



Original Article

# Apply Machine Learning to Predict Saltwater Intrusion in the Ham Luong River, Ben Tre Province

Pham Ngoc Hoai<sup>1</sup>, Pham Bao Quoc<sup>1</sup>, Tran Thanh Thai<sup>2,\*</sup>

<sup>1</sup>*Institute of Applied Technology, Thu Dau Mot University,*

*6 Tran Van On, Phu Hoa, Thu Dau Mot City, Binh Duong, Vietnam*

<sup>2</sup>*Institute of Tropical Biology, Vietnam Academy of Science and Technology,*

*85 Tran Quoc Toan, Vo Thi Sau, District 3, Ho Chi Minh City, Vietnam*

Received 08 October 2021

Revised 15 December 2021; Accepted 14 February 2022

**Abstract:** Saltwater intrusion is a major problem particularly in the Mekong Delta, Việt Nam. In order to better manage the salinity problem, it is important to be able to predict the saltwater intrusion in rivers. The objective of this research is to apply several machine learning algorithms, including Multiple Linear Regression (MLR), Random Forest Regression (RFR), Artificial Neural Networks (ANN) for predicting the saltwater intrusion in Ham Luong River, Ben Tre Province. The input data is composed of 207 weekly saltwater intrusion data points from 2012 to 2020. Yearly salinity was measured during the 23 weeks of the dry season, from January to June. The Nash - Sutcliffe efficiency coefficient (NSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) are used to evaluate the performances of machine learning algorithms. The research results indicated that the ANN model achieved a high performance for salinity forecasting with NSE = 0.907, RMSE = 0.11, MAE = 0.08 for training period, NSE = 0.842, RMSE = 1.16, MAE = 0.11 for testing period. The findings of this study suggest that the ANN algorithm is a promising tool to forecast salinity in Ham Luong River.

**Keywords:** Artificial intelligence, climate change, Mekong Delta, saltwater intrusion.

\* Corresponding author.

*E-mail address:* [thanhthai.bentrect@gmail.com](mailto:thanhthai.bentrect@gmail.com)

<https://doi.org/10.25073/2588-1094/vnuees.4852>

# Áp dụng thuật toán học máy để dự báo độ mặn trên sông Hàm Luông, tỉnh Bến Tre

Phạm Ngọc Hoài<sup>1</sup>, Phạm Bảo Quốc<sup>1</sup>, Trần Thành Thái<sup>2,\*</sup>

<sup>1</sup>*Viện Công nghệ Ứng dụng, Trường Đại học Thủ Dầu Một,*

*6 Trần Văn Ôn, Phú Hòa, Thành phố Thủ Dầu Một, Bình Dương, Việt Nam*

<sup>2</sup>*Viện Sinh học Nhiệt đới, Viện Hàn lâm Khoa học và Công nghệ Việt Nam,*

*85 Trần Quốc Toản, Võ Thị Sáu, Quận 3, Thành phố Hồ Chí Minh, Việt Nam*

Nhận ngày 08 tháng 10 năm 2021

Chỉnh sửa ngày 15 tháng 12 năm 2021; Chấp nhận đăng ngày 14 tháng 12 năm 2022

**Tóm tắt:** Xâm nhập mặn (XNM) là vấn đề rất đáng lưu tâm ở vùng đồng bằng sông Cửu Long (ĐBSCL). Để chủ động trong công tác quản lý nguồn nước ngọt và giảm thiểu tác động của xâm nhập mặn, dự báo chính xác độ mặn trên sông được xem là một trong những giải pháp. Từ đây, mục tiêu của nghiên cứu là đánh giá khả năng áp dụng một số thuật toán học máy, bao gồm hồi quy đa biến (Multiple Linear Regression, MLR), rừng ngẫu nhiên (Random Forest Regression, RFR), mạng nơ-ron nhân tạo (Artificial Neural Networks, ANN) trong dự báo độ mặn trên sông Hàm Luông, tỉnh Bến Tre. Dữ liệu độ mặn sử dụng trong nghiên cứu được thu thập theo tuần, từ năm 2012 đến 2020. Mỗi năm đo đạc trong 23 tuần mùa khô, từ tháng 1 đến tháng 6. Các chỉ số thống kê như Hệ số Nash - Sutcliffe efficiency (NSE), Sai số bình phương trung bình (Root Mean Squared Error, RMSE), và Sai số tuyệt đối trung bình (Mean Absolute Error, MAE), được sử dụng để đánh giá tính chính xác của thuật toán dự báo. Kết quả cho thấy thuật toán mạng nơ-ron nhân tạo dự báo độ mặn tốt nhất trong 3 thuật toán, với NSE = 0,907, RMSE = 0,11, MAE = 0,08 cho tập huấn luyện, NSE = 0,842, RMSE = 1,16, MAE = 0,11 cho tập kiểm tra. Thuật toán mạng nơ-ron nhân tạo hiệu quả trong dự báo mặn trên sông Hàm Luông, tỉnh Bến Tre.

*Từ khóa:* Biến đổi khí hậu, ĐBSCL, trí thông minh nhân tạo, XNM.

## 1. Mở đầu

ĐBSCL nằm ở vùng hạ lưu sông Mê Kông, từ biên giới Việt Nam - Campuchia đến Biên Đông, đây là vùng đồng bằng rộng lớn, màu mỡ lớn thứ ba trên thế giới với 3,9 triệu hecta [1]. ĐBSCL là nơi sinh sống của hơn 18 triệu dân Việt Nam (chiếm hơn 22% dân số cả nước), vùng đồng bằng sản xuất hơn 50% lượng lương thực phẩm và đóng góp vào hơn 85% lượng lúa gạo cho cả nước [2]. Do đặc điểm địa hình trũng

thấp với độ cao trung bình chỉ khoảng 0,8 m trên bề mặt nước biển, ĐBSCL là khu vực chịu tác động rất mạnh của biến đổi khí hậu và đặc biệt là hiện tượng nước biển dâng [2]. Với điều kiện đó, nền sản xuất nông nghiệp của vùng ĐBSCL phải đối mặt với thách thức rất lớn từ các thiên tai như khô hạn và XNM [1, 2]. Mặc dù XNM là hiện tượng thường xuyên của ĐBSCL vào mùa khô; tuy nhiên trong vài năm trở lại đây, hiện tượng này đã trở nên nghiêm trọng do mặn xâm nhập sâu, kéo dài và độ mặn cao [3-5].

\* Tác giả liên hệ.

Địa chỉ email: thanhthai.bentrect@gmail.com

<https://doi.org/10.25073/2588-1094/vnuees.4852>

XNM là một trong những vấn đề chính của quản lý nguồn nước vùng cửa sông ven biển [6, 7]. XNM làm giảm khả năng lọc và gia tăng các loại độc tố trong đất, dẫn đến năng suất cây trồng thấp [8]. Hơn nữa, độ mặn cao làm cây trồng mất nhiều năng lượng để hút nước từ đất làm cây trồng chậm phát triển [8]. Ở ĐBSCL, XNM là một vấn đề sinh thái - xã hội cần được nghiên cứu và giải quyết, vấn đề này trở nên rất nghiêm trọng trong điều kiện biến đổi khí hậu hiện nay [4]. Chính trên tổng số mười ba tỉnh vùng ĐBSCL đang chịu ảnh hưởng từ XNM [4, 7]. Từ đây, hàng nghìn hecta hoa màu, cây ăn trái, lúa gạo, nuôi trồng thủy sản bị tác động [7].

Nghiên cứu cho thấy XNM chịu ảnh hưởng từ nhiều yếu tố như: chế độ triều, chế độ dòng chảy - thủy văn, cấu trúc địa hình lòng sông, gió, nhiệt độ, ... [4]. Để phục vụ việc cảnh báo sớm XNM cũng như quản lý tốt nguồn nước ngọt, nhiều nghiên cứu đã cố gắng đưa ra các dự báo về XNM. Hiện tại, mô hình tiến trình (process-based models) được sử dụng phổ biến, đây là loại mô hình kết hợp toán - vật lý để đưa ra dự báo. Các thuật toán này dự báo và mô tả rất chính xác các quy luật thủy văn (ví dụ XNM) nhờ các quy luật vật lý được nghiên cứu và tích hợp sẵn trong thuật toán. Tuy nhiên, cần có những chuyên gia để khai thác được những thuật toán này vì chúng vận hành rất phức tạp. Hơn nữa, số lượng đầu vào, là dữ liệu của các yếu tố ảnh hưởng đến XNM, phải rất lớn mới đảm bảo tính chính xác [9, 10]. Một cách tiếp cận khác là sử dụng các thuật toán máy học (machine learning) trong dự báo mặn. Phương pháp này có ưu điểm là dễ áp dụng, độ chính xác cao, không đòi hỏi số lượng dữ liệu lớn. Thực tế cho thấy thuật toán học máy đã được sử dụng rộng rãi trong các nghiên cứu dự báo thủy văn như chất lượng nước [11], mực nước [12, 13]. Tác giả Lin và cộng sự [10] đã sử dụng thuật toán Random Forest để dự báo mặn vùng cửa sông Modaomen, đồng bằng Pearl River, Trung Quốc, kết quả cho thấy độ chính xác rất cao, lên đến 91%. Thu thập thông tin về toàn bộ các yếu tố ảnh hưởng đến độ mặn là vô cùng khó khăn và thường không đầy đủ [9]. Cho

nên, lựa chọn các thuật toán học máy để dự báo mặn trong trường hợp này là phù hợp. Tuy nhiên, hiện tại, có rất ít nghiên cứu về dự báo mặn ở ĐBSCL sử dụng các thuật toán máy học.

