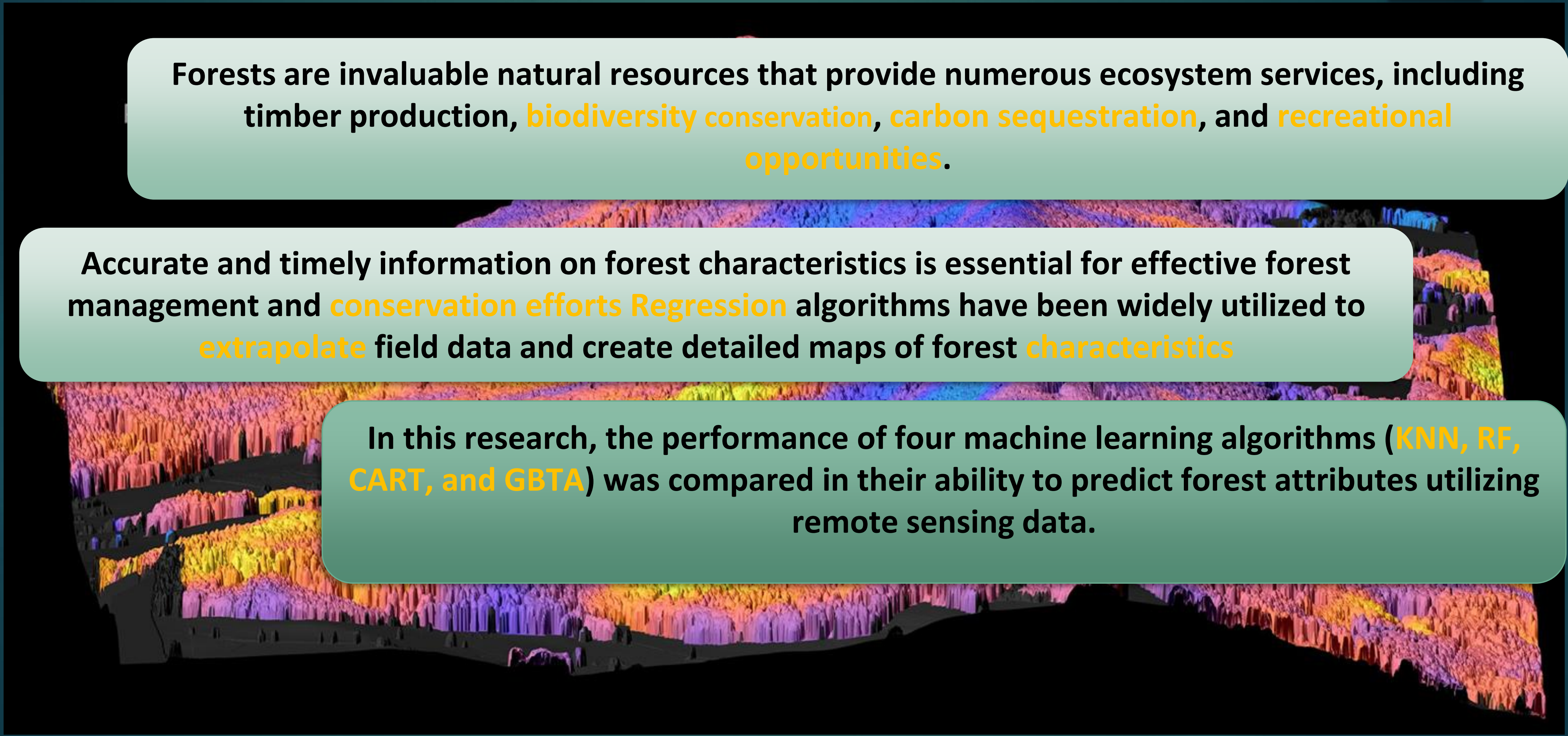
Comparing performance of machine learning algorithms in modelling forest characteristics in Carpathian Mountains using remote sensing data

MOHAMED KESKES

MIHAI DANIEL NITA

Transilvania University of Brasov

Introduction



Forests are invaluable natural resources that provide numerous ecosystem services, including timber production, biodiversity conservation, carbon sequestration, and recreational opportunities.

Accurate and timely information on forest characteristics is essential for effective forest management and conservation efforts Regression algorithms have been widely utilized to extrapolate field data and create detailed maps of forest characteristics

In this research, the performance of four machine learning algorithms (KNN, RF, CART, and GBTA) was compared in their ability to predict forest attributes utilizing remote sensing data.

Aims and Goals:

