ĐÁNH GIÁ KHẢ NĂNG DỰ BÁO MẶN TRÊN SÔNG HÀM LUÔNG CỦA THUẬT TOÁN K-NEAREST NEIGHBORS

Phạm Ngọc Hoài, Phan Thị Thanh Huyền
Trường Đại học Thủ Dầu Một; Học viện Khoa học và Công nghệ - Viện Hàn lâm Khoa học và Công nghệ Việt Nam,
Lê Nguyễn
Trường Đại học Nguyễn Tất Thành
Nguyễn Thu Hiền
Trường Đại học Công nghiệp Thực phẩm
Trần Thành Thái
Viện Sinh học Nhiệt đới - Viện Hàn lâm Khoa học và Công nghệ Việt Nam
Lương Lê Lâm
Trường Đại học Công nghiệp thành phố Hồ Chí Minh

Tóm tắt: *Xâm nhập mặn là vấn đề rất đáng quan tâm ở vùng đồng bằng sông Cửu Long. 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 hữu ích. Từ đây, mục tiêu của nghiên cứu là đánh giá khả năng áp dụng phương pháp K-Nearest Neighbors (KNN), một thuật toán đơn giản và dễ áp dụng của học máy, 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 (tổng cộng 207 tuần). Các chỉ số thống kê như Hệ số Nash - Sutcliffe efficiency (NSE), Lỗi trung bình bình phương gốc (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 mô hình dự báo. Kết quả cho thấy mô hình KNN dự báo độ mặn khá tốt với NSE = 0,960, RMSE = 0,842, MAE = 0,541 cho tập huấn luyện, NSE = 0,904, RMSE = 1,448, MAE = 0,914 cho tập kiểm tra. Mô hình KNN là một mô hình đơn giản, dễ thực thi nhưng cho kết quả dự báo khá chính xác, cho nên mô hình rất tiềm năng trong ứng dụng dự báo mặn ở sông Hàm Luông nói riêng và một số nhánh sông của sông Mê Kông nói chung.*
Từ khóa: *Biến đổi khí hậu, Đồng bằng sông Cửu Long, Trí thông minh nhân tạo, Xâm nhập mặn, K-Nearest Neighbors.*

Summary: *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 create a K-Nearest Neighbors (KNN) model for predicting the saltwater intrusion in Ham Luong River, Ben Tre Province. The input data 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 performances of KNN model. The research results indicated that the KNN model achieved a high performance for salinity forecasting. with NSE = 0,960, RMSE = 0,842, MAE = 0,541 for training period, NSE = 0,904, RMSE = 1,448, MAE = 0,914 for testing period. The findings of this study suggest that the KNN model has promised as a potential tool in salinity forecasting with salinity data characteristics in Ham Luong River.*
Keywords: *Artificial intelligence, Climate change, Mekong Delta, Saltwater intrusion, K-Nearest Neighbors.*

1. GIỚI THIỆU

Đồng bằng sông Cửu Long (ĐBSCL) nằm ở vùng hạ lưu sông Mê Kông, đây là vùng đồng bằng rộng lớn, màu mỡ lớn thứ ba trên thế giới với 3,9 triệu héc-ta [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 lượng thực thực phẩm và đóng góp

Ngày nhận bài: 16/6/2022
Ngày thông qua phản biện: 28/8/2022
Ngày duyệt đăng: 04/10/2022

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 mặn với thách thức rất lớn từ các thiên tai như khô hạn và xâm nhập mặn [1, 2]. Mặc dù xâm nhập mặn (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].

Xâm nhập mặn là một trong những vấn đề chính của quản lý nguồn nước vùng cửa sông ven biển [4]. XNM làm giảm khả năng lọc và gia tăng các loại độc tố trong đất, dẫn đến năng suất cây trồng thấp [5]. Hơn nữa, độ mặn cao làm cây trồng mất nhiều năng lượng để hút nước từ đất làm cây trồng chậm phát triển [6]. 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]. Chín trên tổng số mười ba tỉnh vùng ĐBSCL đang chịu ảnh hưởng từ XNM, từ đây, hàng nghìn héc-ta hoa màu, cây ăn trái, lúa gạo, nuôi trồng thủy sản bị tác động [7].

