PHÍ ĐÌNH TÚ ANH

HỆ THỐNG THÔNG TIN

2024

BỘ CÔNG THƯƠNG

TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI

PHÍ ĐÌNH TÚ ANH

DỰ BÁO CHỈ SỐ CHẤT LƯỢNG KHÔNG KHÍ TẠI NHA TRANG SỬ DỤNG MÔ HÌNH MẠNG HỌC SÂU LSTM

ĐỀ ÁN TỐT NGHIỆP THẠC SĨ NGÀNH HỆ THỐNG THÔNG TIN

Hà Nội – 2024

BỘ CÔNG THƯƠNG
TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI

PHÍ ĐÌNH TÚ ANH

DỰ BÁO CHỈ SỐ CHẤT LƯỢNG KHÔNG KHÍ TẠI NHA TRANG SỬ DỤNG MÔ HÌNH MẠNG HỌC SÂU LSTM

Ngành: Hệ thống thông tin

Mã số: 8480104

ĐỀ ÁN TỐT NGHIỆP THẠC SĨ NGÀNH HỆ THỐNG THÔNG TIN

NGƯỜI HƯỚNG DẪN: TS. NGUYỄN MẠNH CƯỜNG

Hà Nội – 2024

LỜI CAM ĐOAN

Em xin cam đoan đề án tốt nghiệp thạc sĩ ngành Hệ thống thông tin với đề tài "**Dự báo chỉ số chất lượng không khí tại nha trang sử dụng mô hình học sâu LSTM**" là do em nghiên cứu, tổng hợp và thực hiện dưới sự hướng dẫn của TS. Nguyễn Mạnh Cường.

Toàn bộ nội dung của đề án, những điều được trình bày là của chính cá nhân em hoặc là được tham khảo, tổng hợp từ nhiều nguồn tại liệu khác nhau. Tất cả các tài liệu tham khảo, tổng hợp đều được trích xuất nguồn gốc rỏ ràng. Các số liệu, kết quả được nêu trong luận văn là trung thực và được nghiên cứu thực nghiệm.

Học viên

Phí Đình Tú Anh

LỜI CẢM ƠN

Trước hết, em xin bày tỏ tình cảm và lòng biết ơn của em tới Thầy TS. Nguyễn Mạnh Cường. Người đã từng bước hướng dẫn, giúp đỡ em trong quá trình thực hiện đề án tốt nghiệp của mình.

Em xin chân thành cảm ơn Thầy Cô của Khoa Công nghệ thông tin và Khoa đào tạo sau Đại học Trường Đại học Công nghiệp Hà Nội đã dìu dắt, dạy dỗ em cả về kiến thức chuyên môn và tinh thần học tập để em có được những kiến thức thực hiện đồ án tốt nghiệp của mình.

Em xin chân thành cảm ơn Lãnh đạo Nhà trường, các phòng ban và quý Thầy Cô đã giúp đỡ tạo điều kiện tốt nhất cho em trong suốt thời gian học tập tại trường.

Tuy có nhiều cố gắng trong quá trình học tập, cũng như trong quá trình làm đề án tốt nghiệp không thể tránh khỏi những thiếu sót, em rất mong được sự góp ý quý báu của tất cả các thầy cô giáo cũng như tất cả các anh chị để kết quả của em được hoàn thiện hơn.

Em xin trân trọng cảm ơn!

Học viên

Phí Đình Tú Anh

DANH MỤC CÁC KÝ HIỆU VÀ TỪ VIẾT TẮT

Viết tắt	Tiếng Anh	Tiếng Việt
AQI	Air Quality Index	Chỉ số chất lượng không khí
LSTM	Long Short-Term Memory	Mạng bộ nhớ ngắn-dài hạn
GRU	Gated Recurrent Unit	Bộ nhớ hồi tiếp có cổng
ANN	Artificial Neural Network	Mạng nơ-ron nhân tạo
RNN	Recurrent Neural Network	Mạng nơ-ron hồi quy
CNN	Convolutional Neural Network	Mạng nơ-ron tích chập
RF	Random Forest	Thuật toán rừng ngẫu nhiên
SVM	Support Vector Machine	Thuật toán máy vector hỗ trợ
KNN	K-Nearest Neighbors	Thuật toán K-láng giềng gần nhất
RMSE	Root Mean Square Error	Căn bậc hai sai số bình phương trung bình
MSE	Mean Square Error	Sai số bình phương trung bình
MAE	Mean Absolute Error	Sai số trung bình tuyệt đối
MAPE	Mean Absoulute Percentage Error	Phần trăm sai số trung bình tuyệt đối
GDP	Gross Domestic Product	Tổng sản phẩm quốc nội
USEPA	United States Environmental Protection Agency	Cơ quan Bảo vệ Môi trường Hoa Kỳ
HW	Holt-Winters	Mô hình Holt-Winters

ARIMA	AutoRegressive Integrated Moving Average	Mô hình trung bình trượt tích hợp tự hồi quy
SARIMA	Seasonal AutoRegressive Integrated Moving Average	Mô hình ARIMA có yếu tố mùa vụ
GLM	Generalized Linear Model	Mô hình tuyến tính tổng quát
LR	Linear Regression	Hồi quy tuyến tính
RBF	Radial Basis Function	Hàm cơ sở xuyên tâm
XGBoost	Extreme Gradient Boosting	Mô hình tăng cường độ dốc cực đại
FNN	Feedforward Neural Network	Mạng nơ-ron truyền thẳng
ReLU	Rectified Linear Unit	Hàm kích hoạt ReLU
SGD	Stochastic Gradient Descent	Thuật toán tối ưu giảm dần độ dốc ngẫu nhiên
BRNN	Bidirectional Recurrent Neural Network	Mạng nơ-ron hồi tiếp hai chiều
NLP	Natural Language Processing	Xử lý ngôn ngữ tự nhiên
GNMT	Google Neural Machine Translation	Hệ thống dịch máy của Google
API	Application Programming Interface	Giao diện lập trình ứng dụng
HTTP	HyperText Transfer Protocol	Giao thức truyền tải siêu văn bản
JSON	JavaScript Object Notation	Định dạng dữ liệu dựa trên cặp key-value
URL	Uniform Resource Locator	Địa chỉ website

DANH MỤC CÁC BẢNG

DANH MỤC HÌNH VẼ

MỤC LỤC

MỞ ĐẦU

1. Giới thiệu

Dự báo chất lượng không khí là một vấn đề cấp bách và quan trọng, đặc biệt tại các địa điểm đang phát triển nhanh như Thành phố Nha Trang. Chất lượng không khí không chỉ ảnh hưởng trực tiếp đến sức khỏe con người mà còn tác động tiêu cực đến môi trường tự nhiên và kinh tế xã hội. Trong bối cảnh đó, việc ứng dụng công nghệ hiện đại để dự báo và kiểm soát chất lượng không khí trở nên thiết yếu. Đề án tốt nghiệp này sẽ tập trung vào việc sử dụng mô hình mạng học sâu Long Short-Term Memory để dự báo chỉ số chất lượng không khí tại Nha Trang. Mô hình mạng học sâu LSTM có khả năng ghi nhớ và học từ dữ liệu chuỗi thời gian, được kỳ vọng sẽ cung cấp các dự báo chính xác và hữu ích. Ứng dụng mô hình sẽ giúp các cơ quan chức năng và địa phương có thể đưa ra những biện pháp phù hợp để cải thiện chất lượng không khí và bảo vệ sức khỏe cộng đồng.

Đề tài "Dự báo chỉ số chất lượng không khí tại Nha Trang sử dụng mô hình mạng học sâu LSTM" xuất phát từ nhu cầu cấp thiết trong việc bảo vệ sức khỏe cộng đồng và môi trường sống trước tình trạng ô nhiễm không khí ngày càng nghiêm trọng. Thành phố Nha Trang là một điểm đến du lịch nổi tiếng và là trung tâm kinh tế đang phát triển, đang phải đối mặt với thách thức về ô nhiễm không khí do sự gia tăng nhanh chóng của dân số, phương tiện giao thông và các hoạt động công nghiệp. Mặc dù các cơ quan chức năng đã có những nỗ lực trong việc giám sát và quản lý chất lượng không khí nhưng khả năng dự báo chính xác về chỉ số chất lượng không khí vẫn cần được cải thiện.

Việc áp dụng các công nghệ tiên tiến như trí tuệ nhân tạo và mạng học sâu, đặc biệt là mô hình LSTM trở nên cần thiết. Mô hình có khả năng xử lý và phân tích các dữ liệu chuỗi thời gian, giúp nó phù hợp để dự báo các yếu tố phức tạp như chất lượng không khí. Bằng cách sử dụng mô hình này, em mong

muốn cung cấp các dự báo có độ chính xác cao về chỉ số chất lượng không khí. Từ đó hỗ trợ chính quyền địa phương và các nhà hoạch định chính sách trong việc thực hiện các biện pháp bảo vệ môi trường và sức khỏe cộng đồng.

Việc chọn lựa đề tài này không chỉ nhằm mục tiêu khoa học mà còn có ý nghĩa thực tiễn sâu sắc, đóng góp vào việc nâng cao nhận thức của xã hội về tầm quan trọng của môi trường và sự cần thiết của các biện pháp bảo vệ sức khỏe trước tác động của ô nhiễm không khí.

2. Tính cấp thiết

Trong bối cảnh phát triển kinh tế - xã hội hiện đại, việc bảo vệ và nâng cao chất lượng cuộc sống của người dân ngày càng trở nên quan trọng. Môi trường không khí là một trong những yếu tố thiết yếu ảnh hưởng trực tiếp đến sức khỏe con người. Ô nhiễm không khí đã trở thành vấn đề nghiêm trọng ở nhiều khu vực đặc biệt là các thành phố du lịch như Nha Trang. Tình trạng ô nhiễm không khí có thể gây ra nhiều tác động tiêu cực đến sức khỏe con người, làm gia tăng khả năng mắc các bệnh về hô hấp, tim mạch và giảm chất lượng cuộc sống của người dân, du khách cũng như ảnh hưởng tới hình ảnh của thành phố.

Trong thời đại của cuộc Cách mạng Công nghiệp 4.0, các công nghệ tiên tiến như học sâu (Deep Learning) đang mang lại những đột phá lớn trong việc xử lý và phân tích dữ liệu. Sử dụng các mô hình học sâu như LSTM không chỉ giúp nâng cao độ chính xác trong việc dự báo chỉ số chất lượng không khí mà còn hỗ trợ chính quyền và người dân có các biện pháp phòng ngừa kịp thời. Đặc biệt đối với Nha Trang, một thành phố du lịch nổi tiếng của Việt Nam, việc duy trì chất lượng không khí tốt là điều kiện tiên quyết để thu hút khách du lịch và đảm bảo sự phát triển bền vững.

Dù đã có nhiều nghiên cứu về ô nhiễm không khí ở các thành phố lớn, nhưng tại Nha Trang, số liệu và mô hình dự báo chất lượng không khí vẫn cần được cải thiện. Việc ứng dụng mô hình LSTM để dự báo chỉ số chất lượng không khí là cần thiết để đáp ứng nhu cầu hiện tại và lâu dài. Bài toán này không chỉ có ý nghĩa trong việc bảo vệ sức khỏe cộng đồng mà còn góp phần duy trì, phát triển kinh tế bền vững, giúp Nha Trang tiếp tục là điểm đến lý tưởng cho du khách trong và ngoài nước. Vì vậy, nghiên cứu đề tài này là cấp thiết, vừa có giá trị lý luận vừa có ý nghĩa thực tiễn cao trong việc giải quyết một trong những vấn đề quan trọng của xã hội.

3. Mục tiêu

Mục tiêu nghiên cứu của đề tài là xây dựng một mô hình dự báo chỉ số chất lượng không khí (AQI) tại Nha Trang bằng cách sử dụng mô hình mạng nơ-ron học sâu LSTM. Qua đó, nghiên cứu sẽ đánh giá hiệu quả của mô hình LSTM trong việc xử lý dữ liệu phi tuyến và phụ thuộc theo thời gian, từ đó cung cấp các giải pháp dự báo chính xác hơn cho vấn đề ô nhiễm không khí tại khu vực. Điều này không chỉ góp phần nâng cao nhận thức về chất lượng không khí mà còn hỗ trợ các chính sách quản lý môi trường hiệu quả hơn đồng thời có thể mở rộng, áp dụng cho nhiều vùng miền khác nhau trên cả nước.

4. Phạm vi nghiên cứu

Nghiên cứu tập trung vào địa điểm thành phố Nha Trang có mức độ phát triển đô thị và du lịch cao, khiến chất lượng không khí chịu tác động lớn từ các hoạt động kinh tế - xã hội trong giai đoạn từ năm 2023 đến tháng 10 năm 2024.

5. Phương pháp nghiên cứu

Phương pháp nghiên cứu đề tài bao gồm các bước thu thập và tiền xử lý dữ liệu về các chỉ số chất lượng không khí như AQI, $PM_{2.5}$, PM_{10}, CO, NO_2, O_3, SO_2 từ các nguồn uy tín nhằm đảm bảo độ chính xác của thông tin. Sau đó, mô hình LSTM được chọn do khả năng xử lý tốt dữ liệu chuỗi thời gian và dự

đoán các mẫu phi tuyến tính. Mô hình được huấn luyện và đánh giá bằng các chỉ số như MSE, RMSE, MAE để xác định hiệu quả. Kết quả mô hình sẽ được so sánh với các phương pháp khác để tìm ra phương pháp dự báo tối ưu.

6. Bố cục đề án

Ngoài mục lục, phần mở đầu, kết luận, danh mục tham khảo, nội dung chính của đề tài được chia làm ba chương:

- Chương 1: Giới thiệu tổng quan về bài toán dự báo chỉ số không khí, dữ liệu chuỗi thời gian, cách tính toán chỉ số chất lượng không khí (AQI) và các phương pháp tiếp cận bài toán hiện nay
- Chương 2: Trình bày cơ sở lý thuyết về mạng nơ-ron
- Chương 3: Trình bày mô hình LSTM dự báo chỉ số chất lượng không khí, đánh giá và so sánh kết quả với các mô hình tiếp cận khác.

Em hy vọng rằng kết quả nghiên cứu của mình sẽ mang lại lợi ích thiết thực trong việc dự đoán chất lượng không khí thông qua việc ứng dụng mô hình mạng học sâu.

CHƯƠNG 1: TỔNG QUAN BÀI TOÁN DỰ BÁO CHỈ SỐ KHÔNG KHÍ

1.1. TỔNG QUAN VỀ DỮ LIỆU CHUỖI THỜI GIAN

Phân tích chuỗi thời gian là một lĩnh vực nghiên cứu quan trọng trong thống kê và học máy, cho phép khai thác thông tin từ các dữ liệu được thu thập theo thời gian. Theo Box và Jenkins (1976), chuỗi thời gian được định nghĩa là một tập hợp các quan sát được sắp xếp theo thời gian và việc phân tích chuỗi thời gian nhằm mục đích hiểu rõ cấu trúc, hành vi của dữ liệu. Dữ liệu chuỗi thời gian có thể biểu diễn thành một dãy các vector phụ thuộc vào thời gian như sau:

$$\{x(t_0), x(t_1), \dots, x(t_{i-1}), x(t_i), x(t_{i+1}), \dots\}$$

Hình 1.1. Chuỗi thời gian chỉ số VNIndex

Các thành phần chính trong phân tích chuỗi thời gian bao gồm xu hướng, mùa vụ và các biến động ngẫu nhiên. Mỗi thành phần này cung cấp một bức tranh toàn cảnh, cái nhìn sâu sắc về dữ liệu. Dữ liệu chuỗi thời gian có thể được phân loại thành hai dạng: rời rạc và liên tục. Dữ liệu rời rạc là các chuỗi dữ liệu mà thời gian thu thập không liền mạch. Trong khi đó, dữ liệu liên tục được thu

thập theo các khoảng thời gian đều đặn phản ánh sự biến đổi liên tục của dữ liệu theo thời gian.

Đây là một công cụ quan trọng trong nhiều lĩnh vực như kinh tế, môi trường, y tế và xã hội. Trong lĩnh vực kinh tế, các chỉ số như GDP, tỷ lệ thất nghiệp và chỉ số giá tiêu dùng thường được phân tích để dự đoán xu hướng phát triển, đánh giá sức khỏe của nền kinh tế. Nhà nghiên cứu sử dụng nhiều mô hình thống kê học máy khác nhau để nhận diện các mẫu biến động và xác định các yếu tố ảnh hưởng đến những chỉ số này.

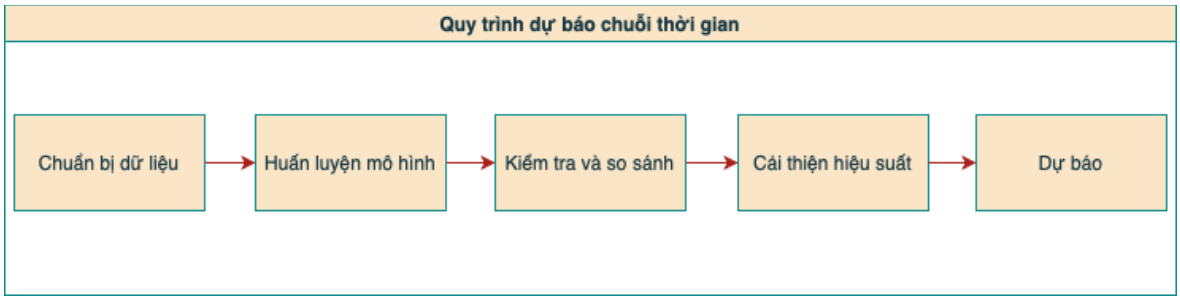
Trong lĩnh vực môi trường, phân tích chuỗi thời gian giúp theo dõi các thông số như nhiệt độ, độ ẩm và ô nhiễm không khí. Những phân tích này không chỉ giúp hiểu rõ sự biến động theo mùa mà còn hỗ trợ trong việc phát hiện những thay đổi dài hạn như biến đổi khí hậu. Thông qua việc phân tích các dữ liệu lịch sử, các nhà khoa học có thể dự đoán xu hướng tương lai từ đó đề xuất các biện pháp ứng phó hiệu quả hơn.

Trong lĩnh vực y tế, phân tích chuỗi thời gian có thể được áp dụng để theo dõi sự lây lan của các dịch bệnh, phân tích các mẫu bệnh tật theo thời gian. Nhà nghiên cứu sử dụng dữ liệu về số ca nhiễm bệnh, tỷ lệ tử vong, các chỉ số sức khỏe khác để xây dựng mô hình dự đoán hỗ trợ cơ quan y tế trong việc đưa ra quyết định kịp thời và hợp lý.

Tương tự trong lĩnh vực xã hội, phân tích chuỗi thời gian giúp nghiên cứu các hiện tượng xã hội như tỷ lệ tội phạm, hành vi tiêu dùng hay biến động dân số. Các nhà nghiên cứu có thể phân tích dữ liệu theo thời gian để phát hiện xu hướng, mối quan hệ và yếu tố ảnh hưởng đến sự thay đổi trong xã hội từ đó xây dựng các chính sách và chương trình can thiệp phù hợp.

1.2. QUY TRÌNH DỰ BÁO CHUỖI THỜI GIAN

Quy trình dự báo chuỗi thời gian, tương tự như các lĩnh vực phân tích dự báo khác bao gồm các bước như hình 1.2.



Hình 1.2. Quy trình dự báo chuỗi thời gian

Bước 1. Chuẩn bị dữ liệu: Bao gồm các công việc có thể như làm sạch dữ liệu sau khi thu thập, mã hóa lại các biến, chuẩn hóa, biến đổi hoặc tạo những biến mới và chia chuỗi dữ liệu thành tập huấn luyện và kiểm tra. Thông thường, tập kiểm tra chiếm tối đa 30% dữ liệu để đảm bảo hiệu suất tính toán.

Bước 2. Huấn luyện mô hình: Tập huấn luyện được sử dụng để thiết lập các mô hình khác nhau từ các mô hình được lựa chọn.

Bước 3. Kiểm tra và so sánh: Độ chính xác hoặc các tiêu chí chấm điểm sai số trên cả tập huấn luyện và kiểm tra của từng mô hình được so sánh để lựa chọn mô hình tiềm năng.

Bước 4. Cải thiện hiệu suất: Các mô hình tiềm năng sẽ được điều chỉnh các tham số và lặp lại quá trình huấn luyện để cải thiện hiệu suất.

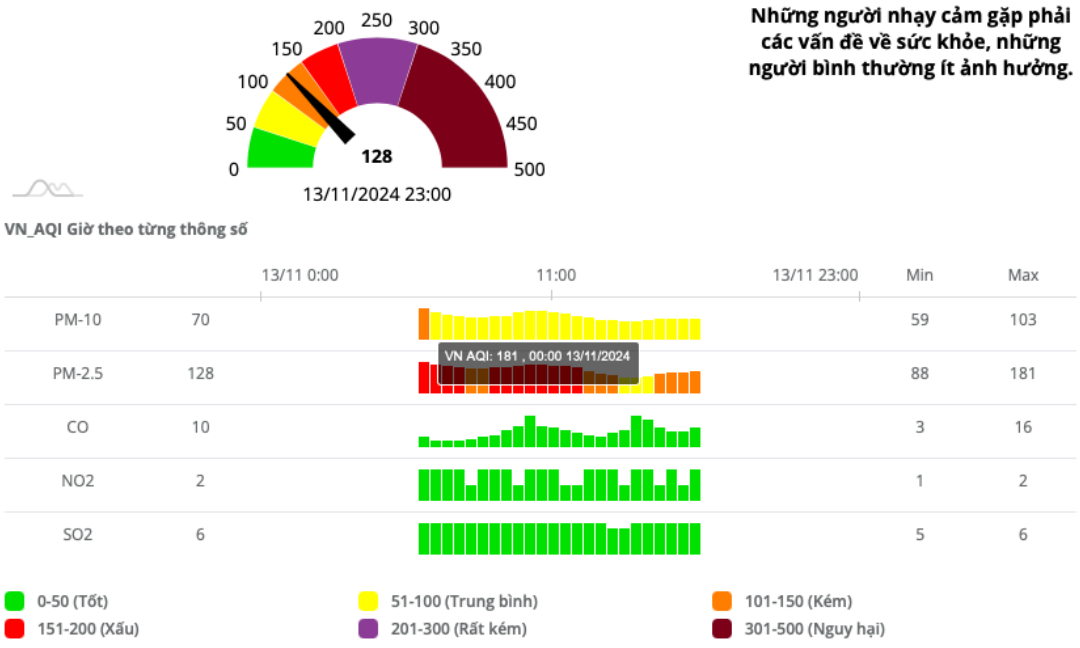
Bước 5. Dự báo: Mô hình tối ưu nhất được sử dụng cho toàn bộ dữ liệu ban đầu và đưa ra dự báo cho các bước thời gian tiếp theo.

