

So sánh mô hình hồi quy CORVI với các thuật toán học máy và học sâu trong ước tính sinh khối rừng ngập mặn trên mặt đất từ dữ liệu Sentinel-1 và Sentinel-2 tại miền Bắc Việt Nam

Nguyễn Văn Thị⁽¹⁾, Nguyễn Hải Hòa^{(2)✉}, Đinh Bá Duy⁽³⁾, Ngô Thành Đạt⁽³⁾

⁽¹⁾ Công ty Cổ phần Thương mại Công nghệ Xuân Mai Green, Hà Nội, Việt Nam

⁽²⁾ Trường Đại học Lâm Nghiệp, Hà Nội, Việt Nam

⁽³⁾ Trung tâm Nhiệt đới Việt - Nga, Hà Nội, Việt Nam

✉ Tác giả liên hệ

E-mail: hoanh@vnuf.edu.vn

ORCID: 0000-0002-8619-1025

Trích dẫn bài báo

Nguyễn, V. T., Nguyễn, H. H., Đinh, B. D., Ngô, T. Đ. (2026). So sánh mô hình hồi quy CORVI với các thuật toán học máy và học sâu trong ước tính sinh khối rừng ngập mặn trên mặt đất từ dữ liệu Sentinel-1 và Sentinel-2 tại miền Bắc Việt Nam. *Tạp chí Khoa học và Công nghệ Nhiệt đới*, 15(3), 24–40.
<https://doi.org/10.58334/vrtc.jtst.vol.015.963>

Tóm tắt

Rừng ngập mặn (RNM) là hệ sinh thái ven biển có chức năng bảo vệ bờ biển, duy trì đa dạng sinh học và tích lũy các-bon xanh. Sinh khối trên mặt đất (AGB) là chỉ tiêu cơ bản để ước tính trữ lượng các-bon rừng, nhưng điều tra thực địa cần nhiều thời gian, chi phí lớn và khó mở rộng trên phạm vi lớn. Nghiên cứu so sánh mô hình hồi quy CORVI dạng lũy thừa với các thuật toán học máy và học sâu gồm Random Forest (RF), Cubist, Support Vector Machine (SVM), eXtreme Gradient Boosting (XGBoost) và Convolutional Neural Network (CNN) trong ước tính AGB RNM tại miền Bắc Việt Nam. Dữ liệu thực địa gồm 322 OTC phân bố tại Quảng Ninh, Hải Phòng, Hưng Yên và Ninh Bình. AGB được tính từ các chỉ tiêu đo đặc thực địa theo loài hoặc nhóm loài và quy đổi về đơn vị Mg/ha. Các biến đầu vào gồm dữ liệu Sentinel-1, Sentinel-2 và chỉ số kết hợp quang học–Radar CORVI. Độ chính xác mô hình được kiểm định chéo 5-fold thông qua thông qua các chỉ tiêu: R^2 , RMSE, MAE và Bias. Kết quả cho thấy, trong kịch bản CORVI đơn biến với mô hình hồi quy lũy thừa đạt hiệu quả kiểm định ổn định nhất hoặc tương đương nhóm mô hình tốt nhất, với $R^2 = 0,6056 \pm 0,0647$, $RMSE = 27,9263 \pm 1,2988$ Mg/ha và $MAE = 23,0128 \pm 1,4570$ Mg/ha. RF và XGBoost có nguy cơ quá khớp khi thông tin đầu vào bị giới hạn. Nghiên cứu khẳng định CORVI là biến tích hợp có giá trị trong mô hình hóa AGB rừng ngập mặn, nhưng cần được kết hợp với các biến quang học, radar, đặc trưng không gian và điều kiện lập địa để nâng cao khả năng dự báo ở quy mô rộng.

Từ khoá

các-bon xanh; chỉ số thực vật radar; học máy; học sâu; kiểm định chéo; mô hình hồi quy lũy thừa; tích hợp quang học–radar.

Đặt vấn đề

Hệ sinh thái (HST) rừng ngập mặn (RNM) ven biển có tính đặc thù, phân bố chủ yếu tại các vùng nhiệt đới và cận nhiệt đới, nơi chịu ảnh hưởng đồng thời của thủy triều, độ mặn, trầm tích và các quá trình động lực ven bờ. HST này đóng góp phần lớn trong việc bảo vệ, chống sạt lở ven bờ, duy trì đa dạng sinh học (ĐDSH), hỗ trợ sinh kế cộng đồng và điều hòa khí hậu (Alongi, 2002; Duke và cộng sự, 2007; Giri và cộng sự, 2011). Trước tình hình biến đổi khí hậu (BĐKH) hiện nay, RNM được quan tâm đặc biệt do khả năng tích lũy và lưu giữ các-bon cao trong sinh khối thực vật và trầm tích. Donato et al. (2011) cho thấy RNM thuộc nhóm HST rừng giàu các-bon ở vùng nhiệt đới, trong khi Alongi (2014) nhấn mạnh vai trò của RNM trong chu trình các-bon ven biển. Các nghiên cứu sau đó tiếp tục khẳng định tiềm năng của RNM trong

giảm nhẹ BĐKH, đặc biệt tại các nước nhiệt đới có diện tích RNM lớn (Murdiyarso và cộng sự, 2015; Rovai và cộng sự, 2018).

Sinh khối trên mặt đất (Above-ground Biomass - AGB) biểu thị lượng vật chất hữu cơ tích lũy trong thân, cành và lá cây, là cơ sở để đánh giá cấu trúc rừng và ước tính trữ lượng các-bon. Việc ước tính chính xác AGB RNM góp phần quan trọng đối với kiểm kê khí nhà kính, giám sát tài nguyên rừng, REDD+, chi trả dịch vụ môi trường rừng (DVMTR) và các hệ thống đo đạc - báo cáo - thẩm định (MRV) trong lâm nghiệp (Kauffman & Donato, 2012; Murdiyarso và cộng sự, 2015; Pham và cộng sự, 2020a). Phương pháp truyền thống xác định AGB thường dựa trên điều tra ô tiêu chuẩn kết hợp phương trình sinh khối theo loài hoặc

nhóm loài. Cách tiếp cận này có ưu điểm là phản ánh trực tiếp đặc điểm lâm học tại hiện trường, tuy nhiên, phương pháp này khó áp dụng ở quy mô lớn do quá trình đo đạc thực địa kéo dài và tốn kém, nhất là tại các khu vực bãi triều, nền bùn lầy và khó tiếp cận (Kauffman & Donato, 2012).

Dữ liệu viễn thám có ưu thế trong ước tính sinh khối và trữ lượng các-bon rừng nhờ khả năng quan sát theo chu kỳ, bao phủ diện rộng và cung cấp các biến đầu vào phục vụ xây dựng mô hình định lượng từ số liệu ô tiêu chuẩn (Fatoyinbo & Simard, 2013; Kuenzer và cộng sự, 2011). Đối với RNM, các nguồn dữ liệu quang học, Radar và LiDAR đã được sử dụng để mô tả phân bố, chiều cao tán, cấu trúc rừng và sinh khối. Dữ liệu quang học (Landsat và Sentinel-2) có ưu thế trong phản ánh trạng thái sinh trưởng, độ xanh và độ ẩm tán rừng thông qua các kênh phổ và chỉ số thực vật. Tuy nhiên, ảnh quang học thường bị ảnh hưởng bởi mây, bóng mây, hơi nước khí quyển và hiện tượng trộn phổ giữa rừng, đất bùn, mặt nước và các lớp phủ ven biển khác. Ngược lại, ảnh Radar (Sentinel-1) có khả năng được cung cấp trong điều kiện mây che phủ và hoạt động cả ngày lẫn đêm, đồng thời nhạy với cấu trúc tán, độ nhám bề mặt và độ ẩm nền. Dù vậy, tín hiệu Radar cũng chịu ảnh hưởng của nhiễu speckle, thủy triều, độ ẩm bề mặt và giới hạn xuyên tán của băng C tại các khu rừng có sinh khối cao (Ghosh & Behera, 2021; Lagomasino và cộng sự, 2019).

Nhiều nghiên cứu xác nhận khả năng ước tính AGB RNM dựa trên dữ liệu viễn thám. Castillo và cộng sự (2017) sử dụng ảnh Sentinel để ước tính AGB của RNM tại Philippines, qua đó cho thấy dữ liệu Sentinel có khả năng hỗ trợ mô hình hóa sinh khối tại các HST ven biển nhiệt đới. Fatoyinbo & Simard (2013) kết hợp dữ liệu ICESat/GLAS và SRTM để ước tính chiều cao và sinh khối RNM ở châu Phi. Simard và cộng sự (2019) xây dựng bản đồ chiều cao tán RNM toàn cầu, cho thấy chiều cao tán có liên hệ chặt

với điều kiện sinh thái và tiềm năng sinh khối. Tại Trung Quốc, Huang và cộng sự (2022) sử dụng Sentinel-1, Sentinel-2 và thuật toán ML để ước tính AGB RNM cho thấy việc tích hợp ảnh vệ tinh quang học và Radar có thể cải thiện độ mô hình so với sử dụng riêng lẻ từng nguồn dữ liệu. Ở Việt Nam, Pham và cộng sự (2018) đã kết hợp ALOS-2 PALSAR-2 với Sentinel-2A thông qua thuật toán học máy để ước tính AGB RNM trồng ở ven biển phía Bắc; tiếp đó, Pham và cộng sự (2020) so sánh nhiều thuật toán ML để ước tính AGB RNM tại Khu DTSQ châu thổ sông Hồng. Kết quả nghiên cứu chỉ ra hướng tích hợp dữ liệu đa nguồn và mô hình học máy là phù hợp với đặc điểm RNM Việt Nam.

Các công bố gần đây giai đoạn 2024–2026 tiếp tục cho thấy xu hướng sử dụng dữ liệu viễn thám đa nguồn kết hợp thuật toán ML trong ước tính sinh khối/các-bon RNM. Farzanmanesh và cộng sự (2024) dùng ảnh Sentinel-2 và mô hình hồi quy vector để xác định biến động AGB trong các dự án *blue forests*, qua đó cho thấy khả năng ứng dụng dữ liệu quang học trong theo dõi hiệu quả bảo tồn và phục hồi RNM. Một số nghiên cứu mới hơn đã mở rộng theo hướng tích hợp Sentinel-1/Sentinel-2, UAV-LiDAR, dữ liệu đa thời gian và mô hình ML nhằm cải thiện khả năng mô tả cấu trúc tán, sinh khối và trữ lượng các-bon RNM (Liu và cộng sự, 2026; Xue và cộng sự, 2026). Mặc dù vậy, các công trình này tập trung vào việc tối ưu độ chính xác bằng cách mở rộng nguồn dữ liệu và số lượng biến đầu vào, trong khi vai trò riêng của một chỉ số tích hợp quang học-radar trong kịch bản đơn biến vẫn cần được kiểm định rõ hơn.

