Tạp chí Phát triển Khoa học và Công nghệ – Khoa học Tự nhiên, 6(1):1884-1896

Open Access Full Text Article

Bài nghiên cứu

Áp dụng mạng trí nhớ ngắn hạn định hướng dài hạn (Long Short–Term Memory) để dự báo lưu lượng nước tại trạm thủy văn Mỹ Thuận trên sông Tiền

Trần Thành Thái[1,*], Phạm Ngọc Hoài[2], Phạm Bảo Quốc[2], Nguyễn Phan Nhân[3], Nguyễn Phương Đông[4], Nguyễn Duy Liêm[5], Khuất Thùy Phương[6]

Use your smartphone to scan this QR code and download this article

[1]*Viện Sinh học nhiệt đới, Viện Hàn lâm Khoa học và Công nghệ Việt Nam. TP. Hồ Chí Minh, Việt Nam*

[2]*Viện Phát triển Ứng dụng, Trường Đại học Thủ Dầu Một. TP. Thủ Dầu Một, tỉnh Bình Dương, Việt Nam*

[3]*Sở Tài nguyên và Môi trường Bến Tre. TP. Bến Tre, tỉnh Bến Tre*

[4]*Phân Viện Khoa học Khí tượng Thủy văn và Biến đổi khí hậu. TP. Hồ Chí Minh, Việt Nam*

[5]*Trường Đại học Nông Lâm. TP. Hồ Chí Minh, Việt Nam*

[6]*Trung tâm Tin học, Trường Đại học Khoa học Tự nhiên – ĐHQG-HCM, Việt Nam*

Liên hệ

Trần Thành Thái, Viện Sinh học nhiệt đới, Viện Hàn lâm Khoa học và Công nghệ Việt Nam. TP. Hồ Chí Minh, Việt Nam

Email: thanhthai.bentrect@gmai.com

Lịch sử

- Ngày nhận: 13-09-2021
- Ngày chấp nhận: 13-01-2022
- Ngày đăng: 28-02-2022

DOI : 10.32508/stdjns.v6i1.1129

Check for updates

TÓM TẮT

Dự báo lưu lượng nước rất quan trọng trong công tác quản lý nguồn nước để cảnh báo lũ lụt và chủ động nước cho nông nghiệp. Tuy nhiên, dự báo chính xác lưu lượng nước là việc rất khó khăn do có nhiều yếu tố thủy văn tác động. Cho nên, mục tiêu của nghiên cứu là bước đầu kiểm tra khả năng áp dụng thuật toán Mạng trí nhớ ngắn hạn định hướng dài hạn (LSTM, Long Short–Term Memory), là một thuật toán học sâu của học máy, trong dự báo lưu lượng nước tại trạm thủy văn Mỹ Thuận trên sông Tiền. Dữ liệu lưu lượng nước sử dụng trong nghiên cứu được thu thập theo giờ trong năm 2018. Các chỉ số thống kê như Hệ số xác định (Coefficient of determination, R^2), 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. Nghiên cứu đánh giá khả năng dự báo lưu lượng nước của thuật toán LSTM với số nơron khác nhau (1, 2, 3, 4) ở các thời gian dự báo khác nhau: 1, 2, 3, 4, 5 giờ tiếp theo (t + 1, t + 2, t + 3, t + 4, t + 5, tương ứng). Kết quả cho thấy mô hình mạng LSTM có 3 nơron dự báo tối ưu nhất ở 1 giờ tiếp theo (t + 1) với R^2 = 0,937, RMSE = 2294,60, MAE = 1738,33 cho tập huấn luyện, R^2 = 0,884, RMSE = 2655,66, MAE = 2064,30 cho tập kiểm tra. Mô hình mạng LSTM dự báo khá tốt lưu lượng nước trạm Mỹ Thuận và tiềm năng ứng dụng cho các trạm thủy văn khác.

Từ khoá: Biến đổi khí hậu, chuỗi thời gian, đồng bằng sông Cửu Long, học sâu, học máy

GIỚI THIỆU

Lũ lụt là hiện tượng tự nhiên nguy hiểm, có thể ảnh hưởng đến đời sống người dân, nhất là cộng đồng dân cư vùng hạ lưu. Dự báo chính xác thời điểm và dấu hiệu của lũ lụt là rất quan trọng trong công tác quản lý nguồn nước để giảm thiểu tác động và có kế hoạch ứng phó kịp thời với tình huống lũ lụt bất ngờ. Tuy nhiên, dự báo chính xác lưu lượng nước là vấn đề vô cùng khó khăn, do lưu lượng dòng chảy chịu tác động của rất nhiều yếu tố thủy văn như chế độ triều, chế độ dòng chảy, lượng mưa, địa hình lòng sông[1]. Ngoài ra, hoạt động tích/xả nước của hệ thống đập thủy điện vùng thượng nguồn cũng tác động rất lớn đến lưu lượng dòng chảy. Lấy trường hợp hệ thống sông Mê Kông làm điển hình, sáu đập thủy điện ở thượng nguồn sông Mê Kông đã quyết định đến hơn 40% lưu lượng dòng chảy vùng hạ nguồn[2,3]. Nhiều nghiên cứu đã cố gắng đưa ra các phương pháp để dự báo lưu lượng nước. Các dự báo về lưu lượng nước chủ yếu dự vào các mô hình toán và có thể chia các mô hình này thành hai nhóm: (1) Mô hình tiến trình (Process–based models) là loại mô hình kết hợp toán–vật lý để đưa ra dự báo, các quy luật thủy văn được mô tả và ước tính chính xác nhờ các quy luật vật lý được nghiên cứu và tích hợp sẵn trong mô hình. Ưu điểm của nhóm này là tính chính xác rất cao và các quy luật thủy văn có thể được giải thích thông qua các quy luật toán–vật lý; tuy nhiên, nhóm này thường đòi hỏi phải có chuyên gia vì mô hình vận hành rất phức tạp. Hơn nữa, số lượng đầu vào, là dữ liệu của các yếu tố ảnh hưởng đến lưu lượng, phải rất lớn mới đảm bảo tính chính xác[4,5]. Một cách tiếp cận khác là sử dụng (2) các thuật toán học máy (machine learning) trong dự báo mặn, ví dụ K–Nearest Neighbors, Decision Tree, Random Forest, Support Vector Machine, Artificial eural etwork, và Long Short–Term Memory (LSTM). Phương pháp này có ưu điểm là dễ áp dụng, độ chính xác cao, không đòi hỏi số lượng dữ liệu lớn; tuy nhiên, do các mô hình học máy thuộc nhóm "black–box" nên đôi khi rất khó để giải thích kết quả[4,6,7]. Thu thập thông tin về toàn bộ các yếu tố ảnh hưởng đến lưu lượng dòng chảy là vô cùng khó khăn và thường không đầy đủ[2]. Cho nên, lựa chọn các mô hình học máy để dự báo lưu lượng dòng chảy trong trường hợp này là phù hợp.

Trích dẫn bài báo này: Thái T T, Hoài P N, Quốc P B, Nhân N P, Đông N P, Liêm N D, Phương K T. **Áp dụng mạng trí nhớ ngắn hạn định hướng dài hạn (Long Short–Term Memory) để dự báo lưu lượng nước tại trạm thủy văn Mỹ Thuận trên sông Tiền**. *Sci. Tech. Dev. J. - Nat. Sci.;* 6(1):1884-1896.

Hiện nay, thuật toán mạng trí nhớ ngắn hạn định hướng dài hạn LSTM, là một trong những thuật toán học sâu (deep learning) mới nhất, mô hình này đã ứng dụng thành công vào nhiều lĩnh vực: Nhận diện giọng nói[8]; chuyển đổi ngôn ngữ[9], dự báo lượng khách du lịch[10], chứng khoán[11], lượng mưa[12]. Cho nên, LSTM cũng tiềm năng trong dự báo các đặc điểm thủy văn học. Mô hình LSTM rất thành công trong dự báo biến động mực nước ở các thủy vực: Ở một số hồ ở Phần Lan với độ chính xác R^2 từ 0,84 đến 0,97[13], vịnh Narragansett (Mỹ) với R^2 là 0,99[14], hồ Dongting (Trung Quốc) với R^2 là 0,99[15]. Tác giả Le và cộng sự đã áp dụng mô hình LSTM trong dự báo lưu lượng nước tại trạm thủy văn Sơn Tây trên sông Hồng, kết quả dự báo rất chính xác, gần như tuyệt đối với chỉ số NSE (Nash–Sutcliffe Efficiency) đạt từ 0,992 đến 0,993 ở thời gian dự báo t + 1[1]. Mặc dù được ứng dụng rộng rãi và thành công ở rất nhiều nơi trên thế giới, mạng LSTM còn chưa được chú ý nghiên cứu nhiều ở Việt Nam. Đặc biệt trong mô phỏng các đặc điểm thủy văn, mô hình mạng LSTM cho thấy chúng có thể dự báo ở độ chính xác cao.