1.3. DỮ LIỆU CHẤT LƯỢNG KHÔNG KHÍ

Dữ liệu chất lượng không khí là một dạng dữ liệu chuỗi thời gian để theo dõi và phân tích sự biến động của các chỉ số ô nhiễm như AQI, SO_2, NO_2, O_3, $PM_{2.5}$, PM_{10}, CO trong khoảng thời gian dài. Được thu thập liên tục theo ngày, giờ hoặc phút phản ánh sự thay đổi do tác động của các yếu tố như thời tiết, giao thông hay hoạt động công nghiệp. Nhờ có dữ liệu này, các nhà nghiên cứu có thể xác định xu hướng dài hạn trong chất lượng không khí, giúp cảnh báo sớm khi mức độ ô nhiễm vượt quá ngưỡng an toàn.

$PM_{2.5}$ và PM_{10} là các hạt bụi mịn có kích thước siêu nhỏ, dễ dàng xâm nhập sâu vào hệ hô hấp gây hại cho sức khỏe. Sự biến động của chúng trong dữ liệu chuỗi thời gian phụ thuộc vào nguồn phát sinh như xe cộ và công nghiệp. CO (carbon monoxide) là một chất khí độc hại thường có giá trị cao ở khu vực đô thị, khu vực có mật độ giao thông cao với nồng độ thay đổi nhanh chóng tùy thuộc vào luồng giao thông và các hoạt động khác.

NO_2 và SO_2 là các khí thường sinh ra từ quá trình đốt nhiên liệu hóa thạch, công nghiệp và giao thông. NO_2 có thể tác động mạnh đến phổi, hệ hô hấp trong khi SO_2 thường gây kích ứng da, niêm mạc mắt, mũi, họng và phổi. Ozone mặt đất (O_3) là một khí ô nhiễm thứ cấp hình thành từ phản ứng giữa ánh sáng mặt trời và các chất ô nhiễm khác. Nồng độ ozone thường tăng cao vào mùa hè và ban ngày khi cường độ ánh sáng mạnh gây tác động xấu đến sức khỏe và môi trường xung quanh.



Hình 1.3. Ví dụ dữ liệu chất lượng không khí

Hình 1.3 cung cấp biểu đồ về chỉ số chất lượng không khí tại một địa điểm được phân tích theo từng thông số như PM_{10}, $PM_{2.5}$, CO, NO_2 và SO_2

trong ngày 13/11/2024. Chỉ số AQI lúc 23:00 đạt giá trị 128 thuộc mức độ kém (màu cam) thể hiện cho những người nhạy cảm có thể gặp phải các vấn đề về sức khỏe, trong khi người bình thường ít bị ảnh hưởng.

Các chỉ số CO, NO_2, SO_2 duy trì ở mức độ tốt với các giá trị tối đa lần lượt bằng 16, 2 và 6. Chỉ số PM_{10} có giá trị trong khoảng 59 đến 103 thuộc mức độ trung bình. Chỉ số $PM_{2.5}$ có giá trị cao nhất trong các chỉ số, có thời điểm chạm mốc 181 thể hiện tình trạng ô nhiễm không khí ở mức rất xấu. Biểu đồ còn cho thấy các khoảng màu khác nhau (xanh, vàng, cam, đỏ, tím) tương ứng với mức độ ô nhiễm không khí. Những màu sắc khác nhau này hỗ trợ hiển thị thông tin một cách trực quan, dễ dàng quan sát trong quá trình phân tích.

1.4. CHỈ SỐ AQI

Những nghiên cứu quốc tế đã sử dụng chỉ số chất lượng không khí (Air Quality Index – AQI) để đánh giá toàn diện chất lượng môi trường không khí. Hiện nay, nhiều chính phủ trên thế giới đã áp dụng AQI như một công cụ tiêu chuẩn để mô tả tình trạng không khí và coi đây là quy định bắt buộc trong việc quản lý môi trường. Cơ quan Bảo vệ Môi trường Hoa Kỳ (United States Environmental Protection Agency – USEPA) đã áp dụng các thang đo AQI khác nhau để đánh giá các tiêu chí cụ thể liên quan đến một số chất gây ô nhiễm, bao gồm bụi lơ lửng có thể xâm nhập đường hô hấp (RSPM), dioxit lưu huỳnh (SO_2), nitơ dioxit (NO_2) và bụi dạng hạt lơ lửng (SPM).

Đây là một chỉ số tổng hợp biểu thị mức độ ô nhiễm không khí dựa trên nhiều thành phần như $PM_{2.5}$, PM_{10}, NO_2, SO_2, O_3 và CO. Chỉ số này phản ánh mức độ ô nhiễm chung thường dao động đáng kể trong ngày do ảnh hưởng của các yếu tố như thời tiết và giao thông.

Chỉ số chất lượng không khí được tính theo thang điểm (khoảng giá trị AQI) tương ứng với những biểu tượng, màu sắc khác nhau để cảnh báo chất lượng không khí và mức độ ảnh hưởng tới sức khỏe con người.

Bảng 1.1. Khoảng giá trị AQI và đánh giá chất lượng không khí

Khoảng giá trị AQI	Chất lượng không khí	Màu sắc
0 – 50	Tốt	Xanh
51 – 100	Trung bình	Vàng
101 – 150	Kém	Da cam
151 – 200	Xấu	Đỏ
201 – 300	Rất xấu	Tím
301 - 500	Nguy hại	Nâu

Để tính toán chỉ số AQI một cách chính xác, dữ liệu quan trắc cần được xử lý, đảm bảo đã loại bỏ các giá trị sai lệch, đạt yêu cầu đối với quy trình đảm bảo và kiểm soát chất lượng số liệu. Chỉ số VN_AQI được tính toán bao gồm AQI giờ và ngày. Số liệu sử dụng để tính toán VN_AQI là giá trị quan trắc trung bình một giờ, trung bình tám giờ và trung bình 24 giờ.

Chỉ số này được tính toán tự động, liên tục với dữ liệu môi trường không khí xung quanh từng trạm quan trắc. Đối với mỗi trạm quan trắc, cần tính toán giá trị AQI_x với x là từng thông số quan trắc đo được. Giá trị AQI cuối cùng là giá trị lớn nhất trong các giá trị AQI_x của mỗi thông số.

Các thông số được sử dụng để tính toán chỉ số VN_AQI bao gồm: SO_2, CO, NO_2, O_3, PM_{10} và $PM_{2.5}$. Phương pháp tính toán này yêu cầu bắt buộc phải có tối thiểu một trong hai thông số PM_{10}, $PM_{2.5}$ trong công thức tính.

1.4.1. Tính toán chỉ số chất lượng không khí AQI theo giờ (AQI^h)

a) Tính giá trị Nowcast đối với thông số $PM_{2.5}$ và PM_{10}

Gọi c_1, c_2, …c_{12} là 12 giá trị quan trắc trung bình một giờ (với c_1 là giá trị quan trắc trung bình một giờ hiện tại, c_{12} là giá trị quan trắc trung bình một giờ cách 12 giờ so với hiện tại). Thực hiện tính giá trị trọng số theo công thức sau:

$$W^* = \frac{C_{min}}{C_{max}} \quad (1.1)$$

Trong đó:

- $C_{\min}$ là giá trị nhỏ nhất trong số 12 giá trị trung bình một giờ.
- C_{max} là giá trị lớn nhất trong số 12 giá trị trung bình một giờ.

Nếu $W^* \leq \frac{1}{2}$ lấy $W = \frac{1}{2}$

$$Nowcast = \frac{1}{2}C_1 + \left(\frac{1}{2}\right)^2 C_2 + \cdots + \left(\frac{1}{2}\right)^{12} C_{12} \quad (1.2)$$

Nếu $W^* \geq \frac{1}{2}$ lấy $W = W^*$

$$Nowcast = \frac{\sum_{i=1}^{12} w^{i-1} c_i}{\sum_{i=1}^{12} w^{i-1}} \quad (1.3)$$

Lưu ý:

- Nếu có ít nhất hai trong ba giá trị c_1, c_2, c_3 có dữ liệu thì mới tính được giá trị Nowcast, ngược lại coi như "không có dữ liệu" (không tính được giá trị Nowcast).
- Nếu c_i không có giá trị thì lấy $w^{i-1} = 0$.

b) Tính giá trị AQI^h của từng thông số (AQI_x)

Giá trị AQI^h của các thông số SO_2, CO, NO_2, O_3 được tính toán theo công thức 1.4, giá trị AQI^h của các thông số PM_{10}, $PM_{2.5}$ được tính toán theo công thức 1.5.

$$AQI_x = \frac{I_{i+1} - I_i}{BP_{i+1} - BP_i}(C_x - BP_i) + I_i \quad (1.4)$$

$$AQI_x = \frac{I_{i+1} - I_i}{BP_{i+1} - BP_i}(Nowcast_x - BP_i) + I_i \quad (1.5)$$

Trong đó:

- AQI_x là giá trị AQI của thông số x.

- BP_i là nồng độ giới hạn dưới của giá trị thông số quan trắc được quy định trong bảng 1.2 tương ứng với mức i.
- BP_{i+1} là nồng độ giới hạn trên của giá trị thông số quan trắc được quy định trong bảng 1.2 tương ứng với mức $i + 1$.
- I_i là giá trị AQI ở mức i đã cho trong bảng tương ứng với giá trị BP_i.
- I_{i+1} là giá trị AQI ở mức $i + 1$ đã cho trong bảng tương ứng với giá trị BP_{i+1}.
- C_x là giá trị quan trắc trung bình một giờ của thông số x.
- $Nowcast_x$ là giá trị Nowcast được tính toán.

Bảng 1.2. Bảng giá trị BP_i đối với các thông số

i	I_i	**Giá trị BP_i quy định đối với từng thông số** *(Đơn vị: µg/m³)*						
		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 có giá trị AQI_x của mỗi thông số, chọn giá trị AQI lớn nhất của các thông số để lấy làm giá trị *AQI giờ tổng hợp*.

$AQI^h = max(AQI_x)$ *(làm tròn thành số nguyên)*

1.4.2. Tính toán chỉ số chất lượng không khí AQI theo ngày (AQI^d)

Giá trị AQI ngày được tính toán dựa trên các giá trị:

- Thông số $PM_{2.5}$ và PM_{10} trung bình 24 giờ.

- Thông số O_3 trung bình một giờ lớn nhất trong ngày và giá trị trung bình tám giờ lớn nhất trong ngày.
- Thông số SO_2, NO_2 và CO trung bình một giờ lớn nhất trong ngày.
- Giá trị trung bình một giờ lớn nhất trong ngày là giá trị lớn nhất trong số các giá trị quan trắc trung bình một giờ.
- Giá trị quan trắc trung bình tám giờ lớn nhất trong ngày là giá trị lớn nhất trong số các giá trị trung bình tám giờ. Giá trị trung bình tám giờ là trung bình cộng các giá trị trung bình một giờ trong tám giờ liên tiếp.

Bảng 1.3. Ví dụ giá trị trung bình một giờ của chỉ số ozone

Giờ	18:00	19:00	20:00	21:00	22:00	23:00	00:00	1:00
O_3(TB1h)	15,7	14,2	17,7	18,9	19,3	15,7	19,7	22,6
Giờ	2:00	3:00	4:00	5:00	6:00	7:00	8:00	9:00
O_3(TB1h)	27,1	29,0	31,9	25,3	34,7	35,2	41,6	45,7
Giờ	10:00	11:00	12:00	13:00	14:00	15:00	16:00	17:00
O_3(TB1h)	49,2	55,8	69,6	78,3	91,5	97,7	81,2	71,8
Giờ	18:00	19:00	20:00	21:00	22:00	23:00	00:00	
O_3(TB1h)	43,5	34,3	21,5	20,5	19,4	20,4	21,3	

Bảng 1.4. Ví dụ giá trị trung bình tám giờ của chỉ số ozone trong một ngày

Giờ	1:00	2:00	3:00	4:00	5:00	6:00
O_3(TB8h)	18,0	19,4	21,2	23,0	23,8	25,8
Giờ	7:00	8:00	9:00	10:00	11:00	12:00
O_3(TB8h)	28,2	30,9	33,8	36,6	39,9	44,7
Giờ	13:00	14:00	15:00	16:00	17:00	18:00

O_3(TB8h)	51,3	58,4	66,2	71,1	74,4	73,7
Giờ	19:00	20:00	21:00	22:00	23:00	0:00
O_3(TB8h)	71,0	65,0	57,8	48,7	39,1	31,6

Tính giá trị AQI^d của từng thông số (AQI_x)

Giá trị AQI ngày của các thông số SO_2, CO, NO_2, O_3, PM_{10}, $PM_{2.5}$ được tính toán theo *công thức 1.* Sau khi đã có giá trị AQI_x ngày của mỗi thông số, chọn giá trị AQI lớn nhất của các thông số để lấy làm giá trị AQI ngày tổng hợp.

$$AQI^d = max(AQI_x) \text{ (làm tròn thành số nguyên)}$$

1.5. BÀI TOÁN DỰ BÁO CHẤT LƯỢNG KHÔNG KHÍ

Các bài toán liên quan đến xây dựng mô hình đánh giá và dự báo chất lượng không khí luôn nhận được sự quan tâm trong lĩnh vực khoa học môi trường. Điều này xuất phát từ những tác động tiêu cực của ô nhiễm không khí đối với sức khỏe con người, đặc biệt tại các khu vực đô thị lớn. Phân vùng ô nhiễm không khí cho một tỉnh hoặc thành phố đã trở thành trọng tâm của nhiều nghiên cứu trong nước, nhằm đưa ra các giải pháp quản lý và cải thiện chất lượng môi trường không khí.

Bài toán dự báo chất lượng không khí tại Nha Trang sẽ xây dựng mô hình có khả năng dự báo chỉ số AQI trong tương lai dựa trên dữ liệu lịch sử đã được thu thập. Mục tiêu bài toán là xây dựng một mô hình dự báo chính xác, hiệu quả để đưa ra các cảnh báo kịp thời, hỗ trợ người dân, cơ quan quản lý và các tổ chức liên quan.

Quá trình xây dựng mô hình yêu cầu lượng dữ liệu lớn được thu thập từ quá khứ. Dữ liệu này bao gồm nồng độ các chất gây ô nhiễm không khí ($PM_{2.5}$, PM_{10}, CO, NO_2, SO_2) ghi nhận theo giờ hoặc ngày. Đầu ra của mô hình là giá trị dự báo chỉ số chất lượng không khí trong khoảng thời gian tương lai.

Bài toán yêu cầu mô hình phải đạt độ chính xác cao thông qua việc giảm thiểu các sai số dự báo như MSE, RMSE, MAE và MAPE. Mô hình cần có khả năng dự báo ngắn hạn và trung hạn cung cấp các kết quả chính xác trong khoảng thời gian từ vài giờ đến vài ngày. Ngoài ra mô hình phải đảm bảo tính khả thi trong thực tế với khả năng triển khai rộng rãi ở nhiều khu vực khác nhau.

Để đáp ứng yêu cầu đó bài toán cần sử dụng những tập dữ liễu đầu vào chất lượng không bị đồng nhất, nhiễu hoặc thiếu lượng giá trị lớn. Tập dữ liệu cần thể hiện được tính biến động phức tạp của chỉ số AQI chịu tác động từ nhiều yếu tố như thời tiết, giao thông hay hoạt động công nghiệp. Đây là khó khăn đồng thời là thách thức lớn đối với bài toán do sự hạn chế của các trạm quan trắc, thiết bị hay nguồn nhân lực tại những địa điểm có vị trí địa lý khó khăn dẫn đến thiếu hụt lượng lớn thông tin cần đo lường. Một khó khăn khác đối với bài toán khi xây dựng mô hình học sâu, với lượng dữ liệu tăng theo từng ngày từng giờ sẽ yêu cầu lượng tài nguyên tính toán đáng kể để xử lý và huấn luyện mô hình.

Bài toán có ý nghĩa ứng dụng thực tiễn cao, đặc biệt trong việc cảnh báo sớm về tình trạng ô nhiễm không khí. Điều này giúp người dân chủ động bảo vệ sức khỏe trước các nguy cơ tiềm ẩn từ môi trường. Ứng dụng mô hình của bài toán hỗ trợ hiệu quả cho các cơ quan quản lý trong việc quy hoạch và triển khai các biện pháp bảo vệ môi trường đô thị. Mô hình dự báo được sử dụng cho bài toán đóng góp một phần tài liệu cho nghiên cứu khoa học, cung cấp dữ liệu và phân tích nhằm hiểu rõ hơn tác động của ô nhiễm không khí đối với sức khỏe cộng đồng.

1.6. KHẢO SÁT CÁC PHƯƠNG PHÁP HIỆN CÓ

Dự báo chỉ số chất lượng không khí đóng vai trò quan trọng trong việc giảm thiểu tác động tiêu cực của ô nhiễm không khí đến sức khỏe cộng đồng và môi trường. Hiện nay, các phương pháp dự báo đã không ngừng phát triển

từ các mô hình thống kê truyền thống đến các phương pháp học máy hiện đại. Dưới đây là chi tiết về từng phương pháp bao gồm các đặc điểm, ứng dụng, ưu và nhược điểm.

1.6.1. Holt-Winters

Holt (2004) và Winters (1960) đã mở rộng mô hình Holt cho các dữ liệu chuỗi thời gian với cả hai thành phần xu hướng và mùa, gọi là mô hình HW. Hai dạng mô hình HW cơ bản bao gồm một phương trình dự báo và ba phương trình làm trơn được cho sau đây:

- Mô hình HW^{+} được biểu diễn:

$$\hat{Y}_{t+h|t} = (l_t + hb_t)s_{t+h-m(k+1)} \quad (1.6)$$

$$l_t = \alpha(Y_t - s_{t-m}) + (1-\alpha)(l_{t-1} + b_{t-1}) \quad (1.7)$$

$$b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1} \quad (1.8)$$

$$s_t = \gamma(Y_t - l_{t-1} - b_{t-1}) + (1-\gamma)s_{t-m} \quad (1.9)$$

Trong đó:

- $0 \le \alpha \le 1;\ 0 \le \beta \le 1;\ 0 \le \gamma \le 1.$
- m là tần suất mùa.
- k là phần nguyên của $(h-1)/m.$
- h là số bước dự đoán tiếp theo tính từ thời điểm *t.*
- l_t, b_t và s_t lần lượt là các thành phần mức độ, xu hướng và mùa của chuỗi.

- Mô hình HW* được biểu diễn:

$$\hat{Y}_{t+h|t} = l_t + hb_t + s_{t+h-m(k+1)} \quad (1.10)$$

$$l_t = \alpha\left(\frac{Y_t}{s_{t-m}}\right) + (1-\alpha)(l_{t-1} + b_{t-1}) \quad (1.11)$$

$$b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1} \quad (1.12)$$

$$s_t = \gamma\left(\frac{Y_t}{l_{t-1} - b_{t-1}}\right) + (1-\gamma)s_{t-m} \quad (1.13)$$

1.6.2. Autoregressive Integrated Moving Average

Hai tác giả George Box và Gwilym Jenkins (1976) đã nghiên cứu mô hình Autoregressive Integrated Moving Average, viết tắt là ARIMA. Tên của họ (Box-Jenkins) được dùng để gọi cho các quá trình ARIMA tổng quát áp dụng vào phân tích và dự báo các chuỗi thời gian. Bản chất của mô hình ARIMA là dự báo giá trị tương lai của một biến số (biểu thị theo chuỗi thời gian) dựa trên giá trị quá khứ và các sai số ngẫu nhiên. Trong nghiên cứu của P. Goyal (2006) chỉ ra rằng mô hình này có thể dự báo chất lượng không khí tại Delhi một cách chính xác. Tuy nhiên, mô hình này cũng cho thấy giới hạn trong việc xử lý dữ liệu có sự biến động phức tạp do đó chỉ thích hợp cho việc phân tích, dự báo dữ liệu chuỗi thời gian không có yếu tố mùa vụ.

Mô hình có kí hiệu ARIMA(*p, d, q*) được kết hợp từ các thành phần tự hồi quy AR(*p*), tích hợp I(*d*) với *d* là số lần sai phân chuỗi đến khi đạt trạng thái dừng, và trung bình trượt MA(*q*) để loại bỏ tính không ổn định và mô hình hóa cấu trúc của chuỗi thời gian. Mô hình có thể biểu diễn như sau:

$$Y_t' = c + \phi_1 Y_{t-1}' + \phi_2 Y_{t-2}' + \cdots + \phi_p Y_{t-p}' + \in_t + \theta_1 \in_{t-1} + \theta_2 \in_{t-2} + \cdots + \theta_q \in_{t-q} \quad (1.14)$$

Trong đó:

- Y_t' là sai phân của chuỗi Y_t (sai phân đến *d* lần).
- ϕ_i và θ_j $(i = 1, \dots, p, j = 1, \dots, q)$ lần lượt là các tham số của mô hình AR(*p*) và MA(*q*).
- $\in_t$ là nhiễu trắng.

1.6.3. Seasonal ARIMA

Mô hình SARIMA được phát triển tiếp từ mô hình ARIMA phù hợp với bất kỳ dữ liệu chuỗi thời gian mùa vụ nào, có thể là bốn quý trong năm; bảy ngày trong tuần v.v. Nghiên cứu của Zhang et al. (2018), SARIMA đã được áp dụng để dự báo AQI tại Bắc Kinh và đạt kết quả khả quan trong việc nắm bắt

xu hướng theo mùa. Mặc dù vậy, hiệu suất của mô hình này lại giảm đáng kể khi áp dụng cho các dữ liệu dài hạn hoặc có nhiều yếu tố phi tuyến.