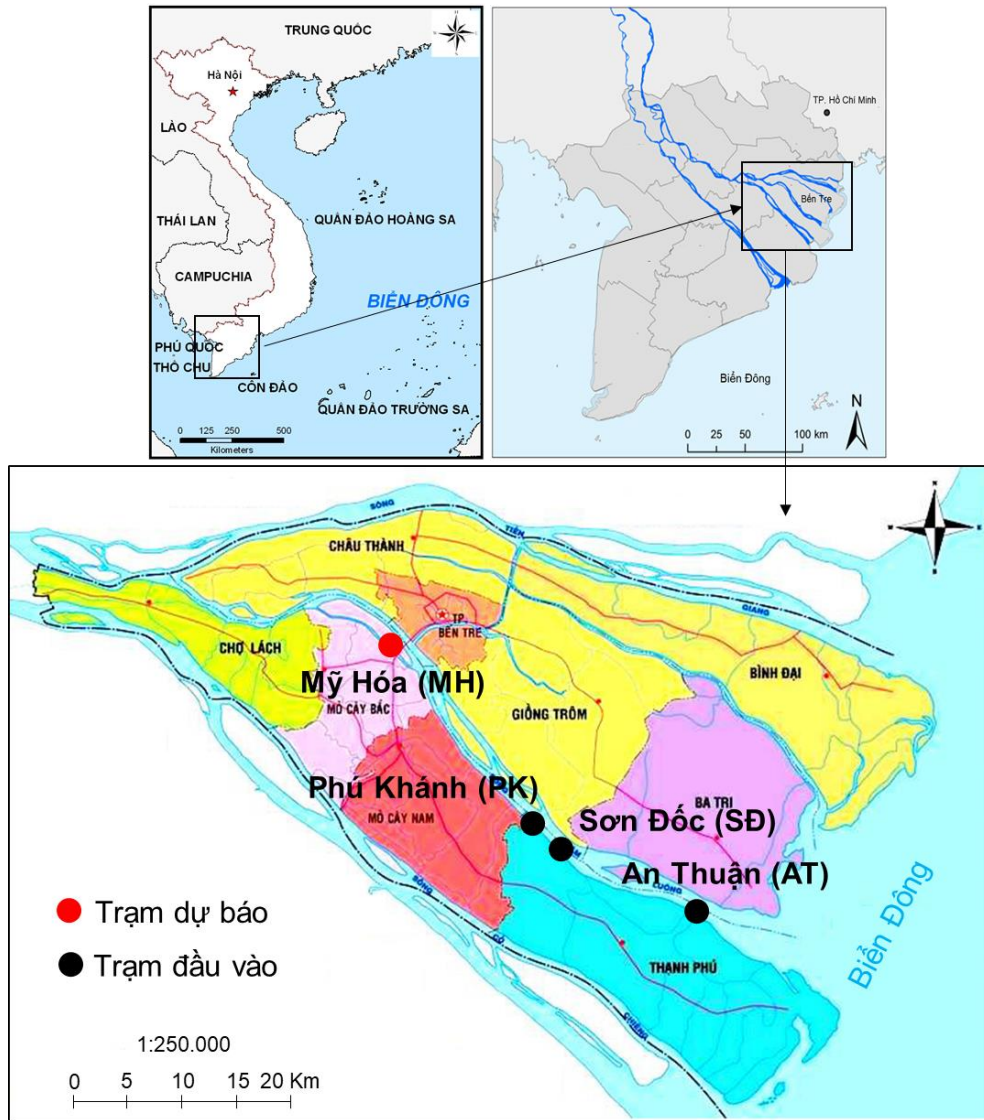
Một nhược điểm của các thuật toán máy học là chúng thuộc nhóm "black - box" nên đôi khi rất khó để giải thích kết quả [9, 14, 15]. Hiện tại, các thuật toán học máy như hồi quy đa biến (Multiple Linear Regression, MLR), rừng ngẫu nhiên (Random Forest Regression, RFR), mạng nơ-ron nhân tạo (Artificial Neural Networks, ANN) được sử dụng rộng rãi trong các nghiên cứu phân tích, dự báo, và bước đầu cho kết quả khá khả quan [16-18].

Do đó, nghiên cứu được thực hiện với mục tiêu đánh giá khả năng của thuật toán MLR, RFR, và ANN trong dự báo mặn ở sông Hàm Luông, tỉnh Bến Tre. Đây là một trong những nhánh sông lớn của hệ thống sông Mê Kông và đang bị mặn xâm nhập sâu, từ đó ảnh hưởng đến sinh hoạt và sản xuất của người dân trong vùng. Kết quả từ nghiên cứu có thể cung cấp thêm một cách tiếp cận đơn giản, hữu hiệu trong quản lý tài nguyên nước và giảm thiểu tác động của XNM.

## 2. Phương pháp nghiên cứu

### 2.1. Khu vực nghiên cứu

Sông Hàm Luông là một trong 4 nhánh sông thuộc sông Tiền, sông chảy trọn vẹn trong địa phận tỉnh Bến Tre. Sông có chiều dài, rộng, và sâu lần lượt là 70 km, 1.200 - 1.500 m, và 12 - 16 m [19]. Sông Hàm Luông đóng vai trò quan trọng trong cung cấp nguồn nước cho sinh hoạt, phát triển công - nông nghiệp, và các hoạt động kinh tế khác như vận tải đường sông, du lịch [19]. Có 4 trạm quan trắc mặn trên sông Hàm Luông, lần lượt từ cửa sông lên thượng nguồn là: An Thuận (AT), Sơn Đốc (SĐ), Phú Khánh (PK), Mỹ Hóa (MH) (Hình 1). Hiện tại, mặn xâm nhập sâu vào sông Hàm Luông hơn so với các sông khác như Mỹ Tho, Cổ Chiên [20].



Hình 1. Bản đồ vị trí các trạm quan trắc mặn trên sông Hàm Luông, tỉnh Bến Tre.

Bảng 1. Thống kê mô tả bộ dữ liệu về độ mặn tại các trạm quan trắc từ năm 2012 đến 2020

| Đặc điểm dữ liệu         | An Thuận | Sơn Đốc | Phú Khánh | Mỹ Hóa |
|--------------------------|----------|---------|-----------|--------|
| Số dữ liệu (Count)       | 207      | 207     | 207       | 207    |
| Trung bình (Mean, PSU)   | 21,58    | 10,06   | 7,37      | 3,15   |
| Độ lệch chuẩn (Std, PSU) | 4,67     | 6,51    | 5,51      | 4,36   |
| Cực tiểu (Min, PSU)      | 11,10    | 0,10    | 0,10      | 0,10   |
| 25% (PSU)                | 18,00    | 5,05    | 4,00      | 0,30   |
| 50% (PSU)                | 21,60    | 8,60    | 5,90      | 1,20   |
| 75% (PSU)                | 25,45    | 13,10   | 8,60      | 4,15   |
| Cực đại (Max, PSU)       | 31,50    | 28,20   | 26,70     | 17,20  |

## 2.2. Thu thập và tiền xử lý dữ liệu

Dữ liệu độ mặn (PSU) từ năm 2012 đến 2020 tại 4 trạm quan trên sông Hàm Luông được thu thập từ Đài khí tượng thủy văn tỉnh Bến Tre (<https://bentre.gov.vn/thong-tin-can-biet/du-bao-do-man>). Ở các trạm, độ mặn được đo theo tuần, và chỉ đo trong 23 tuần của mùa khô (từ tháng 1 đến tháng 6). Bảng 1 mô tả thống kê bộ dữ liệu độ mặn về số lượng dữ liệu, trung bình, độ lệch chuẩn, giá trị nhỏ - lớn nhất, điểm phân vị thứ 25, 50, và 75.

