TẠP CHÍ
KHÍ TƯỢNG THỦY VĂN

Bài báo khoa học

Mô phỏng nồng độ bụi $PM_{2.5}$ tại khu vực trung tâm Thành phố Hồ Chí Minh bằng thuật toán học máy và học sâu

Nguyễn Phúc Hiếu[1], Nguyễn Nhật Dương[1], Đỗ Quang Lĩnh[1], Đào Nguyên Khôi[1]*

[1] Khoa Môi trường, Trường ĐH Khoa học tự nhiên, ĐHQG-HCM;
phuchieu50@gmail.com; 19170139@student.hcmus.edu.vn; dqlinh@hcmus.edu.vn; dnkhoi@hcmus.edu.vn

*Tác giả liên hệ: dnkhoi@hcmus.edu.vn; Tel.: +84–088304379

Ban Biên tập nhận bài: 15/2/2024; Ngày phản biện xong: 20/3/2024; Ngày đăng bài: 25/6/2024

Tóm tắt: Nghiên cứu áp dụng ba thuật toán học máy *Random Forest Regression (RFR), XGBoost Regression (XGBR), Multilayer Perceptron Regression (MLPR)* và một thuật toán học sâu *Convolutional Neural Network (CNN)* để mô phỏng nồng độ bụi $PM_{2.5}$ tại khu vực trung tâm Thành phố Hồ Chí Minh. Bộ dữ liệu được sử dụng trong nghiên cứu là dữ liệu ngày trong giai đoạn từ 2016-2021 bao gồm nồng độ bụi $PM_{2.5}$ thu thập từ trạm Lãnh Sự Quán Mỹ và sáu thông số khí tượng bao gồm nhiệt độ trung bình, hướng gió, tốc độ gió, độ ẩm, số giờ nắng và lượng mưa tại trạm Tân Sơn Hòa. Bộ dữ liệu được chuẩn hóa và phân chia với tỷ lệ 80:20 phục vụ quá trình huấn luyện và kiểm tra các thuật toán. Sau đó, sáu kịch bản các thông số đầu vào khác nhau được xây dựng dựa trên kết quả phân tích tương quan riêng phần giữa các thông số khí tượng với nồng độ bụi $PM_{2.5}$. Kết quả nghiên cứu cho thấy cả ba thuật toán học máy đều có khả năng mô phỏng tốt nồng độ $PM_{2.5}$ với giá trị hệ số tương quan r dao động trong khoảng 0,770 đến 0,854, trong đó thuật toán XGBR với sáu thông số khí tượng đầu vào cho hiệu quả mô phỏng tốt nhất với r = 0,854, IOA = 0,922 và NMB = 6,711. Bên cạnh đó, kết quả mô phỏng nồng độ $PM_{2.5}$của thuật toán CNN là chưa đạt với giá trị r nhỏ hơn 0,5 ở tất cả kịch bản mô phỏng.

Từ khóa: Bụi $PM_{2.5}$; Học máy; Học sâu; TP. Hồ Chí Minh.

1. Đặt vấn đề

Hiện nay, ô nhiễm không khí đã trở thành một trong những vấn đề môi trường có ảnh hưởng lớn đến sức khỏe cộng đồng, đặc biệt dưới ảnh hưởng của quá trình đô thị hóa và công nghiệp hóa [1–6]. Theo Tổ chức Y tế Thế giới (WHO), có 7 triệu ca tử vong sớm do ô nhiễm không khí cả bên ngoài và trong nhà trên toàn cầu mỗi năm [7]. Đặc biệt là ô nhiễm do bụi $PM_{2.5}$, đang trở thành một trong những vấn đề tác động tiêu cực đối với sức khỏe toàn cầu, trong đó có Việt Nam [8]. Bụi $PM_{2.5}$ được định nghĩa là các hạt bụi mịn có đường kính nhỏ hơn 2,5 µm [9]. Một số nghiên cứu đã thực hiện [10–13] cho thấy mối liên hệ chặt chẽ giữa nồng độ $PM_{2.5}$ và các bệnh như ung thư, tim mạch, hô hấp, chuyển hóa và béo phì. Tại Việt Nam, nồng độ bụi $PM_{2.5}$ năm 2021 cao thứ 36 trong 117 quốc gia [14] và mức độ ô nhiễm bụi $PM_{2.5}$ cũng thể hiện sự phân hóa theo mức độ đô thị hóa. Thành phố Hồ Chí Minh (TP.HCM) là trung tâm kinh tế của khu vực phía nam, cùng với sự phát triển kinh tế, thành phố có số dân cao nhất cả nước với mật độ dân số 4.375 người/km^2 (năm 2021) đang phải đối mặt với nguy cơ ảnh hưởng sức khỏe người dân do ô nhiễm không khí. Theo kết quả thống kê [15–16], 12/24 quận/huyện ở TP.HCM có nồng độ bụi $PM_{2.5}$ năm 2020 vượt quy chuẩn QCVN 05:2013/ BTNMT. Đến năm 2021, mặc dù hầu hết các quận huyện có nồng độ

$PM_{2.5}$ nằm trong ngưỡng cho phép về chất lượng không khí theo quy chuẩn quốc gia, tuy nhiên các giá trị này vẫn lớn hơn so với mức khuyến nghị của WHO. Bên cạnh đó, số ca tử vong sớm do phơi nhiễm $PM_{2.5}$ năm 2019 tại TP. Hồ Chí Minh là 4.130 ca, đứng thứ hai cả nước, tập chủ yếu tại những quận trung tâm thành phố, lớn nhất tại quận Bình Tân với 370 ca, theo sau là quận Gò Vấp, huyện Bình Chánh và Quận 12 (khoảng 280-320 ca). Nồng độ bụi $PM_{2.5}$ cao ở các quận trung tâm, thấp ở các huyện như Củ Chi, Cần Giờ [15–16]. Vì vậy, việc mô phỏng và dự báo nồng độ bụi $PM_{2.5}$ tại khu vực trung tâm TP.HCM (Hình 1) là cần thiết nhằm phục vụ cho công tác quản lí và kiểm soát ô nhiễm, cũng như giảm thiểu rủi ro gây ra do ô nhiễm bụi $PM_{2.5}$.

