Advanced Spatio-Temporal Modeling of Seagrass Meadows Through Machine Learning Techniques

Hamdi Braiek[1,2(✉)], Nadim Nagati[1], Mayssa Trabelsi[1], Amir Ben Ayed[1], Nidhal Mezni[1], and Mohamed Amine Askri[1]

[1] ESPRIT School of Engineering, 18 rue de l'Usine Charguia II 2035, Ariana, Tunisia
hamdi.houichet@gmail.com,
{nadim.nagati,mayssa.trabelsi,amir.benayed,nidhal.mezni,
mohamed.askri}@esprit.tn

[2] Laboratory for Mathematical and Numerical Modeling in Engineering Science, National Engineering School at Tunis, University of Tunis El Manar, B.P. 37, 1002 Tunis-Belvédère, Tunisia

Abstract. Seagrass meadows are critical coastal ecosystems that provide invaluable services: carbon sequestration, habitat for marine biodiversity, and shoreline protection, but they are declining at alarming rates. Recent studies estimate arround 29% global loss of seagrass cover since the 1700s and project large carbon emissions if this loss continues. Conventional monitoring methods struggle to capture these complex dynamics, motivating data-driven approaches. In this work, we develop an integrated machine learning framework for spatio-temporal modeling of seagrass health and biomass. We apply iterative imputation to handle missing environmental data, and train both an XGBoost regressor and a stacked ensemble of ML models (Ridge, Random Forest, LGBM) on geospatial features. Model hyperparameters are tuned via the Tree-structured Parzen Estimator (TPE), a Bayesian optimization technique that efficiently explores the hyperparameters. Our results show that the optimized XGBoost model outperforms the ensemble across key seagrass metrics, achieving superior accuracy in predicting biomass and productivity. We identify the most influential environmental drivers such as light availability and water temperature through feature importance analysis, providing actionable ecological insights. This enhanced predictive framework, supported by a clear methodology flow, offers a robust tool for proactive seagrass conservation and decision-making.

Keywords: Seagrass ecosystems · XGBoost · Random forest · Iterative imputation · Marine biodiversity · Ecological monitoring

Hamdi Braiek—The author has legally changed their name from Hamdi Houichet to Hamdi Braiek.

F. Kamoun et al. (Eds.): AFRICATEK 2025, LNICST 677, pp. 216–234, 2026.
https://doi.org/10.1007/978-3-032-16638-8_15

1 Introduction

Seagrass meadows are often called the "lungs of the sea" for their outsized role in marine carbon sequestration and biodiversity support [1]. These submerged flowering plants occupy over 160,000 km^2 of coastal zones and store large amounts of organic carbon in sediments [2]. They provide nursery habitat for fisheries, stabilize shorelines, and sustain marine food webs [1]. However, mounting evidence indicates accelerated seagrass loss worldwide. Since the mid-1700s, global seagrass coverage has decreased by approximately 29%, equivalent to around 51,000 km^2 [2]. In the Mediterranean basin alone, the endemic *Posidonia oceanica* has lost an estimated 13–50% of its extent in recent decades [2].

These declines are driven by a range of anthropogenic stressors, including coastal development, pollution, eutrophication, and climate change impacts such as sea-level rise and marine heatwaves [1]. The loss of seagrass ecosystems has cascading effects on marine biodiversity, fisheries, and the global carbon cycle. Current estimates suggest that ongoing seagrass degradation may be releasing hundreds of teragrams of carbon per year back into the atmosphere [2]. These findings emphasize the urgent need for more effective monitoring systems and predictive tools to support seagrass conservation efforts.

Traditional monitoring approaches, including field surveys and remote sensing, offer valuable insights but often fall short in capturing the full spatial and temporal dynamics of seagrass ecosystems. Many datasets remain sparse, heterogeneous, and incomplete [3]. Although recent advances in satellite imagery, such as Sentinel-2, and computer vision techniques have improved large-scale seagrass mapping, significant gaps and biases still limit their effectiveness [2].

Machine learning (ML) offers a powerful alternative by identifying patterns from complex and diverse data sources, including light availability, water temperature, nutrient concentrations, and spatial context. Recent studies have demonstrated high accuracy in seagrass classification and biomass prediction using ML algorithms. Moreover, hybrid models that combine ensemble learning with metaheuristic optimization have shown promise in forecasting habitat suitability. Despite these advances, key challenges remain in handling noisy and missing data and in translating model predictions into actionable ecological insights.

In this study, we propose a systematic machine learning framework for modeling seagrass ecosystems. Our approach integrates several components: data preparation with robust imputation techniques for missing values; model training using both XGBoost regressors and stacked ensemble methods; hyperparameter tuning through Tree-structured Parzen Estimation (TPE); and comprehensive evaluation using predictive metrics such as coefficient of determination (R^2) and root mean square error (RMSE). Each stage of the pipeline is designed to enhance model accuracy, interpretability, and ecological relevance.

Through this integrated methodology, we aim to advance the use of machine learning in coastal ecosystem monitoring and provide a scalable, data-driven approach for supporting the conservation of these vital underwater landscapes.

2 Related Works

Recent years have seen a surge in machine learning and remote sensing approaches for seagrass mapping and ecological modeling. Sentinel-2 and Landsat imagery combined with ML classifiers have enabled broad-scale seagrass density estimation, such as using Random Forest on satellite spectral data, with promising accuracies [4,5]. In one study, Sentinel-2 data processed in Google Earth Engine and classified using Random Forest achieved approximately 87% accuracy for mapping Caribbean seagrass extent [5]. Other work has leveraged UAV-acquired images with deep convolutional neural networks, particularly U-Net architectures, to classify multiple seagrass species with high precision [6]. Hybrid modeling approaches are also emerging. A recent work, in [7], reviews numerous ML techniques applied to seagrass monitoring and highlights that ensemble methods often improve generalization and robustness.

