XÂY DỰNG MÔ HÌNH MÔ PHỎNG COD VÀ BOD CHO LƯU VỰC SÔNG BUÔNG BẰNG THUẬT TOÁN TRÍ TUỆ NHÂN TẠO

Building Artificial Intelligence-based model for COD and BOD simulation in the Buong River Basin

Nguyễn Thị Diễm Thúy[1,2], Phạm Ngọc Đông Thi[3], Nguyễn Phúc Hiếu[3,4], Phạm Thị Thảo Nhi [5], Đào Nguyên Khôi[3]

(1) Viện Môi Trường và Tài Nguyên, ĐHQG-HCM

(2) Viện Khí Tượng Thủy Văn Hải Văn và Môi Trường

(3) Trường ĐH Khoa học Tự nhiên, ĐHQG-HCM

(4) Công ty TNHH ERM Việt Nam

(5) Viện Khoa học và Công nghệ Tính toán

Tóm tắt

Mục tiêu của nghiên cứu là xây dựng mô hình trí tuệ nhân tạo (AI) mô phỏng BOD và COD cho sông Buông giai đoạn 2010-2017 trên địa bàn tỉnh Đồng Nai. Nghiên cứu đã áp dụng và đánh giá 8 thuật toán trí tuệ nhân tạo, bao gồm LR, SVR, LinearSVR, SGDR, RidgeCV, xgboost, RFR và mạng nơ-ron MLP, và lựa chọn thuật toán tốt nhất để mô phỏng BOD và COD cho khu vực nghiên cứu. Kết quả nghiên cứu cho thấy thuật toán RFR cho kết quả mô phỏng tốt nhất cho cả 2 mô hình mô phỏng BOD và COD, và các mô hình này đều cho kết quả mô phỏng BOD và COD ở mức khá.

Từ khóa: trí tuệ nhân tạo, BOD, COD, sông Buông, Đồng Nai

Abstract

The aim of the study was to build an artificial intelligence (AI) model for simulating BOD and COD in the period 2010-2017 for the Buong River in Dong Nai province. The study applied and evaluated 8 AI algorithms, including LR, SVR, LinearSVR, SGDR, RidgeCV, xgboost, RFR, and MLP neural network, and selected the best algorithm to simulate BOD and COD for the study area. The results show that the RFR algorithm gives the best simulation results for both AI models for simulating BOD and COD, and these models give good simulation results of BOD and COD.

Keywords: artificial intelligence, BOD, COD, Buong River, Dong Nai province

1. Đặt vấn đề

Trong những năm gần đây, các công cụ mô phỏng và dự báo chất lượng nước với độ chính xác cao đóng vai trò rất quan trọng trong công tác quản lý tài nguyên nước (trữ lượng và chất lượng) [1]. Mặc dù có nhiều phương pháp khác nhau để mô phỏng và dự đoán chất lượng nước như mô hình khái niệm, mô hình vật lý, mô hình số, mô hình thống kê, v.v…, mô hình

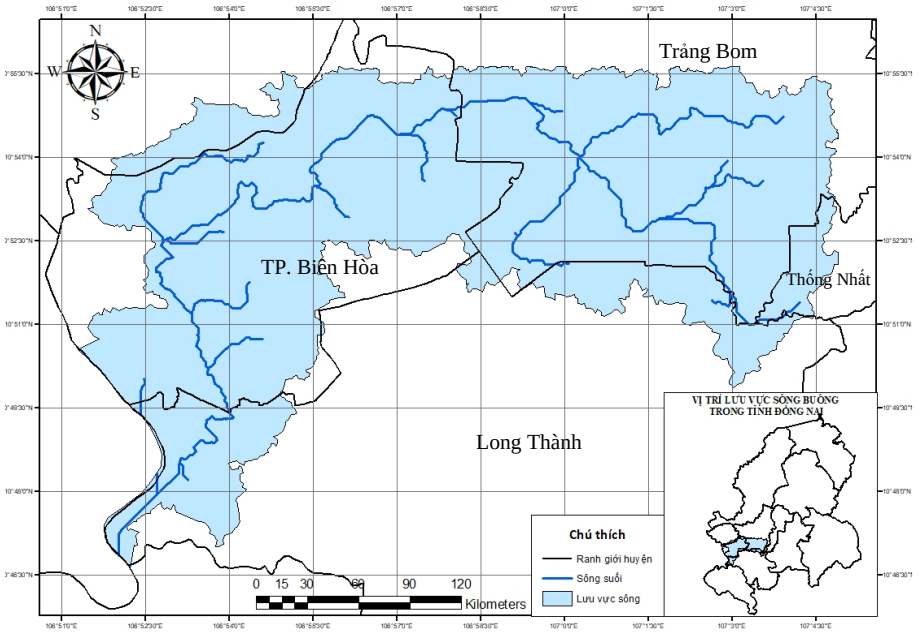
trí tuệ nhân tạo (AI) là một lựa chọn vì tính đơn giản và tính chính xác của kết quả mô phỏng. Một điểm mạnh là mô hình AI có khả năng mô phỏng các hiện tượng phức tạp mà không cần hiểu rõ về bản chất vấn đề. Do đó, việc sử dụng phương pháp tiếp cận AI trong mô phỏng chất lượng nước ngày càng gia tăng và thu hút sự quan tâm của các nhà nghiên cứu đặc biệt trong kỷ nguyên dữ liệu lớn và cuộc cách mạng công nghệ 4.0.