Chất lượng dữ liệu (số lượng đủ lớn, liên tục, ít giá trị ngoại lai,...) quyết định đến tính chính xác của các thuật toán dự báo [21]. Cho nên, dữ liệu được tiền xử lý qua ba bước trước khi được đưa vào thuật toán để huấn luyện thuật toán:

i) Loại bỏ các giá trị Null, đồng thời thay thế các giá trị đó bằng giá trị nội suy (theo phương pháp Linear Interpolation trong thư viện Pandas của Python);

ii) Giá trị ngoại lai khác thường trong bộ số liệu cần được kiểm tra lại, nếu đó là giá trị lỗi thì thay thế bằng trung bình của 4 giá trị gần đó [22]. Dữ liệu được mô tả ở 5 vị trí: giá trị nhỏ nhất (min), tứ phân vị thứ nhất (Q1), trung vị (median), tứ phân vị thứ 3 (Q3) và giá trị lớn nhất (max) của biểu đồ hộp. Giá trị ngoại lai là giá trị nằm ngoài giới hạn trên ( $Q3 + 1,5 * \text{Độ trải giữa (IQR, Interquartile Range)}$ ) và giới hạn dưới ( $Q1 - 1,5 * \text{IQR}$ ) của biểu đồ hộp [23];

iii) Tất cả các số liệu được chuẩn hóa dạng Logarit hóa.

Tương quan về độ mặn giữa các trạm quan trắc được đánh giá bằng tương quan hạng Spearman. Ngoài ra, mức độ ảnh hưởng (về phương sai) của các biến độc lập lên biến phụ thuộc được xác định bằng phương pháp Extra Trees Classifier (Extremely Randomized Trees Classifier) trong thư viện scikit - learn của Python [24].

## 2.3. Thuật toán học máy

*Thuật toán hồi quy đa biến (Multiple Linear Regression, MLR)*

Thuật toán hồi quy đa biến thể hiện mối liên hệ giữa các biến độc lập  $x(x_1, x_2, \dots, x_n)$  và biến

phụ thuộc  $y$ . Phương trình toán học của MLR được thể hiện như sau [25]:

$$y = f(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \varepsilon$$

Trong đó,  $\beta_0$  là hệ số chặn (intercept),  $\beta_n$  hệ số hồi quy (regression parameters) cho  $n$  biến độc lập,  $\varepsilon$  là sai số (error term).

*Thuật toán rừng ngẫu nhiên (Random Forest Regression, RFR)*

RFR là một thuật toán học có giám sát, được phát triển bởi Breiman vào năm 2001 [26]. RFR là thuật toán phi tham số (nonparametric) cho nên dữ liệu không cần tuân theo quy luật phân bố chuẩn [27]. Bản chất RFR dùng kỹ thuật bagging, lựa chọn một nhóm nhỏ các thuộc tính tại mỗi nút của cây phân lớp để phân chia thành các mức tiếp theo.

Thuật toán RFR được tiến hành như sau: i) Chọn số lượng các cây thành phần sẽ được xây dựng (ntree); ii) Chọn số lượng các thuộc tính sẽ được dùng tại mỗi node của cây (mtry); iii) Dùng các cây quyết định bằng cách hoán vị các tập mẫu khởi động (bootstrap). Khi xây dựng cây, tại mỗi node sẽ chọn ra  $mtry$  thuộc tính, và sử dụng các thuộc tính này để tìm ra cách phân chia tốt nhất; và iv) Thu thập kết quả phân tích trên tất cả các cây quyết định và sử dụng kết quả được chọn nhiều nhất làm kết quả cuối cùng của thuật toán [28].

Khoảng 2/3 các phần tử trong tập huấn luyện tham gia vào trong các tính toán và 1/3 các phần tử còn lại, được gọi là dữ liệu out-of-bag, để ước tính lỗi dự báo và tầm quan trọng của biến [26].

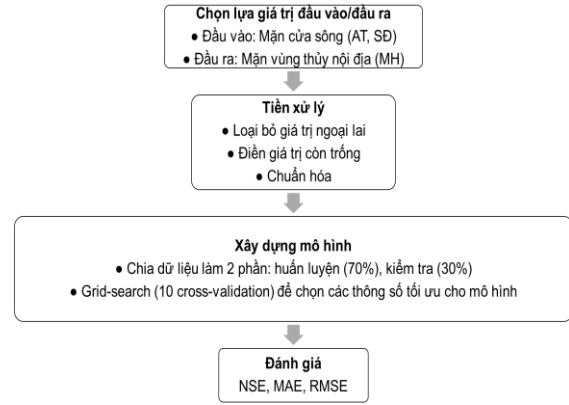
*Thuật toán mạng nơ-ron nhân tạo (Artificial Neural Networks, ANN)*

ANN là thuật toán mô phỏng hoạt động của các tế bào thần kinh trong não người, bao gồm một mạng lưới các nơ-ron được liên kết để xử lý thông tin. Cấu trúc của một mạng ANN gồm ba thành phần: Lớp đầu vào (input layer), các lớp ẩn (hidden layer, một ANN có thể có nhiều lớp ẩn), và lớp đầu ra (output layer). Trong đó, các lớp ẩn gồm các nơ-ron nhận dữ liệu vào (inputs) xử lý chúng và cho ra một kết quả (output) duy nhất. Kết quả xử lý của một nơ-ron có thể làm input cho các nơ-ron khác.

2.4. Xây dựng và đánh giá thuật toán

Để dự báo độ mặn ở thượng nguồn, độ mặn tại các trạm An Thuận, Sơn Đốc được dùng làm đầu vào (input, hay là các biến độc lập), độ mặn ở trạm Mỹ Hóa làm đầu ra (output, hay là biến phụ thuộc). Toàn bộ dữ liệu được chia làm 2 phần: 70% cho tập huấn luyện (training), 30% cho tập kiểm tra (testing). Phương pháp Cross Validation (CV) được áp dụng để hạn chế overfitting trong huấn luyện thuật toán. CV là phương pháp chia nhỏ tập training ra thành nhiều phần (n phần). Với mỗi lần huấn luyện, thuật toán sẽ sử dụng n-1 phần cho huấn luyện, sau đó kiểm tra dựa trên 1 phần còn lại, điều này sẽ giúp cho thuật toán hạn chế gặp phải overfitting. Nghiên cứu sử dụng n = 10, đây là giá trị phổ biến trong huấn luyện thuật toán máy học [29]. Các thuật toán được thực thi trong môi trường

Python (<https://www.python.org/>) thông qua 5 bước (Hình 2). Các thông số tối ưu ở từng thuật toán được lựa chọn bằng phương pháp Grid-search qua 10 cross-validation (Bảng 2).



Hình 2. Quá trình xây dựng thuật toán dự đoán mặn.

Bảng 2. Các thông số được sử dụng trong Grid-search để chọn thông số tối ưu

| Thuật toán | Hyper - parameters   |
|------------|--|
| RFR        | 'n_estimators': [10, 15, 20, 25, 30, 50, 70, 100, 200], 'max_depth': [2, 4, 6, 8]      |
| ANN        | Loss: [MSE, MAE], optimizer: Adam, epochs: [50, 100, 200, 300], batch size: [1, 5, 10] |

Ba chỉ số thống kê là Hệ số xác định NSE (Nash - Sutcliffe efficiency coefficient), Sai số bình phương trung bình (Root Mean Squared Error, RMSE), và Sai số tuyệt đối trung bình (Mean Absolute Error, MAE), được sử dụng để đánh giá tính chính xác của thuật toán dự báo. NSE phản ánh mức độ giải thích của các biến độc lập đối với các biến phụ thuộc, giá trị NSE càng cao thì thuật toán càng tốt (NSE dao động từ -∞ đến 1) [30]. Cả MAE và RMSE đều thể đo sự khác biệt giữa các giá trị dự đoán và giá trị thực tế, chúng nằm trong khoảng từ 0 đến ∞ và giá trị càng thấp thì thuật toán sẽ càng tốt hơn.

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

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

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i|$$

Trong đó, n là số mẫu,  $\hat{y}_i$ ,  $y_i$ ,  $\bar{y}$  tương ứng là giá trị dự báo, giá trị thực, trung bình giá trị thực.