Mặc dù các nghiên cứu trước đã chứng minh tiềm năng của dữ liệu viễn thám và học máy trong ước tính AGB rừng ngập mặn, vẫn còn một số khoảng trống cần tiếp tục làm rõ. Thứ nhất, mô hình hồi quy hoặc mô hình bán thực nghiệm có cấu trúc đơn giản, dễ diễn giải nhưng khó mô tả đầy đủ quan hệ phi tuyến giữa tín hiệu viễn thám và AGB. Thứ hai, các nghiên

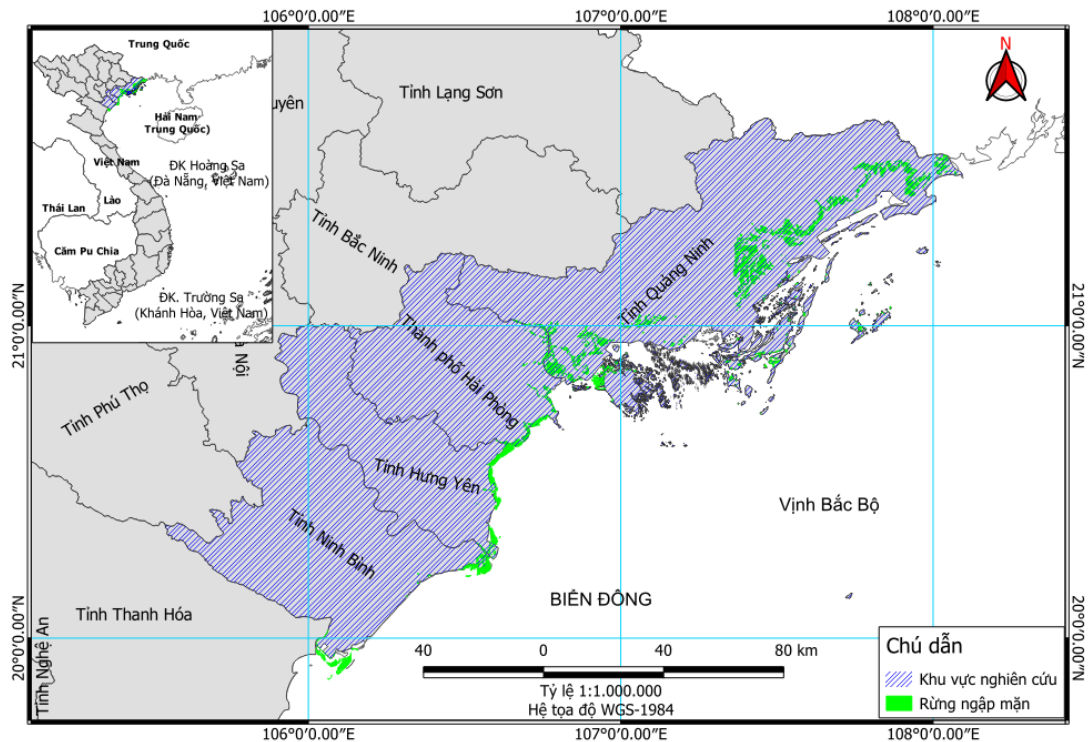
cứ học máy thường sử dụng tập biến đầu vào phong phú, nhưng chưa làm rõ vai trò riêng của các chỉ số tích hợp quang học - Radar trong mô hình. Thứ ba, phần lớn mô hình dạng bảng chỉ trích xuất giá trị tại vị trí ô mẫu hoặc pixel trung tâm, trong khi RNM có cấu trúc không gian phức tạp, thường xen kẽ với bãi bùn, kênh rạch, mặt nước nông, ao nuôi thủy sản và các lớp phủ ven biển khác. Vì vậy, mô hình ước tính AGB RNM cần được tiếp cận theo hướng vừa tận dụng các chỉ số tích hợp từ Sentinel-1/Sentinel-2, vừa kiểm định khả năng cải thiện độ chính xác của các thuật toán ML/DL, đặc biệt là các mô hình có thể khai thác thông tin không gian lân cận.

Các thuật toán ML như RF, Cubist, SVM và XGBoost có ưu thế trong việc mô hình hóa quan hệ phi tuyến và tương tác phức tạp giữa các biến viễn thám với AGB. RF là mô hình tổ hợp dựa trên nhiều cây quyết định, có tính ổn định và chỉ ra mức độ đóng góp của biến đầu vào (Breiman, 2001). Cubist là mô hình hồi quy dựa trên luật, kết hợp phân chia không gian biến với các mô hình hồi quy tuyến tính cục bộ, do đó có ưu thế nhất định về khả năng diễn giải. Support Vector Machine, đặc biệt ở dạng hồi quy vector hỗ trợ, sử dụng cho bài toán hồi quy phi tuyến và tập mẫu không quá lớn (Smola & Schölkopf, 2004). XGBoost là thuật toán boosting dựa trên cây quyết định, có khả năng giảm thiểu sai số dự báo qua quá trình học tuần tự và thường đạt kết quả tối ưu trong nhiều bài toán dữ liệu có cấu trúc (Chen & Guestrin, 2016). Bên cạnh đó, CNN có khả năng học các đặc trưng phi tuyến từ biến đầu vào thông qua phép tích chập, phù hợp để kiểm tra xem bối cảnh không gian quanh ô mẫu có giúp cải thiện ước tính AGB hay không (Le Cun và cộng sự, 2015).

Mặc dù Sentinel-1, Sentinel-2 và ML đã được ứng dụng hiệu quả trong ước tính AGB rừng ngập mặn, nhưng hầu hết các nghiên cứu vẫn dựa trên dữ liệu đa nguồn hoặc nhiều chỉ số viễn thám. Trong bối cảnh đó, vẫn còn thiếu các đánh giá có hệ thống về mức độ hữu ích của một chỉ số tích hợp quang học-radar khi được sử

dụng như biến đầu vào duy nhất. Việc kiểm định kịch bản đơn biến có ý nghĩa phương pháp luận vì giúp làm rõ liệu độ phức tạp của thuật toán có thể bù đắp cho sự giới hạn thông tin đầu vào hay không. Do đó, nghiên cứu này không chú trọng phát triển một thuật toán mới, mà tập trung xem xét sự cân bằng giữa độ chính xác dự báo, khả năng khái quát hóa và tính ổn định của mô hình, nguy cơ quá khớp và tính diễn giải của các mô hình ước tính AGB RNM.

Trên cơ sở đó, nghiên cứu này so sánh mô hình hồi quy CORVI dạng lũy thừa với các mô hình RF, Cubist, SVM, XGBoost và CNN trong ước tính AGB RNM tại miền Bắc Việt Nam (Hình 1), nhằm kiểm tra giả thuyết rằng trong kịch bản chỉ sử dụng một biến tích hợp quang học-radar, mô hình hồi quy đơn giản có thể đạt khả năng kiểm định ổn định hơn hoặc tương đương các thuật toán ML/DL phức tạp. Trong phạm vi nghiên cứu này, CORVI được sử dụng với nghĩa là *Combined Optical-Radar Vegetation Index*, một chỉ số tích hợp dữ liệu vệ tinh quang học và Radar do nhóm tác giả đề xuất trong nghiên cứu này. CORVI được xây dựng từ chỉ số RVI của Sentinel-1 và thông tin hồng ngoại sóng ngắn SWIR của Sentinel-2 nhằm kết hợp tín hiệu tán xạ radar liên quan đến cấu trúc tán, độ nhám bề mặt với thông tin quang học liên quan đến độ ẩm và trạng thái nền bề mặt trong môi trường rừng ngập mặn ven biển. Do đây là chỉ số được đề xuất trong nghiên cứu này, nguồn gốc và công thức tính CORVI được trình bày cụ thể trong phần Phương pháp. Thuật ngữ này không liên quan đến *Climate and Ocean Risk Vulnerability Index* cũng được viết tắt là CORVI trong một số tài liệu quốc tế. Các mô hình được so sánh gồm hồi quy CORVI, RF, Cubist, SVM, XGBoost và CNN. Cách tiếp cận giải quyết ba vấn đề: (1) mức độ cải thiện của mô hình ML/DL so với hồi quy CORVI; (2) làm sáng tỏ vai trò của chỉ số tích hợp quang học - Radar trong mô hình hóa AGB; và (3) tăng cường khả năng ứng dụng mô hình tốt nhất cho lập bản đồ sinh khối, ước tính trữ lượng các-bon và giám sát RNM ven biển.



Hình 1. Vị trí khu vực nghiên cứu

Vật liệu và phương pháp nghiên cứu

Vật liệu

Vật liệu đầu vào gồm dữ liệu điều tra thực địa từ 322 ô tiêu chuẩn (OTC) RNM, trong đó AGB được ước tính từ chỉ tiêu cấu trúc rừng như loài cây, đường kính thân tại vị trí 1,3 m, chiều cao vút ngọn, đường kính tán và mật độ cây. Dữ liệu viễn thám được sử dụng gồm Sentinel-1 SAR với hai phân cực VV, VH và Sentinel-2 MSI với các kênh phổ liên quan, đặc biệt là B11 và B12, thời gian thu nhận ảnh từ đầu tháng 10 đến cuối tháng 12 năm 2025. Từ các dữ liệu này, các biến và chỉ số viễn thám như RVI (Radar Vegetation Index - Chỉ số thực vật Radar) và CORVI (Combined Optical - Radar Vegetation Index - Chỉ số thực vật kết hợp quang học và Radar) được tính toán, trích xuất tại vị trí ô tiêu chuẩn và sử dụng làm đầu vào cho các mô hình ước tính AGB.

Dữ liệu Sentinel-1 và Sentinel-2 được lựa chọn trong giai đoạn từ đầu tháng 10 đến cuối tháng 12 nhằm bảo đảm tính tương đồng về thời gian với dữ liệu thực địa và giảm ảnh hưởng của biến động mùa vụ. Đối với mỗi nguồn dữ liệu, ảnh được tổng hợp theo giá trị trung vị trong giai đoạn nghiên cứu nhằm hạn chế ảnh hưởng của

hiều cục bộ, mây, bóng mây và các biến động tức thời của bề mặt. Tuy nhiên, do RNM tại miền Bắc Việt Nam phân bố trong vùng bãi triều, tín hiệu radar và quang học có thể chịu ảnh hưởng của thủy triều tại thời điểm thu nhận ảnh. Trong phạm vi nghiên cứu này, ảnh hưởng thủy triều chưa được chuẩn hóa hoàn toàn theo từng thời điểm chụp ảnh và được xem là một nguồn bất định cần được thảo luận khi diễn giải kết quả mô hình.

Phương pháp nghiên cứu

Phương pháp thu thập và xử lý số liệu

Các OTC được bố trí đại diện cho các trạng thái rừng, nhóm loài chính và điều kiện lập địa của RNM trong khu vực nghiên cứu. Diện tích ô mẫu được thiết kế theo đặc điểm sinh trưởng, mật độ và kích thước cây của từng nhóm loài. Đối với rừng Bần, do cây thường có kích thước thân lớn, tán rộng và mật độ phân bố thưa hơn, diện tích ô tiêu chuẩn được thiết lập là 500 m² (bán kính ≈ 12,62 m) để phản ánh đầy đủ cấu trúc tầng cây và hạn chế sai số do biến động cá thể trong ô. Đối với các loài cây ngập mặn khác như Trảng, Vẹt, Sú, Mắm, do cây thường có kích thước nhỏ hơn, mật độ dày hơn và cấu trúc quần thụ tương đối đồng đều hơn, diện tích ô mẫu được thiết lập là 100 m² (bán kính ≈ 5,64 m).

Trong mỗi OTC, thực hiện đo đếm đường kính ngang ngực ($D_{1.3}$) đối với những cây đủ điều kiện đo tại vị trí 1,3 m được đo bằng thước dây có độ chính xác đến mm; chiều cao vút ngọn (H_{vn}) được đo bằng thước đo cao SUNTO có độ chính xác đến dm; đường kính tán cây (D_t) được đo bằng thước dây có độ chính xác đến cm.

Tổng số mẫu thực địa gồm 322 mẫu trong đó 282 mẫu điều tra năm 2024 và 40 mẫu điều tra năm 2025 tại Hưng Yên. Các OTC được phân bố ngẫu nhiên phân tầng (theo loài cây chính) ở các tỉnh như sau: Quảng Ninh 99 mẫu, Hưng Yên 43 mẫu, Ninh Bình 130 mẫu và TP. Hải Phòng 50 mẫu.

AGB của từng cá thể được tính bằng các phương trình sinh khối phù hợp với loài hoặc nhóm dạng sống.

