

ỨNG DỤNG GOOGLE EARTH ENGINE VÀ DỮ LIỆU GEDI ĐỂ ƯỚC TÍNH TRỮ LƯỢNG CARBON TRÊN MẶT ĐẤT TẠI KHU VỰC ĐẦM THỊ NẠI, TỈNH GIA LAI

Vũ Thị Hoài Thu^{1,2}, Đặng Thị Kiều Oanh¹, Triệu Ánh Ngọc^{3,*}

¹ Viện Công nghệ Tài nguyên nước và Môi trường,

Liên hiệp các Hội Khoa học và Kỹ thuật Việt Nam

² Trường Đại học Giao thông vận tải Thành phố Hồ Chí Minh

³ Trường Đại học Thủy lợi

*Email: ngocta@tlu.edu.vn

TÓM TẮT

Đầm Thị Nại là hệ đầm phá ven biển lớn ở miền Trung, nơi tập trung các kiểu lớp phủ và thảm thực vật đặc trưng của cảnh quan ven đầm. Nghiên cứu này tích hợp ảnh Sentinel-2 mức phản xạ bề mặt, dữ liệu GEDI L2A/L4A V002 và DEM SRTM trên nền tảng Google Earth Engine nhằm: (i) phân loại lớp phủ sử dụng đất năm 2024; (ii) mô hình hóa chiều cao tán (TCH), mật độ sinh khối trên mặt đất (AGBD); và (iii) ước tính trữ lượng carbon trên mặt đất của các lớp phủ có thảm thực vật. Bộ biến đầu vào gồm các băng phổ gốc, NDVI, NDWI, NDBI, ba thành phần chính PCA và hai biến địa hình; bốn thuật toán RF, CART, GBT và SVM được so sánh. Dữ liệu GEDI được lọc theo các chỉ tiêu chất lượng trước khi liên kết với Sentinel-2. Kết quả phân loại cho thấy, mặt nước chiếm 53,81 km²; tiếp theo là cây hàng năm 44,78 km², rừng 39,88 km², đất xây dựng 30,88 km², đất trống 21,03 km² và cây lâu năm 10,26 km²; độ chính xác phân loại đạt OA=99,00% và Kappa=0,9877. Trong mô hình hồi quy, RF cho kết quả tốt nhất với RMSE=3,82 m đối với TCH và 7,77 Mg/ha đối với AGBD, tương ứng R²=0,62 và R²=0,64. Tổng trữ lượng carbon trên mặt đất của các lớp phủ có thảm thực vật ước tính đạt 71.583,47 Mg C. Kết quả khẳng định tiềm năng của việc kết hợp GEE, Sentinel-2 và GEDI trong giám sát carbon trên mặt đất ở hệ sinh thái ven đầm; đồng thời chỉ ra nhu cầu bổ sung kiểm chứng thực địa độc lập và lượng hóa carbon dưới mặt đất/trầm tích trong các nghiên cứu tiếp theo.

Từ khóa: Đầm Thị Nại, carbon trên mặt đất, GEDI, Sentinel-2, học máy.

1. ĐẶT VẤN ĐỀ

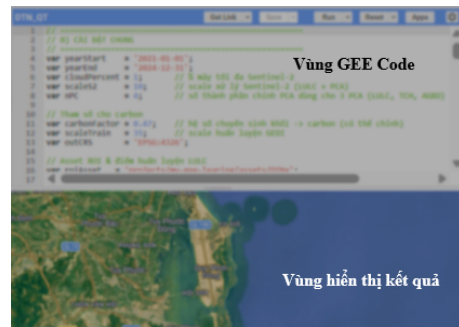
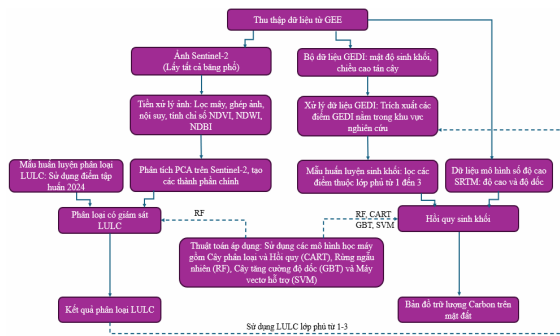
Các hệ sinh thái ven biển như rừng ngập mặn, bãi triều và cỏ biển là những bể chứa carbon quan trọng nhờ khả năng tích lũy vật chất hữu cơ trong môi trường yếm khí và lưu giữ carbon trong thời gian dài [1, 2]. Viễn thám kết hợp điện toán đám mây là công cụ hiệu quả để giám sát sinh khối và carbon trên diện rộng. Google Earth Engine (GEE) cho phép xử lý khối lượng lớn dữ liệu viễn thám [3], trong khi GEDI cung cấp thông tin chiều cao tán và mật độ sinh khối trên mặt đất (AGBD) ở kích thước footprint khoảng 25 m [4]. GEDI có ưu điểm cung cấp thông tin trực tiếp về cấu trúc và sinh khối thảm thực vật, nhưng dữ liệu phân bố không liên tục, do đó cần kết hợp với Sentinel-2 để mở rộng khả năng ước tính theo không gian.

Đầm Thị Nại là hệ đầm phá ven biển lớn ở miền Trung, với nhiều loại lớp phủ và thảm thực vật phân bố đan xen [5, 6]. Các nghiên cứu tại khu vực này chủ yếu tập trung vào cỏ biển và chất lượng môi trường [5, 6], trong khi ứng dụng tích hợp Sentinel-2, GEDI và GEE để phân loại lớp phủ và ước tính carbon trên mặt đất ở quy mô cảnh quan còn hạn chế. Các mô hình học máy có khả năng mô tả quan hệ phi tuyến giữa dữ liệu viễn thám và sinh khối, cho phép so sánh và lựa chọn mô hình phù hợp với khu vực nghiên cứu. Vì vậy, nghiên cứu này được thực hiện nhằm: (i) phân loại lớp phủ sử dụng đất năm 2024; (ii) so sánh các thuật toán học máy trong dự báo TCH và AGBD; và (iii) ước tính trữ lượng carbon trên mặt đất của các lớp phủ có thảm thực vật tại khu vực đầm Thị Nại.

cứ sử dụng chuỗi ảnh Sentinel-2 năm 2024, kết hợp lọc mây và tổng hợp theo giá trị trung vị; tại các footprint GEDI, ưu tiên ảnh trong cửa sổ gần thời điểm quan sát ($\pm 15\text{-}30$ ngày) khi có đủ dữ liệu không mây. Cách xử lý này giúp giảm nhiễu khí quyển, hạn chế trộn phổ cục bộ và tăng tính tương thích giữa dữ liệu LiDAR GEDI với ảnh quang học Sentinel-2.