Do đó, nghiên cứu được thực hiện với mục tiêu: (i) Đánh giá tính hiệu quả của mô hình mạng LSTM trong dự báo lưu lượng nước ở trạm Mỹ Thuận trên sông Tiền, (ii) tìm ra các thông số tối ưu của mô hình, phù hợp với bộ dữ liệu hiện có. Kết quả từ nghiên cứu có thể cung cấp một cách tiếp cận mới, tiềm năng trong quản lý tài nguyên nước và nghiên cứu dự báo các đặc điểm thủy văn, biến động môi trường.

VẬT LIỆU VÀ PHƯƠNG PHÁP NGHIÊN CỨU

Thu thập và tiền xử lý dữ liệu

Trạm thủy văn Mỹ Thuận nằm trên nhánh sông Tiền, thuộc hệ thống sông Mê Kông, có địa chỉ tại 192 Vĩnh Thuận Tân, xã Tân Ngãi, thành phố Vĩnh Long (Hình 1). Dữ liệu về lưu lượng nước (m^3/s) tại trạm Mỹ Thuận được thu thập từ Đài Khí tượng Thủy văn khu vực Nam Bộ (http://www.kttv-nb.org.vn/), số liệu được đo theo giờ trong năm 2018. Bảng 1 mô tả thống kê bộ dữ liệu lưu lượng nước về số lượng dữ liệu, trung bình, độ lệch chuẩn, giá trị nhỏ–lớn nhất, điểm phân vị thứ 25, 50, và 75.

Đối với dữ liệu chuỗi thời gian, chất lượng dữ liệu (số lượng đủ lớn, liên tục, ít giá trị ngoại lai…) quyết định đến tính chính xác của các mô hình dự báo[16]. Cho nên, dữ liệu được tiền xử lý qua hai bước trước khi được đưa vào mô hình để huấn luyện thuật toán:

1) 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[17]. Biểu đồ hộp dùng để mô tả phân bố của số liệu 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). 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[18].

2) Tất cả các số liệu được chuẩn hóa về chung một phạm vi từ 0 đến 1 bằng phương pháp bình thường hóa dữ liệu (normalization scaling, công thức 1), được thực hiện trong thư viện scikit–learn của Python[19]:

$$z = \frac{x - min(x)}{max(x) - min(x)} \quad (1)$$

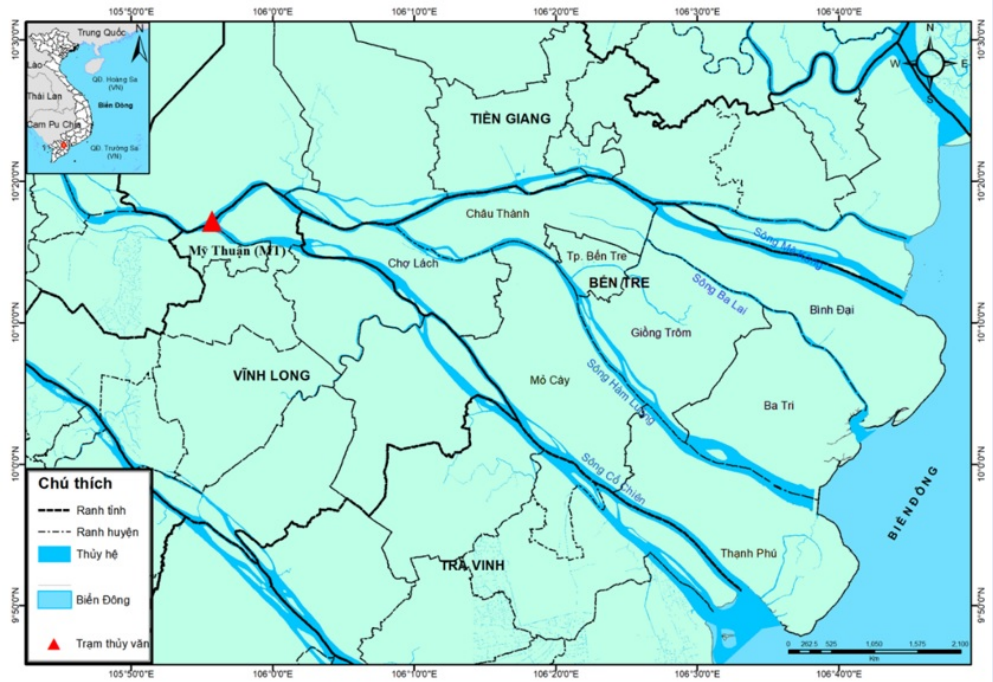
Trong đó, *max(x)*, *min(x)* là giá trị lớn nhất/nhỏ nhất của bộ dữ liệu, x là giá trị dữ liệu, z là giá trị đã được chuẩn hóa từ x.

Chuẩn hóa giúp hội tụ và ngăn cản sự phân tán của dữ liệu[20] (Bảng 1). Ngoài ra, chuẩn hóa giúp tăng tốc quá trình huấn luyện thuật toán mà không ảnh hưởng đến tính chính xác[21].

Mạng trí nhớ ngắn hạn định hướng dài hạn (LSTM, Long Short - Term Memory)

Mạng nơron hồi quy (RNN, Recurrent Neural Network) được thiết kế cho việc xử lý các loại dữ liệu có dạng chuỗi tuần tự. Mạng RNN nhận một đầu vào x_t, tiến hành xử lý và đưa ra đầu ra h_t . RNN sẽ lưu lại giá trị của h_t để sử dụng cho đầu vào tiếp theo. Có thể coi một mạng RNN là một chuỗi những mạng con giống hệt nhau, mỗi mạng sẽ truyền thông tin nó vừa xử lý cho mạng phía sau nó. Nếu ta tách từng vòng lặp xử lý trong RNN ra thành từng mạng nhỏ thì sẽ có một mạng RNN có kiến trúc như Hình 2a. Chuỗi các đầu vào x_{t-1}, x_t, x_{t+1} là những sự kiện xảy ra theo thứ tự thời gian và có mối liên hệ về thông tin với nhau. Thông tin của chúng sẽ được giữ lại để xử lý sự kiện tiếp theo trong mạng RNN. Về mặt lý thuyết thì RNN có thể xử lý và lưu trữ thông tin của một chuỗi dữ liệu với độ dài bất kỳ. Tuy nhiên trong thực tế thì RNN chỉ tỏ ra hiệu quả với chuỗi dữ liệu có độ dài không quá lớn. Nguyên nhân do vấn đề triệt tiêu/bùng nổ đạo hàm (vanishing/ exploding gradient problems) khi xử lý các chuỗi dữ liệu có độ dài lớn[22].

Mạng LSTM là một kiến trúc nâng cấp đặc biệt của mạng RNN được phát minh bởi Hochreiter và Schmidhuber năm 1997[23]. Mạng LSTM khắc phục được nhược điểm của mạng RNN truyền thống về phụ thuộc xa (long–term dependencies)[22]. Cấu trúc của mạng LSTM bao gồm nhiều tế bào LSTM (LSTM memory cell) liên kết với nhau (Hình 2b). Mạng LSTM bổ sung thêm trạng thái bên trong tế bào và ba cổng sàng lọc các thông tin đầu vào và đầu ra cho tế bào bao gồm forget gate, input gate, và output gate.



