Remote Sensing Applications: Society and Environment 27 (2022) 100817

Contents lists available at ScienceDirect

Remote Sensing Applications: Society and Environment

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

Mixed tropical forests canopy height mapping from spaceborne LiDAR GEDI and multisensor imagery using machine learning models

Rajit Gupta, Laxmi Kant Sharma *

Remote Sensing & GIS Lab, Department of Environmental Science, School of Earth Sciences, Central University of Rajasthan, NH-8, Bandarsindri, 305817, Ajmer, Rajasthan, India

ARTICLE INFO

Keywords:
GEDI
Optical
SAR
Canopy height
Machine learning
Tropical mixed forests

ABSTRACT

Spatial mapping of forests canopy height (Hcanopy) provides an opportunity to assess above-ground biomass, net primary productivity, carbon dioxide (CO_2) sequestration, biodiversity conservation and forest fire risks. This study incorporated a continuous coverage of multi-spectral optical and synthetic aperture radar (SAR) along with sparsely global ecosystem dynamics investigation (GEDI) spaceborne Light Detection and Ranging (LiDAR) data in the machine learning (ML) models for mapping Hcanopy in the mixed tropical forests of Shoolpaneshwar wildlife sanctuary (SWLS), Gujarat, India. We trained seven ML models, including quantile random forest (QRF), support vector machine (SVM), Bayesian regularization for feed-forward neural networks (BRNN), conditional inference random forest (Cforest), Extreme gradient boosting (Xgbtree), multivariate adaptive regression splines (MARS), and k-nearest neighbors (KNN) using GEDI_02A extracted Hcanopy as training data. We used predictors which were extracted from LiDAR (GEDI metrics), multispectral optical (Landsat -8, Sentinel-2), and SAR (ALOS-2/PALSAR-2, Sentinel-1). A 10-fold cross-validation (CV) resampling was used to avoid overfitting or underfitting. The comparison of the models performances shows that the BRNN model has the highest satisfactory accuracy metrics, such as root mean square error (RMSE) of 4.686 m, R-squared (R^2) of 0.49 and mean absolute error (MAE) of 3.66 m. Low training samples of tall canopies (>25 m), presence of mixed vegetation, geometric and structural variability and sloppy terrain of SWLS possibly restricted models from performing well. Field validation shows an R^2 of 0.55, satisfactory for mixed tropical forests using spaceborne LiDAR. The present work provides insights into using spaceborne LiDAR GEDI data with optical and SAR data for Hcanopy mapping through ML models, which help to manage SWLS and further implications of forest Hcanopy mapping over large spatial scales.

1. Introduction

Quantitative mapping of forest canopy height (Hcanopy) helps to determine above-ground biomass, net primary productivity, carbon dioxide (CO_2) sequestration, biodiversity conservation, forests health, restoration and management, and forest fire risks (Simard et al., 2011; Wang et al., 2018, 2021). Manual measurements of Hcanopy have been hindered by topographical and climatic

* Corresponding author.
E-mail addresses: 2017phdes03@curaj.ac.in (R. Gupta), laxmikant_evs@curaj.ac.in (L.K. Sharma).

https://doi.org/10.1016/j.rsase.2022.100817
Received 29 October 2021; Received in revised form 5 July 2022; Accepted 25 July 2022
Available online 31 July 2022

variability across the geographical areas, producing gaps and uncertainties (Chen et al., 2021a). Additionally, measuring Hcanopy in the field is time-consuming, labour-intensive, and offers limited information which is generally specific to forest plot level (Stojanova et al., 2010). Using remote sensing & geographical information system (GIS), it is possible to monitor forests and their attributes at a spatial scale with repeated observations (Lin et al., 2020). Light detection and ranging (LiDAR) remote sensing can measure discrete tree stand structures in three-dimension, including horizontal and vertical profile distribution (Lefsky et al., 1999; Zhang et al., 2017; Sharma et al., 2021). Evidently, terrestrial and airborne LiDAR sensors provide more detailed and accurate forest structural measurements at a small scale than spaceborne LiDAR due to very high scanning density (Larue et al., 2020). However, high cost, large processing time, and less data availability cause the narrow down the adoptability of terrestrial and airborne LiDAR in the public domain.

GEDI LiDAR instrument was launched in late 2018 and mounted on the international space station (ISS), providing billions of discrete point measurements to monitor tropical and temperate forest ecosystems since 2019 (Qi and Dubayah, 2016; Dubayah et al., 2020a). GEDI measures forest structural characteristics by providing vertical distributions of forest canopies in waveforms (Dubayah et al., 2020b). GEDI gives main information about surface topography, elevation, Hcanopy, relative height metrics, plant area index (PAI), and gridded above-ground biomass (Tang and Armston, 2019; Potapov et al., 2021). GEDI data comprises discrete footprints/shots of spatial resolution 25 m, sparsely distributed over the coverage regions (Tang et al., 2019). Therefore, sparse coverage of GEDI Hcanopy over spatial scales can be continuously imputed with remote sensing optical and SAR data using ML models (Jiang et al., 2021).

SAR Advanced land observing satellite-2/phased array type L-band synthetic aperture radar-2 (ALOS-2/PALSAR-2) and Sentinel-1 ground range detected (GRD) analysis-ready data are publicly available. SAR has high spatial resolution and temporal coverage and is not affected due to cloudy weather conditions (Chen et al., 2021a). Optical sensors (Landsat-8 and Sentinel-2) provide inadequate information on forest structures, while Synthetic Aperture Radar (SAR) sensor such as Sentinel-1 C-band data gives information only about canopy top cover due to the low penetration power. Therefore, these sensors are insufficient to acquire in-depth forest structural information (Spracklen and Spracklen, 2021a). Spaceborne LiDAR data beams can penetrate the dense, multi-layered canopy, which can better estimate forest dynamics and canopy structure than optical and SAR (Spracklen and Spracklen, 2021a).

ML models have emerged as widely used tools for remote sensing data classification and regression modelling (Wang et al., 2021). Machine learning (ML) models can be effective for the continuous mapping of sparsely distributed GEDI Hcanopy. ML models are broadly categorized into parametric and non-parametric. Parametric ML models only consider standard linear relationships, whereas non-parametric ML models could explain complex and non-linear relationships between dependent and independent variables of interest (Maxwell et al., 2018). Examples of commonly used non-parametric ML models are neural networks, support vector machines (SVM), K-nearest neighbors (KNN), boosting models, and random forest (RF) (Jiang et al., 2021). Previous studies by Ota et al., (2014); Staben et al., (2018); Pourshamsi et al., (2021); Wang et al., (2021) explored the different non-parametric ML models for the prediction of Hcanopy with reasonable accuracy.

Certainly, spaceborne LiDAR remote sensing is an innovative and promising opportunity for predictive mapping of forest Hcanopy. However, only a few studies have explored the integrated use of space-borne LiDAR GEDI with predictors from multisensor imagery for

Table 1
List of studies that used GEDI for Hcanopy mapping, including author's contribution/findings.

Author's name	ML model	Predictors	Contribution/Findings
Fayad et al., (2021)	Four convolutional neural networks (CNN) variants	Footprint geolocation, waveform samples, Relative Hcanopy (0%–100%)	They used deep learning CNN models to automatically extract useful predicting information from GEDI LiDAR waveforms to estimate dominant Hcanopy. The results showed that the CNN variants could predict dominant Hcanopy with a root mean square error (RMSE) (1.54 m–1.94 m) and R-squared (R^2) (0.86–0.91).
Lang et al. (2021a)	Bayesian CNN	GEDI L1B waveforms	They mapped global Hcanopy estimates with RMSE of 2.7 m, low bias and the underestimation bias from −1.0 m to −0.1 m (mean error)
Lang et al. (2021b)	Deep CNN	Sentinel-2	They estimated canopy height for each 10 m Sentinel-2 pixel using GEDI LiDAR as training data and achieved an overall RMSE of 6.3 m.
Potapov et al., (2021)	Per-pixel ML algorithm (regression tree)	Multi-temporal Landsat-8, vegetative indices and Shuttle Radar Topography Mission (SRTM) DEM (90 m)	They integrated GEDI data and Landsat-8 to map global Hcanopy. RMSE of 6.6 m, MAE of 4.45 m, and R^2 of 0.62 were obtained by comparing predicted Hcanopy with GEDI validation data, while RMSE of 9.07 m, MAE of 6.36 m, and R^2 of 0.61 was obtained on comparing with airborne LiDAR.
Rishmawi et al., (2021)	Random forest regression	Annually metrics from Visible Infrared Imaging Radiometer Suite (VIIRS) data	They integrated GEDI-derived canopy structure data with VIIRS-derived multi-temporal phenology metrics to produce contiguous maps of Hcanopy, canopy fraction cover, PAI, and foliage height diversity. The predicted outcomes were validated against the independent GEDI subset. An accuracy (R^2 = 0.8 and RMSE = 3.35 m) was obtained for Hcanopy.
Spracklen and Spracklen, (2021b)	Random forest	GEDI retrieved metrics such as relative height (RH), thickest canopy, plant area index (PAI), and foliage height diversity	They investigated old-growth forests in the Ukrainian Carpathian Mountains, Europe. The random forest algorithm shows an accuracy of 73% in classifying old-growth forests.

predictive mapping of forest Hcanopy using ML models (Table 1). Therefore, we aimed to employ the capabilities of robust ML models in examining their utility on spatially continuous multisensor optical, SAR, topographical and climate data variables for spaceborne LiDAR GEDI in the spatial mapping of Hcanopy over the mixed tropical forests of Shoolpaneshwar wildlife sanctuary (SWLS), Gujarat, India. Further, we identified the importance of predictor variables to predict forest Hcanopy using ML models. Further, we showed how prediction errors are distributed according to several ML modelling results. Gupta and Sharma (2020) revealed that the forests in the SWLS faced a high anthropogenic burden and showed a trend of forest cover loss over the years. Therefore, mapping of Hcanopy is helpful for the better management of ecological diversity-rich mixed tropical forests of SWLS.

2. Materials and methods

2.1. Study site

The study was conducted in the mixed tropical forest of the SWLS, located in the southern part of the Narmada district in Gujarat, India. The extent of SWLS is 21°.55′–21°.84′ N, 73°.50′–73°.89′ E. SWLS area is around 675 sq. Km which is covered under Gora, Piplod, Dediapada, Sagai and Fulsar ranges. The surface topography in SWLS is mainly mountainous, with an elevation range between 31 m and 864 m above mean sea level (Fig. 1). SWLS has a tropical climate with a maximum and minimum mean temperature is 43 °C and 8 °C, respectively. The mean rainfall recorded was 1000 mm during the monsoon season. SWLS has a rich diversity of mixed forests comprising slightly moist teak forests, southern moist mixed deciduous forests, dry deciduous scrubs, dry bamboo brakes and tropical riverine forests (Champion and Seth, 1968). The number of young trees with short-range Hcanopy is high due to plantation activities. Moderate to dense tall canopies are located in the inner zones of SWLS. *Tectona grandis* is the dominant tree species in SWLS. Other major tree species are *Terminalia tomentosa*, *Butea monosperma*, *Acacia chundra*, and *Mitragyna parvifolia* (Gupta and Sharma, 2020). Fig. 1 shows the map of the geographical location of SWLS, which was created using ArcGIS Desktop version 10.1. Fig. 1 also shows GEDI footprints/shots, Shuttle Radar Topography Mission (SRTM), digital elevation model (DEM), and field sampling sites.