2.3. Phương pháp nghiên cứu

Quy trình nghiên cứu gồm bốn bước chính: (i) tiền xử lý ảnh Sentinel-2 và xây dựng ảnh tổ hợp năm 2024; (ii) trích xuất tập biến đặc trưng từ băng phổ, chỉ số phổ, thành phần chính PCA và địa hình; (iii) phân loại lớp phủ sử dụng đất và đánh giá độ chính xác bằng ma trận nhầm lẫn; và (iv) xây dựng mô hình hồi quy TCH/AGBD từ GEDI, chuyển đổi sang carbon trên mặt đất và lập bản đồ trữ lượng carbon (Hình 2).



(a) Sơ đồ quy trình tính toán

(b) Triển khai tính toán trên GEE

Hình 2. Sơ đồ khung phương pháp nghiên cứu

2.3.1. Tiền xử lý dữ liệu và trích xuất chỉ số phổ

Dữ liệu Sentinel-2 về mức phản xạ bề mặt được tiền xử lý trên nền tảng Google Earth Engine nhằm đảm bảo tính nhất quán của quang phổ và giảm thiểu sai số khí quyển. Ảnh được loại bỏ các pixel chịu ảnh hưởng bởi mây và bóng mây thông qua mặt nạ chất lượng QA60 kết hợp

ngưỡng che phủ mây, nhằm đảm bảo chỉ giữ lại các quan sát có độ tin cậy cao.

Sau bước lọc mây, chuỗi ảnh được tổng hợp theo phương pháp median composite nhằm giảm nhiễu ngắn hạn do điều kiện khí quyển và thủy văn biến động, đồng thời tăng tính ổn định của tín hiệu phản xạ bề mặt phục vụ các bước trích xuất chỉ số và mô hình hóa tiếp theo.

Bảng 2. Các chỉ số phổ sử dụng trong nghiên cứu

Chỉ số	Công thức tính	Mục đích
Chỉ số thực vật (NDVI)	$NDVI = \frac{NIR - RED}{NIR + RED} \quad (1)$	Đánh giá sức khỏe và mật độ cây xanh
Chỉ số nước (NDWI)	$NDWI = \frac{GREEN - NIR}{GREEN + NIR} \quad (2)$	Tách biệt vùng mặt nước và các khu vực bãi triều/bãi bồi
Chỉ số xây dựng (NDBI)	$NDBI = \frac{SWIR - NIR}{SWIR + NIR} \quad (3)$	Hỗ trợ nhận diện vùng đất trống và khu xây dựng ven đầm

Ghi chú: GREEN, RED, NIR và SWIR lần lượt tương ứng với các băng 3, 4, 8 và 11 của Sentinel-2.

Ba chỉ số viễn thám được sử dụng gồm: NDVI để tăng cường tín hiệu thực vật, NDWI để nhận dạng mặt nước [8] và NDBI để hỗ trợ phân biệt khu xây dựng/đất trống [9]. Bộ chỉ số này được kết hợp với các băng phổ gốc nhằm mô tả tốt hơn sự khác biệt giữa thực vật ven đầm, mặt nước, bãi bồi và khu xây dựng (Bảng 2).

2.3.2. Phân tích thành phần chính (PCA) và thiết lập hệ thống biến đầu vào

Do các băng phổ và chỉ số có tương quan nội tại cao, nghiên cứu áp dụng phân tích thành phần chính (PCA) để giảm chiều dữ liệu và hạn chế đa cộng tuyến [10]. PCA cho phép nén thông tin phổ thành các trục không tương quan, giúp rút gọn bộ

biến đầu vào mà vẫn giữ được cấu trúc biến thiên chính của cảnh quan.

$$PC_i = a_{i1}X_1 + a_{i2}X_2 + \dots + a_{im}X_m \quad (4)$$

Bảng 3. Hệ thống các biến đặc trưng sử dụng trong mô hình

Nhóm biến số	Ký hiệu	Ý nghĩa đặc trưng
Băng phổ góc	B2, B3, B4, B8, B11, B12	Phản xạ phổ cơ bản từ dải nhìn thấy đến hồng ngoại
Chỉ số viễn thám	NDVI, NDWI, NDBI	Khuếch đại đặc tính thực vật, mặt nước và đất trống
Biến nén PCA	PC1, PC2, PC3	Tối ưu hóa dữ liệu, loại bỏ nhiễu và đa cộng tuyến
Biến địa hình	Elevation, Slope	Phân tách sinh thái theo độ cao và độ dốc

Bảng 4. Cấu hình các thuật toán học máy sử dụng cho phân loại LULC

Thuật toán	Cấu hình chính	Ứng dụng
RF	100 cây cho LULC; 50 cây cho hồi quy TCH/AGBD; variablesPerSplit = $\sqrt{p}$; minLeafPopulation = 1; bagFraction = 0,5; seed cố định	Phân loại và hồi quy
CART	Cây quyết định nhị phân; minLeafPopulation = 1; điều chỉnh độ phức tạp theo tập kiểm định	So sánh mô hình nền
GBT	200 cây; shrinkage = 0,05; samplingRate = 0,7	Hồi quy và phân loại phi tuyến
SVM	Kernel RBF; gamma = 0,1; cost = 10	Hồi quy và phân loại

Các thuật toán được thiết lập và tối ưu theo cùng một tập biến đầu vào. RF sử dụng 100 cây quyết định cho phân loại lớp phủ; số biến xét tại mỗi nút được đặt theo căn bậc hai của số biến đầu vào, minLeafPopulation = 1, bagFraction = 0,5 và seed cố định để tăng khả năng tái lập. GBT được thiết lập với 200 cây, shrinkage = 0,05 và samplingRate = 0,7; SVM sử dụng kernel RBF với gamma = 0,1 và cost = 10, CART được sử dụng làm mô hình nền. Tập dữ liệu được chia theo tỷ lệ 70% huấn luyện và 30% kiểm định, đồng thời áp dụng kiểm định chéo 5-fold trong tập huấn luyện để giảm nguy cơ overfitting và lựa chọn cấu hình ổn định (Bảng 4). Mô hình có OA và Kappa cao được lựa chọn để xây dựng bản đồ LULC năm 2024.

2.3.4. Xây dựng mô hình hồi quy TCH và AGBD

Bộ biến sử dụng cho mô hình gồm sáu băng phổ góc, ba chỉ số viễn thám, ba thành phần chính PCA và hai biến địa hình. Cách tổ hợp này nhằm tận dụng đồng thời thông tin quang phổ, cấu trúc cảnh quan và bối cảnh địa hình của khu vực ven đầm (Bảng 3).