Một số nghiên cứu ứng dụng AI trong đánh giá chất lượng nước điển hình có thể kể đến nghiên cứu của Abba và cộng sự (2017) về mô phỏng chất lượng nước sông ở thành phố Agra (Ấn Độ) sủ dụng kỹ thuật hồi quy MLR (hồi quy tuyến tính đa biến), mạng nơ-ron nhân tạo (ANN), và hệ mờ ANFIS. Kết quả cho thấy cả mô hình ANN và ANFIS đều cho kết quả mô phỏng DO tốt hơn mô hình MLR [2]. Một nghiên cứu khác của Ahmed và cộng sự (2019) về dự báo chất lượng nước cho hồ nước Rawal (Pakistan) dùng các thuật toán hồi quy (Gaussian Naïve Bayes, Logistic Regression, Stochastic Gradient Descent, KNN, Decision Tree, Random Forest, SVM, Gradient Boosting Classifier, và Bagging Classifier) và học máy (mạng nơ-ron MLP) cũng cho thấy mô hình mạng nơ-ron MLP cho kết quả mô phỏng WQI là tốt nhất với độ chính xác R^2 = 0,8507 [3]. Hussain và Khan (2020) sử dụng các thuật toán học máy (mạng nơ-ron MLP, support vector regression (SVM), và random forest (RF)) dự báo dòng chảy tháng cho khu vực sông Hunza (Pakistan) giai đoạn 1962-2008. Kết quả cho thấy mô hình RF cho kết quả mô phỏng tốt hơn so với mô hình SVR và MLP [4].

Mặc dù ứng dụng AI trong mô phỏng chất lượng nước được nghiên cứu nhiều trên thế giới, nhưng các nghiên cứu này ở Việt Nam vẫn còn hạn chế, đặc biệt là ở khu vực hệ thống sông Đồng Nai hiện nay vẫn chưa có nghiên cứu nào được thực hiện. Mục tiêu của nghiên cứu là xây dựng mô hình dự báo chất lượng nước (thông số COD và BOD) dựa vào thuật toán trí tuệ nhân tạo với trường hợp nghiên cứu điển hình ở lưu vực sông Buông.

2. Khu vực nghiên cứu

Sông Buông là một phụ lưu của sông Đồng Nai có chiều dài 56 km và tổng lượng dòng chảy năm khoảng 412×10^6 m^3 (Hình 1).