Hình 1: Bản đồ vị trí trạm thủy văn Mỹ Thuận trên sông Tiền

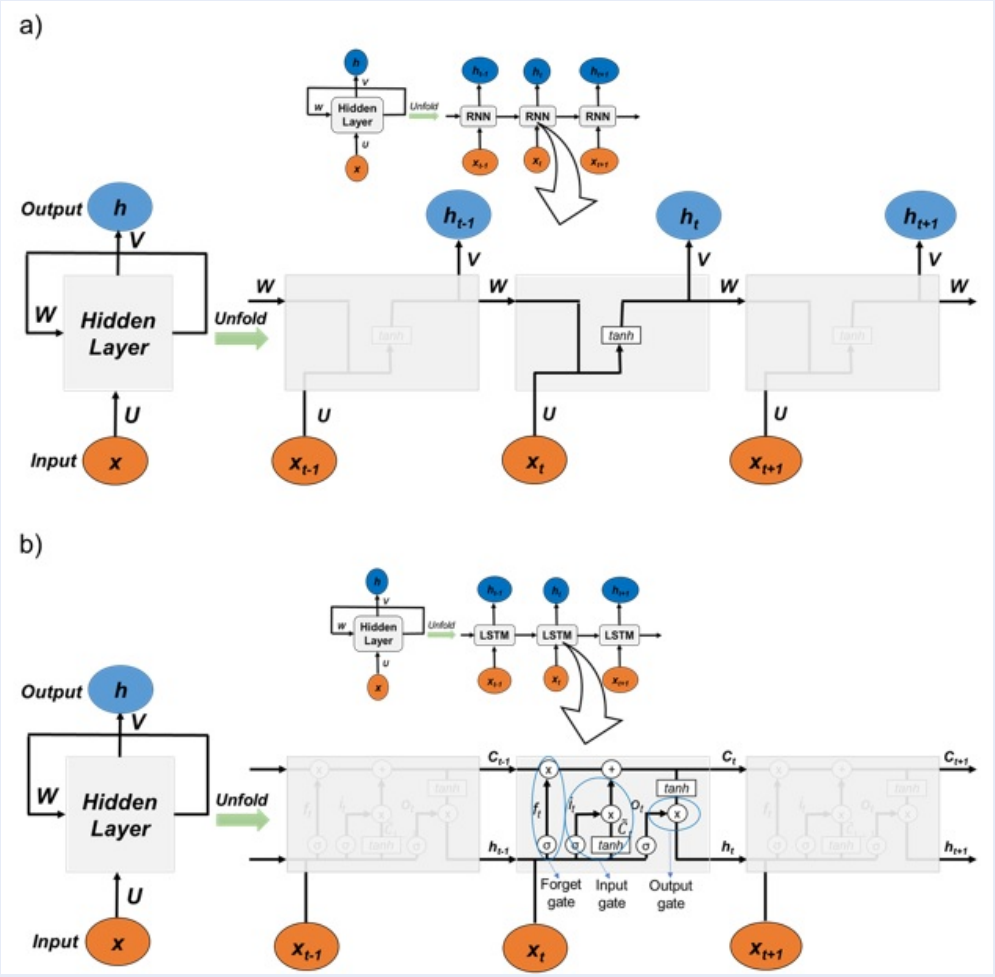
Bảng 1: Thống kê mô tả bộ dữ liệu về lưu lượng nước tại trạm thủy văn Mỹ Thuận trong năm 2018

Đặc điểm dữ liệu	Lưu lượng tại Mỹ Thuận	
	Trước chuẩn hóa	Sau chuẩn hóa
Số dữ liệu (Count)	8.760	8.760
Trung bình (Mean, m^3/s)	5.652,06	0,54
Độ lệch chuẩn (Std, m^3/s)	9.107,25	0,25
Cực tiểu (Min, m^3/s)	-14.500	0,00
Phân vị 25% (m^3/s)	-2.110	0,33
Phân vị 50% (m^3/s)	7.340	0,59
Phân vị 75% (m^3/s)	13.200	0,75
Cực đại (Max, m^3/s)	22.500	1,00

Tại mỗi bước thời gian, các cổng đều lần lượt nhận giá trị đầu vào (đại diện cho một phần tử trong chuỗi đầu vào) và giá trị có được từ đầu ra của tế bào từ bước thời gian trước đó. Forget gate có nhiệm vụ loại bỏ những thông tin không cần thiết, input gate chọn lọc những thông tin cần thiết, và output gate xác định những thông tin nào từ tế bào được sử dụng như đầu ra. Cơ chế hoạt động bên trong của một tế bào LSTM được mô tả chi tiết trong Nguyen Truong Long năm 2018[24].

Xây dựng và đánh giá mô hình

Toàn bộ dữ liệu được chia làm 2 phần: 70% cho tập huấn luyện (training), 30% cho tập kiểm tra (testing). Phương pháp đánh giá độ chính xác (trung bình) của mô hình phân lớp (Cross Validation, CV) được áp dụng để hạn chế hiện tượng mô hình dự đoán quá khớp với dữ liệu huấn luyện (overfitting). CV là phương pháp chia nhỏ tập training ra thành N phần. Với mỗi lần huấn luyện, mô hình sẽ sử dụng N–1 phần cho huấn luyện, sau đó kiểm tra dựa trên 1 phần còn lại, điều này sẽ giúp cho mô hình hạn chế gặp phải



Hình 2: Cấu trúc mạng RNN truyền thống (a) và mạng LSTM (b)

overfittings. Nghiên cứu sử dụng CV = 10, đây là số phổ biến trong huấn luyện thuật toán học máy[25].

Ba chỉ số thống kê là Hệ số xác định (Coefficient of determination, R^2, công thứ 2), Lỗi trung bình bình phương gốc (Root Mean Squared Error, RMSE, công thứ 3), và Sai số tuyệt đối trung bình (Mean Absolute Error, MAE, công thứ 4), được sử dụng để đánh giá tính chính xác của mô hình dự báo. R^2 phản ánh mức độ giải thích của các biến độc lập đối với các biến phụ thuộc, giá trị R^2 càng cao thì mô hình càng chính xác (R^2 dao động từ 0 đến 1)[26]. Cả MAE và RMSE đều thể đo sự khác biệt giữa các giá trị dự đoán và giá trị thực tế, chúng nằm trong khoảng từ 0 đến ∞ và giá trị càng thấp thì mô hình sẽ sẽ tốt hơn[27].

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(yi - \widehat{yi})^2}{\sum_{i=1}^{n}(yi - \bar{y})^2} \quad (2)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widehat{yi} - yi)^2} \quad (3)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\widehat{yi} - yi| \quad (4)$$

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

Nghiên cứu sẽ đánh giá khả năng dự báo lưu lượng nước của thuật toán LSTM với số nơron khác nhau (1, 2, 3, 4) ở các thời gian dự báo khác nhau: 1, 2, 3, 4, 5 giờ tiếp theo (t + 1, t + 2, t + 3, t + 4, t + 5, tương ứng).

KẾT QUẢ

Lựa chọn số nơron cho mạng LSTM

Đối với mạng LSTM có 1 nơron, chỉ số R^2 ở thời gian dự báo t +1 là cao nhất với 0,785. Thời gian dự báo tăng lên t + 2, t + 3, t + 4 thì giá trị R^2 giảm dần, tương ứng là 0,463; 0,324; 0,391. Tuy nhiên, R^2 lại tăng lên 0,453 khi thời gian dự báo là t + 5. Giá trị RMSE và

MAE đạt thấp nhất với thời gian dự báo t + 1 (3619,54 và 2818,47, tương ứng), sau đó tăng dần khi thời gian dự báo tăng từ t + 2, t + 3, từ 5732,48 đến 6435,11 cho RMSE, từ 2818,47 đến 5010,18 cho MAE. Đến thời gian t + 4, t + 5, RMSE giảm từ 6106,65 xuống 5785,13, trong khi MAE giảm từ 4762,33 xuống 4619,48. Mạng LSTM có 2–4 nơron, ở thời gian t + 1, chỉ số R^2 đạt cao nhất, dao động từ 0,912 đến 0,917. Khác với mạng LSTM có 1 nơron, khi R^2 giảm rồi tăng khi tăng thời gian dự báo, mạng LSTM có 2–4 nơron cho thấy chỉ số R^2 giảm liên tục khi tăng thời gian dự báo từ t + 1 lên t + 5. Cụ thể, mạng LSTM có nơron = 2, R^2 giảm từ 0,915 xuống 0,341 từ t + 1 đến t + 5, nơron = 3 có R^2 giảm từ 0,917 xuống 0,361 và khi nơron = 4 có R^2 giảm từ 0,912 xuống 0,299. Mạng 1 nơron, chỉ số RMSE và MAE tăng rồi giảm khi tăng thời gian dự báo, trong khi mạng 2–4 nơron, chỉ số RMSE và MAE tăng dần. Ví dụ khi tăng thời gian dự báo từ t + 1 lên t + 5, RMSE tăng từ 2272,76 đến 6348,02 (2 nơron), từ 2249,64 lên 6251,10 (3 nơron), từ 2318,36 lên 6545,82 (4 nơron) (Bảng 2).

