

**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC BÁCH KHOA
KHOA KHOA HỌC VÀ KỸ THUẬT MÁY TÍNH**

**Xây dựng mô hình AI sử dụng dữ liệu
quan trắc môi trường dự đoán chất
lượng không khí theo thời gian thực**

VÕ BÁ DAT
MSHV: 2170307

LUẬN VĂN THẠC SĨ

Chuyên ngành
KHOA HỌC MÁY TÍNH

Tp Hồ Chí Minh, Ngày 21 tháng 2 năm 2025

Lời cảm ơn

Để hoàn thành luận văn này, tôi xin bày tỏ lòng biết ơn sâu sắc đến Phó giáo sư, Tiến sĩ Phạm Trần Vũ, người đã tận tình hướng dẫn, dùi dắt tôi trong suốt quá trình nghiên cứu. Thầy không chỉ truyền đạt kiến thức uyên bác mà còn khơi dậy trong tôi niềm đam mê khoa học, thôi thúc tôi không ngừng cố gắng và sáng tạo.

Tôi xin chân thành cảm ơn quý thầy cô giáo trong Khoa Khoa học và Kỹ thuật Máy tính, Trường Đại học Bách Khoa Thành phố Hồ Chí Minh đã trang bị cho tôi nền tảng kiến thức vững chắc trong những năm học tập tại trường. Những kiến thức quý báu này không chỉ là hành trang giúp tôi thực hiện đề tài luận văn này mà còn là hành trang quan trọng để tôi tự tin bước vào tương lai.

Cuối cùng, tôi xin gửi lời chúc tốt đẹp nhất đến quý thầy cô. Kính chúc quý thầy cô luôn mạnh khỏe, hạnh phúc và đạt được nhiều thành công hơn nữa trong sự nghiệp trồng người cao quý.

Tóm tắt

Trong thời đại số, sự phát triển nhanh chóng của công nghệ thông tin và trí tuệ nhân tạo đã mở ra những khả năng mới cho việc theo dõi và đánh giá những thay đổi của môi trường. Đặc biệt, việc thu thập dữ liệu chất lượng không khí theo thời gian thực đóng một vai trò quan trọng, nhất là trong bối cảnh ô nhiễm nghiêm trọng và sự nóng lên toàn cầu do các hoạt động công nghiệp, giao thông và đô thị gây ra. Việc theo dõi và quản lý chất lượng không khí đã trở thành tối quan trọng.

Luận văn này nghiên cứu và phát triển một quy trình AI để phân loại và dự đoán chất lượng không khí theo thời gian thực. Hệ thống bao gồm các mô hình học sâu, một lĩnh vực đang nổi lên mạnh mẽ trong trí tuệ nhân tạo. Tập trung vào việc phân tích và sử dụng dữ liệu quan trắc môi trường - dữ liệu chuỗi thời gian - hệ thống phân loại và dự đoán chất lượng không khí.

Hơn nữa, nghiên cứu này đi sâu vào việc đánh giá hiệu quả của việc triển khai hệ thống trong thực tế, phân tích ưu nhược điểm của phương pháp nghiên cứu để đề xuất các cải tiến cho hệ thống trong tương lai. Nghiên cứu mong muốn đóng góp vào việc bảo vệ môi trường và nâng cao chất lượng cuộc sống của chúng ta. Đồng thời, nghiên cứu khám phá những thách thức của việc tích hợp học sâu vào quy trình giám sát và quan trắc thời gian thực, định hình tương lai của công nghệ này trong lĩnh vực quan trắc dữ liệu.

Abstract

In the digital age, the rapid advancement of information technology and artificial intelligence has unlocked new possibilities for monitoring and evaluating environmental changes. Particularly, the collection of real-time air quality data plays a crucial role, especially amidst severe pollution and global warming exacerbated by industrial, transportation, and urban activities. Monitoring and managing air quality have become paramount.

This thesis investigates and develops an AI pipeline for real-time air quality classification and prediction. The system comprises deep learning models, a rapidly emerging field within artificial intelligence. Focusing on analyzing and utilizing environmental monitoring data—time-series data—the system classifies and predicts air quality.

Furthermore, this research delves into evaluating the effectiveness of real-world system deployment, analyzing the advantages and disadvantages of the research methodology to propose future system improvements. The study aspires to contribute to environmental protection and enhance our quality of life. Additionally, it explores the challenges of integrating deep learning into real-time monitoring and observation processes, shaping the future of this technology in data observation.

Lời cam đoan

Ở Việt Nam, đã có nhiều nghiên cứu ứng dụng các mô hình AI để dự đoán chất lượng nguồn nước, tuy nhiên đối với chất lượng không khí thì chưa được phổ biến. Tôi xin cam đoan luận văn "Xây dựng mô hình AI sử dụng dữ liệu quan trắc môi trường dự đoán chất lượng không khí theo thời gian thực" là công trình nghiên cứu độc lập của tôi dưới sự hướng dẫn của thầy Phạm Trần Vũ.

- Mọi thông tin, số liệu và kết quả nghiên cứu trong luận văn này đều trung thực và chính xác.
- Tôi đã thực hiện nghiên cứu này một cách độc lập và tự giác, tuân thủ đầy đủ các quy định về đạo đức nghiên cứu khoa học.
- Các số liệu, ý tưởng, hoặc nội dung của tác giả khác được sử dụng trong luận văn đều đã được trích dẫn nguồn đầy đủ và rõ ràng theo quy định.

Tôi xin chịu hoàn toàn trách nhiệm về tính chính xác và trung thực của nội dung luận văn này.

Xác nhận của tác giả:

Ngày:

Mục lục

Lời cảm ơn	i
Tóm tắt	ii
Lời cam đoan	iv
Danh sách hình vẽ	vii
Danh sách bảng	viii
Danh mục chữ viết tắt	ix
1 Tổng quan	1
1.1 Đặt vấn đề	1
1.2 Mục tiêu nghiên cứu	2
1.3 Hướng tiếp cận và giải quyết bài toán	3
2 Các nghiên cứu liên quan	4
2.1 Phát hiện và hiệu chỉnh dữ liệu quan trắc môi trường nước ngoại lai	6
2.2 Diền dữ liệu bị khuyết sử dụng mô hình lai theo không gian và thời gian	7
2.3 Dự báo chất lượng không khí bằng mô hình lai LSTM-MA	8
3 Phương pháp thực hiện	9
3.1 Nền tảng lý thuyết	9
3.1.1 Dữ liệu quan trắc không khí	9
3.1.2 Chỉ số chất lượng không khí	10
3.1.3 Phương pháp nội suy không gian IDW	13
3.2 Chuẩn bị dữ liệu	14
3.2.1 Nguồn dữ liệu thô	14
3.2.2 Phân tích và tiền xử lý	16
3.3 Phát hiện và hiệu chỉnh dữ liệu ngoại lai	21
3.3.1 Xác định độ phân tán và IQR factor	22
3.3.2 Các bước hiệu chỉnh dữ liệu ngoại lai	23
3.3.3 Làm mượt dữ liệu	23
3.4 Diền dữ liệu bị khuyết	24
3.4.1 Phân tích tính khả thi phương pháp nội suy không gian	24
3.4.2 Thủ nghiệm với Random Forest	26
3.5 Mô hình dự đoán chất lượng không khí	28
3.5.1 Tiền xử lý và chuẩn bị dữ liệu huấn luyện	28
3.5.2 Xây dựng mô hình lai LSTM - CNN	29
3.6 Triển khai hệ thống theo thời gian thực	31
3.6.1 Thiết kế hệ thống	31

3.6.2 Chi tiết về các khói	32
4 Kết quả và nhận xét	34
4.1 Quá trình phát hiện và hiệu chỉnh dữ liệu ngoại lai	34
4.2 Quá trình lắp đầy phần dữ liệu khuyết	39
4.3 Kết quả mô hình dự đoán AQI	45
4.4 Thực nghiệm hệ thống	53
4.4.1 Triển khai hệ thống trên AWS	53
4.4.2 Thiết kế chi tiết ứng dụng web	55
5 Kết luận và hướng nghiên cứu tiếp theo	58
5.1 Kết luận	58
5.2 Đề xuất hướng nghiên cứu tiếp theo	59
Tài liệu tham khảo	61

Danh sách hình vẽ

1.1	Điễn biến nồng độ trung bình năm của hai chất PM _{2.5} và PM ₁₀ ở thành phố Hồ Chí Minh	2
3.1	Biểu đồ chuỗi thời gian hiển thị nồng độ PM _{2.5} trung bình hàng ngày	10
3.2	Vị trí 6 trạm quan trắc trên bản đồ	15
3.3	Ma trận tương quan giữa các chất ở trạm số 2	16
3.4	Biểu đồ nồng độ PM _{2.5} theo thời gian ở các trạm	17
3.5	Biểu đồ nồng độ O ₃ theo thời gian ở các trạm	18
3.6	Biểu đồ nồng độ CO theo thời gian ở các trạm	19
3.7	Biểu đồ nồng độ NO ₂ theo thời gian ở các trạm	20
3.8	Minh họa MA với window size = 3	24
3.9	Ma trận tương quan của các chất ô nhiễm giữa các trạm	25
3.10	Cách hoạt động của Random Forest	27
3.11	Biểu đồ phân bố chất ô nhiễm chính của các trạm	29
3.12	Mô hình lai CNN-LSTM áp dụng vào hệ thống	30
3.13	Sơ đồ khái niệm	31
4.1	Dữ liệu theo thời gian của các chất ô nhiễm đã được hiệu chỉnh outlier	34
4.2	Dữ liệu theo thời gian của các chất ô nhiễm qua MA	36
4.3	Dữ liệu được hiệu chỉnh dùng Random Forest	40
4.4	Ma trận tương quan của các chất ô nhiễm sau khi điền dữ liệu khuyết	41
4.5	Biểu đồ dự đoán AQI và chất ô nhiễm chính trên trạm 2	48
4.6	Biểu đồ dự đoán AQI và chất ô nhiễm chính trên trạm 2	49
4.7	Biểu đồ dự đoán AQI và chất ô nhiễm chính trên trạm 2	50
4.8	Các biểu đồ dự đoán AQI trong 8 giờ tiếp theo	51
4.9	Các biểu đồ dự đoán AQI trong 16 giờ tiếp theo	51
4.10	Các biểu đồ dự đoán AQI trong 24 giờ tiếp theo	52
4.11	Thiết kế hệ thống dự đoán AQI trên AWS	53
4.12	Giao diện main page của hệ thống	55
4.13	Giao diện forecast page	56
4.14	Giao diện login page	57

Danh sách bảng

2.1	Tổng hợp một số nghiên cứu ứng dụng mô hình LSTM trên thế giới và Việt Nam.	5
3.1	Khoảng giá trị AQI và chất lượng không khí tương ứng	11
3.2	Các giá trị điểm ngắt tương ứng với các chất ô nhiễm	13
3.3	Tọa độ các trạm quan trắc	15
3.4	Độ lệch chuẩn cho từng chất ô nhiễm tại mỗi trạm	22
4.1	Thống kê số lượng điểm dữ liệu ngoại lai được hiệu chỉnh	35
4.2	Bảng so sánh khoảng dao động dữ liệu trung bình 24h trước và sau IQR .	35
4.3	Bảng so sánh khoảng dao động dữ liệu trung bình 24h trước và sau MA. .	37
4.4	Bảng so sánh độ lệch chuẩn và trung bình sau hiệu chỉnh dữ liệu	38
4.5	So sánh kết quả điền dữ liệu khuyết PM _{2.5} tại trạm 2	42
4.6	So sánh kết quả điền dữ liệu khuyết O ₃ tại trạm 3	42
4.7	So sánh kết quả điền dữ liệu khuyết CO tại trạm 4	43
4.8	So sánh kết quả điền dữ liệu khuyết NO ₂ tại trạm 6	43
4.9	Các chỉ số đánh giá của mô hình Random Forest và KNN của các chất .	44
4.10	So sánh kết quả AQI dự đoán với thực tế	46
4.11	So sánh hiệu suất giữa các mô hình về dự báo AQI	47
4.12	Các chỉ số đánh giá của trạm 2	48
4.13	Các chỉ số đánh giá của trạm 3	49
4.14	Các chỉ số đánh giá của trạm 6	50
4.15	So sánh các chỉ số đánh giá giữa các mô hình trên trạm số 4	52

Danh mục chữ viết tắt

AWS	Amazon Web Services
AI	Artificial Intelligence
AQI	Air Quality Index
ARIMA	AutoRegressive Integrated Moving Average
CNN	Convolutional Neural Network
CO	Carbon Monoxide
CSV	Comma Separated Values
DNS	Domain Name System
EPA	Environmental Protection Agency
GAN	Generative Adversarial Networks
GRU	Gated Recurrent Unit
IDW	Inverse Distance Weighting
IoT	Internet of Things
IQR	InterQuartile Range
JSON	JavaScript Object Notation
K-NN	K-Nearest Neighbors
LSTM	Long Short-Term Memory
MA	Moving Average
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MICE	Multiple Imputation by Chained Equations
MSE	Mean Squared Error
NO₂	Nitrogen Dioxide
O₃	Ozone
PM2.5	Particulate Matter 2.5
PM10	Particulate Matter 10
RMSE	Root Mean Square Error
SaaS	Software as a Service
SO₂	Sulfur Dioxide

TSP	Total Suspended Particulates
TadGAN	Time-series Anomaly Detection using Generative Adversarial Networks
TCN	Temporal Convolutional Networks
UV	UltraViolet

Chương 1

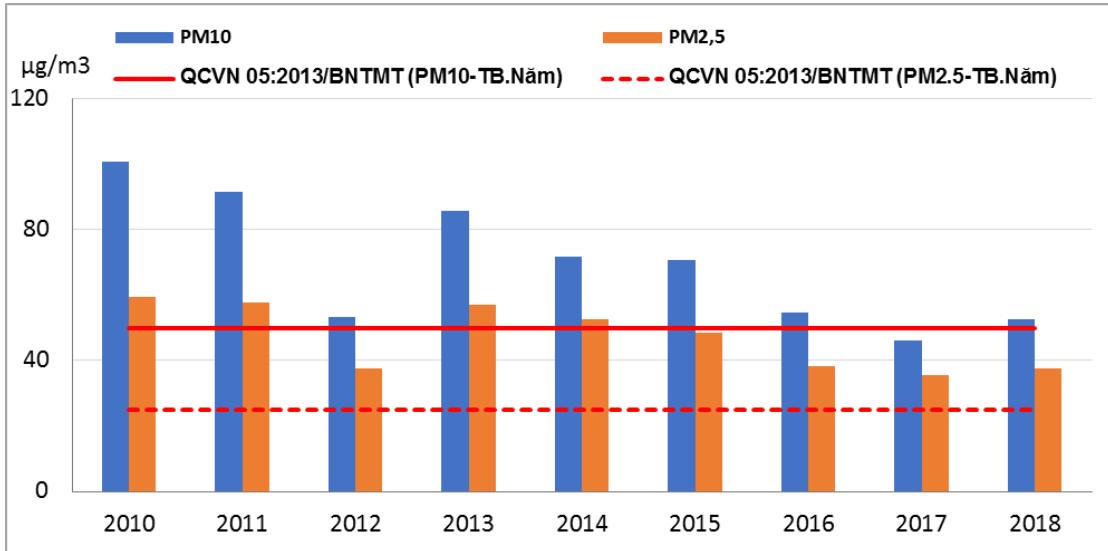
Tổng quan

1.1 Đặt vấn đề

Quan trắc môi trường là quá trình theo dõi và quan sát một cách có hệ thống các yếu tố môi trường như đất, nước, không khí, tiếng ồn và ánh sáng, nhằm đánh giá hiện trạng và diễn biến chất lượng môi trường. Ngày nay, ô nhiễm không khí là một vấn đề sức khỏe cộng đồng ngày càng trở nên nghiêm trọng, gây ra hàng triệu ca tử vong mỗi năm [1]. Ô nhiễm không khí không chỉ đe dọa con người mà còn ảnh hưởng đến toàn bộ hệ sinh thái Trái đất. Với sự phát triển của công nghệ, trí tuệ nhân tạo đang được ứng dụng rộng rãi trong lĩnh vực quan trắc môi trường, đặc biệt là giám sát và đo lường chất lượng không khí với dữ liệu được thu thập từ nhiều trạm quan trắc khác nhau. Đây là bài toán không gian khó, gồm hai vấn đề cần phải giải quyết:

- Loại bỏ nhiễu và xử lý dữ liệu bị thiếu, hiệu chỉnh dữ liệu theo thời gian thực cho các trạm quan trắc.
- Xây dựng các mô hình dự đoán AQI hiệu quả bằng cách kết hợp dữ liệu từ nhiều trạm quan trắc khác nhau (bài toán không gian).

Ngoài ra với dữ liệu quan trắc không khí, còn những bài toán khác như: Phát triển hệ thống cảnh báo sớm, có khả năng dự đoán và cảnh báo người dân về các sự kiện ô nhiễm không khí nghiêm trọng; phân tích mối liên hệ giữa ô nhiễm không khí và các bệnh lý về hô hấp, tim mạch, từ đó đưa ra các khuyến cáo về bảo vệ sức khỏe; đánh giá tác động của ô nhiễm không khí đến hệ sinh thái, biến đổi khí hậu, năng suất nông nghiệp...



HÌNH 1.1: Diễn biến nồng độ trung bình năm của hai chất PM_{2.5} và PM₁₀ ở thành phố Hồ Chí Minh

Sự phát triển của Internet of Things đã tạo điều kiện thuận lợi cho việc thu thập và truyền tải dữ liệu quan trắc môi trường. Tuy nhiên, việc xử lý dữ liệu từ nhiều trạm quan trắc trong một địa bàn thành phố như Thành phố Hồ Chí Minh còn gặp phải những thách thức như dữ liệu bị thiếu do các trạm quan trắc bị lỗi hoặc bị tắt định kỳ trong một khoảng thời gian dài hoặc dữ liệu bị nhiễu vì cảm biến hoạt động không ổn định. Những vấn đề này, cùng với sự phức tạp của dữ liệu từ nhiều trạm quan trắc, đặt ra nhiều thách thức cho việc hiệu chỉnh dữ liệu, xây dựng mô hình dự đoán, và phân tích xu hướng ô nhiễm không khí.

Để góp phần giải quyết những thách thức trên, luận văn của tôi tập trung vào việc nghiên cứu giải pháp và thiết kế một hệ thống AI nhằm hiệu chỉnh dữ liệu và dự đoán chỉ số AQI cho thành phố Hồ Chí Minh dựa trên dữ liệu thu thập được từ các trạm quan trắc môi trường xung quanh thành phố.

1.2 Mục tiêu nghiên cứu

Nhằm góp phần giải quyết những thách thức đã trình bày, luận văn này tập trung vào việc đạt được các mục tiêu sau:

1. Xử lý và hiệu chỉnh dữ liệu quan trắc không khí, giải quyết các vấn đề như nhiễu và thiếu sót trong dữ liệu quan trắc không khí.

2. Xây dựng một hệ thống thông minh sử dụng các thuật toán học máy để dự đoán chất lượng không khí dựa trên chỉ số AQI theo thời gian thực.

Phạm vi của luận văn này được giới hạn trong việc sử dụng dữ liệu thu thập từ sáu trạm quan trắc có sẵn trên địa bàn thành phố Hồ Chí Minh. Trọng tâm của luận văn là nghiên cứu, phát triển các phương pháp hiệu chỉnh dữ liệu và xây dựng mô hình dự đoán AQI. Các bài toán khác sẽ được xem xét trong các nghiên cứu tiếp theo.

1.3 Hướng tiếp cận và giải quyết bài toán

Để đạt được các mục tiêu nghiên cứu, tôi sẽ bắt đầu từ việc nghiên cứu nguồn dữ liệu thu được từ trạm quan trắc, sau đó sẽ xây dựng mô hình và hệ thống nhằm giải quyết bài toán đã đặt ra, cụ thể như sau:

1. Phân tích dữ liệu: Tiền xử lý dữ liệu quan trắc không khí từ sáu trạm quan trắc phố Hồ Chí Minh. Phân tích các đặc trưng thống kê của dữ liệu, bao gồm trung bình, độ lệch chuẩn, phân phối, và mối tương quan giữa các chỉ số ô nhiễm. Phân tích xu hướng biến động của chất lượng không khí theo thời gian và không gian.

2. Xử lý dữ liệu: Áp dụng các phương pháp thống kê để phát hiện và xử lý các dữ liệu ngoại lai, sau đó sử dụng các kỹ thuật nội suy hoặc thay thế bằng giá trị trung bình để xử lý dữ liệu bị khuyết.

3. Lựa chọn mô hình: Nghiên cứu và đánh giá các mô hình học máy phù hợp để dự đoán chỉ số AQI. Lựa chọn mô hình tối ưu dựa trên các tiêu chí về độ chính xác, và độ phức tạp.

4. Xây dựng hệ thống pipeline: Tích hợp các bước xử lý dữ liệu và mô hình dự đoán vào một hệ thống pipeline để hoạt động theo thời gian thực.

5. Đánh giá và cải thiện hệ thống: So sánh kết quả với những nghiên cứu liên quan và đề xuất các phương án cải thiện hệ thống.

Hướng tiếp cận này giúp tôi giải quyết bài toán một cách khoa học và hiệu quả, từ việc phân tích, xử lý dữ liệu đến xây dựng mô hình dự đoán và triển khai hệ thống theo thời gian thực.

Chương 2

Các nghiên cứu liên quan

Nghiên cứu về quan trắc không khí và dự đoán chất lượng không khí (AQI) đang thu hút sự quan tâm ngày càng lớn trên thế giới, thể hiện qua sự phát triển của các phương pháp học sâu trong dự đoán ô nhiễm không khí [2]. Chỉ số AQI đóng vai trò quan trọng trong việc đánh giá và quản lý chất lượng không khí [3]. Các phương pháp dự đoán AQI cũng ngày càng đa dạng và tiên tiến, từ các mô hình truyền thống như mô hình phân tán, mô hình thống kê (hồi quy tuyến tính, ARIMA), và mô hình thụ động, đến các mô hình sử dụng AI [4].

Ở Việt Nam, lĩnh vực quan trắc môi trường nói chung và quan trắc không khí nói riêng đang trong giai đoạn phát triển và chủ yếu chỉ mới áp dụng nhiều trong lĩnh vực tài nguyên nước. Dối với dự báo chất lượng không khí, các nghiên cứu sử dụng mô hình LSTM còn hạn chế về số lượng, một phần bởi vì Bộ dữ liệu đầu vào cho mô hình LSTM có yêu cầu cao nên cần phải có các giải pháp tiền xử lý số liệu phù hợp để cho ra dự báo tối ưu [5]. Mặc dù đã có những nỗ lực trong việc nghiên cứu và ứng dụng AQI [6], nhưng việc triển khai hệ thống quan trắc không khí vẫn gặp nhiều khó khăn do số lượng trạm quan trắc còn hạn chế, dẫn đến thiếu hụt dữ liệu đầy đủ và đại diện cho việc nghiên cứu và dự đoán AQI [7].

