VNUHCM Journal of Environment and Earth Sciences 2026, 10(2): 1454-1467

Open Access Full Text Article

Bài báo nghiên cứu

XÂY DỰNG MÔ HÌNH LAI LSTM-ARIMA DỰ BÁO CHẤT LƯỢNG KHÔNG KHÍ TẠI TRẠM QUAN TRẮC NGÃ TƯ GIẾNG NƯỚC, TP.HCM

Hồ Minh Dũng[1,*], Tôn Minh Hiền[1,2]

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

TÓM TẮT

Ô nhiễm không khí là một trong những nguyên nhân chính gây ra các vấn đề sức khỏe nghiêm trọng như bệnh phổi, tim mạch và làm gia tăng nguy cơ mắc các bệnh khác. Do đó, dự báo sớm chất lượng không khí đóng vai trò quan trọng trong việc cảnh báo cộng đồng và giúp người dân chủ động thực hiện các biện pháp bảo vệ sức khỏe. Trong nghiên cứu này, mô hình LSTM được lựa chọn nhờ khả năng nhận diện tốt các mối quan hệ phi tuyến trong dữ liệu, phù hợp với cả dự báo ngắn hạn và dài hạn. Nhằm nâng cao hiệu quả, mô hình LSTM được kết hợp với ARIMA để tạo thành mô hình lai LSTM-ARIMA, mang lại kết quả dự báo khả quan. Thống kê giá trị của các thông số CO, NO_2, SO_2, $PM_{2.5}$ và PM_{10} tại trạm quan trắc ngã tư Giếng Nước (phường Vũng Tàu, TP.HCM) trong khoảng 2020-2024 luôn nằm dưới ngưỡng cho phép QCVN 05:2023/BTNMT. Tuy nhiên, nồng độ O_3 trung bình 8 giờ tại 866 thời điểm và O_3 trung bình 1 giờ tại 74 thời điểm vượt qui chuẩn. Chất lượng không khí tại trạm quan trắc ngã tư Giếng Nước chủ yếu nằm trong nhóm ``Tốt'' và ``Trung bình'' theo chỉ số AQI. Mô hình LSTM-ARIMA cho kết quả dự báo 1 ngày với độ chính xác cao (RMSE = 14,57; MAE = 7,81; MAPE = 16,3%) và vẫn giữ hiệu suất ổn định khi dự báo mở rộng đến 7 và 14 ngày, cho thấy tiềm năng ứng dụng trong công tác dự báo chất lượng không khí. Kết quả nghiên cứu cho thấy mô hình LSTM-ARIMA có hiệu suất dự báo tốt hơn so với các mô hình đơn, có thể mở rộng và áp dụng cho các trạm quan trắc khác góp phần xây dựng các hệ thống dự báo linh hoạt, phù hợp với điều kiện dữ liệu quan trắc thực tế tại các đô thị đang phát triển.

Từ khoá: chất lượng không khí, LSTM-ARIMA, trạm ngã tư Giếng nước

[1]*Viện Môi trường và Tài nguyên, ĐHQG-HCM, 142 Tô Hiến Thành, Phường Diên Hồng, Thành phố Hồ Chí Minh, Việt Nam*

[2]*Viện Khoa học An toàn Vệ sinh Lao động Thành phố, 348 Nguyễn Thái Sơn, Phường Hạnh Thông, Thành phố Hồ Chí Minh, Việt Nam*

Liên hệ

Hồ Minh Dũng, Viện Môi trường và Tài nguyên, ĐHQG-HCM, 142 Tô Hiến Thành, Phường Diên Hồng, Thành phố Hồ Chí Minh, Việt Nam

Email: H_minhdung@yahoo.com

Lịch sử

- Ngày nhận: 19-11-2025
- Ngày sửa đổi: 04-03-2026
- Ngày chấp nhận: 03-08-2026
- Ngày đăng: 25-09-2026

DOI : https://doi.org/10.32508/vnuhcmjees.v10i2.858

Check for updates

Bản quyền

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

ĐẶT VẤN ĐỀ

Ô nhiễm không khí tại các đô thị và thành phố lớn đang trở thành vấn đề nghiêm trọng trong bối cảnh công nghiệp hóa và đô thị hóa nhanh chóng. Việc dự báo chất lượng không khí là một nhiệm vụ phức tạp do nồng độ các chất ô nhiễm như $PM_{2.5}$, PM_{10}, SO_2, NO_2, O_3 và CO biến động nhanh và khó kiểm soát. Trong bối cảnh đó, việc ứng dụng trí tuệ nhân tạo (AI) trong dự báo chất lượng không khí là cần thiết, giúp tăng độ chính xác và khả năng phản ứng nhanh trong công tác quản lý môi trường [1,2].

Các mô hình dự báo chất lượng không khí được chia thành hai nhóm chính là mô hình giải tích (Deterministic model) và mô hình thống kê (Statistical model) [2]. Trong đó, mô hình giải tích dựa trên các quá trình vật lý, hóa học và khí tượng để mô phỏng sự biến đổi của chất lượng không khí theo không gian và thời gian. Từ giữa thế kỷ 20, nhiều mô hình giải tích đã được phát triển và ứng dụng rộng rãi như CMAQ, WRF-Chem, NAQPMS, AERMOD và TAPOM [3,4]. Mặc dù có độ chính xác cao, các mô hình giải tích vẫn tồn tại hạn chế như thời gian tính toán dài, yêu cầu dữ liệu đầu vào chi tiết về nguồn thải, và việc ước lượng thông số mô hình dựa vào kinh nghiệm, làm giảm độ tin cậy của kết quả [1,5].Các mô hình thống kê truyền thống trong dự báo chất lượng không khí tập trung chủ yếu vào phân tích mối quan hệ thống kê giữa dữ liệu đầu vào như yếu tố khí tượng và nồng độ ô nhiễm trong quá khứ với nồng độ ô nhiễm trong tương lai, mà ít quan tâm đến các cơ chế vật lý và hóa học của chất ô nhiễm [4]. Các phương pháp phổ biến bao gồm ARMA, ARIMA, GWR, MLR và SVR [6,7], thường là các mô hình hồi quy mô tả mối quan hệ giữa biến độc lập và biến phụ thuộc. Ví dụ, các nghiên cứu sử dụng ARMA, ARIMA để dự báo O_3, NO, NO_2 và CO 3, hay MLR để dự báo NOx tại Athens và Helsinki [8]. Tuy nhiên, các mô hình này gặp hạn chế khi phải tích hợp dữ liệu đa nguồn và không đồng nhất như lưu lượng giao thông, điều kiện khí tượng và sử dụng đất, những yếu tố quan trọng ảnh hưởng đến chất lượng không khí nhưng không được mô hình hóa đầy đủ [1]. Bên cạnh đó, khi dữ liệu trở nên phức tạp và lớn, các phương pháp thống kê truyền thống thường có hiệu suất xử lý kém, tốn nhiều thời gian và tài nguyên tính toán, khiến chúng khó đáp ứng yêu cầu ứng dụng thực tế hiện đại [9].

Trích dẫn bài báo này: Dũng H M, Hiền T M. **XÂY DỰNG MÔ HÌNH LAI LSTM-ARIMA DỰ BÁO CHẤT LƯỢNG KHÔNG KHÍ TẠI TRẠM QUAN TRẮC NGÃ TƯ GIẾNG NƯỚC, TP.HCM**. *VNUHCM J. Environ. Earth Sci.2026; 10(2): 1454-1467.*

Các nghiên cứu gần đây tập trung ứng dụng mô hình học sâu, đặc biệt là LSTM và các mô hình lai, nhằm nâng cao độ chính xác trong dự báo chất lượng không khí. Yu và cộng sự (2019) cho thấy LSTM dự báo AQI tốt hơn ARIMA và hồi quy tuyến tính [10]; Sagar và cộng sự (2020) kết hợp LSTM-RNN với mạng cảm biến không dây giúp tăng độ chính xác [11]; Yan Rui và cộng sự (2021) kết hợp CNN-LSTM và phân cụm không – thời gian dự báo AQI đa điểm tại Bắc Kinh hiệu quả [12]. Một số nghiên cứu khác cũng khẳng định ưu thế của mô hình lai: Aishah và Zeyar (2023) chứng minh LSTM và NAR-NN vượt trội hơn mô hình thống kê [13]; Ricardo (2020) cho thấy RNN-LSTM hiệu quả hơn khi kết hợp [14]; Yue và cộng sự (2020) đề xuất LSTM đa biến ổn định hơn [15]; Congcong (2018) phát triển CNN-LSTM dự báo $PM_{2.5}$ chính xác hơn LSTM đơn [16]; Jiahui và cộng sự (2023) kết hợp ARIMA, CNN, LSTM và thuật toán DBO giúp giảm RMSE 84,71% và tăng R^2 92,41% [17]. Ở Việt Nam, học sâu được áp dụng chủ yếu trong lĩnh vực tài nguyên nước như dự báo độ mặn [18], nhiệt độ – lượng mưa [19] hay mực nước sông [20]. Trong lĩnh vực chất lượng không khí, một số nghiên cứu tiên phong đã đạt kết quả tích cực như Hồ Minh Dũng (2024) phát triển mô hình LSTM-MA dự báo 1 ngày tại Bà Rịa – Vũng Tàu [21]; Cao Minh Quý (2024) kết hợp LSTM, ANN và RNN dự báo các chất ô nhiễm [22]; Nguyễn Việt Hưng (2021) dùng CNN-LSTM vượt trội hơn KNN và Linear Regression [23]; Đỗ Văn Triều (2022) kết hợp LSTM và thuật toán NA đạt độ chính xác 99,48% [24]; Rajnish Rakholia (2022) so sánh các mô hình và xác định SGDRegressor hiệu quả nhất tại TP.HCM [25]. Nhìn chung, học sâu và mô hình lai đang trở thành xu hướng quan trọng, giúp nâng cao độ chính xác và hiệu quả trong dự báo chất lượng không khí.