Đối với loài Bần, sinh khối cá thể được tính theo công thức (Komiya và cộng sự, 2005):

$$AGB_{B\text{ần}} = 0,251 \times \rho \times D_{1.3}^{2,46} \quad (1)$$

Đối với Trang dạng bụi, sinh khối cá thể được tính theo công thức (Fu & Wu, 2011):

$$AGB_{Trang} = 2,5904 \times (D_t^2 \times H_{vn})^{0,9987} \quad (2)$$

Đối với Vẹt dù, sinh khối cá thể được tính theo công thức (Clough & Scott, 1989):

$$AGB_{V\text{ẹt dù}} = 0,186 \times D_{1.3}^{2,31} \quad (3)$$

Đối với Sú và Vẹt dạng bụi, sinh khối cá thể được tính theo công thức (Fu & Wu, 2011):

$$AGB_{S\text{ú/Vẹt}} = 3,1253 \times (D_t^2 \times H_{vn})^{0,9063} \quad (4)$$

Đối với Mắm biển dạng bụi, sinh khối cá thể được tính theo công thức (Fu & Wu, 2011):

$$AGB_{M\text{ắm}} = 1,8247 \times (D_t^2 \times H_{vn})^{1,0202} \quad (5)$$

Sinh khối ô tiêu chuẩn được tính bằng tổng sinh khối của toàn bộ cá thể trong ô (Kauffman & Donato, 2012):

$$AGB_{\hat{o}} = \sum_{i=1}^n AGB_i \quad (6)$$

Trong đó: $AGB_{\hat{o}}$ là tổng AGB đất của ô mẫu, kg/ô; AGB_i là sinh khối cá thể thứ i, kg/cây; n là tổng số cây trong ô.

Sinh khối ô mẫu được quy đổi về đơn vị Mg/ha theo công thức (Kauffman & Donato, 2012):

$$AGB_{ha} = \frac{AGB_{\hat{o}} \times 10.000}{1.000 \times S_{\hat{o}}} = \frac{AGB_{\hat{o}}}{S_{\hat{o}}} \times 10 \quad (7)$$

Trong đó: AGB_{ha} là AGB của 1 ha, Mg/ha; $S_{\hat{o}}$ là diện tích ô tiêu chuẩn, m²; 10.000 là diện tích 1 ha; 1000 là hệ số đổi từ kg sang Mg.

Phương pháp tiền xử lý dữ liệu viễn thám

Quy trình nghiên cứu gồm các bước chính: chuẩn hóa dữ liệu ô tiêu chuẩn; tiền xử lý ảnh Sentinel-1 và Sentinel-2; tính toán các biến viễn thám; trích xuất giá trị biến tại vị trí ô mẫu; xây dựng mô hình ước tính AGB; kiểm định, đánh giá, so sánh và lựa chọn mô hình..

Đối với Sentinel-1, chỉ số RVI được tính toán từ hai phân cực VV và VH. Do dữ liệu Sentinel-1 GRD trên GEE đã được xử lý thành hệ số tán xạ ngược σ^0 ở đơn vị decibel (dB), giá trị này cần được chuyển về dạng tuyến tính theo công thức (8) (Google Earth Engine, n.d.) trước khi tính toán các chỉ số dựa trên phép tỷ số như RVI và CORVI:

$$\sigma_{linear}^0 = 10^{\frac{\sigma_{dB}^0}{10}} \quad (8)$$

Chỉ số RVI tính theo công thức (Nasirzadehdizaji et al., 2019):

$$RVI = \frac{4 \times VH}{VV + VH} \quad (9)$$

Trong đó: VH và VV là hệ số tán xạ radar ở dạng tuyến tính (tính theo công thức 1).

Đối với Sentinel-2, hai kênh hồng ngoại sóng ngắn B11 và B12 được sử dụng để tính chỉ số SWIRS bên cạnh đó B4 và B8 dùng để tính chỉ số NDVI (Rouse và cộng sự, 1974):

$$SWIRS = B11 + B12 \quad (10)$$

$$NDVI = \frac{B8 + B4}{B8 - B4} \quad (11)$$

CORVI (Combined Optical–Radar Vegetation Index) là chỉ số tích hợp quang học–radar do nhóm tác giả đề xuất trong nghiên cứu này nhằm phục vụ ước tính AGB rừng ngập mặn từ dữ liệu Sentinel-1 và Sentinel-2. Chỉ số này được xây dựng trên cơ sở kết hợp RVI từ Sentinel-1 với thông tin SWIR từ Sentinel-2. Công thức CORVI được xác định như sau:

$$CORVI = RVI^{(1-SWIRS)} \quad (12)$$

Trong đó: B11 và B12 là giá trị phản xạ bề mặt của hai kênh SWIR Sentinel-2 đã được chuẩn hóa về thang 0–1. Thành phần RVI phản ánh đặc

điểm tán xạ radar và cấu trúc bề mặt từ Sentinel-1, trong khi *SWIRS* phản ánh thông tin độ ẩm, nền bề mặt và trạng thái quang phổ vùng hồng ngoại sóng ngắn từ Sentinel-2. Việc đưa *SWIRS* vào số mũ của *RVI* cho phép điều chỉnh tín hiệu radar theo điều kiện quang học - độ ẩm của bề mặt, qua đó hình thành chỉ số tích hợp quang học - radar phù hợp với môi trường RNM ven biển.

Phương pháp so sánh mô hình

Các biến viễn thám được trích xuất tại vị trí 322 ô tiêu chuẩn và ghép nối với giá trị AGB thực địa để xây dựng bộ dữ liệu đầu vào. Độ ổn định và khả năng khái quát hóa của các mô hình được đánh giá bằng kiểm định chéo k-fold với k=5. Cụ thể, bộ dữ liệu được chia ngẫu nhiên thành 5 phần tương đương; tại mỗi vòng lặp, 4 phần được dùng để huấn luyện và phần còn lại dùng để kiểm định. Quy trình được lặp lại 5 lần để mỗi phần dữ liệu đều một lần đóng vai trò là tập kiểm định. Cách đánh giá này được áp dụng thống nhất cho tất cả các mô hình, gồm hồi quy CORVI, học máy và học sâu. Kết quả cuối cùng được trình bày dưới dạng trung bình $\pm$ độ lệch chuẩn của các chỉ tiêu đánh giá qua 5 fold.

Mô hình hồi quy CORVI dạng lũy thừa được sử dụng làm mô hình nền:

$$AGB = a \times CORVI^b \quad (13)$$

Trong đó: *AGB* là sinh khối trên mặt đất RNM, đơn vị Mg/ha; *a* và *b* là các tham số hồi quy. Mô hình này được sử dụng làm mốc so sánh nhằm đánh giá mức độ cải thiện của các thuật toán học máy và học sâu.

Các mô hình so sánh là thuật toán ML/DL, gồm Random Forest (RF), Cubist, Support Vector Machine (SVM), eXtreme Gradient Boosting (XGBoost) và Convolutional Neural Network (CNN). Random Forest là thuật toán học máy tổ hợp dựa trên nhiều cây quyết định, trong đó kết quả dự báo được tổng hợp từ nhiều cây nhằm tăng tính ổn định của mô hình. Cubist là mô hình hồi quy dựa trên luật, kết hợp phân chia không gian biến đầu vào với các mô hình hồi quy tuyến tính cục bộ. Support Vector Machine, trong bài toán hồi quy thường được triển khai dưới dạng hồi quy vector hỗ trợ, có khả năng xử lý quan hệ phi tuyến thông qua hàm nhân. eXtreme Gradient Boosting là thuật toán boosting dựa trên cây quyết định, trong đó các

cây được xây dựng tuần tự để giảm sai số dự báo. CNN là mô hình DL được triển khai dưới dạng mạng tích chập một chiều để xây dựng mô hình quan hệ giữa các biến viễn thám và AGB.

Đánh giá độ chính xác của các mô hình được thực hiện trên tập kiểm định bằng các chỉ tiêu R^2 , RMSE, MAE và Bias. Các công thức tính như sau (Congalton & Green, 2019; Foody, 2002):

$$R^2 = 1 - \frac{\sum_{i=1}^n (AG B_i - \overline{AGB_i})^2}{\sum_{i=1}^n (AG B_i - \overline{AGB})^2} \quad (14)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (AG B_i - \overline{AGB_i})^2} \quad (15)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |AG B_i - \overline{AGB_i}| \quad (16)$$

$$Bias = \frac{1}{n} \sum_{i=1}^n (\overline{AGB_i} - AG B_i) \quad (17)$$

Trong đó: AGB_i là giá trị AGB quan sát tại ô mẫu thứ *i*; $\overline{AGB_i}$ là giá trị AGB dự báo; $\overline{AGB}$ là AGB trung bình của tập kiểm định; *n* là số ô mẫu kiểm định. Việc lựa chọn mô hình tốt nhất dựa trên R^2 cao, RMSE và MAE thấp và Bias gần 0.

Bên cạnh kịch bản sử dụng một biến đầu vào là CORVI, nghiên cứu bổ sung kịch bản đa biến CORVI + NDVI nhằm đánh giá vai trò tương đối của CORVI so với một chỉ số quang học truyền thống trong ước tính AGB rừng ngập mặn.

Đối với mô hình DL, nghiên cứu sử dụng mạng nơ-ron tích chập một chiều (1D-CNN) nhằm kiểm tra khả năng mô hình hóa quan hệ giữa các biến viễn thám đầu vào và AGB. Dữ liệu đầu vào của CNN được chuẩn hóa theo thống kê của tập huấn luyện, sau đó được tổ chức dưới dạng (n_{sample} , $n_{predictor}$, 1). Trong kịch bản CORVI đơn biến, mỗi mẫu có một biến đầu vào là CORVI; trong kịch bản đa biến, mỗi mẫu gồm hai biến CORVI và NDVI. Kiến trúc CNN gồm một lớp tích chập một chiều với 16 bộ lọc, hàm kích hoạt ReLU, tiếp theo là lớp làm phẳng, một lớp Dense gồm 16 nút với hàm kích hoạt ReLU, lớp Dropout với tỷ lệ 0,1 và lớp đầu ra tuyến tính để dự báo AGB. Mô hình được huấn luyện bằng thuật toán Adam với learning rate = 0,005, hàm mất mát MSE, số epoch = 200 và batch size = 8.

Tỷ lệ validation split được đặt là 0,2 trong quá trình huấn luyện. Các tham số này được giữ thống nhất trong các fold nhằm bảo đảm khả năng so sánh giữa các mô hình.

Đối với các mô hình ML có khả năng đánh giá đóng góp biến, nghiên cứu đánh giá mức độ quan trọng của biến đầu vào trong mô hình đa biến CORVI + NDVI. Với mô hình Random Forest, đánh giá thông qua hai chỉ tiêu %IncMSE và IncNodePurity. Trong đó, %IncMSE phản ánh mức tăng sai số dự báo khi giá trị của một biến bị hoán vị; biến có %IncMSE càng cao thì vai trò của biến đó trong mô hình càng lớn. IncNodePurity phản ánh tổng mức giảm độ không thuần nhất tại các nút phân chia do biến đó tạo ra trong toàn bộ rừng cây. Với mô hình XGBoost, mức độ đóng góp của biến được đánh giá bằng chỉ tiêu Gain, thể hiện mức giảm hàm mất mát trung bình do biến đó tạo ra khi được sử dụng để phân chia cây. Các chỉ tiêu này được dùng để so sánh vai trò tương đối giữa CORVI và NDVI trong mô hình đa biến.

Công cụ nghiên cứu

Sử dụng Google Earth Engine (GEE) để xử lý

dữ liệu ảnh Sentinel-1 và Sentinel-2, phần mềm QGIS để quản lý dữ liệu không gian, kiểm tra vị trí ô tiêu chuẩn và biên tập bản đồ thành quả. Các mô hình hồi quy, học máy và học sâu được xây dựng, huấn luyện và đánh giá trong môi trường R. Trong môi trường R, các mô hình được triển khai bằng các gói chuyên dụng được cài đặt sẵn cho hồi quy, học máy và học sâu; toàn bộ quy trình huấn luyện và kiểm định được thực hiện thống nhất theo kiểm định chéo 5-fold.

