

performance than the CatBoost model with lower RMSE, higher R^2 , lower MAE, and lower bias. However, both models achieved reasonably good accuracy in capturing variations in AGB (Table 2).

Variable Importance Analysis

The variable importance analysis identified several significant predictors that strongly influenced the models used for estimating AGB. The variables that made the most significant contributions to the models were ranked based on their respective levels of importance. P2_HV, P2_SQRT(HH_HV), S1_Multi(VV_VH), NDVI, P2_HH, B5, B6, EVI, SAVI, MSAVI, S1_VV, S1_VH, S1_Avg(VV_VH), B8, S2_ARVI, S1_Div(VV_VH), B8A, and P2_Div(HH_HV) were found to be more significant variables. The significance of radar backscatter measurements in estimating AGB was highlighted by incorporating variables such as P2_HV and P2_SQRT(HH_HV) obtained from PALSAR-2 data, as depicted in Fig. 3. The variables mentioned above have facilitated a deeper understanding of the interplay between radar signals and vegetation, thereby aiding in characterising vegetation structure and biomass distribution.

Similarly, Sentinel-2 spectral bands such as B5, B6, and B8 were of moderate importance because of their sensitivity to vegetation and canopy reflectance properties. VIs like NDVI, EVI, SAVI, and MSAVI were critical in capturing

