DOI: https://doi.org/10.62985/j.huit_ojs.vol25.no2.182

TIẾP CẬN CÁC PHƯƠNG PHÁP PHÂN LỚP DỮ LIỆU TRONG DỰ BÁO CHẤT LƯỢNG NƯỚC

Bùi Công Danh, Phạm Nguyễn Huy Phương*

Trường Đại học Công Thương Thành phố Hồ Chí Minh

*Email: *phuongpnh@huit.edu.vn*

Ngày nhận bài: 10/10/2024; Ngày chấp nhận đăng: 03/3/2025

TÓM TẮT

Chất lượng nguồn nước là một vấn đề quan trọng vì mối quan hệ của nó với con người và các sinh vật sống khác trong thế giới tự nhiên. Vấn đề đặt ra làm thế nào dự đoán các thông số chất lượng nước một cách chính xác nhằm đảm bảo việc quản lý tài nguyên nước đạt hiệu quả cao. Ngoài ra, trên thực tế cũng chưa có các giải pháp áp dụng kĩ thuật phân loại dựa trên mô hình học sâu vào lĩnh vực quản lý tài nguyên nước. Xuất phát từ những thực tiễn nêu trên, trong bài báo này nhóm tác giả giới thiệu một cách tiếp cận dùng các kỹ thuật phân lớp như: SVM, Random Forest, Logistic Regression, Kết quả thực nghiệm của bài báo cho thấy mô hình học sâu CNN của nhóm tác giả đề xuất có độ chính xác cao hơn so với các phương pháp phân lớp truyền thống khác.

Từ khóa: Phân loại, học sâu, chất lượng nước.

1. GIỚI THIỆU

Nước là nguồn tài nguyên quan trọng cho sự sống còn của con người và chất lượng của nó ảnh hưởng trực tiếp đến phát triển và sử dụng tài nguyên nước. Vì sự thay đổi chất lượng nước có liên quan mật thiết với môi trường khí hậu, sự thay đổi theo mùa và các hoạt động của con người, sự thay đổi chất lượng nước sông có đặc điểm là thay đổi dần dần, phi tuyến tính và không chắc chắn [1], rất khó để chính xác đoán sự thay đổi chất lượng nước. Tuy nhiên, dự báo chất lượng nước có ý nghĩa rất lớn đối với việc lập kế hoạch quản lý tài nguyên nước và môi trường. Theo kết quả dự báo, tình hình ô nhiễm nước có thể được dự đoán trước, do đó sự cố ô nhiễm nước có thể được ngăn chặn trước thông qua các mô hình dự báo. Chất lượng nước là một vấn đề quan trọng vì mối quan hệ của nó với con người và các sinh vật sống khác.

Sự phát triển nhanh chóng của ngành công nghiệp đã làm suy giảm chất lượng nước một cách đáng lo ngại. Thêm vào đó, cơ sở hạ tầng, cùng với việc công chúng thiếu nhận thức và tình trạng vệ sinh không tốt, có tác động lớn đến chất lượng nước uống [2]. Thực tế cho thấy, những hậu quả từ việc uống nước bị ô nhiễm rất nguy hiểm, có thể gây ảnh hưởng tiêu cực đến sức khỏe, môi trường cũng như cơ sở hạ tầng. Theo báo cáo của Liên Hợp Quốc, mỗi năm có khoảng 1,5 triệu người chết vì các bệnh liên quan đến nước ô nhiễm. Tại các nước đang phát triển, 80% các vấn đề về sức khỏe được cho là do nước ô nhiễm gây nên. Đã có 5 triệu ca tử vong và 2,5 tỷ ca bệnh được ghi nhận mỗi năm [3]. Tỷ lệ tử vong như vậy còn cao hơn số ca tử vong do tai nạn, tội phạm và khủng bố [4].

Do đó dự báo chất lượng nước đóng vai trò rất quan trọng trong việc bảo vệ nguồn nước, giữ gìn sức khỏe cộng đồng và phát triển bền vững ngành nước. Trong bài báo này đề xuất thuật toán dựa trên các phương pháp học máy như SVM, Random Forest, và Logistic Regression, CNN để dự báo chất lượng nước. Kết quả thực nghiệm của bài báo cho thấy tính hiệu quả của mô hình học sâu CNN của nhóm tác giả đề xuất so với các phương pháp phân lớp truyền thống khác về độ chính xác trong việc phân tích và dự báo chỉ số chất lượng nước (Water Quality Index - WQI) dựa trên bộ dữ liệu gồm 3276 mẫu [5]. Nghiên cứu này tiến hành đánh giá dựa trên một số tiêu chí chính bao gồm độ chính xác của dự đoán, hiệu suất tính toán, khả năng diễn giải kết quả, và tính tổng quát hóa của các mô hình. Cụ thể hơn, phương pháp đề xuất không chỉ ít tốn kém hơn về tài nguyên tính toán mà còn có khả năng cung cấp dự đoán đáng tin cậy và dễ diễn giải hơn khi làm việc với kích thước dữ liệu vừa phải như trong bộ dữ liệu được đề cập. Điều này có thể hỗ trợ việc ra quyết định nhanh chóng và hiệu quả trong các ứng

dụng thực tế liên quan đến quản lý và bảo vệ nguồn nước. Trong các phần sau bài báo trình bày: Các công trình liên quan tại mục 2, mục 3 tổng quan các phương pháp đề xuất, mục 4 Mô tả bài toán dự báo chất lượng nước, mục 5 trình bày các kết quả thực nghiệm và mục 6 kết luận nội dung của bài báo.

2. CÁC NGHIÊN CỨU LIÊN QUAN

Hiện nay, có hai loại chính để mô hình hóa và dự đoán chất lượng nước: các mô hình hướng cơ chế và không hướng cơ chế. Mô hình cơ chế tương đối phức tạp; nó sử dụng dữ liệu cấu trúc hệ thống tiên tiến để mô phỏng chất lượng nước (WQ), do đó, nó được coi là một mô hình đa chức năng có thể được sử dụng cho bất kỳ thể nước nào. Ngoài ra, mô hình Streeter–Phelos (S–P), một trong những mô hình mô phỏng WQ sớm nhất, đã được sử dụng rộng rãi.