Các nghiên cứu cho thấy mô hình lai cho kết quả dự báo chính xác và ổn định hơn mô hình đơn. Mục tiêu của nghiên cứu này là xây dựng mô hình lai LSTM-ARIMA để dự báo chất lượng không khí nhờ vào việc tận dụng ưu điểm của cả hai mô hình đơn, ARIMA (AutoRegressive Integrated Moving Average) là mô hình thống kê truyền thống, phù hợp với chuỗi thời gian tuyến tính, có xu hướng ổn định và mối quan hệ rõ ràng giữa các giá trị quá khứ và hiện tại, nhưng hạn chế trong xử lý dữ liệu phi tuyến và biến động phức tạp. Ngược lại, LSTM (Long Short-Term Memory) là mạng nơ-ron hồi quy có khả năng học các mối quan hệ phi tuyến, phụ thuộc dài hạn, giúp dự báo hiệu quả với dữ liệu biến động mạnh, song đòi hỏi thời gian huấn luyện dài và dễ overfitting khi dữ liệu hạn chế. Việc kết hợp hai mô hình này giúp bù trừ điểm yếu của từng mô hình, ARIMA đảm nhiệm phần tuyến tính và xu hướng dài hạn, còn LSTM xử lý phần phi tuyến và phụ thuộc phức tạp, đồng thời hạn chế các nhược điểm khi sử dụng mô hình riêng lẻ. Nhờ việc tận dụng được ưu thế của cả hai mà mô hình lai LSTM-ARIMA nâng cao độ chính xác, giảm sai số và tăng tính ổn định cho dự báo chất lượng không khí trong cả ngắn hạn và dài hạn.

NỘI DUNG VÀ PHƯƠNG PHÁP NGHIÊN CỨU

Nội dung nghiên cứu

Để đáp ứng mục tiêu đặt ra, nghiên cứu sẽ thực hiện hiện các nội dung sau: (i) Đánh giá chất lượng không khí thông qua dữ liệu thu thập về quan trắc chất lượng không khí xung quanh tại trạm ngã tư Giếng nước, bao gồm: Thu thập các tài liệu, số liệu quan trắc chất lượng không khí tại trạm quan trắc tự động, liên tục ngã tư Giếng nước; Đánh giá hiện trạng chất lượng không khí dựa trên các dữ liệu thông tin thu thập, quan trắc; (ii) Tiền xử lý số liệu và xây dựng mô hình dự báo chất lượng không khí, bao gồm: Tiền xử lý dữ liệu đầu vào; Tính toán chỉ số chất lượng không khí AQI; Chuẩn hóa dữ liệu; Xây dựng mô hình dự báo chất lượng không khí; Tiến hành huấn luyện mô hình LSTM-ARIMA; (iii) Dự báo diễn biến chất lượng không khí từ mô hình tối ưu cho khu vực Bà Rịa - Vũng Tàu: Sử dụng mô hình tối ưu từ quá trình đánh giá dự báo diễn biến chất lượng không khí cho khu vực Bà Rịa – Vũng Tàu; So sánh kết quả từ mô hình dự báo với kết quả thực tế thu thập.

Dữ liệu nghiên cứu

Bộ dữ liệu sử dụng trong nghiên cứu này là nồng độ trung bình giờ của các thông số chất lượng không khí được đo tại trạm quan trắc tự động chất lượng không khí xung quanh ngã tư Giếng nước đặt tại phường 7, thành phố Vũng Tàu tỉnh Bà Rịa – Vũng Tàu (nay thuộc phường Vũng Tàu, TP.HCM), với tọa độ X: 426939; Y:146298. Thời gian thu thập của bộ dữ liệu trong 5 năm từ 18/1/2020 đến 31/12/2024. Bộ dữ liệu bao gồm 15 thông số: 06 thông số khí tượng (bao gồm: tốc độ gió, hướng gió, nhiệt độ, độ ẩm, áp suất, bức xạ mặt trời); 09 thông số chất lượng không khí ($\mu g/m^3$) (bao gồm: NO, NO_2, NO_x, SO_2, CO, O_3, PM_{10}, $PM_{2.5}$, TSP).

Phương pháp nghiên cứu

Phương pháp 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. Dữ liệu đầu vào là phần quan trọng dẫn đến kết quả chính xác. Tuy nhiên, các tập dữ liệu thu thập được phân phối không đồng đều và có thể thiếu

do lỗi thiết bị. Do đó, các tập dữ liệu cần được thực hiện loại bỏ dữ liệu ngoại lai và 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.

- Loại bỏ dữ liệu ngoại lai: Dữ liệu ngoại lai là những giá trị dữ liệu có sự sai lệch lớn so với phần còn lại của tập dữ liệu. Quá trình huấn luyện mô hình học máy rất phụ thuộc vào phạm vi và phân phối của các điểm dữ liệu. Các giá trị ngoại lai có thể gây ảnh hưởng tiêu cực đến quá trình huấn luyện, làm cho quá trình huấn luyện diễn ra lâu hơn và giảm độ chính xác của mô hình. Do đó, việc loại bỏ các giá trị ngoại lai là rất quan trọng để cải thiện hiệu suất của mô hình. Một phương pháp phổ biến để xác định các giá trị ngoại lai là sử dụng Điểm số Z (Z-scores). Điểm số Z của một mẫu là chỉ số đo lường số độ lệch chuẩn mà giá trị của mẫu này khác biệt so với giá trị trung bình của tập dữ liệu. Điểm số Z của một mẫu x được xác định:

$$Z_i = \frac{(x_i - \mu)}{\sigma}$$

Trong đó, Z_i là điểm số Z của điểm dữ liệu thứ i, x_i là điểm dữ liệu thứ i, μ là giá trị trung bình và σ là độ lệch chuẩn của dữ liệu. Nếu giá trị tuyệt đối của chỉ số Z của bất kỳ mẫu nào lớn hơn 3, thì mẫu đó được coi là ngoại lệ và bị xóa khỏi tập dữ liệu.

- Điền giá trị thay thế (dữ liệu khuyết): Trong quá trình thu thập dữ liệu từ các trạm quan trắc, có thể xảy ra tình trạng thiếu dữ liệu do các sự cố kỹ thuật như trục trặc thiết bị đo, mất điện hoặc sự cố mạng, dẫn đến gián đoạn trong quá trình truyền tải thông tin. Các giá trị thiếu này có thể gây ảnh hưởng đến tính liên tục của tập dữ liệu và làm giảm hiệu quả trong việc huấn luyện mô hình. Để giải quyết vấn đề này, các giá trị thiếu sẽ được thay thế bằng giá trị đo được vào cùng thời điểm trong tuần trước, bởi vì các giá trị này thường tuân theo cùng một phân phối theo thời gian. Trong trường hợp không có giá trị quan trắc tại cùng thời điểm trong tuần trước, giá trị trung bình của các quan trắc trong cùng một biến sẽ được sử dụng để thay thế các giá trị thiếu, nhằm đảm bảo tính đồng nhất và không làm gián đoạn quá trình phân tích dữ liệu. Trong nghiên cứu này sử dụng phần mềm SPSS để thực hiện việc điền dữ liệu khuyết.

- Chuẩn hóa dữ liệu (Data normalization): Để nâng cao hiệu quả tính toán và tối ưu hóa quá trình huấn luyện mô hình, chuẩn hóa dữ liệu là một bước quan trọng giúp biến tập dữ liệu chứa nhiễu thành dữ liệu có cấu trúc hoàn chỉnh. Các đặc trưng dữ liệu thường có sự khác biệt về đơn vị đo (scale) và phân phối (distribution), điều này ảnh hưởng trực tiếp đến hiệu quả phân loại của mô hình và khả năng hội tụ của các thuật toán tối ưu hóa, như gradient descent. Khi các đặc trưng có đơn vị quá khác biệt, gradient có thể không hội tụ về cực trị toàn cục, dẫn đến việc mô hình không đạt được kết quả tối ưu. Hơn nữa, sự khác biệt về đơn vị còn gây sai lệch trong việc đánh giá ảnh hưởng của từng đặc trưng đối với kết quả dự đoán. Để khắc phục tình trạng này và cải thiện tốc độ hội tụ của mô hình, dữ liệu sẽ được chuẩn hóa về phạm vi [0,1] thông qua phương pháp chuẩn hóa Min-Max:

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

Trong đó, x' là giá trị dữ liệu được chuyển đổi (giá trị 0 – 1), x là giá trị gốc.

