

XÂY DỰNG MÔ HÌNH HỌC SÂU DỰ BÁO ĐẶC TRƯNG LŨ TẠI TRẠM THỦY VĂN YÊN BÁI

Ngô Thị Thủy⁽¹⁾, Lương Hữu Dũng⁽¹⁾, Trịnh Thu Phương⁽²⁾,

Lương Tuấn Trung⁽¹⁾, Chu Nguyễn Ngọc Sơn⁽¹⁾

⁽¹⁾Viện Khoa học Khí tượng Thủy văn và Biến đổi khí hậu

⁽²⁾Ủy hội sông Mê Công Việt Nam

Ngày nhận bài: 21/5/2025; ngày chuyển phản biện: 22/5/2025; ngày chấp nhận đăng: 08/6/2025

Tóm tắt: Trí tuệ nhân tạo mà điển hình là các mô hình học sâu đang ngày càng được ứng dụng rộng rãi trong các nghiên cứu về thủy văn và tài nguyên nước. Trong bài toán dự báo lũ lớn trên các sông, các mô hình học máy, học sâu đã thể hiện được tiềm năng với tốc độ xử lý nhanh, cho kết quả với độ chính xác cao ngay cả với những chuỗi dữ liệu ngắn, tập trung vào mô tả các sự kiện lũ và không có tính liên tục về thời gian. Tuy có nhiều ưu điểm nhưng các mô hình học sâu hay trí tuệ nhân tạo nói chung đều cần kiến thức chuyên gia để định hướng cách học, tránh việc học máy móc. Nghiên cứu này được thực hiện nhằm xây dựng được mô hình học sâu dựa trên nền tảng mạng nơ-ron truyền thẳng đa tầng (Multilayer Perceptron-MLP) phục vụ dự báo đặc trưng lũ tại Trạm thủy văn Yên Bái trên Sông Thao. Từ dữ liệu lưu lượng chân lũ, lượng mưa tích lũy 5 ngày tại các trạm Sa Pa, Văn Chấn và Yên Bái, một mô hình MLP được xây dựng với cấu trúc 11x11. Mô hình được huấn luyện bởi 62 số liệu (1961-2022) và cho kết quả tốt với hệ số $R^2=0,69-0,84$, sai số MAE<20% lưu lượng đỉnh lũ trung bình. Dữ liệu năm 2023-2024 được sử dụng để dự báo thử nghiệm đỉnh lũ tại Yên Bái, kết quả cho sai số đỉnh lũ đạt 9,8% (năm 2023) và 20,9% (năm 2024).

Từ khóa: Multilayer Perceptron, dự báo lũ, Sông Thao, học sâu.

1. Tổng quan chung

Dự báo lũ là một trong những bài toán quan trọng trong quản lý tài nguyên nước và phòng chống thiên tai. Với đặc trưng dữ liệu phi tuyến, đa chiều, và chịu ảnh hưởng bởi nhiều yếu tố thời gian - không gian, các mô hình thống kê truyền thống thường gặp khó khăn trong việc nắm bắt đầy đủ mối quan hệ giữa đầu vào và đầu ra. Trong bối cảnh đó, học sâu (Deep Learning) đã nổi lên như một công cụ mạnh mẽ, linh hoạt, và có khả năng tự học từ dữ liệu, hỗ trợ hiệu quả cho bài toán dự báo lũ.

Các mô hình học sâu như Multilayer Perceptron (MLP), Long Short-Term Memory (LSTM) và Convolutional Neural Networks (CNN) đã được ứng dụng rộng rãi trong nhiều nghiên cứu gần đây. MLP là kiến trúc đơn giản nhưng hiệu quả trong việc xây dựng mô hình quan hệ

giữa các biến khí tượng - thủy văn (như lượng mưa, mực nước, độ ẩm đất) và lưu lượng đỉnh tại các trạm đo [1], [2]. MLP đặc biệt phù hợp với các bài toán thủy văn khi dữ liệu đầu vào có mối quan hệ phức tạp và không tuân theo phân phối tuyến tính. Theo nghiên cứu của Dawson & Wilby [3], mạng nơ-ron nhân tạo (ANN), trong đó có MLP, cho thấy khả năng vượt trội trong việc dự báo dòng chảy so với các mô hình thống kê truyền thống. Trong khi đó, LSTM - một dạng mạng hồi tiếp (RNN) - được sử dụng phổ biến cho các bài toán chuỗi thời gian do khả năng ghi nhớ thông tin dài hạn, rất phù hợp để mô hình hóa quá trình tích tụ nước và truyền lũ trong lưu vực [4], [5], [6].

Tại Việt Nam, một số nghiên cứu gần đây cũng đã thử nghiệm áp dụng mạng trí tuệ nhân tạo như ANN và LSTM để dự báo lưu lượng lũ tại các khu vực như Tây Nguyên [7], Đồng bằng Sông Hồng [8], Vu Gia - Thu Bồn [9] và đạt được kết quả khả quan. Tuy nhiên, những nghiên cứu này phần lớn sử dụng nguồn dữ liệu dài và mang

Liên hệ tác giả: Ngô Thị Thủy

Email: tide4586@gmail.com

tính liên tục về thời gian (timeseries), với những dữ liệu mang tính đặc trưng như lưu lượng đỉnh lũ và tổng lượng lũ thì việc ứng dụng các mô hình như ANN (NARX) hay LSTM thường không phù hợp về mặt vật lý (do sử dụng bước thời gian trễ). Đối với các loại dữ liệu này, những mô hình học sâu như MLP (mạng nơ-ron truyền thẳng đa tầng) là cấu trúc phù hợp để thực hiện hồi quy theo trạng thái hiện tại (lượng mưa tích lũy và nền lũ hiện tại).

