

## LỜI CAM ĐOAN

Tôi cam đoan: Khóa luận tốt nghiệp này là kết quả nghiên cứu của riêng tôi, được thực hiện dưới sự hướng dẫn khoa học của TS. Huỳnh Đệ Thủ, đảm bảo tính trung thực và tuân thủ các quy định về trích dẫn, chú thích tài liệu tham khảo. Tôi xin chịu hoàn toàn trách nhiệm về lời cam đoan này

Tp. Hồ Chí Minh, ngày      tháng      năm 2025

Sinh viên

## LỜI CẢM ƠN

Trong suốt quá trình nghiên cứu và thực hiện tôi luôn được sự quan tâm, hướng dẫn và giúp đỡ tận tình của các giảng viên trong khoa Kỹ thuật và Khoa học máy tính.

Lời đầu tiên tôi xin được bày tỏ lòng biết ơn sâu sắc đến Ban giám hiệu trường đại học Quốc tế Sài Gòn đã tạo điều kiện cho tôi được học tập tại môi trường rất hiện đại này, Tôi xin chân thành cảm ơn sâu sắc đến khoa Kỹ thuật và Khoa học máy tính đã tạo điều kiện và hỗ trợ tôi tiếp cận với mảng kiến thức đầy hữu ích này.

Đặc biệt, tôi xin gửi lời cảm ơn chân thành đến Tiến sĩ Huỳnh Đệ Thủ - giảng viên Khoa Kỹ thuật và Khoa học máy tính đại học Quốc tế Sài Gòn - Người hướng dẫn chính đã tận tình chỉ bảo và hướng dẫn để tôi hoàn thành tốt đề tài này.

Trong quá trình làm đồ án còn nhiều thiếu sót, em rất mong nhận được sự chỉ bảo, đóng góp ý kiến của các thầy cô để tôi có điều kiện bổ sung, khắc phục những hạn chế của bài đồ án này.

Tp. Hồ Chí Minh, ngày      tháng      năm 2025

Sinh viên

## NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

- Kết cấu, phương pháp trình bày:

.....

.....

.....

.....

.....

Cơ sở lý luận: .....

.....

.....

.....

.....

- Tính thực tiễn và khả năng ứng dụng của khóa luận:

.....

.....

- Các hướng nghiên cứu phát triển của đề tài:

.....

.....

- Kết quả: (Đạt hoặc không đạt).

.....

.....

*Thành phố Hồ Chí Minh, ngày    tháng    năm 2025*

**Giảng viên hướng dẫn**

## MỤC LỤC

|  |            |
|--|------------|
| <b>LỜI CAM ĐOAN .....</b>                                | <b>i</b>   |
| <b>LỜI CẢM ƠN .....</b>                                  | <b>ii</b>  |
| <b>NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN.....</b>            | <b>iii</b> |
| <b>MỤC LỤC .....</b>                                     | <b>iv</b>  |
| <b>DANH MỤC KÝ HIỆU, TỪ VIẾT TẮT.....</b>                | <b>vii</b> |
| <b>DANH MỤC CÁC BẢNG.....</b>                            | <b>ix</b>  |
| <b>DANH MỤC HÌNH ẢNH, BIỂU ĐỒ, SƠ ĐỒ .....</b>           | <b>x</b>   |
| <b>MỞ ĐẦU .....</b>                                      | <b>1</b>   |
| 1. Lý do chọn đề tài .....                               | 1          |
| 2. Các nghiên cứu liên quan .....                        | 5          |
| 3. Mục tiêu nghiên cứu .....                             | 13         |
| 4. Phát biểu bài toán .....                              | 14         |
| 5. Đối tượng và phạm vi nghiên cứu .....                 | 15         |
| 6. Phương pháp nghiên cứu .....                          | 15         |
| 7. Cấu trúc khóa luận.....                               | 16         |
| <b>Chương 1. CƠ SỞ LÝ THUYẾT.....</b>                    | <b>17</b>  |
| 1.1 Các thuật toán sử dụng: .....                        | 17         |
| 1.1.1 Kỹ thuật học máy trực tuyến .....                  | 17         |
| 1.1.2 Phương pháp Inverse Distance Weighting (IDW) ..... | 18         |
| 1.2 Các thuật toán học máy: .....                        | 20         |
| 1.2.1 Long-Short Term Memory (LSTM) [6] .....            | 20         |
| 1.2.2 LightGBM [7].....                                  | 22         |
| 1.2.3 XGBoost [8] .....                                  | 23         |
| <b>KẾT LUẬN CHƯƠNG .....</b>                             | <b>24</b>  |
| <b>Chương 2 THIẾT KẾ HỆ THỐNG .....</b>                  | <b>25</b>  |

