

Dự báo chất lượng không khí bằng mô hình LSTM-MA: Nghiên cứu điển hình tại tỉnh Bà Rịa – Vũng Tàu, Việt Nam

Hồ Minh Dũng^{1,*}, Nguyễn Hiền Thân²



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

TÓM TẮT

Ô nhiễm không khí góp phần đáng kể tác động tiêu cực đến các vấn đề sức khỏe cộng đồng, làm trầm trọng thêm tỷ lệ mắc các bệnh về hô hấp và tim mạch. Dự báo kịp thời sự biến động chất lượng không khí là rất quan trọng để nâng cao nhận thức cộng đồng và quản lý ô nhiễm hiệu quả trong môi trường đô thị. Nghiên cứu này áp dụng công cụ trí tuệ nhân tạo để dự báo chất lượng không khí tại tỉnh Bà Rịa - Vũng Tàu, Việt Nam. Mô hình LSTM (Long Short-Term Memory) được sử dụng và được tăng cường bằng bộ lọc Trung bình trượt (MA) để giảm thiểu nhiễu dữ liệu. Phân tích thống kê cho thấy chất lượng không khí trong khu vực nghiên cứu vẫn ở mức tương đối chấp nhận được, với nồng độ CO, NO₂, SO₂, PM₁₀ và PM_{2.5} đều dưới mức cho phép. Nồng độ ozon thường xuyên vượt ngưỡng cho phép, gây ảnh hưởng đáng kể đến chỉ số chất lượng không khí (AQI). Mô hình LSTM-MA được phát triển đã chứng minh độ chính xác dự báo vượt trội trong khoảng thời gian 7 ngày, đạt sai số bình phương trung bình (RMSE) từ 0,157 đến 0,271 và sai số phần trăm tuyệt đối trung bình (MAPE) từ 30,0 đến 31,0. Đối với các dự báo mở rộng kéo dài từ hai tuần đến một tháng, mô hình duy trì hiệu suất khả quan, với các giá trị RMSE và MAPE lần lượt nằm trong khoảng 0,186 - 0,497 và 46,3 - 84,4. Ngoài ra, các đánh giá so sánh đã được thực hiện với các mô hình dự báo thay thế khác, bao gồm LSTM và ARIMA.

Từ khoá: Dự báo, LSTM-MA, mô hình lai, chất lượng không khí, Bà Rịa-Vũng Tàu

¹Viện Môi trường và Tài nguyên, ĐHQG-HCM, P. Linh Xuân, TP.HCM, Việt Nam

²Trung tâm tư vấn trung hòa Carbon, Trường ĐH Thủ Dầu Một, P. Phú Lợi, TP.HCM, Việt Nam

Liên hệ

Hồ Minh Dũng, Viện Môi trường và Tài nguyên, ĐHQG-HCM, P. Linh Xuân, TP.HCM, Việt Nam

Email: H_minhdung@yahoo.com

Lịch sử

- Ngày nhận: 24-12-2025
- Ngày sửa đổi: 06-01-2026
- Ngày chấp nhận: 31-07-2026
- Ngày đăng: 21-10-2026

DOI : <https://doi.org/10.32508/vnuhcmj-es.v10i2.871>



Bản quyền

© Tạp chí ĐHQG-HCM. Đây là bài báo công bố mở được phát hành theo các điều khoản của the Creative Commons Attribution 4.0 International license.

MỞ ĐẦU

Đô thị hóa và công nghiệp hóa nhanh chóng là những yếu tố chính gây ra ô nhiễm không khí. Ô nhiễm không khí là một trong những nguyên nhân gây ra nhiều vấn đề về sức khỏe. Do đó, việc dự đoán những thay đổi về chất lượng không khí sớm sẽ giúp cảnh báo cộng đồng về mức độ ô nhiễm, một nhiệm vụ quan trọng trong việc quản lý chất lượng không khí ở môi trường đô thị. Hiện nay, các mô hình dự báo chất lượng không khí có thể được chia thành hai phương pháp: mô hình truyền thống và mô hình thống kê¹: Các mô hình truyền thống dựa trên các quá trình vật lý và hóa học trong khí quyển, kết hợp với các yếu tố khí tượng, phát thải chất gây ô nhiễm và các công cụ toán học để mô phỏng chất lượng không khí ở nhiều quy mô khác nhau, chẳng hạn như CMAQ^{2,3}, WRF-Chem⁴, AERMOD⁵ và TAPOM^{6,7}, NAQPMS, v.v. Ngược lại, các mô hình thống kê ít tập trung vào các quá trình vật lý và hóa học cơ bản gây ra ô nhiễm không khí, mà chủ yếu nhấn mạnh mối quan hệ giữa dữ liệu đầu vào, chẳng hạn như các yếu tố khí tượng và mức độ ô nhiễm trong quá khứ, và nồng độ ô nhiễm trong tương lai⁸. Một số mô hình thống kê để dự báo chất lượng không khí bao gồm: AR, ARMA, ARIMA⁹, GWR¹⁰, MLR¹¹, SVR¹², và các mô hình

lai¹³.

Ứng dụng trí tuệ nhân tạo (AI) trong dự báo chất lượng không khí đã được thực hiện rộng rãi trong các mô hình thống kê. Trong những năm gần đây, nghiên cứu về dự báo chất lượng không khí dựa trên mạng nơ-ron đã có những bước tiến đáng kể. Các nhà nghiên cứu đã chứng minh rằng mạng nơ-ron nhân tạo (ANN) có thể đưa ra dự đoán tốt hơn so với các mô hình truyền thống. Theo nghiên cứu¹⁴, nhiều nghiên cứu đã áp dụng học máy (ML) và học sâu (DL) để phát triển các mô hình dự báo chất lượng không khí. Mạng nơ-ron hồi quy bộ nhớ dài hạn ngắn hạn (LSTM) là một loại mạng nơ-ron hồi quy (RNN) và là một thuật toán học sâu. Trong những năm gần đây, LSTM đã trở thành một mô hình hiệu quả để học dữ liệu tuần tự. Nghiên cứu gần đây đã chỉ ra rằng LSTM có các đặc tính thống kê tốt hơn so với các mạng RNN truyền thống. Nghiên cứu cụ thể cho thấy LSTM là một thuật toán chính xác hơn khi so sánh với nhiều thuật toán học sâu và hồi quy khác nhau¹⁵⁻¹⁷. Do đó, phương pháp LSTM cho phép chúng ta dự đoán xu hướng chất lượng không khí tốt hơn. Mô hình LSTM có thể nắm bắt được sự phân bố các thay đổi về chất lượng không khí theo thời gian.

Tất cả các kỹ thuật dự đoán chất lượng không khí nêu trên đều dựa trên các mô hình riêng lẻ, và độ chính xác

Trích dẫn bài báo này: Dũng H M, Thân N H. Dự báo chất lượng không khí bằng mô hình LSTM-MA: Nghiên cứu điển hình tại tỉnh Bà Rịa – Vũng Tàu, Việt Nam. *VNUHCM J. Environ. Earth Sci.* 2026;10(2): 1485-1498.

dự đoán giảm là do các yếu tố như đặc trưng không gian, quy mô mô hình và lựa chọn tham số. Để bù đắp cho những hạn chế của một mô hình đơn lẻ, các mô hình dự đoán lai đã trở thành xu hướng trong các nghiên cứu mới. Mô hình lai được tạo ra bằng cách kết hợp các kỹ thuật với các mô hình dự đoán khác và bằng cách tổng hợp các giá trị dự đoán của tất cả các chuỗi con để có được dự đoán cuối cùng. Một số nghiên cứu cho thấy rằng các mô hình lai thường hoạt động tốt hơn các mô hình đơn lẻ^{14,18}. Các nghiên cứu cho thấy mô hình LSTM lai cho kết quả tốt hơn so với mô hình LSTM đơn lẻ hoặc mô hình dự báo truyền thống. Do đó, nhiều nhà nghiên cứu cũng đã sử dụng mô hình LSTM lai để dự đoán chất lượng không khí ở các quốc gia khác nhau hoặc các địa phương trong cùng một quốc gia (Bảng 1).

Chỉ số chất lượng không khí (AQI) là thước đo ô nhiễm không khí dựa trên nồng độ của các chất ô nhiễm khác nhau trong khí quyển. Các chất ô nhiễm chính ảnh hưởng đến AQI là NO_2 , SO_2 , CO, $\text{PM}_{2.5}$, PM_{10} và O_3 . AQI nhằm mục đích giáo dục cộng đồng về tác động của ô nhiễm không khí đến sức khỏe. Do đó, việc giám sát và dự báo AQI đã trở thành một công cụ quan trọng cho sự phát triển bền vững của quốc gia. Các nghiên cứu dự báo chất lượng không khí bằng trí tuệ nhân tạo dựa trên chỉ số AQI ít phổ biến hơn so với các nghiên cứu dựa trên từng chất ô nhiễm riêng lẻ.

