Ứng dụng mô hình Transformer để dự báo nồng độ bụi mịn PM2.5 tại Hà Nội trong giai đoạn 2022-2025

Đặng Hữu Nghị[1], Bùi Thị Vân Anh[1], Phạm Đức Hậu[1]

[1]*Trường ĐH Mỏ - Địa chất, Bắc Từ Liêm, Hà Nội.*

Email tác giả liên hệ email: danghuunghi@humg.edu.vn, buithivananh@humg.edu.vn

https://doi.org/10.5281/zenodo.xxxxxxxx

Tóm tắt :

Ô nhiễm không khí do bụi mịn PM2.5 đang trở thành một vấn đề môi trường nghiêm trọng tại nhiều đô thị lớn, đặc biệt ở các khu vực có tốc độ đô thị hóa nhanh. Việc dự báo chính xác nồng độ PM2.5 có ý nghĩa quan trọng trong công tác quản lý chất lượng không khí và xây dựng các hệ thống cảnh báo ô nhiễm môi trường. Nghiên cứu này nhằm đánh giá khả năng ứng dụng của các phương pháp học máy và học sâu trong dự báo nồng độ PM2.5 dựa trên dữ liệu chuỗi thời gian kết hợp với các yếu tố khí tượng. Dữ liệu sử dụng trong nghiên cứu bao gồm nồng độ PM2.5 cùng các biến khí tượng như nhiệt độ, độ ẩm và tốc độ gió tại khu vực Hà Nội. Các bước tiền xử lý dữ liệu được thực hiện bao gồm phát hiện và xử lý ngoại lai bằng phương pháp khoảng tứ phân vị (IQR), chuẩn hóa dữ liệu theo phương pháp Z-score và xây dựng các đặc trưng chuỗi thời gian. Các mô hình dự báo được xem xét gồm ARIMA, Random Forest, LSTM, GRU và Transformer. Kết quả thực nghiệm cho thấy các mô hình học sâu đạt hiệu suất dự báo cao hơn so với các phương pháp truyền thống. Trong đó, mô hình Transformer cho kết quả tốt nhất với sai số dự báo thấp và khả năng tái hiện xu hướng biến động của PM2.5 hiệu quả hơn so với các mô hình còn lại. Kết quả nghiên cứu cho thấy tiềm năng ứng dụng của các mô hình học sâu trong dự báo chất lượng không khí và cung cấp cơ sở khoa học cho việc xây dựng các hệ thống cảnh báo sớm ô nhiễm không khí tại các đô thị lớn.

Từ khóa: *PM2.5, dự báo chuỗi thời gian, học sâu, Transformer, chất lượng không khí.*

Ngày nhận bài: 26/02/2026 Ngày sửa lại: 14/03/2026 Ngày chấp nhận đăng: 15/03/2026 Ngày xuất bản: 30/04/2026

Forecasting PM2.5 Concentrations Using a Transformer Model

Dang Huu Nghi[1], Bui Thi Van Anh[1], Pham Duc Hau[1]

[1]*Hanoi University of Mining and Geology, Bac Tu Liem, Hanoi.*

Corresponding Author Email: danghuunghi@humg.edu.vn, buithivananh@humg.edu.vn

Abstract:

Air pollution caused by fine particulate matter (PM2.5) has become a serious environmental problem in many large urban areas, especially in rapidly urbanizing regions. Accurate prediction of PM2.5 concentrations plays an important role in air quality management and the development of early warning systems for air pollution. This study evaluates the applicability of machine learning and deep learning approaches for forecasting PM2.5 concentrations using time-series data combined with meteorological variables. The dataset includes PM2.5 concentrations together with meteorological variables such as temperature, relative humidity, and wind speed collected in Hanoi. Data preprocessing steps include outlier detection using the Interquartile Range (IQR) method, data normalization using the Z-score approach, and the construction of time-series features. Several forecasting models were implemented and compared, including ARIMA, Random Forest, LSTM, GRU, and Transformer models. The experimental results show that deep learning models outperform traditional statistical approaches in PM2.5 prediction. Among the evaluated models, the Transformer model achieved the best performance with lower prediction errors and a better ability to capture temporal variations in PM2.5 concentrations. The results demonstrate the potential of deep learning techniques for air quality forecasting and provide a scientific basis for developing early warning systems for air pollution in large urban areas.

Keywords: *PM2.5, time-series forecasting, deep learning, Transformer, air quality.*

Submission received: 26/02/2026 Revised: 14/03/2026 Accepted: 15/03/2026 Published: 30/04/2026

1. Giới thiệu

Ô nhiễm không khí do bụi mịn PM2.5 đang trở thành một trong những vấn đề môi trường nghiêm trọng tại nhiều đô thị lớn trên thế giới. Với kích thước khí động học nhỏ hơn 2,5 µm, các hạt bụi PM2.5 có khả năng xâm nhập sâu vào hệ hô hấp và hệ tuần hoàn của con người, gây ra nhiều bệnh lý nguy hiểm như bệnh tim mạch, viêm phổi và ung thư phổi. Nhiều nghiên cứu đã chỉ ra rằng nồng độ PM2.5 trong môi trường đô thị chịu ảnh hưởng của nhiều yếu tố khác nhau như điều kiện khí tượng, hoạt động giao thông, công nghiệp và các nguồn phát thải sinh khối. Do đó, việc dự báo chính xác nồng độ PM2.5 có ý nghĩa quan trọng trong công tác cảnh báo ô nhiễm không khí, hỗ trợ quản lý môi trường và bảo vệ sức khỏe cộng đồng [1].