Sau đó, một số quốc gia đã phát triển nhiều mô hình chất lượng nước (WQ), bao gồm mô hình QUAL [6] và mô hình WASP [7], đã được sử dụng rộng rãi để mô phỏng chất lượng nước của các con sông. Tiếp theo, Warren và Bach [8] đã đề xuất sử dụng mô hình MIKE21 để thiết kế các hệ thống mô hình hóa các cửa sông, vùng nước ven biển và biển. MIKE21 là một hệ thống mô hình hóa toàn diện, có thể được sử dụng để mô phỏng các quá trình thủy động lực học và chất lượng nước trong các môi trường ven biển và đại dương.

Liao và Sun đã phát triển một mô hình kết hợp giữa mạng neural nhân tạo (ANN) và cây quyết định để dự báo WQI của Hồ Chao ở Trung Quốc [9]. Yan và Qian đề xuất một mô hình cụm lan truyền ái lực dựa trên máy vectơ hỗ trợ bình phương nhỏ nhất (AP-LSSVM), tuy nhiên mô hình này nhạy cảm với các giá trị bị thiếu [10]. Solanki và cộng sự [11] sử dụng mô hình mạng học sâu để phân tích và dự đoán các thông số hóa học của nước như oxy hòa tan và độ pH, và báo cáo kết quả chính xác hơn các kỹ thuật học có giám sát truyền thống. Li và cộng sự [12] phát triển một mô hình lai kết hợp mạng neural và chuỗi Markov để dự đoán oxy hòa tan, một thước đo quan trọng của WQI [13]. Khan và See [14] sử dụng mạng neural nhân tạo (ANN) để dự báo WQI dựa trên các thông số như oxy hòa tan, diệp lục, độ dẫn điện và độ đục. Yan và cộng sự đề xuất sử dụng thuật toán di truyền (GA) và tối ưu hóa bầy hạt (PSO) để tối ưu hóa mạng neural truyền ngược (BP) nhằm dự đoán nhu cầu oxy, cải thiện độ chính xác của kết quả dự báo [15].

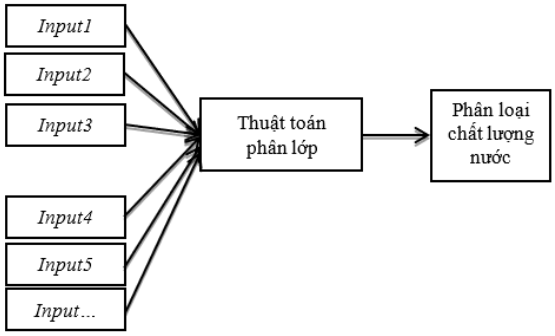
Ngoài ra, Ahmed và cộng sự đã đề xuất một phương pháp sử dụng một tập hợp giới hạn các thông số chất lượng nước, bao gồm nhiệt độ, độ đục, độ pH và tổng chất rắn hòa tan, để ước tính cả Chỉ số chất lượng nước (WQI) và Lớp chất lượng nước (WQC) [13]. Bài báo nêu bật những hạn chế của các phương pháp đánh giá chất lượng nước truyền thống trong phòng thí nghiệm, đặc biệt là về mặt chi phí và thời gian, đồng thời trình bày tiềm năng của học máy để giải quyết những thách thức này. Các tác giả nhận thấy rằng tăng cường độ dốc và hồi quy đa thức là các thuật toán hiệu quả nhất để dự đoán WQI.

Trong một nghiên cứu khác, Wang và cộng sự đã đề xuất một phương pháp học máy tổng hợp được gọi là xếp chồng mô hình có cấu trúc học hai lớp, trong đó đầu ra của năm mô hình học máy riêng lẻ được sử dụng rộng rãi (hồi quy tuyến tính bội, bình phương nhỏ nhất một phần, bình phương nhỏ nhất một phần thưa thớt , rừng ngẫu nhiên và mạng Bayesian) được lấy làm các tính năng đầu vào cho một mô hình khác tạo ra dự đoán cuối cùng [16]. Áp dụng phương pháp này cho ba bãi biển dọc theo phía đông Hồ Erie, New York, Hoa Kỳ và so sánh với tất cả năm mô hình cơ sở. Theo từng năm, điểm chính xác của mô hình xếp chồng luôn ở hoặc gần đầu bảng xếp hạng, với độ chính xác trung bình theo năm là 78%, 81% và 82,3% tại ba bãi biển được thử nghiệm. Còn trong một nghiên cứu khác của Sillberg và cộng sự xem xét việc sử dụng phương pháp tiếp cận dựa trên máy học, cụ thể là máy nhận dạng thuộc tính và SVM, để phân loại chất lượng nước ở Sông Chao Phraya, con sông lớn nhất ở Thái Lan [17]. Các tác giả nghiên cứu việc sử dụng một số thông số chất lượng nước, bao gồm nhu cầu oxy sinh học, độ dẫn điện, oxy hòa tan và vi khuẩn coliform trong phân, để phát triển chỉ số chất lượng nước có thể phân loại chính xác chất lượng nước của sông. Họ thấy rằng phương pháp tiếp cận này thành công trong việc phân loại chất lượng nước với độ chính xác cao và cũng được xác thực bằng cách sử dụng dữ liệu từ một con sông khác, sông Tha Chin. Nghiên cứu chứng minh tiềm năng của các kỹ thuật máy học trong việc cải thiện việc giám sát và quản lý chất lượng nước. Kết quả phương pháp tuyến tính SVM đã cho phép các kết quả phân loại tốt nhất được biểu thị là độ chính xác là 0,94, độ chính xác trung bình là 0,84, độ thu hồi trung bình là 0,84 và điểm F1 trung bình là 0,84. Việc xác thực cho thấy AR-SVM là một phương pháp mạnh mẽ để xác định chất lượng nước sông với độ chính xác 0,86–0,95 khi áp dụng cho ba đến sáu đặc điểm. Vì vậy, có thể thấy rằng sử dụng các phương pháp học máy để

phân tích chất lượng nguồn nước và các chỉ số liên quan đến nguồn nước là vấn đề đang được nhiều nhà nghiên cứu quan tâm trong thời gian gần đây.