Tại Việt Nam, số lượng trạm quan trắc không khí xung quanh tự động vẫn còn thấp³². Do đó, bên cạnh việc quan trắc chất lượng không khí hiện tại, cần phải sử dụng các công cụ để dự báo chất lượng không khí địa phương. Bên cạnh đó, các nghiên cứu về ứng dụng trí tuệ nhân tạo trong dự báo chất lượng không khí tại Việt Nam chủ yếu tập trung vào các chất ô nhiễm không khí đơn lẻ như PM_{10} và $\text{PM}_{2.5}$, v.v.^{33,34}. Hiện nay, có rất ít nghiên cứu về dự báo chất lượng không khí dựa trên chỉ số AQI. Ngoài ra, cũng có ít nghiên cứu trước đây sử dụng mô hình lai LSTM-MA để dự báo chất lượng không khí. Nghiên cứu này đã thực hiện xây dựng và đánh giá mô hình lai LSTM-MA cho bài toán dự báo chất lượng không khí dựa trên dữ liệu chỉ số AQI (từ các trạm quan trắc chất lượng không khí tự động tại tỉnh Bà Rịa - Vũng Tàu). Mô hình hướng tới cải thiện độ ổn định dự báo ngắn hạn thông qua việc giảm nhiễu đầu vào.

DỮ LIỆU VÀ PHƯƠNG PHÁP NGHIÊN CỨU

Dữ liệu nghiên cứu

Bộ dữ liệu quan trắc chất lượng không khí được sử dụng trong nghiên cứu này là nồng độ trung bình giờ

của $\text{PM}_{2.5}$, PM_{10} , TSP, CO, NO_2 , O_3 và SO_2 được đo đạc tại các trạm quan trắc không khí tự động tại tỉnh Bà Rịa-Vũng Tàu (Bảng 2). Bộ dữ liệu bao gồm giai đoạn từ ngày 18 tháng 1 năm 2020 đến ngày 31 tháng 12 năm 2022. Dữ liệu thu thập tại các trạm KK1, KK2 và KK3 (Hình 1) bao gồm 25.799-25.878 giá trị cho mỗi thông số. Lượng dữ liệu trống (không có dữ liệu) nhìn chung không nhiều (từ 0,1% - 3,7% đối với dữ liệu tại KK1; từ 0,1% - 7,8% đối với dữ liệu tại KK2; từ 0,1% - 4,8% đối với dữ liệu tại KK3). Dữ liệu trống trong tập dữ liệu thu thập được có thể do nhiều yếu tố khách quan liên quan đến nguồn điện, thiết bị đo lường, thiết bị lưu trữ dữ liệu, v.v. Tuy nhiên, tỷ lệ dữ liệu trống nêu trên tương đối nhỏ so với tổng tập dữ liệu thu thập được. Kết quả thống kê cho thấy vẫn còn các giá trị bằng không, và giá trị lớn nhất của hầu hết các tham số quan sát được đều cách xa giá trị trung vị và giá trị trung bình. Điều này cho thấy tập dữ liệu chứa các giá trị ngoại lai (giá trị bất thường), vì vậy chúng ta nên loại bỏ chúng để thu được thông tin có ý nghĩa. Sau khi loại bỏ các giá trị này, chúng ta sử dụng kỹ thuật tái tạo dữ liệu thiếu để điền vào các giá trị thiếu.

Phương pháp nghiên cứu

Tiền xử lý số liệu

Dữ liệu giám sát chất lượng không khí tự động thường biến động mạnh và bị ảnh hưởng bởi nhiều yếu tố tự nhiên (khí tượng) và nhân tạo (thiết bị, v.v.). Trước tiên, dữ liệu giám sát được kiểm tra dữ liệu khuyết bằng phương pháp phân tích dữ liệu khuyết trong phần mềm SPSS (dữ liệu thiếu). Dữ liệu thiếu có thể được phân loại thành ba loại: hoàn toàn ngẫu nhiên, ngẫu nhiên hoặc không ngẫu nhiên. Dữ liệu từ hệ thống quan trắc chất lượng không khí tự động tại tỉnh Bà Rịa-Vũng Tàu bị thiếu do nhiều yếu tố chủ quan, bao gồm đầu dò bị lỗi và quy trình nhập dữ liệu tự động. Nếu không xử lý các giá trị này, chúng sẽ ảnh hưởng đến kết quả đánh giá. Phương pháp phát hiện các giá trị bất thường bằng kỹ thuật phi tham số được thực hiện trong SPSS.

Các giá trị bất thường được phát hiện, và nguyên nhân cũng như quy luật của dữ liệu cũng được xem xét. Các giá trị bất thường có giá trị quá lớn so với dữ liệu trong cùng kỳ, và nguyên nhân không thể xác định được sẽ bị loại bỏ. Có nhiều cách để xử lý các giá trị thiếu và giá trị ngoại lai bằng cách sử dụng phân tích dữ liệu thiếu. Một cách tiếp cận phổ biến là bỏ qua dữ liệu thiếu, điều này làm giảm hiệu quả và gây ra sai lệch trong quá trình mô hình hóa. Cách tiếp cận thứ hai là tái cấu trúc dữ liệu, bao gồm nhiều cơ chế, từ đơn giản (ví dụ: thay thế các giá trị thiếu bằng giá trị gần

Bảng 1: Một số nghiên cứu sử dụng mô hình LSTM trong dự báo chất lượng không khí

Nghiên cứu	Chất ô nhiễm	Quốc gia	Dữ liệu	Mô hình sử dụng
1. Chen Y., 2021 ¹⁰	PM ₁₀	China	2015 - 2017	LSTM
2. Bakar M.A.A., 2022 ¹⁹	PM ₁₀	Malaysia	7/2017 - 6/2019	LSTM, ARIMA
3. Ho C. H., 2022 ²⁰	PM _{2.5}	Korea	2015 - 2021	CMAQ-LSTM
4. Xu S., 2022 ²¹	PM _{2.5} , SO ₂ , NO ₂ , CO, O ₃ , PM ₁₀	China	2016 - 2020	LSTM, CNN-LSTM, CEEMDAN-CNN-LSTM,
5. Spyrou E.D., 2022 ²²	CO	Greece	-	LSTM, ARIMA
6. Tsokov S., 2022 ²³	PM _{2.5}	China	3/2013 - 02/2017	LSTM, CNN, CNN-LSTN
7. Duan J., 2023 ²⁴	AQI	China	01/2015 - 3/2022	LSTM, ARIMA-CNN-LSTM, CEEMDAN-CNN-LSTM, CEEMDAN-LSTM
8. Li X., 2017 ¹	PM _{2.5}	China	01/2014 - 5/2016	LSTME
9. Seng D., 2021 ¹²	PM _{2.5} , SO ₂ , NO ₂ , CO, O ₃	China	01/2016 - 12/2017	LSTM, MMSL
10. Drewil G.I., 2022 ²⁵	PM _{2.5} , PM ₁₀ , CO, NOx	India	2017 - 2020	LSTM, GA-LSTM
11. Chen H., 2021 ²⁶	PM _{2.5} , PM ₁₀ , NO ₂ , CO, O ₃ and SO ₂	China	2013 - 2018	LSTM, Seq2Seq technology, XG-Boosting tree
12. Raheja S., 2022 ²⁷	PM _{2.5} , PM ₁₀ , CO, O ₃ , NO ₂ , SO ₂	India	2019	LSTM, A-LSTM
13. Li T., 2020 ²⁸	PM _{2.5}	China	01/2015-12/2027	CNN-LSTM
14. Pak U., 2018 ²⁹	O ₃	China	-	CNN-LSTM
15. Qi Y., 2019 ³⁰	PM _{2.5}	China	01/2015-04/2016	GC-LSTM
16. Fan S., 2021 ³¹	SO ₂ , NO ₂ , CO, O ₃ , PM ₁₀ , PM _{2.5} , AQI	China	05/2018-08/2010	LSTM, ARMA, CEEMDAN-LSTM, WD-LSTM

[Nguồn: Nhóm tác giả]

Bảng 2: Vị trí các trạm quan trắc không khí xung quanh tự động ở tỉnh Bà Rịa – Vũng tàu

Tên trạm	Tọa độ	Sơ đồ vị trí
1. Trạm ngã tư Giếng nước (KK1)	10°22'04.7"N 107°05'03.6"E	
2. Trạm ngã tư đường Phan Đăng Lưu – đường 27/4 (KK2)	10°30'08.7"N 107°10'09.7"E	
3. Trạm khu xử lý chất thải Tóc tiên (KK3)	10°35'23.1"N 107°07'54.2"E	

[Nguồn: Nhóm tác giả]

Bảng 3: Thống kê mô tả dữ liệu sau khi xử lý sơ bộ dữ liệu

Thông số	N	Trung bình	Độ lệch chuẩn	Khuyết		Số lượng trị bất thường	
				Count	Phần trăm	Thấp	Cao
NO ₂	77,509	15.84	12.41	0	0	0	2.898
O ₃	77,509	46.24	35.46	0	0	0	2.441
SO ₂	77,509	11.81	18.24	0	0	0	4.583
CO	77,509	580.65	496.04	0	0	0	3.153
PM ₁₀	77,509	29.04	20.36	0	0	0	2.776
PM _{2.5}	77,509	16.23	12.25	0	0	0	2.469

[Nguồn: Nhóm tác giả]

nhất, trung bình và trung vị) đến phức tạp hơn (ví dụ: chuỗi Markov Monte Carlo, NIPALS). Trong nghiên cứu này, tác giả đã sử dụng phương pháp tái cấu trúc thông tin thiếu dựa trên hồi quy đa biến và phương pháp bình phương tối thiểu từng phần lặp phi tuyến tính. Kết quả sau khi xử lý dữ liệu được thể hiện trong Bảng 3.