Despite these advances, a comparative examination reveals several trade-offs. Deep CNNs can achieve fine-scale classification accuracy but require large labeled datasets and high computational resources [6]. Simpler models, such as linear regression, are easier to interpret but often underperform due to the non-linear nature of ecological data. Missing or sparse data continue to pose a major bottleneck. Many studies acknowledge the necessity of imputation or bias correction when working with incomplete global seagrass databases. The research gap this study addresses is the integration of robust data imputation with advanced ensemble learning models under a unified spatio-temporal framework.

3 Methodology

The dataset used in this study is publicly available from [8]. It contains 6,658 rows and 17 columns, capturing various attributes related to seagrass ecosystems. The dataset includes both numerical and categorical variables, covering aspects such as geospatial location, biological properties, productivity, and structural characteristics. However, some features have significant gaps, with certain columns missing up to 90% of their data. To address these missing values, imputation strategies will be employed, including methods such as filling based on averages, interpolation, or other statistical techniques.

Our architecture is designed to transform raw ecological data from the seagrass database file into actionable insights for seagrass conservation. Figure 1 provides an overview of our paper's architecture.

Our framework, illustrated in Fig. 2 proceeds through three main phases: Data Preparation, Modeling & Optimization, and Evaluation. The process is structured into several key phases, each contributing to the overall goal of understanding and predicting seagrass health, biomass, and productivity.

We start with data Summarization and Analysis to examine spatial and temporal trends, categorize the data by bioregion and genus, and investigate patterns in missing data. This phase allows us to identify gaps and inconsistencies in the dataset, providing a clearer understanding of the distribution and health of seagrass ecosystems across various regions and time periods.

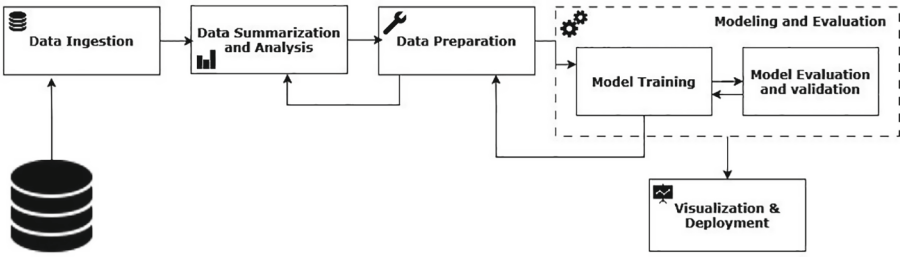

Fig. 1. An overview of our work's architecture

Once the data is summarized and analyzed, we proceed to the Preparation phase. Where we address the missing values, encode categorical variables, and standardize numerical features. In this step, duplicate rows are removed, and any inconsistent or incomplete data is cleaned and refined. To ensure quality, we standardize numeric features and apply one-hot encoding to categorical variables. Temporal and spatial coverage is explicitly handled (Sect. 4). Crucially, missing values are addressed using an iterative imputation method. The imputation iterates multiple passes over the data, refining the predicted values each round. We split the data into 70% training and 30% testing before imputation to prevent information leak, then impute missing values separately on each split. As shown in Sect. 6.1, this yields more realistic estimates than simple mean/median fill and leads to significant gains in predictive R^2.

For regression, we evaluate a stacked ensemble and a single model approach. The stacked model comprises three base learners: Ridge Regression, Random Forest Regressor, and LGMBRegressor. This stacking models captures diverse patterns while mitigating individual model biases. The XGBoost model is used as a strong baseline due to its robustness on tabular environmental data.

We tune model hyperparameters using the Tree-structured Parzen Estimator (TPE). In practice, TPE builds separate density estimates for the best-performing (good) and worst-performing (bad) hyperparameters and then samples from regions likely to improve performance. This sequential approach efficiently hones in optimal settings without exhaustive grid search. The objective function is the coefficient of determination R^2 on a cross-validation split, and we run the optimization for each target metric. The tuned hyperparameters are reported in Table 2.

Model performance is assessed on the held test set using relevant metrics. We compare the stacked ensemble versus XGBoost: in our experiments XGBoost consistently achieves higher R^2 on total biomass and density (see Sect. 6.2). We also examine the effect of iterative imputation by comparing scores across different numbers of imputation iterations (see Sect. 6.1). Feature importances are extracted from the models to interpret ecological drivers: for example, light attenuation and water temperature emerged as top predictors of biomass, aligning with established ecological knowledge. All methodological steps above are

directly linked to the results: e.g., Table 1 reports how imputation reduced error compared to naive filling, and Table 2 shows the final tuned hyperparameters and resulting improvements.

