



Rừng là vàng, nếu mình biết bảo vệ và xây dựng thì rừng rất quý

(Lời Hồ Chủ tịch)

Rừng & Môi trường

ISSN 1859-1248

HỘI KHOA HỌC KỸ THUẬT LÂM NGHIỆP VIỆT NAM



Số 123

Năm 2024

NGHIÊN CỨU ỨNG DỤNG MÔ HÌNH HỌC MÁY NHẪM NÂNG CAO HIỆU QUẢ DỰ BÁO CHẤT LƯỢNG NƯỚC SÔNG BA CHẾ, TỈNH QUẢNG NINH

● Nguyễn Thị Hồng^{1*}, Đào Trung Thành¹

TÓM TẮT:

Ứng dụng các mô hình học máy trong dự báo chất lượng nước mặt có thể mang lại hiệu quả cao trong quản lý tài nguyên nước. Một trong những mô hình học máy là LSTM có khả năng phân tích dữ liệu từ các chỉ số của môi trường nước để dự báo chất lượng nước mặt trong tương lai. Ngoài ra, mô hình này có thể được kết nối với các thiết bị cảm biến để thu thập dữ liệu về chất lượng nước mặt trong thời gian thực. Khi dữ liệu được cập nhật liên tục, mô hình có thể phân tích và dự báo chất lượng nước mặt ngay lập tức, giúp nhà quản lý có thể phát hiện sớm các vấn đề về ô nhiễm nguồn nước và đưa ra biện pháp khắc phục kịp thời. Báo cáo nghiên cứu ứng dụng tại sông Ba Chế của huyện Ba Chế, đây là huyện miền núi của tỉnh Quảng Ninh với số lượng lớn đồng bào dân tộc thiểu số sinh sống tại nhiều khu vực có địa hình phức tạp. Nhiều năm qua, người dân vẫn sử dụng nguồn nước tự chảy từ các bể chứa, từ khe, suối, nên chất lượng nước khó đảm bảo. Cần thiết phải có biện pháp quản lý hiệu quả tài nguyên nước trên địa bàn để người dân yên tâm sinh hoạt, sản xuất, ổn định đời sống, vươn lên thoát nghèo.

Từ khóa: Học máy, LSTM, WQI, sông Ba Chế.

I. ĐẶT VẤN ĐỀ

Sông Ba Chế là một trong bốn con sông lớn của tỉnh Quảng Ninh, không chỉ cung cấp nguồn nước quan trọng cho sinh hoạt, nông nghiệp và công nghiệp, mà còn là một tuyến giao thông thủy quan trọng và có tầm ảnh hưởng lớn đến phát triển kinh tế và văn hóa của khu vực. Dòng sông đã tạo ra một môi trường sống đa dạng và phong phú, góp phần vào sự hình thành và phát triển của các cộng đồng dân cư xung quanh. Tuy nhiên, bên cạnh việc phát triển kinh tế và đô thị một cách ồ ạt của địa phương cũng làm ảnh hưởng đến chất lượng nước sông. Kết quả quan trắc giám sát chất lượng nước cho thấy, sông Ba Chế đang bị ô nhiễm hữu cơ, giá trị COD, BOD₅ tại hầu hết các điểm quan trắc đều vượt giới hạn cho phép từ 2 đến 2,3 lần. Phương pháp đánh giá hiện tại thường dựa trên kết quả đo nhanh tại chỗ, lấy mẫu và phân tích trong phòng thí nghiệm với các thông số vật lý, hóa học và sinh học trong nước. Phương pháp này tuy có độ chính xác cao nhưng khá tốn thời gian, kinh phí, trang thiết bị và nhân lực kỹ thuật cao.

Hiện nay, việc áp dụng mô hình học máy (Machine Learning) vào nhiều lĩnh vực khoa học khác nhau đang trở thành xu thế nhằm tăng độ chính xác của các phép toán dự báo và phân loại. Mô hình học máy các mối quan hệ thống kê bậc cao, do đó thường vượt trội so

