TẠP CHÍ
KHÍ TƯỢNG THỦY VĂN

Bài báo khoa học

Dự báo chất lượng không khí bằng mô hình LSTM-MA trường hợp sử dụng dữ liệu tại trạm quan trắc tự động ngã tư Giếng Nước, tỉnh Bà Rịa - Vũng Tàu

Hồ Minh Dũng[1]*, Khổng Doãn An Khang[1]

[1] Viện Môi trường và Tài nguyên, ĐHQG-HCM; H_minhdung@yahoo.com; ankhang28040506@gmail.com

*Tác giả liên hệ: H_minhdung@yahoo.com; Tel.: +84–903605245

Ban Biên tập nhận bài: 20/3/2024; Ngày phản biện xong: 2/5/2024; Ngày đăng bài: 25/9/2024

Tóm tắt: Ô nhiễm không khí là một trong những nguyên nhân làm tăng nguy cơ mắc các bệnh về hô hấp và tim mạch. Dự báo diễn biến chất lượng không khí giúp cảnh báo cho cộng đồng về mức độ ô nhiễm. Nghiên cứu này ứng dụng trí thông minh nhân tạo cho dự báo chất lượng không khí tại khu vực trạm quan trắc tự động Ngã tư Giếng Nước, tỉnh Bà Rịa - Vũng Tàu. Mô hình bộ nhớ dài ngắn (LSTM) được lựa chọn cho nghiên cứu và để tối ưu khả năng dự báo, bộ lọc trung bình trượt (MA) được sử dụng. Kết quả nghiên cứu cho thấy chất lượng không khí khu vực nghiên cứu tương đối tốt khi nồng độ của CO, NO_2, SO_2, PM_{10} và $PM_{2.5}$ đều dưới ngưỡng cho phép. Ozon là thông số có số lần vượt ngưỡng cho phép cao nhất, cũng là thông số có tác động chính đến chỉ số chất lượng không khí; Mô hình LSTM–MA đã được xây dựng thành công với khả năng dự báo có độ chính xác cao nhất cho thời gian 1 ngày tiếp theo với giá trị căn bậc 2 sai số bình phương trung bình (RMSE) là 3,05; sai số trung bình tuyệt đối (MAE) là 2,17 và sai số phần trăm tuyệt đối trung bình (MAPE) là 3,19%. Khi dự báo trong thời gian dài hơn với 2 tuần tiếp theo, mô hình cho kết quả khả quan khi các chỉ số RMSE, MAE và MAPE lần lượt đạt 22,79; 15,74 và 24,38%.

Từ khóa: LSTM-MA; Bà Rịa - Vũng Tàu; Dự báo; Chất lượng không khí.

1. Mở đầu

Hiện nay, do quá trình phát triển kinh tế - xã hội, nên tỉnh Bà Rịa - Vũng Tàu đang phải đối mặt với nhiều vấn đề về môi trường, đã xuất hiện một số khu vực có dấu hiệu ô nhiễm môi trường không khí, ảnh hưởng trực tiếp đến sức khỏe người dân. Bên cạnh công tác quan trắc chất lượng không khí thì hoạt động dự báo cũng rất cần thiết. Sử dụng trí tuệ nhân tạo (AI) trong dự báo chất lượng không khí xung quanh là một trong những hướng nghiên cứu mới để giải quyết vấn đề này. Các mô hình dự báo chất lượng không khí hiện nay có thể được phân thành 2 loại là mô hình giải tích và mô hình thống kê [1].

Các mô hình giải tích dựa trên các quá trình vật lý và hóa học trong khí quyển, kết hợp với yếu tố khí tượng và công cụ toán học để mô phỏng chất lượng không khí ở nhiều qui mô khác nhau, có thể kể đến như CMAQ (*Community Multiscale Air Quality*) [2–4], WRF–Chem (*Weather Research and Forecasting Model with Chemistry*) [5,6], AERMOD (*AMS/EPA Regulatory Model*) [7–10] và TAPOM (*Transport and Air Pollution Model*) [11,12],... Các mô hình thống kê lại ít chú ý đến các cơ chế vật lý và hóa học của các chất ô nhiễm mà tập trung chủ yếu vào sự tương quan của dữ liệu đầu vào như các biến về khí tượng và dữ liệu các chất ô nhiễm trong khoảng thời gian trước đó với nồng độ chất ô nhiễm trong tương lai [13]. Các mô hình thống kê để dự báo chất lượng không khí bao gồm: ARMA (*Auto Regression Moving Average*), ARIMA (*Auto Regression Integrated Moving Average*) [14],

GWR (*Geographically Weighted Regression*), MLR (*Multiple Linear Regression*) [15], SVR (*Support Vector Regression*) [16]. Các mô hình này đều là mô hình hồi quy với các hàm mô tả mối quan hệ giữa một hoặc nhiều biến độc lập, biến phản hồi, biến phụ thuộc hoặc biến mục tiêu.

Trên thế giới, các nghiên cứu ứng dụng AI cho việc dự báo chất lượng không khí đã được tiến hành rộng rãi. Các kết quả thu thập được đều cho thấy kết quả khả quan khi kết hợp một hoặc nhiều các mô hình lại với nhau. Mô hình LSTM có khả năng thể hiện tốt các phân bố về diễn biến chất lượng không khí theo thời gian nên thường được các tác giả kết hợp với mô hình diễn biến theo không gian như CNN. Kết quả từ các nghiên cứu đều cho thấy các mô hình đơn LSTM hoặc lai của LSTM đều cho hiệu quả tốt hơn so với các mô hình dự báo truyền thống.

Bảng 1. Tổng hợp một số nghiên cứu ứng dụng mô hình LSTM trên thế giới và Việt Nam.

Nghiên cứu	Chất ô nhiễm	Quốc gia	Khoảng thời gian của dữ liệu	Mô hình sử dụng
Yan và cs [17]	AQI	Trung Quốc	2015 - 2016	BPNN, CNN, LSTM CNN-LSTM
Jiao và cs [18]	AQI	Trung Quốc	10/2023 - 9/2018	LSTM
Belavadi và cs [19]	AQI	Ấn Độ	9/3/2019 - 13/4/2019	LSTM
Duan và cs [20]	AQI	Trung Quốc	01/2015 - 03/2022	LSTM, CNN-LSTM DBO-LSTM, CEEMDAN-LSTM
Yammahi và cs [21]	NO_2	UAE	2019 - 2020	LSTM, NAR-NN ARIMA, SARIMA
Navares và cs [22]	CO, NO_2, O_3, PM_{10}, và SO_2	Tây Ban Nha	2001 - 2013	LSTM-RNN
Chang và cs [23]	$PM_{2.5}$	Đài Loan	2013 - 2017	Aggregated -LSTM
Wen và cs [24]	$PM_{2.5}$	Trung Quốc	2016 - 2017	CNN-LSTM
Jung và cs [25]	PM_{10}	Hàn Quốc	2009 - 2019	DNN, RNN LSTM
Rakholia và cs [26]	$PM_{2.5}$	Việt Nam	02/2021 - 12/2021	XGBoost, GDRegressor 1D CNN-LSTM, Prophet
Hưng [27]	CO, NO_x, O_3, $PM_{2.5}$, PM_{10}, SO_2	Việt Nam	2010 - 2018	CNN-LSTM

Hiện nay tại Việt Nam, ứng dụng AI với các mô hình học sâu chủ yếu chỉ mới áp dụng nhiều trong lĩnh vực tài nguyên nước [28, 29]. Đối với dự báo chất lượng không khí, các nghiên cứu sử dụng mô hình LSTM còn hạn chế về số lượng. Ngoài ra, trong các nghiên cứu này còn một số hạn chế: Tập dữ liệu sử dụng bị hạn chế về địa điểm cũng như số lượng do các trạm quan trắc không khí tự động, liên tục chỉ mới được lắp đặt trong thời gian gần đây; Bộ dữ liệu đầu vào cho mô hình LSTM có yêu cầu cao nên cần phải có các giải pháp tiền xử lý số liệu phù hợp để cho ra dự báo tối ưu; Dự báo ô nhiễm không khí theo AI thì chỉ dựa vào diễn biến số liệu quan trắc thu thập trong quá khứ nên khó có thể kết hợp với các kịch bản phát triển kinh tế xã hội. Điều này sẽ dẫn đến việc khó xác định biến đầu vào cũng như việc thu thập các số liệu có liên quan. Vì vậy, để góp phần vào nghiên cứu ứng dụng AI trong dự báo chất lượng không khí, nghiên cứu này được thực hiện với mục tiêu sử dụng mô hình LSTM kết hợp MA để dự báo chất lượng không khí dựa trên dữ liệu quan trắc tại trạm quan trắc tự động Ngã tư Giếng Nước, TP. Vũng Tàu, tỉnh Bà Rịa - Vũng Tàu.

