environmental sciences proceedings

MDPI

Proceedings

Simulation of DEM based on ICESat-2 data using openly accessible topographic datasets†

Shruti Pancholi [1], *, Abhinav A [1], Sandeep Maithani[1] and Ashutosh Bhardwaj [1]

[1] Indian Institute of Remote Sensing; pancholis724@gmail.com
[1] Indian Institute of Remote Sensing; abhinavalangadan@gmail.com
[1] Indian Institute of Remote Sensing; maithani@iirs.gov.in
[1] Indian Institute of Remote Sensing; ashutosh@iirs.gov.in
* Correspondence: pancholis724@gmail.com

† Presented at the 5th Electronic Conference on Remote Sensing.

Abstract: Digital Elevation Model (DEM) is a 3-dimensional digital representation of the terrain or the Earth's surface. For determining topography, DEMs are the most used and ideal method with (i.e., Digital Surface Model) or without the objects (i.e., Digital Terrain Model). Various techniques are used to create DEMs, including traditional surveying methods, photogrammetry, InSAR, lidar, clinometry and radargrammetry. DEMs generated by LiDAR tend to be the most accurate except for the VHR datasets acquired from UAVs having spatial resolution of a few centimeters. In many parts of the region, LiDAR data is not available, which limits researchers' access to high-resolution and accurate DEMs. Having a beam footprint of 13 meters and a pulse interval of 0.7 meters, ICESat-2 promises high orbital precision and high accuracy. ICESat-2 can produce high-accuracy DEMs in complex topographies with an accuracy of a few centimeters. Earth's surface elevations are provided by discrete photon data from ICESat-2. It is difficult to justify the continuity of the topographical data using traditional interpolation techniques since they over-smooth the estimated space. Geospatial data can be analyzed with machine learning algorithms to extract patterns and spatial extents. To estimate a DEM from LiDAR point data from ICESat-2 using CartoDEM, machine learning regression algorithms are used in this study V3 R1. This study was conducted over a hilly terrain of Dehradun region in the foothills of Himalayas in India. The applicability and robustness of these algorithms has been tested for a plain region of Ghaziabad, India in an earlier study. The interpolation of DEM from ICESat-2 data was analyzed using regression-based machine learning techniques. Interpolated DEMs were evaluated against the TANDEM-X DEM of the same region with RMSEs of 7.13m, 7.01m, 7.15m, and, 3.76m respectively, using Gradient Boosting Regressors, Random Forest Regressors, Decision Tree Regressors, and Multi-Layer Perceptron (MLP) Regressors. Based on the four algorithms tested, the MLP Regressor shows the best performance in the previous study. The accuracy of the simulated ICESat-2 DEM using MLP Regressor was assessed in this study using the DGPS points over the Dehradun region. The RMSE was of the order of 6.58m for the DGPS reference data.

Keywords: Digital Elevation Model, Machine Learning, Multi-Layer Perceptron, Spaceborne LiDAR, Differential GPS

Citation: To be added by editorial staff during production.

Academic Editor: Firstname Lastname

Published: date

1. Introduction

Digital Elevation Model is a visualization of the bare Earth's surface elevations [1]. DEMs are generated from numerous sources including contour lines, topographic maps, stereo photogrammetry, SAR Interferometry, DGPS points, etc. Amongst all the techniques to create DEMs high-resolution Laser Altimetry (LiDAR) is proven to generate higher accuracy DEMs [2]. Various terrain related studies including hydrological modelling, flood inundation mapping, monitoring volcanic activities, etc. use DEM as an integral input data. Therefore, the accuracy of the input DEMs for various applications is an important parameter to yield good quality results [3]. Systematic errors in DEM products are still possible due to equipment precision limitations, which is time consuming, costly and difficult to rectify [4]. To enhance the quality and accuracy of the available open-source DEMs various studies have been conducted [5].