Ở mạng LSTM có 1 nơron, R^2 khá thấp, chỉ từ 0,453 đến 0,785; tuy nhiên khi tăng số nơron (từ 2 đến 4), chỉ số R^2 tăng lên rõ rệt. Cụ thể, R^2 từ 0,341 đến 0,915 (2 nơron), từ 0,361 đến 0,917 (3 nơron), từ 0,299 đến 0,912 (4 nơron). Ngược lại, số nơron trong mạng tăng lên làm giảm đáng kể giá trị RMSE và MAE. Ví dụ, RMSE từ 3619,54–6106,65 (1 nơron) giảm xuống 2272,76–5916,54 (2 nơron). Ở thời gian t + 5, giá trị R^2 ở mạng 1 nơron cao hơn các mạng còn lại có nhiều nơron hơn; hơn nữa, RMSE va MAE ở mạng 1 nơron thường thấp hơn các mạng có nhiều nơron. Ngoài ra, giá trị R^2, RMSE và MAE không khác biệt nhiều ở các mạng 2, 3, 4 nơron (Bảng 2).

Hình 3 so sánh các chỉ số đánh giá tính hiệu quả mô hình LSTM với số lượng các nơron và thời gian dự báo khác nhau. Nhìn chung, khi tăng thời gian dự báo, chỉ số R^2 giảm trong khi RMSE và MAE tăng. Ngoài ra, khi tăng số nơron trong mạng, chỉ số R^2 tăng trong khi RMSE và MAE giảm. Cho nên, tính chính xác của mô hình dự báo tăng khi tăng số nơron trong mạng. Hơn nữa, khi tăng thời gian dự báo, mạng LSTM trong nghiên cứu có tính chính xác kém đi, thời gian dự báo 1 giờ tiếp theo (t + 1) là thời gian dự báo tối ưu.

Áp dụng mô hình mạng LSTM vào dự báo lưu lượng nước ở thời gian t + 1

Hàm loss dùng cho cả 3 mạng LSTM (nơron là 2, 3, 4) là loss = MSE, epochs = 100, batch size = 10, và optimizer = Adam. Đối với mạng nơron = 3, không ghi nhận trình trạng overfittings ở 100 epochs do giá trị loss của tập huấn luyện và kiểm tra giảm dần khi tăng epochs, ngoài ra, chúng gần như nằm trùng lên nhau. Tuy nhiên, có overfittings xuất hiện ở 100 epochs với mạng nơron = 2, 4, mặc dù rất thấp (Hình 4).

Trong quá trình huấn luyện, tương quan giữa giá trị dự báo và giá trị thực tế rất cao ($r > 96\%$, $p < 0,0001$). Cụ thể, hệ số tương quan r là 0,969; 0,968; 0,966 với mạng 2, 3, 4 nơron, tương ứng. Tương tự, trong quá trình kiểm tra, hệ số tương quan r ở mạng 2, 3, 4 nơron cũng rất cao, tương ứng đạt 0,957; 0,956; 0,955 (Hình 5). Điều này cho thấy, giá trị dự báo từ mô hình rất gần với giá trị thực tế, cho nên mô hình với số nơron 2, 3, 4 khá chính xác. Ngoài ra, hệ số tương quan r trong tập huấn luyện và kiểm tra của 3 mạng LSTM không quá khác biệt.

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 LSTM dự báo khá chính xác và đáng tin cậy. Świątek và Okruszko năm 2011 đề xuất thang đánh giá tính chính xác của mô hình dự báo dựa vào giá trị R^2, cụ thể như sau: Mô hình xuất sắc ($0,99 \leq R^2 < 1,00$), rất tốt ($0,95 \leq R^2 < 0,99$), tốt ($0,90 \leq R^2 < 0,95$), khá tốt ($0,85 \leq R^2 < 0,90$), trung bình ($0,80 \leq R^2 < 0,85$), chấp nhận được ($0,70 \leq R^2 < 0,80$), không đáng tin cậy ($R^2 < 0,70$) [26]. Như vậy, mô hình LSTM ứng với số lượng nơron 2, 3, 4 dự báo lưu lượng nước tại trạm thủy văn Mỹ Thuận được đánh giá tốt trong giai đoạn huấn luyện và khá tốt trong giai đoạn kiểm tra. Tuy nhiên, các mô hình này vẫn có xuất hiện overfittings nhưng ở mức thấp do R^2 tập huấn luyện cao hơn tập kiểm tra (Bảng 3).

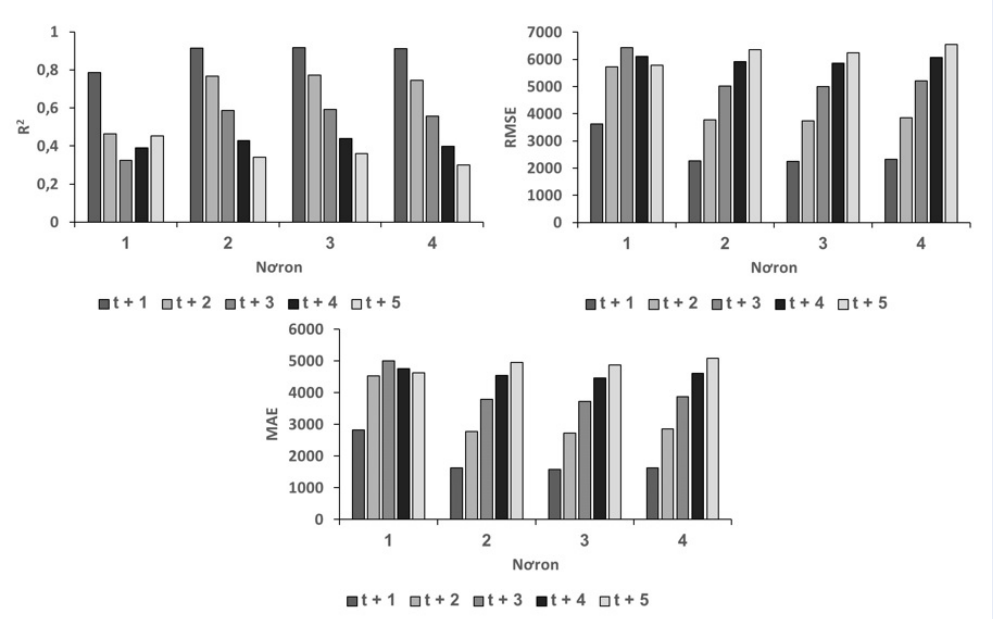
THẢO LUẬN

Mạng LSTM tối ưu cho dự báo lưu lượng trạm thủy văn Mỹ Thuận

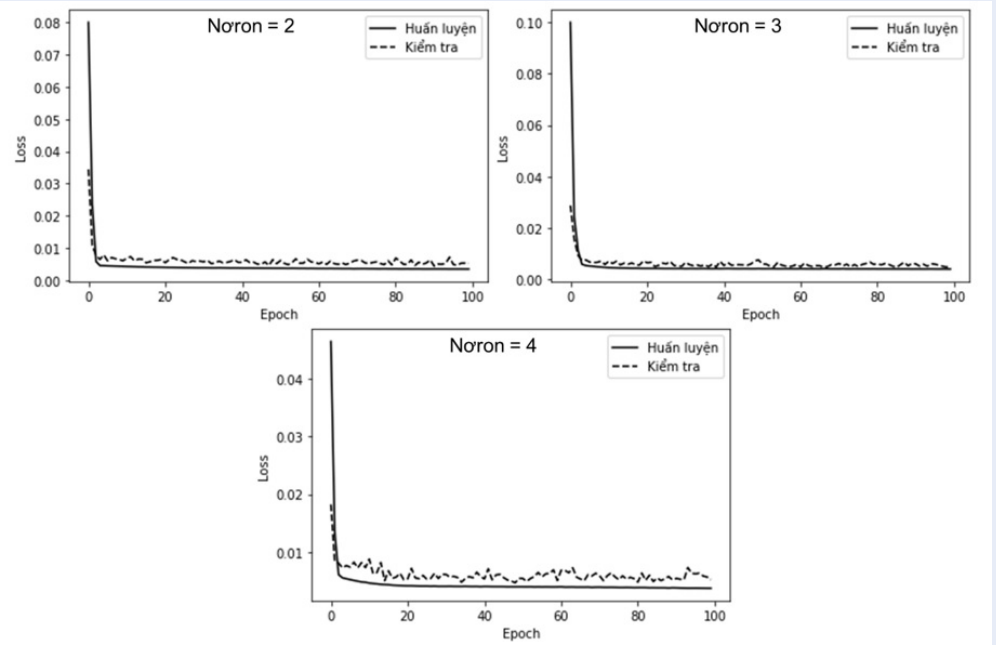
Nhìn chung độ chính xác trong dự báo của mô hình LSTM với số lượng nơron 1 không cao, do mạng không đủ nơron để học và bóc tách quy luật của chuỗi số liệu thời gian, nhất là với các chuỗi dữ liệu lớn. Chuỗi dữ liệu về mực nước của trạm Mỹ Thuận với 8760 giá trị, nhìn chung tương đối lớn, nên mạng LSTM đơn giản với 1 nơron không thể đáp ứng tốt. Tuy nhiên, chỉ cần thêm 1 nơron vào mạng, kết quả đã cải thiện rõ rệt tính chính xác của mô hình dự báo. Độ tin cậy của mô Hình 2, 3 và 4 nơron là khá cao và không có sự khác biệt lớn. Tuy nhiên, mạng LSTM với 3 nơron là thích hợp nhất vì (i) giá trị loss của mạng này ít overfitting nhất (Hình 4), (ii) giá trị dự báo và thực tế gần như trùng nhau (Hình 6), (iii) trong quá trình huấn luyện và kiểm tra, R^2 của mô hình này cao nhất và RMSE, MAE thấp nhất (trừ RMSE huấn luyện) (Bảng 3).