BẢNG 2.1: Tổng hợp một số nghiên cứu ứng dụng mô hình LSTM trên thế giới và Việt Nam. [7]

Nghiên cứu	Chất ô nhiễm	Quốc gia	Khoảng thời gian của dữ liệu	Mô hình sử dụng
Yan và cs	AQI	Trung Quốc	2015 - 2016	BPNN, CNN, LSTM
Jiao và cs	AQI	Trung Quốc	10/2023 - 9/2018	CNN-LSTM
Belavadi và cs	AQI	Ấn Độ	9/3/2019 - 13/4/2019	LSTM
Duan và cs	AQI	Trung Quốc	01/2015 - 03/2022	LSTM, CNN-LSTM, DBO-LSTM, CEEMDAN-LSTM
Yammahi và cs	NO ₂	UAE	2019 - 2020	LSTM, NAR-NN ARIMA, SARIMA
Navares và cs	CO, NO ₂ , O ₃ , PM ₁₀ , và SO ₂	Tây Ban Nha	2001 - 2013	LSTM-RNN
Chang và cs	PM _{2.5}	Đài Loan	2013 - 2017	Aggregated-LSTM
Wen và cs	PM _{2.5}	Trung Quốc	2016 - 2017	CNN-LSTM
Jung và cs	PM ₁₀	Hàn Quốc	2009 - 2019	DNN, RNN, LSTM
Rakholia và cs	PM _{2.5}	Việt Nam	02/2021 - 12/2021	XGBoost, GDRegressor, 1D CNN-LSTM, Prophet
Hung	CO, NO _x , O ₃ , PM _{2.5} , PM ₁₀ , SO ₂	Việt Nam	2010 - 2018	CNN-LSTM

Để giải quyết bài toán dự đoán AQI, cần tập trung vào một số vấn đề quan trọng. Đầu tiên, việc phát hiện và hiệu chỉnh các giá trị ngoại lai là bước then chốt để đảm bảo tính chính xác của dữ liệu [8]. Tiếp theo, xử lý giá trị thiếu (missing value) là một thách thức không thể tránh khỏi trong dữ liệu quan trắc thời gian thực, đặc biệt là khi dữ liệu được thu thập từ nhiều trạm quan trắc. Cần lưu ý đến hai dạng missing value: missing value rời rạc và missing value có tính chu kỳ. Việc xử lý missing value với tỷ lệ lớn và chu kỳ thường xuyên là một thách thức lớn, đòi hỏi các kỹ thuật phức tạp như nội suy không gian kết hợp với mô hình chuỗi thời gian hoặc các mô hình học máy [9]. Sau khi đã xử lý dữ liệu, bước tiếp theo là lựa chọn mô hình phù hợp để dự đoán AQI. Các mô hình học máy như LSTM (Long Short-Term Memory) đang được ứng dụng rộng rãi và cho thấy hiệu quả cao trong việc dự đoán AQI [5]. Cuối cùng, cần xây dựng hệ thống pipeline để vận hành hệ thống dự đoán theo thời gian thực, tận dụng tối đa sức mạnh của phần cứng và phần mềm.

2.1 Phát hiện và hiệu chỉnh dữ liệu quan trắc môi trường nước ngoại lai

Trong luận văn "Xây dựng giải pháp phát hiện bất thường và hiệu chỉnh dữ liệu quan trắc theo thời gian thực" của tác giả Tống Quốc Sang đã nghiên cứu và đề xuất một giải pháp để phát hiện và hiệu chỉnh dữ liệu bất thường. Bộ dữ liệu sử dụng trong nghiên cứu của tác giả đến từ hệ thống quan trắc nước, cũng có nhiều điểm tương đồng với dữ liệu quan trắc không khí [8].

Tác giả đã sử dụng mô hình Time-series Anomaly Detection using Generative Adversarial Networks (TadGAN) để phát hiện các điểm bất thường trong dữ liệu chuỗi thời gian. Mô hình này bao gồm bốn thành phần chính: encoder, generator, criticX và criticZ. Encoder có nhiệm vụ chuyển đổi dữ liệu đầu vào sang không gian tiềm ẩn, generator tạo dữ liệu giả từ biểu diễn tiềm ẩn, criticX phân biệt giữa dữ liệu thật và dữ liệu giả trong không gian đầu vào, và criticZ phân biệt giữa vector tiềm ẩn thật và vector tiềm ẩn được mã hóa từ dữ liệu.

Để hiệu chỉnh dữ liệu ngoại lai, tác giả sử dụng generator của mô hình TadGAN để dự đoán lại giá trị tại các thời điểm được xác định là bất thường. Kết quả thực nghiệm cho thấy mô hình TadGAN có thể phát hiện các điểm bất thường với độ chính xác cao

và hiệu chỉnh dữ liệu với sai số khoảng 2% . Mặc dù đạt được kết quả khả quan, tác giả cũng chỉ ra một số yếu tố cần cải thiện trong tương lai:

- Cần nghiên cứu thêm để tối ưu hóa các siêu tham số của mô hình TadGAN nhằm nâng cao hiệu suất.
- Bổ sung đặc trưng: Nên bổ sung thêm các đặc trưng từ bộ dữ liệu để cải thiện khả năng học và dự đoán của mô hình.
- Việc mở rộng phạm vi dữ liệu huấn luyện sẽ giúp mô hình trở nên mạnh mẽ và tổng quát hơn.

2.2 Điền dữ liệu bị khuyết sử dụng mô hình lai theo không gian và thời gian

Nhiều nghiên cứu đã đề xuất các phương pháp khác nhau để giải quyết vấn đề này, từ các kỹ thuật nội suy không gian đơn giản đến các mô hình học máy phức tạp.

Jun và cộng sự đã kết hợp phương pháp nội suy không gian IDW với mô hình chuỗi thời gian ARIMA để dự đoán giá trị PM_{2.5} bị thiếu trong bài báo "A Hybrid Model for Spatiotemporal Imputation of Missing Air Quality Data" [10]. IDW được sử dụng để ước tính giá trị PM_{2.5} tại các vị trí không có dữ liệu dựa trên giá trị PM_{2.5} của các trạm lân cận. Sau đó, mô hình ARIMA được sử dụng để dự đoán giá trị PM_{2.5} tại các thời điểm bị thiếu dựa trên chuỗi thời gian PM_{2.5} đã được nội suy bằng IDW. Kết quả thực nghiệm cho thấy mô hình lai IDW-ARIMA đạt hiệu quả cao trong việc lấp đầy giá trị bị thiếu, vượt trội hơn so với việc sử dụng IDW hoặc ARIMA riêng lẻ. Mô hình này đặc biệt hiệu quả khi dữ liệu bị thiếu có tính chất không gian và thời gian rõ rệt. Tuy nhiên, mô hình này có độ phức tạp cao và yêu cầu dữ liệu phải có tính chất không gian và thời gian rõ rệt, đồng thời, việc áp dụng cho dữ liệu có số lượng trạm quan trắc lớn có thể gặp khó khăn về mặt tính toán.

Trong bài báo "Missing Value Imputation for Air Quality Data Using a Hybrid Spatiotemporal Model" của Li và cộng sự [9], nhóm tác giả đề xuất một phương pháp khác, sử dụng mô hình lai kết hợp k-NN và Random Forest để lấp đầy các giá trị bị thiếu trong dữ liệu quan trắc không khí. k-NN được sử dụng để tìm kiếm các trạm quan

trắc lân cận có dữ liệu đầy đủ tại thời điểm cần điền giá trị thiếu. Random Forest được sử dụng để xây dựng mô hình dự đoán giá trị thiếu dựa trên dữ liệu của các trạm lân cận và các thông tin thời gian. Kết quả thực nghiệm cho thấy mô hình lai này đạt hiệu quả cao trong việc xử lý dữ liệu bị thiếu với tỷ lệ cao và phân bố ngẫu nhiên. Tuy nhiên, mô hình này có thể gặp khó khăn khi dữ liệu bị thiếu tập trung theo một xu hướng nhất định.

Trong nghiên cứu này, tôi sẽ đánh giá và lựa chọn phương pháp phù hợp để xử lý dữ liệu bị thiếu, nhằm xây dựng một mô hình dự đoán AQI hiệu quả cho các trạm quan trắc ở Hồ Chí Minh.

2.3 Dự báo chất lượng không khí bằng mô hình lai LSTM-MA

Trong bài báo "Dự báo chất lượng không khí bằng mô hình LSTM-MA trường hợp sử dụng dữ liệu tại trạm quan trắc tự động Ngã tư Giếng Nước, tỉnh Bà Rịa - Vũng Tàu" [5], tác giả Hồ Minh Dũng và Khổng Doãn An Khang đã sử dụng bộ dữ liệu là nồng độ trung bình giờ của các thông số PM_{2.5}, PM₁₀, CO, NO₂, O₃ và SO₂, được đo tại trạm quan trắc tự động Ngã tư Giếng Nước, tỉnh Bà Rịa - Vũng Tàu. Thời gian thu thập dữ liệu từ ngày 18/01/2020 đến ngày 31/12/2022. Dữ liệu này có tỉ lệ khuyết thấp và vẫn chứa các giá trị ngoại lai, được tác giả chia thành ba tập: huấn luyện, kiểm định và thử nghiệm theo tỷ lệ 70:15:15.

Mô hình được tác giả sử dụng là mô hình LSTM kết hợp với bộ lọc trung bình trượt (MA). Tông đó bộ lọc MA được sử dụng để làm mịn dữ liệu đầu vào, giúp cải thiện hiệu suất của mô hình LSTM. Nghiên cứu cho ra kết quả dự báo chính xác chất lượng không khí trong thời gian ngắn hạn khoảng một ngày với RMSE = 3.05, MAE = 2.17 và MAPE = 3.19% . Trong khi đó mô hình cho kết quả khả quan khi dự báo trong thời gian dài hơn (hai tuần) với RMSE = 22.79, MAE = 15.74 và MAPE = 24.38% .

Chương 3

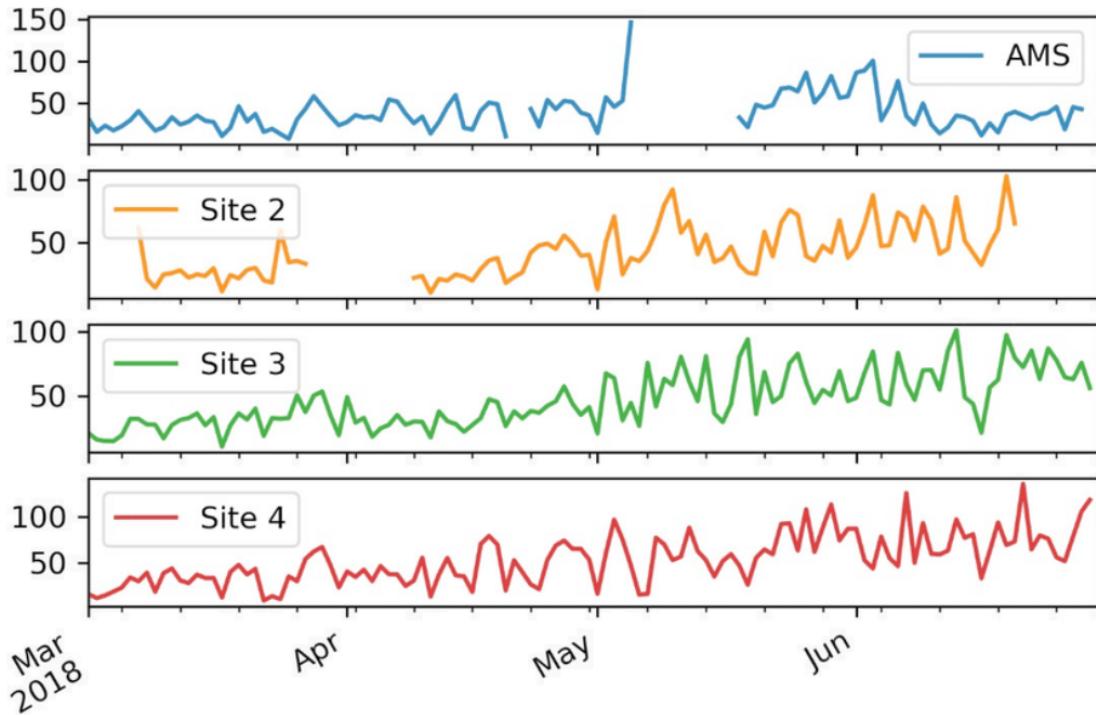
Phương pháp thực hiện

3.1 Nền tảng lý thuyết

3.1.1 Dữ liệu quan trắc không khí

Dữ liệu quan trắc là một tập hợp các giá trị đo lường được thu thập từ các thiết bị cảm biến hoặc các phương pháp quan sát khác, nhằm theo dõi và đánh giá các hiện tượng, đặc điểm, hoặc sự thay đổi của một đối tượng hoặc môi trường cụ thể [3]. Ví dụ, trong quan trắc không khí, các thiết bị cảm biến được sử dụng để đo lường nồng độ các chất ô nhiễm như PM_{2.5} (bụi mịn), PM₁₀ (bụi thô), CO (carbon monoxide), NO₂ (nitrogen dioxide), O₃ (ozone), và SO₂ (sulfur dioxide). Ngoài ra, các thông số khí tượng như nhiệt độ, độ ẩm, áp suất, chỉ số UV, tốc độ gió và hướng gió cũng được thu thập để phân tích và đánh giá chất lượng không khí một cách toàn diện.

Một dạng quan trọng của dữ liệu quan trắc là dữ liệu chuỗi thời gian (time series data), bao gồm các giá trị đo lường được ghi nhận theo thời gian, có thể theo phút, theo giờ hoặc theo ngày. Dữ liệu chuỗi thời gian quan trắc không khí bao gồm các quan sát được ghi lại từ các trạm quan trắc khí tượng, cho phép phân tích sự biến động và dự đoán các xu hướng môi trường trong tương lai.



HÌNH 3.1: Biểu đồ chuỗi thời gian hiển thị nồng độ PM_{2.5} trung bình hàng ngày được đo tại các trạm AMS [11]

3.1.2 Chỉ số chất lượng không khí

i) Khái niệm và phân loại

Chỉ số chất lượng không khí tên khoa học là Air Quality Index (AQI) là một chỉ số báo cáo chất lượng không khí trong một khoảng thời gian nhất định. Đây được coi là một thước đo đơn giản hóa mức độ ô nhiễm không khí, cho biết không khí xung quanh ta là sạch hay ô nhiễm, ô nhiễm đến mức độ nào. Rủi ro đối với sức khỏe cộng đồng càng cao khi chỉ số AQI càng lớn. Chỉ số AQI tập trung vào sự ảnh hưởng tới sức khỏe người dân có thể gặp trong vòng vài giờ hoặc vài ngày sau khi hít thở không khí ô nhiễm [12].

Chỉ số chất lượng không khí được tính dựa trên chỉ số nồng độ các chất ô nhiễm không khí (PM_{2.5}, PM₁₀, CO, NO₂, O₃, và SO₂ (sulfur dioxide)) được đo tại các trạm quan trắc theo giờ hoặc theo ngày. AQI được phân loại theo thang điểm tương ứng với biểu tượng và các màu sắc để cảnh báo chất lượng không khí và mức độ ảnh hưởng tới sức khỏe con người [13].

BẢNG 3.1: Khoảng giá trị AQI và chất lượng không khí tương ứng

Khoảng giá trị AQI	Chất lượng không khí	Màu sắc	Mã màu RBG	Ảnh hưởng tới sức khỏe
0 - 50	Tốt	Xanh	0;228;0	Chất lượng không khí tốt, không ảnh hưởng tới sức khỏe
51 - 100	Trung bình	Vàng	255;255;0	Chất lượng không khí ở mức chấp nhận được. Tuy nhiên, đối với những người nhạy cảm (người già, trẻ em, người mắc các bệnh hô hấp, tim mạch...) có thể chịu những tác động nhất định tới sức khỏe.
101 - 150	Kém	Da cam	255;126;0	Những người nhạy cảm gặp phải các vấn đề về sức khỏe, những người bình thường ít ảnh hưởng.
151 - 200	Xấu	Đỏ	255;0;0	Những người bình thường bắt đầu có các ảnh hưởng tới sức khỏe, nhóm người nhạy cảm có thể gặp những vấn đề sức khỏe nghiêm trọng hơn.
201 - 300	Rất xấu	Tím	143;63;151	Cảnh báo hưởng tới sức khỏe: mọi người bị ảnh hưởng tới sức khỏe nghiêm trọng hơn.
301 - 500	Nguy hại	Nâu	126;0;35	Cảnh báo khẩn cấp về sức khỏe: Toàn bộ dân số bị ảnh hưởng tới sức khỏe tới mức nghiêm trọng.

ii) Công thức tính toán

Phương pháp tính toán chỉ số chất lượng không khí Việt Nam (VN-AQI) từ dữ liệu quan trắc của trạm quan trắc không khí tự động, liên tục được ban hành kèm theo Quyết định số 1459/QĐ-TCMT ngày 12/11/2019 của Tổng Cục Môi trường [13]. Cơ quan Bảo vệ Môi trường Hoa Kỳ (EPA) thông qua sự công nhận của quốc tế đã đưa ra công thức tính toán chỉ số AQI trung bình trong một giờ được áp dụng chung trên thế giới, cụ thể như sau:

$$I_{\text{AQI}} = \frac{(I_{\text{Hi}} - I_{\text{Lo}})}{(BP_{\text{Hi}} - BP_{\text{Lo}})} \times (C_p - BP_{\text{Lo}}) + I_{\text{Lo}}$$

Trong đó:

- Cp: Nồng độ của chất ô nhiễm trong không khí (trung bình 1 giờ).
- BP_{Hi}: Điểm ngắt nồng độ cao hơn cho chất ô nhiễm tương ứng.
- BP_{Lo}: Điểm ngắt nồng độ thấp hơn cho chất ô nhiễm tương ứng.
- I_{Hi}: Chỉ số chất lượng không khí tương ứng với BP_{Hi}.
- I_{Lo}: Chỉ số chất lượng không khí tương ứng với BP_{Lo}.

Theo Quyết định 1459/QĐ-TCMT của Tổng cục Môi trường Việt Nam, điểm ngắt nồng độ và chỉ số chất lượng không khí tương ứng theo bảng dưới đây [13]:

BẢNG 3.2: Các giá trị điểm ngắt tương ứng với các chất ô nhiễm

i	I_i	Giá trị BP_i quy định đối với từng thông số (Đơn vị: $\mu\text{g}/\text{m}^3$)						
		O ₃ (1h)	O ₃ (8h)	CO	SO ₂	NO ₂	PM ₁₀	PM _{2.5}
1	0	0	0	0	0	0	0	0
2	50	160	100	10.000	125	100	50	25
3	100	200	120	30.000	350	200	150	50
4	150	300	170	45.000	550	700	250	80
5	200	400	210	60.000	800	1.200	350	150
6	300	800	400	90.000	1.600	2.350	420	250
7	400	1.000	-	120.000	2.100	3.100	500	350
8	500	≥ 1.200	-	≥ 150.000	≥ 2.630	≥ 3.850	≥ 600	≥ 500

Sau khi đã tính được I_{AQI} của các chất ô nhiễm (CO, NO₂, O₃, PM_{2.5}), giá trị AQI của mỗi trạm sẽ là giá trị I_{AQI} cao nhất. Tuy nhiên, để tính được chỉ số AQI cho thành phố Hồ Chí Minh dựa trên AQI của các trạm quan trắc cũng là một bài toán nan giải. Trang web của Tổng cục Môi trường đưa ra AQI của thành phố Hồ Chí Minh thường trùng khớp với giá trị AQI cao nhất được ghi nhận tại các trạm quan trắc trong thành phố. Cách tính này trực quan, đơn giản và dễ thực hiện nhất. Tuy nhiên vẫn những mặt hạn chế như chỉ số AQI của toàn thành phố sẽ bị ảnh hưởng nhiều bởi trạm quan trắc đặt tại khu vực điểm nóng ô nhiễm nên không phản ánh đúng tình hình chung của thành phố, hay thiếu thông tin về sự phân bố ô nhiễm trong thành phố.

3.1.3 Phương pháp nội suy không gian IDW

Nội suy không gian là một phương pháp được sử dụng để ước tính giá trị của một biến tại các vị trí không có dữ liệu dựa trên các giá trị đã biết tại các vị trí lân cận. Phương pháp này đặc biệt hữu ích trong các trường hợp dữ liệu bị thiếu hoặc phân bố không đều [9].

IDW (Inverse Distance Weighting) là một trong những phương pháp nội suy không gian đơn giản và phổ biến nhất. Phương pháp này dựa trên giả định rằng các điểm dữ liệu gần nhau có sự tương quan cao hơn so với các điểm dữ liệu ở xa nhau. Do

đó, trọng số được gán cho mỗi điểm dữ liệu dựa trên khoảng cách nghịch đảo giữa điểm đó và điểm cần ước tính [10], cụ thể theo công thức sau:

$$Z(x_0) = \frac{\sum_{i=1}^n w_i z_i}{\sum_{i=1}^n w_i}$$

Trong đó:

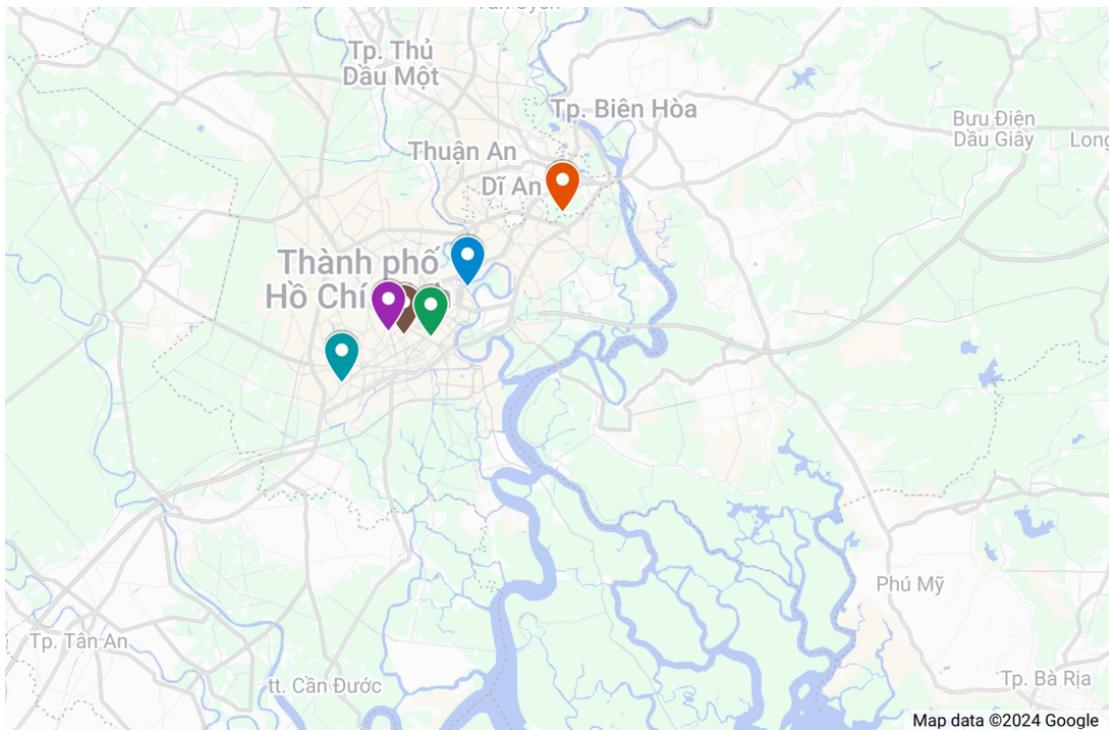
- $Z(x_0)$ là giá trị ước tính tại vị trí x_0 .
- z_i là giá trị đã biết tại vị trí x_i .
- w_i là trọng số được gán cho điểm x_i , được tính bằng công thức: $w_i = \frac{1}{d(x_0, x_i)^p}$.
- $d(x_0, x_i)$ là khoảng cách giữa vị trí x_0 và x_i .
- p là tham số kiểm soát ảnh hưởng của khoảng cách, thường được chọn là 2.

Trong trường hợp dữ liệu chuỗi thời gian quan trắc không khí bị thiếu, IDW có thể được áp dụng để ước tính giá trị tại các thời điểm bị thiếu dựa trên dữ liệu của các trạm quan trắc lân cận [9]. Đối với mỗi trạm quan trắc có dữ liệu bị thiếu, đầu tiên xác định các trạm lân cận có dữ liệu đầy đủ tại thời điểm đó. Sau đó tính toán khoảng cách giữa trạm có dữ liệu bị thiếu và các trạm lân cận. Công thức IDW sẽ được áp dụng để tính toán trọng số cho mỗi trạm lân cận dựa trên khoảng cách. Cuối cùng là tính toán giá trị ước tính tại trạm có dữ liệu bị thiếu bằng cách sử dụng công thức IDW.

3.2 Chuẩn bị dữ liệu

3.2.1 Nguồn dữ liệu thô

Dữ liệu của đề tài được thu thập từ sáu trạm quan trắc không khí trên địa bàn Thành phố Hồ Chí Minh. Mỗi trạm ghi nhận thông tin nồng độ các chất ô nhiễm (TSP, PM_{2.5}, O₃, CO, NO₂), các thông số khí tượng (nhiệt độ, độ ẩm, điểm sương, áp suất, lượng mưa, tốc độ gió, hướng gió) và chỉ số UV. Thông tin chi tiết về vị trí và tọa độ của sáu trạm quan trắc được thể hiện trong hình vẽ và bảng dưới đây.