An Earth Observation System satellite, the Ice, Cloud, and Land Elevation Satellite (ICESat-2), was launched by NASA. Highly accurate data from ICESat-2 provides extensive and sufficient reference data for quality analyzing different DEMs [6].

Interpolation is the process of estimating the value of attributes at unsampled sites from measurements made at point locations within the same area or region but it often leads to over smoothening [7]. Simulation technique can be defined as a statistical way to generate data, where unavailable based on the statistical models like linear regression which correlates the input and output of the sample/training data and calculates the statistical relationship between the two and implements the same for other input points to generate their corresponding output. This study hence utilizes the CartoDEM and ICESat-2 LiDAR data to simulate a higher accuracy DEM using machine learning algorithms.

Various studies have shown that for the Indian region, good quality and best accuracy terrain data is available by Cartosat-1 DEM [8]. This study focusses on simulating a higher accuracy spaceborne LiDAR DEM by correlating it with the CartoDEM measurements. The simulated DEM is then validated using DGPS data. The accuracy of the simulated output DEM is higher than the CartoDEM closer and to the LiDAR measurements.

2. Methods

2.1. Study Area

This study was conducted over the hilly terrain of Dehradun region in the foothills of Himalayas. The study area lies between latitudes 30°01' N and 31°2'N and longitudes 77°34' E and 78°18'E (Figure 1).

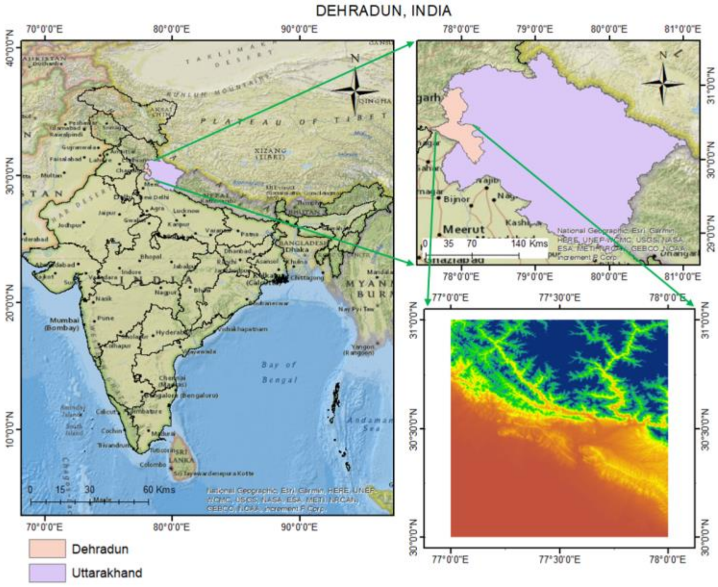

Figure 1. Study area showing Dehradun region

2.2. Datasets

2.2.1. CartoDEM V3 R1

Using the CartoDEM V3 R1 product, the corresponding LiDAR DEM was generated to enhance the vertical accuracy of the CartoDEM. The Cartosat-1 satellite is the first Indian remote sensing satellite that can provide stereo visualization in orbit. A number of products derived from Cartosat-1 can be used for various Geographical Information Systems (GIS) applications, including Digital Elevation Models (Figure 1), Ortho Image products, and Value-added products for GIS.

2.2.2. ICESat-2

In ICESat-2, the ATLAS instrument provides all of the topographic data through its advanced topographic laser altimetry system. A total of three relatively strong beams and three relatively weak beams are present [9]. In the context of accurate analysis of different DEMs, it provides enough and high-quality reference data [8].

2.2.3. Ground Control Points (GCPs)

The Trimble R7 GNSS receivers and Leica 500 series receivers were used for collection of the field data. A total of 16 GCPs were collected over the Dehradun region and utilized for the validation of the simulated DEM.

2.3. Methodology

The overall methodology followed for this study is depicted in Figure 2. Pancholi et al. has successfully generated DEM using the machine learning models of Decision Tree (DT), Random Forest, Gradient Boosting Machine (GBM) and Multi-Layer Perceptron (MLP) [10], out of which the MLP model gave minimal error output.