Kết quả

Đặc điểm phân bố sinh khối khu vực nghiên cứu

Tổng số mẫu thực địa sử dụng trong nghiên cứu là 322 ô tiêu chuẩn, phân bố tại bốn địa phương ven biển phía Bắc Việt Nam, gồm Quảng Ninh, Hải Phòng, Hưng Yên và Ninh Bình. Giá trị AGB của toàn bộ bộ mẫu dao động từ 15,40 Mg/ha đến 282,80 Mg/ha, trung bình đạt 116,99 Mg/ha, độ lệch chuẩn 45,47 Mg/ha và hệ số biến động 38,87% (Bảng 1). Kết quả này cho thấy AGB của RNM trong khu vực nghiên cứu có mức biến động tương đối lớn, thể hiện tính khác biệt về loài cây, tuổi rừng, mật độ, điều kiện lập địa và trạng thái phát triển của rừng.

Bảng 1. Phân bố sinh khối trong khu vực nghiên cứu

TT	Tỉnh/Thành phố	Số mẫu	AGB nhỏ nhất (Mg/ha)	AGB lớn nhất (Mg/ha)	AGB trung bình (Mg/ha)	Độ lệch chuẩn (Mg/ha)	CV (%)
1	Quảng Ninh	99	15,4	282,8	97,3	56,92	58,5
2	Hải Phòng	50	39,9	195,9	119,84	33,63	28,06
3	Hưng Yên	43	44,14	212,5	115,64	29,22	25,27
4	Ninh Bình	130	28,8	227,2	131,34	38,3	29,16
	Toàn khu vực nghiên cứu	322	15,4	282,8	116,99	45,47	38,87

Kết quả kiểm định mô hình đơn biến

Kết quả kiểm định chéo 5-fold cho thấy các mô hình ước tính AGB rừng ngập mặn có mức độ phù hợp khác nhau giữa tập huấn luyện và tập kiểm định (Bảng 2). Trong kịch bản sử dụng một biến đầu vào là CORVI, mô hình Random Forest đạt độ khớp cao nhất trên tập huấn luyện, với $R^2 = 0,8733 \pm 0,0062$, RMSE = $16,1377 \pm 0,4133$ Mg/ha và MAE = $13,0312 \pm 0,2918$ Mg/ha. Tiếp theo là XGBoost với $R^2 = 0,7848 \pm 0,0106$, RMSE =

$21,0306 \pm 0,2642$ Mg/ha và MAE = $17,0808 \pm 0,2498$ Mg/ha. Các mô hình còn lại gồm SVM, hồi quy lũy thừa, CNN và Cubist có R^2 huấn luyện dao động từ 0,6181 đến 0,6389, cho thấy mức độ khớp thấp hơn nhưng tương đối ổn định.

Trên tập kiểm định, mô hình hồi quy lũy thừa dựa trên CORVI cho kết quả tốt nhất nếu xét đồng thời R^2 , RMSE và MAE, với $R^2 = 0,6056 \pm 0,0647$, RMSE = $27,9263 \pm 1,2988$ Mg/ha và MAE = $23,0128 \pm 1,4570$ Mg/ha. Cubist có kết

quả rất gần với mô hình lũy thừa, đạt $R^2 = 0,6033 \pm 0,0369$, RMSE = $28,1432 \pm 1,8926$ Mg/ha và MAE = $23,1868 \pm 1,5797$ Mg/ha. CNN đạt $R^2 = 0,6006 \pm 0,0378$, RMSE = $29,0414 \pm 1,8616$ Mg/ha và MAE = $23,9007 \pm 1,7371$ Mg/ha. SVM có $R^2 = 0,5905 \pm 0,0551$, RMSE = $28,5606 \pm 2,2129$ Mg/ha và MAE = $23,4317 \pm 1,7618$ Mg/ha.

Kết quả chỉ ra các mô hình hồi quy lũy thừa, Cubist, SVM và CNN có độ chính xác kiểm định khá gần nhau, trong đó R^2 dao động từ 0,5905 đến 0,6056. Tuy nhiên, mô hình hồi quy lũy thừa có RMSE và MAE thấp nhất trên tập kiểm định, đồng thời có cấu trúc đơn giản và dễ diễn giải. Điều này cho thấy, trong trường hợp chỉ sử dụng một chỉ số tích hợp quang học - Radar là CORVI, mô hình hồi quy truyền thống vẫn có khả năng khái quát hóa tốt và không thua kém các mô hình ML/DL.

Ngược lại, Random Forest và XGBoost thể hiện sự chênh lệch rõ giữa kết quả huấn luyện và kiểm định. Random Forest đạt $R^2 = 0,8733 \pm 0,0062$ trên tập huấn luyện nhưng giảm xuống

$0,4569 \pm 0,1293$ trên tập kiểm định; RMSE tăng từ $16,1377 \pm 0,4133$ Mg/ha lên $32,6203 \pm 2,5298$ Mg/ha. Tương tự, XGBoost giảm từ $R^2 = 0,7848 \pm 0,0106$ trên tập huấn luyện xuống $0,5455 \pm 0,0921$ trên tập kiểm định; RMSE tăng từ $21,0306 \pm 0,2642$ Mg/ha lên $29,9196 \pm 1,7132$ Mg/ha. Sự chênh lệch này cho thấy các mô hình cây tổ hợp có nguy cơ quá khớp khi chỉ sử dụng một biến đầu vào, dẫn đến khả năng khái quát hóa kém hơn trên tập kiểm định.

Về sai lệch hệ thống, SVM có Bias trung bình gần 0 nhất trên tập kiểm định, đạt $-0,0365 \pm 7,2372$ Mg/ha, tiếp theo là XGBoost ($-1,2134 \pm 4,5775$ Mg/ha), Random Forest ($-1,5240 \pm 4,1710$ Mg/ha), CNN ($-1,8760 \pm 4,5377$ Mg/ha), hồi quy lũy thừa ($-3,2768 \pm 5,9509$ Mg/ha) và Cubist ($1,0480 \pm 5,4127$ Mg/ha). Mặc dù mô hình hồi quy lũy thừa không có Bias gần 0 nhất, nhưng xét tổng hợp các tiêu chí R^2 , RMSE, MAE, tính ổn định và khả năng diễn giải lại là tốt nhất trong các mô hình sử dụng CORVI đơn biến để ước tính AGB rừng ngập mặn.

Bảng 2. Kết quả kiểm định chéo 5-fold đối với mô hình đơn biến

Mô hình	Dataset	R^2	RMSE (Mg/ha)	MAE (Mg/ha)	Bias (Mg/ha)
Lũy thừa	Train	$0,6262 \pm 0,0129$	$27,7232 \pm 0,3593$	$22,8353 \pm 0,3301$	$-3,2895 \pm 0,1043$
Lũy thừa	Test	$0,6056 \pm 0,0647$	$27,9263 \pm 1,2988$	$23,0128 \pm 1,4570$	$-3,2768 \pm 5,9509$
Cubist	Train	$0,6181 \pm 0,0099$	$28,0258 \pm 0,4217$	$23,0186 \pm 0,3614$	$1,0587 \pm 0,4224$
Cubist	Test	$0,6033 \pm 0,0369$	$28,1432 \pm 1,8926$	$23,1868 \pm 1,5797$	$1,0480 \pm 5,4127$
Random Forest	Train	$0,8733 \pm 0,0062$	$16,1377 \pm 0,4133$	$13,0312 \pm 0,2918$	$-0,2245 \pm 0,1309$
Random Forest	Test	$0,4569 \pm 0,1293$	$32,6203 \pm 2,5298$	$25,7364 \pm 1,9691$	$-1,5240 \pm 4,1710$
SVM	Train	$0,6389 \pm 0,0206$	$27,2405 \pm 0,4455$	$22,1722 \pm 0,3362$	$0,2024 \pm 1,3115$
SVM	Test	$0,5905 \pm 0,0551$	$28,5606 \pm 2,2129$	$23,4317 \pm 1,7618$	$-0,0365 \pm 7,2372$
XGBoost	Train	$0,7848 \pm 0,0106$	$21,0306 \pm 0,2642$	$17,0808 \pm 0,2498$	$-0,0004 \pm 0,0530$
XGBoost	Test	$0,5455 \pm 0,0921$	$29,9196 \pm 1,7132$	$24,0812 \pm 0,9961$	$-1,2134 \pm 4,5775$
CNN	Train	$0,6243 \pm 0,0090$	$27,9725 \pm 0,3251$	$23,4962 \pm 0,2740$	$-1,2778 \pm 1,9427$
CNN	Test	$0,6006 \pm 0,0378$	$29,0414 \pm 1,8616$	$23,9007 \pm 1,7371$	$-1,8760 \pm 4,5377$

Kết quả kiểm định mô hình đa biến

Kết quả kiểm định chéo 5-fold đối với kịch bản đa biến CORVI + NDVI chỉ ra việc kết hợp NDVI vào CORVI làm thay đổi hiệu quả của các mô hình ML/DL, nhưng mức cải thiện không đồng đều giữa các thuật toán (Bảng 3). Trên tập huấn luyện, Random Forest đạt độ khớp cao nhất với $R^2 = 0,8949 \pm 0,0046$, RMSE = $14,7026 \pm 0,3021$ Mg/ha và MAE = $11,7919 \pm 0,2389$ Mg/ha. Tiếp theo là XGBoost với $R^2 = 0,8172 \pm 0,0118$, RMSE = $19,3829 \pm 0,5422$ Mg/ha và MAE = $15,6740 \pm 0,3922$ Mg/ha. Các mô hình còn lại gồm CNN, SVM và Cubist có R^2 huấn luyện dao động từ 0,6181 đến 0,6380, thấp hơn rõ so với hai mô hình cây tổ hợp.

Trên tập kiểm định, kết quả cho thấy Cubist đạt $R^2 = 0,6033 \pm 0,0369$, RMSE = $28,1432 \pm 1,8926$ Mg/ha và MAE = $23,1868 \pm 1,5797$ Mg/ha; đây là mô hình có R^2 kiểm định cao nhất trong kịch bản CORVI + NDVI. CNN đạt RMSE thấp nhất với $28,3705 \pm 1,3145$ Mg/ha và MAE = $23,3890 \pm 1,3341$ Mg/ha, đồng thời có $R^2 = 0,5942 \pm 0,0563$. XGBoost đạt $R^2 = 0,5624 \pm 0,0885$, RMSE = $29,4078 \pm 2,4912$ Mg/ha và MAE = $24,0368 \pm 1,3218$ Mg/ha. Random Forest đạt $R^2 = 0,5452 \pm 0,0510$, RMSE = $30,1395 \pm 2,5921$ Mg/ha và MAE = $24,4473 \pm 1,5215$ Mg/ha. SVM có kết quả kiểm định thấp hơn, với $R^2 = 0,5675 \pm 0,0307$, RMSE = $29,4954 \pm 3,2349$ Mg/ha và MAE = $24,0150 \pm 2,1775$ Mg/ha.

Về sai lệch hệ thống, Random Forest có Bias gần 0 nhất trên tập kiểm định, đạt $-0,2462 \pm 5,0570$ Mg/ha, tiếp theo là CNN với $0,3543 \pm 6,2809$ Mg/ha, SVM với $0,7602 \pm 6,9717$ Mg/ha, XGBoost với $-0,7851 \pm 3,5826$ Mg/ha và Cubist với $1,0480 \pm$

$5,4127$ Mg/ha. Điều này phản ánh các mô hình đa biến nhìn chung không tạo ra sai lệch hệ thống lớn, nhưng độ phân tán của Bias giữa các fold còn khá cao, phản ánh sự nhạy cảm nhất định của mô hình đối với cách chia dữ liệu.

