A comprehensive Approach to Above-Ground Biomass Estimation using multi-source data integration and Advanced Learning Techniques

Abstract

The measurement of biomass across multiple crops is critical for optimizing resource utilization, projecting yields, and soil fertility. As part of this research on the development of remote sensing systems for biomass estimation, we have conducted multiple studies related to modelling and methods for the analysis of multisource data. The research objective presented in this paper is the estimation of above-ground biomass using multisource remote sensing data. This study examines the effect of integrating data from several sources, comprising spectral reflectance from Sentinel 2 and spectral vegetation indices (NDVI, GDVI, SIPI, NDRE, SAVI, and RECL) sensitive to canopy structure, chlorophyll content, and soil variables (nitrogen content, organic carbon, and water content). We have used various machine-learning and deep-learning approaches to estimate the biomass. The results of our investigation showed that all the models significantly improved their training accuracy when used with soil data. The results show slight improvement in Random Forest, where the R2 score improved from 0.75 to 0.78 and RMSE decreased from 42.26 Mg ha-1 to 38.75 Mg ha-1 followed by PNN, BNN, and XGBoost. However, in the test dataset, the BNN showed a significant improvement in the R2 value from 0.54 to 0.66. Interestingly, the BNN model achieved an average RMSE test accuracy of 59–60 after multiple runs. While, the XGBoost model showed only a slight improvement in its performance. This depicts the importance of using a complex multimodal dataset and its relevance in the context of precision farming. Also, the ability of the BNN model to uncover complicated correlations and generalizability in varied agroecosystems.

1. Introduction

Demands for food production is soon expected to skyrocket – it will reportedly surpass the pre-COVID-19 levels of nine billion people by the year 2050 (The World Population Prospects, United Nations, 2017), resulting in an increased need for farming productivity. As a result, there is an acute need to meet the expanding demands. There is global stress on agriculture to boost production with merely more resources. Crop yield is the most important element that influences production. First and foremost, agricultural production follows significant seasonal patterns based on crop biological life cycles. Production depends secondarily on the physical landscape (e.g., soil type), climatic variables and agricultural practices. All these variables are highly variable in space and time. Monitoring crop growth is essential for agriculture or cropland reclamation, helping to understand the physiological status of crops, forecast agricultural production[1], and evaluate reclaimed cropland. Various characteristics like above-ground biomass (AGB)[3], monitoring to reflects the crop's response to the growing environment and cultivation practices[2], and further courses of action can be designed and implemented accordingly.

Precision agriculture helps to enhance crop yield by efficiently allocating resources by leveraging technology and data-driven approaches. AGB estimation can benefit precision agriculture with its focused data-driven decision-making approaches. Precision agriculture uses modern technologies such as satellite imagery, field mapping, and sensor data to improve crop quality and profitability. In traditional agriculture, where, the farming manages fields as a single block, the precision farming works by dividing them into separate areas. This zoning allows for capturing details and making tailored management decisions for individual field parts. It keeps the farmers informed about the stressed areas of the fields and helps them to optimize their resources to achieve their objective. It promotes sustainable use of traditional resources (such as water, fertilizers, and pesticides) while minimizing the waste. Also, it helps in mitigating the impacts of using excessive chemical pesticides and fertilizers and promotes environmental and soil health. Thus, precision agriculture is omnidirectionally useful in terms of cost savings, resource efficiency, and providing global food security.

1.1 Above Ground Biomass

The use of precision agriculture has increased over the past few decades. It is essential for agriculture as it has pronounced effect on sustainability, productivity, and environmental effect. Biomass estimation helps to keep a vigilant eye on the agricultural plots and is directly contributing to achieve the objective of precision farming. Biomass measurement, also known as above-ground biomass (AGB), is an important measure for yield and grain quality prediction. Above-ground biomass is a key indicator of crop growth, health, and potential yield. It mainly indicates the foliage that a particular crop has during different stages of its growth. Healthy crops generally exhibit higher AGB due to better biochemical features associated with them such as adequate photosynthesis, efficient nutrient uptake, and favourable conditions. Monitoring AGB allows farmers to assess crop health, identify crop stressors of which some are pests, diseases, or nutrient deficiencies and take corrective measures. Moreover, the timely decisions made helps to comprehend the status of crops before significant damages to the yield. Hence, it helps to avoid crop failure This, in turn, helps the farmers predict and optimize the required genetic selection, fertilization, irrigation, and pest control[11]. Thus, resulting in building a robust ecosystem that can shield the farmers from unexpected results and lead to overall profitability. Hence, AGB estimate is one of the most crucial agricultural factors as its estimation can improve crop monitoring and yield prediction (Bendig et al., 2015; Brocks and Bareth, 2018; Yue et al., 2017). Moreover, farmers can use this information to improve crop management practices, resulting in higher-quality produce that directs better market prices and overall profitability[11].

Within the past 10 years, there has been extensive research to explore the use of AGB in the estimation of crop yield. The existing techniques used to quantify AGB are majorly characterized into two groups i.e. ground-based and remote sensing (RS). Destruction techniques are implemented on the ground as oppose to the airborne ones. Destructive methods traditionally involve cutting the plants in the field and then drying and weighing them in the laboratory[4]. To complete the forest AGB calculation of sample plots and prevent tree damage, the allometric equation measures forest parameters (e.g., diameter at breast height, height, and stock volume). Although these measurements generate the most accurate estimates of plant biomass, they are time-consuming and labour-intensive[5]. Also, these are only applicable to small-scale areas because it is challenging to gather enough sample plots for large-scale areas. Ground-based methods for non-destructive measurement of AGB have been studied for decades[6,7]. These approaches estimate AGB using equations relating biomass to measurable biophysical factors such as plant height and plant density[8]. Handheld devices are the most straightforward instruments for measuring these biophysical factors[9]. The most widely used and well-documented ground-based method for the non-destructive measurement of AGB in grasslands is the rising plate meter (RPM)[10].

The development of remote sensing satellites has made it feasible to quickly and effectively estimate AGB at a range of geographic resolutions. Since remote sensing is ideal for capturing data across large areas with a high return frequency, it can play a vital role in giving a timely and accurate picture of the agricultural sector. The primary uses of remote sensing are outlined in this study, with an emphasis on regional and worldwide uses. It offers justifications for increasing financial support for agricultural monitoring systems. It is based on the firm belief that intensive oversight of agricultural production systems is required. Since agriculture must significantly boost output to meet rising demand. The environmental impact of agriculture must be kept to a minimum to achieve this productivity rise. To achieve this goal, agriculture must adapt to climate change and compete with land users who are not engaged in food production (such as the production of biofuels, urban growth, etc.) in order to achieve this goal. To give decision-makers input on their investments and policies, it is imperative to closely monitor the necessary adjustments and transitions.