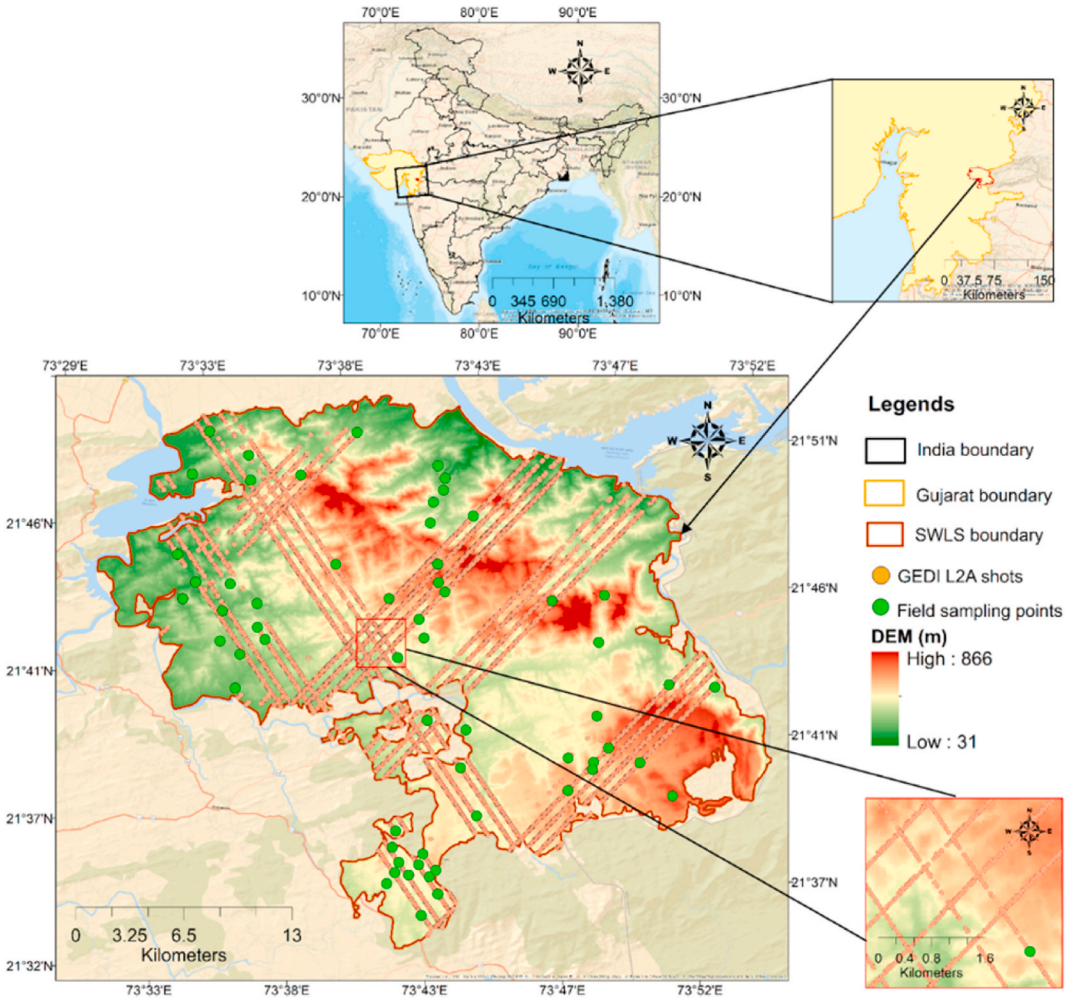

Fig. 1. The geographical location of Shoolpaneshwar wildlife sanctuary (SWLS), Gujarat, India.

2.2. Data collection

We used spaceborne GEDI LiDAR, optical (Landsat-8 and Sentinel-2) multispectral and SAR (ALOS-2/PALSAR-2, Sentinel-1) imageries. In addition, we used surface topographic SRTM 30 m DEM and climatic variables to extract predictors variables (Table 2). Then, we performed data subsetting, resampled all raster layers to a spatial resolution of 30 m and clipped them to match the extent. Field observed data was used for independent validation purposes. The methodological framework comprised of five stages is shown in Fig. 2.

2.2.1. Field sampling

SWLS is a mountainous and sloppy terrain; therefore, sampling points were selected based on site accessibility, forest density and available GEDI footprints. We visited SWLS in November 2019 and 2020 to collect ground observations of Hcanopy (m) (Fig. 3). A total of sixty field plots of 30 m × 30 m were laid for field Hcanopy observations using the Haga altimeter (Bharat Emporium, Haridwar, India), while Hcanopy was randomly measured in inaccessible areas. A handheld global positioning system (GPS) device (Garmin Etrex 10x, Taiwan) was used to record the sampling point's latitude, longitude and elevation. The descriptive statistics for the field measured Hcanopy are shown in Table 3.

2.2.2. Spaceborne GEDI LiDAR data

To download GEDI data granules, we used the GEDI Finder Service Version 1.0 (https://lpdaacsvc.cr.usgs.gov/services/gedifinder) tool for one year (01-04-2019 to 01-04-2020) over SWLS. GEDI instrument emits three laser beams, one is a coverage laser beam, and the other two are the full power laser beams, which further split into four beams and then eight ground tracks comprised of discrete footprints/shots having 25 m spatial resolution. The total ground swath of GEDI laser beams is 4.2 km (https://lpdaac.usgs.gov/). A total of 21 granules were downloaded for GEDI01_B, GEDI02_A, and GEDI02_B (Table 4).

2.2.3. Multispectral sensors optical data (Landsat-8, Sentinel-2)

Surface reflectance (SR) bands of Landsat-8 with mean pixel values from the growing season (June to December 2019) were downloaded from the cloud computing platform Climate Engine (Huntington et al., 2017). In addition, we downloaded already computed band ratios from the Climate Engine platform, while some band ratios were derived using a raster calculator in QGIS Desktop 3.18.0 with Grass 7.8.5. A cloud-free SR Sentinel-2 satellite image was downloaded from Google Earth Engine (GEE) image collection 'Copernicus/S2_SR' for SWLS in November 2019 (Gorelick et al., 2017). We also extracted the Rededge-1 (B5), Rededge-2 (B6), Rededge-3 (B7), and Rededge-4 (B8A) bands having a spatial resolution of 20 m and then resampled them to 30 m.

2.2.4. Synthetic aperture radar (ALOS-2/PALSAR-2, Sentinel-1)

We used the global mosaic and forest/non-forest (FNF) map product of L-band ALOS-2/PALSAR-2 SAR, which is freely available in GeoTIFF format from the Japan Aerospace Exploration Agency (JAXA) webpage (https://www.eorc.jaxa.jp/ALOS/en/palsar_fnf/data/index.htm). The FNF map for 2019 and onwards was not released yet; therefore, we used the FNF map for 2018. We downloaded horizontal polarized transmit/horizontal receive (HH) and horizontal transmit/vertical receive (HV) bands for 2019. The Sentinel-1 mission provides data from a dual-polarization vertically transmission/vertically receiving (VV) and a vertically transmission/horizontally receiving (VH) C-band SAR instrument at center frequency 5.405 GHz. We collected a Sentinel-1 interferometric wide (IW) swath level 1 single look complex (S1A_IW_SLC) image for August 2020 and a Sentinel-1 IW swath level ground range detected (S1A_IW_GRD) image for September 2019. S1A_IW_SLC was downloaded from Copernicus open access hub (https://scihub.copernicus.eu/), while the S1A_IW_GRD image was extracted from GEE 'Copernicus/S1_GRD Sentinel-1 Image Collection' (Gorelick et al., 2017).

2.2.5. Ancillary data

NASA's Shuttle Radar Topography Mission (SRTM) data was used for the DEM and to derive the slope and aspect map of SWLS. The spatial resolution of SRTM V003 data is global 1 arc-second (30 m). In addition, we took monthly mean 30 years (1990–2019) temperature (Tav, °C), precipitation (PPT, mm), and climate water deficit (CWD, mm) as climatic data variables, which were downloaded using the ClimateEngine platform from TerraClimate (Abatzoglou et al., 2018).

Table 2
List of the various datasets, their spatial resolution and predictors used in the current study.

Data type	Sensors	Spatial resolution (m)	Predictors
SAR	ALOS-2/PALSAR-2	25	Backscatter (HH, HV, HH/HV) and RFDI
	Sentinel-1	5 × 20	Backscatter (VV, VH, VV + VH), polarimetric decomposition (alpha, anisotropy and entropy)
Optical	Landsat-8	30	SR bands, vegetative indices, band ratios
	Sentinel-2	20	Rededge-1 (B5), Rededge-2 (B6), Rededge-3 (B8) and Rededge-4 (B8A)
LiDAR	GEDI (L1B, L2A, L2B)	25	Rh25, Rh35, Rh45, Rh55, Rh65, Rh75, Rh85, Rh95, Rh98, Plant area index (PAI)
Training	GEDI L2A	25	Canopy height (Rh100) (Hcanopy)
Validation	Field observations	30	Hcanopy
Ancillary data	SRTM	30	DEM, slope, aspect
	Terraclimate	4000	Climate water deficit (CWD), mean temperature (Tav) and precipitation (PPT)
	ALOS-2/PALSAR-2	25	Forest/non-forest mask

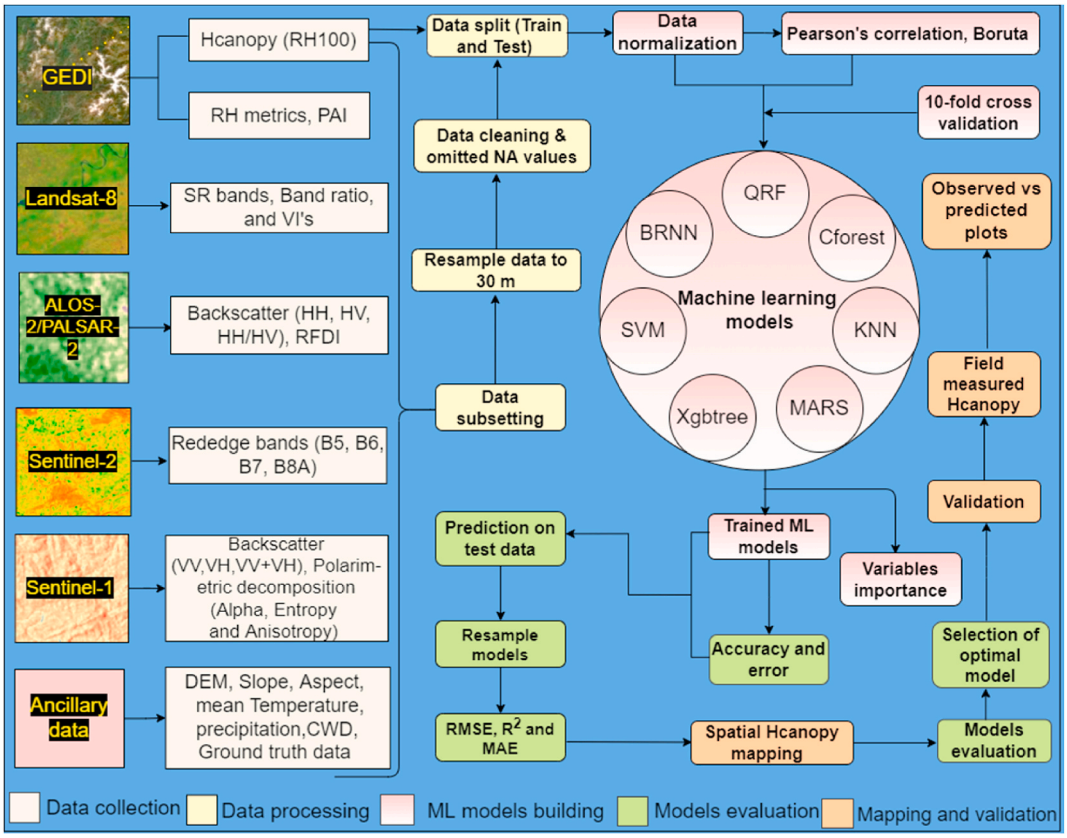

Fig. 2. Outline of methodological framework followed in this study. There are five primary stages: 1) data collection, 2) data processing, 3) ML models building, 4) Models evaluation and 5) Mapping and validation.

2.3. Data pre-processing

2.3.1. GEDI (L1B, L2A, L2B) data pre-processing

We processed GEDI (L1B, L2A, and L2B) granules in Jupyter notebook Python version 3.7. GEDI data has a unique shot number for every shot location in the full orbit represented in the form of the waveform (Fig. 4). On processing GEDI L2A data, we plotted RH metrics (Fig. 5) and quality filtering to remove low-quality shots. Finally, the processed Hcanopy distribution from GEDI L2A products is shown in Fig. 6, and descriptive statistics of retrieved Hcanopy are shown in Table 5. We used the kriging interpolation method to produce spatially continuous estimates of sparsely distributed RH metrics except for RH100. The source code for pre-processing GEDI (L1B, L2A, L2B) products is available through the link https://git.earthdata.nasa.gov/projects/LPDUR/repos/gedi-tutorials/browse.

2.3.2. Optical and SAR pre-processing

Optical Landsat-8 SR bands were used to retrieve a list of vegetative indices, and band ratios are shown in Table 6. ALOS-2/PALSAR-2 digital number values were converted to backscattering coefficient (dB) (Gamma naught) using the equation (1) (Zhang et al., 2019)

$$\gamma^{\circ} = 10*\log(10) * (DN^2) - 83 \quad (1)$$

γ° = Backscattering coefficient (Gamma naught) in dB.
DN = Digital number.