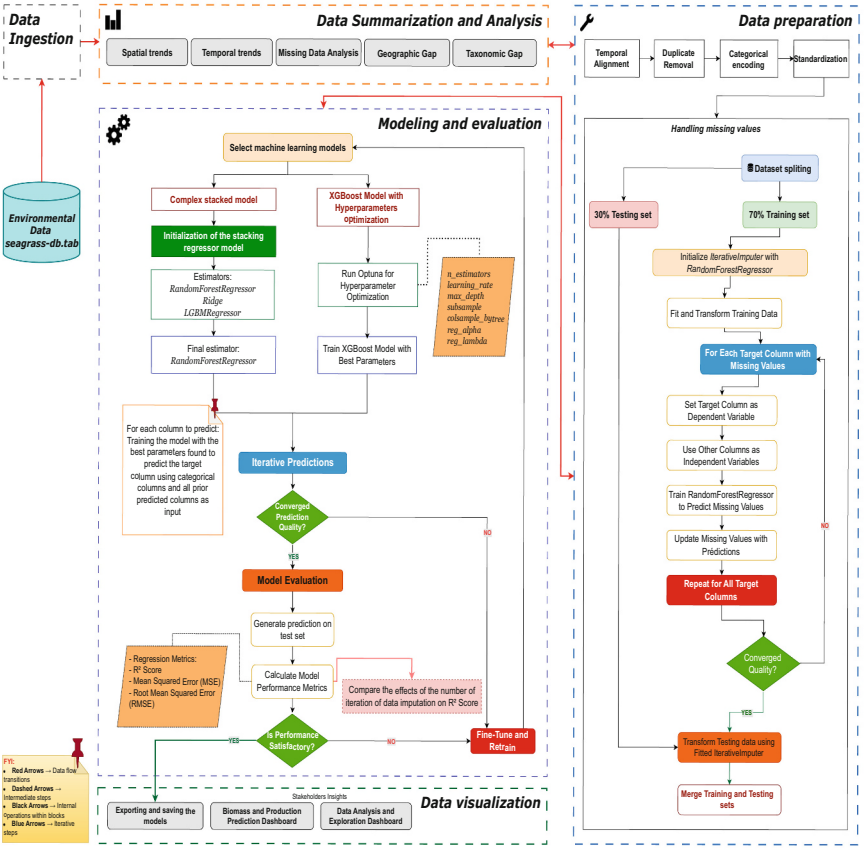

Fig. 2. A detailed breakdown of our data architecture

4 Data Summarization and Analysis

The aim of this section is to understand data limitations and ensure that subsequent modeling efforts are based on consistent and high-quality inputs. The main features of the dataset are summarized in the Table 1.

Table 1. Summary of key attributes in the dataset

Attribute	Type	Description
Latitude	Numerical	Location latitude
Longitude	Numerical	Location longitude
Bioregion	Categorical	Biogeographic classification
Habitat	Categorical	Seagrass environment type
Genus	Categorical	Seagrass genus classification
Year obs [a AD]	Numerical	Observation year
Seagr biom above [g/m^2]	Numerical	Above-ground dry biomass
Seagr biom below [g/m^2]	Numerical	Below-ground dry biomass
Seagr biom tot [g/m^2]	Numerical	Total dry biomass
Seagr shoot dens [#/m^2]	Numerical	Shoot density per m^2
Seagr cov [%]	Numerical	Percentage of area covered
Seagr leaf dens [#/m^2]	Numerical	Leaf density per m^2
Seagr prod above [g/m^2/day]	Numerical	Above-ground daily production
Seagr prod below [g/m^2/day]	Numerical	Below-ground daily production
Seagr prod total [g/m^2/day]	Numerical	Total daily biomass production
Seagr shoot prod [g/m^2/day]	Numerical	Daily shoot biomass production
Seagr leaf prod [g/m^2/day]	Numerical	Daily leaf biomass production

4.1 Data Summarization

The dataset shows significant heterogeneity, with some columns missing between 70% and 90% of their data like *Seagr leaf dens*, *Seagr prod below dm* and *Seagr shoot prod dm*. In Fig. 3, we show the distribution of missing values across all columns in the dataset. Darker shades indicate a higher percentage of missing data. This heatmap illustrates that the majority of missing values are concentrated in the biological and environmental sections of our dataset, with the highest concentration in the production section.

We also show, in Fig. 4, a bar plot the average proportion of missing data across different bioregions. We observe that the *Temperate Southern* region has the highest proportion of missing data, while the *Temperate North Atlantic* and *Temperate North Pacific* regions show relatively lower but still significant proportions of missing data.

Figure 5 presents the trend of missing data over time, from the early 1970s to around 2020. We see some clear periods of improvement, particularly in the late 1970s and late 1990s, where missing data dropped significantly. However, there are also spikes indicating times when data completeness worsened.

Furthermore, the dataset includes geographical information through the *Latitude* and *Longitude* columns, which pinpoint the locations of seagrass meadows. The *Bioregion* column classifies these locations into broader ecological regions

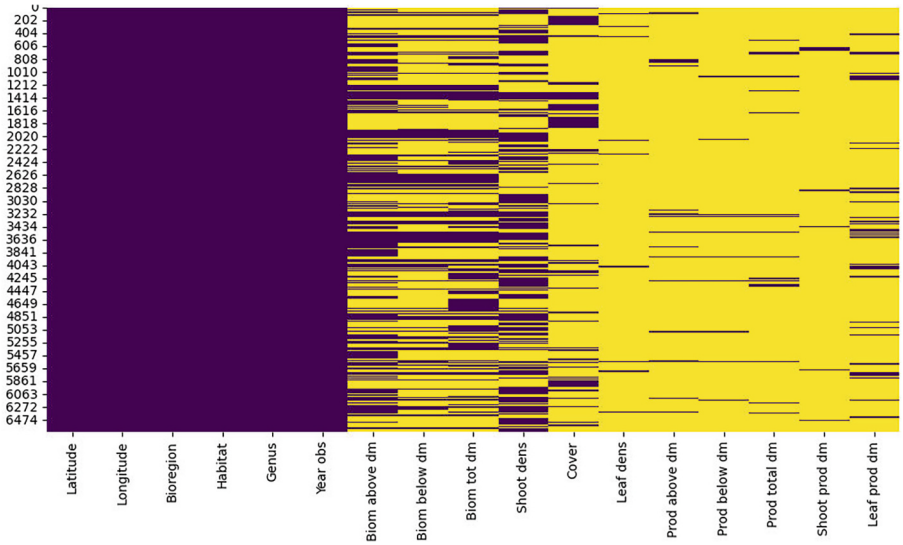

Fig. 3. Heatmap of missing values in the dataset

such as the *Mediterranean* and *Tropical Indo-Pacific*. The dataset shows a significant bias in seagrass distribution studies. The *Tropical Indo-Pacific* region is overrepresented, with over 1750 rows of data, compared to regions like the *Temperate North Atlantic*, *Temperate North Pacific*, and *Mediterranean*, each with fewer than 1000 rows. In contrast, regions like the *Temperate North Atlantic* and *Mediterranean* have fewer species.

4.2 Data Analysis

As shown, in Fig. 6, a worldwide observation of the global distribution of seagrass across different bioregions. We find that regions with harsher weather conditions and poorer economic situations tend to have less seagrass. Environmental factors also play a crucial role: for example, extreme cold, with temperatures consistently below 0 °C in shallow waters, can limit seagrass growth. This is because most species are better suited to temperate or tropical climates.