A breakthrough in optical remote sensing technology, Sentinel-2A's launch on June 23, 2015, combined high spatial resolution, multispectral, and high temporal resolution to obtain fine observation information. Since then, the technology has been used more and more in a variety of fields. Sentinel-2 data has yet to be completely utilized in practice for AGB estimation; this is especially true when combined with other remote sensing data over wide regions

2. Literature Survey

Numerous modelling techniques are frequently used to evaluate AGB. Choosing the best ones have a direct impact on the precision and dependability of AGB estimation. Both parametric and nonparametric models are the common approaches for estimating AGB using field AGB samples and variables obtained from remotely sensed data. Stepwise multiple regression (SMR) is a frequently used method of biomass estimate in parametric linear models, which performs better at fewer sample points. In actuality, several factors may prevent a straightforward linear relationship between variables and forest biomass. Machine learning algorithms (MLAs) may combine several inputs, learn extremely complicated nonlinear relationships, and hence produce better simulation results than linear models[6]. Support Vector machine (SVM)[8] , artificial neural network (ANN[12]), and Random Forest (RF)[14] models are machine learning models that are commonly used to for assessing biomass. Most research uses a variety of models for comparative analysis because of the variations in model. The scope of combining the results of various models along with different remote sensing data has also been explored to improve the model performance in various scenarios. Combining RF and KNN models improved the accuracy of calculating AGB in the Qilian mountains using multisource remotely sensed data. Furthermore, In Shangri-La the accuracy of models_ was improved by combining the time series multisource data of sentinel-1, sentinel-2[13]. Similar work, in the Dabie mountain region showed the combination of Geofen-1 with Sentinel-1 showed the random forest achieved highest prediction accuracy[14]. To find the optimal model technique for estimating forest AGB using multisource high-resolution remote sensing data, more comparative study is required.

The literature survey reveals substantial advancements in understanding the Above Ground biomass of an area using multispectral data. These datasets are combined with several other data including SAR, LiDAR which further helps in increasing the accuracy of estimation. While other researchers have used different machine learning algorithms for the Above Ground Biomass Estimation ranging from linear regression to Random Forest, XGBoost and Artificial Neural Networks. However, some crucial research gaps persist which needs to be addressed with utmost clarity of concept.

Though there are several other biophysical factors apart from canopy of the trees contributing to above ground biomass estimation. Very Few research address the need of including the soil properties data for estimation of above ground biomass. Various applications like integration of satellite image data are suggested in the literature, there's a need for detailed investigations showcasing the practical integration of soil properties data in real-world devices.

While several methods like Random Forest, XGBoost, Multiple Linear Regression have been widely used and explored there has been a lack of Neural Network based deep learning approaches for the purpose of Above Ground Biomass Estimation. Also, the inventory data collected only few plot for calculating the observed above Ground Biomass Dataset which is not sufficient for the training of Machine Learning Models. The researchers seem to be less interested in exploring the non-destructive approach for above ground biomass estimation. This study takes into account the predictive capabilities of machine learning and deep Learning models to predict the biomass of an area, fields and avoid any kind of destructive approach and hence preventing unreasonable fall of trees

Moreover, various studies show that the study done is particularly based on the specific area. There is no universal approach or algorithm developed serving the large space beyond the area of research or the region of interest. Our study is focused on developing a universal approach for above ground biomass estimation which is revolution for mankind. Further benefits involve the prevention of large budget spent for above ground biomass estimation in the vicinity of a universal model for above ground biomass estimation.

The works done so far included only few point upon which the researchers have trained their model and tested, which lacks the usability for comprehensive applications. The work done in this study has taken into account thousands of data points for training and testing of models leading to a more generalizable comprehensive approach.

The researches used several features such as bands separately as well as combined obtained from several satellites and commercially available datasets but they lack the use of feature selection for building of the models. The models have comprehensively taken into account the use of sequential feature selection technique for selecting the subset of most important features. The hyperparameter optimisation, which is important for the performance of models but very few researchers have described it in their studies. This study has comprehensively taken into account all the necessary hyperparameter optimisation for obtaining a robust predictive model

Addressing these research gaps could pave the way for a more nuanced understanding of Above Ground Biomass Estimation using multispectral Dataset and their effective utilization in various practical applications.

Most of the studies done before this involved the use of single-source or multi-source satellite data. The combination of multiple remote sensing sources[13] data can improve the accuracy of forest biomass calculation, a promising approach that many academics have been trying to utilize consistently, and lessen the drawbacks of a single data source. However, there is very limited research demonstrating the use of soil properties data along with multispectral data. Apart from that, there is very limited information on the generalizability of models used so far. This study gives an eminent picture of the importance of physical, chemical, and derived properties of soil for the estimation of above-ground biomass. Moreover, the research gives insight into the generalizability of the models used, for the estimation of above-ground biomass of any region near or far away place. This is beneficial as the above-ground biomass data is not readily available for most parts of the world. The research further explores the comparison of machine learning models for above-ground biomass estimation using Sentinel -L2A combined with soil properties variables, which is new and less explored.

Estimating AGB in an agricultural area has multi-dimensional benefits. It helps in yield prediction, harvest scheduling, tailored fertilizer application thus, minimizing wastage and environmental pollution. In this study, we evaluate the potential of topographic data, model methods, and Sentinel-2 along with vegetation indices and soil properties data to extract aboveground biomass from the agricultural area in the Yellapur region of Uttara Kannada district in Karnataka. Soil has a great influence on the quality of produce in an area and is indicative of the above-ground biomass that can be obtained from that area. Various nutrients present in the soil, nourish the vegetation differently to plant growth and can be used for the determination of above-ground biomass estimation purposes. The study explores the importance of integrating the soil properties data along with the remotely sensed satellite data for the estimation of above-ground biomass which most of the previous studies ignored.

We compared the performance of models namely, Random Forest, XGBoost, PNN, and BNN on the dataset of sentinel-2 and the vegetation indices combined and excluded with soil properties data respectively. Upon analysing the models, we got that BNN model showed an improvement in their results as their R2 got increased from 0.54 to 0.66 respectively, while the RMSE got decreased from 69.88 to 59.75 respectively while XGBoost showed slight improvement in its performance. The Random Forest and Progressive Neural Networks models showed a decline in their results as their R2 decreased from 0.63 to 0.60 and R2 = 0.65 to 0.59 respectively, while their RMSE increased from 62.39 to 64.95 and 60.59 to 66.20 respectively. The test dataset is kept completely out of the training dataset to get the models that can generalize on any such test dataset. The models result overall decrease in average R2 value from training to testing dataset without soil properties data except the PNN model and an overall decrease in R2 value from training to testing dataset with soil properties data except the BNN model. This is due to the spatial variability of the vegetation in training and testing datasets and depicts their significance for general use throughout different patches for above-ground biomass estimation.