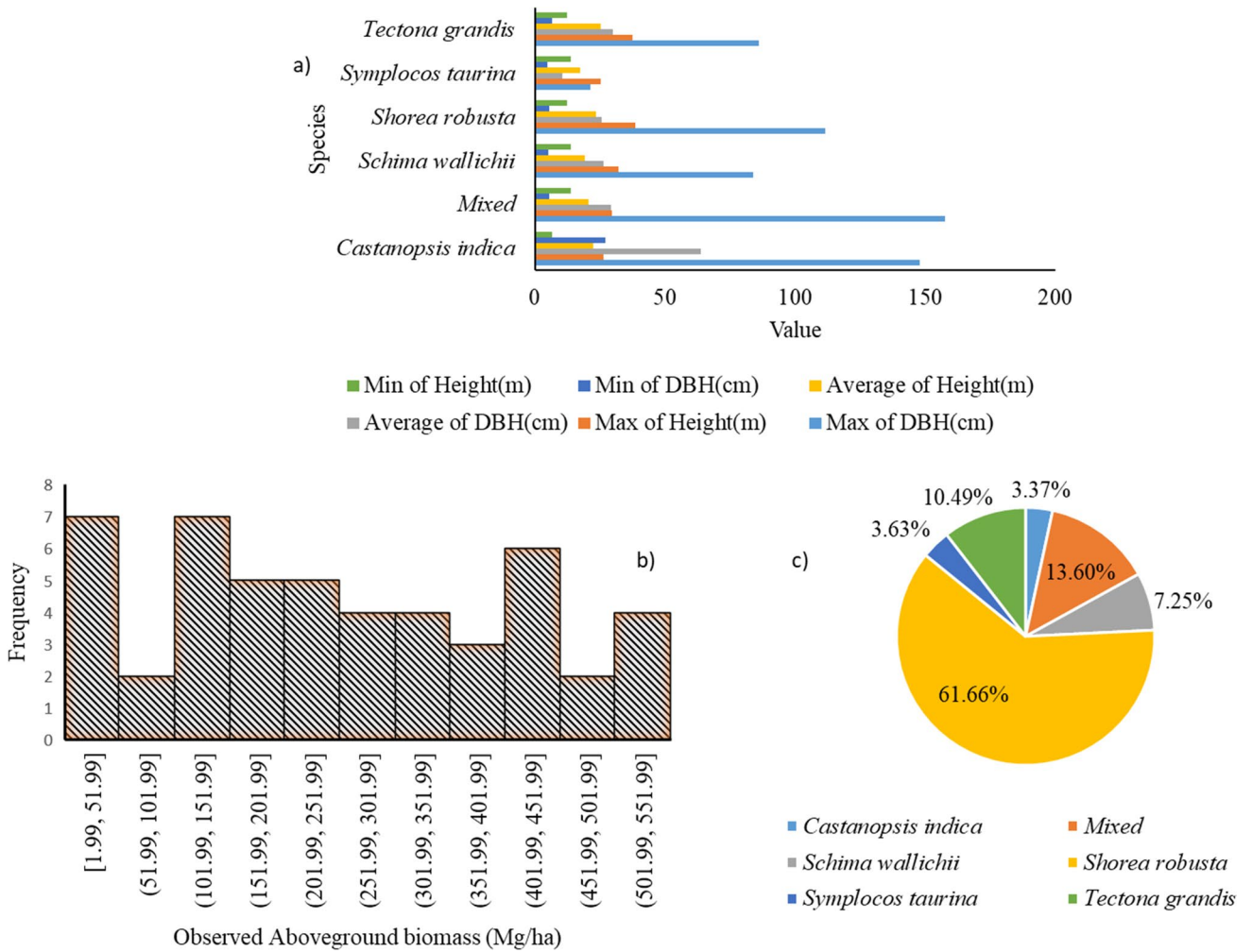


Fig. 4 a DBH of different species, b field-based AGB analysis and its frequency. It shows the AGB distribution and frequency of field sampling points, c percentage of individual species composition

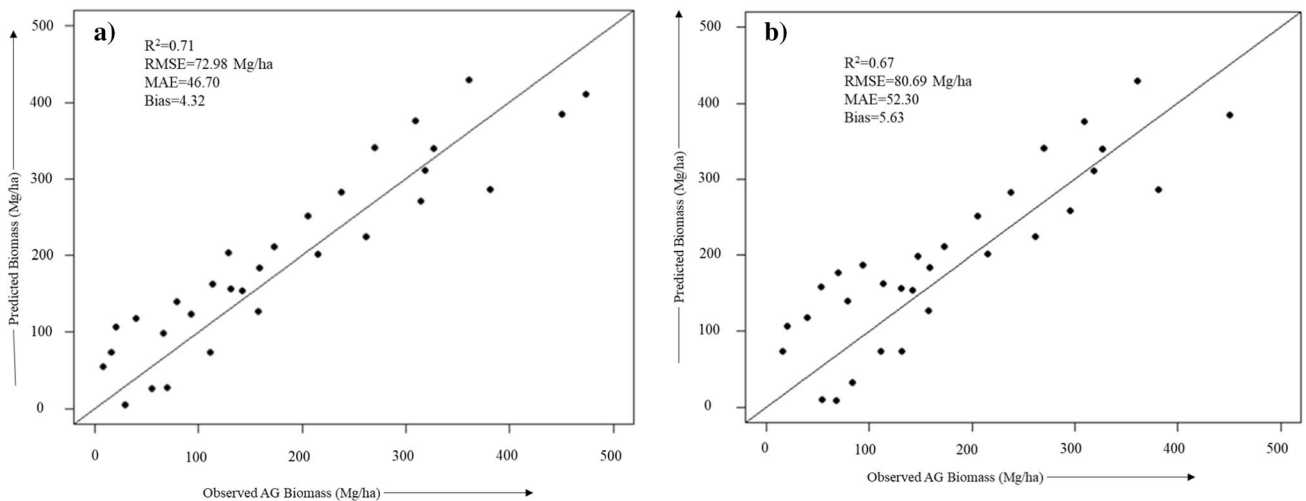


Fig. 5 a The RF and b CatBoost model predicted vs field-estimated AGB

Table 2 Details of regression error values using RF, CatBoost models run for AGB estimation

Satellite data	Model	AGB range (Mg/ha)	Cross validation RMSE (Mg/ha)	Coefficient of determination (R^2)	Mean absolute error (MAE)	Bias (Mg/ha)
Sentinel-2 + Sentinel-1 + PALSAR-2	RF	51–477	72.98	0.71	46.70	4.32
Sentinel-2 + Sentinel-1 + PALSAR-2	CatBoost	61–467	80.69	0.67	52.30	5.63

vegetation vigour and chlorophyll content, allowing for a more accurate estimation of AGB variations. Furthermore, the use of radar variables such as S1_Multi(VV_VH) and S1_Avg(VV_VH) demonstrated the importance of analysing different polarisations and their interactions. P2_HV was the most important variable, contributing 92.36% to the overall model performance, thus capturing the variations in AGB. The variable P2_SQRT(HH_HV) followed closely with a percentage importance of 89.36%, indicating its strong influence. S1_Multi(VV_VH) and NDVI were important variables, contributing 86.98% and 85.6%, respectively. These variables capture information from Sentinel-1 and Sentinel-2 sensors and are valuable for understanding model performance. Other variables such as P2_HH, B5, B6, EVI, SAVI, MSAVI, S1_VV, S1_VH, S1_Avg(VV_VH), B8, S2_ARVI, S1_Div(VV_VH), B8A, and P2_Div(HH_HV) exhibited

varying degrees of importance, contributing to the overall model performance.