Trong những năm gần đây, nhiều phương pháp khác nhau đã được áp dụng để dự báo nồng độ PM2.5 theo chuỗi thời gian. Các phương pháp thống kê truyền thống là một trong những hướng tiếp cận sớm nhất trong lĩnh vực này. Trong đó, mô hình ARIMA (AutoRegressive Integrated Moving Average) được sử dụng rộng rãi để phân tích và dự báo các chuỗi dữ liệu môi trường. Mô hình này dựa trên mối quan hệ tự hồi quy và trung bình trượt của chuỗi dữ liệu quá khứ để dự báo các giá trị tương lai. Mahajan và cộng sự đã áp dụng phương pháp làm mịn theo hàm mũ kết hợp với ARIMA để dự báo nồng độ PM2.5 trong ngắn hạn và cho thấy mô hình có thể đạt được độ chính xác nhất định trong các dự báo ngắn hạn [2]. Tuy nhiên, các mô hình thống kê truyền thống thường giả định mối quan hệ tuyến tính giữa các biến và gặp hạn chế khi xử lý các chuỗi dữ liệu môi trường phức tạp chịu ảnh hưởng của nhiều yếu tố phi tuyến.

Sự phát triển của các phương pháp học máy (machine learning) đã mở ra nhiều hướng tiếp cận mới trong dự báo ô nhiễm không khí. Các thuật toán như Random Forest, Support Vector Machine và Gradient Boosting có khả năng mô hình hóa các mối quan hệ phi tuyến giữa các biến môi trường và khí tượng. Random Forest là một thuật toán học tập thể dựa trên tập hợp nhiều cây quyết định, sử dụng kỹ thuật lấy mẫu bootstrap và lựa chọn ngẫu nhiên các thuộc tính trong quá trình xây dựng cây nhằm giảm sai số phương sai và tăng khả năng tổng quát hóa của mô hình [3]. Các nghiên cứu trong lĩnh vực trí tuệ tính toán cho thấy các phương pháp học máy có thể cải thiện đáng kể độ chính xác dự báo so với các mô hình thống kê truyền thống khi xử lý dữ liệu môi trường đa biến [4]. Tuy nhiên, các mô hình học máy truyền thống thường không được thiết kế chuyên biệt cho dữ liệu chuỗi thời gian, do đó khả năng mô hình hóa các phụ thuộc theo thời gian vẫn còn hạn chế.

Các mô hình học sâu (deep learning) dựa trên mạng nơ-ron đã được áp dụng rộng rãi trong dự báo chuỗi thời gian môi trường. Các kiến trúc mạng nơ-ron hồi quy như Long Short-Term Memory (LSTM) và Gated Recurrent Unit (GRU) được thiết kế nhằm giải quyết vấn đề phụ thuộc dài hạn trong dữ liệu chuỗi thời gian. LSTM sử dụng các cơ chế cổng (gate mechanisms) để kiểm soát dòng thông tin trong mạng, giúp mô hình ghi nhớ các thông tin quan trọng trong chuỗi dữ liệu dài. GRU là một biến thể đơn giản hơn của LSTM nhưng vẫn giữ được khả năng học các phụ thuộc theo thời gian hiệu quả. Naz và cộng sự đã thực hiện phân tích so sánh giữa các mô hình thống kê và các mô hình học sâu trong dự báo ô nhiễm không khí tại các khu vực đô thị và cho thấy các mô hình học sâu thường đạt độ chính xác cao hơn so với các phương pháp truyền thống [1]. Tuy nhiên, các mô hình mạng

nơ-ron hồi quy vẫn tồn tại một số hạn chế khi xử lý các chuỗi dữ liệu dài do quá trình tính toán diễn ra tuần tự theo thời gian và khó tận dụng hiệu quả khả năng tính toán song song của các hệ thống hiện đại.

Gần đây, mô hình Transformer đã thu hút sự quan tâm lớn trong các nghiên cứu dự báo chuỗi thời gian nhờ khả năng xử lý hiệu quả các phụ thuộc dài hạn trong dữ liệu. Transformer được giới thiệu lần đầu tiên trong nghiên cứu của Vaswani và cộng sự với kiến trúc dựa trên cơ chế self-attention, cho phép mô hình đánh giá mức độ quan trọng của các phần tử trong chuỗi dữ liệu đối với nhau [5]. Khác với các mạng nơ-ron hồi quy truyền thống, Transformer có khả năng xử lý toàn bộ chuỗi dữ liệu cùng lúc, nhờ đó cải thiện hiệu quả tính toán và khả năng học các mối quan hệ phức tạp trong dữ liệu. Trong lĩnh vực dự báo ô nhiễm không khí, các nghiên cứu gần đây cho thấy các mô hình dựa trên Transformer có thể đạt độ chính xác cao hơn so với các mô hình LSTM và GRU, đặc biệt khi xử lý các tập dữ liệu lớn và đa biến [6].