Có thể thấy, công tác mô phỏng và dự báo chất lượng không khí có vai trò quan trọng trong việc ứng phó với ô nhiễm khí và bảo vệ sức khỏe con người. Tuy nhiên, việc dự báo chất lượng không khí là khá phức tạp và bị chi phối bởi nhiều yếu tố, trong đó có điều kiện khí tượng và tải lượng phát thải. Hiện nay, các phương pháp nghiên cứu dự đoán ô nhiễm không khí chủ yếu bao gồm phương pháp mô hình số và phương pháp thống kê [17, 18]. Tuy nhiên, phương pháp mô hình số thường đòi hỏi nhiều dữ liệu và người dùng cần hiểu biết sâu sắc về cơ chế lan truyền và bản chất của các chất gây ô nhiễm không khí để có thể lựa chọn các sơ đồ vật lý và hóa học phù hợp được sử dụng trong cấu hình của mô hình [19]. Phương pháp thống kê thì tương đối đơn giản, tiết kiệm thời gian và tài nguyên tính toán, và dễ thực hiện. Tuy nhiên, hiệu quả mô phỏng sẽ phụ thuộc vào số lượng các biến số và dữ liệu sẵn có, kết quả dự báo sẽ phụ thuộc rất nhiều vào mối tương quan giữa biến đầu ra và các yếu tố đầu vào. Bên cạnh đó, hiện nay với xu thế của cách mạng công nghiệp 4.0, đã có nhiều nghiên cứu sử dụng các thuật toán trí tuệ nhân tạo bao gồm cả học máy và học sâu nhằm tăng hiệu quả mô phỏng, dự đoán chất lượng không khí.

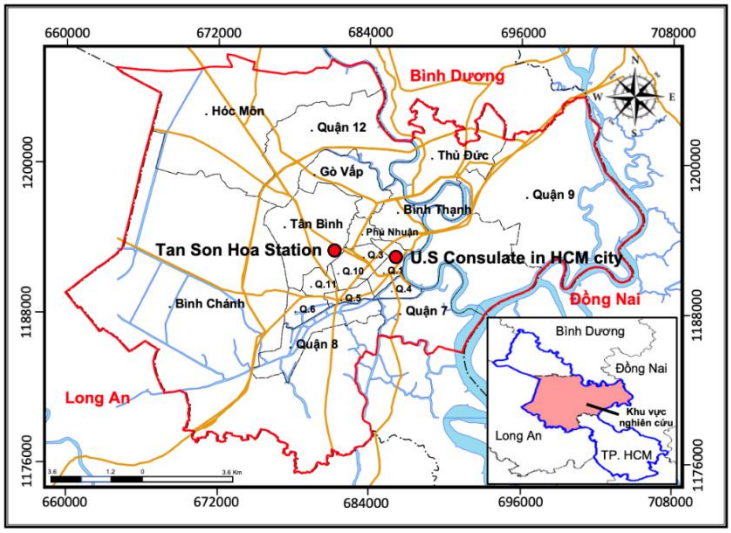
Một số nghiên cứu sử dụng thuật toán học máy để dự đoán bụi mịn đã được thực hiện có thể kể đến như nghiên cứu [20] đã sử dụng các thuật toán học máy *Random Forest (RF), eXtreme Gradient Boosting (XGBoost)*, và học sâu *Deep Neural Network (DNN)* để nghiên cứu dự đoán nồng độ $PM_{2.5}$ ở khu vực đô thị của Tehran, sử dụng bộ dữ liệu khí tượng từ trạm quan trắc và dữ liệu độ dày quang học sol khí (AOD) từ ảnh MODIS. Kết quả cho thấy thuật toán XGB cho khả năng dự báo tốt nhất trong 3 phương pháp. Nghiên cứu [21] dự báo nồng độ PM_{10} khu vực Caribe bằng sáu thuật toán học máy bao gồm: *Support Vector Machine (SVM), RFR, k-nearest Neighbor Regression (kNN), Gradient Boosting Regression (GBR), Tweedie Regression (TR)* và *Bayesian Ridge Regression (BRR)*. Các phương pháp này đã được áp dụng để xây dựng thuật toán dự đoán dựa trên mối quan hệ giữa nồng độ PM_{10} và các yếu tố thời tiết của khu vực nghiên cứu và kết quả cho thấy thuật toán GBR cho hiệu quả dự báo tốt nhất. Nghiên cứu [22] đã sử dụng các phương pháp học máy như RF, GBR, *Support Vector Regression (SVR)* và *Multilayer Regression (MLR)* để dự đoán PM_{10} và $PM_{2.5}$ ở Ma cao, Trung Quốc. Dữ liệu khí tượng và chất lượng không khí từ năm 2013 đến năm 2018 được sử dụng để dự đoán. Nghiên cứu này cho thấy RF là phương pháp dự đoán đáng tin cậy nhất về nồng độ chất ô nhiễm, thuật toán này cũng chứng minh được tính hiệu quả khi dự báo nồng độ $PM_{2.5}$ trong vùng Paso Del Norte với độ chính xác đạt 92% [23]. Bên cạnh các thuật toán học máy, các thuật toán học sâu như *Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN)* cũng được sử dụng trong nhiều nghiên cứu nhằm dự báo nồng độ bụi $PM_{2.5}$ [24–27].

Từ các nghiên cứu kể trên, có thể thấy các thuật toán học máy và học sâu được sử dụng phổ biến và có hiệu quả cao trong mô phỏng, dự báo nồng độ bụi mịn tại nhiều quốc gia trên thế giới. Nghiên cứu này sẽ sử dụng các thuật toán được đánh giá có hiệu quả tốt trong các nghiên cứu đã thực hiện là RF và XGB, bên cạnh đó, hai thuật toán học máy và học sâu phổ biến là MLP và CNN cũng được áp dụng để thử nghiệm trong nghiên cứu này. Ngoài ra, phần lớn các nghiên cứu đã thực hiện đều sử dụng các dữ liệu khí tượng để dự báo $PM_{2.5}$. Mục tiêu của nghiên cứu nhằm xác định thuật toán và bộ thông số tối ưu phục vụ mô phỏng nồng độ bụi $PM_{2.5}$ dựa trên dữ liệu về khí tượng tại khu vực trung tâm TP.HCM.