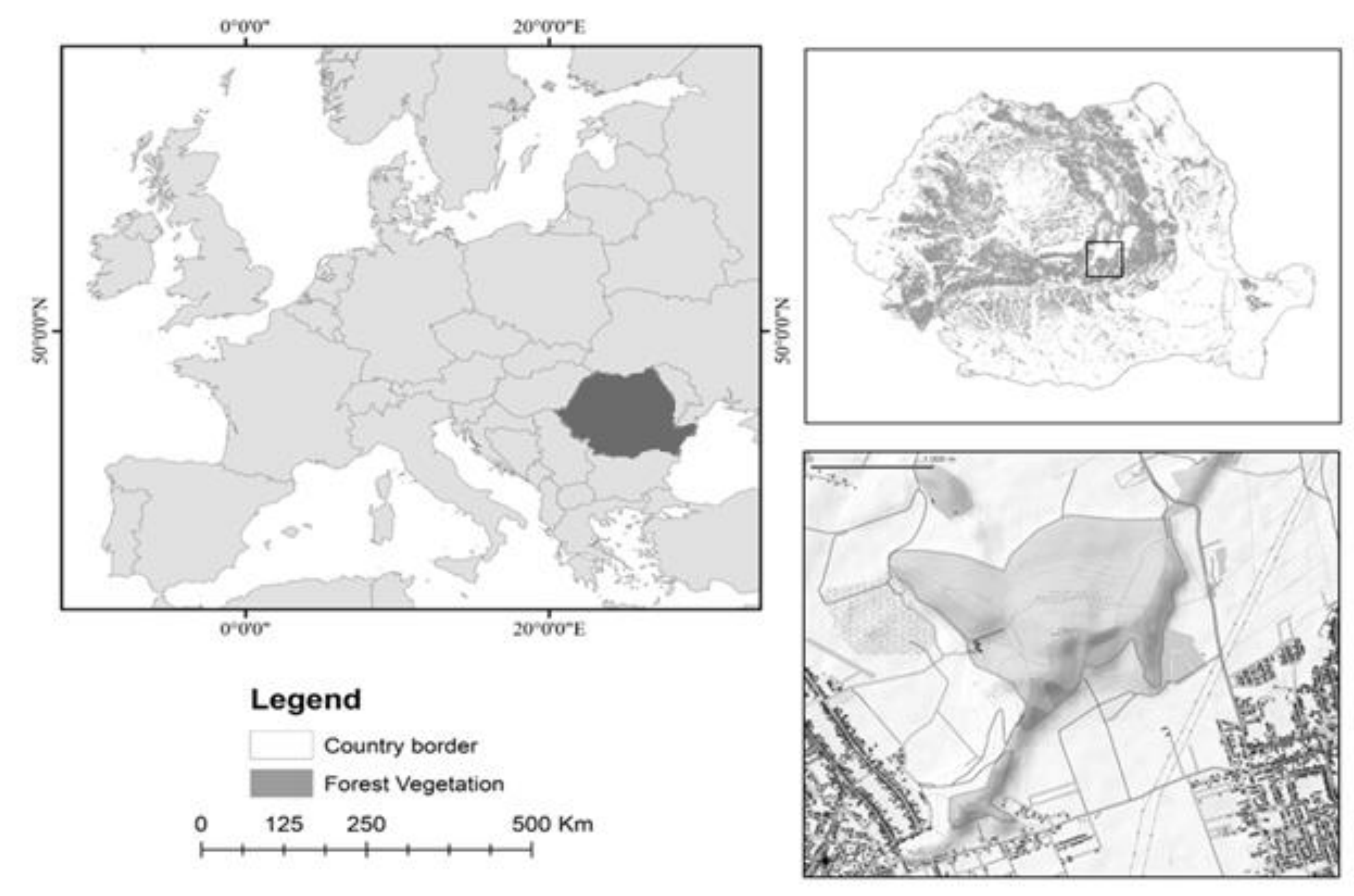
- Analyze and compare different scenarios for the CART and RF models by varying the number of trees in each model, ranging from a minimum suggested number to a maximum suggested number, in order to assess the influence of model complexity on the accuracy of forest characteristic predictions

- Evaluate the impact of different resolutions (10, 50, and 100) of the integrated remote sensing and field-based data on the accuracy of the model predictions to determine the optimal resolution for capturing fine-scale forest attributes.

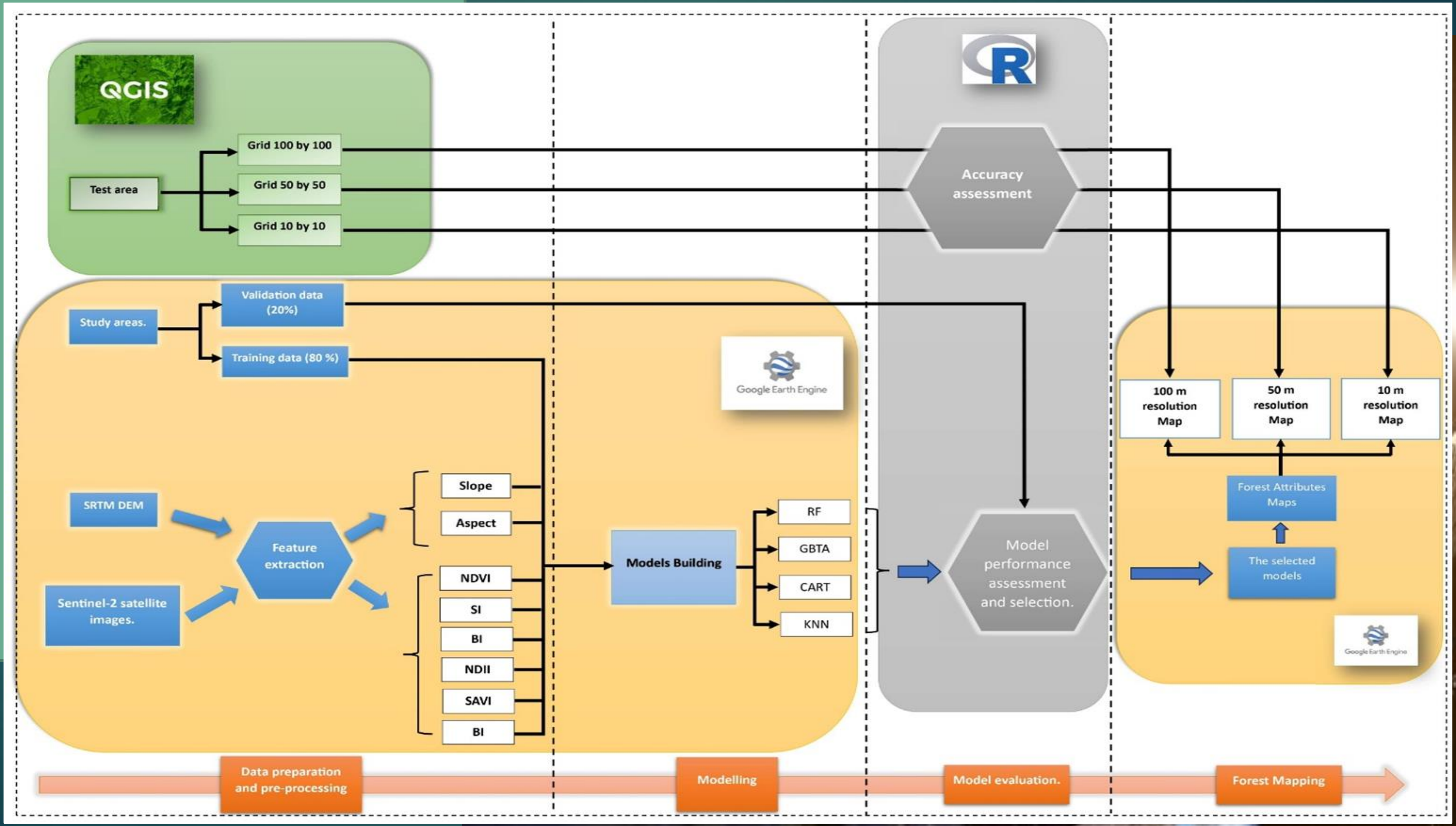
- Validate the predicted forest characteristics by comparing them with independent ground truth data collected from representative forest plots, ensuring the reliability and accuracy of the model predictions and providing robust support for their application in forest management and conservation decision-making processes.