Bảng 2: So sánh các giá trị R^2, RMSE, và MAE ở thời gian dự báo và số lượng nơron khác nhau

Chỉ số	Thời gian dự báo	Số nơron			
		1	2	3	4
R^2	t + 1	0,785	0,915	0,917	0,912
	t + 2	0,463	0,767	0,771	0,745
	t + 3	0,324	0,587	0,593	0,556
	t + 4	0,391	0,428	0,439	0,399
	t + 5	0,453	0,341	0,361	0,299
RMSE	t + 1	3619,54	2272,76	2249,64	2318,36
	t + 2	5732,48	3770,60	3739,86	3848,28
	t + 3	6435,11	5027,57	4991,21	5210,06
	t + 4	6106,65	5916,54	5857,71	6062,59
	t + 5	5785,13	6348,02	6251,10	6545,82
MAE	t + 1	2818,47	1628,23	1585,02	1632,95
	t + 2	4525,23	2775,04	2720,43	2857,96
	t + 3	5010,18	3784,60	3724,07	3869,40
	t + 4	4762,33	4540,97	4469,01	4602,73
	t + 5	4619,48	4957,45	4866,19	5075,89



Hình 3: So sánh các chỉ số đánh giá tính hiệu quả mô hình LSTM với số lượng các nơron và thời gian dự báo khác nhau



Hình 4: Hàm loss cho mô hình LSTM dự báo lưu lượng nước với số lượng nơron khác nhau ở thời gian t + 1

Bảng 3: Hiệu quả dự đoán lưu lượng của mô hình LSTM với số lượng nơron khác nhau ở thời gian t + 1. T: Tốt, KT: Khá tốt

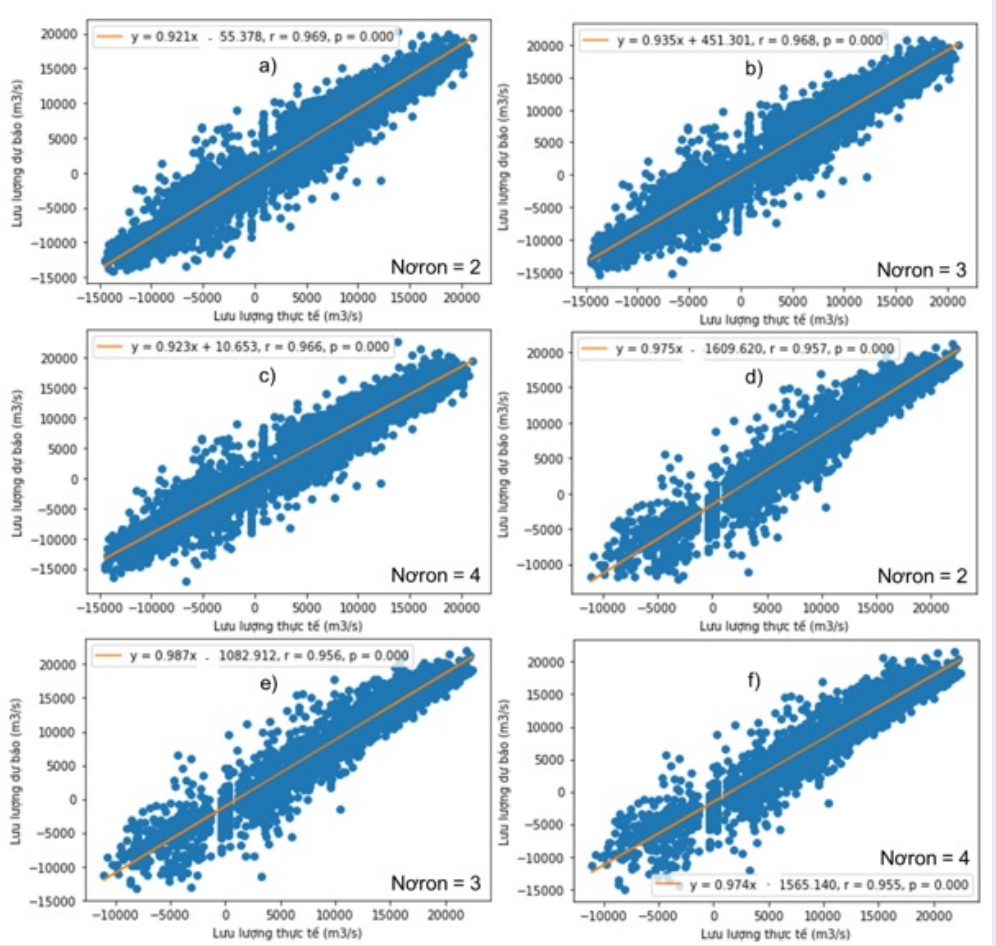
Nơron	Huấn luyện			Kiểm tra		
	R^2	RMSE	MAE	R^2	RMSE	MAE
2	$0,936^{T}$	2287,04	1754,48	$0,856^{KT}$	2964,49	2400,15
3	$0,937^{T}$	2294,60	1738,33	$0,884^{KT}$	2655,66	2064,30
4	$0,932^{T}$	2359,79	1828,71	$0,853^{KT}$	2999,59	2423,90

Độ chính xác của mô hình dự báo sẽ giảm khi tăng thời gian dự báo và ở thời gian t + 1 sẽ có kết quả dự báo chính xác nhất, điều này đã ghi nhận trong các nghiên cứu của Le và cộng sự (dự báo lưu lượng sông Hồng)[1], Sudriani và cộng sự (dự báo lưu lượng sông Cimandiri, Indonesia)[28], tác giả Phan và Nguyen (dự báo mực nước sông Hồng)[29], Ren và cộng sự[30]. Trong nghiên cứu, với mạng 2, 3, 4 nơron ở thời gian dự báo t + 2 có R^2 tương ứng là 0,767, 0,771, 0,745. Tính chính xác của mô hình với giá trị R^2 này vẫn ở mức chấp nhận được theo Świątek và Okruszko năm 2011[26]. Tuy nhiên, ở thời gian dự báo t + 1, R^2 ở mạng 2, 3, 4 nơron rất cao, từ 0,932 đến 0,937 ở tập huấn luyện, từ 0,853 đến 0,884 ở tập kiểm tra. Nhìn chung, nghiên cứu xác định số nơron = 3, thời gian dự báo t + 1 là tối ưu với bộ dữ liệu lưu lượng theo giờ ở trạm Mỹ Thuận năm 2018.

Mạng LSTM: Đơn giản và hiệu quả

Rõ ràng, mạng LSTM trong nghiên cứu có hiệu quả và tính chính xác tương đối cao (đã được thảo luận ở trên). Điều đáng ghi nhận ở đây là cấu trúc mạng LSTM trong nghiên cứu rất đơn giản với [1 lớp LSTM có 2–3 nơron và tangent activation + 1 lớp Dense với 1 nơron và tangent activation + Compile (MSE loss, Adam optimizer với learning rate 0.001)]. Mạng LSTM này được thi hành với chỉ máy tính cấu hình phổ thông (Intel(R) Core(TM) i5–4310U CPU 2.60 GHz, 4GB RAM) nhưng thời gian thực thi rất nhanh, tương ứng là 146,674 giây, 158,450 giây, và 160,959 giây với mạng LSTM 2, 3, 4 nơron.