2. Dữ liệu và phương pháp nghiên cứu

2.1. Thu thập và xử lí dữ liệu

Bộ dữ liệu được sử dụng trong nghiên cứu là dữ liệu ngày trong giai đoạn từ 5/2/2016 đến 30/4/2021 bao gồm nồng độ bụi $PM_{2.5}$ trung bình ngày thu thập từ trạm Lãnh Sự Quán Mỹ và sáu thông số khí tượng bao gồm nhiệt độ trung bình (T), hướng gió (WD), tốc độ gió (W), độ ẩm (H), số giờ nắng (S) và lượng mưa trung bình (R) tại trạm Tân Sơn Hòa. Các thông số này được chọn dựa theo Báo cáo hiện trạng môi trường quốc gia 2021 của Bộ Tài nguyên và môi trường, nồng độ chất ô nhiễm trong không khí gần mặt đất phụ thuộc rất lớn vào yếu tố khí tượng (hướng gió, tốc độ gió, nhiệt độ, độ ẩm tương đối, lượng mưa), các yếu tố khí tượng có liên quan mật thiết đến sự hình thành, tích tụ và phân tán các chất ô nhiễm không khí và bụi vào môi trường xung quanh [7]. Vị trí các trạm được thể hiện trong Hình 1 và đặc trưng bộ dữ liệu đầu vào được thống kê và trình bày trong Bảng 1.



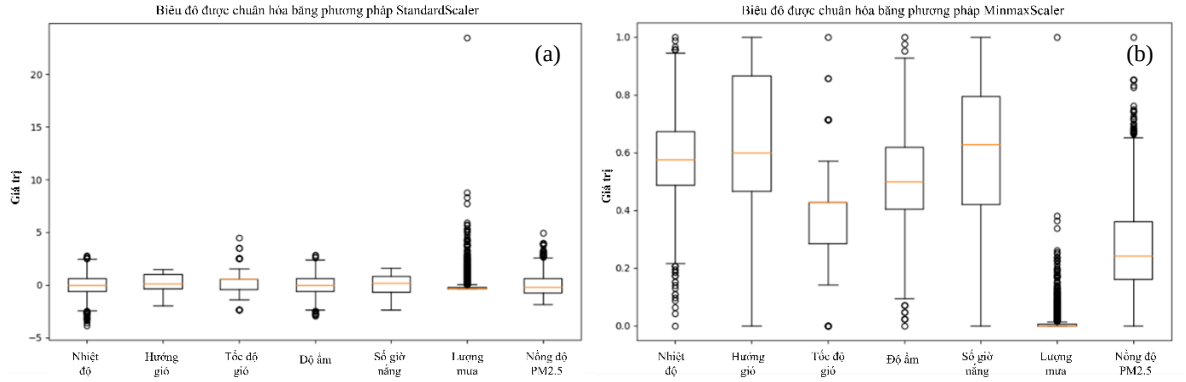
Hình 1. Khu vực nghiên cứu.

Bảng 1. Đặc trưng bộ dữ liệu đầu vào.

Biến	Đơn vị	Nhỏ nhất	Lớn nhất	Trung bình	Trung vị	Độ lệch chuẩn
Nhệt độ (T)	°C	23,4	32,6	28,73	28,7	1,38
Hướng gió (WD)	-	0,0	15,0	8,53	9,0	4,38
Tốc độ gió (W)	m/s	3,0	10,0	5,40	6,0	1,02
Độ ẩm (H)	%	51	93	72,31	72,0	7,40
Số giờ nắng (S)	h	0,0	10,2	5,98	6,4	2,57
Lượng mưa (R)	mm	0,0	366,0	5,02	0,0	15,38
Nồng độ PM2.5	µg/m3	5,0	77,65	24,75	22,63	10,73

Bộ dữ liệu trong nghiên cứu có sự chênh lệch về độ lớn cũng như không đồng nhất về đơn vị, do đó việc chuẩn hóa bộ dữ liệu cần được thực hiện nhằm đưa bộ dữ liệu về cùng khoảng giá trị phục vụ cho các phép tính trong thuật toán. Trong nghiên cứu này, phương pháp chuẩn hóa StandardScaler nhằm đưa giá trị trong bộ dữ liệu có trung bình bằng 0 và độ lệch chuẩn bằng 1 (Hình 2a), phương pháp này được áp dụng cho các thuật toán hồi quy như RFR, XGBR, MLPR và phương pháp MinMaxScaler nhằm đưa các giá trị về khoảng [0, 1] phục vụ tính toán cho thuật toán CNN (Hình 2b).

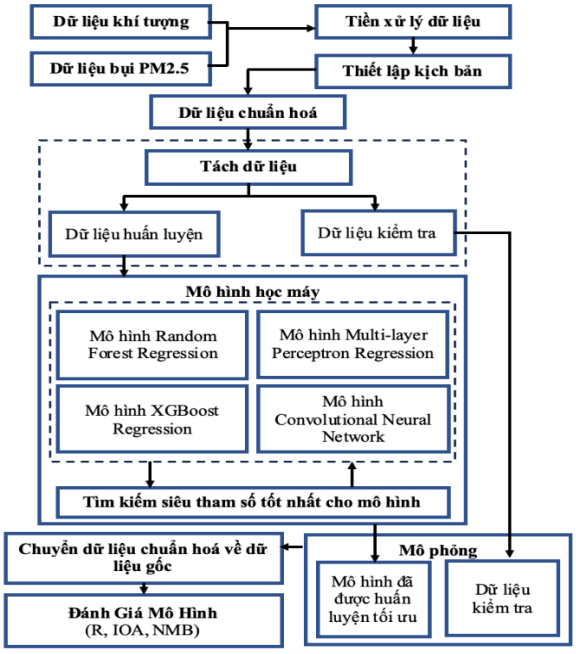
Sau khi đã chuẩn hóa dữ liệu, nghiên cứu tiến hành phân chia dữ liệu thành hai phần phục vụ quá trình huấn luyện và kiểm tra mô hình theo tỷ lệ 80:20, tức là 80% dữ liệu được sử dụng cho quá trình huấn luyện nhằm tìm bộ thông số tối ưu của mô hình và 20% dữ liệu phục vụ quá trình kiểm tra nhằm đánh giá hiệu quả, độ tin cậy của các mô hình.



Hình 2. Đặc trưng bộ dữ liệu sau khi chuẩn hóa: (a) Phương pháp StandardScaler, (b) Phương pháp MinMaxScaler.