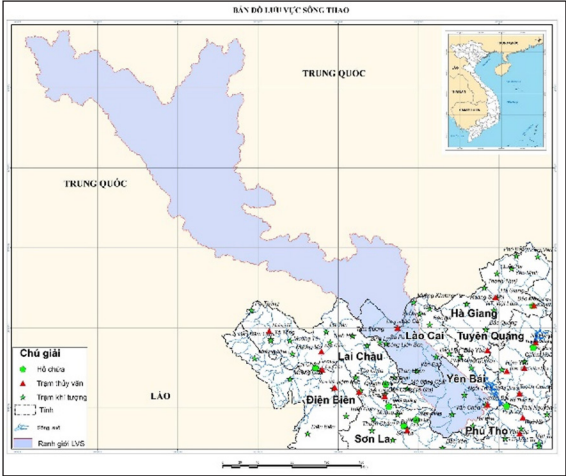
Mặc dù có nhiều ưu điểm, việc ứng dụng học sâu cũng đối mặt với một số thách thức, bao gồm yêu cầu khắt khe về lượng dữ liệu huấn luyện, nguy cơ quá khớp (overfitting), và mức độ “hộp đen” (black-box) của mô hình khi các thuật toán học một cách máy móc, chỉ dựa vào dữ liệu mà thiếu kiến thức về hệ thống thủy văn. Do đó, cần thiết phải xây dựng các mô hình học sâu giải thích được và kết hợp với kiến thức chuyên ngành để nâng cao độ tin cậy trong các ứng dụng thực tiễn. Chính vì vậy, nghiên cứu này được thực hiện nhằm xây dựng một mô hình học sâu MLP phù hợp với bài toán dự báo dòng chảy đỉnh lũ tại trạm thủy văn dựa vào các đặc trưng về khí tượng và thủy văn của lưu vực lân cận. Trong nghiên cứu, bộ thông số mô hình được phân tích và lựa chọn để phù hợp với đặc trưng gián đoạn của dữ liệu và mục đích dự báo. Để kiểm định mô hình học sâu MLP đã xây dựng, mô hình tiến hành dự báo dòng chảy lũ (thời kỳ lũ chính vụ) tại Trạm thủy văn Yên Bái từ dữ liệu khí tượng và thủy văn của các trạm lân cận.

2. Khu vực nghiên cứu và dữ liệu sử dụng

Nghiên cứu tiến hành dự báo dòng chảy đỉnh lũ đến Trạm thủy văn Yên Bái trên Sông Thao, dòng chính của Sông Hồng. Sông Thao bắt nguồn từ dãy núi Ngụ Sơn ở độ cao 1.776 m thuộc tỉnh Vân Nam Trung Quốc, chảy theo hướng Tây Bắc - Đông Nam song song với Sông Đà. Chiều dài sông 843 km trong đó, phần chảy qua địa phận Việt Nam là 332 km [10]. Trạm Yên Bái là trạm thủy văn cấp 1 đo đạc các yếu tố mực nước, nhiệt độ nước, lượng mưa, lưu lượng và phù sa. Lưu vực khống chế đến Trạm thủy văn Yên Bái có diện tích 48.000 km² chiếm 94% diện tích lưu vực Sông Thao (Hình 1). Dòng chảy đến trạm Yên Bái, vì thế, có thể xem như đại diện cho nguồn nước của lưu vực Sông Thao.

Để dự báo dòng chảy lũ đến trạm Yên Bái, các dữ liệu khí tượng, thủy văn được thu thập, sử dụng bao gồm:

- Dữ liệu thủy văn: Lưu lượng chân lũ Q_c và lưu lượng đỉnh lũ Q_{max} tại Trạm thủy văn Yên Bái từ năm 1961 đến 2024. Trong đó dữ liệu từ 1961-2022 dùng để xây dựng mô hình học sâu MLP, dữ liệu năm 2023-2024 được sử dụng để dự báo thử nghiệm.
- Dữ liệu khí tượng: Lượng mưa 1, 3, 5 ngày tích lũy tại các trạm khí tượng trên lưu vực khống chế bởi trạm Yên Bái (Sa Pa-SP, Văn Chấn-VC, Yên Bái-YB). Số liệu mưa được thu thập từ 1961-2024 với 62 năm số liệu phục vụ xây dựng mô hình, 2 năm số liệu dùng để dự báo thử nghiệm.