HÌNH 3.2: Vị trí 6 trạm quan trắc trên bản đồ

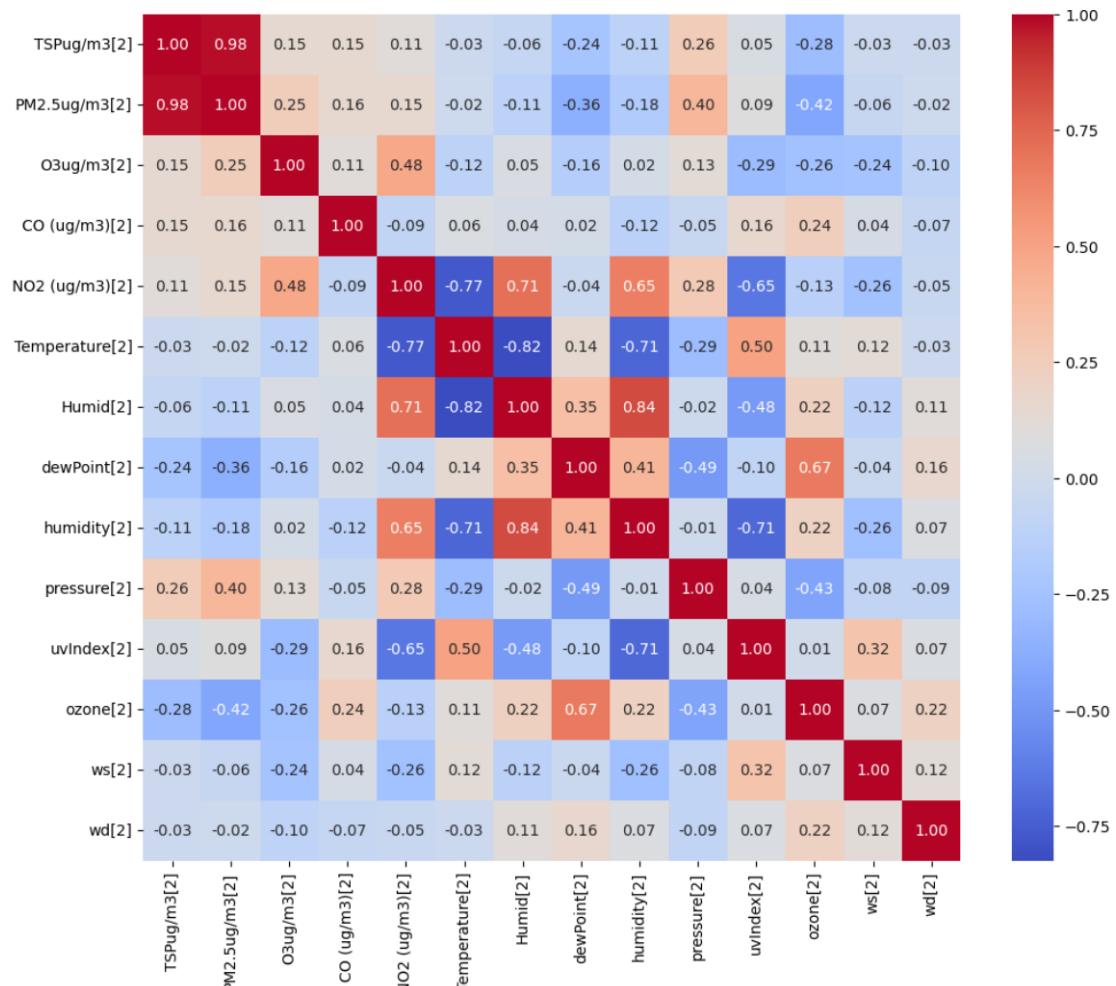
BẢNG 3.3: Tọa độ các trạm quan trắc

Station_id	Latitude	Longitude	Station Name
1	10.8744262	106.7936593	Khu đô thị Đại học Quốc gia TP. Hồ Chí Minh, TP Thủ Đức
2	10.7400843	106.6165755	Cạnh Phòng GDDT Q. Bình Tân- Q. Bình Tân
3	10.78061	106.654032	Trạm phát sóng MobiFone trong khuôn viên KCN Tân Bình, Q. Tân Bình
4	10.8158591	106.7173921	Trường THCS Cù Chính Lan, Thanh Đa, Phường 27, Q. Bình Thạnh
5	10.7763798	106.687783	Tòa soạn Báo Thanh Niên – Nguyễn Đình Chiểu - Q3
6	10.7781075	106.6661818	MobiFone Thành Thái, Q10

Định dạng của dữ liệu là chuỗi thời gian theo giờ, và được thu thập trong khoảng thời gian từ ngày 23/02/2021 đến ngày 26/01/2023. Trong luận văn này, tôi chia dữ liệu làm hai tập dữ liệu con. Tập con đầu tiên chiếm đa số, bắt đầu từ ngày

23/02/2021 đến hết ngày 14/01/2022 nhằm huấn luyện và thử nghiệm mô hình. Tập còn lại được sử dụng để mô phỏng dữ liệu theo thời gian thật nhằm mục đích kiểm thử và mô phỏng hệ thống khi đã hoàn thiện.

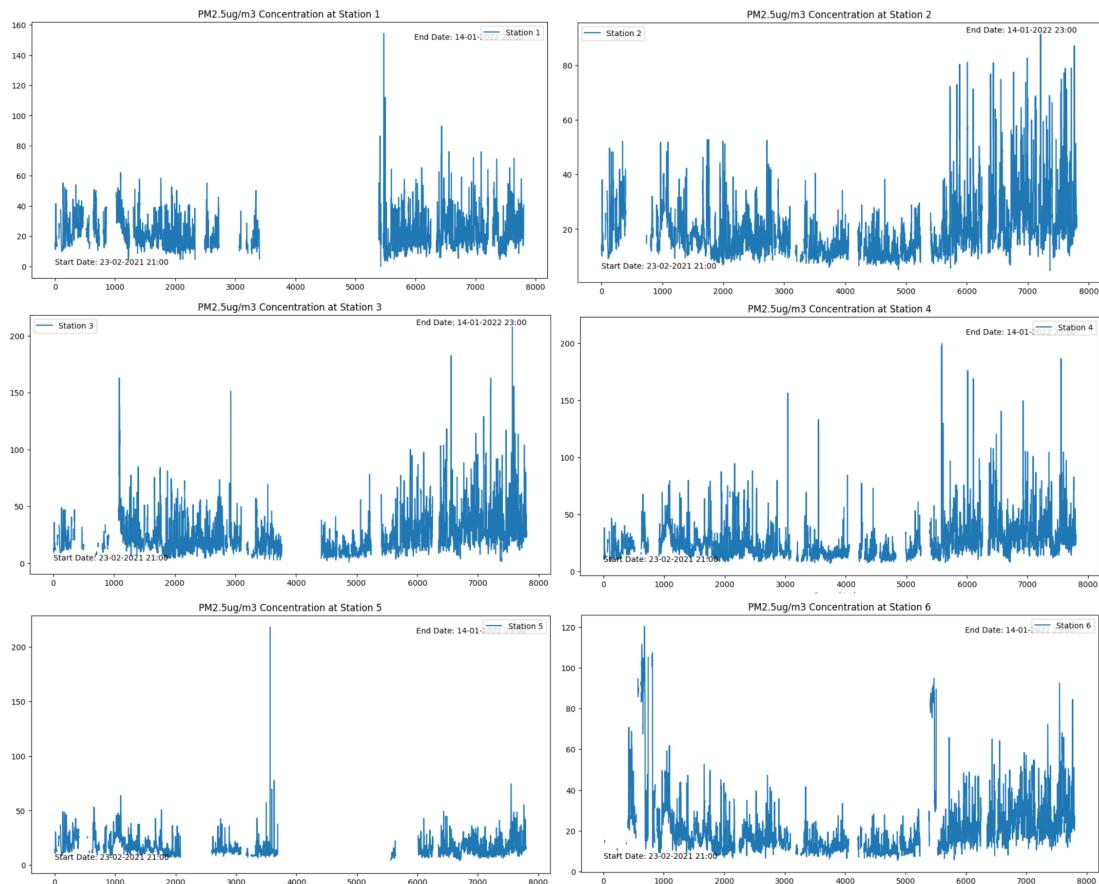
3.2.2 Phân tích và tiền xử lý



HÌNH 3.3: Ma trận tương quan giữa các chất ở trạm số 2

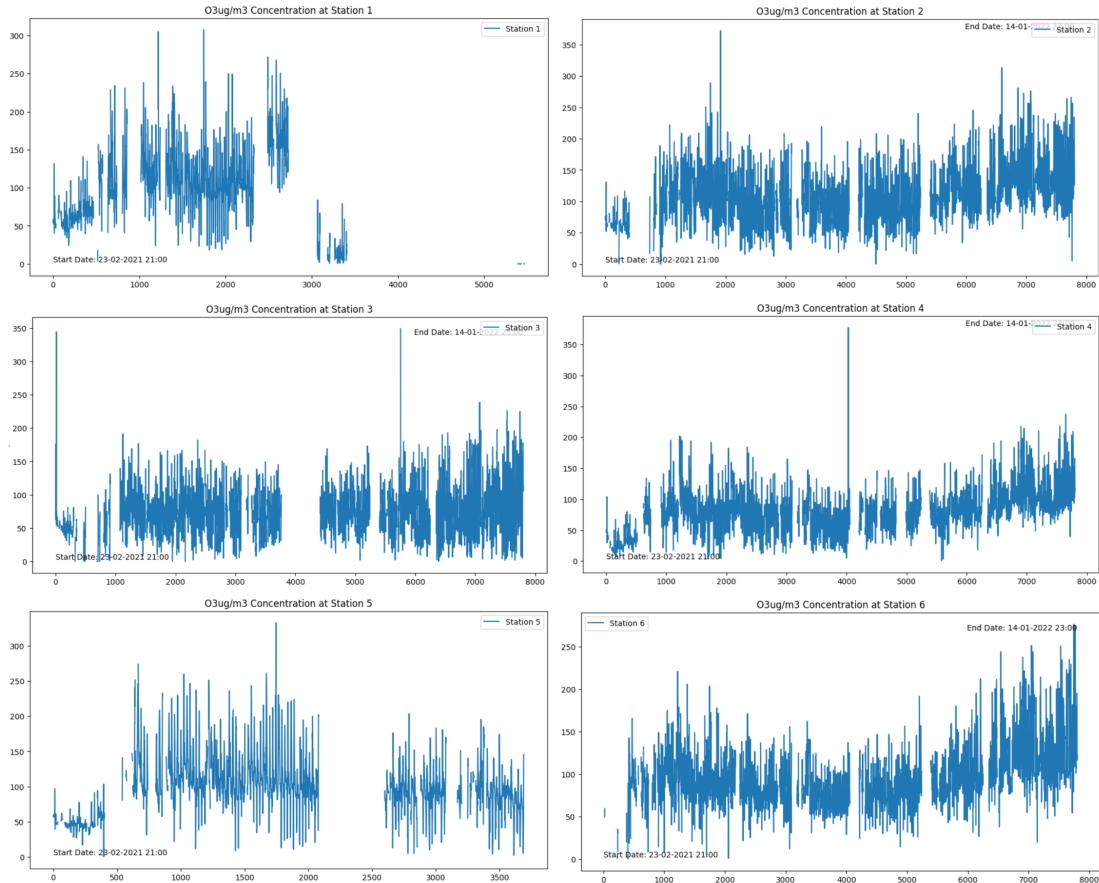
Ma trận tương quan cho thấy các chất có mối tương quan từ trung bình tới yếu. Dễ nhận thấy TSP (Total Suspended Particulates) và PM_{2.5} (Particulate Matter 2.5) có mối tương quan rất cao (gần chạm mức 1). TSP bao gồm tất cả các hạt vật chất lơ lửng trong không khí, có kích thước đa dạng, từ rất nhỏ đến rất lớn. PM_{2.5} bao gồm các hạt có đường kính khí động học nhỏ hơn hoặc bằng 2.5 micromet, là tập hợp con của TSP và ảnh hưởng trực tiếp đến chất lượng không khí. Vì vậy, trong luận văn này tôi sẽ loại bỏ dữ liệu TSP và chỉ sử dụng PM_{2.5}.

Các thông số khí tượng như nhiệt độ, độ ẩm,... và tia UV không phục vụ cho việc tính toán AQI. Do đó, nhằm giảm kích thước dữ liệu và số chiều dữ liệu, tôi chỉ sử dụng dữ liệu của bốn chất ô nhiễm là PM_{2.5}, O₃, NO₂ và CO.



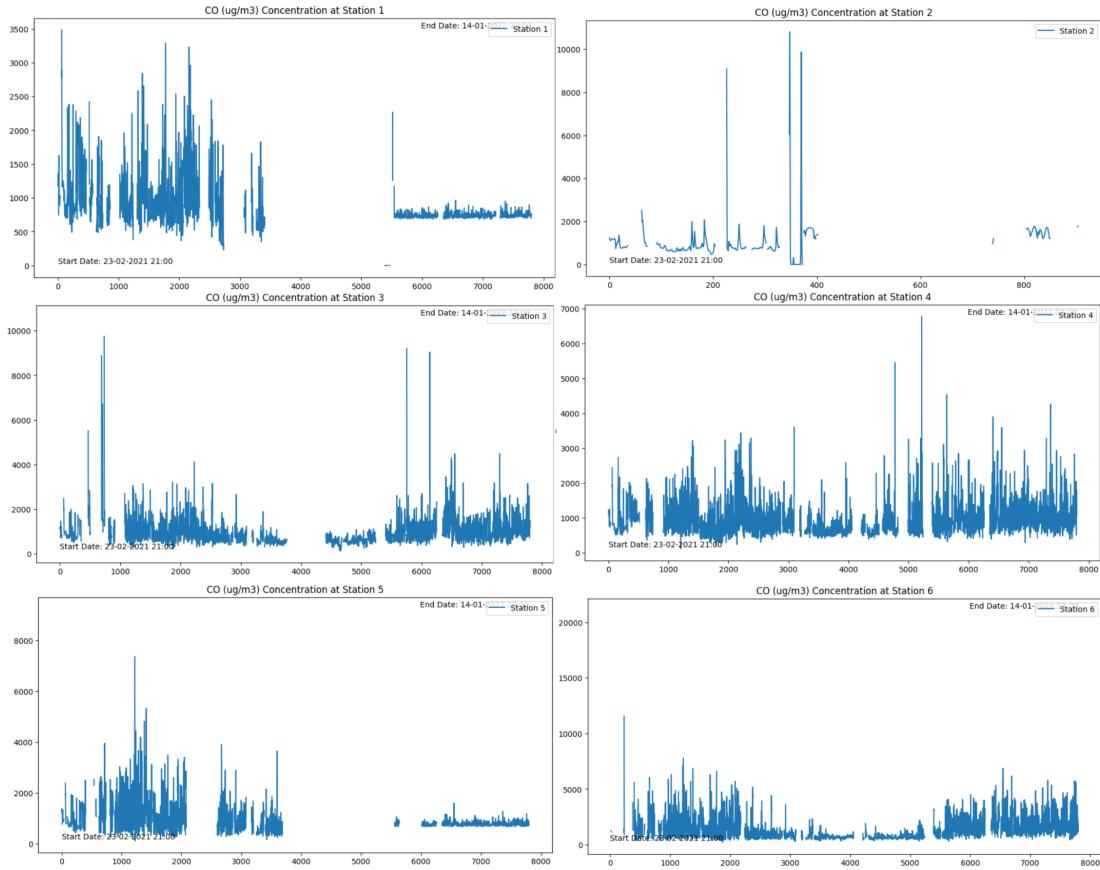
HÌNH 3.4: Biểu đồ nồng độ PM_{2.5} theo thời gian ở các trạm

Dữ liệu PM_{2.5} không cho thấy xu hướng tăng hoặc giảm rõ ràng trong suốt khoảng thời gian quan sát. Nồng độ PM_{2.5} dao động ở mức cao, với một số thời điểm tăng cao đột biến. Nồng độ PM_{2.5} thường cao hơn vào ban ngày và thấp hơn vào ban đêm. Điều này có thể liên quan đến hoạt động giao thông và các hoạt động sản xuất công nghiệp. Ngoài dữ liệu còn thể hiện sự biến động theo mùa, đặc biệt là vào những tháng cuối năm, nồng độ PM_{2.5} có xu hướng cao hơn so với đầu năm. Bên cạnh đó trạm 1 và trạm 5 có dữ liệu bị thiếu nhiều hơn so với các trạm khác và có sự xuất hiện của nhiều giá trị ngoại lai trong dữ liệu.



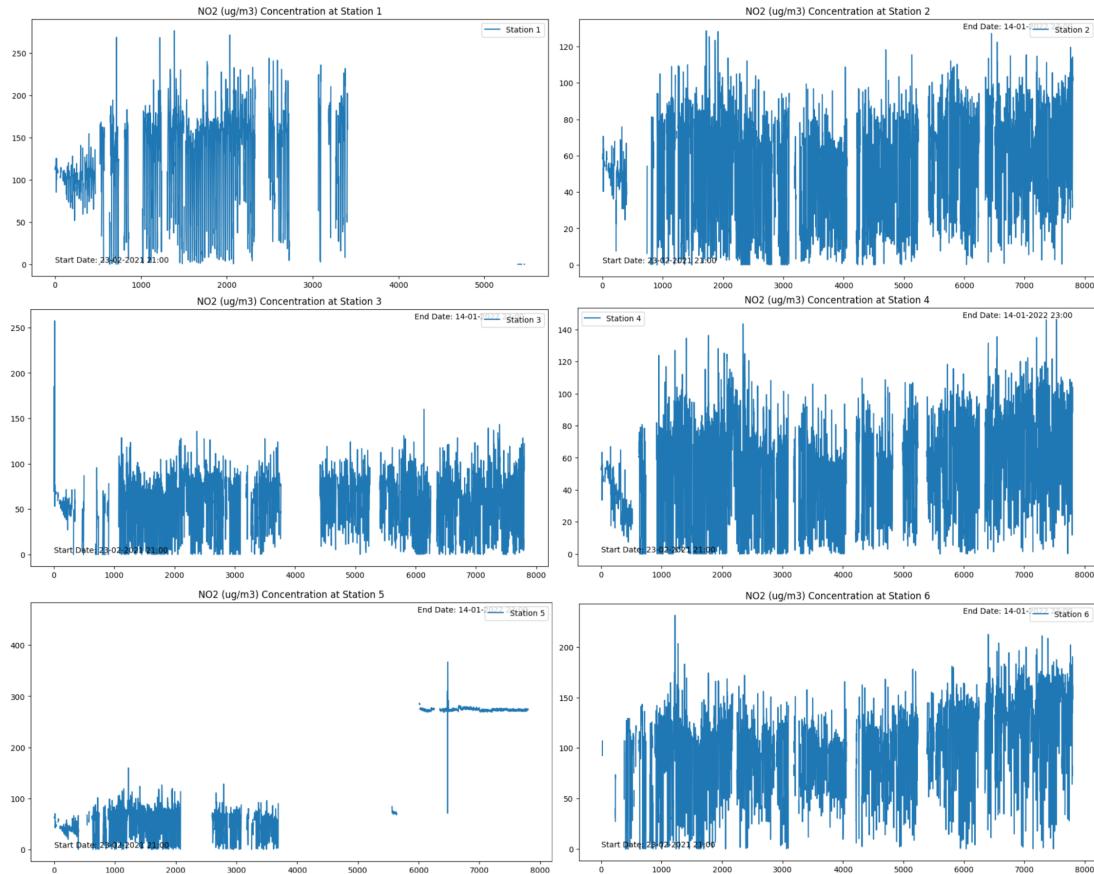
HÌNH 3.5: Biểu đồ nồng độ O₃ theo thời gian ở các trạm

Nhìn chung dữ liệu O₃ dao động quanh một mức trung bình. Có sự xuất hiện của các giá trị ngoại lai trong dữ liệu O₃ nhưng không nhiều như PM_{2.5}, số giá trị ngoại lai vượt biên trên của O₃ khá thấp. Về vấn đề khuyết dữ liệu, trạm 1 và trạm 5 vẫn dẫn đầu với tỉ lệ khuyết cực cao khoảng 40 đến 50 % tuy nhiên có sự khác biệt: trạm 1 bị khuyết trong một khoảng thời gian dài, bắt đầu từ giữa năm 2021 tới đầu năm 2022; trạm 5 bị khuyết dữ liệu ngẫu nhiên và theo chu kỳ ngắn hạn.



HÌNH 3.6: Biểu đồ nồng độ CO theo thời gian ở các trạm

Nồng độ của CO dao động rất lớn và nằm trong khoảng từ 750 đến $3500 \mu\text{g}/\text{m}^3$. Điều này cũng bình thường vì nguồn gốc của CO phát sinh tại khu vực chủ yếu đến từ việc đốt nguyên liệu hóa thạch và hoạt động giao thông trong khu vực. Tuy nhiên khi tính toán, AQI của CO rất thấp do đó ít ảnh hưởng đến kết quả tính toán AQI tổng thể. CO là chất rất độc và ảnh hưởng nghiêm trọng đến sức khỏe của con người, điều này có thể giải thích bởi nồng độ CO trong không khí thường thấp hơn nhiều so với các chất ô nhiễm khác như PM_{2.5} hay NO₂. Trong số bốn chất ô nhiễm, CO là chất có tỉ lệ khuyết cao nhất lên tới gần 60 %. Trạm 1 và trạm 5 bị khuyết ở khoảng thời gian giữa năm, trạm 2 thì bị mất đến 90 % dữ liệu.

HÌNH 3.7: Biểu đồ nồng độ NO₂ theo thời gian ở các trạm

Nồng độ của NO₂ nhìn chung dao động ở mức thấp đối với các trạm từ 1 đến 5, trạm 6 có mức độ dao động cao hơn. Có thể nói NO₂ là chất được đo tốt nhất trong tất cả các chất ô nhiễm, với tỉ lệ bị dữ liệu bị khuyết thấp và ít xuất hiện ngoại lai.

Tổng quan lại, tôi nhận thấy dữ liệu của trạm 1 và trạm 5 có tỷ lệ thiếu hụt dữ liệu rất cao, lên tới xấp xỉ 40% , trong khi các trạm còn lại chỉ từ 10-15% .Vậy nên để đảm bảo chất lượng dữ liệu và kết quả dự đoán của hệ thống, tôi quyết định không sử dụng dữ liệu của hai trạm quan trắc này. Bên cạnh đó, tôi quyết định loại bỏ các bản ghi có dữ liệu null liên tục trong hơn 24 giờ. Tuy giảm khoảng 10% kích thước dữ liệu nhưng việc loại bỏ dữ liệu null kéo dài mang lại những lợi ích sau:

- Đầu tiên là giúp tập trung vào các khoảng thời gian có dữ liệu đầy đủ và đáng tin cậy hơn, giảm thiểu nhiều và sai số trong quá trình phân tích.

- Các mô hình dự đoán thường hoạt động tốt hơn với dữ liệu liên tục và đầy đủ. Loại bỏ các chuỗi missing value dài có thể giúp mô hình học được các mẫu và xu hướng rõ ràng hơn, từ đó dự đoán chính xác hơn..
- Các chuỗi khuyết dữ liệu dài có thể đại diện cho các sự kiện bất thường hoặc các vấn đề kỹ thuật trong quá trình thu thập dữ liệu (cambio bảo trì định kì, gặp hư hỏng do thiên tai,...). Loại bỏ chúng giúp giảm thiểu sai lệch trong kết quả phân tích và dự đoán.

3.3 Phát hiện và hiệu chỉnh dữ liệu ngoại lai

Trong quá trình tiền xử lý dữ liệu, việc xử lý dữ liệu ngoại lai (outlier) là một bước quan trọng để đảm bảo tính chính xác và độ tin cậy của kết quả phân tích. Trong nghiên cứu này, tôi lựa chọn xử lý dữ liệu ngoại lai trước khi điền dữ liệu bị thiếu nhằm đảm bảo tính chính xác của dữ liệu trước khi điền giá trị thiếu, từ đó giảm thiểu sai lệch trong kết quả phân tích và dự đoán. Hơn nữa việc xử lý ngoại lai trước khi điền dữ liệu giúp tránh việc các giá trị ngoại lai bị "lạm truyền" trong quá trình nội suy. Nếu điền dữ liệu trước, các giá trị ngoại lai có thể ảnh hưởng đến việc ước tính giá trị thiếu, dẫn đến việc tạo ra thêm các giá trị ngoại lai mới.