2. Dữ liệu và Phương pháp nghiên cứu

2.1. Dữ liệu nghiên cứu

Bộ dữ liệu quan trắc chất lượng không khí sử dụng trong nghiên cứu này là nồng độ trung bình giờ của các thông số $PM_{2.5}$, PM_{10}, CO, NO_2, O_3 và SO_2, được đo tại trạm quan

trắc tự động (QTTĐ) Ngã tư Giếng Nước (đặt tại phường 7, thành phố Vũng Tàu, tọa độ X: 426939; Y:146298), thuộc quyền quản lý của Sở Tài nguyên và Môi trường tỉnh Bà Rịa – Vũng Tàu. Thời gian của bộ dữ liệu kéo dài từ 18/1/2020 đến 31/12/2022.

Hình 1. Vị trí trạm quan trắc tự động Ngã tư Giếng nước, TP Vũng Tàu.

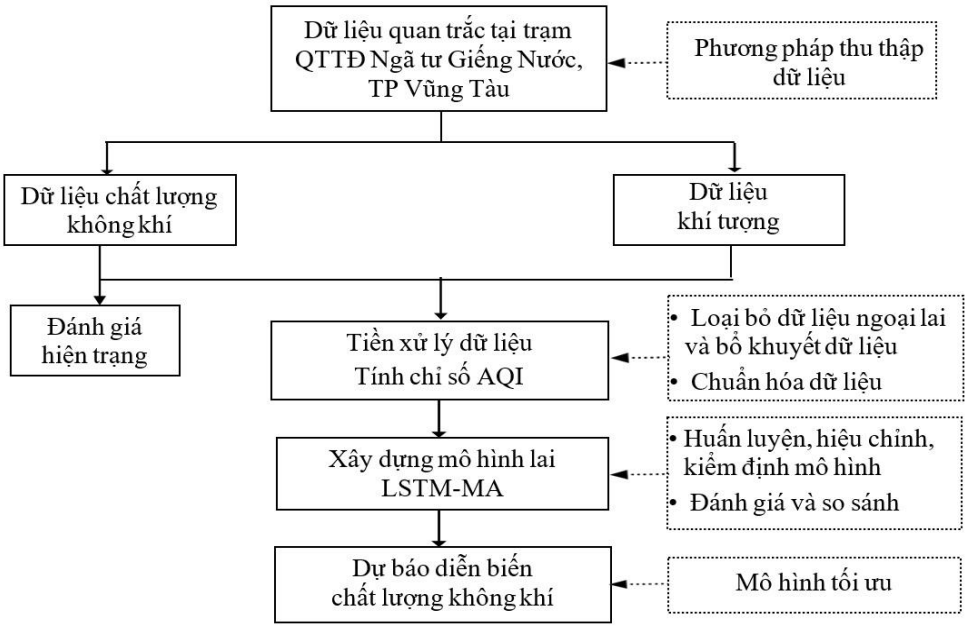
2.2. Phương pháp nghiên cứu

2.2.1. Tiền xử lý số liệu

Dự báo chất lượng không khí dựa trên mô hình học sâu đòi hỏi một lượng lớn tập dữ liệu để huấn luyện mô hình. Tuy nhiên, các tập dữ liệu thu thập được phân phối không đồng đều và không được ghi nhận (dữ liệu trống) do lỗi thiết bị hoặc sự cố mất điện và bảo trì. Sự tồn tại của những dữ liệu này ảnh hưởng đến độ tin cậy của bộ dữ liệu nghiên cứu, chúng có thể làm sai lệch kết quả phân tích dẫn đến giảm độ chính xác trong dự báo của các mô hình dự báo. Do đó, các tập dữ liệu cần được điền đầy đủ dữ liệu bị thiếu hoặc nội suy dữ liệu để các mô hình học sâu có thể được huấn luyện tốt hơn [13].

a) Loại bỏ dữ liệu ngoại lai

Các giá trị ngoại lai có thể dẫn đến quá trình huấn luyện mô hình diễn ra lâu hơn và mô hình kém chính xác hơn. Do đó, các dữ liệu ngoại lai cần phải được loại bỏ [19]. Có nhiều phương pháp để phát hiện dữ liệu ngoại lai nhưng cần phải dựa trên đặc tính của bộ dữ liệu, có thể kể đến một số phương pháp như: Phương pháp phát hiện dữ liệu ngoại vi theo độ lệch trung bình tuyệt đối (*mean absolute deviation - MAD*); Phương pháp đánh giá điểm số Z (Z - scores). Tuy nhiên, 2 phương pháp trên chỉ phù hợp cho xử lý dữ liệu tuân theo nguyên tắc phân phối chuẩn trong khi theo các nghiên cứu [30–32], bộ dữ liệu chất lượng không khí không phải là phân bố chuẩn. Để giải quyết vấn đề trên, phương pháp phát hiện dữ liệu ngoại lai bằng biểu đồ hộp (Box và Whisker) sẽ được sử dụng vì nó có khả năng làm việc tốt với



Hình 2. Sơ đồ cấu trúc nghiên cứu.

các bộ dữ liệu không tuân theo nguyên tắc phân bố chuẩn. Đây là một trong những phương pháp sử dụng công cụ đồ họa đơn giản nhưng có thể biểu diễn các thông tin về một biến liên tục, bao gồm trung vị, trung bình, các phân vị (phân vị 25% - Q1 và phân vị 75% - Q3) cùng với các cực trị. Trong đó, khoảng phân vị (*IQR - Inter quartile range*) là khoảng giá trị từ phân vị 25% (Q1) đến phân vị 75% (Q3). Khoảng tin cậy là khoảng giá trị nằm trong khoảng từ giới hạn dưới có giá trị là Q1 - 1,5IQR đến giới hạn trên có giá trị là Q3+1,5IQR. Một giá trị nếu nằm ngoài khoảng này có thể coi là giá trị ngoại vi. Tuy nhiên, trong thực tế có thể các giá trị được phát hiện là ngoại lai này có thể là các điểm dữ liệu ghi nhận được nồng độ ô nhiễm cao bất thường bởi ô nhiễm môi trường nào đó mà không phải do lỗi đo đạc hoặc tăng cao vào một số thời điểm nhất định trong ngày phụ thuộc theo điều kiện thời tiết và đặc điểm tại khu vực. Do đó, để chắc chắn không loại bỏ mất các dữ liệu ô nhiễm bất thường, các điểm dữ liệu được cho là ngoại lai phát hiện bởi biểu đồ hộp sẽ một lần nữa được kiểm tra so sánh. Quá trình phân tích và loại bỏ dữ liệu ngoại lai này sẽ được xử lý thông qua công cụ Microsoft Excel.

b) Điền dữ liệu khuyết

Tập dữ liệu quan trắc thu được có thể chứa các giá trị bị thiếu do thiết bị quan trắc bị trục trặc hoặc gặp sự cố về điện hoặc mạng gây ảnh hưởng cho quá trình truyền dữ liệu tại các trạm quan trắc. Các giá trị bị thiếu này gây ra sự gián đoạn tạm thời trong tập dữ liệu và cản trở quá trình huấn luyện. Do đó, để giải quyết vấn đề này, các giá trị bị thiếu sẽ được thay thế bằng giá trị được ghi vào cùng thời điểm một tuần trước đó vì chúng thuộc cùng một phân phối. Nếu không có gì được ghi lại vào cùng thời điểm một tuần trước đó, thì giá trị trung bình của dữ liệu sẽ được thay thế cho giá trị bị thiếu. Quá trình điền dữ liệu khuyết này sẽ được xử lý thông qua công cụ Microsoft Excel và SPSS.

c) Chuẩn hóa dữ liệu

Các chiều dữ liệu thường có sự khác biệt về đơn vị, phân phối và điều đó tác động không nhỏ lên hiệu quả phân loại của mô hình và khả năng hội tụ của các thuật toán trượt gradient. Một tập hợp dữ liệu có đơn vị quá khác biệt giữa các biến thường khiến gradient không hội tụ tới cực trị toàn cục. Các khác biệt về đơn vị cũng khiến việc đánh giá ảnh hưởng của các biến bị sai lệch nhiều hơn. Mục tiêu của việc chuẩn hóa là đưa các giá trị về gần hơn giá trị trung bình của các biến chứ không làm thay đổi hình dạng phân phối dữ liệu. Thông thường, phương pháp này đưa các giá trị về một khoảng đặc biệt, thường là [0,1] hoặc [−1,1] [33]. Để loại bỏ ảnh hưởng của sự khác biệt về chiều và cải thiện tốc độ hội tụ của các mô hình, tất cả dữ liệu sẽ được chuyển đổi theo phạm vi [0,1] dựa theo phương pháp chuẩn hóa Min-Max sau:

$$x' = \frac{x - \min_{(x)}}{\max_{(x)} - \min_{(x)}} \quad (1)$$

Trong đó x’ là giá trị dữ liệu được chuyển đổi (giá trị 0 - 1); x là giá trị gốc.

d) Bộ lọc trung bình trượt (*Moving average filter*)

Như đã biết, diễn biến của các thông số chất lượng không khí còn là một hàm số có quy luật theo thời gian và thường được sử dụng rộng rãi trong việc dự báo dài hạn [16−19]. Chất lượng dữ liệu theo thời gian này có tác động rất đáng kể đến khả năng dự báo chính xác của mô hình LSTM. Dữ liệu chính xác và đa dạng giúp mô hình học được nhiều mẫu và xu hướng, từ đó cải thiện khả năng dự đoán. So với các phương pháp xử lý dữ liệu khác, thuật toán sử dụng cho bộ lọc trung bình trượt đơn giản và phù hợp xử lý dữ liệu chuỗi thời gian [34].