AGB mapping and Uncertainty Analysis

This study conducted the AGB mapping and validation process for the RF and CatBoost model's performance in estimating AGB. The RF model ranges from 51.37 to 477.63 Mg/ha (Fig. 6a), while the CatBoost model ranges from 61.37 to 467.63 Mg/ha (Fig. 6b). These ranges demonstrate the ability of both models to capture a wide range of AGB. In our study, we employed the coefficient of variation (CV) as a key measure to assess AGB uncertainty. The CV was calculated as the ratio of the standard deviation to the mean of the AGB estimates, quantifying the relative variability in the data. The observed range of CV values from 0 to 20% within our study area provides valuable insights into variability in our AGB estimates.

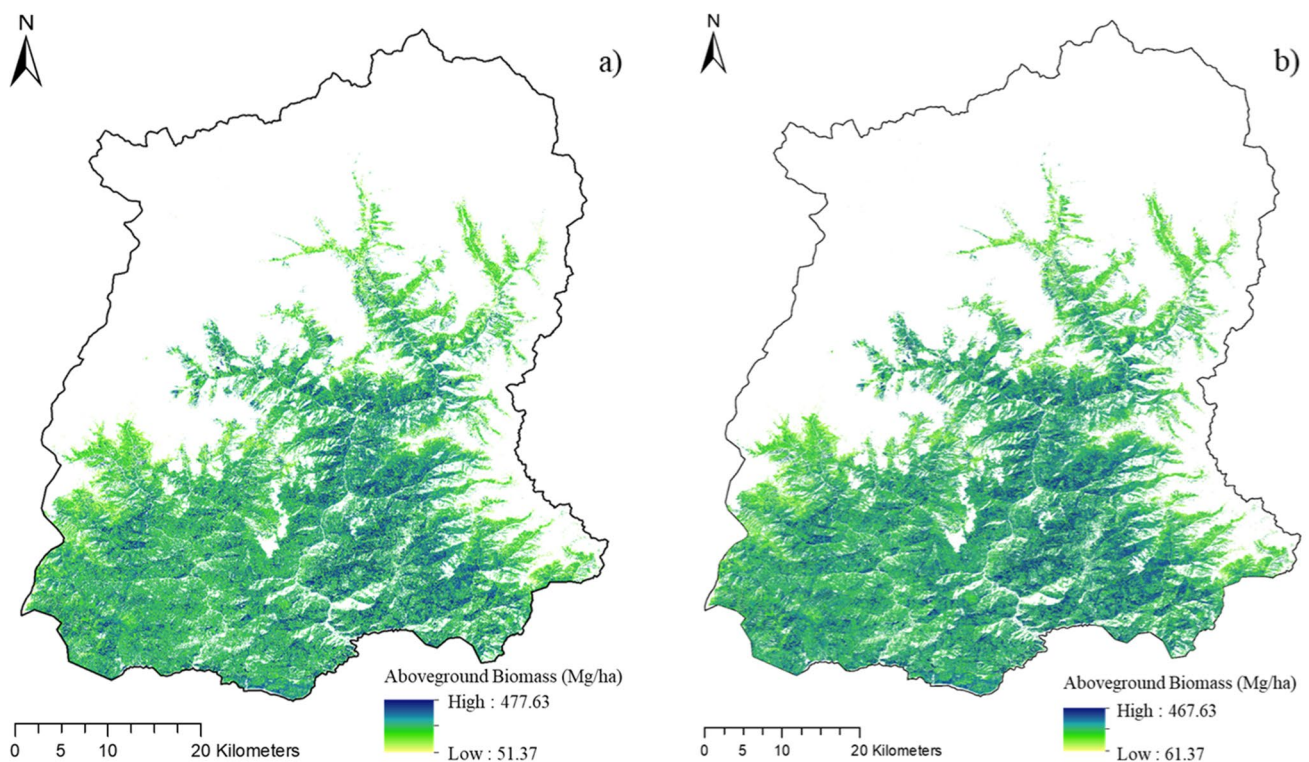


Fig. 6 a RF Predicted map. b CatBoost predicted map. This map presents the predicted AGB values obtained from the RF and CatBoost model