Để phát hiện và hiệu chỉnh outlier, tôi sử dụng phương pháp kết hợp giữa bộ khoảng tứ phân vị (Interquartile Range - IQR). IQR là một đại lượng thống kê đo lường sự phân tán của dữ liệu bằng cách tính toán hiệu số giữa tứ phân vị thứ ba (Q_3) và tứ phân vị thứ nhất (Q_1). Q_1 đại diện cho giá trị tại vị trí 25% của dữ liệu đã được sắp xếp, trong khi Q_3 đại diện cho giá trị tại vị trí 75%. Công thức tính IQR là: $IQR = Q_3 - Q_1$

Để xác định ngưỡng trên và ngưỡng dưới cho việc phát hiện ngoại lai, tôi sử dụng IQR factor - là một hệ số nhân với IQR để xác định khoảng cách từ Q_1 và Q_3 đến ngưỡng dưới và ngưỡng trên. IQR là tổng của hai thành phần base factor và phần dynamic. Phần base factor cố định dựa trên độ phân tán của dữ liệu từng chất ô nhiễm.

3.3.1 Xác định độ phân tán và IQR factor

Phần base factor cố định dựa trên độ phân tán của dữ liệu từng chất ô nhiễm. Độ phân tán cho biết mức độ dàn trải của dữ liệu quanh giá trị trung tâm. Độ phân tán càng lớn, dữ liệu càng dàn trải, và khả năng xuất hiện ngoại lai càng cao. Độ phân tán của dữ liệu được xác định dựa trên độ lệch chuẩn, giá trị trung bình, và tỷ lệ độ lệch chuẩn/trung bình. Quan sát kết quả tính toán trong bảng cho thấy PM_{2.5} và CO có độ phân tán lớn, trong khi O₃ và NO₂ phân tán ở mức trung bình. Từ đó tôi lựa chọn base factor cụ thể cho từng chất ô nhiễm như sau: PM_{2.5}: 1.2, O₃: 2.0, CO: 1.0, NO₂: 2.0. Việc lựa chọn base factor khác nhau cho từng chất ô nhiễm cho phép tối điều chỉnh mức độ "nhạy cảm" của phương pháp IQR trong việc phát hiện ngoại lai, phù hợp với đặc điểm phân tán của từng chất.

BẢNG 3.4: Độ lệch chuẩn cho từng chất ô nhiễm tại mỗi trạm

Pollutant	Station	Std _ Dev	Mean	Std _ Dev/Mean
PM _{2.5}	2	10.900296	19.356782	0.563125
	3	17.979660	24.510773	0.733541
	4	15.365807	25.733665	0.597109
	6	13.718709	20.270895	0.676769
O ₃	2	38.633513	110.279721	0.350323
	3	32.306497	75.434685	0.428271
	4	31.237826	79.848554	0.391213
	6	31.106008	93.870055	0.331373
CO	2	968.363205	1050.801502	0.921547
	3	543.992430	909.330626	0.598234
	4	445.754274	931.463868	0.478552
	6	927.941350	1304.201516	0.711502
NO ₂	2	25.916244	58.403746	0.443743
	3	29.323248	59.695309	0.491215
	4	28.422140	56.761385	0.500730
	6	36.896324	105.140943	0.350923

Dynamic factor là một chỉ số động nhằm tăng tính linh hoạt trong việc phát hiện ngoại lai, được tính toán dựa trên tỷ lệ giữa độ lệch chuẩn và trung bình của từng

chất ô nhiễm trong vòng 24 giờ. Việc này nhằm tránh lọc đi mất xu hướng của dữ liệu cũng như bỏ sót outlier vào những ngày biến độ dao động không lớn. Dynamic factor sẽ có giá trị lớn hơn khi dữ liệu có biến động lớn, giúp nới rộng ngưỡng phát hiện ngoại lai.

3.3.2 Các bước hiệu chỉnh dữ liệu ngoại lai

Sau khi xác định được base factor và Dynamic factor, phương pháp IQR được áp dụng để hiệu chỉnh ngoại lai theo các bước sau:

Bước 1. Tính toán ngưỡng trên và ngưỡng dưới:

- Ngưỡng dưới = $Q_1 - (\text{base factor} + \text{Dynamic factor}) \cdot \text{IQR}$
- Ngưỡng trên = $Q_3 + (\text{base factor} + \text{Dynamic factor}) \cdot \text{IQR}$

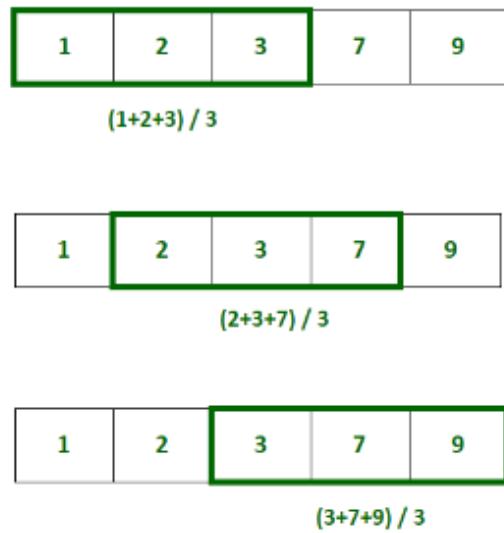
Bước 2. Xác định ngoại lai: Bất kỳ giá trị nào nhỏ hơn ngưỡng dưới hoặc lớn hơn ngưỡng trên đều được coi là ngoại lai.

Bước 3. Hiệu chỉnh ngoại lai: Thay thế các giá trị ngoại lai bằng giá trị trung vị của dữ liệu trong 24 giờ.

3.3.3 Làm mượt dữ liệu

MA là một kỹ thuật xử lý tín hiệu được sử dụng để làm mịn dữ liệu chuỗi thời gian bằng cách tính toán giá trị trung bình của một số điểm dữ liệu liền kề, giúp giảm nhiễu và các biến động ngẫu nhiên nhỏ. Mặc dù IQR đã hiệu chỉnh phần lớn dữ liệu ngoại lai, tuy nhiên vẫn còn một số lượng điểm dữ liệu chưa được xử lý. MA làm giảm ảnh hưởng của chúng và giúp cải thiện kết quả dự đoán của mô hình.

Để áp dụng bộ lọc MA, cần xác định cửa sổ trượt (window size), là số lượng điểm dữ liệu liền kề được sử dụng để tính trung bình. Cửa sổ trượt càng lớn, dữ liệu càng được làm mịn, nhưng cũng có thể làm mất đi các chi tiết nhỏ của dữ liệu.



HÌNH 3.8: Minh họa MA với window size = 3

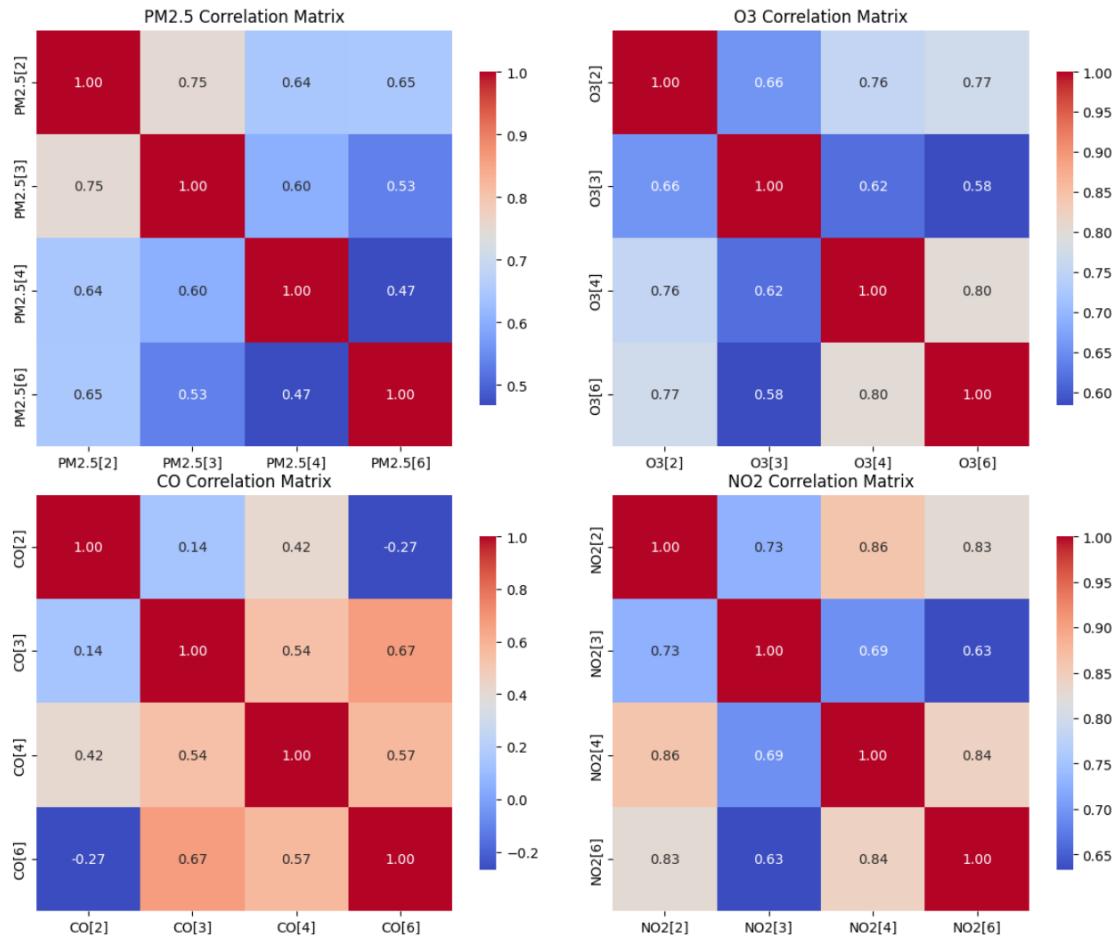
Trong nghiên cứu này, tôi chọn cửa sổ trượt bằng 3 nhằm lấy ba giờ liền kề để tính trung bình. Lựa chọn chọn nhằm tránh làm mất đi đặc trưng của các chỉ số ô nhiễm theo các buổi trong ngày.

3.4 Điền dữ liệu bị khuyết

3.4.1 Phân tích tính khả thi phương pháp nội suy không gian

Trong hệ thống quan trắc nhiều trạm, việc thu thập dữ liệu liên tục và đầy đủ từ tất cả các trạm là một thách thức do nhiều yếu tố khách quan như điều kiện thời tiết khắc nghiệt, sự cố thiết bị, hoặc các hoạt động bảo trì định kỳ. Các phương pháp nội suy không gian tận dụng thông tin từ các trạm lân cận để ước tính giá trị bị thiếu, dựa trên giả định rằng các trạm gần nhau về mặt không gian thường có tương quan cao hơn về các biến quan trắc. Khái niệm 'tương quan không gian' không chỉ đơn thuần là khoảng cách vật lý mà còn bao gồm các yếu tố địa hình, đặc điểm môi trường, hướng gió và các yếu tố vi khí hậu khác. Một số phương pháp nội suy không gian phổ biến bao gồm IDW, nội suy Kriging và Natural Neighbor. Mỗi phương pháp có những ưu nhược điểm riêng và phù hợp với từng loại dữ liệu và đặc điểm phân bố của các trạm quan trắc. Tuy nhiên, điều kiện tiên quyết để áp dụng hiệu quả các phương pháp nội

suy không gian, bên cạnh việc dữ liệu giữa các trạm có tương quan lớn, còn đòi hỏi sự tương quan chặt chẽ về mặt không gian giữa các trạm quan trắc. Ví dụ, các trạm quan trắc nằm gần nhau thường thể hiện mối tương quan dữ liệu cao hơn so với các trạm cách xa nhau, do chịu ảnh hưởng của các yếu tố môi trường tương đồng hơn.



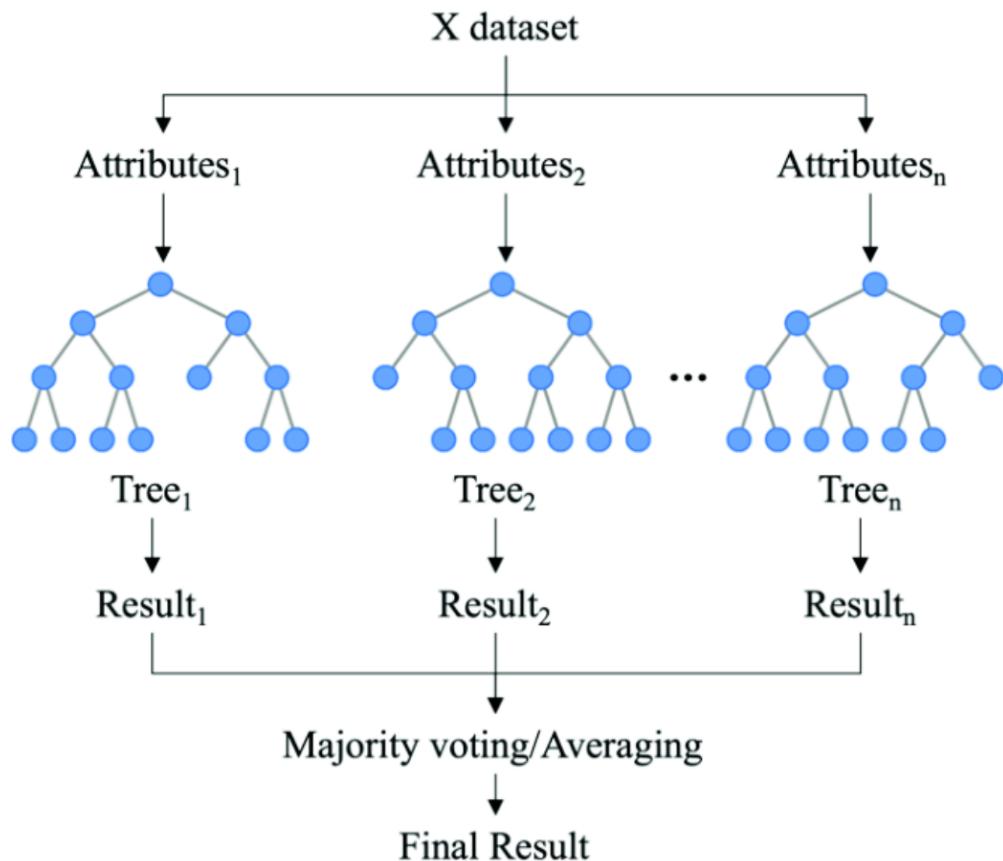
HÌNH 3.9: Ma trận tương quan của các chất ô nhiễm giữa các trạm

Sau khi phân tích ma trận tương quan giữa các trạm 2, 3, 4 và 6, kết quả cho thấy mức độ tương quan giữa PM_{2.5} và CO là tương đối thấp, trong khi O₃ và NO₂ có tương quan khả quan hơn. Tuy nhiên, phân tích tương quan đơn thuần chưa thể hiện đầy đủ mối liên hệ giữa khoảng cách vật lý giữa các trạm và mức độ tương quan dữ liệu. Dựa trên tọa độ các trạm ở bảng 3.3, có thể thấy khoảng cách từ trạm 6 đến trạm 3 ngắn hơn so với khoảng cách đến trạm 4. Mặc dù vậy, dữ liệu quan trắc tại trạm 6 lại thể hiện tương quan chặt chẽ hơn với trạm 4. Điều này cho thấy khoảng cách địa lý không phải là yếu tố quyết định duy nhất đến sự tương quan của dữ liệu, mà còn bị ảnh hưởng bởi các yếu tố khác như đặc điểm khí tượng, địa hình khu vực và các nguồn phát thải cục bộ. Do đó, việc áp dụng các phương pháp nội suy không gian chỉ dựa trên

khoảng cách, ví dụ như IDW, có thể không hiệu quả trong bài toán này. Thật vậy, tôi đã thử thực nghiệm sử dụng IDW và kết quả cho thấy dữ liệu được điền có phân bố không đồng đều và làm mất đi các đặc tính biến đổi của dữ liệu theo thời gian ngắn. Cụ thể, IDW ước tính giá trị tại một vị trí chưa biết bằng cách tính trung bình trọng số của các giá trị tại các trạm lân cận. Trọng số của mỗi trạm được tính tỷ lệ nghịch với một lũy thừa của khoảng cách từ trạm đó đến vị trí cần nội suy. Ví dụ, nếu khoảng cách tăng gấp đôi, trọng số sẽ giảm đi bốn lần (với lũy thừa mặc định là 2). Do đó, các trạm càng gần vị trí cần nội suy sẽ có ảnh hưởng càng lớn đến giá trị được ước tính. Trong trường hợp dữ liệu quan trắc của bài toán này, việc áp dụng IDW đã dẫn đến hiện tượng dữ liệu được "kéo" về giá trị của các trạm gần nhất một cách quá mức, làm mất đi sự biến thiên tự nhiên của dữ liệu theo thời gian ngắn và tạo ra phân bố không đồng đều giữa các khoảng thời gian.

3.4.2 Thủ nghiệm với Random Forest

Để khắc phục những hạn chế của các phương pháp nội suy dựa trên khoảng cách, nghiên cứu này đã áp dụng mô hình Random Forest Regressor để điền dữ liệu khuyết. Phương pháp này tận dụng thông tin từ các biến quan trắc khác nhau tại mỗi trạm để dự đoán giá trị bị thiếu, cho phép nắm bắt các mối quan hệ phức tạp và phi tuyến giữa các biến. Khác với IDW chỉ dựa vào khoảng cách, Random Forest xem xét đồng thời ảnh hưởng của nhiều yếu tố, phản ánh chính xác hơn sự biến thiên của dữ liệu trong không gian và thời gian. Trong các bài toán với số lượng lớn trạm quan trắc phân bố rộng khắp một khu vực, việc kết hợp Random Forest với thuật toán KNN có thể mang lại hiệu quả cao hơn[9]. KNN được sử dụng để xác định k trạm lân cận có dữ liệu tương quan cao nhất với trạm cần điền dữ liệu khuyết, từ đó tập trung việc huấn luyện Random Forest vào những trạm có ảnh hưởng lớn nhất. Điều này giúp giảm độ phức tạp tính toán và có thể cải thiện độ chính xác của mô hình. Tuy nhiên, trong bài toán cụ thể của nghiên cứu này, số lượng trạm quan trắc chỉ là bốn và chúng tập trung chủ yếu ở khu vực trung tâm thành phố. Do đó, việc áp dụng KNN để lựa chọn trạm lân cận là chưa thực sự cần thiết, và Random Forest được áp dụng trực tiếp trên dữ liệu từ tất cả các trạm còn lại.



HÌNH 3.10: Cách hoạt động của Random Forest

Cụ thể, quá trình triển khai Random Forest được thực hiện như sau: Với mỗi biến quan trắc có dữ liệu bị thiếu, một mô hình Random Forest riêng biệt được huấn luyện. Dữ liệu từ các biến quan trắc khác tại các trạm được sử dụng làm biến đầu vào, và giá trị của biến mục tiêu tại các thời điểm có dữ liệu được sử dụng làm biến mục tiêu. Các tham số của mô hình Random Forest, bao gồm số lượng cây, độ sâu tối đa của cây, số mẫu tối thiểu để phân chia một nút, và số mẫu tối thiểu ở một nút lá, đã được lựa chọn thông qua quá trình thử nghiệm và đánh giá để tối ưu hóa hiệu suất mô hình, cụ thể như sau:

- `n_estimators = 300`: Tham số này quy định số lượng cây quyết định được xây dựng trong rừng. Việc sử dụng 300 cây cho thấy một nỗ lực cân bằng giữa hiệu suất và thời gian tính toán. Số lượng cây càng nhiều thường giúp mô hình mạnh mẽ hơn và ít bị overfitting hơn, nhưng cũng tốn nhiều thời gian tính toán hơn.
- `max_depth = 10`: Tham số này giới hạn độ sâu tối đa của mỗi cây. Giới hạn độ sâu giúp ngăn ngừa overfitting, đặc biệt khi dữ liệu huấn luyện có nhiều nhiễu.

Giá trị 10 cho phép cây có độ phức tạp vừa phải để nắm bắt các mối quan hệ quan trọng trong dữ liệu mà không bị quá "học thuộc" dữ liệu huấn luyện.

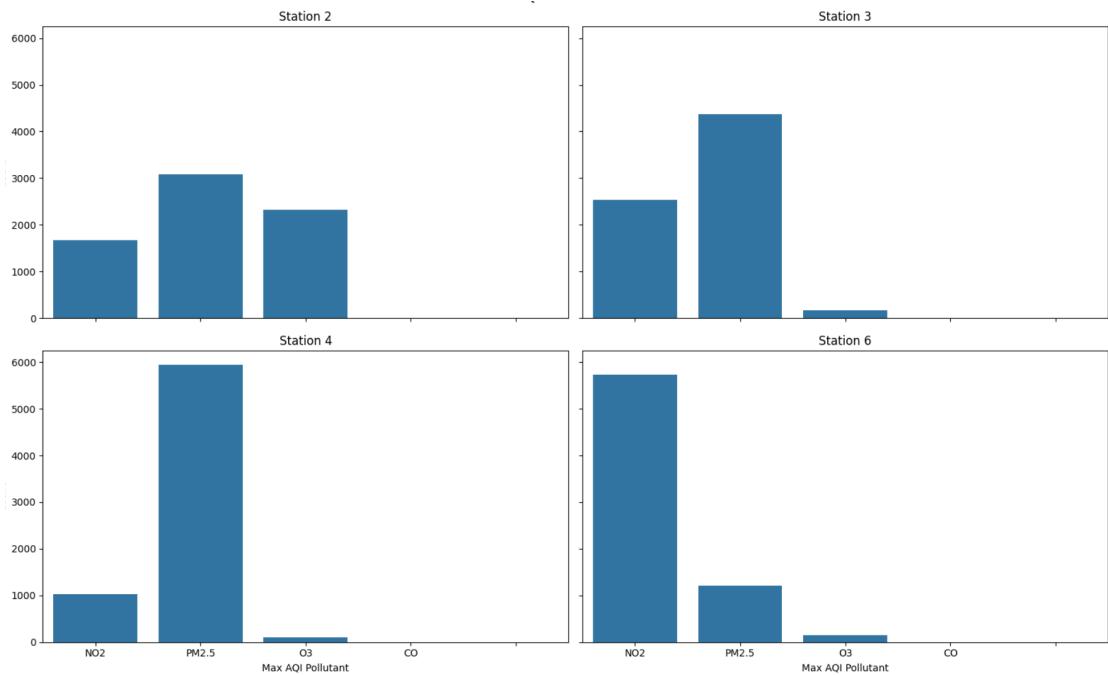
- `min_samples_split = 7`: Tham số này quy định số mẫu tối thiểu cần có trong một nút để nút đó có thể được phân chia tiếp. Giá trị 7 cho thấy một sự cân nhắc rằng mỗi nút cần có đủ dữ liệu để việc phân chia là có ý nghĩa thống kê. Nếu giá trị này quá nhỏ, mô hình có thể tạo ra các nút lá chỉ chứa một vài mẫu, dẫn đến overfitting.
- `min_samples_leaf = 3`: Tham số này quy định số mẫu tối thiểu cần có ở một nút lá. Tương tự như `min_samples_split`, tham số này cũng giúp ngăn ngừa overfitting bằng cách đảm bảo rằng mỗi nút lá đại diện cho một số lượng mẫu đủ lớn. Giá trị 3 cho thấy mỗi nút lá cần ít nhất 3 mẫu để được coi là hợp lệ.

3.5 Mô hình dự đoán chất lượng không khí

3.5.1 Tiền xử lý và chuẩn bị dữ liệu huấn luyện

Dữ liệu lúc này đã được hiệu chỉnh và không còn khoảng khuyết, sẵn sàng cho việc tính toán Chỉ số Chất lượng Không khí. AQI được tính toán dựa trên nồng độ của từng chất ô nhiễm theo công thức và các ngưỡng giá trị theo Quyết định số 1459/QĐ-TTgMT ngày 12/11/2019 của Tổng Cục Môi trường. Dữ liệu lúc này sẽ được gom lại theo trạm, thay vì theo chất ô nhiễm và áp dụng công thức tính AQI cho từng chất ô nhiễm. AQI tổng thể cho mỗi trạm được xác định bằng cách lấy giá trị lớn nhất trong số các AQI của từng chất ô nhiễm tại trạm đó.

Để xác định chất ô nhiễm nào gây ra mức AQI cao nhất tại mỗi thời điểm, một đặc trưng mới ược tạo ra cho mỗi trạm. Cột này lưu trữ số hiệu tương ứng với chất ô nhiễm có AQI lớn nhất. Việc xác định chất ô nhiễm chính được thực hiện bằng cách sử dụng hàm `idxmax` để tìm cột có giá trị lớn nhất trong số các cột AQI của từng chất ô nhiễm, sau đó sử dụng ánh xạ để chuyển đổi tên cột thành số hiệu tương ứng.