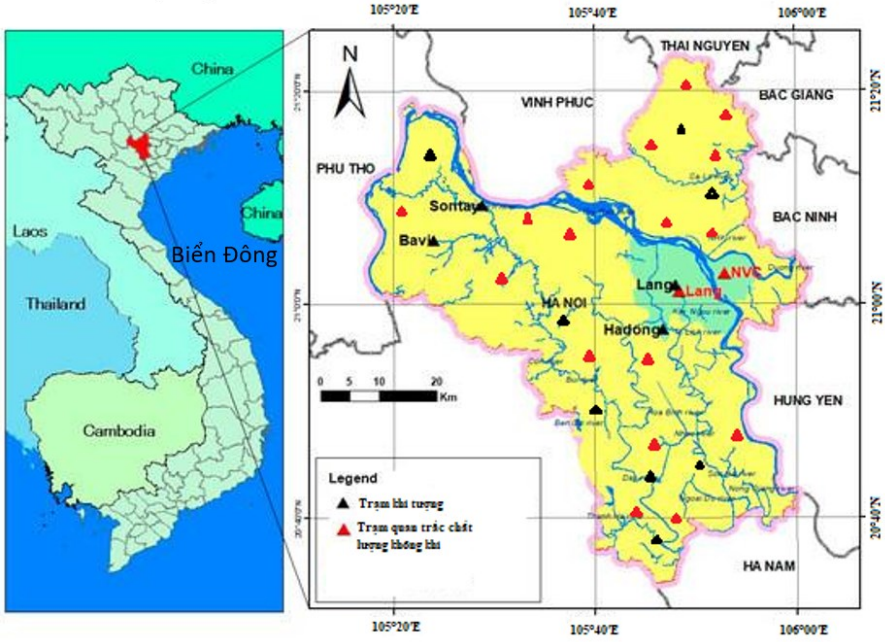
Ngoài các nghiên cứu về phương pháp dự báo, nhiều nghiên cứu cũng tập trung phân tích mối quan hệ giữa nồng độ PM2.5 và các yếu tố khí tượng. Hien và cộng sự đã nghiên cứu ảnh hưởng của các điều kiện khí tượng đến nồng độ PM2.5 và PM10 trong mùa gió mùa tại Hà Nội và cho thấy nhiệt độ, độ ẩm và tốc độ gió có ảnh hưởng đáng kể đến sự biến động của các hạt bụi trong khí quyển [7]. Các nghiên cứu tại Trung Quốc cũng cho thấy sự phân bố nồng độ PM2.5 và PM10 có mối liên hệ chặt chẽ với điều kiện khí tượng và cấu trúc đô thị [8]. Zhao và cộng sự đã phân tích tỷ lệ PM2.5/PM10 tại các khu vực kinh tế khác nhau và chỉ ra rằng các yếu tố khí tượng và nguồn phát thải đóng vai trò quan trọng trong việc kiểm soát sự biến động của bụi mịn trong môi trường đô thị [9].

Bên cạnh đó, các mô hình học sâu dựa trên kiến trúc encoder–decoder cũng đã được áp dụng trong dự báo nồng độ PM2.5 theo giờ. Yan và cộng sự đã đề xuất mô hình encoder–decoder để dự báo nồng độ PM2.5 theo chuỗi thời gian và kết quả cho thấy phương pháp này có khả năng cải thiện đáng kể độ chính xác dự báo so với các mô hình truyền thống [10].

Mặc dù đã có nhiều nghiên cứu về dự báo PM2.5 trên thế giới, việc ứng dụng các mô hình học sâu hiện đại, đặc biệt là các mô hình Transformer, cho dữ liệu ô nhiễm không khí tại Việt Nam vẫn còn hạn chế. Phần lớn các nghiên cứu trước đây tập trung vào các mô hình thống kê hoặc các kiến trúc mạng nơ-ron hồi quy như LSTM. Do đó, việc nghiên cứu và đánh giá hiệu quả của mô hình Transformer trong dự báo nồng độ PM2.5 tại khu vực Hà Nội là cần thiết nhằm làm rõ tiềm năng ứng dụng của các phương pháp học sâu hiện đại trong quản lý chất lượng không khí đô thị.

2. Khu vực nghiên cứu và dữ liệu

2.1. Khu vực nghiên cứu



Hình 1. Khu vực nghiên cứu

Khu vực nghiên cứu được lựa chọn là thành phố Hà Nội, thủ đô của Việt Nam và là một trong những trung tâm kinh tế – xã hội lớn của cả nước. Hà Nội nằm ở khu vực đồng bằng sông Hồng với tọa độ địa lý khoảng từ 20°40′ đến 21°20′ vĩ độ Bắc và từ 105°20′ đến 106°00′ kinh độ Đông. Thành phố có diện tích tự nhiên khoảng 3.358 km² và dân số hơn 8 triệu người. Với tốc độ đô thị hóa nhanh, mật độ dân cư cao và hệ thống giao thông dày đặc, Hà Nội đang phải đối mặt với nhiều vấn đề môi trường, trong đó ô nhiễm không khí là một trong những thách thức nghiêm trọng.

Theo các nghiên cứu trước đây, nồng độ bụi mịn PM2.5 tại Hà Nội thường xuyên vượt ngưỡng khuyến nghị của Tổ chức Y tế Thế giới (WHO). Nhiều đợt ô nhiễm nghiêm trọng xảy ra vào mùa đông khi điều kiện khí tượng bất lợi, đặc biệt là hiện tượng nghịch nhiệt và tốc độ gió thấp, làm hạn chế quá trình khuếch tán các chất ô nhiễm trong khí quyển. Bên cạnh đó, các nguồn phát thải từ giao thông đô thị, hoạt động xây dựng, công nghiệp và đốt sinh khối cũng góp phần làm gia tăng nồng độ bụi mịn trong không khí.

Các nghiên cứu trước đây cho thấy nồng độ PM2.5 tại Hà Nội chịu ảnh hưởng đáng kể của các yếu tố khí tượng như nhiệt độ, độ ẩm tương đối và tốc độ gió. Hien và cộng sự đã chỉ ra rằng sự biến động của PM2.5 và PM10 tại Hà Nội có mối liên hệ chặt chẽ với các điều kiện khí tượng trong mùa gió mùa [7]. Do đó, việc kết hợp dữ liệu quan trắc môi trường với các biến khí tượng được xem là cần thiết trong các nghiên cứu dự báo ô nhiễm không khí.

Trong nghiên cứu này, Hà Nội được lựa chọn làm khu vực nghiên cứu nhằm đánh giá khả năng ứng dụng của các mô hình học sâu trong dự báo nồng độ PM2.5 tại các đô thị lớn có mức độ ô nhiễm không khí cao.

2.2. Dữ liệu nghiên cứu

Dữ liệu sử dụng trong nghiên cứu bao gồm chuỗi thời gian nồng độ PM2.5 và các biến khí tượng liên quan đến điều kiện môi trường đô thị. Các biến khí tượng được xem xét trong nghiên cứu gồm nhiệt độ không khí, độ ẩm tương đối và tốc độ gió. Những yếu tố này được chứng minh có ảnh hưởng đáng kể đến quá trình phát tán và tích tụ của các hạt bụi mịn trong khí quyển.