Hình 1. Bản đồ và mạng lưới trạm khí tượng thủy văn lưu vực đến trạm Yên Bái

3. Xây dựng mô hình học sâu dự báo đặc trưng lũ

3.1. Xử lý dữ liệu đầu vào

Để xây dựng được mô hình dự báo tốt, có hiệu suất cao, đảm bảo mô tả đúng đặc trưng vật lý của hệ thống thủy văn, cần thiết phải xử lý dữ liệu đầu vào. Nghiên cứu sử dụng 10 dữ liệu đầu vào với độ dài 62 năm (tương đương 62 số liệu) để dự báo đặc trưng dòng chảy lũ đến trạm Yên Bái. Tương quan giữa các biến đầu vào (mưa 1, 3, 5 ngày max các trạm và Q_c) và với biến mục tiêu (Q_{max}) được phân tích dựa vào hệ số tương quan Pearson (r) [11] theo công thức sau:

r = (Σ(x - x̄)(y - ȳ)) / (√Σ(x - x̄)² Σ(y - ȳ)²) (1)

Trong đó, x̄ và ȳ là giá trị trung bình của biến x và y. Hệ số tương quan r biến thiên từ -1 đến +1 với +1 thể hiện tương quan đồng biến hoàn hảo, -1 thể hiện tương quan nghịch biến hoàn hảo và 0 thể hiện không có tương quan tuyến tính. Kết quả xây dựng ma trận tương quan giữa các biến được thể hiện trong Hình 2.

Có thể thấy rằng lượng mưa tích lũy 1, 3, 5 ngày tại các trạm Sa Pa (SP1, SP3, SP5) và Văn Chấn (VC1, VC3, VC5) có mối tương quan khá tốt với dòng chảy đỉnh lũ tại Yên Bái, hệ số tương quan từ 0,49-0,65. Lượng mưa 1, 3, 5 ngày tại Yên Bái (YB1, YB3, YB5) có tương quan kém hơn với dòng chảy lũ tại Yên Bái, hệ số tương quan cao nhất là 0,28 (YB5). Để hạn chế tương quan bội giữa các biến đầu vào, gây nhiễu cho quá trình học của mô hình MLP, tiến hành sàng lọc, lựa chọn các biến đóng vai trò quan trọng trong việc hình thành lũ tại trạm Yên Bái. Các điều kiện để lựa chọn các biến đầu vào như sau:

- Biến đầu vào phải có tương quan với biến đầu ra (r/>0,2);
- Các biến đầu vào được lựa chọn phải tương đối độc lập với nhau;
- Nếu các cụm biến cùng có tương quan tốt với biến đầu ra, ưu tiên biến có tương quan tốt nhất.

Trên cơ sở đó, các biến được lựa chọn làm đầu vào của mô hình học sâu MLP dự báo đỉnh lũ tại Yên Bái bao gồm: Q_c, SP5, VC5, YB5 (Bảng 1).

r	Qmax	Qc	SP1	SP3	SP5	VC1	VC3	VC5	YB1	YB3	YB5
Qmax	1										
Qc	0.51	1									
SP1	0.57	0.15	1								
SP3	0.62	0.19	0.90	1							
SP5	0.65	0.15	0.85	0.95	1						
VC1	0.49	0.07	0.62	0.53	0.52	1					
VC3	0.49	0.03	0.59	0.58	0.58	0.88	1				
VC5	0.57	0.03	0.59	0.58	0.63	0.83	0.95	1			
YB1	0.06	-0.14	0.19	0.19	0.19	0.06	0.15	0.23	1		
YB3	0.22	-0.03	0.28	0.34	0.36	0.07	0.21	0.29	0.87	1	
YB5	0.28	0.04	0.30	0.37	0.41	0.15	0.27	0.37	0.82	0.94	1

Hình 2. Ma trận tương quan giữa các biến đầu vào và biến đầu ra (Q_{max}). Giá trị trong ma trận thể hiện hệ số tương quan Pearson

Bảng 1. Dữ liệu khí tượng thủy văn sử dụng trong dự báo đỉnh lũ tại Yên Bái

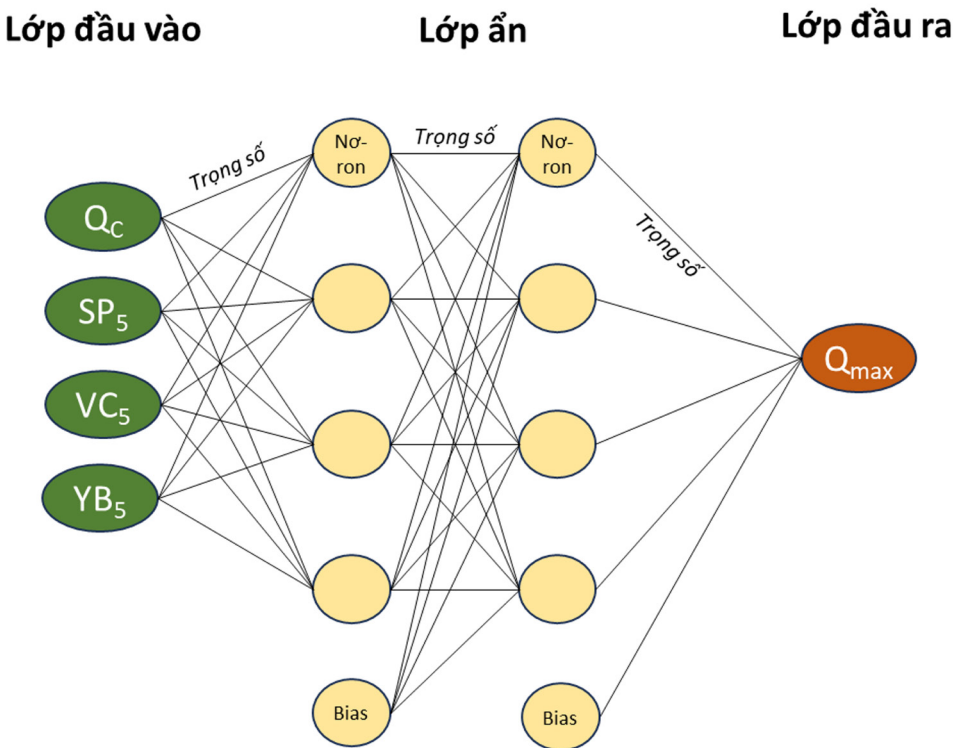
Loại biến	Ký hiệu	Giải thích
Đầu ra	Q _{max}	Lưu lượng đỉnh lũ tại trạm Yên Bái (m³/s)
Đầu vào	Q _c	Lưu lượng chân lũ lên tại trạm Yên Bái (m³/s)
	SP ₅	Lượng mưa tích lũy 5 ngày tại trạm Sa Pa (mm)
	VC ₅	Lượng mưa tích lũy 5 ngày tại trạm Văn Chấn (mm)
	YB ₅	Lượng mưa tích lũy 5 ngày tại trạm Yên Bái (mm)