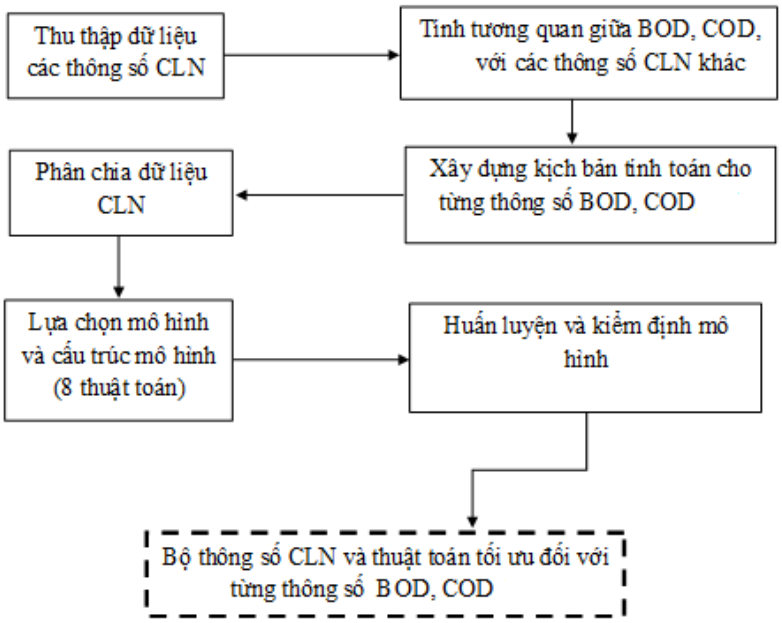
Hình 1. Vị trí lưu vực sông Buông

Tổng diện tích lưu vực sông Buông khoảng 478,5 km2 với mật độ sông ngòi khoảng 0,67 km/km^2. Lưu vực có lượng mưa trung bình khoảng 1.800 mm và nhiệt độ khoảng 25-26^{o}C. Lưu vực nằm trên khu vực ảnh hưởng bởi khí hậu nhiệt đới gió mùa nên khí hậu được chia làm 2 mùa: mùa mưa (tháng 11 đến tháng 04) và mùa khô (tháng 05 đến tháng 10). Bên cạnh đó, hình thái lưu vực sông Buông có hình lá và địa hình khá phẳng. Thổ nhưỡng chủ yếu là đất đỏ bazan nên rất phù hợp cho phát triển nông nghiệp (diện tích đất nông nghiệp chiếm 80% diện tích lưu vực). [5]

3. Phương pháp nghiên cứu

3.1. Các bước xây dựng mô hình AI

Trong nghiên cứu này, 08 thuật toán trí tuệ nhân tạo bao gồm LR, SVR, LinearSVR, SGDR, RFR, xgboost, RidgeCV, và MLP được sử dụng để xây dựng mô hình mô phỏng BOD, COD. Hình 2 trình bày quy trình xây dựng mô hình AI mô phỏng BOD và COD. Đầu tiên, số liệu chất lượng nước thu thập được sẽ được xử lý và đánh giá chất lượng dữ liệu. Sau đó, ma trận tương quan giữa các thông số chất lượng nước được thực hiện để tìm ra mối quan hệ giữa thông số BOD và COD với các thông số chất lượng nước khác (DO, BOD, COD, nhiệt độ, pH, độ đục, độ dẫn, TSS, NH_4, độ mặn, NO_2, NO_3, PO_4, Fe, Ecoli, và coliform). Dựa vào ma trận tương quan này, xây dựng kịch bản tính toán với các thông số đầu vào cho các mô hình mô phỏng chất lượng nước cho thông số BOD và COD. Sau đó, thực hiện phân chia dữ liệu thành 2 phần bao gồm dữ liệu huấn luyện và dữ liệu kiểm định. Cuối cùng, thực hiện huấn luyện mô hình (trainning phase) dựa trên bộ dữ liệu huấn luyện nhằm tìm ra các tham số của mô hình, sau đó tiến hành kiểm tra lại kết quả mô phỏng (testing phase) dựa trên bộ dữ liệu kiểm định nhằm đánh giá hiệu quả mô phỏng của các mô hình.



Hình 2. Trình tự xây dựng mô hình AI mô phỏng BOD và COD

Trong nghiên cứu này, bộ dữ liệu chất lượng nước giai đoạn 2009-2017 ở lưu vực sông Buông thu thập từ Trung tâm Quan trắc và Kỹ thuật môi trường tỉnh Đồng Nai được sử dụng để xây dựng mô hình AI.

3.2. Đánh giá hiệu quả mô phỏng

Hiệu quả mô phỏng của mô hình được đánh giá bằng phương pháp đồ thị và phương pháp thống kê để so sánh chất lượng và độ tin cậy của kết quả mô phỏng với số liệu thực đo. Trong nghiên cứu này, ba chỉ số thống kê được sử dụng đển đánh giá hiệu quả mô phỏng của mô hình bao gồm: hệ số xác định (R^2), sai số tương đối RE, và sai số quân phương RMSE. R^2 càng tiến đến 1 thì mô hình càng đạt hiệu quả cao, RE và RMSE càng tiến đến 0 thì sai lệch mô hình càng thấp.