Chúng tôi thử nghiệm các mô hình dự đoán dựa trên tập dữ liệu về Chỉ số chất lượng không khí vùng Hà Nội trong giai đoạn từ năm 2022 đến năm 2025, tập dữ liệu được tải về từ [11]. Tập dữ liệu bao gồm các chuỗi quan trắc PM2.5 kết hợp với các biến khí tượng tương ứng tại cùng thời điểm.

Các dữ liệu ban đầu được chuẩn hóa về cùng một hệ thời gian nhằm đảm bảo tính đồng bộ giữa các biến đầu vào của mô hình. Trong quá trình xây dựng tập dữ liệu, các bước xử lý sơ bộ được thực hiện bao gồm kiểm tra dữ liệu thiếu, loại bỏ các giá trị bất thường và chuẩn hóa dữ liệu trước khi đưa vào quá trình huấn luyện mô hình.

Các biến dữ liệu chính được sử dụng trong nghiên cứu được trình bày trong Bảng 1.

Bảng 1. Các biến dữ liệu sử dụng trong nghiên cứu

Biến dữ liệu	Ký hiệu	Đơn vị	Mô tả
Nồng độ bụi mịn	PM2.5	$\mu g/m^3$	Nồng độ bụi min trong không khí
Nhệt độ	Temp	^0C	Nhiệt độ không khí
Độ ẩm tương đối	RH	%	Độ ẩm không khí
Tốc độ gió	WS	m/s	Tốc độ gió tại thời điểm quan trắc
Thời gian	Time	giờ	Mốc thời gian quan trắc

Ngoài các biến môi trường, các đặc trưng thời gian như giờ trong ngày, ngày trong tuần và tháng trong năm cũng được sử dụng nhằm phản ánh các chu kỳ biến động theo thời gian của nồng độ PM2.5.

Tập dữ liệu sau khi xử lý được chia thành ba phần bao gồm tập huấn luyện (training set), tập kiểm định (validation set) và tập kiểm tra (test set). Việc chia dữ liệu được thực hiện theo thứ tự thời gian nhằm đảm bảo tính liên tục của chuỗi dữ liệu và tránh hiện tượng rò rỉ thông tin giữa các tập dữ liệu.

3. Phương pháp nghiên cứu

Nghiên cứu này áp dụng phương pháp học sâu để dự báo nồng độ bụi mịn PM2.5 theo chuỗi thời gian tại khu vực Hà Nội. Quy trình nghiên cứu bao gồm các bước chính: thu thập và tiền xử lý dữ liệu, xây dựng đặc trưng chuỗi thời gian, thiết kế mô hình dự báo và đánh giá hiệu suất mô hình.

3.1. Tiền xử lý dữ liệu

Dữ liệu quan trắc môi trường thường chứa các giá trị thiếu, nhiễu và sai số do các nguyên nhân như lỗi thiết bị đo, gián đoạn truyền dữ liệu hoặc điều kiện vận hành trạm quan trắc. Do đó, việc tiền xử lý dữ liệu là bước quan trọng nhằm đảm bảo chất lượng dữ liệu trước khi đưa vào huấn luyện mô hình.

Đồng bộ thời gian

Các chuỗi dữ liệu PM2.5 và dữ liệu khí tượng được chuẩn hóa theo cùng một bước thời gian với độ phân giải 1 giờ. Tất cả các quan trắc được chuyển đổi về cùng một hệ thời gian (UTC+7) để đảm bảo tính đồng bộ giữa các biến.

Xử lý dữ liệu thiếu

Trong quá trình quan trắc môi trường, dữ liệu có thể bị thiếu do gián đoạn trong quá trình đo hoặc lỗi cảm biến. Các đoạn dữ liệu thiếu ngắn được xử lý bằng phương pháp nội suy tuyến tính theo thời gian. Nếu khoảng thiếu dữ liệu lớn hơn một ngưỡng nhất định (ví dụ 6 giờ), các đoạn dữ liệu này sẽ được loại bỏ khỏi tập huấn luyện nhằm tránh ảnh hưởng đến kết quả dự báo.

Phát hiện và xử lý ngoại lai

Các giá trị ngoại lai (outliers) trong dữ liệu PM2.5 có thể xuất hiện do lỗi thiết bị hoặc nhiễu đo. Phương pháp khoảng tứ phân vị (Interquartile Range – IQR) được sử dụng để phát hiện các giá trị bất thường. Một giá trị được xem là ngoại lai nếu thỏa mãn điều kiện:

$$x < Q1 - 1.5IQR \text{ hoặc} x > Q3 + 1.5IQR \quad (1)$$

trong đó:

- Q1 là tứ phân vị thứ nhất
- Q3 là tứ phân vị thứ ba

$$\text{IQR} = \text{Q3} - \text{Q1} \quad (2)$$

Các giá trị ngoại lai sau khi phát hiện sẽ được thay thế bằng giá trị trung vị hoặc nội suy từ các giá trị lân cận.

Chuẩn hóa dữ liệu

Các biến môi trường và khí tượng có đơn vị và thang đo khác nhau. Do đó, dữ liệu được chuẩn hóa nhằm đưa các biến về cùng một phạm vi giá trị. Trong nghiên cứu này, phương pháp chuẩn hóa Z-score được sử dụng:

$$x_{norm} = \frac{x - \mu}{\sigma} \quad (3)$$

trong đó:

- x là giá trị ban đầu
- μ là giá trị trung bình của biến
- σ là độ lệch chuẩn

Việc chuẩn hóa dữ liệu giúp tăng tính ổn định và hiệu quả của quá trình huấn luyện mô hình học sâu.

