Enhanced Prediction of Tropical Tree Biomass Using Ensemble Models

Qiucai Dang
Zhumadian Preschool Education College, Zhumadian 463000, China
E-mail: Dqc336699@163.com

Keywords: above-ground biomass, below-ground biomass, ensemble stacking, grid search optimization

Received: July 3, 2025

The present paper aims to propose a novel model to investigate its utility in evaluating the beneficial effects of tropical forest biomass. To address the multiplicity of variables, as well as the complexity and nonlinear relationships between them, five Machine Learning (ML) models, namely Gradient Boosting (GB), Extra Trees (ET), XGB, ElasticNet, and Poisson Regression, were employed to concurrently predict both the below-ground and above-ground tree biomass (BGB and AGB, respectively), as well as the total biomass (TB = BGB + AGB). Since the results of the aforementioned models were not entirely satisfactory, an additional model called the Stacking Ensemble (SE) was introduced. Each model can have its parameters optimized by Grid Search with cross-validation to make sure that there is generalization and consistent performance. The data collected were based on 175 trees from 27 ecoregional plots located in the Central Highlands ecoregion of Vietnam. The dataset was processed to investigate the proposed model's ability to predict tree biomass. The study's findings revealed that the proposed method demonstrated strong and efficient predictive capabilities for biomass estimation in forest ecoregions. The Stacking model showed the most significant improvements in the highest R 2 (0.968) and VAF (0.971), and the lowest errors, and MDAPE (23.081 percent), which means that it has a strong ability to predict and minimal bias. However, STD (105.763) was marginally higher; nevertheless, the error and strength of this variation exceeded this variance. Thus, incorporating a Stacking Ensemble (SE) model strengthens the ML approach in predicting forest tree biomass amounts.

Povzetek: Študija predlaga ansambelski model za napoved tropske drevesne biomase, ki združuje pet ML-modelov in optimizacijo z iskanjem po mreži. Stacking Ensemble doseže najboljša napovedovanja ter najnižje napake, kar občutno izboljša oceno nadzemne, podzemne in skupne biomase.

1 Introduction

1.1 The role of biomass

Given that biomass plays an unquestionable role as one of the world's vital sources of energy [1]. The disputing matter is what appropriate model would be able to recognize and prove its traits. Zhantao Song et al. (2024) in their work discussed original visions about the concept of the pyrolysis process of biomass. They argued the contribution of various factors to the challenging anticipation of physicochemical traits by applying machine learning techniques such as Random Forest, gradient boosting decision tree, extreme gradient boosting, in which R^2 was higher than 0.97 for particular surface area biochar anticipation as well as analysis, involving yield as well as N content of biochar [1].

In another study, Jia et al. (2024) exploited machine learning methods to anticipate zeolite-catalyzed biomass pyrolysis, and as a result, the Random Forest algorithm performed the highest prediction with R^2 >0.91 for their suggested models. They concluded that their selected factors and methods based on biomass characteristics can be taken into account as a plausible reference [2].

1.2 Above-ground biomass (AGB)

The term above-ground biomass (AGB) refers to the product of above-ground volume (AGV) and vegetation mass. It is also closely linked to the carbon cycle in global grassland ecosystems. Additionally, accurate estimation of AGB variations is essential for assessing carbon decomposition and its impact on climate change. It is also crucial to screen in situ-harvested AGB data before modeling [3]. Furthermore, AGB is an indispensable factor for evaluating ecosystem health and carbon storage. To estimate AGB, the above-ground volume (AGV) of vegetation is considered a high-priority parameter in research [4].

To estimate AGB variations of China's grassland ecosystems, machine learning algorithms, among which the Random Forest model with R2 = 0.83 (i.e., 83 % of the harvesting AGB variations), and RMSE = 43.84 gm−2, revealed accurate performance in estimating grassland AGB [3]. Mao et al. (2021) in their proposed model proved that structural, textural, and spectral metrics contribute to shrub AGV models. They also suggested a direct reference to specify proper vegetation metrics to screen shrub AGV. The efficiency, accuracy, and low cost

are considered to be the pros of their proposed approach for digital terrain model (DTM) output and AGV estimation; thus, it can bridge the gap between ground-based research and satellite remote sensing [4]. May et al. 2024 obtained spatially complete predictions of biomass in a tropical area. They state that this sort of spatially coherent data about AGB supplied by their model is useful to validate the eco-friendly forest handling, carbon decomposition innovations, and climate change alleviation [5].

1.3 Below-ground biomass (BGB)

Below-ground biomass (BGB) is a significant part of forest tree biomass; however, fewer studies have focused on BGB about forest biomass and carbon. This is largely because the process of measuring BGB in large trees is costly and time-consuming. As a result, researchers often use Above-Ground Biomass (AGB) to estimate BGB by applying a root-to-shoot ratio. For different forest types, researchers have also developed specific direct BGB equations [6].

In a recent study, Oliveira et al. (2024) suggested that predicting peanut BGB using their proposed alternative method—i.e., the multi-output regression (MTR) approach—would enable both researchers and farmers to quantify BGB more accurately. They proposed this method to predict multiple peanut maturity indices at the field level, helping to reduce subjectivity in determining peanut maturity [7].

1.4 Ensemble approaches

Ensemble learning is a potent machine learning technique that reduces overfitting, boosts robustness, and enhances overall performance by combining predictions from several models. Ensemble approaches combine the advantages of multiple algorithms to improve generalization rather than depending on a single model [8]. Stacking, also known as stacked generalization, is a versatile and successful ensemble technique. Stacking mixes different kinds of models, possibly with different architectures and learning strategies [9]. In contrast to bagging (e.g., Random Forest) or boosting (e.g., Gradient Boosting, XGBoost), which combine similar models (typically decision trees). Naik et al. (2022) utilized automated stacked ensemble modelling powered by machine learning for predicting aboveground biomass in forests using multitemporal Sentinel-2 data [10]. A stacking ensemble algorithm was used by Zhang et al. (2022) to reduce the biases in estimates of forest aboveground biomass derived from several remotely sensed datasets [11]. Besides, Jin et al. (2025) evaluated the impact of validation techniques and ensemble learning algorithms on estimating aboveground biomass in forests: a case study of natural secondary forests [12]. To this end, they developed models based on various outcomes, qualified to synchronously anticipate AGB, BGB, and the total amount of tree biomass, i.e., TB, in forest areas, solving the problem of carbon estimation for various forest sites.

1.5 Regression models

It is appropriate to take a brief glimpse at the regression models proposed in the present article:

The Gradient Boosting (GB) model is regarded as a strong ML algorithm for numerical optimization problems. Thanks to Leo Breiman (1998) and Jerome Friedman (2001), GB has been developed. The former used GB for decreasing variance for categorization, and the latter improved it for regression and categorization models. GB algorithms carry out numerical optimization for the models of regression and categorization, repeatedly being approximately directed towards the loss function negative gradient. Due to some complexity, it is impossible to direct precisely towards a negative gradient; normally, a weak learner is applied by a GB model to estimate the extreme decline direction [13].

Extra Trees (ET), a recently developed regression model, is considered to be an ensemble ML algorithm related to decision trees. Originally, ET is the improved form of the Random Forest algorithm for the purpose of regression or categorization performance. The reason that makes the ET regression algorithm more competitive for small-sized sample ML is that it utilizes all data to improve the branches of nodes in decision trees effectively [14]. Wang et al. (2023) in their study provided an efficient ML model utilizing an ET regression algorithm for anticipating the relevant synthesis gas traits in the process of biomass chemical looping gasification, and then compared its ability in prediction between the ET model and traditional ones. In another study, using both RF along with ET al algorithm models, researchers developed a general model to precisely predict the co-pyrolysis of coal and biomass, in which ET performed better [15]. ET is advantageous due to achieving more efficient performance than the Random Forest. Compared to RF, ET does not perform bootstrap accumulation like, i.e. it takes a random subset of data without replacement. Hence, nodes are divided randomly, but not based on the best divisions. Therefore, in the ET regression model, randomness doesn't come from bootstrap accumulation but from the random divisions of the data [16]. According to Roy (2021), RF was introduced to overcome the Decision Tree problems, giving medium variance. Accordingly, ET was proposed when accuracy was more crucial than a generalized model. It also delivers low variance.

Extreme Gradient Boosting (XGB) is another strong, multifaceted ML algorithm used for regression and taxonomy jobs. It is well-known for its exceptional capability to predict performances and deal with intricate datasets. GB involves a series of procedures, preparing models in sequence, based on which the previously produced errors are reformed by each new model. XGBoost is a type of ensemble learning technique that mixes the predictions of various ML models to yield an ultimate prediction that is more precise. Besides that, this algorithm also makes use of decision trees like basic learners during its process. To add more, XGB is intended to efficiently influence processors of high-capacity and approaches of the distribution system [17]. Ayub et al.

(2023) applied an XGB algorithm model on a multi-level factorial design outcome to predict and improve the gasification product, in which the XGB model depicts a good prediction accuracy as well as model optimization analysis. The key characteristics of the XGB are explained as an ability to handle complicated relations in data with regularizing techniques, effectively preventing overfitting; thus, it performs the calculation efficiently due to parallel processing. It considers the usage of decision trees as base learners and then makes use of regularizing techniques for model generalization at a higher dimensionality. XGB, more popularly acknowledged for its efficiency in computations, provides processing efficiency with perceptive analysis of feature significance, as well as deals with missing values smoothly [18].

ElasticNet, being a powerful linear regression technique highly beneficial in ML and statistical modeling, excels traditional linear regression models. It bears the ability to mix the penalties created by both Lasso and Ridge regressions. It is useful in particular when traditional linear regressions struggle with multicollinearity, i.e., when predictors are highly correlated [19]. That is to say, ElasticNet is advantageous due to bearing multi-dimensional datasets, selecting significant traits, and being a more consistent and reliable model where there exists collinearity. Aimed to help in solving problems of regression and developing models' performance, ElasticNet offers effective analytical means for handling multi-dimensional regression. Its common applications include characteristics selection, analysis of regression, and modeling for prediction [20]. The significance of ElasticNet regression includes multicollinearity handling, automatic feature selection, aiding in model interpretability and reducing overfitting, flexible regularization, allowing researchers to control the balance between Lasso and Ridge penalties, robustness in high-dimensional data, appropriateness for a variety of regression problems [15].

Poisson is a regression analysis where its answer is based on the distribution called Poisson. The regression suffers from a limitation of the variance equaling the mean, called Equi dispersion. As a consequence of the assumption being violated, resulting in the biased standard error, the less exact test statistics drawn from the model, and consequently, the obtained conclusions will be less valid. The Poisson regression model, therefore, cannot be used under occurrences of over-dispersion or under-dispersion. Poisson regression is one of the generalized linear models. It finds its main application because it usually happens to model occurrences of the kind that are rarely occurring [21].

The Stacking Ensemble (SE) model makes use of an ensemble generalizing approach through learning, despite the fact that it may lack instructions for appropriate non-hyperparameterized meta-learners. The necessity of applying stacking is when multiple ML methods reveal various advantages for a certain task. In this case, the stacking ensemble method employs a discrete ML technique for specifying the efficient application of various algorithms [22]. For this reason, Arif et al. (2024) developed a model of stacking ensemble, by a non-homogeneous mixture of fundamental models, for accurate yet, at the same time, interpretable prediction of lung cancer prognosis so as to recognize crucial risk factors [18].

The use of DL methods is unquestionably dominant over other traditional methods, particularly in tropical forests biomass research [6]. Although many studies have investigated tree biomass anticipation by applying various models [6], the applied models are well-established. But lack combining models as ML, ensemble, and optimization of hyperparameters approaches. This work adds value by combining them using a meticulously designed Stacking Ensemble specifically designed for predicting AGB, BGB, and TB using a small, real-world dataset from 27 eco-regions in Vietnam. The Fit Index (FI), a stability-focused evaluation metric that hasn't been used in biomass prediction before, is introduced in this study. The proposed approach provides new methodological insights that improve prediction accuracy and generalizability in tropical biomass estimation by combining rigorous preprocessing, multi-target modelling within an ecological context, and systematic hyperparameter tuning through Grid Search. Furthermore, this work differs from earlier black-box DL applications in that it incorporates Shapley Additive Explanations (SHAP) for ecological feature interpretation, which offers important ecological insight. Hence, this study was conducted to serve the purpose of bridging this gap. This subject is an expansion of an ongoing strategy to integrate remote sensing inputs acquired using a satellite or a drone and a source of biomass determinations as measured on the ground in order to develop a spatially superior, and rooted business-time dynamic biomass forecast model. Besides the otherwise plausible analytical foundation of the process, the model is capable of capturing some facets of complex nonlinear responses and enhancing the accuracy of predicting biomass over wider geographical areas and timeframes due to the use of sophisticated Stacking ensembles, enabled by Grid Search and cross-validation. Besides, the climatic variables can be included to forecast the change in biomass distribution in the case of a future climate change scenario, which can provide a significant insight both in forest management and on carbon budgeting. That is to say, designing a new model qualified to anticipate tree BGB, AGB, as well as the total of tree biomass TB (i.e., TB = BGB + AGB) concurrently, will fulfil the requirement of estimating forest carbon. On this account, making use of a community of up-to-date regression algorithms to increase the reliability for the aforementioned parameters estimation, as well as that for the newly proposed model, will assist the progressing literature in the realm of forestry science. The study proposes that integrated ensemble models will anticipate tropical tree biomass better than traditional modeling systems; as a result, the model will be dominant over conventional ones. The study objectives are twofold: firstly, designing a model to concurrently anticipate tree AGB, BGB, and TB, guaranteeing additivity of tropical forests in Vietnam by the names of Dipterocarp and Evergreen Broadleaf, and secondly, cross-validating

errors compared to a traditional model, applying the same dataset as well as anticipators in the mentioned forests.

The rest of this paper is structured as follows. That is, Section 2 discusses detailed methodology, including materials and data used in this work. Section 3 presents numerical analyses, graphical analyses, and experimental results under the heading of results and discussions. Lastly, section 4 summarizes the concluding points in the study.

2 Methodology

The present paper aimed to investigate the efficiency of a state-of-the-art model to be qualified for predicting tropical forest tree biomass effects.

This study was conducted in one of Vietnam's eight highest tropical forests, called the Central Highlands ecoregion. Two main tropical forest categories were selected for the focus of the research, i.e., Dipterocarp and Evergreen Broadleaf (See Fig. 1).