2.3.3. Phân loại lớp phủ

Nghiên cứu so sánh bốn thuật toán học máy gồm: Random Forest (RF), CART, Gradient Boosting Tree (GBT) và Support Vector Machine (SVM) [11-14] để phân loại sáu lớp phủ. Đối với bài toán phân loại LULC, các điểm mẫu được xây dựng độc lập cho từng lớp phủ thông qua giải đoán ảnh độ phân giải cao. Mẫu được gán nhãn theo lớp phủ tương ứng và trích xuất các biến dự báo từ Sentinel-2, chỉ số phổ, PCA và địa hình. Bộ mẫu được chia 70% cho huấn luyện và 30% cho kiểm định; tập kiểm định được sử dụng để đánh giá kết quả phân loại bằng OA và Kappa.

Mô hình hồi quy được xây dựng độc lập với phân loại LULC. GEDI L2A được sử dụng để xác định chiều cao tán (TCH), còn GEDI L4A cung cấp dữ liệu tham chiếu AGBD [4]. Các footprint GEDI đạt yêu cầu chất lượng được liên kết với các biến Sentinel-2, chỉ số phổ, PC1-PC3, Elevation và Slope. TCH và AGBD là biến mục tiêu, các biến viễn thám và địa hình là biến dự báo.

Bốn thuật toán RF, CART, GBT và SVM được áp dụng cho hồi quy; RF sử dụng 50 cây, các cấu hình GBT và SVM tương tự bảng 4. Dữ liệu được chia 70% huấn luyện và 30% kiểm định, kết hợp 5-fold CV trong tập huấn luyện. Mô hình được đánh giá bằng R^2 và RMSE; mô hình có R^2 cao và RMSE thấp được lựa chọn để dự báo TCH và AGBD toàn khu vực. Việc kết hợp GEDI với Sentinel-2 và học máy giúp mở rộng

kết quả AGBD từ các footprint GEDI không liên tục sang phạm vi không gian nghiên cứu [15, 16].

2.3.5. Ước tính trữ lượng carbon trên mặt đất

Mật độ carbon trên mặt đất (AGCD) được suy ra từ AGBD bằng cách nhân với hệ số chuyển đổi carbon 0,47, tham chiếu giá trị mặc định của IPCC [17]. Trên cơ sở đó, trữ lượng carbon của từng pixel được xác định bằng AGCD nhân với hệ số carbon và diện tích pixel:

$$C_i = AGBD_i \times f \times A_i \quad (5)$$

Trong đó: C_i là trữ lượng carbon trên mặt đất của pixel i (Mg C); $AGBD_i$ là mật độ sinh khối trên mặt đất (Mg/ha); A_i là diện tích pixel (ha); $f = 0,47$ là hệ số carbon [17]. Các pixel mặt nước, đất trống và đất xây dựng được loại khỏi quá trình tính tổng. Nghiên cứu chỉ xem xét carbon trên mặt đất của thảm thực vật, không bao gồm carbon dưới mặt đất và carbon trong đất/trầm tích.

Toàn bộ quy trình được triển khai trên Google Earth Engine (GEE), cho phép đồng bộ hóa tiền xử lý dữ liệu, trích xuất biến và huấn luyện mô hình trong cùng môi trường tính toán [3].

2.3.6. Đánh giá độ chính xác

Hiệu suất mô hình được đánh giá bằng hai nhóm chỉ tiêu. Đối với phân loại lớp phủ, độ chính xác được đánh giá thông qua ma trận nhầm lẫn tổng hợp từ quy trình kiểm định chéo trên bộ mẫu phân loại, độ chính xác tổng thể (Overall Accuracy - OA) và hệ số Kappa nhằm phản ánh mức độ phù hợp giữa kết quả phân loại và dữ liệu tham chiếu. Do hạn chế về số liệu khảo sát thực

địa độc lập, các điểm tham chiếu được giải đoán từ ảnh độ phân giải cao trên Google Earth, kiểm tra chéo nhiều thời điểm ảnh và đối chiếu với bối cảnh địa hình - sinh thái trên GEE; hạn chế này được nêu rõ trong phần thảo luận bất định.

$$OA = \frac{\sum_{i=1}^k n_{ii}}{N} \quad (6)$$

$$\hat{K} = \frac{N \sum_{i=1}^k n_{ii} - \sum_{i=1}^k (n_{i+} \times n_{+i})}{N^2 - \sum_{i=1}^k (n_{i+} \times n_{+i})} \quad (7)$$

Trong đó: n_{ii} là số mẫu được phân loại đúng của lớp i ; n_{i+} và n_{+i} lần lượt là tổng số mẫu theo hàng và theo cột của lớp i ; N là tổng số mẫu kiểm định. OA phản ánh tỷ lệ phân loại đúng tổng thể, còn Kappa đánh giá mức độ phù hợp sau khi loại trừ khả năng trùng khớp ngẫu nhiên.

Đối với bài toán hồi quy, hai chỉ tiêu sử dụng là hệ số xác định (R^2) và sai số căn trung bình bình phương (RMSE).

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (8)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}} \quad (9)$$

Trong đó: R^2 phản ánh mức độ giải thích biến thiên của mô hình, còn RMSE cho biết sai số dự đoán tuyệt đối so với giá trị tham chiếu GEDI. Mô hình được xem là tốt hơn khi có R^2 cao và RMSE thấp.

2.4. Thiết lập bộ mẫu huấn luyện và kiểm định

Bảng 5. Thống kê bộ mẫu huấn luyện và kiểm chứng cho các lớp phủ (LULC)

STT	Lớp phủ	Số lượng mẫu	Mô tả đặc điểm tại khu vực
1	Mặt nước	59	Diện tích mặt nước đầm, cửa sông và hệ thống kênh rạch nội vùng
2	Rừng	57	Quản thể rừng ngập mặn tự nhiên và trồng mới (Mắm, Đước, Bần...)
3	Cây hàng năm	77	Đất canh tác lúa và hoa màu ngắn ngày tại vùng đệm ven đầm
4	Cây lâu năm	19	Các khu vực trồng cây ăn quả, cây lâu năm và vườn tạp của hộ gia đình
5	Đất trống	49	Các bãi bồi, bãi cát ven bờ hoặc đất trống chưa có thảm thực vật che phủ
6	Đô thị/khu xây dựng	39	Các khu dân cư tập trung, cơ sở hạ tầng giao thông và bãi nuôi trồng thủy sản có mái che

Hệ thống phân loại LULC được xây dựng với sáu lớp chính: Mặt nước, rừng, cây hàng năm, cây lâu năm, đất trống và đất xây dựng. Bộ mẫu được xây dựng bằng cách xác định các khu vực có đặc trưng đồng nhất và đại diện cho từng lớp phủ trên ảnh vệ tinh độ phân giải cao từ Google Earth. Tại các khu vực được xác định, các điểm mẫu được lựa chọn và gán nhãn theo lớp phủ tương ứng; đồng thời đối chiếu với đặc điểm địa hình - sinh thái và kiểm tra trên nhiều thời điểm ảnh để hạn chế sai số giải đoán. Tổng cộng 300 điểm mẫu được sử dụng cho huấn luyện, kiểm định và kiểm tra chéo mô hình phân loại; số lượng và đặc trưng của từng lớp phủ được tổng hợp trong bảng 5.