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 mô hình này dự báo và mô tả rất chính xác các quy luật thủy văn (ví dụ xâm nhập mặn) nhờ các quy luật vật lý được nghiên cứu và tích hợp sẵn trong mô hình. Tuy nhiên, cần có những chuyên gia để khai thác được những mô hình 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 [8]. Một cách tiếp cận khác là sử dụng các thuật toán học máy (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 mô hình học máy đã được sử dụng rộng rãi trong các nghiên cứu dự báo thủy văn như chất lượng nước [9], độ cao cột nước [10]. Tác giả Lin và cộng sự năm 2019 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% [11]. 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 [8]. Cho nên, lựa chọn các mô hình 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 học máy.

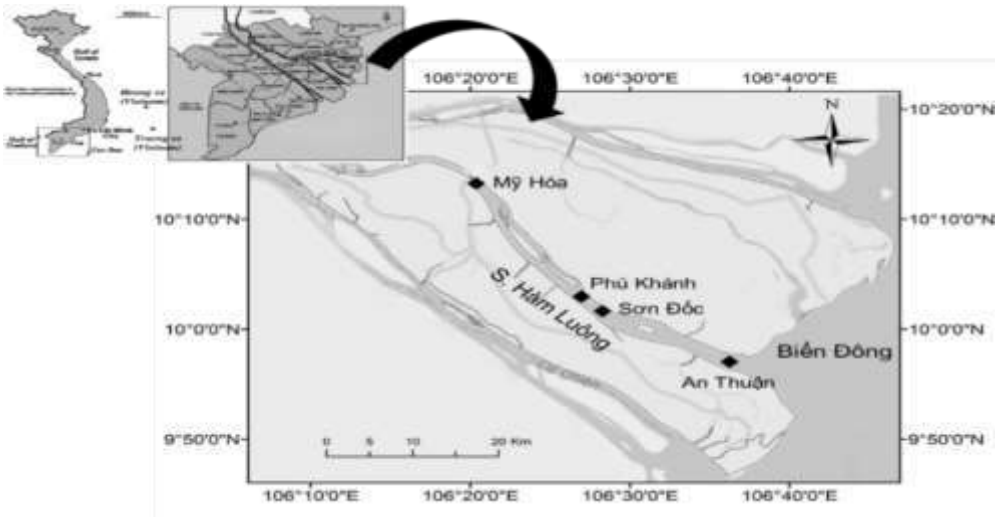
Do đó, nghiên cứu được thực hiện với mục tiêu đánh giá khả năng của mô hình K láng giềng gần nhất (K- Nearest Neighbors, KNN), là mô hình rất đơn giản, để áp dụng, trong dự báo mặn ở sông Hàm Luông (SHL), tỉnh Bến Tre. Đây là một trong những nhánh sông lớn của hệ thống sông Mê Kông và đang bị mặn xâm 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 con sông chảy trọn vẹn trong địa bàn tỉnh Bến Tre, đây là một phân lưu của sông Tiền. Sông có chiều dài khoảng 70 km, bắt đầu từ địa phận xã Tân Phú (huyện Châu Thành), chảy theo hướng Đông Nam, đi qua địa phận các huyện như Chợ Lách, Mỏ Cày Bắc, thành phố Bến Tre, Giồng Trôm, Mỏ Cày Nam, Thạnh Phú, Ba Tri đổ ra Biển Đông tại cửa Hàm Luông. Chiều rộng lớn nhất tại cửa sông khoảng 1.200 - 1.500 m. Sông có độ sâu trung bình từ 12 đến 16 m [12]. SHL là một

sông lớn của tỉnh Bến Tre, đóng vai trò quan trọng trong cung cấp nguồn nước cho sinh hoạt, sản xuất công - nông nghiệp, du lịch và các hoạt động vận tải đường sông [12]. Trên SHL có 4 trạm quan trắc mặn, lần lượt từ cửa sông lên thượng nguồn là: An Thuận (AT), Sơn Đốc (SĐ), Phú Khánh (PK), Mỹ Hóa (MH) (Hình 1).