3. TỔNG QUAN CÁC PHƯƠNG PHÁP ĐỀ XUẤT

Trong mô đề xuất nghiên cứu mong muốn thực hiện phân loại chất lượng nước từ các thành phần hoá lý đáng quan tâm tác động lớn đến chất lượng nước. Trước khi áp dụng các mô hình học máy để phân tích dữ liệu, các bước chuẩn bị dữ liệu đã được thực hiện để tạo ra đầu vào thích hợp cho mô hình. Quá trình này bao gồm việc chia dữ liệu thành các tập huấn luyện và kiểm thử để huấn luyện 3 mô hình và đánh giá hiệu suất của chúng trên tập dữ liệu có tính tương đồng với chất lượng nước tại Việt Nam. Đồng thời, dữ liệu đã được làm sạch bằng cách loại bỏ các giá trị không hợp lệ và thay thế các ô trống bằng giá trị trung vị của các biến đầu vào. Sau đó, các mô hình học máy khác nhau đã được sử dụng để dự đoán phân loại chất lượng nước (WQC) dựa trên các yếu tố đã được xác định.

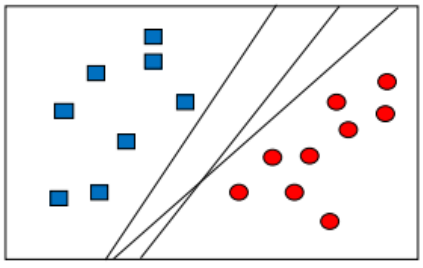


Hình 1. Mô hình đề xuất phân loại chất lượng nước tại Việt Nam

Thuật toán phân lớp trong lưu đồ hình 1 bao gồm các thuật toán như: SVM, Logistic Regression và Random Forest, CNN.

3.1. Support Vector Machine (SVM)

Máy vectơ hỗ trợ (SVM) là một thuật toán học máy được giám sát có thể được sử dụng cho cả các thử thách phân loại hoặc hồi quy. Tuy nhiên, nó chủ yếu được sử dụng trong các bài toán phân loại. Trong thuật toán SVM, vẽ mỗi mục dữ liệu dưới dạng một điểm trong không gian n chiều (với n là một số đối tượng) với giá trị của mỗi đối tượng là giá trị của một tọa độ cụ thể. Sau đó, thực hiện phân loại bằng cách tìm siêu mặt phẳng phân biệt hai lớp [18].

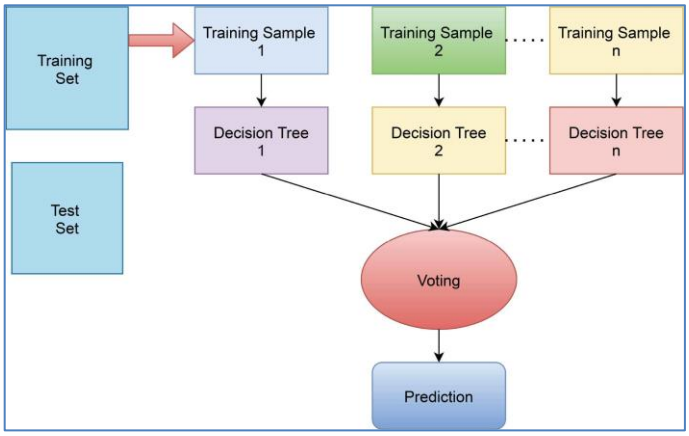


Hình 2. Có vô số đường thẳng phân tách hai lớp dữ liệu [18]

Vectơ hỗ trợ chỉ đơn giản là tọa độ của quan sát cá nhân. Bộ phân loại SVM là biên giới phân tách tốt nhất hai lớp (siêu mặt phẳng / dòng).

3.2. Random Forest

Random Forest là một phương pháp học hồi quy, phân loại và các vấn đề phức tạp. Random Forest hoạt động bằng cách huấn luyện một số lượng lớn dữ liệu mẫu. Random Forest xây dựng cây quyết định từ các mẫu và sử dụng quyết định đa số cho phân loại và hồi quy. Bởi vì Random Forest làm việc với tập con dữ liệu, chúng hoạt động nhanh hơn cây quyết định. Vì vậy, chúng ta có thể giải quyết dễ dàng hàng trăm đặc điểm mà không gặp bất kỳ sự khó khăn nào [19].



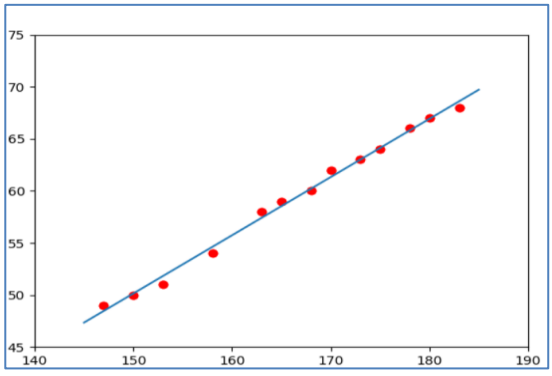
Hình 3. Lưu đồ thuật toán Random Forest [19]

3.3. Logistic Regression

Sử dụng Logistic Regression để đánh giá dữ liệu phân loại trong các lĩnh vực khác nhau, bao gồm sức khỏe, xã hội và học thuật. Một biến phản hồi sau đó đã được dự đoán bằng cách sử dụng các biến giải thích tích hợp hoặc phân loại. Tức là nó đo lường tỷ lệ ý nghĩa tương đối của các biến độc lập, đánh giá các mối tương quan và giúp chúng ta hiểu được tác động của các biến kiểm soát tương quan [20]. Trong nghiên cứu này muốn biết các sửa đổi đã ảnh hưởng như thế nào đến các biến giải thích đối với xác suất của trong Phương trình (1), được biểu diễn dưới dạng [21]:

P(Y=j/X1,X2,...,Xk) = P(Y=j/K); j= 0, 1, ..., J (1)

Ta có: Y = Biến phản hồi, X = Biến giải thích.