+ Hệ số tương quan (R^2): $R^2 = \left[\frac{\sum_{i=1}^{n}(O_i - \bar{O})(P_i - \bar{P})}{\sqrt{\sum_{i=1}^{n}(O_i - \bar{O})^2}\sqrt{\sum_{i=1}^{n}(P_i - \bar{P})^2}}\right]^2$
+ Sai số tương đối (RE): $RE = \left|\frac{O_i - P_i}{O_i}\right| x\ 100$
+ Sai số quân phương (RMSE): $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(O_i - P_i)^2}$

Trong đó: O_i: số liệu thực đo; $\bar{O}$: trung bình số liệu thực đo; P_i: số liệu mô phỏng; $\bar{P}$: trung bình số liệu mô phỏng; n: tổng số dữ liệu thực đo/mô phỏng

4. Kết quả và thảo luận

4.1. Đánh giá mối tương quan của BOD, COD với các thông số chất lượng nước

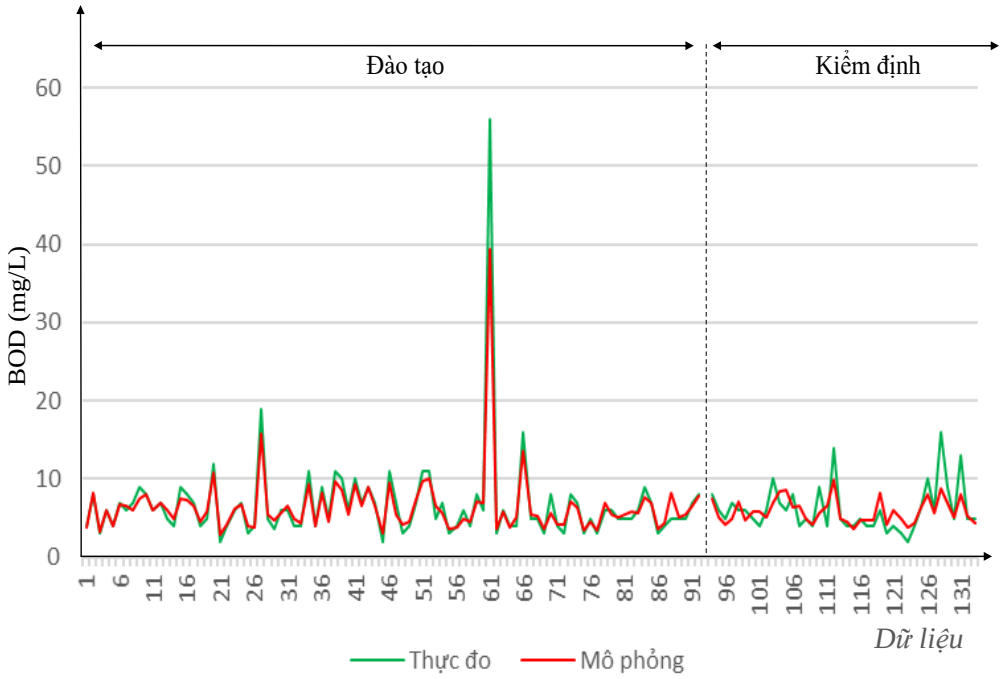
Mối quan hệ giữa BOD, COD và các thông số chất lượng nước khác (BOD, COD, DO, nhiệt độ, pH, độ đục, độ dẫn, TSS, NH_4, độ mặn, NO_2, NO_3, PO_4, Fe, Ecoli, và Coliform) được xác định bằng ước tính hệ số tương quan riêng phần. Kết quả cho thấy 04 thông số chất lượng nước có mối tương quan mạnh với BOD như sau: độ mặn (R = 0,83), NH_4 (R = 0,72), PO4 (R = 0,71), và Fe (R = 0,70). Dựa trên mối tương quan với 4 thông số đã xác định, mô hình AI mô phỏng BOD được xây dựng với 4 thông số đầu vào (độ mặn, NH_4, PO_4, và Fe). Bên cạnh đó, 2 thông số chất lượng nước có mối tương quan mạnh với COD được xác định là TSS (R = 0,83) và độ đục (R = 0,79).