Tính toán chỉ số chất lượng không khí (AQI)

Chỉ số chất lượng không khí (AQI) mô tả định lượng chất lượng không khí. Các chất ô nhiễm cấu thành AQI bao gồm PM₁₀, PM_{2.5}, SO₂, NO₂, CO và O₃. Phương pháp tính chỉ số chất lượng không khí Việt Nam (VN_AQI) từ dữ liệu giám sát của các trạm giám sát chất lượng không khí tự động được ban hành tại Quyết định số 1459/QĐ-TCMT ngày 12 tháng 11 năm 2019. Chỉ số AQI được tính bằng hai giá trị: AQI giờ (AQI^h) và AQI ngày (AQI^d). Chỉ số VN_AQI được chia thành sáu mức: Tốt (0-50), Trung bình (51-100), Kém (101-150), Xấu (151-200), Rất xấu (201-300) và Nguy hại (301-500)³⁵.

Phương pháp trung bình trượt

Mô hình trung bình trượt (moving average – MA) dùng để điều chỉnh các mức độ trong một dãy số có biến động tăng giảm thất thường, nhằm loại bỏ ảnh hưởng của các nhân tố ngẫu nhiên, vạch rõ xu thế phát triển cơ bản của dữ liệu^{1,10}. Quá trình $\{X_t, t \in Z\}$ được gọi là quá trình trung bình trượt cấp q, ký hiệu $X_t \sim MA(q)$ thỏa mãn biểu thức sau:

$$X_t = \varepsilon_t + b_1 \varepsilon_{t-1} + \dots + b_q \varepsilon_{t-q} \quad (1)$$

Trong đó, $b_1, b_2, \dots, b_q$ là các hệ số của mô hình, ε_t là nhiễu trắng. Một chuỗi thời gian là nhiễu trắng nếu các biến độc lập và phân phối giống hệt nhau với giá trị trung bình bằng không³⁶.

Mô hình LSTM

Mô hình LSTM là một mạng nơ-ron hồi quy được phát triển bởi Hochreiter và Schmidhuber vào năm 1997. Mikolov đã cải tiến mô hình này vào năm 2019 để nâng cao hiệu suất của nó³⁷. Trái ngược với mô hình mạng nơ-ron hồi quy, mô hình LSTM có dung lượng bộ nhớ lớn hơn và thiết kế phức tạp hơn³⁸. Mô hình LSTM có thể giải quyết các vấn đề phức tạp³⁷. Cấu trúc mạng LSTM được mô tả trong nghiên cứu của Yuan (2019)³⁸. Mô hình LSTM dự báo chất lượng không khí dựa trên dữ liệu từ 3 trạm tại tỉnh Bà Rịa - Vũng Tàu. Như đã đề cập ở trên, sự biến động ngẫu nhiên, các giá trị ngoại lai và nhiễu trong dữ liệu ảnh hưởng đáng kể đến độ chính xác dự báo của mô hình LSTM. Do đó, nghiên cứu này sử dụng bộ lọc trung bình trượt (MA) để làm mịn dữ liệu với các biến động tự nhiên theo thời gian. Johnston đã chứng minh rằng MA sẽ tăng tính ổn định của mô hình dự báo và giảm phương sai lỗi dự báo ít nhất 3% khi sử dụng bộ lọc này³⁹. Trong nghiên cứu này, nhóm tác giả đã kết hợp hai phương pháp, LSTM và MA, để nâng cao hiệu suất dự đoán của mô hình. Cách tiếp cận này được gọi là mô hình lai.

Dữ liệu đầu vào: Chỉ số chất lượng không khí (AQI) được tính toán dựa trên 6 thông số: CO, NO₂, O₃, SO₂, PM_{2.5} và PM₁₀. Giá trị AQI của mỗi trạm quan trắc được sử dụng làm đặc trưng đầu vào cho mô hình LSTM-MA. Dữ liệu được chia thành hai tập dữ liệu: huấn luyện (98-99%) và dự báo (1-2%). Tập dữ liệu huấn luyện cũng được chia thành hai tập dữ liệu: 90% để huấn luyện và 10% để kiểm chứng.

Phương pháp thống kê

Nghiên cứu này sử dụng một số phương pháp thống kê để xác định các đặc điểm của dữ liệu chuỗi thời gian nhằm hỗ trợ việc xây dựng mô hình dự báo chỉ số VN_AQI, bao gồm kiểm định Augmented Dickey-Fuller, tự tương quan riêng phần, tự tương quan và

kiểm định White. Trong phân tích dữ liệu chuỗi thời gian, một mô hình tốt được hình thành khi phân tích được thực hiện trên dữ liệu dừng. Theo Gujarati và Porter (2012), một chuỗi thời gian được gọi là dừng khi giá trị trung bình, phương sai và tự tương quan (ở các độ trễ khác nhau) của nó không đổi bất kể thời điểm quan sát chuỗi⁴⁰. Dữ liệu chất lượng không khí quan sát được đã được kiểm tra tính ổn định bằng cách sử dụng phép thử nghiệm nghiệm đơn vị và phép thử tính ổn định. Các mô hình dự báo chuỗi thời gian sẽ thất bại nếu dữ liệu không ổn định. Tiêu chí để kiểm tra tính ổn định của dữ liệu là thống kê ADF (thống kê T) nhỏ hơn ADF quan sát được (T quan sát được), và giá trị p nhỏ hơn 0,05.

Hàm tự tương quan (ACF) và hàm tự tương quan riêng phần (PACF) được sử dụng để xác định bậc trễ thích hợp của chuỗi thời gian thông qua phân tích mô tả chuỗi thời gian. Phân tích ACF và PACF được sử dụng tạm thời để xây dựng mô hình ARIMA. ACF và PACF có ý nghĩa thống kê khi độ trễ vượt quá giới hạn trên 95%. Để đảm bảo tính hợp lệ và phù hợp của mô hình, phần dư của mô hình phải thể hiện đặc điểm của nhiễu trắng. Do đó, kiểm định White được sử dụng để đánh giá tính không đồng nhất phương sai (phương sai sai số không đổi).

Các giá trị thống kê được sử dụng để đánh giá độ chính xác của mô hình dự báo bao gồm hệ số tương quan R^2 , sai số bình phương trung bình gốc (RMSE) và sai số phần trăm tuyệt đối trung bình (MAPE). Các tham số thống kê này thường được áp dụng trong nghiên cứu dự báo. RMSE, MAPE và R^2 lần lượt biểu thị sai số, sai số phần trăm và mối tương quan giữa các giá trị dự đoán và giá trị thực tế⁴¹⁻⁴⁴. Giá trị RMSE và MAPE nên càng gần 0 càng tốt, và giá trị R^2 nên càng gần 1 càng tốt.

KẾT QUẢ VÀ THẢO LUẬN

Đánh giá chất lượng không khí tại tỉnh Bà Rịa – Vũng tàu

Dựa trên các dữ liệu thu thập được cho thấy giá trị trung bình theo giờ của CO, NO₂ và SO₂ luôn nằm trong giới hạn cho phép. Đối với giá trị trung bình theo giờ của O₃, có 69 giá trị (tại KK1), 124 giá trị (tại KK2) và 58 giá trị (tại KK3) vượt QCVN 05:2023/BT-NMT (Quy chuẩn kỹ thuật quốc gia về chất lượng không khí tại Việt Nam), với giá trị cao nhất xảy ra từ 12 giờ trưa đến 2 giờ chiều. Điều này phù hợp với quy luật ozon tăng dần vào buổi sáng, đạt mức cao nhất vào buổi trưa và sau đó giảm dần vào buổi tối. Cường độ bức xạ mặt trời là nguyên nhân chính gây ra sự dao động nồng độ ozon bề mặt. Tại các trạm quan trắc, nồng độ trung bình 24 giờ của PM₁₀ và PM_{2.5}

nhìn chung nằm trong giới hạn cho phép. Nồng độ PM₁₀ đạt đỉnh điểm từ 7 giờ đến 8 giờ sáng (Hình 1). Chỉ số chất lượng không khí (AQI) cho thấy mối tương quan cao nhất với giá trị O₃ 8 giờ và 1 giờ, lần lượt là 0,95 và 0,81 (tại KK1), 0,97 và 0,80 (tại KK2), và 0,49 và 0,43 (tại KK3). Điều này cũng phù hợp với thực tế là giá trị O₃ 8 giờ và 1 giờ có nhiều lần vượt quá giới hạn cho phép hơn so với các thông số còn lại. Phép tính tương quan cũng cho thấy CO và SO₂ có tác động ít nhất đến AQI (0,18-0,20 đối với CO; 0,08-0,28 đối với SO₂). Ngoài ra, tất cả các giá trị chất ô nhiễm trong tập dữ liệu đều có tương quan dương với AQI^d (Hình 2), nhấn mạnh tầm quan trọng của việc tính đến các yếu tố này khi phân tích và dự đoán sự thay đổi chất lượng không khí trong khu vực nghiên cứu. Chúng ta nên chọn các yếu tố thiết yếu có mối tương quan đáng kể với AQI^d.