Ngoài ra mô hình SARIMA còn được sử dụng trong nhiều lĩnh vực khác. Reininger & Fingerlos (2007) sử dụng dữ liệu chuỗi GDP thực trong thời gian từ quý 1/1980 tới quý 4/2006 và tìm ra được mô hình phù hợp với mục đích giải thích các đặc điểm mùa vụ của bài toán. Wongkoon và cộng sự (2008) áp dụng mô hình để dự báo ca sốt xuất huyết ở miền Bắc Thái Lan. K. Rajendran và cộng sự (2011) sử dụng mô hình SARIAM và tuyến tính tổng quát (GLM) để nghiên cứu mối tương quan giữa số ca bệnh dịch tả với thời tiết. Savas (2013) sử dụng mô hình SARIAM và Kalman để dự báo lạm phát hàng tháng ở Luxembourg, Mexico, Bồ Đào Nha và Thụy Sỹ. Tại Việt Nam, cũng đã có những công trình sử dụng mô hình SARIAM để dự báo như Nguyễn Khắc Hiếu (2014) sử dụng mô hình ARIAM để dự báo lạm phát sáu tháng cuối năm 2014.

1.6.4. Linear Regression

LR là một mô hình toán học mô tả mối quan hệ tuyến tính giữa một biến phụ thuộc và một hoặc nhiều biến độc lập. Mô hình này được biểu diễn qua phương trình:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon \qquad (1.15)$$

Trong đó:

- Y là biến phụ thuộc.
- $X_1, X_2, \dots, X_p$ là các biến độc lập.
- $\beta_0, \beta_1, \dots, \beta_p$ là các hệ số hồi quy.
- ϵ là sai số.

Hồi quy tuyến tính là một kỹ thuật thống kê đã được sử dụng từ lâu, hiệu quả và dễ dàng áp dụng cho phần mềm, tính toán, cung cấp một công thức toán học dễ giải thích để đưa ra các dự đoán.

1.6.5. Random Forest

RF là một thuật toán supervised learning, có thể giải quyết cả bài toán regression và classification. Trong nghiên cứu ở Delhi, Ấn Độ [1] đã nghiên cứu ảnh hưởng của nhận thức về chất lượng không khí đối với lựa chọn phương tiện giao thông bằng cách sử dụng các mô hình máy học. Các tác giả đã thu thập dữ liệu từ 346 người tham gia khảo sát và phân tích cách ô nhiễm không khí tác động đến hành vi di chuyển. Các mô hình được sử dụng bao gồm Random Forest, XGBoost, Naive Bayes, K-Nearest Neighbor, Support Vector Machine và Logistic Regression.

Nghiên cứu cho thấy Random Forest có độ chính xác cao nhất trong việc dự đoán hành vi di chuyển. Khi chất lượng không khí xấu đi, người đi lại có xu hướng chọn các phương tiện giao thông kín (như xe hơi có điều hòa và tàu điện ngầm) thay vì các phương tiện mở (như đi bộ và xe đạp) và ưu tiên giao thông công cộng hơn giao thông cá nhân do hệ thống thông gió và lọc khí.

1.6.6. Recurrent Neural Network

Recurrent Neural Networks (RNN) là một phương pháp học sâu phổ biến, đặc biệt hiệu quả trong việc phân tích và dự báo chuỗi thời gian nhờ khả năng ghi nhớ trạng thái của dữ liệu trước đó. Trong các nghiên cứu về dự báo chỉ số chất lượng không khí, RNN đã được áp dụng để phân tích tác động của biến động không khí theo thời gian. Một nghiên cứu tại Bắc Kinh đã sử dụng mô hình RNN để dự báo AQI hàng giờ, cho thấy khả năng nắm bắt tốt các chu kỳ ngắn hạn của chất lượng không khí. Mô hình này cũng đã được so sánh với các phương pháp truyền thống như hồi quy tuyến tính và ARIMA, với kết quả vượt trội hơn trong hầu hết các trường hợp. Tuy nhiên, để xử lý dữ liệu có phụ thuộc dài hạn hoặc phi tuyến tính cao, các biến thể như LSTM hoặc GRU thường được sử dụng thay thế để khắc phục nhược điểm của RNN cơ bản.

1.6.7. Long Short-Term Memory

LSTM là một mô hình hiện đại được sử dụng rộng rãi trong các bài toán phân tích và dự báo chuỗi thời gian. Trong các nghiên cứu liên quan đến dự báo chất lượng không khí, LSTM được đánh giá cao nhờ khả năng xử lý dữ liệu có phụ thuộc dài hạn và phi tuyến tính. Các nghiên cứu đã chứng minh rằng việc áp dụng các mô hình LSTM riêng lẻ hoặc kết hợp LSTM với các phương pháp khác đều đạt hiệu quả vượt trội so với các mô hình dự báo truyền thống.

Bảng 1.5. Tổng hợp các nghiên cứu ứng dụng mô hình LSTM

Nghiên cứu	**Chất ô nhiễm**	**Quốc gia**	**Mô hình sử dụng**
Yan và cs [15]	AQI	Trung Quốc	BPNN, CNN, LSTM CNN-LSTM
Jiao và cs [16]	AQI	Trung Quốc	LSTM
Belavadi và cs [17]	AQI	Ấn Độ	LSTM
Duan và cs [18]	AQI	Trung Quốc	LSTM, CNN-LSTM DBO-LSTM
Yammahi và cs [19]	NO_2	UAE	LSTM, NAR-NN ARIMA, SARIMA
Navares và cs [20]	CO, NO_2, O_3, PM_{10}, SO_2	Tây Ban Nha	LSTM-RNN
Chang và cs [21]	$PM_{2.5}$	Đài Loan	Aggregated-LSTM
Wen và cs [22]	$PM_{2.5}$	Trung Quốc	CNN-LSTM
Jung và cs [23]	PM_{10}	Hàn Quốc	DNN, RNN LSTM
Rakholia và cs [24]	$PM_{2.5}$	Việt Nam	XGBoost, GDRegressor 1D

			CNN-LSTM, Prophet
Hưng [25]	CO, NO_x, O_3, $PM_{2.5}$, PM_{10}, SO_2	Việt Nam	CNN-LSTM

1.7. KẾT LUẬN CHƯƠNG

Chương 1 đã trình bày tổng quan về bài toán dự báo chỉ số chất lượng không khí (AQI) bao gồm khái niệm dữ liệu chuỗi thời gian, quy trình xử lý, và các phương pháp hiện có. Những nội dung này cung cấp cái nhìn tổng quan cần thiết để nghiên cứu bài toán dự báo chỉ số chất lượng không khí, làm cơ sở cho việc triển khai các mô hình phân tích và dự báo trong các chương sau.

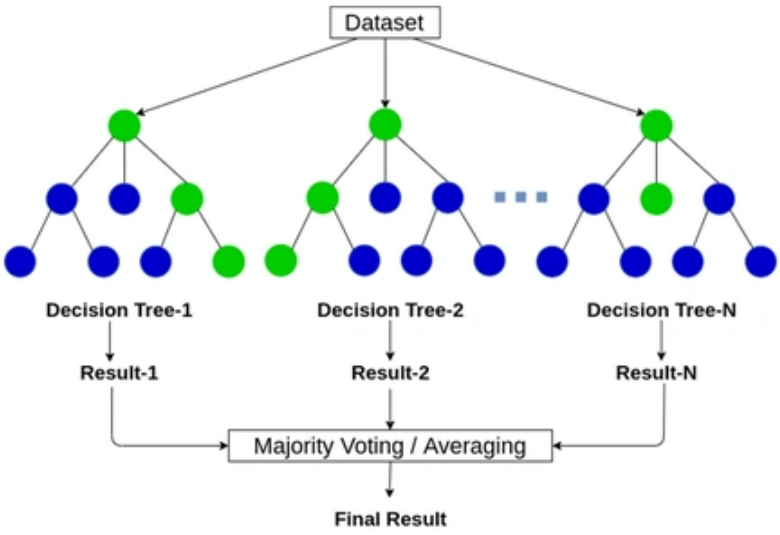
CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

Chương 2 trình bày các cơ sở lý thuyết làm nền tảng cho việc nghiên cứu và áp dụng các mô hình dự báo chỉ số chất lượng không khí. Nội dung bao gồm tổng quan về thuật toán học máy và các mô hình mạng nơ-ron, trong đó nhấn mạnh các đặc điểm và ứng dụng của mạng nơ-ron nhân tạo (ANN), mạng nơ-ron hồi quy (RNN) và mạng nơ-ron bộ nhớ dài ngắn hạn (LSTM).

2.1. THUẬT TOÁN HỌC MÁY

a) Random Forest

Thuật toán Random Forest là một thuật toán học máy phổ biến, được ứng dụng rộng rãi trong các bài toán phân loại và hồi quy [13].



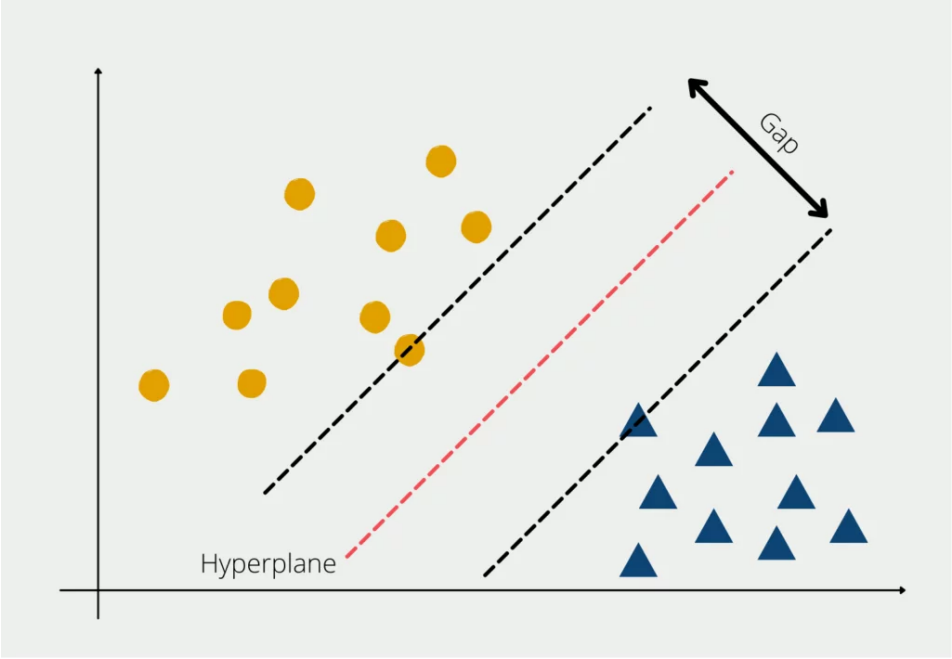
Hình 2.1. Mô hình ra quyết định của Random Forest

Nguyên lý hoạt động của thuật toán là xây dựng một tập hợp các cây quyết định (decision trees), trong đó, số lượng cây quyết định tăng lên sẽ làm tăng độ chính xác của dự đoán. Mỗi cây trong mô hình có các node, trong đó các node hình chữ nhật đại diện cho câu hỏi, còn node hình tròn đại diện cho kết quả. Kết quả phân loại được xác định dựa trên số phiếu bầu của các cây

quyết định. Random Forest có khả năng xử lý dữ liệu thiếu và tính toán hiệu quả với dữ liệu lớn, đa chiều, đem lại độ chính xác cao trong các tình huống thực tế.

b) Support Vector Machine

Support Vector Machine là một thuật toán học máy phổ biến có khả năng áp dụng hiệu quả cho cả bài toán phân loại và hồi quy [13] với mục tiêu tối ưu hóa khả năng phân tách dữ liệu dựa trên không gian đặc trưng (feature space). Mỗi dữ liệu được biểu diễn như một điểm trong không gian với số chiều tương ứng với số lượng đặc trưng của dữ liệu. Thuật toán này xây dựng một hyperplane để phân chia các lớp dữ liệu sao cho khoảng cách giữa hyperplane và các điểm gần nhất thuộc hai lớp dữ liệu (được gọi là các vector hỗ trợ - support vectors) là lớn nhất. Điều này đảm bảo tính tổng quát hóa của mô hình giúp nó hoạt động tốt trên dữ liệu chưa biết.



Hình 2.2. Minh họa hyperplane trong SVM

SVM hoạt động dựa trên khái niệm tối đa hóa biên (margin) được định nghĩa là khoảng cách từ các vector hỗ trợ tới hyperplane. Một hyperplane tối ưu không chỉ phân tách các lớp dữ liệu một cách chính xác mà còn đạt được

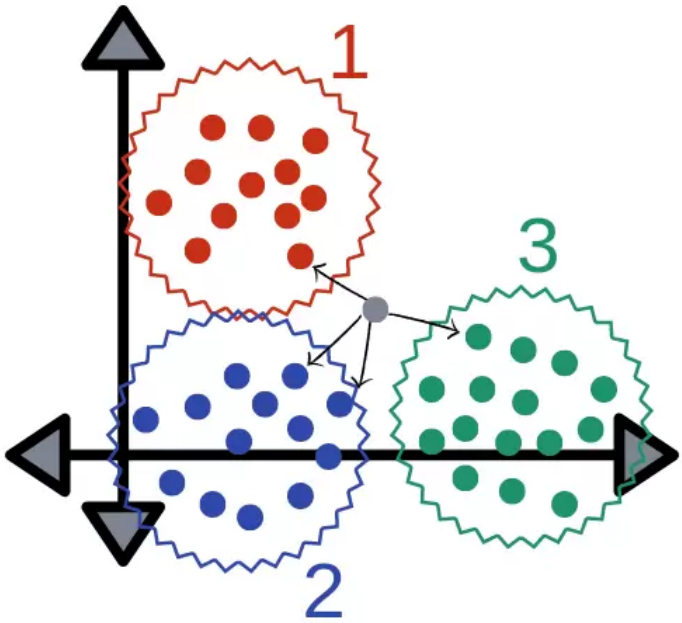
khoảng cách biên lớn nhất, giúp giảm thiểu khả năng overfitting. Đối với các bài toán mà dữ liệu không thể phân tách tuyến tính, SVM sử dụng các hàm kernel để ánh xạ dữ liệu từ không gian gốc lên không gian đặc trưng cao hơn. Các hàm kernel phổ biến bao gồm kernel tuyến tính (linear kernel), Gaussian RBF (Radial Basis Function) và polynomial kernel. Nhờ vào khả năng này, SVM có thể giải quyết hiệu quả các bài toán phức tạp và phi tuyến.

c) K-Nearest Neighbor

Thuật toán K-Nearest Neighbors (KNN) là một thuật toán học máy dựa trên nguyên lý đo lường khoảng cách giữa các điểm dữ liệu trong không gian đặc trưng. KNN thuộc nhóm thuật toán không tham số và có thể được áp dụng trong cả bài toán phân loại và hồi quy.

Trong bài toán phân loại, KNN hoạt động bằng cách tìm kiếm K điểm dữ liệu gần nhất (theo khoảng cách Euclid hoặc các phương pháp đo lường khác như khoảng cách Manhattan hoặc Minkowski) trong không gian đặc trưng của dữ liệu huấn luyện. Dựa trên nhãn của K điểm gần nhất này, điểm dữ liệu mới sẽ được phân loại theo nguyên tắc majority voting. Ví dụ, nếu trong K điểm gần nhất phần lớn thuộc về một nhãn cụ thể thì điểm dữ liệu mới cũng sẽ được gán nhãn tương tự. Điều này làm cho KNN trở thành một thuật toán rất trực quan và hiệu quả trong các trường hợp mà ranh giới phân loại không rõ ràng hoặc không tuyến tính.

Trong bài toán hồi quy, KNN dự đoán giá trị của điểm dữ liệu mới dựa trên trung bình giá trị đầu ra tương ứng từ K điểm dữ liệu gần nhất. KNN yêu cầu lựa chọn cẩn thận số lượng K cũng như hàm đo khoảng cách phù hợp để đạt được kết quả tốt nhất. Một giá trị K quá nhỏ có thể làm tăng nguy cơ overfitting, trong khi một giá trị K quá lớn có thể làm giảm độ chính xác của mô hình do ảnh hưởng của các điểm dữ liệu xa.



Hình 2.3. Minh họa thuật toán KNN

Nhược điểm lớn của KNN là chi phí tính toán cao trong giai đoạn dự đoán, đặc biệt đối với các tập dữ liệu lớn hoặc có số chiều cao (high-dimensional data). Thuật toán này cũng nhạy cảm với dữ liệu nhiễu và yêu cầu bước tiền xử lý dữ liệu cẩn thận, bao gồm việc chuẩn hóa để đảm bảo các đặc trưng có thang đo tương đương. Tuy nhiên nhờ tính dễ hiểu và hiệu quả trong các bài toán không yêu cầu mô hình hóa phức tạp, KNN vẫn được áp dụng rộng rãi trong nhiều lĩnh vực.

d) Extreme Gradient Boosting

XGBoost là một thuật toán học máy mạnh mẽ dựa trên phương pháp cây quyết định. Được phát triển từ năm 2016 tại Đại học Washington, Hoa Kỳ, thuật toán này nhanh chóng trở thành một công cụ phổ biến trong lĩnh vực khoa học dữ liệu nhờ tính linh hoạt và hiệu quả cao trong xử lý dữ liệu lớn và đa chiều.

Ưu điểm của XGBoost là khả năng xử lý tốt dữ liệu thiếu (missing values) và giảm thiểu hiện tượng overfitting nhờ các cơ chế như regularization và pruning. Thuật toán hỗ trợ cả hai loại bài toán học máy phổ biến là hồi quy

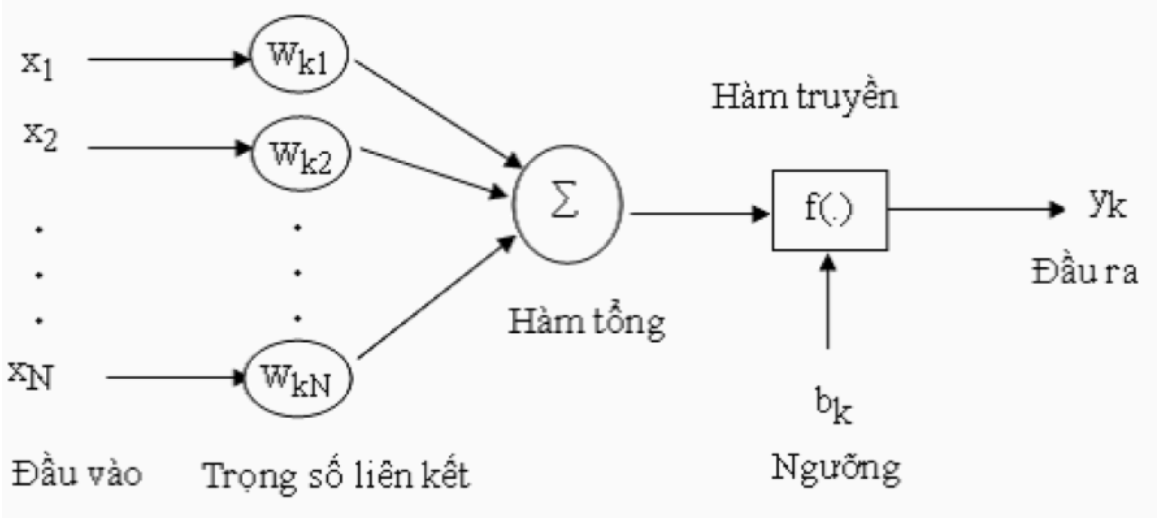
và phân loại với ứng dụng rộng rãi trong các lĩnh vực như dự đoán, phân loại, xếp hạng và khuyến nghị. Ngoài ra XGBoost còn có khả năng tính toán song song, giúp giảm thời gian huấn luyện đáng kể trên các tập dữ liệu lớn và phức tạp.

2.2. MẠNG NƠ-RON NHÂN TẠO

Nơ-ron là đơn vị cấu trúc cơ bản của hệ thống thần kinh và giữ vai trò trung tâm trong hoạt động của não bộ. Con người có khoảng 10 triệu nơ-ron, mỗi nơ-ron liên kết với khoảng 10.000 nơ-ron khác. Mỗi nơ-ron gồm thân (soma) chứa nhân, sợi nhánh (dendrites) để nhận tín hiệu đầu vào, và sợi trục (axon) truyền tín hiệu đầu ra tới nơ-ron khác. Khi nhận xung điện đủ lớn qua sợi nhánh, nơ-ron sẽ kích hoạt, truyền tín hiệu qua sợi trục, thực hiện chức năng tương tự như hàm sigmoid trong mạng nơ-ron nhân tạo, đảm bảo xử lý và truyền tải thông tin hiệu quả. Mạng nơ-ron nhân tạo được lấy cảm hứng từ cấu trúc và cách hoạt động của não bộ nhưng nó không nhằm mục đích mô phỏng toàn bộ các chức năng phức tạp của bộ não. Mục đích chính của mạng nơ-ron nhân tạo là được sử dụng để giải quyết các bài toán cụ thể trong thực tế.

Mô hình Artificial Neural Networks (ANN) là một hệ thống xử lý thông tin được thiết kế dựa trên nguyên lý hoạt động của hệ thần kinh, bao gồm một số lượng lớn các nơ-ron liên kết chặt chẽ để thực hiện việc xử lý thông tin. Tương tự như não bộ con người, ANN học hỏi thông qua kinh nghiệm (quá trình huấn luyện), lưu trữ các tri thức đã học và sử dụng chúng để đưa ra dự đoán cho những dữ liệu chưa được biết đến.

a) Hoạt động của một nơ-ron nhân tạo



Hình 2.4. Cấu trúc của một nơ-ron nhân tạo

Đầu vào $(x_1, x_2, \dots, x_N)$: Một nơ-ron trong mạng nơ-ron nhân tạo nhận các tín hiệu đầu vào, được biểu diễn dưới dạng x_i $(i = 1,2, \dots, N)$. Các tín hiệu này có thể là dữ liệu thô từ tập dữ liệu đầu vào hoặc là đầu ra từ các nơ-ron ở lớp trước. Số lượng đầu vào *N* phụ thuộc vào số lượng đặc trưng của dữ liệu.