Theo Moriasi và cộng sự [31], hiệu quả của thuật toán máy học được đánh giá qua chỉ số NSE như sau: Rất tốt ( $NSE > 0,80$ ), tốt ( $0,7 < NSE \leq 0,8$ ), chấp nhận được ( $0,50 < NSE \leq 0,70$ ), hoặc không chấp nhận được ( $NSE \leq 0,50$ ).

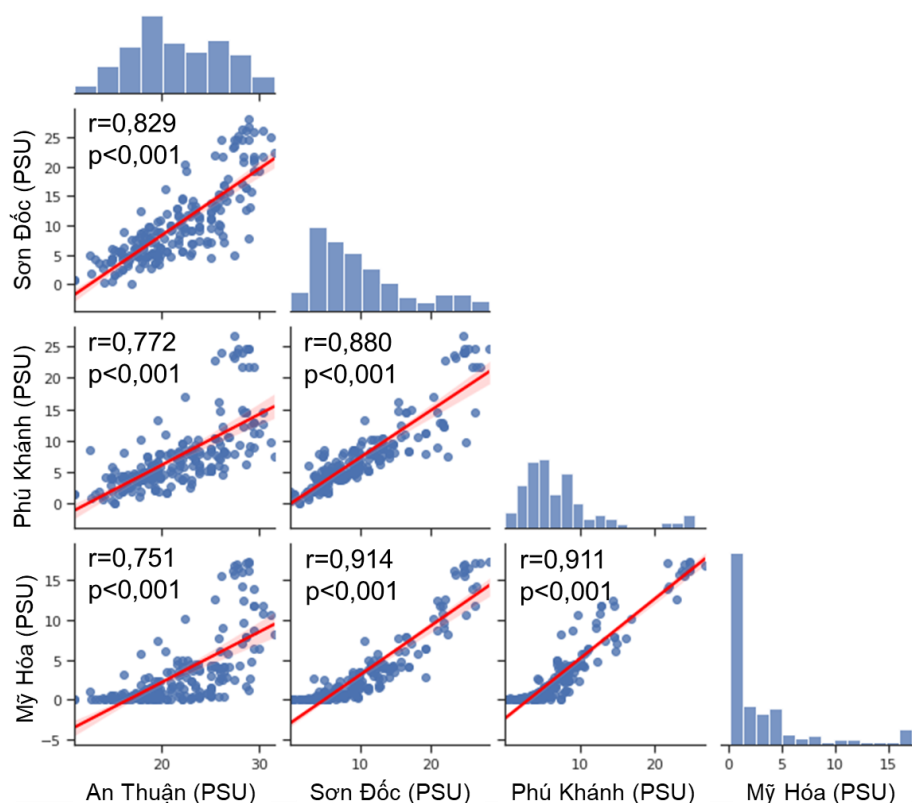
3. Kết quả và thảo luận

3.1. Tương quan về độ mặn giữa các trạm quan trắc

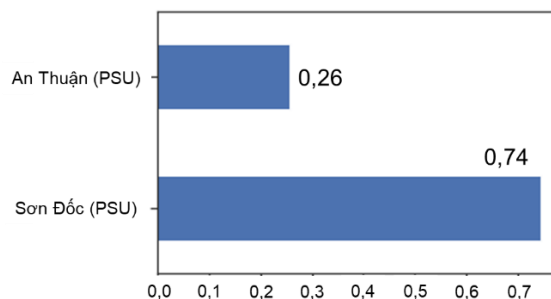
Kết quả phân tích tương quan Spearman rank cho thấy độ mặn ở các trạm quan trắc điều có tương quan thuận ý nghĩa thống kê ( $p < 0,05$ ). Độ mặn ở trạm Mỹ Hóa tương quan thuận mạnh với độ mặn ở trạm Sơn Đốc ( $r = 0,914$ ,  $p < 0,001$ ), và trạm Phú Khánh ( $r = 0,911$ ,  $p < 0,001$ ). Độ mặn ở trạm cửa sông An Thuận cũng ghi nhận có tương quan thuận với độ mặn trạm thượng nguồn Mỹ Hoa với  $r = 0,751$ ,  $p < 0,001$ . Mức độ tương quan giữa các biến độ

lập với nhau cũng khá cao, ví dụ độ mặn ở Sơn Đốc tương quan thuận rất chặt chẽ với độ mặn ở Phú Khánh ( $r = 0,880$ ,  $p < 0,001$ ) (Hình 3). Những biến độc lập có tương quan chặt cần được loại bỏ khỏi thuật toán, do tương quan giữa biến phụ thuộc (Mỹ Hóa) với biến Sơn Sóc cao hơn khi so với Phú Khánh ( $r = 0,914 > 0,911$ ) nên độ mặn ở Phú Khánh sẽ bị loại bỏ khỏi thuật toán.

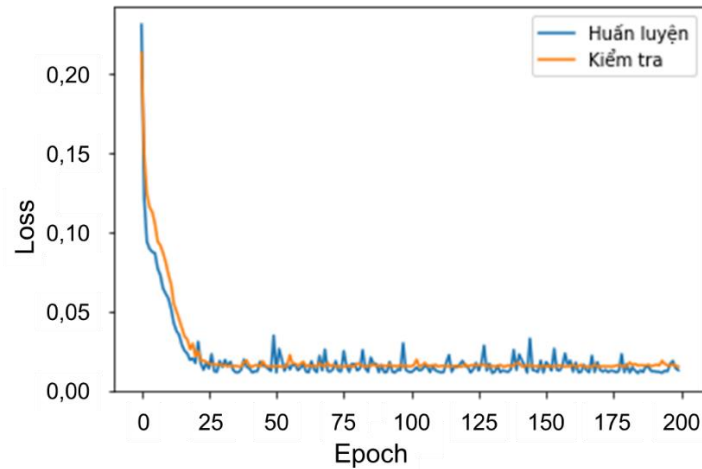
Tóm lại, độ mặn ở An Thuận và Sơn Đốc sẽ là các biến đầu vào (biến độc lập) để dự báo mặn ở thượng nguồn Mỹ Hóa (biến phụ thuộc). Mức độ ảnh hưởng lên độ mặn trạm Mỹ Hóa là khác nhau giữa 2 trạm. Cụ thể, mặn ở trạm Sơn Đốc tác động mạnh nhất đến mặn ở Mỹ Hóa với chỉ số ảnh hưởng lên đến 74%, trong khi mặn ở An Thuận chỉ 26% (Hình 4).



Hình 3. Tương quan giữa độ mặn ở các trạm quan trắc



Hình 4. Mức độ ảnh hưởng lên độ mặn ở trạm thượng nguồn (Mỹ Hóa) của các trạm hạ nguồn (An Thuận, Sơn Đốc).



Hình 5. Hàm loss cho thuật toán ANN ở các epoch khác nhau.

### 3.2. Hiệu quả dự báo mặn của các thuật toán học máy

Kết quả Grid-search cho thấy các thông số tốt nhất cho thuật toán ANN như sau: loss = MSE, epochs = 200, batch size = 10 and optimizer = Adam. Không ghi nhận hiện tượng overfitting do giá trị loss của tập huấn luyện và kiểm tra giảm dần và gần như nằm trùng lên nhau (Hình 5). Cấu trúc mạng ANN dùng trong nghiên cứu bao gồm: 1 lớp đầu vào (2 inputs), 4 lớp ẩn (mỗi lớp 10 nơ ron), và 1 lớp đầu ra (1 output). Ngoài ra, thông số tốt nhất cho thuật toán RFR như sau: max\_depth = 6, n\_estimators = 50. Ở giai đoạn huấn luyện, thuật toán MLR có chỉ số NSE thấp nhất (0,738), trong khi RMSE và MAE cao nhất (0,18 và 0,15, tương ứng). Ngược lại, RFR có NSE cao nhất (0,950), trong khi RMSE và MAE thấp nhất (0,08 và 0,06, tương ứng).