So với kịch bản CORVI đơn biến, việc bổ sung NDVI giúp cải thiện rõ hơn đối với các mô hình cây tổ hợp. Random Forest tăng R^2 kiểm định từ $0,4569 \pm 0,1293$ lên $0,5452 \pm 0,0510$, đồng thời RMSE giảm từ $32,6203 \pm 2,5298$ Mg/ha xuống $30,1395 \pm 2,5921$ Mg/ha. XGBoost cũng cải thiện từ $R^2 = 0,5455 \pm 0,0921$ lên $0,5624 \pm 0,0885$, RMSE giảm từ $29,9196 \pm 1,7132$ Mg/ha xuống $29,4078 \pm 2,4912$ Mg/ha. CNN có RMSE giảm từ $29,0414 \pm 1,8616$ Mg/ha xuống $28,3705 \pm 1,3145$ Mg/ha, nhưng R^2 giảm nhẹ từ $0,6006 \pm 0,0378$ xuống $0,5942 \pm 0,0563$. Trong khi đó, SVM giảm hiệu quả kiểm định khi bổ sung NDVI, với R^2 giảm từ $0,5905 \pm 0,0551$ xuống $0,5675 \pm 0,0307$.

Nhìn chung, kết quả kiểm định chéo 5-fold cho thấy kịch bản CORVI + NDVI có thể cải thiện một phần khả năng dự báo của một số thuật toán, nhất là Random Forest và XGBoost, nhưng chưa tạo ra sự cải thiện đồng đều cho tất cả các mô hình. Cubist và CNN là hai mô hình có hiệu quả kiểm định tốt hơn trong kịch bản đa biến, trong đó Cubist có R^2 cao nhất, còn CNN có RMSE và MAE thấp nhất. Kết quả này cho thấy NDVI đóng vai trò bổ trợ cho CORVI, nhưng hiệu quả của việc bổ sung NDVI phụ thuộc vào đặc điểm thuật toán. Vì vậy, CORVI vẫn nên được xem là nền tảng trong mô hình hóa AGB rừng ngập mặn, trong khi NDVI và các chỉ số quang học khác cần được sử dụng có chọn lọc nhằm cải thiện độ chính xác nhưng đồng thời kiểm soát nguy cơ quá khớp.

Bảng 3. Kết quả kiểm định chéo 5-fold đối với mô hình đa biến

Mô hình	Dataset	R ²	RMSE (Mg/ha)	MAE (Mg/ha)	Bias (Mg/ha)
Cubist	Train	0,6181 ± 0,0099	28,0258 ± 0,4217	23,0186 ± 0,3614	1,0587 ± 0,4224
Cubist	Test	0,6033 ± 0,0369	28,1432 ± 1,8926	23,1868 ± 1,5797	1,0480 ± 5,4127
Random Forest	Train	0,8949 ± 0,0046	14,7026 ± 0,3021	11,7919 ± 0,2389	-0,0296 ± 0,1103
Random Forest	Test	0,5452 ± 0,0510	30,1395 ± 2,5921	24,4473 ± 1,5215	-0,2462 ± 5,0570
SVM	Train	0,6380 ± 0,0177	27,2766 ± 0,3532	21,7256 ± 0,3840	0,7866 ± 0,8920
SVM	Test	0,5675 ± 0,0307	29,4954 ± 3,2349	24,0150 ± 2,1775	0,7602 ± 6,9717
XGBoost	Train	0,8172 ± 0,0118	19,3829 ± 0,5422	15,6740 ± 0,3922	0,0070 ± 0,0498
XGBoost	Test	0,5624 ± 0,0885	29,4078 ± 2,4912	24,0368 ± 1,3218	-0,7851 ± 3,5826
CNN	Train	0,6310 ± 0,0139	27,5447 ± 0,3833	22,4353 ± 0,2617	0,1198 ± 0,7520
CNN	Test	0,5942 ± 0,0563	28,3705 ± 1,3145	23,3890 ± 1,3341	0,3543 ± 6,2809

So sánh mô hình hồi quy với các mô hình học máy, học sâu

Kết quả kiểm định chéo 5-fold chỉ ra các mô hình phức tạp hơn không nhất thiết cho khả năng dự báo tốt hơn khi đầu vào chỉ gồm một biến CORVI. Trong kịch bản đơn biến, mô hình hồi quy lũy thừa đạt hiệu quả kiểm định ổn định, trong khi các mô hình ML/DL không tạo ra cải thiện rõ rệt. Điều này chứng tỏ quan hệ giữa CORVI và AGB có thể được mô tả tương đối phù hợp bằng một hàm phi tuyến đơn giản, thay vì cần đến các thuật toán có cấu trúc phức tạp (Hình 2).

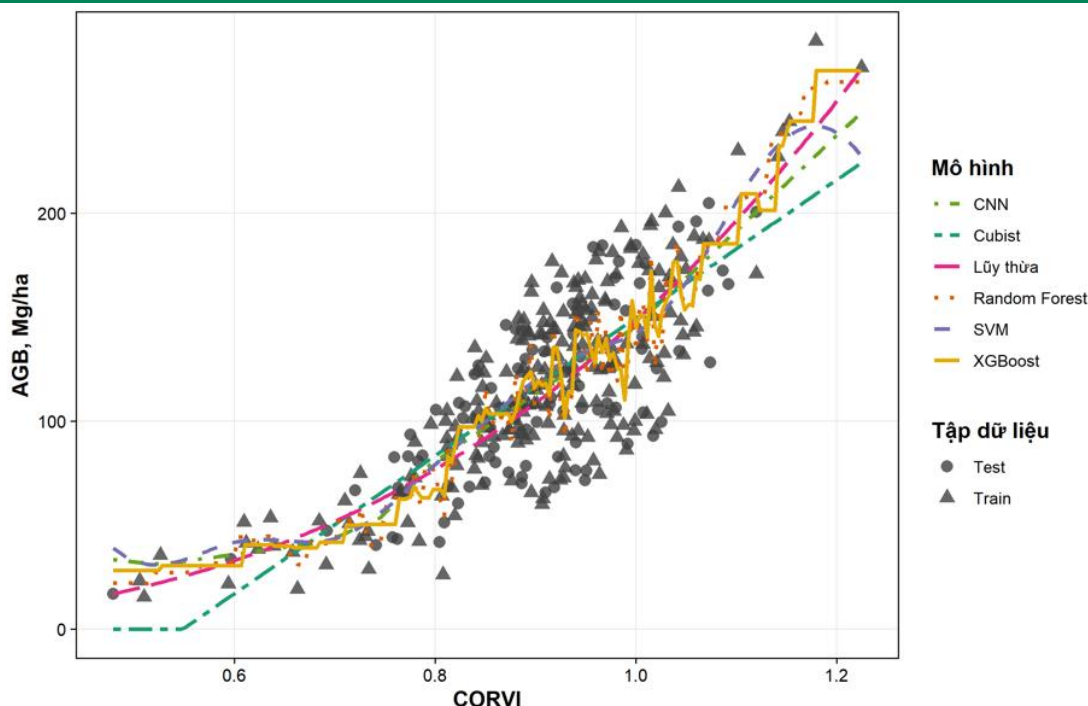
Sự khác biệt rõ nhất thể hiện ở nhóm mô hình cây tổ hợp. RF và XGBoost có khả năng khớp rất tốt trên tập huấn luyện, nhưng hiệu quả giảm khi kiểm định chéo. Nguyên nhân là khi chỉ có một biến độc lập, các thuật toán này chủ yếu chia nhỏ không gian biến một chiều thành nhiều khoảng giá trị, làm tăng khả năng bám sát dữ liệu huấn luyện nhưng giảm tính khái quát hóa. Do đó, kết quả huấn luyện cao không đồng nghĩa với khả năng dự báo ổn định trên dữ liệu mới.

Đối với CNN, lợi thế khai thác đặc trưng không gian chưa thể hiện rõ trong kịch bản CORVI đơn biến. Mặc dù mô hình này có khả năng học thông

tin lân cận từ cấu trúc patch, số lượng biến đầu vào còn hạn chế nên CNN chưa phát huy đầy đủ ưu thế của học sâu. Mô hình này phù hợp hơn với các kịch bản đa kênh, trong đó có thể kết hợp đồng thời các kênh phổ, phân cực radar, chỉ số thực vật, chỉ số độ ẩm và đặc trưng không gian.

Khi bổ sung NDVI vào CORVI, một số mô hình học máy, đặc biệt là RF và XGBoost, có cải thiện nhất định; tuy nhiên, xu hướng này không đồng đều ở tất cả các thuật toán. Điều đó cho thấy NDVI có vai trò hỗ trợ nhưng chưa đủ để làm thay đổi rõ rệt hiệu quả mô hình. Vì vậy, CORVI vẫn có thể được xem là biến nền tảng trong mô hình hóa AGB RNM, còn các biến quang học hoặc radar khác cần được lựa chọn có kiểm soát để tránh tăng độ phức tạp và nguy cơ quá khớp.

Nhìn chung, mô hình hồi quy lũy thừa dựa trên CORVI là lựa chọn phù hợp trong kịch bản đơn biến nhờ cấu trúc đơn giản, dễ diễn giải và khả năng kiểm định ổn định. Các mô hình ML/DL vẫn có tiềm năng cải thiện ước tính AGB, nhưng cần được phát triển trong các kịch bản đa biến hoặc đa kênh ảnh để khai thác tốt hơn thông tin quang học, radar và bối cảnh không gian của RNM.



Hình 2. Biểu đồ các mô hình ước tính sinh khối rừng ngập mặn (AGB)

Vai trò của CORVI trong các mô hình đa biến

Trong mô hình đa biến CORVI + NDVI, kết quả chỉ ra CORVI giữ vai trò chi phối rõ rệt trong các mô hình học máy, đặc biệt ở nhóm mô hình cây quyết định. Đối với RF, mức độ quan trọng của CORVI đạt %IncMSE = 79,20, trong khi NDVI chỉ đạt %IncMSE = 9,65; đồng thời chỉ tiêu IncNodePurity của CORVI đạt 343.900,66, gần gấp 2,8 lần so với NDVI (123.231,07). Kết quả này cho thấy khi CORVI bị loại bỏ hoặc bị hoán vị, sai số dự báo của mô hình tăng mạnh hơn nhiều so với NDVI, chứng tỏ CORVI đóng góp lớn hơn trong việc biến động AGB. Với XGBoost, CORVI cũng có mức đóng góp cao hơn xét theo chỉ tiêu Gain, đạt 0,687, trong khi NDVI đạt 0,313. Như vậy, CORVI là biến tạo ra mức giảm sai số lớn hơn trong quá trình phân chia cây, phản ánh khả năng cung cấp thông tin hữu ích hơn cho dự báo sinh khối. Tuy nhiên, việc thêm biến NDVI vào CORVI không làm cải thiện đồng đều kết quả kiểm định của tất cả mô hình. Trên tập huấn luyện, mô hình RF tăng R^2 từ 0,891 ở mô hình CORVI đơn biến lên 0,907 ở kịch bản CORVI + NDVI; XGBoost tăng từ 0,816 lên 0,842; SVM tăng từ 0,680 lên 0,684; và CNN tăng từ 0,672 lên 0,674. Điều này cho thấy NDVI có bổ sung

thêm thông tin cho quá trình khớp mô hình, nhưng mức tăng chủ yếu thể hiện ở tập huấn luyện. Trên tập kiểm định, hiệu quả đa biến lại không ổn định: RF tăng R^2 từ 0,378 lên 0,446, XGBoost tăng từ 0,417 lên 0,449, nhưng SVM giảm từ 0,506 xuống 0,373, CNN giảm từ 0,530 xuống 0,507, còn Cubist hầu như không thay đổi. Điều này cho thấy việc thêm NDVI có thể giúp một số mô hình cây tổ hợp cải thiện khả năng dự báo, nhưng cũng có thể làm tăng độ phức tạp và giảm tính khái quát hóa ở một số thuật toán khác.