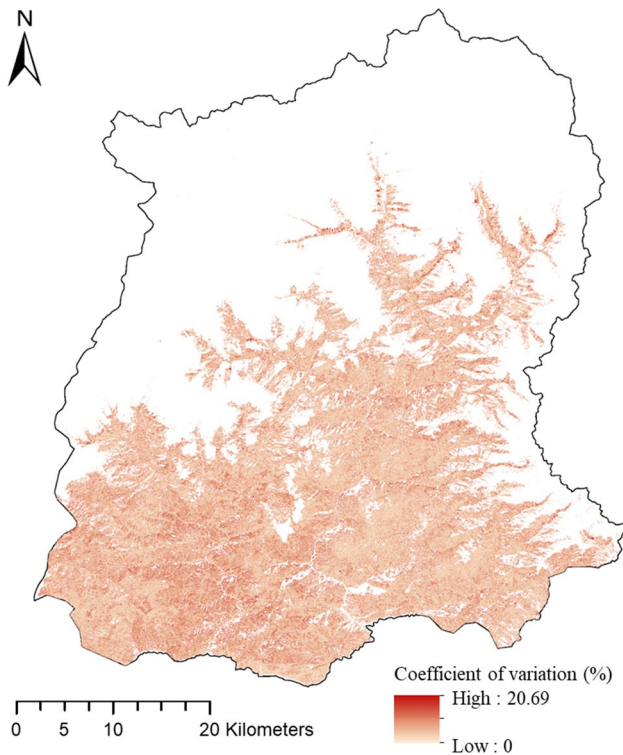


Fig. 7 Coefficient of variation (CV) map. This map visualises the coefficient of variation, CV as a key measure to assess AGB uncertainty

A CV of 0% indicated highly consistent and stable AGB estimates, reflecting homogeneity in forest conditions and AGB distribution. Conversely, a CV of 20% suggested a higher degree of variability, signifying diverse forest types, topography, and other factors influencing AGB distribution. Therefore, CV serves as a robust tool for quantifying and communicating the uncertainty in our AGB estimations, offering a valuable perspective on the reliability of our AGB mapping efforts (Fig. 7).

Discussion

ML Models and Feature Importance

Both RF and CatBoost models exhibited strong performance in AGB estimation, as indicated by their respective R^2 values of 0.71 and 0.67. These values suggest that the input variables and model predictions explain approximately 71 and 67% of the variance, respectively. Additionally, RMSE of 72.98 Mg/ha and 80.69 Mg/ha highlights the models' ability to minimise prediction errors. The efficacy of ML models resides in their capacity to effectively manage nonlinear associations and capture intricate interactions among numerous predictor variables. These models can

effectively estimate AGB across the study area by leveraging the rich information provided by multi-sensor data. The variable importance analysis further underscores the significance of various spectral indices, radar backscatter values, and VIs in influencing AGB estimates.

The results of the variable importance analysis indicated that certain variables had a more significant influence on AGB estimation than others. For instance, P2_HV, P2_SQRT(HH_HV), and S1_Multi(VV_VH) emerged as the most important predictors, highlighting the importance of polarimetric SAR data in AGB modelling. These variables capture the interaction between radar waves and forest structure, allowing for a more accurate estimation of AGB. NDVI, P2_HH, and B5 also demonstrated considerable importance, emphasising the significance of spectral reflectance and VIs in capturing vegetation density and AGB variations. However, it is essential to address the site-specific nature of these findings. The importance of variables in AGB estimation can vary across different geographical locations and ecosystems. The most influential variables may depend on local environmental conditions, vegetation types, and forest structure (Ghosh et al., 2022a, 2022b). While our study emphasises the relevance of these variables in our specific study area, we acknowledge that their significance can differ in other contexts. Variability in variable importance underscores the diverse nature of forests and landscapes worldwide.