HÌNH 3.11: Biểu đồ phân bố chất ô nhiễm chính của các trạm

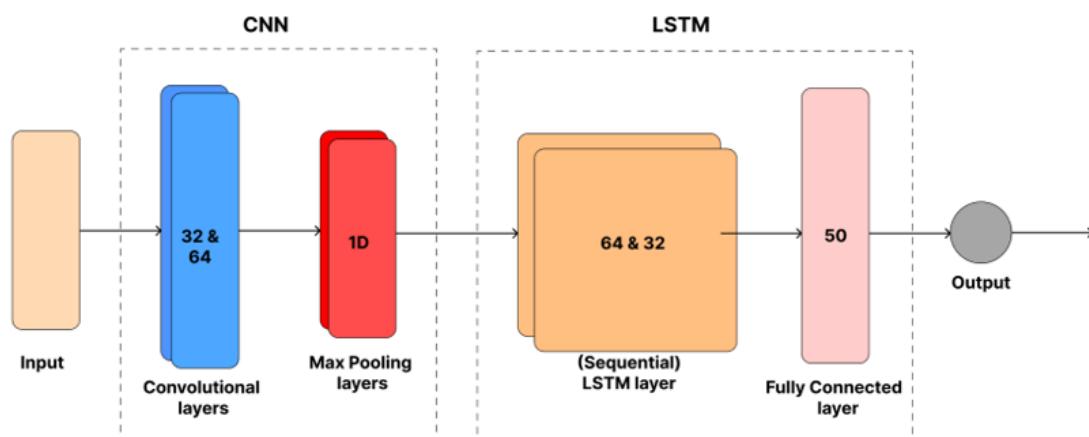
Dựa trên phân tích biểu đồ phân bố các chất ô nhiễm chính tại từng trạm theo thời gian. PM_{2.5} là chất ô nhiễm chiếm ưu thế tại hầu hết các trạm, đặc biệt là tại trạm 3 và 4, cho thấy đây là nguồn ô nhiễm đáng quan tâm nhất. Ngược lại, nồng độ CO ghi nhận không có thời điểm nào là chất ô nhiễm chính ở tất cả các trạm, cho thấy CO hầu như không ảnh hưởng đến chỉ số AQI của trạm. NO₂ là chất ô nhiễm đứng sau PM_{2.5}, đặc biệt ở trạm 6 khi là chất ô nhiễm chính trong hầu hết khoảng thời gian. O₃ nhìn chung ít ảnh hưởng đến chất lượng không khí, khi thực sự ảnh hưởng ở trạm 2. Nhìn chung với đặc trưng mới này được sử dụng để dự đoán chất ô nhiễm chiếm ưu thế cùng với chỉ số AQI cho từng trạm.

3.5.2 Xây dựng mô hình lai LSTM - CNN

Mạng nơ-ron hồi quy dài-ngắn hạn (LSTM) đã chứng minh được tính hiệu quả vượt trội trong việc xử lý và dự báo dữ liệu chuỗi thời gian, nhờ khả năng ghi nhớ thông tin từ các bước thời gian trước đó và nắm bắt các phụ thuộc dài hạn trong dữ liệu. Điều này khiến LSTM trở thành lựa chọn phổ biến trong nhiều bài toán như dự báo giá chứng khoán [14], nhận dạng giọng nói [15], và xử lý ngôn ngữ tự nhiên [16]. Bên cạnh đó, mạng nơ-ron tích chập (CNN) nổi tiếng với khả năng trích xuất các đặc trưng không gian hiệu quả từ dữ liệu, ban đầu chủ yếu được áp dụng trong xử lý ảnh [17].

Tuy nhiên, gần đây, CNN cũng được chứng minh là hữu ích trong việc trích xuất các đặc trưng cục bộ từ dữ liệu chuỗi thời gian, ví dụ như tìm kiếm các mẫu lặp lại hoặc các đoạn xu hướng ngắn [18]. Việc kết hợp LSTM và CNN trong một mô hình lai cho phép tận dụng điểm mạnh của cả hai: CNN trích xuất đặc trưng quan trọng từ dữ liệu đầu vào, sau đó LSTM xử lý chuỗi đặc trưng này để nắm bắt các phụ thuộc thời gian và đưa ra dự đoán. Trong bài toán dự đoán AQI cho các trạm quan trắc, việc kết hợp này đặc biệt hữu ích vì dữ liệu AQI thể hiện cả tính chất chuỗi thời gian và các đặc trưng cục bộ như sự tăng đột biến nồng độ ô nhiễm do một sự kiện cụ thể. So với các phương pháp truyền thống như ARIMA hay các mô hình học máy cổ điển, mô hình lai LSTM-CNN có khả năng nắm bắt các mối quan hệ phức tạp và phi tuyến tính trong dữ liệu tốt hơn, từ đó mang lại độ chính xác dự báo cao hơn [19].

Trong nghiên cứu này, mô hình lai LSTM-CNN được triển khai để dự báo AQI cho từng trạm quan trắc dựa trên dữ liệu lịch sử về nồng độ các chất ô nhiễm. Cụ thể, dữ liệu đầu vào là chỉ số AQI của bốn chất ô nhiễm PM_{2.5}, O₃, CO, và NO₂ của mỗi trạm. Mô hình được xây dựng với kiến trúc như sau: lớp đầu vào được định hình để phù hợp với dữ liệu chuỗi thời gian, tiếp theo là hai lớp tích chập 1D với kích thước kernel là 3 để trích xuất các đặc trưng cục bộ. Sau đó, hai lớp MaxPooling 1D được sử dụng để giảm chiều dữ liệu. Phần LSTM của mô hình bao gồm hai lớp LSTM với 50 và 25 đơn vị, kết hợp với các lớp Dropout với tỷ lệ 0.2 để ngăn chặn hiện tượng quá khớp. Cuối cùng, các lớp Dense (Fully connected layer) được sử dụng để đưa ra dự đoán cho AQI trạm và chất ô nhiễm chính. Số lượng đơn vị của lớp Dense cuối cùng được thiết lập bằng số giờ dự báo nhân với số lượng biến mục tiêu (trong trường hợp này là hai, AQI trạm và chất ô nhiễm chính).



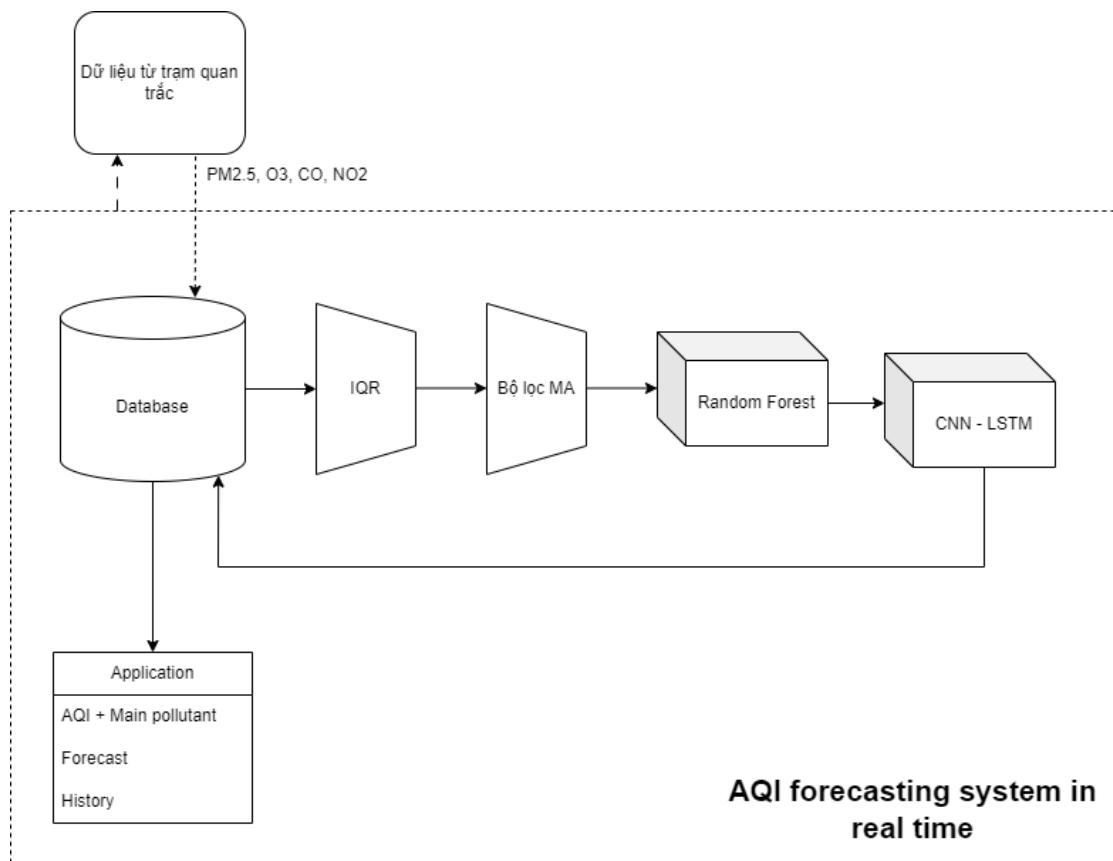
HÌNH 3.12: Mô hình lai CNN-LSTM áp dụng vào hệ thống

Mô hình được huấn luyện với thuật toán tối ưu Adam và hàm mất mát trung bình phương sai (MSE), với độ đo MAE được sử dụng để đánh giá hiệu suất. Việc lựa chọn các tham số như số lượng lớp, số lượng đơn vị trong mỗi lớp, kích thước kernel, và tỷ lệ dropout được thực hiện dựa trên kinh nghiệm và thử nghiệm để đạt được hiệu suất tốt nhất trên dữ liệu. Cụ thể, giá trị look_được đặt là 24, nghĩa là mô hình sử dụng 24 giờ dữ liệu quá khứ để dự đoán. Số giờ dự báo (forecast _) được thiết lập là 48. Mô hình được huấn luyện trong 50 epochs với batch size là 32. Các giá trị này có thể được điều chỉnh để tối ưu hóa hiệu suất cho từng bài toán cụ thể.

3.6 Triển khai hệ thống theo thời gian thực

3.6.1 Thiết kế hệ thống

Trong nghiên cứu này, khôi xử lý dữ liệu bị thiếu và khôi dự đoán là trái tim và lá phổi của hệ thống, bên cạnh đó là những khối phụ có chức năng thu thập, xử lý, lưu trữ dữ liệu và quản trị hệ thống.



HÌNH 3.13: Sơ đồ khái hệ thống

3.6.2 Chi tiết về các khối

1. Khối thu thập dữ liệu: Thu thập dữ liệu đầu vào của hệ thống được thu thập từ các trạm quan trắc chất lượng không khí. Dữ liệu này bao gồm thông tin về thời gian theo từng giờ và nồng độ của các chất ô nhiễm chính trong không khí, bao gồm PM_{2.5}, O₃, CO, và NO₂. Dữ liệu thô này có thể được thu thập dưới nhiều định dạng khác nhau, ví dụ như tệp CSV (Comma Separated Values), JSON (JavaScript Object Notation), hoặc thông qua các giao thức truyền dữ liệu trực tiếp từ hệ thống giám sát của các trạm quan trắc. Trong nghiên cứu này vì không tiếp cận được với dữ liệu theo thời gian thực của các trạm nên tôi quyết định sử dụng dữ liệu giả - là khoảng thời gian cuối cùng được trích trong tập dữ liệu ban đầu.

2. Khối lưu trữ: Là nơi lưu trữ dữ liệu của toàn bộ hệ thống, bao gồm dữ liệu thô, dữ liệu đã qua hiệu chỉnh và kết quả dự đoán AQI.

3. Khối tiền xử lý dữ liệu gồm các thành phần chính sau:

- Khối IQR: Dữ liệu thô từ các trạm quan trắc thu thập theo thời gian thực đôi lúc chứa những điểm ngoại lai. Khối IQR xác định và hiệu chỉnh những điểm này.
- Bộ lọc MA: Phương pháp trung bình trượt được sử dụng để làm mượt dữ liệu.
- Khối Random Forest: Dữ liệu sau khi đã được xử lý ngoại lai và làm mượt, khối Random Forest được sử dụng để ước tính và điền các giá trị bị thiếu.
- Khối tiền Xử Lý Dữ Liệu Trước Huấn Luyện: Dựa dữ liệu sạch về dạng AQI theo từng chất ô nhiễm và kết hợp với dữ liệu trong quá khứ để thành các chuỗi thời gian với độ dài look back 24 giờ để làm đầu vào cho mô hình LSTM.

4. Khối CNN-LSTM: Mô hình đã được huấn luyện trên tập huấn luyện trước đó và bắt đầu quá trình dự đoán trên tập dữ liệu mới bao gồm dữ liệu vừa nhận được từ khối tiền xử lý dữ liệu. Kết quả là chỉ số AQI và chất ô nhiễm chính của các trạm trong những giờ tiếp theo.

5. Khối ứng dụng: Là một ứng dụng web bao gồm backend và frontend hoàn chỉnh nhằm đưa kết quả dự đoán đến người dùng.

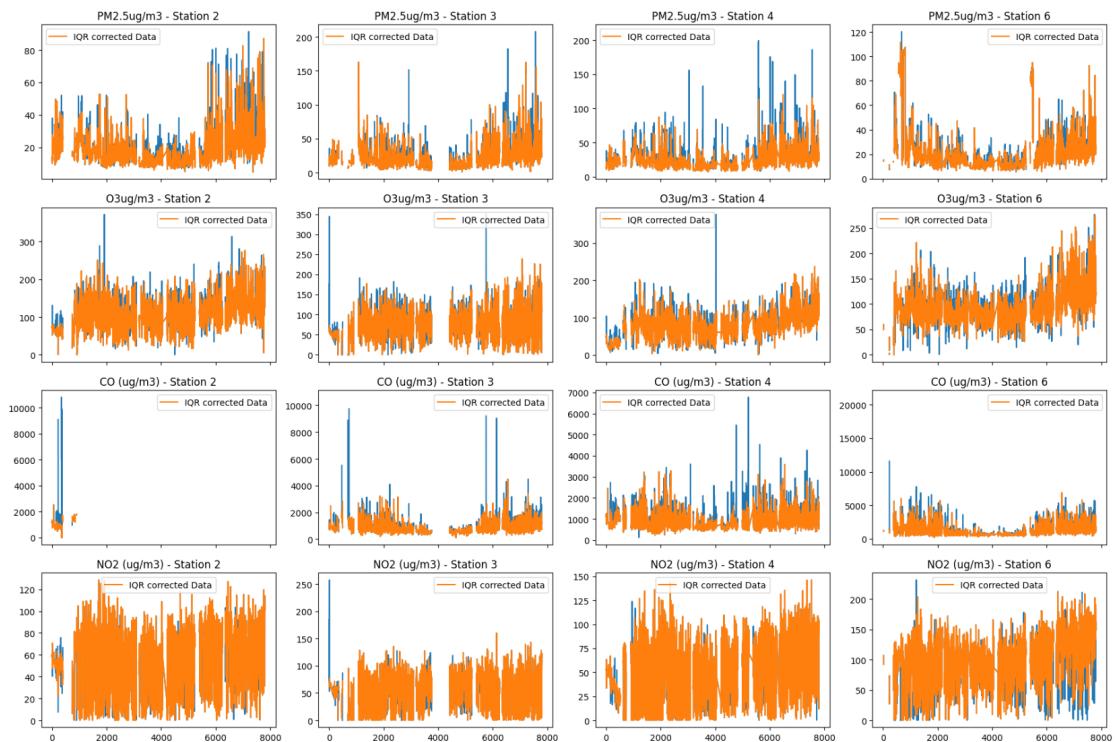
Hệ thống được xây dựng trên nền tảng Amazon Web Service, một nền tảng điện toán đám mây phát triển toàn diện được cung cấp bởi Amazon. Trong dự án này tôi sẽ thực thi toàn bộ các tác vụ từ thu thập dữ liệu, xử lý dữ liệu, lưu trữ, huấn luyện mô hình qua các service của AWS.

Chương 4

Kết quả và nhận xét

4.1 Quá trình phát hiện và hiệu chỉnh dữ liệu ngoại lai

Xử lý dữ liệu ngoại lai là bước tiên quyết để diệt dữ liệu khuyết hiệu quả và tạo ra các mô hình dự báo hiệu quả. Dưới đây là dữ liệu thời gian của các chất ô nhiễm sau khi hiệu chỉnh bằng phương pháp IQR.



HINH 4.1: Dữ liệu theo thời gian của các chất ô nhiễm đã được hiệu chỉnh outlier

BẢNG 4.1: Thống kê số lượng điểm dữ liệu ngoại lai được hiệu chỉnh

Chất ô nhiễm	Trạm 2	Trạm 3	Trạm 4	Trạm 6
PM _{2.5}	290	230	328	301
O ₃	204	170	257	321
CO	30	250	290	247
NO ₂	115	97	85	274

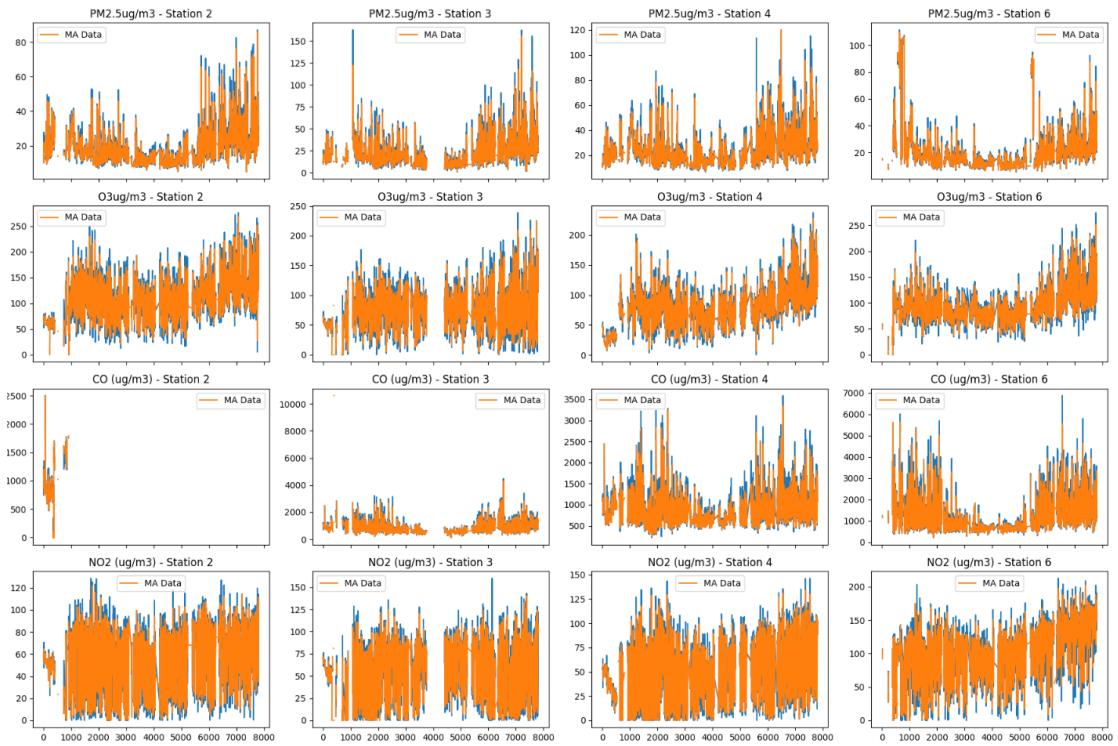
Dựa vào số lượng điểm ngoại lai được phát hiện, tổng quan thì số điểm ngoại lai tìm được chiếm khoảng từ 2% đến 4.057% tổng số dữ liệu. Đây là một con số tốt khi so sánh với đề kết quả trong bài báo "Dự báo chất lượng không khí bằng mô hình LSTM-MA trường hợp sử dụng dữ liệu tại trạm quan trắc tự động ngã tư Giồng Nước, tỉnh Bà Rịa - Vũng Tàu" [5]. Cùng là phát hiện ngoại lai của tập dữ liệu quan trắc không khí, tác giả sử dụng phương pháp Box-Whisker và cho ra kết quả 3,2%, có thể thấy kết quả trên phản ứng đúng với phân tích ban đầu với số điểm ngoại lai tập trung nhiều ở PM_{2.5} và CO (ngoại trừ trạm 2 vì dữ liệu quá ít) và ít nhất ở NO₂ (trạm 6 là ngoại lệ vì có nhiều điểm dữ liệu cận zero). Như phân tích ở chương 3, PM_{2.5} và CO có độ tương quan giữa các trạm thấp, dữ liệu phân tán cao và khoảng dao động rất lớn đối với những dữ liệu vượt biên trên. O₃ và NO₂ thể hiện sự ổn định qua khoảng dao động nhỏ hơn, có độ phân tán thấp hơn và sự tương quan giữa các trạm cũng cao hơn.

BẢNG 4.2: Bảng so sánh khoảng dao động dữ liệu trung bình 24h trước và sau IQR

Chất ô nhiễm	Trạm 2	Trạm 3	Trạm 4	Trạm 6
PM _{2.5} ($\mu\text{g}/\text{m}^3$)	Original: 23.25 Corrected: 19.09	Original: 42.62 Corrected: 34.89	Original: 35.21 Corrected: 25.6	Original: 20.18 Corrected: 16.82
O ₃ ($\mu\text{g}/\text{m}^3$)	Original: 115.86 Corrected: 102.14	Original: 104.84 Corrected: 94.33	Original: 80.39 Corrected: 67.71	Original: 83.77 Corrected: 67.97
CO ($\mu\text{g}/\text{m}^3$)	Original: 1047.39 Corrected: 450.95	Original: 1216.83 Corrected: 986.26	Original: 1252.28 Corrected: 963.31	Original: 2148.42 Corrected: 1722.23
NO ₂ ($\mu\text{g}/\text{m}^3$)	Original: 77.18 Corrected: 75.01	Original: 87.01 Corrected: 83.6	Original: 81.42 Corrected: 79.37	Original: 112.91 Corrected: 95.52

Ngoài ra IQR còn làm tốt trong việc hiệu chỉnh những dữ liệu nhiễu vượt cực đại bằng cách thay chúng bằng giá trị trung giúp giảm khoảng dao động của dữ liệu trong 24 giờ. Những chất phân tán rộng hơn như PMO_{2.5} hay CO thì khoảng dao động

giảm nhiều hơn so với O_3 hay NO_2 . Điều này làm cải thiện kết quả dự đoán AQI, vì công thức tính AQI là lấy giá trị AQI lớn nhất của các chất ô nhiễm dẫn đến những điểm nhiễu cực đại này ảnh hưởng rất xấu đến mô hình dự đoán.



HÌNH 4.2: Dữ liệu theo thời gian của các chất ô nhiễm qua MA

BẢNG 4.3: Bảng so sánh khoảng dao động dữ liệu trung bình 24h trước và sau MA.

Chất ô nhiễm	Trạm 2	Trạm 3	Trạm 4	Trạm 6
PM2.5 ($\mu g/m^3$)	Before: 19.09 After MA: 15.4	Before: 34.89 After MA: 26.94	Before: 25.6 After MA: 20.51	Before: 16.82 After MA: 13.75
O₃ ($\mu g/m^3$)	Before: 101.95 After MA: 80.28	Before: 94.09 After MA: 74.66	Before: 67.71 After MA: 52.69	Before: 67.96 After MA: 51.63
CO ($\mu g/m^3$)	Before: 442.75 After MA: 339.53	Before: 986.26 After MA: 771.45	Before: 963.31 After MA: 759.16	Before: 1722.23 After MA: 1347.52
NO₂ ($\mu g/m^3$)	Before: 74.36 After MA: 63.22	Before: 82.59 After MA: 71.14	Before: 77.4 After MA: 68.03	Before: 93.98 After MA: 74.21

Đối với số lượng dữ liệu ngoại lai còn lại chưa được xử, bộ lọc MA giúp làm giảm ảnh hưởng của những điểm dữ liệu này. Dựa vào kết quả trên có thể thấy dữ liệu đã mượt mà hơn. Khoảng dao động của dữ liệu các chất ô nhiễm cũng giảm đều và đáng kể so với trước khi qua bộ lọc MA.

BẢNG 4.4: Bảng so sánh độ lệch chuẩn và trung bình sau hiệu chỉnh dữ liệu.