2.2. Phương pháp nghiên cứu

Trong nghiên cứu này, các thuật toán Random Forest Regression (RFR), XGBoost Regression (XGBR), Multilayer Perceptron Regression (MLPR) và Convolutional Neural Network (CNN) sẽ được áp dụng để xây dựng mô hình mô phỏng nồng độ bụi $PM_{2.5}$. Hình 3 thể hiện quy trình thực hiện tổng quát nhằm xác định cấu trúc và bộ thông số mô hình tối ưu mô phỏng nồng độ bụi $PM_{2.5}$ tại khu vực trung tâm TP.HCM, các bước chính cụ thể như sau: (1) thu thập và tiền xử lý dữ liệu; (2) chuẩn hóa dữ liệu, (3) phân chia dữ liệu, (4) tính tương quan riêng phần giữa nồng độ $PM_{2.5}$ và các thông số khí tượng, từ đó xây dựng các kịch bản tính, (5) lựa chọn và xây dựng các thuật toán học máy và học sâu bằng ngôn ngữ lập trình Python, (6) huấn luyện và kiểm tra mô hình, xác định thuật toán và bộ thông số tối ưu mô phỏng nồng độ bụi $PM_{2.5}$ tại khu vực nghiên cứu.



Hình 3. Quy trình xây dựng mô hình.

2.2.1. Thuật toán Random Forest Regression (RFR)

Thuật toán RFR được giới thiệu bởi [28], là một phương pháp học tập tổng thể có giám sát hoạt động dựa trên cây quyết định. Thuật toán này có thể được sử dụng cho cả phân loại và hồi quy, rất linh hoạt và nhanh chóng. Nghiên cứu này sử dụng thuật toán hồi quy để mô phỏng nồng độ $PM_{2.5}$. Cách hoạt động của thuật toán RFR bao gồm các bước như sau: (1) Chọn ngẫu nhiên một số mẫu từ tập dữ liệu huấn luyện ban đầu để tạo ra các tập dữ liệu con

khác nhau; (2) Xây dựng một cây quyết định trên mỗi tập dữ liệu con; (3) Kết hợp mô phỏng của các cây quyết định bằng cách tính trung bình hoặc biểu quyết theo đa số các mô phỏng độc lập từ các cây quyết định.

2.2.2. Thuật toán XGBoost Regression (XGBR)

Thuật toán XGBR là một trong những thuật toán học máy phổ biến được sử dụng trong bài toán mô phỏng giá trị liên tục (hồi quy). Đây là một thuật toán học máy dựa trên kỹ thuật gradient boosting. Thuật toán XGBR sử dụng nhiều cây quyết định để học, trong đó mỗi cây quyết định được xây dựng dựa trên các trọng số của các cây trước đó. XGBR sử dụng các hàm mất mát để tối ưu hóa thuật toán, đồng thời áp dụng các kỹ thuật regularization để tránh hiện tượng quá khớp (overfiting) [29].

2.2.3. Thuật toán Multilayer Perceptron Regression (MLPR)

Thuật toán MLPR là một thuật toán mạng nơ-ron nhân tạo được sử dụng cho bài toán hồi quy, cấu trúc gồm lớp đầu vào, các lớp ẩn, và lớp đầu ra. Cụ thể, lớp đầu vào nhận các đặc trưng của dữ liệu đầu vào và chuyển chúng vào mạng nơ-ron. Số lượng nơ-ron trong lớp đầu vào phụ thuộc vào số lượng đặc trưng trong dữ liệu. Các lớp ẩn nằm giữa lớp đầu vào và lớp đầu ra. Mỗi lớp ẩn chứa một số lượng nơ-ron được chọn trước, số lượng và kích thước của các lớp ẩn có thể khác nhau tùy thuộc vào độ phức tạp của bài toán và khả năng học của thuật toán. Lớp đầu ra chứa một số lượng nơ-ron tương ứng với số lượng biến mục tiêu trong bài toán hồi quy, mỗi nút trong lớp đầu ra tính toán giá trị mô phỏng của biến mục tiêu.

Các nơ-ron trong các lớp đầu vào, lớp ẩn và lớp đầu ra kết nối với nhau thông qua các trọng số và hàm kích hoạt. Quá trình tính toán trong MLPR được thực hiện bằng cách lan truyền thuận, trong đó thông tin được truyền từ lớp đầu vào qua các lớp ẩn và cuối cùng đến lớp đầu ra để tạo ra dự đoán.

2.2.4. Thuật toán Convolutional Neural Network CNN

Thuật toán CNN đã được phát triển với bốn ý tưởng: trường tiếp nhận cục bộ (*Local receptive field*), trọng số chung (*Shared weights*), lấy mẫu con không gian (*Spatial subsampling*) và sử dụng nhiều lớp (*Pooling layer*). Một trong những lợi ích của mạng này là trọng số được chia sẻ giúp giảm số lượng tham số. Một thuật toán CNN điển hình bao gồm ba loại lớp: lớp tích chập (*Convolutional layer*), lớp lấy mẫu con (*Subsampling layer*) và lớp kết nối đầy đủ (*Fully connected layer*) [30].

Trong lĩnh vực môi trường, thuật toán CNN 1D có thể được áp dụng để phân tích và mô phỏng các dữ liệu liên quan đến môi trường như dữ liệu khí quyển, chất lượng không khí, và dữ liệu địa chất. Cụ thể, CNN 1D có khả năng xử lý các chuỗi dữ liệu không gian và thời gian, như dữ liệu về nồng độ ô nhiễm không khí theo thời gian, dữ liệu về thay đổi khí hậu, hay dữ liệu về sự biến đổi địa chất trong một khu vực.

2.3. Đánh giá hiệu quả mô phỏng

Hiệu quả mô phỏng của các mô hình được đánh giá bằng phương pháp đồ thị và phương pháp thống kê nhằm so sánh chất lượng và độ tin cậy của kết quả mô phỏng từ các mô hình với số liệu thực đo. Trong nghiên cứu này, các chỉ số được dùng để đánh giá độ hiệu quả của các mô hình bao gồm hệ số tương quan riêng phần pearson (r), chỉ số tương đồng (IOA) và độ lệch trung bình chuẩn hoá (NMB). Cách tính của từng chỉ số được trình bày lần lượt trong các công thức 1, công thức 2 và công thức 3. Tiêu chuẩn đánh giá hiệu quả mô phỏng của mô hình dựa trên ba chỉ số thống kê r, IOA và NMB được thể hiện ở Bảng 2.

$$R = \frac{\Sigma[(P_j-\overline{P})\times(O_j-\overline{O})]}{\sqrt{\Sigma(P_j-\overline{P})^2\times\Sigma(O_j-\overline{O})^2}} \text{ với, } -1 \le R \le 1 \quad (1)$$

$$IOA = 1 - \frac{\sum(P_j - O_j)^2}{\sum(|P_j - \bar{O}| + |O_j - \bar{O}|)^2} \text{ với}, 0 \le IOA \le 1 \quad (2)$$

$$NMB = \frac{\sum(P_j - O_j)}{\sum O_j} \times 100 \text{ với}, -100\% \le NMB \le +\infty \quad (3)$$

Bảng 2. Tiêu chuẩn đánh giá cho các mô hình thuật toán với ba chỉ số R, IOA và NMB [31].