3.2. Mô hình mạng nơ-ron đa tầng MLP

Multilayer Perceptron (MLP) là một loại mạng nơ-ron nhân tạo thuộc nhóm mạng nơ-ron truyền thẳng (feedforward neural networks), được sử dụng phổ biến trong các bài toán hồi quy và phân loại phi tuyến. MLP bao gồm ít nhất ba lớp: Một lớp đầu vào (input layer), một hoặc nhiều lớp ẩn (hidden layers), và một lớp đầu ra (output layer) (Hình 3). Mỗi nơ-ron trong một lớp được kết nối với tất cả các nơ-ron ở lớp kế tiếp thông qua các trọng số, và sử dụng một hàm kích hoạt phi tuyến như ReLU, sigmoid, hoặc tanh để học các mối quan hệ phức tạp trong dữ liệu [12]. Ngoài các nơ-ron đầu vào và nơ-ron ẩn, mỗi lớp (trừ lớp đầu vào) thường bao gồm một nơ-ron thiên vị (bias) cung cấp đầu vào không đổi cho các nơ-ron ở lớp tiếp theo. Các nơ-ron thiên vị có trọng số riêng liên quan đến từng kết nối, cũng được học trong quá trình đào tạo. Nơ-ron thiên vị dịch chuyển hiệu quả hàm kích hoạt của các nơ-ron ở lớp tiếp theo, cho phép mạng học được độ lệch hoặc độ lệch trong ranh giới

quyết định. Bằng cách điều chỉnh trọng số được kết nối với nơ-ron thiên vị, MLP có thể học cách kiểm soát ngưỡng kích hoạt và phù hợp hơn với dữ liệu đào tạo.

Trong quá trình huấn luyện, thuật toán lan truyền ngược (backpropagation) được sử dụng để điều chỉnh trọng số thông qua phương pháp tối ưu hóa như “gradient descent” hoặc “Adam optimizer” nhằm giảm sai số dự báo giữa giá trị thực đo và dự báo. MLP đặc biệt phù hợp với các bài toán thủy văn khi dữ liệu đầu vào có mối quan hệ phức tạp và không tuân theo phân phối tuyến tính. Do khả năng biểu diễn các hàm phi tuyến bất kỳ, MLP được xem là một bộ xấp xỉ hàm toàn năng (universal function approximator) nếu có đủ số lượng nơ-ron và lớp ẩn [13]. Tuy nhiên, hiệu quả của MLP phụ thuộc mạnh vào số lượng dữ liệu, việc lựa chọn cấu trúc mạng (số lớp, số nơ-ron), và các kỹ thuật điều chỉnh như regularization hoặc dropout để tránh hiện tượng quá khớp (overfitting), đặc biệt là khi tập dữ liệu nhỏ.



Hình 3. Ví dụ về cấu trúc mạng MLP dự báo đỉnh lũ tại Yên Bái (Trong đó: Bias là nơ-ron thiên vị, các ký hiệu khác được giải thích ở Bảng 1)

3.2.1. Kỹ thuật xử lý dữ liệu phục vụ huấn luyện mô hình

Với chuỗi dữ liệu gồm 62 trận lũ, không có sự liên tục về thời gian và chứa những quan hệ phi tuyến phức tạp giữa mưa và dòng chảy, cần thực hiện một số kỹ thuật để đảm bảo mô hình MLP không bị quá khớp (overfitting) và học lệch (gradient learning). Các kỹ thuật được ứng dụng trong nghiên cứu này bao gồm:

- Chuẩn hóa dữ liệu bằng hàm “StandardScaler” trong ngôn ngữ lập trình Python. Với mỗi biến đầu vào “StandardScaler” sẽ biến đổi theo công thức:

$$x_{scaled} = \frac{x - \mu}{\sigma} \quad (2)$$

Trong đó, μ là trung bình của dữ liệu và σ là độ lệch chuẩn của dữ liệu đó.