These study's specific goals are to: (1) assess the potential of variables extracted from high spatial resolution remotely sensed data from Sentinel-2; (2) determine the best combination of variables; (3)

determine the most accurate modelling techniques for estimating AGB using integrated Sentinel images, soil Properties, vegetation indices, topographic metrics, and forest inventory data; and (4) create an accurate and finer-resolution (40m) AGB map.

3. Data Collection

3.1. Experimental design and data collection

Yellapur being one of the towns located in Karnataka's Uttara Kannada district is endowed with bountiful sources of nature which makes it even more special, as the entire scene of local economy revolves around this only. The farmers grow variety of crops such as rice, areca nut, coconut. The huge demand is reason why this business is the reason for flourishment of spice cultivation. Arecanut plantations dominate the landscape and, reflects its importance as a major cash crop. Coconut farming is also widespread, alongside rice cultivation which thrives in both irrigated and rain-fed fields. Spice cultivation, especially of black pepper and cardamom is adding to the agricultural tapestry of the region. Additionally, horticulture is gaining major traction as farmers are increasingly turning for cultivation of fruits like banana, pineapple, and papaya. Joint efforts by the government and NGO's aim to promote sustainable agricultural practices despite challenges like water scarcity and market fluctuations. Thus, bolstering the resilience of Yellapur's farming community and ensuring its contribution to the regional prosperity.

Latitude (N)	74.55 N
Longitude (E)	14.84 E
Elevation (m)	459 m (above mean sea level)
Slope	8.7 ± 7.8 degrees
Plant Functional Types (PFTs)	EBT, DBT, GSW
Mean Annual Precipitation	2383 ± 421 mm yr-1
Mean Annual Temperature	24.4 ± 0.4 C

Table 1. Geographical features of the Area of Interest

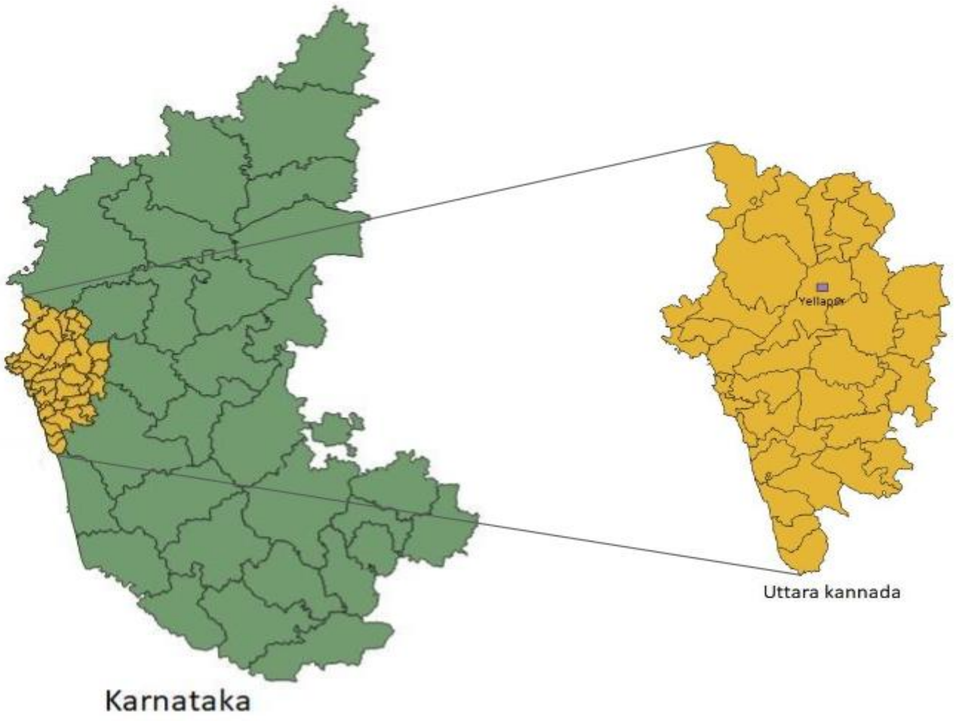


Figure 1. Uttara Kannada District of Karnataka state

3.2 Data

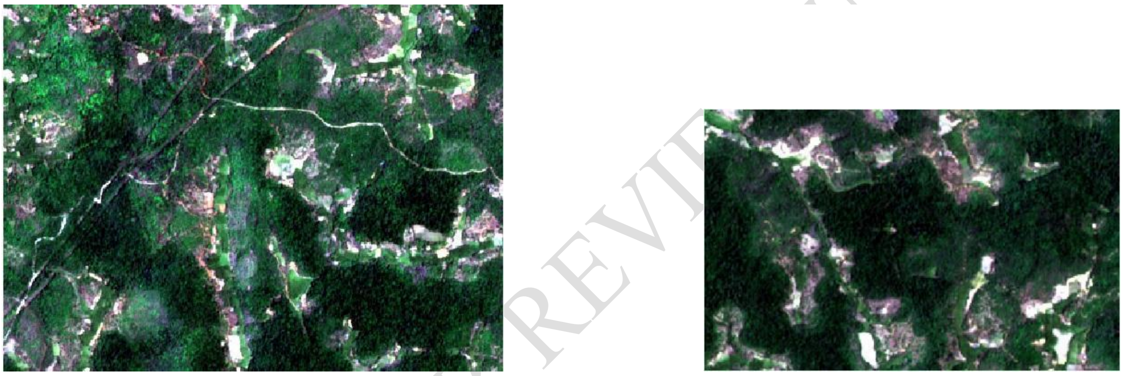
3.2.1. Field Data

The inventory data is collected from https://bhuvan-app3.nrsc.gov.in/ ISRO Geosphere-Biosphere Programme (IGBP) above-ground biomass (AGB) Data. The dataset includes aerial LiDAR data and field inventory plots that were used to create maps of the aboveground biomass. The dataset contains aboveground biomass prediction maps derived from field inventory plots and airborne LiDAR data. The visualization of Indian AGB maps at 100m and 40m spatial resolution is facilitated. The dataset collected covers an area of 12.974 Km^2 and a test area of 4.967 Km^2 in the region of Yellapur.