Study area

We focused on all forests in Romania and we used the ”Fortress-hill Lempes – Harman marsh” site (ROSCI 0055 Dealul Cetăţii Lempeş - Mlaştina Hărman) as area for independent validation



Methodology



Methodology

This study utilized both traditional forest inventory tools and a mobile LIDAR device, including the vertex logger IV for tree height and forestry tape for measuring diameter at breast height (DBH).

There were 1326 samples of different areas (e.g. 500 sqm., 300 sqm.) chosen based on criteria such as tree density and spatial distribution, with a specific emphasis on forests designated for thinning and selective logging activities.

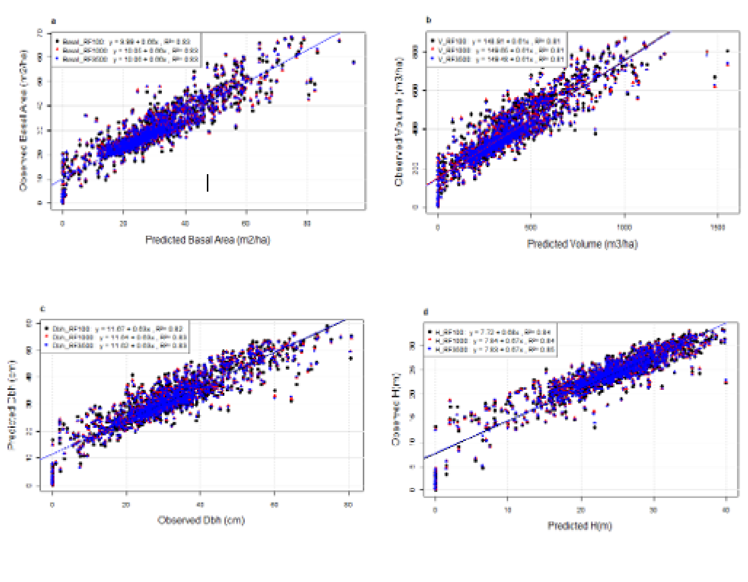


Figure 4: Scatterplots for the RF regression model of forest attributes: (a) basal area (b) tree height (H) (c) tree volume (Vol) (d) diameter at breast height (DBH)

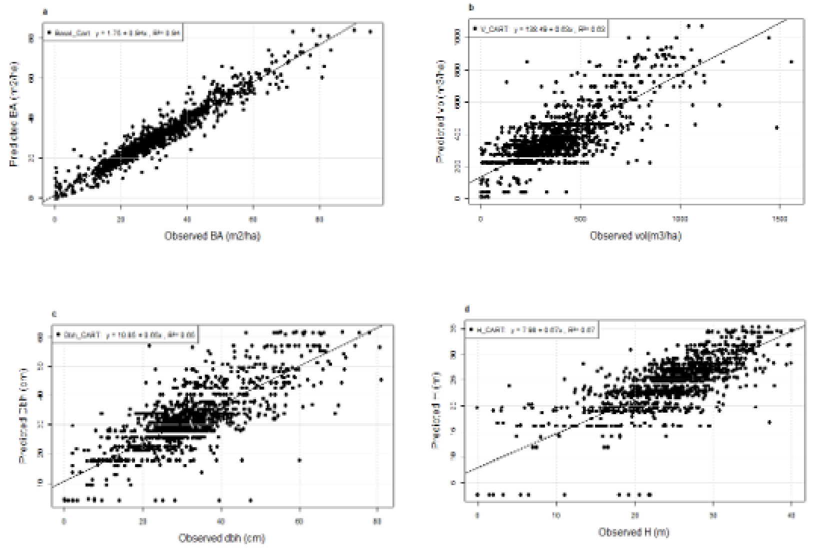
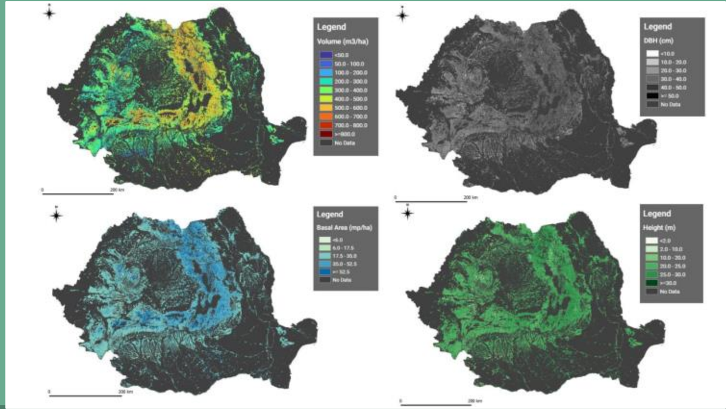