Radar forest degradation index (RFDI) was calculated from speckle filtered HH and HV backscatter ALOS-2/PALSAR-2 bands using the formula shown in Table 6. Sentinel-1 Level-1 GRD preprocessed scenes obtained from GEE as backscatters coefficient (σ°) in decibels (dB). The equation used to calculate backscatter (dB) is $10*\log10\sigma^{\circ}$ (https://earthengine.google.com/). SAR dual-pol polarimetric decomposition like alpha, anisotropy and entropy was retrieved from Sentinel-1 Level-1 SLC in SNAP vs 8.0.0.

2.4. Machine learning models building

We obtained sparsely and uniformly distributed 6916 GEDI Hcanopy point observations, which were used to extract pixel values from predictors. Some observations contain missing values, which were omitted, and we remained with 6855 observations to use in ML models. A target (dependent variable) is predicted from an available set of predictors (independent variables). Model training continues until it reaches the desired level of accuracy. The target variable is GEDI RH100 Hcanopy and various predictors retrieved from

Fig. 3. Field photographs of mixed tropical forests of SWLS captured during field surveys.

Table 3
The descriptive statistics of field-measured Hcanopy in the mixed tropical forests of SWLS.

Descriptive statistics	Values
Mean	12.60
Standard Error	0.24
Median	12
Mode	14
Standard Deviation	5.76
Sample Variance	33.13
Kurtosis	−0.149
Skewness	0.614
Minimum	2.5
Maximum	29
Count	549

Table 4
List of the used GEDI products, their description, data duration, number of granules and sources.

Products	Description	Data duration	No. of granules	Sources
GEDI01_B	Level 1B geolocated waveforms	One year	7	LP DAAC/NASA Earth data search
GEDI02_A	Level 2A elevation and Hcanopy metrics	One year	7	LP DAAC/NASA Earth data search
GEDI02_B	Level 2B canopy cover and vertical profile metrics, PAI	One year	7	LP DAAC/NASA Earth data search

continuous remote sensing datasets, GEDI RH metrics and ancillary data.

Quantile random forests (QRF) are non-parametric, consistent, and accurately estimate conditional quantiles for high-dimensional predictor variables (Meinshausen and Ridgeway, 2006). Support vector machines (SVMs) are among the most robust and accurate machine learning techniques, which separate the classes present in the data by identifying the optimal hyperplane of separation (Rodrigues and de la Riva, 2014). The BRNN model fits a two-layer neural network and uses the Nguyen and Widrow (1990) algorithm to assign initial weights and the Gauss-Newton algorithm to optimise (Foresee and Hagan, 1997; Pérez-Rodríguez et al., 2013). Cforest uses conditional inference trees as base learners. This algorithm creates unbiased decision trees based on sub-sampling without

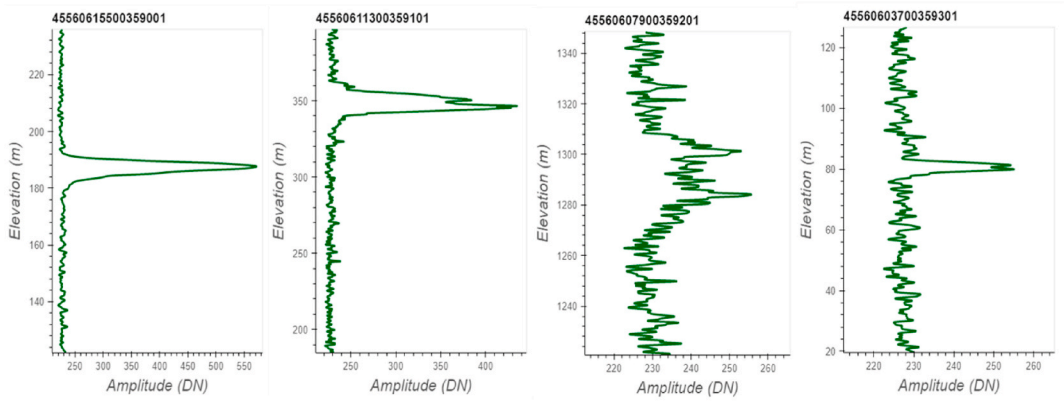

Fig. 4. GEDI L1B waveform for specific shot number from SWLS.

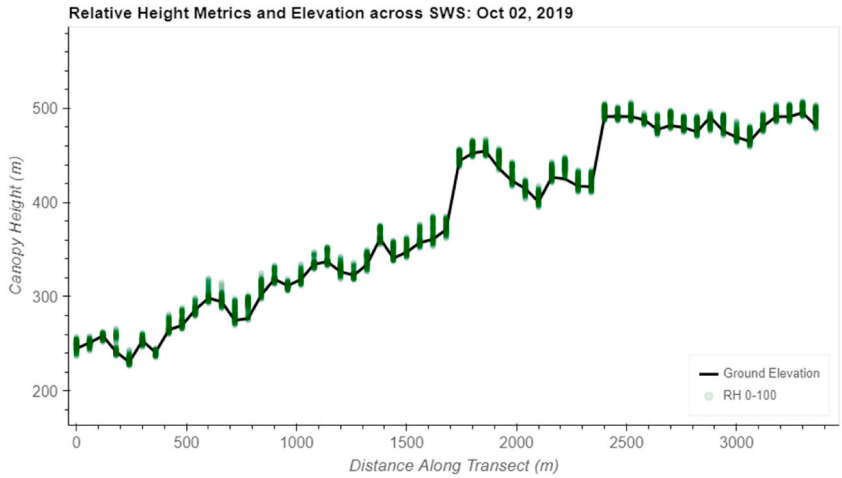

Fig. 5. Canopy RH (0–100) metrics and ground elevation (m) retrieved from GEDI L2A.

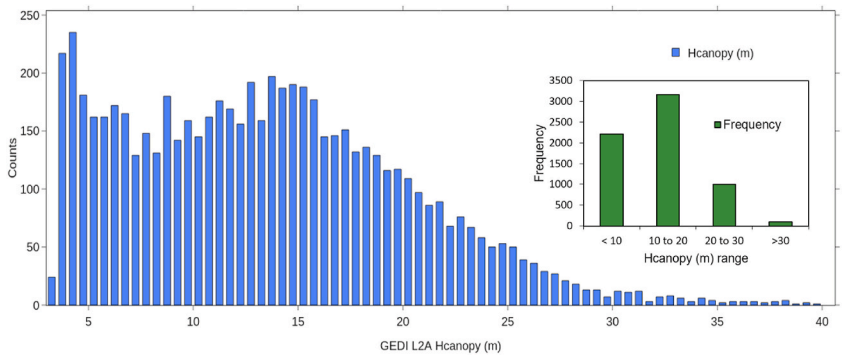

Fig. 6. Distribution and frequency of GEDI02_A extracted Hcanopy over the SWLS.

replacement rather than bootstrap samples. Two parameters, randomly selected variables (mtry) and the number of trees (ntree), were adjusted to develop the Cforest (Strobl et al., 2009; Fletcher, 2016). Xgboost model tree-based learners (Chen and Guestrin, 2016). MARS model finds a set of simple linear functions that gives the best overall prediction (https://machinelearningmastery.com/). Two tuning parameters are associated with the MARS algorithm, the degree of interactions (ndegree) and the number of retained terms (nprune) (Friedman, 1991). K-nearest neighbour (KNN) use previously memorised data to classify new data points into the target class, depending on the k nearest available points (Wu et al., 2018).

Table 5
Descriptive statistics of retrieved Hcanopy (m) from GEDI L2A over SWLS.

Descriptive statistics	Values
Mean	13.55
Standard error	0.083
Median	13.14
Mode	4.3
Standard deviation	6.73
Sample variance	45.31
Kurtosis	0.44
Skewness	0.65
Minimum	3.14
Maximum	52.68
Count	6486

Table 6
List of remote sensing data bands, band ratios and vegetation indices used in this study.

Data/predictors	Bands/variables/equation	Source/Reference
Landsat SR bands	Blue, Green, Red, NIR, SWIR1, SWIR2	ClimateEngine
Sentinel-2	Rededge-1 (B5), Rededge-2 (B6), Rededge-3 (B8), Rededge-4 (B8A)	GEE, *sentinel.esa.int*
ALOS-2/PALSAR-2	HH, HV, HH/HV	JAXA (https://www.eorc.jaxa.jp/ALOS/en/palsar_fnf/data/index.htm)
Sentinel-1 Level-1 GRD	VH, VV, VV + VH	GEE, Copernicus Open Access Hub
Sentinel-1 Level-1 SLC	Entropy, Anisotropy, Alpha	
Normalized difference vegetation index (NDVI)	$\frac{(NIR - Red)}{(NIR + Red)}$	Tucker, (1979)
Enhanced Vegetation Index (EVI)	$2.5\,\frac{NIR - RED}{(NIR + 6RED - 7.5Blue) + 1}$	Huete et al., (1997)
Chlorophyll vegetation index (CVI)	$NIR\frac{Red}{Green^2}$	Vincini et al., (2008)
Normalized difference greenness index (NDGI)	$\frac{(Green - Red)}{(Green + Red)}$	Bannari et al., (1995); Yang et al., (2019)
Normalized burn ratio SWIR2 (NBR)	$\frac{(NIR - SWIR2)}{(NIR + SWIR2)}$	Ji et al., (2011)
Normalized difference infrared index (NDII)	$\frac{(NIR - SWIR1)}{(NIR + SWIR1)}$	Ji et al., (2011)
Green difference vegetation index (GDVI)	$NIR - Green$	Sripada et al., (2006)
Modified soil adjusted vegetation index (MSAVI2)	$\frac{2NIR + 1 - \sqrt{(2NIR + 1)^2 - 8(NIR - RED)}}{2}$	Qi et al., (1994)
Difference vegetation index (DVI)	$NIR - Red$	Tucker, (1979)
Green atmospherically resistant vegetation index (GARI)	$\frac{NIR - (Green - (Blue - Red))}{NIR - (Green + (Blue - Red))}$	Gitelson et al., (1996)
Atmospherically resistant vegetation indices (ARVI)	$\frac{(NIR - (2{*}Red - Blue))}{(NIR + (2{*}Red - Blue))}$	Kaufman and Tanre, (1996)
Modified simple ratio (MSR)	$\frac{\left(\frac{NIR}{Red}\right) - 1}{\sqrt{(NIR/RED) + 1}}$	Chen,1996
Radar forest degradation index (RFDI)	$\frac{\gamma^0_{HH} - \gamma^0_{HV}}{\gamma^0_{HH} + \gamma^0_{HV}}$	Mitchard et al., (2012)
Band ratios (Landsat-8 SR bands)	Green_over_NIR, Red_over_Blue, SWIR1_over_Red, SWIR2_over_NIR, SWIR2_over_SWIR1, SWIR2_over_Green	ClimateEngine

2.4.1. Pearson correlation and variables selection

Variable selection techniques are crucial for effective ML model building by improving accuracy, low time consumption and dimensionality reduction. We tested the normal distribution of the training dataset. The Pearson correlation was used to determine the correlation among normalized predictors. Fig. 7 shows the Pearson correlation among thirty predictors and the target variable Hcanopy.

After that, we used the Boruta package in RStudio (RStudio Team, 2021) to assess the predictors importance. Boruta acts as a wrapper algorithm that initially adds randomness to the given data set by creating shadow predictors. Boruta algorithm gives better results on variable importance and is easy to analyse (Kursa and Rudnicki, 2010).