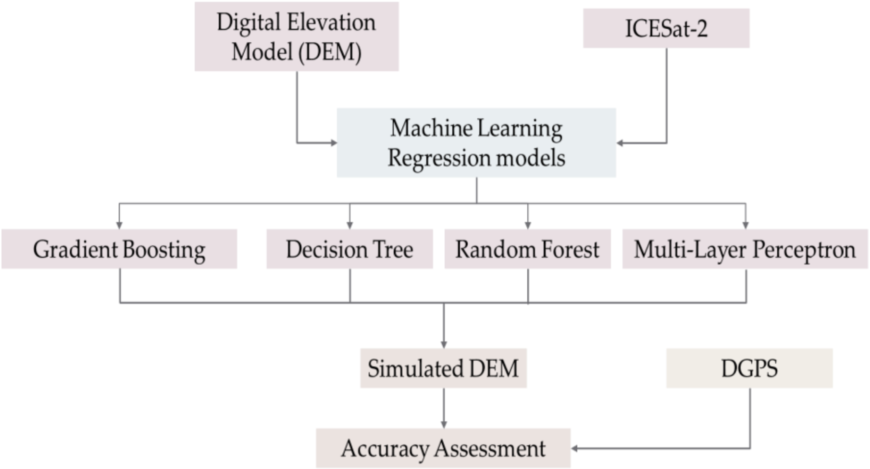

Figure 2. Methodology followed for the simulation of DEM

2.3.1. Machine Learning Models

This section describes the machine learning models used for this study.

- Decision Tree

The decision tree model, which finds a foundation in machine learning theory, is a potent tool for dealing with regression and classification challenges. In contrast to other classification approaches that use a group of features (or bands) together to complete classification/regression in a single decision step, it relies on a multilevel or hierarchical decision strategy or a tree-like structure. It consists of leaves, internal nodes, and the root node. Each decision tree node uses a top-down technique to perform binary classification, separating one or more classes from the others by progressing down the tree until the leaf node is reached. In essence, a complicated problem statement is divided into lesser problems by a decision tree, and the simpler decisions that follow lead to the complex conclusion. The decision tree model is chosen for the study because it effectively resolves problems involving both linear and non-linear interactions [11].

- Random Forest

The ensemble machine learning model random forest has two or more decision trees, which together form a "decision forest". Finding the majority by voting on the individual decision tree outcomes yields the random forest model's outcome. The design of each decision tree that makes up Random Forest affects how well it performs. There are two steps in this process that include random selection. The first step uses a bootstrap technique to randomly select about two-thirds of the training dataset before beginning to build each decision tree. Out-of-bag (OOB) data, which make up the final third of the dataset, are utilized for inner cross-validation to assess the precision of the mode [12].

- Gradient Boosting Machine

Gradient boosting is a unique ensemble machine learning approach that utilizes the predictive capability of boosting on a decision tree. It has several decision trees constructed sequentially, each of which is a "weak" learner. These following learners draw lessons from the preceding model's errors to create the final model, which is a "strong" one. The first model is given some initial constant values that are calculated by averaging all of the target values. Residuals are the calculated differences between the anticipated value and the actual target values. The goal values for the following decision tree are these residuals r1, and the residuals r2 are computed from the anticipated value and r1. This keeps on until every decision tree is trained [13].

- Artificial Neural Network

An artificial neural network (ANN) is a nonlinear nonparametric framework that uses neural network propagation across layers based on gradient learning techniques to simulate human brain receptors and information processing. The input layer, hidden layer, and output layer are the three layers that make it up. Through synapses, the input layer receives the input and transmits it to the hidden levels; likewise, the hidden layers transmit the data to the output layer. The weights that the synapses hold regulate how information moves from one layer to the next. Equation (1) mathematically describes a neuron in the hidden layer or output layer.

$$u = \sum_{i=1}^{n} w_i x_{\ i} \tag{1}$$

and, y=$\varphi(u + b)$

where w denotes synaptic weights, x denotes input to neurons, y denotes output from neurons, u denotes a linear combiner of input signals, b denotes bias, and $\varphi()$ is the activation function used to restrict the input range.