Chất ô nhiễm	Trạm 2	Trạm 3	Trạm 4	Trạm 6
PM _{2.5}	Org Std_Dev: 10.9 Org Mean: 19.36 Org Std_Dev/Mean: 0.56 Std_Dev: 9.56 Mean: 18.73 Std_Dev/Mean: 0.51	Org Std_Dev: 17.98 Org Mean: 24.51 Org Std_Dev/Mean: 0.73 Std_Dev: 15.22 Mean: 23.27 Std_Dev/Mean: 0.65	Org Std_Dev: 15.37 Org Mean: 25.73 Org Std_Dev/Mean: 0.6 Std_Dev: 12.0 Mean: 24.55 Std_Dev/Mean: 0.49	Org Std_Dev: 13.72 Org Mean: 20.27 Org Std_Dev/Mean: 0.68 Std_Dev: 13.4 Mean: 19.87 Std_Dev/Mean: 0.67
	Org Std_Dev: 38.55 Org Mean: 110.35 Org Std_Dev/Mean: 0.35 Std_Dev: 34.32 Mean: 109.03 Std_Dev/Mean: 0.31	Org Std_Dev: 32.14 Org Mean: 75.61 Org Std_Dev/Mean: 0.43 Std_Dev: 27.61 Mean: 74.77 Std_Dev/Mean: 0.37	Org Std_Dev: 31.24 Org Mean: 79.85 Org Std_Dev/Mean: 0.39 Std_Dev: 28.0 Mean: 79.23 Std_Dev/Mean: 0.35	Org Std_Dev: 31.09 Org Mean: 93.89 Org Std_Dev/Mean: 0.33 Std_Dev: 27.26 Mean: 92.64 Std_Dev/Mean: 0.29
	Org Std_Dev: 961.71 Org Mean: 1109.7 Org Std_Dev/Mean: 0.87 Std_Dev: 690.84 Mean: 1053.02 Std_Dev/Mean: 0.66	Org Std_Dev: 543.99 Org Mean: 909.33 Org Std_Dev/Mean: 0.6 Std_Dev: 439.71 Mean: 870.53 Std_Dev/Mean: 0.51	Org Std_Dev: 445.75 Org Mean: 931.46 Org Std_Dev/Mean: 0.48 Std_Dev: 345.49 Mean: 891.06 Std_Dev/Mean: 0.39	Org Std_Dev: 927.94 Org Mean: 1304.2 Org Std_Dev/Mean: 0.71 Std_Dev: 713.71 Mean: 1235.57 Std_Dev/Mean: 0.58
	Org Std_Dev: 25.66 Org Mean: 58.68 Org Std_Dev/Mean: 0.44 Std_Dev: 23.51 Mean: 59.17 Std_Dev/Mean: 0.4	Org Std_Dev: 28.65 Org Mean: 60.54 Org Std_Dev/Mean: 0.47 Std_Dev: 26.13 Mean: 60.55 Std_Dev/Mean: 0.43	Org Std_Dev: 27.79 Org Mean: 57.59 Org Std_Dev/Mean: 0.48 Std_Dev: 25.84 Mean: 57.87 Std_Dev/Mean: 0.45	Org Std_Dev: 36.4 Org Mean: 105.53 Org Std_Dev/Mean: 0.34 Std_Dev: 31.43 Mean: 108.57 Std_Dev/Mean: 0.29

Nhìn chung, dữ liệu đã được làm sạch đáng kể, thể hiện qua việc giảm tỷ số giữa độ lệch chuẩn và giá trị trung bình. Tỷ số Std_Dev/Mean là một chỉ số quan trọng đánh giá mức độ biến động tương đối của dữ liệu. Việc giảm tỷ số này cho thấy dữ liệu đã trở nên tập trung hơn quanh giá trị trung bình, giảm thiểu ảnh hưởng của các giá trị ngoại lai và nhiễu. Điều này chứng tỏ sự kết hợp của hai phương pháp này đã mang lại hiệu quả trong việc xử lý dữ liệu ngoại lai và làm mượt dữ liệu.

Tuy nhiên, phương pháp kết hợp IQR và MA cũng có những hạn chế nhất định. IQR chỉ dựa trên phân vị của dữ liệu để xác định dữ liệu ngoại lai, bỏ qua thông tin về phân bố của dữ liệu. Điều này có thể dẫn đến việc bỏ sót một số ngoại lai nằm gần (Q_1) hoặc (Q_3) nhưng vẫn có ảnh hưởng đáng kể. Hơn nữa, MA chỉ đơn giản tính trung bình các giá trị trong một cửa sổ trượt, không xem xét đến các yếu tố thời gian hoặc xu hướng phức tạp trong dữ liệu [20].

So với các phương pháp thống kê khác như Z-score, phương pháp kết hợp IQR và MA có ưu điểm là ít bị ảnh hưởng bởi phân bố của dữ liệu. Z-score giả định dữ liệu

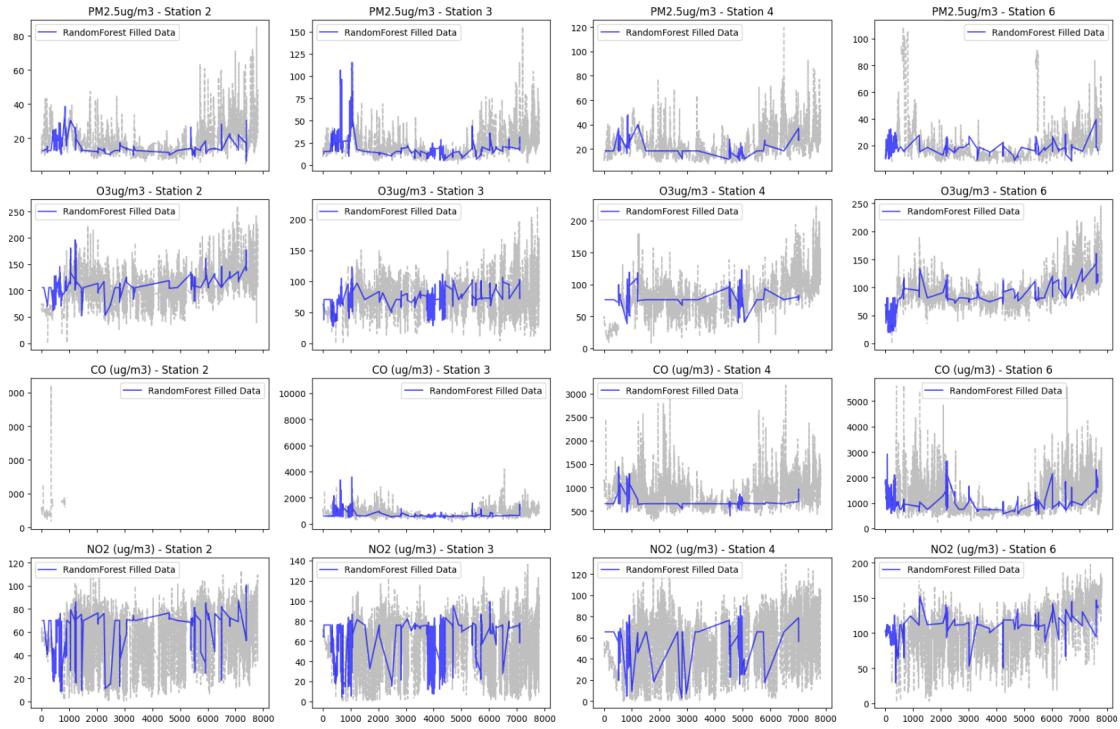
tuân theo phân bố chuẩn, trong khi IQR không yêu cầu điều này. Tuy nhiên, Z-score có thể phát hiện ngoại lai ở cả hai phía của phân bố (cả giá trị quá lớn và quá nhỏ), trong khi IQR tập trung vào các giá trị nằm ngoài khoảng tứ phân vị.

Khi so sánh với các phương pháp học máy và deep learning, phương pháp kết hợp IQR và MA thể hiện sự đơn giản và dễ dàng triển khai. Các phương pháp học máy như Isolation Forest [21], hay TadGAN được sử dụng trong đề tài của tác giả Sang [8] có khả năng phát hiện ngoại lai phức tạp hơn, dựa trên nhiều đặc trưng của dữ liệu và không phụ thuộc vào các giả định về phân bố. Các mô hình deep learning như Autoencoder có thể học biểu diễn nén của dữ liệu và phát hiện ngoại lai dựa trên lỗi tái tạo [22]. Tuy nhiên, các phương pháp này đòi hỏi lượng dữ liệu lớn hơn để huấn luyện, chi phí tính toán cao hơn.

Tóm lại, phương pháp kết hợp IQR và MA là một lựa chọn tốt cho việc xử lý dữ liệu ngoại lai và làm mượt dữ liệu khi yêu cầu về tốc độ xử lý và tính dễ dàng triển khai được ưu tiên. Tuy nhiên, khi dữ liệu phức tạp hoặc yêu cầu độ chính xác cao hơn, các phương pháp học máy và deep learning có thể mang lại kết quả tốt hơn, mặc dù với chi phí tính toán và độ phức tạp cao hơn.

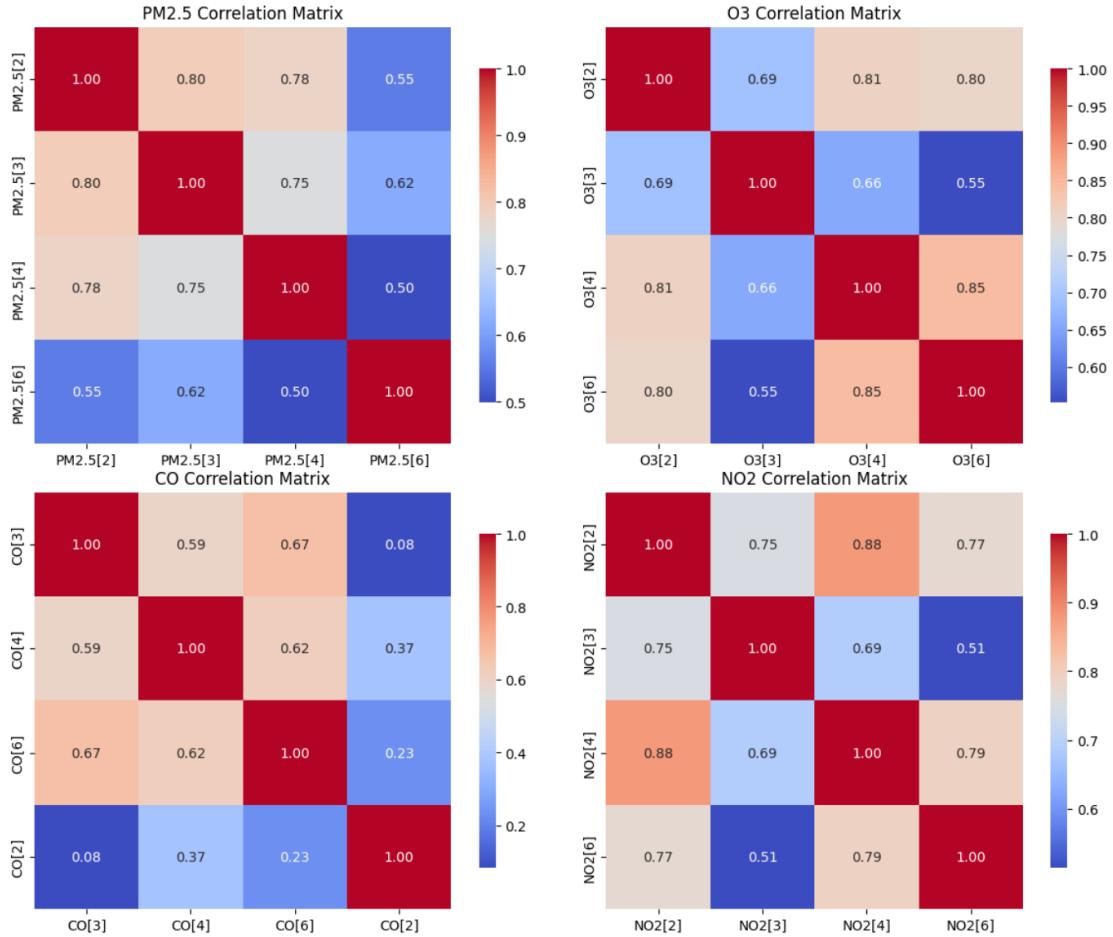
4.2 Quá trình lắp đầy phần dữ liệu khuyết

Trong nghiên cứu này, thuật toán Random Forest được áp dụng để xử lý dữ liệu bị khuyết tại bốn trạm quan trắc, trừ dữ liệu CO ở trạm 2 vì dữ liệu bị khuyết quá nhiều, kết quả thể hiện dưới đây.



HÌNH 4.3: Dữ liệu được hiệu chỉnh dùng Random Forest

Kết quả cho thấy Random Forest đã thể hiện khả năng điền dữ liệu khá tốt, đặc biệt là đối với hai chất có tương quan tốt là NO_2 và O_3 . Quan sát trên hình ảnh cho thấy đường biểu diễn của dữ liệu được điền (màu xanh) bám khá sát với xu hướng của dữ liệu gốc (màu xám), cho thấy Random Forest đã nắm bắt được xu hướng biến động của dữ liệu. Đối với $\text{PM}_{2.5}$ và CO , mặc dù dữ liệu được điền vẫn thể hiện được một phần xu hướng chung, nhưng độ chính xác có vẻ thấp hơn so với NO_2 và O_3 , có thể do tương quan của chúng với các biến khác không mạnh bằng. Ma trận tương quan sau khi làm sạch dữ liệu được thể hiện ở hình bên dưới.



HÌNH 4.4: Ma trận tương quan của các chất ô nhiễm sau khi điền dữ liệu khuyết

Phân tích ma trận tương quan trước (hình 3.9) và sau khi điền dữ liệu cho thấy Random Forest đã bảo toàn rất tốt mối tương quan giữa NO_2 và O_3 , hai chất có tương quan cao. Điều này chứng tỏ phương pháp hiệu quả trong việc điền dữ liệu mà không làm sai lệch mối quan hệ giữa các biến. Đối với $\text{PM}_{2.5}$ và CO , mặc dù tương quan vẫn ở mức thấp sau khi điền, việc loại bỏ các giá trị tương quan âm cho thấy Random Forest đã phần nào "làm mượt" dữ liệu. Tuy nhiên, cần lưu ý rằng phương pháp này có thể gặp khó khăn trong việc khôi phục mối quan hệ giữa các biến vốn dĩ không tương quan mạnh. So sánh ma trận tương quan cũng cho thấy Random Forest không làm xáo trộn cấu trúc tương quan giữa các trạm quan trắc đối với từng chất ô nhiễm. Tuy nhiên, để biết được Random Forest có hoạt động chính xác hay không, tôi đã thử nghiệm mô hình bằng cách tự làm khuyết đi tập dữ liệu hiện tại trong một khoảng thời gian ngắn, trải đều ra các chất ô nhiễm và thuộc các trạm khác nhau. Dữ liệu bị khuyết giả sau khi được điền được so sánh với phương pháp KNN imputation theo chiều thời gian để lấy k láng giềng gần.

BẢNG 4.5: So sánh kết quả điền dữ liệu khuyết PM_{2.5} tại trạm 2

Index	Original	Random Forest	KNN
7344	22.268750	22.730186	15.503889
7397	44.278889	35.296159	38.756667
7445	11.589095	13.898515	12.989306
7475	18.586667	21.023585	27.970900
7478	23.007500	23.612664	25.504301
7480	24.211667	22.078440	17.829006
7493	21.048807	20.892108	17.617529
7548	56.207149	56.419231	50.360694
7583	20.944167	20.457044	21.573333
7619	60.877083	62.228751	60.335972

BẢNG 4.6: So sánh kết quả điền dữ liệu khuyết O₃ tại trạm 3

Index	Original	Random Forest	KNN
7346	96.252433	73.495367	95.909284
7347	93.326510	69.867788	94.925442
7390	116.348646	93.057225	116.027295
7393	126.777083	85.120786	95.657309
7414	73.080854	77.854582	101.549687
7418	67.592633	71.741047	77.252229
7444	47.701732	63.658526	86.751344
7503	56.123998	67.685427	83.779204
7509	127.349625	126.698865	110.149966
7611	50.432742	56.634300	80.837184

BẢNG 4.7: So sánh kết quả điền dữ liệu khuyết CO tại trạm 4

Index	Original	Random Forest	KNN
7353	1631.044500	1116.137084	1336.602083
7403	1578.069951	1385.993775	1278.777382
7417	968.226750	961.297035	1021.786333
7439	831.107067	976.973820	1064.519046
7508	952.100893	1030.918515	887.863500
7511	988.663750	1228.848293	1169.314244
7549	1842.594364	1375.908296	1029.203500
7554	1450.036188	1197.642645	1142.124292
7608	1239.783513	1302.445517	1468.694500
7621	1330.759820	1223.804814	965.938161

BẢNG 4.8: So sánh kết quả điền dữ liệu khuyết NO₂ tại trạm 6

Index	Original	Random Forest	KNN
7339	112.127992	127.719303	101.602402
7386	118.613050	95.260609	99.114435
7391	194.049883	162.694074	162.847892
7404	139.074933	101.178926	113.493844
7457	89.171513	98.784375	103.120444
7498	139.871808	135.057842	111.892742
7530	114.433442	104.357264	97.581700
7542	155.271721	153.740170	140.475396
7562	166.392325	158.535159	144.731722
7580	143.447608	113.607638	122.759866

BẢNG 4.9: Các chỉ số đánh giá của mô hình Random Forest và KNN của các chất

Chất ô nhiễm	Random Forest	KNN
PM _{2.5}	RMSE: 3.15 MAE: 1.91 MAPE: 7.25%	RMSE: 5.10 MAE: 4.24 MAPE: 17.32%
O ₃	RMSE: 19.53 MAE: 15.45 MAPE: 18.12%	RMSE: 23.21 MAE: 18.58 MAPE: 28.51%
CO	RMSE: 261.56 MAE: 206.75 MAPE: 15.04%	RMSE: 348.18 MAE: 284.06 MAPE: 20.69%
NO ₂	RMSE: 20.93 MAE: 17.19 MAPE: 12.65%	RMSE: 21.20 MAE: 20.27 MAPE: 14.76%

Kết quả cho thấy Random Forest vượt trội hơn KNN về các chỉ số RMSE, MAE và MAPE, cho thấy khả năng nắm bắt các mối quan hệ phức tạp trong dữ liệu tốt hơn. Tuy nhiên, một vấn đề được ghi nhận là số lượng trạm quan trắc khác nhau giữa các chất ô nhiễm (ba trạm cho CO so với bốn trạm cho các chất còn lại) đã ảnh hưởng đến các chỉ số đánh giá. Việc ít trạm hơn đồng nghĩa với ít dữ liệu hơn để huấn luyện mô hình, dẫn đến khó khăn trong việc nắm bắt sự biến động theo không gian và thời gian của CO, từ đó làm giảm độ chính xác của việc điền dữ liệu và khiến các chỉ số RMSE, MAE và MAPE của CO có xu hướng cao hơn.

Xét đến sự phân bố của dữ liệu trước và sau khi sử dụng hai phương pháp Random Forest và KNN để điền dữ liệu khuyết. Biểu đồ Histogram cho thấy sau khi áp dụng KNN, biểu đồ histogram cho thấy Random Forest thường tạo ra biểu đồ histogram mượt mà hơn, gần giống với phân bố ban đầu của dữ liệu trước khi bị khuyết. Tuy nhiên, đối với dữ liệu của CO, có thể thấy KNN làm tốt hơn Random Forest vì KNN dựa vào trung vị của các lảng giềng gần nhất để điền giá trị, trong khi Random Forest dựa vào số trạm lân cận, và CO bị khuyết mất dữ liệu của trạm 2, dẫn đến hiệu suất kém hơn so với những chất còn lại.

Một ưu điểm nổi bật của Random Forest là khả năng xử lý tốt với dữ liệu nhiều và ít bị ảnh hưởng bởi dữ liệu ngoại lai [23]. Điều này đặc biệt hữu ích trong bối cảnh dữ liệu quan trắc môi trường thường chứa nhiều nhiễu và biến động. Tuy nhiên, phương pháp này cũng tồn tại một số nhược điểm. Thứ nhất, Random Forest có thể gặp khó khăn khi điền dữ liệu cho các biến có tỷ lệ khuyết quá cao hoặc ít tương quan với các biến khác. Trong trường hợp này, các phương pháp khác như điền bằng giá trị trung bình, trung vị hoặc k-NN có thể đơn giản và hiệu quả hơn, trường hợp chất CO là minh chứng cụ thể. Thứ hai, Random Forest có thể tốn kém về mặt tính toán, đặc biệt là với dữ liệu lớn và số lượng cây quyết định nhiều.

So với các phương pháp điền dữ liệu dựa trên không gian như Kriging hoặc IDW, Random Forest có ưu điểm là không yêu cầu giả định về cấu trúc không gian của dữ liệu. Kriging và IDW dựa trên giả định rằng các điểm gần nhau sẽ có giá trị tương tự nhau, và có thể không phù hợp nếu giả định này không đúng, cụ thể là trong bài toán này. So với các phương pháp dựa trên thời gian như ARIMA, Random Forest có thể xử lý dữ liệu đa biến và không yêu cầu dữ liệu phải là chuỗi thời gian đều đặn. ARIMA tập trung vào việc nắm bắt xu hướng và tính chu kỳ trong dữ liệu thời gian, và có thể không hiệu quả khi dữ liệu bị khuyết một cách ngẫu nhiên.

Khi so sánh với các phương pháp deep learning như Autoencoder hoặc GAN, Random Forest có ưu điểm là dễ triển khai và ít tốn tài nguyên hơn, đặc biệt là với những hệ thống thời gian thực. Các mô hình deep learning có khả năng học các biểu diễn phức tạp của dữ liệu và có thể cho kết quả điền dữ liệu rất tốt với dữ liệu có cấu trúc phức tạp. Tuy nhiên, chúng đòi hỏi lượng dữ liệu lớn hơn để huấn luyện, chi phí tính toán cao hơn, và khó nắm bắt hơn. Một số nghiên cứu đã chỉ ra rằng, trong một số trường hợp, Random Forest có thể đạt được hiệu suất tương đương với deep learning trong bài toán điền dữ liệu, đặc biệt khi dữ liệu không quá lớn [24].

4.3 Kết quả mô hình dự đoán AQI

Mô hình dự báo AQI sử dụng CNN-LSTM đã được xây dựng thành công để dự báo AQI theo trạm và chất ô nhiễm chính của trạm đó. Dữ liệu được sử dụng để huấn luyện mô hình bao gồm các giá trị AQI của các chất ô nhiễm khác nhau cùng với AQI

tổng hợp, tập trung vào một trạm quan trắc cụ thể. Mô hình dự báo cho ba khung thời gian khác nhau: 1 giờ, 4 giờ và 8 giờ tới.

Dựa trên kết quả huấn luyện ở epoch cuối cùng (epoch 50):

Loss: 0.0341 MAE: 0.1144 Val_/_/Loss: 0.0449 Val_/_/MAE: 0.1219

Nhìn chung, mô hình đạt được kết quả huấn luyện chấp nhận được với giá trị loss và MAE tương đối nhỏ. Tuy nhiên, để đánh giá chính xác hơn về hiệu suất dự báo, cần phải thực nghiệm trên tập dữ liệu kiểm thử đã được phân chia và chưa từng được sử dụng để huấn luyện model này.