Data collection for ISRO's IGBP involved a variety of methods including Remote Sensing, In situ measurements, Field Experiments, and Modelling. The data collected had a spatial resolution of 40*40 meters and ranged from 0.1 Mg ha-1 to 549.7 Mg ha-1 with a Mean AGB of 164.38 Mg ha-1 & Standard Deviation of 102.99 Mg ha-1 with a Mean Standard Deviation of 31.136 Mg ha-1

3.2.2 Image Acquisition And Preprocessing

Sentinel-2 is an Earth observation mission from the Copernicus Programme designed for the acquisition of optical image at high spatial resolution (10 m to 60 m) and suitable for land and coastal waters coverage. The mission is currently a constellation with two satellites, Sentinel-2A and Sentinel-2B. The multispectral data. It has 13 spectral Bands with four bands having 10 meters of spatial resolution, seven Bands having a 20m spatial Resolution, and two bands having 60m spatial resolution. The Sentinel-2 images were downloaded from https://sentinels.copernicus.eu/web/sentinel/sentinel-data-access (acquired on 15 November 2023). The 10m and 20m bands are resampled to 40 meters to correspond to that of inventory Data.



(a) (b)

Figure 2. True colour composite of a) Training Dataset and b) Test Dataset

Vegetation Indices like NDVI, GNDVI, SIPI, NDRE, SAVI, and RECL are used.

Vegetation Indices	Formula
NDVI	(NIR – RED) / (NIR + RED)
GNDVI	(NIR – GREEN) / (NIR + GREEN)
SIPI	(NIR – BLUE) / (NIR – RED)
NDRE	(NIR – RED EDGE) / (NIR + RED EDGE)
SAVI	((NIR – RED) / (NIR + RED + 0.428)) * (1 + 0.428)
RECL	(NIR / RED) – 1

Table 2 – Standard Indices

Along with Sentinel L-2A data and vegetation indices, we collected derived soil properties, i.e., organic carbon density; physical soil properties, i.e., bulk density and volume of water content at -10 kPa; chemical soil properties, i.e., soil nitrogen content, and derived soil properties, i.e., organic carbon density, from https://soilgrids.org/ . The data collected was initially at 253 m by 253 m spatial resolution. The layers are reprojected and then resampled to 40 m by 40 m to correspond to the inventory data.

3.3 Data Pre-Processing

The sentinel bands were filtered based on their relevance to the work required to be carried out. We collected the sentinel L2A dataset from https://sentinels.copernicus.eu/web/sentinel/sentinel-data-access(acquired on November 15, 2023). For modelling, band_02, band_03, band_04, band_05, band_08, band_8A, band_11, and band_12 are taken. These bands were resampled to 40m by 40m to correspond to the MAGB data. The digital number value (DN value) is collected pixelwise for each band. Furthermore, we calculated standard vegetation Indices including NDVI, NDRE, SIPI, GNDVI, SAVI, and RECL. Moreover, we collected derived soil properties, i.e., organic carbon density; physical soil properties, i.e., bulk density and volume of water content at -10 kPa; chemical soil properties, i.e., soil nitrogen content, and derived soil properties, i.e., organic carbon density, from https://soilgrids.org/. These data were reprojected to _ Coordinate system and resampled to 40 m by 40 m to correspond to that of MAGB data. For the purpose of modelling, we performed the data standardization technique,

4 Methodology and Modelling

4.1 Methodology

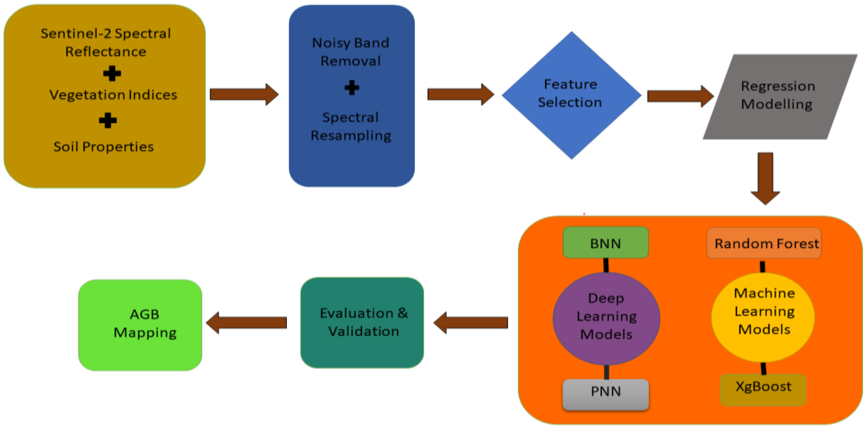


Figure 3. Workflow for the AGB estimation modelling.

Our study used an integrated case-study approach to thoroughly investigate the benefits of combining multiple data sources inside a remote sensing regression framework. We specifically combined Sentinel 2 satellite data along with soil data. This is aimed at integrating several types of data that have the potential to improve the precision and strength of regression models in the context of Biomass estimation.

Case 1:

The purpose of this Case 1 regression analysis is centred on the relationship between surface spectral characteristics and above-ground biomass. Sentinel-2's spectral bands and derived VIs (NDVI, GDVI, SIPI, NDRE, SAVI, and RECL) were used in capturing vegetation-related biophysical properties. This method serves as a benchmark to determine how much information can be gained or how much predictive ability can be /is added by using soil data. This case acts as a reference point that isolates the predictive power before integrating models with soil variables. This supports further analysis of whether the inclusion of soil data translates into significantly enhanced model performance.

Case 2:

In this case, we further consider soil properties (nitrogen content, organic carbon, and water content) which are obtained from ISRIC (World Soil Information Database). Incorporating soil attributes seeks to provide a wider understanding of the variables affecting the target variable. The objective is to investigate whether such additional complexity results in significantly enhanced performances of the regression model compared to those in Case 1.

Both case studies involve thorough preprocessing that includes resampling and spatial/temporal alignment of all data sources. This helps achieve data quality and consistency for machine learning regression models. In addition to these established regression metrics, model performance is compared between both the cases. Therefore, a comparative analysis aims at quantifying the value of multisource data fusion for improved prediction and gives actionable insights into factors driving the dependent variables.

4.2 Modelling Techniques

Most previous studies used the default value of parameters for modelling, and few studies have described the details of the feature selection process. We did sequential feature selection, which is a technique used in machine learning and statistics to improve the performance of a model by selecting a subset of features from the original set. The idea behind sequential feature selection is to iteratively add or remove features based on the criteria to find the most relevant and informative subset for a particular task. This process helps in reducing the dimensionality of the data, which can lead to improved model interpretability, reduced computational complexity, and better generalization performance. We performed sequential feature selection integrated with the respective model. The feature selection is performed in both directions, i.e., forward and backward sequential feature selection.

