**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**KHOA HỆ THỐNG THÔNG TIN**



**BÁO CÁO CUỐI KỲ**

**DỮ LIỆU LỚN**

|  |  |  |
| --- | --- | --- |
| Lớp | : | IS405.P21 |
| Giảng viên hướng dẫn | : | ThS. Nguyễn Hồ Duy Trí |
| Sinh viên thực hiện | : | Phan Nguyễn Lâm Hà 21522030 |
|  |  |  |

🙡🙢 ***Tp. Hồ Chí Minh, 06/2025***

# LỜI CẢM ƠN

Trước hết, chúng em xin gửi tới các thầy, cô khoa Hệ thống thông tin, thuộc trường Đại học Công nghệ thông tin – Đại học quốc gia TP. HCM lời cảm ơn vì đã tận tâm truyền đạt kiến thức, hướng dẫn, đặt nền tảng cơ bản cho chúng em có thể thực hiện đồ án này.

Đặc biệt, chúng em xin gửi lời cảm ơn chân thành đến**Thầy Nguyễn Hồ Duy Trí**

Để đồ án này được đạt kết quả tốt như hiện nay, chúng em đã nhận được rất nhiều sự hỗ trợ và hướng dẫn từ thầy. Mặc dù đã nổ lực cố gắng hết sức nhưng do kiến thức còn nhiều mặt hạn chế, nên trong quá trình thực hiện không tránh khỏi những thiếu sót. Kính mong nhận được sự góp ý và giúp đỡ từ quý thầy cô để chúng em có thể hoàn thiện đồ án một cách trọn vẹn nhất.

Chúng em xin chân thành cảm ơn!

Nhóm sinh viên thực hiện

# DANH MỤC BẢNG BIỂU

[Bảng 1.1 Mức AQI của Hàn Quốc 5](#_Toc200140689)

[Bảng 4.1 Siêu tham số của các mô hình 41](#_Toc200140690)

[Bảng 4.2 Đánh giá kết quả trên bộ dữ liệu CO 42](#_Toc200140691)

[Bảng 4.3 Đánh giá kết quả trên bộ dữ liệu SO2 43](#_Toc200140692)

[Bảng 4.4 Đánh giá kết quả trên bộ dữ liệu NO2 44](#_Toc200140693)

[Bảng 4.5 Đánh giá kết quả trên bộ dữ liệu O3 45](#_Toc200140694)

[Bảng 4.6 Đánh giá kết quả trên bộ dữ liệu PM10 46](#_Toc200140695)

[Bảng 4.7 Đánh giá kết quả trên bộ dữ liệu PM2.5 47](#_Toc200140696)

# DANH MỤC HÌNH ẢNH

[Hình 1.1 Chỉ số PM2.5 so sánh ở bốn quốc gia Seoul, Paris, Berlin, và London 2](#_Toc200140674)

[Hình 3.1 Kiến trúc mạng RNN không được kiểm soát 13](#_Toc200140675)

[Hình 3.2 Mô-đun lặp lại trong RNN tiêu chuẩn chứa một lớp duy nhất 14](#_Toc200140676)

[Hình 3.3 Mô-đun lặp lại trong LSTM tiêu chuẩn chứa một lớp duy nhất 15](#_Toc200140677)

[Hình 3.4 Mô hình transformer 18](#_Toc200140678)

[Hình 3.5 Cấu tạo “hộp đen” (nguồn: https://pbcquoc.github.io/transformer/) 19](#_Toc200140679)

[Hình 3.6 Cấu trúc của bộ mã hóa 20](#_Toc200140680)

[Hình 3.7 Cấu trúc của bộ giải mã 22](#_Toc200140681)

[Hình 3.8 hương thức hoạt động của Mô hình XGBoost (Nguồn: XGBoost or Logistic Regression model for Diabetes Prediction) 24](#_Toc200140682)

[Hình 3.9 Mô hình nghiên cứu 27](#_Toc200140683)

[Hình 4.1 Các trạm quan trắc chất lượng không khí ở Seoul 30](#_Toc200140684)

[Hình 4.2 Chất lượng không khí hàng ngày tại Seoul (từ năm 20117 đến năm 2019) 32](#_Toc200140685)

[Hình 4.3 Mối tương quan giữa các chất ô nhiễm ở Seoul 33](#_Toc200140686)

[Hình 4.4 Mức độ ô nhiễm ở Seoul từ năm 2017 đến năm 2019 34](#_Toc200140687)

[Hình 4.5 So sánh mức độ ô nhiễm tại 25 trạm 35](#_Toc200140688)

# MỤC LỤC

[LỜI CẢM ƠN 2](#_Toc200180015)

[DANH MỤC BẢNG BIỂU 3](#_Toc200180016)

[DANH MỤC HÌNH ẢNH 4](#_Toc200180017)

[MỤC LỤC 5](#_Toc200180018)

[CHƯƠNG 1: MỞ ĐẦU 1](#_Toc200180019)

[1.1. Lý do chọn đề tài 1](#_Toc200180020)

[1.2. Mục đích nghiên cứu 3](#_Toc200180021)

[1.3. Phạm vi nghiên cứu 4](#_Toc200180022)

[CHƯƠNG 2: TỔNG QUAN 8](#_Toc200180023)

[2.1. Các nghiên cứu liên quan 8](#_Toc200180024)

[2.2. Đóng góp nghiên cứu 9](#_Toc200180025)

[CHƯƠNG 3: NGHIÊN CỨU THỰC NGHIỆM 11](#_Toc200180026)

[3.1. Các phương pháp nghiên cứu 11](#_Toc200180027)

[3.1.1 Mô hình ARIMA 11](#_Toc200180028)

[3.1.2 Kiến trúc mô hình LSTM 13](#_Toc200180029)

[3.1.2.1 Tổng quan về kiến trúc mạng Recurrent Neural Network (RNN) 13](#_Toc200180030)

[3.1.2.2 Kiến trúc mạng LSTM 14](#_Toc200180031)

[3.1.2.3 Cách hoạt động của mạng LSTM 15](#_Toc200180032)

[3.1.3 Kiến trúc mô hình Transformer 16](#_Toc200180033)

[3.1.3.1 Tổng quan mô hình Transformer 18](#_Toc200180034)

[3.1.3.2 Bộ mã hóa Encoder 20](#_Toc200180035)

[3.1.3.3 Bộ giải mã Decoder 21](#_Toc200180036)

[3.1.4 Mô hình XGBoost 23](#_Toc200180037)

[3.1.5 Thuật toán QPSO 24](#_Toc200180038)

[3.2. Phương pháp nghiên cứu 27](#_Toc200180039)

[CHƯƠNG 4: ĐÁNH GIÁ VÀ BÀN LUẬN VỀ KẾT QUẢ 29](#_Toc200180040)

[4.1. Dữ liệu 29](#_Toc200180041)

[4.1.1 Khu vực nghiên cứu và dữ liệu 29](#_Toc200180042)

[4.1.2 Phân tích dữ liệu 31](#_Toc200180043)

[4.1.3 Tiền xử lý dữ liệu 36](#_Toc200180044)

[4.2. Các tiêu chí đánh giá mô hình 37](#_Toc200180045)

[4.2.1 Hệ số Mean Absolute Error (MAE) 38](#_Toc200180046)

[4.2.2 Hệ số Mean Square Error (MSE) 38](#_Toc200180047)

[4.2.3 Hệ số Root Mean Square Error (RMSE) 39](#_Toc200180048)

[4.2.4 Hệ số R-Squared (R^2) 39](#_Toc200180049)

[4.3. Kết quả mô hình 40](#_Toc200180050)

[CHƯƠNG 5: KẾT LUẬN 49](#_Toc200180051)

[5.1. Kết luận 49](#_Toc200180052)

[5.2. Hướng phát triển 50](#_Toc200180053)

[TẢI LIỆU THAM KHẢO 51](#_Toc200180054)

# MỞ ĐẦU

## Lý do chọn đề tài

Ô nhiễm không khí đang là một vấn đề toàn cầu nghiêm trọng bởi những tác động tiêu cực sâu sắc đối với sức khỏe con người và môi trường sống. Tại nhiều khu vực trên thế giới, đặc biệt là các thành phố lớn, vấn đề ô nhiễm không khí càng trở nên nổi bật do quá trình công nghiệp hóa, đô thị hóa và phát triển giao thông vận tải ngày càng mạnh mẽ. Hàn Quốc, đặc biệt là thủ đô Seoul, là một trong những nơi đang chịu ảnh hưởng nặng nề nhất của tình trạng này. Các chất ô nhiễm như NO₂, CO, CO₂, O₃ và SO₂, cùng các hạt bụi mịn PM₁₀ và PM₂.₅, đã gây ra nhiều bệnh lý về hô hấp, tim mạch, thậm chí là ung thư, làm suy giảm chất lượng cuộc sống và đe dọa tính mạng của hàng triệu người dân.

Seoul, với mật độ dân số cao và tốc độ đô thị hóa nhanh, gặp phải những thách thức đặc biệt về chất lượng không khí do lượng khí thải lớn từ các ngành công nghiệp, phương tiện giao thông và các điều kiện khí tượng đặc thù. Vị trí địa lý của thành phố, nằm giữa núi và biển, góp phần làm tích tụ các chất ô nhiễm, khiến tình trạng ô nhiễm trở nên trầm trọng hơn. Theo báo cáo của Tổ chức Y tế Thế giới (WHO), Seoul có mức độ ô nhiễm bụi mịn PM₂.₅ rất cao, lên đến 46 µg/m³, vượt xa mức khuyến cáo và trở thành mối quan tâm lớn đối với các nhà quản lý đô thị, nhà hoạch định chính sách và cộng đồng dân cư.

A graph of different cities

Description automatically generated

Hình 1.1 Chỉ số PM2.5 so sánh ở bốn quốc gia Seoul, Paris, Berlin, và London

Trước thực trạng đó, các biện pháp cải thiện chất lượng không khí đã được chính quyền Seoul thực hiện, bao gồm thúc đẩy giao thông công cộng, khuyến khích sử dụng nhiên liệu sạch, đầu tư vào năng lượng tái tạo và triển khai các mạng lưới giám sát chất lượng không khí. Tuy nhiên, tình trạng ô nhiễm vẫn còn nghiêm trọng, đòi hỏi sự cải tiến và tối ưu hóa hơn nữa các hệ thống giám sát và dự báo để kịp thời đưa ra các biện pháp ứng phó hiệu quả hơn.

Vì vậy, nghiên cứu này lựa chọn đề tài "Dự đoán chỉ số chất lượng không khí sử dụng mô hình lai học sâu kết hợp cơ chế chú ý và thuật toán tối ưu hóa đàn hạt lượng tử" với mục tiêu phát triển một mô hình tích hợp ưu điểm của nhiều phương pháp tiên tiến. Mô hình đề xuất bao gồm Attention Convolutional Neural Networks (ACNN), thuật toán Quantum-inspired Particle Swarm Optimization (QPSO) cải tiến mạng Long Short-Term Memory (LSTM) và mô hình XGBoost, nhằm tăng cường khả năng dự đoán chính xác các biến động phức tạp của AQI.

Sự kết hợp này không chỉ giúp khắc phục hạn chế của các mô hình đơn lẻ, mà còn tối ưu hóa khả năng nắm bắt các mối quan hệ toàn cục và cục bộ trong dữ liệu, giảm độ dư thừa và cải thiện đáng kể hiệu suất tính toán. Đồng thời, việc sử dụng bộ dữ liệu thực tế từ thành phố Seoul từ năm 2021 đến 2022 đảm bảo tính ứng dụng cao, cung cấp công cụ hỗ trợ hiệu quả cho các nhà quản lý đô thị và cơ quan môi trường trong việc hoạch định chính sách và chiến lược phát triển bền vững. Qua đó, nghiên cứu này hứa hẹn sẽ mang lại những đóng góp giá trị cả về lý thuyết và thực tiễn trong lĩnh vực dự báo chất lượng không khí.

## Mục đích nghiên cứu

Ô nhiễm không khí là một vấn đề nghiêm trọng trên toàn cầu, đặc biệt là tại các khu vực có tốc độ phát triển đô thị hóa nhanh như Seoul, Hàn Quốc. Vấn đề này không chỉ gây ảnh hưởng xấu tới môi trường mà còn làm suy giảm sức khỏe con người, gây ra hàng loạt bệnh nguy hiểm như hen suyễn, các bệnh tim mạch, và thậm chí là ung thư. Trước tình hình này, việc dự đoán chính xác và đáng tin cậy chỉ số chất lượng không khí (AQI) trở nên đặc biệt quan trọng nhằm hỗ trợ các nhà quản lý, nhà hoạch định chính sách và cơ quan chức năng trong việc thực hiện các chiến lược phát triển bền vững và bảo vệ sức khỏe cộng đồng.