BẢNG 4.10: So sánh kết quả AQI dự đoán với thực tế

Thời gian dự báo	Trạm	AQI dự đoán	Chất ô nhiễm chính dự đoán	AQI thực tế	Chất ô nhiễm chính thực tế
T+1	2	48.4735	NO ₂	51	NO ₂
T+1	3	64.86291	PM _{2.5}	55	PM _{2.5}
T+1	4	50.586693	PM _{2.5}	60	PM _{2.5}
T+1	6	80.397606	NO ₂	77	NO ₂

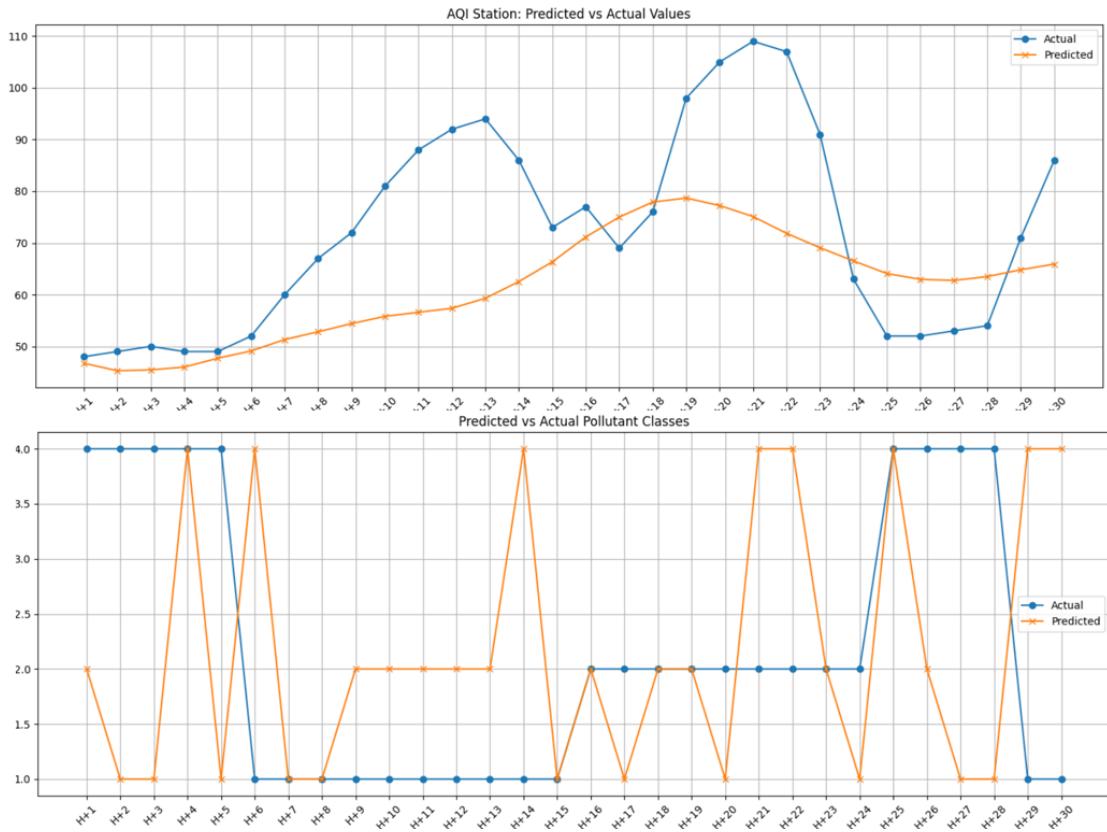
Kết quả cho thấy mô hình dự đoán khả quan trong khoảng thời gian T+1, tuy nhiên khi kéo dài thời gian dự báo lên đến T+4 hay T+8 thì hiệu suất dự báo giảm đi xấp xỉ 4 lần.

BẢNG 4.11: So sánh hiệu suất giữa các mô hình về dự báo AQI

Nghiên cứu	Mô hình	Thời gian dự báo	RMSE	MAE
Trong nghiên cứu này	CNN-LSTM	1h	2.98	2.97
		2h	5.91	5.24
		4h	11.07	9.46
		8h	26.38	21.91
		24h	26.18	20.33
		48h	18.58	30.30
Dũng và cs	LSTM	1 ngày	24.24	13.97
		14 ngày	27.17	19.06
	LSTM-MA	1 ngày	3.05	2.17
		14 ngày	22.79	15.74
Duan và cs	LSTM	Không đề cập	12.3-48.0	9.1-32.9
	CEEMDAN-LSTM	Không đề cập	6.6-24.8	4.6-17.9

Bảng 4.11 trình bày hiệu suất giữa mô hình CNN-LSTM trong nghiên cứu này và một số nghiên cứu khác. Có thể thấy nghiên cứu cho ra kết quả kém hơn so với nghiên cứu Nghiên cứu "Dự báo chất lượng không khí bằng mô hình LSTM-MA trường hợp sử dụng dữ liệu tại trạm quan trắc tự động ngã tư Giếng Nước, tỉnh Bà Rịa - Vũng Tàu" của tác giả Hồ Minh Dũng [5]. Có một số lưu ý như sau:

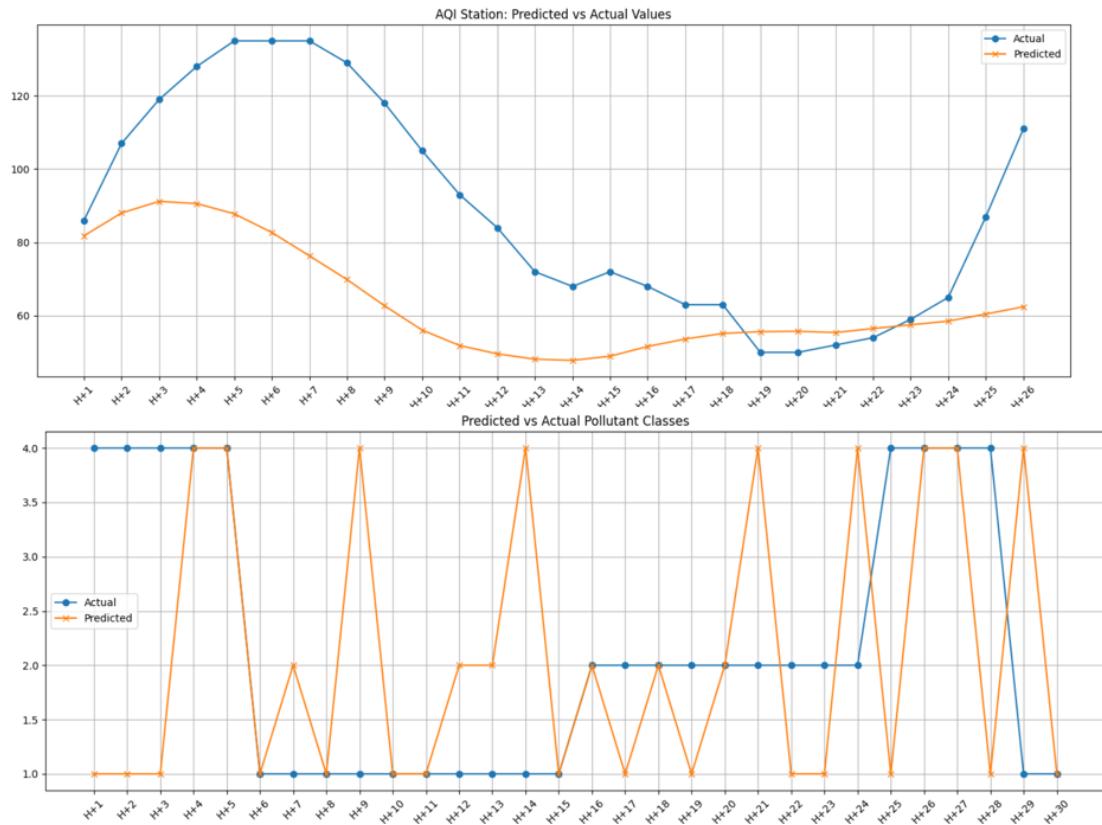
- Dữ liệu trong nghiên cứu chỉ chứa khoảng 1% tỉ lệ dữ liệu khuyết [5].
- Số trạm quan trắc được sử dụng để dự đoán trong nghiên cứu là 1 và định hướng của bài toán là theo chiều thời gian [5].
- Vị trí của trạm quan trắc là tỉnh Bà Rịa - Vũng Tàu, nơi chỉ số AQI tốt hơn và khoảng dao động ít hơn so với thành phố Hồ Chí Minh [5].
- Bài nghiên cứu hướng tới dự đoán AQI theo ngày, đây là một chỉ số tổng quan và ít biến động hơn so với AQI theo giờ [5].



HÌNH 4.5: Biểu đồ dự đoán AQI và chất ô nhiễm chính trên trạm 2

BẢNG 4.12: Các chỉ số đánh giá của trạm 2

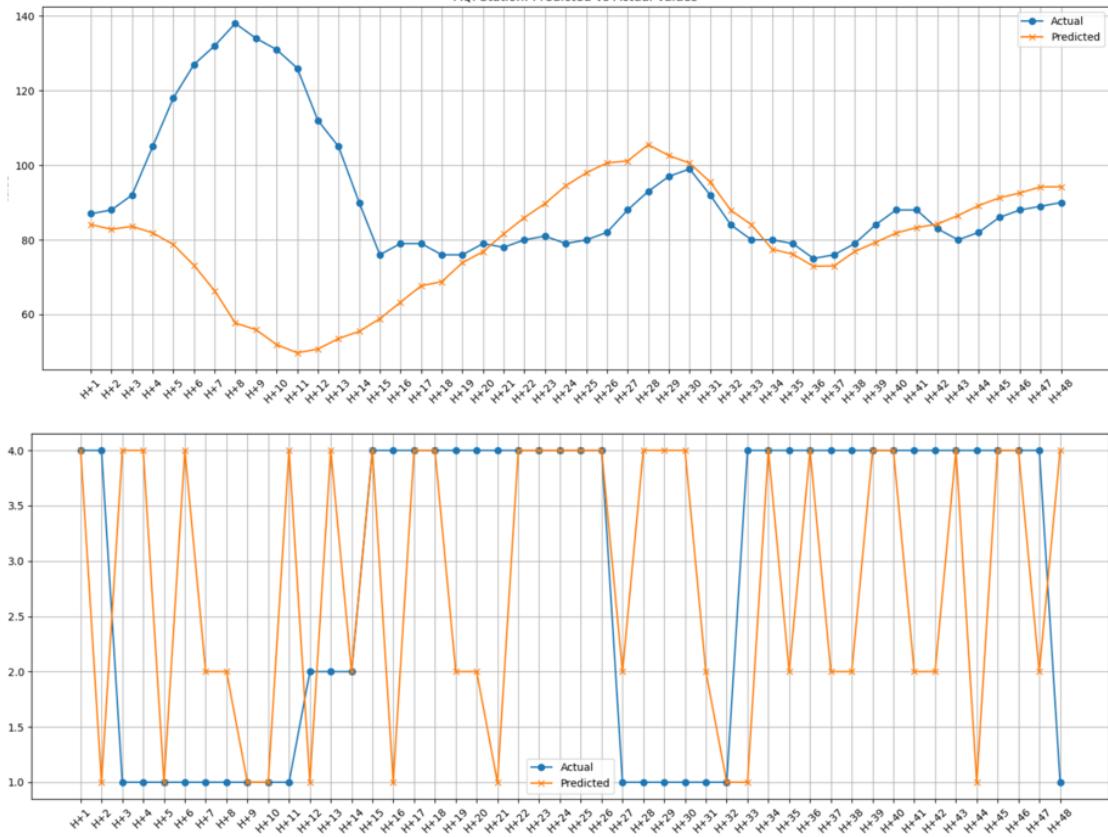
Thời gian	MAE	RMSE	MAPE	sMAPE	Correlation	DA (%)	MASE
1h	1.2673	1.2673	2.6406	2.6753	nan	nan	0.1115
2h	2.4880	2.7714	5.1485	5.2711	-1.0	0.0	0.2189
4h	3.1299	3.3544	6.3547	6.5945	-0.8019	33.33	0.2753
8h	4.9443	6.4356	9.2519	9.2519	0.9230	57.14	0.4254
24h	15.3572	19.7415	35.1657	20.6912	0.1472	65.21	1.3587



HÌNH 4.6: Biểu đồ dự đoán AQI và chất ô nhiễm chính trên trạm 2

BẢNG 4.13: Các chỉ số đánh giá của trạm 3

Thời gian	MAE	RMSE	MAPE	sMAPE	Correlation	DA (%)	MASE
1h	4.1825	4.1825	4.8634	4.9846	nan	nan	0.2510
2h	11.5752	13.7345	11.2952	12.2178	1.0	100.00	0.6948
4h	22.0856	25.2405	18.7910	21.2740	0.9579	66.67	1.3256
8h	38.2013	42.4949	29.7676	36.3084	-0.1436	42.86	2.2929
24h	25.4719	32.1632	25.3950	30.6502	0.7489	56.52	1.5289



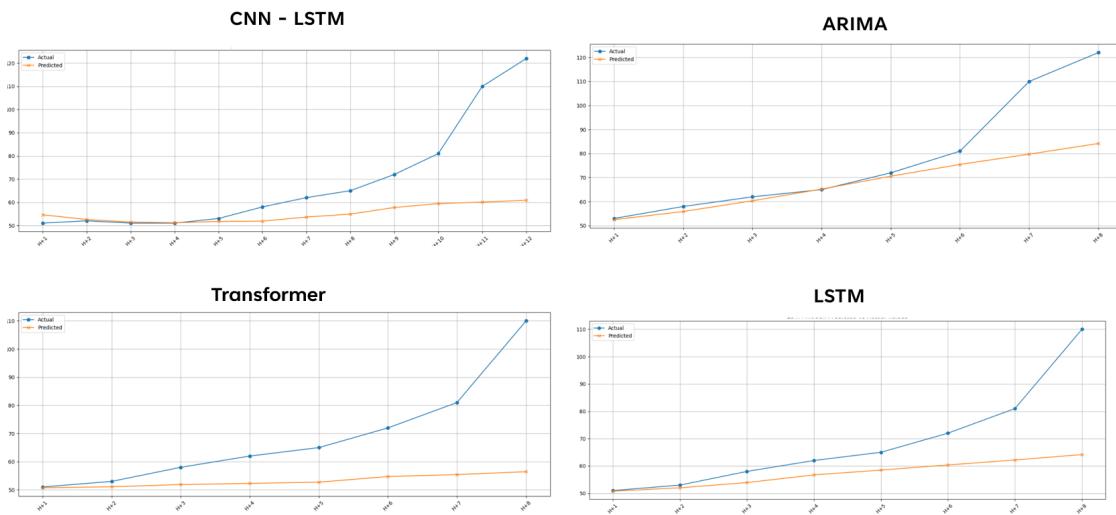
HÌNH 4.7: Biểu đồ dự đoán AQI và chất ô nhiễm chính trên trạm 2

BẢNG 4.14: Các chỉ số đánh giá của trạm 6

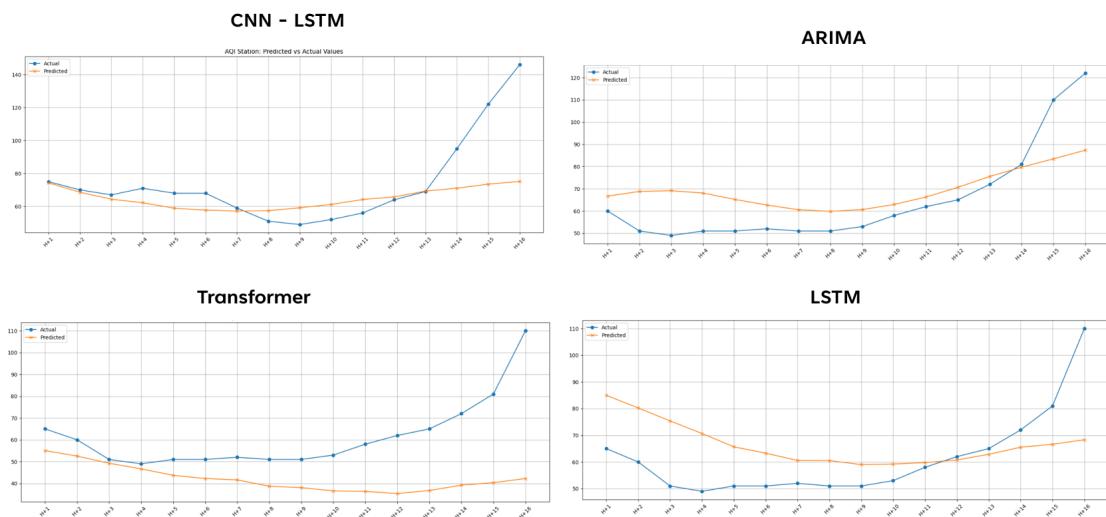
Thời gian	MAE	RMSE	MAPE	sMAPE	Correlation	DA (%)	MASE
1h	2.8669	2.8869	3.3184	3.3744	nan	nan	0.2692
2h	4.0305	4.1896	4.5989	4.7159	-0.9999	0.0	0.3758
4h	9.8934	12.6494	10.0849	12.9945	-0.8198	33.33	0.9225
8h	34.8280	44.4919	27.9914	35.7121	-0.9033	14.29	3.2476
24h	31.1995	42.1769	27.4771	35.3003	-0.5149	34.78	2.9093
48h	18.5803	30.3012	17.2516	21.0094	-0.4471	55.32	1.7326

Dựa trên những phân tích về kết quả và những hạn chế của mô hình CNN-LSTM hiện tại, một số hướng tiếp cận được đề xuất để cải thiện độ chính xác dự báo trong tương lai, tập trung vào việc tối ưu cấu trúc mô hình và áp dụng các kỹ thuật tiên tiến hơn. Cụ thể, việc tối ưu kiến trúc CNN-LSTM sẽ được thực hiện bằng cách thử nghiệm với các cấu hình khác nhau về số lượng lớp, số lượng hidden units, kích thước kernel, kích thước pooling và các tham số khác, nhằm tìm ra cấu trúc tối ưu nhất cho bài toán dự báo AQI. Bên cạnh đó, việc bổ sung các lớp Batch Normalization hoặc Dropout cũng được xem xét để giảm thiểu tình trạng overfitting và tăng tốc độ huấn luyện mô hình.

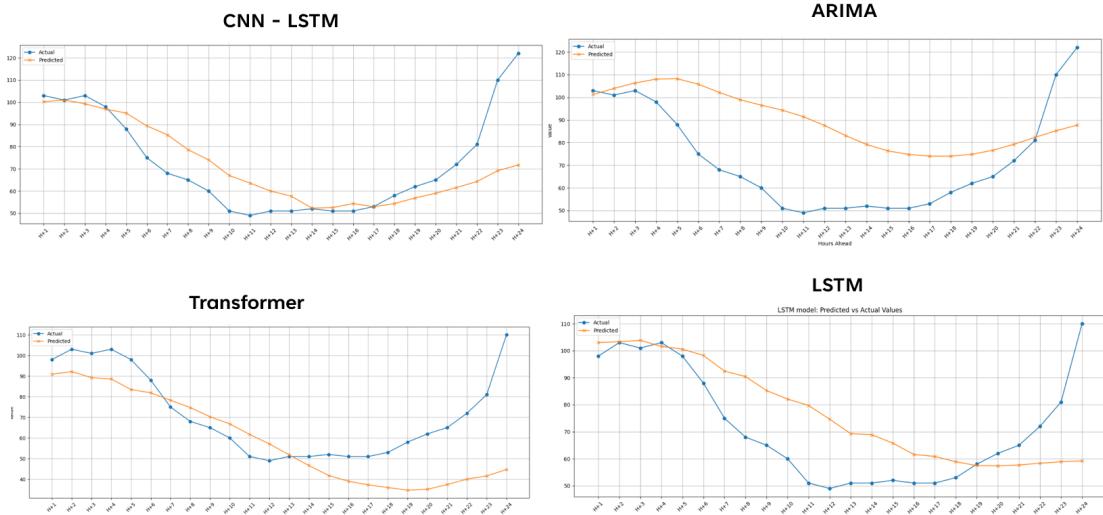
Ngoài ra, luận văn còn thực hiện so sánh kết quả giữa mô hình lai LSTM-CNN với các mô hình khác ở trạm quan trắc số 4 với khoảng dự đoán là 8 giờ, 16 giờ và 24 giờ tiếp theo. Cụ thể các mô hình được nhắc đến là mô hình truyền thống ARIMA, LSTM và mô hình Transformer đang nhận được nhiều sự chú ý hiện nay. Kết quả cho thấy, tại trạm quan trắc này, mô hình LSTM-CNN tiếp tục thể hiện ưu thế vượt trội so với các mô hình LSTM, ARIMA và Transformer. Cụ thể, LSTM-CNN đạt MAE và RMSE thấp nhất, đồng thời có hệ số tương quan cao nhất. Điều này cho thấy khả năng của LSTM-CNN trong việc nắm bắt các đặc trưng phức tạp của dữ liệu AQI, vốn có nhiều biến động và chịu ảnh hưởng của nhiều yếu tố.



HÌNH 4.8: Các biểu đồ dự đoán AQI trong 8 giờ tiếp theo



HÌNH 4.9: Các biểu đồ dự đoán AQI trong 16 giờ tiếp theo



HÌNH 4.10: Các biểu đồ dự đoán AQI trong 24 giờ tiếp theo

Model	hour	MAE	RMSE	Correlation	DA (%)
Arima	4h	4.52	5.57	0.58	33.3
	8h	10.586	14.53	0.9	71.4
	24h	17.4	20.95	0.67	69.6
LSTM	4h	2.39	2.98	-0.32	33.3
	8h	7.79	10.92	0.97	71.4
	24h	13.93	17.98	0.6	60.8
Transformer	4h	11.04	11.34	-0.1	66.7
	8h	9.35	10.12	0.93	85.7
	12h	8.81	9.47	0.97	90.9
	24h	15.76	21.1	0.65	69.6
CNN-LSTM	4h	1.89	2.37	0.7	33.3
	8h	7.5	9.78	0.96	71.4
	12h	9.47	11.22	0.96	72.7
	24h	10.77	16.1	0.69	65.2
	48h	18.58	30.3	0.44	55.3

BẢNG 4.15: So sánh các chỉ số đánh giá giữa các mô hình trên trạm số 4

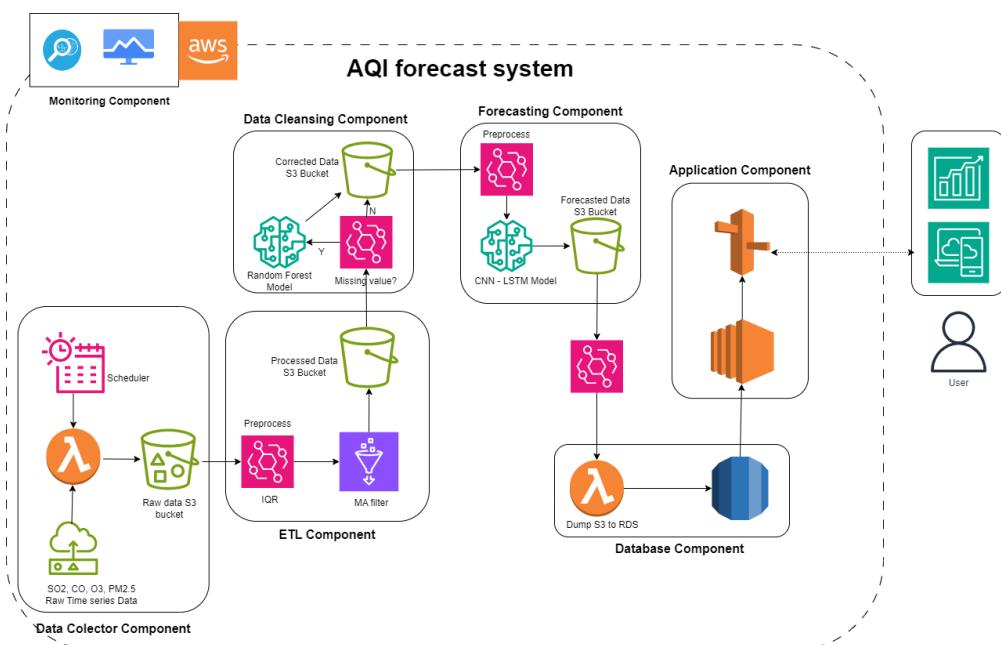
Tuy nhiên, mỗi mô hình đều có những ưu nhược điểm riêng. LSTM có thể gặp khó khăn trong việc xử lý các đặc trưng không liên tục, trong khi ARIMA có thể không hiệu quả với dữ liệu phi tuyến tính. Transformer, mặc dù mạnh mẽ, lại đòi hỏi chi phí

tính toán lớn. Do đó, LSTM-CNN, với khả năng cân bằng giữa độ chính xác và hiệu quả tính toán, là lựa chọn tối ưu cho bài toán này.

4.4 Thực nghiệm hệ thống

4.4.1 Triển khai hệ thống trên AWS

Dựa trên thiết kế tổng quan hệ thống, hệ thống được triển khai thiết kế chi tiết và xây dựng trên nền tảng Amazon Web Service.