In Fig. 7, we show the distribution of seagrass in the Temperate North Pacific region, illustrating how challenging it is for seagrasses to thrive in colder areas where photosynthesis and growth are inhibited. Figure 8 presents the global distribution of total seagrass biomass. Darker areas indicate regions with minimal biomass. The Mediterranean Sea stands out, showing the highest total biomass, thanks to its warm climate and abundant sunlight, which offer ideal conditions for seagrass growth. However, Fig. 9 reveals a nuanced picture that despite the high biomass in the Mediterranean, some areas exhibit low daily production, likely due to nutrient limitations.

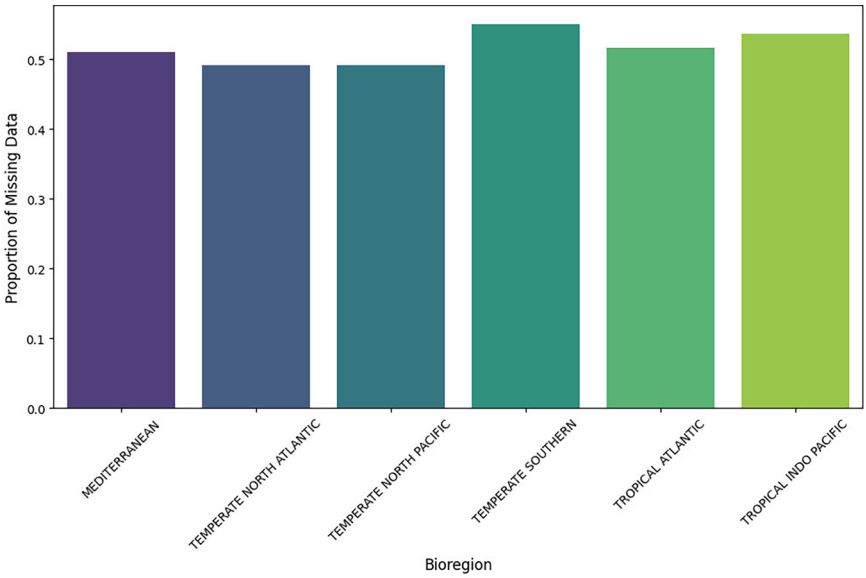

Fig. 4. Average proportion of missing data by bioregion

In Fig. 10, we can observe that seagrass biomass is generally dominated by below-ground mass. However, in certain years: 1979, 2019, 2001–2005, and 2009–2016 we observe dramatic increases in above-ground biomass, sometimes reaching up to three times higher than usual. We believe this could be due to a combination of environmental factors like water temperature, light availability, and nutrient levels, along with human interventions such as restoration projects and recovery efforts following disturbances [16,17].

Looking at Fig. 11, we present that the Tropical Atlantic region has the highest total biomass but relatively low production. This pattern is also evident in the *Mediterranean*, where high biomass doesn't always correlate with high productivity, likely due to environmental stressors. On the other hand, in the *Tropical Indo-Pacific*, both biomass production and total biomass rank second highest among all bioregions, suggesting the region is in its peak development phase. In contrast, the *Temperate North Pacific* shows high production rates but lower total biomass, indicating that it is still developing and has potential for future growth. Finally, the *Temperate North Atlantic*, which has the lowest production and biomass, is constrained by its harsh environmental conditions, including cold temperatures and limited sunlight.

Lastly, Fig. 12 compares the structure of seagrass across bioregions, focusing on the ratio of shoot density to leaf density. In the *Mediterranean*, shoot density is, on average, 5.2 times higher than leaf density, making it unique in this regard. In other regions, leaf density dominates, with the *Tropical Indo-Pacific* displaying the most extreme ratio, with 20 times more leaves than shoots.

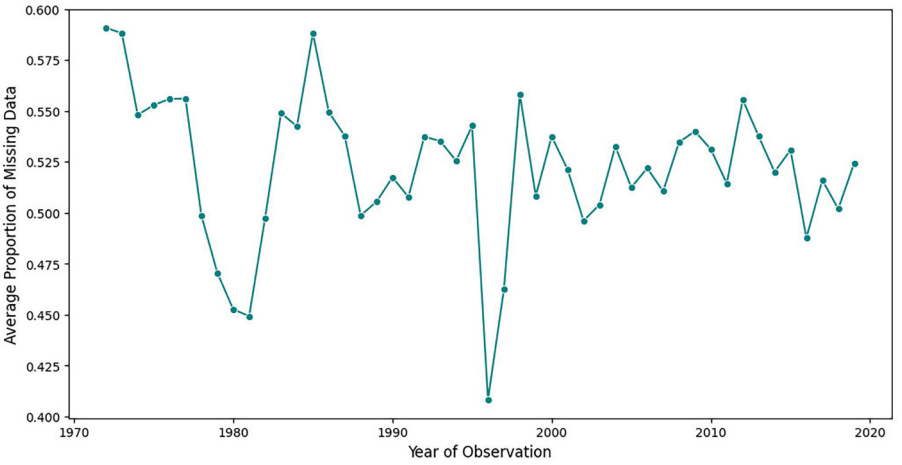

Fig. 5. Trend of missing data over time

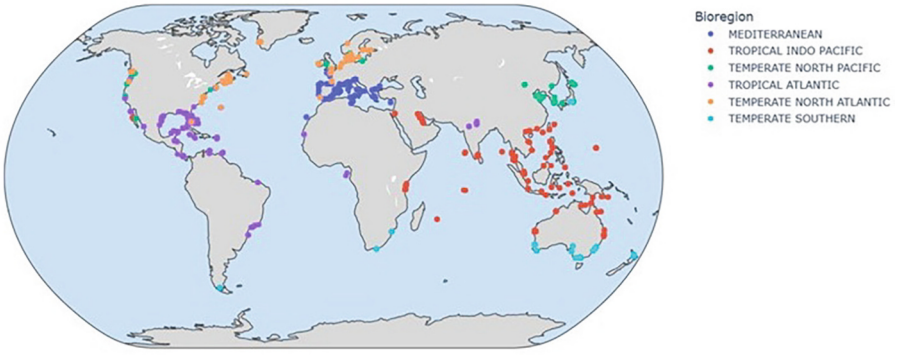

Fig. 6. Seagrass observations accross bioregions

Fig. 7. Seagrass observations in the temperate north pacific bioregion

Fig. 8. Total biomass distribution around the world.