1. Random Forest

Random Forest is an ensemble learning algorithm that combines multiple decision trees to enhance predictive accuracy. Leo Breiman, along with Adele Cutler[21], further developed the random forest algorithm. During training, it constructs a multitude of trees, each utilizing a random subset of the dataset and features. Bagging helps to evade overfitting and provides diversity since it uses bootstrapped samples for training trees. The model here uses a voting system during predictions, with the most vote of classification and averaging for regression. Its parallelizability boosts efficiency as well as the ensemble nature makes it resistant to noises and outliers. Random Forest is flexible as it is applicable to the classification and regression tasks.

2. XGBoost

First proposed by Chen et al.[11], XGBoost is the algorithm that utilizes the second-order Taylor expansion and a regularization term to increase the convergence speed of machine learning and prevent overfitting. With its scalability and ability to handle sparse data efficiently, XGBoost has gained popularity in large-scale model training. On the one hand, the regularization term is added to the objective function to restrain the complexity of the tree so that a simpler model is obtained as well as overfitting is avoided. To summarize, XGBoost is a highly scalable tree structure enhancement model that can deal with sparse data with great speed, and scale down the training memory used, especially in very large-scale data training.

3. Progressive Neural Network (PNN)

PNN is based on incremental learning. It starts with the simple neural network, which can be further expanded and refined by the availability of new data. The architecture adjusts and adapts to accommodate and deal with the growing complexity of new tasks. This approach helps in efficient utilization of computational resources. Prog NN can be particularly useful in the cases that require constant learning. It is also a powerful and responsive tool for dealing with the varied and dynamic datasets. Also, it allows the model to achieve a continuous progress on both the existing and the new tasks. This is because the neural network model is able to grow and adjust as the time goes by[16]. On the other hand, it reduces the complexity that is associated with catastrophic forgetting in conventional neural networks[17].

4. Bayesian Neural Network (BNN)

The Bayesian Neural Network (BNN) is the neural networks' extension that uses the Bayesian in its structure. In contrast to standard neural networks where the weights are fixed, BNNs hold a probability distribution over its weights. Being able to account for the fraction of error in the predictions the BNNs are ideal for situations where limited data or noisy inputs is available[18]. And while training BNNs update both the weights and the corresponding uncertainties, the model representation becomes complex. This Bayesian method not only helps with more accurate predictions but also provides the foundation for principled model calibration.

5. Results and Validation

5.1 Validation

5.1.1 RMSE

One of the statistical metrics used when forecasting the usability of models in regression analysis is the Mean Root Squared Error. It measures the average error of series forecast at the point of intersection of the two curves. The square root of the mean of the squared differences between the observed value and the predicted value is the RMSE. It is computed as the root mean square error. RMSE mathematical formula is provided below:

$$\text{RMSE} = \sqrt{\frac{\sum(yi - \hat{y}i)^\wedge 2}{n}}$$

5.1.2 MAE

This statistic is a typical method of assessing how well regression models predict the behavior of a "response" variable. The MAE, which is the difference between the actual and the expected ground truth values, can be used to indicate the variation. In the nutshell, it means a finding out to what extent the model tends to over-predict or under-predict real data. MAE is mathematically represented as,

This is a typical statistical measure for assessing how well regression models perform. The mean absolute error (MAE) denotes the variation between the expected and actual ground truth values. In simpler terms, it determines the average deviation between the model's predictions and the actual data. MAE is mathematically represented as,

$$\text{MAE} = \frac{\sum|yi - \hat{y}i|}{n}$$

5.1.3 R-Squared

The most important measure of model fitting available in regression analysis is the coefficient of determination, or (R^2). It indicates the percentage of the dependent variable's variance explained by the independent variables in the model. R2 essentially gauges how well the model's predictions account for the variability in the observed data. R2 is given as,

$$R2 = 1 - \frac{\sum(yi - \hat{y}\ i)\hat{}2}{\sum(yi - \bar{Y}i)\hat{}2}$$

5.2 Model Performance

The table [3,4] shows the R2, RMSE, and MAE values of different models. Table[3] shows the performance of models without soil properties, the best-performing model was PNN and the worst was BNN, while Table[4] shows the performance of models with soil properties data, the best-performing model was BNN and the worst was PNN. In both of these datasets, Random Forest and XGBoost showed average performance where Random Forest performed better than XGBoost without soil properties data and XGBoost performed slightly better than Random Forest with soil properties data.

Comparative analysis of the performance of optimized AGB models. After tuning the parameters, we obtained the best models of the RF, XGBoost, PNN, and BNN using the same dataset. The performance of models could be explained by the scatterplots. Thus, showing the relationship between the observed AGB values and predicted AGB values. Figure (4) shows the model performance excluding soil properties data. They showed that the Random Forest performed best followed by PNN, XGBoost and BNN models with the same test dataset (i.e. excluding soil properties data). Figure (5) shows the model performance (including soil properties data). They showed that the BNN performed best followed by XGBoost, Random Forest and PNN models with the same test dataset (i.e. including soil properties data).

5.3. Results

5.3.1 Case 1 Analysis

Algorithms	Dataset	RMSE	R2	MAE
Random Forest	Training	42.26	0.75	32.4
	Testing	62.40	0.63	48.52
XgBoost	Training	49.33	0.65	37.89
	Testing	62.90	0.62	47.46
PNN	Training	51.71	0.62	39.37
	Testing	60.59	0.65	45.68
BNN	Training	52.64	0.60	40.14
	Testing	69.88	0.54	55.80

Table 3. The performance of models Random Forest, XGBoost, PNN, and BNN (excluding Soil Properties data)

5.3.2 Case 2 Analysis

Algorithms	Dataset	RMSE	R2	MAE
Random Forest	Training	38.75	0.78	28.88
	Testing	64.95	0.60	50.07
XgBoost	Training	49.18	0.65	37.67
	Testing	62.82	0.63	48.53
PNN	Training	50.19	0.64	38.11
	Testing	66.20	0.58	49.87
BNN	Training	52.00	0.61	39.30
	Testing	59.75	0.66	45.20

Table 4. The performance of models Random Forest, XGBoost, PNN, and BNN (including Soil Properties data)