Identifying these influential variables will provide valuable insights for future research and practical applications. By understanding the relative importance of different variables, researchers and practitioners can optimise their data collection strategies, prioritise relevant features for AGB estimation, and streamline the modelling process (David et al., 2022). Furthermore, variable importance analysis helps to identify potential data gaps and areas for improvement. This enables researchers to focus on acquiring or enhancing specific data types or variables that significantly impact accuracy. Variable importance analysis is not only restricted to our study but has been extensively employed in various AGB estimation studies. Similar findings regarding the importance of SAR data, spectral indices, and vegetation-related variables were reported by other researchers (Li et al., 2020; Nandy et al., 2021; Rosenqvist et al., 2014). This consensus further strengthens the reliability and generalisability of our results and highlights the consistent role of these variables in AGB modelling across different regions and ecosystems.

Sampling Variations in Density Classes within Species Distribution

In our study, we laid 49 ESUs across the site, strategically chosen based on the heterogeneity observed in AGB

distribution. Our sampling approach considered the minimum and maximum ranges, encompassing dominant species and their associates, resulting in estimated AGB values ranging from 2 to 552 Mg/ha. Our sampling efforts might have slightly underestimated the maximum AGB in Sikkim, as denser forests and additional sampling could potentially reveal higher values. Our RF and CatBoost models, which predicted AGB, demonstrated robust performance, yielding estimates within the ranges of 51–478 and 61–468 Mg/ha, respectively. However, we acknowledge the inherent diversity of forests, even within a single geographical region, such as Sikkim. Variations in tree species, age, health, and local environmental conditions contribute to differences in forest density and structure, thereby influencing the AGB estimates. Recognising these nuances is crucial for refining AGB estimation models. Despite the challenges posed by the topographically complex terrain of Sikkim, our study aimed to generate an indicative and maiden AGB map for the state, with a primary focus on demonstrating the integration of SAR data and ML algorithms for predicting the AGB.

To ensure a robust model, we adopted a 60–40 split strategy, allocating 60% (31 ESUs) for model training and 40% (18 ESUs) for testing and validation. This approach aims to balance the model's accuracy and generalisation. While this study serves as a technological demonstration, showcasing the potential of SAR data and ML algorithms in AGB prediction, we recognise that more extensive sampling could enhance AGB estimates. Future studies should consider incorporating a broader array of environmental variables, expanding datasets to include various density classes within species formations, and employing more tree-based allometric equations to minimise bias. In conclusion, our study lays the foundation for an indicative AGB map of Sikkim, emphasising the technological demonstration of SAR data and ML algorithms. However, we advocate future research to address the challenges posed by the region's topography by including more comprehensive data collection, increasing the sample size, and refining environmental variables. These improvements will undoubtedly contribute to more accurate and nuanced AGB predictions across different forest types and conditions in Sikkim.

Comparison to Similar Studies

Our findings are consistent with previous research, which also observed similar patterns in the AGB estimation. Recent studies, such as those conducted by Dang et al. (2019) and Ghosh et al., (2022a, 2022b), have emphasised the effectiveness of integrating ML models with multi-sensor data for accurate AGB estimation. For example, Guerra-Hernández et al. (2022) utilised an RF model with

ICESat-2, Sentinel-1, Sentinel-2, ALOS2/PALSAR-2, and topographic data to estimate AGB in tropical forests. Their results demonstrated reasonable agreement with ICESat-2- and ALS-based AGB observations, with $R^2 = 0.63$ and 0.64 and RMSE values of 11.10 and 12.28 Mg/ha, respectively. Another study by Luo et al. (2021) employed various ML models to estimate AGB using Landsat data, among which CatBoost exhibited the highest accuracy. The RMSE values obtained were 26.54 Mg/ha for coniferous forests, 24.67 Mg/ha for broad-leaved forests, 22.62 Mg/ha for mixed forests, and 25.77 Mg/ha for all forests. These findings corroborate our results and further underscore the efficacy of ML models in accurately estimating AGB. Furthermore, Ghosh et al. (2018) and Malhi et al. (2021) support variable importance, emphasising the significance of VIs, radar backscatter values, and VIs in accurately estimating the AGB. Collectively, these comparisons highlight the robustness and reliability of our approach, substantiating the advancements achieved in AGB estimation through the integration of ML models and multi-sensor data. This study has limitations in terms of the sample size used for model training, which may have constrained the representation of AGB variability across the Sikkim Himalaya region. To enhance the predictive capabilities of ML models, future research could consider incorporating a more comprehensive range of environmental variables and increasing the sample size. This leads to an improved accuracy in the predictions made by the model.