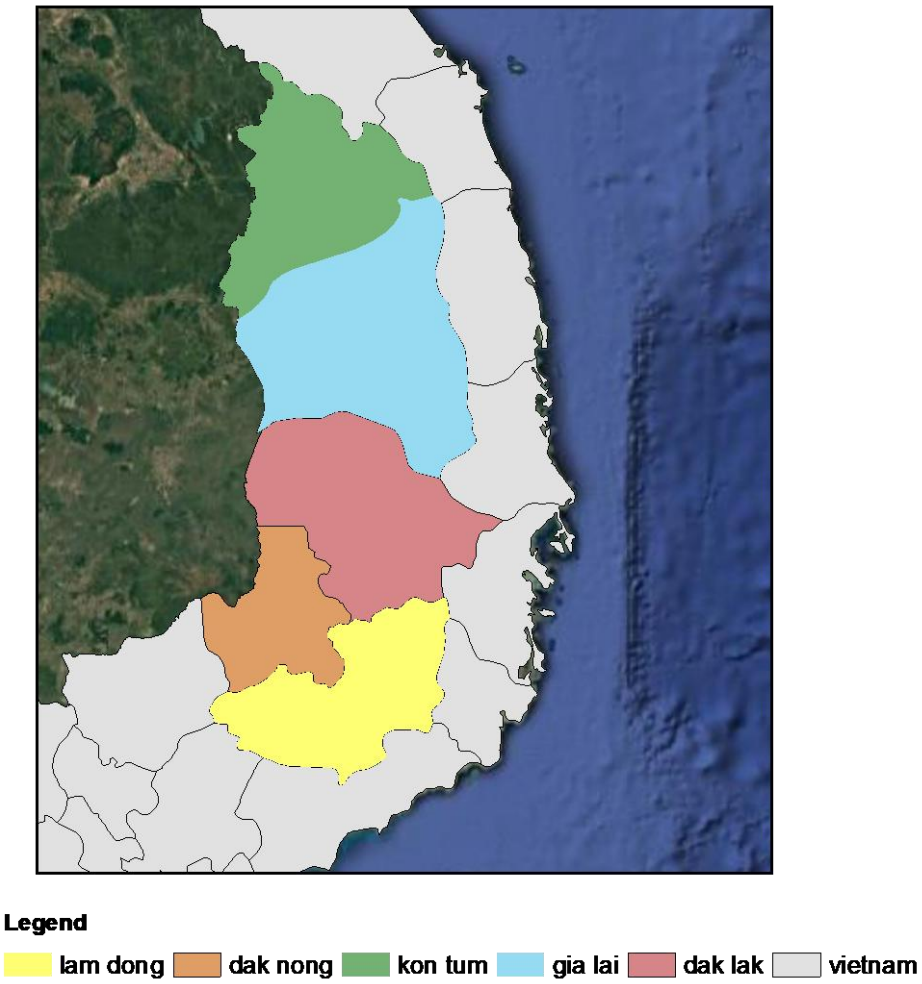


Figure 1: Sample plots, Locations for the forests Dipterocarp and Evergreen Broadleaf in the ecoregion of Central Highlands, Vietnam

In this work, the dataset was exploited in a research study conducted by Huy et al. [6]. The collected data were based on 175 trees from 27 ecoregional lots located in the Central Highlands, Vietnam. We clearly define the dataset partitioning strategy to ensure reproducibility: the entire dataset of 175 samples was randomly divided into training (80%) and testing (20%) sets. Cross-validation was used over iterations to ensure robust evaluation and minimize sampling bias. To ensure compatibility across models and better convergence during training, feature preprocessing involved removing outliers and normalizing all input variables to a [0,1] range using Min-Max scaling. The hyperparameter tuning process was carried out using Grid Search with 5-fold internal cross-validation for each machine learning model: Poisson regression, ElasticNet, XGB, Extra Trees (ET), and Gradient Boosting (GB). This allowed us to systematically explore parameter combinations and choose those that produced the best performance on training data. Based on the results of cross-validation, the Grid Search methodically investigates a predetermined set of hyperparameter values to determine which combination produces the best model performance.

A customized grid of important hyperparameters was built for every model. For instance, tree-based models such as GB, ET, and XGB had their learning rate, maximum depth, and number of estimators adjusted. We adjusted the L1 ratio and alpha (regularization strength)

for ElasticNet. Likewise, pertinent parameters for the stacking meta-learner and Poisson regression were adjusted. In order to maximize generalization and performance consistency, Grid Search was used in a cross-validation framework to guarantee that each model was trained with the best parameter settings. This method greatly enhanced both the Stacking Ensemble's overall performance and the accuracy of the individual models.

The desired targets in this dataset included the amount of above-ground tropical biomass (AGB), the amount of below-ground tropical biomass (BGB), and (TB), namely the total tropical tree biomass; equaling the summation of the below-ground and above-ground tree biomass (i.e., TB = BGB + AGB). Moreover, preprocessing and normalization operations were done on the data.

To serve the purpose of the study, five ML models, as a base learner including GB, ET, XGBoost, ElasticNet, and Poisson, were employed to synchronously anticipate both the amount of below and above-ground tree biomass (BGB and AGB, respectively) as well as the total amount of tree biomass, i.e., TB = BGB + AGB.

Owing to the individual models' mediocre performance, these five models were used as base learners to create a Stacking Ensemble (SE). Following that, a meta-learner was trained using their predictions to generate the final prediction for every biomass component.

For the purpose of assessing as well as selecting the most efficient model able to concurrently anticipate tropical tree BGB, AGB, and TB, a powerful process of cross-validation was carried out.

The Total number of the data was 175 which was randomly split ten times into two sections, involving 140 (80%) for training data, and 35 (20%) for testing data, evaluating impartially. The reason why the data was altered into data testing and training data was to conduct a data analysis satisfying accuracy and reliability in this research. A wide range of assessment metrics, such as MSE, RMSE, MAE, R2, STD, NMSE, MDAPE, and VAF, were used to evaluate performance.

the Fit Index (FI), a goodness-of-fit metric intended to assess the quality of predictions across several cross-validation realizations. A higher FI value indicates a better fit, with values approaching 1. The formula for calculating the FI is presented below.

$$FI = \frac{1}{k}\sum_{1}^{k}(1 - \frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{m}(y_i - \bar{y}_i)^2} \quad (1)$$

In the equation above k stands for the realizations number (in this study k = 10), m is the number of trees sampled in the validation dataset; and yi is the observed value. $\hat{y}$i represents the predicted value, and $\bar{y}$ shows the averaged value for BGB, AGB, and TB of the ith validated tree in the realization of kth.

The study goal was to evaluate accuracy and model consistency in light of the ecological context and the small dataset size. Metrics like R2 and VAF measure the percentage of variance explained by the models, while MSE, RMSE, and MAE quantify absolute prediction errors. Understanding normalization effects and error distribution is aided by STD and NMSE. MDAPE is a reliable percentage-based metric that works especially well with data that contains outliers or skewness, which is typical in biomass measurements. A new and comprehensible metric designed for model comparison across several validation folds, the Fit Index (FI) was introduced to reward accuracy and stability. Ultimately, when combined, these metrics make sure that the assessment covers robustness, interpretability, and predictive accuracy—all of which are critical components for ecological modelling and decision-making, demonstrating that the Stacking Ensemble model was more efficient than the other compared models. Evaluation methods for error metrics criteria are exhibited in Table 1.

Table 1: Equations for evaluation of statistical metrics criteria

Statistics	Name	Equation
MSE	Mean Squared Error	$MSE(y,\hat{y}) = \frac{\sum_{N-1}^{i=0}(y_i - \hat{y}_i)^2}{N}$
RMSE	Root Mean Square Error	$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$
MAE	Mean Absolute Error	$MAE = \frac{\sum_{n}^{i=1}\lvert y_i - \hat{y}_i\rvert}{n}$
R^2	Determination Coefficient	$R^2(y,\hat{y}) = 1 - \frac{\sum_{N}^{i=1}(y_i - \hat{y}_i)^2}{\sum_{N}^{i=1}(y_i - \bar{y})^2}$
STD	Standard Deviation	$STD = \sqrt{\frac{\sum_{n}^{i=1}(x_i - \bar{x})^2}{n-1}}$
NMSE	Normalized Mean Square Error	$1 - \frac{\lVert x - y\rVert_2}{\lVert x - \bar{x}\rVert}$
MDAPE	Median Absolute Percentage Error	$median\left(\left\lVert\frac{e_1^{abs} - e_1^{\widehat{pre}}}{e_1^{abs}}\right\rVert\right) * 100\%$
VAF	Variance Account Factor	$\left(1 - \frac{var(t_n - y_n)}{var(t_n)}\right) * 100$

n represents observations number, y_i is ith observed value, $\hat{y}_i$ shows ith predicted value, and $\bar{y}$ is the observations average.

Graphical analyses were also carried out to assess the accuracy of the recommended model performance. Illustrated in different plots, they give the reader illuminating perceptions of the suitability and accuracy of the models, all which have been discussed in section 3 of this paper.

As an overview of the research, the general flowcharts of the whole study have been demonstrated below (See Fig. 2 and Fig. 3). Fig. 2 also illustrates a brief comprehension of the step-by-step research methodology. That is to say, the research process begins with the dataset, going through analyzing and normalizing them, next, dividing the normalized data into train and test. More important part is here where the proposed ML models are evaluated based on an array of specific metrics to opt an appropriate model which is appeared to be Stacking Ensemble. Finally, ensemble models are also assessed on the basis of evaluation metrics to choose the best one. Hence, the results are saved for future usage.

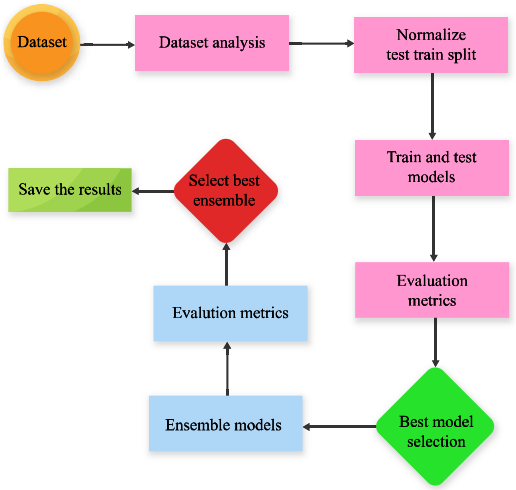


Figure 2: General flowchart of the whole research process for applying the proposed model

Figure 3 shows the modeling procedure involving data collection process for the purpose of theory, and then applying six ML models to concurrently predict tropical forest tree biomass and specifying the best reliable model for such prediction by comparing the selected ML models with the aid of evaluation metrics.

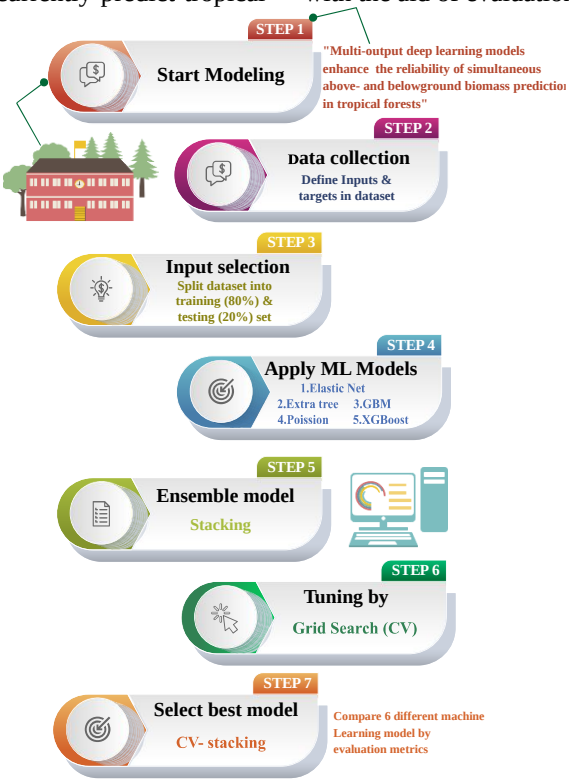


Figure 3: Flowchart of modeling procedure showing the process of employing ML models concurrently

The flowcharts of the six regression models, including ElasticNet, GB, ET, XGB, Poisson, and SE have been illustrated respectively in the following figure. The optimization of hyperparameters is also utilized through Grid Search tuning. These models were utilized with the goal of synchronously predicting ABG, BGB, and TB = BGB + AGB. Additionally, each proposed regression will also be discussed briefly below.

2.1 Base machine learning models

ElasticNet regression is an extension of linear regression that integrates both regularizing penalties of Lasso (abbreviated as L1) and Ridge (abbreviated as L2) into the loss function. This combination allows ElasticNet to deal with circumstances where there are a large number of characteristics, and they are also highly correlated. Its mathematical formulation is shown below.

$$ElasticNet = \sum_{i=1}^{n}(y_i - y(x_i))^2 + \alpha \sum_{j=1}^{p}|w_j| + \alpha \sum_{j=1}^{p}(w_i)^2 \quad (2)$$

In Elastic Net regression, the parameters alpha and l1_ratio bear significant roles in specifying the regularization technique used in the model. These parameters control the trade-off between the L1 and L2 penalties. In the presented formula α is the regularization strength parameter in ElasticNet. It supervises the whole strength of regularization applied to the model. For $\alpha = 0$, no regularization is applied, and Elastic Net equals Ordinary Least Squares (OLS) regression. For $\alpha = 1$, regularizations of both L1 and L2 are applied, blending their penalties. For 0<α<1, this model employs a mixture of L1 and L2 regularization, permitting a flexible mixture of penalties. L1 Ratio (l1_ratio) is the blending parameter that identifies the balance between L1 and L2 penalties. It controls the proportion of the penalty determined to the L1 norm relative to the L2 norm. For l1_ratio=0, the model applies only regularization of L2 (which equals Ridge regression). For l1_ratio=1, it uses merely regularization of L1 (which equals Lasso regression). For 0<l1_ratio<1, Elastic Net applies a mixture of both L1 and L2 regularization, allowing for a combination of penalties [23].