Mục tiêu của phương pháp 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. Phương pháp này đưa các giá trị về một khoảng đặc biệt, thường là [0,1] hoặc [−1,1].

Phương pháp tính toán chỉ số chất lượng không khí AQI

Kỹ thuật tính toán và công bố chỉ số chất lượng không khí Việt Nam (VN_AQI) được quy định dựa trên dữ liệu quan trắc từ các trạm quan trắc không khí tự động, liên tục, theo Quyết định số 1459/QĐ-TCMT ngày 12/11/2019 của Tổng Cục Môi trường[26]. Chỉ số VN_AQI được biểu thị thông qua một thang điểm (hay khoảng giá trị AQI) tương ứng với các biểu tượng và màu sắc nhằm cung cấp thông tin cảnh báo về chất lượng không khí và mức độ ảnh hưởng đến sức khỏe con người.

Phương pháp học sâu (deep learning)

XÂY DỰNG MÔ HÌNH LSTM

Cấu trúc chính của mạng LSTM là thành phần ô trạng thái là dạng băng chuyền chạy thẳng xuyên suốt toàn bộ chuỗi với chỉ một vài tương tác tuyến tính nhỏ giúp cho thông tin có thể truyền dọc theo đồ thị mạng nơ ron ổn định, còn cổng là nơi sàng lọc thông tin đi qua nó, chúng được kết hợp bởi một tầng mạng sigmoid. Đầu tiên, thông tin sẽ được quyết định loại bỏ hoặc giữ lại ở ô trạng thái thông qua tầng cổng quên (forget gate layer). "Tầng cổng quên" nhận đầu vào là h_{t-1} (giá trị đầu ra tại thời điểm t-1) và x_t (dữ liệu đầu vào hiện tại) và đầu ra là một số nằm trong khoảng [0,1]. Nếu đầu ra là 0 thì tất cả các thông tin bị loại bỏ. Nếu là 1 thì tất cả các thông tin được đi qua. Ưu điểm của mô hình LSTM chính là khả năng phát hiện được mối quan hệ phi tuyến trong dữ liệu nên phù hợp cho cả dự báo trong ngắn và dài hạn[27,28]. Tuy nhiên, mô hình này có yêu cầu cao về bộ dữ liệu sử dụng cho huấn luyện, nếu trong bộ dữ liệu có các giá trị ngoại

lai sẽ khiến cho mô hình tốn nhiều thời gian để huấn luyện và tính chính xác của dự báo từ mô hình sẽ bị giảm đi. Do đó, để tài nghiên cứu sử dụng bộ tự hồi quy tích hợp trung bình trượt (ARIMA) để lọc các dữ liệu có trong quá khứ cũng như để nhận biết các xu hướng thay đổi trong tương lai của chuỗi thời gian.

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 đánh giá mức độ phù hợp, Sai số trung bình bình phương gốc (Root Means Squared Error – RMSE), Độ lệch tuyệt đối (Mean Absolute Error – MAE) và Phần trăm sai số tuyệt đối trung bình (Mean Absolute Percentage Error – MAPE) được sử dụng để đánh giá hiệu quả của mô hình.

Phương pháp xử lý số liệu

Các dữ liệu được thu thập từ quan trắc ghi nhận được xử lý bằng phương pháp thống kê, phân tích và đánh giá. Phương pháp thống kê có thể dụng Excel thường sử dụng phổ biến và dễ sử dụng cho các phân tích thống kê cơ bản, R và Python ngôn ngữ lập trình mạnh mẽ cho phân tích và trực quan hóa dữ liệu.

• Thí nghiệm 1: xác định tham số khả thi cho mô hình ARIMA

Trong thí nghiệm đầu tiên, mục tiêu chính là xác định bộ tham số tối ưu cho mô hình ARIMA nhằm đảm bảo khả năng mô phỏng chính xác đặc trưng động lực của chuỗi dữ liệu thời gian. Mô hình ARIMA được đặc trưng bởi ba tham số cơ bản: bậc tự hồi quy (AR – p), bậc sai phân (I – d) và bậc trung bình trượt (MA – q). Việc lựa chọn chính xác các tham số này đóng vai trò quyết định đến hiệu quả dự báo, bởi chúng phản ánh trực tiếp mối quan hệ phụ thuộc theo thời gian và mức độ ổn định của chuỗi.

Quy trình xác định tham số bao gồm:

Kiểm định tính dừng của chuỗi: Sử dụng kiểm định Augmented Dickey-Fuller (ADF test) để đánh giá đặc tính dừng. Nếu chuỗi không dừng, thao tác sai phân được áp dụng nhiều lần cho đến khi dữ liệu đạt tính ổn định, qua đó xác định được giá trị tham số d. Nếu chuỗi thời gian dừng ở bậc 0 ta ký hiệu I (d=0), nếu sai phân bậc 1 của chuỗi dừng ký hiệu I (d=1), …

Việc xác định tính dừng của mô hình nhằm để loại bỏ phương sai và ổn định xu hướng. Chuỗi dừng có đặc tính không thay đổi theo thời gian. Nếu không xác định tính dừng của chuỗi dữ liệu thì chuỗi có tính thay đổi theo thời gian xu hướng không ổn định, dẫn đến việc phân tích và dự báo sẽ không ổn định, sai lệch và kết quả không đáng tin cậy.

Xác định bậc AR (thành phần tự hồi quy) và MA (thành phần trung bình trượt): Sau khi kiểm định tính dừng, xác định bậc của quy trình tự hồi quy (AR) và quy trình trung bình trượt (MA) thông qua phân tích đồ thị hàm tự tương quan (ACF) và hàm tự tương quan riêng phần (PACF). Nhận dạng mô hình ARIMA (p, d, q) là tìm các giá trị thích hợp của p, d, q, với d là bậc sai phân của chuỗi thời gian được khảo sát, p là bậc tự hồi quy và q là bậc trung bình trượt. Việc xác định p và q sẽ phụ thuộc vào các đồ thị PACF = f(t) và ACF = f(t)

Ngoài ra, nhiều mô hình ARIMA với các tổ hợp tham số khác nhau được ước lượng, sau đó so sánh dựa trên các tiêu chí thống kê như Akaike Information Criterion (AIC) và Bayesian Information Criterion (BIC) nhằm cân bằng độ khớp của mô hình với dữ liệu và độ phức tạp của mô hình. Mô hình có giá trị AIC có điểm AIC thấp hơn tốt hơn điểm AIC cao hơn. Bên cạnh đó, BIC là tiêu chí khác để lựa chọn mô hình tốt, hạn chế tính phức tạp hơn AIC. Các mô hình có giá trị BIC thấp thường được chọn.

• 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 (adaptive moment estimation algorithm – Adam). Sai số RMSE được chọn là hàm mất mát để xác định mô hình tối ưu.

• 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 lai LSTM-ARIMA, nghiên cứu tiến hành xác định khả năng dự báo chất lượng không khí tại trạm quan trắc ngã tư Giếng nước. Thời gian dự báo (bước nhảy) của mô hình được tiếp tục khảo sát để đánh giá khả năng dự báo của mô hình. Nghiên cứu đã tiến hành khảo sát bước nhảy dự báo lần lượt k =1, k = 7, k = 14.

KẾT QUẢ VÀ THẢO LUẬN

Đánh giá hiện trạng chất lượng không khí tại trạm Ngã tư Giếng nước

Hiện trạng bộ dữ liệu thu thập

Bộ dữ liệu thu thập tại trạm ngã tư Giếng nước bao gồm tổng cộng 43.250 mẫu dữ liệu, tỷ lệ thu thập cho các thông số cho thời gian từ 18/1/2020 đến

31/12/2024 đạt khoảng từ 93% đến 100%. Tỉ lệ số liệu thu thập được cho cả 6 chất ô nhiễm chính đều đạt tỉ lệ cao trên 93%. Trong tất cả các thông số được sử dụng để tính toán AQI, CO là thông số có sự dao động lớn nhất. Độ lệch chuẩn của CO lên đến 582,3 $\mu g/m^3$. 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 tại khu vực đô thị. 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 khi dao động từ 0,1 đến 139,2 $\mu g/m^3$ với NO_2 và từ 0,1 đến 194 $\mu g/m^3$ với SO_2. Giá trị trung bình cho 2 thông số này cũng không có sự khác biệt đáng kể khi lần lượt là 15,7 $\mu g/m^3$và 33,6 $\mu g/m^3$. Giá trị trung bình của $PM_{2.5}$ và PM_{10} là 13,73 $\mu g/m^3$ và 22,69 $\mu g/m^3$. Tuy nhiên, giá trị cực đại ghi nhận được ở 2 thông số này đều lớn hơn 300 $\mu g/m^3$, điều này tiềm ẩn khả năng chất lượng không khí bị biến động lớn từ nguồn bụi mịn có trong không khí.