Nhìn chung, các kết quả trên khẳng định CORVI là biến nền tảng trong mô hình ước tính AGB RNM, vì nó duy trì mức đóng góp cao hơn NDVI trong cả RF và XGBoost. Kết quả này xuất phát từ khả năng kết hợp nhiều nguồn thông tin của CORVI: kết hợp thông tin tán xạ radar từ Sentinel-1, liên quan đến cấu trúc tán và độ nhám bề mặt, với thông tin SWIR từ Sentinel-2, liên quan đến độ ẩm và trạng thái nền bề mặt. Trong môi trường RNM ven biển, nơi tín hiệu quang học dễ bị ảnh hưởng bởi bùn nước, thủy triều và pixel hỗn hợp, CORVI cung cấp thông tin tổng hợp hơn so với NDVI đơn thuần. Vì vậy, trong các kịch bản mô hình đa biến tiếp theo, CORVI nên được giữ như biến

đầu vào chủ đạo, còn NDVI và các biến quang học/radar khác nên được xem là biến bổ trợ nhằm cải thiện thêm độ chính xác nhưng cần kiểm soát nguy cơ quá khớp.

Thảo luận

Kết quả của nghiên cứu này có thể đặt trong tương quan với một số nghiên cứu trước về ước tính AGB rừng ngập mặn bằng dữ liệu viễn thám đa nguồn. Castillo và cộng sự (2017) sử dụng Sentinel-1 SAR và Sentinel-2 đa phổ để ước tính, lập bản đồ AGB rừng ngập mặn tại Philippines, đạt $R^2 = 0,72$ và $RMSE = 28,38$ Mg/ha, với khoảng AGB từ 0 đến 346 Mg/ha. Tại Việt Nam, Pham và cộng sự (2018) tích hợp ALOS-2 PALSAR-2 và Sentinel-2A với mô hình SVR cho 25 ô mẫu RNM trồng ven biển Hải Phòng, đạt $R^2 = 0,596$, $RMSE = 0,187$ và $MAE = 0,123$. Trong nghiên cứu hiện tại, mô hình hồi quy CORVI dạng lũy thừa đạt $R^2 = 0,6056 \pm 0,0647$, $RMSE = 27,9263 \pm 1,2988$ Mg/ha và $MAE = 23,0128 \pm 1,4570$ Mg/ha. So với Castillo và cộng sự (2017), R^2 của nghiên cứu này thấp hơn, nhưng RMSE gần tương đương; so với Pham và cộng sự (2018), R^2 đạt mức tương tự. Tuy nhiên, các so sánh này cần được hiểu thận trọng do khác biệt về cảm biến, thuật toán, phạm vi không gian, số lượng ô mẫu và cách đánh giá mô hình. Điểm đáng chú ý là mô hình CORVI chỉ sử dụng một chỉ số tích hợp quang học–radar nhưng vẫn đạt độ chính xác tương đương một số nghiên cứu sử dụng tập biến đa nguồn, qua đó cho thấy tiềm năng của CORVI như một biến nền tảng trong ước tính AGB rừng ngập mặn.

Một số công bố gần đây cho thấy XGBoost và các biến thể boosting thường đạt kết quả tốt khi được xây dựng trên tập biến đầu vào phong phú. Pham và cộng sự (2020a) sử dụng dữ liệu đa nguồn gồm Sentinel-2, Sentinel-1, ALOS-2 PALSAR-2 và các chỉ số thực vật để ước tính AGB rừng ngập mặn tại Khu dự trữ sinh quyển châu thổ sông Hồng. Trong nghiên cứu này, mô hình XGBR-GA đạt $R^2 = 0,683$ và $RMSE = 25,08$ Mg/ha, đồng thời cho kết quả tốt hơn các mô hình học máy khác được so sánh (Pham và cộng sự 2020a). Tại Khu dự trữ sinh quyển Cần

Giờ, Pham và cộng sự (2020b) tích hợp Sentinel-2 với ALOS-2 PALSAR-2 trong mô hình XGBoost và đạt $R^2 = 0,805$, $RMSE = 28,13$ Mg/ha trên tập kiểm định (Pham và cộng sự 2020b). Huang và cộng sự (2022) sử dụng 227 ô mẫu kết hợp dữ liệu Sentinel-1, Sentinel-2 và các thuật toán học máy để ước tính AGB rừng ngập mặn tại Maowei Sea, Trung Quốc; mô hình LGBM-PSO cho kết quả tốt nhất với $R^2 = 0,7807$, $RMSE = 24,6864$ Mg/ha và khoảng AGB dự báo từ 4,623 đến 206,975 Mg/ha.

So với các nghiên cứu trên, kết quả của nghiên cứu này có R^2 thấp hơn các mô hình boosting đa biến trong Pham và cộng sự (2020a, 2020b) và Huang và cộng sự (2022), nhưng RMSE ở mức tương đương. RMSE của nghiên cứu này gần với Pham và cộng sự (2020b) tại Cần Giờ, nơi mô hình XGBoost đạt $RMSE = 28,13$ Mg/ha, và chỉ cao hơn khoảng 2,85 Mg/ha so với Pham và cộng sự (2020a) tại châu thổ sông Hồng. Tuy nhiên, cần nhấn mạnh rằng các nghiên cứu trước sử dụng tập biến đa nguồn và quy trình lựa chọn đặc trưng, trong khi nghiên cứu này chủ yếu kiểm định năng lực của một chỉ số tích hợp quang học–radar là CORVI. Do đó, kết quả hiện tại cho thấy CORVI có tiềm năng làm biến nền tảng trong ước tính AGB RNM, nhưng để đạt mức R^2 cao tương tự các mô hình boosting đa biến, cần bổ sung thêm các biến quang học, radar, cấu trúc rừng, điều kiện lập địa và yếu tố thủy triều.

Các nghiên cứu giai đoạn 2024–2026 tiếp tục cho thấy xu hướng ước tính AGB rừng ngập mặn đang chuyển mạnh sang tích hợp dữ liệu đa nguồn, dữ liệu đa thời gian và các mô hình học máy/học sâu. Các nghiên cứu sử dụng Sentinel-2, Sentinel-1/Sentinel-2, UAV-LiDAR hoặc dữ liệu đa thời gian đều nhấn mạnh vai trò của việc bổ sung nhiều nhóm biến đầu vào nhằm mô tả tốt hơn cấu trúc tán, trạng thái sinh trưởng, độ ẩm và điều kiện nền bề mặt (Farzanmanesh và cộng sự, 2024, Liu và cộng sự, 2026, Xue và cộng sự, 2026). So với các hướng tiếp cận này, nghiên cứu hiện tại có phạm vi hẹp hơn về số lượng biến đầu vào, nhưng lại tập trung làm rõ giá trị và giới hạn

của CORVI khi được sử dụng như một biến tích hợp duy nhất. Đây là khác biệt quan trọng về mục tiêu phương pháp luận: Nghiên cứu không nhằm tối ưu hóa độ chính xác bằng mọi nguồn biến có thể, mà kiểm định liệu một chỉ số tích hợp quang học–radar có đủ thông tin để hỗ trợ mô hình hóa AGB và liệu mô hình phức tạp có thực sự vượt mô hình hồi quy đơn giản trong điều kiện đầu vào bị giới hạn hay không.

So với các nghiên cứu trên, kết quả của nghiên cứu này có điểm khác biệt quan trọng: Các mô hình được so sánh ở đây chủ yếu được thiết kế theo hướng sử dụng một biến đầu vào là CORVI, nhằm kiểm tra riêng năng lực của chỉ số tích hợp quang học–radar này. Do đó, việc RF và XGBoost không vượt mô hình hồi quy lũy thừa trên tập kiểm định không mâu thuẫn với các công bố trước đó. Nguyên nhân là các nghiên cứu đạt độ chính xác cao thường sử dụng nhiều biến đầu vào, bao gồm kênh phổ, chỉ số thực vật, thông tin radar, biến cấu trúc hoặc quy trình lựa chọn đặc trưng (Huang và cộng sự, 2022, Pham và cộng sự, 2018, 2020a, 2020b). Trong khi đó, khi chỉ sử dụng một biến độc lập, các thuật toán cây tổ hợp như RF và XGBoost không khai thác được đầy đủ ưu thế xử lý tương tác giữa nhiều biến, mà chủ yếu phân chia không gian biến một chiều thành nhiều khoảng giá trị nhỏ. Điều này tương thích với đặc điểm của RF là mô hình tổ hợp dựa trên nhiều cây quyết định (Breiman, 2001) và XGBoost là mô hình boosting tuần tự dựa trên cây quyết định (Chen & Guestrin, 2016), nhưng cũng làm tăng khả năng khớp sát dữ liệu huấn luyện nếu thông tin đầu vào bị giới hạn. Kết quả kiểm định chéo 5-fold trong nghiên cứu này xác nhận xu hướng đó: RF và XGBoost có độ khớp cao trên tập huấn luyện nhưng suy giảm rõ trên tập kiểm định, trong khi mô hình hồi quy lũy thừa có cấu trúc đơn giản hơn nhưng cho kết quả kiểm định ổn định hơn. Vì vậy, trong kịch bản CORVI đơn biến, mô hình hồi quy lũy thừa là lựa chọn hợp lý hơn xét đồng thời về độ chính xác kiểm định, tính ổn định, khả năng diễn giải và nguy cơ quá khớp.

Khi so sánh giữa các nghiên cứu, cần xem xét đồng thời cỡ mẫu thực địa, nguồn dữ liệu viễn

thám, phương pháp chọn biến và bối cảnh sinh thái của từng khu vực. Các nghiên cứu của Pham và cộng sự (2018, 2020a, 2020b) thường dựa trên dữ liệu đa nguồn, trong đó radar băng L từ ALOS-2 PALSAR-2 có khả năng phản ánh thông tin cấu trúc rừng khác với radar băng C của Sentinel-1. Huang và cộng sự (2022) sử dụng đồng thời Sentinel-1 và Sentinel-2 với thuật toán ML khi nghiên cứu tại Maowei Sea, nhưng điều kiện địa lý, thành phần loài, tuổi rừng và cấu trúc rừng có thể khác với hệ thống RNM miền Bắc Việt Nam. Trong khi đó, nghiên cứu này sử dụng bộ dữ liệu thực địa gồm 322 OTC phân bố tại RNM miền Bắc Việt Nam và tập trung vào mô hình CORVI đơn biến. Vì vậy, sự chênh lệch về độ chính xác giữa các công trình không nên được diễn giải đơn thuần là do thuật toán, mà cần gắn với sự khác biệt về dữ liệu đầu vào, phương pháp chọn biến, kiểu rừng, điều kiện lập địa và mức độ không đồng nhất không gian.

Đối với mô hình DL, kết quả nghiên cứu cũng cho thấy CNN chưa tạo ra cải thiện rõ rệt so với hồi quy lũy thừa trong kịch bản CORVI đơn biến. Ghosh và Behera (2021) cho thấy mô hình DL có thể vượt mô hình bán thực nghiệm trong ước tính AGB rừng ngập mặn nhiệt đới khi sử dụng dữ liệu Sentinel-1 SAR coherence và các biến phù hợp với cấu trúc rừng. Tuy nhiên, trong nghiên cứu hiện tại, đầu vào của CNN chỉ gồm một hoặc một số ít biến viễn thám, nên mô hình chưa có đủ thông tin để phát huy đầy đủ ưu thế học đặc trưng. Điều này cho thấy CNN có tiềm năng trong các mô hình đa biến hoặc đa kênh ảnh, nhưng cần được kết hợp với nhiều nguồn thông tin hơn như VV, VH, B11, B12, NDVI, NDMI, RVI, CORVI và các đặc trưng không gian để tăng hiệu quả dự báo.