As shown in Fig. 4, applying the Elastic Net model in this study involves several linear steps. Because of the multicollinearity between predictors that are specific to trees and sites, ElasticNet regression was used. It made feature selection and coefficient shrinkage possible at the same time by combining L1 and L2 penalties. Grid Search with 5-fold cross-validation was used to optimize the regularization parameters (l1_ratio and alpha), which enhanced the generalization and stability of the model.

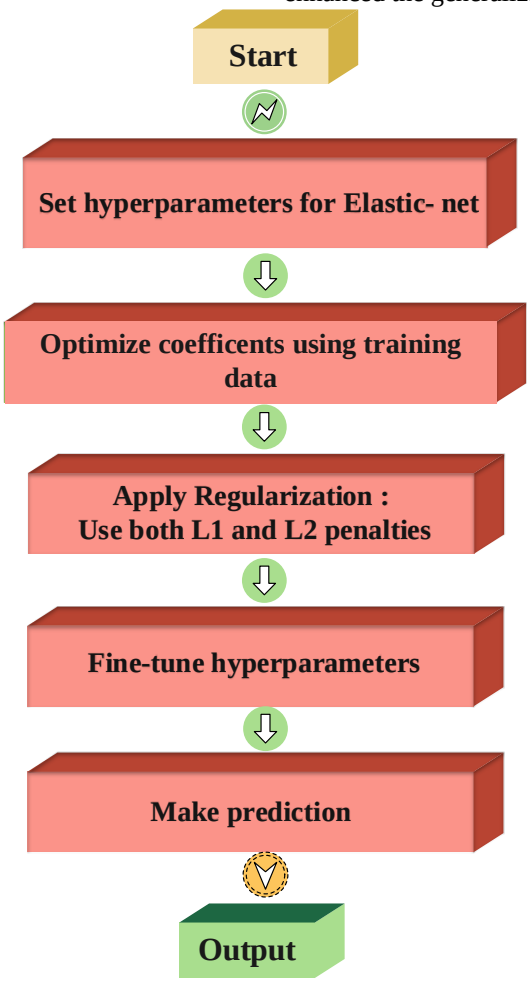


Figure 4: The steps of the ElasticNet model applied for predicting tropical tree biomass

Extra Trees (ET) regression is a type of ensemble ML strategy that accumulates the outcomes of various decision trees decorrelated, similar to Random Forest regression [24]. ET regression employs a conventional top-down technique to create a collection of regression trees. RF model applies two stages, respectively, including bootstrapping and bagging [25].

Hameed et al. (2021) have discussed the Random Forest model as an array of decision trees in their article, and used its equation as follows:

$$T(x, \theta_1, \ldots, \theta_r) = \frac{1}{R}\sum_{r=1}^{R} T(x, \theta_r) \quad (3)$$

In which T (x, θr) demonstrates the Tth tree prediction, in which θ presents a uniform independent distribution vector appointed before the tree growth. All these trees are blended and averaged in an ensemble of them (i.e., shaping forest), named T(x).

According to Hameed et al. (2021), two main differences between the ET and the RF systems are cited as follows. Firstly, two main differences between the ET and the RF systems are cited as follows. Firstly, ET exploits all the divided nodes as well as cutting points, selecting randomly from the cut points. Secondly, the algorithm applies all the samples to help the tree grow so as to limit bias.

Two parameters involved in the ET model for controlling the splitting process are k and nmin; k represents the characteristic number, chosen randomly in the node, while nmin refers to the minimum size of the sample anticipated nodes division. Moreover, respectively via k and nmin, the feature selection strength and the average strength of output noise are specified. The abovementioned two factors enhance the accuracy and decrease the ET model overfitting [19].

Figure 5 demonstrates the design of the ET model proposed in the present study. According to Fig. 6, training data are processed through the ET model as follows. N predicted outputs result from N tree decisions. Consequently, the obtained output predictions are averaged to result in the optimal output.

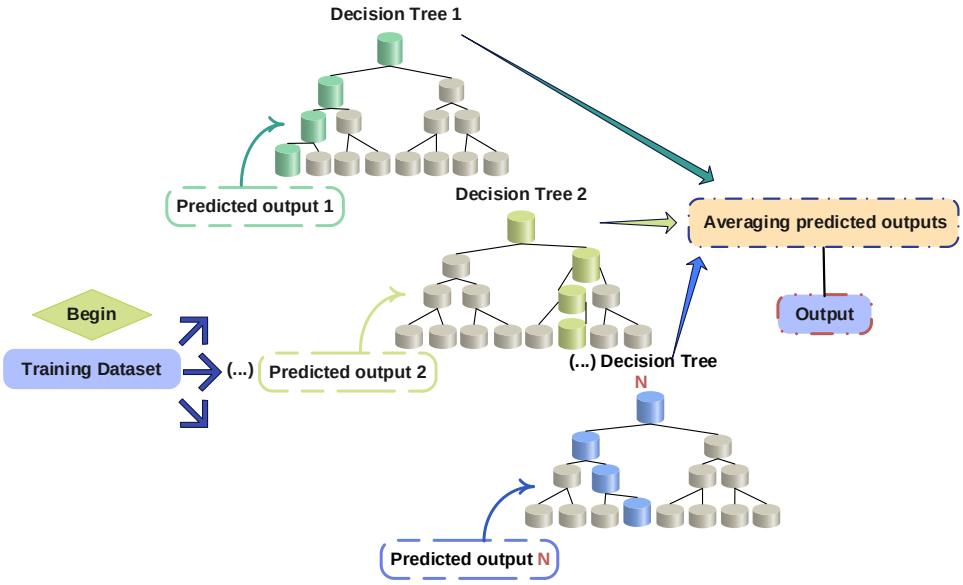


Figure 5: The steps of Extra Trees (ET) model used for predicting tropical tree biomass

The Gradient Boosting (GB) algorithm's job is to find a function $T(x_i)$, minimizing some loss function $\mathscr{L}$ [$T(x_1),\ldots, T(x_n)$], in which x_i is a vector with k dimensions for i=1,…, n. This algorithm begins with a primary prediction $T_0(x_i)$ and carries on repeatedly so that $T_m(x_i) = T_{m-1}(x_i) + h_m(x_i)$. Supposedly, $h_m(x_i)$ would experience extreme reduction direction in $\mathscr{L}$ [$T_{m-1}(x_1),\ldots,T_{m-1}(x_n)$]. Such direction is delivered via $\mathscr{L}$ with a negative gradient assessed in [$T_{m-1}(x_1)\ldots, T_{m-1}(x_n)$]. Despite being very demanding or sometimes impossible for a function detection of $h(x_i)$ to approximate $\mathscr{L}$ assessed in [$T_{m-1}(x_1),\ldots, T_{m-1}(x_n)$], when $h(x_i)$ approximation is estimated, GB algorithm progresses through $T_m(x_i) = T_{m-1}(x_n) + \alpha h_m(x_i)$ for $\alpha>0$ which is a supposed learning rate [9].

As the GB model is represented in Fig. 6, the data first goes through a bootstrap sampling to be split into T data subsets for which there would be hT tree decisions. Thereafter, there is a one-to-one result for each tree decision; altogether, hT(x) results. Ultimately, the attained results' average is calculated to produce the final result, namely H(x). The ability of Extra Trees (ET) and Gradient Boosting (GB) to model intricate non-linear relationships—which are common in ecological systems—led to their selection. With a small dataset size (n=175), ET's randomized split selection was especially helpful in reducing overfitting. To attain the best bias-variance tradeoffs, we employed Grid Search to adjust variables like the number of estimators, tree depth, and leaf size.

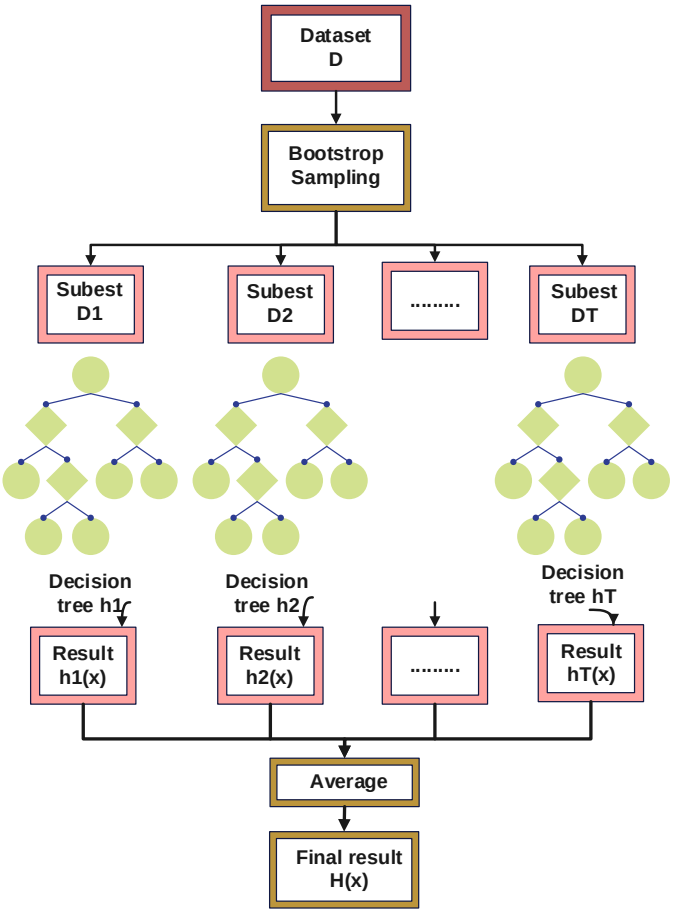


Figure 6: The stages of Gradient Boosting (GB) model employed for tropical tree biomass predictions

Given that the Poisson distribution is the basis for Poisson regression, it identifies the probability of the occurrence of any number of events in a steady interval of time or space, with the assumption that the events occur at a constant rate and are independent of each other. The formula for the Poisson distribution is given by:

The Poisson distribution could be calculated from the formula below:

$$P(X = k) = \left(\frac{\lambda^k e^{-\lambda}}{k!}\right) \quad (4)$$

In the above equation, X is the number of random occurrences, and λ represents the average or mean of the events. Poisson regression exploits its distribution to provide a comprehension of the predictor variables' relationships along with that of the count data in the dataset. In this regression, the expected value (mean) of the count variable (namely Y) is designed as a model of a linear mixture of predictor variables (namely X):

$$\lambda = \exp(\beta 0 + \beta 1X1 + \beta 2X2 + \ldots + \beta nXn) \quad (5)$$

in which: λ is the expected count, which represents the occurrence proportion, β0 is the intercept term, β1, β2, …, βn represent coefficients related to each predictor variable. The link function in Poisson regression is the natural logarithm (log-link), ensuring the predicted values are not negative. This model is evaluated via maximum likelihood estimation, and the coefficients (β) are specified to maximize the probability of observing the actual count data in the model [26].

The approach for the application of the Poisson model is well-illustrated in Fig. 7, which experiences various stages in a linear pattern. To begin with, a point cloud is taken as an input; second, the surface normal of all the points is detected by computing the eigenvector over the k-nearest neighbors of each point. Third, an octree with a predefined depth d is selected for categorizing the reconstructed surface. Then, the Gradient of the indicator function (Vx) equated to the vector V is defined by the point cloud. The next stage involves defining an indicator function X with the value of 1 inside and 0 outside the surface. Thus, Vx=V and the divergence operator is applied to either side; i.e. $\Delta x \equiv \nabla.\nabla x = \nabla.V$. On the next stage, the indicator function x is solved as a standard Poisson problem. The marching cube algorithm is used to extract the surface from the solved indicator function x. Eventually, the reconstructed surface is stored in the octree of depth d. Since AGB, BGB, and TB are skewed and non-negative, Poisson regression was employed. Although it was initially created for count data, its formulation fits biomass distributions quite nicely. To make sure the Poisson model's assumptions held true in this situation, diagnostic tests were conducted.

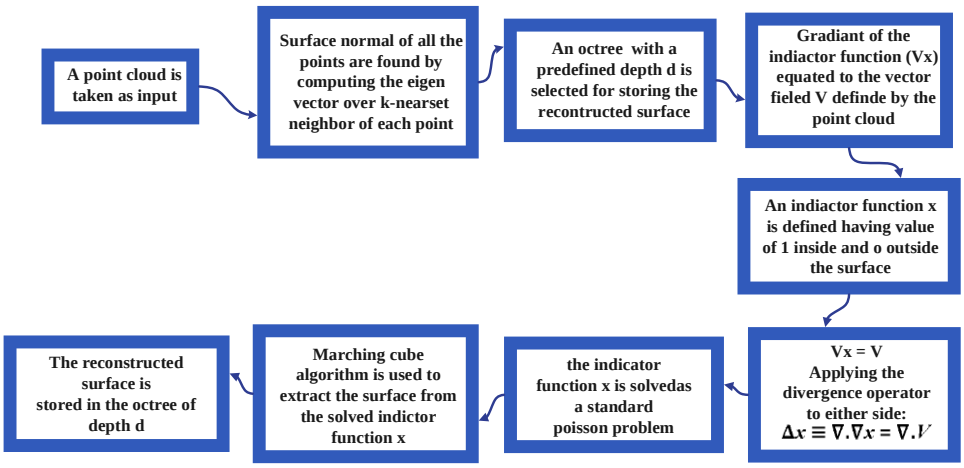


Figure 7: The stages of the Poisson model employed for tropical tree biomass predictions

2.2. Hyperparameter tuning

Grid Search was applied in a 5-fold cross-validation framework to find optimal hyperparameters for all the models. For tree-based models like Gradient Boosting, Extra Trees, and XGB, the number of estimators, learning rate, and max tree depth were systematically changed to optimize model complexity and predictability. In the ElasticNet scenario, the regularization parameter (alpha) and L1 fraction were tuned to prevent overfitting and cause sparsity. In Poisson regression, tuning was for the regularization parameters and the number of iterations to achieve better convergence. After separately tuning each of the base models, their outputs were fed into a meta-learner in the Stacking Ensemble, whose parameters also were tuned via Grid Search. This broad tuning process ensured that all models, including the ensemble, reached optimal generalization and performance [27].