- Bổ sung đặc trưng tương tác giữa các biến đầu vào, ví dụ $X1 * X2$;

- Sử dụng kỹ thuật k-fold cross validation (k=5) để đánh giá mô hình thay vì chỉ chia tập huấn luyện/kiểm tra thông thường. Kỹ thuật này chia dữ liệu thành nhiều phần nhỏ để huấn luyện và kiểm tra nhiều lần [14], từ đó giảm tính ngẫu nhiên của việc lựa chọn tập huấn luyện/kiểm tra.

3.2.2. Lựa chọn cấu trúc mạng nơ-ron phù hợp

Rõ ràng, hiệu suất của mô hình MLP phụ thuộc rất nhiều vào việc lựa chọn cấu trúc mạng nơ-ron phù hợp. Với bài toán dự báo tập dữ liệu nhỏ, cấu trúc nơ-ron càng phức tạp, càng nhiều

trọng số mô hình thì càng gây nhiễu cho quá trình huấn luyện. Chính vì vậy, nghiên cứu tiến hành đánh giá, lựa chọn mô hình phù hợp với bài toán dự báo đỉnh lũ tại Yên Bái như sau:

- Do kích thước chuỗi dữ liệu khá ngắn nên mô hình mạng MLP được đề xuất bao gồm 2 lớp ẩn.

- Xây dựng mạng nơ-ron MLP với số lượng nơ-ron ở mỗi lớp từ 2 đến 25 nơ-ron theo khuyến nghị của [15] ứng với trường hợp chuỗi dữ liệu ngắn, số nơ-ron lớp ẩn thứ nhất bằng với lớp ẩn thứ hai.

- Với mỗi cấu trúc mạng tiến hành chạy mô hình 100 lần, hiệu suất mô hình được đánh giá thông qua hệ số R^2 (hệ số xác định [0-1]) và MAE (sai số tuyệt đối trung bình, m^3/s) theo các công thức sau:

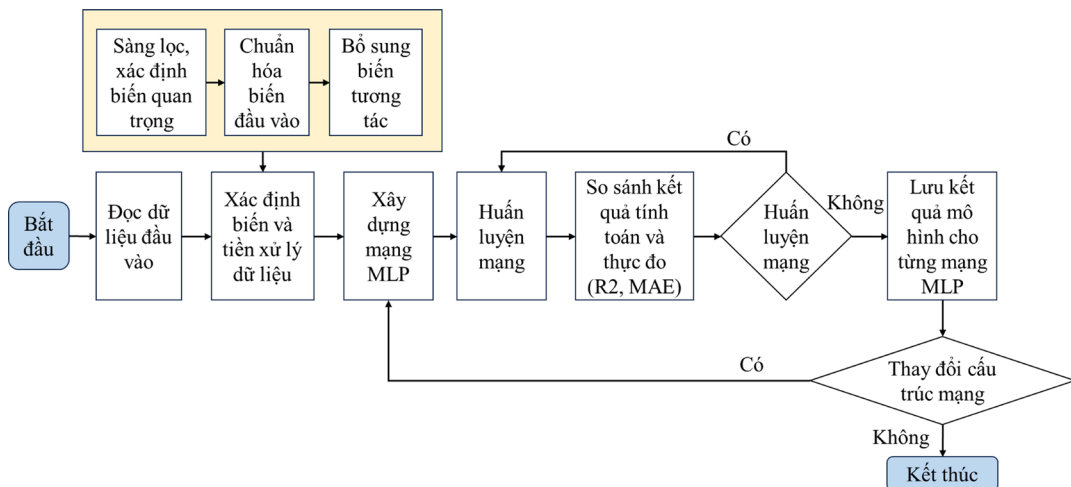
$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (3)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (4)$$

Trong đó y_i là giá trị lưu lượng thực đo, $\hat{y}_i$ là giá trị lưu lượng dự báo (m^3/s) và $\bar{y}$ là giá trị trung bình của lưu lượng thực đo (m^3/s).

Đánh giá phân phối của kết quả chạy mạng MLP nhiều lần để lựa chọn cấu trúc mạng tốt nhất, cho kết quả huấn luyện và kiểm tra ổn định nhất.

Sơ đồ khối thực hiện xây dựng mô hình mạng MLP được thể hiện trong Hình 4.