Mặc dù mô hình hồi quy lũy thừa dựa trên CORVI đạt kết quả kiểm định ổn định trong kiểm định chéo 5-fold, giá trị R^2 trên tập kiểm định cho thấy CORVI đơn biến chỉ giải thích được khoảng 60% biến động AGB RNM. Điều này phản ánh rằng AGB là chỉ tiêu chịu tác động tổng hợp của nhiều yếu tố sinh thái và cấu trúc rừng, bao gồm tổ thành loài, tuổi rừng, mật

độ cây, chiều cao tán, đường kính thân, điều kiện lập địa, chế độ ngập triều và mức độ không đồng nhất không gian. Trong khi đó, CORVI chủ yếu tổng hợp thông tin tán xạ radar Sentinel-1 và thông tin SWIR Sentinel-2, nên chưa thể phản ánh đầy đủ toàn bộ biến thiên sinh khối trong điều kiện rừng ngập mặn ven biển phức tạp. Vì vậy, kết quả nghiên cứu cần được hiểu theo hướng CORVI là một biến tích hợp có giá trị và có khả năng làm mô hình nền, nhưng chưa đủ để thay thế các kịch bản đa biến trong ước tính AGB rừng ngập mặn ở quy mô rộng.

Ảnh hưởng của thủy triều được xem là một nguồn bất định quan trọng trong ước tính AGB rừng ngập mặn bằng dữ liệu Sentinel-1/Sentinel-2 tại miền Bắc Việt Nam (Huang và cộng sự, 2022, Kuenzer và cộng sự, 2011, Lagomasino và cộng sự, 2019). Khi mực nước triều thay đổi, tỷ lệ nền nước, bùn ướt và thảm thực vật trong pixel có thể thay đổi, làm biến động cả tín hiệu tán xạ ngược radar và phản xạ phổ quang học (Kuenzer và cộng sự, 2011, White và cộng sự, 2015). Đối với Sentinel-1, nước triều và nền bùn ướt có thể làm thay đổi cơ chế tán xạ bề mặt và tán xạ kép giữa thân cây, rễ chống và mặt nước (White và cộng sự, 2015). Đối với Sentinel-2, sự thay đổi mực nước và độ ẩm nền có thể ảnh hưởng đến các kênh SWIR Nam (Huang và cộng sự, 2022, Kuenzer và cộng sự, 2011), vốn là thành phần được sử dụng trong chỉ số CORVI. Vì vậy, mặc dù việc sử dụng ảnh tổng hợp trung vị giúp giảm một phần biến động tức thời, kết quả ước tính AGB vẫn có thể chịu ảnh hưởng của trạng thái thủy triều (Ghosh & Behera, 2021, Huang và cộng sự, 2022, Lagomasino và cộng sự, 2019). Các nghiên cứu tiếp theo cần bổ sung dữ liệu mực nước triều, lựa chọn ảnh theo cùng pha triều hoặc kiểm tra độ nhạy của CORVI theo điều kiện triều nhằm tăng độ ổn định và khả năng chuyển giao của mô hình (Kuenzer và cộng sự, 2011, White và cộng sự, 2015).

Một nguồn bất định khác của nghiên cứu là độ lệch thời gian giữa dữ liệu thực địa và dữ liệu ảnh vệ tinh. Bộ dữ liệu AGB gồm 322 OTC, trong đó 282 OTC được điều tra năm 2024 và

40 OTC được điều tra năm 2025, trong khi ảnh Sentinel-1/Sentinel-2 được tổng hợp trong giai đoạn tháng 10–12/2025. Đối với 40 OTC điều tra năm 2025, sự tương đồng thời gian giữa dữ liệu thực địa và ảnh vệ tinh tương đối tốt hơn. Tuy nhiên, đối với 282 OTC điều tra năm 2024, khoảng lệch thời gian có thể làm phát sinh sai số nếu khu vực ô mẫu có biến động về sinh trưởng, mật độ, tán cây, suy thoái, khai thác, bồi tụ/xói lở hoặc tác động của bão trong giai đoạn 2024–2025. Đây là dạng bất định thường được đề cập trong các nghiên cứu hiệu chỉnh mô hình sinh khối bằng dữ liệu thực địa và dữ liệu viễn thám, bên cạnh sai số đo đạc, sai số phương trình sinh khối và sai khác vị trí giữa ô mẫu với pixel ảnh (Kauffman & Donato, 2012, Réjou-Méchain và cộng sự, 2019) [18, 29]. Do AGB là chỉ tiêu phản ánh cấu trúc tích lũy của quần thụ, biến động trong một năm thường nhỏ hơn so với các chỉ số quang phổ tức thời; tuy nhiên, yếu tố này không thể bỏ qua trong môi trường rừng ngập mặn ven biển có động lực mạnh, nơi tín hiệu Sentinel còn chịu ảnh hưởng của thủy triều, nền bùn nước và điều kiện bề mặt (Huang và cộng sự, 2022, Kuenzer và cộng sự, 2011, Lagomasino và cộng sự, 2019, White và cộng sự, 2015). Vì vậy, kết quả mô hình cần được hiểu trong bối cảnh có độ lệch thời gian nhất định giữa dữ liệu thực địa và dữ liệu viễn thám. Các nghiên cứu tiếp theo nên ưu tiên thu nhận ảnh vệ tinh trùng hoặc gần trùng thời điểm điều tra thực địa, hoặc phân tích riêng các nhóm ô theo năm điều tra để đánh giá độ nhạy của mô hình đối với sai lệch thời gian.

Từ các nhận định trên, có thể nói đóng góp chính của nghiên cứu không nằm ở việc khẳng định các thuật toán học máy hoặc học sâu luôn vượt trội so với hồi quy truyền thống, mà ở việc chỉ ra giới hạn của các thuật toán phức tạp khi thông tin đầu vào bị giới hạn. Trong mô hình CORVI đơn biến, Random Forest và XGBoost đạt độ khớp cao trên tập huấn luyện nhưng suy giảm rõ trên tập kiểm định chéo, thể hiện nguy cơ quá khớp trong không gian biến một chiều. Ngược lại, mô hình hồi quy lũy thừa có cấu trúc đơn giản hơn nhưng cho kết quả kiểm định ổn

định, cho thấy quan hệ giữa CORVI và AGB có thể được mô tả hiệu quả bằng một dạng hàm phi tuyến đơn giản. Đây là phát hiện có ý nghĩa thực tiễn, vì trong nhiều trường hợp lập bản đồ sinh khối/các-bon RNM, mô hình dễ diễn giải, ổn định và ít yêu cầu biến đầu vào có thể phù hợp hơn so với mô hình phức tạp nhưng nhạy cảm với dữ liệu huấn luyện.

Để nâng cao khả năng dự báo thực tế, cần tiếp tục mở rộng tập biến đầu vào theo hướng kết hợp nhiều nhóm thông tin. Các biến này có thể bao gồm các kênh phổ Sentinel-2, chỉ số vật lý như NDVI, EVI, NDMI; các phân cực và tỷ số Radar như VV, VH, VH/VV; đặc trưng kết cấu ảnh; dữ liệu địa hình; khoảng cách đến bờ hoặc kênh rạch; thông tin thủy triều và các biến mô tả cấu trúc rừng hoặc điều kiện lập địa. Ngoài ra, cần áp dụng kiểm định chéo không gian hoặc kiểm định độc lập theo khu vực nhằm đánh giá khả năng chuyển giao của mô hình khi lập bản đồ sinh khối/trữ lượng các-bon RNM trên phạm vi lớn.

Kết luận

Nghiên cứu đã so sánh mô hình hồi quy CORVI dạng lũy thừa với các mô hình học máy và học sâu gồm RF, Cubist, SVM, XGBoost và CNN trong ước tính sinh khối trên mặt đất rừng ngập mặn tại miền Bắc Việt Nam. Khác với cách tiếp cận sử dụng nhiều biến đầu vào thường gặp trong các nghiên cứu trước, nghiên cứu này tập trung kiểm định khả năng của một chỉ số tích hợp quang học-radar khi được sử dụng với vai trò biến đầu vào chính, qua đó đánh giá độ chính xác kiểm định, tính ổn định, khả năng diễn giải và nguy cơ quá khớp của mô hình.

Kết quả kiểm định chéo 5-fold cho thấy, trong kịch bản CORVI đơn biến, mô hình hồi quy lũy thừa đạt hiệu quả kiểm định ổn định nhất hoặc tương đương nhóm mô hình tốt nhất, với $R^2 = 0,6056 \pm 0,0647$, $RMSE = 27,9263 \pm 1,2988$ Mg/ha và $MAE = 23,0128 \pm 1,4570$ Mg/ha. Các mô hình ML/DL không tạo ra cải thiện rõ rệt so với mô hình hồi quy truyền thống khi thông tin đầu vào chỉ giới hạn ở một biến CORVI. Đặc biệt, RF và XGBoost cho kết quả huấn luyện tốt nhưng suy giảm rõ khi kiểm định, cho thấy nguy cơ quá khớp khi áp dụng các thuật toán cây tổ hợp trong không gian biến một chiều.

Khi bổ sung NDVI vào CORVI, hiệu quả mô hình thay đổi nhưng hiệu quả không đồng đều giữa các thuật toán. NDVI có vai trò hỗ trợ nhất định, đặc biệt đối với một số mô hình cây tổ hợp, song chưa đủ để tạo ra sự cải thiện ổn định cho toàn bộ mô hình. Phân tích mức độ đóng góp biến cho thấy CORVI vẫn giữ vai trò nổi bật hơn NDVI trong RF và XGBoost, phản ánh giá trị của chỉ số này trong việc tích hợp thông tin tán xạ radar Sentinel-1 với thông tin SWIR Sentinel-2 để mô tả biến động AGB trong môi trường rừng ngập mặn ven biển.

Kết quả nghiên cứu chỉ ra rằng CORVI có tiềm năng làm biến nền trong ước tính AGB rừng ngập mặn, nhưng mô hình đơn biến mới giải thích được khoảng 60% biến động AGB. Vì vậy, mô hình hồi quy lũy thừa CORVI nên được xem là mô hình nền có cấu trúc đơn giản, dễ diễn giải và ổn định, thay vì mô hình tối ưu tuyệt đối cho mọi điều kiện. Các nghiên cứu tiếp theo cần mở rộng theo hướng đa biến hoặc đa kênh ảnh, kết hợp CORVI với các kênh phổ, chỉ số thực vật, biến radar, đặc trưng không gian, điều kiện thủy triều và thông tin lập địa nhằm nâng cao khả năng dự báo và chuyển giao mô hình phục vụ lập bản đồ sinh khối, ước tính trữ lượng các-bon và giám sát RNM mặn ven biển.

Lời cảm ơn: Đây là một phần kết quả nghiên cứu thuộc Luận án tiến sĩ “Nghiên cứu cơ sở khoa học và xây dựng mô hình ước tính trữ lượng các-bon RNM tại phía Bắc Việt Nam” của NCS Nguyễn Văn Thị. Tập thể tác giả xin được gửi lời cảm ơn đến Viện Sinh thái rừng và Môi trường đã cung cấp dữ liệu thực địa, Khoa QLTNR&MT, Trường Đại học Lâm nghiệp cùng quý đồng nghiệp đã tạo điều kiện và hỗ trợ chúng tôi hoàn thành nghiên cứu này.

Tuyên bố khả năng tiếp cận dữ liệu nghiên cứu: Dữ liệu hỗ trợ cho các kết quả của nghiên cứu này có thể được cung cấp bởi tác giả liên hệ theo yêu cầu hợp lý

Tuyên bố về đóng góp của tác giả: Nguyễn Văn Thị: Xây dựng mô hình và trình bày phân kết quả. Kết luận; Nguyễn Hải Hòa: Đặt vấn đề, thảo luận và hiệu chỉnh bản thảo; Đình Bá Duy: Vật liệu và phương pháp nghiên cứu; Ngô Thành Đạt: Chuẩn bị bản thảo gốc.

Tuyên bố về việc sử dụng Generative AI (GenAI): Nhóm tác giả xin khẳng định rằng không sử dụng bất kỳ công cụ GenAI nào để tạo dữ liệu, xử lý số liệu, phân tích kết quả hoặc xây dựng các kết luận khoa học trong nghiên cứu này. Toàn bộ dữ liệu, kết quả phân tích và nội dung khoa học của bài báo được thực hiện trực tiếp bởi các tác giả và chịu trách nhiệm hoàn toàn về tính chính xác, trung thực học thuật của công trình.