The Stacking Ensemble was selected due to the meta-learner included, as it blends heterogeneous base learners with varying predictive ability and error behaviors, as well as generalizes well. Due to its nature of broad application in addressing multicollinearity on regression prediction of the base models and capturing of non-linear relationships, a tree-based learner was applied as the meta-model in this research. Stacking model showing an RMSE of 18.298, MAE of 12.422, and R2 of 0.968 performed significantly better compared to any of the base models on the test data, meaning that the ensemble was able to selectively leverage the strengths of each of the related models to generate more stable and accurate predictions of biomass.

In the Stacking model, presented in Fig. 8**,** training data are processed on the basis of three level 0 models separately. Each model's prediction results are gathered as other processed training data in the study. All of the base learners' predictions (GB, ET, XGB, ElasticNet, and Poisson) were aggregated by the Stacking Ensemble. To avoid overfitting and information leakage, the meta-learner, a Ridge regression model, was trained on out-of-fold predictions. We were able to improve overall predictive performance by combining the complementary strengths of all models—capturing distribution-specific, linear, and non-linear trends—into this ensemble. Table 2 provides the hyperparameters chosen by the stacking meta learner for the models.

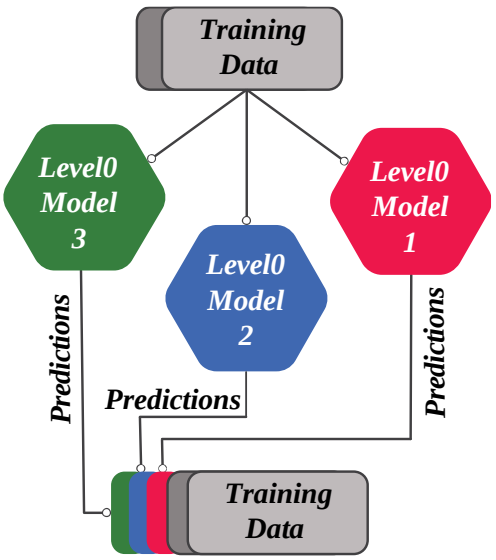


Figure 8: Stacking model procedures used for tropical tree biomass

Last but not least, XGB is one of the common algorithms in ML. It is based on the ensemble learning framework, following the gradient boosting algorithm. Thus, it is applicable for the tasks of supervised learning, i.e., regression, ranking, and categorization. XGB is a predictive model that combines multiple individual models' predictions iteratively. It works by adding weak learners into the ensemble one after another, such that at every step, a new learner tries to correct the errors of the prior ones. It also minimizes a prespecified loss function during training data using some sort of gradient descent optimization [13].

In summary, the XGB is developed in three stages straightforwardly: First, a primary model, namely F0, was used to predict, i.e. the aimed variable. The XGB model is related to a residual (y–F0). Second, the residuals obtained in the prior stage are adapted to a new model called h1. Third, the combination of F_0 and h_1 delivers F_1, which is the promoted form of F_0. Consequently, the MSE metric system from F1 will be lower than that from F_0.

$$F_1(X) < -F_0(x) + h_1(x) \tag{6}$$

For improving F1's performance, a residuals model of F1 can be designed, and an original model called F2 is presented.

$$F_2(X) < -F_1(x) + h_2(x) \tag{7}$$

This process would be iterated for a number of 'n' stages up until potentially minimizing residuals as much as probable, i.e.

$$F_n(X) < -F_{n-1}(x) + h_n(x) \tag{8}$$

It is worth mentioning that additive learners would not mess with the functions developed in prior iterations but add information of their own in order to bring down the error values. First, the model begins with some function F0(x). This F0(x) needs to minimize the loss function or MSE, hence:

$$\begin{aligned} &F_0(x) = argmin_\Upsilon \textstyle\sum_{i=1}^{n} L(y_i, \Upsilon) \\ &argmin_\Upsilon \textstyle\sum_{i=1}^{n} L(y_i, \Upsilon) = \\ &argmin_\Upsilon \textstyle\sum_{i=1}^{n} (y_i - \Upsilon)^2 \end{aligned} \tag{9}$$

Regarding the prime differential of this equation with γ, it is observed the function is minimized at the mean i=1, …, n. Thus, the promoting model can proceed with:

$$F_0(x) = \frac{\sum_{i=1}^{n} y_i}{n} \tag{10}$$

F0(x) presents the first step of predictions in this model. Next, for each instance, the residual error is expressed as: (yi – F0(x) [28].

In Fig. 9, the XGB model employs a multifaceted approach to make predictions about input data. Afterwards, the average of predictions is calculated and an ultimate XGB prediction is thus generated .Because of its exceptional performance with structured tabular data and its integrated regularization, which helps avoid overfitting, XGB was included. It was well-suited for this task because of its efficient handling of non-linearities, support for missing values, and robustness to noise, even with the small sample size. Grid Search was used to optimize important hyperparameters, such as learning rate, maximum depth, and gamma.

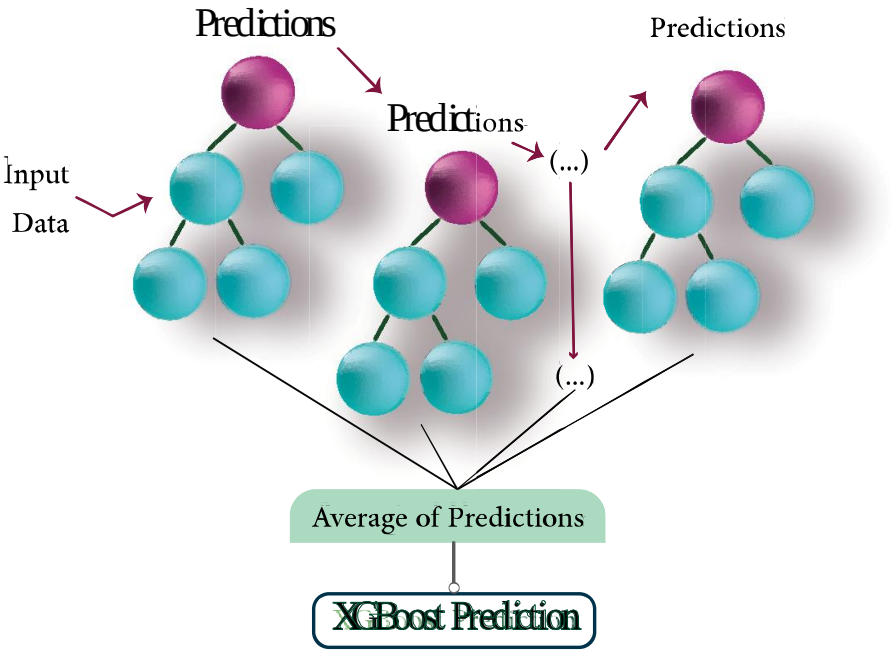


Figure 9: The procedure used in the XGB model for predicting tropical tree biomass

3 Results and discussion

3.1 Exploratory data analysis

To display how closely related the multiple variables of the study data are, a Pearson correlation heatmap is exploited as an effective color-coded visual matrix (See Fig. 10). Variables are demonstrated with rows and columns, and the cells define the relationship between variables two by two. The color shading for each cell indicates the correlations' direction and their strength: the darker the color of a cell, the stronger the correlation of the related variables. As it is obvious from this tabulated heatmap, the colors are darker for stronger correlations and lighter for weaker ones. Additionally, the green colors represent the positive correlations; that is, when one variable increases, the other variable tends to go up, whereas in the case of negative correlations, when one variable increases, the other variable tends to drop. Purple colors have been used.

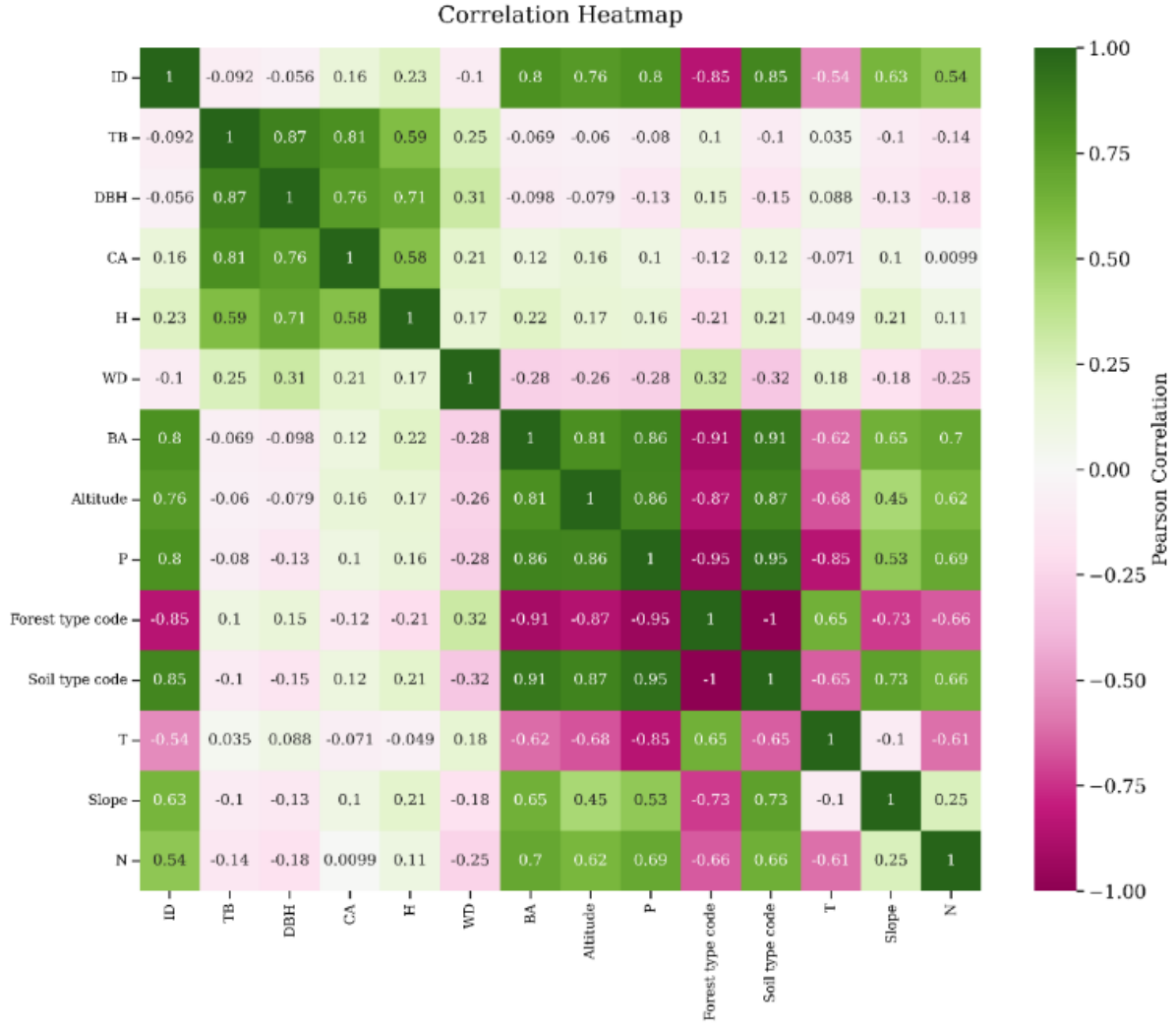


Figure 10: Pearson Correlation Heatmap for detecting the relationship between studied variables.

In Fig. 11, a pair plot visualization for the distribution of dataset parameters is shown for exploring the analysis of the data. In a pair plot, the data is visualized to find the relation between them, where variables are continuous as well as categorical, or form the most divided clusters. Dispersions of the parameters indicate the fact that most features are not evenly distributed. CA, WT, and P are skewed or clustered, and the values of these variables are focused on particular ranges. Scatter plot graphs such as CA versus WT or HA versus CA show positive relationships, which hold good, indicating potential multicollinearity that could be important to model. As a contrast, the variable types like forest type code and soil type code arrive in horizontal bands or discrete groups, as they are categorical. These trends suggest that the explanatory power of the data set is in part due to a mixture of continuous gradations in combination with categorical differences. The distributions in classes are depicted by the colors, and it can be observed that there is clearly a grouping in the plots, either of altitude, or of CA, or of WT. Following is a pair plot providing a high interface level to derive enlightening statistical information about the dataset; i.e., the variations in each plot can be observed, and the crucial diagonal secondary plots show each variable distribution. This pair plot for the relationship between variables of total amount of biomass, namely TB, is also demonstrated in Fig. 12, which more explicitly explores how the CTB classes are distinguished in terms of predictors. In this instance, the scatter plots show that for most variable pairs, the CTB categories are

predominantly overlapping, indicating that no set of variables can completely separate the CTB classes. However, there are regions, particularly in pairings like CA vs. WT or CA vs. P, where some CTB groups are grouped more closely together or are bunched into more constricted value ranges. The histograms on the diagonal also emphasize the bunched character of observations within given intervals, further underscoring that the dataset is skewed in variable distribution. This class grouping within specific regions suggests that individual variables may not always be able to differentiate CTB, but groups of predictors likely have predictive value. Further, the mixture of continuous and discrete variables introduces difficulty, as seen in the scatter plots, where some categories of CTB extend across different bands, with others overlapping.