Chỉ số thống kê	Mức tiêu chuẩn	$PM_{2.5}$
r	Mục tiêu	> 0,70
	Tiêu chuẩn	> 0,60
IOA	Mục tiêu	> 0,80
	Tiêu chuẩn	> 0,70
NMB	Mục tiêu	< ±10%
	Tiêu chuẩn	<±20%

Trong đó P_j và O_j là giá trị quan sát thứ j của giá trị mô phỏng và giá trị thực tế, $\bar{P}$ và $\bar{O}$ là giá trị trung bình của giá trị mô phỏng và giá trị thực tế.

2.4. Xây dựng kịch bản mô phỏng

Các kịch bản mô phỏng được xác định dựa trên mức độ tương quan giữa từng thông số khí tượng với nồng độ bụi $PM_{2.5}$, theo đó thông số nhiệt độ có mức độ tương quan cao nhất với r = 0,296, tiếp theo lần lượt là tốc độ gió, số giờ nắng, độ ẩm, hướng gió và lượng mưa là thông số có mức độ tương quan thấp nhất với nồng độ $PM_{2.5}$. Kết quả tính toán giá trị tương quan đối với từng thông số được trình bày trong Bảng 3. Các biến đầu vào của các kịch bản được xác định dựa trên số lượng biến đầu vào và mức độ tương quan từ cao đến thấp, theo đó 6 kịch bản mô phỏng được xây dựng và thể hiện trong Bảng 4.

Bảng 3. Hệ số tương quan Pearson (r) giữa các biến đầu vào với bụi $PM_{2.5}$.

Thông số	Nhiệt độ (T)	Hướng gió (WD)	Tốc độ gió (W)	Độ ẩm (H)	Số giờ nắng (S)	Lượng mưa (R)
r	0,296	0,072	0,282	0,108	0,152	0,066

Bảng 4. Các kịch bản mô phỏng.

Kịch Bản	Biến đầu vào
KB1	Nhiệt độ (T)
KB2	Nhiệt độ (T), Tốc độ gió (W)
KB3	Nhiệt độ (T), Tốc độ gió (W), Số giờ nắng (S)
KB4	Nhiệt độ (T), Tốc độ gió (W), Số giờ nắng (S), độ ẩm (H)
KB5	Nhiệt độ (T), Tốc độ gió (W), Số giờ nắng (S), độ ẩm (H), Hướng gió (WD)
KB6	Nhiệt độ (T), Tốc độ gió (W), Số giờ nắng (S), độ ẩm (H), Hướng gió (WD), Lượng mưa (R)

3. Kết quả và thảo luận

3.1. Đánh giá hiệu quả mô phỏng của mô hình

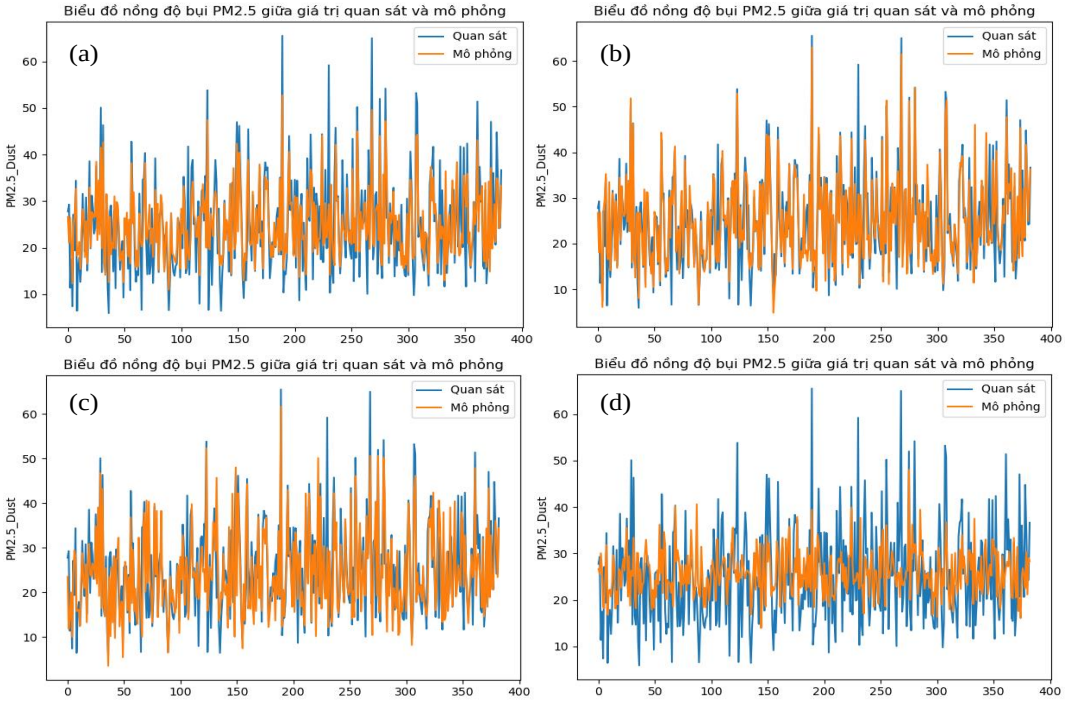
Kết quả đánh giá hiệu quả mô phỏng nồng độ bụi $PM_{2.5}$ theo 6 kịch bản bằng các thuật toán dựa theo các chỉ số r, IOA và NMB được thể hiện trong Bảng 5. Thuật toán có hiệu quả mô phỏng tốt nhất nghĩa là có sự khác biệt nhỏ nhất với nồng độ bụi $PM_{2.5}$ quan trắc.