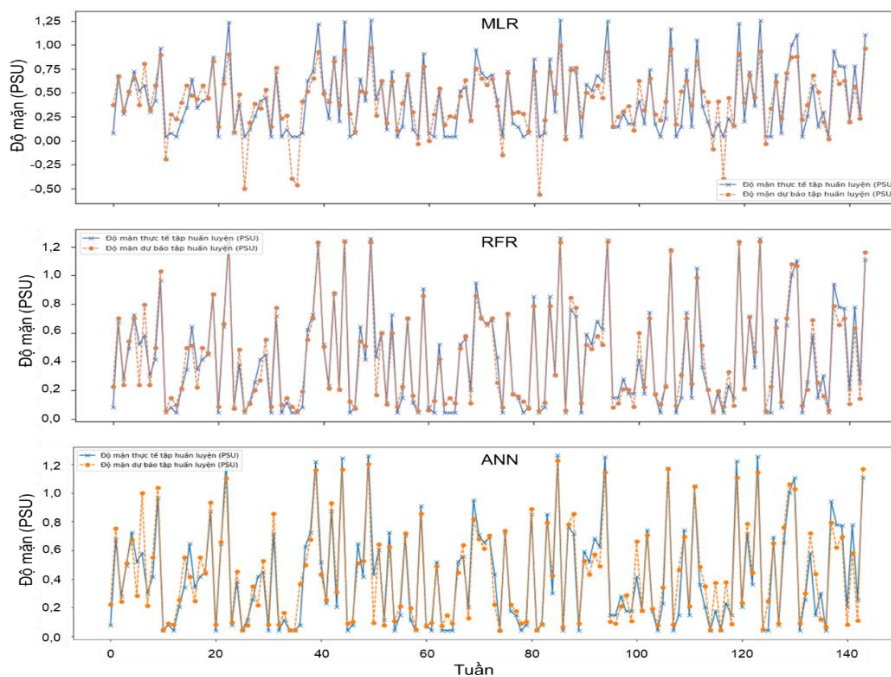
Tương tự, ở giai đoạn kiểm tra, thuật toán MLR có chỉ số NSE thấp nhất (0,756), trong khi RMSE và MAE cao nhất (0,19 và 0,16, tương ứng). Ngược lại, ANN có NSE cao nhất (0,842), trong khi RMSE và MAE thấp nhất (0,16 và 0,11, tương ứng). Như vậy, thuật toán MLR cho kết quả dự báo kém chính xác nhất trong 3 thuật toán. Hiện tượng overfitting đã xuất hiện với thuật toán RFR và ANN khi có NSE huấn luyện cao hơn NSE kiểm tra. Tuy nhiên, thuật toán RFR có overfitting khá mạnh khi NSE của huấn luyện là 0,950 lớn hơn khá nhiều với NSE của kiểm tra là 0,840. Cho nên, lựa chọn ANN là thuật toán dự báo tốt nhất trong 3 thuật toán là phù hợp trong tính huống này. Kết quả kiểm tra giá trị dự báo so với giá trị thực tế của 3 thuật toán được thể hiện ở Hình 6, 7. Ngoài ra, tương quan giữa giá trị dự báo và giá trị thực tế được thể hiện ở Hình 8, 9.

Bảng 3. Hiệu quả dự đoán độ mặn trạm thượng nguồn Mỹ Hóa của thuật toán MLR, RFR, và ANN. T: Tốt, KT: Khá tốt. Giá trị lớn nhất được in đậm, giá trị nhỏ nhất được gạch chân

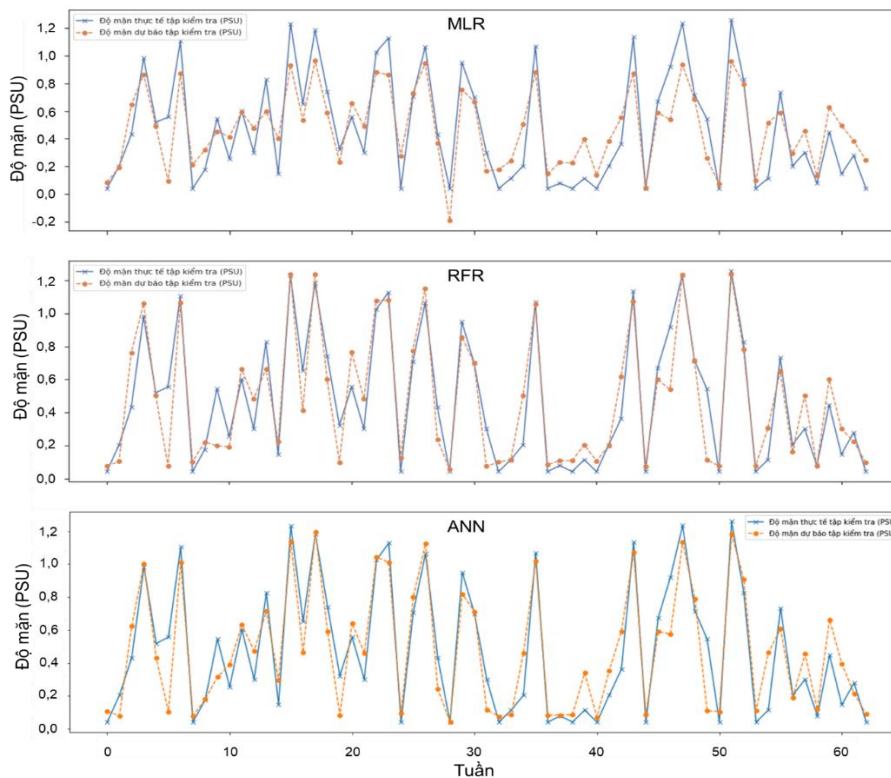
| Thuật toán | Huấn luyện                |             |             | Kiểm Tra                  |             |             |
|------------|---------------------------|-------------|-------------|---------------------------|-------------|-------------|
|            | NSE                       | RMSE        | MAE         | NSE                       | RMSE        | MAE         |
| MLR*       | <u>0,738<sup>T</sup></u>  | <b>0,18</b> | <b>0,15</b> | <u>0,756<sup>T</sup></u>  | <b>0,19</b> | <b>0,16</b> |
| RFR        | <b>0,950<sup>RT</sup></b> | <u>0,08</u> | <u>0,06</u> | 0,840 <sup>RT</sup>       | <u>0,16</u> | 0,12        |
| ANN        | 0,907 <sup>RT</sup>       | 0,11        | 0,08        | <b>0,842<sup>RT</sup></b> | <u>0,16</u> | <u>0,11</u> |

\* Phương trình tuyến tính đa biến: Mặn MH = 0,18 \* Mặn AT + 1,07 \* Mặn SD - 0,84