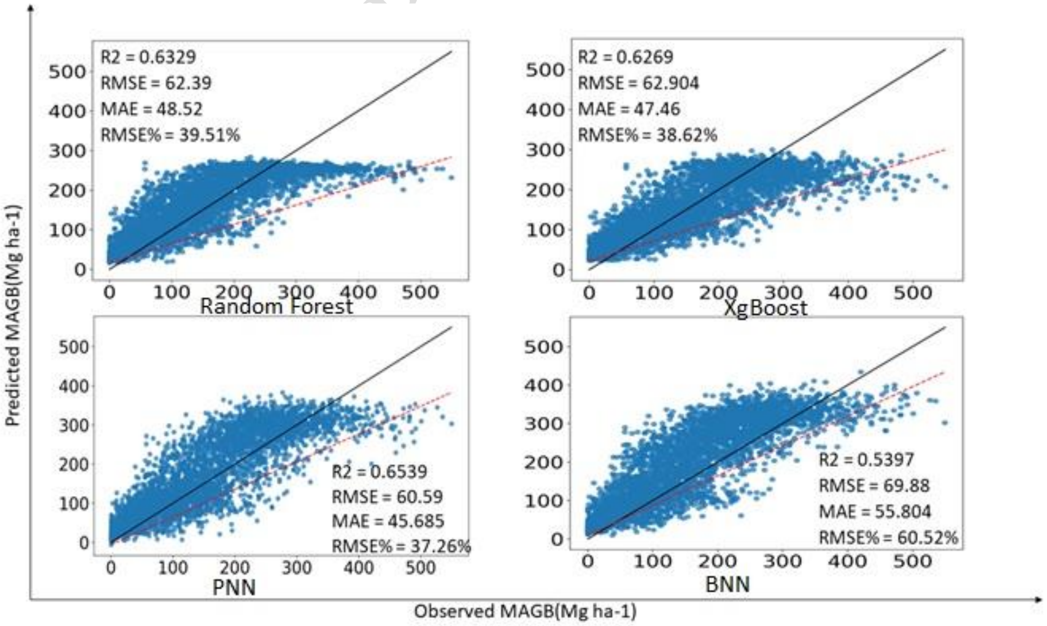


Figure 4. Y predicted values vs. Y plot of AGB (excluding soil properties data) for XGBoost, Random Forest, PNN, and BNN

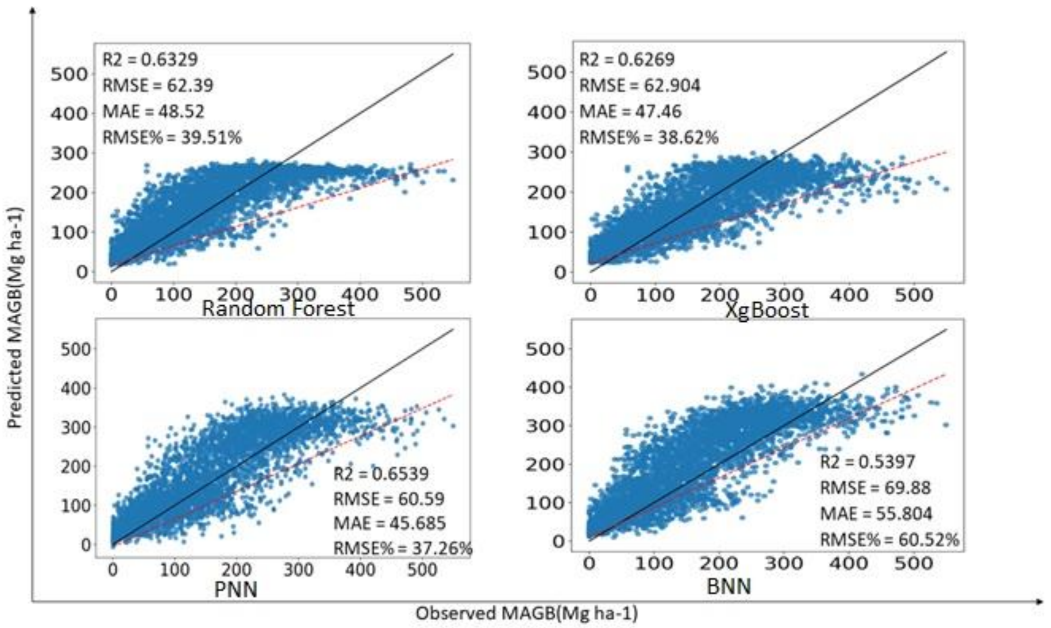


Figure 5. Y predicted values vs Y plot of AGB (including soil properties data) for XGBoost, Random Forest, PNN, BNN

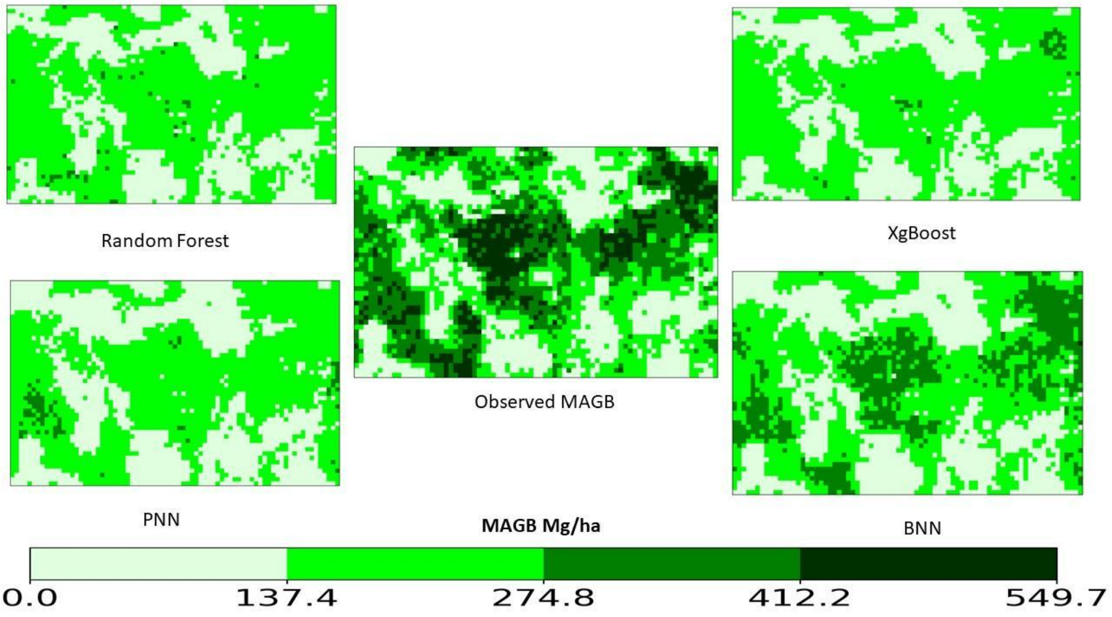


Figure 6. AGB plot of the predicted AGB values (including soil properties data) by the models Random Forest, XGBoost, PNN, BNN

6. Discussion

A neural network is a bulk computational structure that is inspired by structure and functioning of a human brain. It usually has connected nodes, also called as neurons/artificial neurons, that are organized in layers. The layers' architecture is the neural network's defining feature- the arrangement and the connecting pattern of these layers. The static picture of neural networks architecture involves of input layers, hidden layers, output layers, weight and bias, and activation functions. As a basic Bayesian neural network (BNN) composition closely resembles that of classical neural networks, though with a fundamental difference in the way weights and biases are handled. Instead of having a deterministic weights and biases parameters values in an ordinary neural network, in a probabilistic model, these parameters values are treated as probability distribution. This incorporation is based on an uncertainty and therefore an additional conditions must be taken into account when it comes to the structure as well as training methodology. One of the major disparities between the traditional neural networks and the newer types is that the latter can observe, and be more sensitive, to the spatial information in the input.