Thực tế cho thấy các đặc điểm thủy văn hệ thống sông Mê Kông (trong đó có lưu lượng nước) là vô cùng phức tạp do chịu ảnh hưởng từ nhiều yếu tố[31,32]. Cho nên đặc điểm thủy văn sông Mê Kông đã được phân tích và dự báo bằng nhiều phương pháp hiện đại như MIKE[33–35], viễn thám[36,37]. Ưu điểm của

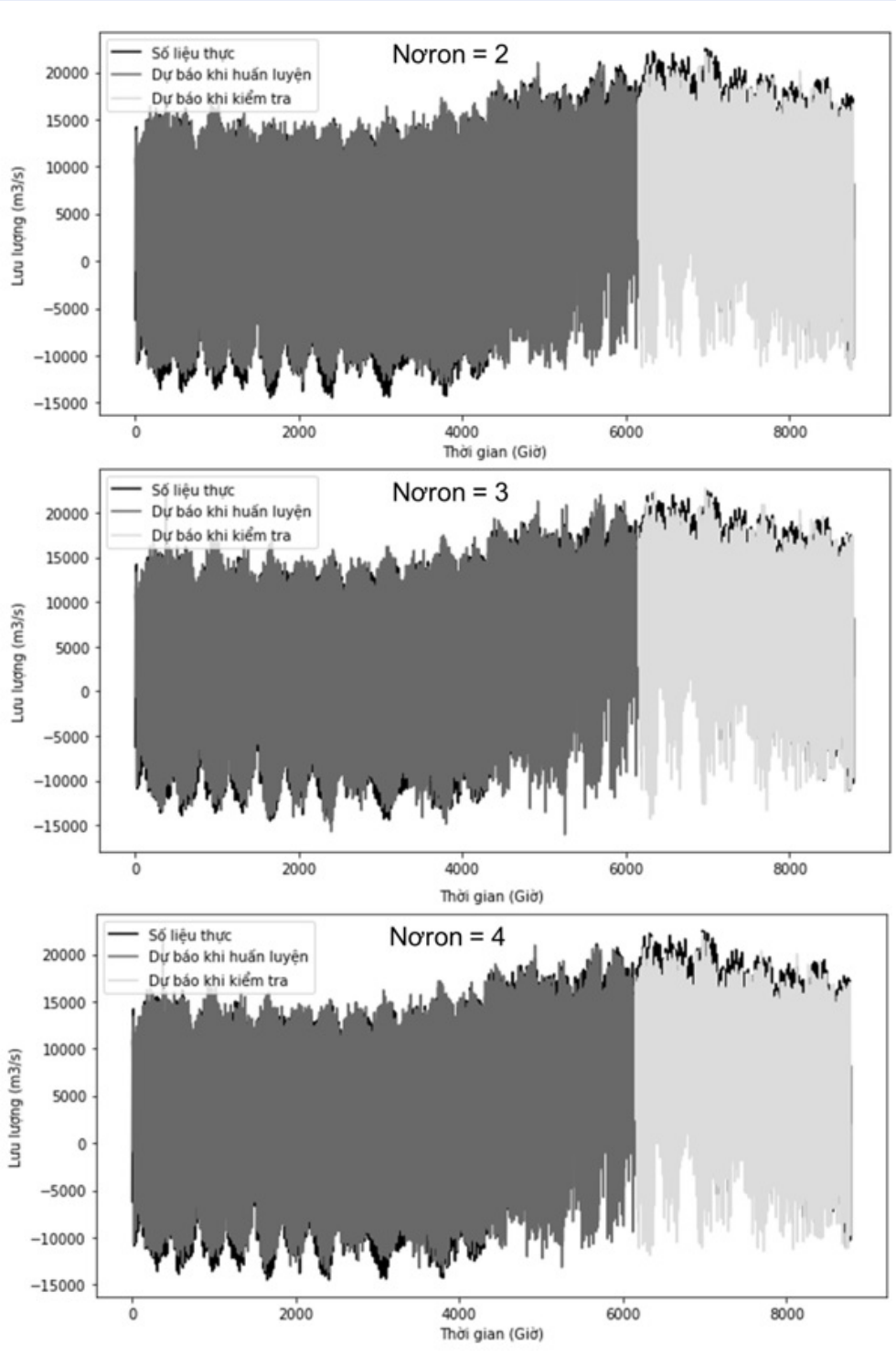


Hình 5: Quan hệ tuyến tính giữa lưu lượng thực tế và dự báo bằng mô hình LSTM với số lượng nơron khác nhau ở thời gian t + 1. Giai đoạn huấn luyên (a–c), giai đoạn kiểm tra (e–f)

các phương pháp này là tính chính xác rất cao và các quy luật thủy văn (ví dụ lưu lượng nước) có thể được giải thích thông qua các phương trình toán–vật lý; tuy nhiên chúng thường vận hành rất phức tạp và yêu cầu số lượng dữ liệu đầu vào rất lớn và đầy đủ mới đảm bảo tính chính xác[4,5]. Thu thập thông tin về toàn bộ các yếu tố ảnh hưởng đến độ lưu lượng là vô cùng khó khăn và thường không đầy đủ[4]. Cho nên, lựa chọn các mô hình học máy để dự báo lưu lượng trong trường hợp này là phù hợp. Phương pháp này có ưu điểm là dễ áp dụng, độ chính xác cao, không đòi hỏi số lượng dữ liệu lớn[4,6,7].

Mạng LSTM đã được áp dụng rộng rãi trên thế với và tiềm năng ứng dụng ở Việt Nam

Các mô hình dự báo có thể chia làm 2 nhóm: định tính (qualitative models), định lượng (quantitative models). Mô hình định tính là dựa vào kinh nghiệm và quyết định của các chuyên gia còn mô hình định lượng là dựa vào các phương trình toán học - vật lý. Mô hình định lượng tiếp tục chia làm 2 nhóm nhỏ: Mô hình nguyên nhân - kết quả (causal models) và mô hình chuỗi thời gian (time series models). Mô hình nguyên nhân - kết quả mô tả mối quan hệ giữa biến phụ thuộc (biến dự báo) và biến độc lập (các yếu tố tác động). Mô hình chuỗi thời gian dựa vào sự tương tác, quan hệ giữa các số liệu trong quá khứ để làm cơ sở dự báo cho tương lai[38]. Do có khả năng thực hiện cả 2 kiểu dự báo (nguyên nhân - kết quả và chuỗi thời gian) nên mạng LSTM được ứng dụng rộng rãi trong các nghiên cứu mô phỏng, nhất là dự báo các đặc điểm thủy văn. Nói về dự báo chuỗi thời gian, khi cung cấp số liệu đủ tốt và đủ lớn, mạng LSTM đã cho thấy khả năng dự báo rất chính xác (Bảng 4). Mạng LSTM cần số lượng đủ lớn dữ liệu để huấn luyện các nơron, khi cung cấp ít số liệu, quá trình huấn luyện



Hình 6: So sánh giữa lưu lượng thực tế và dự báo bằng mô hình LSTM với số lượng nơron khác nhau ở thời gian t + 1

Bảng 4: Ứng dụng mạng LSTM vào dự báo lưu lượng nước trên thế giới

Địa điểm	Dữ liệu	Hiệu quả	Tham khảo
Lưu lượng			
Mỹ Thuận, sông Tiền, Việt Nam	8760 giá trị đo theo giờ trong năm 2018	R^2 huấn luyện: 0,937, R^2 kiểm tra: 0,884, RMSE huấn luyện: 2294,60 m^3/s, RMSE kiểm tra: 2655,66 m^3/s (mạng 3 nơron, t + 1)	Nghiên cứu này
Sơn Tây, sông Hồng, Việt Nam	5475 giá trị đo theo ngày trong 15 năm: 1972–1978; 1997–2004	NSE: 0,992–0,993 (t + 1)	1
Sông Cimandiri, Indonesia	Trạm Leuwilisung: 6205 giá trị theo ngày trong 17 năm. Trạm Tegaldatar: 4745 giá trị theo ngày trong 13 năm	RMSE Leuwilisung: 4182 m^3/s, RMSE Tegaldatar: 2625 $m^3/2$ (t + 1)	28
Thí nghiệm mô phỏng	4464 giá trị	R^2 huấn luyện: 0,991–0,992 R^2 kiểm tra: 0,978–0,986	35
Mực nước			
69 hồ ở Phần Lan	42 hồ: 12775 giá trị theo ngày trong 35 năm (1984–2018). 27 hồ: 11315 giá trị trong 31 năm (1984–2014)	R^2: 0,84–0,97, RMSE: 0,02–0,07 m	13
Vịnh Narragansett, Mỹ	26280 giá trị theo giờ trong 3 năm (2013–2015)	R^2: 0,99, RMSE: 0,05 m, MAE: 0,03 m	14
Hồ Dongting, Trung Quốc	3650 giá trị theo ngày trong 11 năm (2003–2013)	R^2: 0,99, RMSE: 0,85–0,99 m	15