Figure 6: Scatterplots for the CART regression model of forest attributes(a) basal area (b) tree volume (Vol) (c) diameter at breast height (DBH) (d) tree height (H)

Results at national level



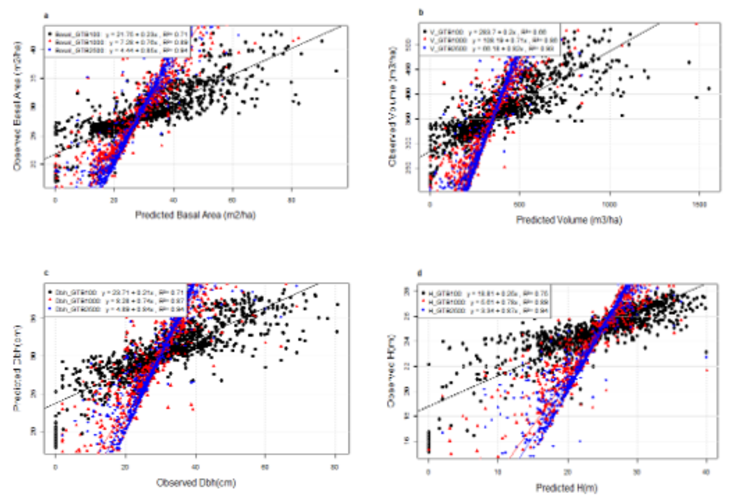


Figure 5: Scatterplots for the GBTA regression model of forest attributes with the variation of the number of trees: (a) basal area (b) tree volume (Vol) (c) diameter at breast height (DBH) (d) tree height (H)

This improvement was indicated by the increase in R-squared values. For instance, when the algorithm employed 100 trees for volume estimation, it achieved an R-squared value of approximately 0.66. However, when the number of trees was increased to 2500, the R-squared value substantially improved to 0.92. This increase implies that as the GBTA algorithm utilizes a higher number of trees, it becomes more proficient at capturing complex relationships and patterns inherent in the data.

ce of 0%

display distinct levels of performance when it comes to forecasting forest attributes

Table 2: Performance of forest attributes estimation models with independent datasets.

attributes	Pixel size	Algorithm	Statistical analysis				
			R2	P Value	RMSE(m)	rRMSE(%)	MAE (m)
Volume	10	RF	0.234	< 2.2e-16	120.943	0.297	102.554
		GBTA	0.222	< 2.2e-16	134.598	0.344	100.610
		CART	0.217	< 2.2e-16	120.067	0.294	109.795
	50	RF	0.215	2.67E-09	59.666	0.149	49.566
		GBTA	0.367	2.68E-16	87.015	0.229	39.646
		CART	0.061	0.002339	50.781	0.148	65.647
	100	RF	0.286	6.15E-06	66.809	408.478	56.783
		GBTA	0.388	4.89E-08	64.431	390.531	44.834
		CART	0.360	2.05E-07	59.374	403.605	53.118
BA	10	RF	0.286	< 2.2e-16	6.592	0.217	5.433
		GBTA	0.281	< 2.2e-16	9.285	0.323	5.723
		CART	0.351	< 2.2e-16	6.993	0.228	7.484
	50	RF	0.343	< 2.2e-16	6.203	0.202	5.424
		GBTA	0.358	< 2.2e-16	6.626	0.227	6.600
		CART	0.256	4.43E-13	7.392	0.238	5.614
	100	RF	0.194	0.000974	7.897	0.260	7.046
		GBTA	0.153	0.003786	8.171	0.283	8.587
		CART	0.102	0.01976	9.451	0.309	7.430
DBH	10	RF	0.285	< 2.2e-16	9.200	0.288	7.921
		GBTA	0.278	< 2.2e-16	9.218	0.293	7.885
		CART	0.244	< 2.2e-16	9.179	0.306	7.754
	50	RF	0.297	1.79E-11	6.935	0.212	6.241
		GBTA	0.312	4.28E-12	6.037	0.186	7.377
		CART	0.220	1.59E-08	8.974	0.310	4.752
	100	RF	0.578	2.03E-13	5.248	0.162	4.483
		GBTA	0.596	5.11E-14	4.219	0.138	4.326
		CART	0.577	2.24E-13	4.982	0.155	3.498
H	10	RF	0.207	< 2.2e-16	6.062	0.245	5.254
		GBTA	0.201	< 2.2e-16	6.091	0.242	5.418
		CART	0.176	< 2.2e-16	6.270	0.260	5.192
	50	RF	0.419	< 2.2e-16	3.359	0.135	2.910
		GBTA	0.466	< 2.2e-16	3.299	0.131	3.028
		CART	0.349	< 2.2e-16	3.484	0.135	2.837
	100	RF	0.504	4.28E-10	4.300	0.173	3.865
		GBTA	0.555	1.99E-11	4.155	0.165	4.103
		CART	0.417	4.36E-08	4.507	0.175	3.723

The KNN algorithm achieved a significantly lower R-squared value of 18.62% for volume prediction, indicating a limited ability to explain the variance in volume predictions. The KNN algorithm operates by first preparing a labeled dataset with input features and corresponding target values.