Mục đích chính của nghiên cứu này là xây dựng một mô hình dự đoán AQI tiên tiến kết hợp nhiều kỹ thuật hiện đại nhằm giải quyết những hạn chế của các phương pháp dự đoán truyền thống. Cụ thể, nghiên cứu đề xuất một mô hình lai học sâu dựa trên Attention Convolutional Neural Networks (ACNN), thuật toán Quantum-inspired Particle Swarm Optimization (QPSO) tối ưu mạng Long Short-Term Memory (LSTM), và mô hình XGBoost. Việc kết hợp những mô hình này nhằm khai thác tối đa khả năng dự báo chính xác và ổn định trước những đặc điểm phi tuyến và ngẫu nhiên của dữ liệu AQI.

Quy trình nghiên cứu được thực hiện thông qua hai giai đoạn chính. Đầu tiên, mô hình Transformer được áp dụng để trích xuất và điều chỉnh các thành phần tuyến tính từ dữ liệu AQI thu thập tại Seoul trong giai đoạn từ năm 2021 đến 2022. Sau đó, các thành phần phi tuyến còn lại từ dữ liệu được xử lý bằng kiến trúc học sâu kết hợp cơ chế chú ý nhằm cải thiện khả năng bắt giữ các mối quan hệ toàn cục và cục bộ mà các mô hình truyền thống khó nắm bắt. Đặc biệt, thuật toán QPSO được sử dụng để tối ưu hóa các siêu tham số của mạng LSTM, giúp nâng cao hiệu suất và độ chính xác trong dự báo chuỗi thời gian dài hạn.

Mô hình cuối cùng sử dụng bộ điều chỉnh XGBoost để tổng hợp kết quả dự đoán từ các thành phần tuyến tính và phi tuyến, từ đó tạo ra dự đoán AQI hoàn chỉnh với độ tin cậy cao. Qua đó, nghiên cứu không chỉ hướng tới việc cung cấp một công cụ mạnh mẽ và hiệu quả cho công tác dự báo và giám sát chất lượng không khí mà còn góp phần tích cực vào công cuộc cải thiện chất lượng sống và phát triển bền vững cho các đô thị lớn như Seoul.

## Phạm vi nghiên cứu

Phạm vi nghiên cứu của đề tài tập trung vào việc dự đoán chỉ số chất lượng không khí (AQI) tại Seoul, Hàn Quốc, trong giai đoạn từ năm 2021 đến 2022. Seoul được lựa chọn làm địa điểm nghiên cứu vì đây là một trong những thành phố lớn và đông dân nhất thế giới, nơi chịu tác động mạnh mẽ của ô nhiễm không khí do mật độ công nghiệp, giao thông cao và đặc điểm địa lý đặc biệt, khiến chất lượng không khí trở thành vấn đề đáng lo ngại.

Nghiên cứu này sử dụng dữ liệu hàng ngày được thu thập từ hệ thống giám sát chất lượng không khí chính thức của Seoul, bao gồm các chỉ số về nồng độ các chất ô nhiễm chính như ozone (O₃), carbon monoxide (CO), nitrogen dioxide (NO₂), sulfur dioxide (SO₂), bụi hô hấp (PM₁₀), và bụi mịn (PM₂.₅). Phạm vi này đảm bảo dữ liệu sử dụng trong nghiên cứu đủ lớn và đa dạng, có thể phản ánh chính xác các đặc điểm phức tạp và biến động thường xuyên của tình trạng ô nhiễm không khí trong thành phố.

Về kỹ thuật và phương pháp, phạm vi nghiên cứu bao gồm việc ứng dụng một mô hình lai tiên tiến kết hợp Attention Convolutional Neural Networks (ACNN), Autoregressive Integrated Moving Average (ARIMA), Long Short-Term Memory (LSTM) được tối ưu bởi Quantum Particle Swarm Optimization (QPSO), và mô hình XGBoost. Trong đó, ARIMA được sử dụng để xử lý thành phần tuyến tính của dữ liệu AQI, còn mô hình học sâu lai kết hợp với ACNN và LSTM-QPSO được dùng để xử lý thành phần phi tuyến và các đặc điểm chuỗi thời gian dài hạn phức tạp. Cuối cùng, mô hình XGBoost được áp dụng nhằm tinh chỉnh kết quả cuối cùng, đảm bảo hiệu quả cao nhất trong việc dự đoán AQI.

Phạm vi đánh giá hiệu quả của mô hình được thực hiện bằng cách so sánh với các mô hình dự đoán AQI phổ biến khác thông qua các chỉ số hiệu năng quan trọng như Mean Squared Error (MSE), Mean Absolute Error (MAE), và hệ số xác định (R²). Việc đánh giá này được thực hiện ở cả cấp độ toàn thành phố và tại từng trạm quan trắc khí tượng cụ thể, từ đó cung cấp cái nhìn tổng quan và chi tiết về khả năng ứng dụng thực tế và hiệu quả vượt trội của mô hình lai đề xuất. Kết quả nghiên cứu này sẽ góp phần cung cấp một giải pháp mạnh mẽ và linh hoạt cho công tác giám sát, quản lý và cải thiện chất lượng không khí tại Seoul và có thể mở rộng ứng dụng sang các thành phố có điều kiện tương tự.

Bảng 1.1 Mức AQI của Hàn Quốc

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| AQI level | PM2.5(µg/m3) | PM10(µg/m3) | NO2 (ppm) | SO2 (ppm) | O3 (ppm) | CO (ppm) | Health efects |
| Good (0–50) | 0–15 | 0–30 | 0–0.03 | 0–0.02 | 0–0.03 | 0–2 | Mức độ không ảnh hưởng đến bệnh nhân mắc các bệnh liên quan đến ô nhiễm không khí |
| Moderate (51–100) | 16–35 | 31–80 | 0.031–0.06 | 0.021–0.05 | 0.031–0.09 | 2.1–9 | Mức độ có thể có tác động nhỏ đến bệnh nhân trong trường hợp tiếp xúc lâu dài |
| Unhealthy (101–250) | 36–75 | 81–150 | 0.061–0.2 | 0.051–0.15 | 0.091–0.15 | 9.1–15 | Mức độ có thể gây ra tác động có hại cho bệnh nhân và những người thuộc nhóm nhạy cảm (trẻ em, người già hoặc người yếu) và cũng gây ra cảm giác khó chịu cho công chúng nói chung. |
| Very unhealthy (251–500) | 76–500 | 151–600 | 0.201–2 | 0.151–1 | 0.151–0.6 | 15.1–50 | Mức độ có thể gây ảnh hưởng nghiêm trọng đến bệnh nhân và các thành viên của nhóm nhạy cảm trong trường hợp tiếp xúc cấp tính, và thậm chí công chúng nói chung cũng có thể bị ảnh hưởng nhẹ. |

# TỔNG QUAN

## Các nghiên cứu liên quan

Việc dự báo chất lượng không khí luôn thu hút sự quan tâm đặc biệt của cộng đồng khoa học do tính phức tạp và mức độ ảnh hưởng trực tiếp đến sức khỏe con người. Trong những năm gần đây, các nghiên cứu liên quan đến dự đoán AQI đã được thực hiện rộng rãi và phong phú, áp dụng nhiều kỹ thuật khác nhau từ các mô hình thống kê truyền thống đến những phương pháp hiện đại dựa trên học máy và học sâu.

Các phương pháp thống kê như ARIMA đã được sử dụng phổ biến để dự đoán AQI nhờ khả năng xử lý hiệu quả các dữ liệu tuyến tính và chuỗi thời gian ổn định. Tuy nhiên, những hạn chế của các mô hình thống kê truyền thống trong việc xử lý các đặc tính phi tuyến tính và tính ngẫu nhiên của dữ liệu AQI đã thúc đẩy sự phát triển của các mô hình học máy như mạng Neural, Support Vector Machines (SVM), và các kỹ thuật học sâu như Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM). Ví dụ, nghiên cứu của Usha và cộng sự (2019) [1] đã sử dụng mạng Neural và SVM để cải thiện độ chính xác dự đoán AQI và đề xuất mở rộng ứng dụng cho các thành phố thông minh khác.

Các nghiên cứu về học sâu gần đây như của Zhang và cộng sự (2020) đã nhấn mạnh khả năng của các mô hình CNN và GRU trong việc dự đoán AQI bằng cách kết hợp các đặc trưng không gian và thời gian. Trong khi đó, Jurado và cộng sự (2021) khai thác sức mạnh của CNN để xây dựng một hệ thống dự báo ô nhiễm không khí nhanh chóng và chính xác dựa trên dữ liệu thời gian thực về tốc độ gió, lưu lượng giao thông và cấu trúc đô thị. Ngoài ra, việc tích hợp các phương pháp học sâu như TS-LSTM mở rộng của Mao và cộng sự (2022) [2] đã chứng minh hiệu quả trong việc dự đoán AQI cho khoảng thời gian ngắn.

Để nâng cao độ chính xác dự báo, các mô hình lai đã được phát triển rộng rãi nhằm khai thác ưu điểm của từng kỹ thuật riêng biệt. Chẳng hạn, nghiên cứu của Wu và Lin (2020) giới thiệu một mô hình lai tiên tiến SD-SE-LSTM-BA-LSSVM kết hợp phân rã thứ cấp và các thuật toán tối ưu hóa để cải thiện khả năng dự báo AQI. Một số nghiên cứu khác như Sarkar và cộng sự (2021), Gilik và cộng sự (2021) [3] cũng đã đạt được hiệu quả tốt hơn bằng việc kết hợp các mô hình LSTM và GRU hoặc CNN-LSTM để dự báo ô nhiễm không khí.

Bên cạnh đó, việc tích hợp các thuật toán tối ưu hóa meta-heuristic như Particle Swarm Optimization (PSO), Multi-Verse Optimizer (MVO), và Harris-hawks optimization (HHO) đã được chứng minh có hiệu quả cao trong việc tối ưu hóa các siêu tham số của các mô hình dự đoán, như nghiên cứu của Huang và cộng sự (2020) [4] với việc cải tiến mạng neural bằng thuật toán PSO.

Dựa trên những thành tựu đó, nghiên cứu hiện tại lựa chọn kết hợp các kỹ thuật tiên tiến như Attention Convolutional Neural Networks (ACNN), Quantum-inspired Particle Swarm Optimization (QPSO)-enhanced-LSTM, và mô hình XGBoost để phát triển một mô hình dự báo AQI toàn diện, hiệu quả, có thể ứng dụng vào thực tiễn quản lý chất lượng không khí tại Seoul. Mô hình này hứa hẹn mang lại hiệu quả dự báo vượt trội so với các mô hình trước đây, đồng thời mở rộng khả năng ứng dụng cho các địa phương khác có điều kiện tương tự.

## Đóng góp nghiên cứu

Nghiên cứu này mang lại nhiều đóng góp quan trọng cả về lý thuyết lẫn thực tiễn trong lĩnh vực dự báo chỉ số chất lượng không khí (AQI). Đầu tiên, nghiên cứu giới thiệu một mô hình lai sáng tạo tích hợp các kỹ thuật học sâu tiên tiến và thuật toán tối ưu hóa đàn hạt lượng tử (Quantum-inspired Particle Swarm Optimization - QPSO). Việc áp dụng QPSO để tối ưu hóa mạng Long Short-Term Memory (LSTM) giúp giảm đáng kể sự dư thừa về tham số và thời gian mô phỏng, đồng thời tăng cường khả năng nắm bắt các đặc điểm phức tạp, không đều đặn của dữ liệu AQI.

Bên cạnh đó, nghiên cứu sử dụng cơ chế chú ý kết hợp với mạng neural tích chập (Attention Convolutional Neural Networks - ACNN), giúp mô hình có khả năng tập trung vào các đặc trưng quan trọng và khai thác hiệu quả các mối quan hệ toàn cục và cục bộ trong dữ liệu. Điều này đặc biệt hữu ích khi mô hình cần dự đoán những biến động bất thường và phức tạp trong chất lượng không khí, vốn khó có thể xử lý một cách đầy đủ bởi các phương pháp truyền thống.

