BAB 1

INTRODUCTION

The quantity of carbon stored in an ecosystem's above-ground and below-ground biomass is known as its carbon stocks [1], [2]. Carbon stocks are significant in addressing climate change due to the ability of plants to absorb or retain carbon from the atmosphere and store it in the form of biomass, thereby reducing greenhouse gas concentrations [3], [4], [5]. Forests, as one of the largest carbon sinks, play a crucial role in regulating the global climate [6], [7]. Therefore, studies related to carbon stock estimation are crucial in supporting environmental conservation and climate change mitigation [8].

However, a significant challenge in carbon stock studies is the management of complex data and the need for efficient analysis methods. The use of technologies such as machine learning (ML) has become increasingly popular due to its ability to analyze large-scale data and discover patterns that are difficult to recognize with traditional methods [9]. In the context of carbon stocks, many studies have explored the potential of ML to analyze imagery and other datasets to improve the quality of analysis results. For example, in [10] a study was conducted using Landsat 8 OLI data using several regression algorithms such as Support Vector Machine (SVM), Random Forest (RF), k-Nearest Neighbors (kNN), and XGBoost and using the Boruta Method feature selection method. This research shows that XGBoost gives the best performance R^2 = 0,89. Another study [11] used the XGBoost model with the Gradient Boosting selection feature to predict Soil Organic Carbon Stock (SOCS) on sentinel-1 and sentinel-2 datasets and field data and showed the results of the above model were R^2 = 0,59. In addition, research [12] applies various ML models to predict soil organic carbon content (SOC), one of which is XGBoost with a feature selection Genetic Algorithm. This study uses several types of datasets, such as soil data and Auxiliary variables, which include 105 predictor variables derived from various sources, including 60 variables generated from Landsat 8 and MODIS satellite images. The results of the XGBoost model in this study Mean Absolute Error (MAE) = 0.66%, Root Mean Square Error (RMSE) = 0.82%, R^2 = 0,57.

Many studies have explored the use of XGBoost in environmental data analysis. However, few have integrated the advanced feature extraction capabilities of VGG16, especially in the context of complex data such as carbon stock estimation. Based on the above research, this study uses XGBoost and VGG16 due to their advantages in complex data analysis tasks. VGG16 is used for its ability to extract visual features automatically and efficiently, especially on images with complex structures such as satellite and drone images [13], [14]. The VGG16 model pre-trained with the ImageNet dataset is used to process and extract important features from the image dataset. The obtained features were then utilized as input for the XGBoost model, which was selected on the basis of its superior ability to manage high-dimensional regression data. XGBoost's ability to utilize these high-dimensional features is critical as it allows the model to make more accurate predictions about carbon stocks by utilizing the extracted features. It has built-in features for feature selection, which helps to reduce noise, prevent overfitting, and improve model accuracy [12], [15], [16], [17], [18]. In addition, previous research shows that XGBoost consistently provides the best results compared to other algorithms, such as Random Forest and SVM [10], especially in tasks involving environmental data. As far as the researchers know, no studies have explored feature selection on XGBoost models with features extracted using the VGG16 architecture, particularly in regression models with imagery datasets related to carbon stock. This study not only uses imagery datasets but also integrates field data that measures the total carbon content at locations corresponding to the imagery datasets. The field data is used to verify and accurately label the imagery dataset, improving the accuracy of the carbon stock estimation model. By combining direct field measurements with imagery datasets, the developed model is able to provide more accurate and reliable predictions, reflecting actual conditions on the ground. This study evaluates the impact of feature selection techniques on the performance of carbon stock estimation models by implementing four different scenarios: a baseline model without feature selection, a model that uses Information Gain, a model that uses Feature Importance, and a model that will use Recursive Feature Elimination (RFE). The selection of these feature selection techniques is based on their potential to improve model accuracy and efficiency. Information Gain is used as a feature selection technique to reduce data dimensionality by prioritizing features that have a high level of importance and help sort the most informative features [19], [20]. Feature Importance, generated through the trained XGBoost model, allocates a score to each feature based on its contribution to model accuracy [21]. This technique allows the selection and focus on the

most significant features, and later, Information Gain and Feature Importance will use Top N Features starting from 500 to 5000. Recursive Feature Elimination (RFE) will be implemented to iteratively reduce the number of features, eliminating variables that contribute the least to the predictive power of the model and aiming to build a more compact and efficient model without compromising its performance. The contribution of the research is to propose the use of XGBoost and VGG16 algorithms for feature extraction complemented by the application of feature selection techniques to enhance model accuracy. This research also makes an important contribution in the form of a comprehensive comparison between various models developed using different feature selection techniques. This analysis aims to identify the most effective models for accurately estimating carbon stocks. By conducting an in-depth evaluation, we were able to determine the optimal feature selection, which significantly improved the accuracy of carbon stock prediction. This method utilizes field data and image data to improve accuracy in carbon stock estimation.