Hình 1: Vị trí các trạm quan trắc mặn trên sông Hàm Luông, tỉnh Bến Tre

2.2. Thu thập và tiền xử lý số liệu

Dữ liệu độ mặn nước mặt (đơn vị: PSU) của SHL từ năm 2012 đến 2020 tại 4 trạm quan trắc mặn được thu thập từ Đài Khí tượng Thủy văn tỉnh Bến Tre (http://www.bentre.gov.vn/Lists/ThongTinCanBiet/TongQuat.aspx). Ở mỗi 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 các đặc điểm thống kê của bộ dữ liệu độ mặn trên SHL về số lượng dữ liệu, trung bình, độ lệch chuẩn, giá trị nhỏ - lớn nhất, giá trị phân vị thứ 25, 50, và 75.

Bảng 1: Thống kê mô tả bộ dữ liệu về độ mặn tại 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
Phân vị 25% (PSU)	18,00	5,05	4,00	0,30
Phân vị 50% (PSU)	21,60	8,60	5,90	1,20
Phân vị 75% (PSU)	25,45	13,10	8,60	4,15
Cực đại (Max, PSU)	31,50	28,20	26,70	17,20

Dữ liệu mặn tại các trạm được tiền xử lý qua ba bước trước khi được đưa vào mô hình để huấn luyện thuật toán:

1) Loại bỏ các giá trị trống (Null), đồng thời thay thế các giá trị đó bằng thuật toán nội suy của Python.

2) 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 [13]. 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 * Độ trải giữa (IQR, Interquartile Range)) và giới hạn dưới (Q1 - 1.5 * IQR) của biểu đồ hộp.

3) Số liệu được chuẩn hóa về chung một tỷ lệ từ 0 đến 1 bằng phương pháp normalization scaling, được thực hiện trorng thư viện scikit - learn của Python

Ngoài ra, tương quan về độ mặn giữa các trạm quan trắc được đánh giá bằng phân tích Pearson.

2.3. Mô hình K láng giềng gần nhất (K-Nearest Neighbors, KNN)

Thuật toán K - láng giềng gần nhất (K- Nearest Neighbors, KNN) là thuật toán học máy có giám sát (supervised-learning) đơn giản và dễ triển khai, thường được dùng trong các bài toán phân loại (classification) và hồi quy (regression).

Ý tưởng của thuật toán KNN là dự đoán dữ liệu mới dựa trên dữ liệu của K điểm gần nhất xung quanh nó (K láng giềng). Trong bài toán hồi quy, giá trị dự báo bằng chính giá trị của điểm dữ liệu đã biết gần nhất (nếu K=1), hoặc là trung bình có trọng số của các điểm dữ liệu gần nhất. Trong thuật toán KNN, việc xác định (i) phương pháp đo khoảng cách giữa các điểm dữ liệu và (ii) bao nhiêu số K láng giềng là quan trọng nhất. Trong không gian một chiều, khoảng cách giữa hai điểm là trị tuyệt đối giữa hiệu giá trị của hai điểm đó. Trong thực tế, việc đo khoảng cách giữa các điểm dữ liệu có thể sử dụng rất nhiều phương pháp đo, nghiên cứu này áp dụng phương pháp phổ biến nhất là Euclidean. Về số K láng giềng, khi K nhỏ làm cho mô hình phức tạp, gây ra trình trạng quá khớp (overfitting) nhưng khi K lớn làm cho mô hình bị nhiễu. Số K láng giềng được xác định bằng phương pháp GridSearchCV (10 CV, n_neighbors: 1→12).