Sau khi gán nhãn, bộ mẫu được chia ngẫu nhiên thành 70% mẫu huấn luyện và 30% mẫu kiểm định trên nền tảng GEE. Tập huấn luyện được sử dụng để xây dựng mô hình phân loại, trong khi tập kiểm định độc lập được sử dụng để đánh giá độ chính xác thông qua ma trận nhầm

lẫn, OA và Kappa. Ngoài ra, kiểm định chéo 5-fold được thực hiện trên tập huấn luyện nhằm đánh giá độ ổn định của mô hình và hạn chế overfitting. Để giảm ảnh hưởng của tương quan giữa các biến đầu vào, mô hình sử dụng các biến Sentinel-2, chỉ số phổ, các thành phần PCA và biến địa hình. Việc xây dựng mẫu theo các lớp phủ đại diện và kiểm tra chéo nhiều thời điểm ảnh nhằm nâng cao độ tin cậy của nhãn mẫu trong điều kiện chưa có bộ mẫu khảo sát thực địa độc lập.

3. KẾT QUẢ VÀ THẢO LUẬN

3.1. Kết quả nghiên cứu

3.1.1. Phân tích kết quả PCA và tầm quan trọng của biến

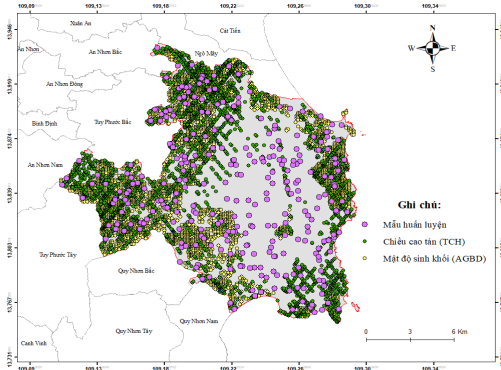
Kết quả cho thấy, ba thành phần chính đầu tiên giải thích 97,99% phương sai toàn bộ tập dữ liệu, trong đó PC1 và PC2 lần lượt đạt 66,28% và 28,21% (Bảng 6). Vì vậy, PC1-PC3 được giữ lại cho các mô hình học máy, vừa giảm trùng lặp thông tin, vừa hạn chế nhiễu từ bộ biến ban đầu.

Bảng 6. Phân bố phương sai tích lũy của các thành phần chính (PCA)

Chỉ số phân tích	PC1	PC2	PC3	PC4	PC5
Độ lệch chuẩn	0.5103	0.3329	0.1172	0.0613	0.0389
Tỷ lệ phương sai (%)	66.2800	28.2100	3.5000	0.9561	0.3855
Phương sai tích lũy (%)	66.2800	94.4900	97.9900	98.9447	99.3301

Ba thành phần chính giữ lại đại diện tương đối tốt cho tương phản phổ của nước, thực vật và bề mặt trống/khu xây dựng. Việc sử dụng PCA vì vậy giúp mô hình ổn định hơn trong bối cảnh dữ liệu đầu vào có nhiều biến và độ tương quan cao.

3.1.2. Phân loại lớp phủ sử dụng đất (LULC) năm 2024



Hình 3. Các điểm mẫu sử dụng cho phân loại LULC và mô hình hóa AGBD, TCH

Bảng 7. Diện tích các loại hình lớp phủ tại khu vực nghiên cứu năm 2024

STT	Lớp phủ (LULC)	Diện tích (km ²)	Tỷ lệ (%)
1	Mặt nước (Water)	53,81	26,82
2	Đất trồng cây hàng năm (Annual crop)	44,78	22,32
3	Rừng (Forest)	39,88	19,88
4	Đất xây dựng (Built-up)	30,88	15,39
5	Đất trống (Barren land)	21,03	10,48
6	Cây lâu năm (Perennial crop)	10,26	5,11
	Tổng cộng	200,64	100,00

Tổng số 300 điểm mẫu LULC được phân bổ theo nguyên tắc đại diện cho sáu kiểu phủ bề mặt

chủ đạo; riêng mô hình hồi quy sử dụng 5.239 footprint AGBD và 2.875 footprint TCH sau lọc chất lượng từ GEDI (Hình 3). Hai nhóm dữ liệu này được sử dụng cho hai nhiệm vụ khác nhau: Điểm LULC phục vụ phân loại lớp phủ, còn footprint GEDI phục vụ mô hình hóa cấu trúc tán và sinh khối. Cách tách biệt này giúp tránh nhầm lẫn giữa dữ liệu phân loại lớp phủ và dữ liệu tham chiếu carbon, đồng thời tăng tính minh bạch của quy trình trước khi tính tổng trữ lượng carbon.

Kết quả phân loại cho thấy, cảnh quan nghiên cứu có sự phân bố đan xen rõ rệt giữa các loại lớp phủ và thảm thực vật, trong đó mặt nước vẫn là lớp phủ chiếm ưu thế. Các lớp thực vật và đất canh tác phân bố thành dải quanh đầm, đan xen

với khu xây dựng và bãi bồi. Diện tích mặt nước đạt 53,81 km² (26,82%), phản ánh đặc trưng của một hệ đầm phá ven biển mở (Bảng 7). Lớp rừng đạt 39,88 km² (19,88%); trong bối cảnh ven đầm, lớp này chủ yếu gồm các dải cây gỗ, rừng ngập mặn ven bờ và mảng thực vật thân gỗ phân bố rời rạc. Hai lớp nông nghiệp (cây hàng năm và cây lâu năm) chiếm tổng cộng 55,04 km², cho thấy áp lực sử dụng đất sản xuất vẫn rất lớn xung quanh đầm. Đất trống và đất xây dựng lần lượt đạt 21,03 km² và 30,88 km², phản ánh các bãi bồi, khu chuyển tiếp và không gian hạ tầng/đô thị ven đầm. Tổng diện tích các lớp phủ là 200,64 km², thống nhất với ranh giới phân tích đã chuẩn hóa ở mục 2.1.