Conclusion

This study highlights the performance of the ML RF and CatBoost models in accurately estimating AGB in the Indian state of Sikkim. Using multi-sensor data from Sentinel-2, Sentinel-1, and PALSAR-2, we achieved high accuracy, with the RF model exhibiting an R^2 of 0.71 and an RMSE of 72.98 Mg/ha and the CatBoost model producing an R^2 of 0.67 and an RMSE of 80.69 Mg/ha. The variable importance analysis revealed significant predictors, such as P2_HV, P2_SQRT(HH_HV), S1_Multi(VV_VH), NDVI, EVI, and spectral bands B5 and B6, emphasising the importance of SAR data, spectral reflectance, and vegetation-related parameters. The similarity between our findings and prior research strengthens their validity and applicability. The precise mapping and estimation of AGB made possible by ML algorithms and RS data has significant implications for ecological studies, carbon accounting, and nature-based solutions within the context of climate change. These findings provide valuable insights for forest management, carbon credit programmes, and developing nature-based solutions.

Acknowledgements AJP and SP thank the Ministry of Education, Government of India, for the grant of PhD Research Fellowships. SS thanks Ministry of Education, Government of India, New Delhi, for providing fellowship for M.Tech study. All Authors acknowledge the authorities of IIT Kharagpur for instrumental support and facilities provided; the Sikkim state forest and wildlife department is thanked for support in conducting fieldwork.

Author's Contribution AJP and MDB contributed to the conceptualization, data curation, formal analysis, methodology, supervision, and writing—original draft. SP and SM were involved in the investigation, methodology, resources, supervision, and writing—review and editing. BRP and SS assisted in the visualization and writing—review and editing. NS contributed to writing—review and editing.

Funding No external funding was received for this research.

Declarations

Conflict of interest Authors declare that there is no potential conflict of interest.