Fig. 9. Total biomass production distribution around the world

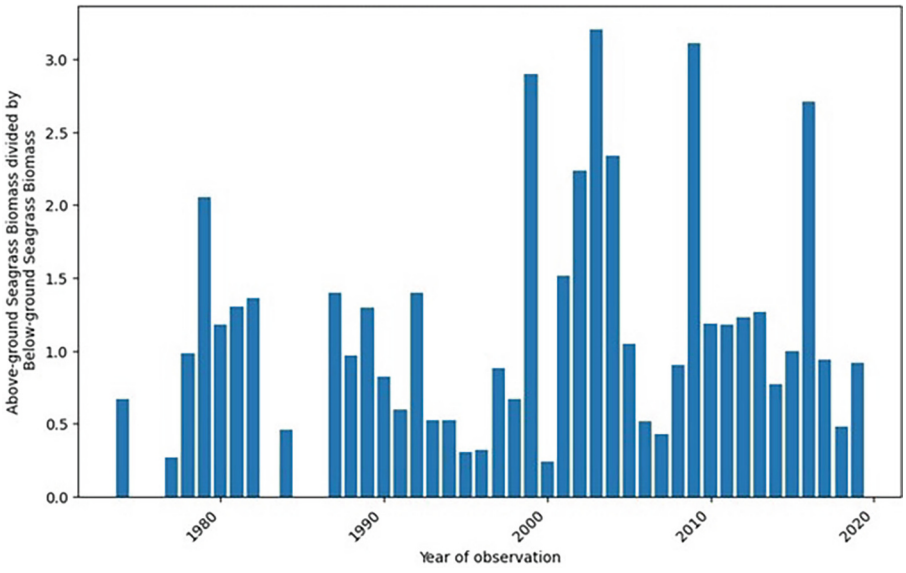

Fig. 10. Proportional study of above/below ground seagrass mass by year

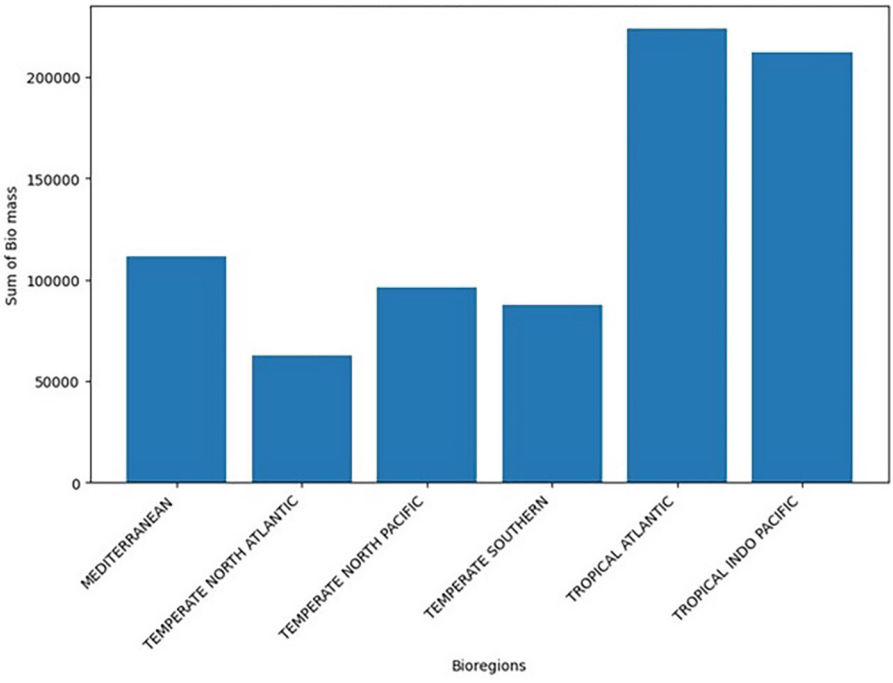

Fig. 11. Distribution of total biomass by bioregion

5 Data Preparation

This section outlines the steps taken to clean, transform, and prepare the data, along with the rationale behind each step. The process includes data cleaning, encoding, handling missing values, standardization, and merging of datasets. First, in order to ensure consistency in the dataset, all string values were converted to uppercase to avoid discrepancies during data processing and analysis. Second, all rows with missing values in the *Year of observation* column were dropped. The duplicate rows were removed to ensure that the dataset does not contain redundant information. We find in this datasat only 146 duplicate rows. As mentioned in Sect. 3, the dataset contains categorical variables, like (*Bioregion*, *Habitat*, and *Genus*), we applied one-hot encoding to convert these categories into binary vectors [9]. One-hot encoding is necessary because most machine learning library require numerical input, and categorical data cannot be directly used in these models. By creating binary columns for each category, one-hot encoding allows the model to interpret categorical data effectively. In the description process of our dataset, we find that some data, such as the *biomass* or *leaf density* columns, require normalization. We scale the data to center it