¹ Trường Đại học Mỏ - Địa chất

* Tác giả liên hệ Email: nguyenthihong@humg.edu.vn

với các phương pháp mô hình hóa và dự báo truyền thống. Trong dự báo chất lượng nước, các thuật toán học máy có thể ước lượng và phân tích các thông số, dựa trên mối quan hệ tuyến tính và phi tuyến tính. Các kỹ thuật học máy cũng đóng một vai trò quan trọng trong việc dự báo chất lượng nước sông nhờ mạng nơ ron nhân tạo (ANN), vectơ hỗ trợ (SVM), rừng ngẫu nhiên (RF), cây quyết định (DT) và học sâu (DL) [1]. Học máy làm tăng hiệu quả đánh giá chất lượng nước mặt bằng chỉ số chất lượng nước (WQI) thông qua các thuật toán có giám sát như dự báo đa lớp đối với thuật toán AdaBoost (AB), thuật toán K-nearest neighbors (KNN), quy trình Gaussian (GP) và hồi quy tăng cường độ dốc cực đại (XGBR) [2, 3]. Ngoài ra, các phương pháp tiếp cận dựa trên máy học như rừng ngẫu nhiên đã thành công trong việc phân loại chất lượng nước sông thành các loại ô nhiễm hoặc ô nhiễm nhẹ với độ chính xác cao, nhấn mạnh tầm quan trọng của thông số DO trong dự đoán chất lượng nước [4]. Hơn nữa, việc tích hợp các thuật toán học sâu như mạng tích chập đồ thị (Graph Convolutional Network - GCN) và mạng thần kinh bộ nhớ ngắn hạn dài (Long Short-Term Memory - LSTM) đã cải thiện đáng kể độ chính xác của các mô hình dự báo chất lượng nước, hỗ trợ quản lý môi trường nước hiệu quả [5].

Ở Việt Nam, nhiều nghiên cứu thường dùng chỉ số chất lượng nước (WQI) để đánh giá chất lượng nước mặt và khả năng sử dụng của nguồn nước đối với từng mục đích khác nhau. Tuy nhiên, việc tính toán WQI từ các thông số quan trắc chất lượng nước còn gặp nhiều hạn chế do cần nhiều thông số quan trắc và tính toán còn tương đối phức tạp. Trong những năm gần đây, nhiều nghiên cứu tại Việt Nam áp dụng thuật toán học máy để dự báo chỉ số chất lượng nước mặt, nâng cao độ chính xác và giảm chi phí giám sát. Các thuật toán học máy, đặc biệt là tăng cường độ dốc (GBoost), được ứng dụng hiệu quả trong ước tính chỉ số chất lượng nước mặt ở đồng bằng sông Hồng, Việt Nam, nâng cao độ chính xác dự báo và hiệu quả chi phí [6]. Kể cả với môi trường nước biển, các thuật toán cây quyết định (DT), rừng ngẫu nhiên (RF), hồi quy tăng cường độ dốc (GBR) và

hồi quy tăng cường độ dốc Ada (ABR) sử dụng ảnh viễn thám Sentinel-2 đối với khu vực ven biển tỉnh Bình Định. Nghiên cứu cho thấy thuật toán rừng ngẫu nhiên (RF) dự báo các thông số chất lượng nước biển chính xác nhất so với dữ liệu đo tại hiện trường ($R^2 = 0,59 \div 0,74$) [7]. Việc tìm ra thuật toán để dự báo hiệu quả WQI là rất quan trọng và cần thiết nhằm phục vụ tốt hơn cho công tác đánh giá chất lượng nước mặt cho vùng nghiên cứu. Nghiên cứu tại bán đảo Cà Mau đã áp dụng phương pháp Bayes (BMA) để lựa chọn các thông số chất lượng nước tối ưu (pH, BOD₅, PO₄ và Coliform) để tính toán WQI. Kết quả cho thấy sử dụng các thuật toán tăng cường độ dốc (GBoost) có kết quả dự báo chính xác nhất ($R^2=0,973$), tăng cường độ dốc cực đại (XGBoost) có $R^2=0,966$, cây quyết định có $R^2=0,944$ và tăng cường độ dốc nhẹ (LGBoost) có $R^2=0,928$ [8].

Như vậy, các nghiên cứu trước đây về ứng dụng học máy trong dự báo chất lượng nước chủ yếu tập trung vào dự báo các thông số chất lượng nước và tính toán chỉ số chất lượng nước (WQI). Một số nghiên cứu còn kết hợp thuật toán học máy với mạng giám sát thời gian thực và cho kết quả rất khả quan. Tuy nhiên, các nghiên cứu này chủ yếu trên thế giới, còn tại Việt Nam, có rất ít nghiên cứu đánh giá tiềm năng của thuật toán học máy trong dự báo chỉ số chất lượng nước mặt (WQI) dựa trên dữ liệu đầu vào nhằm giảm chi phí giám sát chất lượng nước mặt. Vì vậy, việc nghiên cứu ứng dụng các mô hình học máy để dự báo chỉ số chất lượng nước mặt tại sông Ba Chẽ là cần thiết và góp phần đưa ra các phương pháp tính toán chỉ số chất lượng nước mặt hiệu quả và tiết kiệm hơn.