3.2. Tạo đặc trưng chuỗi thời gian

Nồng độ PM2.5 trong môi trường đô thị thường có sự biến động theo chu kỳ ngày – đêm và theo mùa. Do đó, việc xây dựng các đặc trưng thời gian giúp mô hình học được các quy luật biến động này.

Các đặc trưng thời gian được sử dụng trong nghiên cứu bao gồm:

- giờ trong ngày (hour of day)
- ngày trong tuần (day of week)
- tháng trong năm (month)

Để biểu diễn các đặc trưng tuần hoàn, các biến thời gian được mã hóa bằng hàm sin và cos:

$$\sin\left(\frac{2\pi t}{T}\right), cos\left(\frac{2\pi t}{T}\right) \quad (4)$$

trong đó:

- t là thời điểm trong chu kỳ
- T là độ dài chu kỳ

Ngoài ra, các đặc trưng trễ (lag features) cũng được xây dựng nhằm phản ánh sự phụ thuộc của PM2.5 vào các giá trị trong quá khứ. Các độ trễ được sử dụng trong nghiên cứu gồm:

- PM2.5(t−1h)
- PM2.5(t−3h)
- PM2.5(t−24h)
- PM2.5(t−168h)

Những đặc trưng này giúp mô hình nắm bắt được xu hướng biến động ngắn hạn và dài hạn của chuỗi dữ liệu.

3.3. Xây dựng mô hình dự báo

Trong nghiên cứu này, mô hình Transformer được sử dụng để dự báo nồng độ PM2.5 theo chuỗi thời gian. Transformer là một kiến trúc học sâu dựa trên cơ chế **self-attention**, cho phép mô hình học các mối quan hệ giữa các phần tử trong chuỗi dữ liệu.

Khác với các mạng nơ-ron hồi quy truyền thống như LSTM hoặc GRU, Transformer không xử lý dữ liệu theo thứ tự tuần tự mà xử lý toàn bộ chuỗi dữ liệu cùng lúc. Điều này giúp mô hình có khả năng khai thác hiệu quả các phụ thuộc dài hạn trong chuỗi thời gian.

Kiến trúc Transformer bao gồm hai thành phần chính:

- **Encoder**: mã hóa chuỗi dữ liệu đầu vào thành các biểu diễn đặc trưng
- **Decoder**: tạo ra giá trị dự báo dựa trên các biểu diễn đã mã hóa

Trong cơ chế self-attention, ba ma trận đặc trưng được sử dụng bao gồm:

$$Q - XW_Q,\ K - XW_K,\ V - XW_V \tag{5}$$

trong đó:

- Q là ma trận truy vấn (Query)
- K là ma trận khóa (Key)
- V là ma trận giá trị (Value)

Trọng số attention được tính theo công thức:

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{6}$$

Cơ chế này cho phép mô hình đánh giá mức độ quan trọng của từng phần tử trong chuỗi dữ liệu đối với các phần tử khác.

3.4. Đánh giá mô hình

Hiệu suất của các mô hình dự báo được đánh giá bằng các chỉ số sai số phổ biến trong dự báo chuỗi thời gian, bao gồm:

Sai số tuyệt đối trung bình (MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{7}$$

trong đó:

- y_i là giá trị quan trắc
- $\hat{y}_i$ là giá trị dự báo
- n là số lượng quan sát

Sai số bình phương trung bình căn bậc hai (RMSE)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (8)$$

RMSE nhạy cảm hơn với các sai số lớn và thường được sử dụng để đánh giá độ ổn định của mô hình.

Hệ số xác định (R^2)

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2} \quad (9)$$

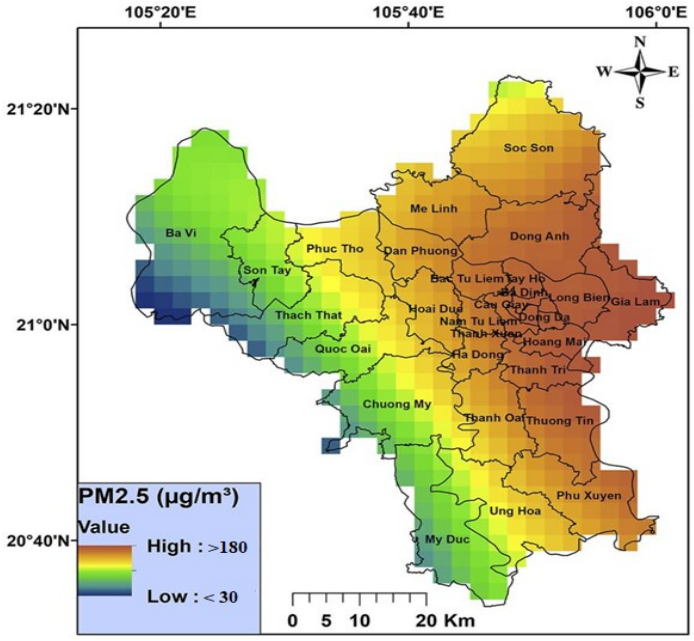
trong đó $\bar{y}$ là giá trị trung bình của chuỗi quan trắc.

Các chỉ số này được sử dụng để so sánh hiệu suất của mô hình Transformer với các phương pháp khác như ARIMA, Random Forest, LSTM và GRU.

4. Kết quả và thảo luận

4.1. Kết quả dự báo nồng độ PM2.5

Để cung cấp cái nhìn tổng quan về đặc điểm phân bố không gian của ô nhiễm không khí tại khu vực nghiên cứu, bản đồ phân bố nồng độ PM2.5 tại Hà Nội được thể hiện trong Hình 2.



Hình 2. Phân bố không gian nồng độ PM2.5 tại Hà Nội (xây dựng từ dữ liệu tổng hợp)

Bản đồ cho thấy nồng độ PM2.5 có xu hướng phân bố không đồng đều trong không gian, với các khu vực đô thị trung tâm và phía Đông – Nam thành phố ghi nhận mức ô nhiễm cao hơn so với các khu vực ngoại thành. Điều này phản ánh ảnh hưởng của mật độ giao thông, hoạt động xây dựng và các nguồn phát thải đô thị đến chất lượng không khí.