Bộ lọc trung bình trượt là một bộ lọc phản hồi xung hữu hạn (*FIR - Finite Impulse Response*) thường được sử dụng để làm trơn một mảng dữ liệu thu thập được. Bộ lọc sẽ có N mẫu dữ liệu đầu vào tại một thời điểm và cho một dữ liệu đầu ra duy nhất lấy giá trị trung bình của N mẫu đó. Dữ liệu đầu ra có độ mượt tăng dần theo độ dài của N mẫu dữ liệu, tuy nhiên đáp ứng tần suất thể hiện của bộ dữ liệu sẽ trở nên kém đi. Phương trình sai phân cho

bộ lọc trung bình di chuyển theo thời gian rời rạc N điểm với đầu vào được biểu thị bằng vectơ y_i và vectơ đầu ra trung bình y_{tr} được thể hiện theo công thức sau [35]:

$$y_{tr} = \frac{1}{m} \sum_{i=t-m+1}^{t} y_i \quad (2)$$

Theo nghiên cứu của Babu, sử dụng bộ lọc MA để chia bộ dữ liệu thành 2 phần: 1 phần với độ biến động thấp làm dữ liệu đầu vào cho mô hình ARIMA, phần còn lại mang tính chất biến động cao để sử dụng cho mô hình ANN [36].

2.2.2. Tính toán chỉ số chất lượng không khí AQI

Phương pháp tính toán chỉ số chất lượng không khí Việt Nam (VN_AQI) từ dữ liệu quan trắc của trạm quan trắc không khí tự động, liên tục được ban hành kèm theo Quyết định số 1459/QĐ-TCMT ngày 12/11/2019 của Tổng Cục Môi trường [37]. Chỉ số chất lượng không khí được tính theo thang điểm (khoảng giá trị AQI) tương ứng với biểu tượng và các màu sắc để cảnh báo chất lượng không khí và mức độ ảnh hưởng tới sức khỏe con người.

2.2.3. Xây dựng mô hình dự báo chất lượng không khí LSTM-MA

Mô hình LSTM được sử dụng để dự báo chất lượng không khí AQI cho trạm Ngã tư Giếng nước ở tỉnh Bà Rịa - Vũng Tàu. Mô hình LSTM là một loại mạng nơ-ron phản hồi được xây dựng bởi [38]. Mô hình này được cải tiến để tăng hiệu quả hoạt động bởi [39]. Như đã đề cập ở trên, các biến động ngẫu nhiên, trị bất thường, yếu tố gây nhiễu của dữ liệu ảnh hưởng đáng kể đến độ chính xác dự báo của mô hình LSTM. Do đó, nghiên cứu sử dụng kết hợp bộ lọc trung bình trượt MA để lọc các dữ liệu có sự biến động tự nhiên theo thời gian. Johnston đã chứng minh MA sẽ làm tăng tính ổn định của mô hình dự báo và làm giảm phương sai sai số dự báo ít nhất 3% khi sử dụng bộ lọc này [36].

Tiến trình thực hiện xây dựng mô hình nghiên cứu được mô tả tóm tắt theo Hình 3.

Kịch bản dự báo: kịch bản dự báo trong nghiên cứu là điều kiện phát triển kinh tế, xã hội tại khu vực phát triển bình thường và không có sự biến động đột biến.

Dữ liệu đầu vào: Dữ liệu chỉ số chất lượng không khí AQI được tính toán dựa trên 06 thông số: CO, NO_2, O_3, SO_2, $PM_{2.5}$ và PM_{10}. Theo nhiều nghiên cứu chỉ ra tập luyện có tỷ lệ từ 60-80% và các tập dữ liệu còn lại được sử dụng cho đánh giá mô hình nhằm hạn chế hiện tượng quá khớp (*overfitting*). Chỉ số chất lượng không khí sau khi được tính toán sẽ được chia thành 3 tập dữ liệu gồm: huấn luyện, kiểm định và thử nghiệm theo tỉ lệ 70:15:15.

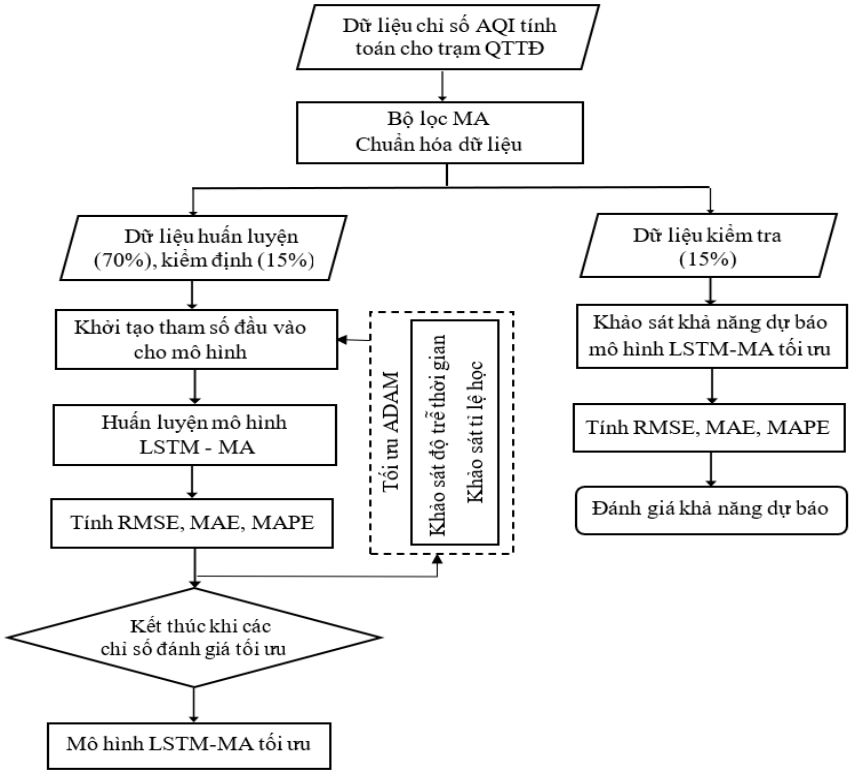
Để xác định cấu trúc mạng LSTM-MA tối ưu, 03 thí nghiệm được thiết lập:

Thí nghiệm 1 (xác định độ trễ thời gian *r*): Vì giá trị N tối ưu cho bộ lọc MA chỉ được áp dụng khi bộ dữ liệu đầu ra phù hợp với nhu cầu của mô hình [33, 40] do đó nghiên cứu đề xuất khảo sát giá trị N cho bộ lọc là 7 và 28 tương ứng với biến động chỉ số chất lượng không khí trong ngắn hạn (theo tuần) và dài hạn (theo tháng). Theo Li và cs (2017), độ trễ thời gian bộ dữ liệu có tác động đáng kể đến khả năng dự báo của mô hình. Với độ trễ thời gian nhỏ, dữ liệu đầu vào không thể đảm bảo đầy đủ; Do đó, mô hình không thể khai thác triệt để LSTM cho dự báo dài hạn. Độ trễ thời gian lớn cho phép tăng số lượng đầu vào không liên quan, tuy nhiên nó đồng thời làm tăng độ phức tạp của mô hình và khó khăn trong việc huấn luyện từ các tính năng hữu ích [1]. Nghiên cứu đề xuất xác định độ trễ thời gian r tối ưu thay đổi: 4, 6, 8, 12.

Bộ dữ liệu đầu vào với giá trị N thay đổi từ bộ lọc MA sẽ được sử dụng cho mô hình. Các tham số khác trong mô hình số nơ-ron ẩn là 100, số vòng lặp cực đại (*epoch*) được điều chỉnh ở mức 100, tỷ lệ học ban đầu 0,001, thuật toán tối ưu ước tính thời điểm điều chỉnh (*adaptive moment estimation algorithm - Adam*).

Thí nghiệm 2 (xác định tỷ lệ học tối ưu): Trong thí nghiệm này, độ trễ thời gian sẽ được giữ nguyên theo kết quả đã được xác định ở thí nghiệm 1. Nghiên cứu thực hiện luyện mạng LSTM với tỷ lệ học ban đầu biến thiên trong khoảng 0,001-0,009. Các tham số khác trong

mô hình số nơ-ron ẩn là 100, số vòng lặp cực đại là 100, ngưỡng gradient là 1, thuật toán tối ưu ước tính thời điểm điều chỉnh - Adam.