Hiện trạng chất lượng không khí

Tại trạm Ngã tư Giếng Nước, giá trị các thông số CO, NO_2 và SO_2 luôn ổn định trong giới hạn cho phép. Nồng độ O_3 trung bình 1 giờ có 74 thời điểm vượt quy chuẩn từ 1,1–1,6 lần, chủ yếu vào khung 12–14 giờ; O_3 trung bình 8 giờ có 866 thời điểm vượt từ 1,1–1,8 lần, tập trung vào 16–20 giờ – thời gian nồng độ O_3 đạt cực đại do cường độ ánh sáng mạnh. Bụi PM_{10} và $PM_{2.5}$ phần lớn nằm trong ngưỡng cho phép, chỉ có một số thời điểm cận ngưỡng vào giờ cao điểm giao thông (6–9 giờ và 16–21 giờ). Nhìn chung, các thông số ô nhiễm có xu hướng tăng vào buổi trưa và chiều, chịu ảnh hưởng mạnh bởi bức xạ mặt trời và hoạt động giao thông.

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. Trong đó, những điểm dữ liệu có giá trị lớn hơn Q3+1,5*IQR và nhỏ hơn Q1-1,5*IQR (trong đó, Q1 và Q3 lần lượt là phân vị 25% và phân vị 75%, IQR = Q3-Q1) tạm được nghi ngờ là dữ liệu ngoại lai.

Nồng độ NO_2 qua các năm đều nằm trong ngưỡng cho phép và dao động từ 0-25 $\mu g/m^3$. Kết quả về phân bố SO_2 qua các năm cho thấy các giá trị đều nằm trong ngưỡng cho phép dao động từ 0-30 $\mu g/m^3$. Năm 2023, có thể thấy thông số SO_2 có ngưỡng trung bình bất thường so với các năm còn lại khoảng gần 100 $\mu g/m^3$; Phân bố nồng độ O_3 trung bình 1 giờ qua các năm đều có các giá trị vượt ngưỡng, trong đó giá trị vượt ngưỡng cao nhất 320 $\mu g/m^3$vào năm 2022. Ngưỡng trung bình dao động 40-90 $\mu g/m^3$; Với thông số CO cho thấy các giá trị các năm đều nằm trong ngưỡng giá trị cho phép, giá trị đạt cao nhất 7.261,7 $\mu g/m^3$năm 2021, giá trị trung bình đạt 664,77 $\mu g/m^3$. Kết quả nồng độ $PM_{2.5}$ trung bình 24 giờ qua các năm hầu như nằm trong ngưỡng cho phép so với QCVN 05:2023/BTNMT, giá trị trung bình đạt 12,7 $\mu g/m^3$ và có 1 thời điểm xấp xỉ giới hạn. Phân bố nồng độ $PM_{2.5}$ tập trung chủ yếu trong khoảng 5 - 20 $\mu g/m^3$; Kết quả nồng độ PM_{10} trung bình 24 giờ qua các năm nằm trong ngưỡng cho phép so với QCVN 05/2023/BTNMT, giá trị trung bình đạt 21,7 $\mu g/m^3$. Phân bố nồng độ tập trung chủ yếu trong khoảng 17-23 $\mu g/m^3$ (Bảng 1).

Từ kết quả thống kê, ta thấy rằng số lượng các điểm được cho là giá trị ngoại lai là không nhiều. Ngoài ra, giá trị của các thông số ô nhiễm trong kết quả so sánh với QCVN ở trên đều nhận thấy rằng, giá trị nồng độ của NO_2, SO_2, CO, $PM_{2.5}$ và PM_{10} đều không vượt giới hạn cho phép (Hình 1). Ngoài ra, bộ dữ liệu thu thập được cho các thông số ô nhiễm đạt tỉ lệ cao trên 93% nên việc loại bỏ những dữ liệu ngoại lai này sẽ làm mất đi tính khách quan và thực tế của trạm quan trắc tự động. Do đó, bộ dữ liệu sẽ chỉ được bổ sung những dữ liệu còn thiếu.

Tính toán và đánh giá chất lượng không khí tại ngã tư Giếng nước

Dữ liệu sau khi được xử lý sẽ được tính toán chỉ số AQI[d]. Các kết quả phân tích, đánh giá diễn biến theo từng năm và tháng sẽ được thể hiện:

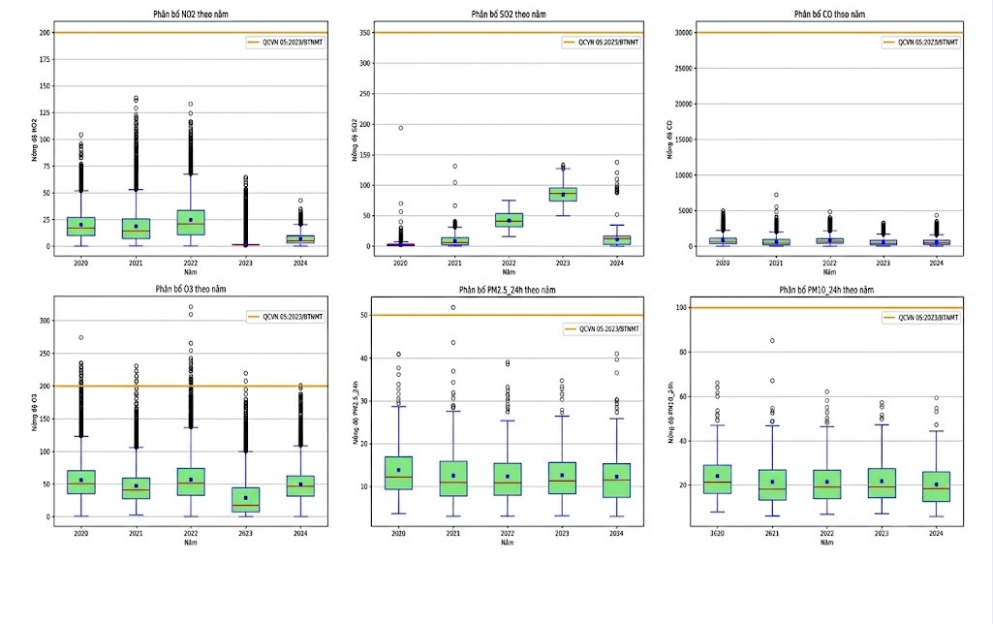
Đánh giá diễn biến chất lượng không khí theo năm

Trong giai đoạn 2020–2024, giá trị AQI[d] tại trạm quan trắc không khí ngã tư Giếng Nước dao động trong khoảng từ 14-196 (Hình 2). Năm 2020, chỉ số AQI nằm trong khoảng 18–177, với 80,1% số ngày có chất lượng không khí Tốt, 9,2% Trung bình và 3,56% Kém. Sang năm 2021, chất lượng không khí được cải thiện khi 84,1% số ngày đạt mức Tốt, 12,4% Trung bình và chỉ 3,6% Kém; giá trị AQI cao nhất rơi vào tháng 1 và giảm mạnh trong giai đoạn tháng 6–11 do ảnh hưởng của giãn cách xã hội (đại dịch Covid 19). Năm 2022, AQI dao động từ 17–196, trong đó 73,6% số ngày có chất lượng không khí Tốt, 12,4% Trung bình và 14,0% Kém, cho thấy mức độ ô nhiễm tăng trở lại.

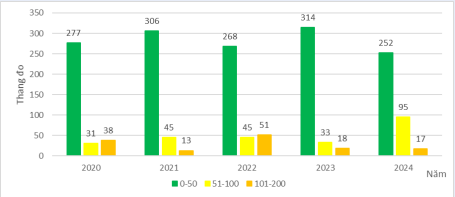
Bảng 1: Thống kê theo phương pháp Box-Whisker cho bộ dữ liệu thô

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	239,9	510,19	664,60	917,23	7261,7	677,33	0	1933,25
NO_2	0,1	2,93	10,6	15,24	22,4	139,2	19,47	0	51,60
O_3	0	25,2	43,34	47,81	63,82	320	38,615	0	121,74
SO_2	0	2,56	16,2	31,32	58,4	194	55,84	0	142,16
$PM_{2,5}$	0	7,06	10,7	12,71	15,8	324,8	8,74	0	28,91
PM_{10}	0	12,4	18,5	21,68	27,2	348,9	14,8	0	49,4

[Nguồn: Nhóm tác giả]
Ghi chú:
Min: Giá trị nhỏ nhất
Max: Giá trị lớn nhất
IQR (Interquartile Range): Khoảng tứ phân vị.



Hình 1: Phân bố nồng độ các chất ô nhiễm qua các năm tại trạm ngã tư Giếng nước [Nguồn: Nhóm tác giả]



Hình 2: Phân loại chất lượng không khí tại trạm ngã tư Giếng nước theo năm [Nguồn: Nhóm tác giả]

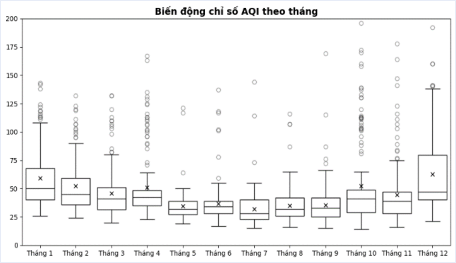
Năm 2023 ghi nhận chất lượng không khí cải thiện với 86,02% số ngày Tốt, 9,04% Trung bình và 4,93% Kém, AQI dao động trong khoảng 28–132. Tuy nhiên, đến năm 2024, chất lượng không khí có xu hướng giảm khi chỉ còn 69,23% số ngày đạt mức Tốt, 26,1% Trung bình và 4,67% Kém, với giá trị AQI nằm trong khoảng 17–141.