Giá trị AQI^d trung bình trong giai đoạn 2020 - 2022 tại ba trạm quan trắc nhìn chung ở mức thấp, như sau: Trạm KK1 có chất lượng không khí ở mức Tốt (0-50) với 79,2%; chất lượng không khí ở mức Trung bình (51-100) với 11,4%; chất lượng không khí ở mức Kém (101-150) với 8,2% và chất lượng không khí ở mức Kém (151-200) với 1,2%; Trạm KK2 có chất lượng không khí ở mức Tốt với 70,0%; ở mức Trung bình với 15,8%; ở mức Kém với 12,3%, ở mức Kém với 1,4%; ở mức Rất xấu (201-300) với 0,5%; Trạm KK3 có chất lượng không khí ở mức Tốt với 71,4%; ở mức Trung bình với 25%; ở mức Kém với 2,9%, ở mức Kém với 0,5% và ở mức Rất xấu với 0,2%. Chỉ số AQI^d thường cao vào mùa khô (đầu và cuối năm, từ tháng 11 đến tháng 4) và thấp hơn vào mùa mưa (giữa năm, từ tháng 5 đến tháng 10). Điều này cho thấy chất lượng không khí vào mùa khô ô nhiễm hơn so với mùa mưa.

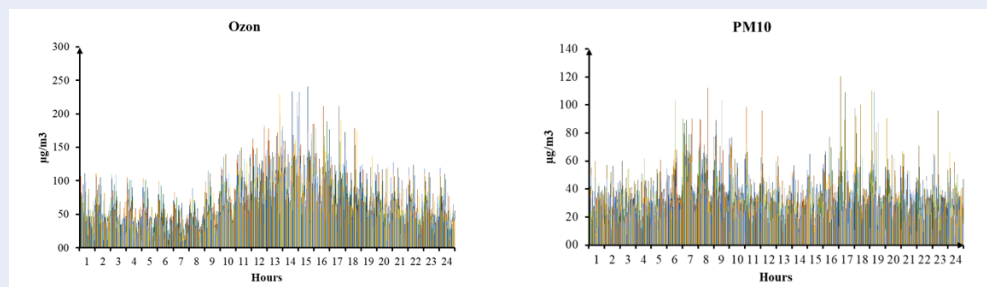
Mô hình dự báo chất lượng không khí

Khảo sát về tính ổn định của dữ liệu

Bảng 4: Kiểm định tính dừng dữ liệu tại các trạm quan trắc

Stt	Trạm	T quan sát	T thống kê	P-Value	Nhận xét
1	KK1	-5,40	-0,91	< 0,0001	Có tính dừng
2	KK2	-5,62	-0,91	< 0,0001	Có tính dừng
3	KK3	-5,45	-0,91	< 0,0001	Có tính dừng

[Nguồn: Nhóm tác giả]



Hình 1: Các giá trị trung bình giờ của O₃ và PM₁₀ tại KK2 [Nguồn: Nhóm tác giả]



Hình 2: Mỗi tương quan giữa AQI của từng chất ô nhiễm và chỉ số AQI^d [Nguồn: Nhóm tác giả]

Phân tích kiểm định Augmented Dickey-Fuller (ADF) được thực hiện tại vị trí KK1 cho thấy giá trị ADF tuyệt đối cao hơn đáng kể là -0,91, so với giá trị ADF quan sát được là -5,40, với giá trị p nhỏ hơn 0,0001. Điều này cho thấy dữ liệu chuỗi AQI tại vị trí KK2 là ổn định. Tương tự, tại vị trí KK2, thống kê ADF là -0,91 vượt quá thống kê ADF quan sát được là -5,62, với giá trị p là 0,0001, cho thấy dữ liệu AQI tại vị trí này cũng ổn định. Kết luận tương tự cũng áp dụng cho trạm quan trắc KK3, cho thấy đặc điểm ổn định (Bảng 4).

Nghiên cứu độ trễ dữ liệu của các trạm quan trắc.

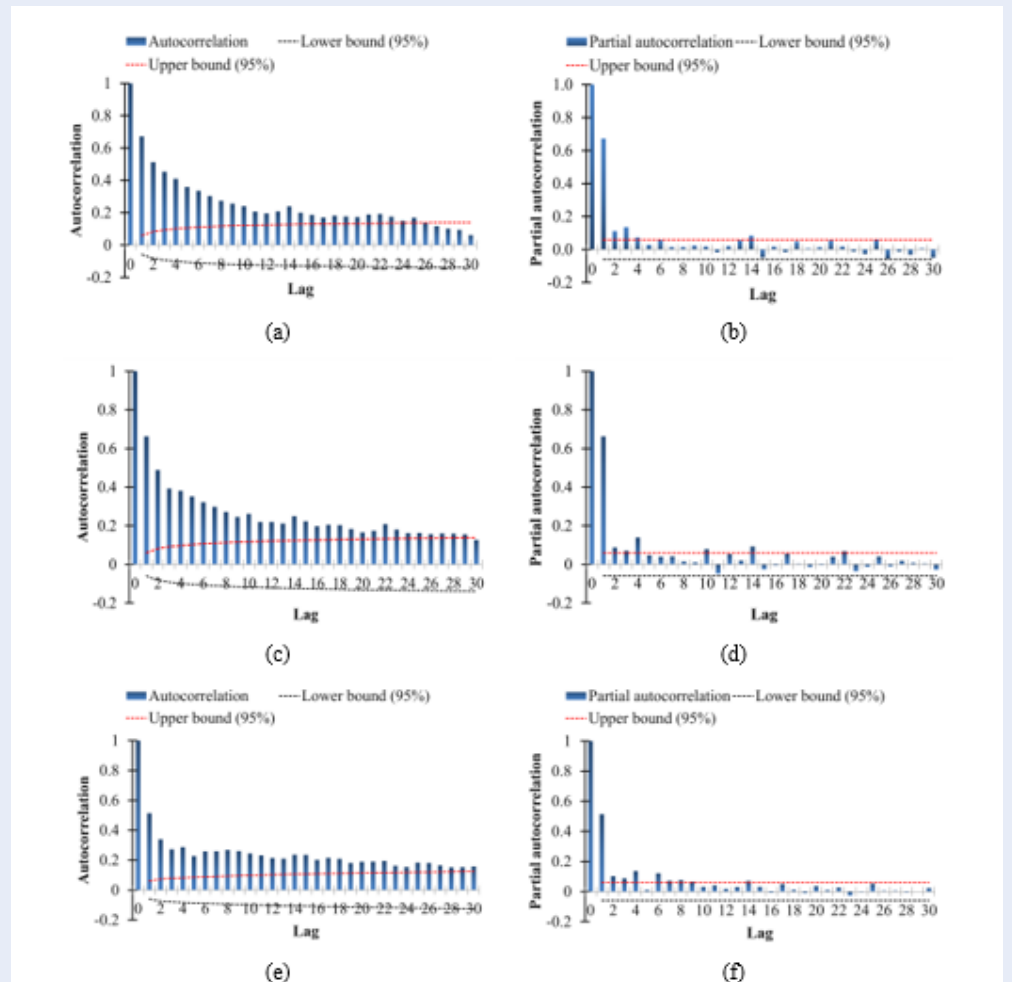
Hơn nữa, việc nghiên cứu hàm tự tương quan (ACF) và hàm tự tương quan riêng phần (PACF) được sử dụng để đánh giá cấu trúc độ trễ của chuỗi dữ liệu AQI cho tỉnh Bà Rịa - Vũng Tàu cho thấy ACF giảm nhanh chóng, như minh họa trong Hình 3. Phân tích chỉ ra rằng ACF có 30 độ trễ đáng kể, trong đó KK3 có độ trễ cao nhất với 30, tiếp theo là KK2 với 29 và KK1 có độ trễ thấp nhất với 26.

Việc kiểm tra kết quả PACF đã xác nhận rằng hệ số tương quan tại các vị trí giám sát AQI đạt mức có ý nghĩa thống kê ở các độ trễ 1, 2, 3 và 4. Các mối tương quan thể hiện mô hình sóng sin, với các độ trễ có ý nghĩa được phân bố không đều. Mặc dù KK1 cho thấy PACF có ý nghĩa ở bậc 14 với mối tương quan tương

đối yếu, KK3 ghi nhận các độ trễ có ý nghĩa ở vị trí 6-8 và 14, trong khi phần tử KK2 có ý nghĩa ở các độ trễ 10, 14 và 22. Sau đó, thứ tự thích hợp cho các mô hình Trung bình trượt ($MA(q)$) và ARIMA (p, q) được xác định bằng cách phân tích biểu đồ Hàm tự tương quan (ACF) và Hàm tự tương quan riêng phần (PACF). Việc xác định các tham số p và q được hướng dẫn bởi các đặc điểm hiển thị trong các biểu đồ này. Theo ACF và PACF thu được, hệ số q trong các mô hình MA và ARIMA được xác định nằm trong khoảng từ 1 đến 4.