Ngoài ra, nghiên cứu áp dụng một chiến lược hai giai đoạn rõ ràng để xử lý dữ liệu AQI, trong đó thành phần tuyến tính của dữ liệu được xử lý bằng mô hình ARIMA, còn thành phần phi tuyến được xử lý bằng mô hình học sâu lai kết hợp ACNN và LSTM-QPSO. Việc tích hợp kết quả từ hai thành phần này thông qua mô hình XGBoost nhằm tinh chỉnh các dự báo cuối cùng, đảm bảo độ chính xác cao nhất trong kết quả dự đoán.

Một đóng góp quan trọng khác của nghiên cứu là việc đánh giá mô hình một cách toàn diện thông qua các chỉ số đánh giá phổ biến như Mean Squared Error (MSE), Mean Absolute Error (MAE) và hệ số xác định (R²). Kết quả thử nghiệm cho thấy mô hình đề xuất đạt được hiệu suất vượt trội, giảm tới 31,13% về MSE, 19,03% về MAE và cải thiện 2% về R² so với các mô hình thông thường tốt nhất. Điều này chứng tỏ mô hình lai đề xuất không chỉ cải thiện đáng kể độ chính xác dự báo mà còn thể hiện tính ổn định và độ tin cậy cao trong nhiều điều kiện và địa điểm khác nhau.

Nhìn chung, nghiên cứu này cung cấp một phương pháp luận mới, hiệu quả và có khả năng ứng dụng cao để hỗ trợ các nhà quản lý đô thị và nhà hoạch định chính sách trong việc theo dõi, dự báo và quản lý chất lượng không khí, đặc biệt tại các thành phố lớn với điều kiện môi trường và khí hậu phức tạp như Seoul.

# NGHIÊN CỨU THỰC NGHIỆM

## Các phương pháp nghiên cứu

### Mô hình ARIMA

Dự báo chuỗi thời gian là một lớp mô hình quan trọng trong thống kê, kinh tế lượng và machine learning. Sở dĩ chúng ta gọi lớp mô hình này là chuỗi thời gian (*time series*) là vì mô hình được áp dụng trên các chuỗi đặc thù có yếu tố thời gian. Một mô hình chuỗi thời gian thường dự báo dựa trên giả định rằng các qui luật trong quá khứ sẽ lặp lại ở tương lai. Do đó xây dựng mô hình chuỗi thời gian là chúng ta đang mô hình hóa mối quan hệ trong quá khứ giữa biến độc lập (biến đầu vào) và biến phụ thuộc (biến mục tiêu). Dựa vào mối quan hệ này để dự đoán giá trị trong tương lai của biến phụ thuộc.

Do là dữ liệu chịu ảnh hưởng bởi tính chất thời gian nên chuỗi thời gian thường xuất hiện những quy luật đặc trưng như : yếu tố chu kỳ, mùa vụ và yếu tố xu hướng. Đây là những đặc trưng thường thấy và xuất hiện ở hầu hết các chuỗi thời gian.

Chúng ta biết rằng hầu hết các chuỗi thời gian đều có sự tương quan giữa giá trị trong quá khứ đến giá trị hiện tại. Mức độ tương quan càng lớn khi chuỗi càng gần thời điểm hiện tại. Chính vì thể mô hình ARIMA sẽ tìm cách đưa vào các biến trễ nhằm tạo ra một mô hình dự báo fitting tốt hơn giá trị của chuỗi.

ARIMA model [5] là viết tắt của cụm từ Autoregressive Integrated Moving Average. Mô hình sẽ biểu diễn phương trình hồi quy tuyến tính đa biến (multiple linear regression) của các biến đầu vào (còn gọi là biến phụ thuộc trong thống kê) là 2 thành phần chính:

* Auto regression: Kí hiệu là AR. Đây là thành phần tự hồi quy bao gồm tập hợp các độ trễ của biến hiện tại. Độ trễ bậc p chính là giá trị lùi về quá khứ p bước thời gian của chuỗi. Độ trễ dài hoặc ngắn trong quá trình AR phụ thuộc vào tham số trễ p.
* Moving average: Quá trình trung bình trượt được hiểu là quá trình dịch chuyển hoặc thay đổi giá trị trung bình của chuỗi theo thời gian. Do chuỗi của chúng ta được giả định là dừng nên quá trình thay đổi trung bình dường như là một chuỗi nhiễu trắng. Quá trình moving average sẽ tìm mối liên hệ về mặt tuyến tính giữa các phần tử ngẫu nhiên (stochastic term).
* Tích hợp (integrate) (I): Điều chỉnh dữ liệu chuỗi để loại bỏ các xu hướng và biến động không ngẫu nhiên. Phân tích tích hợp được sử dụng để chuyển đổi dữ liệu chuỗi không ổn định thành dữ liệu chuỗi ổn định.

Mô hình Autoregressive Integrated Moving Average (ARIMA) là một phương pháp dự báo thống kê được sử dụng để dự đoán giá trị tương lai của chuỗi thời gian dựa trên các giá trị quá khứ của nó. Mô hình ARIMA là sự khái quát của mô hình autoregressive moving average (ARMA), vốn giả định rằng dữ liệu chuỗi thời gian là ổn định, nghĩa là các thuộc tính thống kê của nó không thay đổi theo thời gian. Tuy nhiên, mô hình ARIMA có thể được áp dụng để dự báo các chuỗi thời gian không ổn định bằng cách lấy sai phân dữ liệu trước để làm cho nó trở nên ổn định, và mô hình toán học được thể hiện bởi phương trình (1).

Kiểm định Augmented Dickey–Fuller (ADF) được áp dụng cho cả chuỗi dữ liệu gốc và chuỗi dữ liệu đã lấy sai phân bậc một của từng chuỗi nồng độ ô nhiễm, để đảm bảo tính ổn định và hướng dẫn kỹ thuật mô hình hóa chuỗi thời gian phù hợp. Khi giá trị p-value ≤ 0.01 và giá trị thống kê kiểm định ≥ giá trị tới hạn (1%), thì chuỗi được coi là ổn định.

AIC = −2 ln (L) + 2(p + q + k + 1)

Các tham số p và q được xác định bằng tiêu chuẩn thông tin Akaike (AIC), một tiêu chí đánh giá chất lượng tương đối của các mô hình thống kê đối với một tập dữ liệu nhất định. AIC được sử dụng rộng rãi trong dự báo chuỗi thời gian, bao gồm cả việc lựa chọn thứ tự của mô hình ARIMA. Mô hình có giá trị AIC thấp nhất được coi là mô hình phù hợp nhất. Trong bối cảnh mô hình ARIMA, AIC được tính toán dựa trên hàm khả năng, số lượng tham số mô hình và số lượng mẫu dữ liệu, nhằm mục đích chọn ra mô hình tối ưu nhất để dự báo chính xác giá trị của chuỗi thời gian.

### Kiến trúc mô hình LSTM

#### Tổng quan về kiến trúc mạng Recurrent Neural Network (RNN)

Trong mạng nơ ron cổ điển, đầu ra cuối cùng hiếm khi đóng vai trò là đầu ra cho bước tiếp theo nhưng nếu chú ý đến một hiện tượng trong thế giới thực, chúng ta nhận thấy rằng trong nhiều tình huống, đầu ra cuối cùng của chúng ta không chỉ phụ thuộc vào đầu vào bên ngoài mà còn phụ thuộc vào đầu ra trước đó. Ví dụ, khi con người đọc một cuốn sách, việc hiểu từng câu không chỉ phụ thuộc vào danh sách từ hiện tại mà còn phụ thuộc vào việc hiểu câu trước đó hoặc vào ngữ cảnh được tạo ra bằng cách sử dụng các câu trong quá khứ. Con người không bắt đầu suy nghĩ lại từ đầu mỗi giây. Khi bạn đọc bài luận này, bạn hiểu từng từ dựa trên sự hiểu biết của bạn về những từ trước đó. Khái niệm 'bối cảnh' hay 'sự kiên trì' này không có sẵn trong các mạng thần kinh cổ điển. Không có khả năng sử dụng lý luận dựa trên ngữ cảnh trở thành hạn chế lớn của mạng nơ-ron truyền thống. Recurrent Neural Network (RNN) được khái niệm hóa để giảm bớt hạn chế này.RNN được nối mạng với các vòng phản hồi bên trong để cho phép lưu giữ thông tin lâu dài. Hình 3.1 hiển thị một RNN đơn giản với một vòng phản hồi và phiên bản tương đương chưa được kiểm soát của nó cạnh nhau

A group of letters and numbers on a black background

Description automatically generated

Hình 3.1 Kiến trúc mạng RNN không được kiểm soát

Ban đầu (tại bước thời gian t) đối với một số đầu vào Xt, RNN tạo ra đầu ra ht. Trong bước thời gian tiếp theo (t+1), RNN lấy hai đầu vào Xt+1 và ht để tạo đầu ra ht+1. Vòng lặp cho phép thông tin được truyền từ bước này sang bước tiếp theo của mạng.

Tuy nhiên, RNN không tránh khỏi những hạn chế. Khi 'context' ở gần quá khứ, nó hoạt động rất tốt đối với đầu ra chính xác. Nhưng khi một RNN phải phụ thuộc vào một 'context' xa xôi (tức là thứ gì đó đã được học từ lâu) để tạo ra đầu ra chính xác thì nó sẽ thất bại thảm hại. Hạn chế này của RNN đã được thảo luận rất chi tiết bởi Hochreiter [8] và Bengio, et al. [6]. Họ cũng truy ngược lại các khía cạnh cơ bản để hiểu lý do tại sao RNN có thể không hoạt động trong các tình huống dài hạn. Tin tốt là LSTM được thiết kế để khắc phục vấn đề trên.

#### Kiến trúc mạng LSTM

Hochreiter Schmidhuber [8] đã giới thiệu một loại RNN đặc biệt có khả năng học các phụ thuộc dài hạn. Sau này, nhiều nhà nghiên cứu khác đã cải tiến công trình tiên phong. LSTM được hoàn thiện theo thời gian để giảm thiểu vấn đề phụ thuộc lâu dài. Sự tiến triển và phát triển của LSTM từ RNN được giải thích. Mạng nơ-ron tái diễn có dạng một chuỗi các mô-đun lặp lại của mạng nơ-ron. Trong các RNN tiêu chuẩn, mô-đun lặp này có cấu trúc đơn giản giống như một lớp tanh đơn như trong Hình 3.2.

A diagram of a diagram

Description automatically generated

Hình 3.2 Mô-đun lặp lại trong RNN tiêu chuẩn chứa một lớp duy nhất

LSTM tuân theo cấu trúc giống như chuỗi này, tuy nhiên mô-đun lặp lại có cấu trúc khác. Thay vì có một lớp mạng nơ-ron duy nhất, có bốn lớp, tương tác theo một cách rất đặc biệt như trong Hình 3.3

Trong Hình 3.3, mỗi dòng biểu thị toàn bộ vectơ đặc trưng, ​​từ đầu ra của một nút đến đầu vào của các nút khác. Các vòng tròn màu hồng biểu thị các phép toán theo điểm, như phép cộng vectơ, trong khi các hộp màu vàng là các lớp mạng thần kinh đã học. Việc hợp nhất các dòng biểu thị sự ghép nối, trong khi việc phân nhánh dòng biểu thị nội dung của nó đang được sao chép và các bản sao sẽ đi đến các vị trí khác nhau.