Hình 3. Sơ đồ các bước xây dựng mô hình LSTM-MA.

Thí nghiệm 3 (xác định khả năng dự báo của mô hình): Sau khi đã xác định được các thông số tối ưu cho mô hình LSTM-MA, nghiên cứu tiến hành xác định khả năng dự báo chất lượng không khí. Thời gian dự báo (bước nhảy) của mô hình được tiếp tục khảo sát để đánh giá khả năng dự báo của mô hình, lần lượt k = 2, k = 4, k = 6 và k = 14.

Huấn luyện, đánh giá, kiểm định hiệu suất của mô hình*:*

Tập dữ liệu được chia thành các tập huấn luyện, đánh giá và kiểm định. Mô hình sẽ được tiến hành đánh giá thông qua các chỉ số như RMSE, MAE, MAPE như sau:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y_i^*)^2} \;;\quad \mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - y_i^*| \;;\quad \mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - y_i^*|}{y_i^*} \qquad (3)$$

Trong đó y_i^* là giá trị chất lượng không khí thực tế, y_i là giá trị chất lượng không khí dự đoán và n là số lượng mẫu dữ liệu.

Nền tảng sử dụng và thông số kỹ thuật máy tính để xây dựng mô hình:

Mô hình LSTM và LSTM-MA được nghiên cứu và phát triển thông qua sử dụng ngôn ngữ Python. Thông số kỹ thuật của nền tảng cũng như tài nguyên máy đã sử dụng trong nghiên cứu này bao gồm: Ngôn ngữ sử dụng (Python phiên bản 3.10); Thư viện AI sử dụng (pandas, numpy, tensorflow, matplotlib, sk.learn); Thông số GPU sử dụng để huấn luyện các mô hình nghiên cứu (12th Gen Intel(R) Core(TM) i5–1240P 1.70 GHz, RAM 8 GB).

3. Kết quả và thảo luận

3.1. Đánh giá hiện trạng chất lượng không khí tại khu vực Ngã tư Giếng Nước

3.1.1. Hiện trạng bộ dữ liệu

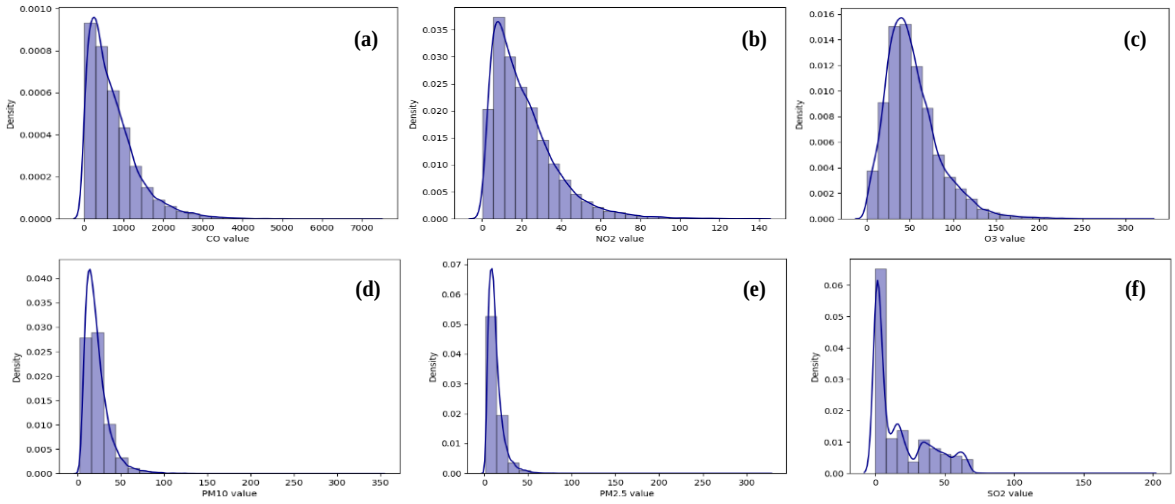
Bộ dữ liệu thu thập tại trạm QTTĐ Ngã tư Giếng nước bao gồm tổng cộng 25.764 dữ liệu cho từng thông số cho thời gian từ 18/1/2020 đến 31/12/2022. Tỉ lệ số liệu thu thập được cho cả 6 chất ô nhiễm đều cao trên 99%. Trong tất cả các thông số được sử dụng, CO là thông số có sự biến thiên lớn nhất có độ lệch chuẩn lên đến 623,5 µg/Nm^3. Biến động của CO trong

tập dữ liệu cũng là điều bình thường vì nguồn gốc của CO phát sinh tại khu vực chủ yếu đến từ việc đốt nguyên liệu hóa thạch và hoạt động giao thông trong khu vực. Giá trị NO_2 và SO_2 có cùng xu hướng khi diễn biến ổn định và không có sự chênh lệch nhiều giữa các ngày trong năm. Giá trị trung bình của $PM_{2.5}$ và PM_{10} là 12,9 µg/Nm^3 và 22,2 µg/Nm^3, tuy nhiên, giá trị cực đại ghi nhận được ở 2 thông số này đều lớn hơn 300 µg/Nm^3, điều này tiềm ẩn khả năng chất lượng không khí bị ô nhiễm từ nguồn bụi mịn có trong không khí.

3.1.2. Hiện trạng chất lượng không khí

Giá trị trung bình giờ của các thông số CO, NO_2 và SO_2 luôn nằm trong giới hạn cho phép. Đối với O_3 trung bình giờ, có 64 thời điểm vượt QCVN 05:2023/BTNMT, chủ yếu vào các thời điểm từ 12-14 giờ trong ngày. Với O_3 trung bình 8 giờ, 660 thời điểm vượt ngưỡng cho phép, chủ yếu rơi vào thời điểm từ 16-20 giờ do đã đạt giá cực đại từ thời điểm buổi trưa. Điều này phù hợp với quy luật O_3 sẽ tăng lên vào buổi sáng, đạt giá trị cực đại vào buổi trưa và giảm dần vào chiều tối. Cường độ bức xạ mặt trời là nguyên nhân chính ảnh hưởng tới sự biến động nồng độ O_3 tầng mặt. Đối với thông số PM_{10} và $PM_{2.5}$, mặc dù ghi nhận giá trị trung bình giờ tại một số thời điểm trong ngày cao nhưng nồng độ trung bình 24 giờ đều đạt trong ngưỡng cho phép. Nhìn chung, nồng độ bụi sẽ đạt cao nhất vào khung giờ 7-8 giờ sáng. Thời điểm này thường là giờ cao điểm của buổi sáng bởi nhu cầu di chuyển của phương tiện giao thông cao và góp phần tạo ra lượng lớn chất khí ô nhiễm. Sau giờ cao điểm buổi sáng, nồng độ bụi giảm dần và có xu hướng thấp nhất vào lúc 15-16 giờ. Sau đó, nồng độ bụi sẽ tăng nhẹ vào giờ cao điểm buổi chiều (17-19 giờ). Sự phát thải từ giao thông được xem là nguyên nhân chính dẫn đến sự gia tăng hàm lượng bụi trong không khí vào các giờ cao điểm trong khu vực.

Theo Hình 3, dữ liệu cho các chất ô nhiễm đều có phân phối lệch dương. Việc dữ liệu thu thập trong nghiên cứu không phải là phân bố chuẩn cùng phù hợp theo các báo cáo đã thực hiện trên thế giới [30–32]. Theo Kumar, nhiều mô hình hoạt động tốt hơn khi dữ liệu có phân phối bình thường và hoạt động kém hơn khi dữ liệu có phân phối lệch [41]. Do đó, dữ liệu phải được chuẩn hóa sự khác biệt về cường độ để giảm tác động của các giá trị ngoại lai này.



Hình 4. Phân phối dữ liệu cho các chất ô nhiễm không khí: (a) CO, (b) NO_2, (c) O_3, (d) PM_{10}, (e) $PM_{2.5}$, (f) SO_2.

3.2. Xử lý dữ liệu ngoại lai và điền dữ liệu khuyết

Chất lượng dữ liệu là điều kiện tiên quyết đầu tiên và quan trọng nhất để trực quan hóa và tạo ra các mô hình dự báo hiệu quả. Các bước tiền xử lý giúp giảm nhiễu có trong dữ liệu, từ đó tăng tốc độ xử lý và khả năng tổng quát hóa cho các thuật toán. Bộ dữ liệu sẽ được tiến hành tiền xử lý dữ liệu với phương pháp Box Whisker để loại bỏ dữ liệu ngoại lai.

Bảng 2. Thống kê theo phương pháp Box-Whisker cho bộ dữ liệu thô.