When predicting a new data point, it identifies the k nearest neighbors based on a similarity metric, such as Euclidean distance. The predicted value is then determined either by taking a majority vote or calculating the average of the target values of the k neighbors

Table04: results of the ANOVA test of the forest attributes predictions for all resolutions

attribute	Pixel size	Anova test				
		Df	Sum Sq	Mean Sq	F value	Pr(>F)
VOLUME	10	2	1853703	926852	91.17	<2e-16 ***
	50	2	345413	172706	40.5	<2e-16 ***
	100	2	12180	6090	1.361	0.259
BA	10	2	9772	4886	79.75	<2e-16 ***
	50	2	355	177.7	8.078	0.000349 ***
	100	2	93.9	46.96	2.888	0.0586
DBH	10	2	11020	5510	129.1	<2e-16 ***
	50	2	903	451.4	19.3	1.3e-08 ***
	100	2	137.7	68.86	6.092	0.00272 **
H	10	2	2797	1398.5	122.3	<2e-16 ***
	50	2	101	50.46	10.36	3.73e-05 ***
	100	2	24.4	12.182	2.657	0.073

Independent validation at local level

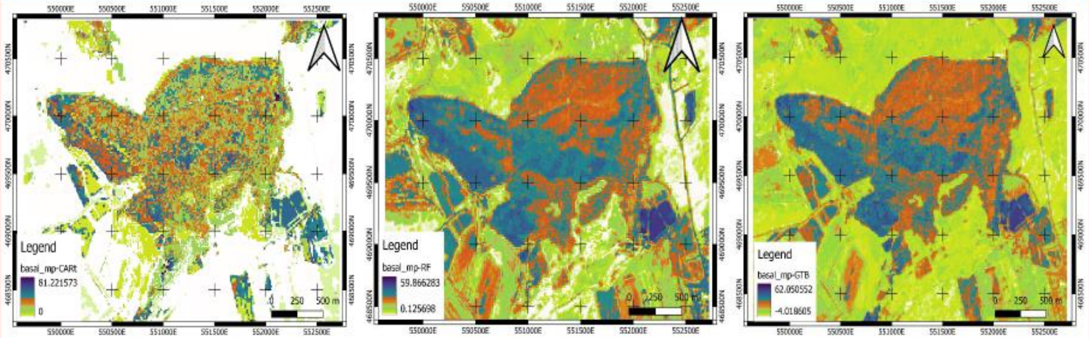


Fig.E1: Forest Characteristic Mapping for the Basal area using: (a) CART (b) RF (c) GTBA

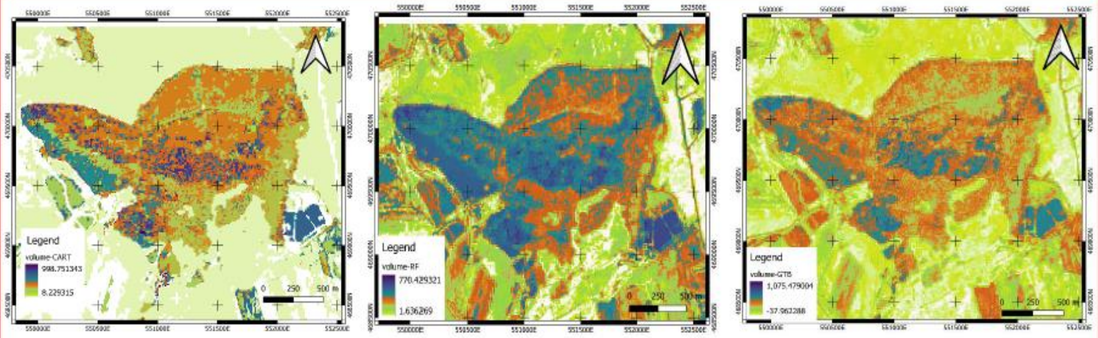


Fig.E4: Forest Characteristic Mapping for the volume using: (a) CART (b) RF (c) GTBA

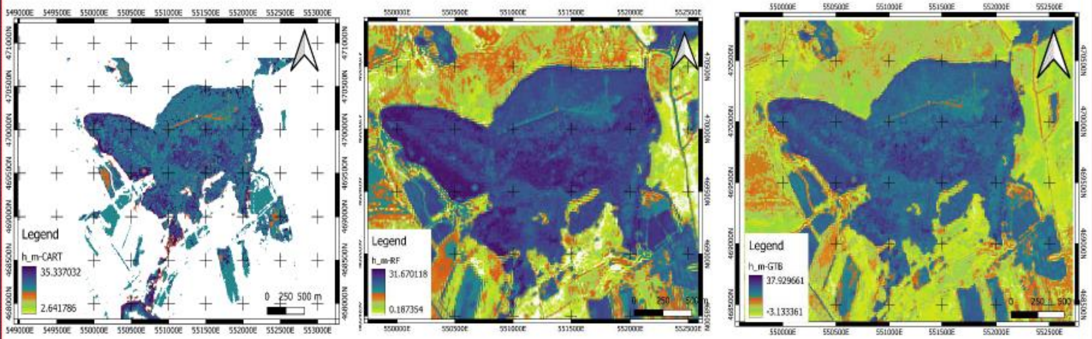


Fig.E2: Forest Characteristic Mapping for the canopy height using: (a) CART (b) RF (c) GTBA