Sau khi tiền xử lý dữ liệu, tập dữ liệu được chia thành ba phần theo thứ tự thời gian, bao gồm tập huấn luyện (70%), tập kiểm định (15%) và tập kiểm tra

(15%). Trên cơ sở đó, các mô hình dự báo chuỗi thời gian được xây dựng nhằm dự báo nồng độ PM2.5.

Trong nghiên cứu này, năm phương pháp được xem xét bao gồm: ARIMA, Random Forest, LSTM, GRU và Transformer. Hiệu suất của các mô hình được đánh giá thông qua các chỉ số MAE, RMSE và hệ số xác định R^2.

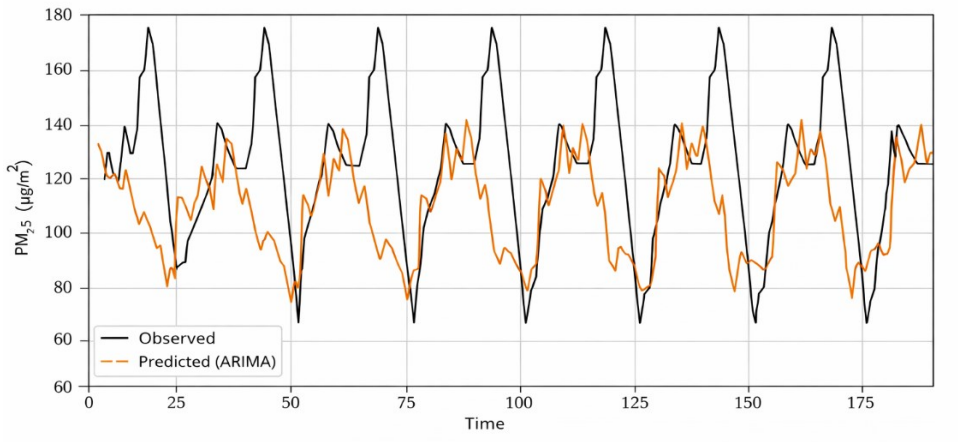
Kết quả đánh giá hiệu suất dự báo của các mô hình được trình bày trong Bảng 2.

Bảng 2. So sánh hiệu suất dự báo của các mô hình

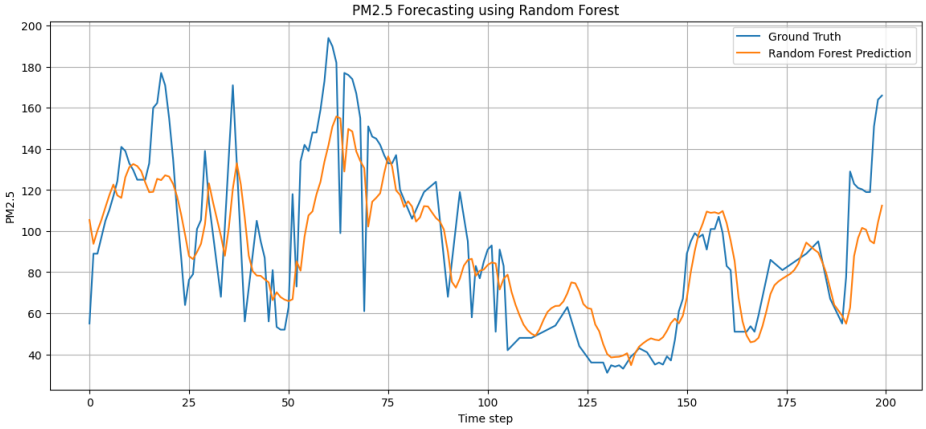
Mô hình	MAE (µg/m³)	RMSE (µg/m³)	R^2
ARIMA	11.84	15.62	0.71
Random Forest	9.76	13.25	0.79
LSTM	8.43	11.74	0.84
GRU	8.15	11.36	0.86
Transformer	**7.42**	**10.28**	**0.90**

Kết quả cho thấy mô hình **Transformer đạt hiệu suất dự báo tốt nhất**, với sai số thấp nhất và hệ số xác định cao nhất. Điều này chứng tỏ khả năng của mô hình trong việc nắm bắt các mối quan hệ phi tuyến và phụ thuộc dài hạn trong chuỗi dữ liệu PM2.5.

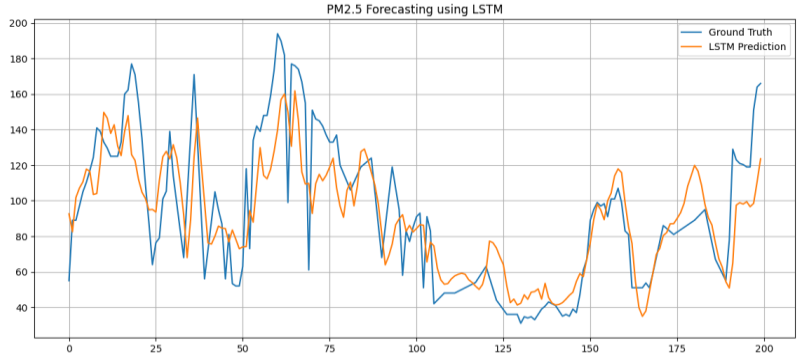
Sự so sánh giữa giá trị quan trắc và giá trị dự báo của các mô hình được thể hiện trong Hình 3 đến Hình 7.