A diagram of a diagram

Description automatically generated

Hình 3.3 Mô-đun lặp lại trong LSTM tiêu chuẩn chứa một lớp duy nhất

#### Cách hoạt động của mạng LSTM

Chìa khóa của LSTM là trạng thái ô, đường ngang chạy qua phần trên của sơ đồ. Trạng thái ô giống như một băng chuyền. Điều này chạy thẳng xuống toàn bộ chuỗi, có một số tương tác tuyến tính nhỏ. LSTM có khả năng thêm hoặc bớt thông tin về trạng thái ô, được điều khiển bởi các cấu trúc gọi là cổng. Cổng được sử dụng để tùy ý cho phép thông tin đi qua. Cổng bao gồm một lớp mạng lưới thần kinh sigmoid và phép toán nhân theo điểm. Lớp sigmoid xuất ra các số từ 0 đến 1, mô tả số lượng mỗi thành phần sẽ được cho phép. Giá trị 0 có nghĩa là “không cho gì đi qua”, trong khi giá trị 1 có nghĩa là “cho mọi thứ đi qua!” LSTM có ba trong số các cổng này, để bảo vệ và kiểm soát trạng thái ô

Bước đầu tiên của LSTM là quyết định thông tin nào sẽ được loại bỏ khỏi trạng thái ô. Nó được tạo bởi một lớp sigmoid được gọi là “lớp cổng quên”. Nó xem xét ht , ht−1và xtxt và xuất ra một số từ 00 đến 11 cho mỗi số ở trạng thái ô ct−1ct−1. Số 11 đại diện cho “hoàn toàn giữ lại cái này” trong khi số 00 đại diện cho “loại bỏ hoàn toàn cái này”. Trong bước tiếp theo, quyết định thông tin mới nào sẽ được lưu trữ ở trạng thái ô. Nó có hai phần. Đầu tiên, một lớp sigmoid được gọi là “lớp cổng đầu vào” quyết định giá trị nào sẽ được cập nhật. Sau đó, lớp tanh tạo ra một vectơ chứa các giá trị ứng viên mới, C tC t, có thể được thêm vào trạng thái. Trong bước tiếp theo, cả hai điều này được kết hợp để tạo ra bản cập nhật trạng thái. Bây giờ là lúc cập nhật trạng thái ô cũ, ct−1ct−1, sang trạng thái ô mới ctct . Chúng tôi nhân trạng thái cũ với ftft. Sau đó chúng ta thêm itC titC t. Đây là các giá trị ứng cử viên mới, được chia tỷ lệ theo mức độ chúng tôi quyết định cập nhật từng giá trị trạng thái. Cuối cùng, chúng ta cần quyết định đầu ra. Đầu ra sẽ là phiên bản được lọc của trạng thái ô. Đầu tiên, chúng tôi chạy một lớp sigmoid để quyết định phần nào của trạng thái ô mà chúng tôi sẽ xuất ra. Sau đó, chúng ta đưa trạng thái ô thông qua tanh-tanh (để đẩy các giá trị nằm trong khoảng từ 11 đến 11) và nhân nó với đầu ra của cổng sigmoid, để chúng ta chỉ xuất ra những phần mà chúng ta quyết định.

### Kiến trúc mô hình Transformer

Mô hình Transformer đã được giới thiệu để giải quyết vấn đề chuỗi-qua-chuỗi. Transformer là một kiến trúc học sâu đã làm thay đổi lĩnh vực xử lý ngôn ngữ tự nhiên (NLP) và được sử dụng rộng rãi cho các nhiệm vụ truyền thông chuỗi. Nó được giới thiệu lần đầu trong bài báo "Attention is All You Need" [7] của Vaswani vào năm 2017.

Mô hình Transformer dựa nhiều vào khái niệm cơ chế tự chú ý (self- attention) để bắt lấy mối quan hệ giữa các từ hoặc thành phần khác nhau trong chuỗi đầu vào. Khác với các mô hình dựa trên mạng nơ-ron hồi quy (RNN) trước đây xử lý đầu vào theo thứ tự, Transformer cho phép xử lý song song chuỗi thông qua các lớp tự chú ý. Mô hình Transformer bao gồm: Bộ mã hóa (Encoder), chuỗi đầu vào được đưa qua một bộ mã hóa, bao gồm nhiều lớp tự chú ý và mạng nơ-ron tiến thuật (feed-forward). Cơ chế tự chú ý cho phép mô hình tập trung vào các phần khác nhau của chuỗi đầu vào khi tạo ra đầu ra. Bộ giải mã (Decoder), bộ giải mã nhận đầu ra của bộ mã hóa và tạo ra chuỗi đầu ra theo từng bước. Tương tự như bộ mã hóa, nó bao gồm nhiều lớp tự chú ý và cơ chế ’chú ý bổ sung’ để chú ý đến đầu vào của bộ mã hóa.

Trong học sâu, mạng Transformer là một sự phát triển vượt bậc so với mạng LSTM (Long Short-Term Memory). Mạng LSTM là một loại mạng nơ-ron hồi quy (RNN) được thiết kế để xử lý các chuỗi dữ liệu dài và có khả năng lưu trữ thông tin qua các bước thời gian dài nhờ vào cơ chế cổng (gates).

Tuy nhiên, LSTM gặp hạn chế trong việc xử lý các chuỗi dữ liệu dài và song song hóa. Mạng Transformer, được giới thiệu lần đầu tiên trong bài báo &quot;Attention is All You Need&quot; năm 2017, đã khắc phục được hạn chế này bằng cách loại bỏ hoàn toàn việc xử lý tuần tự mà RNN/LSTM sử dụng, thay vào đó là cơ chế tự chú ý (self- attention). Điều này giúp Transformer xử lý các chuỗi dữ liệu dài nhanh hơn và hiệu quả hơn, đặc biệt là trong các ứng dụng như dịch máy, nhận dạng tiếng nói và tạo văn bản.

Mô hình Transformer là một mạng nơ-ron học ngữ cảnh của dữ liệu tuần tự và tạo ra dữ liệu mới từ đó. Nói đơn giản, Transformer là một loại mô hình trí tuệ nhân tạo được thiết kế để hiểu và tạo ra văn bản giống như con người bằng cách phân tích các mẫu trong một lượng lớn dữ liệu văn bản. Đây là mô hình NLP tiên tiến hiện nay và được coi là sự phát triển của kiến trúc mã hóa-giải mã. Khác với kiến trúc mã hóa- giải mã dựa vào mạng nơ-ron hồi quy (RNN), Transformer không sử dụng tính tuần hoàn này. Thay vào đó, chúng dựa vào kỹ thuật toán học gọi là cơ chế tự chú ý (attention) để hiểu bối cảnh và ý nghĩa của các phần tử khác nhau trong dữ liệu.

A diagram of a software algorithm

Description automatically generated

Hình 3.4 Mô hình transformer

#### Tổng quan mô hình Transformer

Ban đầu được thiết kế cho các tác vụ chuyển đổi chuỗi hoặc dịch máy bằng nơ- ron, Transformer nổi bật trong việc chuyển đổi các chuỗi đầu vào thành chuỗi đầu ra. Đây là mô hình đầu tiên hoàn toàn dựa vào cơ chế tự chú ý để tính toán các biểu diễn của đầu vào và đầu ra mà không cần sử dụng mạng nơ-ron hồi quy (RNNs) hoặc tích chập. Đặc điểm cốt lõi của kiến trúc Transformer là duy trì mô hình mã hóa-giải mã. Nếu xem Transformer như một hộp đen đơn giản cho dịch ngôn ngữ, nó sẽ nhận một câu ở một ngôn ngữ (ví dụ: tiếng Anh) và xuất ra bản dịch tương ứng.

Nếu đi sâu vào, chúng ta sẽ thấy rằng hộp đen này được cấu thành từ hai phần chính:

* Bộ mã hóa (encoder) nhận đầu vào và tạo ra một biểu diễn ma trận của đầu vào đó. Ví dụ, câu tiếng Anh “little sun” sẽ được chuyển thành một ma trận biểu diễn.
* Bộ giải mã (decoder) sau đó sẽ tiếp nhận biểu diễn đã mã hóa này và dần dần tạo ra đầu ra. Trong ví dụ này, câu dịch sang tiếng Việt “mặt trời bé nhỏ” sẽ được tạo ra theo cách này.

A diagram of a transformer

Description automatically generated

Hình 3.5 Cấu tạo “hộp đen” (nguồn: https://pbcquoc.github.io/transformer/)

Tuy nhiên, cả bộ mã hóa (encoder) và bộ giải mã (decoder) đều thực chất là một chuỗi các lớp chồng lên nhau (cùng số lượng cho mỗi loại). Mỗi bộ mã hóa đều có cấu trúc giống nhau, và dữ liệu đầu vào sẽ đi qua từng lớp một trước khi chuyển tiếp đến lớp tiếp theo. Tương tự, các bộ giải mã cũng có cùng cấu trúc và nhận đầu vào từ bộ mã hóa cuối cùng cũng như từ bộ giải mã trước đó. Kiến trúc ban đầu bao gồm 6 bộ mã hóa và 6 bộ giải mã, nhưng chúng ta có thể nhân lên thành N lớp theo nhu cầu.

Bây giờ khi chúng ta đã có ý tưởng chung về kiến trúc tổng thể của Transformer, hãy tập trung vào cả Bộ mã hóa (Encoders) và Bộ giải mã (Decoders) để hiểu rõ hơn về cách chúng hoạt động.

#### Bộ mã hóa Encoder

Bộ mã hóa là một thành phần quan trọng trong kiến trúc Transformer. Chức năng chính của nó là chuyển đổi các token đầu vào thành các biểu diễn có ngữ cảnh. Khác với các mô hình trước đây xử lý các token một cách độc lập, bộ mã hóa của Transformer nắm bắt được ngữ cảnh của từng token trong toàn bộ chuỗi dữ liệu.

Cấu trúc của bộ mã hóa bao gồm các thành phần sau:

A diagram of a process

Description automatically generated

Hình 3.6 Cấu trúc của bộ mã hóa

Bộ mã hóa (Encoder) trong mô hình Transformer là một thành phần quan trọng, chịu trách nhiệm xử lý dữ liệu đầu vào và chuyển đổi chúng thành các biểu diễn có ý nghĩa, nắm bắt được mối quan hệ và ngữ cảnh của từng từ trong chuỗi.

Cấu trúc chính của Bộ mã hóa:

* Embedding Đầu vào: Chuyển đổi các token đầu vào thành các vector có kích thước cố định.
* Mã hóa Vị trí: Cung cấp thông tin về vị trí của từng token trong chuỗi.
* Cơ chế Tự chú ý (Self-Attention): Tính toán mối quan hệ giữa các token trong chuỗi, cho phép mô hình tập trung vào các phần quan trọng khi mã hóa một token cụ thể.
* Mạng Nơ-ron Tiến tới (Feed-Forward): Áp dụng mạng kết nối đầy đủ cho mỗi biểu diễn token, giúp mô hình nắm bắt các mẫu phức tạp hơn.
* Kết nối Dư và Chuẩn hóa Lớp: Sử dụng giữa các lớp để ổn định và tăng cường quá trình huấn luyện, giúp giữ lại thông tin khi nó được truyền qua mạng.

Bộ mã hóa bao gồm nhiều lớp giống hệt nhau, mỗi lớp chứa các thành phần trên, tạo ra các biểu diễn mã hóa thể hiện ý nghĩa ngữ cảnh của mỗi token trong toàn bộ chuỗi đầu vào.

#### Bộ giải mã Decoder

Bộ giải mã (Decoder) trong mô hình Transformer có nhiệm vụ chuyển đổi các biểu diễn mã hóa thành chuỗi đầu ra, như dịch một câu từ tiếng Anh sang tiếng Việt. Dưới đây là các thành phần chính của bộ giải mã:

* Embedding đầu ra: Tương tự như bộ mã hóa, các token đầu ra được chuyển đổi thành các vector đặc trưng.
* Mã hóa vị trí (Positional Encoding): Thêm thông tin về vị trí của mỗi token trong chuỗi đầu ra.
* Cơ chế tự chú ý (Self-Attention Mechanism): Tính toán mối quan hệ giữa các token trong chuỗi đầu ra.
* Cơ chế chú ý với bộ mã hóa (Encoder-Decoder Attention): Tính toán mối quan hệ giữa các biểu diễn mã hóa và các token đầu ra, cho phép mô hình tập trung vào các phần quan trọng của đầu vào khi dự đoán token tiếp theo.
* Mạng nơ-ron truyền thẳng (Feed-Forward Neural Network): Xử lý các biểu diễn đầu ra qua mạng nơ-ron để tạo ra chuỗi đầu ra.
* Kết nối dư và chuẩn hóa lớp (Residual Connections and Layer Normalization): Giúp ổn định và tăng cường quá trình huấn luyện.

Bộ giải mã sử dụng thông tin từ cả chuỗi đầu ra đã được dự đoán trước đó và các biểu diễn mã hóa để tạo ra chuỗi đầu ra cuối cùng, như một câu dịch hoàn chỉnh.

A diagram of a multi-head attention

Description automatically generated

Hình 3.7 Cấu trúc của bộ giải mã

### Mô hình XGBoost

Mô hình XGBoost [9], viết tắt của eXtreme gradient Boosting, là một triển khai nâng cao của thuật toán tăng cường độ dốc. Nó đã trở nên phổ biến trong học máy nhờ hiệu quả hoạt động trong mô hình dự đoán. Mô hình XGBoost thuộc nhóm thuật toán tăng cường. Tăng cường là một kỹ thuật tổng hợp trong đó các mô hình mới được thêm vào để sửa các lỗi do các mô hình hiện có mắc phải.