Nhìn chung, chất lượng không khí tại khu vực trạm quan trắc ngã tư Giếng nước là tương đối tốt với phần bậc chất lượng không khí đều nằm trong nhóm "Tốt" và "Trung bình" khi xếp loại theo AQI. Tuy nhiên, vào những năm 2021, 2022 và 2024, 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 sau đại dịch Covid 19 và trạm QTTĐ mang đặc tính của khu vực giao thông đô thị.

Đánh giá diễn biến chất lượng không khí theo tháng

Giá trị AQI trung bình là 45,3, chỉ số AQI có xu hướng cao hơn vào thời điểm đầu năm 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 (15-41) và tháng 12 là thời điểm chỉ số AQI cao nhất (Hình 3).



Hình 3: Biến động chỉ số AQI theo tháng tại trạm ngã tư Giếng nước [Nguồn: Nhóm tác giả]

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, có lợi cho sự khuếch tán và lắng đọng các chất ô nhiễm không khí[29]. Nhìn chung, các thông số ô nhiễm đa số đều nằm trong ngưỡng giới hạn cho phép theo QCVN 05:2023/BTNMT và Ozon là thông số chính tác động tới giá trị AQI ngày tại khu vực.

Tính toán tương quan giữa AQI^d và AQI thành phần

Kết quả cho thấy AQI^d tương quan với AQI_O_3 trung bình 8 giờ và AQI_O_3 trung bình 1 giờ rất cao tương ứng 0,91 và 0,84 (Hình 4). Điều này phù hợp với thực tế giá trị O_3 trong 8 giờ và O_3 trong 1 giờ các thời điểm tăng cao và vượt ngưỡng cho phép so với các thông số còn lại. Bên cạnh đó, PM_{10} và $PM_{2.5}$ có tương quan khá cao với AQI^d tương ứng 0,66 và 0,69. 3 chất ô nhiễm NO_2, CO và SO_2 có tác động ít nhất đối với chỉ số AQI tương ứng 0,32; 0,19 và 0,12. Ngoài ra, các giá trị của các chất gây ô nhiễm không khí 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.

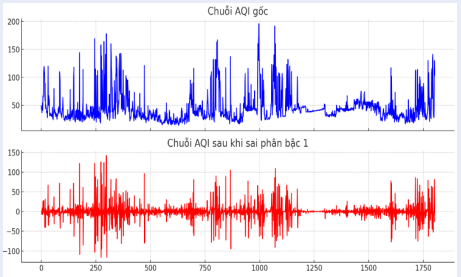


Hình 4: Tương quan giữa chỉ số AQI ngày và các chỉ số AQI thành phần [Nguồn: Nhóm tác giả]

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

Lựa chọn mô hình ARIMA tối ưu

Phân tích chuỗi thời gian gốc biểu đồ biểu thị chuỗi thời gian của chỉ số chất lượng không khí AQI theo ngày. Có thể thấy các biến động không xoay quanh một đường giá trị trung bình cố định và biên độ dao động cũng bất thường. Biểu đồ có nhiều đỉnh nhọn, cho thấy các biến động đột ngột và mạnh mẽ, trái ngược với chuỗi dừng.



Hình 5: Chuỗi dữ liệu khi sai phân bậc 1 [Nguồn: Nhóm tác giả]

Hình 5 đã thực hiện bước sai phân trong việc biến đổi chuỗi dữ liệu không dừng thành chuỗi dừng với d=1. Đây là bước bắt buộc để dữ liệu phù hợp cho việc xây dựng và hiệu chỉnh các mô hình dự báo. Có thể thấy các giá trị dao động xung quanh một giá trị trung bình ổn định (trung bình gần bằng 0). Biên độ biến động của chuỗi đã được làm phẳng không còn các cụm biến động quá lớn. Việc này ổn định xu hướng và loại bỏ phương sai.

Để đảm bảo mô hình ARIMA được chọn sẽ là mô hình khả thi để thực hiện mô hình lai, ta xác định giá

trị p và q của mô hình ARIMA dựa vào biểu đồ ACF và PACF, cho thấy ACF có giá trị giảm dần, không cắt rõ rệt, PACF có spike mạnh tại lag 1, sau đó nhỏ dần vì thế chọn p là 1 và q là 1.

Ta thấy kết quả AC ở lag 1 là 0,66 và các ở lag khác giảm dần cho thấy chuỗi có tính phụ thuộc theo thời gian, PAC lag 1 là 0,66 sau đó giảm mạnh đây là đặc trưng của AR, p-value ở mọi lag rất nhỏ gần bằng 0 mô hình ARIMA (1,1,1) là mô hình phù hợp thực hiện nghiên cứu. Ngoài ra, ta xét tiếp mô hình theo các tiêu chí giá trị AIC và BIC thấp hơn sẽ là mô hình được ưu tiên. Với dữ liệu được chọn ta đạt được kết theo Bảng 5.

Mô hình ARIMA được chọn với tham số p, d, q tương ứng là (1, 1, 1) cho ta thấy hệ số tương quan (ACF) phần dư đều nằm trong khoảng tin cậy gần bằng 0,005, không có lag nào vượt ngưỡng đáng kể. Vì thế phần dư không còn tự tương quan. Tương tự ACF các giá trị PACF phần dư đều nằm trong khoảng tin cậy gần bằng 0,005, không có bất thường ở một độ cụ thể. Các lag khác (2-5) dao động quanh giá trị 0, giá trị nhỏ ±0,005 vì thế các giá trị nằm trong khoảng này là nằm trong ngưỡng nhiễu. Mô hình với tham số khả thi.

Để tăng độ tin cậy của mô hình, nghiên cứu đã tính MAPE của mô hình và đạt được 23,88%, giá trị này ở mức chấp nhận được.

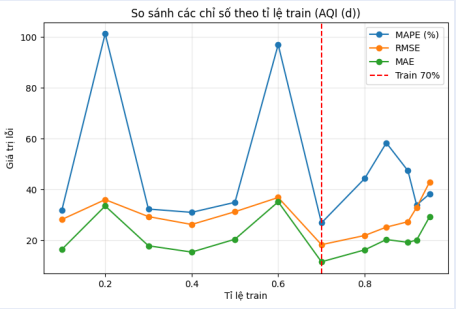
Kết luận mô hình ARIMA dự báo khá tốt xu hướng tổng thể và các dao động lớn trong dữ liệu theo thời gian tuy nhiên mô hình không mạnh ở dự báo biến động ngắn hạn vì thế việc kết hợp với LSTM có thể tăng khả năng dự đoán các biến động nhỏ, bất thường hơn.

Mô hình LSTM

Tỉ lệ học là một trong số những tham số quan trọng nhất của mô hình. Độ lớn của tỷ lệ học sẽ ảnh hưởng trực tiếp tới tốc độ mô hình thay đổi các trọng số để phù hợp với thuật toán và kết quả của mô hình. Tốc độ học lớn có thể giúp mạng nơ-ron được huấn luyện nhanh hơn gấp 10 lần nhưng cũng có thể làm giảm độ chính xác.

Kết quả Hình 6 cho thấy có rất nhiều tỷ lệ train khác nhau và đạt kết quả tốt như 40%, 70%. Tỷ lệ train 70% sẽ cho kết quả ổn định và tốt hơn so với các tỷ lệ khác qua các giá trị RMSE, MAE và MAPE thấp nhất tương ứng 18,30; 11,56 và 26,86%. Sau khi tính toán và đánh giá hiện trạng chất lượng không khí ngày, lấy 70% dữ liệu sau tính toán đưa vào huấn luyện và 30% dữ liệu được kiểm chứng.

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 mô 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 không thể nắm bắt được các giá trị phức tạp có trong dữ liệu như đã phân tích ở trên và cần thiết quá trình loại nhiễu khỏi bộ dữ liệu. Hình 7 cho ta thấy độ mất mát từ quá trình train cho LSTM-ARIMA được cải thiện đáng kể so với LSTM gốc. Đường train và đường validation của mô hình LSTM-ARIMA rất thấp, dao động 0,01 đến 0,025 và gần tiệm cận nhau.