4.2. Đánh giá mô hình BOD

Hiệu quả mô phỏng BOD với 4 thông số đầu vào (độ mặn, NH4, PO4, và Fe) bằng 8 thuật toán trí tuệ nhân tạo được trình bày ở Bảng 1. Bộ dữ liệu chất lượng nước giai đoạn 2010 – 2017 được chia thành hai phần bao gồm 70% dữ liệu được dùng để huấn luyện và 30% dữ liệu được dùng để kiểm định mô hình. Quá trình phân chia dữ liệu trong đề tài được thực hiện bằng việc thử và sai để tìm ra bộ dữ liệu huấn luyện và kiểm định tối ưu nhất cho các mô hình. Kết quả cho thấy 8 thuật toán đều cho kết quả mô phỏng ở mức khá tốt ở giai đoạn huấn luyện, với các chỉ số thống kê R^2 = 0,73 ÷ 0,99; RE = 0,07 ÷ 0,32 và RMSE = 0,62 ÷ 3,10. Ở giai đoạn kiểm định, chỉ có RFR cho kết quả mô phỏng BOD ở mức thỏa mãn. Hình 2 trình bày đồ thị so sánh giá trị BOD mô phỏng và BOD thực đo bằng thuật toán RFR (thuật toán cho kết quả mô phỏng tốt nhất). Nhìn chung, thuật toán RFR cho kết quả phù hợp mô phỏng BOD với 4 thông số đầu vào là độ mặn, NH4, PO4 và Fe.

Bảng 1. Hiệu quả mô phỏng BOD với 8 thuật toán AI

Thuật toán AI	Huấn luyện (70% dữ liệu)			Kiểm định (30% dữ liệu)		
	R^2	RE	RMSE	R^2	RE	RMSE
LR	0,90	0,22	1,9	0,45	0,27	2,20
SVR	0,89	0,22	1,98	0,40	0,27	2,31
LinearSVR	0,88	0,24	2,01	0,39	0,28	2,31
SGDR	0,87	0,26	2,16	0,27	0,30	2,54
RidgeCV	0,84	0,28	2,40	0,12	0,33	2,79
MLP	0,73	0,32	3,10	0,05	0,35	2,90
Xgboost	0,99	0,07	0,62	0,45	0,27	2,21
RFR	0,90	0,13	1,87	0,52	0,25	2,07



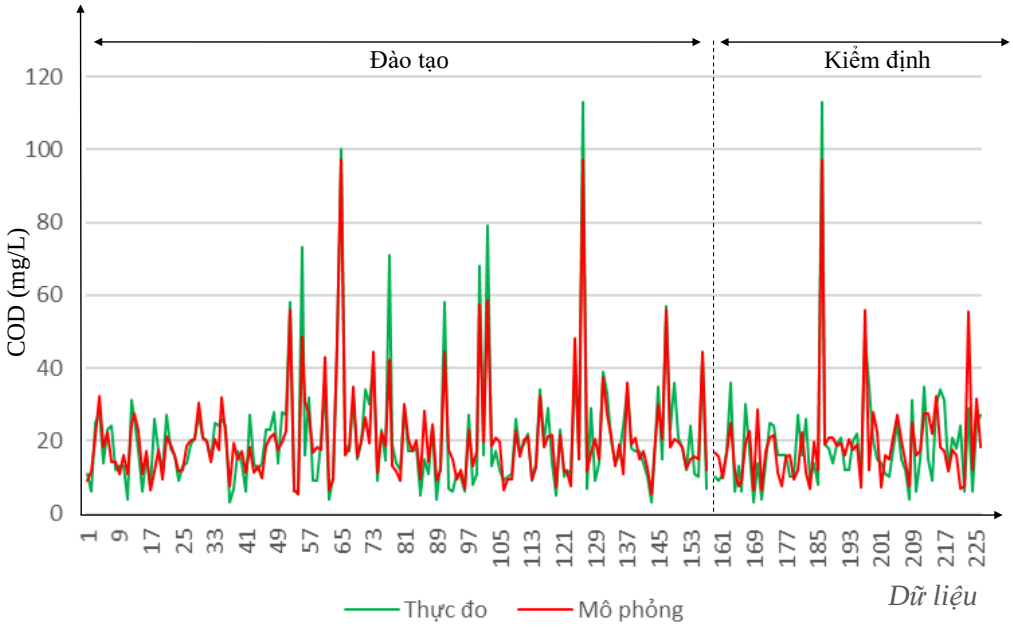
Hình 2. Đồ thị so sánh BOD mô phỏng và thực đo với thuật toán RFR