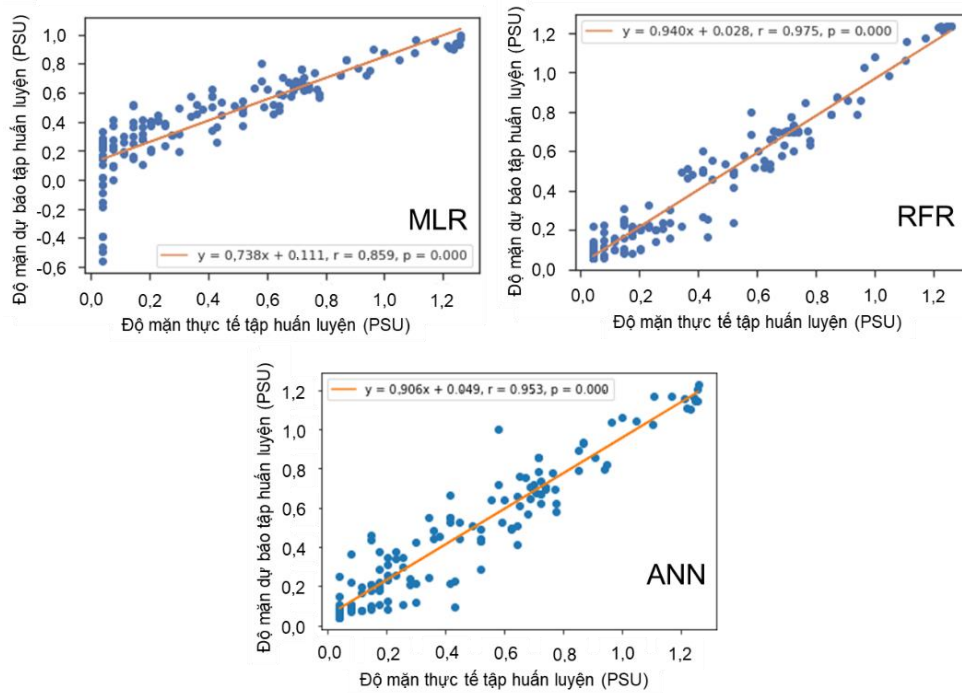




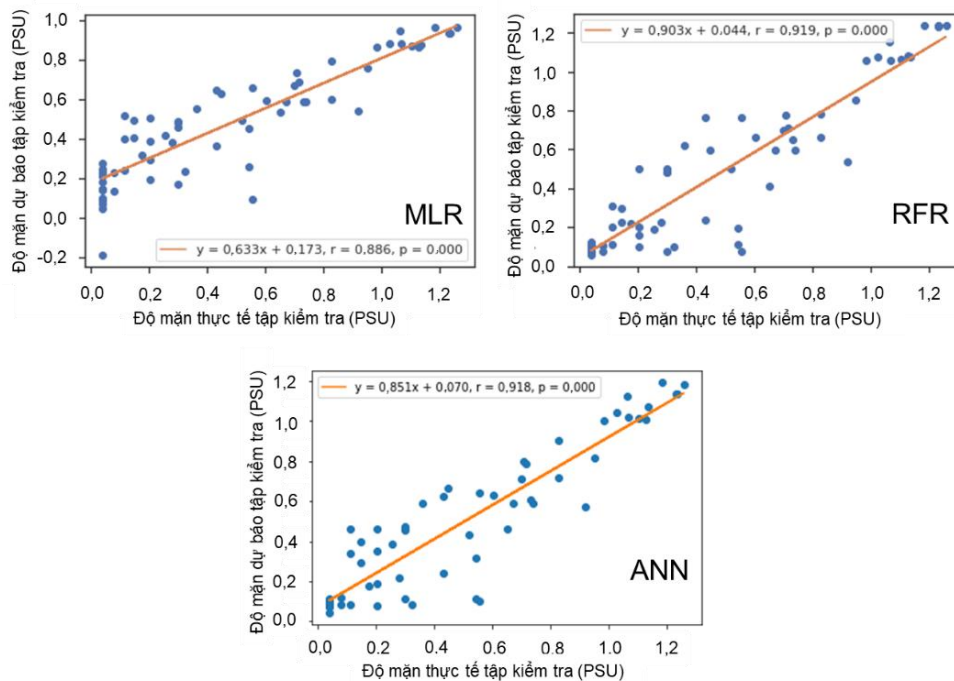
Hình 6. So sánh giữa độ mặn thực tế và dự báo bằng thuật toán MLR, RFR, và ANN ở tập huấn luyện.



Hình 7. So sánh giữa độ mặn thực tế và dự báo bằng thuật toán MLR, RFR, và ANN ở tập kiểm tra.



Hình 8. Quan hệ tuyến tính giữa độ mặn thực tế và dự báo bằng thuật toán MLR, RFR, và ANN ở tập huấn luyện.



Hình 9. Quan hệ tuyến tính giữa độ mặn thực tế và dự báo bằng thuật toán MLR, RFR, và ANN ở tập kiểm tra.

### 3.3. So sánh hiệu quả dự báo mặn trên sông Hàm Luông của thuật toán tuyến tính và phi tuyến tính; giữa thuật toán học máy truyền thống và thuật toán học sâu

Không giống như các thuật toán máy học khác dạng “black - box”, thường rất khó để giải thích kết quả dự báo [9, 14, 15]. Thuật toán hồi quy tuyến tính đa biến nói riêng và thuật toán hồi quy tuyến tính nói chung có ưu điểm là rất đơn giản, dễ áp dụng, dễ giải thích kết quả, quá trình thực thi nhanh chóng [16] nên chúng thường được áp dụng trong các nghiên cứu dự báo. Tuy nhiên, hiệu quả dự báo thường không cao, do có độ chính xác thấp. Qiu và Tang [32], dùng thuật toán MLR để dự báo chất lượng polymer điều chế (đầu ra) từ nhiệt độ (đầu vào), kết quả cho thấy hiệu quả dự báo của thuật toán MLR với  $R^2$  chỉ từ 0,553 đến 0,649. Abba và cộng sự [33] so sánh khả năng dự báo oxy hòa tan (đầu ra) từ pH, nhu cầu oxy sinh học, nhiệt độ (đầu vào) của một số thuật toán như MLR, ANN (Artificial Neural Network), ANFIS (Adaptive Neuro - Fuzzy Inference System), kết quả cho thấy thuật toán MLR dự báo kém nhất với  $R^2$  từ 0,06 đến 0,62. Ngoài ra, thuật toán MLR cũng cho kết quả dự báo ít tin cậy khi so với thuật toán RBFN (Radial Basis Function Neural Network) và ANFIS trong nghiên cứu của Chen và Liu năm 2015 [34]. Garcia và Eldeiry [35], dùng thuật toán GLM (Generalized Linear Model, là một biến thể của thuật toán hồi quy), OLS (Ordinary Least Squares), MARS (Multivariate Adaptive Regression Spline), ANN để dự báo nhiễm mặn trong đất. Kết quả cho thấy thuật toán GLM dự báo kém chính xác nhất với NSE chỉ từ 0,13 đến 0,52.

Thực tế cho thấy, tính đơn giản của MLR vừa là ưu điểm vừa là nhược điểm, do mối quan hệ giữa các biến độc lập và phụ thuộc thường ở dạng phi tuyến tính (non - linear) [36]. Các hiện tượng tự nhiên, trong đó có độ mặn, thường chịu ảnh hưởng đa dạng của nhiều yếu tố, các yếu tố này thường xuất hiện ở dạng chu kỳ, và có quan hệ phi tuyến phức tạp [36, 37]. Cho nên, thuật toán MLR, vốn dùng cho quan hệ tuyến tính, thường khó chính xác khi dự báo các hiện tượng tự nhiên. Điều này tiếp tục được thể hiện qua kết quả nghiên cứu, rõ ràng thuật toán phi tuyến tính

(ANN và RFR) cho kết quả dự báo tốt hơn thuật toán tuyến tính MLR.

Các thuật toán học máy học sâu (ANN, RNN, LSTM,...) thường có kết quả dự báo rất chính xác; tuy nhiên, chúng đòi hỏi phải cung cấp lượng dữ liệu đầu vào rất lớn mới đáp ứng được nhu cầu huấn luyện và đảm bảo tính chính xác [38, 39]. Các thuật toán trong nghiên cứu được cung cấp số lượng dữ liệu đầu vào khá hạn chế (207 dữ liệu); tuy nhiên, thuật toán học sâu vẫn cho thấy ưu điểm về tính chính xác so với thuật toán truyền thống trong cả điều kiện hạn chế dữ liệu đầu vào.

### 3.4. Tăng tính tin cậy của thuật toán và ứng dụng kết quả nghiên cứu trong dự báo mặn trên sông Hàm Luông

Khi sử dụng thuật toán MLR kết hợp với một số thuật toán phi tuyến khác, hiệu quả của thuật toán kết hợp được nâng lên rõ rệt. Nguyên nhân là các hiện tượng tự nhiên thường bao gồm 2 xu hướng: tuyến tính và phi tuyến [13], thuật toán MLR sẽ phân tích phần tuyến tính, còn thuật toán còn lại phụ trách phần phi tuyến. Hiệu quả dự báo của thuật toán kết hợp ANN - MLR cao hơn so với thuật toán đơn lẻ ANN và MLR khi dự báo chlorophyll-*a* ở Terengganu, Malaysia [40]. Tác giả Ebrahimi và Rajaei [41], kết hợp thuật toán MLR và Wavelet để dự báo biến động mực nước ngầm, kết quả cho thấy thuật toán kết hợp có độ chính xác cao hơn nhiều so với thuật toán đơn lẻ (RMSE của W - MLR là 0,058, của MLR là 0,576). Cho nên, các “thuật toán lai” có thể rất hữu dụng để phân tích đầy đủ một hiện tượng tự nhiên.

Đã có nhiều nghiên cứu dự báo xâm nhập mặn ở đồng bằng sông Cửu Long với kết quả rất triển vọng. Tác giả Tran và cộng sự [42] dùng thuật toán MIKE để dự báo xâm nhập mặn trên sông Hậu. Nghiên cứu sử dụng 8 yếu tố đầu vào như lượng mưa hằng ngày của 7 trạm quan trắc trên sông từ năm 1978 đến 2011, lưu lượng nước theo ngày tại trạm Kratie (2010 - 2011), lưu lượng nước theo giờ tại trạm Cần Thơ (2010 - 2011), mực nước tại 10 trạm trên sông (2005 - 2011), thủy triều (2005 - 2011), mạng lưới thủy vực (2005-2011), chế độ triều (2010-2011), lưu lượng nước của các nhánh sông nhỏ (2010-

2011). Kết quả dự báo rất chính xác với  $R^2$  từ 0,92 đến 0,99 (tập huấn luyện), 0,91 - 0,96 (tập kiểm tra). Tuy nhiên, những thuật toán thuộc nhóm thuật toán tiến trình (Process - based models) thường rất phức tạp, số lượng đầu vào phải rất lớn mới đảm bảo tính chính xác [9, 10]. Thông thường, những dữ liệu về toàn bộ các yếu tố ảnh hưởng đến độ mặn là vô cùng khó khăn và thường không đầy đủ [9]. Đối với trường hợp sông Hàm Luông, thuật toán ANN không đòi hỏi nhiều yếu tố đầu vào, dễ áp dụng, đã dự báo thành công và tương đối chính xác độ mặn thượng nguồn Mỹ Hóa. Nên có tiềm năng trong ứng dụng để cảnh báo sớm xâm nhập mặn trên sông Hàm Luông, tỉnh Bến Tre.