II. THU THẬP MẪU VÀ PHƯƠNG PHÁP NGHIÊN CỨU

1. Khu vực nghiên cứu và lấy mẫu

Vị trí quan trắc lấy mẫu được thống kê trong bảng 1 và thể hiện trên hình 2 trong 2 đợt khảo sát mùa khô và mùa mưa giai đoạn 2017-2022. Mẫu được lấy theo TCVN 6663-6-2018 (ISO 5667-6-2014) trong lấy mẫu nước sông và TCVN 6663-3-2016 (ISO 5667-3-2012) trong bảo quản và xử lý mẫu nước (Bảng 1, Hình 1).

2. Phương pháp thu thập dữ liệu

Dữ liệu cho bài báo này chủ yếu tập trung vào việc thu thập thông số chất lượng nước như pH, TSS, NTU, BOD₅, COD, DO, NO₃⁻, NH₄⁺, NO₂⁻, PO₄³⁻, Zn, Fe, Cu, Pb, Hg, và Coliform.

3. Xử lý số liệu

Đánh giá mẫu nước theo QCVN 08:2023/BTNMT, mức A - Quy chuẩn kỹ thuật Quốc gia về chất lượng nước phục vụ mục đích cấp nước sinh hoạt.

Quy trình tính toán chỉ số WQI theo hướng dẫn trong Quyết định số 1460/QĐ-TCMT của Tổng cục Môi trường trong giai đoạn từ 2017 đến 2023.

4. Phương pháp học máy

Mạng nơ ron hồi quy (RNN) là một loại thuật toán học máy phù hợp với dữ liệu liên tục, đa biến có thể xử lý dữ liệu tuần tự và đã được chứng minh hiệu quả trong dự báo chất lượng nước.

Mạng nơ ron bộ nhớ ngắn-dài (LSTM) kế thừa thuật toán của RNN để xử lý các chuỗi dữ liệu theo thời gian (sequence/time-series), phù hợp trong việc dự báo chất lượng nước thời gian thực. LSTM vừa học được các thông tin từ state gần nó (short term memory) giống RNN và các thông tin từ state ở trước nó rất xa (long term memory) điều mà RNN không làm được. Ở state thứ t của mô hình LSTM:

Output: C_t , h_t , trong đó c là cell state, h là hidden state.

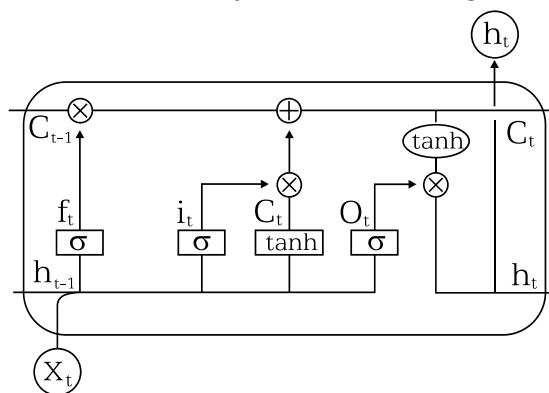
Input: C_{t-1} , h_{t-1} , x_t trong đó x_t là input ở state thứ t của

Bảng 1. Tọa độ điểm khảo sát thu thập mẫu nước mặt trên sông Ba Chẽ

SHM	Vị trí	Tọa độ VN2000 (107°45', múi chiều 3°)	
		X	Y
NM01	Chân cầu ngầm xã Thanh Sơn	2356132	441516
NM02	Phía sau điểm hợp lưu với sông Làng Cống tại xã Đồn Đạc	2353555	448881
NM03	Điểm lấy nước về nhà máy nước Ba Chẽ	2353478	450465
NM04	Đập tràn thị trấn Ba Chẽ	2353426	451224
NM05	Chân cầu Ba Chẽ 2	2353207	452971



Hình 1. Các điểm lấy mẫu nước trên sông Ba Chẽ



Hình 2. Cấu trúc của mô hình LSTM

mô hình, c_{t-1} , h_{t-1} là output của layer trước.

(Hình 2).