References

- Avitabile, V., Herold, M., Heuvelink, G. B. M., Lewis, S. L., Phillips, O. L., Asner, G. P., et al. (2016). An integrated pan-tropical biomass map using multiple reference datasets. *Global Change Biology*, 22(4), 1406–1420. <https://doi.org/10.1111/gcb.13139>
- Baccini, A., Walker, W., Carvalho, L., Farina, M., & Houghton, R. A. (2019). Tropical forests are a net carbon source based on above-ground measurements of gain and loss. *Science*, 363(6423), 230–234. <https://doi.org/10.1126/science.aat1205>
- Behera, M. D., Tripathi, P., Mishra, B., Kumar, S., Chitale, V. S., & Behera, S. K. (2016). Above-ground biomass and carbon estimates of Shorea robusta and Tectona grandis forests using QuadPOL ALOS PALSAR data. *Advances in Space Research*, 57(2), 552–561. <https://doi.org/10.1016/j.asr.2015.11.010>
- Belgiu, M., & Drăgu, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 114, 24–31. <https://doi.org/10.1016/j.isprsjprs.2016.01.011>
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Dang, A. T. N., Nandy, S., Srinet, R., Luong, N. V., Ghosh, S., & Senthil Kumar, A. (2019). Forest aboveground biomass estimation using machine learning regression algorithm in Yok Don National Park. *Vietnam. Ecological Informatics*, 50, 24–32. <https://doi.org/10.1016/j.ecoinf.2018.12.010>
- David, R. M., Rosser, N. J., & Donoghue, D. N. M. (2022). Improving above ground biomass estimates of Southern Africa dryland forests by combining Sentinel-1 SAR and Sentinel-2 multispectral imagery. *Remote Sensing of Environment*, 282, 113232. <https://doi.org/10.1016/j.rse.2022.113232>
- Dorogush, A. V., Ershov, V., & Gulin, A. (2018). CatBoost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*
- Fassnacht, F. E., Latifi, H., Stereńczak, K., Modzelewska, A., Lefsky, M., Waser, L. T., et al. (2016). Review of studies on tree species classification from remotely sensed data. *Remote Sensing of Environment*, 186, 64–87. <https://doi.org/10.1016/j.rse.2016.08.013>
- Forkuor, G., Benewinde Zoungrana, J.-B., Dimobe, K., Ouattara, B., Vadrevu, K. P., & Tondoh, J. E. (2020). Above-ground biomass mapping in West African dryland forest using Sentinel-1 and 2 datasets - A case study. *Remote Sensing of Environment*, 236, 111496. <https://doi.org/10.1016/j.rse.2019.111496>
- FSI. (1996). Volume equations for forests of India, Nepal and Bhutan. Forest Survey of India, Ministry of Environment and Forests, Govt. of India, Dehradun
- FSI. (2019). India state of forest report. *Forest Survey of India*
- FSI. (2021). India State of Forest Report. Dehradun: Forest Survey of India, Ministry of Environment Forest and Climate Change
- Ghosh, S. M., & Behera, M. D. (2018). Aboveground biomass estimation using multi-sensor data synergy and machine learning algorithms in a dense tropical forest. *Applied Geography*, 96, 29–40. <https://doi.org/10.1016/j.apgeog.2018.05.011>
- Ghosh, S. M., Behera, M. D., Jagadish, B., Das, A. K., & Mishra, D. R. (2021). A novel approach for estimation of aboveground biomass of a carbon-rich mangrove site in India. *Journal of Environmental Management*, 292, 112816. <https://doi.org/10.1016/j.jenvman.2021.112816>
- Ghosh, S. M., Behera, M. D., Kumar, S., Das, P., Prakash, A. J., Bhaskaran, P. K., et al. (2022a). Predicting the forest canopy height from lidar and multi-sensor data using machine learning over India. *Remote Sensing*. <https://doi.org/10.3390/rs14235968>
- Ghosh, S. M., Behera, M. D., Kumar, S., Das, P., Prakash, A. J., Bhaskaran, P. K., & Behera, S. K. (2022b). Predicting the forest canopy height from LiDAR and multi-sensor data using machine learning over India. *Remote Sensing*, 14(23), 5968.
- Guerra-Hernández, J., Narine, L. L., Pascual, A., Gonzalez-Ferreiro, E., Botequim, B., Malambo, L., et al. (2022). Aboveground biomass mapping by integrating ICESat-2, SENTINEL-1, SENTINEL-2, ALOS2/PALSAR2, and topographic information in Mediterranean forests. *Giscience & Remote Sensing*, 59(1), 1509–1533. <https://doi.org/10.1080/15481603.2022.2115599>
- Hancock, J. T., & Khoshgoftaar, T. M. (2020). CatBoost for big data: An interdisciplinary review. *Journal of Big Data*, 7(1), 94. <https://doi.org/10.1186/s40537-020-00369-8>
- Jha, N., Tripathi, N. K., Barbier, N., Virdis, S. G. P., Chanthorn, W., Viennois, G., et al. (2021). The real potential of current passive satellite data to map aboveground biomass in tropical forests. *Remote Sensing in Ecology and Conservation*, 7(3), 504–520. <https://doi.org/10.1002/rse2.203>
- Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, 28(5), 1–26. <https://doi.org/10.18637/jss.v028.i05>
- Li, Y., Li, M., Li, C., & Liu, Z. (2020). Forest aboveground biomass estimation using Landsat 8 and Sentinel-1A data with machine learning algorithms. *Scientific Reports*, 10(1), 9952. <https://doi.org/10.1038/s41598-020-67024-3>
- Lu, D., Chen, Q., Wang, G., Liu, L., Li, G., & Moran, E. (2016). A survey of remote sensing-based aboveground biomass estimation methods in forest ecosystems. *International Journal of Digital Earth*, 9(1), 63–105. <https://doi.org/10.1080/17538947.2014.990526>
- Luo, M., Wang, Y., Xie, Y., Zhou, L., Qiao, J., Qiu, S., & Sun, Y. (2021). Combination of feature selection and catboost for prediction: The first application to the estimation of aboveground biomass. *Forests*. <https://doi.org/10.3390/f12020216>
- Malhi, R. K. M., Anand, A., Srivastava, P. K., Chaudhary, S. K., Pandey, M. K., Behera, M. D., et al. (2021). Synergistic evaluation of Sentinel 1 and 2 for biomass estimation in a tropical forest of India. *Advances in Space Research*. <https://doi.org/10.1016/j.asr.2021.03.035>
- Mutanga, O., Masenyama, A., & Sibanda, M. (2023). Spectral saturation in the remote sensing of high-density vegetation traits: A systematic review of progress, challenges, and prospects. *ISPRS*