2.3.2. Hyper parameters Used

Some variables must be put up in advance and cannot be changed while training. These variables or parameters are called hyper-parameters. They are the factors that control how a learning algorithm learns and determine the final outcome of the models [14]. The goal of hyperparameter optimisation is to find the optimal settings for hyperparameters to provide good results from data as rapidly as feasible. Hyper parameter optimisation is performed as the parameters tuned during this process are not optimized by the models during training and has to be provided to the models before the training actually begins.

Table 1. Hyperparameters tuned for the regression models

Model	Parameter	Selected Value
DT	Maximum depth	None
	Criterion	Mean Absolute Error
RF	Number of trees	100
	Maximum depth	None
GBM	Loss	Squared Error

	Number of Estimators	100
ANN	Learning Rate	0.001
	Maximum number of iteration	200
	Activation	ReLU

2.3.3. Accuracy Assessment

Utilizing the coefficient of determination (R2), root mean square error (RMSE), and mean absolute error (MAE) in comparison to the simulated DEM and DGPS data, the machine learning model was statistically evaluated for the Dehradun region. Regression model performance is often evaluated using the R2 and RMSE of the predicted and actual values. For estimating accuracy metrics over an area's elevation values, higher R2 and lower MAE and RMSE are correlated with higher precision and accuracy, respectively. To get a clearer result, LE90 value was also calculated for the simulated DEM using MLP regressor model. The formula extensively used for LE90 is given in Equation (2) [15], [16].

$$LE90 = 1.6449 * RMSE \tag{2}$$

3. Results and Discussion

In this study an implementation of machine learning models was done **t**o simulate a higher accuracy DEM providing elevation values closer to the ICEsat-2. The accuracy of the simulation was evaluated primarily using the 20% testing data that is unseen by the model and is shown in Table 1.

Table 2. Accuracy metrics of machine learning models

	DT	**RF**	**GBM**	**MLP**
R2	0.99	0.99	0.99	0.99
RMSE	3.61	3.41	3.42	3.12
MAE	2.28	2.21	2.25	2.20

ANN model displayed the best results in terms of RMSE and MAE followed by RF, GBM, and DT Figure 3 shows the simulated DEM using the four models. The validation of the simulated output using MLP was done using DGPS GCPs (shown in Figure 4). The accuracy of the simulated DEM using DGPS yielded an RMSE of 6.58m which is very promising in a hilly terrain in the foothills of Himalayas for simulated DEM product. The LE90 score for the simulated DEM was 10.82m, signifying the confidence that minimum 90% of the vertical error fall within the limit of 10.82m. The variation in RMSE while comparing the RMSE derived from ICESat-2 and DGPS can be attributed to the lower uncertainty of DGPS on collecting the elevation data when compared to ICESat-2 points, which need filtering of footprints (elevation values) based on the deviations. Furthermore, ICESat-2 footprints are not evenly distributed throughout the study area, and are more concentrated in plane area and less concentrated in hilly area.

The highest values of elevation are 1950.87m, 1975m, 1964.77m, and 1967.78m for DT, GBM, RF, and MLP machine learning models, whereas the highest elevation value in the ICESat-2 footprint is 1976.87m. This is a realistic representation of elevation with respect to the training data used in the model. However, since the ICESat-2 data points are not densely distributed in the study area and very sparsely distributed in the high elevation zones, there are possibilities of under-representation of elevation in zones higher than 1976.87m.