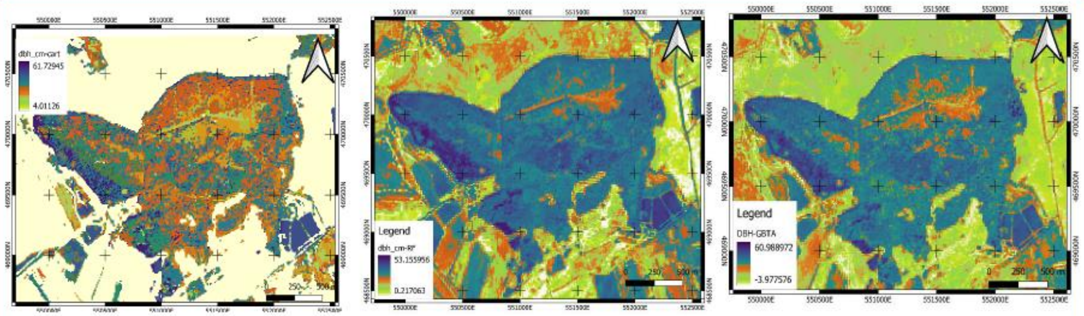


Fig.E3: Forest Characteristic Mapping for the DBH using: (a) CART (b) RF (c) GTBA

RF's very good performance in predicting basal area can be attributed to its unique capabilities in capturing and modeling the complex relationships and patterns specific to this attribute.

Among the algorithms considered, the RF, GBTA algorithms with a maximum number of suggested trees demonstrated outstanding performance in predicting various forest attributes during the evaluation with initial validations data. These models outperformed the other algorithms in terms of R-squared values, MAE, and rRMSE.

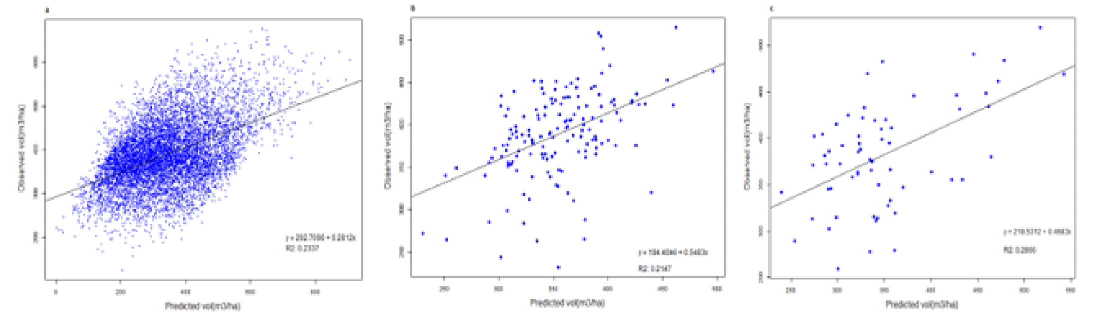
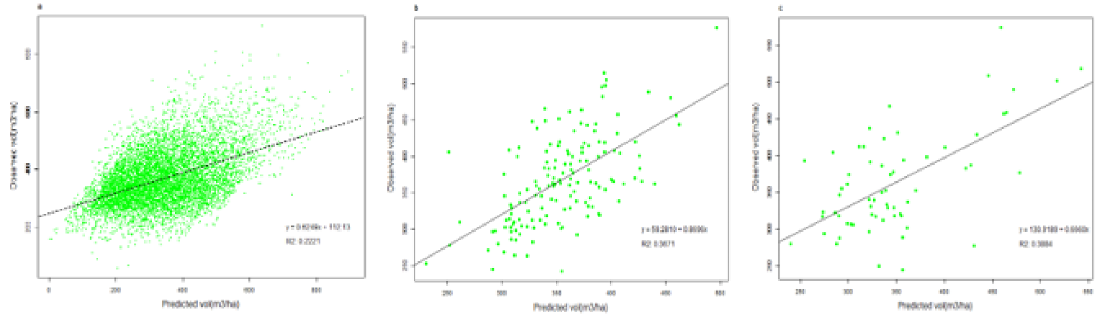
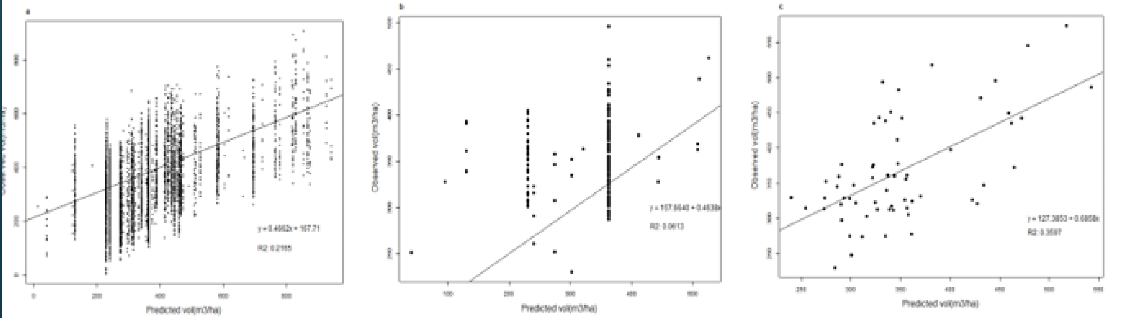