Chất ô nhiễm	Min	Phân vị 25%–Q1	Trung vị	Trung bình	Phân vị 75%–Q3	Max	IQR (Q3–Q1)	Q1–1,5* IQR	Q3+1,5*IQR
CO	0,1	270,7	567,4	725,9	1.004,4	7.261,7	733,7	0,1	2.104
NO_2	0,1	9,3	17,2	21,3	28,6	139,2	19,3	0,1	57,5
O_3	0	31,6	47,5	53,5	68,7	320	37,1	0	124,3
SO_2	0,1	1,7	7,4	17,7	32,7	194	31,0	0,1	75,1
$PM_{2,5}$	1,1	7,1	10,7	12,9	13,9	324	6,8	1,1	29,1
PM_{10}	2	12,7	18,8	22,2	27,7	348,9	15	2	50,2

Từ kết quả thống kê ở Bảng 2 cho thấy số lượng các điểm được cho là giá trị ngoại lai là không nhiều. Khi lọc các dữ liệu cho thông số O_3, phát hiện 844 dữ liệu (chiếm khoảng 3,2% tổng bộ dữ liệu) có giá trị nồng độ lớn hơn khoảng Q3+1,5×IQR tương ứng là 124,3 µg/Nm^3. Tuy nhiên, đa số các dữ liệu này đều rơi vào thời điểm buổi trưa khi nhiệt độ và bức xạ mặt trời tăng cao khiến cho xảy ra việc hình thành O_3. Điều này củng cố cho việc điểm dữ liệu ghi nhận được nồng độ ô nhiễm cao hơn so với đa số bộ dữ liệu bởi ô nhiễm môi trường nào đó mà không phải do lỗi đo đạc hoặc tăng cao vào một số thời điểm nhất định trong ngày phụ thuộc theo điều kiện thời tiết và hoạt động tại khu vực. Ngoài ra, bộ dữ liệu thu thập được cho các thông số ô nhiễm đạt tỉ lệ cao trên 99% nên việc loại bỏ những dữ liệu ngoại lai này sẽ làm mất đi tính khách quan và thực tế của trạm QTTĐ. Do đó, bộ dữ liệu sẽ chỉ được bổ sung những dữ liệu còn thiếu.

3.3. Đánh giá kết quả huấn luyện mô hình dự báo chất lượng không khí

3.3.1. Tính toán và đánh giá dữ liệu cho mô hình dự báo.

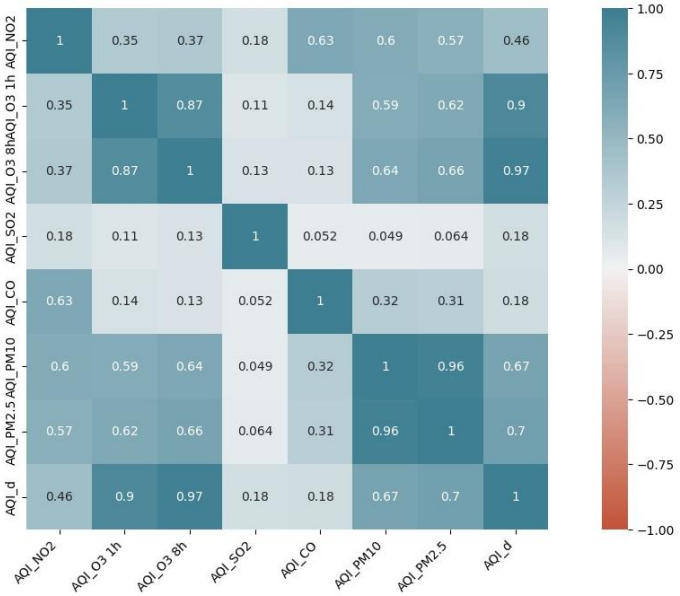
a) Tính toán chỉ số AQI

Theo Hình 5, chỉ số AQI có mức độ tương quan cao nhất lần lượt với giá trị O_3 trong 8 giờ và O_3 trong 1 giờ, tương ứng với mức 0,97 và 0,9. Điều này cũng phù hợp với thực tế giá trị O_3 8 giờ và O_3 1 giờ có số thời điểm tăng cao và vượt ngưỡng cho phép so với các thông số còn lại. Việc tính toán mối tương quan cũng chỉ ra rằng CO và SO_2 có tác động ít nhất đối với chỉ số AQI. Ngoài ra, có thể thấy tất cả các giá trị của các chất gây ô nhiễm trong tập dữ liệu đều có mối tương quan dương với chỉ số AQI. Điều này cho thấy rằng khi nồng độ các chất ô nhiễm tăng lên, giá trị AQI cũng tăng theo, phản ánh chất lượng không khí kém hơn. Nhận thức này nhấn mạnh tầm quan trọng của việc xem xét các yếu tố này trong việc phân tích và dự đoán sự thay đổi chất lượng không khí trong khu vực nghiên cứu. Chúng ta nên chọn những yếu tố quan trọng có mối tương quan đáng kể với AQI.

b) Đánh giá diễn biến chất lượng không khí theo năm

Các kết quả phân tích, đánh giá diễn biến chất lượng không khí tại trạm Ngã tư Giếng nước cho thời gian nghiên cứu theo từng năm như sau: Giá trị AQI ngày trong năm 2020 dao động từ 18-177, trong đó có 277 ngày chỉ số AQI nằm trong khoảng 0-50 (chất lượng không khí Tốt) chiếm 80,1 % số ngày trong năm, 32 ngày chỉ số AQI nằm trong khoảng 51-100 (chất lượng không khí Trung bình) chiếm 9,2 % và 37 ngày chỉ số AQI nằm trong khoảng 101-200 (chất lượng không khí Kém) chiếm 3,56 %; Giá trị AQI ngày trong năm 2021 dao động trong khoảng từ 14-143, trong đó, có 306 ngày chỉ số AQI nằm trong khoảng 0-50 chiếm 84,1 %, 45 ngày chỉ số AQI nằm trong khoảng 51-100 chiếm 12,4 % và 13 ngày chỉ số AQI nằm trong khoảng 101-200 chiếm 3,6 %. Giá trị AQI ngày cao nhất trong tháng 1 và thấp vào các tháng từ tháng 6 đến tháng 11 do giãn cách xã hội; Giá trị AQI ngày trong năm 2022 dao động trong khoảng từ 17-196, trong đó có 268 ngày chỉ số AQI nằm trong khoảng 0-50 chiếm 73,6 %, 45 ngày chỉ số AQI nằm trong khoảng 51-100 chiếm 12,4 % và 51 ngày chỉ số AQI nằm trong khoảng 101-200 chiếm 14,0 %. Nhìn chung, chất lượng không khí tại khu vực trạm QTTĐ Ngã tư Giếng nước tương đối tốt với phân bậc chất lượng không khí

đều nằm trong nhóm "Tốt" và "Trung bình". Tuy nhiên, sau khoảng thời gian dịch vào năm 2021, xu hướng chất lượng không khí có thể chuyển biến theo xu hướng suy giảm về chất lượng do các hoạt động đã trở lại bình thường và trạm QTTĐ mang đặc tính của khu vực giao thông đô thị.



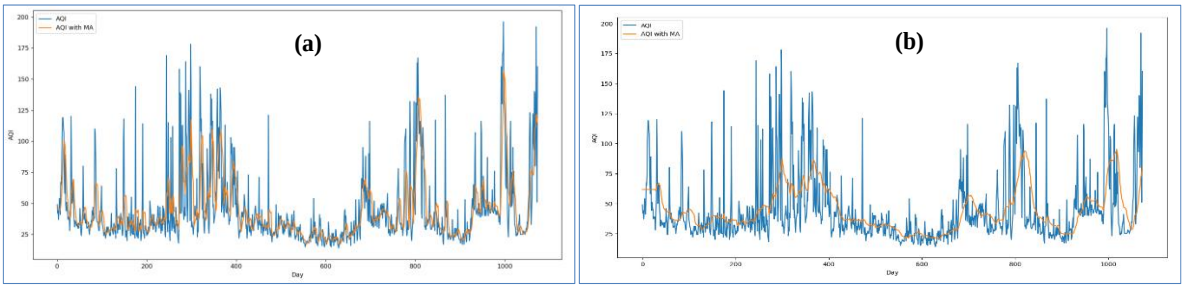
Hình 5. Tương quan giữa chỉ số AQI ngày và các chỉ số AQI thành phần.

c) Đánh giá diễn biến chất lượng không khí theo tháng

Giá trị AQI ngày trong thời gian khảo sát có xu hướng cao hơn vào thời điểm đầu và cuối năm (từ tháng 10 đến tháng 4 năm sau), thấp hơn vào giữa năm (tháng 5 đến tháng 9). Trong đó, tháng 7 là thời điểm chỉ số AQI thấp nhất và tháng 12 là thời điểm chỉ số AQI diễn biến cao nhất. Nguyên nhân là vào mùa hè, bức xạ mặt trời cao khiến nhiệt độ bề mặt tăng mạnh và làm nóng không khí gần bề mặt, điều này dẫn đến sự đối lưu gia tăng, khí quyển không ổn định, có lợi cho sự khuếch tán và lắng đọng các chất ô nhiễm không khí [17]. Ngoài ra, thời điểm này thường cũng sẽ có gió mạnh hơn có thể giúp phân tán các chất ô nhiễm này.