Forget gate: $f_t = \sigma(W_f \times [h_{t-1}, x_t] + b_f)$, $\tilde{c}_t = \tanh(W_c \times [h_{t-1}, x_t] + b_c)$. (1)

Input gate:

$i_t = \sigma(W_i \times [h_{t-1}, x_t] + b_i)$. (2)

Output gate:

$o_t = \sigma(W_o \times [h_{t-1}, x_t] + b_o)$. (3)

Long memory:

$$c_t = f_t \times c_{t-1} + i_t \times \tilde{c}_t. \quad (4)$$

Short memory:

$$h_t = o_t \times \tanh(c_t). \quad (5)$$

Ghi chú: σ , $\tanh$ là bước dùng sigma, $\tanh$ activation function; $\times$, $+$ là phép nhân, phép cộng ma trận; f_t , i_t , và o_t tương ứng với forget gate, input gate và output gate.

Việc đánh giá độ chính xác của mô hình trong dự báo chỉ

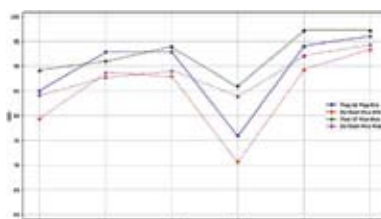
số chất lượng nước tại khu vực nghiên cứu bằng các chỉ số như: Sai số tuyệt đối trung bình (MAE); sai số bình phương trung bình (MSE); căn bậc hai của giá trị trung bình của sai số bình phương (RMSE); R-square (R^2).

III. KẾT QUẢ VÀ THẢO LUẬN

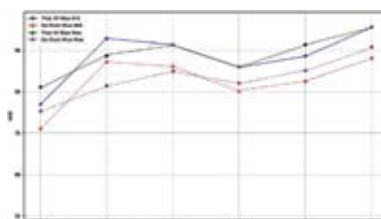
1. Kết quả tính toán chỉ số chất lượng nước mặt

Kết quả tính toán giá trị WQI của các mẫu phục vụ huấn luyện mô hình được thể hiện trong giai đoạn từ năm 2017 đến 2022 cho thấy, diễn biến chất lượng nước sông Ba Chẽ có sự biến động theo mùa, chất lượng nước mùa khô thấp hơn mùa mưa. Trong giai đoạn giám sát chất lượng nước khá ổn định ở mức tốt đến rất tốt với giá trị WQI dao động trong khoảng 76 đến 97, đáp ứng tiêu chuẩn cấp nước sinh hoạt. Trong đó, tại các điểm quan trắc NM-02 là điểm trước khu vực lấy nước về nhà máy nước sạch Ba Chẽ có chất lượng tốt hơn ở các điểm phía hạ nguồn khi chưa chịu ảnh hưởng nhiều của việc xả nước thải sinh hoạt, dịch vụ và công nghiệp trên địa bàn.

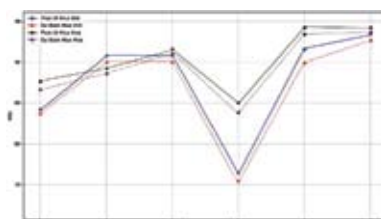
Dựa trên kết quả thu được, lựa chọn các thông số để xây dựng mô hình học máy. Nghiên cứu đã xây dựng mô hình từ các thông số (BOD_5 , NH_4^+ , PO_4^{3-} , TSS, NTU và Coliform) gồm dữ liệu đào tạo (24 mẫu) và thử nghiệm (36 mẫu) để dự báo chất lượng nước mặt (WQI) theo các mô hình là Mạng nơ ron hồi quy (RNN) và Mạng nơ ron bộ nhớ ngắn-dài (LSTM).



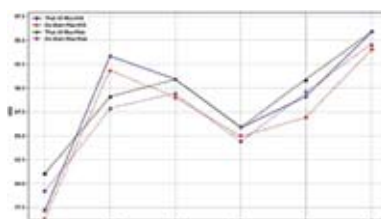
a-RNN-NM02



b-RNN-NM05



c-LSTM-NM02



d-LSTM-NM05

Hình 3. So sánh giữa WQI dự báo và thực tế đối với dữ liệu thử nghiệm theo các mô hình học máy

Bảng 2. Đánh giá các mô hình dự báo chỉ số chất lượng nước mặt khu vực nghiên cứu

Học máy	Thông số đầu vào	Đầu ra	Đánh giá			
			MAE	MSE	RMSE	R^2
RNN	BOD_5 , NH_4^+ , PO_4^{3-} , TSS, NTU, Coliform	WQI	5,7	78,6	8,85	0,82
LSTM	BOD_5 , NH_4^+ , PO_4^{3-} , TSS, NTU, Coliform	WQI	5,5	73,0	8,52	0,86

Kết quả báo cáo WQI và biểu đồ so sánh giữa giá trị WQI dự đoán và đo được đối với tập dữ liệu thực nghiệm theo các mô hình học máy (Hình 3).