2.4. Xây dựng và đánh giá mô hình

Để dự báo độ mặn ở thượng nguồn, độ mặn tại các trạm An Thuận, Sơn Đốc, Phú Khánh 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 overfittings trong huấn luyện thuật toán. Các chỉ số thống kê như Hệ số xác định NSE (Nash - Sutcliffe efficiency coefficient), Lỗi trung bình bình phương gốc (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 mô hình dự báo:

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

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

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

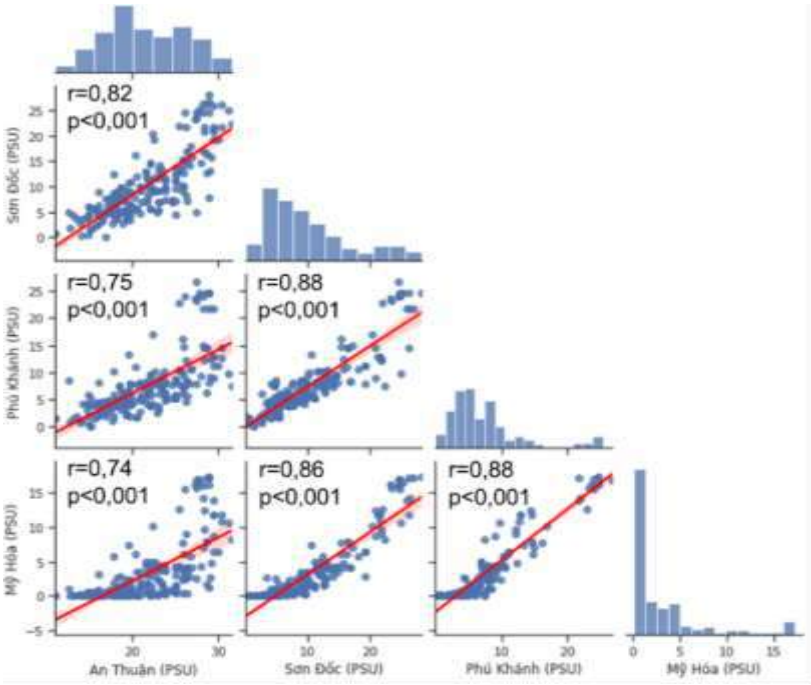
Trong đó, n là số mẫu, $\hat{y}i$, yi, $\bar{y}$ tương ứng là giá trị dự báo, giá trị thực, trung bình giá trị thực

Moriasi và cộng sự năm 2015 đề xuất thang đánh giá tính chính xác của mô hình dự báo dựa vào giá trị NSE, cụ thể như sau: Mô hình rất tốt (NSE > 0.80), tốt (0.7 < NSE ≤ 0.8), chấp nhận được (0.50 < NSE ≤ 0.70), không đáng tin cậy (NSE ≤ 0.50) [14].