Bảng 8. Ma trận nhầm lẫn tổng hợp của mô hình phân loại LULC (OA=99,00%; Kappa=0,9877)

Tham chiếu\ dự báo	Nước	Rừng	Cây hàng năm	Cây lâu năm	Đất trống	Đất xây dựng	Tổng
Mặt nước	58	0	0	0	0	1	59
Rừng	0	57	0	0	0	0	57
Cây hàng năm	0	0	76	1	0	0	77
Cây lâu năm	0	0	0	19	0	0	19
Đất trống	0	0	0	0	48	1	49
Đất xây dựng	0	0	0	0	0	39	39
Tổng	58	57	76	20	48	41	300

3.1.3. Đánh giá hiệu năng mô hình hồi quy

Bốn thuật toán học máy cho kết quả khác nhau đối với hai biến mục tiêu TCH và AGBD. Trước hết, ma trận nhầm lẫn của mô hình phân loại LULC cho thấy độ chính xác tổng thể đạt OA=99,00% và Kappa=0,9877 (Bảng 8), phản ánh mức độ tách lớp tốt giữa các lớp phủ chủ đạo trong khu vực nghiên cứu. Đối với hồi quy TCH/AGBD, RF cho sai số thấp nhất, với RMSE = 3,82 m đối với TCH và 7,77 Mg/ha đối với AGBD (Bảng 9). Mức chính xác này cho thấy RF phù hợp với bối cảnh dữ liệu ven đầm có tính không đồng nhất cao, đồng thời cân bằng tốt giữa độ chính xác và chi phí tính toán trên nền tảng GEE.

Các mô hình còn lại có RMSE AGBD từ 7,96-9,48 Mg/ha; trong đó CART có sai số lớn nhất, cho thấy khả năng khái quát hóa kém hơn khi áp dụng cho cảnh quan khảm nhiều loại phủ bề mặt. Kết quả này phù hợp với đặc điểm của RF: Thuật toán tổ hợp nhiều cây quyết định nên

ít nhạy với nhiễu cục bộ và có khả năng mô tả tốt quan hệ phi tuyến giữa biến phổ, địa hình và cấu trúc tán.

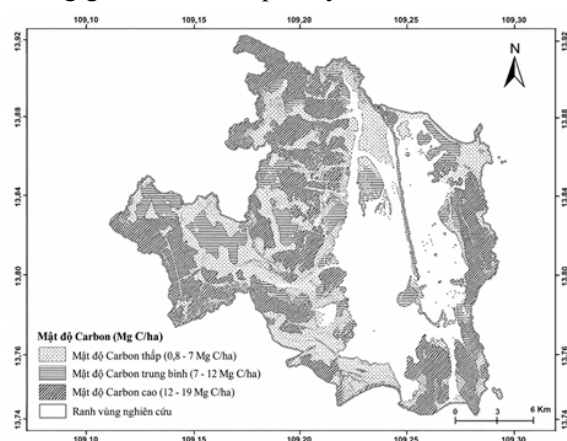
Bảng 9. So sánh sai số (RMSE) giữa các mô hình học máy

Thuật toán	RMSE (TCH - m)	RMSE (AGBD - Mg/ha)	Ghi chú
Random Forest (RF)	3,82	7,77	Độ ổn định cao nhất
CART	5,13	9,48	Dễ bị quá khớp
GBT	3,95	7,96	Cân bằng giữa độ chính xác và tốc độ
SVM	4,22	8,33	Phù hợp tập mẫu nhỏ

Kết quả của nghiên cứu nhìn chung tương thích với xu hướng đã được ghi nhận ở Việt Nam: Mô hình học máy sử dụng dữ liệu Sentinel-2 hoặc tích hợp đa nguồn thường cho khả năng ước tính

sinh khối tốt hơn các mô hình tuyến tính truyền thống. Ở vùng Cần Giờ (xã Cần Giờ, Thành phố Hồ Chí Minh), kết quả nghiên cứu của Phạm và cs (2020) [18] cho thấy, $R^2 = 0,805$ và $RMSE = 28,13$ Mg/ha khi tích hợp Sentinel-2 với ALOS-2 PALSAR-2; trong Khu Dự trữ sinh quyển Đồng bằng sông Hồng, mô hình XGBR-GA đạt $R^2 = 0,683$ và $RMSE = 25,08$ Mg/ha [15]; kết quả nghiên cứu của Quang và cs (2022) [16] ghi nhận, $RMSE = 5,42$ Mg/ha cho mô hình MARS ở cửa sông ven biển miền Bắc.

Sự khác biệt giữa các nghiên cứu chủ yếu xuất phát từ cấu trúc rừng, dải biến thiên sinh khối, loại dữ liệu tham chiếu và mức độ đồng nhất của đối tượng nghiên cứu. Trong nghiên cứu này, mục tiêu là cảnh quan ven đầm có tính khảm mạnh và dữ liệu tham chiếu là footprint GEDI thay vì ô đo thực địa thuần nhất; do đó $R^2 = 0,62$ đối với TCH và $R^2 = 0,64$ đối với AGBD được xem là mức chấp nhận được cho bài toán sàng lọc không gian và hỗ trợ quản lý.

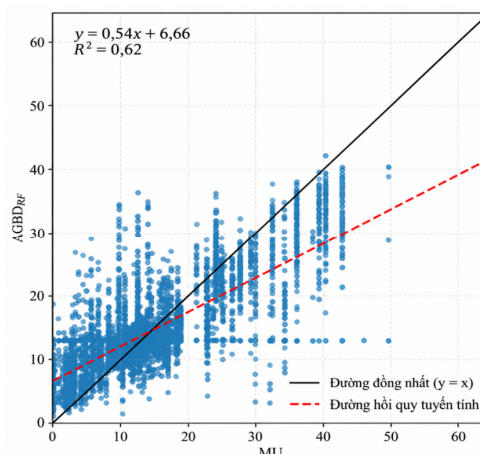


Hình 4. Bản đồ mật độ carbon trên mặt đất (Mg C/ha) theo mô hình Random Forest

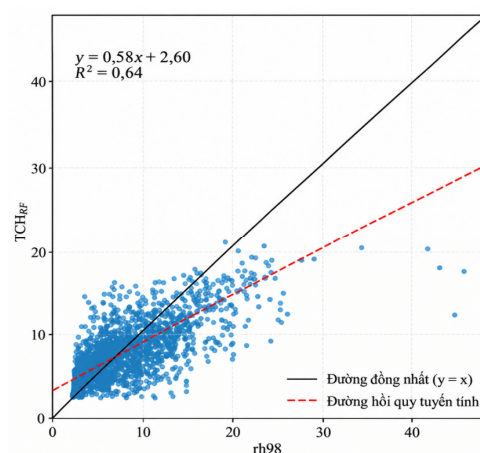
Bản đồ mật độ carbon trên mặt đất cho thấy, các vùng giá trị cao tập trung chủ yếu ở các dải thực vật ven đầm và các khu vực có độ che phủ cây gỗ tốt, trong khi mặt nước, đất trống và đất xây dựng gần như không đóng góp đáng kể (Hình 4). Kết quả này phù hợp với logic sinh thái của hệ đầm phá, nơi sinh khối trên mặt đất phân bố không đều và phụ thuộc mạnh vào dạng phủ bề mặt.