Kiểm định nhiều trắng và phân phối dữ liệu

Việc áp dụng phương pháp kiểm định White để giám sát chất lượng không khí tại tỉnh Bà Rịa - Vũng Tàu cho thấy chỉ số AQI không phù hợp với đặc điểm của nhiều trạm. Tại điểm KK1, giá trị p từ các phép kiểm định khác nhau, bao gồm Jarque-Bera, Box-Pierce, Ljung-Box và McLeod-Li, đều nhỏ hơn 0,0001, cho thấy sự sai lệch đáng kể so với phân bố chuẩn. Phát hiện này được chứng thực bởi các quan sát tại các trạm KK3 và KK2, như được trình bày chi tiết trong Bảng 5. Do đó, có thể kết luận rằng dữ liệu chỉ số AQI đầu vào đáp ứng các tiêu chí để được phân loại là dữ liệu chuỗi thời gian.



Hình 3: Hàm tự tương quan toàn phần (ACF) và hàm tự tương quan riêng phần (PACF): (a và b tại trạm KK1; c và d tại trạm KK3; e và f tại trạm KK2) [Nguồn: Nhóm tác giả]

Huấn luyện mô hình LSTM-MA

MÔ HÌNH TRUNG BÌNH TRƯỢT

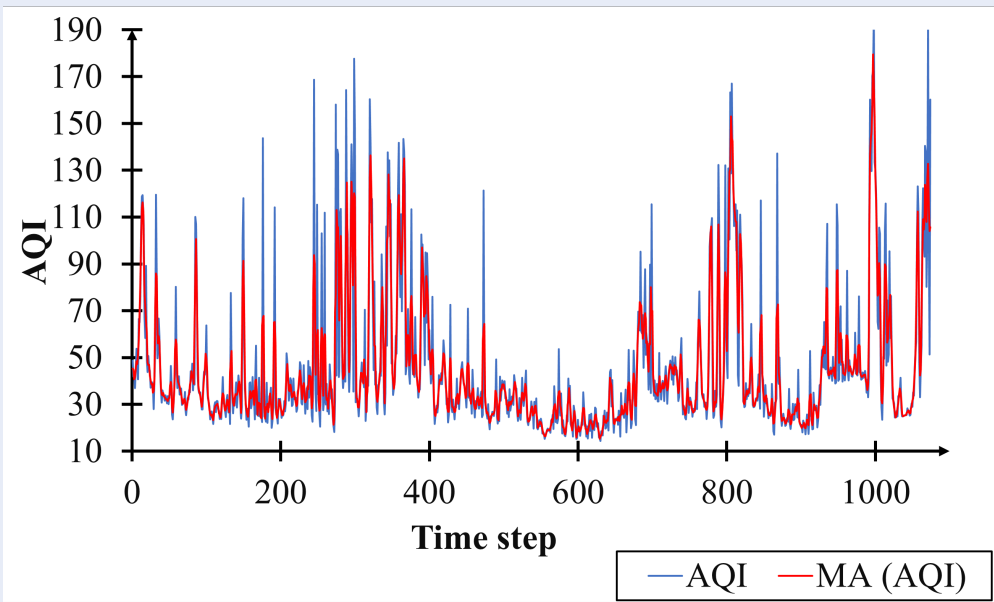
Dữ liệu chất lượng không khí là dữ liệu có tính biến động cao. Theo Haiyan, dữ liệu môi trường thường không ổn định⁴⁵. Do đó, dữ liệu chỉ số AQI^d được xử lý để loại bỏ nhiễu và làm mịn dữ liệu. (Bảng 6, Hình 4-6).

HUẤN LUYỆN MÔ HÌNH LSTM-MA

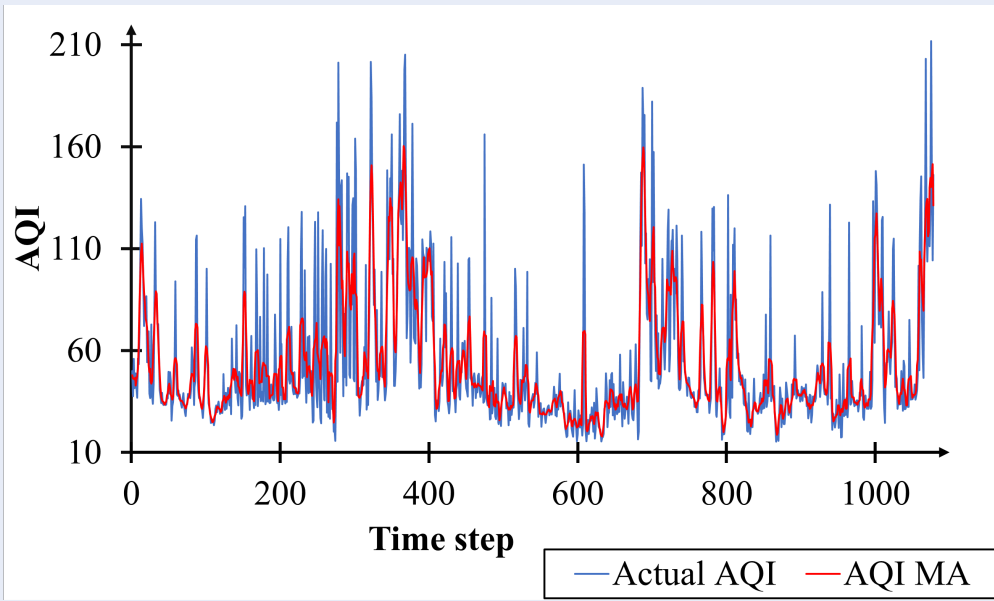
Xác định số lượng nơ-ron tối ưu trong lớp ẩn
Nghiên cứu này sử dụng dữ liệu quan trắc từ trạm KK3 để xác định số lượng nơ-ron lớp ẩn tối ưu trong mạng nơ-ron. Nghiên cứu đã xem xét nhiều số lượng nơ-ron khác nhau, bắt đầu từ 10 và tăng dần cho đến khi mô hình xác định được giá trị tối ưu. Các tham số sau đây đã được thiết lập cho mô hình huấn luyện mạng: thuật toán tối ưu hóa Adam được sử dụng, với

ngưỡng biến đổi gradient là 1, tốc độ học ban đầu là 0,01 và tối đa 200 lần lặp.

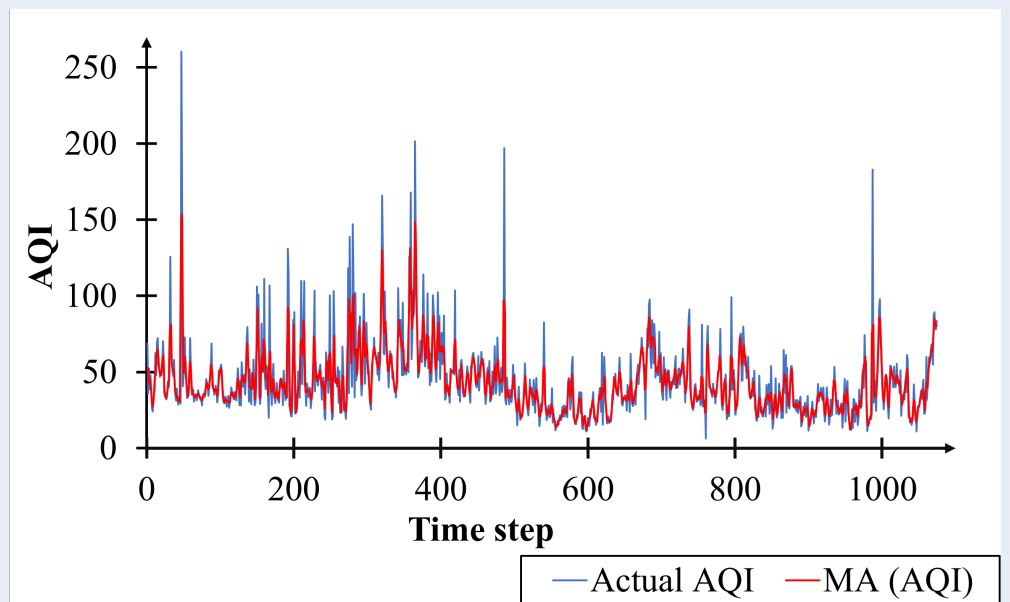
Nghiên cứu bao gồm việc huấn luyện mạng nơ-ron với các tham số mô hình chuẩn trong khi thay đổi số lượng nơ-ron lớp ẩn. Phân tích cho thấy cấu hình tối ưu cho LSTM với mô hình MA là 40 nơ-ron lớp ẩn. Kiến trúc này thể hiện các chỉ số hiệu suất vượt trội trong quá trình huấn luyện mô hình. Cụ thể, RMSE cho tập dữ liệu đánh giá và dự báo lần lượt là 0,0256 và 0,038. Các giá trị này thể hiện tỷ lệ lỗi thấp nhất được quan sát thấy trong số các mô hình được đánh giá. Do đó, các tham số huấn luyện được tinh chỉnh cho mô hình LSTM-MA được áp dụng để dự đoán chất lượng không khí tại tỉnh Bà Rịa - Vũng Tàu được thiết lập như sau: 40 nơ-ron trong lớp ẩn, triển khai thuật toán tối ưu hóa Adam, ngưỡng biến đổi gradient được đặt là 1, tốc độ học ban đầu là 0,01 và giới hạn lặp tối đa là 200 (Bảng 7).