Đầu ra y_k*:* Kết quả của một mạng ANN là một giải pháp cho một vấn đề, ví dụ như với bài toán xem xét chấp nhận cho khách hàng vay tiền hay không thì đầu ra là có hoặc không.

Trọng số liên kết $(w_{k1}, w_{k2}, \dots, w_{kN})$: Mỗi tín hiệu đầu vào x_i được nhân với một trọng số tương ứng w_{ki}. Trọng số này phản ảnh mức độ quan trọng của tín hiệu đầu vào đối với nơ-ron. Trong quá trình huấn luyện, các trọng số này được điều chỉnh để tối ưu kết quả đầu ra.

Hàm tổng: Dùng để tính tổng trọng số của tất cả các giá trị đầu vào được đưa vào mỗi nơ-ron. Hàm tổng của một nơ-ron được tính theo công thức sau:

$$y_k = \sum_{i=1}^{N} X_i W_i \qquad (2.1)$$

Trong đó:

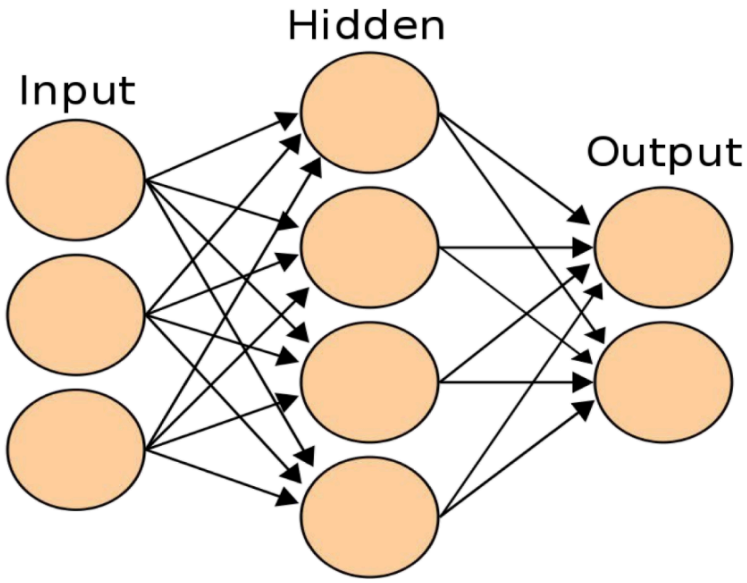
- X_i là giá trị của đầu vào thứ i.

- W_i là giá trị trọng số thứ i.

Sau khi tính toán giá trị y_k, nơ-ron áp dụng một hàm truyền $f(y_k)$ để tính toán đầu ra. Hàm này thường được gọi là hàm kích hoạt (activation function), giúp mạng nơ-ron học các mối quan hệ phi tuyến tính. Việc lựa chọn hàm kích hoạt có tác động lớn đến kết quả đầu ra của mạng ANN. Một số hàm kích hoạt phổ biến là *sigmoid* hoặc *tanh*.

b) Kiến trúc mạng nơ-ron nhân tạo

Một mạng nơ-ron nhân tạo gồm có ba lớp: lớp đầu vào, lớp ẩn và lớp đầu ra. Đây là cấu trúc thường thấy trong các mô hình học sâu và được sử dụng để giải quyết nhiều bài toán phức tạp.



Hình 2.5. Mạng nơ-ron nhân tạo

Lớp đầu vào là nơi nhận dữ liệu đầu vào từ bài toán, với mỗi nơ-ron đại diện cho một đặc trưng hoặc biến số trong tập dữ liệu. Số lượng nơ-ron của lớp này được xác định bởi số chiều của dữ liệu đầu vào. Dữ liệu từ lớp này sẽ được truyền đến lớp ẩn thông qua các liên kết có trọng số.

Lớp ẩn đóng vai trò quan trọng trong việc học các mối quan hệ phi tuyến tính giữa dữ liệu đầu vào và đầu ra. Các kết nối này mang trọng số là những tham số được điều chỉnh trong quá trình huấn luyện. Tại lớp ẩn, các tín hiệu đầu vào được tổng hợp, áp dụng hàm kích hoạt (activation function) và tạo ra đầu ra phi tuyến tính để truyền đến lớp tiếp theo.

Lớp đầu ra là nơi biểu diễn kết quả cuối cùng của mạng. Số lượng nơ-ron tại lớp này phụ thuộc vào yêu cầu của bài toán. Với bài toán phân loại, số nơ-ron tương ứng với số lớp (classes) cần phân loại, trong khi với bài toán hồi quy, thường chỉ có một nơ-ron đầu ra.

Các nơ-ron trong mạng được kết nối hoàn toàn (fully connected), mỗi nơ-ron ở một lớp được liên kết với tất cả các nơ-ron ở lớp liền kề. Điều này giúp đảm bảo rằng thông tin từ mỗi đặc trưng đầu vào đều được truyền đi và ảnh hưởng đến quá trình đưa ra kết quả. Trọng số liên kết và hệ số bias sẽ được tối ưu hóa trong quá trình huấn luyện để giảm thiểu sai số, thông qua các thuật toán khác nhau.

c) Mạng nơ-ron lan truyền thẳng

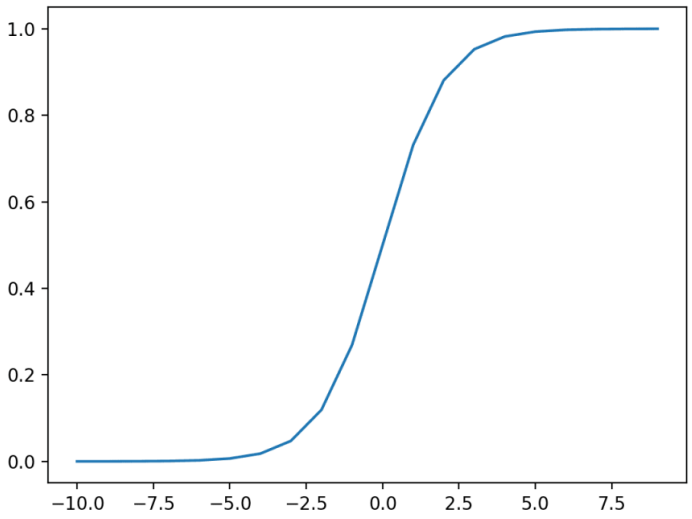
Mạng nơ-ron lan truyền thẳng (Feedforward Neural Network - FNN) là một trong những kiến trúc cơ bản của mạng nơ-ron nhân tạo, trong đó dữ liệu di chuyển theo một hướng từ lớp đầu vào qua các lớp ẩn đến lớp đầu ra mà không có vòng lặp hay kết nối ngược. Mô hình này được sử dụng rộng rãi trong các bài toán phân loại, hồi quy và dự báo.

Cấu trúc của FNN bao gồm ba thành phần chính: lớp đầu vào (input layer), một hoặc nhiều lớp ẩn (hidden layers) và lớp đầu ra (output layer). Mỗi lớp bao gồm một tập hợp các nơ-ron, trong đó mỗi nơ-ron nhận đầu vào từ các nơ-ron của lớp trước đó, thực hiện phép biến đổi tuyến tính thông qua trọng số

(weights) và hệ số điều chỉnh (bias), sau đó áp dụng một hàm kích hoạt phi tuyến như ReLU, *sigmoid* hoặc *tanh* để tạo đầu ra.

Mạng nơ-ron lan truyền thẳng học cách ánh xạ dữ liệu đầu vào sang đầu ra bằng cách điều chỉnh trọng số và bias thông qua thuật toán lan truyền ngược (Backpropagation) kết hợp với các phương pháp tối ưu hóa như Gradient Descent. Quá trình huấn luyện này nhằm mục tiêu giảm thiểu hàm mất mát, giúp mô hình cải thiện khả năng dự đoán trên dữ liệu chưa thấy trước.

d) Hàm *sigmoid*



Hình 2.6. Hàm kích hoạt sigmoid

Hàm *sigmoid* là hàm kích hoạt phi tuyến được sử dụng phổ biến nhất trong mạng nơ-ron nhân tạo được định nghĩa:

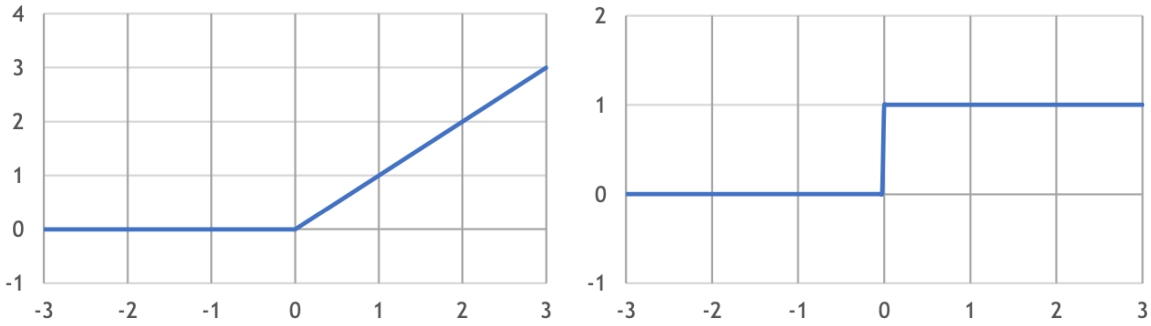
$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{2.2}$$

Hàm này nhận một giá trị đầu vào bất kỳ vào cho giá trị đầu ra biến thiên trong khoảng [0, 1]. Nhờ đặc tính này, hàm *sigmoid* thường được dùng để chuyển đổi các giá trị thực thành xác suất. Khi đầu vào x là số âm có giá trị lớn, đầu ra của hàm *sigmoid* sẽ tiến gần về 0; ngược lại, khi x là số dương có giá trị lớn, đầu ra của hàm sẽ tiến gần về 1. Tại thời điểm $x = 0$, giá trị của hàm

sigmoid là $\sigma(0) = 0.5$. Giá trị này thường được coi là đường biên phân loại đầu ra theo dạng nhị phân.

Một nhược điểm của hàm *sigmoid* là hiện tượng bão hòa xảy ra khi trị tuyệt đối của *x* lớn, làm cho đầu ra tiệm cận gần 0 hoặc 1. Dẫn đến giá trị đạo hàm của hàm *sigmoid* gần như bằng 0, gây ra hiện tượng *vanishing gradient* trong quá trình huấn luyện. Hệ quả là tốc độ hội tụ của mô hình giảm đáng kể, đặc biệt trong các mô hình mạng nơ-ron.

e) Hàm ReLU



Hình 2.7. Hàm ReLU và đạo hàm

Hàm Rectified Linear Unit (ReLU) được định nghĩa như sau:

$$ReLU(x) = \max(0, x) \tag{2.3}$$

Hàm ReLU là một hàm không tuyến tính và không có đạo hàm tại $x = 0$ Tuy nhiên, trong thực tế tính toán, đạo hàm của ReLU tại điểm này thường được mặc định là 0 hoặc 1. ReLU có ưu điểm tính toán đơn giản hơn các hàm như *sigmoid* hay *tanh*, do không yêu cầu phép toán lũy thừa. Theo nghiên cứu, đạo hàm của ReLU bằng 1 khi nơ-ron được kích hoạt, giúp giảm hiện tượng vanishing gradient. Tuy nhiên, nhược điểm của ReLU là gradient bằng 0 khi nơ-ron không kích hoạt [12], dẫn đến khả năng một nơ-ron không bao giờ được điều chỉnh trọng số nếu không kích hoạt ngay từ đầu.

2.3. MẠNG NƠ-RON HỒI QUY

Kiến trúc mạng nơ-ron nhân tạo (Artificial Neural Networks - ANN) được thiết kế để xử lý các dữ liệu đầu vào có số chiều cố định và không yêu cầu thứ tự cụ thể. Tuy nhiên trong thực tế, nhiều loại dữ liệu như chuỗi thời gian hoặc dữ liệu có thứ tự đòi hỏi mô hình có khả năng xử lý số chiều biến đổi và tính phụ thuộc thứ tự. Để giải quyết hạn chế này, mạng nơ-ron hồi quy (Recurrent Neural Networks - RNN) đã được phát triển.

RNN là một biến thể đặc biệt của mạng nơ-ron nhân tạo có cấu trúc liên kết các nút thành đồ thị hướng theo thời gian. Mạng này cho phép truyền tải thông tin từ các bước trước đó như một cơ chế ghi nhớ, giúp xử lý hiệu quả dữ liệu có tính liên kết như văn bản, chuỗi thời gian hay ngôn ngữ nói. Với khả năng điều chỉnh số lượng nơ-ron linh hoạt, RNN có thể xử lý dữ liệu với bất kỳ độ dài nào, vượt trội so với mạng ANN thông thường.

Được lấy cảm hứng từ hoạt động của não người, RNN sử dụng các lớp thời gian liên tiếp để mô phỏng quá trình xử lý thông tin và nhận diện các khuôn mẫu phức tạp. Điều này giúp RNN trở thành một công cụ hiệu quả để dự đoán và xử lý dữ liệu chuỗi, đáp ứng các bài toán mà mạng nơ-ron truyền thống không thể giải quyết một cách tối ưu.

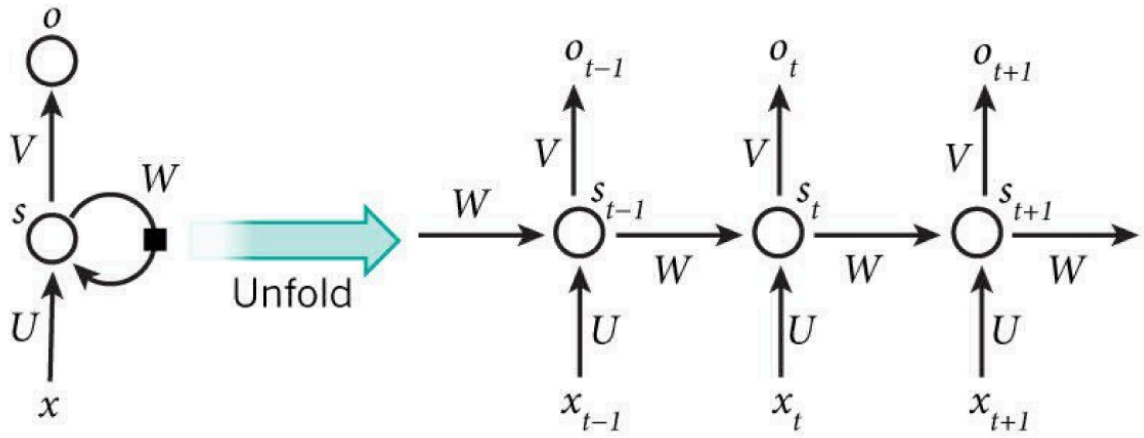
2.3.1. Phân loại

Kiến trúc của mạng nơ-ron hồi quy (RNN) có thể thay đổi linh hoạt tùy theo đặc thù của bài toán cần xử lý, với các cấu trúc phổ biến như sau:

- Một-một (one-to-one): Áp dụng khi có một cặp dữ liệu đầu vào và đầu ra, thường được sử dụng trong các mạng nơ-ron truyền thống.
- Một-nhiều (one-to-many): Một đầu vào duy nhất tạo ra nhiều đầu ra, thường thấy trong các bài toán như sinh nhạc.
- Nhiều-một (many-to-one): Kết hợp nhiều đầu vào qua các mốc thời gian để tạo ra một đầu ra duy nhất, ứng dụng trong nhận dạng cảm xúc hoặc phân tích chuỗi.

- Nhiều-nhiều (many-to-many): Chuyển đổi một chuỗi đầu vào thành chuỗi đầu ra, tiêu biểu trong các hệ thống dịch thuật ngôn ngữ.

2.3.2. Mô hình mạng RNN



Hình 2.8. Mô hình RNN

RNN bao gồm các nơ-ron được kết nối liên tục, cho khả năng duy trì thông tin từ các bước thời gian trước đó nhờ cơ chế lan truyền trạng thái ẩn. Ở mỗi bước *t,* RNN bao gồm đầu vào x_t và trạng thái ẩn s_{t-1} từ bước trước đó. Trạng thái ẩn mới s_t được tính toán như sau:

$$s_t = f(Ux_t + Ws_{t-1}) \quad (2.4)$$

Trong đó:

- W là trọng số liên kết giữa các trạng thái ẩn.
- U là trọng số liên kết giữa đầu vào và trạng thái ẩn.
- f là hàm kích hoạt, thường sử dụng *tanh* hoặc ReLU.

Đầu ra tại bước *t* được tính dựa trên trạng thái ẩn hiện tại:

$$o_t = g(Vs_t) \quad (2.5)$$

Trong đó:

- V là trọng số liên kết giữa trạng thái ẩn và đầu ra.
- g là hàm kích hoạt.

2.3.3. Mở rộng mạng nơ-ron hồi quy

a) Truncated Backpropagation Through Time

Huấn luyện mạng hồi quy truyền thống gặp khó khăn khi chuỗi thời gian quá dài, dẫn đến chi phí tính toán và lưu trữ tăng cao cùng với tốc độ hội tụ chậm. TBPTT, một biến thể của thuật toán Stochastic Gradient Descent (SGD), giải quyết vấn đề này bằng cách giới hạn lan truyền ngược chỉ trong *k* lớp thời gian gần nhất, giảm đáng kể yêu cầu tính toán và bộ nhớ mà vẫn đảm bảo hiệu quả trong xử lý chuỗi dài.

b) Bidirectional Recurrent Neural Network

Một hạn chế lớn của RNN truyền thống là các trạng thái ẩn chỉ có thể nắm bắt thông tin từ quá khứ mà không có khả năng khai thác thông tin từ tương lai. Trong các ứng dụng như phân tích ngữ nghĩa, việc tích hợp thông tin cả trước và sau thời điểm hiện tại cải thiện đáng kể độ chính xác của mô hình. BRNN khắc phục hạn chế này bằng cách học từ cả hai chiều thời gian, ngay cả trong những trường hợp không yêu cầu thông tin tương lai, giúp cải thiện kết quả đáng kể.

c) Multilayer RNN

Kiến trúc nhiều lớp của RNN cho phép mô hình hóa các mối quan hệ phức tạp hơn trong dữ liệu. Trong thực tế, các mô hình với hai đến ba lớp ẩn thường đủ để xử lý hầu hết các bài toán mà không gặp phải hiện tượng overfitting. Tuy nhiên, số lớp ẩn nên được điều chỉnh tương quan với kích thước dữ liệu để đảm bảo hiệu suất tối ưu, đặc biệt trong các bài toán phức tạp đòi hỏi phải phân tích sâu.

2.3.4. Ứng dụng của RNN

Mạng nơ-ron hồi quy đã ghi nhận được nhiều thành công trong nhiều lĩnh vực, đặc biệt trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP – Natural Language Processing).

a) Mô hình hóa ngôn ngữ và sinh văn bản

Mô hình ngôn ngữ có thể hỗ trợ trong việc dự đoán tỉ lệ xuất hiện của một từ cụ thể sau chuỗi các từ đã xuất hiện trước đó. Với khả năng ước lượng độ tương tự giữa các câu, mô hình này còn được ứng dụng rất nhiều trong dịch máy. Khả năng dự đoán từ tiếp theo là khả năng giúp cho mô hình có thể tự sinh từ. Tùy thuộc vào mô hình ngôn ngữ mà có thể tạo ra nhiều văn bản khác nhau. Trong cấu trúc mô hình ngôn ngữ, đầu vào thường là một chuỗi các từ (được biểu diễn bằng vector onehot), trong khi đầu ra là chuỗi các từ được dự đoán. Trong quá trình huấn luyện, gán $o_t = x_{t+1}$ với mục tiêu là đầu ra tại bước t chính là từ tiếp theo của câu.

Từ bài toán này có thể mở rộng thành bài toán sinh văn bản (generating text/generative model) cho phép tạo ra văn bản mới dựa vào tập dữ liệu huấn luyện. Ví dụ, khi huấn luyện mô hình này bằng các văn bản truyện Kiều thì mô hình có thể tạo ra được các đoạn văn tựa truyện Kiều. Tùy theo loại dữ liệu huấn luyện sẽ có nhiều loại ứng dụng khác nhau.

b) Dịch thuật

Dịch thuật là một ứng dụng quan trọng của mạng nơ-ron hồi quy (RNN), đặc biệt trong lĩnh vực xử lý ngôn ngữ tự nhiên (Natural Language Processing - NLP). Về cơ bản, quá trình dịch thuật tương tự với mô hình hóa ngôn ngữ, trong đó đầu vào là một chuỗi các từ thuộc ngôn ngữ nguồn (ví dụ: tiếng Việt), và đầu ra là một chuỗi các từ thuộc ngôn ngữ đích (ví dụ: tiếng Anh). Tuy nhiên, khác với việc tạo ra từ tiếp theo trong mô hình hóa ngôn ngữ, dịch thuật đòi hỏi toàn bộ chuỗi đầu vào phải được xử lý trước khi bắt đầu suy luận và sinh ra chuỗi đầu ra. Lý do là từ đầu tiên trong câu dịch phải dựa trên bối cảnh đầy đủ của chuỗi đầu vào để đảm bảo ý nghĩa chính xác.

Theo nghiên cứu, các mô hình RNN đặc biệt là các biến thể như LSTM (Long Short-Term Memory) và GRU (Gated Recurrent Unit), đã chứng minh hiệu quả trong dịch thuật nhờ khả năng ghi nhớ và xử lý các mối quan hệ dài

hạn trong dữ liệu chuỗi. Nghiên cứu của Bahdanau et al. (2014) đã đề xuất cơ chế chú ý (Attention Mechanism), giúp mạng nơ-ron tập trung vào các thông tin quan trọng của chuỗi đầu vào trong quá trình dịch. Đây là bước ngoặt quan trọng nâng cao hiệu suất dịch thuật, đặc biệt đối với các câu dài và phức tạp.

Các nghiên cứu đã so sánh giữa những phương pháp thống kê truyền thống và phương pháp học sâu chỉ ra rằng mô hình dựa trên RNN mang lại độ chính xác cao hơn đáng kể. Các hệ thống hiện đại như Google Neural Machine Translation (GNMT) đã tận dụng kiến trúc của RNN để cải thiện tối đa chất lượng dịch thuật của chương trình.