HÌNH 4.11: Thiết kế hệ thống dự đoán AQI trên AWS

Nhìn chung, các thành phần của các khối cơ bản giống như trong thiết kế tổng quát. Trong đó, các service được Amazon cung cấp nhằm phục vụ cho quá trình xây dựng và vận hành hệ thống, chi tiết như sau:

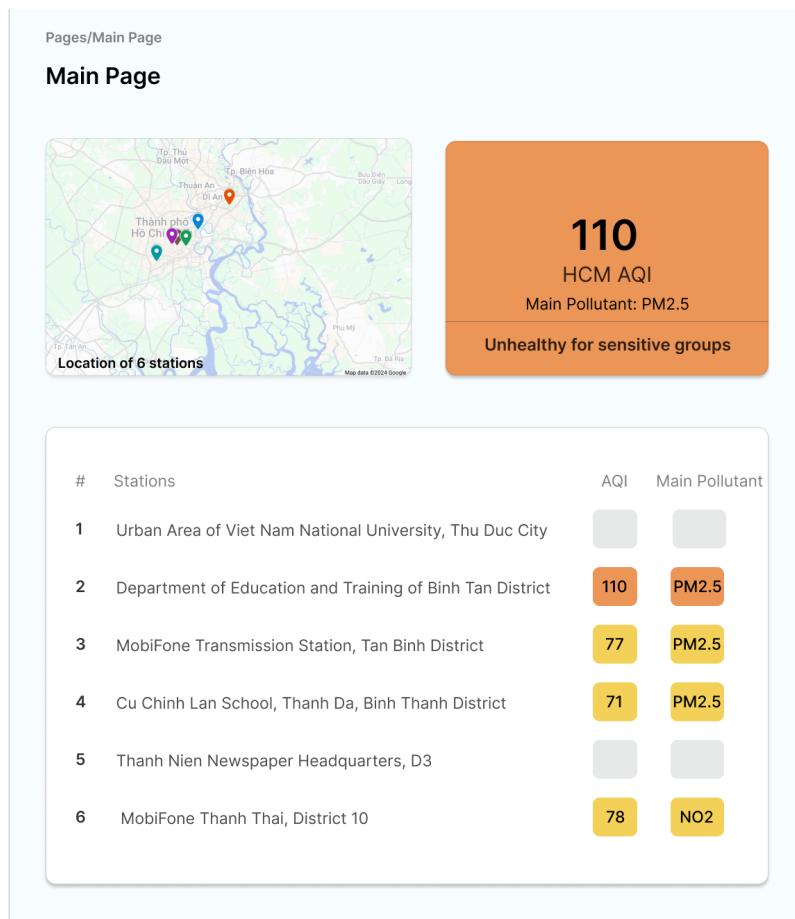
- AWS Lambda là một dịch vụ điện toán phi máy chủ, theo định hướng sự kiện, giúp vận hành code cho hầu hết mọi loại ứng dụng hoặc dịch vụ backend mà không cần cung cấp hay quản lý máy chủ. Trong nghiên cứu này, việc thu thập dữ liệu không xảy ra với tần suất liên tục mà chạy theo một lịch trình cố định. Do đó AWS Lambda là sự lựa chọn hợp lí nhằm tiết kiệm chi phí thay vì sử dụng một máy chủ thường trực.

- Amazon EventBridge Scheduler được sử dụng để tạo các lịch trình kích hoạt việc thực thi các service khác. Trong nghiên cứu này, scheduler được sử dụng để thiết lập chu kì chạy cho các AWS Lambda.
- Amazon S3 là dịch vụ lưu trữ đối tượng được xây dựng để lưu trữ và truy xuất bất kỳ lượng dữ liệu nào từ bất cứ nơi nào. Hệ thống sẽ sử dụng các S3 bucket để lưu trữ dữ liệu thô được lấy về từ AWS Lambda trước khi đưa vào giai đoạn tiền xử lí.
- AWS Glue dùng để tích hợp dữ liệu phi máy chủ, giúp người dùng dễ dàng khám phá, chuẩn bị, di chuyển và tích hợp dữ liệu từ nhiều nguồn cho hoạt động phân tích, máy học và phát triển ứng dụng. Trong nghiên cứu, AWS Glue được sử dụng để chuẩn hóa và làm sạch các dữ liệu thô được lưu bên trong S3 bucket. Ngoài ra trong quá trình tiền xử lí dữ liệu, AWS Glue còn có nhiệm vụ chỉnh sửa dữ liệu và tạo ra các feature cho mô hình máy học.
- Amazon EventBridge cho phép truy cập theo thời gian thực vào những thay đổi dữ liệu trong các dịch vụ AWS và ứng dụng phần mềm dưới dạng dịch vụ (SaaS) mà không cần viết mã.
- SageMaker là nền tảng học máy dựa trên điện toán đám mây, hỗ trợ việc xây dựng, huấn luyện và triển khai các mô hình học máy. Trong nghiên cứu này, các SageMaker dùng để huấn luyện mô hình Random Forest và CNN-LSTM.
- Amazon RDS là cơ sở dữ liệu phân tán chạy trên hệ thống của AWS. RDS hỗ trợ nhiều loại cơ sở dữ liệu khác nhau như MySQL, PostgreSQL hay Oracle. Trong hệ thống này, PostgreSQL được sử dụng làm cơ sở dữ liệu chính.
- Amazon EC2 là dịch vụ web cung cấp năng lực điện toán có kích cỡ linh hoạt trên đám mây. Dịch vụ này được thiết kế để giúp các nhà phát triển dễ sử dụng điện toán đám mây ở quy mô web hơn. Trong nghiên cứu này, Amazon EC2 dùng để deploy ứng dụng lên cloud.
- Amazon Route 53 là một dịch vụ web về Hệ thống phân giải tên miền (DNS) có tính sẵn sàng và khả năng mở rộng cao. Route 53 kết nối yêu cầu của người dùng với các ứng dụng Internet chạy trên AWS hoặc tại chỗ. Hệ thống sử dụng dịch vụ này để tạo public domain cho hệ thống.

- Hệ thống sử dụng NodeJS xây dựng một backend đơn giản nhằm phục vụ các REST API. Backend sẽ có nhiệm vụ kết nối với RDS, sau đó lấy các dữ liệu và gửi lên frontend.
- Bên cạnh đó, React được sử dụng để xây dựng Frontend cho ứng dụng. Khi người dùng request vào ứng dụng này, Frontend sẽ gọi các API của Backend để lấy dữ liệu và render cho người dùng.

4.4.2 Thiết kế chi tiết ứng dụng web

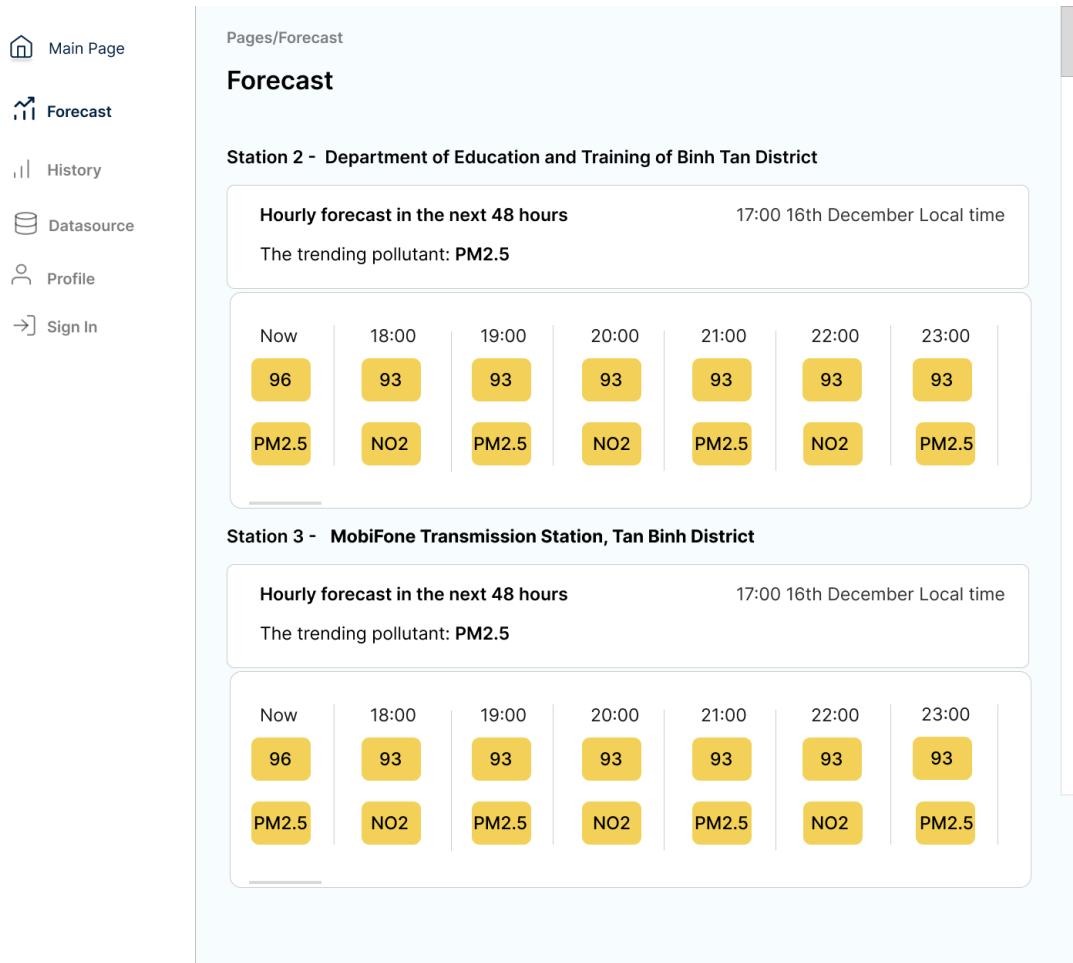
Ứng dụng của hệ thống gồm bốn trang chính: Main Page, Forecast, History và Datasource với thiết kế giao diện như sau:



HÌNH 4.12: Giao diện main page của hệ thống

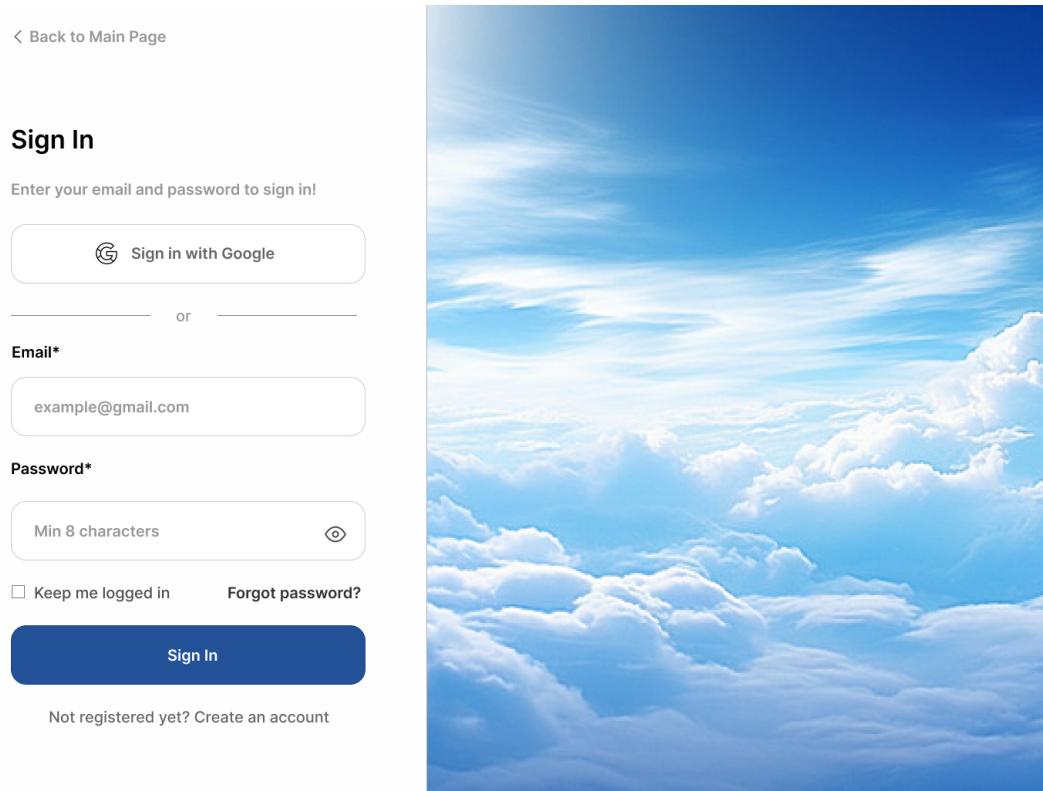
Main page thể hiện chỉ số AQI hiện tại, trạng thái ô nhiễm và chất ô nhiễm chính của sáu trạm quan trắc được sử dụng trong nghiên cứu này. Bên cạnh đó là chỉ

số AQI tạm tính của thành phố Hồ Chí Minh, được lấy theo chỉ số lớn nhất trong số sáu trạm quan trắc



HÌNH 4.13: Giao diện forecast page

Trang forecast thể hiện kết quả dự đoán chỉ số AQI trong 48 giờ tiếp theo của sáu trạm quan trắc. Trang History dùng để tra cứu lịch sử dự đoán trong quá khứ, cũng như là phân tích tỉ lệ chính xác của dự đoán. Trang Datasource dùng để kiểm soát nguồn dữ liệu đầu vào. Trong luận văn này vì chưa thể tiếp cận với nguồn dữ liệu theo thời gian thực của sáu trạm nên dữ liệu giả sẽ được dùng làm dữ liệu đầu vào.



HÌNH 4.14: Giao diện login page

Trang Profile và Sign in dùng cho quá trình xác thực và quản trị hệ thống. Người dùng là khách vãng lai chỉ được tham khảo kết quả dự đoán của hệ thống và không có quyền can thiệp vào nguồn dữ liệu còn lại, nhằm đảm bảo tính bí mật và toàn vẹn của hệ thống.

Chương 5

Kết luận và hướng nghiên cứu tiếp theo

5.1 Kết luận

Nghiên cứu này đã triển khai một hệ thống dự báo chất lượng không khí thời gian thực cho các trạm quan trắc tại Thành phố Hồ Chí Minh, tập trung vào dữ liệu từ sáu trạm đặt tại các quận trung tâm. Dữ liệu thô từ các trạm này gặp phải nhiều thách thức, bao gồm nhiều giá trị ngoại lai và tỉ lệ khuyết dữ liệu cao. Để giải quyết vấn đề này, một quy trình tiền xử lý dữ liệu toàn diện đã được áp dụng. Cụ thể, phương pháp IQR kết hợp với trung bình trượt (MA) đã được sử dụng để hiệu chỉnh và làm mượt dữ liệu ngoại lai, giúp loại bỏ nhiễu và cải thiện tính ổn định của dữ liệu. Bên cạnh đó, thuật toán Random Forest đã được lựa chọn để điều chỉnh các giá trị khuyết, tận dụng khả năng xử lý dữ liệu đa chiều và nắm bắt mối quan hệ phức tạp giữa các biến. Việc sử dụng Random Forest đã cho thấy ưu điểm trong việc bảo toàn cấu trúc phân bố của dữ liệu, tạo ra dữ liệu được điều chỉnh tự nhiên hơn so với các phương pháp khác như KNN (như đã được phân tích ở các phần trước). Điều này đặc biệt quan trọng trong việc duy trì tính chính xác của các phân tích và dự báo tiếp theo. Hệ thống dự báo được xây dựng dựa trên kiến trúc kết hợp CNN-LSTM, tập trung vào dự đoán chỉ số AQI và chất ô nhiễm chính tại khu vực quan trắc. Kết quả đánh giá cho thấy hệ thống đã đạt được mục tiêu dự đoán được xu hướng ô nhiễm một cách khá chính xác và vận hành ổn định trong môi trường thời gian thực, ít tiêu tốn tài nguyên trong quá trình xử lý dữ liệu.

Tuy nhiên, nghiên cứu vẫn còn một số hạn chế cần được tiếp tục nghiên cứu và cải thiện trong tương lai. Đầu tiên, mặc dù Random Forest đã thể hiện khả năng điều chỉnh dữ liệu khuyết tốt, nhưng trong một số trường hợp, phương pháp này vẫn có thể làm giảm độ biến động của dữ liệu, đặc biệt là ở những khu vực có nhiều dữ liệu khuyết. Việc lựa chọn tham số cho Random Forest và đánh giá ảnh hưởng của việc điều chỉnh dữ liệu lên các đặc trưng thống kê của dữ liệu cần được nghiên cứu sâu hơn. Tiếp đến, việc sử dụng kết hợp IQR và MA có thể làm mất đi một số thông tin quan trọng, đặc biệt là các biến

động ngắn hạn. Việc áp dụng các phương pháp xử lý ngoại lai khác dựa trên phân bố thống kê hoặc học sâu, có thể giúp cải thiện hiệu suất của hệ thống mặc dù sẽ tiêu tốn tài nguyên hơn. Bên cạnh đó, việc lựa chọn các tham số tối ưu cho mô hình CNN-LSTM, cũng như việc thử nghiệm các kiến trúc mạng khác (ví dụ như Transformer), cũng là một hướng nghiên cứu tiềm năng để nâng cao độ chính xác của dự báo. Cuối cùng, việc mở rộng phạm vi nghiên cứu bằng cách bổ sung dữ liệu từ các trạm quan trắc khác, cũng như tích hợp thêm các yếu tố ảnh hưởng đến chất lượng không khí như yếu tố khí tượng (nhiệt độ, độ ẩm, gió) và các hoạt động giao thông, sẽ giúp hệ thống dự báo trở nên toàn diện và chính xác hơn.

5.2 Đề xuất hướng nghiên cứu tiếp theo

Để tiếp tục hoàn thiện và nâng cao hiệu suất của hệ thống dự báo chất lượng không khí, đặc biệt là cho bài toán không gian, một số hướng nghiên cứu tiếp theo được đề xuất như sau:

- Tối ưu hóa phương pháp điền dữ liệu khuyết. Mặc dù Random Forest đã chứng minh được hiệu quả trong việc điền dữ liệu khuyết, việc giảm độ biến động của dữ liệu trong một số trường hợp vẫn là một vấn đề cần được giải quyết. Do đó, cần sử dụng các phương pháp tối ưu như Grid Search hoặc Randomized Search để tìm ra bộ tham số tối ưu cho từng biến hoặc từng trạm quan trắc. Bên cạnh đó là tiến hành nghiên cứu những phương pháp khác như MICE (Multiple Imputation by Chained Equations), Kalman filter hoặc các phương pháp deep learning để lựa chọn phương pháp phù hợp nhất.
- Cải thiện phương pháp xử lý ngoại lai. Việc sử dụng kết hợp IQR và MA tuy đơn giản và hiệu quả trong việc loại bỏ ngoại lai và làm mượt dữ liệu, nhưng có thể làm mất đi thông tin quan trọng, đặc biệt là các biến động ngắn hạn. Do đó, cần nghiên cứu và áp dụng các phương pháp xử lý ngoại lai tiên tiến hơn, chẳng hạn Wavelet transform nhằm phân tích tín hiệu thành các tần số khác nhau để loại bỏ nhiễu mà vẫn giữ lại các đặc trưng quan trọng của tín hiệu. Các phương pháp dựa trên phân bố thống kê: Sử dụng các phương pháp như Z-score, modified Z-score hoặc Hampel filter để phát hiện và xử lý ngoại lai dựa trên phân bố của dữ liệu.

- Tối ưu kiến trúc CNN-LSTM để khai thác hiệu quả cả đặc trưng không gian và đặc trưng thời gian của dữ liệu. Khám phá các kiến trúc mạng deep learning mới hơn như Transformer, Temporal Convolutional Networks (TCN) hoặc các biến thể của LSTM như GRU để so sánh và lựa chọn mô hình phù hợp nhất.
- Thử nghiệm các hàm mất mát khác nhau như Huber loss, quantile loss hoặc hàm mất mát vật lý và các thuật toán tối ưu như AdamW, RMSprop để cải thiện quá trình huấn luyện mô hình.
- Mở rộng phạm vi nghiên cứu, cần bổ sung dữ liệu từ các trạm quan trắc khác, mở rộng hệ thống quan trắc để hình thành mạng lưới quan trắc dày đặc
- Tích hợp thêm các yếu tố ảnh hưởng đến chất lượng không khí như yếu tố khí tượng (nhiệt độ, độ ẩm, gió, bức xạ mặt trời), dữ liệu giao thông (lưu lượng xe, tốc độ xe), hoạt động công nghiệp và các nguồn ô nhiễm khác nhằm tạo thêm đặc trưng quan hệ gần giúp cải thiện mô hình dự đoán. Bên cạnh đó còn giúp xây dựng mô hình đa biến để dự đoán đồng thời nhiều chất ô nhiễm và AQI.

Bằng việc triển khai các hướng nghiên cứu này, hệ thống dự báo chất lượng không khí sẽ ngày càng hoàn thiện, cung cấp thông tin chính xác và kịp thời, góp phần vào công tác quản lý chất lượng không khí và bảo vệ sức khỏe cộng đồng.

Tài liệu tham khảo

- [1] C. Kahraman, “Risk analysis and crisis response,” *Stochastic Environmental Research and Risk Assessment*, vol. 23, pp. 413–414, 2009.
- [2] S. Vakulenko, A. Avlocin, and F. Marais, “Air pollution prediction using deep learning: A survey,” *Environmental Modelling & Software*, vol. 140, p. 105029, 2021.
- [3] P. Broomandi, S. Karimi, and A. Nikfal, “A review of air quality index (aqi) and its applications in assessment of air quality,” *Air Quality, Atmosphere & Health*, vol. 10, no. 3, pp. 365–381, 2017.
- [4] U. Kumar, J. Kumar, and P. Joshi, “A review on air quality modeling – past, present and future,” *Resources and Environment*, vol. 5, no. 4, pp. 108–119, 2015.
- [5] M. D. Hồ and K. D. A. Khang, “Dự báo chất lượng không khí bằng mô hình lstm-ma truong hợp sử dụng dữ liệu tại trạm quan trắc tự động ngã tư giềng nước, tỉnh Bà rịa - vũng tàu,” *Tạp chí Khí tượng Thủy Văn*, vol. 765, pp. 75–89, 2024.
- [6] T. L. Nghiêm, “Nghiên cứu ứng dụng chỉ số chất lượng không khí (aqi) để phục vụ cho công tác quản lý chất lượng không khí,” *Tạp chí Khoa học và Công nghệ*, vol. 46, no. 5, 2008.
- [7] Bộ Tài nguyên và Môi trường, “Báo cáo hiện trạng môi trường quốc gia giai đoạn 2016-2020,” 2021.
- [8] T. Q. Sang, “Xây dựng giải pháp phát hiện bất thường và hiệu chỉnh dữ liệu quan trắc theo thời gian thực,” 2024.
- [9] X. Li, Y. Zhou, J. Zhu, and W. Sun, “Missing value imputation for air quality data using a hybrid spatiotemporal model,” *Science of The Total Environment*, vol. 740, p. 140150, 2020.
- [10] C. Jun, S. Yoon, and J. Kim, “A hybrid model for spatiotemporal imputation of missing air quality data,” *Atmosphere*, vol. 12, no. 1, p. 113, 2021.

- [11] J. Sinclair, R. Garland, A. D. McKerrow, S. J. Loots, and G. Breetzke, "Intra-urban variability of pm 2.5 in a dense low-income settlement on the south african highveld," *Air Quality, Atmosphere Health*, vol. 14, no. 7, pp. 1287–1299, 2021.
- [12] Vinmec, "Chỉ số chất lượng không khí là gì và mối liên hệ tới sức khỏe," n.d., truy cập ngày 10 tháng 12 năm 2024. [Online]. Available: <https://www.vinmec.com/>
- [13] T. C. M. Trưởng, "Hướng dẫn kỹ thuật tính toán và công bố chỉ số chất lượng không khí việt nam (vn_aqi)," 2019, truy cập ngày 10 tháng 12 năm 2024.
- [14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [15] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural Networks*, vol. 18, no. 5-6, pp. 602–610, 2005.
- [16] M. Sundermeyer, R. Schlüter, and H. Ney, "LSTM neural networks for language modeling," in *Proceedings of Interspeech*, 2012, pp. 194–197.
- [17] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [18] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.
- [19] X. Shi, Z. Chen, H. Wang, D. Y. Ye, Z. Wu, and Y. Ying, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," *Advances in Neural Information Processing Systems*, vol. 28, 2015.
- [20] R. J. Hyndman and G. Athanasopoulos, *Forecasting: principles and practice*. OTexts, 2021.
- [21] F. T. Liu, K. M. Ting, and Z. H. Zhou, "Isolation forest," in *2008 Eighth IEEE International Conference on Data Mining*. IEEE, December 2008, pp. 413–422.
- [22] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

- [23] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [24] S. Shahdad *et al.*, “A comparison of machine learning methods for imputation of missing data in environmental datasets,” *Environmental Modelling & Software*, vol. 145, p. 105204, 2021.