2. Kết quả đánh giá mô hình

Kết quả đánh giá mô hình học máy dự vào 4 tiêu chí (MAE, MSE và RMSE và R^2) để dự đoán chỉ số chất lượng nước mặt tại khu vực nghiên cứu (Bảng 2).

Theo kết quả tính toán, mô hình học máy LSTM có độ chính xác cao hơn mô hình RNN. Trong đó, mô hình mô hình RNN có R^2 là 0,82; giá trị lỗi là 5,70; 78,6; 8,85 và mô hình LSTM có R^2 là 0,86; giá trị lỗi là 5,50; 73,0; và 8,52. Có thể thấy, các mô hình học máy áp dụng trong

bài báo đều có thể dự đoán tốt WQI cho khu vực nghiên cứu (R^2 đều lớn hơn 80%). Do đó, việc ứng dụng các mô hình học máy trong tính toán WQI cho sông Ba Chẽ là phù hợp và tiết kiệm chi phí. Từ đó, giúp tăng hiệu quả trong công tác giám sát chất lượng nước sông phục vụ cấp nước sinh hoạt, ổn định đời sống và phát triển kinh tế-xã hội trên địa bàn huyện. Đồng thời, nâng cao công tác quản lý tài nguyên nước, đặc biệt là tại các vùng có điều kiện khó khăn như huyện miền núi Ba Chẽ.

3. Kết quả phân vùng chất lượng nước sông Ba Chẽ

Kết quả nội suy lưới số liệu chất lượng nước tính toán chỉ số WQI và dự báo bằng mô

hình LSTM tại các điểm khảo sát tạo lưới số liệu chất lượng nước tương ứng. Các mặt cắt đại diện trên sông Ba Chẽ được tính toán nội suy hiển thị theo đa giác theo các màu tương ứng với chất lượng nước quy ước. Tiến hành chồng các lớp bản đồ trên cơ sở nền địa hình tại khu vực nghiên cứu thu được phân vùng chất lượng nước sông Ba Chẽ theo hai mùa (Hình 4).



a-Mùa khô



b-Mùa mưa

Hình 4. Phân vùng dự báo chất lượng nước sông Ba Chẽ theo hai mùa tại khu vực nghiên cứu

Phân bố mức chất lượng nước sông Ba Chẽ cả hai mùa cho thấy chất lượng nước tốt đến rất tốt. Trong đó, chất lượng nước sông vào mùa mưa được cải thiện hơn do một lượng nước lớn từ các nhánh sông Quách, sông Đoảng, sông Làng Cống và thượng nguồn chảy về và khả năng tự làm sạch của sông cao. Nhìn chung chất lượng nước sông vẫn đảm bảo về chất lượng nước mặt phục vụ mục đích cấp nước sinh hoạt trên địa bàn huyện Ba Chẽ theo QCVN 08:2023/BTNMT (mức A). Tuy nhiên, vào mùa khô cần các biện pháp xử lý phù hợp trước khi dùng làm nước cấp sinh hoạt cho người dân.

IV. KẾT LUẬN

Kết quả nghiên cứu đã đưa ra cơ sở khoa học cho việc tính toán chỉ số chất lượng nước mặt bằng học máy và đã phát huy hiệu quả trong dự báo chỉ số chất lượng nước (WQI) sông Ba Chẽ. Báo cáo đã xây dựng mô hình học máy dự báo WQI với 5 tham số được chọn, bao gồm: BOD_5 , NH_4^+ , PO_4^{3-} , TSS, NTU và Coliform. Theo kết quả tính toán dự báo WQI của các mô hình học máy cho thấy, mô hình LSTM có độ chính xác hơn mô hình RNN và cả hai mô hình đều có độ chính xác cao, dao động từ 82-86%.

Tuy nhiên, bộ dữ liệu sử dụng trong bài báo còn hạn chế chưa đạt được độ chính xác cao như mong muốn. Vì vậy, cần tiếp tục bổ sung thêm bộ dữ liệu về nguồn xả thải, thủy động lực học dòng chảy và đặc biệt là những yếu tố nhân tạo gây ảnh hưởng đến chất lượng nước sông. Từ đó, giúp nâng cao độ chính xác hơn nữa của mô hình học máy trong công tác dự báo và nâng cao hiệu quả quản lý môi trường sông để đảm bảo chất lượng nước phục vụ mục đích cấp nước sinh hoạt bền vững trên địa bàn huyện Ba Chẽ.