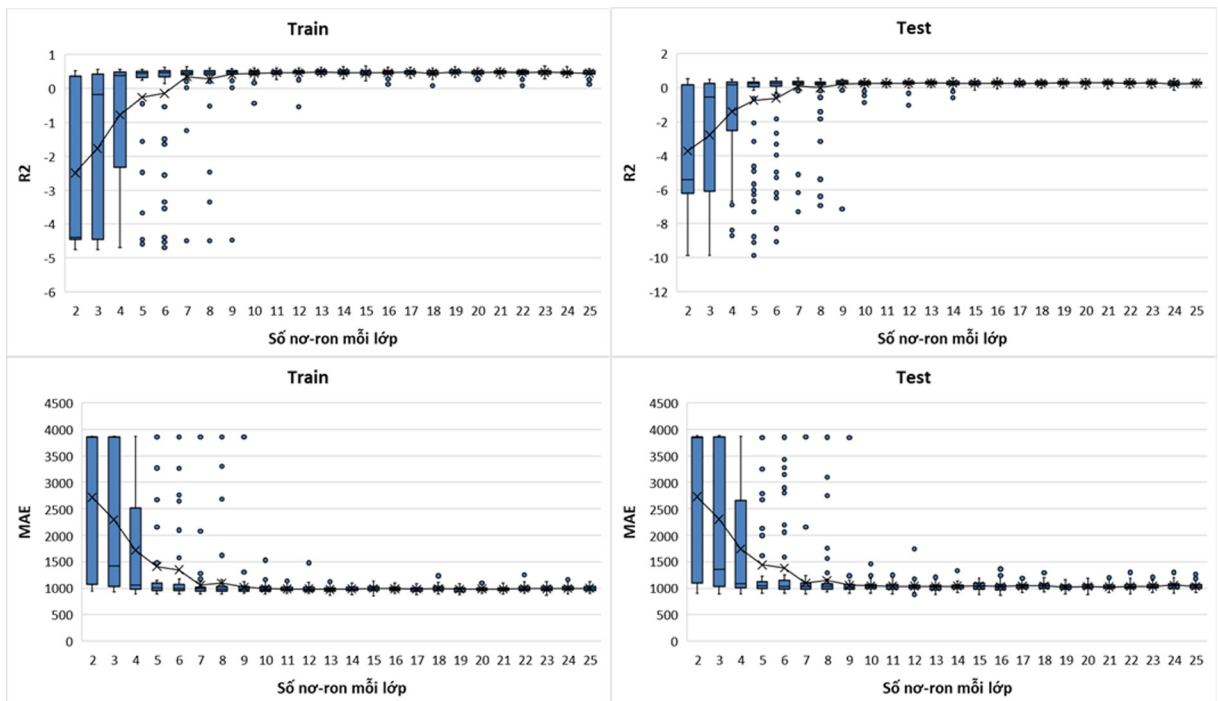
Hình 4. Sơ đồ khối xây dựng mô hình MLP dự báo đỉnh lũ tại Yên Bái

4. Kết quả dự báo đỉnh lũ tại Yên Bái

4.1. Lựa chọn cấu trúc mô hình MLP

Phân phối của 100 lần chạy huấn luyện 24 cấu trúc mạng MLP (từ 2-25 nơ-ron) đối với các tập dữ liệu huấn luyện và kiểm tra được thể hiện trong Hình 5. Kết quả luyện mạng cho thấy, cấu trúc mạng từ 9x9 (mỗi lớp ẩn có từ 9 nơ-ron) có hệ số R^2 và MAE trung bình tốt nhất và kết quả luyện mạng ít phân tán nhất. Về cơ bản, mạng MLP càng nhiều nơ-ron thì hiệu suất mô hình càng tăng, tuy nhiên, mức tăng hiệu suất không đáng kể với

nhiều cấu trúc mạng từ 11 nơ-ron trở đi. Bên cạnh đó, có thể thấy rằng phân phối kết quả huấn luyện và kiểm tra của các cấu trúc mạng 10x10, 12x12 và ngay cả 25x25 cũng tồn tại những điểm ngoại biên. Điều này thể hiện cấu trúc mô hình không ổn định, phụ thuộc vào đặc điểm ngẫu nhiên của lần huấn luyện. Căn cứ vào kết quả huấn luyện trung bình và tính ổn định của mô hình, nghiên cứu lựa chọn cấu trúc 11x11 (2 lớp ẩn, mỗi lớp có 11 nơ-ron) là cấu trúc phù hợp nhất để dự báo đỉnh lũ cho trạm Yên Bái.

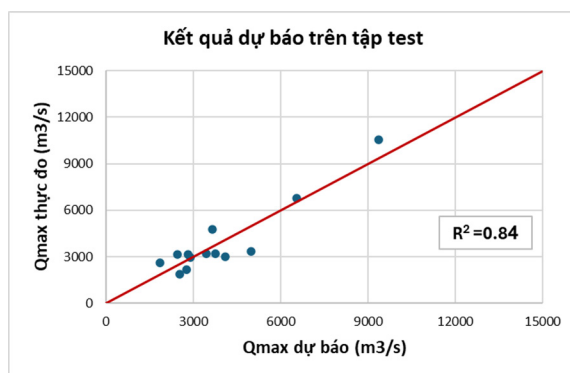
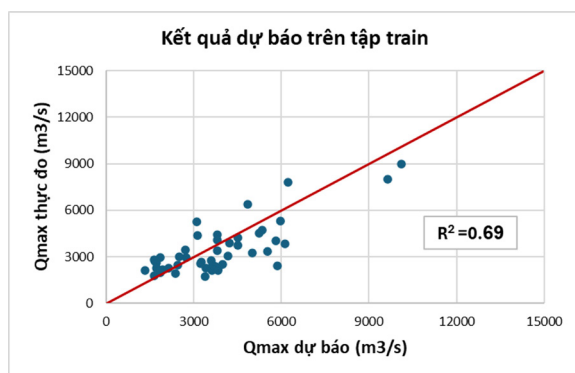


Hình 5. Phân phối hệ số R^2 và MAE (m^3/s) của các cấu trúc mạng MLP trên cả tập train và test