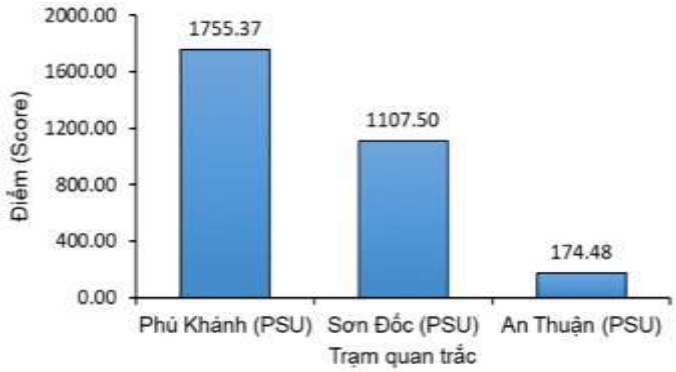
3. KẾT QUẢ VÀ BÀN LUẬN

3.1. Quan hệ giữa độ mặn cửa sông và thượng nguồn

Phân tích Pearson ghi nhận độ mặn ở các trạm quan trắc điều có tương quan thuận ý nghĩa thống kê. Độ mặn ở trạm Mỹ Hóa tương quan thuận mạnh với độ mặn ở trạm Sơn Đốc ($r = 0,82$, $p < 0,001$). Ngoài ra, độ mặn ở trạm cửa sông Phú Khánh và An Thuận cũng ghi nhận có tương quan thuận với độ mặn trạm thượng nguồn Mỹ Hoa với r là 0,75; 0,74, tương ứng (Hình 2). Điều này cho thấy độ mặn ở trạm Mỹ Hóa chịu ảnh hưởng của độ mặn ở 3 trạm ngoài cửa sông là An Thuận, Sơn Đốc, và Phú Khánh. Tuy nhiên, mức độ ảnh hưởng lên độ mặn trạm Mỹ Hóa của 3 trạm là khác nhau. Cụ thể, độ mặn ở trạm Phú Khánh và Sơn Đốc tác động mạnh nhất đến mặn ở Mỹ Hóa với điểm số ảnh hưởng lên đến 1755,37 và 1107,50, tương ứng. Mặn ở trạm An Thuận ít ảnh hưởng đến mặn ở Mỹ Hóa với điểm số ảnh hưởng chỉ 174,48 (Hình 3).



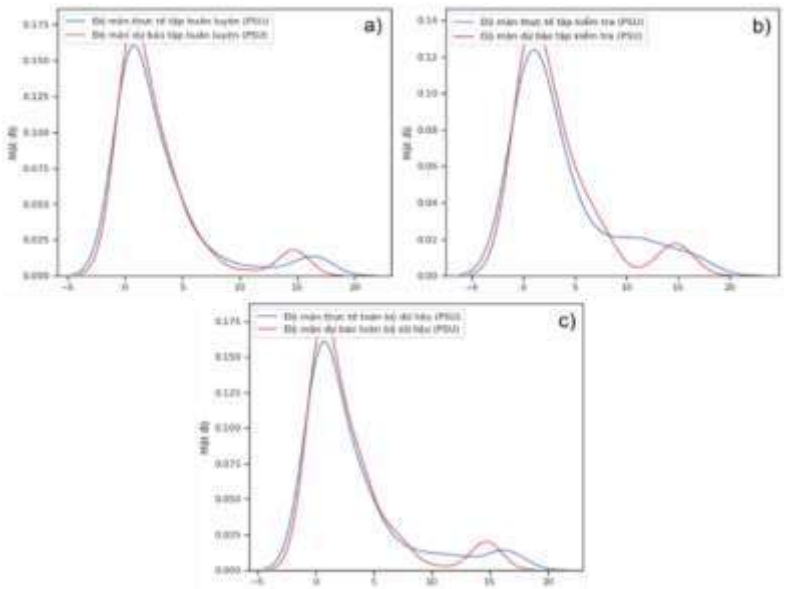
Hình 2: Mối quan hệ giữa độ mặn ở các trạm quan trắc trên sông Hàm Luông



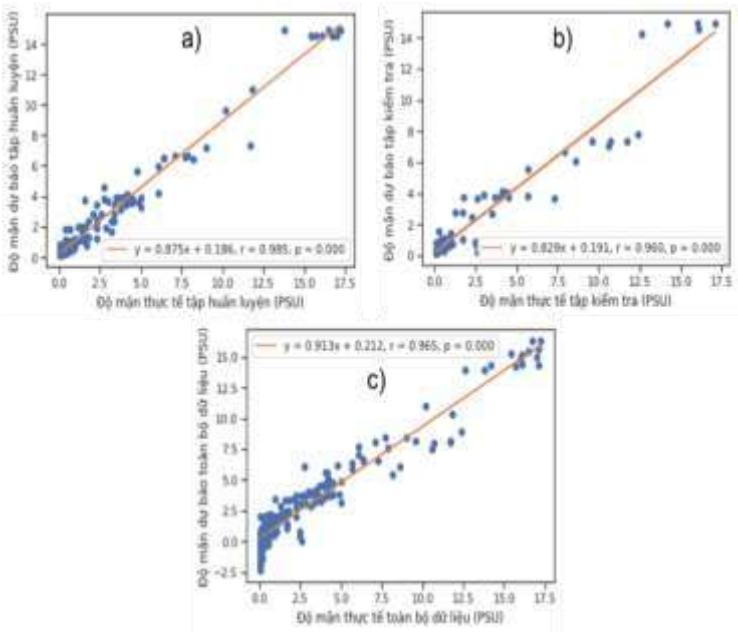
Hình 3: Mức độ ảnh hưởng (về phương sai) của độ mặn ở các trạm cửa sông (An Thuận, Sơn Đốc, Phú Khánh) lên độ mặn trạm thượng nguồn (Mỹ Hóa) xác định theo phương pháp SelectKBest