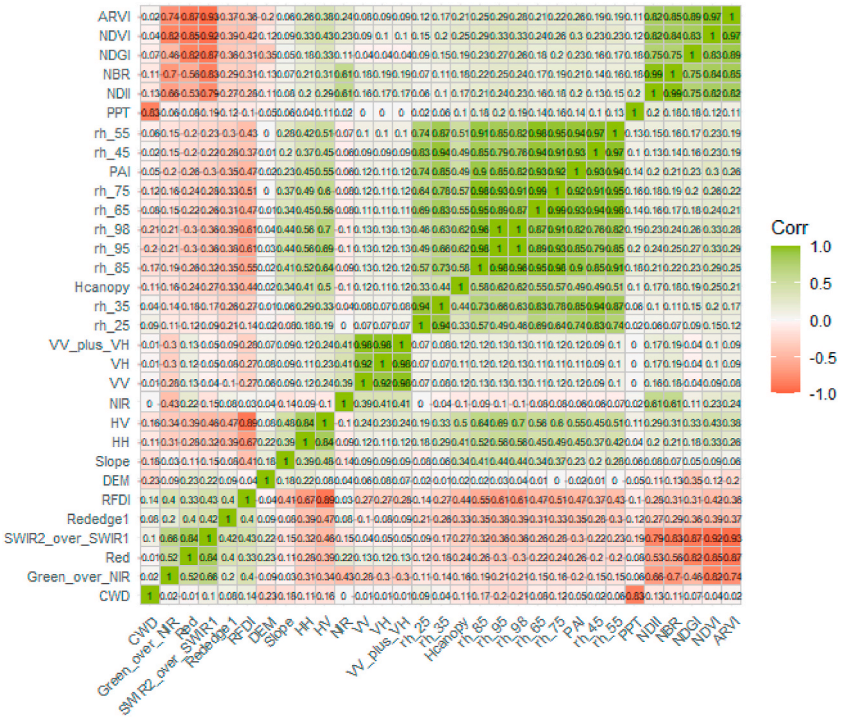

Fig. 7. Pearson correlation among predictors and target variable Hcanopy.

2.4.2. Training and testing

We followed the Pareto principle for Hcanopy data splitting into a train (80%) and test (20%) (Dunford et al., 2014) using the 'createDataPartition' function in the caret package. Hcanopy observations in train and test were 5486 and 1369, respectively. We used a powerful machine learning caret package (Kuhn, 2008) in RStudio version 1.4.1717 to train and test each ML model (RStudio Team, 2021). In the train control function, we used a 10-fold cross-validation resampling approach to evaluate trained models for unseen conditions and minimize overfitting or underfitting. We used the 'preprocess' function to center and scale train data to range the data between 0 and 1. A tuneLength value of '10' was used for hyperparameter tuning, which automatically searches for the optimal accuracy of models. Further, we assessed each predictor's relative (scaled 0 to 100) importance in the ML models.

2.4.3. Accuracy and error estimation

The trained regression model's accuracy and error estimation was performed using RMSE, R^2 and MAE metrics. These metrics are widely used for assessing model evaluation in machine learning. MAE is more robust to outliers than a mean squared error as it does not use square. RMSE, MAE and R^2 were obtained using the equations (2)–(4)

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_i)^2} \tag{2}$$

x_i is an observed data;

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\bar{x}_i - x_i| \tag{3}$$

$\bar{x}_i$ is the predicted data; 'n' is the number of observations.

RSS is the sum of squares of residuals

$$\mathrm{R}_2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} \tag{4}$$

TSS is the total sum of squares.

2.5. Resample and prediction

We compared trained model performances using resample function in RStudio to select the optimal model. In this function, we just listed the names of the trained model and ran them. The summary function gives each regression ML model's RMSE, R^2 and MAE. Low RMSE and MAE, while high R^2 was the criteria for selecting an optimal model. Finally, the prediction was performed on test data and

over SWLS raster stack layers using predict function from the caret package.

2.6. Predictive mapping and field validation

Predicted Hcanopy raster maps from each ML model were exported from RStudio in GeoTIFF format. FNF raster map was used to mask out areas in SWLS which is not under forest cover. Then, the field validation of predicted Hcanopy was performed using field observed Hcanopy. RMSE, R^2, and MAE measures were used to evaluate the performances of the optimal ML model against field data.

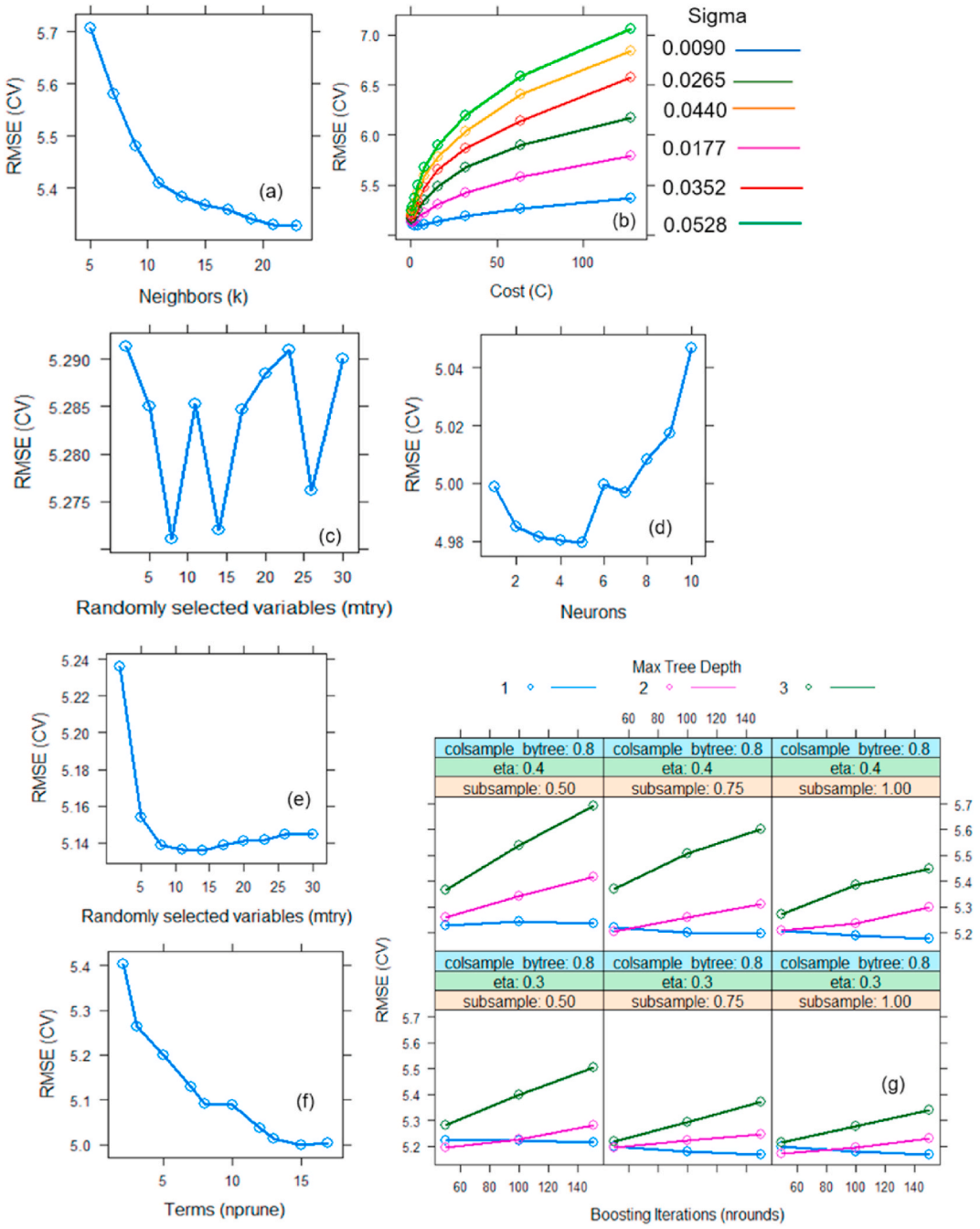

Fig. 8. Variations in the training accuracy with hyperparameters of the ML models: KNN a), SVM b), QRF c), BRNN d), Cforest e), MARS f), and Xgbtree g).

3. Results

3.1. GEDI canopy height and its accuracy

Maximum, mean and minimum Hcanopy retrieved from GEDI_02A data for SWLS was 52.68 m, 13.55 m, and 3.14 m respectively. The Hcanopy observations for less than 10 m are 2216, 10–20 m are 3165, 20–30 m are 1003, and greater than 30 m are 102. RMSE metric was used to select the final ML model accuracy based on the smallest value obtained for the model tuning parameter. The final value used for the optimal KNN model was based on tuning parameter 'k' at 23 when the RMSE = 5.32 m was the lowest (Fig. 8a). An optimal SVM trained model was selected when the RMSE value for the tuning parameter sigma is 0.00900, and the cost function is 2. The final RMSE for optimal SVM was 5.10 m (Fig. 8b). The lowest RMSE value of 5.271 m was achieved when the tuning parameter 'mtry' was 8 for the QRF model (Fig. 8c). The final value used for the optimal BRNN model was at tuning parameter 'neurons' at 5 on obtaining RMSE of 4.98 m (Fig. 8d). For optimal Cforest, tuning parameter 'mtry' at 14 was used to select the final lowest value of RMSE = 5.135 m (Fig. 8e). For the optimal MARS model, the final values used were 'nprune' at 15 and 'degree' at 1 for the lowest RMSE = 5.0 m (Fig. 8f). The final values used for the Xgbtree were at 'nrounds' is 150, 'max_depth' is 1, 'eta' is 0.3, 'gamma' is 0, 'colsample_bytree' is 0.8, 'min_child_weight' is 1, and 'subsample' is 0.75. The final value of RMSE = 5.165 m was used for optimal Xgbtree (Fig. 8g).

3.2. Variable importance

Derived predictors from Sentinel-1 polarimetric decomposition such as alpha, anisotropy and entropy show very low importance in the Boruta method; therefore, they were omitted in the initial screening. Also, highly correlated vegetative indices observed in Pearson correlation were omitted to reduce the number of predictors. Finally, ML models were run with 30 predictors, and the variables importance was derived from Boruta (Fig. 9) and individual model (Fig. 10 a-g) inbuilt importance function. Results showed that Rh_98 has the most importance, while rededge-1 was the second most important predictor for Hcanopy prediction (Table 7). ALOS-2/PALSAR-2 backscatter HV and HH are the fourth and fifth top predictors. Among Landsat-8 vegetative indices NDVI, NDGI, and ARVI have shown importance in all models to predict Hcanopy. Sentinel-1 VH backscatter is the least important predictor shown in Fig. 9. Fig. 10a–g revealed that GEDI derived RH metrics and PAI are the top important variables to predict Hcanopy. Rh_25 is the lowest importance among RH metrics for all models except the MARS algorithm. ALOS-2/PALSAR-2 backscatter HV, HH, and derived indices RFDI are non-GEDI variables that show the main importance in Hcanopy prediction. The topographic variable slope is more important in Hcanopy prediction than DEM and aspect. Among Sentinel-2 Rededge bands, only rededge-1 is an important predictor of Hcanopy in all models. Sentinel-1 backscatter VV, VH and Landsat-8 vegetative indices show low importance. Landsat-8 red band and ratio of SWIR-2 and SWIR-1 are more significant predictors than vegetative indices. Climatic variable PPT and CWD showed the importance for Hcanopy prediction in KNN, SVM, QRF and BRNN models.

3.3. Resample models

Fig. 11 shows that the KNN model has the highest, and BRNN has the smallest mean MAE. KNN has a mean MAE of 4.17 m, whereas BRNN has a mean MAE of 3.84 m. Mean MAE for QRF, Cforest, Xgbtree, MARS and SVM models is 4.00 m, 4.00 m, 4.03 m, 3.87 m and 3.87 m respectively, as shown in Table 8. The mean RMSE for the BRNN model is 4.98 m, the smallest, whereas the mean RMSE for the KNN model is 5.33 m, which is the highest among ML models. Mean RMSE for QRF, Cforest, Xgbtree, MARS and SVM models is 5.27 m, 5.14 m, 5.17 m, 5.00 m and 5.10 m, respectively. R^2 for the BRNN model is 0.454, the highest, whereas R^2 for the KNN model is 0.38, the smallest among ML models. R^2 for QRF, Cforest, Xgbtree, MARS and SVM models is 0.40, 0.42, 0.41, 0.45 and 0.45 respectively.

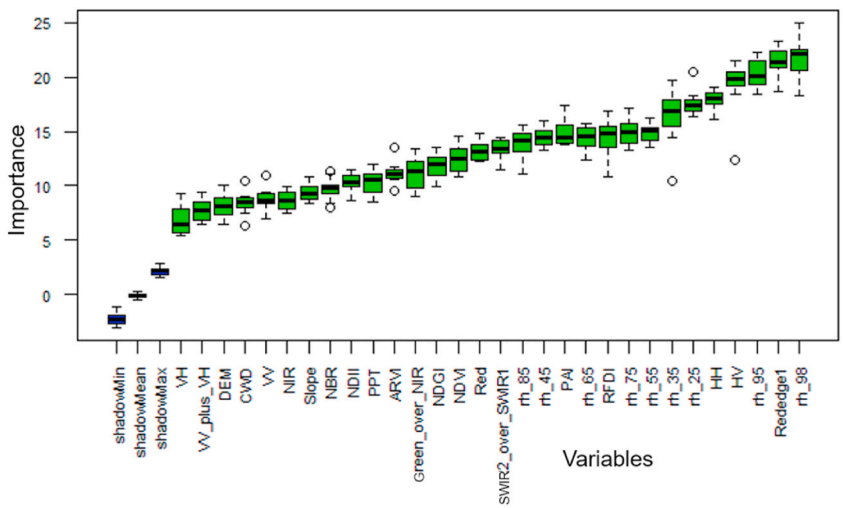

Fig. 9. Importance of predictor variables generated from the Boruta method.