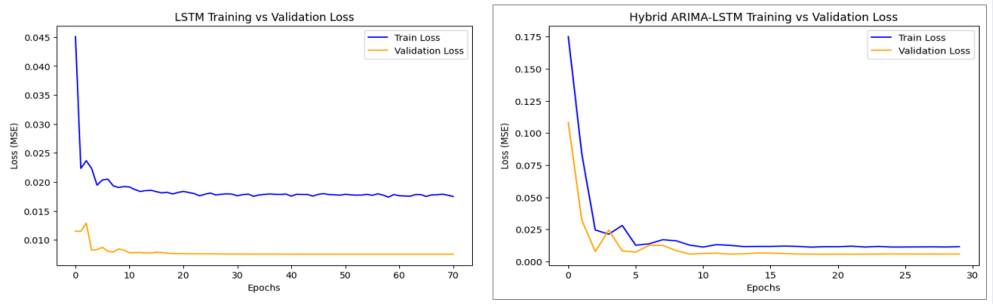


Hình 6: Sai số RMSE, MAE và %MAPE theo tỉ lệ học trên bộ dữ liệu huấn luyện [Nguồn: Nhóm tác giả]

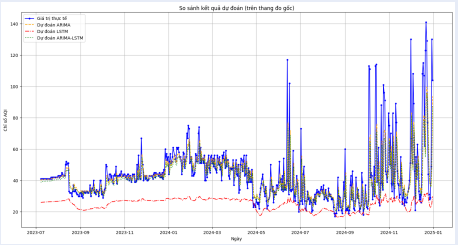
Dự báo chất lượng không khí cho khu vực ngã tư Giếng nước

Nghiên cứu tiến hành đánh giá khả năng dự báo của mô hình lai LSTM-ARIMA. Khi thời gian dự báo bắt đầu tăng lên, khả năng dự báo của mô hình giảm dần. 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 14,57; 7,81; 16,3%. Khi tăng khả năng dự báo cho 1 và 2 tuần tiếp theo, mô hình LSTM-ARIMA vẫn đạt độ chính xác trong khả năng dự báo khi sai lệch MAPE trong kết quả từ dự báo và tập kiểm tra ở mức 19,23% và 20,39%, sai số trung bình tuyệt đối ghi nhận được mặc dù có tăng nhưng đều ở mức chấp nhận. Kết quả cho thấy được mô hình lai LSTM-ARIMA đã nắm bắt tốt biến động từ quá khứ của chỉ số chất lượng không khí để có thể dự báo trong tương lai.

Để có so sánh một cách trực quan giữa mô hình lai LSTM-ARIMA và mô hình LSTM, các kết quả dự báo như sau: Hình 8 cho thấy đường dự đoán của mô hình LSTM-ARIMA và ARIMA sát với đường thực tế và nắm bắt được những điểm và đáy dao động ngắn. Hình 9 cho thấy đường dự đoán của mô hình LSTM-ARIMA và ARIMA vẫn sát với đường thực tế, tuy



Hình 7: Độ mất mát từ quá trình training 70% qua số epochs huấn luyện cho mô hình LSTM và LSTM-ARIMA [Nguồn: Nhóm tác giả]



Hình 8: Mô phỏng kết quả dự báo chất lượng không khí chu kỳ 1 ngày [Nguồn: Nhóm tác giả]

nhiên các đáy và đỉnh các thời điểm không nắm bắt được chính xác trong dự báo chỉ số AQI.

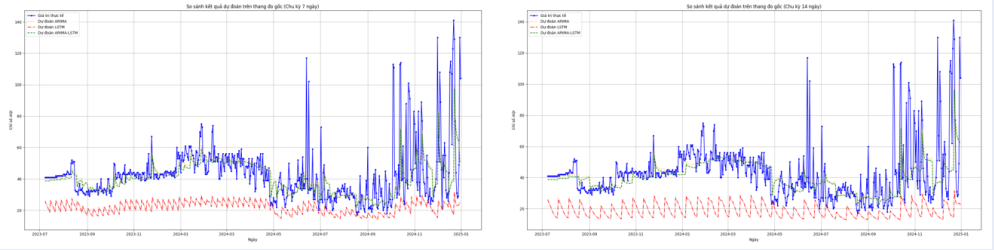
Kết quả dự báo chất lượng không khí chu kỳ 14 ngày trên tập dữ liệu kiểm tra ta thấy ARIMA và LSTM-ARIMA vẫn bám sát với đường thực tế tuy nhiên việc bám sát với giá trị AQI có dấu hiệu giảm dần; Kết quả dự báo chất lượng không khí 1 ngày, 7 ngày và 14 ngày trên tập dữ liệu kiểm tra ta thấy ARIMA và LSTM-ARIMA luôn bám sát với đường thực tế và nắm bắt được xu thế tuy nhiên việc nắm bắt các đáy, đỉnh giảm dần qua các chu kỳ.

Theo các chỉ số đánh giá, mô hình ARIMA cho kết quả ổn định nhất với RMSE thấp (14,37–16,41), trong khi LSTM có sai số cao hơn (24,60-30,17), cho thấy hạn chế khi dự báo dữ liệu AQI biến động mạnh. Tuy nhiên, mô hình lai LSTM–ARIMA cho kết quả tiệm cận, thậm chí vượt ARIMA ở chu kỳ 1 ngày, chứng minh tính hiệu quả khi kết hợp khả năng xử lý tuyến tính của ARIMA và phi tuyến của LSTM. Với chỉ số MAE, ARIMA tiếp tục đạt sai số thấp nhất (7,63-9,25), LSTM cao nhất (18,26-24,24), còn LSTM–ARIMA đạt giá trị gần ARIMA và cải thiện rõ rệt so với LSTM. Ở chỉ số MAPE, LSTM–ARIMA cho kết quả tốt nhất (16,3%-20,39%), trong khi LSTM có sai số cao (36,9%-51,32%), thể hiện sự kém ổn định với dữ liệu AQI. Nhìn chung, mô hình lai LSTM–ARIMA cho hiệu quả vượt trội, duy trì độ ổn định của ARIMA, tận dụng ưu thế học phi tuyến của LSTM và nâng cao độ chính xác trong dự báo chất lượng không khí.

Sự khác biệt giữa các chỉ số RMSE và MAPE xuất phát từ bản chất của từng thước đo sai số cũng như đặc điểm phân bố giá trị của chỉ số AQI. Cụ thể, RMSE phản ánh sai số tuyệt đối và nhạy với các giá trị AQI lớn, do đó mô hình ARIMA có thể đạt RMSE thấp hơn nhẹ khi mô tả tốt xu hướng tổng thể và các giá trị AQI cao. Ngược lại, MAPE là thước đo sai số tương đối, phụ thuộc lớn vào giá trị AQI thực tế. Tại các thời điểm có giá trị AQI thấp (thường < 30), chỉ một sai lệch tuyệt đối nhỏ cũng có thể dẫn đến sai số phần trăm lớn. Ngoài ra, với đặc tính làm mượt và cấu trúc tuyến tính, mô hình ARIMA có xu hướng dự báo kém chính xác hơn tại các thời điểm có giá trị AQI thấp và biến động nhanh, từ đó làm gia tăng MAPE. Trong khi đó, mô hình lai LSTM–ARIMA có khả năng nắm bắt tốt hơn các dao động phi tuyến và biến động ngắn hạn nhờ thành phần LSTM, giúp giảm sai số tương đối tại các thời điểm AQI thấp. Vì vậy, mặc dù RMSE của mô hình lai thấp hơn không đáng kể, nhưng MAPE lại được cải thiện rõ rệt, phản ánh độ chính xác tương đối tốt hơn trên toàn dải giá trị AQI.

So với mô hình ARIMA, mô hình lai LSTM–ARIMA đòi hỏi tài nguyên tính toán lớn hơn trong giai đoạn huấn luyện do cấu trúc mạng nơ-ron và số lượng nơ-ron ẩn được sử dụng. Ngoài ra, cấu hình 100 nơ-ron ẩn được lựa chọn nhằm đạt sự cân bằng giữa độ chính xác và độ ổn định của mô hình; trong các ứng dụng thực tế, cấu hình này có thể được điều chỉnh hoặc mô hình có thể được triển khai trên máy chủ trung tâm để phục vụ nhiều trạm quan trắc.

Các kết quả đạt được từ mô hình lai LSTM–ARIMA cho thấy sự cải thiện rõ rệt về độ chính xác dự báo, đặc biệt là đối với sai số tương đối trong dự báo ngắn



Hình 9: Mô phỏng kết quả dự báo chất lượng không khí chu kỳ 7 ngày và 14 ngày [Nguồn: Nhóm tác giả]

hạn. Kết quả này trực tiếp giải quyết khoảng trống nghiên cứu đã nêu trong phần Đặt vấn đề, nơi các mô hình ARIMA đơn lẻ thường gặp hạn chế trong việc mô tả các biến động phi tuyến và cho sai số lớn tại các thời điểm có AQI thấp, trong khi các mô hình LSTM thuần túy có xu hướng thiếu ổn định khi dữ liệu có tính nhiễu cao. Việc tách rõ thành phần tuyến tính bằng ARIMA và mô hình hóa phần dư phi tuyến bằng LSTM đã giúp cải thiện khả năng nắm bắt các dao động ngắn hạn của AQI, qua đó làm giảm MAPE một cách đáng kể. Những phân tích này giúp làm nổi bật vai trò của mô hình lai không chỉ về mặt cải thiện độ chính xác, mà còn trong việc giải quyết các hạn chế cốt lõi của các nghiên cứu hiện có về dự báo chất lượng không khí.