Hai biểu đồ tương quan giữa giá trị tham chiếu GEDI và giá trị dự báo RF có hệ số xác định lần lượt là $R^2 = 0,62$ đối với TCH và $R^2 = 0,64$ đối với AGBD, cho thấy mô hình nắm bắt

được xu thế biến thiên chung của cấu trúc tán và sinh khối trên mặt đất (Hình 5 và 6). Tuy nhiên, độ phân tán vẫn còn hiện diện ở các điểm có giá trị cao, cho thấy nhu cầu bổ sung dữ liệu thực địa và điều chỉnh tham số theo từng sinh cảnh trong các nghiên cứu tiếp theo.



Hình 5. Tương quan giữa AGBD tham chiếu GEDI và AGBD dự báo theo mô hình RF



Hình 6. Tương quan giữa TCH tham chiếu GEDI và TCH dự báo theo mô hình RF

Với mục tiêu lập bản đồ giám sát ở quy mô cảnh quan, RF là lựa chọn hợp lý nhất trong bốn thuật toán thử nghiệm và được chọn cho bước chuyển đổi sang carbon trên mặt đất.

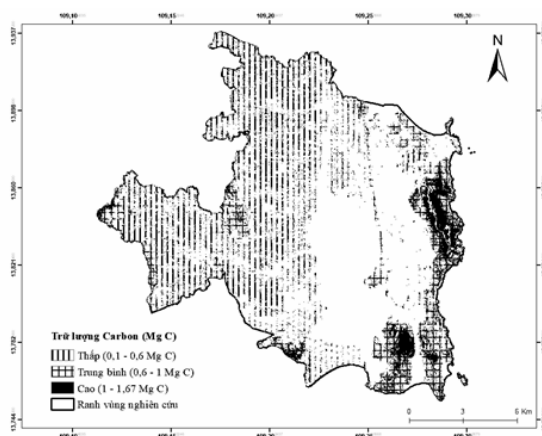
3.1.4. Ước tính trữ lượng carbon trên mặt đất

Dựa trên mô hình RF, bản đồ carbon trên mặt đất được xây dựng cho các lớp phủ có thảm thực vật. Giá trị carbon trung bình đạt 7,46 Mg C/ha và tổng trữ lượng carbon trên mặt đất ước tính đạt 71.583,47 Mg C (Bảng 10).

Bảng 10. Trữ lượng carbon trên mặt đất của thảm thực vật

Thuật toán	Carbon lớn nhất (Mg C/ha)	Carbon trung bình (Mg C/ha)	Carbon nhỏ nhất (Mg C/ha)	Tổng trữ lượng carbon (Mg C)
Random Forest (RF)	19,23	7,46	0,83	71.583,47
CART	30,35	7,41	0,0	71.139,93
GBT	21,95	7,45	0,0	71.496,44
SVM	22,25	7,20	0,0	69.083,70

Giá trị carbon trung bình thu được nhìn chung thấp hơn nhiều nghiên cứu tại các hệ sinh thái rừng ngập mặn điển hình ở Việt Nam và Đông Nam Á, nơi AGBD thường đạt trên 50-150 Mg/ha [1, 15-16, 18-19]. Sự khác biệt này chủ yếu xuất phát từ đặc điểm lớp phủ của khu vực nghiên cứu, vốn không phải là rừng ngập mặn liên tục mà là cảnh quan ven đầm dạng khảm gồm: Mặt nước, bãi triều, đất trồng, cây nông nghiệp và các dải thực vật phân bố rời rạc. Vì vậy, giá trị trung bình 7,46 Mg C/ha không nên diễn giải như mật độ carbon riêng của rừng ngập mặn, mà là giá trị trung bình trên các lớp có thảm thực vật trong toàn cảnh quan ven đầm. Kết quả này cũng chưa đại diện cho toàn bộ “carbon xanh” của hệ đầm phá, vì chưa bao gồm carbon dưới mặt đất và carbon trong đất/trầm tích - các bể chứa thường chiếm tỷ trọng rất lớn ở hệ sinh thái ngập mặn, bãi triều và cỏ biển [1-2].



Hình 7. Bản đồ tổng trữ lượng carbon trên mặt đất theo mô hình Random Forest

Kết quả mô phỏng tại hình 7 phản ánh chi tiết tổng trữ lượng tích lũy trên quy mô pixel 10×10 m, với trữ lượng được phân cấp từ thấp (0,1-0,6 Mg C/pixel) đến cao (1,0-1,67 Mg C/pixel). Các pixel đạt ngưỡng cao tập trung tại vùng lõi rừng

ngập mặn và các dải cây gỗ ven đầm, đóng góp chủ chốt vào tổng trữ lượng 71.583,47 Mg C của toàn khu vực. Cách biểu diễn theo pixel giúp nhận diện các điểm nóng tích lũy carbon, nhưng không thay thế cho kiểm kê carbon đầy đủ theo ô đo thực địa.

3.2. Thảo luận

Kết quả nghiên cứu cho thấy, mô hình RF đạt hiệu quả tốt nhất trong dự báo TCH và AGBD, với RMSE = 3,82 m và 7,77 Mg ha⁻¹, tương ứng R² = 0,62 và 0,64. Kết quả này phù hợp với các nghiên cứu tại Việt Nam về ứng dụng viễn thám và học máy trong đánh giá sinh khối và carbon. Nghiên cứu của Pham và cs (2020) [18] cho thấy, việc kết hợp Sentinel-2, ALOS-2 PALSAR-2 và XGBoost có hiệu quả trong ước tính AGB rừng ngập mặn tại Cần Giờ (xã Cần Giờ, Thành phố Hồ Chí Minh); nghiên cứu của Pham và cs (2020) [15] cũng chứng minh khả năng ứng dụng và so sánh nhiều thuật toán học máy trong ước tính AGB rừng ngập mặn. Bên cạnh đó, nghiên cứu của Pham và cs (2021) [20] cho thấy, Sentinel-2 kết hợp học máy có thể hỗ trợ ước tính carbon đất rừng ngập mặn trên diện rộng, với mô hình CBR-GA đạt R² = 0,665 và RMSE = 18,41 Mg ha⁻¹. Trong khi đó, nghiên cứu của Nguyen Huu Tam và cs (2025) [21] cũng cho thấy tiềm năng của Sentinel-2 trong xây dựng mô hình dự báo carbon trên mặt đất ở vùng Nam Trung Bộ. Các kết quả trên cho thấy, dữ liệu quang học kết hợp học máy có khả năng mô tả biến động sinh khối và carbon trong các hệ sinh thái ven biển.