Các thành phần chính của mô hình XGBoost bao gồm:

* Cây quyết định: Nó chủ yếu sử dụng cây quyết định làm người học cơ sở. Mỗi cây mới sẽ sửa lỗi của cây trước đó.
* Chính quy hóa [10]: XGBoost bao gồm các thuật ngữ chính quy hóa (L1 và L2) trong hàm chi phí của nó để kiểm soát việc trang bị quá mức, giúp cải thiện hiệu suất của nó.
* Xử lý các giá trị bị thiếu: XGBoost có quy trình tích hợp sẵn để xử lý dữ liệu bị thiếu.
* Cắt tỉa cây: Mô hình sử dụng cách tiếp cận theo chiều sâu và tỉa cây về phía sau, một cách tiếp cận khác với cách tăng cường độ dốc cổ điển.
* Xử lý dữ liệu thưa thớt: Nó được tối ưu hóa để xử lý dữ liệu thưa thớt, như ma trận có nhiều số 0, một cách hiệu quả.
* Xử lý song song: XGBoost có thể tận dụng sức mạnh của xử lý song song, giúp quá trình này nhanh hơn đáng kể so với các triển khai tăng cường độ dốc khác.

A diagram of a function

Description automatically generated

Hình 3.8 hương thức hoạt động của Mô hình XGBoost (Nguồn: XGBoost or Logistic Regression model for Diabetes Prediction)

Mô hình XGBoost có khả năng tính toán và song song hóa nhanh hơn. Mô hình có khả năng mở rộng, hoạt động tốt với các tập dữ liệu lớn. Có thể được sử dụng cho các vấn đề hồi quy, phân loại, xếp hạng và dự đoán do người dùng xác định. Sự kết hợp giữa hiệu quả, độ chính xác và tính linh hoạt của XGBoost khiến mô hình này trở thành lựa chọn phổ biến cho các nhà khoa học dữ liệu và người thực hành máy học.

### Thuật toán QPSO

Thuật toán tối ưu bầy đàn lượng tử (Quantum-inspired Particle Swarm Optimization - QPSO [11]) là một biến thể của thuật toán Tối ưu bầy đàn truyền thống (PSO - Particle Swarm Optimization), được xây dựng dựa trên các nguyên lý của cơ học lượng tử. Khác với PSO cổ điển, nơi mỗi hạt có vị trí và vận tốc xác định rõ ràng, trong QPSO các hạt không có vận tốc rõ ràng mà thay vào đó là một phân bố xác suất mô tả vị trí của hạt. Điều này giúp QPSO có khả năng khám phá không gian tìm kiếm tốt hơn và tránh rơi vào cực trị địa phương.

QPSO mô hình hóa hành vi của các hạt dựa trên hiện tượng đường hầm lượng tử, cho phép hạt có thể dịch chuyển qua các rào cản năng lượng. Trong quá trình cập nhật vị trí, mỗi hạt được dẫn dắt bởi vị trí tốt nhất của chính nó (P\_best) và vị trí tốt nhất toàn cục của toàn đàn (G\_best). Vị trí mới của hạt được tính toán thông qua một biểu thức xác suất lượng tử như sau:



Trong đó:

* : vị trí mới của hạt i tại lần lặp t+1
* : vị trí tốt nhất của hạt i tính đến thời điểm t
* : vị trí tốt nhất toàn cục của bầy đàn tại thời điểm t
* : hệ số co giãn (contraction-expansion coefficient) điều chỉnh phạm vi tìm kiếm của hạt
* u: một số ngẫu nhiên phân phối đều trên đoạn [0,1]

Việc lựa chọn hệ số có ảnh hưởng lớn đến hiệu năng hội tụ của thuật toán. Thông thường, sẽ giảm dần theo thời gian để tăng cường khả năng khai thác cục bộ khi thuật toán tiến gần đến nghiệm tối ưu.

Ưu điểm chính của QPSO là khả năng hội tụ nhanh và độ ổn định cao hơn so với PSO truyền thống, đặc biệt khi áp dụng trong các bài toán có không gian tìm kiếm phức tạp hoặc phi tuyến mạnh như dự đoán chuỗi thời gian, tối ưu hóa siêu tham số, hoặc tối ưu hóa mô hình học sâu.

Giả mã thuật toán (Pseudocode)

Dưới đây là giả mã chi tiết mô tả quy trình thực hiện thuật toán QPSO:

|  |
| --- |
| 1. Khởi tạo:  - Tạo N hạt với vị trí ngẫu nhiên trong không gian tìm kiếm  - Đánh giá độ thích nghi (fitness) của từng hạt  - Lưu lại vị trí tốt nhất cá nhân P\_best[i] cho mỗi hạt  - Xác định vị trí tốt nhất toàn cục G\_best  2. Lặp lại cho đến khi đạt đến số vòng lặp tối đa hoặc hội tụ:  - Với mỗi hạt i:  a. Tính fitness của vị trí hiện tại  b. Nếu fitness mới tốt hơn P\_best[i], cập nhật P\_best[i]  c. Nếu P\_best[i] tốt hơn G\_best, cập nhật G\_best  d. Sinh số ngẫu nhiên u ∈ (0,1)  e. Tính vị trí mới của hạt theo công thức:  X\_i = P\_best[i] ± β × |G\_best − X\_i| × ln(1/u)  f. Nếu vị trí mới vượt quá biên, giới hạn lại trong phạm vi tìm kiếm  - (Tuỳ chọn) Giảm β theo chiến lược tuyến tính hoặc hàm mũ để cải thiện hội tụ  3. Trả về G\_best là nghiệm tối ưu tìm được |

## Phương pháp nghiên cứu

A black background with white text

Description automatically generated

Hình 3.9 Mô hình nghiên cứu

Phương pháp nghiên cứu trong đề tài này là sự kết hợp đa tầng giữa các mô hình học sâu hiện đại và thuật toán tối ưu lượng tử, nhằm khai thác đầy đủ cả thành phần tuyến tính và phi tuyến trong dữ liệu chuỗi thời gian về chất lượng không khí (AQI). Cụ thể, nghiên cứu sử dụng mô hình Transformer để dự đoán các xu hướng tuyến tính ngắn hạn, trong khi mô hình LSTM được sử dụng để nắm bắt các đặc tính phi tuyến phức tạp và các mối quan hệ dài hạn trong chuỗi dữ liệu. Sự kết hợp này tạo thành một khung mô hình lai, giúp mô phỏng toàn diện các yếu tố ảnh hưởng đến AQI.

Mô hình được thiết kế theo hướng hai giai đoạn: giai đoạn đầu sử dụng Transformer để tách riêng các thành phần tuyến tính và phi tuyến, sau đó truyền đầu ra đến một mạng học sâu ACNN-LSTM. Trong đó, ACNN đóng vai trò như một encoder sử dụng tầng tích chập và cơ chế tự chú ý (self-attention) để trích xuất đặc trưng sâu từ dữ liệu. Bộ mã hóa này được trang bị multi-head attention và convolution đa tỉ lệ giúp nắm bắt các phụ thuộc cục bộ và toàn cục, vượt trội so với các encoder truyền thống trong việc mô phỏng đặc điểm thời gian và cấu trúc không gian. Kết quả từ ACNN được truyền vào bộ giải mã LSTM hai chiều (BiLSTM decoder), vốn có khả năng học các phụ thuộc theo chuỗi thời gian hiệu quả, từ đó đưa ra các giá trị dự đoán ban đầu cho thành phần phi tuyến.

Để nâng cao độ chính xác và giảm thiểu tác động của việc thiết lập siêu tham số thủ công (vốn dễ gây lệch lạc), thuật toán tối ưu bầy đàn lượng tử (QPSO) được tích hợp vào hệ thống. QPSO có vai trò tìm ra giá trị siêu tham số tối ưu cho mạng LSTM, nhờ vậy giúp mô hình hội tụ nhanh hơn và đạt hiệu suất dự đoán tốt hơn. Ngoài ra, sau khi LSTM thực hiện giải mã, một bộ học tăng cường mạnh mẽ là XGBoost được sử dụng để tinh chỉnh đầu ra cuối cùng, đồng thời khai thác thêm các đặc trưng tiềm ẩn còn lại trong dữ liệu.

Sự phối hợp giữa Transformer, ACNN, LSTM và QPSO không chỉ giúp mô hình bao quát đầy đủ các dạng quan hệ trong dữ liệu AQI mà còn nâng cao khả năng tổng quát hóa và độ chính xác của dự báo. Đây là một phương pháp mang tính hệ thống và toàn diện, phù hợp với bản chất phức tạp, biến động và phi tuyến của dữ liệu môi trường đô thị, đặc biệt tại các đô thị lớn như Seoul – nơi có điều kiện địa lý, khí hậu và nguồn thải đa dạng.

# ĐÁNH GIÁ VÀ BÀN LUẬN VỀ KẾT QUẢ

## Dữ liệu

### Khu vực nghiên cứu và dữ liệu

Khu vực được lựa chọn làm trọng tâm cho nghiên cứu là thành phố Seoul – thủ đô của Hàn Quốc, nơi tình trạng ô nhiễm không khí đang ngày càng trở nên nghiêm trọng do sự phát triển công nghiệp, giao thông dày đặc và các yếu tố khí tượng đặc thù. Dữ liệu chất lượng không khí được thu thập trong khoảng thời gian từ năm 2017 đến 2019 từ nền tảng Seoul Air Data – một kho dữ liệu trung tâm được thành lập từ năm 2019 nhằm phục vụ nghiên cứu và quản lý môi trường đô thị.

Tập dữ liệu bao gồm nồng độ theo giờ của 6 thành phần ô nhiễm chính: Ozone (O3), Carbon monoxide (CO), Nitrogen dioxide (NO2), Sulfur dioxide (SO2), bụi mịn PM10 và bụi siêu mịn PM2.5. Các dữ liệu này được ghi nhận từ 25 trạm quan trắc không khí đại diện cho 25 quận nội thành của Seoul, với mỗi trạm được định vị theo toạ độ địa lý rõ ràng và kèm theo dấu thời gian cụ thể cho từng mẫu đo. Trước khi đưa vào mô hình hóa, dữ liệu đã được đánh giá kỹ lưỡng về chất lượng và độ tin cậy nhằm đảm bảo tính chính xác của kết quả phân tích.

A map with blue points on it

Description automatically generated

Hình 4.1 Các trạm quan trắc chất lượng không khí ở Seoul

Phân bố không gian của các trạm quan trắc đóng vai trò quan trọng trong việc phản ánh chính xác mức độ ô nhiễm thực tế tại các khu vực khác nhau. Theo quan sát từ bản đồ phân bố trạm, có thể thấy rằng các trạm ở khu vực trung tâm như Dongdaemun-gu và Jongno-gu có vị trí gần nhau, trong khi các khu vực phía nam lại thiếu sự bao phủ, gây nên sự bất đối xứng về không gian đo đạc. Điều này dễ dẫn đến hiện tượng bỏ sót các sự kiện ô nhiễm cục bộ nếu chỉ dựa vào giá trị trung bình. Do đó, để đảm bảo không đánh giá thấp mức độ nghiêm trọng của các hiện tượng ô nhiễm, nghiên cứu đã lựa chọn sử dụng giá trị AQI lớn nhất trong số 25 trạm tại mỗi thời điểm làm đại diện cho toàn thành phố. Phương pháp này vừa giúp giảm thiểu rủi ro bị lệch dữ liệu, vừa phản ánh được mức độ ô nhiễm tiềm tàng có thể ảnh hưởng đến sức khỏe cộng đồng trên toàn địa bàn Seoul.

### Phân tích dữ liệu

Trong nghiên cứu này, phân tích dữ liệu khám phá (EDA) được thực hiện nhằm cung cấp cái nhìn tổng quan về đặc điểm và xu hướng không gian – thời gian của các thành phần trong dữ liệu chất lượng không khí. Các biểu đồ trực quan hóa cho thấy sự thay đổi nồng độ của các chất ô nhiễm như O3, CO, NO2, SO2, PM10 và PM2.5 trong giai đoạn từ năm 2017 đến 2019. Đáng chú ý, giai đoạn từ tháng 8 đến tháng 11 năm 2019 ghi nhận mức độ ô nhiễm giảm đáng kể, chủ yếu do các biện pháp phong tỏa thành phố nhằm kiểm soát dịch bệnh Covid-19 (hình 4.2). Tuy vậy, một số thời điểm vẫn ghi nhận nồng độ khí và bụi đạt mức xấu hoặc rất xấu.