2.4. LONG SHORT-TERM MEMORY

Long Short-Term Memory là một loại mạng nơ-ron hồi quy được thiết kế để xử lý các vấn đề liên quan đến chuỗi dữ liệu dài hạn. LSTM giới thiệu bởi Hochreiter và Schmidhuber vào năm 1997, mô hình khắc phục được vấn đề vanishing gradient và exploding gradient trong các mạng RNN truyền thống nhờ cấu trúc đặc biệt của các cổng (gates). Cấu trúc này bao gồm cổng đầu vào (input gate), cổng quên (forget gate) và cổng đầu ra (output gate) phối hợp cùng nhau để lựa chọn và lưu trữ thông tin có ý nghĩa từ các bước thời gian trước đó, đồng thời loại bỏ các thông tin không liên quan. Cơ chế lưu trữ sử dụng ô trạng thái (*cell state*) cho phép LSTM duy trì khả năng học các phụ thuộc dài hạn, làm cho mô hình trở thành công cụ mạnh mẽ trong nhiều ứng dụng.

So với các mô hình truyền thống như ARIMA hay mạng RNN cơ bản, LSTM không yêu cầu các giả định tuyến tính hoặc điều kiện tĩnh, cho phép mô hình áp dụng linh hoạt vào các tập dữ liệu có tính phi tuyến và không ổn định. Điều này đặc biệt quan trọng trong các lĩnh vực như kinh tế, khí hậu hay dự báo chỉ số chất lượng không khí.

Mô hình này đã chứng minh tính hiệu quả trong nhiều ứng dụng nhưng vẫn tồn tại một số hạn chế. Chi phí tính toán cao và khó khăn trong việc tối ưu

hóa các tham số là những thách thức lớn đối với các nhà nghiên cứu. Để khắc phục, nhiều biến thể của LSTM đã được phát triển như GRU (Gated Recurrent Unit) và các kiến trúc lai kết hợp LSTM với các mạng khác như CNN hoặc Transformer đã mở ra triển vọng cải thiện hiệu suất và giảm thiểu thời gian tính toán.

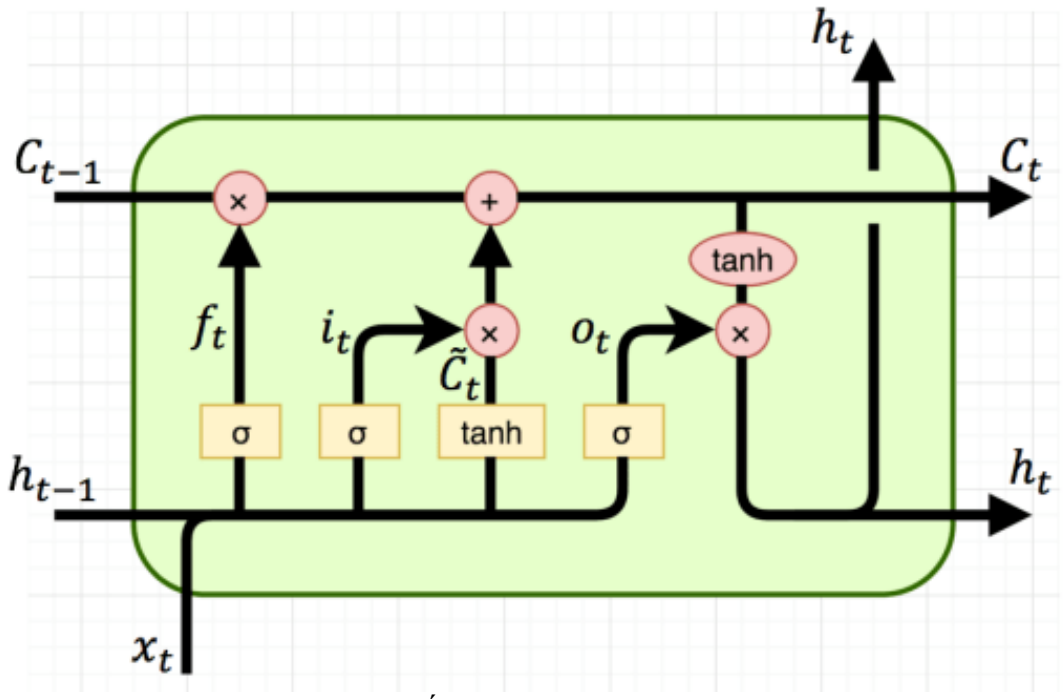
Với khả năng học các mẫu phức tạp từ dữ liệu chuỗi thời gian, LSTM đã vượt qua các giới hạn của các phương pháp truyền thống, thúc đẩy sự phát triển của các công nghệ trí tuệ nhân tạo hiện đại. Sự phát triển liên tục các biến thể của LSTM giúp mô hình này ngày càng mở rộng thêm phạm vi ứng dụng, từ nghiên cứu khoa học đến giải quyết các vấn đề thực tiễn phức tạp trong nhiều lĩnh vực khác nhau.

2.4.1. Mô hình mạng LSTM

LSTM có kiến trúc tương tự với RNN, nhưng các module trong mô hình này có cấu trúc khác với mạng RNN chuẩn. Mô hình LSTM có thể chia thành hai phần tương tác chặt chẽ với nhau là bộ nhớ ngắn hạn (short-term memory) và bộ nhớ dài hạn (long-term memory).

Bộ nhớ ngắn hạn là khả năng lưu trữ thông tin tạm thời, được sử dụng để xử lý và tạo ra các output tại mỗi bước. Bộ nhớ này tương ứng với trạng thái ẩn (h_t) là thông tin được sử dụng để xác định đầu ra hay chuyển tiếp đến các bước tiếp theo.

Bộ nhớ dài hạn lưu trữ thông tin giữa các bước cách xa nhau. Trong LSTM, ô trạng thái (*cell state*) đại diện cho bộ nhớ này hoạt động như một "băng truyền" để đưa thông tin đi qua các nút một cách toàn vẹn giảm thiểu vấn đề vanishing gradient, đảm bảo việc học của mô hình không bị suy giảm khi chuỗi dữ liệu kéo dài.

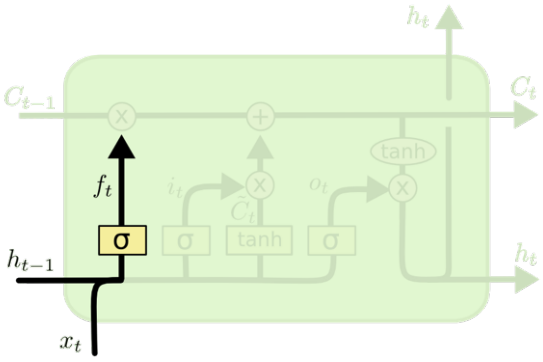


Hình 2.9. Cấu trúc của một nút trong LSTM

Bộ nhớ ngắn hạn cung cấp thông tin để tính toán đầu ra, cập nhật bộ nhớ dài hạn khi cần thiết. Bộ nhớ dài hạn đảm bảo các thông tin quan trọng không bị mất đi khi chuỗi dữ liệu kéo dài. Sự phối hợp này được điều chỉnh thông qua các cổng (*gates*), giúp mô hình quyết định thông tin nào cần lưu trữ hay những thông tin nào cần loại bỏ.

a) Forget Gate

Cổng quên có nhiệm vụ điều chỉnh việc lưu giữ hay loại bỏ thông tin trong bộ nhớ dài hạn. Vai trò chính của cổng này là quyết định thông tin nào từ *cell state* trước đó (C_{t-1}) sẽ được giữ lại và thông tin nào sẽ bị loại bỏ dựa trên dữ liệu tại bước hiện tại.



Hình 2.10. Forget Gate

Cổng quên hoạt động thông qua một hàm *sigmoid*, được viết theo công thức sau:

$$f_t = \sigma\left(W_f * [h_{t-1}, x_t] + b_f\right) \tag{2.6}$$

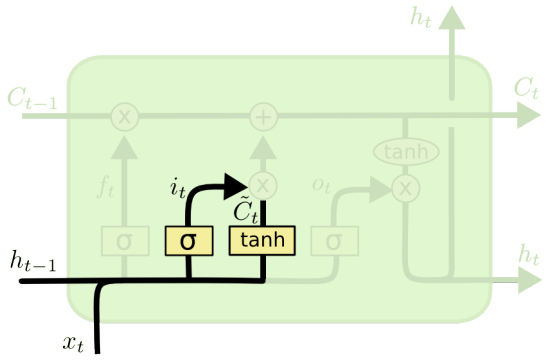
Trong đó:

- σ là hàm kích hoạt *sigmoid*
- W_f là ma trận trọng số của cổng quên
- h_{t-1} là trạng thái ẩn từ bước trước
- x_t là dữ liệu đầu vào tại bước hiện tại
- b_f là hệ số bias

Kết quả f_t có giá trị nằm trong khoảng từ 0 đến 1. Với giá trị 0 nghĩa là loại bỏ toàn bộ thông tin, còn 1 là giữ lại toàn toàn bộ thông tin. Giá trị này sau đó sẽ được nhân với *cell state* trước (C_{t-1}) để xác định phần thông tin được giữ lại.

b) Input Gate

Cổng đầu vào có vai trò xác định thông tin mới nào sẽ được thêm vào bộ nhớ dài hạn tại mỗi bước. Cổng này giúp mô hình LSTM tiếp nhận thông tin mới từ dữ liệu đầu vào (x_t) và trạng thái ẩn trước đó (h_{t-1}) một cách chọn lọc. Nhờ cơ chế này, mô hình có khả năng duy trì và cập nhật thông tin cần thiết để học các phụ thuộc dài hạn trong chuỗi thời gian.



Hình 2.11. Input Gate

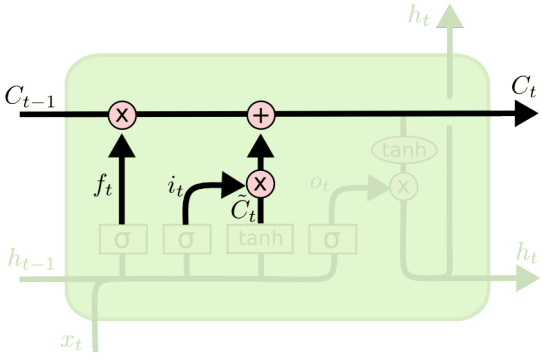
Cổng đầu vào sử dụng hàm *sigmoid* để xác định mức độ thông tin cần thêm vào *cell state* và hàm *tanh* để tạo một vector cho giá trị mới ($\widetilde{C_t}$). Giá trị hai hàm được viết theo công thức sau:

$$i_t = \sigma(W_i * [h_{t-1}, x_t] + b_i) \quad (2.7)$$

$$\tilde{C}_t = \tanh(W_C * [h_{t-1}, x_t] + b_C) \quad (2.8)$$

Trong đó:

- h_{t-1} là trạng thái ẩn từ bước trước
- x_t là dữ liệu đầu vào tại bước hiện tại
- b_i, b_C là các hệ số bias
- W_i, W_C là các ma trận trọng số



Hình 2.12. Quy trình cập nhật cell state

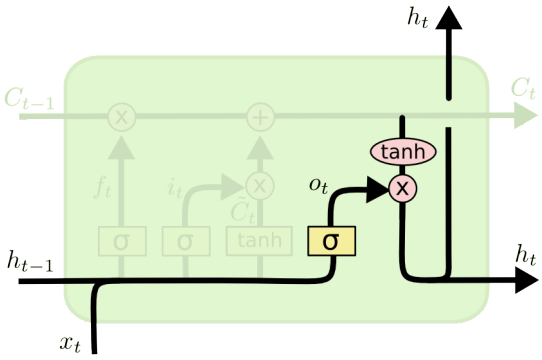
Sau đó *cell state* cũ C_{t-1} được cập nhật thành *cell state* mới C_t. Mô hình thực hiện nhân *cell state* cũ với f_t để bỏ đi những thông tin quyết định quên.

Giá trị này được cộng thêm với $i_t * \widetilde{C}_t$ để tạo thành *cell state* mới. Quá trình cập nhật được biểu diễn theo công thức sau:

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (2.9)$$

c) Output Gate

Cổng đầu ra chịu trách nhiệm tính toán trạng thái ẩn tại mỗi bước. Đây là phần cuối cùng trong chuỗi xử lý của một *time step*, nơi các thông tin từ *cell state* được lọc và chuyển đổi thành đầu ra phù hợp để phục vụ cho bài toán hoặc truyền sang nút tiếp theo. Cổng đầu ra đảm bảo rằng chỉ những thông tin quan trọng nhất từ *cell state* mới được chuyển tiếp, giúp giảm thiểu nhiễu và cải thiện khả năng dự đoán của mô hình.



Hình 2.13. Output Gate

Cổng đầu ra hoạt động qua hai giai đoạn để chuyển đổi thông tin từ *cell state* thành trạng thái ẩn. Đầu tiên, cổng đầu ra tính toán giá trị o_t theo công thức:

$$o_t = \sigma(W_o * [h_{t-1}, x_t] + b_o) \quad (2.10)$$

Trong đó:

- W_o là ma trận trọng số của cổng đầu ra
- b_o là hệ số bias của cổng đầu ra

Sau đó, trạng thái ẩn h_t tại bước hiện tại được tính bằng cách kết hợp *cell state* (C_t) và giá trị o_t. Hàm *tanh* được sử dụng để điều chỉnh giá trị của *cell state*, sau đó nhân với o_t để lọc thông tin cần thiết:

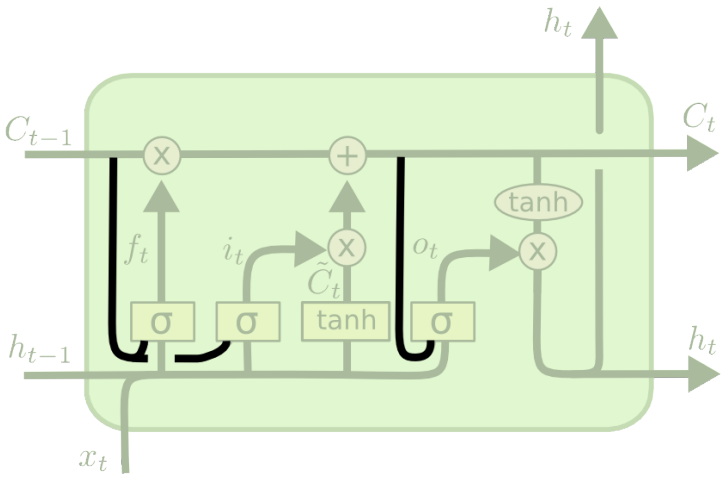
$$h_t = o_t * \tanh(C_t) \quad (2.11)$$

2.4.2. Biến thể của mô hình LSTM

Mô hình LSTM đã chứng minh hiệu quả trong việc giải quyết các bài toán liên quan đến chuỗi thời gian dài và xử lý các phụ thuộc xa, những hạn chế trong cấu trúc chuẩn của mô hình đã thúc đẩy sự phát triển của các biến thể khác nhau. Các biến thể này được thiết kế nhằm tối ưu hóa khả năng học và giảm bớt các vấn đề như độ phức tạp khi tính toán.

a) Peephole

Peephole LSTM là một cải tiến của kiến trúc LSTM tiêu chuẩn, được giới thiệu bởi Gers và Schmidhuber vào năm 2000. Điểm khác biệt của Peephole LSTM nằm ở các cổng trong mô hình (cổng quên, cổng đầu vào và cổng đầu ra) được phép kết nối trực tiếp với *cell state* (C_t). Nhờ đó, các cổng có khả năng truy cập trực tiếp vào thông tin lưu trữ trong *cell state*, thay vì chỉ dựa vào trạng thái ẩn (h_t) và đầu vào hiện tại (x_t).

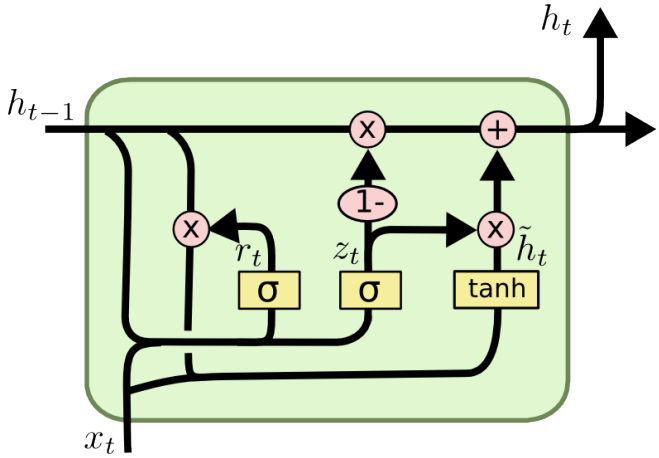


Hình 2.14. Mô hình Peephole LSTM

b) Gated Recurrent Unit

Mô hình GRU được giới thiệu bởi Cho et al. (2014) có thể xem là một biến thể thế hệ mới của LSTM, về cấu trúc GRU đơn giản chỉ với hai cổng: cổng khởi tạo (*reset gate*) và cổng cập nhật (*update gate*). Với thiết kế đơn giản

như vậy mô hình GRU giảm số lượng tham số so với LSTM làm cho mô hình nhanh hơn trong việc huấn luyện và dự báo. GRU phù hợp với tập dữ liệu nhỏ, đòi hỏi tốc độ học nhanh trong quá trình huấn luyện và giảm thiểu nguy cơ *overfitting*.



Hình 2.15. Cấu trúc mô hình GRU

Nguyên tắc hoạt động của GRU: Cổng khởi tạo r_t kiểm soát mức độ hợp nhất thông tin trạng thái hiện tại với thông tin trạng thái trước đó và cổng cập nhật z_t (kết hợp của cổng đầu vào và cổng quên) kiểm soát xem thông tin trạng thái tại thời điểm trước đó có được giữ lại hay không và mức độ lưu giữ ở trạng thái hiện tại. Nếu đầu ra thời điểm trước đó là h_{t-1} và đầu vào tại thời điểm hiện tại là x_t, quá trình tính toán đơn vị lớp ẩn h_t dựa trên GRU được viết dưới dạng phương trình như sau:

$$r_t = \sigma(W_r[h_{t-1}, x_t] + b_r) \quad (2.12)$$

$$z_t = \sigma(W_z[h_{t-1}, x_t] + b_z) \quad (2.13)$$

$$\tilde{h}_t = tanh(W_{\tilde{h}}[r_t * h_{t-1}, x_t] + b_{\tilde{h}}) \quad (2.14)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \quad (2.15)$$

Trong đó:

- $W_r, W_z, W_{\tilde{h}}$ lần lượt là ma trận trọng số của $r_t, z_t, \tilde{h}_t$.
- $b_r, b_z, b_{\tilde{h}}$ lần lượt là hệ số bias của $r_t, z_t, \tilde{h}_t$.

c) CNN-LSTM

Mô hình CNN-LSTM là một biến thể lai mạnh mẽ kết hợp mạng nơ-ron tích chập (CNN) và mạng nơ-ron hồi quy dài ngắn hạn (LSTM), tận dụng các ưu điểm của cả hai để xử lý dữ liệu một cách hiệu quả. CNN đóng vai trò trích xuất các đặc trưng không gian quan trọng từ dữ liệu đầu vào, trong khi LSTM được sử dụng để phân tích và dự báo các mối quan hệ phụ thuộc dài hạn trong chuỗi thời gian.

Trong thiết kế của CNN-LSTM, các lớp CNN đầu tiên xử lý dữ liệu đầu vào, như ảnh hoặc chuỗi thời gian đa chiều, để phát hiện các mẫu không gian phức tạp. Đầu ra từ CNN sau đó được chuyển qua các lớp LSTM để nắm bắt các quan hệ theo thời gian, giúp tối ưu hóa khả năng dự báo. Sự kết hợp này làm cho CNN-LSTM đặc biệt phù hợp trong các bài toán như dự báo chất lượng không khí, dự đoán thời tiết, xử lý tín hiệu cảm biến và phân tích video.

Theo nghiên cứu thực nghiệm, CNN-LSTM đã cho thấy hiệu quả vượt trội hơn so với việc sử dụng riêng CNN hoặc LSTM. Trong bài toán dự báo chỉ số chất lượng không khí, mô hình này có thể xử lý đồng thời các đặc trưng không gian từ nhiều trạm quan trắc và mối quan hệ thời gian giữa các chuỗi dữ liệu, mang lại kết quả dự báo chính xác hơn. Một nghiên cứu điển hình đã cho thấy CNN-LSTM đạt được độ chính xác cao hơn so với các phương pháp truyền thống như ARIMA hoặc mô hình mạng nơ-ron đơn lẻ, nhờ khả năng trích xuất đặc trưng tốt và xử lý thông tin đa chiều.