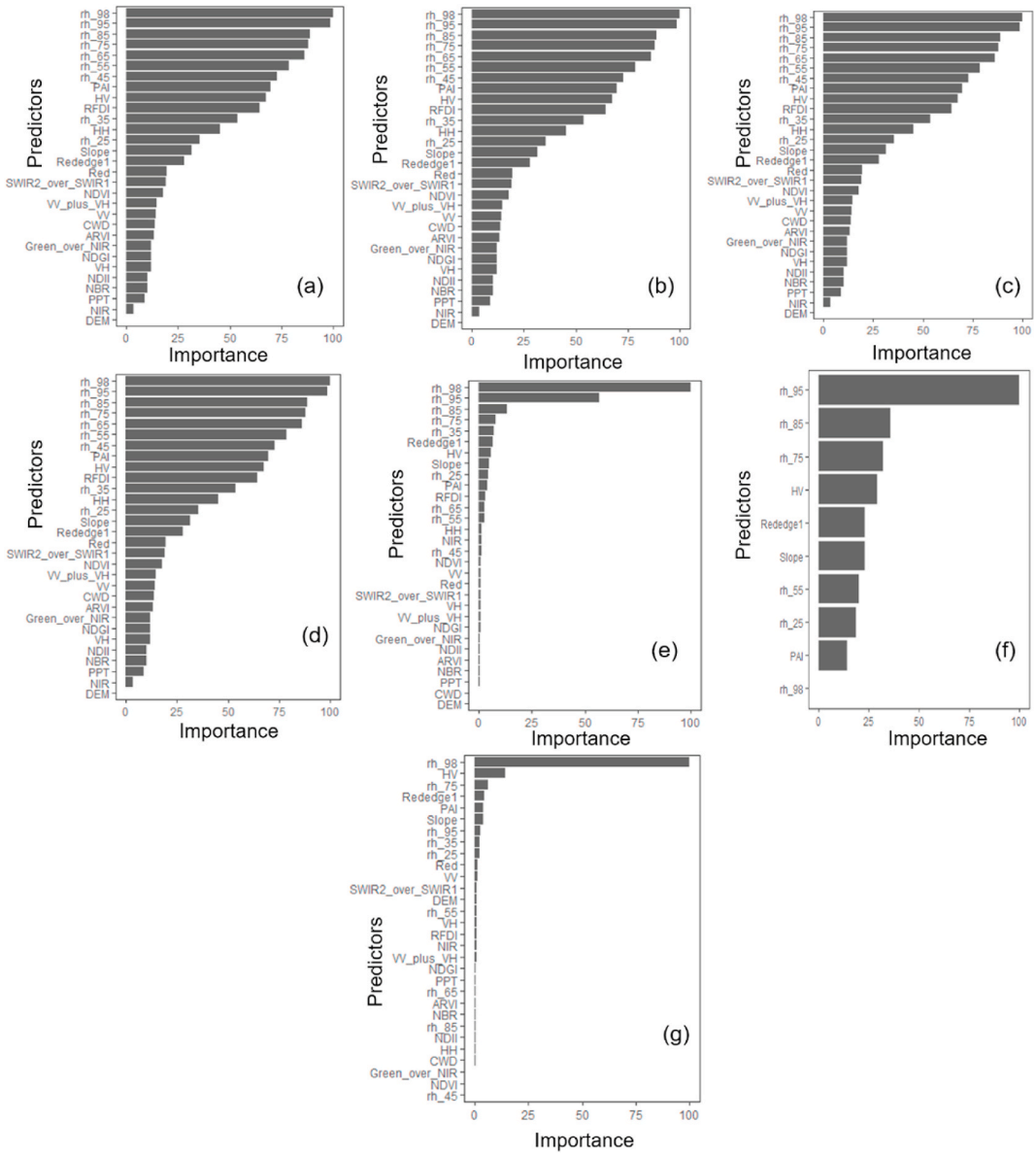

Fig. 10. Relative importance (x-axis) of predictors (y-axis) in different ML models for the prediction of Hcanopy; KNN a), SVM b), QRF c), BRNN d), Cforest e), MARS f), and Xgbtree g).

3.4. Hcanopy spatial predictive mapping

Fig. 12a–g is the predictive spatial maps of Hcanopy for the seven ML models (KNN, SVM, QRF, BRNN, Cforest, MARS and Xgbtree) over the SWLS. We categorized Hcanopy values into five sub-categories (<10 m, 10–15 m, 15–20 m, 20–25 m, and >25 m) and masked non-forest cover. It is observed that the ML models provide better prediction outcomes on the availability of a good number of sample data in that particular height range. Fig. 12a–g shows that the Hcanopy range of 10–15 m and 15–20 m is mainly predicted over SWLS as there was a good number of available true GEDI RH100 training height samples in those ranges. Less samples for tall canopies result from low prediction for tall canopies from ML models. Tall Hcanopy in the range 20–25 m and above 30 m is generally found in inner zones of dense forest in SWLS.

Hcanopy less than 10 m is mainly observed in the outer parts and edges of SWLS as plantation activities have been occurring recently. Hcanopy in the range of 10–15 m is distributed mainly in middle zones, whereas medium Hcanopy in the range of 15–20 m is mainly distributed in the inner zones of SWLS. Fig. 13a–g shows the spatial maps of Hcanopy standard deviation (SD) from the seven ML models (KNN, SVM, QRF, BRNN, Cforest, MARS and Xgbtree) over the SWLS. It is observed that SD of higher than 3.5 m is generally distributed across the edges of SWLS, while the inner dense forest zones of SWLS have SD of 0.5–1.5 m.

Table 7
List of predictors with their mean, median, minimum and maximum importance scores.

Predictors	Mean	Median	Minimum	Maximum
Rh_98	21.595	21.804	19.294	23.215
Rh_95	20.245	20.435	18.821	21.712
HV	20.126	19.622	18.268	23.877
Rededge1	20.097	20.266	18.442	21.611
Rh_35	16.707	16.686	13.894	18.456
HH	16.679	16.920	14.939	17.990
PAI	16.380	16.330	15.347	17.399
RFDI	15.972	16.811	11.773	17.830
Rh_75	14.159	14.073	12.879	15.838
Rh_85	13.741	13.474	9.683	16.722
Rh_65	13.648	13.655	11.850	15.618
Rh_45	13.573	13.580	11.679	15.184
Red	12.509	12.363	11.616	13.726
Rh_55	12.092	12.224	10.799	13.360
SWIR2_over_SWIR2	12.081	12.447	8.969	13.644
NDVI	12.005	12.139	9.228	13.380
NDGI	11.892	12.262	9.824	12.886
Green_over_NIR	11.508	12.238	7.235	13.649
ARVI	11.083	11.086	9.468	11.999
Rh_25	10.702	10.655	8.976	12.976
PPT	10.032	10.261	4.185	11.663
NBR	9.066	9.310	3.399	11.638
Slope	8.941	9.024	6.781	10.155
NDII	8.140	8.528	4.482	10.836
CWD	7.321	7.669	3.230	9.905
NIR	7.233	7.308	5.494	9.312
DEM	6.879	6.878	5.425	9.032
VV	6.717	6.912	3.353	9.851
VV_plus_VH	5.461	5.138	3.206	7.959
VH	4.652	4.650	2.674	6.206

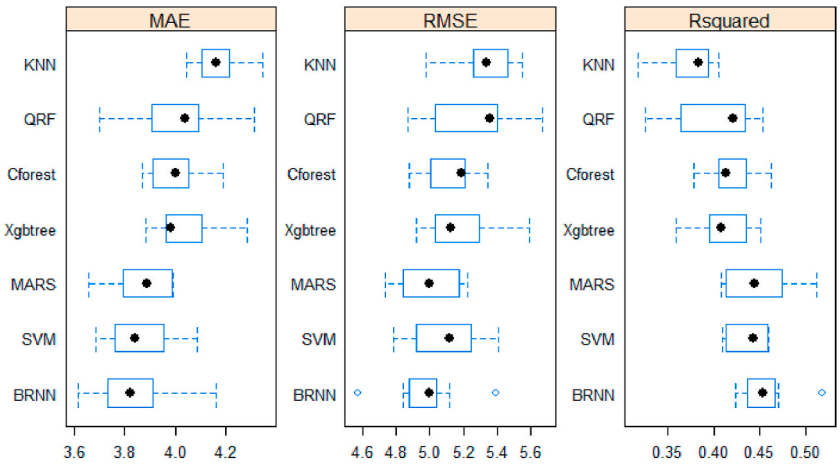

Fig. 11. Box plots show accuracy metrics (MAE, RMSE, and R^2) obtained for the trained ML models.

3.5. Models evaluation and validation

Scatterplots between observed and predicted Hcanopy shows RMSE obtained from BRNN was 4.686 m, the lowest, and from KNN was 5.035 m, the highest. RMSE obtained from SVM, QRF, Cforest, MARS, and Xgbtree was 4.778 m, 4.971 m, 4.845 m, 4.720 m and 4.856 m respectively (Fig. 14 a-g). The BRNN model gives the highest $R^2 = 0.49$, while KNN has the lowest $R^2 = 0.41$. SVM and MARS model show an R^2 value of 0.48, while QRF, Cforest, and Xgbtree show R^2 values of 0.44, 0.45 and 0.45 respectively. The MAE obtained for KNN, SVM, QRF, BRNN, Cforest, MARS, and Xgbtree was 3.94 m, 3.62 m, 3.74 m, 3.66 m, 3.76 m, 3.66 m and 3.79 m respectively. Validation scatterplot between field and BRNN predicted Hcanopy shows R^2 of 0.55, RMSE = 3.94 m, and MAE = 3.12 m (Fig. 14h).

Table 8
Accuracy metrics (MAE, RMSE, and R^2) values of Hcanopy for the trained ML models.

Model	Minimum	Median	Mean	Maximum
	MAE (m)			
QRF	3.70	4.04	4.00	4.31
Cforest	3.87	4.00	4.00	4.19
Xgbtree	3.88	3.98	4.03	4.28
MARS	3.66	3.89	3.87	3.99
BRNN	3.62	3.82	3.84	4.16
SVM	3.69	3.84	3.87	4.08
KNN	4.04	4.16	4.17	4.35
	RMSE (m)			
QRF	4.87	5.36	5.27	5.67
Cforest	4.88	5.19	5.14	5.34
Xgbtree	4.92	5.12	5.17	5.59
MARS	4.74	5.00	5.00	5.23
BRNN	4.58	5.00	4.98	5.39
SVM	4.79	5.12	5.10	5.41
KNN	4.98	5.34	5.33	5.55
	R^2			
QRF	0.33	0.42	0.40	0.45
Cforest	0.38	0.41	0.42	0.46
Xgbtree	0.36	0.41	0.41	0.45
MARS	0.41	0.44	0.45	0.51
BRNN	0.42	0.45	0.45	0.52
SVM	0.41	0.44	0.44	0.46
KNN	0.32	0.38	0.38	0.41

4. Discussion

Integrating spaceborne LiDAR GEDI data with remote sensing optical and SAR imagery using the ML provides the continuous mapping of Hcanopy. Therefore, we attempted to predict Hcanopy over the mixed tropical forests of SWLS in India, using ML models on GEDI retrieved true Hcanopy and a set of predictors derived from multisensor remote sensing datasets. Retrieved Hcanopy data from GEDI shows that the SWLS has dominant forest Hcanopy in 10 m–20 m. Hcanopy samples from tall canopies (>25 m) are low, and their predicted distribution from ML is also low. We obtained best prediction accuracy of RMSE = 4.686 m, MAE = 3.66 m and R^2 = 0.49 from BRNN model and the validation accuracy of RMSE = 3.94 m, MAE = 3.12 m, and R^2 = 0.55 against field-measured Hcanopy.