3.2. Hiệu quả dự báo của mô hình KNN

Mô hình KNN dùng độ mặn ở An Thuận, Sơn Đốc, Phú Khánh (đầu vào, hay các biến độc lập) để dự báo độ mặn ở Mỹ Hóa (đầu ra, biến phụ thuộc). Kết quả dự báo ghi nhận các giá trị thực tế và giá trị dự báo ở tập huấn luyện, kiểm tra, và toàn bộ dữ liệu có phân phối mật độ gần như trùng lên nhau (Hình 4). Trong quá trình huấn luyện và kiểm tra, tương quan giữa giá trị dự báo và giá trị thực tế rất cao ($r > 96\%$, $p < 0{,}001$). Cụ thể, hệ số tương quan r là 0,985 và 0,96 cho huấn luyện và kiểm tra. Ngoài ra, khi dùng toàn bộ dữ liệu để chạy mô hình dự báo (không phân ra tập huấn luyện và kiểm tra), hệ số tương quan giữa giá trị dự báo và giá trị thực tế cũng rất cao ($r = 0{,}965$, $p < 0{,}001$) (Hình 5). Điều này chứng tỏ kết quả dự báo khá chính xác và có độ tin cậy.

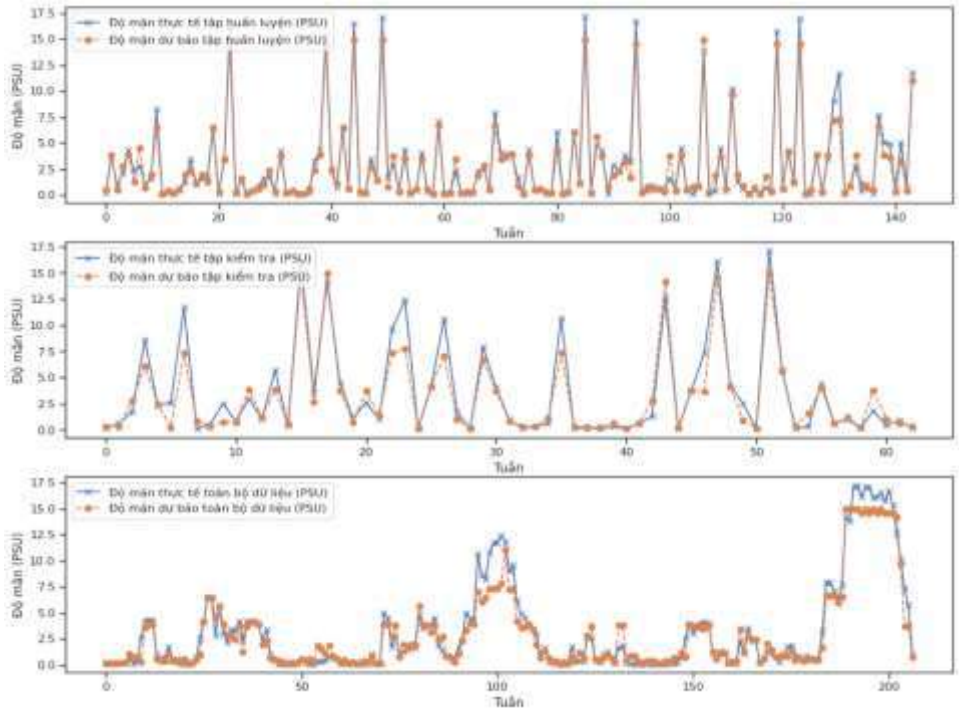


Hình 4: Mật độ phân bố giá trị thực và dự báo ở tập huấn luyện (a), kiểm tra (b), toàn bộ dữ liệu (c)