Nhìn chung, nghiên cứu này có hai hạn chế chính, cần được cân nhắc để cải thiện và áp dụng vào thực tiễn. Thứ nhất, xâm nhập mặn chịu tác động của nhiều yếu tố như chế độ triều, chế độ dòng chảy - thủy văn, cấu trúc địa hình lòng sông, gió, nhiệt độ; tuy nhiên, do hạn chế của trong việc thu thập số liệu, nghiên cứu này chỉ dùng độ mặn trong quá khứ làm biến đầu vào cho mô hình. Hạn chế này có thể được giải quyết khi liên kết nghiên cứu này với Đài khí tượng thủy văn tỉnh Bến Tre để cập nhật các số liệu và hiệu quả của mô hình thông qua việc cung cấp đầy đủ biến đầu vào hơn. Hạn chế thứ hai là khả năng cập nhật số liệu đầu vào cho mô hình một cách kịp thời để dự báo xâm nhập mặn với thời gian thực (real time forecasting) và có các kỹ thuật đồng hóa số liệu (data assimilation) nhằm làm giảm độ sai số của mô hình. Khía cạnh này chưa được xem xét trong nghiên cứu hiện tại và có thể xem xét trong tương lai khi yếu tố cập nhật số liệu kịp thời được đảm bảo.

#### 4. Kết luận

Nghiên cứu đánh giá khả năng của một số thuật toán học máy trong dự báo mặn trên sông Hàm Luông, tỉnh Bến Tre, tính chính xác của thuật toán dựa vào các chỉ số như NSE, RMSE, MAE. Từ những kết quả đạt được, nghiên cứu đi đến kết luận rằng trong trường hợp ít dữ liệu đầu vào: i) Thuật toán ANN cho kết quả dự báo tốt hơn MLR và RFR; ii) Thuật toán học sâu có thể

vẫn đảm bảo tính chính xác khi so với thuật toán truyền thống; và iii) Thuật toán ANN có tiềm năng trong ứng dụng để cảnh báo sớm xâm nhập mặn trên sông Hàm Luông, tỉnh Bến Tre.

#### Lời cảm ơn

Nghiên cứu được tài trợ bởi trường Đại học Thủ Dầu Một trong đề tài mã số “DT.21.2-036”.

#### Tài liệu tham khảo

- [1] S. Eslami, P. Hoekstra, N. N. Trung, S. A. Kantoush, D. V. Binh, T. T. Quang, M. V. D. Vejt, Tidal Amplification and Salt Intrusion in the Mekong Delta Driven by Anthropogenic Sediment Starvation, *Sci. Rep.*, Vol. 9, No. 1, 2019, pp. 1-10, <https://doi.org/10.1038/s41598-019-55018-9>.
- [2] N. V. K. Triet, N. V. Dung, L. P. Hoang, N. L. Duy, D. D. Tran, T. T. Anh, K. Matti, M. Bruno, H. Apel, Future Projections of Flood Dynamics in the Vietnamese Mekong Delta, *Sci. Total Environ.*, Vol. 742, 2020, pp. 140596, <https://doi.org/10.1016/j.scitotenv.2020.140596>.
- [3] H. T. Nguyen, A. D. Gupta, Assessment of Water Resources and Salinity Intrusion in the Mekong Delta, *Water Int.*, Vol. 26, No. 1, 2001, pp. 86-95, <https://doi.org/10.1080/02508060108686889>.
- [4] A. D. Tran, L. P. Hoang, M. D. Bui, P. Rutschmann, Simulating Future Flows and Salinity Intrusion Using Combined One-and Two-dimensional Hydrodynamic Modelling-the Case of Hau River, *Vietnamese Mekong Delta, Water*, Vol. 10, No. 7, 2018, pp. 897, <https://doi.org/10.3390/w10070897>.
- [5] V. B. Doan, S. A. Kantoush, M. Saber, N. P. Mai, S. Maskey, D. T. Phong, T. Sumi, Long-term Alterations of Flow Regimes of the Mekong River and Adaptation Strategies for the Vietnamese Mekong Delta, *J. Hydrol. Reg. Stud.*, Vol. 32, 2020, pp. 100742, <https://doi.org/10.1016/j.ejrh.2020.100742>.
- [6] A. Smajgl, T. Q. Toan, D. K. Nhan, J. Ward, N. H. Trung, L. Q. Tri, P. V. Tri, P. T. Vu, Responding to Rising Sea Levels in the Mekong Delta, *Nat. Clim. Change*, Vol. 5, No. 2, 2015, pp. 167-174, <https://doi.org/10.1038/nclimate2469>.
- [7] H. Apel, M. Khiem, N. H. Quan, T. Q. Toan, Brief Communication: Seasonal Prediction of Salinity Intrusion in the Mekong Delta, *Nat. Hazards Earth*

- Syst. Sci., Vol. 20, No. 6, 2020, pp. 1609-1616, <https://doi.org/10.5194/nhess-2019-276>.
- [8] M. H. Rahman, T. Lund, I. Bryceson, Salinity Impacts on Agro-biodiversity in three Coastal, Rural Villages of Bangladesh, *Ocean Coast Manag.*, Vol. 54, No. 6, 2011, pp. 455-468, <https://doi.org/10.1016/j.ocecoaman.2011.03.003>.
- [9] A. C. Ross, C. A. Stock, An Assessment of the Predictability of Column Minimum Dissolved Oxygen Concentrations in Chesapeake Bay using A Machine Learning Model, *Estuar. Coast. Shelf Sci.*, Vol 221, 2019, pp. 53-65, <https://doi.org/10.1016/j.ecss.2019.03.007>.
- [10] K. Lin, P. Lu, C.Y. Xu, X. Yu, T. Lan, X. Chen, Modeling Saltwater Intrusion using an Integrated Bayesian model Averaging Method in the Pearl River Delta, *J. Hydroinformatics*, Vol. 21, No. 6, 2019, pp. 1147-1162, <https://doi.org/10.2166/hydro.2019.073>.
- [11] ]Z. Liang, R. Zou, X. Chen, T. Ren, H. Su, Y. Liu, Simulate the Forecast Capacity of a Complicated Water Quality Model using the Long Short-term Memory Approach, *J. Hydrol.*, Vol. 581, 2020, pp. 124432.
- [12] J. Zhang, Y. Zhu, X. Zhang, M. Ye, J. Yang, Developing a Long Short-Term Memory (LSTM) Based Model for Predicting Water Table Depth in Agricultural Areas, *J. Hydrol.*, Vol. 561, 2018, pp. 918-929, <https://doi.org/10.1016/j.jhydrol.2018.04.065>.
- [13] T. T. H. Phan, X. H. Nguyen, Combining Statistical Machine Learning Models with ARIMA for Water Level Forecasting: The Case of the Red River, *Adv. Water Resour.*, Vol. 142, 2020, pp. 103656-103692, <https://doi.org/10.1016/j.advwatres.2020.103656>.
- [14] S. Palani, S.Y. Liang, P. Tkalich, An ANN Application for Water Quality Forecasting, *Mar. Pollut. Bull.*, Vol. 56, No. 9, 2008, pp. 1586-1597, <https://doi.org/10.1016/j.marpolbul.2008.05.021>.
- [15] J. M. Hunter, H. R. Maier, M. S. Gibbs, E. R. Foale, N. A. Grosvenor, N. P. Harders, T. C. K. Miller, Framework for Developing Hybrid Process-driven, Artificial Neural Network and Regression Models for Salinity Prediction In River Systems, *Hydrol. Earth Syst. Sci.*, Vol. 22, No. 5, 2018, pp. 2987-3006, <https://doi.org/10.5194/hess-22-2987-2018>.
- [16] A. Stelzl, M. Pointl, D. F. Hanusch, Estimating Future Peak Water Demand with a Regression Model Considering Climate Indices, *Water*, Vol. 13, No. 14, 2021, pp. 1912, <https://doi.org/10.3390/w13141912>.
- [17] T. D. H. Le, M. Kattwinkel, K. Schützenmeister, J. R. Olson, C. P. Hawkins, R. B. Schäfer, Predicting Current and Future Background Ion Concentrations in German Surface Water Under Climate Change, *Philos. Trans. R. Soc. Lond., B. Biol. Sci.*, Vol. 374, No. 1764, 2019, pp. 20180004, <https://doi.org/10.1098/rstb.2018.0004>.
- [18] G. Zhu, S. Zhang, Y. Bian, A. S. Hursthouse, Multi-linear Regression Model for Chlorine Consumption By Waters, *Environ. Eng. Res.*, Vol. 26, No. 4, 2020, pp. 200402, <https://doi.org/10.4491/eer.2020.402>.
- [19] T. T. Tran, Q. X. Ngo, H. H. Ha, N. P. Nguyen, Short-term Forecasting of Salinity Intrusion in Ham Luong River, Ben Tre Province using Simple Exponential Smoothing Method, *J. Viet. Env.*, Vol. 11, 2019, pp. 43-50, <https://doi.org/10.13141/jve.vol11.no2.pp43-50>.
- [20] T. T. Tran, L. N. Duy, T. L. Pham, T. M. Y. Nguyen, T. H. Y. Tran, X. Q. Ngo, V. T. Lam, H. P. Ngoc, Performance Evaluation of Autoregressive Integrated Moving Average Models for Forecasting Saltwater Intrusion Into Mekong River Estuaries of Vietnam, *Vietnam Journal of Earth Sciences*, Vol. 43, No. 4, 2021, pp. 428-443, <https://doi.org/10.15625/2615-9783/16440>.
- [21] H. Liu, G. X. Sun, R. X. Cao, The Application of GM (1, 1) Dynamic Model in the Forecast of Groundwater Level In Wujiang City, *J. Geol. Hazards Environ. Preserv.*, Vol. 19, No. 3, pp. 47-51.
- [22] M. Pan, H. Zhou, J. Cao, Y. Liu, J. Hao, S. Li, C.H. Chen, Water Level Prediction Model Based on GRU and CNN, *IEEE Access*, Vol. 8, 2008, pp. 60090-60100, <https://doi.org/10.1109/ACCESS.2020.2982433>.
- [23] M. Frigge, D. C. Hoaglin, B. Iglewicz, Some Implementations of the Boxplot, *Am Stat.*, Vol. 43, No. 1, 1989, pp. 50-54, <https://doi.org/10.2307/2685173>.
- [24] P. Geurts, D. Ernst, L. Wehenkel, Extremely Randomized Trees, *Mach. Learn.*, Vol. 63, No. 1, 2006, pp. 3-42, <https://doi.org/10.1007/s10994-006-6226-1>.
- [25] J. Adamowski, H. Fung Chan, S. O. Prasher, B. O. Zielinski, A. Sliusarieva, Comparison of Multiple Linear and Nonlinear Regression, Autoregressive Integrated Moving Average, Artificial Neural Network, and Wavelet Artificial Neural Network Methods for Urban Water Demand Forecasting in Montreal, Canada, *Water Resour. Res.*, Vol. 48, No. 1, 2012, pp. 1528, <https://doi.org/10.1029/2010WR009945>.