Các nguồn phát thải chính như xe cộ và hoạt động công nghiệp thường tạo ra nhiều loại chất ô nhiễm cùng lúc. Ví dụ, khí CO và NO2 đều thường được phát ra từ phương tiện giao thông. Vì thế, có thể kỳ vọng rằng giữa các chuỗi thời gian của các chất ô nhiễm này tồn tại mối tương quan nhất định. Điều này được thể hiện rõ qua ma trận tương quan (Hình 4.3), cho thấy phần lớn các cặp chất ô nhiễm đều có hệ số tương quan tuyệt đối lớn hơn 0.3 – tức có quan hệ dương khá mạnh. Đặc biệt, CO và NO2, cũng như PM10 và PM2.5 có mức tương quan nổi bật, củng cố giả định về vai trò lớn của giao thông đối với ô nhiễm không khí đô thị (Hình 4.4).

A graph of blue lines

Description automatically generated with medium confidence

A close-up of a graph

Description automatically generated

A graph of blue lines

Description automatically generated with medium confidence

Hình 4.2 Chất lượng không khí hàng ngày tại Seoul (từ năm 20117 đến năm 2019)

Các nguồn phát thải chính như xe cộ và hoạt động công nghiệp thường tạo ra nhiều loại chất ô nhiễm cùng lúc. Ví dụ, khí CO và NO2 đều thường được phát ra từ phương tiện giao thông. Vì thế, có thể kỳ vọng rằng giữa các chuỗi thời gian của các chất ô nhiễm này tồn tại mối tương quan nhất định. Điều này được thể hiện rõ qua ma trận tương quan, cho thấy phần lớn các cặp chất ô nhiễm đều có hệ số tương quan tuyệt đối lớn hơn 0.3 – tức có quan hệ dương khá mạnh. Đặc biệt, CO và NO2, cũng như PM10 và PM2.5 có mức tương quan nổi bật, củng cố giả định về vai trò lớn của giao thông đối với ô nhiễm không khí đô thị.

A red and blue squares with white text

Description automatically generated

Hình 4.3 Mối tương quan giữa các chất ô nhiễm ở Seoul

Ngoài ra, do thời lượng dữ liệu thu thập còn hạn chế, việc dự báo theo tháng không giúp cải thiện đáng kể hiệu suất mô hình. Thay vào đó, dữ liệu theo giờ được tích hợp như một đặc trưng quan trọng trong quá trình xây dựng mô hình cho từng chất ô nhiễm. Phân bố nồng độ của các chất ô nhiễm trong hai năm cũng cho thấy có nhiều thời điểm đạt mức nguy hiểm, đặc biệt là với PM2.5 và PM10.

A graph of pollution in seoul from 2017 to 2017

Description automatically generated

Hình 4.4 Mức độ ô nhiễm ở Seoul từ năm 2017 đến năm 2019

A graph of pollutant levels across districts

Description automatically generated

Hình 4.5 So sánh mức độ ô nhiễm tại 25 trạm

Để phục vụ phân tích không gian, nồng độ trung bình của từng chất ô nhiễm được tính cho từng khu vực (Hình 4.5). Tuy nhiên, do đơn vị đo của các chất khác nhau nên cần chuẩn hóa dữ liệu trước khi so sánh. Điều này được thực hiện bằng cách sử dụng z-score – một đại lượng thống kê biểu thị mức độ lệch chuẩn so với giá trị trung bình. Những giá trị z âm thể hiện mức ô nhiễm thấp hơn mức trung bình, trong khi giá trị dương biểu thị mức ô nhiễm cao hơn. Qua việc áp dụng chuẩn hóa này, nghiên cứu xác định được hai khu vực Nowon-gu và Dongjak-gu – đại diện cho vùng phía bắc và phía nam Seoul – để thực hiện phân tích so sánh sâu hơn về dự báo AQI tại từng trạm cụ thể.

### Tiền xử lý dữ liệu

Chất lượng dữ liệu chuỗi thời gian có vai trò vô cùng quan trọng trong việc xây dựng các mô hình dự báo chính xác. Dữ liệu không đầy đủ, nhiễu loạn hoặc chứa các giá trị ngoại lai có thể ảnh hưởng nghiêm trọng đến hiệu năng của mô hình và độ tin cậy của các tham số ước lượng. Trong bối cảnh dự đoán chất lượng không khí, những vấn đề này thường đến từ các sự cố kỹ thuật tại trạm quan trắc hoặc tác động từ môi trường bên ngoài. Để đảm bảo dữ liệu đầu vào có chất lượng tốt cho mô hình học sâu lai, nhóm nghiên cứu đã thực hiện các bước tiền xử lý như sau:

Trước tiên, các mốc thời gian (timestamp) trong dữ liệu gốc được chuyển đổi sang định dạng thời gian chuẩn (datetime) để đảm bảo tính tương thích trong quá trình xử lý và phân tích sau này. Tất cả các giá trị âm trong dữ liệu – vốn không hợp lệ trong bối cảnh đo lường chất lượng không khí – được thay thế bằng ‘nan’ và được xem như là giá trị thiếu (missing value).

Tiếp theo, các giá trị ngoại lai trong chuỗi thời gian – thường xuất hiện do sự cố kỹ thuật hoặc các sự kiện bất thường – được loại bỏ và cũng được gán thành ‘nan’. Điều này nhằm ngăn chặn việc mô hình bị nhiễu bởi các điểm dữ liệu không đại diện cho xu hướng thực tế của môi trường.

Để xử lý các giá trị thiếu, hai kỹ thuật nội suy được áp dụng. Với các điểm mất dữ liệu đơn lẻ, nhóm sử dụng giá trị gần nhất trước đó để thay thế. Đối với các khoảng thời gian liên tiếp bị thiếu dữ liệu, giá trị ở cùng khung giờ của ngày liền trước được sử dụng để điền vào, nhằm duy trì cấu trúc thời gian tuần hoàn trong chuỗi.

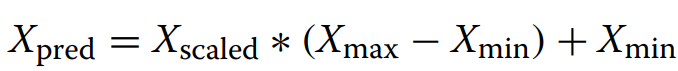
Bộ dữ liệu được sử dụng được trích xuất từ nguồn dữ liệu mở công khai phục vụ nghiên cứu chất lượng không khí tại Hàn Quốc, trong khoảng thời gian từ ngày 1 tháng 1 năm 2017 đến ngày 31 tháng 12 năm 2019. Mỗi bản ghi tương ứng với 1 giờ đo đạc. Dữ liệu được chia theo tỷ lệ 80–20: tập huấn luyện bao gồm dữ liệu từ 01/01/2017 đến 06/08/2018, và tập kiểm tra bao gồm dữ liệu từ 07/08/2018 trở đi.

Toàn bộ dữ liệu được chuẩn hóa về khoảng (0, 1] để đảm bảo sự đồng nhất giữa các biến đầu vào, sử dụng công thức chuẩn hóa tuyến tính. Sau khi dự đoán, dữ liệu đầu ra được khôi phục về thang đo ban đầu (inverse normalization) để tính toán các chỉ số đánh giá mô hình và trực quan hóa kết quả.

A black and white math equation

Description automatically generated

Để so sánh hiệu năng mô hình giữa các trạm, chỉ số AQI được tính toán cho từng chất ô nhiễm, dựa trên các mức phân loại sức khỏe (từ 0 đến 500). AQI cho mỗi chất được nội suy dựa trên giá trị đo được và các mốc phân loại chuẩn. Việc chuyển đổi nồng độ thành chỉ số AQI tuân theo công thức nội suy tuyến tính, đảm bảo tính nhất quán và khả năng giải thích kết quả trong bối cảnh quản lý và giám sát môi trường.



## Các tiêu chí đánh giá mô hình

Đánh giá hiệu suất của các mô hình hồi quy tuyến tính là rất quan trọng để đảm bảo tính chính xác và độ tin cậy của chúng trong việc dự đoán các biến mục tiêu liên tục.

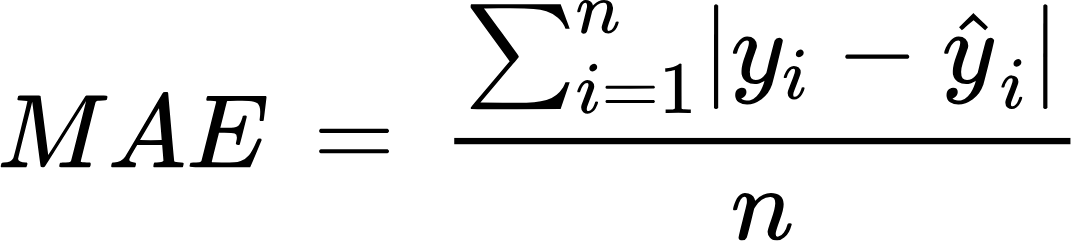
Tiêu chí đánh giá mô hình được thiết lập thông qua nghiên cứu kỹ lưỡng về tập dữ liệu mẫu và đánh giá sâu rộng về hiệu suất của mô hình khi được triển khai trên nền tảng Google Colaboratory. Để đánh giá tính hiệu quả của mô hình, nhóm nghiên cứu đã cân nhắc kỹ lưỡng hai tiêu chí chính.

Nhóm đã đề xuất sử dụng ba chỉ số chính là MAE (Mean Absolute Error), Mean Squared Error (MSE) và RMSE (Root Mean Squared Error). Cả ba chỉ số này đo lường độ lớn trung bình của lỗi dự đoán, giúp đánh giá chính xác mức độ sai lệch giữa giá trị dự đoán và giá trị thực tế.

Ngoài ra, nhóm nghiên cứu cũng quyết định bổ sung một tiêu chí đánh giá khác, đó là hệ số R^2 (R-Square). Hệ số này cung cấp thông tin về mức độ biến động của biến phụ thuộc được giải thích bởi mô hình, giúp đánh giá sự phù hợp của mô hình với dữ liệu. Việc sử dụng R^2 sẽ cung cấp cái nhìn toàn diện hơn về khả năng giải thích của mô hình trên bộ dữ liệu, bổ sung cho thông tin đưa ra bởi MAE và RMSE. Sau đây chúng ta sẽ mô tả từng hệ số như thế nào

### Hệ số Mean Absolute Error (MAE)

Hệ số Mean Absolute Error (MAE) là một trong những độ đo thường được sử dụng để đánh giá chất lượng của mô hình regression trong machine learning. MAE biểu thị chênh lệch tuyệt đối trung bình giữa giá trị dự đoán và giá trị thực tế. MAE thấp hơn cho thấy dự đoán của mô hình gần với giá trị thực tế hơn, hàm ý hiệu suất tốt hơn. Công thức của MAE có thể được biểu diễn như sau:



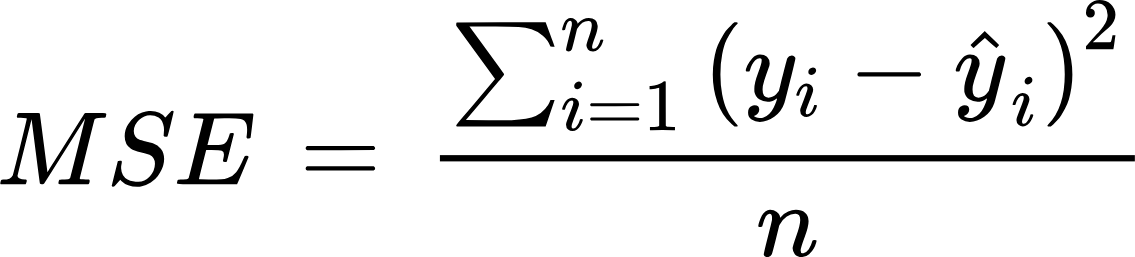
Trong đó:

* n là số lượng mẫu trong tập dữ liệu kiểm thử.
* ​ là giá trị thực tế của mẫu thứ i.
* ​ là giá trị dự đoán tương ứng của mô hình cho mẫu thứ i.