An even distribution of ICESat-2 data in plane and hilly terrain while training the model can potentially improve the accuracy of the models. Including ICESat-2 points in the hilly terrain of nearby area for training the models or using the same for developing a deeper neural network based on transfer learning approach can evenly balance training data in all elevation ranges and improve the results of the model.

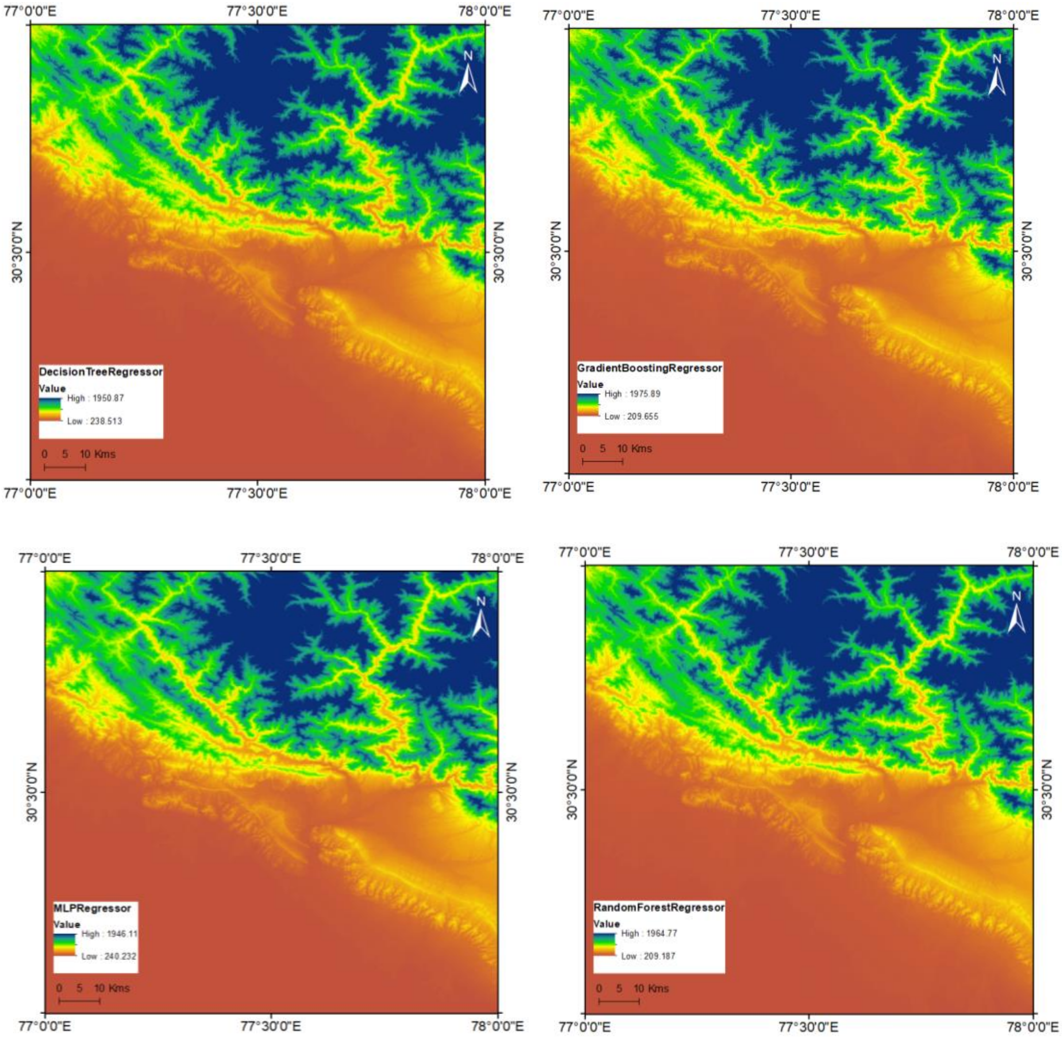

Figure 3. Simulated DEM from CartoDEM and ICESat-2 (a) Decision Tress Regressor (b) Gradient Boosting Regressor (c) Decision Tree Regressor (d) Multi Layer Perceptron