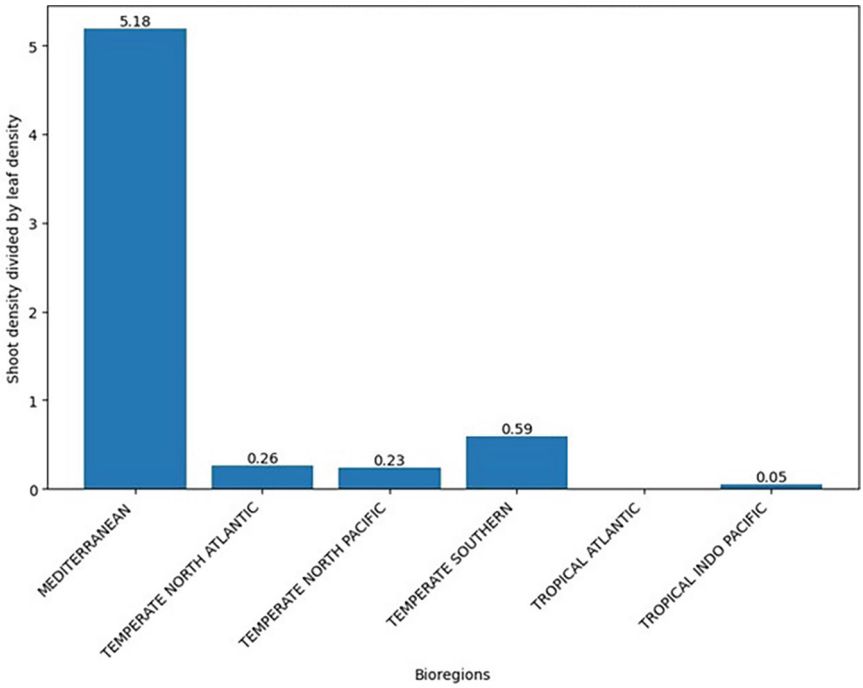

Fig. 12. Seagrass structure comparison by bioregion.

around 0 with a standard deviation of 1. For a given original feature x, the normalized version of x is obtained using the following transformation:

$$x' = \frac{x - \mu_x}{\sigma_x},$$

where μ_x and σ_x are the mean and the standard deviation of x, respectively. The main challenge in this work is how to handle missing values, especially in biological and environmental features. Missing data can introduce bias and reduce the accuracy of model predictions. To address this, the first idea is to use a simple approach based on the median or the mean of the features. However, the main issue with this method is that, upon examining the imputed values, we find that some of them are located in places where they lack relevance or consistency. For this reason, we employ a novel approach based on iterative imputation using a Random Forest Regressor [10]. Random Forest Regressor can capture complex and nonlinear relationships between features. The imputation process refines imputed values by iterating over the dataset multiple times and we evaluated the model across a range of iterations. Furthermore, we split the dataset into a 70% training set and a 30% testing set before imputation, and we analyze the resulting R^2 scores metric for each iteration. From our analysis of the R^2 scores, several key trends emerged. In the early iterations between 0 and 2, we saw

significant improvement in performance for most metrics, particularly for *Biom tot dm* and *Biom above dm*, as the models began to establish better relationships with the data. Although some metrics fluctuated, the overall trend was one of improvement.

As we progressed into the mid-range of iterations between 3 and 5, the R^2 scores for metrics like *Biom above dm* and Cover stabilized, while other metrics showed slight declines, which we interpret as possible signs of overfitting or a performance plateau. Despite these fluctuations, the overall improvements remained evident. On the other hand, production-related metrics such as *Prod total dm* continued to exhibit inconsistency.

Beyond iteration 5, we observed that the R^2 scores for well-performing metrics either plateaued or showed slight declines, likely due to overfitting. For production-related metrics, negative R^2 values persisted, signaling poor generalization, and further iterations offered diminishing returns. This suggests that, after a certain point, additional iterations may no longer contribute meaningfully to model improvement and could even detract from it.

Based on these findings, we conclude that early iterations (0–2) are important for enhancing model accuracy, while additional iterations beyond 5 provide only marginal improvements. After completing the above step, the training and test datasets were merged into a single dataset to obtain a 6502 non-null rows. This datatset will facilitate the analysis and further modeling.

6 Modeling and Evaluation

In this section, we present two algorithms to predict key seagrass metrics, like *biomass*, *density*, and *production*. We used the Stacked Model which is based on multiple three models and the XGBoost model. The detailed architecture for our data modeling is given in Fig. 2. Predictions steps were generated iteratively, with previously predicted variables serving as inputs for subsequent steps. This sequential approach leveraged the interdependencies between *biomass*, *density*, and *production* metrics, allowing the model to dynamically refine its predictions and improve overall accuracy.

6.1 Stacked Model

In the Stacking model, multiple machine learning algorithms have been used, in order to minimize individual biases and variances, resulting in improved generalization and robust performance. By integrating diverse models, stacking captures complex patterns within the data while minimizing the risk of overfitting, making the predictions more reliable when applied to new datasets [11,12]. The stacked model follows a two-level hierarchical structure. At the first level, multiple base models are employed to capture various aspects of the data. Specifically:

– Ridge Regression: Provides a linear approach with regularization to control overfitting.

– Random Forest Regressor: Provides robust predictions through ensemble decision trees with built-in feature importance.
– LGBMRegressor: Optimized for speed and performance, making it suitable for large datasets.

The predictions generated by these base models are then passed to the second-level model, another Random Forest Regressor, which learns from the combined predictions to produce refined and accurate outputs. This layered architecture allows for an optimal fusion of individual model strengths, enhancing both accuracy and generalization. The overall process of stacking is illustrated in Fig. 13.

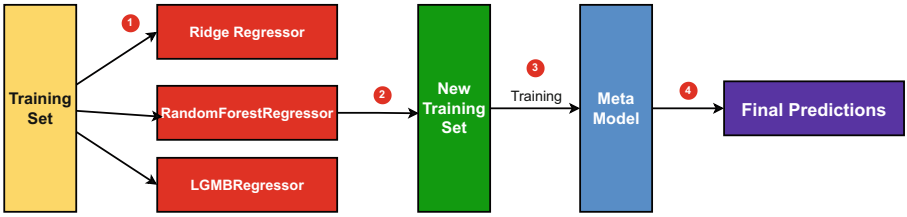

Fig. 13. Stacking model architecture and process

The stacking procedure begin with training each base model independently on the training dataset. Their predictions were then used as inputs for the final-level model, which aggregated the information and generated refined predictions. This hierarchical learning process allowed the final model to make more informed and accurate decisions. Given the interrelated nature of *biomass*, *density*, and *production*, predictions were performed iteratively. The outputs from previous predictions were used as input features for subsequent steps, enabling the model to learn from the evolving context and maintain dynamic adaptability.