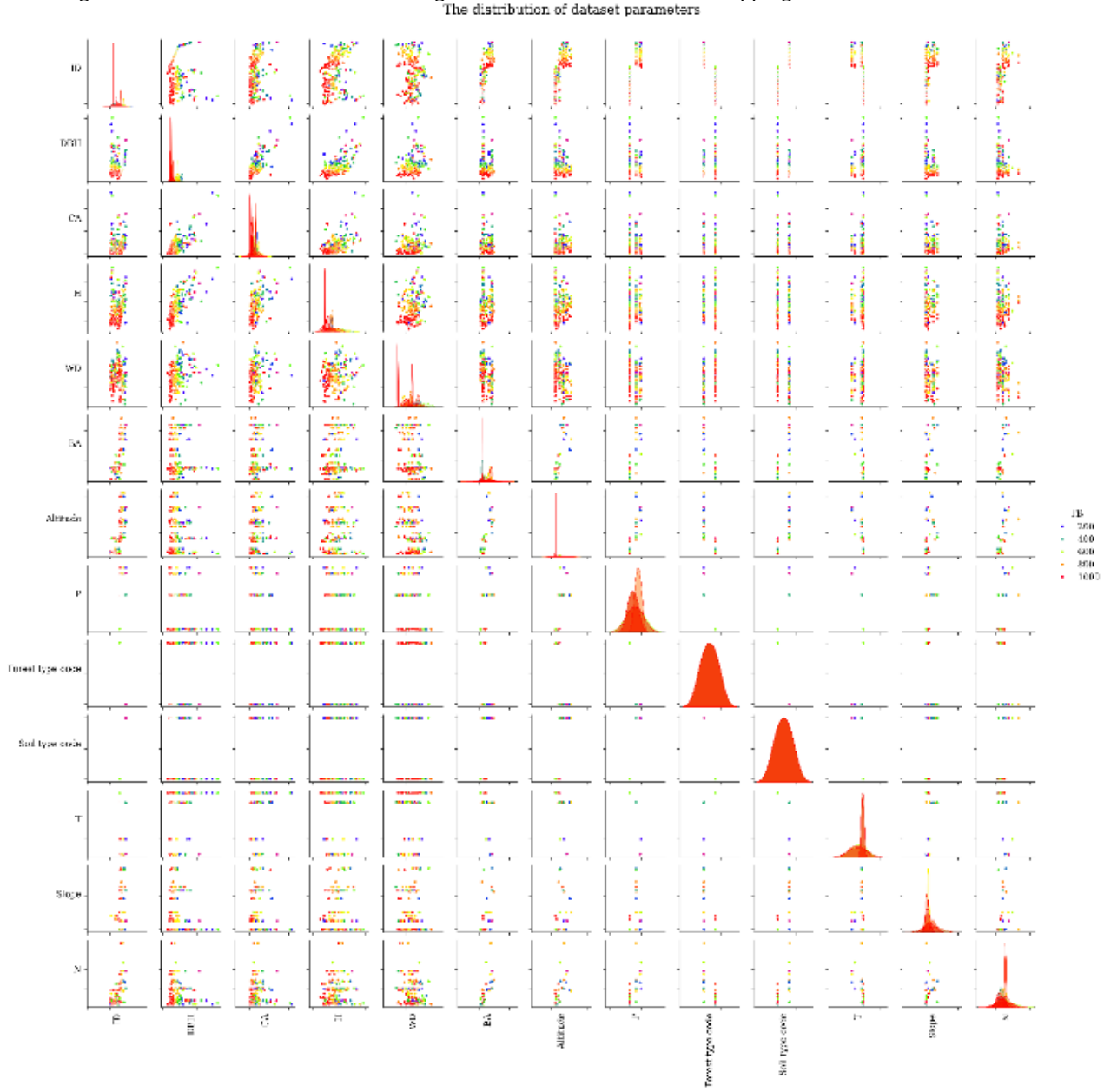


Figure 11: Pair plot for specifying the distribution of dataset parameters as well as their relationship

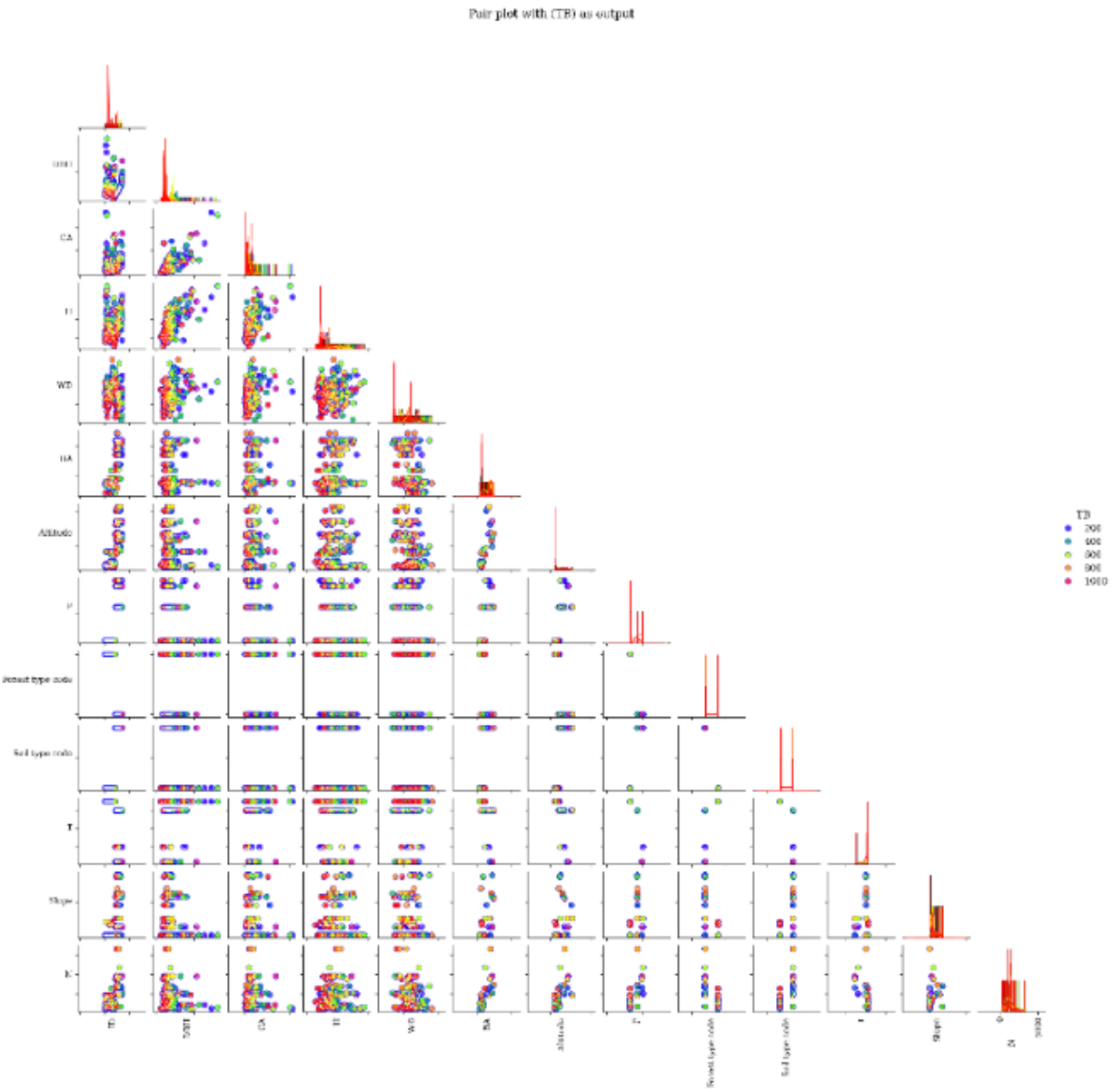


Figure 12: Pair plot for showing the relationship between the variables of total tropical tree biomass (TB) and their distributions

3.2. Machine learning results

In the present investigation, six methods were employed in two forest locations, i.e., Dipterocarp and Evergreen Broadleaf, to concurrently anticipate BGB, AGB, and TB = BGB + AGB. To assess the base models' performance, Table 3 presents the findings of the error metrics criteria for each recommended models considering the train and test data. By comparing these error metrics along with FI, it was detected that the Stacking Ensemble model was optimal than other models. The very large R 2 values close to 0.999 on the training dataset is an indicator of overfitting or data leaks. To prevent this we made sure to have rigorous separation of the training and test data and we optimized our hyperparameters using the Grid Search with cross-validation to prevent overfitting. The rock-bottom R2 values on the testing data (such as 0.962) are indicative of a lack of overfitting, so the overfitting appears to be contained. Further improvements with regard to regularization and data augmentation will be necessary in future computations to minimize the chances of overfitting. The test results indicate that the Stacking model outperforms the others in nearly all metrics, demonstrating higher predictive accuracy and reliability. Its mean squared error (MSE) is considerably low at 334.82, indicating lower average squared discrepancies between the predicted value and actual value compared to other models like ElasticNet (2378.17) and Extra Trees (1216.54). In the same vein, root mean squared error (RMSE) for Stacking is 18.30, a far cry from those of ElasticNet (48.77) and Gradient Boosting (41.75), meaning they were more precise in their predictions. Mean absolute error (MAE) performs the same, at 12.42 for Stacking, a far better performance than for models such as Poisson regression (19.32) and XGB (21.18).

In regard to explained variance and fit, Stacking had the best R^2 value of 0.968 across all models, which means it accounts for nearly 97% of the test data variance. This is

significantly better than ElasticNet's R^2 score of 0.77 and Extra Trees' R^2 score of 0.88. The Stacking model's normalized mean squared error (NMSE) is 0.032, the minimum, with no significant normalized error against data variance. Likewise, its variance accounted for (VAF) is 0.971, indicating impeccable consistency of predicted and actual values. The median absolute percentage error (MDAPE) of 23.08 and the standard deviation (STD_dev) of residuals of 105.76 also illustrate the consistency of the model in its performance.

On the other hand, the other models show higher error measures and lower variance explanation, with the Stacking model remaining the most accurate and consistent for the test set in this comparison. This claim was on the basis of the higher R^2 value of the Stacking model for both train and test data. According to this table, the higher R2 and VAF for each model make them worthy of a better model; in this case SE model. On the other hand, the lower the other metrics such as MSE, RMSE, MAE, NMSE, MDAPE, and STD, the more the model would have the merit of being an efficient predictor. Therefore, the Stacking model is deemed the most efficient and performs better than the other models for both training and testing data. In contrast, ElasticNet shows weaker performance in predicting the variables. Furthermore, the results of the employed evaluation metrics are presented and thoroughly discussed using relevant figures at the end of this section. Improved accuracy and stability on both forest types were indicated by the lower MSE, RMSE, and MDAPE, but higher R2 and VAF of the Stacking Ensemble, which consistently outperformed the other algorithms. ElasticNet performed poorly because of its linear framework, which failed to properly capture the intricate, nonlinear patterns in biomass data. Because Stacking possessed the ability to combine the powers of tree-based models like GB, ET, and XGB, it outperformed them despite the fact that they were moderate.

Figure 13 below is an illustration of the data values obtained via the ML parameters; i.e., ElasticNet, Extra Trees, GB, Poisson, Stacking, XGB; and accordingly, a comparison of these parameters in detail, along with their distance from the target value data, is presented.

Table 3: Error metrics criteria result for the proposed ML models considering the train and test datasets.

Models / Metrics	ElasticNet	Extra Trees	GB	Poisson	Stacking	XGB
Train						
MSE	4026.476	4.933E-26	5.800	1725.207	1153.269	1.72544E-05
RMSE	63.455	2.221E-13	2.408	41.536	33.960	0.004
MAE	32.550	7.82E-14	1.821	16.042	11.562	0.003
R^2	0.788	0.999	0.999	0.909	0.939	0.999
NMSE	0.212	2.601E-30	0.000	0.091	0.061	9.09812E-10
MDAPE	67.280	1.651E-13	5.016	29.535	8.722	0.008
STD_dev	100.279	137.713	137.491	150.171	134.774	137.713
VAF	0.788	0.999	0.999	0.909	0.939	0.999
Test						
MSE	2378.167	1216.538	1743.162	1051.301	334.820	2090.796
RMSE	48.766	34.879	41.751	32.424	18.298	45.725
MAE	36.818	18.068	21.948	19.320	12.422	21.178
R^2	0.770	0.882	0.831	0.898	0.968	0.797
NMSE	0.230	0.118	0.169	0.102	0.032	0.203
MDAPE	68.390	32.640	34.406	31.075	23.081	30.099
STD_dev	100.272	85.241	76.508	78.865	105.763	75.760
VAF	0.777	0.894	0.853	0.924	0.971	0.817

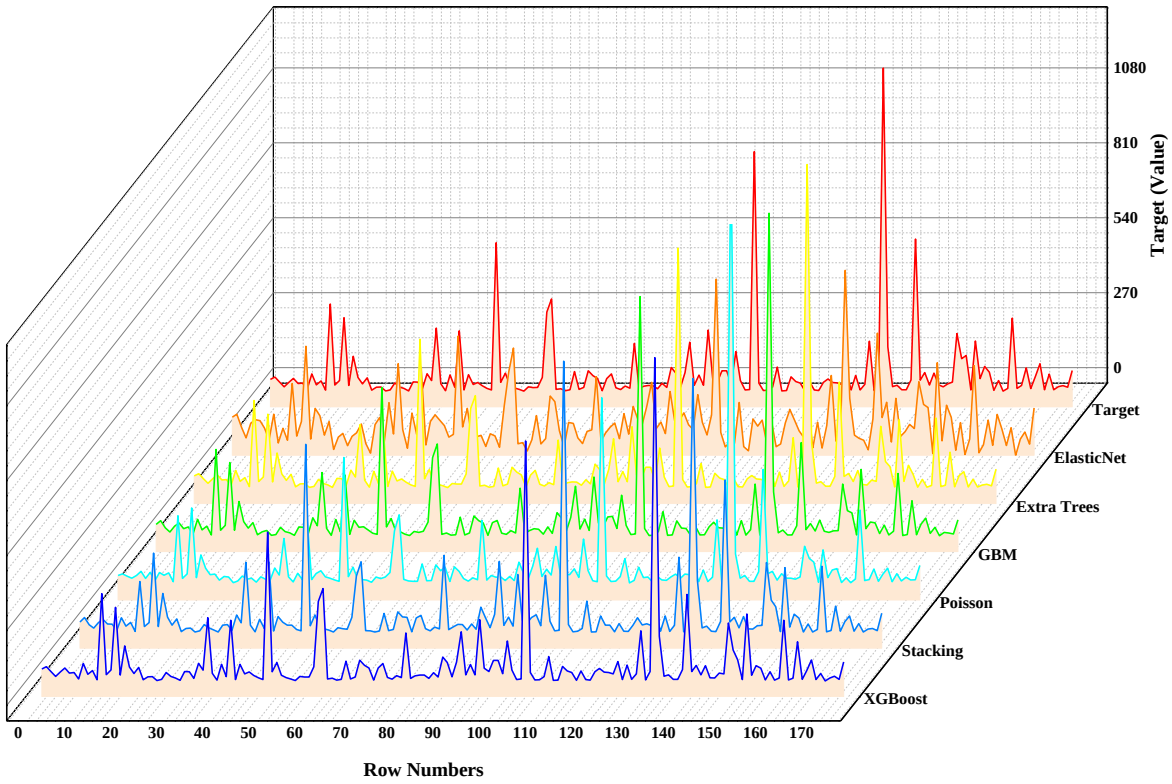


Figure 13: Value plot of comparing ML parameter values with the target value.