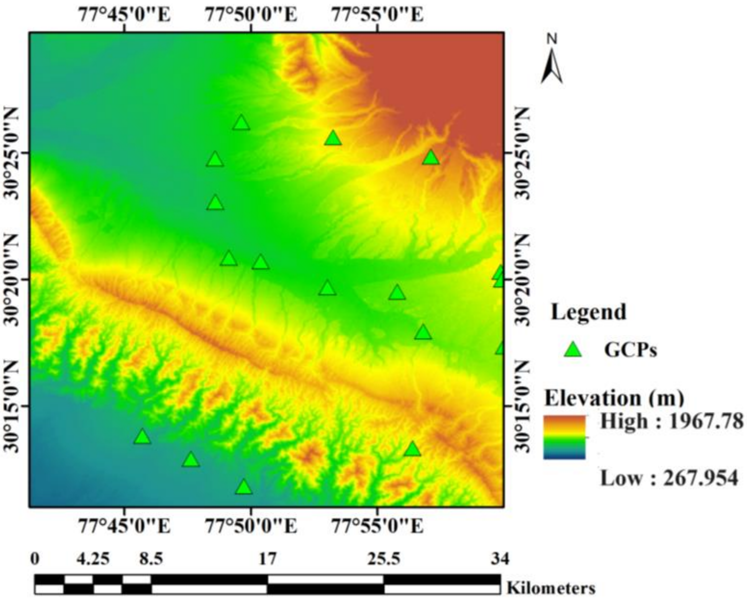

Figure 4. GCPs collected using DGPS survey are overlaid on Simulated MLP DEM product

4. Conclusion

The current study attempted to simulate an ICESat-2 DEM over a 388 km^2 area in the hilly terrain of Dehradun located in the foothills of Himalaya. Four machine learning algorithms- DT, RF, GBM, and MLP was used for the simulation using CartoDEM and ICESat-2 data and produced promising results with MLP performing the best. The accuracy assessment was initially done using ICESat-2 points and validated using DGPS GCPs. The study concluded that although DGPS points provide a planned way of validation of DEMs, however collection of large number of DGPS points is time consuming and costly issue. Whereas the ICESat-2 dataset not only provide large number of high accuracy elevation points for the simulation. Further investigations must be done for improving the accuracy of the DEM in centimeter scale. Increasing the number of training points in all elevation zones and land use land cover areas, transfer learning ML approach are suggested for future improvements.

Author Contributions: Conceptualization, A.A. and A.B..; methodology, A.A. and A.B..; software, S.P. and A.A.; validation, S.P., A.B and S.M.; resources, S.M. and A.B.; writing—original draft preparation, S.P. and A.A.; writing—review and editing, A.B.; visualization, S.P.; supervision, S.M. and A.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: Open-source data was used for this article as explained in the dataset section.

Acknowledgments: The authors would like to thank the Director, Indian Institute of Remote Sensing for providing the laboratory facility for carrying out this study. The authors would like to thank the Indian Space Research Organization (ISRO), and National Aeronautics and Space Administration (NASA), to make the openly accessible data available to the researchers through their data sharing platforms.

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

References

[1] "What is a digital elevation model (DEM)? | U.S. Geological Survey."

https://www.usgs.gov/faqs/what-digital-elevation-model-dem (accessed Sep. 11, 2023).

[2] S. Rayburg, M. Thoms, and M. Neave, "A comparison of digital elevation models generated from different data sources," *Geomorphology*, vol. 106, no. 3–4, pp. 261–270, May 2009, doi: 10.1016/J.GEOMORPH.2008.11.007.

[3] X. Liu, "Airborne LiDAR for DEM generation: some critical issues," *http://dx.doi.org/10.1177/0309133308089496*, vol. 32, no. 1, pp. 31–49, Feb. 2008, doi: 10.1177/0309133308089496.