6.2 XGBoost Model

The XGBoost model is employed because it excels at capturing complex and nonlinear relationships in the data. XGBoost have a high performance and scalability. It uses parallelized execution, making it significantly faster than many other algorithms, especially for large datasets [13].

To enhance the predictive performance of the XGBoost regressor, we leverage the Tree-structured Parzen Estimator (TPE) as a hyperparameter optimization technique [14,15]. TPE is a Bayesian optimization method that efficiently explores the hyperparameter space by modeling the distribution of the objective function. In this work, the objective of the hyperparameter tuning process is to maximize the coefficient of determination R^2 score, aiming to find the optimal hyperparameters θ^* that satisfy:

$$\theta^* = \arg\max_{\theta} R^2(\theta),$$

where $R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$, with y_i is true value of the target variables, $\hat{y}_i$ is the predicted value from the model, $\bar{y}$ is the mean of the true values and n is the number of samples. The TPE algorithm models the posterior distribution of the objective function and splits the observations into two groups based on a threshold score y^* . The probability model is formulated as:

$$p(\theta \mid y) = \begin{cases} l(\theta) = p(\theta \mid y \geq y^*) & \text{(good hyperparameters)} \\ g(\theta) = p(\theta \mid y < y^*) & \text{(bad hyperparameters)} \end{cases}$$

Hence, we use the acquisition function $\gamma(\theta) = \frac{l(\theta)}{g(\theta)}$ to guide the search by focusing on promising hyperparameters. Hyperparameters optimization was conducted for the XGBoost model across all metrics, revealing several key trends. The number of estimators (`n_estimators`) ranged from 300 to 1500, with higher values often associated with production-related metrics such as *Prod below dm* and *Leaf prod dm*. The learning rate (`learning_rate`) typically ranged between 0.01 and 0.3, with lower rates used for metrics requiring finer adjustments, such as *Shoot prod dm*. The tree depth (`max_depth`) varied from 5 to 15, with deeper trees often applied to complex metrics like *Biom below dm* and *Prod below dm*. Additionally, higher regularization values (`reg_alpha` and `reg_lambda`) were used for metrics prone to overfitting, such as *Biom below dm* and *Shoot dens*. These improved optimized hyperparameters are presented in Table 2.

Table 2. Optimized hyperparameters for key metrics

Metric	`n_estimators`	`learning_rate`	`max_depth`	`reg_alpha`
Biom tot dm	800	0.05	7	0.1
Biom below dm	1000	0.015	9	1.0
Prod below dm	1000	0.201	9	0.5
Shoot prod dm	500	0.01	6	0.2

7 Results and Discussion

We now present the main findings, discussing them in relation to prior work and outlining their significance.

The predictive performance of both XGBoost and stacked ensemble models was evaluated across key seagrass metrics, including *biomass*, *production*, and *density*. Table 3 presents the R^2 scores for each model before and after hyperparameter optimization.

The optimized XGBoost model consistently outperformed both the baseline XGBoost and the stacked ensemble across all metrics. After hyperparameter

Table 3. Comparison between XGBoost and stacked models across R^2 scores

Metric	XGB	Optimized XGB	Stacking	Optimized stacking
Biom above dm	0.842	0.882	0.831	0.891
Biom below dm	0.824	0.857	0.812	0.866
Biom tot dm	0.921	0.973	0.918	0.968
Prod below dm	0.913	0.975	0.907	0.963
Shoot prod dm	0.902	0.962	0.915	0.956

tuning, XGBoost achieved an R^2 score of 0.973 for *total biomass* (Biom tot dm), highlighting its effectiveness in capturing complex relationships within ecological data. Similar gains were observed for other key metrics: *Prod below dm* and *Shoot prod dm* exceeded R^2 values of 0.96. In comparison, the optimized stacked model also showed competitive performance (e.g., $R^2 = 0.968$ for *Biom tot dm*), but slightly underperformed relative to XGBoost for below-ground biomass and production metrics. This reinforces the power of gradient boosting in modeling nonlinear dependencies, consistent with prior findings such as those by Bakirman et al.

Hyperparameter optimization via the Tree-structured Parzen Estimator (TPE) significantly improved predictive performance, yielding an average R^2 increase of 4–6%. For *Biom tot dm*, the optimized XGBoost improved from 0.921 to 0.973, a 5.6% gain. Notably, this tuning approach enhanced performance by nearly 10% over XGBoost's default parameters, demonstrating the utility of Bayesian optimization methods in ecological modeling.

To assess the effect of iterative imputation, we analyzed how model performance changed over successive rounds of MissForest. As shown in Table 4, early iterations (0–2) significantly improved R^2 scores for both models, particularly for biomass metrics.

Table 4. R^2 scores of XGBoost and stacking models across imputation iterations

Iterations	Metric	XGB R^2	Stacking R^2
0	Biom above dm	0.612	0.701
0	Biom below dm	0.517	0.412
0	Biom tot dm	0.981	0.978
1	Biom above dm	0.538	0.619
1	Biom below dm	0.390	0.552
1	Biom tot dm	0.972	0.969
2	Biom above dm	0.612	0.701
2	Biom below dm	0.421	0.589
2	Biom tot dm	0.981	0.978

While the stacked model initially exhibited more stable behavior, it later became erratic–especially for production metrics such as *Prod total dm* and *Leaf prod dm*, occasionally returning negative R^2 values. In contrast, XGBoost showed initial variability but stabilized more consistently across iterations. Based on these observations, we selected XGBoost as the final model due to its superior accuracy, computational efficiency, stability, and interpretability. Notably, it handled missing data more robustly and did not suffer from the extreme negative R^2 values sometimes seen in the stacked model.