Hình 5: Tương quan tuyến tính giữa độ mặn thực tế và dự báo bằng mô hình KNN ở tập huấn luyện (a), kiểm tra (b), toàn bộ dữ liệu (c)

Hình 6 so sánh các giá trị dự báo và thực tế ở tập huấn luyện, kiểm tra, và toàn bộ số liệu. Nhìn chung, giá trị dự báo gần như trùng khớp với giá trị thực tế, điều này cho thấy mô hình KNN dự báo khá chính xác. Giá trị NSE cho tập huấn luyện, kiểm tra, và toàn bộ dữ liệu tương ứng là 0,960; 0,904; 0,940. Như vậy, mô hình KNN để dự báo mặn ở thượng nguồn Mỹ Hóa được đánh giá rất tốt trong cả 3 giai đoạn: huấn luyện, kiểm tra, và toàn bộ dữ liệu. Ngoài ra, mô hình gần như không xuất hiện overfittings do NSE ở giai đoạn huấn luyện và kiểm tra cao gần bằng nhau (Bảng 2).



Hình 6: So sánh giữa độ mặn thực tế và dự báo bằng mô hình KNN ở tập huấn luyện (a), kiểm tra (b), toàn bộ dữ liệu (c)

Bảng 2: Hiệu quả dự đoán độ mặn trạm thượng nguồn Mỹ Hóa của mô hình KNN. RT: Rất tốt

Huấn luyện			Kiểm Tra			Toàn bộ số liệu		
NSE	RMSE	MAE	NSE	RMSE	MAE	NSE	RMSE	MAE
0,960^{RT}	0,842	0,541	0,904^{RT}	1,448	0,914	0,940^{RT}	1,063	0,654

Không giống như một số thuật toán học máy truyền thống như Decision Tree, Random Forest, Support Vector Machine hay thuật toán học sâu như Artificial Neural Network, Long Short Term Memory networks, thường phức tạp và khó giải thích. Thuật toán KNN thường rất đơn giản, để áp dụng, và quá trình thực thi nhanh chóng [15]. Tuy nhiên, tính đơn giản của KNN cũng là nhược điểm, thuật toán gán trọng số bằng nhau cho tất cả các biến, mặc dù một số biến nhất định có thể có mức độ ưu tiên cao hơn, điều này có thể dẫn đến dự báo kém chính xác [15].

Mặc dù không được đánh giá cao trong các nghiên cứu dự báo; tuy nhiên, mô hình KNN vẫn thể hiện rất tốt đối với bộ số liệu mặn của

SHL từ 2012 đến 2020. Nguyên nhân do độ mặn giữa các trạm An Thuận, Sơn Đốc, Phú Khánh, và Mỹ Hóa có quan hệ tuyến tính và tương quan khá chặt chẽ. Ví dụ tương quan giữa độ mặn trạm Mỹ Hóa và Sơn Đốc, Phú Khánh, An Thuận khá chặt chẽ, tuyến tính, với r > 0,74. Điều này chứng tỏ, mặc dù KNN là một mô hình thuộc nhóm phi tuyến nhưng khả năng học dữ liệu tuyến tính cũng khá tốt. 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 [16]. Cho nên, mô hình KNN, vốn có thể ứng dụng cho phân tích quan hệ tuyến tính và phi tuyến tính, có thể xem xét ứng dụng để dự báo các 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 năm 2018 dùng mô hình 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) [4]. Tuy nhiên, những mô hình thuộc nhóm mô hình tiến trình thường rất phức tạp, số lượng đầu vào phải rất lớn mới đảm bảo tính chính xác [8]. 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 [8]. Đối với trường hợp SHL, mô hình KNN với đặc điểm là đơn giản, dễ thực hiện, không đòi hỏi nhiều yếu tố đầu vào, dễ áp dụng, đã dự báo thành công và chính xác độ mặn thượng nguồn Mỹ Hóa. Nên có tiềm năng trong ứng dụng để cảnh báo sớm xâm nhập mặn trên SHL, tỉnh Bến Tre.