Có thể thấy rằng độ chính xác của mô hình giảm dần khi dự báo dài hạn do nhiều nguyên nhân. Thứ nhất, sai số tích lũy qua từng bước dự báo khiến tổng sai số tăng, làm giảm độ tin cậy của kết quả. Mặc dù mô hình lai LSTM-ARIMA kết hợp được ưu điểm của ARIMA trong xử lý xu hướng tuyến tính và LSTM trong học phi tuyến, phụ thuộc dài hạn, nhưng vẫn chịu ảnh hưởng lớn bởi sai số tích lũy này. Thứ hai, khả năng nắm bắt các mẫu dài hạn bị hạn chế, vì khi thời gian dự báo kéo dài, thông tin ban đầu dần bị "loãng", khiến LSTM khó duy trì mối quan hệ phụ thuộc, trong khi ARIMA mất khả năng mô hình hóa chính xác các biến động phi tuyến hoặc chu kỳ mới. Thứ ba, đặc tính chuỗi thời gian có thể thay đổi theo thời gian như xuất hiện xu hướng mới, biến động bất thường hoặc các yếu tố môi trường đột biến, khiến mô hình huấn luyện từ dữ liệu quá khứ không thích nghi kịp. Cuối cùng, việc kết hợp ARIMA và LSTM thường tạo hiệu ứng mượt hóa, giúp giảm nhiễu và tránh overfitting nhưng đồng thời làm giảm độ nhạy của mô hình với các biến động mạnh, dẫn đến dự báo dài hạn trở nên kém chính xác hơn, đặc biệt trong các giai đoạn xuất hiện biến động hoặc đột biến lớn.

So sánh tương quan hiệu suất giữa mô hình LSTM-ARIMA trong nghiên cứu này và một số mô hình kết hợp khác đã được ứng dụng (Bảng 2) cho thấy mô hình LSTM-ARIMA vẫn đạt các chỉ số RMSE và MAE thấp hơn so với các mô hình lai khác vì thế có thể được ứng dụng trong dự đoán trong các chu kỳ.

Ngoài ra, mặc dù nghiên cứu được thực hiện tại quy mô một trạm quan trắc và tập trung vào dự báo chỉ số AQI, các kết quả thu được vẫn mang lại những đóng góp mới đối với hiểu biết về cơ chế biến động chất lượng không khí. Cụ thể, việc mô hình lai LSTM–ARIMA cải thiện đáng kể độ chính xác dự báo, đặc biệt trong các giai đoạn AQI thấp và biến động nhanh, cho thấy vai trò quan trọng của các quá trình phi tuyến, điển hình là các phản ứng quang hóa hình thành O_3 và sự tương tác với điều kiện khí tượng. Điều này củng cố nhận định rằng O_3 là tác nhân chi phối chính đối với AQI tại khu vực nghiên cứu, đồng thời giải thích vì sao các mô hình tuyến tính đơn lẻ thường gặp hạn chế trong dự báo chất lượng không khí ngắn hạn.

KẾT LUẬN

Từ những tính toán và thí nghiệm đã thực hiện, nghiên cứu đã đánh giá được hiện trạng chất lượng không khí tại trạm ngã tư Giếng nước ở phường Vũng Tàu, TP.HCM cho giai đoạn năm 2020-2024, cho thấy chất lượng không khí tại khu vực nghiên cứu là tương đối tốt khi giá trị các thông số như CO, NO_2, SO_2, $PM_{2.5}$ và PM_{10} đều dưới ngưỡng cho phép trong cả thời gian nghiên cứu. Nồng độ Ozon trung bình 8 giờ có 866 thời điểm được ghi nhận vượt quy chuẩn cho phép từ 1,1 đến 1,8 lần. Ozon trung bình 1 giờ có 74 thời điểm giá trị quan trắc được vượt quy chuẩn cho phép từ 1,1 đến 1,6 lần; Mô hình lai LSTM-ARIMA đã được xây dựng thành công với khả năng dự báo có độ chính xác cao cho thời gian 1 ngày tiếp theo với các giá trị RMSE, MAE và MAPE tương ứng là 14,57; 7,81 và 16,3%. Khi dự báo trong thời gian dài hơn tương ứng với 7 ngày hay 14 ngày trong tương lai, mô hình vẫn đưa ra kết quả khả thi. Ngoài ra, kết quả nghiên cứu cho thấy mô hình LSTM-ARIMA có hiệu suất dự báo

Bảng 2: So sánh hiệu suất giữa các mô hình về dự báo chỉ số AQI

Chỉ số đánh giá	Mô hình	Thời gian dự báo	RMSE	MAE
Trong nghiên cứu này	LSTM	1 ngày tiếp theo	24,57	18,26
		7 ngày tiếp theo	27,27	21,19
		14 ngày tiếp theo	30,17	24,24
	LSTM-ARIMA	1 ngày tiếp theo	14,57	7,81
		7 ngày tiếp theo	16,7	8,97
		14 ngày tiếp theo	16,6	9,33
Hồ Minh Dũng và cộng sự, 2024 [21]	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
	CEEMDAN-LSTM	1 ngày tiếp theo	16,00	9,85
		14 ngày tiếp theo	23,14	16,17
Rui Yan và cộng sự, 2021 [12]	LSTM	1h tiếp theo	23,7	-
		6h tiếp theo	49,2	-
	CNN-LSTM	1h tiếp theo	22,9	-
		6h tiếp theo	47,1	-
Jiahui Duan và cộng sự, 2023 [30]	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

[Nguồn: Nhóm tác giả]

tốt hơn so với các mô hình đơn. Từ góc độ quản lý, khả năng dự báo ổn định của mô hình trong các chu kỳ 1–14 ngày cho phép cung cấp thông tin sớm phục vụ cảnh báo chất lượng không khí, hỗ trợ cơ quan quản lý trong việc ban hành các biện pháp giảm thiểu tác động đến sức khỏe cộng đồng, đặc biệt trong các giai đoạn có nguy cơ gia tăng nồng độ O_3. Ngoài ra, cách tiếp cận mô hình lai được đề xuất trong nghiên cứu này có thể mở rộng và áp dụng cho các trạm quan trắc khác trên địa bàn khu vực Bà Rịa-Vũng Tàu để có bức tranh chung về dự báo chất lượng không khí xung quanh khu vực Bà Rịa-Vũng Tàu, cũng như các trạm quan trắc khác góp phần xây dựng các hệ thống dự báo linh hoạt, phù hợp với điều kiện dữ liệu thực tế tại các đô thị đang phát triển.

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 DS.C2025-24-01.

DANH MỤC CÁC TỪ VIẾT TẮT

ACF (Autocorrelation Function): Hàm tự tương quan

AERMOD: The ASM/EPA Regulatory Model.

AI (Artificial Intelligence): Trí tuệ nhân tạo

AIC (Akaike Information Criterion): Tiêu chuẩn thông tin Akaike

ANN (Artificial Neural Network): Mạng nơ-ron nhân tạo

AQI (Air Quality Index): Chỉ số chất lượng không khí

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

BIC (Bayesian Information Criterion): Tiêu chuẩn thông tin Bayesian

CEEMDAN (Complete Ensemble Empirical Mode Decomposition with Adaptive Noise): Kỹ thuật phân tích tín hiệu giúp tách chuỗi thời gian thành các thành phần khác nhau

CMAQ: Community Multiscale Air Quality

CNN (Convolutional Neural Network): Mạng nơ-ron tích chập

DBO (Dung Beetle Optimizer): Tối ưu hóa bọ hung

IQR (Interquartile Range): Khoảng tứ phân vị

GWR (Geographically Weighted Regression): Hồi quy trọng số địa lý
KNN: K-Nearest Neighbors
LSTM (Long Short-Term Memory): Mạng nơ-ron hồi quy dài hạn ngắn hạn
MA (Moving Average): Trung bình trượt
MAE (Mean absolute error): Độ lệch tuyệt đối
MAPE (Mean Absolute Percentage Error): Sai số phần trăm tuyệt đối trung bình
MLR (Multiple Linear Regression): Hồi quy tuyến tính bội
NAR (Nonlinear Autoregressive Model): Mô hình Tự hồi quy Phi tuyến
NAQPMS: Nested Air Quality Prediction Modeling System
PACF (Partial Autocorrelation Function): Hàm tự tương quan riêng phần
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
TP.HCM: Thành phố Hồ Chí Minh
SVR (Support Vector Regression): Hồi quy vectơ hỗ trợ
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 Tôn Minh Hiền, được bổ sung nội dung bởi Hồ Minh Dũng và được chỉnh sửa hình ảnh, bảng biểu bởi Hồ Minh Dũng và Tôn Minh Hiền.