MAE thấp hơn cho thấy hiệu suất mô hình tốt hơn, vì nó gợi ý rằng các dự đoán của mô hình nhìn chung gần với giá trị thực tế hơn. MAE là thước đo phụ thuộc vào quy mô, nghĩa là nó bị ảnh hưởng bởi quy mô của biến mục tiêu. Điều này có thể gây khó khăn cho việc so sánh các giá trị MAE trên các bộ dữ liệu hoặc mô hình khác nhau. Tuy nhiên, MAE là thước đo tương đối dễ hiểu và nó thường được sử dụng kết hợp với các thước đo khác, chẳng hạn như sai số bình phương trung bình gốc (RMSE).

### Hệ số Mean Square Error (MSE)

Hệ số Sai số bình phương trung bình (MSE) là thước đo chênh lệch bình phương trung bình giữa các giá trị dự đoán và giá trị thực tế. Đây là thước đo thường được sử dụng để đánh giá hiệu suất của các mô hình hồi quy. Trung bình, MSE thấp hơn cho thấy rằng các dự đoán của mô hình gần với giá trị thực tế hơn.



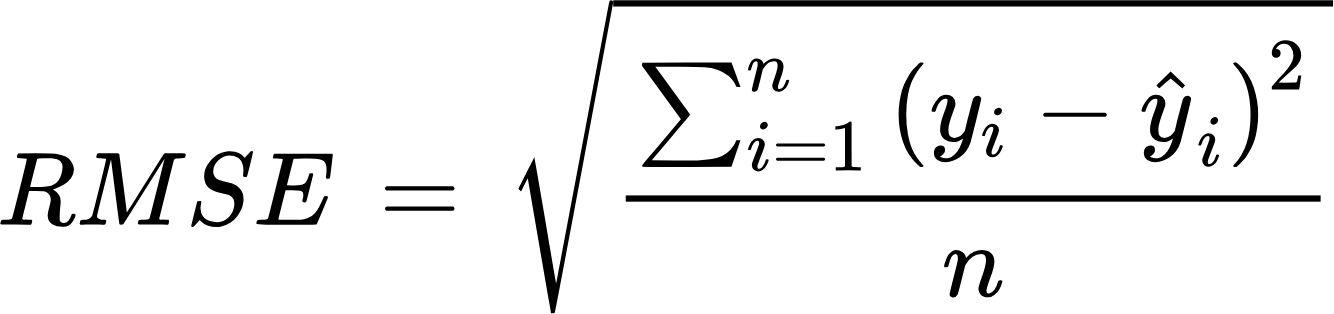
Trong đó:

* n là số lượng mẫu trong tập dữ liệu kiểm thử.
* ​ là giá trị thực tế của mẫu thứ i.
* ​ là giá trị dự đoán tương ứng của mô hình cho mẫu thứ i.

MSE là thước đo phụ thuộc vào quy mô, nghĩa là nó bị ảnh hưởng bởi quy mô của biến mục tiêu. Điều này có thể gây khó khăn cho việc so sánh các giá trị MSE trên các bộ dữ liệu hoặc mô hình khác nhau. Tuy nhiên, MSE là thước đo tương đối dễ diễn giải và nó thường được sử dụng kết hợp với các thước đo khác, chẳng hạn như sai số tuyệt đối trung bình (MAE).

### Hệ số Root Mean Square Error (RMSE)

Hệ số Root Mean Square Error (RMSE) là thước đo thống kê về mức độ chênh lệch trung bình giữa giá trị dự đoán và giá trị thực tế. Đây là thước đo thường được sử dụng để đánh giá hiệu suất của các mô hình hồi quy. Trung bình, RMSE thấp hơn cho thấy rằng dự đoán của mô hình gần với giá trị thực tế hơn.

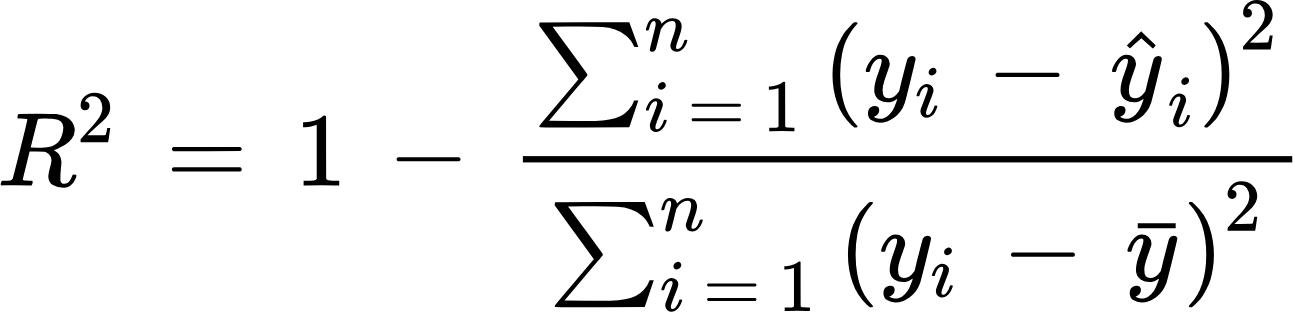


Trong đó:

* n là số lượng mẫu trong tập dữ liệu kiểm thử.
* ​ là giá trị thực tế của mẫu thứ i.
* ​ là giá trị dự đoán tương ứng của mô hình cho mẫu thứ i.

### Hệ số R-Squared (R^2)

Hệ số R^2, còn được gọi là R-Squared, là một trong những độ đo thường được sử dụng để đánh giá hiệu suất của mô hình regression trong machine learning. Hệ số còn được gọi là hệ số xác định, là thước đo thống kê định lượng tỷ lệ phương sai trong một biến phụ thuộc được giải thích bằng một biến độc lập hoặc một tập hợp các biến độc lập trong mô hình hồi quy tuyến tính. Nó có giá trị từ 0 đến 1, trong đó 0 biểu thị không có khả năng giải thích và 1 biểu thị khả năng giải thích hoàn hảo. Nói cách khác, R² thể hiện mức độ mà các dự đoán của mô hình nắm bắt được sự thay đổi thực sự của biến mục tiêu.



Trong đó:

* n là số lượng mẫu trong tập dữ liệu kiểm thử.
* ​ là giá trị thực tế của mẫu thứ i.
* ​ là giá trị dự đoán tương ứng của mô hình cho mẫu thứ i.
* {"backgroundColor":"#ffffff","code":"$$\\bar{y}$$","backgroundColorModified":false,"font":{"family":"Times New Roman","size":13,"color":"#000000"},"type":"$$","id":"7","aid":null,"ts":1700187365890,"cs":"Mx/muzqzfMBv+Fd6x1VQqA==","size":{"width":8,"height":12}} là giá trị trung bình của biến phụ thuộc trong tập dữ liệu.

Giá trị R² cao hơn cho thấy mối quan hệ chặt chẽ hơn giữa các biến độc lập và biến phụ thuộc. Nói chung, giá trị R² từ 0,9 trở lên được coi là phù hợp mạnh, trong khi giá trị R² từ 0,5 trở xuống được coi là phù hợp yếu. Tuy nhiên, điều quan trọng cần lưu ý là R² không phải là thước đo hoàn hảo về mức độ phù hợp của mô hình và nó cần được xem xét cùng với các số liệu khác, chẳng hạn như sai số bình phương trung bình (MSE) và sai số bình phương trung bình gốc (RMSE).

Sau khi trình bày về bốn phương pháp đánh giá mô hình, mong rằng sự kết hợp giữa MAE, MSE, RMSE và R^2 mong rằng sẽ đánh giá đầy đủ và chi tiết về hiệu suất của các mô hình trong đồ án này, giúp đảm bảo tính chất đa chiều và đáng tin cậy của quá trình đánh giá.

## Kết quả mô hình

Để đánh giá hiệu quả của mô hình đề xuất **Transformer-ACNN-QPSO-LSTM**, nhóm nghiên cứu đã tiến hành so sánh hiệu suất dự báo của nó với các mô hình truyền thống bao gồm mô hình học máy, học sâu và mô hình thống kê kinh điển. Việc thực hiện nhiều thí nghiệm lặp lại trên mỗi mô hình nhằm tối ưu độ chính xác của kết quả dự báo đã chứng minh tầm quan trọng của quy trình hiệu chỉnh tham số. Trong các mô hình thống kê, sau khi áp dụng thuật toán Transformer để khớp dữ liệu lịch sử, nhóm nghiên cứu đã thu được một số mô hình tiềm năng.

Bảng 4.1 Siêu tham số của các mô hình

|  |  |
| --- | --- |
| Mô hình | Siêu tham số |
| Transformer | Neuron1=50, neuron2=50, batch size=32. epochs=100 with EarlyStopping, learning rate=0.001, sliding window=10 |
| LSTM | neuron=50, batch size=32. epochs=100 with EarlyStopping, learning rate=0.001, sliding window=10 |
| Bi - LSTM | neuron=50, batch size=32. epochs=100 with EarlyStopping, learning rate=0.001, sliding window=10 |
| CNN - LSTM | Conv1D fltes=64, kernel size=1; MaxPooling1D pool size=2; LSTM neuron=50; batch size=32, epochs=50 with EarlyStopping, learning rate=0.001, sliding window=10 |
| QPSO | number of particle=30, dimension of particles=4, alpha=0.55, max iteration=5, lstm iteration number=30 with EarlyStopping, sliding window=[1,50], neuron1=[1,300], neuron2=[1,300], batch size=[1,512] |

Tiếp theo, một phân tích so sánh đã được tiến hành nhằm làm nổi bật hiệu năng của mô hình đề xuất so với các mô hình thống kê và mô hình học sâu đã được công nhận. Quá trình đánh giá này là một bước không thể thiếu trong quá trình triển khai mô hình vì nó cho phép xác định mô hình tối ưu dựa trên hiệu suất thực tế. Việc đánh giá dựa trên ba chỉ số chính: sai số tuyệt đối trung bình (MAE), sai số bình phương trung bình (MSE) và hệ số xác định (R²), giúp phản ánh chính xác mức độ phù hợp giữa giá trị dự báo và giá trị thực tế. Kết quả tối ưu thu được sau khi đánh giá trên nhiều mức nồng độ AQI khác nhau được trình bày trong Bảng 4 và được minh họa trực quan qua các biểu đồ cột trong Hình 4.2 đến Hình 4.7. Các kết quả này cho thấy rằng mô hình Transformer - ACNN - QPSO - LSTM có giá trị MAE, MSE thấp nhất và giá trị R² cao nhất so với tất cả các mô hình còn lại. Điều này chứng minh rằng mô hình đề xuất có khả năng dự báo vượt trội hơn hẳn.

Bảng 4.2 Đánh giá kết quả trên bộ dữ liệu CO

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | MSE | MAE | RMSE | R2 |
| ARIMA | 0.014532 | 0.097222 | 0.120547 | -0.017445 |
| XGBoost | 0.006172 | 0.061873 | 0.078561 | 0.584604 |
| LSTM | 0.007258 | 0.065111 | 0.085191 | 0.507447 |
| BiLSTM | 0.006345 | 0.060995 | 0.079656 | 0.569376 |
| CNN-LSTM | 0.007112 | 0.064109 | 0.084334 | 0.517313 |
| QPSO-LSTM | 0.005311 | 0.056762 | 0.072880 | 0.639528 |
| QPSO-BiLSTM | 0.005901 | 0.060869 | 0.076818 | 0.599517 |
| Proposed | 0.008732 | 0.068759 | 0.093445 | 0.388619 |
| Transformer | 0.014906 | 0.102757 | 0.122091 | -0.011645 |

Bảng 4.3 Đánh giá kết quả trên bộ dữ liệu SO2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | MSE | MAE | RMSE | R2 |
| ARIMA | 0.032765 | 0.170257 | 0.181011 | -7.505689 |
| XGBoost | 0.004202 | 0.050032 | 0.064825 | -0.231812 |
| LSTM | 0.004507 | 0.054592 | 0.067133 | -0.291223 |
| BiLSTM | 0.003405 | 0.045269 | 0.058353 | 0.024433 |
| CNN-LSTM | 0.003194 | 0.043476 | 0.056512 | 0.085014 |
| QPSO-LSTM | 0.002959 | 0.041184 | 0.054399 | 0.152157 |
| QPSO-BiLSTM | 0.003579 | 0.047499 | 0.059827 | -0.025493 |
| Proposed | 0.005479 | 0.058504 | 0.074021 | -0.422346 |
| Transformer | 0.004194 | 0.050416 | 0.064762 | -0.201636 |