4. KẾT LUẬN

Nghiên cứu kiểm tra khả năng dự báo mặn trên sông Hàm Luông, tỉnh Bến Tre của thuật toán KNN. Từ những kết quả đạt được, nghiên cứu đi đến kết luận rằng mô hình KNN với đặc điểm là đơn giản, dễ thực hiện, không đòi hỏi nhiều yếu tố đầu vào đã dự báo chính xác độ mặn thượng nguồn Mỹ Hóa. Cho nên mô hình rất tiềm năng trong ứng dụng dự báo mặn ở sông Hàm Luông nói riêng và một số nhánh sông thuộc hệ thống sông Mê Kông nói chung.

Lời cảm ơn

Nghiên cứu này được tài trợ kinh phí bởi Trường Đại học Thủ Dầu Một trong đề tài mã số DT.21.2-036. Nhóm tác giả trân trọng cảm ơn những đóng góp và chỉnh sửa của Ban biên tập và Quý phản biện.

TÀI LIỆU THAM KHẢO

[1] S. Eslami, P. Hoekstra, N. N. Trung, S. A. Kantoush, D. V. Binh, T. T. Quang, M. V. D. Vegt, "Tidal amplification and salt intrusion in the Mekong Delta driven by anthropogenic sediment starvation," *Scientific reports*, 9(1), pp.1, 2019.

[2] N. V. K. Triet, N. V. Dung, L. P. Hoang, N. L. Duy, D. D. Tran, T. T. Anh, ... & H. Apel, "Future projections of flood dynamics in the Vietnamese Mekong Delta," *Science of the Total Environment*, Vol. 742, pp.140596, 2020.

[3] N. H. Thoi, & A. D. Gupta, "Assessment of water resources and salinity intrusion in the Mekong Delta," *Water International*, 26(1), pp. 86, 2001.

[4] Q. 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*, 10(7), pp. 897, 2018.

[5] M. H. Rahman, T. Lund, I. Bryceson, "Salinity impacts on agro-biodiversity in three coastal, rural villages of Bangladesh," *Ocean & Coastal Management*, 54(6), pp. 455, 2011.

[6] D. V. Binh, 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," *Journal of Hydrology: Regional Studies*, 32, pp.100742, 2020.

[7] H. Apel, M. Khiem, N. H. Quan, & T. Q. Toan, "Brief communication: Seasonal prediction of salinity intrusion in the Mekong Delta," *Natural Hazards and Earth System Sciences*, 20(6), pp.1609, 2020.

[8] 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," *Estuarine, Coastal and Shelf Science*, vol.221, pp.53, 2019.

[9] 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," *Journal of Hydrology*, Vol. 581, pp. 124432, 2020.

[10] 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," *Journal of hydrology*, vol. 561, pp. 918, 2018.

[11] 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," *Journal of Hydroinformatics*, 21(6), pp.1147, 2019.

[12] 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," *Journal of Vietnamese Environment*, 11(2), 43, 2019.

[13] M. Pan, H. Zhou, J. Cao, Y. Liu, J. Hao, S. Li, & C. H. Chen, "Water level prediction model based on GRU and CNN," In Proc. IEEE Access 8, 2020, pp 60100.

[14] D. N. Moriasi, M. W. Gitau, N. Pai, & P. Daggupati, "Hydrologic and water quality models: Performance measures and evaluation criteria," *Transactions of the ASABE*, 58(6), pp.1763, 2015.

[15] D. Vermeulen, &, A. V. Niekerk, "Machine learning performance for predicting soil salinity using different combinations of geomorphometric covariates," *Geoderma* vol. 299, pp.1, 2017.

[16] 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 Computing and Applications*, 33(9), pp. 4179, 2021.