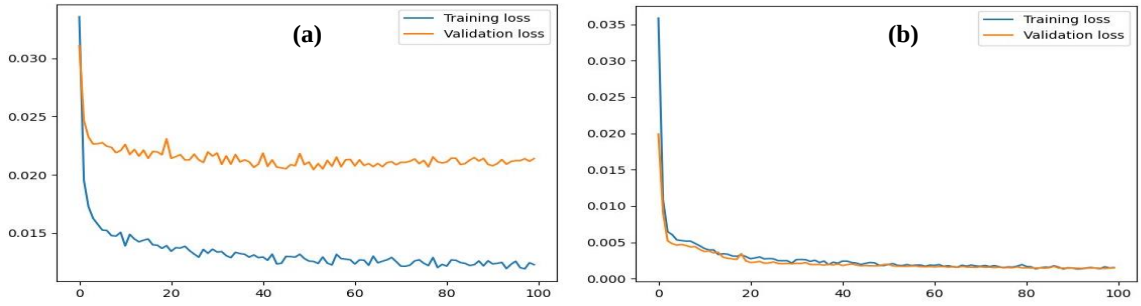
3.3.2. Huấn luyện mô hình dự báo mạng LSTM-MA

Bộ lọc trung bình động MA được sử dụng làm mịn dữ liệu (Hình 6). Đầu ra thu được từ xử lý MA được sử dụng là đầu vào của mô hình LSTM. Để xác định cấu trúc mạng của mô hình LSTM-MA. Nghiên cứu đã thực hiện thí nghiệm độ trễ thời gian r và tỷ lệ học tối ưu. Thiết lập tham số luyện mạng ban đầu cho mô hình gồm số nơ-ron ẩn là 100, số vòng lặp cực đại (*epoch*) được điều chỉnh ở mức 100 với tỷ lệ học ban đầu 0,001. Trong quá trình huấn luyện, qua mỗi epoch độ chính xác của mô hình sẽ tăng dần, tương ứng với sai số giảm dần. Hình 7 miêu tả sự thay đổi của độ mất mát qua các epochs trong quá trình huấn luyện mô hình. Đối với mô hình LSTM, hiện tượng underfitting đã xảy ra khi độ mất mát từ quá trình kiểm định có khoảng trống rõ rệt với quá trình huấn luyện. Điều này mang ý nghĩa mô hình



Hình 6. Kết quả chạy bộ lọc MA: (a) N = 7, (b) N = 28.

không thể nắm bắt được các giá trị phức tạp có trong dữ liệu như đã phân tích ở trên và cần thiết loại nhiễu khỏi bộ dữ liệu.



Hình 7. Độ mất mát từ quá trình huấn luyện cho mô hình: (a) LSTM, (b) LSTM–MA.

Kết quả thí nghiệm độ trễ thời gian r: Đầu tiên, thí nghiệm kiểm tra ảnh hưởng của các độ trễ thời gian khác nhau. Thông qua so sánh chỉ số RMSE, MAE và MAPE để đánh giá hiệu suất dự đoán của mô hình lai LSTM-MA và được thể hiện trong Bảng 3.

Bảng 3. Ảnh hưởng của độ trễ thời gian trên bộ dữ liệu huấn luyện.

Độ trễ r	Bộ lọc MA với N = 7			Bộ lọc MA với N = 28		
	RMSE	MAE	MAPE (%)	RMSE	MAE	MAPE (%)
4	6,33	3,86	7,97	3,95	2,60	4,37
6	6,25	3,88	6,97	5,06	3,42	4,44
8	6,71	4,14	7,99	4,60	3,16	5,06
10	6,13	4,05	8,09	5,21	3,58	4,91
12	6,01	3,78	7,11	4,49	3,16	4,57
14	6,58	4,21	8,79	4,76	3,49	5,13

Kết quả thí nghiệm cho thấy giá trị N của bộ lọc MA có tác động đáng kể đến khả năng hoạt động của mô hình LSTM. Khi lọc nhiễu dữ liệu không khí trong thời gian ngắn hạn, giá trị RMSE khi thay đổi độ trễ r dao động từ 6,01-6,71, giá trị MAE dao động từ 3,86-4,21, MAPE dao động từ 6,97-8,39. Với bộ dữ liệu không khí trong thời gian dài hạn, giá trị RMSE giảm xấp xỉ 50% và dao động từ 3,19-4,05, MAE giảm và dao động từ 2,24-2,90, MAPE cũng giảm tương ứng và dao động từ 3,42-4,04. Tại thời điểm r = 12, các chỉ số đánh giá mô hình RMSE, MAE và MAPE có xu hướng ổn định đạt giá trị tối ưu. Do đó, chọn r = 12 từ bộ dữ liệu lọc MA trong dài hạn làm thông số tối ưu để tiến hành tiếp thí nghiệm khảo sát tỉ lệ học. Thí nghiệm cũng chỉ ra rằng độ trễ thời gian trong quá khứ của mẫu dữ liệu có tác động đáng kể đến khả năng nắm bắt thông tin về diễn biến của dữ liệu theo thời gian trong quá khứ, đặc biệt là dữ liệu có tính chu kỳ.

Kết quả thí nghiệm tỉ lệ học: Tỉ lệ học là một trong số những tham số quan trọng nhất của mô hình. Độ lớn của tỉ lệ học sẽ ảnh hưởng trực tiếp tới tốc độ mô hình thay đổi các trọng số để phù hợp với các thuật toán. Tốc độ học lớn có thể giúp mạng nơ-ron được huấn luyện nhanh hơn gấp 10 lần [42] nhưng cũng có thể làm giảm độ chính xác. Khi tỉ lệ học tăng từ 0,001 đến 0,007, các chỉ số đánh giá RMSE, MAE và MAPE có xu hướng giảm dần và đạt giá trị tối ưu tại 0,007 với RMSE đạt 3,05, MAE đạt 2,17 và MAPE đạt 3,19%. Khi tỉ lệ học tăng dần, các thông số này tăng dần và đạt giá trị cao nhất tại 0,009 cho thấy độ chính xác khi huấn luyện mô hình đã giảm đáng kể. Do đó, chọn tỉ lệ học tại giá trị 0,007 làm tham số tối ưu cho mô hình.

3.4. Dự báo chất lượng không khí cho tỉnh Bà Rịa - Vũng Tàu sử dụng mô hình tối ưu

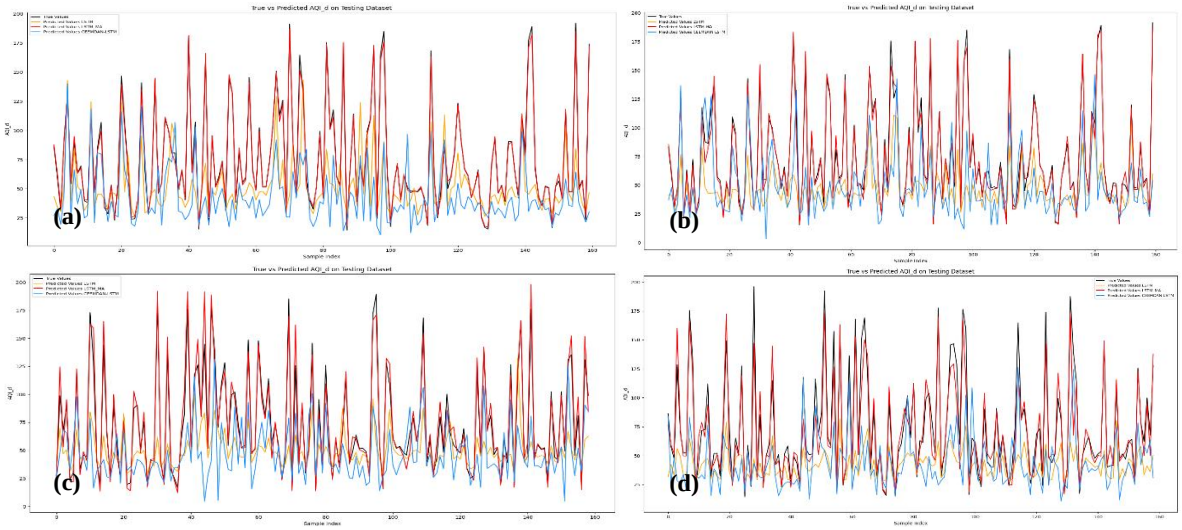
Từ các khảo sát mô hình tối ưu đã thực hiện ở trên, nghiên cứu tiến hành đánh giá khả năng dự báo của mô hình lai LSTM-MA và được tổng hợp ở Bảng 4. Kết quả dự báo cho thời điểm 1 ngày ở tương lai cho thấy kết quả dự báo đạt độ chính xác cao nhất, với các chỉ

số RMSE, MAE và MAPE trên dữ liệu kiểm tra lần lượt là 3,05; 2,17 và 3,19%. Khi tăng thời gian dự báo lên tương ứng 2 và 4 ngày trong tương lai, hiệu suất dự báo lần lượt giảm đi tương ứng xấp xỉ 2 lần và 3 lần, tuy nhiên, giá trị của các chỉ số đánh giá là chấp nhận được khi RMSE đạt 6,71 và 9,36, MAPE lần lượt đạt 6,81% và 8,77%. Khi đánh giá khả năng dự báo cho 1 và 2 tuần tiếp theo, mô hình LSTM-MA vẫn đạt độ chính xác trong khả năng dự báo khi MAPE ở mức 17,10% và 24,38%, MAE ghi nhận được mặc dù có tăng nhưng đều ở mức chấp nhận khi tăng lần lượt lên 10,60 và 15,74. Kết quả cho thấy được mô hình lai LSTM-MA đã nắm bắt tốt biến động từ quá khứ của chỉ số chất lượng không khí để có thể dự báo trong tương lai.