Variable importance shows that the GEDI derived metrics and PAI, which were interpolated for SWLS using kriging, have high importance score; therefore, it plays an important role in Hcanopy prediction. ALOS-2/PALSAR-2 backscatters HH, HV and derived indices RFDI has high importance among non-GEDI products. García et al. (2018) also found that in the temperate broadleaf and mixed forests, HV backscatter from ALOS-2/PALSAR-2 was the most important variable sensitive to forest structure volumetric scattering from foliage and stem branches. According to Chen et al. (2021b), the integrated use of LiDAR retrieved predictors can minimize the saturation issue from optical and SAR C band sensor imageries. This study also used GEDI predictors such as RH metrics and PAI, which improve model prediction accuracy. This study found that Sentinel-1 C band backscatters VV, VH, and derived VV + VH are less influential in Hcanopy estimation due to their inability to penetrate dense multi-layered forests and scattering energy. Also, among Sentinel-2 Rededge bands, only the rededge-1 band has shown importance in Hcanopy prediction. Landsat-8 bands, vegetative indices, and band ratios are influential among non-GEDI metrics in our analysis, improving final prediction accuracy. The poor importance of Landsat-8 retrieved variables is possibly due to forest structure variability and heterogeneous forest, terrain shadow effects, and moisture conditions.

Our ML models did not vary significantly in training accuracy (RMSE, R^2 and MAE). The lowest mean RMSE value of 5 m was obtained from the MARS and BRNN trained model, while the highest mean RMSE value of 5.36 m was obtained from trained QRF. The mean correlation coefficient (R^2 = 0.45) is maximum for BRNN, while for MARS and SVM the R^2 is 0.44. The trained KNN model has shown the lowest R^2 of 0.38. These metrics revealed that the training accuracy did not vary significantly among these models except KNN.

Models evaluation also shows that the ML models did not vary significantly in their prediction accuracy. BRNN is used as an optimal model for Hcanopy prediction due to its low RMSE = 4.686 m, MAE = 3.66 m, and high R^2 = 0.49 than other models. However, SVM (RMSE = 4.686 m, MAE = 3.62 m, R^2 = 0.48 and MARS (RMSE = 4.72 m, MAE = 3.66 m, R^2 = 0.48) also showed similar prediction accuracy. Spatial mapping of Hcanopy over SWLS demonstrates that the tall and medium Hcanopy vegetation are majorly present in the middle and inner zones of SWLS. Short-range (<10 m) Hcanopy is mainly present in the outer zones of SWLS due to the plantation activities in SWLS. Validation results show that the field measured and BRNN model predicted Hcanopy correlates with R^2 of 0.55, satisfactory for mixed tropical forests using sparsely distributed spaceborne LiDAR. A study by Lang et al., (2021b) used CNN deep learning approach against the testing GEDI data and obtained RMSE of 6.3 m and MAE of 4.6 m. Potapov et al., (2021) obtained a validation accuracy (RMSE = 6.6 m; MAE = 4.45 m and R^2 = 0.62) on mapping global forest Hcanopy with GEDI and Landsat-8 using

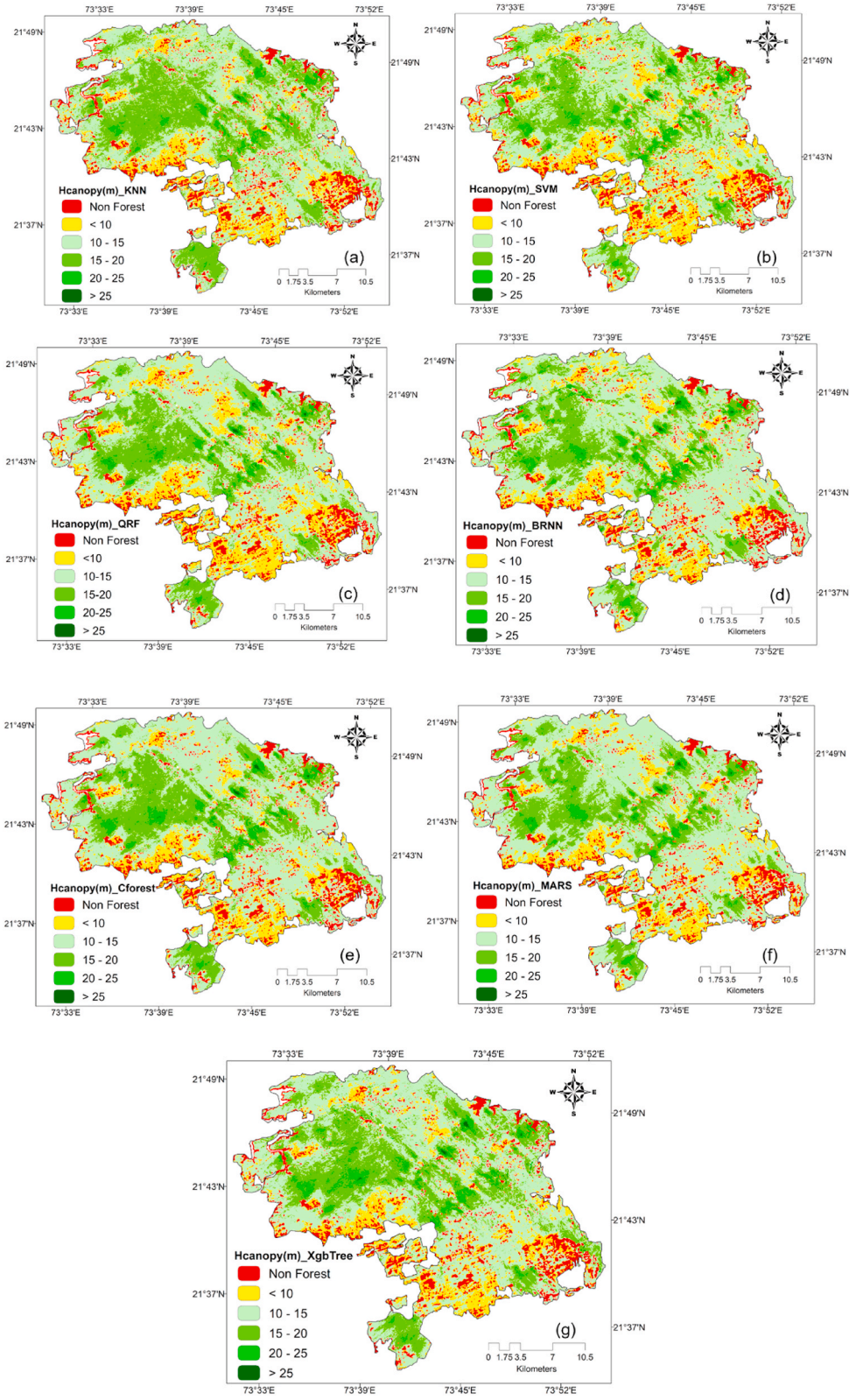

Fig. 12. Spatial mapping of Hcanopy over the SLWS using ML models: KNN a), SVM b), QRF c), BRNN d), Cforest e), MARS f), Xgbtree g).

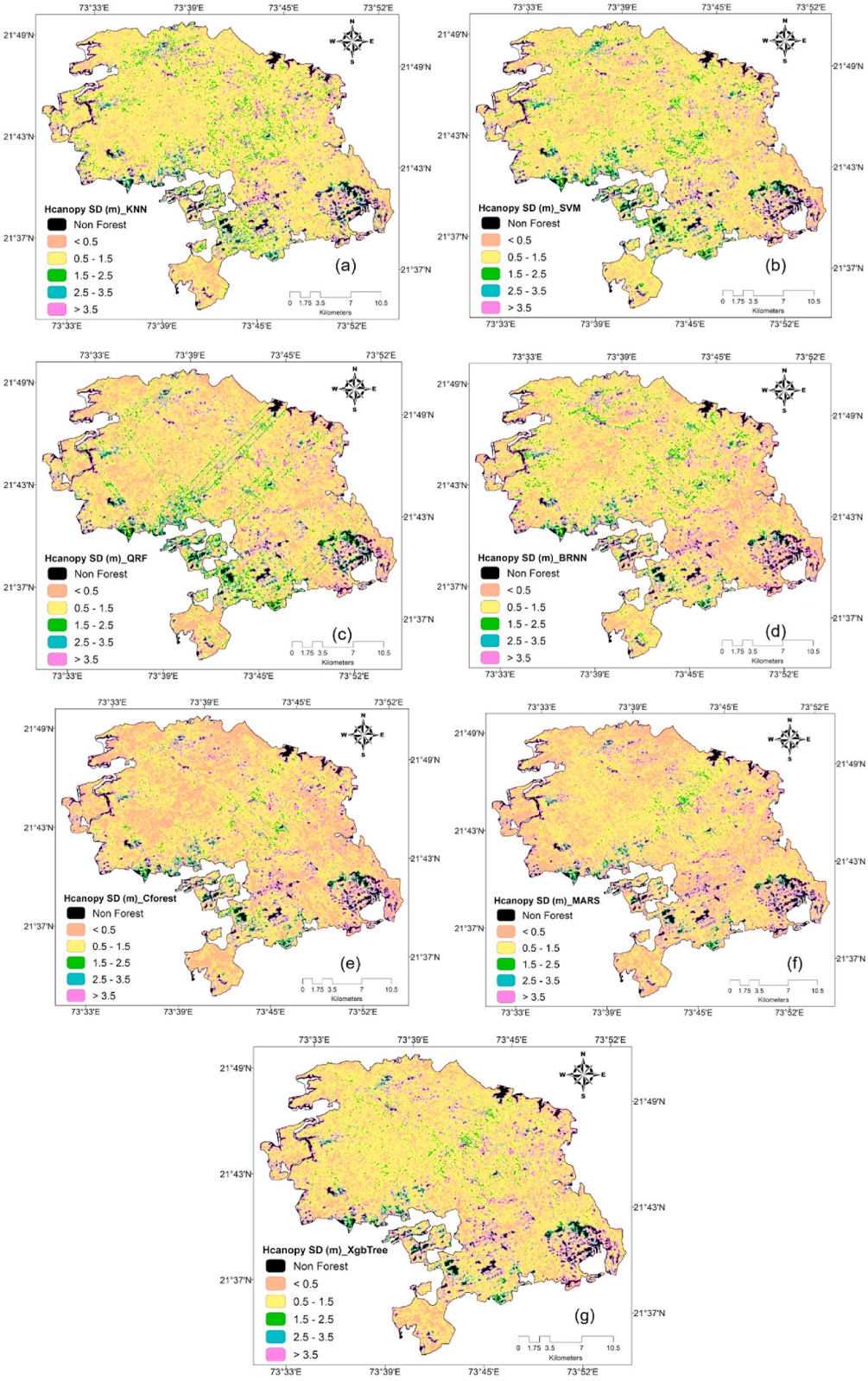

Fig. 13. Standard deviation of predicted Hcanopy over the SLWS for ML models: KNN a), SVM b), QRF c), BRNN d), Cforest e), MARS f), and Xgbtree g).