Ưu điểm của CNN-LSTM bao gồm khả năng xử lý dữ liệu lớn, giảm nguy cơ quá tải thông số, và tính linh hoạt trong ứng dụng với nhiều loại dữ liệu đầu vào. Tuy nhiên, nhược điểm chính của nó là yêu cầu tài nguyên tính toán cao và thời gian huấn luyện lâu hơn so với các mô hình đơn giản hơn. Với sự phát triển của công nghệ phần cứng và các thuật toán tối ưu hóa, CNN-

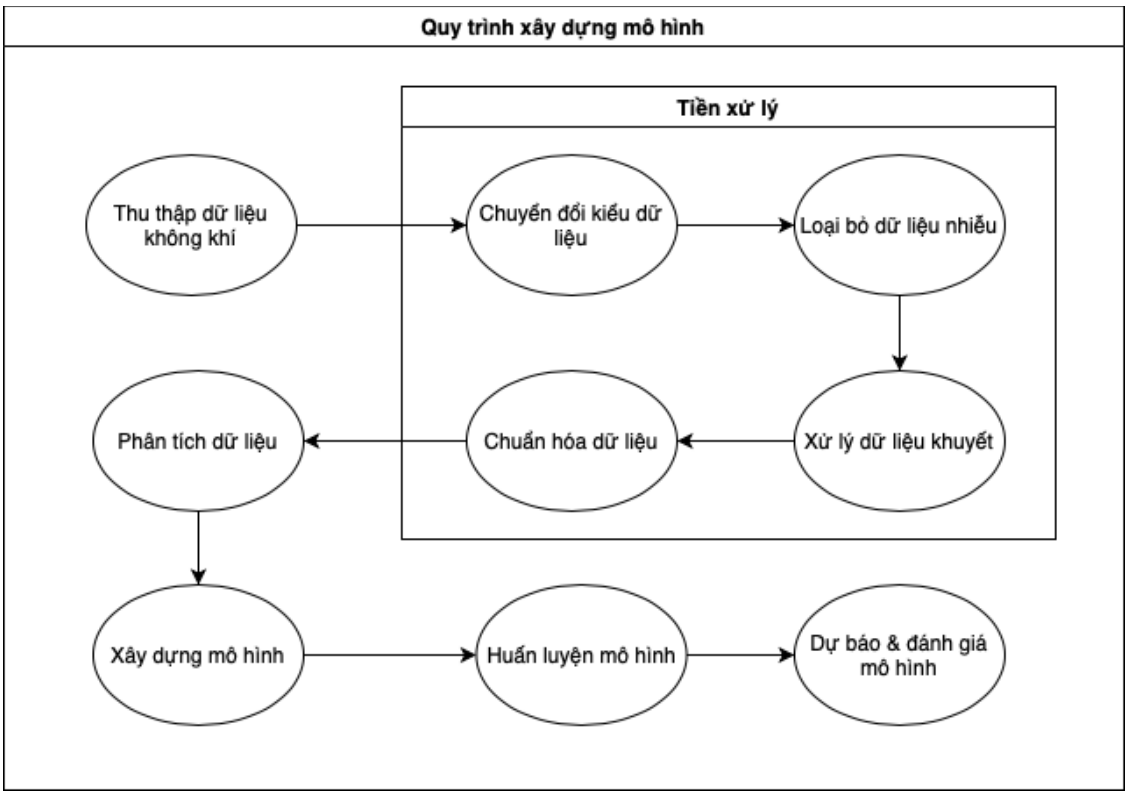
LSTM tiếp tục trở thành lựa chọn hàng đầu trong các bài toán yêu cầu mô hình hóa phức tạp và dự báo chính xác.

2.5. KẾT LUẬN CHƯƠNG

Chương 2 đã trình bày cơ sở lý thuyết làm nền tảng cho nghiên cứu, bao gồm các thuật toán học máy và mạng nơ-ron từ mạng nơ-ron nhân tạo (ANN) đến các kiến trúc phức tạp hơn như RNN và LSTM. Dựa vào thông tin đã được tìm hiểu ở chương này, chương tiếp theo của đề án sẽ thực hiện việc lấy dữ liệu và xây dựng mô hình dự báo cho bài toán.

CHƯƠNG 3: XÂY DỰNG VÀ ĐÁNH GIÁ MÔ HÌNH DỰ BÁO CHỈ SỐ CHẤT LƯỢNG KHÔNG KHÍ

Trong chương ba sẽ trình bày chi tiết về quá trình xây dựng mô hình dự báo chỉ số chất lượng không khí. Sơ đồ tổng quan của quy trình này được mô tả trong hình 3.1.



Hình 3.1. Quy trình xây dựng mô hình

Quy trình xây dựng mô hình dự báo chất lượng không khí được chia thành các bước như sau:

- Thu thập dữ liệu về chất lượng không khí được thu thập từ các nguồn tin cậy.
- Tiền xử lý tập dữ liệu thu thập được như loại bỏ dữ liệu nhiễu, xử lý giá trị khuyết đảm bảo tính nhất quán cho tập dữ liệu.
- Phân tích dữ liệu để khám phá các xu hướng và đặc điểm quan trọng.
- Huấn luyện mô hình trên tập dữ liệu xử lý, sau đó thực hiện dự báo và đánh giá kết quả để dựa trên độ chính xác, hiệu quả của mô hình.

3.1. KHẢO SÁT VÀ THU THẬP BỘ DỮ LIỆU

Trong quá trình thực hiện nghiên cứu dự báo chỉ số chất lượng không khí tại Nha Trang, việc khảo sát và hiểu rõ bộ dữ liệu thực tế là một bước quan trọng. Dữ liệu sử dụng trong nghiên cứu này bao gồm các thông số liên quan đến chất lượng không khí như nồng độ các chất ô nhiễm ($PM_{2.5}$, PM_{10}, CO, NO_2, SO_2, O_3), nhiệt độ, độ ẩm, và một số yếu tố thời tiết khác được thu thập từ các trạm quan trắc môi trường tại Nha Trang.

3.1.1. Nguồn gốc và đặc điểm dữ liệu

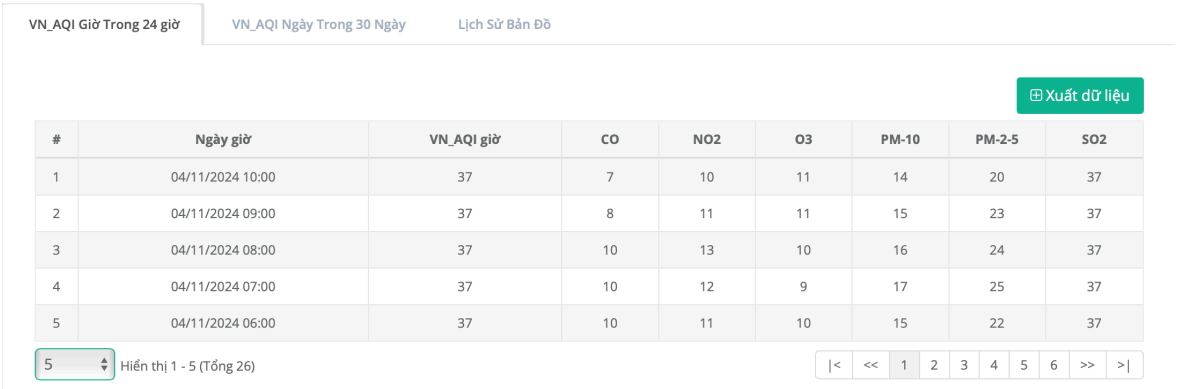
Dữ liệu được thu thập từ phần mềm quản lý dữ liệu quan trắc tự động (Envisoft) được nghiên cứu và phát triển bởi Trung tâm Quan trắc môi trường miền Bắc thuộc Tổng Cục Môi trường – Bộ Tài nguyên và Môi trường**.** Phần mềm Envisoft đã giải quyết được các yêu cầu về: thu thập, xử lý dữ liệu lớn; giám sát dữ liệu, tình trạng thiết bị theo thời gian thực; điều khiển và giám sát lấy mẫu; kiểm duyệt dữ liệu; quản lý truyền nhận dữ liệu giữa Trung ương và địa phương; khai thác, công bố và chia sẻ dữ liệu**.**

Bộ dữ liệu bao gồm các thông số về không khí được thu thập với tần suất theo giờ và theo ngày, giúp theo dõi liên tục các chỉ số chất lượng không khí trong suốt giai đoạn nghiên cứu. Dữ liệu có tính chất chuỗi thời gian, với mỗi quan sát bao gồm nhiều biến độc lập, đóng vai trò quan trọng trong quá trình xây dựng và huấn luyện mô hình dự báo.

3.1.2. Thu thập dữ liệu

Thu thập dữ liệu là bước đầu tiên và quan trọng trong việc xây dựng mô hình. Dữ liệu có thể được lấy từ các cảm biến đo lường tại chỗ, các trạm quan trắc hoặc từ các cơ sở dữ liệu công khai như các tổ chức môi trường. Chất lượng và độ chính xác của dữ liệu thu thập đóng vai trò quyết định đến hiệu quả và độ tin cậy của mô hình.

a) Thu thập dữ liệu trên giao diện



Hình 3.2. Dữ liệu chất lượng không khí dùng để thực nghiệm

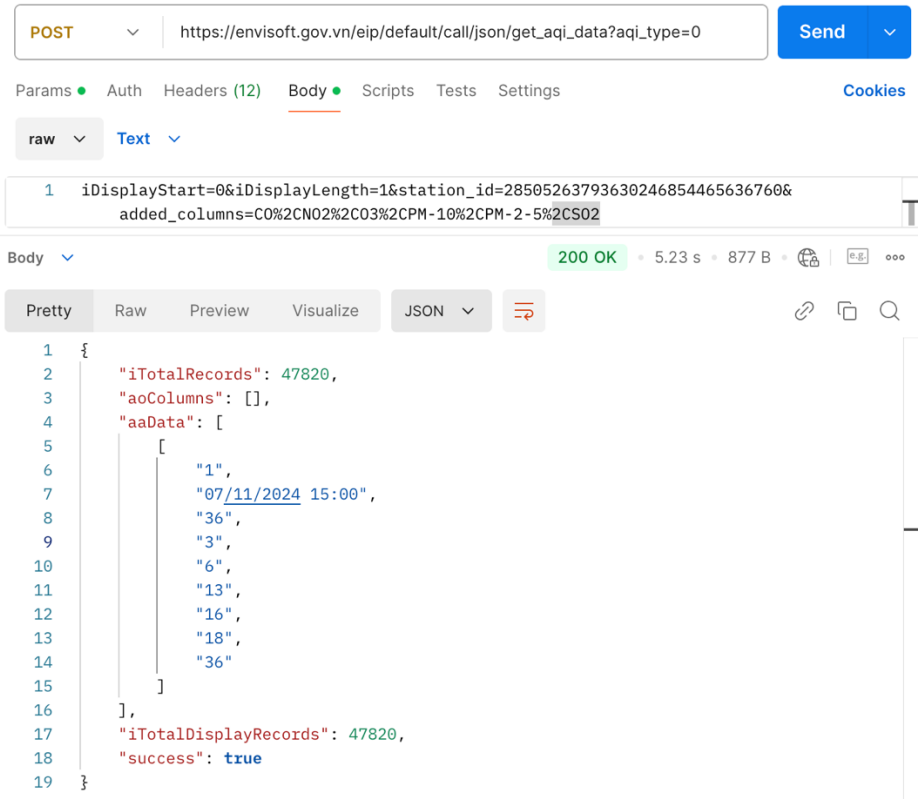
Envisoft cung cấp dữ liệu chất lượng không khí, bao gồm các chỉ số quan trọng như VN_AQI theo giờ, CO, NO_2, O_3, PM_{10}, $PM_{2.5}$ và SO_2 trong khung thời gian cụ thể, giúp người dùng theo dõi và đánh giá tình trạng ô nhiễm không khí. Tuy nhiên, website này có một số hạn chế nhất định trong việc hiển thị dữ liệu, làm giảm khả năng tiếp cận thông tin của người dùng.

Hình 3.2 hiển thị bảng dữ liệu cung cấp các chỉ số cụ thể cho mỗi giờ, trang web chỉ hiển thị giới hạn 26 bản ghi theo giờ tính từ thời điểm tra cứu. Đồng thời trang web không cung cấp các tùy chọn tìm kiếm hay lọc trên bảng dữ liệu gây khó khăn cho việc thu thập dữ liệu giữa các thời điểm hoặc tìm kiếm các giá trị cụ thể nhanh chóng. Ngoài ra, mặc dù có nút "Xuất dữ liệu," website chỉ cho phép người dùng truy xuất thông tin được giới hạn như trên giao diện.

b) Thu thập dữ liệu thông qua web api

Để thu thập dữ liệu chất lượng không khí từ Envisoft bị hạn chế trên giao diện, em đã thực hiện một yêu cầu POST đến API của trang này thông qua công cụ Postman nhằm truy xuất các thông tin cần thiết. Cụ thể, yêu cầu được gửi tới URL https://envisoft.gov.vn/eip/default/call/json/get_aqi_data?aqi_type=1,

với tham số aqi_type=1, nhằm lấy dữ liệu các chỉ số chất lượng không khí theo giờ.



Hình 3.3. Thu thập dữ liệu thông qua web api

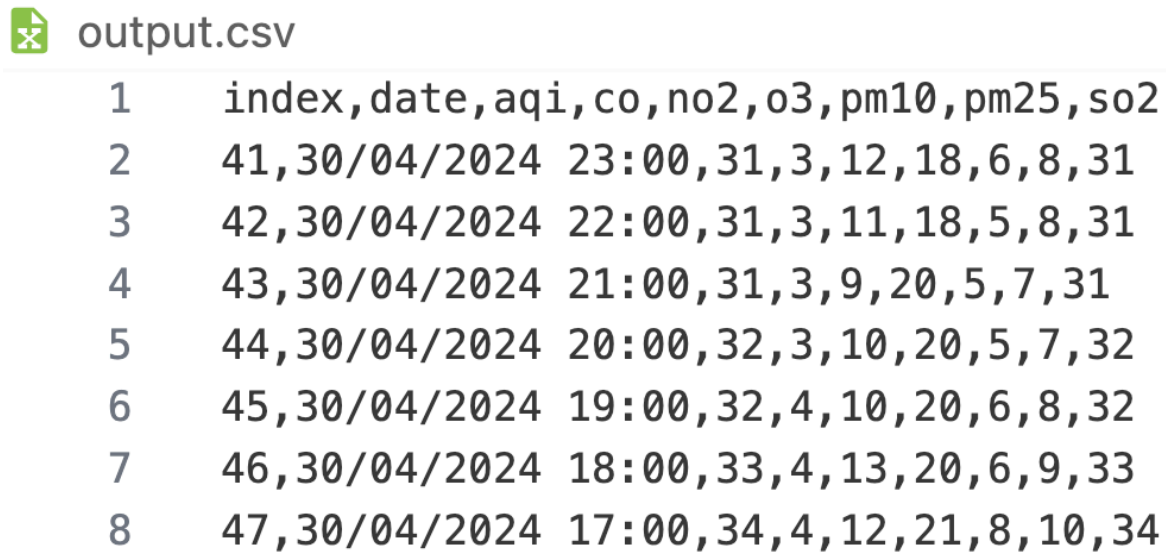
Trong phần body của yêu cầu, em sử dụng định dạng dữ liệu là raw text và truyền các tham số đặc biệt:

- iDisplayStart là số lượng bản ghi bỏ qua bắt đầu từ 0.
- iDisplayLength là số lượng bản ghi cần lấy.
- station_id là ID của địa điểm cần lấy.
- added_columns là các chỉ số về không khí trả về như CO, NO_2, O_3, PM_{10}, $PM_{2.5}$ và SO_2.

Kết quả phản hồi từ API là một đối tượng JSON với các trường quan trọng như iTotalRecords, aaData, và iTotalDisplayRecords. Trong đó:

- Trường iTotalRecords và iTotalDisplayRecords đều có giá trị là 47,820, cho thấy tổng số bản ghi dữ liệu sẵn có trong hệ thống là 47,820 bản ghi.
- Trường aaData chứa dữ liệu thực tế dưới dạng một mảng. Mỗi phần tử trong mảng aaData bao gồm các chỉ số về chất lượng không khí tại một thời điểm cụ thể. Ví dụ, phần tử đầu tiên chứa thời điểm "04/11/2024 10:00" và các chỉ số như VN_AQI là 37, CO là 7, NO_2 là 10, O_3 là 11, PM_{10} là 14, $PM_{2.5}$ là 20 và SO_2 là 37.

Phản hồi trả về có mã trạng thái HTTP là “200 OK”, điều này khẳng định rằng yêu cầu đã được xử lý thành công và dữ liệu đã được lấy về như mong đợi. Dựa vào đây em có thể thực hiện thay đổi hai tham số iDisplayStart và iDisplayLength để yêu cầu để lấy hết dữ liệu. Thông qua quá trình này, em đã truy xuất một lượng lớn thông tin về chất lượng không khí từ hệ thống, tuy nhiên cần xử lý cẩn thận để đảm bảo dữ liệu được lấy đầy đủ và không vượt quá các giới hạn truy xuất của API.


output.csv

```
1   index,date,aqi,co,no2,o3,pm10,pm25,so2
2   41,30/04/2024 23:00,31,3,12,18,6,8,31
3   42,30/04/2024 22:00,31,3,11,18,5,8,31
4   43,30/04/2024 21:00,31,3,9,20,5,7,31
5   44,30/04/2024 20:00,32,3,10,20,5,7,32
6   45,30/04/2024 19:00,32,4,10,20,6,8,32
7   46,30/04/2024 18:00,33,4,13,20,6,9,33
8   47,30/04/2024 17:00,34,4,12,21,8,10,34
```

Hình 3.4. Dữ liệu chất lượng không khí thô theo giờ dạng csv

Sau quá trình thu thập toàn bộ dữ liệu chất lượng không khí dạng JSON từ hệ thống, bao gồm các chỉ số quan trọng như AQI, CO, NO_2, O_3, PM_{10},

$PM_{2.5}$, và SO_2 theo từng thời điểm cụ thể. Các dữ liệu này sau đó được trích xuất và chuẩn hóa để sắp xếp theo định dạng mong muốn trong file output.csv.

Dữ liệu được chuyển đổi từ dạng JSON thành dạng csv với các cột: index, date, aqi, co, no2, o3, pm10, pm25, và so2. Cột index biểu thị số thứ tự của từng bản ghi, trong khi cột date chứa thông tin thời gian, và các cột còn lại chứa các giá trị của từng chỉ số ô nhiễm không khí.

Quá trình này giúp tổ chức dữ liệu thành một định dạng trực quan và dễ dàng sử dụng cho các bước phân tích hoặc ứng dụng vào các mô hình dự đoán chất lượng không khí.

3.1.3. Đánh giá bộ dữ liệu thô

Bộ dữ liệu được thu bao gồm các chỉ số chất lượng không khí đo được theo giờ tại Nha Trang. Dữ liệu bao gồm 47,723 bản ghi và 8 cột được thu thập từ tháng 01 năm 2019 đến tháng 10 năm 2024 bao gồm thời gian đo, chỉ số chất lượng không khí (AQI), và các thành phần khí gây ô nhiễm (CO, NO_2, O_3, PM_{10}, $PM_{2.5}$, và SO_2).

Tất cả các cột đều được lưu dưới dạng chuỗi gây khó khăn trong việc tính toán và phân tích số liệu. Các giá trị khuyết được thể hiện bằng dấu gạch ngang (-). Số lượng giá trị thiếu khá cao, tập trung vào các chỉ như AQI, PM_{10}, $PM_{2.5}$ chiếm tới gần 50% tập dữ liệu. Phân tích sâu vào số lượng giá trị thiếu, em nhận ra dữ liệu này rơi vào khoảng thời gian cuối năm 2020 đến năm 2022 – thời điểm đại dịch COVID-19 xâm nhập và có ảnh hưởng nghiêm trọng tại Việt Nam.

Ngoài dữ liệu bị thiếu do ảnh hưởng bởi đại dịch, bộ dữ liệu cũng có những giá trị bất thường, đột biến. Ví dụ: nồng độ bụi PM_{10} và $PM_{2.5}$ cao đột biến đến 500 dẫn đến sai lệch dữ liệu tính toán AQI. Những dữ liệu này có thể bị đo lường thiếu chính xác do thiết bị hoặc hệ thống xử lý lỗi gây ra.

3.2. TIỀN XỬ LÝ DỮ LIỆU

Sau khi đánh giá bộ dữ liệu thô được thu thập, cần thực hiện tiền xử lý dữ liệu. Mục đích của bước này nhằm đảm bảo chất lượng và độ tin cậy của dữ liệu trước khi áp dụng các mô hình phân tích hoặc dự đoán. Bằng cách làm sạch dữ liệu, xử lý các giá trị thiếu và nhiễu, chuẩn hóa và biến đổi dữ liệu, giúp giảm thiểu sai số tiềm ẩn, làm cho dữ liệu trở nên nhất quán và dễ dàng phân tích hơn, làm cho mô hình học máy hoạt động hiệu quả, tăng độ chính xác, cho ra kết quả phân tích có ý nghĩa và đáng tin cậy hơn.

a) Chuyển đổi kiểu dữ liệu

	date	aqi	co	no2	o3	pm10	pm25	so2
0	03/11/2024 14:00	34	2	3	19	7	11	34
1	03/11/2024 13:00	35	2	2	21	9	12	35
2	03/11/2024 12:00	35	2	2	20	10	14	35
3	03/11/2024 11:00	35	3	4	17	13	19	35
4	03/11/2024 10:00	36	3	4	17	16	24	36

Hình 3.5. Dữ liệu trước khi chuyển đổi kiểu dữ liệu

Đầu tiên sau khi đọc dữ liệu thô em thực hiện chuyển đổi các giá trị kiểu chuỗi này thành các kiểu dữ liệu phù hợp. Cột date chứa thông tin ngày giờ ở dạng chuỗi, nên cần chuyển đổi thành kiểu datetime giúp dễ dàng thao tác trên dữ liệu theo thời gian, như phân tích theo ngày, giờ hoặc tháng.