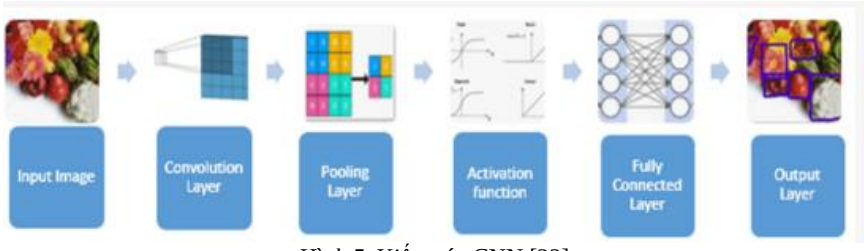


Hình 4. Bản chất của Logistic Regression [20]

3.4. Thuật toán Convolutional Neural Network (CNN)

Là một trong những mô hình Deep Learning vô cùng tiên tiến. CNN sẽ cho phép xây dựng các hệ thống thông minh với độ chính xác vô cùng cao. Mạng CNN được thiết kế với mục đích xử lý dữ liệu thông qua nhiều lớp mảng và duỗi dữ liệu từ nhiều chiều thành một chiều để giảm độ phức tạp của dữ liệu, từ đó thông qua các công thức tính sai số trung bình đưa ra đánh giá độ chính xác…

Một mạng nơ-ron tích tụ bao gồm một lớp đầu vào, các lớp ẩn và một lớp đầu ra. Trong bất kỳ mạng nơron chuyển tiếp nào, bất kỳ lớp giữa nào được gọi là ẩn vì các đầu vào và đầu ra của chúng bị che bởi hàm kích hoạt và tích chập cuối cùng. Trong một mạng nơron tích chập, các lớp ẩn bao gồm các lớp thực hiện các phép chập. Thông thường, điều này bao gồm một lớp thực hiện tích điểm của nhân chập với ma trận đầu vào của lớp. Sản phẩm này thường là sản phẩm bên trong Frobenius và chức năng kích hoạt của nó thường là ReLU. Khi hạt nhân tích chập trượt dọc theo ma trận đầu vào cho lớp, phép toán tích chập tạo ra một bản đồ đặc trưng, bản đồ này sẽ đóng góp vào đầu vào của lớp tiếp theo. Tiếp theo là các lớp khác như lớp gộp, lớp được kết nối đầy đủ và lớp chuẩn hóa [22].



Hình 5. Kiến trúc CNN [22]

Điểm mạnh của CNN trong phân loại dữ liệu có nhãn:

Automatic feature extraction: Không giống như các phương pháp truyền thống trong đó các tính năng cần được thiết kế thủ công, CNN **tự động học và trích xuất các tính năng** từ dữ liệu thô thông qua các lớp tích chập. Điều này loại bỏ nhu cầu về một trình trích xuất tính năng riêng biệt1Khả năng trích xuất tính năng tự động này được trích dẫn là một trong những lý do khiến CNN là thuật toán hiệu quả nhất để hiểu nội dung hình ảnh và đã chứng minh hiệu suất vượt trội trong các tác vụ phân loại, nhận dạng, phân đoạn và truy xuất hình ảnh.

Hierarchical learning: CNN học các đặc điểm theo cách phân cấp, bắt đầu với các đặc điểm đơn giản như cạnh và góc trong các lớp ban đầu và tiến triển đến các đặc điểm phức tạp hơn trong các lớp sâu hơn. Biểu diễn phân cấp này cho phép CNN nắm bắt các mẫu phức tạp và sự phụ thuộc trong dữ liệu, khiến chúng hiệu quả cho các tác vụ phân loại.

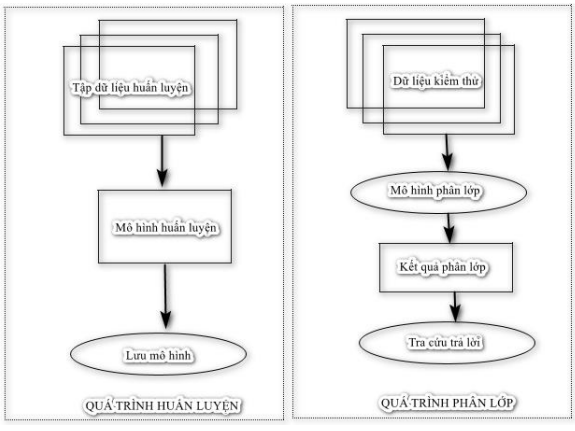
Chia sẻ trọng số: CNN sử dụng cùng một tập hợp trọng số cho các phần khác nhau của hình ảnh đầu vào, giúp giảm đáng kể số lượng tham số so với mạng được kết nối đầy đủ. Chia sẻ trọng số này giúp CNN hiệu quả hơn và ít bị quá khớp hơn một.

Đa nhiệm: CNN có thể được đào tạo để thực hiện nhiều nhiệm vụ cùng lúc. Ví dụ, một CNN duy nhất có thể được sử dụng để phát hiện đối tượng và phân loại hình ảnh.