có thể không hiệu quả [39,40]. Ngoài ra, chất lượng của số liệu cũng ảnh hưởng rất lớn đến tính chính xác của mô hình dự báo chuỗi thời gian nói chung và LSTM nói riêng [16], cho nên số liệu cần phải xem xét và tiền xử lý kỹ càng trước khi đưa vào huấn luyện mô hình. Khi đáp ứng các yêu cầu trên, LSTM cho thấy rất tiềm năng trong ứng dụng để phân tích, dự báo, và mô phỏng các quy luật tự nhiên, nhất là các đặc điểm thủy văn học. Trong thời gian sắp tới, mô hình LSTM cần được quan tâm nhiều hơn vào các dự báo cho thiên tai và các biến động trong môi trường bởi vì tính chính xác và hiệu quả của mô hình mang lại là rất tiềm năng.

KẾT LUẬN

Nghiên cứu đánh giá khả năng của mô hình LSTM trong dự báo lưu lượng nước ở trạm thủy văn Mỹ Thuận trên sông Tiền, tính chính xác của mô hình dựa vào các chỉ số như R^2, RMSE, MAE. Từ dữ liệu hiện tại và những kết quả đạt được, nghiên cứu đi đến một số kết luận như sau: Mô hình LSTM dự báo lưu lượng nước trạm Mỹ Thuận chính xác nhất ở thời gian dự báo t + 1, (ii) mặc dù không có khác biệt lớn ở mạng có 2, 3, 4 nơron nhưng nghiên cứu nhận thấy mô hình tối ưu là 3 nơron và dự báo cho t + 1, (iii) kết quả dự báo rất khả quan cho thấy tiềm năng ứng dụng mô hình LSTM vào các nghiên cứu dự báo, mô phỏng ở Việt Nam, nhất là các quá trình thủy văn và biến động môi trường.

LỜI CẢM ƠN

Nghiên cứu được tài trợ bởi Đại học Thủ Dầu Một trong đề tài mã số "DT.21.2–036".

DANH MỤC TỪ VIẾT TẮT

CV: Cross Validation
IQR: Interquartile Range
LSTM: Long Short–Term Memory
MAE: Mean Absolute Error
MSE: Mean Squared Error
NSE: Nash–Sutcliffe Efficiency
R^2: Coefficient of determination
RMSE: Root Mean Squared Error
RNN: Recurrent Neural Network

XUNG ĐỘT LỢI ÍCH

Các tác giả cam đoan rằng họ không có xung đột lợi ích.

ĐÓNG GÓP TÁC GIẢ

Nghiên cứu này được thiết kế bởi Trần Thành Thái và Phạm Ngọc Hoài. Trần Thành Thái phân tích số liệu và hoàn thiện bản thảo. Phạm Ngọc Hoài, Phạm Bảo Quốc, Nguyễn Phương Đông, Nguyễn Duy Liêm, và Khuất Thùy Phương hỗ trợ phân tích thuật toán và cùng tác giả chính viết phần phương pháp, kết quả–thảo luận. Nguyễn Phan Nhân tổng hợp tài liệu để viết phần Tổng quan. Tất cả tác giả tham gia thảo luận, góp ý và chỉnh sửa hoàn thiện bản thảo.

TÀI LIỆU THAM KHẢO

1. Le XH, Ho HV, Lee G, Jung S. Application of long short-term memory (LSTM) neural network for flood forecasting. Water. 2019;11(7):1387;Available from: https://doi.org/10.3390/w11071387.
2. Mekong River Commission (MRC). Assessment of basin-wide development scenarios, Basin Development Plan Programme, Phase 2. Mekong River Commission (MRC), Vientiane, Lao PDR. 2011;Available from: http://www.mrcmekong.org/assets/Other-Documents/BDP/Assessment-of-Basin-wide-dev-Scenarios-MainReport-110420.pdf.
3. Räsänen TA, Koponen J, Lauri H, Kummu M. Downstream hydrological impacts of hydropower development in the Upper Mekong Basin. Water Resources Management. 2021;26(12):3495-3513;Available from: https://doi.org/10.1007/s11269-012-0087-0.
4. Ross AC, Stock CA. An assessment of the predictability of column minimum dissolved oxygen concentrations in Chesapeake Bay using a machine learning model. Estuarine, Coastal and Shelf Science. 2019;221:53-65;Available from: https://doi.org/10.1016/j.ecss.2019.03.007.
5. Lin K, Lu P, Xu CY, Yu X, Lan T, Chen X. Modeling saltwater intrusion using an integrated Bayesian model averaging method in the Pearl River Delta. Journal of Hydroinformatics. 2019; 21(6):1147-1162;Available from: https://doi.org/10.2166/hydro.2019.073.
6. Palani S, Liong SY, Tkalich P. An ANN application for water quality forecasting. Marine Pollution Bulletin. 2008;56(9):1586-1597;Available from: https://doi.org/10.1016/j.marpolbul.2008.05.021.
7. Hunter JM, Maier HR, Gibbs MS, Foale ER, Grosvenor NA, Harders NP, Kikuchi-Miller T C. Framework for developing hybrid process-driven, artificial neural network and regression models for salinity prediction in river systems. Hydrology and Earth System Sciences. 2018;22(5):2987-3006;Available from: https://doi.org/10.5194/hess-22-2987-2018.
8. Graves A, Mohamed A, Hinton G. Speech recognition with deep recurrent neural networks. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. 2013, Vancouver, Canada, pp. 6645-6649;.
9. Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. In Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 2. 2014, Montreal, Canada, pp. 3104-3112;.
10. Li Y, Cao H. Prediction for tourism flow based on LSTM neural network. Procedia Computer Science. 2018;129:277-283;Available from: https://doi.org/10.1016/j.procs.2018.03.076.
11. Nelson DMQ, Pereira ACM, de Oliveira RA. Stock market's price movement prediction with LSTM neural networks. In Proceedings of the International Joint Conference on Neural Networks (IJCNN). 2017, Anchorage, USA, pp. 1419-1426;.
12. Hu C, Wu Q, Li H, Jian S, Li N, Lou Z. Deep learning with a Long Short-Term Memory networks approach for rainfall-runoff simulation. Water. 2018;10:1543;Available from: https://doi.org/10.3390/w10111543.
13. Zhu S, Hrnjica B, Ptak M, Choiński A, Sivakumar B. Forecasting of water level in multiple temperate lakes using machine learning models. Journal of Hydrology. 2020;585:124819;Available from: https://doi.org/10.1016/j.jhydrol.2020.124819.
14. Tu Z, Gao X, Xu J, Sun W, Sun Y, Su D. A novel method for regional short-term forecasting of water level. Water. 2021;13(6):820;Available from: https://doi.org/10.3390/w13060820.
15. Liang C, Li H, Lei M, Du Q. Dongting lake water level forecast and its relationship with the three gorges dam based on a Long Short-Term Memory network. Water. 2018;10(10):1389;Available from: https://doi.org/10.3390/w10101389.
16. Liu H, Sun GX, Cao RX. The application of GM (1, 1) dynamic model in the forecast of groundwater level in Wujiang city. Journal of Geological Hazards and Environment Preservation. 2008;19(3):47-51;.
17. Pan M, Zhou H, Cao J, Liu Y, Hao J, Li S, Chen CH. Water level prediction model based on GRU and CNN. IEEE Access. 2020;8:60090-60100;Available from: https://doi.org/10.1109/ACCESS.2020.2982433.
18. Frigge M, Hoaglin DC, Iglewicz B. Some implementations of the boxplot. The American Statistician. 1989;43(1):50-54. https://doi.org/10.2307/2685173;Available from: https://doi.org/10.2307/2685173.
19. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J. Scikit-learn: Machine learning in Python. The Journal of Machine Learning Research. 2011;12:2825-2830;.
20. Barzegar R, Aalami MT, Adamowski J. Short-term water quality variable prediction using a hybrid CNN-LSTM deep learning model. Stochastic Environmental Research and Risk Assessment. 2020;34:415-433;Available from: https://doi.org/10.1007/s00477-020-01776-2.
21. Yang CH, Wu CH, Hsieh CM. Long short-term memory recurrent neural network for tidal level forecasting. IEEE Access. 2020;8:159389-159401;Available from: https://doi.org/10.1109/ACCESS.2020.3017089.
22. Hochreiter S. The vanishing gradient problem during learning recurrent neural nets and problem solutions. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems. 1998;6:107-116;Available from: https://doi.org/10.1142/S0218488598000094.
23. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation. 1997;9(8): 1735-1780;Available from: https://doi.org/10.1162/neco.1997.9.8.1735.
24. Long NT. Giai thích chi tiết về mạng Long Short Term Memory (LSTM). 2018;Available from: https://nguyentruonglong.net/giai-thich-chi-tiet-ve-mang-long-short-term-memory-lstm.htmlon2021-09-09.
25. Nguyen HQ, Ha NT, Pham TL. Inland harmful cyanobacterial bloom prediction in the eutrophic Tri An Reservoir using satellite band ratio and machine learning approaches. Environmental Science and Pollution Research. 2020;27(9):9135-9151;Available from: https://doi.org/10.1007/s11356-019-07519-3.
26. Świątek D, Okruszko T. Modelling of Hydrological Processes in the Narew Catchment. 2011, Springer Science & Business Media;Available from: https://doi.org/10.1007/978-3-642-19059-9.
27. Zhu S, Ptak M, Yaseen ZM, Dai J, Sivakumar B. Forecasting surface water temperature in lakes: A comparison of approaches. Journal of Hydrology. 2020;585:124809;Available from: https://doi.org/10.1016/j.jhydrol.2020.124809.
28. Sudriani Y, Ridwansyah I, Rustini HA. Long short term memory (LSTM) recurrent neural network (RNN) for discharge level