Hình 4: Chỉ số AQI thực tế và chỉ số AQI MA tại trạm KK1 [Nguồn: Nhóm tác giả]



Hình 5: Chỉ số AQI thực tế và chỉ số AQI MA tại trạm KK2 [Nguồn: Nhóm tác giả]



Hình 6: Chỉ số AQI thực tế và chỉ số AQI MA tại trạm KK3 [Nguồn: Nhóm tác giả]

Đánh giá hiệu suất dự báo của mô hình LSTM-MA

Mô hình LSTM-MA dự báo tất cả các điểm quan trắc chất lượng không khí bằng cách sử dụng cấu trúc huấn luyện mạng được mô tả ở trên. Mỗi điểm quan trắc có chu kỳ dự báo từ $k = 7$ đến $k = 30$ (Bảng 8).

So sánh hiệu suất dự báo của LSTM-MA với các mô hình khác.

Nghiên cứu đã thực hiện so sánh hiệu suất dự báo của nhiều mô hình khác nhau, bao gồm các phương pháp truyền thống và mô hình mạng LSTM. Để đánh giá độ tin cậy của kết quả, các giá trị thống kê RMSE và MAPE đã được sử dụng. Kết quả cho thấy mô hình LSTM-MA hoạt động tốt hơn các mô hình LSTM và ARIMA tại hầu hết các điểm quan trắc tự động chất lượng không khí.

Tại điểm KK3, giá trị RMSE và MAPE tốt hơn đáng kể so với các mô hình LSTM và ARIMA, lần lượt đạt 0,0950 và 30%. Kết quả tương tự cũng được quan sát thấy tại các điểm KK2 và KK1. Trong số các mô hình được so sánh, mô hình ARIMA có MAPE cao nhất, cho thấy phương pháp dự báo chuỗi thời gian truyền thống kém hiệu quả hơn trong việc dự đoán các chỉ số chất lượng không khí.

Khi so sánh mô hình LSTM-MA và LSTM, mô hình LSTM-MA thể hiện hiệu suất dự báo vượt trội hơn đối với chỉ số chất lượng không khí. Tuy nhiên, sự khác biệt về hiệu suất dự báo giữa hai mô hình không đáng kể (Bảng 9).

Thảo luận

Cấu trúc lai LSTM-MA được tích hợp mô hình LSTM với bộ lọc MA. Vai trò chính của bộ lọc MA là làm mịn dữ liệu đầu vào, giảm thiểu ảnh hưởng của các biến động tự nhiên, các giá trị ngoại lai và các yếu tố nhiễu, những yếu tố thường ảnh hưởng đến độ chính xác dự đoán của mô hình LSTM cơ bản. Điều này cho thấy rằng trong khi bộ lọc MA bổ sung giá trị bằng cách làm mịn dữ liệu đầu vào, thì sức mạnh cốt lõi vẫn nằm ở khả năng xử lý dữ liệu chuỗi thời gian của kiến trúc LSTM. Do đó, mô hình LSTM-MA thể hiện hiệu suất dự đoán vượt trội so với mạng LSTM đơn lẻ. Mô hình ARIMA thể hiện MAPE cao nhất, cho thấy rằng các phương pháp dự báo chuỗi thời gian truyền thống nhìn chung kém hiệu quả hơn trong việc dự đoán các chỉ số chất lượng không khí, vốn chứa các mối quan hệ phi tuyến phức tạp⁴⁶. Hạn chế của ARIMA nằm ở việc nó chủ yếu phân tích thành phần tuyến tính của chuỗi thời gian, trong khi LSTM nắm bắt hiệu quả các mối quan hệ phi tuyến tính và lưu trữ bộ nhớ dài hạn, dẫn đến hiệu suất vượt trội trong môi trường phức tạp¹⁹. Dự báo chất lượng không khí bằng dữ liệu giám sát tự động có đặc điểm cấu trúc dữ liệu phức tạp. Các yếu tố nhiễu có thể dễ dàng ảnh hưởng đến kết quả đo. Mô hình LSTM-MA giúp giảm thiểu ảnh hưởng của các yếu tố chủ quan.

KẾT LUẬN

Nghiên cứu đã đạt được các kết quả sau từ các thống kê, tính toán và thí nghiệm: (i) Đánh giá hiện trạng

Bảng 5: Thống kê các giá trị kiểm định white test

Các giá trị thống kê	DF	Giá trị	p-value
KK1			
Jarque-Bera	2	1.683,8	< 0,0001
Box-Pierce	6	1.427,1	< 0,0001
Ljung-Box	6	1.433,3	< 0,0001
McLeod-Li	6	1.079,0	< 0,0001
Box-Pierce	12	1.825,5	< 0,0001
Ljung-Box	12	1.835,8	< 0,0001
McLeod-Li	12	1.261,5	< 0,0001
KK2			
Jarque-Bera	2	935,1	< 0,0001
Box-Pierce	6	1.298,0	< 0,0001
Ljung-Box	6	1.303,6	< 0,0001
McLeod-Li	6	963,0	< 0,0001
Box-Pierce	12	1.715,7	< 0,0001
Ljung-Box	12	1.725,7	< 0,0001
McLeod-Li	12	1.214,1	< 0,0001
KK3			
Jarque-Bera	2	9.014,8	< 0,0001
Box-Pierce	6	703,6	< 0,0001
Ljung-Box	6	706,6	< 0,0001
McLeod-Li	6	223,6	< 0,0001
Box-Pierce	12	1.098,5	< 0,0001
Ljung-Box	12	1.105,7	< 0,0001
McLeod-Li	12	302,0	< 0,0001

[Nguồn: Nhóm tác giả]

Bảng 6: Mô hình MA tại các trạm quan trắc

Stt	Trạm	q	RMSE	MAPE (%)	R ²	Đánh giá
1	KK1	1	419,0	15,0	0,8	Rất tốt
2	KK2	1	622,0	22,8	0,7	Tốt
3	KK3	1	413,8	17,2	0,7	Tốt

Bảng 7: Nghiên cứu số lượng nơ-ron tối ưu trong lớp ẩn

Stt	Số lượng nơ-ron	RMSE Hiệu chỉnh	RMSE Dự báo
1.	10	0,027	0,043
2.	20	0,028	0,039
3.	30	0,0261	0,038
4.	40	0,0256	0,038
5.	50	0,0275	0,040
6.	60	0,0257	0,040
7.	70	0,0275	0,040

[Nguồn: Nhóm tác giả]

Bảng 8: Hiệu suất dự báo VN_AQI của mô hình LSTM-MA

Trạm	k	RMSE	MAPE (%)
KK1	k = 7	0,256	31,0
	k = 15	0,376	66,6
	k = 30	0,378	84,4
KK2	k = 7	0,271	30,0
	k = 15	0,497	64,0
	k = 30	0,458	74,1
KK3	k = 7	0,157	30,0
	k = 15	0,225	46,3
	k = 30	0,186	61,9

[Nguồn: Nhóm tác giả]

Bảng 9: So sánh hiệu suất dự báo của các mô hình LSTM-MA, LSTM, và ARIMA

Trạm	Mô hình	RMSE	MAPE (%)
KK1	LSTM-MA	0,256	31,0
	LSTM	0,3611	33,2
	ARIMA	0,1216	48,68
KK2	LSTM-MA	0,271	30,0
	LSTM	0,3178	33,4
	ARIMA	0,1337	80,55
KK3	LSTM-MA	0,0950	30,0
	LSTM	0,0997	30,92
	ARIMA	0,0808	40,4

[Nguồn: Nhóm tác giả]

chất lượng không khí tại ba trạm quan trắc chất lượng không khí tự động (KK1-KK3) thuộc tỉnh Bà Rịa-Vũng Tàu trong giai đoạn 2020-2022. Bộ dữ liệu thu thập được đạt mức độ hoàn chỉnh trên 99%. Chất lượng không khí trong khu vực nghiên cứu tương đối tốt, với nồng độ CO, NO₂ và SO₂ dưới mức cho phép trong suốt thời gian nghiên cứu, ngoại trừ PM₁₀ và PM_{2.5}, vượt mức cho phép trong một vài ngày nhưng không đáng kể. Ozon là thông số có số lần vượt mức cho phép cao nhất và là thông số có tác động đáng kể nhất đến chỉ số AQI^d. Chất lượng không khí có xu hướng thay đổi theo mùa, với mức độ thấp vào các tháng mùa hè. Đặc biệt, chất lượng không khí trong khu vực này chủ yếu bị ảnh hưởng bởi chất ô nhiễm Ozon trong khoảng thời gian 1 giờ và 8 giờ; (ii) Mô hình LSTM-MA đã được xây dựng thành công với độ chính xác dự báo cao nhất trong 7 ngày tới, với RMSE từ 0,157 đến 0,271; và MAPE từ 30,0 đến 31,0%. Khi dự báo trong khoảng thời gian dài hơn, tương ứng với 2 tuần hoặc 1 tháng tới, mô hình cho kết quả tích cực, với các chỉ số đánh giá RMSE và MAPE lần lượt nằm trong khoảng từ 0,186 đến 0,497 và từ 46,3 đến 84,4%. Ngoài ra, nghiên cứu này đã so sánh và đánh giá một số mô hình dự báo khác, bao gồm LSTM và ARIMA. Nghiên cứu này đã phát triển thành công sự kết hợp giữa mô hình LSTM và MA để dự báo chất lượng không khí. Điều này được coi là một hướng nghiên cứu mới trong việc ứng dụng các mô hình LSTM lai vào dự báo chất lượng không khí. Tuy nhiên, nghiên cứu vẫn còn một số hạn chế nên cần có các hướng nghiên cứu tiếp theo để hoàn thiện, một số hướng mở rộng nghiên cứu như: Sử dụng dữ liệu phụ trợ làm đầu vào cho mô hình, chẳng hạn như dữ liệu khí tượng và phát triển kinh tế - xã hội, để cải thiện đáng kể hiệu suất dự báo; Ứng dụng các phương pháp xử lý dữ liệu khác, chẳng hạn như CEEMDAN và ARIMA, hoặc kết hợp với các mô hình như CNN-LSTM và LSTM-ARIMA, để cải thiện khả năng dự báo của mô hình.