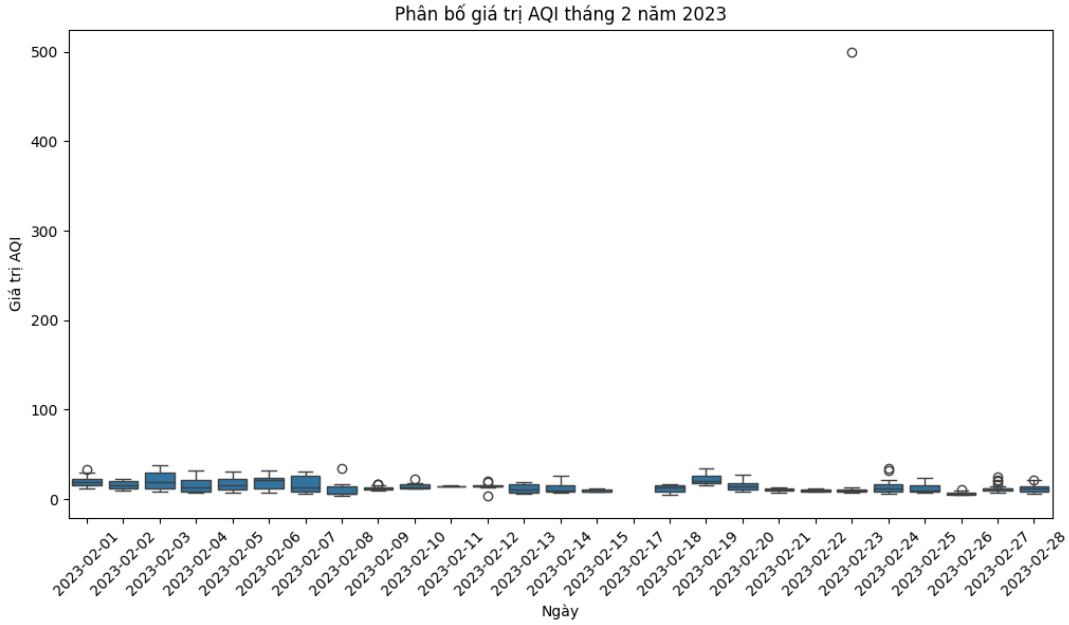
date	aqi	co	no2	o3	pm10	pm25	so2
2024-11-03 14:00:00	34.0	2.0	3.0	19.0	7.0	11.0	34.0
2024-11-03 13:00:00	35.0	2.0	2.0	21.0	9.0	12.0	35.0
2024-11-03 12:00:00	35.0	2.0	2.0	20.0	10.0	14.0	35.0
2024-11-03 11:00:00	35.0	3.0	4.0	17.0	13.0	19.0	35.0
2024-11-03 10:00:00	36.0	3.0	4.0	17.0	16.0	24.0	36.0

Hình 3.6. Dữ liệu sau khi chuyển đổi kiểu dữ liệu

Các cột chứa thông số ô nhiễm không khí như AQI, CO, NO_2, O_3, PM_{10}, $PM_{2.5}$, SO_2 cần được chuyển đổi từ chuỗi thành số thực để đảm bảo tính toán chính xác khi tiến hành các phân tích thống kê. Việc chuyển đổi này cũng giúp phát hiện và xử lý các giá trị không hợp lệ, thường là các giá trị bị nhập sai hoặc thiếu. Những giá trị không hợp lệ sẽ được chuyển thành NaN để xử lý tiếp theo, giúp dữ liệu đầu vào của mô hình đạt tính nhất quán cao.

b) Loại bỏ dữ liệu nhiễu

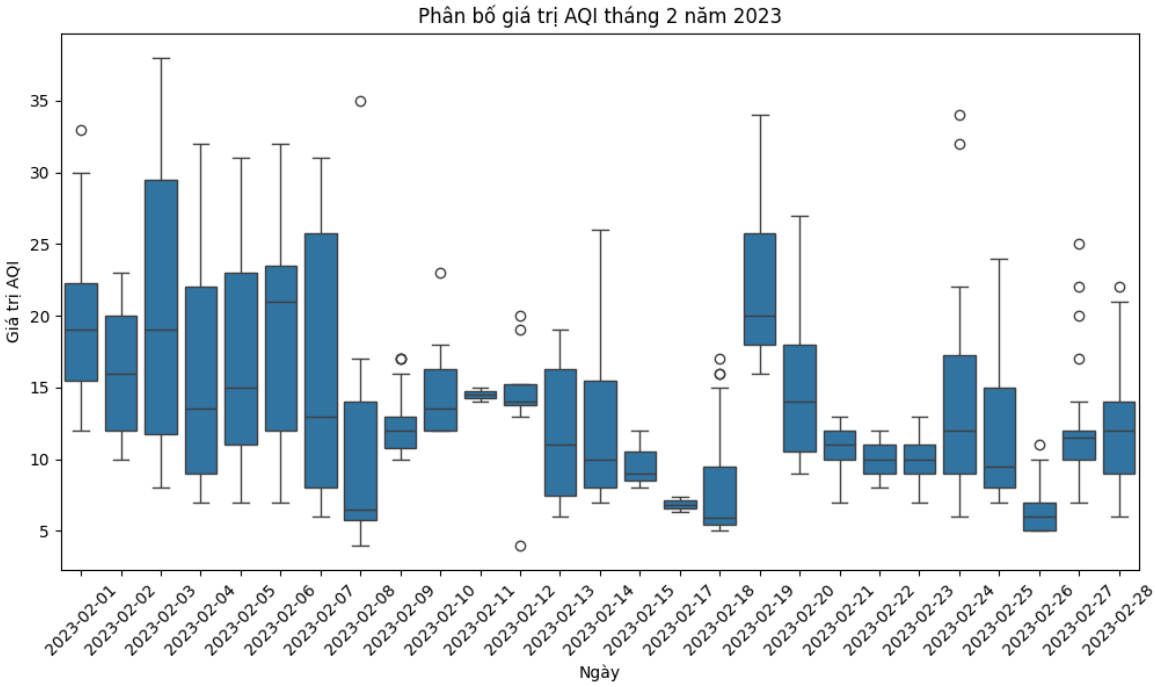
Dữ liệu nhiễu (outliers) trong chuỗi thời gian là những giá trị lệch so với phần còn lại của dữ liệu, có thể gây ra sự sai lệch trong các phân tích và dự báo. Những giá trị này thường phát sinh từ các yếu tố ngoại lai như sự cố kỹ thuật, biến động kinh tế đột ngột hoặc các sự kiện bất thường như thiên tai hoặc dịch bệnh. Nếu không được xử lý đúng cách, outliers có thể làm giảm độ chính xác và độ tin cậy của mô hình dự báo. Vì vậy, việc phát hiện và loại bỏ hoặc điều chỉnh những giá trị này là cần thiết để nâng cao hiệu quả của các mô hình phân tích dữ liệu.



Hình 3.7. Biểu đồ hộp giá trị AQI gây nhiễu

Dựa theo đánh giá thống kê em nhận thấy dữ liệu của 3 giá trị AQI, PM_{10}, $PM_{2.5}$ bị thiếu rất lớn trong khoảng thời gian đại dịch COVID-19, dữ liệu này có được điền khuyết bằng nhiều phương pháp khác nhau nhưng sai số rất cao, không có nhiều giá trị cho mô hình vì vậy em sẽ thực hiện loại bỏ tất cả dữ liệu này và tập trung giữ lại dữ liệu trong hai năm 2023 và 2024.

Nhìn trên biểu đồ hình 3.7, chỉ số AQI có giá trị đạt 500, đây là một giá trị bất thường, không hợp lý gây nhiễu đến tập dữ liệu. Em sẽ thực hiện loại bỏ tất cả những dữ liệu có giá trị gây nhiễu này. Quá trình loại bỏ dữ liệu nhiễu giúp giảm thiểu các tác động tiêu cực mà dữ liệu không chính xác có thể gây ra cho mô hình.



Hình 3.8. Biểu đồ hộp giá trị outliers của AQI đã qua xử lý

c) Xử lý các giá trị bị khuyết

Khi gặp phải dữ liệu thiếu trong chuỗi thời gian, có hai phương pháp phổ biến để xử lý: nội suy tuyến tính (interpolate) và ngoại suy tuyến tính (extrapolate). Nội suy tuyến tính được sử dụng để ước lượng giá trị thiếu trong phạm vi dữ liệu đã có, trong khi ngoại suy tuyến tính dự đoán giá trị ở ngoài

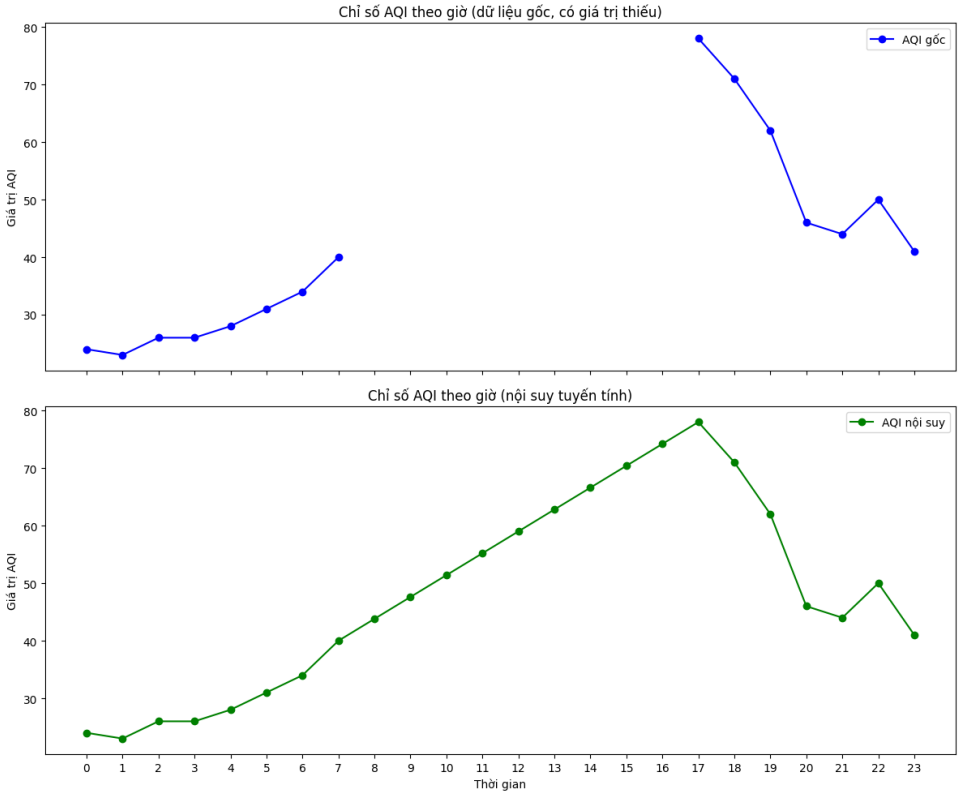
phạm vi dữ liệu hiện tại. Cả hai phương pháp này đều giúp duy trì tính liên tục của chuỗi thời gian và giảm thiểu ảnh hưởng của dữ liệu thiếu đến kết quả phân tích. Với tập dữ liệu thu thập em sẽ dùng phương pháp nội suy tuyến tính để xử lý.

Giả sử có một chuỗi thời gian $x_1, x_2, \dots, x_n$ trong đó x_i là giá trị quan sát tại thời điểm t_i. Nếu dữ liệu bị thiếu tại thời điểm t_k (x_k không xác định), nhưng các giá trị x_{k-1} tại t_{k-1} và x_{k+1} tại t_{k+1} đều được biết, thì giá trị x_k có thể được ước lượng bằng nội suy tuyến tính như sau:

$$x_k = x_{k-1} + \frac{(t_k - t_{k-1})}{(t_{k+1} - t_{k-1})} * (x_{k+1} - x_{k-1}) \quad (3.1)$$

Trong đó:

- x_{k-1} và x_{k+1} : lần lượt là giá trị liền trước và liền sau của x_k
- t_k: thời điểm mà giá trị bị thiếu



Hình 3.9. Dữ liệu trước và sau khi áp dụng nội suy tuyến tính

d) Chuẩn hóa dữ liệu

Dữ liệu trước khi được đưa vào trong mô hình huấn luyện cần được chuẩn hóa nhằm biến đổi các giá trị của mỗi trường về cùng một tỷ lệ, giúp mô hình học máy hội tụ nhanh hơn và làm tăng hiệu suất của mô hình. Trong đề tài sẽ sử dụng phương pháp chuẩn hóa dữ liệu Min-Max Scaling đưa dữ liệu về khoảng [0 – 1] theo công thức:

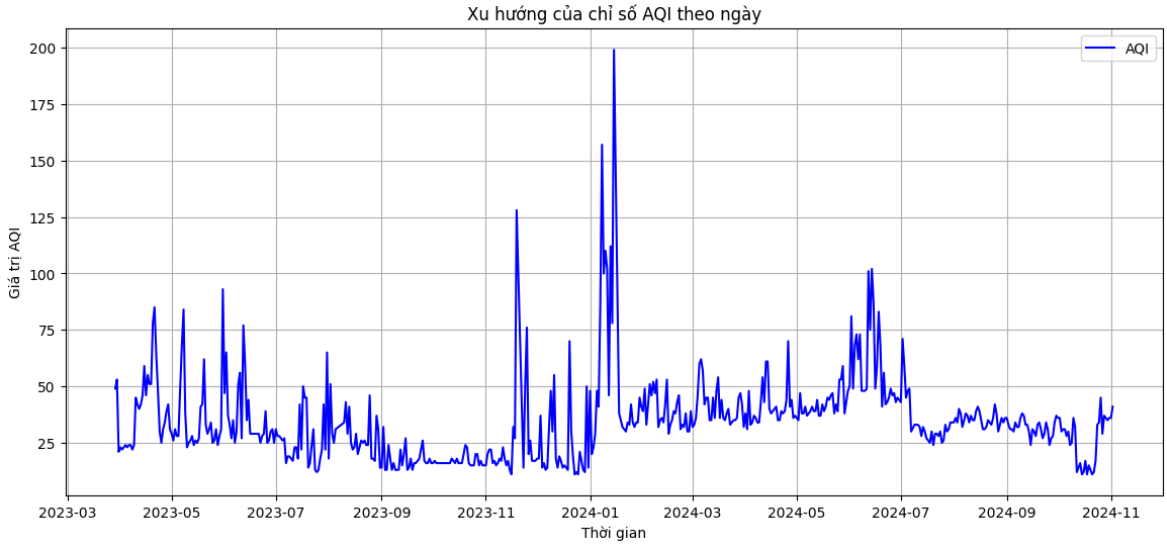
$$X_{norm} = \frac{x - x_{\min}}{x_{max} - x_{min}} \quad (3.2)$$

Trong đó:

- x: giá trị ban đầu
- x_{min}, x_{max}: giá trị nhỏ nhất và lớn nhất
- x_{norm}: giá trị chuẩn hóa sau khi áp dụng Min-Max Scaling

3.3. PHÂN TÍCH XU HƯỚNG

a) Xu hướng chỉ số AQI theo ngày



Hình 3.10. Biểu đồ xu hướng chỉ số AQI theo ngày

Biểu đồ xu hướng chỉ số AQI theo ngày cho thấy sự biến động trong chất lượng không khí tại khu vực nghiên cứu từ tháng 3 năm 2023 đến tháng 11 năm 2024. Dữ liệu biểu diễn các chu kỳ tăng giảm khác nhau của chỉ số AQI, cho

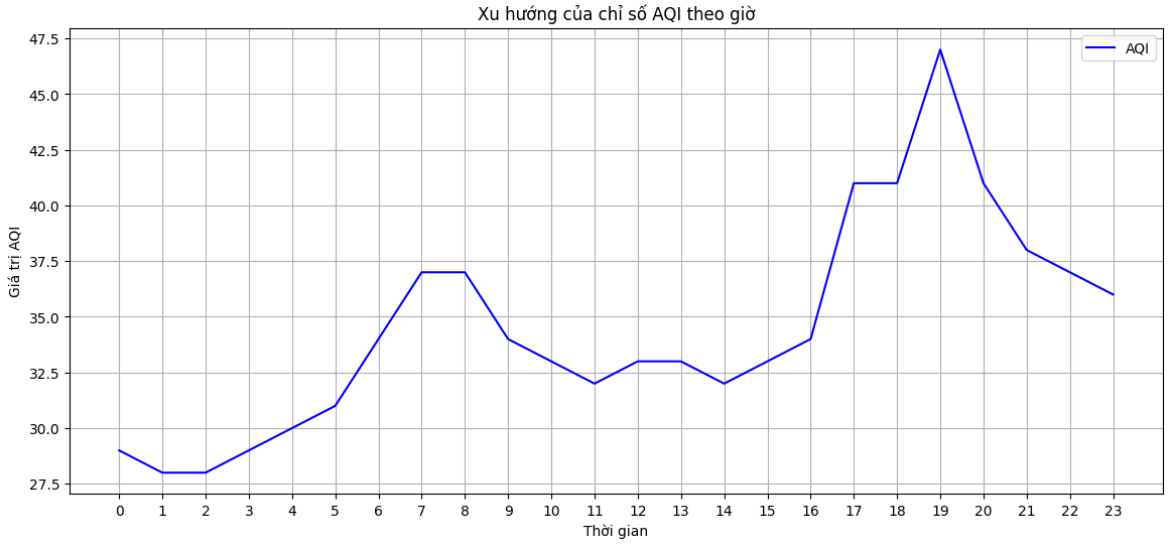
thấy mối liên hệ giữa các yếu tố thời tiết, hoạt động kinh tế xã hội và mùa du lịch đối với tình trạng ô nhiễm không khí.

Giai đoạn từ tháng 3 đến tháng 6 năm 2023, chỉ số AQI có sự dao động mạnh với các đỉnh cao, đặc biệt vào những tháng đầu mùa hè (tháng 5, 6, và 7). Đây cũng là thời điểm cao điểm của mùa du lịch, khi lượng du khách gia tăng đáng kể có thể làm tăng lưu lượng giao thông, hoạt động công cộng và thương mại. Những hoạt động này dễ dẫn đến sự gia tăng các chất ô nhiễm không khí, đặc biệt là các loại khí thải từ phương tiện giao thông và hoạt động công nghiệp. Trong khoảng thời gian này, các đợt tăng đột biến của AQI đạt mức trên 75, cho thấy tình trạng ô nhiễm đáng lo ngại và có khả năng ảnh hưởng tiêu cực đến sức khỏe của cư dân địa phương cũng như du khách.

Bước sang giai đoạn cuối năm 2023, biểu đồ ghi nhận một đợt tăng mạnh vào tháng 11 với đỉnh AQI gần 100, tiếp đó là một đỉnh cao khác vào khoảng tháng 1 năm 2024, với chỉ số AQI vượt quá 200. Đây có thể là kết quả của các yếu tố thời tiết vào mùa đông khi không khí trở nên lạnh và đặc, làm hạn chế khả năng phân tán của các chất ô nhiễm trong không khí. Hiện tượng nhiệt độ thấp và gió yếu thường khiến các hạt bụi mịn và các khí ô nhiễm tích tụ gần mặt đất, gây ra tình trạng ô nhiễm không khí nghiêm trọng.

Trong năm 2024, mặc dù sau đợt đỉnh cao vào tháng 1, chỉ số AQI giảm dần, nhưng dữ liệu cho thấy xu hướng AQI trung bình có dấu hiệu tăng nhẹ so với năm 2023. Điều này được thể hiện qua sự xuất hiện thường xuyên của các đỉnh AQI vượt mức 50 vào các tháng mùa hè và cuối năm, thay vì duy trì ở mức ổn định thấp như giai đoạn giữa năm 2023. Các đỉnh cao của AQI vào các tháng 5, 6, và 7 năm 2024 tiếp tục cho thấy tác động của mùa du lịch và các hoạt động công cộng. Xu hướng ô nhiễm có dấu hiệu tăng cao hơn so với năm trước, nhấn mạnh sự cần thiết phải tăng cường các biện pháp quản lý chất lượng không khí, đặc biệt là vào các thời điểm dễ xảy ra tình trạng ô nhiễm.

b) Xu hướng chỉ số AQI theo giờ



Hình 3.11. Biểu đồ xu hướng chỉ số AQI theo giờ

Biểu đồ hình 3.11 mô tả xu hướng chỉ số chất lượng không khí theo giờ trong ngày. Quan sát từ biểu đồ, em thấy rằng giá trị AQI có xu hướng thay đổi rõ rệt theo từng khoảng thời gian trong ngày, thể hiện qua các đỉnh và đáy của đồ thị.

Trong khoảng thời gian từ 0 giờ đến 5 giờ sáng, chỉ số chất lượng không khí có giá trị thấp, dao động quanh mức 30. Điều này có thể phản ánh tình trạng không khí trong lành hơn vào ban đêm, khi hoạt động của con người và các nguồn phát thải bị hạn chế. Sau đó, từ 6 giờ sáng trở đi, chỉ số AQI bắt đầu tăng, đạt mức cao nhất trong khoảng từ 18 giờ đến 19 giờ với chỉ số khoảng 47.5. Điều này do các hoạt động giao thông và sinh hoạt của con người tăng cao vào thời điểm này trong ngày, làm tăng lượng khí thải và các chất ô nhiễm trong không khí.

Sau giờ cao điểm, chỉ số chất lượng không khí giảm dần vào buổi tối, có xu hướng ổn định trở lại vào cuối ngày. Xu hướng này cho thấy ảnh hưởng của các yếu tố thời gian trong ngày và các hoạt động của con người lên chất lượng không khí.

3.4. XÂY DỰNG MÔ HÌNH

Mô hình dự báo chỉ số AQI được xây dựng trên cơ sở mạng LSTM. Cấu trúc mô hình gồm ba lớp LSTM liên tiếp để học các mối quan hệ giữa các giá trị AQI qua các khoảng thời gian. Các lớp LSTM này sẽ giúp mô hình nhận diện các xu hướng trong chuỗi thời gian và dự đoán giá trị AQI trong tương lai. Cấu trúc chi tiết của mô hình được xây dựng như sau:

- Lớp LSTM đầu tiên: Lớp này nhận dữ liệu đầu vào có size (12, 1). Để giảm thiểu hiện tượng overfitting, lớp này sử dụng kỹ thuật Dropout với tỷ lệ 10%, tức là trong mỗi lần huấn luyện, 10% các đơn vị của lớp này sẽ bị tắt đi.
- Lớp LSTM thứ hai và thứ ba: Các lớp này tiếp tục học các mối quan hệ phức tạp hơn giữa các giá trị AQI trong các khoảng thời gian dài hơn. Mỗi lớp LSTM này cũng đi kèm với lớp Dropout có tỷ lệ 10%.
- Lớp Dense: Cuối cùng, sau ba lớp LSTM, một lớp Dense được sử dụng để tạo ra đầu ra dự đoán duy nhất cho AQI tại thời điểm tiếp theo. Đây là lớp đầu ra của mô hình, với hàm mất mát là Mean Squared Error (MSE), giúp đo lường độ chính xác của mô hình trong việc dự đoán giá trị AQI.

```
model = Sequential()

model.add(LSTM(units = 32, return_sequences = True,
               input_shape=(X_train.shape[1], X_train.shape[2])))
model.add(Dropout(0.1))

model.add(LSTM(units = 32, return_sequences = True))
model.add(Dropout(0.1))

model.add(LSTM(units = 16))
model.add(Dropout(0.1))

model.add(Dense(units = 1))

model.compile(optimizer = 'adam', loss = 'mean_squared_error')
```

Hình 3.12. Xây dựng mô hình LSTM

Mô hình LSTM được huấn luyện sử dụng thuật toán Adam, một phương pháp tối ưu hóa mạnh mẽ trong học sâu. Adam kết hợp ưu điểm của Gradient Descent và Momentum, giúp mô hình hội tụ nhanh hơn trong quá trình huấn luyện.