- Journal of Photogrammetry and Remote Sensing*, 198, 297–309. <https://doi.org/10.1016/j.isprsjprs.2023.03.010>
- Nandy, S., Srinet, R., & Padalia, H. (2021). Mapping forest height and aboveground biomass by integrating ICESat-2, sentinel-1 and sentinel-2 data using random forest algorithm in northwest himalayan foothills of India. *Geophysical Research Letters*. <https://doi.org/10.1029/2021GL093799>
- Prakash, A. J., Behera, M. D., Ghosh, S. M., Das, A., & Mishra, D. R. (2022). A new synergistic approach for Sentinel-1 and PALSAR-2 in a machine learning framework to predict aboveground biomass of a dense mangrove forest. *Ecological Informatics*. <https://doi.org/10.1016/j.ecoinf.2022.101900>
- Rosenqvist, A., Shimada, M., Suzuki, S., Ohgushi, F., Tadono, T., Watanabe, M., et al. (2014). Operational performance of the ALOS global systematic acquisition strategy and observation plans for ALOS-2 PALSAR-2. *Remote Sensing of Environment*, 155, 3–12.
- Sharma, N., Behera, M. D., Das, A. P., & Panda, R. M. (2019). Plant richness pattern in an elevation gradient in the Eastern Himalaya. *Biodiversity and Conservation*, 28(8), 2085–2104. <https://doi.org/10.1007/s10531-019-01699-7>
- Shimada, M., Itoh, T., Motooka, T., Watanabe, M., Shiraishi, T., Thapa, R., & Lucas, R. (2014). New global forest/non-forest maps from ALOS PALSAR data (2007–2010). *Remote Sensing of Environment*, 155, 13–31.
- Singh, C., Karan, S. K., Sardar, P., & Samadder, S. R. (2022). Remote sensing-based biomass estimation of dry deciduous tropical forest using machine learning and ensemble analysis. *Journal of Environmental Management*, 308, 114639. <https://doi.org/10.1016/j.jenvman.2022.114639>
- Singh, R. K., Biradar, C. M., Behera, M. D., Prakash, A. J., Das, P., Mohanta, M. R., & Rizvi, J. (2023). Optimising carbon fixation through agroforestry: Estimation of aboveground biomass using multi-sensor data synergy and machine learning. *Ecological Informatics*, 79, 102408.
- Nations, U. (2016). *The Sustainable Development Goals 2016*. eSocialSciences
- Vaglio Laurin, G., Chen, Q., Lindsell, J. A., Coomes, D. A., Frate, F. D., Guerriero, L., et al. (2014a). Above ground biomass estimation in an African tropical forest with lidar and hyperspectral data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 89, 49–58. <https://doi.org/10.1016/j.isprsjprs.2014.01.001>
- Vaglio Laurin, G., Chen, Q., Lindsell, J. A., Coomes, D. A., Frate, F. D., Guerriero, L., et al. (2014b). Above ground biomass estimation in an African tropical forest with lidar and hyperspectral data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 89, 49–58. <https://doi.org/10.1016/j.isprsjprs.2014.01.001>
- Wang, J., Xiao, X., Bajgain, R., Starks, P., Steiner, J., Doughty, R. B., & Chang, Q. (2019). Estimating leaf area index and aboveground biomass of grazing pastures using Sentinel-1, Sentinel-2 and Landsat images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 154, 189–201. <https://doi.org/10.1016/j.isprsjprs.2019.06.007>
- Xiao, J., Chevallier, F., Gomez, C., Guanter, L., Hicke, J. A., Huete, A. R., et al. (2019). Remote sensing of the terrestrial carbon cycle: A review of advances over 50 years. *Remote Sensing of Environment*, 233, 111383. <https://doi.org/10.1016/j.rse.2019.111383>
- Yu, Y., & Saatchi, S. (2016). Sensitivity of L-band SAR backscatter to aboveground biomass of global forests. *Remote Sensing*. <https://doi.org/10.3390/rs8060522>
- Zhang, Y., Liang, S., & Sun, G. (2014). Forest biomass mapping of northeastern china using GLAS and MODIS data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(1), 140–152. <https://doi.org/10.1109/JSTARS.2013.2256883>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.