Figure 14 shows the results for R^2 as a significant error metric criterion, suggesting how well the employed ML models' predictions fit the real data. Shown in the figure, the model's prediction values align closely to the norm line (when $R^2 = 1$) is considered to be a superior as well as more accurate model. This result is in line with a higher R^2 value (approximately 0.939) for the test data and 0.968 for the training data in the proposed Stacking model.

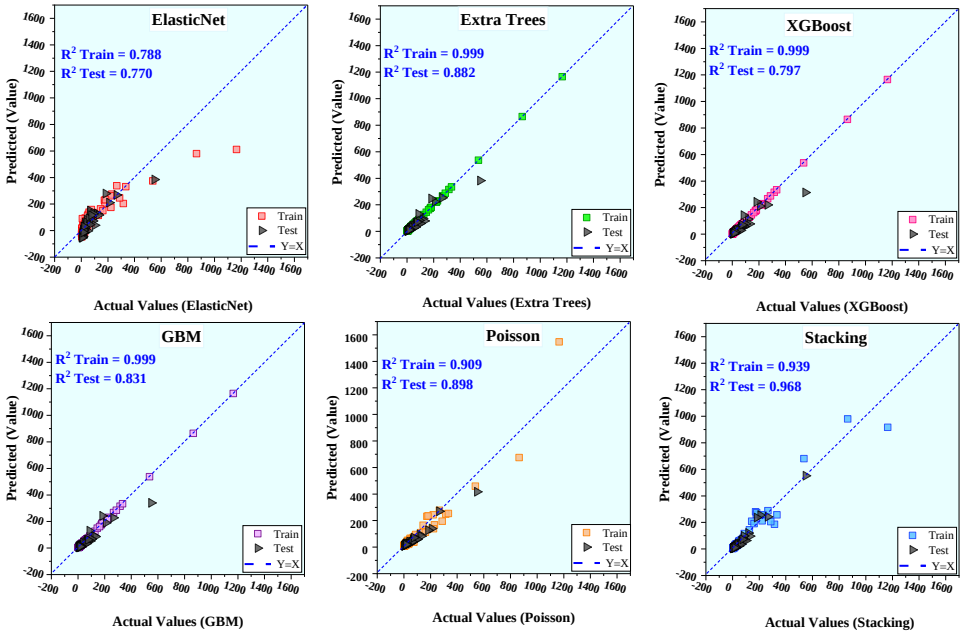


Figure 14: Comparing the coefficient of determination (R2) for each ML model.

The frequency of each error value for each ML method's predictions is represented in Fig. 15. The error analysis was conducted for both train and test parameters, and the ML models were assessed to examine their accuracy. As a result, the error in the ML models' prediction performance ought to be almost zero to be an adequate model for the aim of the study.

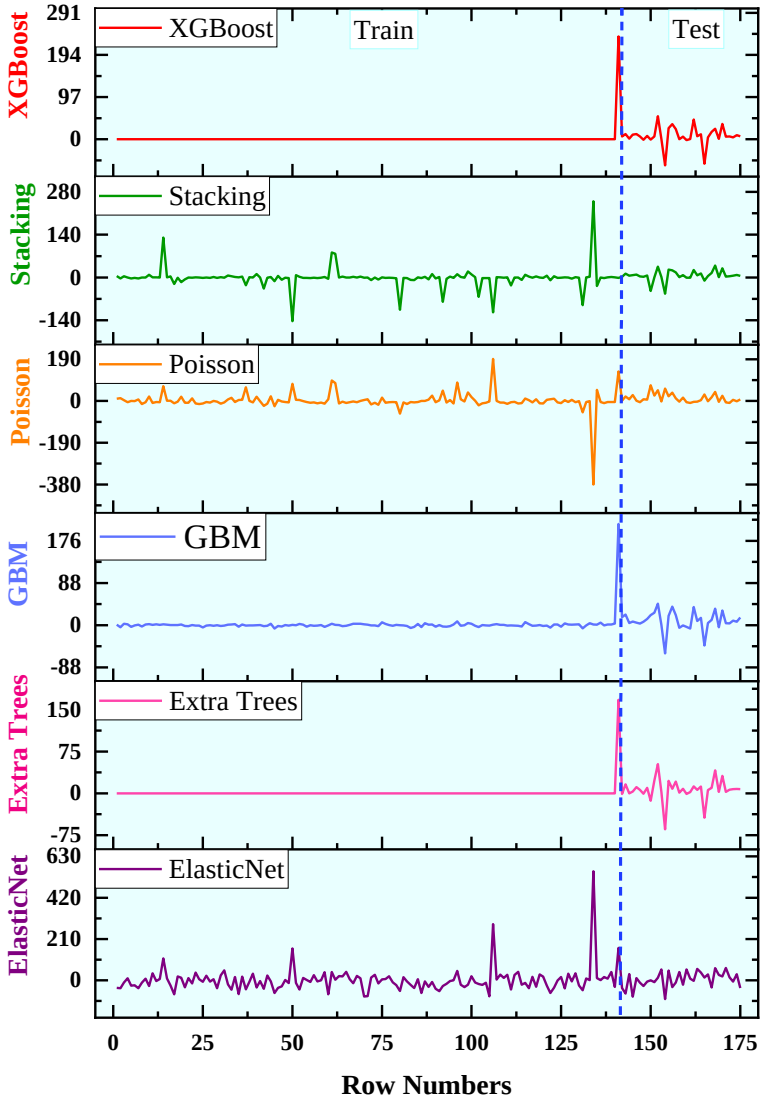


Figure 15: Comparing error values for ML models

Moreover, according to Fig. 16, the error values for the proposed ML models have been illustrated from the least error value to the most errors for both train and test data of each model, moving from left to right. A model with the least error values (i.e., approximately zero) would be the best predictor among the employed ML models. This visualization highlights that the data has intense recurrent peaks—suggesting non-uniform distributions with dominating clusters—and that these patterns persist but evolve subtly across different sections of the dataset. When one of the groups describes a stacking model, its activity can be visually compared with the other groups by looking at how close the mean is to zero and how much and steady the standard deviation is.

Based on the plot, we see that stacking seems to be more accurate than individual models, but by a very small margin. Compared to its predictions, it has fewer errors and reduced variance, indicating that it has a stronger generalization and stability. On the contrary, although other models have also performed adequately, they exhibit some spread or deviations that are a bit higher than the mean.

Overall, it appears that stacking should produce a more consistent and less erratic result than single models, thus making a superior comparison to the single models in terms of performance.

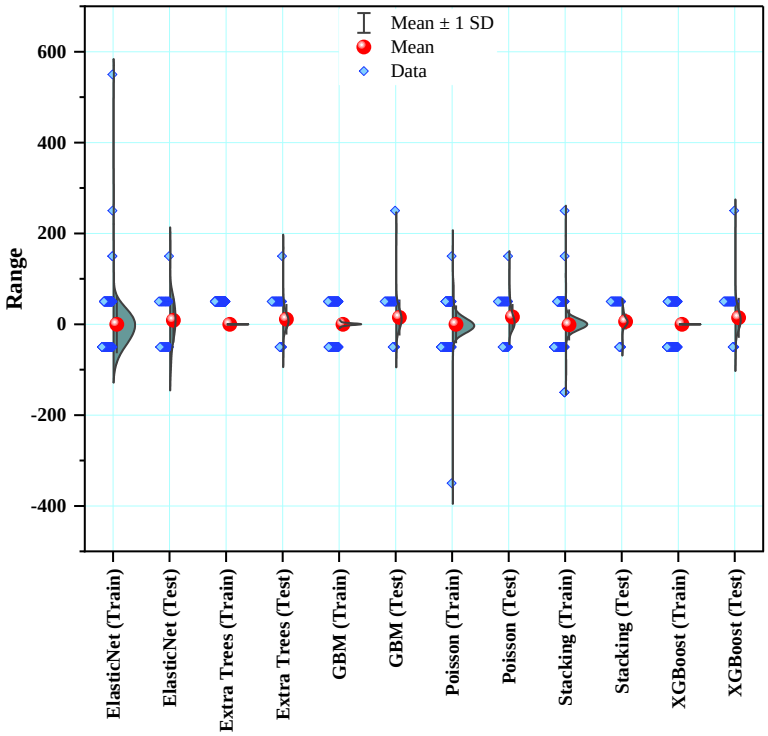


Figure 16: Boxplot of ML models' error values for both train and test data

The following figure (Fig. 17) illustrates a comparison of proposed models in terms of two important statistical evaluation metrics, namely R^2 and VAF, estimated for both the test and train datasets. As the values of these metrics show, all the models perform efficiently in the prediction except the ElasticNet model, which performs weaker than others, having lower VAF and R^2. Stacking and XGB model performances are stronger than the rest of the models, bearing higher VAF and R^2. Based on the plot, we see that stacking seems to be more accurate than individual models, but by a very small margin. Compared to its predictions, it has fewer errors and reduced variance, indicating that it has a stronger generalization and stability. On the contrary, although other models have also performed adequately, they exhibit some spread or deviations that are a bit higher than the mean.

Overall, it appears that stacking should produce a more consistent and less erratic result than single models, thus making a superior comparison to the single models in terms of performance.

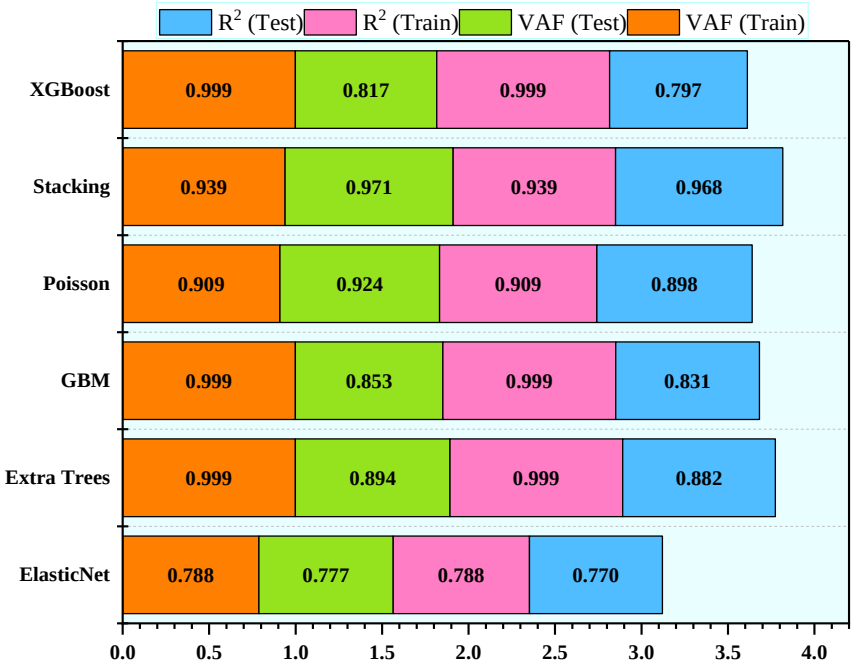


Figure 17: Comparison of Models based on VAF and R2 metrics.

The other evaluation metrics, including MSE, NMSE, MAE, RMSE, STD, and MDAPE, are applied for comparison among the models, supposing that the lowest value of these metrics for each model allows that model to be the best predictor. In this case, the stacking model for both the train and test datasets is the lowest compared to the other models (See Fig. 18).

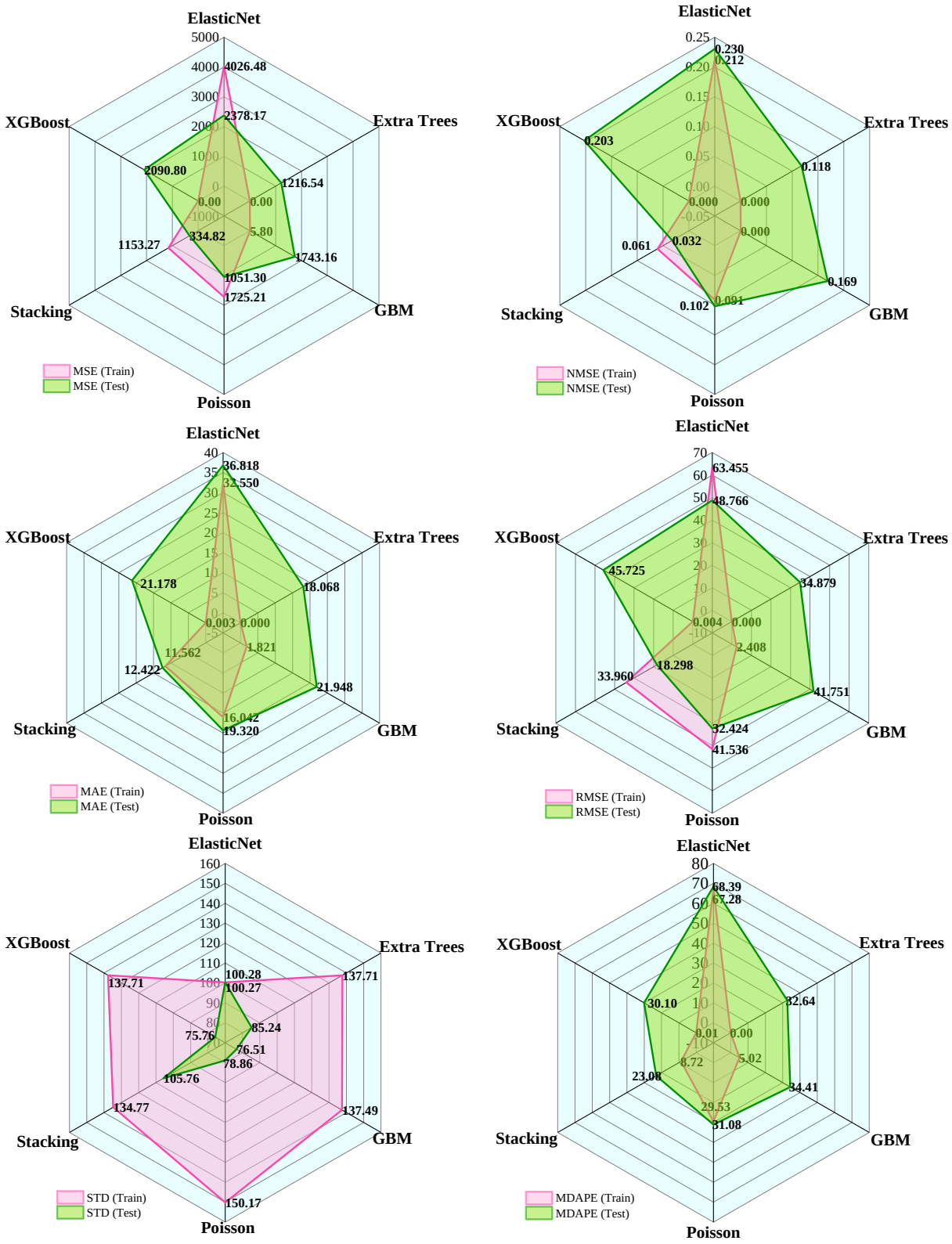
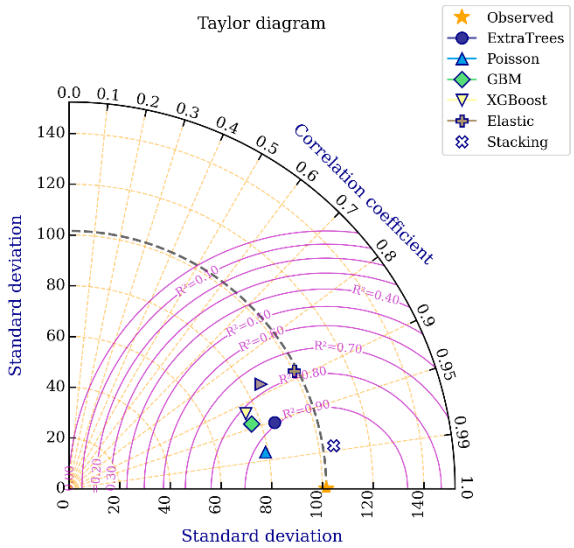