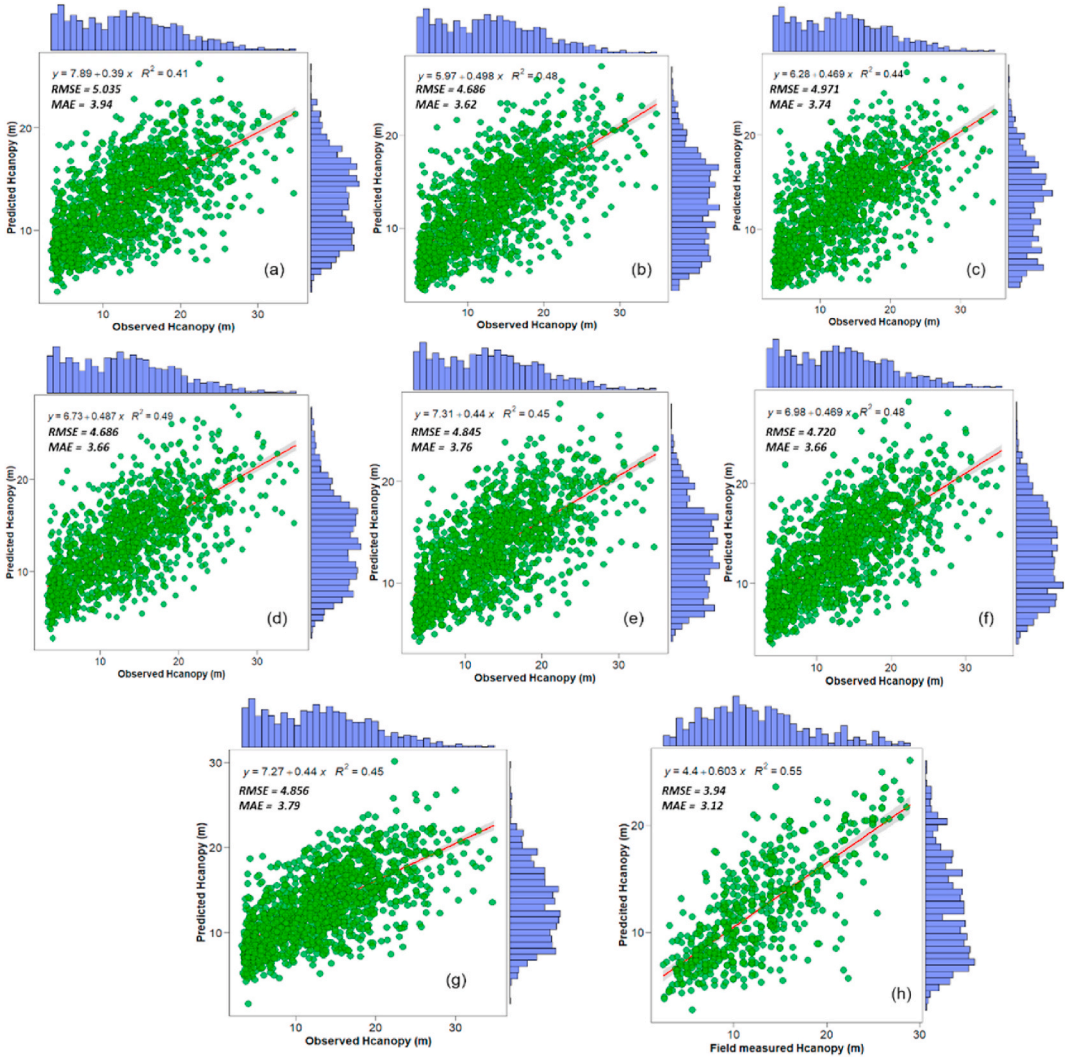

Fig. 14. Scatterplots between observed and predicted Hcanopy for KNN (a), SVM (b), QRF (c), BRNN (d), Cforest (e), MARS (f) and Xgbtree (g), and field measured versus BRNN predicted Hcanopy (h); and their accuracy metrics R^2, RMSE (m) and MAE (m).

regression trees.

The moderate prediction accuracy from spaceborne LiDAR can be from many factors in the perspective of the current study. It may be due to sloppy terrain and surface roughness, mixed species and structures variability, complex biophysical conditions, and a lack of complex predictors that could be more useful in better Hcanopy prediction accuracy. Previous studies showed that prediction from LiDAR could improve by including quad-pole SAR data to derive complex polarimetric decomposition forest structural parameters, which would be more beneficial to predicting Hcanopy than just by taking backscatters. Previous studies also commented that the SAR and its resolution in forest structure measurements had been impacted by the spatial variability in forests and their structures (Saatchi et al., 2011; García et al., 2018). SAR has decreased sensitivity in over-dense forests due to less penetration, absorption, and energy scattering (Saatchi and Moghaddam, 2000; García et al., 2018). Pourshamsi et al. (2021) used polarimetric synthetic aperture radar (PolSAR) decomposition parameters with airborne LiDAR canopy height in heterogeneous forests using Random forest, decisions trees, and SVM, which obtained a validation accuracy of $R^2 = 0.70$ and RMSE = 10 m. Staben et al., (2018) used the random forest to predict mean LiDAR Hcanopy from Landsat-5 and Landsat-7. Their validation on independent data showed an overall accuracy with $R^2 = 0.53$, and RMSE = 2.8 m for Landsat-5 and $R^2 = 0.49$, RMSE = 2.8 m for Landsat-7, similar to our results. It is generally found that the ML model's performance accuracy increases with using more spatial predictors and training data observations; however, saturation reached a particular level (Urbazaev et al., 2018). Moreover, using a large number of predictors will increase the computational time of machines, and it is also not always true that increasing predictors will significantly increase accuracy. Pourshamsi et al. (2021) also revealed that for effective Hcanopy estimation over a heterogeneous forest, the predictors used have a specific set of information on the forests of the study site.

The study site is a relatively small (675 km^2) area of mixed tropical forests. However, this study provides valuable insights about the

used predictors, training and validating model accuracies which will be helpful for researchers to use spaceborne GEDI for regional and large-scale mapping with the integration of multisensor imagery using machine learning. Moreover, from an ecological point of view, this study is the first to map the Hcanopy of SWLS forests and would be helpful in assessing above-ground biomass and CO_2 sequestration rates. Further, Hcanopy mapping is helpful for the better management of SWLS forests, prevention of forest loss, biodiversity conservation and forest fire risks.

5. Conclusion

Integrated use of sparsely distributed Spaceborne LiDAR GEDI waveform Hcanopy data and synoptic remote sensing imageries is useful in the continuous mapping of tropical mixed forests Hcanopy. In this study, we compared the performances of ML models to predict Hcanopy in the tropical mixed forests of SWLS. Our used ML models showed a slightly low but consistent prediction accuracy. However, we used BRNN ($R^2 = 0.49$, RMSE = 4.686 m, MAE = 3.66 m) as the optimal model for validation with field observations due to its slightly better performance than other models. The independent field data validation for the predicted forest Hcanopy shows accuracy ($R^2 = 0.55$, RMSE = 3.94 m, MAE = 3.12 m, n = 550). We identified that GEDI derived RH metrics have high importance in predicting Hcanopy. Mixed tropical forests have high species variability and structural complexity. Therefore, choosing the complex predictors from SAR L-band data and high-resolution optical and vegetative indices is suggested. Further, deep learning models could also be explored for accuracy improvement. The mapping of Hcanopy by integrating GEDI LiDAR Hcanopy data and multisensor imagery using ML in the mixed tropical forest will be helpful proper management of SWLS forests.

Data availability statement

The data supporting this study will be shared upon reasonable request to the corresponding author.

Funding source

This research did not receive any specific funding

Ethical statement

All ethical practices have been followed in relation to the development, writing, and publication of the article.

Author contributions

Rajit Gupta: Software, Methodology design, Modeling, Data Analysis, Reviewing, Writing and Editing, **Laxmi Kant Sharma**: Conceptualization, Supervision, Formal analysis, Reviewing, and Editing.

Declaration of competing interest

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

Acknowledgements

We are highly thankful to the Central University of Rajasthan for the DST-FIST-funded RS & GIS Lab in the Department of Environmental science. The first author is thankful to the University Grants Commission (UGC) for the UGC NET-JRF fellowship (Ref no. 3551/(NET-JAN2017). We also thank forest officials and field staff of SWLS for their support during field surveys. Finally, we would like to thank the anonymous reviewers and editors for their important suggestions and comments.

References

Abatzoglou, J.T., Dobrowski, S.Z., Parks, S.A., Hegewisch, K.C., 2018. Terraclimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958–2015. Sci. Data 5, 170191. https://doi.org/10.1038/sdata.2017.191.

Bannari, A., Morin, D., Bonn, F., Huete, A., 1995. A review of vegetation indices. Rem. Sens. Rev. 13 (1–2), 95–120. https://doi.org/10.1080/02757259509532298.

Champion, H.G., Seth, S.K., 1968. A Revised Survey of the Forest Types of India. Manager of publications.

Chen, J.M., 1996. Evaluation of vegetation indices and a modified simple ratio for boreal applications. Can. J. Remote. Sens. 22 (3), 229–242. https://doi.org/10.1080/07038992.1996.10855178.

Chen, T., Guestrin, C., 2016. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 785–794. San Francisco, CA. arXiv, 1–6.

Chen, W., Zheng, Q., Xiang, H., Chen, X., Sakai, T., 2021a. Forest canopy height estimation using polarimetric interferometric synthetic aperture radar (PolInSAR) technology based on full-polarized ALOS/PALSAR data. Remote. Sens. 13, 174. https://doi.org/10.3390/rs13020174.

Chen, L., Ren, C., Zhang, B., Wang, Z., Liu, M., Man, W., Liu, J., 2021b. Improved estimation of forest stand volume by the integration of GEDI LiDAR data and multisensor imagery in the Changbai Mountains Mixed forests Ecoregion (CMMFE), northeast China. Int. J. Appl. Earth Obs. Geoinf. 100, 102326 https://doi.org/10.1016/j.jag.2021.102326.

Dubayah, R., Blair, J.B., Goetz, S., Fatoyinbo, L., Hansen, M., Healey, S., Hofton, M., Hurtt, G., Kellner, J., Luthcke, S., Armston, J., Tang, H., Duncanson, L., Hancock, S., Jantz, P., Marselis, S., Patterson, P.L., Qi, W., Silva, C., 2020a. The global ecosystem dynamics investigation: high-resolution laser ranging of the earth's forests and topography. Sci. Remote Sens 1, 100002. https://doi.org/10.1016/j.srs.2020.100002.

Dubayah, R., Hofton, M., Blair, J., Armston, J., Tang, H., Luthcke, S., 2020b. GEDI L2A elevation and height metrics data global footprint level v001. NASA EOSDIS Land Processes DAAC (Accessed on 16.10.2020).

Dunford, R., Su, Q., Tamang, E., 2014. The Pareto principle. Plymouth Student Sci. 7 (1), 140–148.