Lời cảm ơn:

Kết quả bài báo được hoàn thành dưới sự hỗ trợ kinh phí và số liệu của đề tài “Ứng dụng mô hình LSTM trong dự báo chất lượng nước sông Ba Chẽ phục vụ cấp nước sinh hoạt cho huyện Ba Chẽ, tỉnh Quảng Ninh” - Mã số T24-24. Các tác giả xin trân trọng cảm ơn.

V. TÀI LIỆU THAM KHẢO

1. Sanja Cojbasic, Sonja Dmitrasinovic, Marija Kostic, Maja Turk Sekulic, Jelena Radonic, Ana Dodig, Milan Stojkovic, (2023). Application of machine learning in river water quality management: a review. *Water Science & Technology*, 88 (9): 2297–2308.
2. Masood, A.; Niazkar, M.; Zakwan, M.; Piraei, R. (2023). A Machine Learning-Based Framework for Water Quality Index Estimation in the Southern Bug River. *Water*, 15 (20), 3543.
3. Anita M, M. G. Dinesh, C. Lakshmipriya, V. Sharmila, A. Muthuram, Siva Subramanian R, (2023). Water Quality Prediction using Machine Learning: A Comparative Study. *Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)*, 348-353.
4. Nasaruddin, N.; Ahmad, A.; Zakaria, S. F.; Ul-Saufie, A. Z.; & Osman, M. S. (2023). Predicting Kereh River's Water Quality: A comparative study of machine learning models. *Environment - Behaviour Proceedings Journal*, 8(SI15), 213–219.
5. Juan Huan, Wenjie Liao, Yongchun Zheng, Xiangen Xu, Hao Zhang, Bing Shi, (2023). A deep learning model with spatio - temporal graph convolutional networks for river water quality prediction. *Water Science & Technology: Water Supply*, 23(7), 2940-2957.
6. Nguyen Quoc Son, Nguyen Cam Linh, Le Thi Phuong Quynh, and Le Phuong Thu, (2023). Application of Machine Learning Algorithms in Studying Water Quality in the Red River System. *VNU Journal Of Science: Earth And Environmental Sciences*, 39 (2), 52-63.
7. Quang Nguyen Hong, Nguyen Tran Dinh, Nguyen Tran Dien, and Le Thanh Son, (2023). Calibration of Sentinel - 2 Surface Reflectance for Water Quality Modelling in Binh Dinh's Coastal Zone of Vietnam. *Sustainability*, 15(2), 1410.
8. Nguyễn Đức Phong, Hà Hải Dương (2023). Nghiên cứu ứng dụng các mô hình học

máy để dự báo chỉ số chất lượng nước mặt vùng bán đảo Cà Mau. *Tạp chí Khoa học và Công nghệ Thủy lợi* 76.

APPLYING MACHINE LEARNING TO IMPROVE THE EFFICIENT WATER QUALITY PREDICTION FOR BA CHE RIVER, QUANG NINH PROVINCE

Nguyen Thi Hong¹, Dao Trung Thanh¹
¹Hanoi University of Mining and Geology

SUMMARY:

Applications of machine learning models in predicting surface water quality have brought efficiency to water resource management. Machine learning models can predict surface water quality through data analysis from physical, chemical, and biological parameters. In addition, it can be connected to sensor devices to collect data on surface water quality in real-time. These models can immediately analyze and predict surface water quality, so managers can detect water pollution problems early and take timely corrective measures. The models used in the article are RNN and LSTM, which have been used by many scientists and have proven their effectiveness. The scope of the article's research is the Ba Che river area of the Ba Che mountainous district with a large number of ethnic minorities living in many areas with complex terrain. The scope of the article's research is the Ba Che river area of the Ba Che mountainous district with complex terrain, sparse population, and difficult life. In recent years, people have used water from reservoirs, streams and streams, so water quality is difficult to ensure. It is necessary to have measures to effectively manage water resources so that people can feel secure in living, producing, stabilizing their lives, and escaping poverty.

Keywords: *Machine learning, LSTM, WQI, Ba Che River.*

Người phản biện: TS. Nguyễn Thị Hòa

Ngày nhận bài: Tháng 10/2024

Ngày phản biện thông qua: Tháng 10/2024

Ngày duyệt đăng: Tháng 10/2024