Ecological interpretation of model outputs revealed that features such as the light attenuation coefficient (secchi depth), water temperature, and nutrient concentrations ranked highest in importance. These findings align with established ecological theory and prior empirical studies, increasing confidence in our model's ecological validity. Unlike earlier studies relying solely on remote sensing, our approach incorporates *in situ* environmental data, enabling a more detailed and biologically meaningful assessment of seagrass ecosystem health.

In summary, this study makes several important contributions. First, we demonstrate state-of-the-art predictive accuracy for global seagrass metrics by combining advanced machine learning models with rigorous imputation strategies. Second, our feature importance analysis provides ecologically interpretable insights relevant to coastal monitoring and management. Third, the entire pipeline–from preprocessing through model training and evaluation–is transparent and reproducible, supporting adoption in other geographic regions or future datasets.

Nonetheless, some limitations exist. Not all relevant ecological drivers (e.g., localized pollution events, wave energy flux) are represented in the data, which may reduce performance for certain metrics. Geographic imbalances in sampling–particularly in tropical areas–could also affect generalization. Computational limits constrained the complexity of our ensemble and cross-validation scheme. Future work could explore deep learning models with larger, balanced datasets or leverage more computational resources to further enhance performance.

We also plan to incorporate new data sources such as bathymetry and satellite-derived chlorophyll time series, and to deploy the models in real-time environmental dashboards. Alternative imputation methods such as deep learning autoencoders may offer further robustness. Finally, linking our model outputs to carbon emission scenarios could help quantify the global climate impacts of seagrass decline, addressing the urgency highlighted in the Introduction.

8 General Conclusion

This work presents the effectiveness of machine learning techniques in predicting seagrass biomass and ecosystem health. We compared two models: an XGBoost-based model and an ensemble model with optimized parameters. The optimized XGBoost model achieved the best performance, demonstrating its ability to handle complex environmental data. Additionally, we employed an iterative imputation strategy, which proved effective in managing missing values and improving prediction reliability. To further enhance predictive accuracy, we plan to extend this work by exploring time series analysis to capture the temporal evolution of seagrass meadows. Implementing Long Short-Term Memory (LSTM) networks could improve predictions by incorporating past trends and seasonal variations, enabling a more dynamic and real-time monitoring system.

Conflict of Interest The author states that the publishing of this paper does not include any conflicts of interest.

Ethics approval and consent to participate Not applicable.

Funding Not applicable.

Data and code availability All data analyzed during this study are available through PANGAEA in [8]. The full code of this study is available upon request from the corresponding author.

References

1. Nawaz, U., Anees-ur-Rahaman, M., Saeed, Z.: A survey of deep learning approaches for the monitoring and classification of seagrass. Ocean Sci. J. **60**(19) (2025)
2. Capistrant-Fossa, K., Dunton, K.H.: Rapid sea level rise causes loss of seagrass meadows. Commun. Earth & Envir. **5**(87) (2024)
3. Li, L., Goshawk, D.P.: Comparison of random forest and multiple imputation for imputing missing data: a case study of the education panel survey of the city of China (2015). https://www.albany.edu/chinanet/events/ucrn2016/papers/18_Comparison%20of%20Random%20Forest%20and%20Multiple%20Imputation%20for%20Imputing%20Missing%20Data.pdf. Accessed 10 July 2025
4. Chowdhury, M., Martínez-Sansigre, A., Mole, M., Alonso-Peleato, E., Basos, N., Blanco, J.-M., Ramirez-Nicolas, M., Caballero, I., de la Calle, I.: Remote sensing integrating deep learning artificial intelligence to monitor Mediterranean seagrass for conservation and ecosystem management. Sci. Rep. **14**(1) (2024)
5. Traganos, D., Aggarwal, B., Poursanidis, D., Topouzelis, K., Chrysoulakis, N., Reinartz, P.: Towards global-scale seagrass mapping and monitoring using Sentinel-2 on Google Earth engine: the case study of the Aegean and Ionian Seas. Remote Sens. **10**(8) (2018)
6. Yamato, C., Ichikawa, K., Arai, N., Tanaka, K., Nishiyama, T., Kittiwattanawong, K.: Deep neural networks based automated extraction of dugong feeding trails from UAV images in the intertidal seagrass beds. PLoS ONE **16**(8) (2021)

7. Effrosynidis, D., Arampatzis, A., Sylaios, G.: Seagrass detection in the mediterranean: a supervised learning approach. Eco. Inform. **48**, 158–170 (2018)
8. Strydom, S., Webster, C.L., McCallum, R., Lafratta, A., O'Dea, C.M., Said, N.E., Inostroza, K., Salinas, C., Billinghurst, S., Phelps, C.M., et al.: Global database of key seagrass structure, biomass and production variables, PANGAEA (2021)
9. Kuhn, M., Johnson, K.: Applied Predictive Modeling. Springer (2013)
10. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
11. Wolpert, D.H.: Stacked generalization. Neural Netw. **5**(2), 241–259 (1992)
12. Breiman, L.: Bagging predictors. Mach. Learn. **24**(2), 123–140 (1996)
13. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016)
14. Erwianda, M.S.F., Kusumawardani, S.S., Santosa, P.I., Rimadana, M.R.: Improving confusion-state classifier model using XGBoost and tree-structured parzen estimator. In: 2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), pp. 309–313. IEEE (2019)
15. Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M.: Optuna: a next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 2623–2631 (2019)
16. Short, F.T., Kosten, S., Morgan, P.A., Malone, S., Moore, G.E.: Impacts of climate change on seagrass growth and distribution. Mar. Ecol. Prog. Ser. **550**, 1–18 (2016)
17. van Katwijk, M.M., et al.: Global analysis of seagrass restoration: the importance of large-scale planting. J. Appl. Ecol. **53**(2), 567–578 (2016)