Ở quy mô quốc tế, kết quả nghiên cứu của Lang và cs (2021) [22] cho thấy, dữ liệu LiDAR không gian kết hợp ảnh quang học có thể hỗ trợ xây dựng bản đồ carbon độ phân giải cao trên diện rộng. Bên cạnh đó, các nghiên cứu ứng dụng dữ liệu viễn thám đa cảm biến cho carbon rừng ngập mặn ở Việt Nam [23] cho thấy, việc kết hợp

nguồn dữ liệu khác nhau có thể cải thiện khả năng mô tả cấu trúc tán và biến thiên sinh khối trong các sinh cảnh ven biển phức tạp. Những kết quả này củng cố tính phù hợp của hướng tiếp cận tích hợp GEDI với Sentinel-2 trên nền tảng GEE cho giám sát carbon xanh ở vùng ven đầm và ven biển Việt Nam.

Tuy nhiên, kết quả AGBD vẫn tồn tại một số nguồn bất định. GEDI cung cấp thông tin trực tiếp về cấu trúc tán và AGBD, nhưng các footprint phân bố không liên tục và có kích thước khoảng 25 m, trong khi Sentinel-2 có độ phân giải 10 m, có thể gây sai khác khi liên kết hai nguồn dữ liệu [4]. Chất lượng và phân bố footprint GEDI cũng ảnh hưởng trực tiếp đến độ tin cậy của mô hình hồi quy. Ngoài ra, đặc điểm lớp phủ dạng khảm tại khu vực ven đầm có thể gây trộn phổ giữa rừng, cây nông nghiệp, bãi bồi và khu xây dựng; các điểm kiểm chứng LULC được giải đoán từ ảnh độ phân giải cao thay vì khảo sát thực địa độc lập cũng là một nguồn bất định. Vì vậy, kết quả AGBD và carbon phù hợp cho giám sát không gian ở quy mô cảnh quan, nhưng cần bổ sung dữ liệu thực địa và đánh giá bất định định lượng trong các nghiên cứu tiếp theo.

4. KẾT LUẬN

Nghiên cứu đã xây dựng quy trình tích hợp ảnh Sentinel-2, DEM SRTM và dữ liệu GEDI L2A/L4A V002 trên nền tảng Google Earth Engine để phân loại lớp phủ và ước tính sinh khối, carbon trên mặt đất tại đầm Thị Nại. Kết quả phân loại năm 2024 đạt OA=99,00% và Kappa=0,9877; mặt nước là lớp phủ chiếm ưu thế, trong khi các lớp thực vật và đất nông nghiệp phân bố thành dải quanh đầm. Trong bốn thuật toán được so sánh, Random Forest cho kết quả tốt nhất với RMSE = 3,82 m đối với TCH và 7,77 Mg/ha đối với AGBD, tương ứng $R^2 = 0,62$ và $R^2 = 0,64$. Tổng trữ lượng carbon trên mặt đất của các lớp phủ có thảm thực vật được ước tính đạt 71.583,47 Mg C. Kết quả cho thấy, việc kết hợp GEDI với Sentinel-2 và học máy có khả năng mở rộng dữ liệu sinh khối điểm thành bản đồ carbon ở quy mô cảnh quan. Tuy nhiên, kết quả vẫn chịu ảnh hưởng bởi sự phân bố không liên tục của footprint GEDI và thiếu dữ liệu kiểm chứng thực địa độc lập. Do đó, cần bổ sung ô đo sinh khối thực địa, dữ liệu carbon dưới mặt đất và trầm tích, đồng thời thực hiện phân tích bất định định

lượng để nâng cao độ tin cậy của kết quả trong các nghiên cứu tiếp theo.

LỜI CẢM ƠN

Bài báo là sản phẩm thuộc đề tài cấp quốc gia KC.09.06/21-30 “Cơ sở khoa học và giải pháp bảo vệ, phục hồi và phát triển bền vững các hệ sinh thái carbon xanh vùng bờ Nam Trung Bộ”.

TÀI LIỆU THAM KHẢO

- [1]. Donato, D. C., Kauffman, J. B., Murdiyarso, D., Kurnianto, S., Stidham, M. & Kanninen, M. (2011). Mangroves among the most carbon-rich forests in the tropics. *Nature Geoscience*, 4, 293-297. <https://doi.org/10.1038/geo1123>.
- [2]. Howard, J., Hoyt, S., Isensee, K., Pidgeon, E. & Telszewski, M. (Eds.). (2014). Coastal blue carbon: Methods for assessing carbon stocks and emissions factors in mangroves, tidal salt marshes, and seagrasses. Conservation International, Intergovernmental Oceanographic Commission of UNESCO, International Union for Conservation of Nature.
- [3]. Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D. & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 202, 18-27. <https://doi.org/10.1016/j.rse.2017.06.031>.
- [4]. Dubayah, R., Blair, J. B., Goetz, S., et al. (2020). The global ecosystem dynamics investigation: High-resolution laser ranging of the Earth's forests and topography. *Science of Remote Sensing*, 1, 100002. <https://doi.org/10.1016/j.srs.2020.100002>.
- [5]. Cao Van Luong and Nguyen Thi Nga (2017). Research on carbon storage capability of seagrass through biomass (in the case of Thi Nai lagoon, Binh Dinh province). *Vietnam Journal of Marine Science and Technology*, 17(1), 63-71. <https://doi.org/10.15625/1859-3097/17/1/7900>.
- [6]. Tran Thi Thanh Dung, Tran Tu Tuan (2021). Application of planetscope-based depth invariant index method in seagrass mapping: The study in Thi Nai lagoon, Binh Dinh province. *Science & Technology Development Journal*, 24(3), 2110-2122. <https://doi.org/10.32508/stdj.v24i3.2737>.