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.

DANH MỤC CÁC TỪ VIẾT TẮT

AERMOD: The ASM/EPA Regulatory Model.
ANN (Artificial Neural Network): Mạng nơ-ron nhân tạo
AQI (Air Quality Index): Chỉ số chất lượng không khí
AR (Autoregressive model): Mô hình tự hồi qui
ARIMA (Autoregressive Integrated Moving Average): Tự hồi quy tích hợp trung bình trượt

ARMA (Autoregressive Moving Average): mô hình tự hồi quy trung bình trượt
CMAQ: Community Multiscale Air Quality
CNN (Convolutional Neural Network): Mạng nơ-ron tích chập
DL (Deep Learning): Học sâu
GWR (Geographically Weighted Regression): Hồi quy trọng số địa lý
LSTM (Long Short-Term Memory): Mạng nơ-ron hồi quy dài hạn ngắn hạn
MA (Moving Average): Trung bình trượt
MAPE (Mean Absolute Percentage Error): Sai số phần trăm tuyệt đối trung bình
ML (Machine Learning): Học máy
MLR (Multiple Linear Regression): Hồi quy tuyến tính bội
RMSE (Root Mean Squared Error): sai số bình phương trung bình gốc
RNN (Recurrent Neural Network): Mạng nơ-ron hồi quy
TAPOM: Transport and Air Pollution Model
WRF-Chem: The Weather Research and Forecasting Model coupled with Chemistry

XUNG ĐỘT LỢI ÍCH

Tập thể tác giả không có xung đột về lợi ích.

ĐÓNG GÓP CỦA CÁC TÁC GIẢ

Bản thảo này được viết bởi Nguyễn Hiền Thân, được bổ sung nội dung bởi Hồ Minh Dũng và được chỉnh sửa hình ảnh, bảng biểu bởi Hồ Minh Dũng và Nguyễn Hiền Thân.

TÀI LIỆU THAM KHẢO

- Li X, Peng L, Yao X, Cui S, Hu Y, You C, et al. Long short-term memory neural network for air pollutant concentration predictions: method development and evaluation. *Environmental Pollution*. 2017;231(Pt 1):997–1004. Available from: <https://doi.org/10.1016/j.envpol.2017.08.114>.
- Buonocore JJ, Dong X, Spengler JD, Fu JS, Levy JJ. Using the Community Multiscale Air Quality (CMAQ) model to estimate public health impacts of PM_{2.5} from individual power plants. *Environment International*. 2014;68:200–8. Available from: <https://doi.org/10.1016/j.envint.2014.03.031>.
- Wang L, Wei Z, Wei W, Fu JS, Meng C, Ma S. Source apportionment of PM_{2.5} in top polluted cities in Hebei, China using the CMAQ model. *Atmospheric Environment*. 2015;122:723–36. Available from: <https://doi.org/10.1016/j.atmosenv.2015.10.041>.
- Saide PE, Carmichael GR, Spak SN, Gallardo L, Osses AE, Mena-Carrasco MA, et al. Forecasting urban PM₁₀ and PM_{2.5} pollution episodes in very stable nocturnal conditions and complex terrain using WRF-Chem tracer model. *Atmospheric Environment*. 2011;45(16):2769–80. Available from: <https://doi.org/10.1016/j.atmosenv.2011.02.001>.
- Pandey G, Venkatram A, Arunachalam S. Evaluating AERMOD with measurements from a major U.S. airport located on a shoreline. *Atmospheric Environment*. 2023;294. Available from: <https://doi.org/10.1016/j.atmosenv.2022.119506>.