TÀI LIỆU THAM KHẢO

1. Jiao ZWY, Zhang Y. Prediction of Air Quality Index Based on LSTM. In: 2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC 2019); 2019.
2. 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. Environ Pollut. 2017;231(Pt 1):997–1004. Available from: https://doi.org/10.1016/j.envpol.2017.08.114.
3. Kumar VKJU. ARIMA forecasting of ambient air pollutants (O3, NO, NO2 and CO). 2009;(24):751–760. Available from: https://doi.org/10.1007/s00477-009-0361-8.
4. Pablo E. Saide GRC, Scott N. Spak, Laura Gallardo, Axel E. Osses, Marcelo A. Mena-Carrasco, Mariusz Pagowski Forecasting urban PM10 and PM2.5 pollution episodes in very stable nocturnal conditions and complex terrain using WRFCO tracer model. Atmos Environ. 2011;45(16):2769–80. Available from: https://doi.org/10.1016/j.atmosenv.2011.02.001.
5. Qi ZML, Lin W, Pan X, Xiao T, Wang Z. Deep Learning for Air Quality Forecasts: a Review. Curr Pollut Rep. 2020;6(4):399–409. Available from: https://doi.org/10.1007/s40726-020-00159-z.
6. Seng QZD, Zhang X, Chen G, Chen X. Spatiotemporal prediction of air quality based on LSTM neural network. Alex Eng J. 2021;60(2):2021–32. Available from: https://doi.org/10.1016/j.aej.2020.12.009.
7. Mallet BSV. Air quality modeling: from deterministic to stochastic approaches. Comput Math Appl. 2008;55(10):2329–37. Available from: https://doi.org/10.1016/j.camwa.2007.11.004.
8. Yammahi ZAAA. Forecasting the concentration of NO2 using statistical and machine learning methods: A case study in the UAE. Heliyon. 2023;Available from: https://doi.org/10.1016/j.heliyon.2022.e12584.
9. Unjin KCP. Ryu Unsok, Sok Kyongjin, Pak Sungnam. A hybrid model based on convolutional neural networks and long short-term memory for ozone concentration prediction. Air Qual Atmos Health. 2018;11(8):883–95. Available from: https://doi.org/10.1007/s11869-018-0585-1.
10. Greff RKK, Koutn'ık J, Steunebrink BR, Schmidhuber J. LSTM: A Search Space Odyssey. Transactions on Neural Networks and Learning Systems; 2018. Available from: https://doi.org/10.1109/TNNLS.2016.2582924.
11. Sagar V, Belavadi SR, Ranjani R. Rajasekar Mohan. Air Quality Forecasting using LSTM RNN and Wireless Sensor Networks. Procedia Comput Sci. 2020;170:241–8. Available from: https://doi.org/10.1016/j.procs.2020.03.036.
12. Rui LJY, Jie Y, Wei S, Nong M, Li F. Multi-hour and multi-site air quality index forecasting in Beijing using CNN, LSTM, CNN-LSTM, and spatiotemporal clustering. Expert Syst Appl. 2021;169:169. Available from: https://doi.org/10.1016/j.eswa.2020.114513.
13. Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Netw. 2005;18(5-6):602–10. Available from: https://doi.org/10.1016/j.neunet.2005.06.042.
14. Navares R, Aznarte JL. Predicting air quality with deep learning LSTM: towards comprehensive models. Ecol Inform. 2020;55:55. Available from: https://doi.org/10.1016/j.ecoinf.2019.101019.
15. Chang HTYS. Satheesh Abimannan, Yo-Ping Huang, Yi-Ting Tsai, Kuan-Ming Lin. An LSTM-based aggregated model for air pollution forecasting. Atmos Pollut Res. 2020;11(8):1451–63. Available from: https://doi.org/10.1016/j.apr.2020.05.015.
16. Wen SLC, Yao X, Peng L, Li X, Hu Y. Tianhe Chi A novel spatiotemporal convolutional long short-term neural network for air pollution prediction. Sci Total Environ. 2018;p. 654. Available from: https://doi.org/10.1016/j.scitotenv.2018.11.086.
17. Sunil NSG. Khare Mukesh. Extreme Events of Reactive Ambient Air Pollutants and their Distribution Pattern at Urban Hotspots. Aerosol Air Qual Res. 2017;17(2):394–405. Available from: https://doi.org/10.4209/aaqr.2016.06.0273.
18. Thanh NC, Giang NT. Building LSTM (Long Short - Term Memory) machine learning model for water salinity forecasting in Dai Ngai. Vietnam Journal of Hydrometeorology. 2022;8:98–104.
19. Ha DT, Nghe NT. Application of long short-term memory multivariate model in temperature and rainfall forecasting. 2022;(58):8–16.
20. Hien LX, Hung HV. Using Long Short-Term Memory Neural Network to Forecast Water level at The Quang Phuc and the Cua Cam stations in Hai Phong, Vietnam. 2018;(62).
21. Dung HM, Khang KD. Forecasting air quality by the LSTM-MA model, using data at the Gieng Nuoc intersection automatic monitoring station, Ba Ria - Vung Tau province. J Hydrometeorol. 2024;9:75–89.
22. Cao MQ, Pham HH. Predicting air pollution from monitoring data using artificial intelligence models. Transport and Communications Science Journal. 2024;75:1404–12.
23. Hung NV, Nam VH, Anh VD. Air quality index prediction using convolutional neural network and long short-term memory. Journal of Science and Technology on Information and Communications. 2021;4:129–36.

24. Phuc LB, Trieu DV. Study on the artificial intelligence model of LSTM - NA in the air quality forecast for deep open pit coal mines. Mining Science and Technology Information; 2022.
25. Rakholia QLR. Khue Vu, Bang Quoc Ho, Ricardo Simon Carbajo. AI-based air quality PM2.5 forecasting models for developing countries: A case study of Ho Chi Minh City, Vietnam. Urban Clim. 2022;46. Available from: https://doi.org/10.1016/j.uclim.2022.101315.
26. trường TCM. Quyết định số 1459/QĐ-TCMT, ngày 12/11/2019 về việc ban hành Hướng dẫn kỹ thuật tính toán và công bố chỉ số chất lượng không khí Việt Nam (VN_AQI). 2019;.
27. Liao MZQ. Deep Learning for Air Quality Forecasts: a Review. Beijing: Institute of Atmospheric Physics; 2020. Available from: https://doi.org/10.1007/s40726-020-00159-z.
28. Pak CKU. Unsok Ryu, Kyongjin Sok. A hybrid model based on convolutional neural networks and long short-term memory for ozone concentration prediction; 2018. Available from: https://doi.org/10.1007/s11869-018-0585-1.
29. Yan R, Liao J, Yang J, Sun W, Nong M, Li F. Multi-hour and multi-site air quality index forecasting in Beijing using CNN, LSTM, CNN-LSTM, and spatiotemporal clustering. Expert Syst Appl. 2021;169:169. Available from: https://doi.org/10.1016/j.eswa.2020.114513.
30. Duan J, Gong Y, Luo J, Zhao Z. Air-quality prediction based on the ARIMA-CNN-LSTM combination model optimized by dung beetle optimizer. Sci Rep. 2023;13(1):12127. Available from: https://doi.org/10.1038/s41598-023-36620-4.

VNUHCM Journal of Environment and Earth Sciences 2026, 10(2): 1454-1467

Research article

Open Access Full Text Article

BUILD OF LSTM-ARIMA HYBRID MODEL TO PREDICTE AIR QUALITY AT GIENG NUOC INTERSECTION STATION, HCMC

Ho Minh Dung[1,*], Ton Minh Hien[2]

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

ABSTRACT

Air pollution is one of the main causes of serious health problems such as lung and cardiovascular diseases and increases the risk of other diseases. Therefore, early forecasting of air quality plays an important role in warning the community and helping people proactively take health protection measures. In this study, the LSTM model was selected due to its ability to identify nonlinear relationships in data, suitable for both short-term and long-term forecasting. In order to improve efficiency, the LSTM model was combined with ARIMA to form a hybrid model LSTM-ARIMA, which brought positive forecasting results. The values of CO, NO_2, SO_2 and PM_{10} parameters at the Gieng Nuoc intersection station (Vung Tau ward, Ho Chi Minh City) were always below the allowable threshold of QCVN 05:2023/BTNMT. The $PM_{2.5}$ concentration was mostly at a safe level, only at one point it almost reached the standard threshold. However, the 8-hour average O_3 concentration exceeded the standard at 866 times, and the 1-hour O_3 exceeded the standard at 74 times. The air quality at the Gieng Nuoc intersection monitoring station is mainly in the ``Good'' and ``Average'' groups according to the AQI index. The LSTM-ARIMA model gives 1-day forecast results with high accuracy (RMSE = 14.57; MAE = 7.81; MAPE = 16.3%) and maintains stable performance when forecasting is extended to 7 and 14 days, showing potential for application in air quality forecasting. The research results show that the LSTM-ARIMA model has better forecasting performance than single models, and can be scaled up and applied to other observation stations, contributing to the development of flexible forecasting systems that are suitable for the actual observational data conditions in developing cities.

Key words: Air quality, LSTM-ARIMA, Gieng Nuoc intersection station

[1]*Institute for Environment and Resources, VNU-HCM, 142 To Hien Thanh Street, Dien Hong Ward, Ho Chi Minh City, Vietnam*

[2]*HCMC Science of Occupational Safety and Health Institute, 348 Nguyen Thai Son Street, Hanh Thong Ward, Ho Chi Minh City, Vietnam*

Correspondence

Ho Minh Dung, Institute for Environment and Resources, VNU-HCM, 142 To Hien Thanh Street, Dien Hong Ward, Ho Chi Minh City, Vietnam

Email: H_minhdung@yahoo.com

History

- Received: 11-09-2025
- Revised: 04-03-2026
- Accepted: 03-08-2026
- Published Online: 25-09-2026

DOI : https://doi.org/10.32508/vnuhcmjees.v10i2.858

Check for updates

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 : Dung H M, Hien T M. **BUILD OF LSTM-ARIMA HYBRID MODEL TO PREDICTE AIR QUALITY AT GIENG NUOC INTERSECTION STATION, HCMC**. *VNUHCM J. Environ. Earth Sci. 2026; 10(2):1454-1467.*