3.5. CHUẨN BỊ DỮ LIỆU

Tập dữ liệu sau khi tiền xử lý là 15,471 bản ghi sẽ được áp dụng phương pháp cửa sổ trượt (Sliding Window). Đây là một kỹ thuật phổ biến trong dự báo chuỗi thời gian, đặc biệt khi áp dụng với mô hình LSTM giúp mô hình học được xu hướng và mẫu dữ liệu theo thời gian bằng cách sử dụng một số lượng điểm dữ liệu nhất định để dự đoán giá trị tiếp theo.

Trong bài toán dự báo này việc chọn số lượng đầu vào là 24 có ý nghĩa quan trọng. AQI thường được đo theo từng giờ, do đó việc sử dụng 24 giờ gần nhất giúp mô hình học được xu hướng theo chu kỳ ngày đêm. Ngoài ra, AQI bị ảnh hưởng bởi nhiều yếu tố như giao thông, thời tiết và hoạt động công nghiệp, thường có tính chu kỳ trong khoảng 24 giờ. Nếu chọn số lượng đầu vào quá lớn, mô hình sẽ khó học do lượng thông tin cần xử lý nhiều hơn, đồng thời có thể gây ra vấn đề dư thừa dữ liệu hoặc overfitting.

Trước khi tiến hành huấn luyện mô hình cần chia tập dữ liệu thành hai phần: tập huấn luyện (training set) và tập kiểm tra (test set). Tập huấn luyện sẽ được sử dụng để huấn luyện mô hình, trong khi tập kiểm tra sẽ được sử dụng để đánh giá hiệu quả của mô hình.

Để thực hiện việc chia tập dữ liệu, em sử dụng hàm *train_test_split* từ thư viện *scikit-learn*. Hàm này chia tập dữ liệu thành các phần huấn luyện và kiểm tra dựa trên tỷ lệ phần trăm xác định. Trong trường hợp này, tập dữ liệu được chia thành 80% dữ liệu huấn luyện và 20% dữ liệu kiểm tra. Quá trình chia dữ liệu được viết như sau:

```
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2)
```

Trong đó:

- X là tập hợp các đặc trưng (features) của dữ liệu.
- Y là giá trị đầu ra cần dự đoán (label).
- test_size=0.2 xác định 20% dữ liệu sử dụng cho tập kiểm tra, 80% còn lại sử dụng cho tập huấn luyện.

Sau khi chia tập dữ liệu, em có tập huấn luyện (X_train ,Y_train) được dùng để huấn luyện mô hình và tập kiểm tra (X_test, Y_test) sẽ được sử dụng để kiểm tra và đánh giá hiệu quả mô hình.

3.6. HUẤN LUYỆN MÔ HÌNH

Sau khi chia tập dữ liệu, mô hình sẽ được huấn luyện trên tập huấn luyện (X_train, Y_train). Trong quá trình huấn luyện, các trọng số của mô hình sẽ được điều chỉnh để giảm thiểu sự sai lệch giữa giá trị thực tế và giá trị dự đoán. Mô hình sử dụng thuật toán Adam để tìm ra các trọng số tối ưu qua nhiều lần lặp (epochs).

Để đảm bảo rằng mô hình không bị overfitting (học quá kỹ các đặc trưng trong tập huấn luyện dẫn đến mất khả năng tổng quát trên dữ liệu mới), kỹ thuật Dropout được sử dụng trong mỗi lớp LSTM. Dropout ngẫu nhiên loại bỏ một tỷ lệ nhất định các kết nối giữa các nơ-ron trong quá trình huấn luyện, giúp mô hình học các đặc trưng tổng quát hơn và giảm thiểu việc nhớ quá chi tiết các mẫu trong tập huấn luyện.

Cụ thể, mô hình được huấn luyện qua 30 epochs, mỗi epoch là một lần duyệt qua toàn bộ tập huấn luyện để cập nhật các trọng số. Mỗi lần huấn luyện, mô hình sử dụng một batch size là 16, tức là 16 mẫu được sử dụng trong mỗi lần tính toán gradient để cập nhật trọng số.

Quá trình huấn luyện được mô tả theo công thức sau:

$$\theta_{t+1} = \theta_t - \eta * \nabla_\theta L(\theta_t) \tag{3.3}$$

Trong đó:

- θ_t là các trọng số của mô hình tại thời điểm t
- η là tốc độ học (learning rate)
- $\nabla_\theta L(\theta_t)$ là gradient của hàm mất mát tại thời điểm t
- $L(\theta_t)$ là hàm mất mát MSE

Trong mỗi lần huấn luyện, mô hình sẽ tính toán gradient của hàm mất mát theo các trọng số hiện tại và cập nhật các trọng số để giảm thiểu lỗi. Quá trình này sẽ được lặp lại cho đến khi sô lần lặp (epochs) hoàn thành.

3.7. ĐÁNH GIÁ MÔ HÌNH DỰ BÁO

Đánh giá là phần quan trọng để kiểm tra độ hiệu quả của mô hình, đảm bảo khả năng áp dụng vào thực tế và đáp ứng các yêu cầu đề ra. Thông qua các chỉ số đánh giá, mô hình có thể được tinh chỉnh và tối ưu để cải thiện độ chính xác cho kết quả dự báo.

a) Mean Squared Error

Sai số bình phương trung bình (MSE) hay lỗi bình phương trung bình là một số liệu đánh giá mô hình học máy được sử dụng rộng rãi. Giá trị này được tính bằng cách đo độ lệch giữa giá trị mà mô hình dự đoán với giá trị thực tế của tập dữ liệu. Ví dụ với bài toán dự đoán độ tuổi của một người, tại thời điểm *i* mô hình dự đoán độ tuổi có giá trị 21 trong khi độ tuổi thực tế là 23 thì sai số bằng 23 – 21 = 2. Do đó, sai số bình phương bằng $2^2 = 4$. Công thức tính MSE được viết như sau:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2 \tag{3.4}$$

Trong đó:

- $\hat{y}_i$ là giá trị dự đoán tại i
- y_i là giá trị thực tế tại i

Giá trị của MSE luôn dương và hiếm khi bằng 0, do dữ liệu thực tế thường bị khuyết hoặc chứa những giá trị gây nhiễu. Một giá trị bất thường trong dữ liệu sẽ tác động đáng kể đến chỉ số này. MSE thể hiện chất lượng của một mô hình dự đoán, với giá trị càng nhỏ thì dự đoán càng chính xác. Giá trị bằng 0 phản ánh sự chính xác tuyệt đối, tuy nhiên điều này rất khó đạt được trong thực tế.

Ưu điểm của MSE là tính phổ biến và dễ triển khai. Tuy nhiên nhược điểm của MSE là làm việc trên các tập dữ liệu thiếu tính đồng nhất – có nhiều giá trị ngoại lệ. Nhược điểm này dẫn đến những đánh giá sai lệch, thiếu tính chính xác cho mô hình.

MSE là một chỉ số đánh giá mô hình cơ bản thường được sử dụng cho các bài toán hồi quy và dự báo trong nhiều lĩnh vực khác nhau. Tuy nhiên, để đánh giá toàn diện hơn về hiệu suất của mô hình, cần sử dụng thêm các chỉ số khác phù hợp với đặc điểm của dữ liệu và bài toán nghiên cứu.

b) Root Mean Squared Error

Chỉ số RMSE được tính toán bằng căn bậc hai của sai số bình phương trung bình với công thức như sau:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.5}$$

Tương tự chỉ số MSE, RMSE luôn mang giá trị dương và với giá trị càng nhỏ thì dự đoán càng chính xác. Khác với MSE, giá trị chỉ số RMSE được biểu diễn trên cùng thang đo với dữ liệu đầu ra giúp dễ dàng hiểu được mức độ sai số của mô hình. Tuy nhiên, RMSE cũng gặp phải nhược điểm khi làm việc trên các tập dữ liệu thiếu tính đồng nhất.

c) Mean Absolute Error

MAE đo lường độ lớn trung bình của các sai số giữa giá trị dự đoán và giá trị thực tế mà không quan tâm đến hướng của sai số. Chỉ số này được tính toán bằng cách lấy giá trị tuyệt đối của từng sai số, sau đó tính trung bình cộng của các giá trị tuyệt đối này. Công thức tính MAE được biểu diễn như sau:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y_i}| \tag{3.6}$$

MAE luôn mang giá trị dương vì giá trị tuyệt đối loại bỏ dấu của sai số. Chỉ số này biểu thị khoảng cách giữa giá trị thực tế và giá trị dự đoán mà không xem xét hướng sai số. Do các sai số không được bình phương, MAE xử lý tốt hơn đối với những tập dữ liệu thiếu tính đồng nhất.

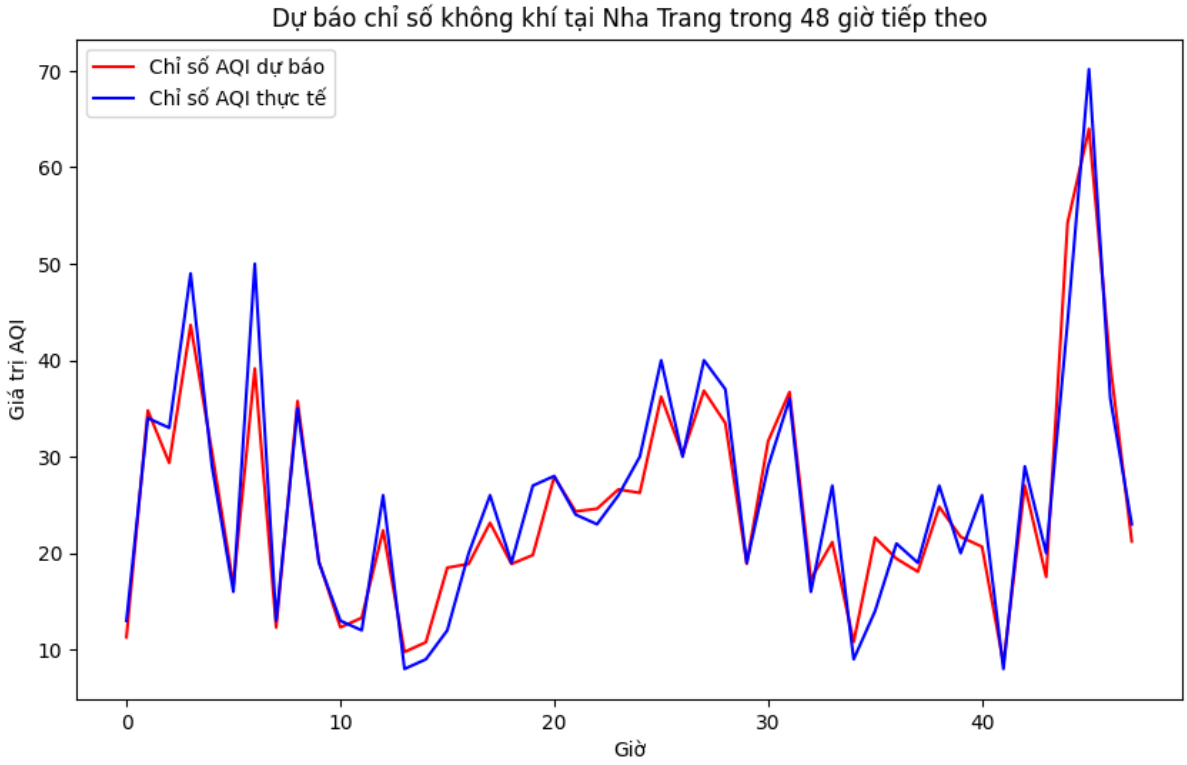
d) Mean Absolute Percentage Error

MAPE hoạt động tương tự như Mean Absolute Error (MAE) nhưng thay vì chỉ sử dụng giá trị tuyệt đối của sai số, chỉ số này chia sai số tuyệt đối cho giá trị thực tế để chuyển đổi thành phần trăm. Sau đó, các sai số phần trăm tuyệt đối được tổng hợp và tính trung bình. Công thức tính MAPE được biểu diễn như sau:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \widehat{y_i}}{y_i}\right| * 100 \tag{3.7}$$

Đặc điểm chính của MAPE là không phụ thuộc vào đơn vị của dữ liệu giúp chỉ số này có thể áp dụng và so sánh trên các tập dữ liệu có thang đo khác nhau. Chỉ số này đặc biệt dễ hiểu với các bên liên quan không chuyên về kỹ thuật, do sai số được biểu diễn dưới dạng tỷ lệ phần trăm. MAPE càng nhỏ thì độ chính xác của mô hình dự báo càng cao.

e) Dự báo và đánh giá mô hình



Hình 3.13. Biểu đồ dự báo chỉ số chất lượng không khí

Nhìn vào biểu đồ, đường dự báo của mô hình bám khá sát đường giá trị thực tế. Có những điểm mà mô hình dự báo không hoàn toàn khớp với thực tế đặc biệt ở các đỉnh và đáy. Đây là những thời điểm chỉ số AQI tăng hoặc giảm đột ngột, mô hình có xu hướng không phản ứng đủ nhanh dẫn đến các sai lệch nhưng nhìn chung mô hình vẫn phản ánh tương đối tốt các biến động của chỉ số AQI.

Bảng 3.1. So sánh chỉ số đánh giá mô hình

Mô hình	RNN	LSTM
MSE	71.91	43.82
RMSE	8.48	6.62
MAE	5.69	3.63
MAPE (%)	28.46	15.54

Dựa trên các chỉ số đánh giá, mô hình LSTM cho thấy sự vượt trội rõ rệt so với RNN trong việc dự báo chỉ số chất lượng không khí (AQI) tại Nha Trang. Cụ thể các chỉ số MSE, RMSE, MAE và MAPE của LSTM lần lượt đạt 43.82, 6.62, 3.63, và 15.54% thấp hơn đáng kể so với các giá trị tương ứng của RNN lần lượt là 71.91, 8.48, 5.69 và 28.46%. Điều này chứng minh rằng mô hình LSTM có khả năng dự báo chính xác và gần với giá trị thực tế hơn, đặc biệt trong bối cảnh dữ liệu AQI thường xuyên biến động lớn và chứa các mẫu phức tạp.

Lý do cho sự vượt trội này dựa vào cấu trúc đặc biệt của LSTM. Nhờ các cơ chế cổng nhớ (forget gate, input gate, và output gate), mô hình LSTM có thể lưu trữ và xử lý thông tin quan trọng trong chuỗi thời gian dài hạn một cách hiệu quả hơn so với RNN truyền thống. Đồng thời, cơ chế này giúp LSTM tập trung vào các thông tin quan trọng và loại bỏ những yếu tố không cần thiết cải thiện khả năng học tập từ dữ liệu chuỗi thời gian phức tạp.

Với những ưu điểm này, LSTM không chỉ phù hợp cho các bài toán dự báo AQI mà còn cho các ứng dụng liên quan đến dữ liệu chuỗi thời gian có tính phi tuyến cao, chứa nhiều dữ liệu nhiễu góp phần nâng cao độ tin cậy của mô hình trong việc đưa ra các dự báo.

3.8. KẾT LUẬN CHƯƠNG

Chương 3 trình bày quá trình xây dựng và đánh giá mô hình dự báo chỉ số chất lượng không khí một cách chi tiết, bao gồm các bước từ khảo sát và thu thập dữ liệu, tiền xử lý, phân tích xu hướng, đến việc xây dựng, huấn luyện và đánh giá mô hình. Thông qua các phân tích và thực nghiệm, chương 3 không chỉ làm rõ cách thức xử lý dữ liệu chuỗi thời gian mà còn minh họa hiệu quả của các mô hình dự báo được sử dụng. Kết quả thu được là nền tảng để so sánh, đánh giá và đề xuất cải tiến mô hình trong các nghiên cứu tiếp theo.

KẾT LUẬN VÀ KIẾN NGHỊ

Nghiên cứu đã áp dụng mô hình mạng nơ-ron hồi quy (RNN) với kiến trúc LSTM để giải quyết bài toán dự báo chỉ số chất lượng không khí (AQI) tại Nha Trang. Kết quả thử nghiệm cho thấy mô hình LSTM có khả năng học và xử lý hiệu quả các chuỗi thời gian, giúp dự đoán xu hướng thay đổi của AQI trong tương lai. Với đặc trưng vượt trội trong việc ghi nhớ thông tin dài hạn và xử lý sự phụ thuộc thời gian, mô hình LSTM khắc phục được vấn đề vanishing gradient, vốn là hạn chế của các mô hình RNN truyền thống.

Dù vậy, mô hình vẫn gặp một số thách thức. Một trong những vấn đề là thời gian huấn luyện khá dài, đặc biệt khi kích thước dữ liệu lớn. Ngoài ra, mô hình cũng nhạy cảm với dữ liệu đầu vào đòi hỏi quá trình tiền xử lý cẩn thận. Mức độ chính xác của dự báo phụ thuộc vào chất lượng và tính đầy đủ của dữ liệu.

Nhìn chung, mô hình LSTM đã thể hiện khả năng mạnh mẽ trong việc dự báo các xu hướng ngắn hạn dựa trên dữ liệu chuỗi thời gian phức tạp. Với sự cải thiện phù hợp, mô hình có thể trở thành công cụ hữu ích trong việc dự báo và quản lý chất lượng không khí, giúp cơ quan quản lý và người dân có những kế hoạch phù hợp để ứng phó với các tình huống xấu.

Dựa trên các kết quả thu được chúng ta có thể cải tiến thêm cho nghiên cứu theo một số phương hướng như sau:

- Để cải thiện độ chính xác của dự báo, việc bổ sung thêm các yếu tố đầu vào như thông tin thời tiết (nhiệt độ, độ ẩm, lượng mưa), lượng khí thải từ giao thông và công nghiệp là rất cần thiết. Những dữ liệu này có thể giúp mô hình hiểu rõ hơn về các yếu tố ảnh hưởng đến chất lượng không khí và đưa ra dự đoán chính xác hơn.

- Thử nghiệm các kiến trúc khác như GRU (Gated Recurrent Unit) hoặc kết hợp các mô hình RNN với các phương pháp khác như CNN (Convolutional Neural Network) để nâng cao khả năng dự báo.
- Với sự phức tạp và lượng dữ liệu lớn, việc huấn luyện mô hình LSTM đòi hỏi thời gian và tài nguyên tính toán lớn. Do đó, việc sử dụng các công nghệ như GPU hoặc TPU sẽ giúp giảm đáng kể thời gian huấn luyện, đồng thời tăng hiệu quả tính toán và cho phép thử nghiệm các mô hình phức tạp hơn.

TÀI LIỆU THAM KHẢO

Tiếng anh

[1] Meena, K.K.; Bairwa, D.; Agarwal, A. A machine learning approach for unraveling the influence of air quality awareness on travel behavior. Decis. Anal. J. 2024, 11, 100459

[2] Yong Yu, Xiaosheng Si, Changhua Hu, Jianxun Zhang (2019). A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures.

[3] Jiaxuan Zhang, Shunyong Li (2022). Air quality index forecast in Beijing based on CNN-LSTM multi-model.

[4] Yu Jiao, Zhifeng Wang, Yang Zhang (2019). Prediction of Air Quality Index Based on LSTM.

[5] Ricardo Navares, José L. Aznarte (2020). Predicting air quality with deep learning LSTM: Towards comprehensive models.

[6] Haiming Z, Xiaoxiao S (2013). Study on prediction of atmo- spheric PM2.5 based on RBF neural network. In: 2013 4th international conference on digital manufacturing automation.

[7] Box GE, Jenkins GM, Reinsel GC, Ljung GM (2015). Time series analysis: forecasting and control. Wiley

[8] Kirchgassner G, Wolters J, Hassler U (2012). Introduction to modern time series analysis. Springer

[9] Jiao K, Xu M, Liu M (2018). Health status and air pollution related socioeconomic concerns in urban china.

[10] Xie J (2017). Deep neural network for PM2.5 pollution forecasting based on manifold learning. In: 2017 international conference on sensing, diagnostics, prognostics, and control (SDPC)

[11] Nair, Vinod, and Geoffrey E. Hinton (2010). Rectified linear units improve restricted boltzamn machines. Proceedings of the 27th international conference on machine learning (ICML-10)

[12] Mass, Andrw L., Awni Y. Hannun, and Andrew Y. Ng. “Rectifier nonlinearities improve neural network acoustic models.” Proc. Icml. Vol. 30. No. 1. 2013.

[13] Aurélien Géron, “Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow”, O’Reilly, 2nd edition, 2019, ISBN: 978-1-492-03264-9.

[14] C. C. Aggarwal, "Neural networks and deep learning: A textbook." Cham, Switzerland: Springer Nature, Jun. 2023.

[15] Yan, R.; et al. Multi-hour and multi-site air quality index forecasting in Beijing using CNN, LSTM, CNN-LSTM, and spatiotemporal clustering.

[16] Jiao, Y.; et al. Prediction of air quality index based on LSTM. Proceeding of the IEEE 8th Joint International Information Technology and Artificial Intelligence Conference 2019.

[17] Belavadi, S.V. Air quality forecasting using LSTM RNN and wireless sensor networks. Proceeding of the 11th International Conference on Ambient Systems. Poland, Elsevier B. V. 2020.

[18] Duan, J.; et al. Air-quality prediction based on the ARIMA-CNN-LSTM combination model optimized by Dung Beetle Optimizer.

[19] Yammahi, A.A.; et al. Forecasting the concentration of NO using statistical and machine learning methods: A case study in the UAE, 2023.

[20] Navares, R.; et al. Predicting air quality with deep learning LSTM: Towards comprehensive models, 2019.

[21] Chang, Y.S.; et al. An LSTM-based aggregated model for air pollution forecasting. Atmos. Pollut. Res, 2020.

[22] Wen, C.; et al. A novel spatiotemporal convolutional long short-term neural network for air pollution prediction, 2019.

[23] Jung, Y.; et al. Concentration separation prediction model to enhance prediction accuracy of particulate matter. J. Inf. Commun. Technol, 2023.

[24] Rakholia, R.; et al. Al-based air quality $PM_{2.5}$ forecasting models for developing countries: A case study of Ho Chi Minh City, Vietnam. Urban Clim, 2022.

[25] Hưng, M.D. Nghiên cứu ứng dụng trí tuệ nhân tạo trong dự báo chất lượng không khí. Luận án tiến sĩ, Đại học Bách khoa Hà Nội, 2020.