Đối với thuật toán RFR, kết quả mô phỏng không tốt ở các kịch bản KB1 và KB2 với r thấp hơn 0,6, IOA thấp hơn 0,7 và NMB lớn hơn 10%, kịch bản KB3 đạt mức tiêu chuẩn. Ngược lại, hiệu quả mô phỏng đạt mức tốt ở các kịch bản KB4, KB5 và KB6 với r lớn hơn 0,7, IOA lớn hơn 0,8 và NMB bé hơn 10%, trong đó KB6 với 6 biến đầu vào cho hiệu quả mô phỏng tốt nhất với các giá trị r = 0,838, IOA = 0,887 và NMB = 9,078. Đồ thị so sánh nồng độ $PM_{2.5}$ quan trắc và mô phỏng với kịch bản tốt nhất KB6 được thể hiện trong Hình 4a.

Đối với thuật toán XGBR, các kịch bản với 3 thông số khí tượng đầu vào là nhiệt độ, tốc độ gió, số giờ nắng cho kết quả mô phỏng chưa đạt, trong khi đó hiệu quả mô phỏng đạt mức tốt ở các kịch bản KB4, KB5 và KB6 với R lớn hơn 0,7, IOA lớn hơn 0,8 và NMB bé hơn 10%, và kịch bản KB6 với 06 thông số đầu vào cho hiệu quả mô phỏng tốt nhất với các giá trị r = 0,854, IOA = 0,922 và NMB = 6,711 (Hình 4b).

Đối với thuật toán MLPR, hiệu quả mô phỏng không tốt ở các kịch bản KB1, KB2, KB3 và KB4 với r thấp hơn 0,6, IOA thấp hơn 0,7 và NMB lớn hơn 10%, và kịch bản KB6 đạt hiệu quả mô phỏng tốt nhất với r = 0,771, IOA = 0,875 và NMB = 3,217 (Hình 4c).

Đối với thuật toán CNN, kết quả cho thấy thuật toán CNN không thể mô phỏng tốt nồng độ bụi $PM_{2.5}$ tại khu vực nghiên cứu, cụ thể kết quả so sánh giữa nồng độ $PM_{2.5}$ quan trắc và mô phỏng cho thấy r thấp hơn 0,5, IOA thấp hơn 0,7 và NMB đều lớn hơn 10% ở tất cả các kịch bản. Ngoài ra, đồ thị thể hiện kết quả mô phỏng trong kịch bản KB6 (Hình 4d) cho thấy, giá trị nồng độ bụi $PM_{2.5}$ mô phỏng từ thuật toán CNN nhỏ hơn rất nhiều so với giá trị quan trắc.



Hình 4. Kết quả mô phỏng nồng độ bụi $PM_{2.5}$ tốt nhất của bốn thuật toán: (a) RFR - KB6; (b) XGBR - KB6; (c) MLPR - KB6; (d) CNN - KB6.

Bảng 5. Hiệu quả mô phỏng nồng độ bụi $PM_{2.5}$ cho các kịch bản giữa các thuật toán trong quá trình kiểm tra.

Thuật toán	Chỉ số	Kịch Bản					
		KB1	**KB2**	**KB3**	**KB4**	**KB5**	**KB6**
RFR	r	0,302	0,476	0,759	0,79	0,835	**0,839**
	IOA	0,483	0,658	0,837	0,862	0,884	**0,887**
	NMB	17,901	15,167	11,21	9,373	8,965	**9,078**
XGBR	r	0,332	0,396	0,392	0,775	0,826	**0,854**
	IOA	0,508	0,556	0,604	0,881	0,909	**0,922**
	NMB	17,185	15,996	15,893	6,106	2,408	**6,711**
MLPR	r	0,253	0,39	0,523	0,568	0,7	**0,771**
	IOA	0,414	0,528	0,701	0,732	0,83	**0,875**
	NMB	18,888	18,463	20,031	5,763	16,87	**3,217**
CNN	r	0,238	0,367	0,388	0,432	0,448	**0,454**
	IOA	0,379	0,491	0,506	0,558	0,571	**0,578**
	NMB	23,636	20,61	19,39	20,607	22,094	**20,091**

3.2. Bộ thông số thuật toán tối ưu

Nhìn chung, các kết quả cho thấy, kịch bản KB6 với 6 thông số khí tượng đầu vào là nhiệt độ (T), tốc độ gió (W), số giờ nắng (S), độ ẩm (H), hướng gió (WD) và lượng mưa (R) ở cả 4 thuật toán đều cho hiệu quả mô phỏng tốt nhất. Bên cạnh đó, khi so sánh hiệu quả mô phỏng tốt nhất giữa bốn thuật toán, kết quả từ Hình 4 và Bảng 5 cho thấy XGBR cho hiệu quả mô phỏng nồng độ bụi $PM_{2.5}$ tại khu vực trung tâm TP.HCM tối ưu nhất. Bộ tham số tối ưu của thuật toán XGBR được thể hiện trong Bảng 6, bao gồm các siêu tham số max_depth, gamma, learning_rate, n_estimators, subsample, giá trị các siêu tham số này được xác định bằng hàm GridSearchCV với thời gian huấn luyện là 30 phút. Kết quả này tương đồng với kết quả trong nghiên cứu [20], nghiên cứu này cũng sử dụng các thuật toán RF, XGB và DNN cùng với bộ dữ liệu khí tượng để dự báo nồng độ $PM_{2.5}$ và cho thấy XGB cho hiệu quả mô phỏng tốt nhất. Xét về bộ tham số tối ưu, giá trị của tham số max_depth ở hai nghiên cứu giống nhau đều bằng 8, riêng các tham số gamma và n_estimators của nghiên cứu [20] lần lượt là 0,7 và 200, có sự chênh lệch so với nghiên cứu này. Điều này có thể do sự khác biệt về đặc trưng và độ lớn của bộ dữ liệu sử dụng.

Bảng 6. Bộ tham số tối ưu của thuật toán XGBR.