Fig.D1: Scatterplots for the regression model of the RF algorithm for the volume: (a) 10m



907

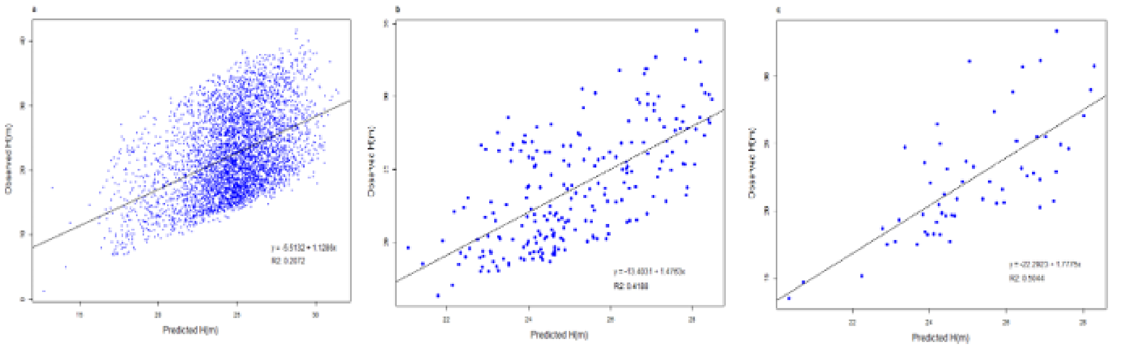
Fig.D2: Scatterplots for the regression model of the GBTA algorithm for the volume: (a) 10m resolution



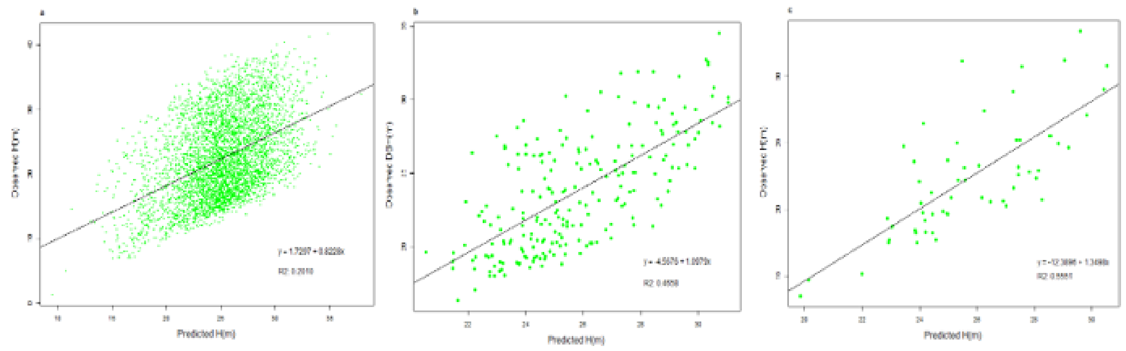
909 **Fig.D2:** Scatterplots for the regression model of the CART algorithm for the volume: (a) 10m resolution

In our study, we observed significant differences in the predictions of all the algorithms for all attributes at both 10m and 50m resolutions. This indicates that the algorithms responded differently to the increased level of detail provided by the smaller pixel sizes. The variations in the predictions suggest that each algorithm processed the large dataset captured at these resolutions in a unique way, leading to more complexity in the relationships between the variables

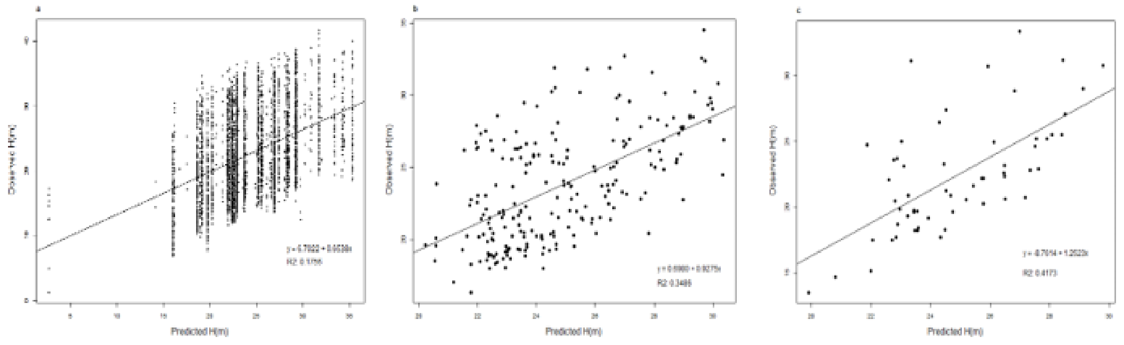
The algorithms may become less sensitive to localized changes and variations, as the larger resolution averages out the characteristics of larger areas. In our study, we found that at a 100m resolution, the predictions of all attributes exhibited notable differences among the algorithms only for the diameter at breast height (DBH) attribute



FigC1: Scatterplots for the regression model of the RF algorithm for the H: (a) 10m resolution (b) 50m