4.2. Kết quả huấn luyện mô hình mạng MLP

Sử dụng mạng MLP tốt nhất đã tìm được, tiến hành huấn luyện và kiểm tra với tập dữ liệu dự báo đỉnh lũ tại Yên Bái trong 1.000 lần chạy và lựa chọn mô hình tốt nhất với hệ số R^2 của tập test lớn nhất và chênh lệch giữa R^2_{train} và $R^2_{\text{test}} < 0,15$ để đảm bảo việc kiểm tra tốt không phải là do “may mắn”. Kết quả cho thấy

hệ số xác định R^2 đạt 0,69 đối với tập huấn luyện và đạt 0,84 đối với tập kiểm tra (Hình 6). Sai số tuyệt đối trung bình đối với tập kiểm tra là 682,71 m^3/s (bằng 17,7% lưu lượng đỉnh lũ trung bình). Như vậy có thể thấy rằng kết quả huấn luyện mô hình mạng MLP để dự báo đỉnh lũ đến Trạm thủy văn Yên Bái là khá tốt, nhất là với nguồn dữ liệu ngắn và có sự biến thiên lớn về mặt giá trị.



Hình 6. Kết quả huấn luyện và kiểm tra mô hình MLP

4.3. Kết quả dự báo thử nghiệm mô hình

Để có thể sử dụng mô hình mạng MLP đã xây dựng trong tương lai, tiến hành dự báo thử nghiệm cho trận lũ năm 2023 và 2024. Đầu vào dự báo bao gồm dòng chảy chân lũ và lượng mưa thực đo tại các trạm. Kết quả dự báo đỉnh

lũ tại Yên Bái cho thấy kết quả dự báo đỉnh lũ năm 2023 khá tốt với sai số dưới 10%,. Trong khi đó, sai số dự báo đỉnh lũ năm 2024 cho kết quả kém hơn, sai số đỉnh lũ 20,9% (Bảng 2). Do sự phức tạp và tính phi tuyến của mưa-dòng chảy nên sai số dự báo này có thể chấp nhận được.

Bảng 2. Kết quả dự báo thử nghiệm đỉnh lũ tại Yên Bái

Năm	$Q_{\max}$ thực đo (m^3/s)	$Q_{\max}$ dự báo (m^3/s)	Sai số (%)
2023	2.670	2.431	9,8
2024	2.410	1.993	20,9

5. Kết luận

Bài báo đã đưa ra một hướng tiếp cận mới trong dự báo đỉnh lũ bằng cách xây dựng và thử nghiệm mô hình học sâu sử dụng mạng trí tuệ nhân tạo đa tầng (MLP) cho chuỗi dữ liệu ngắn và không liên tục. Bằng cách tích hợp các kỹ thuật tiền xử lý và chuẩn hóa dữ liệu phù hợp, mô hình MLP đã được huấn luyện hiệu quả trên tập dữ liệu gồm 62 sự kiện lũ và thử nghiệm dự báo cho hai trận lũ gần đây tại Trạm thủy văn Yên Bái.

Nghiên cứu cũng tiến hành phân tích, lựa chọn cấu trúc mạng MLP phù hợp với 2 lớp ẩn, mỗi lớp có 11 nơ-ron để tiến hành huấn luyện mô hình phục vụ dự báo. Kết quả huấn luyện mô hình cho thấy kết quả khá tốt ở cả tập dữ liệu huấn luyện và kiểm tra, hệ số xác

định R^2 đạt 0,69-0,84, sai số tuyệt đối trung bình nhỏ. Mô hình đã huấn luyện được sử dụng để dự báo thử nghiệm cho năm 2023 và 2024, kết quả dự báo đỉnh lũ tại Yên Bái cho sai số 9,8% với lũ năm 2023 và 20,9% với lũ năm 2024.

Tuy nhiên, nghiên cứu vẫn còn một số hạn chế cần được khắc phục trong các bước tiếp theo. Mặc dù sai số đỉnh lũ có thể chấp nhận được và mô hình có thể được sử dụng trong dự báo nhưng nghiên cứu mới thực hiện thử nghiệm với đầu vào là dữ liệu lưu lượng chân lũ và lượng mưa tích lũy thực đo mà chưa sử dụng kết quả dự báo mưa và nhận định chân lũ. Do đó, trong những nghiên cứu sau này cần bổ sung, xem xét đến dữ liệu dự báo các yếu tố đầu vào.

Đóng góp của các tác giả: Ngô Thị Thủy: Xây dựng code python, phân tích kết quả và viết chính bài báo; Lương Hữu Dũng: Xây dựng ý tưởng, chuẩn bị dữ liệu đầu vào, phân tích kết quả và tham gia viết bài; Trịnh Thu Phương: Xây dựng ý tưởng và chuẩn bị dữ liệu đầu vào; Lương Tuấn Trung: Tham gia viết code và xây dựng ý tưởng; Chu Nguyễn Ngọc Sơn: Chuẩn bị dữ liệu đầu vào và xây dựng bản đồ.

Lời cảm ơn: Nghiên cứu này là sản phẩm của Đề tài “Nghiên cứu khả năng chống chịu của hệ thống công trình phòng chống lũ, xây dựng công nghệ nhận dạng lũ lớn, tổ hợp lũ bất lợi trên lưu vực Sông Hồng và đề xuất giải pháp ứng phó” mã số ĐTĐL.CN-47/22.

Lời cam đoan: Tập thể tác giả cam đoan bài báo này là công trình nghiên cứu của tập thể tác giả, chưa được công bố ở đâu, không được sao chép từ những nghiên cứu trước đây; không có sự tranh chấp lợi ích trong nhóm tác giả.