Bảng 4. Hiệu suất dự báo của mô hình lai LSTM–MA trên tập dữ liệu kiểm tra.

Thời gian dự báo	RMSE	MAE	MAPE (%)
T+1	3,05	2,17	3,19
T+2	6,71	4,66	6,81
T+4	9,36	6,44	8,77
T+7	14,62	10,60	17,10
T+14	22,79	15,74	24,38

Để có so sánh một cách trực quan giữa mô hình lai LSTM-MA và mô hình LSTM, các kết quả dự báo của hai mô hình sẽ được biểu diễn tại Hình 9. Khả năng dự báo của mô hình lai LSTM-MA bỏ xa LSTM khi chúng dễ dàng nắm bắt được các thời điểm mà chỉ số chất lượng không khí biến động cao và thể hiện rõ nhất ở Hình 9a. Điều này là phù hợp như kết quả đã chỉ ra khi mô hình LSTM không thể đạt được hiệu suất tốt ngay từ trong quá trình huấn luyện. Từ Hình 9b-d, mô phỏng từ mô hình LSTM-MA vẫn có khả năng nắm bắt được các giá trị cực trị. Đối với mô phỏng dự báo tương ứng 14 ngày trong tương lai, mô hình LSTM-MA vẫn có thể biểu diễn tốt xu hướng diễn biến chỉ số chất lượng không khí, tuy nhiên lại không thể hiện được khi các giá trị có xu hướng tăng và dữ liệu quá khứ không đủ mô phỏng được diễn biến trong thời gian dài hơn. Kết quả này khá tương đồng với kết quả nghiên cứu [1] khi muốn mô hình đạt kết quả dự báo trong dài hạn thì cần số lượng độ trễ thời gian dài hơn.



Hình 8. Mô phỏng kết quả dự báo chất lượng không khí (a–d) trên tập dữ liệu kiểm tra: (a) Dự báo 1 ngày, (b) 2 ngày, (c) Dự báo 4 ngày, (d) Dự báo 7 ngày tiếp theo.

Ngoài ra, trong nghiên cứu này đồng thời thử nghiệm thuật toán phân rã trạng thái thực nghiệm CEEMDAN như một phương pháp xử lý dữ liệu đầu vào. Thuật toán CEEMDAN có thể phân tách hoàn toàn dữ liệu gốc, có tính biến động mạnh, thành một số thành phần chức năng chế độ nội tại (IMFs) với các đặc tính tần số khác nhau, do đó làm giảm tính biến động

của dữ liệu và cải thiện độ chính xác của dự đoán. Mô hình CEEMDAN-LSTM cho kết quả dự báo trong thời gian 1 ngày tiếp theo với RMSE đạt 16,0 và MAE đạt 9,85, cải thiện kết quả rất tốt so với mô hình đơn LSTM. Kể cả trong dự báo dài hạn là 14 ngày tiếp theo, CEEMDAN-LSTM cho kết quả khá tương đồng với kết quả nhận được từ LSTM-MA với RMSE đạt 23,14 và MAE đạt 16,17. Điều này cho thấy tiềm năng của việc ứng dụng thuật toán CEEMDAN trong dự báo dài hạn cho các nghiên cứu sau này. Nhìn chung, thứ tự độ tin cậy và chính xác của các mô hình dự báo trong nghiên cứu được sắp xếp như sau: LSTM-MA > CEEMDAN-LSTM > LSTM.

Bảng 5. Tương quan hiệu suất giữa các mô hình về dự báo chỉ số AQI.

Nghiên cứu	Mô hình	Thời gian dự báo	RMSE	MAE
Trong nghiên cứu này	LSTM	1 ngày tiếp theo	24,24	13,97
		14 ngày tiếp theo	27,17	19,06
	LSTM-MA	1 ngày tiếp theo	3,05	2,17
		14 ngày tiếp theo	22,79	15,74
Yan và cs [17]	LSTM	1h tiếp theo	23,7	–
		6h tiếp theo	49,2	–
	CNN-LSTM	1h tiếp theo	22,9	–
		6h tiếp theo	47,1	–
Duan và cs [20]	LSTM	Không đề cập	12,3–48,0	9,1 – 32,9
	CEEMDAN-LSTM	Không đề cập	6,6–24,8	4,6 – 17,9

Bảng 5 trình bày hiệu suất giữa mô hình LSTM-MA trong nghiên cứu này và một số nghiên cứu khác, mặc dù so sánh này không thực sự là hợp lý bởi chất lượng và tính chất dữ liệu của các nghiên cứu là khác nhau. Tuy nhiên, nó cung cấp thêm một bức tranh về tiềm năng ứng dụng AI nói chung và các mạng nơron học sâu trong dự báo chất lượng không khí.

4. Kết luận

Ứng dụng trí thông minh nhân tạo trong dự báo chất lượng không khí là xu hướng mới trong thời gian gần đây. Nghiên cứu đã thực hiện và đạt được một số kết quả sau: (i) Đã đánh giá được hiện trạng chất lượng không khí tại trạm QTTĐ Ngã tư Giếng nước, thành phố Vũng tàu, tỉnh Bà Ria - Vũng Tàu giai đoạn năm 2020-2022. Bộ dữ liệu thu thập được đạt mức độ đầy đủ của các thông số trên 99%. Chất lượng không khí tại khu vực nghiên cứu là tương đối tốt khi nồng độ các thông số CO, NO_2, SO_2, PM_{10} và $PM_{2.5}$ đều dưới ngưỡng cho phép. Ozon là thông số có số lần vượt ngưỡng cho phép cao nhất và cũng là thông số có tác động chính đến chỉ số AQI. Chất lượng không khí tại khu vực này chủ yếu bị tác động bởi chất ô nhiễm Ozon trung bình 1 giờ và 8 giờ; (ii) Mô hình lai LSTM-MA đã được xây dựng thành công với khả năng dự báo có độ chính xác cao nhất cho thời gian 1 ngày tiếp theo với RMSE là 3,05; MAE là 2,17 và MAPE là 3,19%. Khi dự báo trong thời gian dài hơn tương ứng 2 tuần trong tương lai, mô hình cho kết quả khả quan khi các chỉ số đánh giá RMSE, MAE và MAPE lần lượt đạt 22,79; 15,74 và 24,38%. Các kết quả nghiên cứu có thể áp dụng cho bộ số liệu từ các trạm QTTĐ khác nhằm đánh giá tổng quát hơn khả năng ứng dụng của mô hình. Bên cạnh đó, cần có những nghiên cứu sử dụng dữ liệu phụ trợ như dữ liệu khí tượng và dữ liệu không gian để có thể cải thiện đáng kể hiệu suất dự đoán; Áp dụng những phương pháp xử lý dữ liệu khác như CEEMDAN hoặc lai hợp với các mô hình khác như CNN-LSTM để cải thiện khả năng dự báo

Đóng góp của tác giả: Xây dựng ý tưởng nghiên cứu: H.M.D.; Xử lý số liệu, chạy mô hình: K.D.A.K., H.M.D.; Viết bản thảo bài báo: H.M.D., K.D.A.K.

Lời cảm ơn: Nghiên cứu được tài trợ bởi Đại học Quốc gia Thành phố Hồ Chí Minh (ĐHQG-HCM) trong khuôn khổ Đề tài mã số C2023-24-02.