earthdata nasa. https://git.earthdata.nasa.gov/projects/LPDUR/repos/gedi-tutorials/browse.
earthengine. https://earthengine.google.com/.
eorcjaxa. https://www.eorc.jaxa.jp/ALOS/en/palsar_fnf/data/index.htm.
Fayad, I., Ienco, D., Baghdadi, N., Gaetano, R., Alvares, C.A., Stape, J.L., Scolforo, H.F., Le Maire, G., 2021. A CNN-based approach for the estimation of canopy heights and wood volume from GEDI waveforms. Remote Sens. Environ. 265, 112652 https://doi.org/10.1016/j.rse.2021.112652.
Fletcher, R.S., 2016. Using vegetation indices as input into random forest for soybean and weed classification. Am. J. Plant Sci. 7, 2186–2198. https://doi.org/10.4236/ajps.2016.715193.
Foresee, F.D., Hagan, M.T., 1997. Gauss-Newton approximation to Bayesian regularization", choosing initial values of the adaptive weights. Proceedings of the IJCNN 3, 21–26.
Friedman, J.H., 1991. Multivariate adaptive regression splines. Ann. Stat. 33, 1–67. https://doi.org/10.1214/aos/1176347963.
García, M., Saatchi, S., Ustin, S., Balzter, H., 2018. Modelling forest canopy height by integrating airborne LiDAR samples with satellite Radar and multispectral imagery. Int. J. Appl. Earth Obs. Geoinf. 66, 159–173. https://doi.org/10.1016/j.jag.2017.11.017.
Gitelson, A.A., Kaufman, Y.J., Merzlyak, M.N., 1996. Use of a green channel in remote sensing of global vegetation from EOS-MODIS. Remote Sens. Environ. 58 (3), 289–298. https://doi.org/10.1016/S0034-4257(96)00072-7.
Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google earth engine: planetary-scale geospatial analysis for everyone. Remote Sens. Environ. 202, 18–27. https://doi.org/10.1016/j.rse.2017.06.031.
Gupta, R., Sharma, L.K., 2020. Efficacy of Spatial Land Change Modeler as a forecasting indicator for anthropogenic change dynamics over five decades: a case study of Shoolpaneshwar Wildlife Sanctuary, Gujarat, India. Ecol. Indicat. 112, 106171 https://doi.org/10.1016/j.ecolind.2020.106171.
Huete, A.R., Liu, H.Q., Batchily, K.V., van Leeuwen, W., 1997. A comparison of vegetation indices over a global set of TM images for EOS-MODIS. Remote Sens. Environ. 59 (3), 440–451. https://doi.org/10.1016/S0034-4257(96)00112-5.
Huntington, J.L., Hegewisch, K.C., Daudert, B., Morton, C.G., Abatzoglou, J.T., McEvoy, D.J., Erickson, T., 2017. Climate engine: cloud computing and visualization of climate and remote sensing data for advanced natural resource monitoring and process understanding. Bull. Am. Meteorol. Soc. 98 (11), 2397–2410. https://doi.org/10.1175/BAMS-D-15-00324.1.
Ji, L., Zhang, L., Wylie, B.K., Rover, J., 2011. On the terminology of the spectral vegetation index (NIR− SWIR)/(NIR+ SWIR). Int. J. Remote Sens. 32 (21), 6901–6909. https://doi.org/10.1080/01431161.2010.510811.
Jiang, F., Zhao, F., Ma, K., Li, D., Sun, H., 2021. Mapping the forest canopy height in Northern China by synergizing ICESat-2 with sentinel-2 using a stacking algorithm. Remote Sens. 13 (8), 1535. https://doi.org/10.3390/rs13081535.
Kaufman, Y.J., Tanre, D., 1996. Strategy for direct and indirect methods for correcting the aerosol effect on remote sensing: from AVHRR to EOS-MODIS. Remote Sens. Environ. 55 (1), 65–79. https://doi.org/10.1016/0034-4257(95)00193-X.
Kuhn, M., 2008. Building predictive models in R using the caret package. J. Stat. Software 28 (1), 1–26. https://doi.org/10.18637/jss.v028.i05.
Kursa, M.B., Rudnicki, W.R., 2010. Feature selection with the Boruta package. J. Stat. Software 36 (11), 1–13. https://doi.org/10.18637/jss.v036.i11.
Lang, N., Kalischek, N., Armston, J., Schindler, K., Dubayah, R., Wegner, J.D., 2021a. Global Canopy Height Estimation with GEDI LIDAR Waveforms and Bayesian Deep Learning arXiv preprint arXiv:2103.03975.
Lang, N., Schindler, K., Wegner, J.D., 2021b. High Carbon Stock Mapping at Large Scale with Optical Satellite Imagery and Spaceborne LIDAR, 07431 arXiv preprint arXiv:2107.
Larue, E.A., Wagner, F.W., Fei, S., Atkins, J.W., Fahey, R.T., Gough, C.M., Hardiman, B.S., 2020. Compatibility of aerial and terrestrial LiDAR for quantifying forest structural diversity. Remote Sens. 12 (9), 1407. https://doi.org/10.3390/rs12091407.
Lefsky, M.A., Cohen, W.B., Acker, S.A., Parker, G.G., Spies, T.A., Harding, D., 1999. LiDAR remote sensing of the canopy structure and biophysical properties of Douglas-fir western hemlock forests. Remote Sens. Environ. 70 (3), 339–361. https://doi.org/10.1016/S0034-4257(99)00052-8.
Lin, C., Labzovskii, L.D., Mak, H.W.L., Fung, J.C., Lau, A.K., Kenea, S.T., et al., 2020. Observation of PM2.5 using a combination of satellite remote sensing and low-cost sensor network in Siberian urban areas with limited reference monitoring. Atmos. Environ. 227, 117410 https://doi.org/10.1016/j.atmosenv.2020.117410.
lpdaac. https://lpdaac.usgs.gov/.
lpdaacsvc. https://lpdaacsvc.cr.usgs.gov/services/gedifinder.
machinelearningmastery. https://machinelearningmastery.com/.
Maxwell, A.E., Warner, T.A., Fang, F., 2018. Implementation of machine-learning classification in remote sensing: an applied review. Int. J. Rem. Sens. 39 (9), 2784–2817. https://doi.org/10.1080/01431161.2018.1433343.
Meinshausen, N., Ridgeway, G., 2006. Quantile regression forests. J. Mach. Learn. Res. 7 (6).
Mitchard, E.T., Saatchi, S.S., White, L.J.T., Abernethy, K.A., Jeffery, K.J., Lewis, S.L., Collins, M., Lefsky, M.A., Leal, M.E., Woodhouse, I.H., Meir, P., 2012. Mapping tropical forest biomass with radar and spaceborne LiDAR in Lopé National Park, Gabon: overcoming problems of high biomass and persistent cloud. Biogeosciences 9 (1), 179–191. https://doi.org/10.5194/bg-9-179-2012.
Nguyen, D., Widrow, B., 1990. Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights. In: 1990 IJCNN International Joint Conference on Neural Networks. IEEE, pp. 21–26.
Ota, T., Ahmed, O.S., Franklin, S.E., Wulder, M.A., Kajisa, T., Mizoue, N., Yoshida, S., Takao, G., Hirata, Y., Furuya, N., Sano, T., 2014. Estimation of airborne lidar-derived tropical forest canopy height using landsat time series in Cambodia. Remote Sens. 6 (11), 10750–10772. https://doi.org/10.3390/rs61110750.
Pérez-Rodríguez, P., Gianola, D., Weigel, K.A., Rosa, G.J.M., Crossa, J., 2013. An R package for fitting Bayesian regularized neural networks with applications in animal breeding. J. Anim. Sci. 91 (8), 3522–3531. https://doi.org/10.2527/jas.2012-6162.
Potapov, P., Li, X., Hernandez-Serna, A., Tyukavina, A., Hansen, M.C., Kommareddy, A., Pickens, A., Turubanova, S., Tang, H., Silva, C.E., Armston, J., 2021. Mapping global forest canopy height through integration of GEDI and Landsat data. Remote Sens. Environ. 253, 112165 https://doi.org/10.1016/j.rse.2020.112165.
Pourshamsi, M., Xia, J., Yokoya, N., Garcia, M., Lavalle, M., Pottier, E., Balzter, H., 2021. Tropical forest canopy height estimation from combined polarimetric SAR and LiDAR using machine-learning. ISPRS J. Photogrammetry Remote Sens. 172, 79–94. https://doi.org/10.1016/j.isprsjprs.2020.11.008.
Qi, W., Dubayah, R.O., 2016. Combining Tandem-X InSAR and simulated GEDI lidar observations for forest structure mapping. Remote Sens. Environ. 187, 253–266. https://doi.org/10.1016/j.rse.2016.10.018.
Qi, J., Chehbouni, A., Huete, A.R., Kerr, Y.H., Sorooshian, S., 1994. A modified soil adjusted vegetation index. Remote Sens. Environ. 48 (2), 119–126. https://doi.org/10.1016/0034-4257(94)90134-1.
Rishmawi, K., Huang, C., Zhan, X., 2021. Monitoring key forest structure attributes across the conterminous United States by integrating GEDI LiDAR measurements and VIIRS Data. Remote Sens. 13 (3), 442. https://doi.org/10.3390/rs13030442.
Rodrigues, M., de la Riva, J., 2014. An insight into machine-learning algorithms to model human-caused wildfire occurrence. Environ. Model. Software 57, 192–201. https://doi.org/10.1016/j.envsoft.2014.03.003.
RStudio Team, 2021. RStudio. Integrated Development for R. RStudio, PBC, Boston, MA. URL. http://www.rstudio.com/.
Saatchi, S.S., Moghaddam, M., 2000. Estimation of crown and stem water content and biomass of boreal forest using polarimetric SAR imagery. IEEE Trans. Geosci. Rem. Sens. 38 (2), 697–709. https://doi.org/10.1109/36.841999.
Saatchi, S., Marlier, M., Chazdon, R.L., Clark, D.B., Russell, A.E., 2011. Impact of spatial variability of tropical forest structure on RADAR estimation of above-ground biomass. Remote Sens. Environ. 115 (11), 2836–2849. https://doi.org/10.1016/j.rse.2010.07.015.
scihub. https://scihub.copernicus.eu/.
Sharma, L.K., Gupta, R., Pandey, P.C., 2021. Future aspects and potential of the remote sensing technology to meet the natural resource needs. In: Advances in Remote Sensing for Natural Resource Monitoring. John Wiley & Sons, Ltd., Hoboken, NJ, USA, ISBN 978-1-119-61601-6, pp. 445–464.
Simard, M., Pinto, N., Fisher, J.B., Baccini, A., 2011. Mapping forest canopy height globally with spaceborne LiDAR. J. Geophys. Res. 116 (G4) https://doi.org/10.1029/2011JG001708.

Spracklen, B., Spracklen, D.V., 2021a. Synergistic use of sentinel-1 and sentinel-2 to map natural forest and Acacia plantation and stand ages in North-Central Vietnam. Remote Sens. 13 (2), 185. https://doi.org/10.3390/rs13020185.

Spracklen, B., Spracklen, D.V., 2021b. Determination of structural characteristics of old-growth forest in Ukraine using spaceborne LiDAR. Remote Sens. 13 (7), 1233. https://doi.org/10.3390/rs13071233.

Sripada, R.P., Heiniger, R.W., White, J.G., Meijer, A.D., 2006. Aerial color infrared photography for determining early in-season nitrogen requirements in corn. Agron. J. 98 (4), 968–977. https://doi.org/10.2134/agronj2005.0200.

Staben, G., Lucieer, A., Scarth, P., 2018. Modelling LiDAR derived tree canopy height from Landsat TM, ETM+ and OLI satellite imagery—a machine learning approach. Int. J. Appl. Earth Obs. Geoinf. 73, 666–681. https://doi.org/10.1016/j.jag.2018.08.013.

Stojanova, D., Panov, P., Gjorgjioski, V., Kobler, A., Džeroski, S., 2010. Estimating vegetation height and canopy cover from remotely sensed data with machine learning. Ecol. Inf. 5 (4), 256–266. https://doi.org/10.1016/j.ecoinf.2010.03.004.

Strobl, C., Hothorn, S., Zeileis, A., 2009. Party on! A New, Conditional Variable Importance Measure for Random Forests Available in the Party Package. Department of Statistics, University of Munich, Munich. Technical Report Number 050.

Tang, H., Armston, J., 2019. Algorithm Theoretical Basis Document (ATBD) for GEDI L2B Footprint Canopy Cover and Vertical Profile Metrics.

Tang, H., Armston, J., Hancock, S., Marselis, S., Goetz, S., Dubayah, R., 2019. Characterizing global forest canopy cover distribution using spaceborne LiDAR. Remote Sens. Environ. 231, 111262 https://doi.org/10.1016/j.rse.2019.111262.

Tucker, C.J., 1979. Red and photographic infrared linear combinations for monitoring vegetation. Remote Sens. Environ. 8, 127–150. https://doi.org/10.1016/0034-4257(79)90013-0.

Urbazaev, M., Thiel, C., Cremer, F., Dubayah, R., Migliavacca, M., Reichstein, M., Schmullius, C., 2018. Estimation of forest above-ground biomass and uncertainties by integration of field measurements, airborne LiDAR, and SAR and optical satellite data in Mexico. Carbon Bal. Manag. 13 (1), 1–20. https://doi.org/10.1186/s13021-018-0093-5.

Vincini, M., Frazzi, E., D'Alessio, P., 2008. A broad-band leaf chlorophyll vegetation index at the canopy scale. Precis. Agric. 9, 303–319. https://doi.org/10.1007/s11119-008-9075-z.

Wang, M., Sun, R., Xiao, Z., 2018. Estimation of forest canopy height and above-ground biomass from spaceborne LiDAR and Landsat imageries in Maryland. Rem. Sens. 10 (2), 344. https://doi.org/10.3390/rs10020344.

Wang, H., Seaborn, T., Wang, Z., Caudill, C.C., Link, T.E., 2021. Modeling tree canopy height using machine learning over mixed vegetation landscapes. Int. J. Appl. Earth Obs. Geoinf. 101, 102353 https://doi.org/10.1016/j.jag.2021.102353.

Wu, N., Zhang, C., Bai, X., Du, X., He, Y., 2018. Discrimination of Chrysanthemum varieties using hyperspectral imaging combined with a deep convolutional neural network. Molecules 23 (11). https://doi.org/10.3390/molecules23112831.

Yang, W., Kobayashi, H., Wang, C., Shen, M., Chen, J., Matsushita, B., Tang, Y., Kim, Y., Bret-Harte, M.S., Zona, D., Oechel, W., 2019. A semi-analytical snow-free vegetation index for improving estimation of plant phenology in tundra and grassland ecosystems. Remote Sens. Environ. 228, 31–44. https://doi.org/10.1016/j.rse.2019.03.028.

Zhang, Z., Cao, L., She, G., 2017. Estimating forest structural parameters using canopy metrics derived from airborne LiDAR data in subtropical forests. Remote Sens. 9 (9), 940. https://doi.org/10.3390/rs9090940.

Zhang, Y., Ling, F., Foody, G.M., Ge, Y., Boyd, D.S., Li, X., et al., 2019. Mapping annual forest cover by fusing PALSAR/PALSAR-2 and MODIS NDVI during 2007-2016. Remote Sens. Environ. 224, 74–91. https://doi.org/10.1016/j.rse.2019.01.038.