6. Bang HQ, Khue VH, Tam NT, Lasko K. Air pollution emission inventory and air quality modeling for Can Tho City, Mekong Delta, Vietnam. *Air Quality, Atmosphere & Health*. 2017;11(1):35–47. Available from: <https://doi.org/10.1007/s11869-017-0512-x>.
7. Dung HM, Khue VH. Study on Load-Carrying Capacity Zoning in Atmospheric Environment in Developing Countries - A Case Study of Can Tho City, Vietnam. *International Journal of Environmental Sciences and Development*. 2021;12(7):193–203. Available from: <https://doi.org/10.18178/ijesd.2021.12.7.1340>.
8. Liao Q, Zhu M, Wu L, Pan X, Tang X, Wang Z. Deep Learning for Air Quality Forecasts: a Review. *Current Pollution Reports*. 2020;6(4):399–409. Available from: <https://doi.org/10.1007/s40726-020-00159-z>.
9. Kumar U, Jain VK. ARIMA forecasting of ambient air pollutants (O₃, NO, NO₂ and CO). *Stochastic Environmental Research and Risk Assessment*. 2010;24(5):751–60. Available from: <https://doi.org/10.1007/s00477-009-0361-8>.
10. Chen SCY, Chen P, Yuan Q, Kang P, Zhu L. An LSTM-based neural network method of particulate pollution forecast in China. *Environmental Research Letters*. 2021;16(4). Available from: <https://doi.org/10.1088/1748-9326/abelf>.
11. Vlachogianni A, Kassomenos P, Karppinen A, Karakitsios S, Kukkonen J. Evaluation of a multiple regression model for the forecasting of the concentrations of NO_x and PM₁₀ in Athens and Helsinki. *The Science of the Total Environment*. 2011;409(8):1559–71. Available from: <https://doi.org/10.1016/j.scitotenv.2010.12.040>.
12. Seng D, Zhang Q, Zhang X, Chen G, Chen X. Spatiotemporal prediction of air quality based on LSTM neural network. *Alexandria Engineering Journal*. 2021;60(2):2021–32. Available from: <https://doi.org/10.1016/j.aej.2020.12.009>.
13. Zhu S, Lian X, Liu H, Hu J, Wang Y, Che J. Daily air quality index forecasting with hybrid models: A case in China. *Environmental Pollution*. 2017;231(Pt 2):1232–44. Available from: <https://doi.org/10.1016/j.envpol.2017.08.069>.
14. Méndez M, Merayo MG, Núñez M. Machine learning algorithms to forecast air quality: a survey. *Artificial Intelligence Review*. 2023;56(9):1–36. Available from: <https://doi.org/10.1007/s10462-023-10424-4>.
15. Chowdhury MSAS, Tanjim MR. Fariha Noor, Rashedur M. Rahman. Application of Data Mining Techniques on Air Pollution of Dhaka City. In: *Proceedings of 2020 (IEEE) 10th International Conference on Intelligent Systems*; 2020. Available from: <https://doi.org/10.1109/IS48319.2020.9200125>.
16. Zhou MCL, Ni Q. A hybrid Prophet-LSTM Model for Prediction of Air Quality Index. In: *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*; 2020. Available from: <https://doi.org/10.1109/SSCI47803.2020.9308543>.
17. Cheng HCJC, IEEE 4th International Conference on Knowledge Innovation and Invention (ICKII). Air Quality Forecast and Evaluation Based on Long Short-Term Memory Network and Fuzzy Algorithm.; 2021. Available from: <https://doi.org/10.1109/ICKII51822.2021.9574707>.
18. Subramaniam NRS. Abbas Ganesan, Nithyaprakash Rajavel, Maheswari Chenniappan, Chander Prakash, Alokesh Pramanik, Animesh Kumar Basak and Saurav Dixit. Artificial Intelligence Technologies for Forecasting Air Pollution and Human Health: A Narrative Review. *Sustainability (Basel)*. 2022;14(16):9951. Available from: <https://doi.org/10.3390/su14169951>.
19. Bakar NMMAA. Mohd Shahrul Mohd Nadzir, Ong Li Wen, Fatin Nur Afifah Suris. Prediction of Multivariate Air Quality Time Series Data using Long Short-Term Memory Network. *Malaysian Journal of Fundamental and Applied Sciences*. 2022;1(1):52–9. Available from: <https://doi.org/10.11113/mjfas.v1i8n1.2393>.
20. Ho CH, Park I, Kim J, Lee JB. Forecast in Korea using the Long Short-Term Memory (LSTM) Model. *Asia Pac J Atmos Sci*. 2022;p. 1–14. Available from: <https://doi.org/10.1007/s13143-022-00293-2>.
21. Xu S, Li W, Zhu Y, Xu A. A novel hybrid model for six main pollutant concentrations forecasting based on improved LSTM neural networks. *Sci Rep*. 2022;Available from: <https://doi.org/10.1038/s41598-022-17754-3>.
22. Spyrou ED, Tsoulos I, Stylios C. Applying and Comparing LSTM and ARIMA to Predict CO Levels for a Time-Series Measurements in a Port Area. *Signals*. 2022;3(2):235–48. Available from: <https://doi.org/10.3390/signals3020015>.
23. Tsokov S, Lazarova M, Aleksieva-Petrova A. A Hybrid Spatiotemporal Deep Model Based on CNN and LSTM for Air Pollution Prediction. *Sustainability (Basel)*. 2022;14(9):5104. Available from: <https://doi.org/10.3390/su14095104>.
24. Drewil GI, Al-Bahadili RJ. Air pollution prediction using LSTM deep learning and metaheuristics algorithms. *Measurement Sensors*. 2022;24. Available from: <https://doi.org/10.1016/j.measen.2022.100546>.
25. Duan J, Gong Y, Luo J, Zhao Z. Air pollution prediction using LSTM deep learning and metaheuristics algorithms. *Measurement: Sensors*. 2023;24. Available from: <https://doi.org/10.1038/s41598-023-36620-4>.
26. Chen H, Guan M, Li H. Air Quality Prediction Based on Integrated Dual LSTM Model. *IEEE Access : Practical Innovations, Open Solutions*. 2021;9:93285–97. Available from: <https://doi.org/10.1109/access.2021.3093430>.
27. Raheja S, Malik S. Prediction of Air Quality Using LSTM Recurrent Neural Network. *International Journal of Software Innovation*. 2022;10(1):1–16. Available from: <https://doi.org/10.4018/ijsi.297982>.
28. Li MHT, Wu X. A Hybrid CNN-LSTM Model for Forecasting Particulate Matter (PM_{2.5}). *IEEE Access : Practical Innovations, Open Solutions*. 2020;8:26933–40. Available from: <https://doi.org/10.1109/ACCESS.2020.2971348>.
29. Pak CKU. Unsok Ryu, Kyongjin Sok, Sungnam Pak. A hybrid model based on convolutional neural networks and long short-term memory for ozone concentration prediction. *Air Quality, Atmosphere & Health*. 2018;11(8):883–95. Available from: <https://doi.org/10.1007/s11869-018-0585-1>.
30. Qi Y, Li Q, Karimian H, Liu D. A hybrid model for spatiotemporal forecasting of PM_{2.5} based on graph convolutional neural network and long short-term memory. *The Science of the Total Environment*. 2019;664:1–10. Available from: <https://doi.org/10.1016/j.scitotenv.2019.01.333>.
31. Fan DHS, Feng Y, Xia K, Yang W. A Hybrid Model for Air Quality Prediction Based on Data Decomposition. *Information (Basel)*. 2021;12(5):210. Available from: <https://doi.org/10.3390/info12050210>.
32. Vu LQTX. A Review on Current Status of Urban Air Pollution in Ho Chi Minh city and Ha Noi. *DTU Journal of Science and Technology*. 2021;4:45–53.
33. Rakholia QLR, Vu K, Ho BQ, Carbajo RS. AI-based air quality PM_{2.5} forecasting models for developing countries: A case study of Ho Chi Minh City, Vietnam. *Urban Climate*. 2022;46. Available from: <https://doi.org/10.1016/j.uclim.2022.101315>.
34. Minh VTT, Hien TT. PM_{2.5} Forecast System by Using Machine Learning and WRF Model, A Case Study: ho Chi Minh City, Vietnam. *Aerosol and Air Quality Research*. 2021;21(12). Available from: <https://doi.org/10.4209/aaqr.210108>.
35. Administration VE. Technical instructions for calculating and publishing Vietnam's air quality index. 2019;.
36. Duong LD. A short - term load forecasting model based on Group method of data handling. *The University of Danang - Journal of Science and Technology*. 2019;17:17.
37. Krishan M, Jha S, Das J, Singh A, Goyal MK, Sekar C. Air quality modelling using long short-term memory (LSTM) over NCT-Delhi, India. *Air Quality, Atmosphere & Health*. 2019;12(8):899–908. Available from: <https://doi.org/10.1007/s11869-019-00696-7>.
38. Yuan X, Li L, Wang Y, Yang C, Gui W. Deep learning for quality prediction of nonlinear dynamic processes with variable attention-based long short-term memory network. *Canadian Journal of Chemical Engineering*. 2019;98(6):1377–89. Available from: <https://doi.org/10.1002/cjce.23665>.

39. Johnston FR, Boyland JE, Meadows M, Shale E, JEB FJ. M. Meadows, E. Shale. Some Properties of a Simple Moving Average when Applied to Forecasting a Time Series. The Journal of the Operational Research Society. 1999;50(12):1267–71. Available from: <https://doi.org/10.1057/palgrave.jors.2600823>.
40. Gujarati DN, Porter DC, Gunasekar S. Basic econometrics. Tata McGraw-Hill Education; 2012.
41. Heydari M, Olyaie E, Mohebzadeh H, Kisi Ö. Development of a neural network technique for prediction of water quality parameters in the Delaware River, Pennsylvania. Middle East Journal of Scientific Research. 2013;13:1367–76. Available from: <https://doi.org/10.5829/idosi.mejsr.2013.13.10.1238>.
42. Dounis AI, Tiropanis P, Tseles D, et al. A comparison of grey model and fuzzy predictive model for time series. International Journal of Information and Mathematical Sciences. 2006;2:176–81.
43. Banejad H, Olyaie E. Application of an artificial neural network model to rivers water quality indexes prediction –a case study. The Journal of American Science. 2011;7:60–5.
44. Panda SS, Garg V, Chaubey I. Artificial neural networks application in lake water quality estimation using satellite imagery. Journal of Environmental Informatics. 2004;4(2):65–74. Available from: <https://doi.org/10.3808/jei.200400038>.
45. Haiyan W. Assessment and prediction of overall environmental quality of Zhuzhou City, Hunan Province, China. Journal of Environmental Management. 2002;66(3):329–40. Available from: <https://doi.org/10.1006/jema.2002.0590>.
46. J K. A short - term load forecasting model based on Group method of data handling. In: E3S Web of Conferences 2021; 2021. Available from: <https://doi.org/10.1051/e3sconf/202125703027>.

Air quality index prediction with LSTM-MA model: A case study in Ba Ria – Vung Tau province, Vietnam

Ho Minh Dung^{1,*}, Nguyen Hien Than²



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

ABSTRACT

Air pollution significantly contributes to public health issues by increasing the incidence of respiratory and cardiovascular diseases; therefore, timely forecasting of air quality variations is essential for enhancing community awareness and supporting effective pollution management in urban areas. This study applies artificial intelligence techniques to predict air quality in Ba Ria–Vung Tau Province, Vietnam, using a Long Short-Term Memory (LSTM) model integrated with a Moving Average (MA) filter to reduce noise in the input dataset. Statistical analysis indicates that air quality in the study area is generally acceptable, with concentrations of CO, NO₂, SO₂, PM₁₀, and PM_{2.5} mostly remaining below regulatory limits, while ozone concentrations frequently exceeded permissible thresholds and exerted a significant influence on the Air Quality Index (AQI). The proposed LSTM–MA model demonstrated strong predictive performance over a 7-day forecasting horizon, achieving Root Mean Square Error (RMSE) values ranging from 0.157 to 0.271 and Mean Absolute Percentage Error (MAPE) values between 30.0% and 31.0%. For longer forecasting periods from two weeks to one month, the model maintained satisfactory accuracy, with RMSE values of 0.186–0.497 and MAPE values of 46.3%–84.4%, and comparative analyses with conventional LSTM and ARIMA models confirmed that the proposed LSTM–MA approach provides improved forecasting accuracy and robustness.

Key words: forecast, LSTM-MA, hybrid model, air quality, Ba Ria-Vung Tau province

¹Institute for Environment and Resources, VNU-HCM, Linh Xuan ward, HCMC, Vietnam

²Carbon neutral consulting center, Thu Dau Mot University, Phu Loi ward, HCMC, Vietnam

Correspondence

Ho Minh Dung, Institute for Environment and Resources, VNU-HCM, Linh Xuan ward, HCMC, Vietnam

Email: H_minhdung@yahoo.com

History

- Received: 24-12-2025
- Revised: 06-01-2026
- Accepted: 31-07-2026
- Published Online: 21-10-2026

DOI : <https://doi.org/10.32508/vnuhcmj-ees.v10i2.871>



Copyright

© VNUHCM Journal. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International license.

Cite this article : Minh Dung H, Hien Than N. **Air quality index prediction with LSTM-MA model: A case study in Ba Ria – Vung Tau province, Vietnam.** *VNUHCM J. Environ. Earth Sci.* 2026; 10(2):1485-1498.