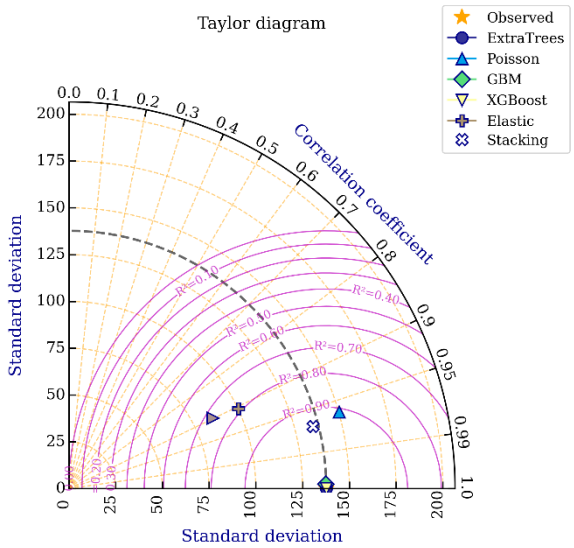
Figure 18: Comparison of Models based on MSE, NMSE, MAE, RMSE, STD, and MDAPE metrics.

Another graphical tool used to compare the models' performance is the Taylor diagram. This diagram evaluates models based on their accuracy, using metrics such as the correlation coefficient, standard deviation (STD), and RMSE. In the diagram, the models' performance is represented by circles, where better

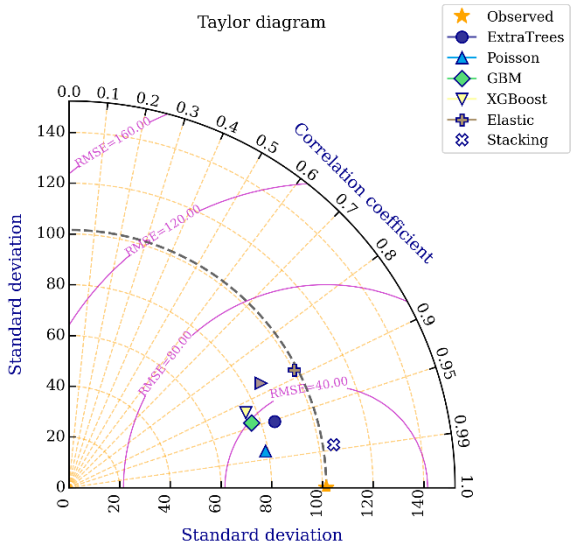
performance is indicated by points closer to the reference point [29]. Taylor diagrams for predicting tropical tree biomass are shown in Fig. 19. As seen in these diagrams, the RMSE of the Stacking model is lower than that of the other models, and its correlation coefficient exceeds 0.9, outperforming the other models in this regard. In contrast, the RMSE of the ElasticNet model is higher than that of the other machine learning (ML) models, and its correlation coefficient is lower. These findings, based on the correlation coefficient, STD, and RMSE, confirm that the Stacking model outperforms the other models.



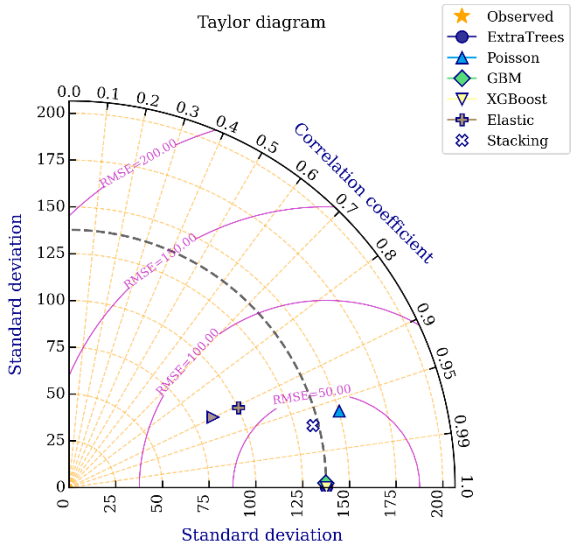
Taylor Diagram_ (R^2, test)



Taylor Diagram_ (R^2, train)



Taylor Diagram_ (RMSE, test)



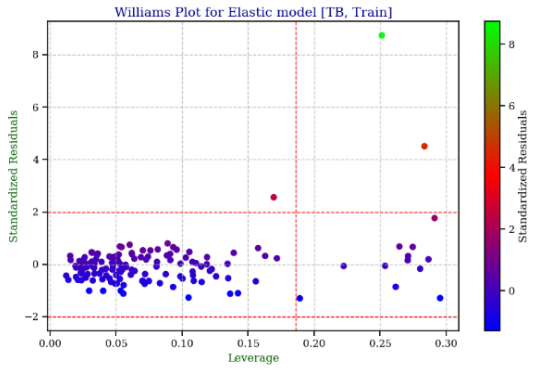
Taylor Diagram_ (RMSE, train)

Figure 19: Taylor diagrams for models' comparison based on RMSE, STD, and R metrics

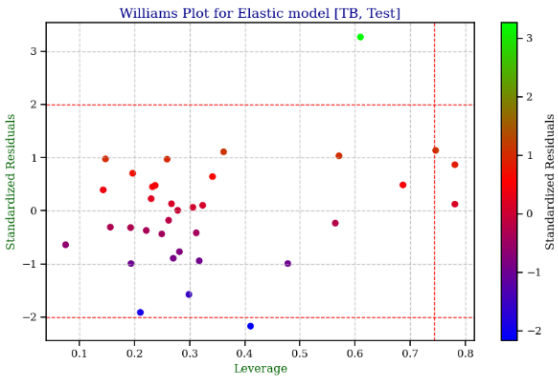
The last plot to be discussed for model comparison is the Williams plot. This plot is used to compare a specific group of compounds in terms of leverage values and standardized residuals [30]. William's plot shows the standardized residuals on the y-axis and leverages on the x-axis of the training and testing datasets. From this plot, the applicability domain is implemented within a squared area inside ±2 standard deviations and a threshold h* in leverage (h* = 3p´/n, being p´ model parameters and n compounds number). The majority of data ought to be located within this area, conceptualizing that they have to be inliers and influential in the model. The Williams plots of the Stacking model on both training and test sets are indicators of good model performance and generalization. Most of the observations in the test plot lie comfortably within the satisfactory limits for leverage and standardized residuals (±2), which is an indicator that the model predictions are not biased and stable and possess minimal outliers. Only a minimal number of observations being out of the ±2 boundary and leverage constraint signifies that there are very few influential or problematic points.

Similarly, the training plot shows tightly clumped residuals around zero with the majority of data points having little leverage, which would mean that the model has not over-fit the training data. Even some of the residuals fall outside ±3 or do possess relatively higher leverage, those are scattered and do not invalidate the model. The similar trend in both plots confirms that the

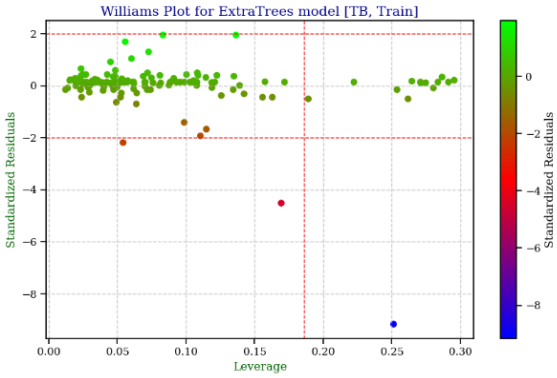
Stacking model works well on unseen data, learning the inherent pattern significantly without being overfit and robust. To be an efficient model predictor, the data must lie within this domain (See Fig. 20) [23].



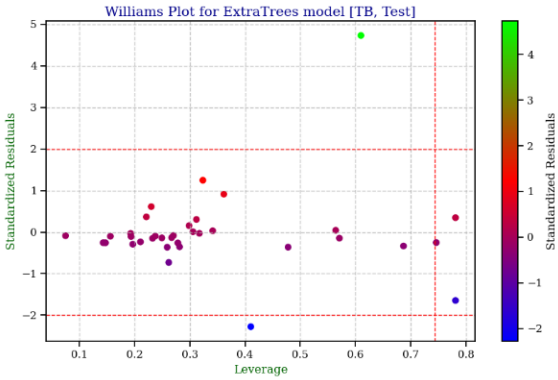
Elastic Net Williams Plot [TB, Train]



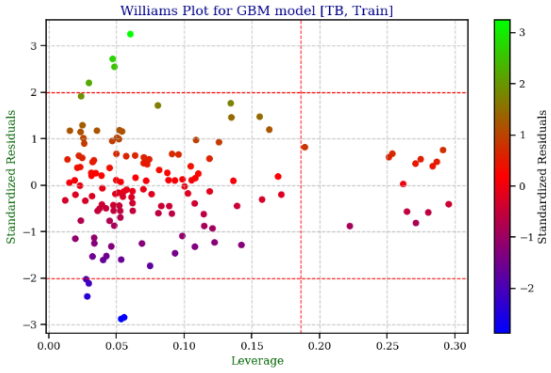
Elastic Net Williams Plot [TB, Test]



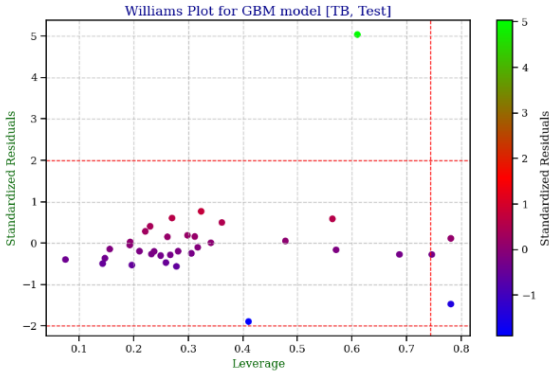
Extra Trees Williams Plot [TB, Train]



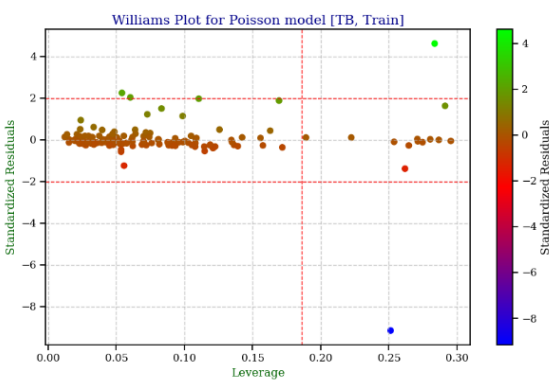
Extra Trees Williams Plot [TB, Test]



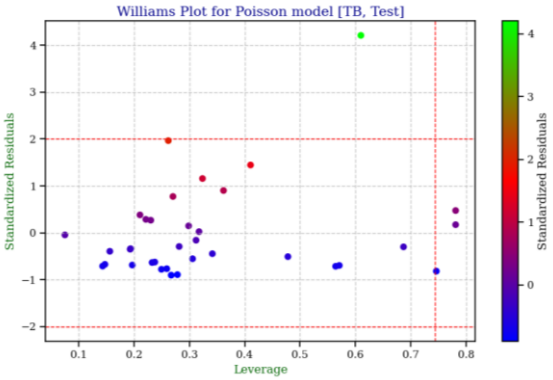
GBM Williams Plot [TB, Train]



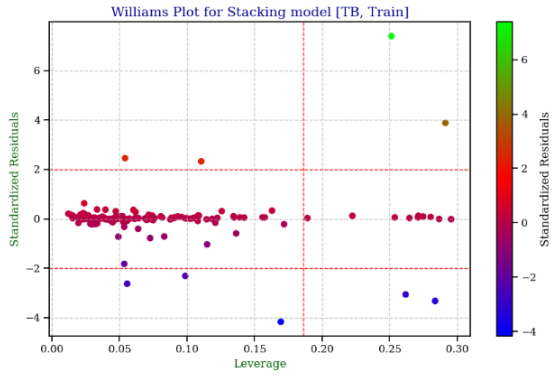
GBM Williams Plot [TB, Test]



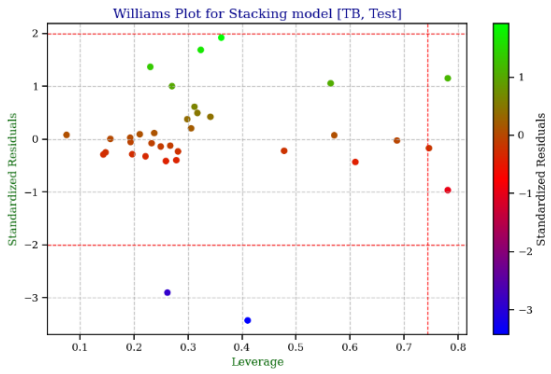
Poisson Williams Plot [TB, Train]



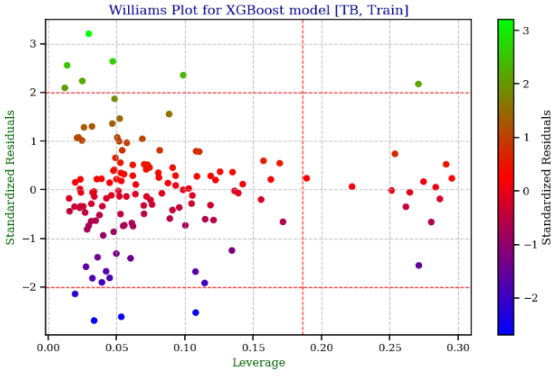
Poisson Williams Plot [TB, Test]



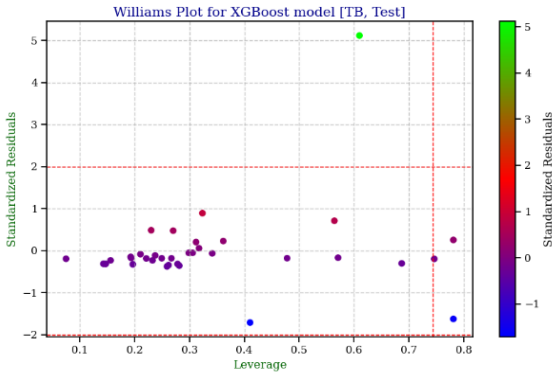
Stacking Williams Plot [TB, Train]



Stacking Williams Plot [TB, Test]



XGBoost Williams Plot [TB, Train]



XGBoost Williams Plot [TB, Test]

Figure 20: Williams plots for models' comparison based on standard residuals and leverage.