prediction and forecast in Cimandiri river, Indonesia. IOP Conference Series: Earth and Environmental Science, IOP Publishing. 2019;299(1): 012037;.
29. Phan TTH, Nguyen XH. Combining statistical machine learning models with ARIMA for water level forecasting: The case of the Red river. Advances in Water Resources. 2020;142: 103656;Available from: https://doi.org/10.1016/j.advwatres.2020.103656.
30. Ren T, Liu X, Niu J, Lei X, Zhang Z. Real-time water level prediction of cascaded channels based on multilayer perception and recurrent neural network. Journal of Hydrology. 2020;585: 124783;Available from: https://doi.org/10.1016/j.jhydrol.2020.124783.
31. Ngoc TA, Letrung T, Hiramatsu K, Nguyen TQ. The effect of simulated sea level on the sedimentation of the Tien River Estuaries, Lower Mekong River, Southern Vietnam. Japan Agricultural Research Quarterly. 2013;47(4):405-415;Available from: https://doi.org/10.6090/jarq.47.405.
32. Dang VH, Tran DD, Pham TBT, Khoi DN, Tran PH, Nguyen NT. Exploring freshwater regimes and impact factors in the coastal estuaries of the Vietnamese Mekong Delta. Water. 2019;11(4):782;Available from: https://doi.org/10.3390/w11040782.
33. Tran Anh D, Hoang LP, Bui MD, Rutschmann P. Simulating future flows and salinity intrusion using combined one-and two-dimensional hydrodynamic modelling-the case of Hau River, Vietnamese Mekong delta. Water. 2018;10(7):897;Available from: https://doi.org/10.3390/w10070897.
34. Khang DN, Kotera A, Sakamoto T, Yokozawa M. Sensitivity of Salinity Intrusion to Sea Level Rise and River Flow Change in Vietnamese Mekong Delta-Impacts on Availability of Irrigation Water for Rice Cropping. Journal of Agricultural and Meteorological. 2008;64:167-176;Available from: https://doi.org/10.2480/agrmet.64.3.4.
35. Kang H, Yang S, Huang J, Oh J. Time series prediction of wastewater flow rate by bidirectional LSTM deep learning. International Journal of Control, Automation and Systems. 2020;18(12):3023-3030;Available from: https://doi.org/10.1007/s12555-019-0984-6.
36. Fok HS, He Q, Chun KP, Zhou Z, Chu T. Application of ENSO and drought indices for water level reconstruction and prediction: A case study in the lower Mekong River estuary. Water. 2018;10(1):58;Available from: https://doi.org/10.3390/w10010058.
37. He Q, Fok HS, Chen Q, Chun KP. Water level reconstruction and prediction based on space-borne sensors: A case study in the Mekong and Yangtze river basins. Sensors. 2018;18(9):3076;Available from: https://doi.org/10.3390/s18093076.
38. Alsharif MH, Younes MK, Kim J. Time series ARIMA model for prediction of daily and monthly average global solar radiation: The case study of Seoul, South Korea. Symmetry. 2019;11:240;Available from: https://doi.org/10.3390/sym11020240.
39. Lago J, De Ridder F, De Schutter B. Forecasting spot electricity prices: Deep learning approaches and empirical comparison of traditional algorithms. Applied Energy. 2015;221:386-405;Available from: https://doi.org/10.1016/j.apenergy.2018.02.069.
40. Schmidt A, Mainwaring DB, Maguire DA. Development of a tailored combination of machine learning approaches to model volumetric soil water content within a mesic forest in the Pacific Northwest. Journal of Hydrology. 2020;588:125044;Available from: https://doi.org/10.1016/j.jhydrol.2020.125044.

Science & Technology Development Journal – Natural Sciences, 6(1):1884-1896

Open Access Full Text Article

Research Article

Application of Long Short–Term Memory neural network for time series prediction of flow rate at My Thuan hydrology station, Tien river

Tran Thanh Thai[1,*], Pham Ngoc Hoai[2], Pham Bao Quoc[2], Nguyen Phan Nhan[3], Nguyen Phuong Dong[4], Nguyen Duy Liem[5], Khuat Thuy Phuong[6]

ABSTRACT

Flow rate prediction has an important role in water resource management to reduce potential damage caused by floods for urban residential areas. However, prediction of flow rate presents great challenges because the task requires a number of information, such as hydrological, geomorphological data. The objective of this paper is to apply an effective approach for flow rate forecasting at My Thuan hydrology station (Tien River), based on the construction of a Long Short–Term Memory (LSTM) neural network model using flow rate monitoring data These data composed of 8760 hourly flow rate data points within 2018. Coefficient of determination (R^2), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) are used to evaluate performances of LSTM model. The study evaluates the ability of LSTM algorithm to predict water flow with different number of neurons (1, 2, 3, 4) at different forecasting time: 1, 2, 3, 4, 5 hours ahead (t + 1, t + 2, t + 3, t + 4, t + 5, respectively). The research results indicated that the LSTM model with 3 neurons achieved a high performance for flow rate forecasting. When forecasting one hour ahead (t + 1), R^2, RMSE, MAE reached 0.937, 2294.60, and 1738.33, respectively for training period, and was 0.884, 2655.66, and 2064.30, respectively for testing period. The findings of this study suggest that the LSTM model has promised as a potential tool in flow rate forecasting at the My Thuan and for other hydrology stations in Vietnam.

Key words: Climate change, deep learning, machine learning, Mekong Delta, time series

[1]*Institute of Tropical Biology, Vietnam Academy of Science and Technology. Ho Chi Minh City, Vietnam*

[2]*Institute of Applied Technology, Thu Dau Mot University. Binh Duong Province, Viet Nam*

[3]*Department of Natural Resources and Environment. Ben Tre Province, Vietnam*

[4]*Sub–Institute of Hydro Meteorology and Climate Change. Ho Chi Minh City, Vietnam*

[5]*Nong Lam University. Ho Chi Minh City, Vietnam*

[6]*Computer Science Center, University of Science–Vietnam National University Ho Chi Minh City. Ho Chi Minh City, Vietnam*

Correspondence

Tran Thanh Thai, Institute of Tropical Biology, Vietnam Academy of Science and Technology. Ho Chi Minh City, Vietnam

Email: thanhthai.bentrect@gmai.com

Cite this article : Thai T T, Hoai P N, Quoc P B, Nhan N P, Dong N P, Liem N D, Phuong K T. **Application of Long Short–Term Memory neural network for time series prediction of flow rate at My Thuan hydrology station, Tien river**. *Sci. Tech. Dev. J. - Nat. Sci.;* 6(1):1884-1896.