In a neural network it is assumed that input values are expressed as constant weights of connections and biases. Through this process, the value involved in this training is adjusted and minimized until a specific loss function has been measured. On the contrary, BNN is defined by effective weights and biases as probability distributions. This brings in an element of uncertainty to the model, enabling it not only to fit the most probable values to parameters but also allow a range of values corresponding to the potential values. The main difference between traditional neural networks and BNNs is in the way they provide predictions. Unlike conventional neural networks that provide point estimates, i.e. outputting single values for given inputs, BNNs output the distributions of the probabilities of the estimates, therefore including measure of uncertainties per each prediction.

On that note, BNNs focus on the probabilistic and nuanced modelling of uncertainty, which means that they are perfect for such applications as those where the evaluation of uncertainty is very crucial.

Although it may seem trivial hyperparameter tuning is one of the essential tool for avoiding both underfitting and overfitting. It replicates human thinking and enhances the capacities of generalization to new examples by the model. The robustness of the model is one of the benefits of hyperparameter tuning, as it optimizes for different sets of the dataset and prepares the model for real-world scenarios. We generally concentrate on essential hyperparameters in our model. For BNN, the hyperparameter tuning done are (1) The number of units in the first Dense layer; (2) The number of units in the second dense layer; (3) the Learning rate (4) Batch size Prior research done for estimation of the above-ground biomass merely focused on the machine learning models like - Random Forest, XgBoost, Support Vector Machines where there was a saturation of AGB values prediction on the higher values of AGB. This limitation of the machine learning model was to a certain extent eliminated in this study, where BNN was successful in predicting the AGB values in the high range. Moreover, the early research used the remote sensing data without consideration of other physical properties which could also impact the AGB. This research further demonstrates the importance of using soil properties data for the estimation of aboveground biomass. The combination of datasets is useful for the estimation of above-ground biomass as is evident from the results of two models namely XGBoost and BNN. However, its effect on the two models is not increasing the model accuracy.

Moreover, using multisource datasets helps to prevent the overfitting on a single-source dataset. Thus, increasing the generalizability of the models. The datasets taken into consideration well across patches for training and validation performed on an independent region of interest to estimate the biomass furthermore adds to the usefulness and acceptability of the deep learning model.

Most studies have used classical statistical regression approaches and machine learning algorithms (MLR, SVM, Random Forest, XGBoost). There exists a complex non-linear relation between the dataset and the vegetation. To address such a relationship we need a more advanced algorithm based on a neural network.

In our study, we have in all 20 predictor variables for above-ground biomass estimation, Using Deep learning models like BNN, addresses the problem of underestimation but the problem of overestimation is still an area of concern. Moreover, it provides improvements in processing the objective of regularized learning, to avoid overfitting. We have also assessed the importance of feature selection on machine learning algorithms to build a more efficient model on the Random Forest and XGBoost algorithms. We obtained the optimal set of features using sequential feature selection. Sequential feature selection which decides the optimal features using multiple rounds of training of algorithms. We did sequential feature selection where we got a subset of features comprising of features (i.e. Sentinel-2 bands, vegetation indices, and soil properties dataset) from the original predictors. This helped in dimensional reduction, assessing the important predictors, faster and more efficient hyperparameter optimization, and better convergence of the model. We have not used feature selection for Deep learning Models as they are robust to noise and the inclusion of irrelevant features. The DNNs can learn to assign low weights to less informative features during training, effectively ignoring them in the final representation.

Earlier research demonstrated the use the machine learning models and validation using small dataset, in contrast our results focus on a comprehensive dataset. Moreover **[14]** earlier results discussed the problem of saturation over higher AGB values using machine learning model discussed by *Yingchang Li et. Al (2020),* BNN was successful in mitigating the problem of saturation in AGB prediction despite of less number of high observed AGB values in training dataset. Also, the BNN's are well generalized models as they offer the higher determination accuracy as compared with other models. In addition, remote sensing data with higher spatial and radiometric resolution, such as LIDAR data and hyperspectral data, or the approach of mixed pixel decomposition, data cleaning, may be solutions for AGB estimation leading to higher accuracy in prediction which we will explore in our future research.

7. Conclusion

This study selected the region of Yellapur, Uttara Kannada, as a case study area to analyse the AGB estimation based on Sentinel-2 along with geological data, and their combination using different modelling algorithms, RF, XGBoost, PNN, BNN. The results indicate the following: (1) In this study, sentinel 2 Multispectral images are successful in predicting the estimated biomass. This can be used by the researchers to describe new areas in terms of biomass estimation in a full-fledged manner . After tuning the R2 score of the BNN model reached 0.66. (2) Machine learning algorithms give good results for biomass estimation. But the problems of high-value underestimation and low-value overestimation for these two algorithms exists, the deep neural network was like the BNN algorithm reduced this problem to a certain extent and made the AGB estimation results closer to that of observed biomass results. Meanwhile, the BNN and XGBoost models significantly improved compared with the PNN and Random Forest model(3). And, Explored the use of Soil properties data along with Sentinel-2 data showed the importance of the use of soil data along with remote sensing data(4). Upon Comparing with Sentinel-2, soil image has more accurate estimation of AGB for models like BNN and XGBoost(5). Thus, their combination helped improve the AGB estimation of two models while for the other two the data combination has not shown significant improvement .

Sentinel-2's data is accessible for a non-financial and comes with a high spatial resolution data as well as higher temporal resolution data. Also, machine learning and deep learning modelling mentioned in this study takes into consideration and also coordinates with the relevant optimizations which are useful while estimating biomass. The software and data implicated in the article are open source and referenced for learning consequently researching it makes it suitable for other students and professionals to employ easily, and basically they would not worry about the costs. We hope the results will turn students as well as researchers into grassroots environmentalists who will assert how biomass data can be useful and develop more appropriate models for biomass estimation. In long term, our team will investigate various data sources including longer path SAR, hyperspectral, high spatial resolution RS images with remote sensing time series data, using geographical information system (GIS) Digital Elevation Maps (DEM), for AGB calculation. In addition, we will focus at the model ensemble methods of AGB estimation..