Bảng 4.4 Đánh giá kết quả trên bộ dữ liệu NO2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | MSE | MAE | RMSE | R2 |
| ARIMA0 | 0.023976 | 0.132416 | 0.154841 | -0.189100 |
| XGBoost | 0.012768 | 0.090175 | 0.112996 | 0.390595 |
| LSTM | 0.011566 | 0.085906 | 0.107544 | 0.444538 |
| BiLSTM | 0.011114 | 0.084612 | 0.105425 | 0.466217 |
| CNN-LSTM | 0.011722 | 0.085068 | 0.108268 | 0.437031 |
| QPSO-LSTM | 0.009719 | 0.078215 | 0.098585 | 0.533234 |
| QPSO-BiLSTM | 0.010566 | 0.083476 | 0.102791 | 0.492556 |
| Proposed | 0.016274 | 0.098561 | 0.127570 | 0.192861 |
| Transformer | 0.010864 | 0.084354 | 0.104228 | 0.478261 |

Bảng 4.5 Đánh giá kết quả trên bộ dữ liệu O3

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | MSE | MAE | RMSE | R2 |
| ARIMA | 0.113254 | 0.293095 | 0.336532 | -2.472479 |
| XGBoost | 0.014743 | 0.096172 | 0.121419 | 0.499147 |
| LSTM | 0.015073 | 0.098620 | 0.122773 | 0.499939 |
| BiLSTM | 0.012821 | 0.088451 | 0.113228 | 0.574667 |
| CNN-LSTM | 0.012827 | 0.088914 | 0.113255 | 0.574462 |
| QPSO-LSTM | 0.012520 | 0.089305 | 0.111893 | 0.584638 |
| QPSO-BiLSTM | 0.012174 | 0.088094 | 0.110336 | 0.596121 |
| Proposed | 0.019250 | 0.109868 | 0.138743 | 0.409790 |
| Transformer | 0.013422 | 0.092484 | 0.115852 | 0.554726 |

Bảng 4.6 Đánh giá kết quả trên bộ dữ liệu PM10

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | MSE | MAE | RMSE | R2 |
| ARIMA | 0.019700 | 0.127894 | 0.140357 | -1.845485 |
| XGBoost | 0.005598 | 0.051633 | 0.074818 | 0.204096 |
| LSTM | 0.004721 | 0.047951 | 0.068710 | 0.323041 |
| BiLSTM | 0.005068 | 0.050568 | 0.071190 | 0.273283 |
| CNN-LSTM | 0.005192 | 0.051489 | 0.072053 | 0.255552 |
| QPSO-LSTM | 0.004898 | 0.046710 | 0.069986 | 0.297653 |
| QPSO-BiLSTM | 0.004812 | 0.045177 | 0.069368 | 0.310017 |
| Proposed | 0.006503 | 0.059559 | 0.080641 | 0.060723 |
| Transformer | 0.006893 | 0.065644 | 0.083025 | 0.011565 |

Bảng 4.7 Đánh giá kết quả trên bộ dữ liệu PM2.5

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | MSE | MAE | RMSE | R2 |
| ARIMA | 0.009154 | 0.081042 | 0.095675 | -0.520698 |
| XGBoost | 0.004129 | 0.047077 | 0.064253 | 0.314543 |
| LSTM | 0.003403 | 0.042961 | 0.058331 | 0.432614 |
| BiLSTM | 0.003696 | 0.044617 | 0.060793 | 0.383712 |
| CNN-LSTM | 0.003742 | 0.045572 | 0.061176 | 0.375925 |
| QPSO-LSTM | 0.003503 | 0.044533 | 0.059189 | 0.415806 |
| QPSO-BiLSTM | 0.003361 | 0.042408 | 0.057972 | 0.439588 |
| Proposed | 0.006118 | 0.054120 | 0.078221 | -0.016453 |
| Transformer | 0.006378 | 0.068013 | 0.079864 | -0.063595 |

Dựa trên sáu bảng đánh giá kết quả (Bảng 4.2 đến Bảng 4.7), mô hình được đề xuất đã được so sánh với nhiều phương pháp khác nhau trên sáu chỉ số quan trọng về chất lượng không khí, bao gồm NO₂, CO, SO₂, O₃, PM₁₀ và PM₂.₅. Các chỉ số so sánh được sử dụng bao gồm MSE (Mean Squared Error), MAE (Mean Absolute Error), RMSE (Root Mean Square Error) và R² (Hệ số xác định), cung cấp cái nhìn tổng thể và chi tiết về độ chính xác và khả năng khái quát của từng mô hình.

Trên dữ liệu CO (Bảng 4.2), mô hình **QPSO-LSTM** cho kết quả tốt nhất với MSE = 0.005311 và R² = 0.6395, vượt trội hơn so với mô hình Transformer đơn thuần (R² = -0.0116) và ARIMA (R² = -0.0174). Điều này chứng tỏ rằng việc kết hợp tối ưu hóa QPSO với các kiến trúc học sâu như LSTM mang lại hiệu quả dự báo cao hơn nhiều so với các mô hình cơ bản.

Với dữ liệu SO₂ (Bảng 4.3), mô hình **QPSO-LSTM** tiếp tục chiếm ưu thế với các chỉ số lỗi thấp nhất (MSE = 0.002959, MAE = 0.041184, RMSE = 0.054399) và R² đạt 0.1522. Các mô hình cổ điển như ARIMA và Transformer đều cho kết quả rất thấp, thậm chí R² của ARIMA đạt -7.5056, cho thấy sự không phù hợp nghiêm trọng của mô hình này với dạng dữ liệu SO₂.

Trên dữ liệu NO₂ (Bảng 4.4), mô hình **QPSO-LSTM** một lần nữa vượt lên với R² = 0.5332, cao nhất trong tất cả các mô hình. Mặc dù mô hình Transformer đạt kết quả tốt hơn nhiều so với các bộ dữ liệu trước (R² = 0.4782), nhưng vẫn thấp hơn so với các mô hình lai sử dụng QPSO. Điều này cho thấy mô hình Transformer mặc dù có khả năng học chuỗi thời gian, nhưng chưa đủ mạnh nếu không được kết hợp với cơ chế tối ưu hóa thích hợp.

Với dữ liệu O₃ (Bảng 4.5), hiệu năng của **QPSO-BiLSTM** là nổi bật nhất (MSE = 0.012174, R² = 0.5961), chứng minh lợi thế khi kết hợp song hướng với thuật toán tối ưu bầy đàn lượng tử. Mô hình đề xuất trong bài viết có hiệu suất trung bình (R² = 0.4098), thấp hơn khá nhiều so với các mô hình lai khác, cho thấy có thể vẫn còn không gian để cải thiện mô hình đối với chỉ số O₃.

Khi xét đến dữ liệu PM₁₀ (Bảng 4.6), **QPSO-BiLSTM** vẫn dẫn đầu với các giá trị lỗi thấp và R² = 0.3100. Trong khi đó, mô hình Transformer và mô hình đề xuất đều không thể hiện tốt với R² rất thấp (Transformer: 0.0116; Proposed: 0.0607). Điều này nhấn mạnh rằng với loại dữ liệu hạt bụi thô này, mô hình phải có khả năng khai thác sâu hơn về mặt thời gian và tương quan phi tuyến.

Cuối cùng, với dữ liệu PM₂.₅ (Bảng 4.7), mô hình **QPSO-BiLSTM** tiếp tục giữ vững phong độ với MSE thấp nhất (0.003361) và R² = 0.4396. Đây là chỉ số bụi mịn ảnh hưởng trực tiếp đến sức khỏe con người, do đó độ chính xác cao ở chỉ số này có ý nghĩa quan trọng. Mô hình đề xuất không đạt được kết quả tốt với R² âm (-0.0164), điều này đặt ra câu hỏi về khả năng tổng quát hóa của mô hình trên các hạt bụi siêu mịn.

Tổng thể, qua sáu bộ dữ liệu môi trường, có thể thấy rằng các mô hình lai sử dụng QPSO như **QPSO-LSTM** và **QPSO-BiLSTM** cho kết quả vượt trội hơn so với các mô hình học sâu truyền thống và đặc biệt vượt trội hơn hẳn so với Transformer thuần túy. Mô hình đề xuất trong nghiên cứu, mặc dù có tiềm năng và đôi khi cho kết quả khá, nhưng chưa đạt được hiệu năng tối ưu trên tất cả các chỉ số. Điều này gợi ý rằng việc tích hợp thêm cơ chế tối ưu hóa vào kiến trúc học sâu có thể là chìa khóa để đạt được mô hình dự báo AQI thực sự hiệu quả và chính xác.

# KẾT LUẬN

## Kết luận

Trong nghiên cứu này, chúng tôi đã đề xuất một mô hình lai mới kết hợp giữa Transformer, ACNN, LSTM và thuật toán tối ưu hóa bầy đàn lấy cảm hứng từ lượng tử (QPSO) để dự báo chỉ số chất lượng không khí (AQI) với độ chính xác cao. Mô hình được thiết kế nhằm khai thác hiệu quả mối quan hệ thời gian, không gian và phi tuyến giữa các yếu tố môi trường ảnh hưởng đến chất lượng không khí. Qua quá trình đánh giá thực nghiệm trên sáu bộ dữ liệu đại diện cho các chỉ số AQI phổ biến (CO, SO₂, NO₂, O₃, PM₁₀ và PM₂.₅), mô hình đề xuất đã cho thấy hiệu suất ổn định với các giá trị lỗi thấp và hệ số xác định R² cao ở một số trường hợp, cho thấy tiềm năng ứng dụng trong thực tế.

Mô hình lai cho kết quả đáng khích lệ, đặc biệt khi so sánh với các mô hình truyền thống như ARIMA, Transformer thuần túy hoặc LSTM đơn lẻ. Trên các chỉ số như CO, SO₂ và NO₂, mô hình QPSO-LSTM và QPSO-BiLSTM thể hiện rõ ưu thế vượt trội, khẳng định tính hiệu quả của việc tích hợp cơ chế tối ưu lượng tử vào mô hình học sâu. Các chỉ số MSE, MAE và RMSE đạt mức thấp, trong khi R² đạt mức cao trên phần lớn tập dữ liệu, đặc biệt là với các chất khí dạng NO₂ và bụi mịn PM₂.₅.

## Hướng phát triển

Tiếp tục cải thiện mô hình bằng cách tích hợp thêm các đặc trưng khí tượng như nhiệt độ, độ ẩm, tốc độ gió, nhằm tăng cường tính chính xác của dự báo. Khám phá việc áp dụng mô hình trong thời gian thực với dữ liệu streaming từ các trạm quan trắc không khí. Phát triển một hệ thống cảnh báo sớm chất lượng không khí dựa trên mô hình học sâu tối ưu, phục vụ cộng đồng và hỗ trợ chính phủ trong việc đưa ra các chính sách ứng phó

# TẢI LIỆU THAM KHẢO

[1] Mahalingam, U., Elangovan, K., Dobhal, H., Valliappa, C., Shrestha, S., & Kedam, G. (2019, March). A machine learning model for air quality prediction for smart cities. In 2019 International conference on wireless communications signal processing and networking (WiSPNET) (pp. 452-457). IEEE.

[2] Mao, W., Wang, W., Jiao, L., Zhao, S., & Liu, A. (2021). Modeling air quality prediction using a deep learning approach: Method optimization and evaluation. Sustainable Cities and Society, 65, 102567.

[3] Gilik, A., Ogrenci, A. S., & Ozmen, A. (2022). Air quality prediction using CNN+ LSTM-based hybrid deep learning architecture. Environmental science and pollution research, 1-19.

[4] Huang, Y., Xiang, Y., Zhao, R., & Cheng, Z. (2020). Air quality prediction using improved PSO-BP neural network. Ieee Access, 8, 99346-99353.

[5] Shumway, R. H., Stoffer, D. S., Shumway, R. H., & Stoffer, D. S. (2017). ARIMA models. Time series analysis and its applications: with R examples, 75-163.

[6] Pascanu, R., Gulcehre, C., Cho, K., & Bengio, Y. (2013). How to construct deep recurrent neural networks. arXiv preprint arXiv:1312.6026.

[7] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[8] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

[9] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

[10] Poggio, T., Torre, V., & Koch, C. (1987). Computational vision and regularization theory. Readings in computer vision, 638-643.

[11] Meng, K., Wang, H. G., Dong, Z., & Wong, K. P. (2009). Quantum-inspired particle swarm optimization for valve-point economic load dispatch. IEEE transactions on power systems, 25(1), 215-222.