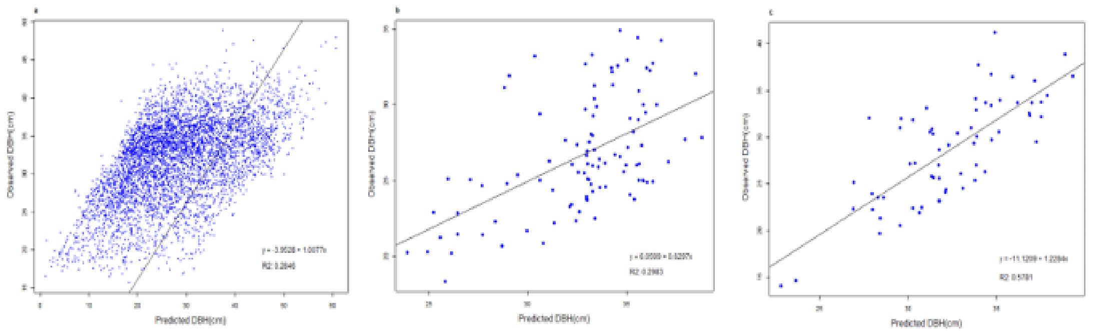
FigC2: Scatterplots for the regression model of the GBTA algorithm for the H: (a) 10m resolution (b) 50m



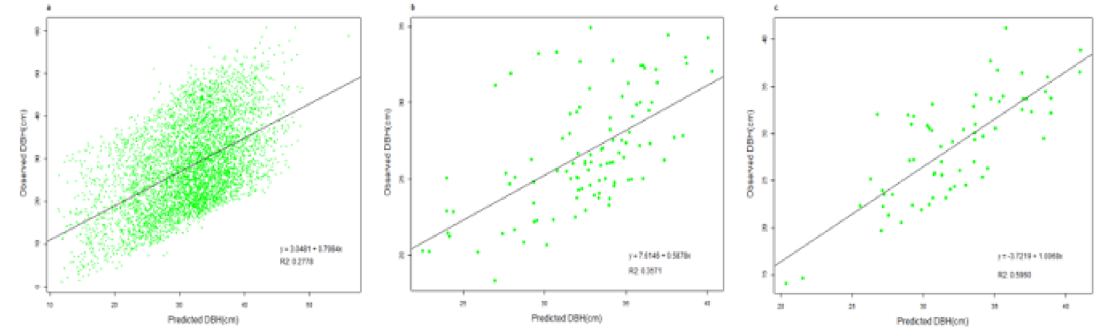
FigC3: Scatterplots for the regression model of the CART algorithm for the H: (a) 10m resolution (b) 50m

Similarly, the Gradient Boosting Algorithm (GBTA) exhibited similar observations when it came to predicting volume and basal area. Regardless of the pixel size, the GBTA model consistently provided accurate predictions for these attributes.

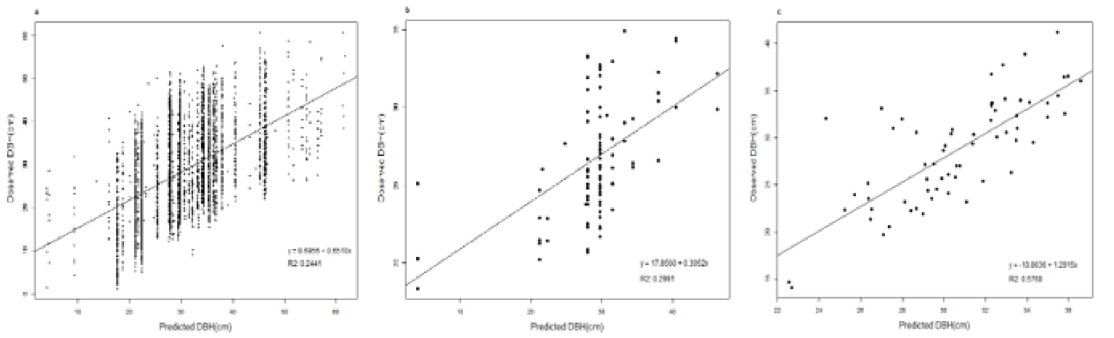
This further reinforces the notion that this algorithm is relatively insensitive to changes. It demonstrates a consistent performance in capturing and analyzing the relevant features and patterns associated with volume and basal area, regardless of the spatial resolution.



FigB1: Scatterplots for the regression model of the RF algorithm for the DBH: (a) 10m resolution (b) 50m



FigB2: Scatterplots for the regression model of the GBTA algorithm for the DBH: (a) 10m resolution (b) 50m



FigB3: Scatterplots for the regression model of the CART algorithm for the DBH: (a) 10m resolution (b) 50m

- **The impact on the predictions became notably evident when examining the DBH and H attributes, particularly in the algorithms CART and GBTA. The predictions for these attributes showed significant differences, indicating that pixel size played a crucial role in the accuracy of these predictions.**

- **The observed variations suggest that further analysis is necessary to delve deeper into the behavior of the CART and GBTA algorithms in predicting DBH and H.**

Summary

This study examines the potential of machine learning algorithms and remote sensing data in predicting forest attributes for improved forest management and conservation decision-making. The random forest regression (RF) and gradient boosting tree algorithms (GBTA) consistently demonstrate high prediction accuracy and strong predictive power across various forest attributes.

Validation data confirm the robustness of RF and GBTA algorithms. According to our study the utilization of the GBTA algorithm occurred as a reliable tool for forest attribute estimation and emphasizes the broader implications of accurate attribute estimation for effective forest management, conservation, and sustainable resource utilization.