- [26] L. Breiman, Random Forests, *Mach. Learn.*, Vol. 45, No. 1, 2001, pp. 5-32.
- [27] T. Francke, J. L. Tarazón, B. Schroder, Estimation of Suspended Sediment Concentration and Yield Using Linear Models, Random Forests and Quantile Regression Forests, *Hydrol. Process.*, No. 22, Vol. 25, 2008, pp. 4892, <https://doi.org/10.1002/hyp.7110>.
- [28] T. T. H. Nguyen, M. T. Doan, Apply the Random Forest Classification Algorithm to Build Building Land Use/Carpet Cover Map of Dak Lak Province Based on Landsat 8 OLI satellite image, *Journal of Agriculture & Rural Development*, Vol. 13, 2018, pp. 122-129 (in Vietnamese).
- [29] H. Q. Nguyen, N. T. Ha, T. L. Pham, Inland Harmful Cyanobacterial Bloom Prediction in the Eutrophic Tri An Reservoir Using Satellite Band Ratio and Machine Learning Approaches, *Environ. Sci. Pollut. Res.*, Vol. 27, No. 9, 2020, pp. 9135-9151, <https://doi.org/10.1007/s11356-019-07519-3>.
- [30] J. E. Nash, J. V. Sutcliffe, River Flow Forecasting Through Conceptual Models Part I – a Discussion of Principles, *J. Hydrol.*, Vol. 10, 1970, pp. 282-290, [https://doi.org/10.1016/0022-1694\(70\)90255-6](https://doi.org/10.1016/0022-1694(70)90255-6).
- [31] D. N. Moriasi, M. W. Gitau, N. Pai, P. Daggupati, Hydrologic and Water Quality Models: Performance Measures and Evaluation Criteria, *Trans. ASABE*, Vol. 58, 2015, pp. 1763-1785, <https://doi.org/10.13031/trans.58.10715>.
- [32] S. B. Qiu, B. Tang, Application of Multiple Linear Regression Analysis in Polymer Modified Mortar Quality Control, in Proceedings of the 2<sup>nd</sup> International Conference on Electronic and Mechanical Engineering and Information Technology, 2012, pp. 1124-1127.
- [33] S. I. Abba, S. J. Hadi, J. Abdullahi, River Water Modelling Prediction Using Multi-Linear Regression, Artificial Neural Network, and Adaptive Neuro-Fuzzy Inference System Techniques, *Procedia Comput. Sci.*, Vol. 120, 2017, pp. 75-82, <https://doi.org/10.1016/j.procs.2017.11.212>.
- [34] W. B. Chen, W.C. Liu, Water Quality Modeling in Reservoirs Using Multivariate Linear Regression and Two Neural Network Models, *Adv. Artif. Neural Syst.*, Vol. 2015, 2015, <https://doi.org/10.1155/2015/521721>.
- [35] L. A. Garcia, A. A. Eldeiry, Evaluating Linear and Nonlinear Regression Models in Mapping Soil Salinity, *Int. J. Res. Agric. For.*, Vol. 7, No. 3, 2020, pp. 21-34, <http://ijraf.org/papers/v7-i3/4.pdf>.
- [36] S. Wolff, F. O'Donncha, B. Chen, Statistical and Machine Learning Ensemble Modelling to Forecast Sea Surface Temperature, *J. Mar. Syst.*, Vol. 208, 2018, pp. 103347, <https://doi.org/10.1016/j.jmarsys.2020.103347>.
- [37] A. Lal, B. Datta, Application of the Group Method of Data Handling and Variable Importance Analysis for Prediction and Modelling of Saltwater Intrusion Processes in Coastal Aquifers, *Neural. Comput. Appl.*, Vol. 33, 2020, pp. 4179-4190, <https://doi.org/10.1007/s00521-020-05232-8>.
- [38] J. Lago, F. D. Ridder, B. D. Schutter, Forecasting Spot Electricity Prices: Deep Learning Approaches and Empirical Comparison of Traditional Algorithms, *Appl. Energy*, Vol. 221, 2015, pp. 386-405, <https://doi.org/10.1016/j.apenergy.2018.02.069>.
- [39] A. Schmidt, D. B. Mainwaring, D. A. Maguire, Development of a Tailored Combination of Machine Learning Approaches to Model Volumetric Soil Water Content Within a Mesic Forest in the Pacific Northwest, *J. Hydrol.*, Vol. 588, 2020, pp. 125044, <https://doi.org/10.1016/j.jhydrol.2020.125044>.
- [40] M. S. Lola, M. N. A. Ramlee, G. S. Gunalan, N. H. Zainuddin, R. Zakariya, M. Idris, I. Khalil, Improved the Prediction of Multiple Linear Regression Model Performance Using the Hybrid Approach: A Case Study of Chlorophyll-a at the Offshore Kuala Terengganu, *Terengganu, Open Journal of Statistics*, Vol. 6, No. 5, 2016, pp. 789-804, <https://doi.org/10.4236/ojs.2016.65065>.
- [41] H. Ebrahimi, T. Rajaei, Simulation of Groundwater Level Variations Using Wavelet Combined with Neural Network, Linear Regression and Support Vector Machine, *Glob. Planet Change*, Vol. 148, 2017, pp. 181-191, <https://doi.org/10.1016/j.gloplacha.2016.11.014>.
- [42] A. D. Tran, L. P. Hoang, M. D. Bui, P. Rutschmann, Simulating Future Flows and Salinity Intrusion Using Combined One- and Two-Dimensional Hydrodynamic Modelling-The Case of Hau River, Vietnamese Mekong Delta, *Water*, Vol. 10, 2018, pp. 897-917, <https://doi.org/10.3390/w10070897>.