- [7]. Drusch, M., Del Bello, U., Carlier, S., *et al.* (2012). Sentinel-2: ESA's optical high-resolution mission for GMES operational services. *Remote Sensing of Environment*, 120, 25-36. <https://doi.org/10.1016/j.rse.2011.11.026>
- [8]. McFeeters, S. K. (1996). The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. *International Journal of Remote Sensing*, 17(7), 1425-1432. <https://doi.org/10.1080/01431169608948714>.
- [9]. Zha, Y., Gao, J. & Ni, S. (2003). Use of normalized difference built-up index in automatically mapping urban areas from TM imagery. *International Journal of Remote Sensing*, 24(3), 583-594. <https://doi.org/10.1080/01431160304987>
- [10]. Jolliffe, I. T. (2002). *Principal Component Analysis* (2nd ed.). Springer Series in Statistics. Springer, New York. <https://doi.org/10.1007/b98835>.
- [11]. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32. <https://doi.org/10.1023/A:1010933404324>
- [12]. Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth International Group, Belmont, California.
- [13]. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189-1232. <https://doi.org/10.1214/aos/1013203451>.
- [14]. Cortes, C. & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20, 273-297. <https://doi.org/10.1007/BF00994018>
- [15]. Pham, T. D., Yokoya, N., Xia, J., *et al.* (2020). Comparison of machine learning methods for estimating mangrove above-ground biomass using multiple source remote sensing data in the Red River Delta Biosphere Reserve, Vietnam. *Remote Sensing*, 12(8), 1334. <https://doi.org/10.3390/rs12081334>
- [16]. Quang, N. H., Quinn, C. H., Carrie, R., Stringer, L. C., Hue, L. T. V., Hackney, C. R. & Tan, D. V. (2022). Comparisons of regression and machine learning methods for estimating mangrove above-ground biomass using multiple remote sensing data in the Red river estuaries of Vietnam. *Remote Sensing Applications: Society and Environment*, 26, 100725. <https://doi.org/10.1016/j.rsase.2022.100725>
- [17]. Eggleston, H. S., Buendia, L., Miwa, K., Ngara, T. & Tanabe, K. (Eds.). (2006). 2006 IPCC Guidelines for National Greenhouse Gas Inventories. IGES, Japan.
- [18]. Pham, T. D., Le, N. N., Ha, N. T., *et al.* (2020). Estimating mangrove above-ground biomass using extreme gradient boosting decision trees algorithm with fused Sentinel-2 and ALOS-2 PALSAR-2 data in Can Gio Biosphere Reserve, Vietnam. *Remote Sensing*, 12(5), 777. <https://doi.org/10.3390/rs12050777>
- [19]. Hai-Hoa Nguyen, Nguyen Thi Ngoc Bich, Nguyen Huu Nghia, Tran Thi Ngoc Lan, Thi Thu Hien Nguyen, Hoang Vuong Dang (2022). Estimation of changes in above-ground biomass and carbon stocks of mangrove forests using Sentinel-2A in Thai Thuy district, Thai Binh province during 2015-2019. *Vietnam Journal of Science and Technology*, 60(1), 73-91. <https://doi.org/10.15625/2525-2518/15755>.
- [20]. Pham, T. D., Yokoya, N., Nguyen, T. T., Le, N. N., Ha, N. T., Xia, J., Takeuchi, W. & Pham, T. D. (2021). Improvement of mangrove soil carbon stocks estimation in North Vietnam using Sentinel-2 data and machine learning approach. *GIScience & Remote Sensing*, 58(1), 68-87. <https://doi.org/10.1080/15481603.2020.1857623>.
- [21]. Nguyen Huu Tam, Nguyen Van Loi & Hoang Huy Tuan (2025). Establishing models for predicting above-ground carbon stock based on Sentinel-2 imagery for evergreen broadleaf forests in South Central Coastal Ecoregion, Vietnam. *Forests*, 16(4), 686. <https://doi.org/10.3390/f16040686>
- [22]. Lang, N., Schindler, K. & Wegner, J. D. (2021). High carbon stock mapping at large scale with optical satellite imagery and spaceborne LIDAR. arXiv preprint arXiv:2107.07431
- [23]. Pham, T. D., *et al.* (2021). Learning from multimodal and multisensor earth observation dataset for improving estimates of mangrove soil organic carbon in Vietnam. *International Journal of Remote Sensing*, 42(22), 8770-8795. <https://doi.org/10.1080/01431161.2021.1945158>.

APPLICATION OF GOOGLE EARTH ENGINE
AND GEDI DATA TO ESTIMATE ABOVEGROUND CARBON STOCK
IN THE THI NAI LAGOON AREA, GIA LAI PROVINCE

Vu Thi Hoai Thu^{1,2}, Dang Thi Kieu Oanh¹, Trieu Anh Ngoc³

¹*Institute for Water Resources and Environment Technology,
Vietnam Union of Science and Technology Associations*

²*University of Transport Ho Chi Minh City*

³*Thuyloi University*

Abstract

Thi Nai lagoon is a large coastal lagoon system in central Vietnam, encompassing land-cover types and vegetation characteristic of lagoon landscapes. This study integrated Sentinel-2 surface reflectance imagery, SRTM DEM and GEDI L2A/L4A V002 data on the Google Earth Engine platform to (i) map 2024 land use/land cover (LULC), (ii) model tree canopy height (TCH) and aboveground biomass density (AGBD), and (iii) estimate aboveground carbon stock for vegetated covers around Thi Nai lagoon. Spectral bands, NDVI, NDWI, NDBI, three PCA components and two terrain variables were used as predictors, while RF, CART, GBT and SVM were compared. GEDI footprints were quality-filtered and harmonized with Sentinel-2. The 2024 LULC map showed water as the dominant class (53.81 km²), followed by annual crops (44.78 km²), forest (39.88 km²), built-up land (30.88 km²), barren land (21.03 km²) and perennial crops (10.26 km²), with OA=99.00% and Kapp=0.9877. RF outperformed the other algorithms, with RMSE = 3.82 m for TCH and 7.77 Mg ha⁻¹ for AGBD, corresponding to R² = 0.62 and R² = 0.64, respectively. The total aboveground carbon stock of vegetated classes was estimated at 71,583.47 Mg C. The study demonstrates the applicability of combining GEE, Sentinel-2 and GEDI for spatial carbon monitoring in lagoon landscapes, while highlighting the need for independent field validation, belowground/sediment carbon assessment and quantitative uncertainty analysis in future studies.

Keywords: *Thi Nai lagoon, aboveground carbon, GEDI, Sentinel-2, machine learning.*

Ngày nhận bài: 5/3/2026

Ngày chuyển phản biện: 25/3/2026

Ngày thông qua phản biện: 12/8/2026

Ngày duyệt đăng: 17/8/2026