Tài liệu tham khảo

1. Widiyari, I. R. et al. (2017), "Deep learning multilayer perceptron (MLP) for flood prediction model using wireless sensor network based hydrology time series data mining", *International Conference on Innovative and Creative Information Technology (ICITech)*, Salatiga, Indonesia, pp. 1-5, doi: 10.1109/INNOCIT.2017.8319150.
2. Valles, J. (2023), "Application of a Multilayer Perceptron Artificial Neural Network (MLP-ANN) in Hydrological Forecasting in El Salvador", in *Advanced Hydroinformatics: Machine Learning and Optimization for Water Resources*, <https://doi.org/10.1002/9781119639268.ch8>
3. Dawson, C. W., & Wilby, R. L. (2001), "Hydrological modelling using artificial neural networks", *Progress in Physical Geography*, 25(1), 80-108.
4. Kratzert, F., et al. (2018), "Rainfall-runoff modelling using Long Short-Term Memory (LSTM) networks", *Hydrol. Earth Syst. Sci.*, 22, 6005–6022, <https://doi.org/10.5194/hess-22-6005-2018>.
5. Zhou, F. et al. (2023), "Application of a New Hybrid Deep Learning Model That Considers Temporal and Feature Dependencies in Rainfall-Runoff Simulation", *Remote Sensing*, 15(5), 1395. <https://doi.org/10.3390/rs15051395>
6. Gasim Hayder, Mahmud Iwan Solihin, M. R. N. Najwa (2022), "Multi-step-ahead prediction of river flow using NARX neural networks and deep learning LSTM", *H2Open Journal* 1 March 2022; 5 (1): 43–60. doi: <https://doi.org/10.2166/h2oj.2022.134>
7. Nguyễn Đăng Tính (2008), "Ứng dụng mạng Noron nhân tạo để dự báo mưa và dòng chảy nhằm phòng tránh và giảm nhẹ thiên tai hạn hán các lưu vực sông thuộc Tây Nguyên", *Tạp chí khoa học kỹ thuật Thủy Lợi và Môi Trường*, số 22 (2008).
8. Hồ Việt Tuấn, Hồ Việt Hùng (2019), "Sử dụng mạng nơ-ron nhân tạo dự báo mực nước sông chịu ảnh hưởng của thủy triều", *Tạp chí Khoa học và Công nghệ Thủy lợi*, số 52 (2019), 1-9.
9. Phạm Văn Chiến và Nguyễn Hoàng Bách (2024), "Mô hình bộ nhớ dài-ngắn LSTM cho mô phỏng dòng chảy lưu vực sông Thu Bồn", *Tạp chí Khoa học và Công nghệ Thủy lợi*, số 83 (2024), 103-111.
10. Trần Thanh Xuân (2016), *Mạng lưới và tài nguyên nước sông Việt Nam, những biến đổi và thách thức*. Nhà xuất bản Khoa học và Kỹ thuật.
11. Stigler, S. M. (1989), "Francis Galton's account of the invention of correlation", *Statistical Science*, 4 (2): 73-79.
12. Goodfellow, I. et al. (2016), *Deep learning*, MIT press.
13. Hornik, K. et al. (1989), "Multilayer feedforward networks are universal approximators", *Neural Networks*, 2(5), 359-366. [https://doi.org/10.1016/0893-6080\(89\)90020-8](https://doi.org/10.1016/0893-6080(89)90020-8)
14. Kohavi, R. (1995), *A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection*, Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI).
15. Ingrassia S. & Morlini I. (2005), "Neural Network Modeling for Small Datasets", *Technometrics*, Vol 47 (3), 297-311.

DEEP LEARNING MODEL FOR FLOOD CHARACTERISTIC PREDICTION AT YEN BAI HYDROLOGICAL STATION

Ngo Thi Thuy⁽¹⁾, Luong Huu Dung⁽¹⁾, Trinh Thu Phuong⁽²⁾,
Luong Tuan Trung⁽¹⁾, Chu Nguyen Ngoc Son⁽¹⁾

⁽¹⁾The Viet Nam Institute of Meteorology, Hydrology and Climate Change

⁽²⁾Mekong River Commission

Received: 21/5/2025; Accepted: 08/6/2025

Abstract: Artificial intelligence (AI), particularly deep learning models, has been widely applied in hydrology and water resources. In river flood forecasting, machine learning and deep learning models have demonstrated promising potential due to their fast computation and high accuracy, even with short and event-specific datasets lacking temporal continuity. Despite these advantages, deep learning models and AI in general require expert knowledge to guide learning and avoid overfitting or misinterpretation. This study aims to develop a deep learning model based on a Multilayer Perceptron (MLP) neural network to forecast flood characteristics at the Yen Bai hydrological station on the Thao River. Using flood recession flow data and 5-day accumulated rainfall from the Sa Pa, Van Chan, and Yen Bai stations, an MLP model with a 11x11 architecture was constructed. The model was trained on 62 historical flood events (1961–2022) and achieved good performance, with R^2 values ranging from 0.69 to 0.84 and a mean absolute error (MAE) of less than 20% of the average flood peak discharge. Validation using 2023–2024 data yielded peak errors of 9.8% and 20.9%, respectively.

Keywords: Multilayer Perceptron, flood forecast, Thao River, deep learning.