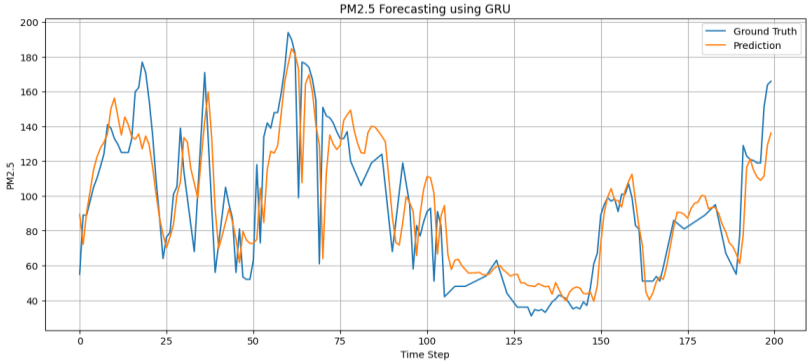
Hình 3: So sánh giữa giá trị thực và dự đoán PM2.5 sử dụng ARIMA



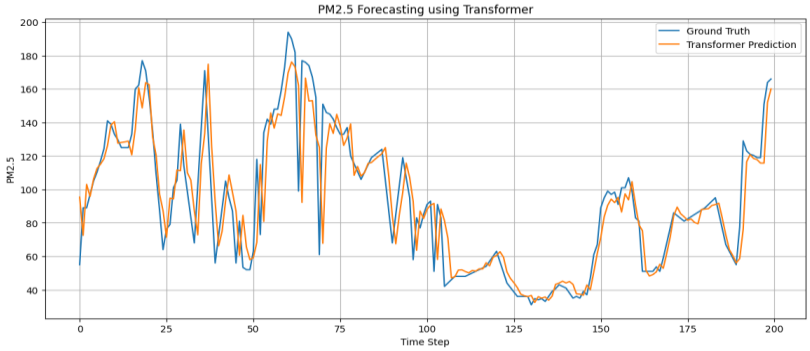
Hình 4: So sánh giữa giá trị thực và dự đoán PM2.5 sử dụng Random Forest



Hình 5: So sánh giữa giá trị thực và dự đoán PM2.5 sử dụng LSTM



Hình 6: So sánh giữa giá trị thực và dự đoán PM2.5 sử dụng GRU



Hình 7: So sánh giữa giá trị thực và dự đoán PM2.5 sử dụng Transformer

4.2. So sánh kết quả dự báo giữa các mô hình

Kết quả trong Bảng 2 cho thấy các mô hình học sâu như LSTM, GRU và Transformer có hiệu suất dự báo cao hơn so với các mô hình thống kê truyền thống. Điều này có thể được giải thích bởi khả năng học các mối quan hệ phi tuyến và phụ thuộc dài hạn trong dữ liệu chuỗi thời gian của các mạng nơ-ron sâu.

Mô hình ARIMA chỉ dựa trên mối quan hệ tuyến tính giữa các giá trị trong chuỗi dữ liệu nên gặp hạn chế khi xử lý các chuỗi dữ liệu môi trường có tính biến động phức tạp. Trong khi đó, Random Forest có khả năng mô hình hóa các mối quan hệ phi tuyến tốt hơn nhưng vẫn chưa khai thác đầy đủ cấu trúc thời gian của dữ liệu.

Các mô hình LSTM và GRU được thiết kế đặc biệt cho dữ liệu chuỗi thời gian và có khả năng ghi nhớ thông tin trong quá khứ thông qua các cơ chế cổng. Do đó, các mô hình này đạt kết quả dự báo tốt hơn so với các phương pháp truyền thống.

Tuy nhiên, quá trình tính toán của các mạng nơ-ron hồi quy vẫn diễn ra tuần tự theo thời gian, dẫn đến hạn chế trong việc xử lý các chuỗi dữ liệu dài.

Mô hình Transformer sử dụng cơ chế self-attention để đánh giá mức độ quan trọng của các phần tử trong chuỗi dữ liệu. Cơ chế này cho phép mô hình khai thác hiệu quả các mối quan hệ giữa các thời điểm khác nhau trong chuỗi thời gian mà không cần xử lý tuần tự. Nhờ đó, Transformer có khả năng nắm bắt các phụ thuộc dài hạn trong dữ liệu PM2.5 và cải thiện độ chính xác của dự báo.

4.3. Phân tích xu hướng dự báo PM2.5

Kết quả dự báo cho thấy mô hình Transformer có khả năng tái hiện tốt xu hướng biến động của nồng độ PM2.5 theo thời gian. Kết quả cho thấy các giá trị dự báo bám sát xu hướng biến động của dữ liệu quan trắc, đặc biệt trong các giai đoạn nồng độ PM2.5 tăng cao. Sai số dự báo chủ yếu xuất hiện trong các thời điểm có biến động đột ngột của nồng độ PM2.5, điều này có thể liên quan đến các yếu tố khí tượng hoặc nguồn phát thải đột biến chưa được mô hình hóa đầy đủ.

Ngoài ra, mô hình Transformer cho thấy khả năng dự báo tốt hơn trong các khoảng thời gian dài so với các mô hình LSTM và GRU. Điều này cho thấy cơ chế attention giúp mô hình khai thác hiệu quả thông tin từ nhiều thời điểm khác nhau trong chuỗi dữ liệu.

4.4. Thảo luận

Kết quả nghiên cứu cho thấy các mô hình học sâu, đặc biệt là Transformer, có tiềm năng lớn trong dự báo nồng độ PM2.5 tại các khu vực đô thị. Việc kết hợp dữ liệu quan trắc môi trường với các biến khí tượng giúp cải thiện đáng kể khả năng dự báo của mô hình.

So với các nghiên cứu trước đây, kết quả của nghiên cứu này phù hợp với xu hướng chung khi các mô hình học sâu thường đạt độ chính xác cao hơn trong các bài toán dự báo ô nhiễm không khí. Tuy nhiên, hiệu suất của mô hình vẫn phụ thuộc vào chất lượng và độ dài của chuỗi dữ liệu đầu vào.