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. Li, X.; et al. Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation. *Environ. Pollut.* **2017**, *231*, 997–1004.
2. Phùng, N.K. Nghiên cứu tính toán và dự báo $PM_{2.5}$ cho khu vực Thành phố Hồ Chí Minh. *Tạp chí khí tượng thủy văn* **2018**, *659*, 1–7.
3. Buonocore, J.J.; et al. Using the Community Multiscale Air Quality (CMAQ) model to estimate public health impacts of $PM_{2.5}$ from individual power plants. *Environ. Int.* **2014**, *68*, 200–208.
4. Wang, L.; et al. Source apportionment of $PM_{2.5}$ in top polluted cities in Hebei, China using the CMAQ model. *Environ. Int.* **2015**, *122*, 723–736.
5. Quy, L.V.; và cs. Ứng dụng công cụ kết nối song song mô hình WRF–CMAQ đánh giá nồng độ một số chất ô nhiễm không khí cho Việt Nam. *Tạp chí môi trường* **2018**.
6. Saide, P. E. et al. Forecasting urban PM_{10} and $PM_{2.5}$ pollution episodes in very stable nocturnal conditions and complex terrain using WRF–Chem CO tracer model. *Environ. Int.* **2011**, *45*, 2769–2780.
7. Hùng, Đ.V.; và cs. Ứng dụng mô hình AERMOD mô phỏng sự lan truyền các chất ô nhiễm không khí từ khu công nghiệp Phú Tài tỉnh Bình Định. *Tạp chí khí tượng thủy văn* **2024**, *758*, 72–86.
8. Khuê, V.H.N.; và cs. Tính toán phát thải khí thải và ứng dụng hệ mô hình TAPM-AERMOD mô phỏng ô nhiễm không khí từ hệ thống bến cảng tại Thành phố Hồ Chí Minh. *Chuyên san Khoa học trái đất và môi trường* **2018**, *2*, 97–106.
9. Long, B.T.; và cs. Mô hình hóa ô nhiễm không khí trong điều kiện địa hình phức tạp – Trường hợp nguồn thải điểm. *Tạp chí Khí tượng Thủy văn* **2019**, *4*, 34–45.
10. Pandey, G. et al. Evaluating AERMOD with measurements from a major U.S. airport located on a shoreline. *Atmos. Environ.* **2023**, *294*, 119506.
11. Bang, H.Q.; et al. Air pollution emission inventory and air quality modeling for Can Tho City, Mekong Delta, Vietnam. *Air Qual. Atmos. Hlth.* **2017**, *11*, 35–47.
12. Dung, H.M.; et al. Study on load–carrying capacity zoning in atmospheric environment in developing countries – a case study of Can Tho City, Vietnam. *Int. J. Environ. Sci. Dev.* **2021**, *12(7)*, 193–203.
13. Liao, Q.; et al. Deep learning for air quality forecasts: A review. *Curr. Pollut. Rep.* **2020**, *60*, 399–409.
14. Kumar, U.; et al. ARIMA forecasting of ambient air pollutants (O_3, NO, NO_2 and CO). *Stoch Env. Res. Risk A.* **2010**, *24*, 751–760.
15. Vlachogianni, A.; et al. Evaluation of a multiple regression model for the forecasting of the concentrations of NOx and PM_{10} in Athens and Helsinki. *Sci. Total Environ.* **2011**, *409*, 1559–1571.
16. Seng, D.; et al. Spatiotemporal prediction of air quality based on LSTM neural network. *Alex. Eng. J.* **2021**, *60*, 2021–2032.
17. Yan, R.; et al. Multi–hour and multi–site air quality index forecasting in Beijing using CNN, LSTM, CNN–LSTM, and spatiotemporal clustering. *Expert Syst. Appl.* **2021**, *169*, 114513.
18. Jiao, Y.; et al. Prediction of air quality index based on LSTM. Proceeding of the IEEE 8th Joint International Information Technology and Artificial Intelligence Conference **2019**.

19. Belavadi, S.V. Air quality forecasting using LSTM RNN and wireless sensor networks. Proceeding of the 11th International Conference on Ambient Systems. Poland, Elsevier B.V. 2020.
20. Duan, J.; et al. Air–quality prediction based on the ARIMA–CNN–LSTM combination model optimized by Dung Beetle Optimizer. *Sci. Rep.* **2023**, *13*, 12127.
21. Yammahi, A.A.; et al. Forecasting the concentration of NO_2 using statistical and machine learning methods: A case study in the UAE. *Heliyon* **2023**, *9*, 12584.
22. Navares, R.; et al. Predicting air quality with deep learning LSTM: Towards comprehensive models.*Ecol. Inform.* **2019**, *55*, 101019.
23. Chang, Y.S.; et al. An LSTM–based aggregated model for air pollution forecasting. *Atmos. Pollut. Res.* **2020**, *11*, 1451–1463.
24. Wen, C.; et al. A novel spatiotemporal convolutional long short–term neural network for air pollution prediction. *Sci. Total Environ.* **2019**, *654*, 1091–1099.
25. Jung, Y.; et al. Concentration separation prediction model to enhance prediction accuracy of particulate matter. *J. Inf. Commun. Technol.* **2023**, *22(1)*, 77–96.
26. Rakholia, R.; et al. AI–based air quality $PM_{2.5}$ forecasting models for developing countries: A case study of Ho Chi Minh City, Vietnam. *Urban Clim.* **2022**, *46*, 1–13.
27. Hưng, M.D. Nghiên cứu ứng dụng trí tuệ nhân tạo trong dự báo chất lượng không khí. Luận án tiến sĩ, Đại học Bách khoa Hà Nội, 2020.
28. Hà, D.T.; và cs. Ứng dụng mô hình đa biến bộ nhớ dài hạn - ngắn hạn trong dự báo nhiệt độ và lượng mưa, *Tạp chí Khoa học Trường ĐH Cần Thơ* **2022**, *58*, 8–16.
29. Thành, N.C.; Giang, N.T. Xây dựng mô hình máy học LSTM (Long Short-Term Memory) phục vụ công tác dự báo mặn tại trạm đo mặn Đại Ngãi. *Tạp chí Khí tượng Thủy văn* **2022**, *740(1)*, 98–104.
30. Nhung, C.T.H.; và cs. Xác định luật phân bố xác suất của dữ liệu chất lượng không khí được quan trắc tại Hà Nội. *Tạp chí Khoa học và Công nghệ Việt Nam* **2012**, *50*, 83–89.
31. Gulia, S.; et al. Extreme events of reactive ambient air pollutants and their distribution pattern at urban hotspots. *Aerosol Air Qual. Res.* **2017**, *17*, 394–405.
32. Sharma, S.; et al. Hybrid modelling approach for effective simulation of reactive pollutants like Ozone. *Atmos. Environ.* **2013**, *80*, 408–414.
33. Babu, N.; et al. A moving-average-filter-based hybrid ARIMA-ANN model for forecasting time series data. *Appl. Soft Comput.* **2014**, *23*, 27–38.
34. Qing, P.; et al. Single-well yield prediction based on LSTM and MA Combination model. *Smart Innovation Syst. Technol.* **2021**, *218*, 143–152.
35. Babu, C.N.; et al. A moving-average-filter-based hybrid ARIMA-ANN model for forecasting time series data. *Appl. Soft Comput.* **2014**, *3*, 27–38.
36. Johnston, F.R.; et al. Some properties of a simple moving average when applied to forecasting a time series. *J. Oper. Res. Soc.* **1999**, *50*, 1267–1271.
37. Tổng Cục Môi Trường. Hướng dẫn kỹ thuật tính toán và công bố chỉ số chất lượng không khí Việt Nam (VN_AQI), **2019**.
38. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9(8)*, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.
39. Mikolov, T.; et al. Learning longer memory in recurrent neural networks. *ArXiv* **2015**, 1–9.
40. Shah, J.; et al. Analytical equations based prediction approach for $PM_{2.5}$ using artifcial neural network. *Appl. Sci.* **2020**, *2*, 1516.
41. Kumar, K.; et al. Air pollution prediction with machine learning: A case study of Indian cities. *Int. J. Environ. Sci. Technol.* **2022**, *20*, 5333–5348.
42. Yan, J.; et al. Water quality prediction in the Luan River based on 1–DRCNN and BiGRU hybrid neural network model. *Water* **2021**, *13*, 1273.

Forecasting air quality by the LSTM-MA model, using data at the Gieng Nuoc intersection automatic monitoring station, Ba Ria - Vung Tau province

Ho Minh Dung[1]*, Khong Doan An Khang[1]

[1] Institute for Environment and Resources, VNU-HCMC; H_minhdung@yahoo.com; ankhang28040506@gmail.com

Abstract: Air pollution is one of the causes that increases the risk of respiratory and cardiovascular diseases. Forecasting air quality developments helps warn the community about pollution levels. This study applies artificial intelligence to forecast air quality at the Gieng Nuoc intersection automatic monitoring station area, Ba Ria - Vung Tau province. The Long Short Memory Model (LSTM) was selected for the study and to optimize the prediction ability, a Moving Average (MA) filter was used. Research results show that the air quality in the study area is relatively good when the concentrations of CO, NO_2, SO_2, PM_{10} and $PM_{2.5}$ are all below the allowed threshold. Ozone is the parameter with the highest number of times exceeding the allowable and is also the parameter that has the main impact on the air quality index; The LSTM-MA model has been successfully built with the highest forecast accuracy for the next 1 day with a root mean square error (RMSE) value of 3.05; the mean absolute error (MAE) was 2.17 and the mean absolute percentage error (MAPE) was 3.19%. When forecasting for a longer period of the next 2 weeks, the model shows positive results with the RMSE, MAE and MAPE are 22.79; 15.74 and 24.38% respectively.

Keywords: LSTM-MA; Ba Ria - Vung Tau; Forecast; Air quality.