CNN là một công cụ mạnh mẽ để phân loại dữ liệu có nhãn. Khả năng tự động trích xuất các tính năng, học các biểu diễn phân cấp và khái quát hóa tốt của chúng khiến chúng phù hợp với nhiều ứng dụng.

4. BÀI TOÁN DỰ BÁO CHẤT LƯỢNG NƯỚC

Các bước xây dựng mô hình phân lớp và thực hiện phân loại dữ liệu dự báo chất lượng nước.



Hình 6. Mô hình hệ thống phân loại chất lượng nước

4.1. Xây dựng tập dữ liệu

Trong bài báo này sử dụng bộ dữ liệu này là kết quả phân tích chất lượng nước với 3276 mẫu nước khác nhau từ các bộ dữ liệu tin cậy. Phân tích đã xác định số lượng của 10 thành phần đáng quan tâm tác động lớn đến chất lượng nước.

Bao gồm: {*pH, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic_carbon, Trihalomethanes, Turbidity, Potability*} [5].

Bảng 1. Mô tả Dữ liệu chất lượng nước theo các thành phần

pH	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
7.64787	160.774	29000.59	7.217409	365.94	438.801	13.1825	67.09997	3.078673	1
2.37676	129.864	11684.11	9.974808	301.42	503.9104	8.741258	76.31069	3.629218	0
5.51773	190.669	17638.32	10.52586	321.59	461.929	12.53168	44.88784	4.523106	0
8.55042	206.522	10453.09	6.482009	326.12	347.8243	13.99747	42.29053	4.90387	0
6.86491	170.827	20464.77	7.074063	365.94	400.9653	14.00387	51.603	4.493747	1
6.38197	235.945	18982.95	8.286514	341.59	511.4852	16.18354	77.83952	3.299047	0
5.09576	273.408	26307.3	10.2198	380.20	513.8776	18.27242	61.42519	4.352193	0
7.12145	204.164	20574.36	7.089146	335.59	353.9276	16.48816	57.02278	3.774601	0
7.0427	162.512	24642.81	7.267573	330.90	414.6507	11.58335	73.94619	3.21574	0
4.08359	201.938	20555.97	1.920271	341.59	435.5115	12.22668	69.28978	2.974871	1

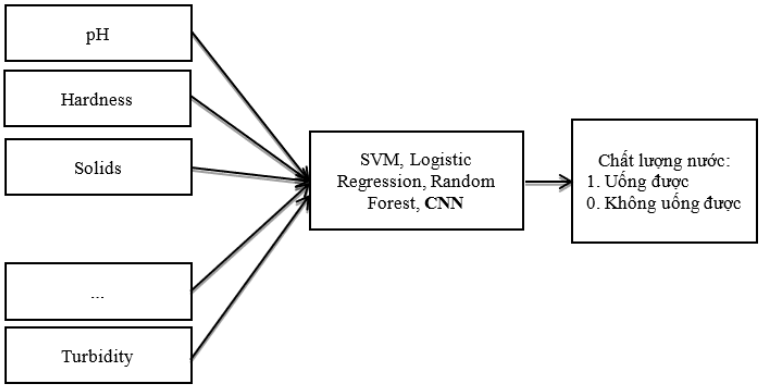
Các chỉ số biểu diễn chất lượng nước được mô tả như sau:

Bảng 2. Mô tả thông số chất lượng nguồn nước.

Chỉ tiêu chất lượng	Mô tả
pH	pH là một thông số quan trọng trong việc đánh giá cân bằng axit-bazơ của nước. Nó cũng là chỉ số về tình trạng axit hoặc kiềm của trạng thái nước. WHO đã khuyến cáo giới hạn tối đa cho phép của pH từ 6,5 đến 8,5. Phạm vi điều tra hiện tại là 6,52–6,83 nằm trong phạm vi tiêu chuẩn của WHO.
Hardness	Độ cứng chủ yếu do muối canxi và magie gây nên. Những muối này được hòa tan từ các trầm tích địa chất mà qua đó nước di chuyển. Khoảng thời gian nước tiếp xúc với vật liệu tạo độ cứng giúp xác định độ cứng có trong nước thô. Độ cứng ban đầu được định nghĩa là khả năng kết tủa xà phòng của nước do canxi và magie gây ra.
Solids	Nước có khả năng hòa tan nhiều loại khoáng chất vô cơ và hữu cơ hoặc muối chẳng hạn như kali, canxi, natri, bicacbonat, clorua, magie, sunfat, v.v. Những khoáng chất này tạo ra mùi vị không mong muốn và màu loãng trong nước. Đây là thông số quan trọng cho việc sử dụng nước. Nước có giá trị TDS cao chứng tỏ nước có nhiều khoáng chất. Giới hạn mong muốn cho TDS là 500 mg/L và giới hạn tối đa là 1000 mg/L được quy định cho mục đích uống.
Chloramines	Clo và chloramine là những chất khử trùng chính được sử dụng trong hệ thống nước công cộng. Chloramine thường được hình thành khi thêm amoniac vào clo để xử lý nước uống. Nồng độ clo lên tới 4 miligam trên lít (mg/L hoặc 4 phần triệu (ppm)) được coi là an toàn trong nước uống.
Sulfate	Sunfat là những chất tự nhiên được tìm thấy trong khoáng chất, đất và đá. Chúng hiện diện trong không khí xung quanh, nước ngầm, thực vật và thực phẩm. Việc sử dụng thương mại chính của sulfat là trong ngành hóa chất. Nồng độ sunfat trong nước biển là khoảng 2.700 miligam trên lít (mg/L). Nằm trong khoảng từ 3 đến 30 mg/L ở hầu hết các nguồn cung cấp nước ngọt, mặc dù nồng độ cao hơn nhiều (1000 mg/L) được tìm thấy ở một số vị trí địa lý.
Conductivity	Nước tinh khiết không phải là chất dẫn điện tốt mà là chất cách điện tốt. Tăng nồng độ ion giúp tăng cường tính dẫn điện của nước. Nói chung, lượng chất rắn hòa tan trong nước quyết định độ dẫn điện. Độ dẫn điện (EC) thực sự đo lường quá trình ion của một dung dịch cho phép nó truyền dòng điện. Theo tiêu chuẩn của WHO, giá trị EC không được vượt quá 400 µS/cm.
Organic_carbon	Tổng lượng Carbon hữu cơ (TOC) trong nguồn nước đến từ các chất hữu cơ tự nhiên đang phân hủy (NOM) cũng như các nguồn tổng hợp. TOC là thước đo tổng lượng carbon trong các hợp chất hữu cơ trong nước tinh khiết. Theo US EPA < 2 mg/L as TOC trong nước đã qua xử lý/nước uống, và < 4 mg/L trong nước nguồn được sử dụng để xử lý.
Trihalomethanes	THMs là hóa chất có thể được tìm thấy trong nước được xử lý bằng clo. Nồng độ THMs trong nước uống thay đổi tùy theo mức độ chất hữu cơ trong nước, lượng clo cần thiết để xử lý nước và nhiệt độ của nước đang được xử lý. Mức THM lên đến 80 ppm được coi là an toàn trong nước uống.
Turbidity	Độ đục của nước phụ thuộc vào lượng chất rắn tồn tại ở trạng thái lơ lửng. Nó là thước đo các đặc tính phát sáng của nước và phép thử được sử dụng để chỉ ra chất lượng xả thải đối với chất keo. Giá trị độ đục trung bình thu được đối với Wondo Genet Campus (0,98 NTU) thấp hơn giá trị khuyến nghị của WHO là 5,00 NTU.
Potability	Cho biết nước có an toàn cho người tiêu dùng hay không trong đó 1 nghĩa là Uống được và 0 nghĩa là Không uống được.