3.3. Comparison with foundation models and LLMs

Large Language Models, or LLMs, are sophisticated AI systems that have been trained on enormous text datasets to comprehend and produce human language. Google created BERT, which is perfect for tasks like classification and question answering because it can analyze words in both directions and understand context. With its emphasis on producing relevant and coherent text, OpenAI's GPT is effective for tasks like content creation, dialogue, and summarization. Both use Transformer architecture, but GPT is more focused on generation and BERT on comprehension. The proposed SE model adds domain-specific efficiency, whereas models such as BERT and GPT are effective for general-purpose NLP tasks. We specifically highlighted how, in contrast to the extensive, data-intensive training of LLMs, SE makes use of structured, domain-relevant features. This study also indicates SE's improved interpretability and reduced computational cost, both of which are important for ecological modelling.

4 conclusions

The study was implemented on eight tropical forests in Vietnam, using the forestry variables, i.e., AGB, BGB, and TB. In an attempt to solve the problem of predicting the mentioned variables, the study used an MGDL regression strategy, which later proved to be an efficient model to bear a strong ability to predict tropical forest biomass. To this, five models were selected as major algorithms to unravel the issue of biomass prediction. These models included Gradient Boosting (GB), Extra Trees (ET), XGB, ElasticNet, and Poisson, all of which were employed to synchronously anticipate both the amount of AGB, BGB, as well as TB = BGB + AGB. Then optimized by Grid Search. Additionally, the SE model was joined to the aforementioned models so as to allow the results to become satisfactory, i.e., mainly for the cross-validation purpose. Therefore, the recommended method's performance was investigated in terms of two sets of actual data, namely training and testing data.

The outcome of this study presented that the recommended method had a vigorous efficacy to estimate the amount of forest biomass. That is to say, employing a simultaneous group of ML models resulted in a significant impact on predicting forestry above- and below ground, as well as the sum of the biomass. The very high R^2 values of near 0.999 in the training set are definitely cause for alarm for overfitting or data leakage. We dealt with this by ensuring strict separation of training and test datasets such that there was no information leakage. We also employed Grid Search with cross-validation during hyperparameter tuning to allow maximum model complexity without over-fitting. The test set results, with R^2 scores considerably lower (e.g., 0.968 for the best model), are a sign of good generalization and suggest that while there may be some overfitting, it is controlled. More regularization and more

data will be tried in future research to reduce the possibility of overfitting even more. Based on the provided metrics, the Stacking ensemble model performed clearly superior to each of the standalone models on the test set. That is because it is capable of leveraging the prediction power of various base learners—ElasticNet, Extra Trees, Gradient Boosting, Poisson Regression, and XGB—and minimizing their respective errors through a meta-learner. Stacking takes into consideration the stand-alone strengths of linear as well as nonlinear models and results in improved generalization and less overfitting.

Quantitatively, the Stacking model achieved the highest coefficient of determination (R^2 = 0.968) and variance accounted for (VAF = 0.971) on the test set, indicating that its predictions were most highly correlated with actual biomass values. It generated the lowest mean squared error (MSE = 334.820), root mean square error (RMSE = 18.298), and mean absolute error (MAE = 12.422), indicating high accuracy and low prediction bias. In terms of normalized error, it also had an NMSE of just 0.032, and the mean directional accuracy percentage error (MDAPE) decreased to 23.081%, significantly better than other models. Although its test standard deviation (STD = 105.763) was slightly greater, this is the natural result with better prediction accuracy and range coverage for both train and test data, where the results showed R2 equals 0.939 for the testing data and 0.968 for the training data in this study data analysis. Therefore, adding the SE model to the proposed models is recommended for predicting forest biomass effects. As a result, this is evidence of the poor performance of this model. The William plots residuals display that its majority corresponds to the tolerant parameters, and very limited outliers or high-leverage points are there in the test and train subsets. This implies that the Stacking model produces reliable, unbiased, and non-overfit predictions, and there is indeed powerful generalization and performance.

Declarations

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Authors' contributions QD

Writing-Original draft preparation, Conceptualization, Supervision, Project administration.

Acknowledgements

I would like to take this opportunity to acknowledge that there are no individuals or organizations that require acknowledgment for their contributions to this work.

Ethical approval

The research paper has received ethical approval from the institutional review board, ensuring the protection of participants' rights and compliance with the relevant ethical guidelines.

References

[1] H. C. Zhantao Song, Xiong Zhang, Xiaoqiang Li, Junjie Zhang, Jingai Shao, Shihong Zhang, Haiping Yang, "Machine learning assisted prediction of specific surface area and nitrogen content of biochar based on biomass type and pyrolysis conditions," *J Anal Appl Pyrolysis*, vol. 183, 2024.

[2] L. Jia, W. Shao, J. Wang, Y. Qian, Y. Chen, and Q. Yang, "Machine learning-aided prediction of bio-BTX and olefins production from zeolite-catalyzed biomass pyrolysis," *Energy*, vol. 306, p. 132478, 2024.

[3] H. Wu, S. An, B. Meng, X. Chen, F. Li, and S. Ren, "Retrieval of grassland aboveground biomass across three ecoregions in China during the past two decades using satellite remote sensing technology and machine learning algorithms," *International Journal of Applied Earth Observation and Geoinformation*, vol. 130, no. November 2023, p. 103925, 2024, doi: 10.1016/j.jag.2024.103925.

[4] P. Mao *et al.*, "An improved approach to estimate above-ground volume and biomass of desert shrub communities based on UAV RGB images," *Ecol Indic*, vol. 125, p. 107494, 2021, doi: 10.1016/j.ecolind.2021.107494.

[5] P. B. May *et al.*, "Mapping aboveground biomass in Indonesian lowland forests using GEDI and hierarchical models," *Remote Sens Environ*, vol. 313, p. 114384, 2024.

[6] B. Huy, N. Quy Truong, K. P. Poudel, H. Temesgen, and N. Quy Khiem, "Multi-output deep learning models for enhanced reliability of simultaneous tree above- and below-ground biomass predictions in tropical forests of Vietnam," *Comput Electron Agric*, vol. 222, no. December 2023, p. 109080, 2024, doi: 10.1016/j.compag.2024.109080.

[7] M. F. Oliveira *et al.*, "Predicting below and above-ground peanut biomass and maturity using multi-target regression," *Comput Electron Agric*, vol. 218, p. 108647, 2024.

[8] G. Kunapuli, *Ensemble methods for machine learning*. Simon and Schuster, 2023.

[9] R. Dey and R. Mathur, "Ensemble learning method using stacking with base learner, a comparison," in *International Conference on Data Analytics and Insights*, Springer, 2023, pp. 159–169.

[10] P. Naik, M. Dalponte, and L. Bruzzone, "Automated machine learning driven stacked ensemble modeling for forest aboveground biomass prediction using multitemporal Sentinel-2 data," *IEEE J Sel Top Appl Earth Obs Remote Sens*, vol. 16, pp. 3442–3454, 2022.

[11] Y. Zhang, J. Ma, S. Liang, X. Li, and J. Liu, "A stacking ensemble algorithm for improving the biases of forest aboveground biomass estimations from multiple remotely sensed datasets," *GIsci Remote Sens*, vol. 59, no. 1, pp. 234–249, 2022.

[12] J. Liu, Y. Niu, Z. Jia, and R. Wang, "Assessing the ethical implications of artificial intelligence integration in media production and its impact on the creative industry," *MEDAAD*, vol. 2023, pp. 32–38, 2023.

[13] R. Huang, C. McMahan, B. Herrin, A. McLain, B. Cai, and S. Self, "Gradient boosting: A computationally efficient alternative to Markov chain Monte Carlo sampling for fitting large Bayesian spatio-temporal binomial regression models," *Infect Dis Model*, vol. 10, no. 1, pp. 189–200, 2025, doi: 10.1016/j.idm.2024.09.008.

[14] Z. Wang, L. Mu, H. Miao, Y. Shang, H. Yin, and M. Dong, "An innovative application of machine learning in prediction of the syngas properties of biomass chemical looping gasification based on extra trees regression algorithm," *Energy*, vol. 275, p. 127438, 2023.

[15] H. Wei, K. Luo, J. Xing, and J. Fan, "Predicting co-pyrolysis of coal and biomass using machine learning approaches," *Fuel*, vol. 310, p. 122248, 2022.

[16] R. (Bob) Roy, "No Title." Accessed: Jun. 27, 2021. [Online]. Available: https://bobrupakroy.medium.com/extra-trees-classifier-regressor-5b5f6abe8228

[17] B. Kıyak, H. F. Öztop, F. Ertam, and İ. G. Aksoy, "An intelligent approach to investigate the effects of container orientation for PCM melting based on an XGBoost regression model," *Eng Anal Bound Elem*, vol. 161, pp. 202–213, 2024.

[18] Y. Ayub, J. Ren, T. Shi, W. Shen, and C. He, "Poultry litter valorization: Development and optimization of an electro-chemical and thermal tri-generation process using an extreme gradient boosting algorithm," *Energy*, vol. 263, p. 125839, 2023.

[19] A. Jain, "No Title." Accessed: Feb. 05, 2024. [Online]. Available: https://medium.com/@abhishekjainindore24/elastic-net-regression-combined-features-of-l1-and-l2-regularization-6181a660c3a5

[20] J. Liu *et al.*, "A new application of Elasticnet regression based near-infrared spectroscopy model: Prediction and analysis of 2, 3, 5, 4′-tetrahydroxy stilbene-2-O-β-D-glucoside and moisture in Polygonum multiflorum," *Microchemical Journal*, vol. 199, p. 110095, 2024.

[21] Purhadi, D. N. Sari, Q. Aini, and Irhamah, "Geographically weighted bivariate zero inflated generalized Poisson regression model and its application," *Heliyon*, vol. 7, no. 7, p. e07491, 2021, doi: 10.1016/j.heliyon.2021.e07491.

[22] U. Arif, C. Zhang, S. Hussain, and A. R. Abbasi, "An Efficient Interpretable Stacking Ensemble Model for Lung Cancer Prognosis," *Comput Biol Chem*, p. 108248, 2024.

[23] H. Yıldırım and M. R. Özkale, "A Novel Regularized Extreme Learning Machine Based on L 1-Norm and L 2-Norm: a Sparsity Solution Alternative to Lasso and Elastic Net," *Cognit Comput*, vol. 16, no. 2, pp. 641–653, 2024.

[24] S. M. Mastelini, F. K. Nakano, C. Vens, and A. C. P. de Leon Ferreira, "Online extra trees regressor," *IEEE Trans Neural Netw Learn Syst*, vol. 34, no. 10, pp. 6755–6767, 2022.

[25] M. M. Hameed, M. K. Alomar, F. Khaleel, and N. Al-Ansari, "An Extra Tree Regression Model for Discharge Coefficient Prediction: Novel, Practical Applications in the Hydraulic Sector and Future Research Directions," *Math Probl Eng*, vol. 2021, 2021, doi: 10.1155/2021/7001710.

[26] "Understanding Poisson Regression." Accessed: Nov. 05, 2023. [Online]. Available: https://medium.com/@data-overload/understanding-poisson-regression-a-powerful-tool-for-count-data-analysis-b7184c61bfde

[27] M. A. Alemayehu, S. D. Kebede, A. D. Walle, D. N. Mamo, E. B. Enyew, and J. B. Adem, "A stacked ensemble machine learning model for the prediction of pentavalent 3 vaccination dropout in East Africa," *Front Big Data*, vol. 8, p. 1522578, 2025.

[28] "XGBoost Algorithm." Accessed: Sep. 04, 2024. [Online]. Available: https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/#:~:text=XGBoost builds a predictive model,made by the existing ones.

[29] M. Ehteram, A. N. Ahmed, P. Kumar, M. Sherif, and A. El-Shafie, "Predicting freshwater production and energy consumption in a seawater greenhouse based on ensemble frameworks using optimized multi-layer perceptron," *Energy Reports*, vol. 7, pp. 6308–6326, 2021, doi: 10.1016/j.egyr.2021.09.079.

[30] A. Beheshti, E. Pourbasheer, M. Nekoei, and S. Vahdani, "QSAR modeling of antimalarial activity of urea derivatives using genetic algorithm-multiple linear regressions," *Journal of Saudi Chemical Society*, vol. 20, no. 3, pp. 282–290, 2016, doi: 10.1016/j.jscs.2012.07.019.