8. Future Scope

Exploring Hyperspectral Dataset can offer precise control over spatial properties, enabling to focus on nitty gritty details throughout the spatial and spectral resolutions. Integrating our dataset with time series dataset can help in giving a clear picture of the most correlated bands among the different seasons which help in estimation of biomass. Also, integrating some biophysical properties like- DEM, Climatic conditions can help improve the model performance and its generalizability. Moreover, incorporating structural texture features along with the multispectral dataset can further push the machine learning models for a more relatable features which can help to estimate the Biomass.

In essence, the future scope in investigating the scope of Multispectral dataset for Biomass Estimation can further help in describing the model and building a robust model to be used in a very general way i.e. universally acceptable.

References

[1] M. L. Avolio, A. M. Hoffman, and M. D. Smith, 'Linking gene regulation, physiology, and plant biomass allocation in Andropogon gerardii in response to drought', Plant Ecol, vol. 219, no. 1, pp. 1–15, Jan. 2018, doi: 10.1007/s11258-017-0773-3.

[2] J. L. Araus and J. E. Cairns, 'Field high-throughput phenotyping: the new crop breeding frontier', *Trends in Plant Science*, vol. 19, no. 1, pp. 52–61, Jan. 2014, doi: 10.1016/j.tplants.2013.09.008.

[3] H. Ren, W. Xiao, Y. Zhao, and Z. Hu, 'Land damage assessment using maize aboveground biomass estimated from unmanned aerial vehicle in high groundwater level regions affected by underground coal mining', *Environ Sci Pollut Res*, vol. 27, no. 17, pp. 21666–21679, Jun. 2020, doi: 10.1007/s11356-020-08695-3.

[4] Yang, X. Assessing Responses of Grasslands to Grazing Management Using Remote Sensing Approaches; Library and Archives Canada Bibliothèque et Archives Canada: Ottawa, ON, Canada, 2013; ISBN 9780494923177.

[5] M.-L. Nordberg and J. Evertson, ‘Monitoring Change in Mountainous Dry-heath Vegetation at a Regional ScaleUsing Multitemporal Landsat TM Data’, *AMBIO: A Journal of the Human Environment*, vol. 32, no. 8, pp. 502–509, Dec. 2003, doi: 10.1579/0044-7447-32.8.502.

[6] R. A. Santillan, W. R. Ocumpaugh, and G. O. Mott, ‘Estimating Forage Yield with a Disk Meter [1]’, *Agronomy Journal*, vol. 71, no. 1, pp. 71–74, Jan. 1979, doi: 10.2134/agronj1979.00021962007100010017x.

[7] U. Lussem, J. Schellberg, and G. Bareth, ‘Monitoring Forage Mass with Low-Cost UAV Data: Case Study at the Rengen Grassland Experiment’, *PFG*, vol. 88, no. 5, pp. 407–422, Oct. 2020, doi: 10.1007/s41064-020-00117-w.

[8] ’t Mannetje, L.; Jones, R.M. Field and Laboratory Methods for Grassland and Animal Production Research; CABI Publishing: Wallingford,UK, 2000; ISBN 9780851993515.

[9] U. Lussem, A. Bolten, J. Menne, M. L. Gnyp, J. Schellberg, and G. Bareth, ‘Estimating biomass in temperate grassland with high resolution canopy surface models from UAV-based RGB images and vegetation indices’, J. Appl. Rem. Sens., vol. 13, no. 03, p. 1, Sep. 2019, doi: 10.1117/1.JRS.13.034525.

[10] M. A. Sanderson, C. A. Rotz, S. W. Fultz, and E. B. Rayburn, ‘Estimating Forage Mass with a Commercial Capacitance Meter, Rising Plate Meter, and Pasture Ruler’, *Agronomy Journal*, vol. 93, no. 6, pp. 1281–1286, Nov. 2001, doi: 10.2134/agronj2001.1281.

[11] D. Cheng, Y. Yao, R. Liu, X. Li, B. Guan, and F. Yu, ‘Precision agriculture management based on a surrogate model assisted multiobjective algorithmic framework’, *Sci Rep*, vol. 13, no. 1, p. 1142, Jan. 2023, doi: 10.1038/s41598-023-27990-w.

[12]X. Tian *et al.*, ‘Modeling forest above-ground biomass dynamics using multi-source data and incorporated models: A case study over the qilian mountains’, *Agricultural and Forest Meteorology*, vol. 246, pp. 1–14, Nov. 2017, doi: 10.1016/j.agrformet.2017.05.026.

[13]C. Chen *et al.*, ‘Estimation of Above-Ground Biomass for Pinus densata Using Multi-Source Time Series in Shangri-La Considering Seasonal Effects’, *Forests*, vol. 14, no. 9, p. 1747, Aug. 2023, doi: 10.3390/f14091747.

[14]H. Han, R. Wan, and B. Li, ‘Estimating Forest Aboveground Biomass Using Gaofen-1 Images, Sentinel-1 Images, and Machine Learning Algorithms: A Case Study of the Dabie Mountain Region, China’, *Remote Sensing*, vol. 14, no. 1, p. 176, Dec. 2021, doi: 10.3390/rs14010176.

[15] L. Breiman, '[No title found]', *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[22] M. Chen, Q. Liu, S. Chen, Y. Liu, C.-H. Zhang, and R. Liu, 'XGBoost-Based Algorithm Interpretation and Application on Post-Fault Transient Stability Status Prediction of Power System', *IEEE Access*, vol. 7, pp. 13149–13158, 2019, doi: 10.1109/ACCESS.2019.2893448.

[16]A. A. Rusu *et al.*, 'Progressive Neural Networks', 2016, doi: 10.48550/ARXIV.1606.04671.

[17] E. Ergün and B. U. Töreyin, 'Sparse Progressive Neural Networks for Continual Learning', in *Advances in Computational Collective Intelligence*, vol. 1463, K. Wojtkiewicz, J. Treur, E. Pimenidis, and M. Maleszka, Eds., in Communications in Computer and Information Science, vol. 1463. , Cham: Springer International Publishing, 2021, pp. 715–725. doi: 10.1007/978-3-030-88113-9_58.

[18] E. Goan and C. Fookes, 'Bayesian Neural Networks: An Introduction and Survey', in *Case Studies in Applied Bayesian Data Science*, vol. 2259, K. L. Mengersen, P. Pudlo, and C. P. Robert, Eds., in Lecture Notes in Mathematics, vol. 2259. , Cham: Springer International Publishing, 2020, pp. 45–87. doi: 10.1007/978-3

[19] Wang, Jie, et al. "Estimating Leaf Area Index and Aboveground Biomass of Grazing Pastures Using Sentinel-1, Sentinel-2 and Landsat Images." *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 154, Aug. 2019, pp. 189–201. *DOI.org (Crossref)*, https://doi.org/10.1016/j.isprsjprs.2019.06.007.