4.2. Đề xuất mô hình dự báo chất lượng nước

Mô hình này sử dụng các thông số hóa học cơ bản của nước làm dữ liệu đầu vào cho mô hình đánh giá để phân loại chất lượng nước, giúp xác định xem nước có an toàn để uống hay không thông qua việc sử dụng các thuật toán như SVM, Logistic Regression và Random Forest, CNN để thực nghiệm cho phép đánh giá và so sánh để chọn ra mô hình tối ưu nhất. Hình 7 mô tả chi tiết quá trình thực hiện mô hình huấn luyện và mô hình phân lớp trong Hình 6.



Hình 7. Mô hình phân loại dự báo chất lượng nước cho bài toán

5. KẾT QUẢ THỰC NGHIỆM

5.1. Môi trường cài đặt

Nhóm tác giả triển khai thuật toán trên môi trường máy tính cá nhân. Kết quả thực nghiệm trên máy tính sử dụng hệ điều hành Windows 8.1 Pro 64bit, RAM 12GB, chip Intel(R) Core(TM) i3-6820HQ, ~2.7 GHz, mô hình huấn luyện và sử dụng ngôn ngữ lập trình Python phiên bản 3.7.7.

5.2. Dữ liệu

Công trình nhóm tác giả đề xuất đã thực nghiệm trên 3276 mẫu nước khác nhau, dữ liệu được chia thành hai phần: một phần dùng để huấn luyện và một phần để dự báo. Nguồn dữ liệu sử dụng từ công trình [5] đã được kiểm chứng về độ tin cậy dữ liệu.

Tập dữ liệu mẫu: water_potability.csv [5].

Thông tin tập dữ liệu:

Mục đích tập: Đánh giá khả năng uống được của nước dựa trên các chỉ số chất lượng nước.

Các chỉ số chất lượng: pH, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic_carbon, Trihalomethanes, Turbidity.

Biến Potability: Chỉ số cuối cùng cho biết nước có an toàn để uống hay không (1 nghĩa là có thể uống, 0 nghĩa là không thể uống).

Trong phần thực nghiệm của bài báo này nhóm tác giả chia tập dữ liệu trên thành các phần huấn luyện, kiểm thử theo các tỷ lệ: 90/10, 80/20, 70/30, 60/40, 50/50 để thực nghiệm so sánh đánh giá các mô hình trên tập kiểm thử.

5.3. Kết quả thực nghiệm so sánh đánh giá các mô hình

Bảng 3 trình bày kết quả về mức độ chính xác trên các tập kiểm thử có kích thước khác nhau. Dựa vào đó có thể thấy, khi tập kiểm thử nhỏ thì độ chính xác cho kết quả tốt và tương đối đồng đều. Khi kích thước tập kiểm thử lớn hơn, bắt đầu có sự chênh lệch đáng kể về mức độ chính xác hơn. Đơn cử trường hợp test set 0,5: phương pháp Logistic Regression và CNN cho độ chính xác tương đối thấp; Random Forest là 0,658; riêng SVM vẫn cho kết quả tương đối tốt là 0,692. Nhìn chung kết quả của CNN luôn đạt kết quả vượt trội trong tất cả tình huống.

Độ chính xác (Độ chính xác): SVM và **CNN** luôn đạt được điểm chính xác trên 0,68 trên tất cả các tỷ lệ tập kiểm tra trong 4 tình huống còn lại, đạt độ chính xác tới 0,75 với tập kiểm tra 10%. Điều

này cho thấy mô hình đã học được các mẫu trong dữ liệu một cách hiệu quả và đưa ra dự đoán chính xác về khả năng uống được của nước.

Độ chuẩn xác (Độ chính xác dự đoán): Từ bảng số liệu, Độ chuẩn xác của SVM và Random Forest thường cao hơn so với Logistic Regression ở các tỷ lệ tập kiểm thử từ 0,1 đến 0,5. Tuy nhiên, đối với Logistic Regression, Độ chuẩn xác đôi khi không được tính, cho thấy model với phương pháp này không cung cấp dự đoán dương tính. **CNN** cho thấy độ chính xác tốt, dao động từ 0,53 đến 0,75. Điều này chỉ ra rằng khi mô hình dự đoán mẫu nước có thể uống được thì nhìn chung là đúng.