Tham số	Ý nghĩa tham số	Giá trị
learning_rate	Kiểm soát sự ảnh hưởng của mỗi cây quyết định đến kết quả và tốc độ học của XGB.	0,6
max_depth	Độ sâu tối đa của mỗi cây. Cây sâu hơn có thể nắm bắt được các tương tác phức tạp hơn nhưng có thể gây overfitting.	8
gamma	Xác định mức độ giảm tối thiểu của hàm mất mát cần đạt được để tiếp tục phân chia một nút cây.	0,2
subsample	Kiểm soát tỷ lệ mẫu được sử dụng để huấn luyện từng cây.	0,6
colsample_bytree	Kiểm soát tỷ lệ cột (biến đầu vào) được sử dụng để huấn luyện từng cây.	0,6
n_estimators	Số lượng cây sẽ được xây dựng trong quá trình huấn luyện. Tăng số lượng cây có thể cải thiện hiệu suất nhưng cũng tăng thời gian huấn luyện.	300

Bên cạnh đó, từ bảng thống kê hiệu quả mô phỏng ở 6 kịch bản trong Bảng 5 có thể thấy, kết quả trong KB6 và KB5 không có sự chênh lệch nhiều ở thuật toán XGBR, và kết quả khi sử dụng 05 thông số đầu vào vẫn đạt hiệu quả mô phỏng cao với r = 0,826, IOA = 0,909 và NMB = 2,408, vì vậy có thể thấy yếu tố lượng mưa không ảnh hưởng nhiều đến nồng độ bụi $PM_{2.5}$ tại khu vực nghiên cứu. Do đó, trong trường hợp không đủ dữ liệu lượng mưa vẫn có thể sử dụng 05 thông số khí tượng là nhiệt độ, tốc độ gió, số giờ nắng, độ ẩm, hướng gió để mô phỏng nồng độ $PM_{2.5}$ mà vẫn đảm bảo hiệu quả mô phỏng.

4. Kết luận

Nghiên cứu đã thực hiện mô phỏng nồng độ bụi $PM_{2.5}$ tại khu vực trung tâm TP.HCM bằng thuật toán học máy và học sâu bao gồm bốn thuật toán RFR, XGBR, MLPR và CNN. Sáu kịch bản mô phỏng được xây dựng dựa trên mức độ tương quan riêng phần giữa nồng độ $PM_{2.5}$ và sáu thông số khí tượng: nhiệt độ, tốc độ gió, số giờ nắng, độ ẩm, hướng gió và lượng mưa. Kết quả mô phỏng nồng độ bụi $PM_{2.5}$ được đánh giá thông qua các chỉ số r, IOA và NMB.

Kết quả cho thấy các thuật toán học máy như RFR, XGBR và MLPR đạt được độ chính xác và hiệu suất tốt trong việc mô phỏng nồng độ bụi $PM_{2.5}$. Đặc biệt, thuật toán XGBR với 06 thông số đầu vào đã đạt hiệu quả mô phỏng cao nhất với các chỉ số r = 0,854, IOA = 0,922 và NMB = 6,711. Kết quả này cho thấy khả năng của thuật toán học máy trong mô phỏng diễn biến chất lượng không khí thông qua nồng độ bụi $PM_{2.5}$ là rất tốt. Kết quả của nghiên cứu có thể được sử dụng trong những bài toán về mô phỏng nồng độ bụi tại khu vực trung tâm TP.HCM, cũng như những khu vực khác có điều kiện tương tự.

Trong những nghiên cứu tiếp theo, để nâng cao hiệu quả mô phỏng sẽ xem xét đến những thông số chất lượng không khí khác như CO_2, SO_2,.. như các biến đầu vào và thử nghiệm các thuật toán khác. Bên cạnh đó, xem xét mở rộng mô phỏng, dự báo các thông số ô nhiễm khác phục vụ quá trình quản lí và kiểm soát ô nhiễm không khí trên địa bàn TP.HCM.

Đóng góp của tác giả: Xây dựng ý tưởng nghiên cứu: N.P.H., D.N.K.; Phương pháp: N.P.H., N.N.D., D.Q.L.; Xử lý số liệu: N.P.H., N.N.D., D.Q.L.; Viết bản thảo bài báo: N.P.H., D.N.K.; Chỉnh sửa bài báo: N.P.H., D.N.K.

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

Tài liệu tham khảo

1. Riojas-Rodríguez, H.; Romieu, I.; Hernández-Ávila, M. Air pollution. Occupational and Environmental Health. Oxford University Press: Oxford, UK, 2017, pp. 345–364.
2. Brunekreef, B.; Holgate, S.T. Air pollution and health. *Lancet* **2002**, *360*, 1233–1242.
3. Guarnieri, M.; Balmes, J.R. Outdoor air pollution and asthma. *Lancet* **2014**, *383*, 1581–1592.
4. Akimoto, H. Global air quality and pollution. *Science* **2003**, *302*, 1716–1719.
5. Wang, Z. Energy and air pollution. Comprehensive Energy Systems. Elsevier: Amsterdam, Netherlands, **2018**, pp. 909–949.
6. Nowak, D.J.; Crane, D.E.; Stevens, J.C. Air pollution removal by urban trees and shrubs in the United States. Urban For. *Urban Green* **2006**, *4*, 115–123.
7. WHO. 7 million premature deaths annually linked to air pollution, 2014.
8. Bộ Tài nguyên và Môi trường. Báo cáo hiện trạng môi trường quốc gia năm 2021 - Môi trường không khí, thực trạng và giải pháp, 2022.
9. Shen, H.; Li, T.; Yuan, Q.; Zhang, L. Estimating regional ground-level $PM_{2.5}$ directly from satellite top-of-atmosphere reflectance using deep belief networks. *J. Atmos. Oceanic Technol.* **2018**, *123*, 13875–13886.
10. Al Hanai, A.H.; Antkiewicz, D.S.; Hemming, J.D.C.; Shafer, M.M.; Lai, A.M.; Arhami, M.; Hosseini, V.; Schauer, J.J. Seasonal variations in the oxidative stress and inflammatory potential of $PM_{2.5}$ in Tehran using an alveolar macrophage model: The role of chemical composition and sources. *Environ. Int.* **2019**, 417–427.
11. Laden, F.; Schwartz, J.; Speizer, F.E.; Dockery, D.W. Reduction in fine particulate air pollution and mortality: Extended follow-up of the Harvard Six Cities Study. *Am. J. Respir. Crit. Care Med.* **2006**, *173*, 667–672.
12. Evans, J.; van Donkelaar, A.; Martin, R.V.; Burnett, R.; Rainham, D.G.; Birkett, N.J.; Krewski, D. Estimates of global mortality attributable to particulate air pollution using satellite imagery. *Environ. Res.* **2013**, *120*, 33–42.
13. Rojas-Rueda, D.; de Nazelle, A.; Teixidó, O.; Nieuwenhuijsen, M.J. Health impact assessment of increasing public transport and cycling use in Barcelona: A morbidity and burden of disease approach. *Prev. Med.* **2013**, *57*, 573–579.
14. IQAir/AirVisual. World Air Quality Report 2021, 2022.
15. VNU-UET, Live&Learn và USAID. Hiện trạng bụi $PM_{2.5}$ ở Việt Nam giai đoạn 2019-2020 sử dụng dữ liệu đa nguồn. Báo cáo được phối hợp thực hiện bởi Trường Đại học Công nghệ - Đại học Quốc gia Hà Nội (VNU-UET), Trung tâm Sống và Học tập vì Môi trường và Cộng đồng (Live&Learn) và Cơ quan Phát triển Quốc tế Hoa Kỳ (USAID), 2021, tr. 34-35.
16. VNU-UET, Live&Learn và USAID. Báo cáo hiện trạng bụi $PM_{2.5}$ và tác động sức khoẻ tại Việt Nam năm 2021. Báo cáo được phối hợp thực hiện bởi Trường Đại học Công nghệ - Đại học Quốc gia Hà Nội (VNU-UET), Trung tâm Sống và Học tập vì

Môi trường và Cộng đồng (Live&Learn) và Cơ quan Phát triển Quốc tế Hoa Kỳ (USAID), 2022.