[4] M. Wang, H. Yu, J. Chen, Y. Zhu, Y. Zhang, and W. Yu, "Comparison of DEM Super-Resolution Methods Based on Interpolation and Neural Networks," *Sensors 2022, Vol. 22, Page 745*, vol. 22, no. 3, p. 745, Jan. 2022, doi: 10.3390/S22030745.

[5] A. Bhardwaj, K. Jain, and R. S. Chatterjee, "Generation of high-quality digital elevation models by assimilation of remote sensing-based DEMs," *https://doi.org/10.1117/1.JRS.13.4.044502*, vol. 13, no. 4, p. 044502, Oct. 2019, doi: 10.1117/1.JRS.13.4.044502.

[6] A. L. Neuenschwander and L. A. Magruder, "Canopy and Terrain Height Retrievals with ICESat-2: A First Look," *Remote Sens. 2019, Vol. 11, Page 1721*, vol. 11, no. 14, p. 1721, Jul. 2019, doi: 10.3390/RS11141721.

[7] A. Setiyoko, A. M. Arymurthy, T. Basaruddin, and R. Arief, "Semivariogram fitting based on SVM and GPR for DEM interpolation," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 311, no. 1, p. 012076, Aug. 2019, doi: 10.1088/1755-1315/311/1/012076.

[8] S. Mariani, G. Pavan, S. Goud, and A. Bhardwaj, "Estimation of Building Heights and DEM Accuracy Assessment Using ICESat-2 Data Products," *Eng. Proc. 2021, Vol. 10, Page 37*, vol. 10, no. 1, p. 37, Nov. 2021, doi: 10.3390/ECSA-8-11442.

[9] T. A. Neumann *et al.*, "The Ice, Cloud, and Land Elevation Satellite – 2 mission: A global geolocated photon product derived from the Advanced Topographic Laser Altimeter System," *Remote Sens. Environ.*, vol. 233, Nov. 2019, doi: 10.1016/J.RSE.2019.111325.

[10] S. Pancholi, A. Abhinav, and A. Bhardwaj, "Simulation of ICESat-2 DEM using Machine Learning Algorithms," Jan. 2023, doi: 10.20944/PREPRINTS202301.0381.V1.

[11] S. B. Kotsiantis, "Decision trees: a recent overview," *Artif. Intell. Rev. 2011 394*, vol. 39, no. 4, pp. 261–283, Jun. 2011, doi: 10.1007/S10462-011-9272-4.

[12] A. Cutler, D. R. Cutler, and J. R. Stevens, "Random Forests," *Ensemble Mach. Learn.*, pp. 157–175, 2012, doi: 10.1007/978-1-4419-9326-7_5.

[13] V. K. Ayyadevara, "Gradient Boosting Machine," *Pro Mach. Learn. Algorithms*, pp. 117–134, 2018, doi: 10.1007/978-1-4842-3564-5_6.

[14] Y. A. Ali, E. M. Awwad, M. Al-Razgan, and A. Maarouf, "Hyperparameter Search for Machine Learning Algorithms for Optimizing the Computational Complexity," *Process. 2023, Vol. 11, Page 349*, vol. 11, no. 2, p. 349, Jan. 2023, doi: 10.3390/PR11020349.

[15] Y. Gorokhovich and A. Voustianiouk, "Accuracy assessment of the processed SRTM-based elevation data by CGIAR using field data from USA and Thailand and its relation to the terrain characteristics," *Remote Sens. Environ.*, vol. 104, no. 4, pp. 409–415, Oct. 2006, doi: 10.1016/J.RSE.2006.05.012.

[16] C. C. Carabajal and D. J. Harding, "ICESat validation of SRTM C-band digital elevation models," *Geophys. Res. Lett.*, vol. 32, no. 22, pp. 1–5, Nov. 2005, doi: 10.1029/2005GL023957.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.