4.3. Đánh giá mô hình COD

Hiệu quả mô phỏng COD với 2 thông số đầu vào (TSS và độ đục) bằng 8 thuật toán AI được trình bày ở Bảng 2. Kết quả cho thấy 8 thuật toán đều cho kết quả mô phỏng ở mức khá cho cả giai đoạn huấn luyện (với các chỉ số thống kê R^2 = 0,64 ÷ 0,86; RE = 0,21 ÷ 0,32 và RMSE = 6,16 ÷10,0) và giai đoạn kiểm định (R^2 = 0,66 ÷ 0,74; RE = 0,3 ÷ 0,33 và RMSE = 7,51 ÷ 8,56). Hình 3 trình bày đồ thị so sánh giá trị COD mô phỏng và COD thực đo bằng thuật toán RFR (thuật toán cho kết quả mô phỏng tốt nhất). Nhìn chung, thuật toán RFR cho kết quả mô phỏng COD tốt nhất với 2 thông số đầu vào là TSS và độ đục.

Bảng 2. Hiệu quả mô phỏng COD với 8 thuật toán AI

Thuật toán AI	Huấn luyện (70% dữ liệu)			Kiểm định (30% dữ liệu)		
	R^2	RE	RMSE	R^2	RE	RMSE
LR	0,68	0,32	9,43	0,74	0,33	7,55
SVR	0,68	0,30	9,42	0,66	0,32	8,56
LinearSVR	0,68	0,32	9,44	0,73	0,33	7,70
SGDR	0,68	0,32	9,43	0,74	0,33	7,55
RidgeCV	0,68	0,32	9,43	0,74	0,33	7,55
MLP	0,64	0,32	10,0	0,70	0,30	8,02
Xgboost	0,74	0,28	8,41	0,74	0,32	7,51
RFR	0,86	0,21	6,16	0,69	0,33	8,21



Hình 3. Đồ thị so sánh COD mô phỏng và thực đo với thuật toán RFR

5. Kết luận

Nghiên cứu đã sử dụng 8 thuật toán trí tuệ nhân tạo bao gồm LR, SVR, LinearSVR, SGDR, RidgeCV, xgboost, RFR và mạng nơ-ron MLP để mô phỏng 2 thông số chất lượng nước là BOD và COD tại lưu vực sông Buông, tỉnh Đồng Nai trong giai đoạn 2010 - 2017. Kết quả nghiên cứu cho thấy thuật toán RFR cho kết quả mô phỏng tốt nhất cho cả 2 mô hình mô phỏng BOD và COD, và các mô hình này đều cho kết quả mô phỏng BOD và COD ở mức khá. Cụ thể, mô hình BOD với 4 thông số đầu vào (độ mặn, NH4, PO4, và Fe) cho kết quả mô phỏng với $R^2 > 0,52$ và mô hình COD với 2 thông số đầu vào (TSS và độ đục) với kết quả mô phỏng với $R^2 > 0,69$.

Lời cám ơn

Nghiên cứu này được tài trợ bởi Sở Khoa Học và Công Nghệ Tp.HCM và được thực hiện bởi Viện Khoa học và Công nghệ Tính toán (ICST) thông qua Hợp đồng thực hiện nhiệm vụ khoa học và công nghệ số 11/2020/HĐ-QPTKHCN ngày 22 tháng 04 năm 2020.

Tài liệu tham khảo

[1] Đào Nguyên Khôi, Giáo trình Cơ sở mô hình hóa chất lượng nước mặt, Nhà Xuất bản xây dựng, 2017.

[2] Abba SI, Hadi SJ, Abdullahi J, "(2017). River water modeling prediction using multi-linear rgression artificial neural network, and adaptive neuro-fuzzy inference system techniques.," *Procedia Computer Science,* vol. 120, pp. 75-82, 2017.

[3] Ahmed U, Mumtaz R, Anwar H, Shah AA, Irfan R, García-Nieto J, "Efficient water quality prediction using supervised machine learning," *Water,* vol. 11, p. 2210, 2019.

[4] Hussain D, Khan AA, "Machine learning techniques for monthly river flow forecasting of Hunza River, Pakistan," *Earth Science Informatics,* vol. 13, pp. 939-949, 2020.

[5] Khoi DN, Nguyen VT, Sam TT, Nhi PTT. , "Evaluation on Effects of Climate and Land-Use Changes on Streamflow and Water Quality in the La Buong River Basin, Southern Vietnam," *Sustainability,* vol. 24, no. 11, p. 7221, 2019.