Tuyên bố về xung đột lợi ích: Các tác giả không có xung đột về lợi ích.

Tài liệu tham khảo

- Alongi, D. M. (2002). Present state and future of the world's mangrove forests. *Environmental Conservation*, 29(3), 331–349. <https://doi.org/10.1017/S0376892902000231>
- Alongi, D. M. (2014). Carbon cycling and storage in mangrove forests. *Annual Review of Marine Science*, 6, 195–219. <https://doi.org/10.1146/annurev-marine-010213-135020>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
- Castillo, J. A. A., Apan, A. A., Maraseni, T. N., & Salmo, S. G. (2017). Estimation and mapping of above-ground biomass of mangrove forests and their replacement land uses in the Philippines using Sentinel imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 134, 70–85. <https://doi.org/10.1016/j.isprsjprs.2017.10.016>
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. <https://doi.org/10.1145/2939672.2939785>
- Clough, B. F., & Scott, K. (1989). Allometric relationships for estimating above-ground biomass in six mangrove species. *Forest Ecology and Management*, 27, 117–127. [https://doi.org/10.1016/0378-1127\(89\)90034-0](https://doi.org/10.1016/0378-1127(89)90034-0)
- Congalton, R. G., & Green, K. (2019). *Assessing the accuracy of remotely sensed data: Principles and practices* (3rd ed.). CRC Press. <https://doi.org/10.1201/9780429052729>
- 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/ngeo1123>
- Duke, N. C., Meynecke, J.-O., Dittmann, S., Ellison, A. M., Anger, K., Berger, U., Cannicci, S., Diele, K., Ewel, K. C., Field, C. D., Koedam, N., Lee, S. Y., Marchand, C., Nordhaus, I., & Dahdouh-Guebas, F. (2007). A world without mangroves? *Science*, 317(5834), 41–42. <https://doi.org/10.1126/science.317.5834.41b>
- Farzanmanesh, R., Khoshelham, K., Volkova, L., Thomas, S., Ravelonjatovo, J., & Weston, C. J. (2024). Quantifying mangrove aboveground biomass changes: Analysis of conservation impact in blue forests projects using Sentinel-2 satellite imagery. *Forest Ecology and Management*, 561(4):121920. <https://doi.org/10.1016/j.foreco.2024.121920>
- Fatoyinbo, T. E., & Simard, M. (2013). Height and biomass of mangroves in Africa from ICESat/GLAS and SRTM. *International Journal of Remote Sensing*, 34(2), 668–681. <https://doi.org/10.1080/01431161.2012.712224>
- Foody, G. M. (2002). Status of land cover classification accuracy assessment. *Remote Sensing of Environment*, 80(1), 185–201. [https://doi.org/10.1016/S0034-4257\(01\)00295-4](https://doi.org/10.1016/S0034-4257(01)00295-4)
- Fu, W., & Wu, Y. (2011). Estimation of aboveground biomass of different mangrove trees based on canopy diameter and tree height. *Procedia Environmental Sciences*, 10, 2189–2194. <https://doi.org/10.1016/j.proenv.2011.09.343>
- Ghosh, S. M., & Behera, M. D. (2021). Aboveground biomass estimates of tropical mangrove forest using Sentinel-1 SAR coherence data: The superiority of deep learning over semi-empirical model. *Computers & Geosciences*, 150, 104737. <https://doi.org/10.1016/j.cageo.2021.104737>
- Giri, C., Ochieng, E., Tieszen, L. L., Zhu, Z., Singh, A., Loveland, T., Masek, J., & Duke, N. (2011). Status and distribution of mangrove forests of the world using earth observation satellite data. *Global Ecology and Biogeography*, 20(1), 154–159. <https://doi.org/10.1111/j.1466-8238.2010.00584.x>
- Google Earth Engine. (n.d.). *Sentinel-1 algorithms*. Google Developers. <https://developers.google.com/earth-engine/guides/sentinel1>
- Huang, Z., Tian, Y., Zhang, Q., Huang, Y., Liu, R., Huang, H., Zhou, G., Wang, J., Tao, J., Yang, Y., Zhang, Y., Lin, J., Tan, Y., Deng, J., & Liu, H. (2022). Estimating mangrove above-ground biomass at Maowei Sea, Beibu Gulf of China using machine learning algorithm with Sentinel-1 and Sentinel-2 data. *Geocarto International*, 37(27), 15778–15805. <https://doi.org/10.1080/10106049.2022.2102226>
- Kauffman, J. B., & Donato, D. C. (2012). *Protocols for the measurement, monitoring and reporting of structure, biomass and carbon stocks in mangrove forests* (CIFOR Working Paper No. 86. Center for International Forestry Research. <https://doi.org/10.17528/cifor/003749>
- Komiyama, A., Ong, J. E., & Pongparn, S. (2005). Common allometric equations for estimating the tree weight of mangroves. *Journal of Tropical Ecology*, 21(4), 471–477. <https://doi.org/10.1017/S0266467405002476>
- Kuenzer, C., Bluemel, A., Gebhardt, S., Quoc, T. V., & Dech, S. (2011). Remote sensing of mangrove ecosystems: A review. *Remote Sensing*, 3(5), 878–928. <https://doi.org/10.3390/rs3050878>
- Lagomasino, D., Fatoyinbo, T. E., Lee, S. K., Feliciano, E., Trettin, C., Shapiro, A., & Mangora, M. M. (2019). Measuring mangrove carbon loss and gain in deltas. *Environmental Research Letters*, 14, 025002. <https://doi.org/10.1088/1748-9326/aaf0de>
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521, 436–444. <https://doi.org/10.1038/nature14539>
- Liu, Z., Yin, Z., Zhao, W., Feng, Z., Pei, H., Grimaldi, P., & Qiu, Z. (2026). Spatiotemporal patterns of aboveground carbon storage in Hainan mangroves based on machine learning and multi-source remote sensing data. *Forests*, 17(1), 131. <https://doi.org/10.3390/f17010131>

- Murdiyarso, D., Purbopuspito, J., Kauffman, J. B., Warren, M. W., Sasmito, S. D., Donato, D. C., Manuri, S., Krisnawati, H., Taberima, S. & Kurnianto, S. (2015). The potential of Indonesian mangrove forests for global climate change mitigation. *Nature Climate Change*, 5, 1089–1092. <https://doi.org/10.1038/nclimate2734>
- Nasirzadehdizaji, R., Sanli, F. B., Abdikan, S., Cakir, Z., Sekertekin, A., & Ustuner, M. (2019). Sensitivity analysis of multi-temporal Sentinel-1 SAR parameters to crop height and canopy coverage. *Applied Sciences*, 9(4), 655. <https://doi.org/10.3390/app9040655>
- Pham, T. D., Le, N. N., Ha, N. T., Nguyen, L. V., Xia, J., Yokoya, N., To, T. T., Trinh, H. X., Kieu, L. Q., & Takeuchi, W. (2020b). Estimating mangrove above-ground biomass using extreme gradient boosting decision trees algorithm with fused Sentinel-2 and AL OS-2 PALSAR-2 data in Can Gio Biosphere Reserve, Vietnam. *Remote Sensing*, 12(5), 777. <https://doi.org/10.3390/rs12050777>
- Pham, T. D., Yokoya, N., Xia, J., Ha, N. T., Le, N. N., Nguyen, T. T. T., Dao, T. H., Vu, T. T. P., Pham, T. D., & Takeuchi, W. (2020a). 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>
- Pham, T. D., Yoshino, K., Le, N. N., & Bui, D. T. (2018). Estimating aboveground biomass of a mangrove plantation on the Northern coast of Vietnam using machine learning techniques with an integration of ALOS-2 PALSAR-2 and Sentinel-2A data. *International Journal of Remote Sensing*, 39(22), 7761–7788. <https://doi.org/10.1080/01431161.2018.1471544>
- Réjou-Méchain, M., Barbier, N., Couteron, P., Ploton, P., Vincent, G., Herold, M., Mermoz, S., Saatchi, S., Chave, J., de Boissieu, F., Féret, J.-B., Takoudjou, S. M., & Pélissier, R. (2019). Upscaling forest biomass from field to satellite measurements: Sources of errors and ways to reduce them. *Surveys in Geophysics*, 40, 881–911. <https://doi.org/10.1007/s10712-019-09532-0>
- Rouse, J. W., Haas, R. H., Schell, J. A., & Deering, D. W. (1974). Monitoring vegetation systems in the Great Plains with ERTS. In *Third Earth Resources Technology Satellite-1 Symposium* (NASA SP-351, Vol. 1, pp. 309–317). NASA.
- Rovai, A. S., Twilley, R. R., Castañeda-Moya, E., Riul, P., Cifuentes-Jara, M., Manrow-Villalobos, M., Horta, P. A., Simonassi, J. C., Fonseca, A. L., & Pagliosa, P. R. (2018). Global controls on carbon storage in mangrove soils. *Nature Climate Change*, 8, 534–538. <https://doi.org/10.1038/s41558-018-0162-5>
- Simard, M., Fatoyinbo, L., Smetanka, C., Rivera-Monroy, V. H., Castaneda-Moya, E., Thomas, N., & Van der Stocken, T. (2019). Mangrove canopy height globally related to precipitation, temperature and cyclone frequency. *Nature Geoscience*, 12, 40–45. <https://doi.org/10.1038/s41561-018-0279-1>
- Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, 14, 199–222. <https://doi.org/10.1023/B:STCO.0000035301.49549.88>
- White, L., Brisco, B., Dabboor, M., Schmitt, A., & Pratt, A. (2015). A collection of SAR methodologies for monitoring wetlands. *Remote Sensing*, 7(6), 7615–7645. <https://doi.org/10.3390/rs70607615>
- Xue, Y., et al. (2026). Inter-annual estimation of mangrove aboveground biomass density using UAV-LiDAR data and Sentinel-1/2 time series imagery. *GIScience and Remote sensing*, 63(1), 2654180. <https://doi.org/10.1080/15481603.2026.2654180>

Abstract

Comparing CORVI-based Regression with Machine Learning And Deep Learning Algorithms for Estimating Mangrove Above-Ground Biomass using Sentinel-1 and Sentinel-2 imageries in Northern Vietnam

Mangrove forests play an important role in shoreline protection, biodiversity conservation, and blue carbon accumulation. Above-ground biomass (AGB) is a key indicator for assessing forest structure and estimating carbon stocks; however, field-based surveys are often time-consuming, costly, and difficult to scale up over large areas. This study compares a CORVI-based power regression model with machine learning and deep learning algorithms, including Random Forest (RF), Cubist, Support Vector Machine (SVM), eXtreme Gradient Boosting (XGBoost), and Convolutional Neural Network (CNN), for estimating mangrove AGB in Northern Vietnam. Field data were collected from 322 sample plots distributed across Quang Ninh, Hai Phong, Hung Yen, and Ninh Binh; AGB was calculated from field-measured variables by species or species groups and converted to Mg/ha. Input variables were extracted from Sentinel-1, Sentinel-2, and the integrated optical–radar index CORVI. Model accuracy was assessed using 5-fold cross-validation based on R^2 , RMSE, MAE, and Bias. The results showed that, in the univariate CORVI scenario, the power regression model achieved the most stable validation performance or was comparable to the best-performing models, with $R^2 = 0.6056 \pm 0.0647$, $RMSE = 27.9263 \pm 1.2988$ Mg/ha, and $MAE = 23.0128 \pm 1.4570$ Mg/ha. RF and XGBoost showed a risk of overfitting when input information was limited. The study confirms that CORVI is a valuable integrated variable for modelling mangrove AGB, but it should be combined with optical, radar, spatial, and site-condition variables to improve predictive performance over larger areas.

Keywords: blue carbon; cross-validation; deep learning; machine learning; optical–radar integration; power regression model; radar vegetation index