Độ bao phủ (Độ bao phủ): Kết quả Độ bao phủ của SVM và Random Forest có vẻ ổn định hơn so với Logistic Regression, đặc biệt khi áp dụng tỷ lệ tập kiểm thử thấp. **CNN**, dao động từ 0,43 đến 0,56, cho thấy khả năng hợp lý trong việc xác định chính xác tất cả các mẫu nước uống trong tập dữ liệu.

Bảng 3. Thông số độ chính xác của các mô hình với các tập kiểm thử

	SVM	Logistic Regression	Random Forest	CNN
Tỷ lệ tập kiểm thử: 0,1				
Độ chính xác	0,744	0,654	0,731	0,754
Độ chuẩn xác	0,78	0,0	0,72	0,79
Độ bao phủ	0,20	0,0	0,30	0,43
Tỷ lệ tập kiểm thử: 0,2				
Độ chính xác	0.6875	0,628	0,689	0,7
Độ chuẩn xác	0,70	0,0	0,67	0,72
Độ bao phủ	0,27	0,0	0,36	0,47
Tỷ lệ tập kiểm thử: 0,3				
Độ chính xác	0,681	0,622	0,682	0,68
Độ chuẩn xác	0,7	1,0	0,67	0,63
Độ bao phủ	0,27	0,005	0,35	0,51
Tỷ lệ tập kiểm thử: 0,4				
Độ chính xác	0,687	0,636	0,679	0,69
Độ chuẩn xác	0,66	0,81	0,59	0,64
Độ bao phủ	0,29	0,01	0,34	0,56
Tỷ lệ tập kiểm thử: 0,5				
Độ chính xác	0,692	0,637	0,658	0,68
Độ chuẩn xác	0,69	0,73	0,59	0,53
Độ bao phủ	0,30	0,04	0,33	0,46

Nhìn chung, từ Bảng dữ liệu 3, CNN và SVM có hiệu suất tốt hơn so với Logistic Regression và Random Forest trong việc dự đoán và phân loại với các chỉ số như độ chính xác, độ chính xác dự đoán và tỷ lệ nhớ tới. Tuy nhiên, để đánh giá một cách toàn diện, cần xem xét thêm nhiều yếu tố khác như overfitting, sử dụng cross-validation.

Với kết quả thực nghiệm như Bảng 3 cho chúng ta thấy rằng việc phân loại dự báo chất lượng nguồn nước cho kết quả khả quan tốt trên các mô hình phân loại khác nhau. Trong đó khả năng học sâu, khái quát cao của CNN và SVM thể hiện khả năng ứng dụng cao nhất cho việc phân loại này. Bài báo cung cấp một hướng đi một cách tiếp cận cho các ứng dụng phân loại chất lượng nước.

6. KẾT LUẬN VÀ KHUYẾN NGHỊ

Bài báo đã phân tích được tầm quan trọng trong việc dự báo chất lượng nguồn nước. Đồng thời qua việc tiến hành thực nghiệm thu được những giá trị kiểm thử có độ chính xác khá cao. Cho thấy tính ưu việt của việc ứng dụng kỹ thuật phân lớp vào việc xác phân loại dự báo chất lượng nguồn nước. Mở ra một hướng đi khả quan cho việc ứng dụng các công cụ thuật toán vào giải quyết bài toán đánh giá môi trường tại Việt Nam cụ thể là chất lượng nước uống.

Nghiên cứu này đã chứng minh tầm quan trọng của việc áp dụng các phương pháp phân loại tiệp cận học máy truyền thống như SVM, Random Forest, và Logistic Regression cũng như thể hiện được hiệu quả trong phương pháp tiếp cận học sâu CNN trong việc dự báo chất lượng nước. Bằng cách sử

dụng các mô hình học máy này, bài báo đã thu được kết quả khả quan về độ chính xác trong việc phân loại nước theo các chỉ số chất lượng như pH, độ cứng, độ đục, và khả năng uống được. Việc dự báo chính xác chất lượng nước không chỉ góp phần quan trọng vào việc bảo vệ sức khỏe cộng đồng mà còn hỗ trợ quản lý hiệu quả tài nguyên nước, giúp các cơ quan quản lý và hoạch định chính sách đưa ra những quyết định kịp thời và chính xác hơn. Trên cơ sở những kết quả đạt được, bài báo kiến nghị tiếp tục nghiên cứu và mở rộng ứng dụng các mô hình học máy tiên tiến trong các lĩnh vực liên quan khác, nhằm cải thiện hơn nữa chất lượng và sự bền vững của nguồn nước. Những phương pháp tiếp cận mới này hứa hẹn sẽ mang lại lợi ích thiết thực trong việc bảo vệ môi trường và đảm bảo an toàn cho con người.

Trong hướng nghiên cứu tiếp theo bài báo khuyến nghị sử dụng các thuật toán học sâu cho dự báo chất lượng nước bằng hình ảnh.