17. Pak, U; Ma, J; Ryu, U; Ryom, K; Juhyok, U; Pak, K; Pak, C. Deep learning-based $PM_{2.5}$ prediction considering the spatiotemporal correlations: A case study of Beijing, China. *Sci. Total Environ.* **2020,** *699*, 133561.
18. Jinghui, M.; Zhongqi, Y.; Qu, Y.; Xu, J.; Cao, Y. Application of the XGBoost Machine Learning Method in $PM_{2.5}$ Prediction: A Case Study of Shanghai. *Aerosol Air Qual. Res.* **2020**, *20*, 128–138.
19. Yumimoto, K.; Uno, I. Adjoint inverse modeling of CO emissions over Eastern Asia using four dimensional variational data assimilation. *Atmos. Environ.* **2006**, *40*, 6836–6845.
20. Zamani Joharestani, M.; Cao, C.; Ni, X.; Bashir, B.; Talebiesfandarani, S. $PM_{2.5}$ Prediction Based on Random Forest, XGBoost, and Deep Learning Using Multisource Remote Sensing Data. *Atmosphere* **2019**, *10*, 373.
21. Plocoste, T.; Laventure, S. Forecasting PM_{10} Concentrations in the Caribbean Area Using Machine Learning Models. *Atmosphere* **2023**, *14*, 134.
22. Lei, T.M.T.; Siu, S.W.I.; Monjardino, J.; Mendes, L.; Ferreira, F. Using machine learning methods to forecast air quality: A case study in Macao. *Atmosphere* **2022**, *13*, 1412.
23. Mahmud, S.; Ridi, T.B.I.; Miah, M.S.; Sarower, F.; Elahee, S. Implementing machine learning algorithms to predict particulate matter ($PM_{2.5}$): A case study in the Paso del Norte Region. *Atmosphere* **2022**, *13*, 2100.
24. Huang, C.J.; Kuo, P.H. A deep CNN-LSTM model for particulate matter ($PM_{2.5}$) forecasting in smart cities. *Sensors* **2018**, *18*, 2220.
25. 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. *Sci. Total Environ.* **2019**, *664*, 1–10.
26. Qin, D.; Yu, J.; Zou, G.; Yong, R.; Zhao, Q.; Zhang, B. A novel combined prediction scheme based on CNN and LSTM for urban $PM_{2.5}$ concentration. *IEEE Access* **2019**, *7*, 20050–20059.
27. Tong, W.; Li, L.; Zhou, X.; Hamilton, A.; Zhang, K. Deep learning $PM_{2.5}$ concentrations with bidirectional LSTM RNN. *Air Qual. Atmos. Health* **2019**, *12*, 411–423.
28. Liaw, A.; Wiener. Classcácification and Regression by RandomForest. *R. News* **2002**, *2(3),* 18–22.
29. Chen, H.; Deng, G.; Liu, Y. Monitoring the influence of industrialization and urbanization on spatiotemporal variations of AQI and $PM_{2.5}$ in three provinces, China. *Atmosphere* **2022**, *13(9)*, 1377.
30. Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. Proceedings of the IEEE, **1998**, *86(11)*, 2278–2324.
31. Huang, L.; Zhu, Y.; Zhai, H.; Xue, S.; Zhu, T.; Shao, Y.; Liu, Z.; Emery, C.; Yarwood, G.; Wang, Y.; Fu, J.; Zhang, K.; Li, L. Recommendations on benchmarks for numerical air quality model applications in China - Part 1: $PM_{2.5}$ and chemical species. *Atmos. Chem. Phys.* **2021**, *21*, 2725–2743.

Simulation of $PM_{2.5}$ concentration in the metropolitan region of Ho Chi Minh City utilizing machine learning and deep learning algorithms

Nguyen Phuc Hieu[1], Nguyen Nhat Duong[1], Do Quang Linh[1], Dao Nguyen Khoi[1]*

[1] Faculty of Environment, University of Science, VNU-HCM; phuchieu50@gmail.com; 19170139@student.hcmus.edu.vn; dqlinh@hcmus.edu.vn; dnkhoi@hcmus.edu.vn

Abstract: The research employs three machine learning algorithms: Random Forest Regression (RFR), XGBoost Regression (XGBR), and Multilayer Perceptron Regression (MLPR), and a Convolutional Neural Network (CNN) deep learning algorithm, to simulate $PM_{2.5}$ concentration in the metropolitan region of Ho Chi Minh City. The dataset used spans from 2016 to 2021 and includes daily $PM_{2.5}$ concentration data from the U.S. Consulate General - Ho Chi Minh City station, as well as six daily meteorological parameters: temperature, wind direction, wind speed, humidity, sunshine hours, and rainfall collected from the Tan Son Hoa station. The dataset is then standardized ad split into an 80:20 ratio for the training and testing phases. Based on the results of Pearson correlation analysis between meteorological parameters and $PM_{2.5}$ concentration, six scenarios are created with different input parameters. The findings reveal that the three machine learning models are effective in simulating $PM_{2.5}$ concentrations with, correlation coefficient (r) values ranging from 0.770 to 0.854. The XGBR model performs the best when all six meteorological parameters are used, attaining an r of 0.771, an Index of Agreement (IOA) of 0.875, and a Normalized Mean Bias (NMB) of 3.217. Nonetheless, it is important to note that the CNN algorithm does not produce satisfactory results, with an r value less than 0.5 in all scenarios.

Keywords: $PM_{2.5}$; Machine learning; Deep learning; Ho Chi Minh City.