Một hạn chế của nghiên cứu là dữ liệu được sử dụng chỉ bao gồm một số biến khí tượng cơ bản. Trong các nghiên cứu tiếp theo, có thể xem xét bổ sung các yếu tố như áp suất khí quyển, bức xạ mặt trời hoặc dữ liệu phát thải để cải thiện độ chính xác của mô hình. Ngoài ra, việc kết hợp dữ liệu vệ tinh hoặc dữ liệu mô hình khí tượng cũng có thể giúp nâng cao khả năng dự báo trong các nghiên cứu tương lai.

5. Kết luận

Nghiên cứu này đã xây dựng và đánh giá mô hình dự báo nồng độ bụi mịn PM2.5 dựa trên các phương pháp học máy và học sâu. Dữ liệu quan trắc PM2.5 kết hợp với các yếu tố khí tượng được sử dụng để xây dựng tập dữ liệu phục vụ huấn luyện và kiểm tra mô hình. Các bước tiền xử lý dữ liệu, phát hiện và xử lý ngoại lai, chuẩn hóa dữ liệu và xây dựng đặc trưng chuỗi thời gian đã được thực hiện nhằm nâng cao chất lượng dữ liệu đầu vào cho các mô hình dự báo.

Kết quả nghiên cứu cho thấy các mô hình học sâu như LSTM, GRU và đặc biệt là Transformer có khả năng dự báo nồng độ PM2.5 tốt hơn so với các phương pháp truyền thống như ARIMA và Random Forest. Trong số các mô hình được xem xét, Transformer đạt hiệu suất dự báo cao nhất với các chỉ số sai số thấp và hệ số xác định cao, cho thấy khả năng khai thác hiệu quả các mối quan hệ phi tuyến và phụ thuộc dài hạn trong chuỗi dữ liệu môi trường.

Ngoài ra, kết quả nghiên cứu cũng cho thấy việc kết hợp các biến khí tượng như nhiệt độ, độ ẩm và tốc độ gió giúp cải thiện đáng kể khả năng dự báo của mô

hình. Điều này cho thấy các yếu tố khí tượng đóng vai trò quan trọng trong sự biến động của nồng độ PM2.5 tại khu vực đô thị.

Mặc dù mô hình Transformer cho kết quả dự báo khả quan, nghiên cứu vẫn còn một số hạn chế nhất định, đặc biệt là số lượng biến đầu vào còn hạn chế và phạm vi dữ liệu nghiên cứu chưa đủ dài để phản ánh đầy đủ các xu thế biến động dài hạn của ô nhiễm không khí. Trong các nghiên cứu tiếp theo, cần xem xét bổ sung thêm các yếu tố ảnh hưởng như dữ liệu phát thải, dữ liệu giao thông, cũng như dữ liệu viễn thám hoặc mô hình khí tượng nhằm nâng cao độ chính xác và khả năng ứng dụng của các mô hình dự báo.

Kết quả của nghiên cứu góp phần làm rõ tiềm năng ứng dụng của các mô hình học sâu trong dự báo chất lượng không khí, đồng thời cung cấp cơ sở khoa học cho việc xây dựng các hệ thống cảnh báo sớm ô nhiễm không khí tại các đô thị lớn ở Việt Nam.

Lời cảm ơn

Các tác giả nên trình bày về sự cảm ơn tới các cơ quan, tổ chức đã đầu tư cho nghiên cứu này, hỗ trợ, cung cấp số liệu nghiên cứu...

Cam kết của các tác giả

Tất cả các tác giả có tên trong bài báo cam kết sự đồng thuận và không có xung đột lợi ích trong công bố khoa học tại bài báo này.

Tài liệu tham khảo

[1] Naz F., Mccann C., Fahim M., Cao T.V., Hunter R., Viet N.T., Nguyen L.D., Duong T.Q. (2023), *Comparative analysis of deep learning and statistical models for air pollutants prediction in urban areas*, IEEE Access, 11, 64016–64025.

[2] Mahajan S., Chen L.J., Tsai T.C. (2018), *Short-term PM2.5 forecasting using exponential smoothing method: A comparative analysis*, Sensors, 18(10), 3223.

[3] Russell S., Norvig P. (2009), *Artificial Intelligence: A Modern Approach*, 3rd Edition, Prentice Hall, Upper Saddle River, NJ.

[4] Oprea M., Mihalache S.F., Popescu M. (2017), *Computational intelligence-based PM2.5 air pollution forecasting*, International Journal of Computers Communications & Control, 12(3), 365–380.

[5] Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I. (2017), *Attention Is All You Need*, Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS), Long Beach, USA.

[6] Nguyen M.H., Le Nguyen P., Nguyen K., Le V.A., Nguyen T.H., Ji Y. (2021), *PM2.5 prediction using genetic algorithm-based feature selection and encoder–decoder model*, IEEE Access, 9, 57338–57350.

[7] Hien P.D., Bac V.T., Tham H.C., Nhan D.D., Vinh L.D. (2002), *Influence of meteorological conditions on PM2.5 and PM2.5–10 concentrations during the monsoon season in Hanoi, Vietnam*, Atmospheric Environment, 36(21), 3473–3484.

[8] Zhou X., Cao Z., Ma Y., Wang L., Wu R., Wang W. (2016), *Concentrations, correlations and chemical species of PM2.5 and PM10 based on published data in China: Potential implications for the revised particulate standard*, Chemosphere, 144, 518–526.

[9] Zhao D., Chen H., Yu E., Luo T. (2019), *PM2.5/PM10 ratios in eight economic regions and their relationship with meteorology in China*, Advances in Meteorology, 2019, 1–15.

[10] Yan L., Wu Y., Yan L., Zhou M. (2018), *Encoder–decoder model for forecast of PM2.5 concentration per hour*, Proceedings of the 1st International Cognitive Cities Conference (IC3), 45–50.

[11] https://www.kaggle.com/datasets/phungdinhdat/aqi-in-hanoi-2022-2025?resource=download&select=2025.csv