TÀI LIỆU THAM KHẢO

1. Nasir N.N., Kansal A., Alshaltone O., Barneih F., Sameer M., Shanableh A., & Al-Shamma'a A. - Water quality classification using machine learning algorithms. Journal of Water Process Engineering **48** (2022) 102920. https://doi.org/10.1016/j.jwpe.2022.102920
2. Zeilhofer P., Zeilhofer L., Hardoim E., Lima Z., & Oliveira C. - GIS applications for mapping and spatial modeling of urban-use water quality: A case study in District of Cuiabá, Mato Grosso, Brazil. Cadernos de Saúde Pública Ministério da Saúde Fundação Oswaldo Cruz Escola Nacional de Saúde Pública **23** (2007) 875-884. https://doi.org/10.1590/S0102-311X2007000400015
3. Kahlown M., Tahir M., & Rasheed H. - National water quality monitoring programme, fifth monitoring report (2005-2006). Pakistan Council of Research in Water Resources Islamabad, Islamabad, Pakistan (2007).
4. Aldhyani T.H., Al-Yaari M., Alkahtani H., & Maashi M. - Water quality prediction using artificial intelligence algorithms. Applied Bionics and Biomechanics **2020** (1) (2020) 6659314. https://doi.org/10.1155/2020/6659314
5. Kadiwal A. - Water quality (2021). [Online]. Available: https://www.kaggle.com/datasets/adityakadiwal/water-potability
6. Lai Y.C., Yang C.P., Hsieh C.Y., Wu C.Y., & Kao C.M. - Evaluation of non-point source pollution and river water quality using a multimedia two-model system. Journal of Hydrology 409 (3) (2011) 583-595. https://doi.org/10.1016/j.jhydrol.2011.08.040
7. Huang J., Liu N., Wang M., & Yan K. - Application WASP model on validation of reservoir-drinking water source protection areas delineation. 2010 3rd International Conference on Biomedical Engineering and Informatics (2010) 3031-3035. https://doi.org/10.1109/BMEI.2010.5639900
8. Warren I.R., & Bach H.K. - MIKE 21: a modelling system for estuaries, coastal waters and seas. Environmental Software **7** (4) (1992) 229-240. https://doi.org/10.1016/0266-9838(92)90006-P
9. Liao H., & Sun W. - Forecasting and evaluating water quality of Chao Lake based on an improved decision tree method. Procedia Environmental Sciences **2** (2010) 970-979. https://doi.org/10.1016/j.proenv.2010.10.109
10. Yan-jun L., & Qian M. - AP-LSSVM modeling for water quality prediction. Proceedings of the 31st Chinese Control Conference, Hefei, China, IEEE (2012) 6928-6932.
11. Solanki A., Agrawal H., & Khare K. - Predictive analysis of water quality parameters using deep learning. International Journal of Computer Applications **125** (2015) 29-34. https://doi.org/10.5120/ijca2015905874
12. Li X., & Song J. - A new ANN-Markov chain methodology for water quality prediction. 2015 International Joint Conference on Neural Networks (IJCNN) (2015) 1-6. https://doi.org/10.1109/IJCNN.2015.7280320
13. Ahmed A.A.M., & Shah S.M.A. - Application of adaptive neuro-fuzzy inference system (ANFIS) to estimate the biochemical oxygen demand (BOD) of Surma River. Journal of King

Saud University - Engineering Sciences **29** (3) (2017) 237-243. https://doi.org/10.1016/j.jksues.2015.02.001

14. Khan Y., & See C.S. - Predicting and analyzing water quality using machine learning: A comprehensive model. 2016 IEEE Long Island Systems, Applications and Technology Conference (LISAT) (2016) 1-6. https://doi.org/10.1109/LISAT.2016.7494106
15. Yan J., Xu Z., Yu Y., Xu H., & Gao K. - Application of a hybrid optimized BP network model to estimate water quality parameters of Beihai Lake in Beijing. Applied Sciences **9** (9) (2019) 1863. https://doi.org/10.3390/app9091863
16. Wang L., Zhu Z., Sassoubre L., Yu G., Liao C., Hu Q., & Wang Y. - Improving the robustness of beach water quality modeling using an ensemble machine learning approach. Science of the Total Environment **765** (2021) 142760. https://doi.org/10.1016/j.scitotenv.2020.142760
17. Sillberg C., Kullavanijaya P., & Chavalparit O. - Water quality classification by integration of attribute-realization and support vector machine for the Chao Phraya River. Journal of Ecological Engineering **22** (9) (2021) 70-86. https://doi.org/10.12911/22998993/141364
18. Xiang Y., & Jiang L. - Water quality prediction using LS-SVM and particle swarm optimization. 2009 Second International Workshop on Knowledge Discovery and Data Mining (2009) 900-904. https://doi.org/10.1109/WKDD.2009.217
19. Hassan M.M., Hassan M.M., Akter L., Rahman M.M., Zaman S., Hasib K.M., Jahan N., Smrity R.N., Farhana J., Raihan M., & Mollick S. - Efficient prediction of water quality index (WQI) using machine learning algorithms. Human-Centric Intelligent Systems **1** (3) (2021) 86-97. https://doi.org/10.2991/hcis.k.211203.001
20. El-Habil A. - An application on multinomial logistic regression model. Pakistan Journal of Statistics and Operation Research **8** (3) (2012) 271-291. https://doi.org/10.18187/pjsor.v8i2.234
21. Devi S. - Random forest advice for water quality prediction in the regions of Kadapa district. International Journal of Innovative Technology and Exploring Engineering **8** (2019) 1464-1466.
22. Bhatt D., Patel C., Talsania H., Patel J., Vaghela R., Pandya S., Modi K., & Ghayvat H. - CNN variants for computer vision: History, architecture, application, challenges and future scope. Electronics **10** (20) (2021) 2470. https://doi.org/10.3390/electronics10202470

ABSTRACT

APPROACHING VARIOUS DATA CLASSIFICATION METHODS IN WATER QUALITY FORECASTING

Bui Cong Danh, Pham Nguyen Huy Phuong*

Ho Chi Minh City University of Industry and Trade

*Email: *phuongpnh@huit.edu.vn*

Water quality is an important issue because of its relationship with humans and other living organisms in the natural world. The problem at hand is how to accurately predict water quality parameters in order to ensure the high effectiveness of water resource management. Additionally, in practice, there are currently no solutions applying classification techniques based on deep learning models in the field of water resource management. Based on the aforementioned practices, in this paper, the authors introduce an approach using classification techniques such as SVM, Random Forest, Logistic Regression. The experimental results of the paper show that the CNN deep learning model proposed by the authors has higher accuracy compared to other traditional classification methods.

Keywords: Classification, deep learning, water quality.