|       |   |           |
|-------|---|-----------|
| 2.1   | Thiết kế mô hình học trực tuyến.....  | 25        |
| 2.1.1 | Thu thập và tiền xử lý dữ liệu: .....   | 26        |
| 2.1.2 | Huấn luyện mô hình trực tuyến:.....   | 28        |
| 2.1.3 | Đánh giá mô hình: .....   | 31        |
| 2.1.4 | Triển khai mô hình: .....   | 32        |
| 2.1.5 | Quy trình cập nhật dữ liệu.....   | 33        |
| 2.2   | Thiết kế mô hình dự đoán chất lượng không khí tại những nơi không có trạm.<br>..... | 34        |
| 2.3   | Xây dựng ứng dụng web.....  | 37        |
| 2.3.1 | Hiển thị dữ liệu lịch sử chất lượng không khí.....                                  | 37        |
| 2.3.2 | Dự đoán chất lượng không khí trong 24 giờ. ....                                     | 39        |
| 2.3.3 | Hiển thị chất lượng không khí tại các trạm trên bản đồ.....                         | 40        |
|       | KẾT LUẬN CHƯƠNG .....   | 42        |
|       | <b>Chương 3 THỰC NGHIỆM VÀ ĐÁNH GIÁ .....</b>                                       | <b>43</b> |
| 3.1   | Môi trường thực nghiệm.....   | 43        |
| 3.1.1 | Môi trường thực nghiệm cho mô hình học máy trực tuyến .....                         | 43        |
| 3.1.2 | Môi trường thực nghiệm cho việc xây dựng ứng dụng:.....                             | 43        |
| 3.2   | Tập dữ liệu .....   | 44        |
| 3.3   | Thang đo và các chỉ số đánh giá.....  | 47        |
| 3.3.1 | Thang đo Root Mean Square Error (RMSE).....   | 47        |
| 3.3.2 | Thang đo Mean Absolute Error (MAE) .....  | 48        |
| 3.3.3 | Chỉ số đánh giá precision .....   | 49        |
| 3.3.4 | Chỉ số đánh giá recall .....  | 50        |
| 3.4   | Kết quả thực nghiệm.....  | 50        |
| 3.4.1 | Kết quả huấn luyện mô hình học trực tuyến .....                                     | 50        |

|   |   |           |
|---|---|-----------|
| 3.4.2   | Kết quả dự đoán dữ liệu chất lượng không khí trong 24 giờ trên thời gian thực ..... | 58        |
| 3.4.3   | Kết quả dự đoán dữ liệu sử dụng phương pháp IDW .....                               | 61        |
| 3.4.4   | Đánh giá hệ thống web .....   | 63        |
| KẾT LUẬN CHƯƠNG .....                             |   | 64        |
| <b>Chương 4 KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN.....</b> |   | <b>65</b> |
| 4.1   | Kết luận.....   | 65        |
| 4.2   | Tác động của đề tài đến xã hội .....  | 66        |
| 4.3   | Hướng phát triển .....  | 66        |
| <b>TÀI LIỆU THAM KHẢO.....</b>                    |   | <b>68</b> |

## **DANH MỤC KÝ HIỆU, TỪ VIẾT TẮT**

|          |  |
|----------|--|
| AI       | Artificial intelligence                  |
| ANN      | Artificial Neural Network                |
| API      | Application Programming Interface        |
| AQI      | Air Quality Index                        |
| CAMS     | Copernicus Atmosphere Monitoring Service |
| CEC      | Constant Error Carousel                  |
| CNN      | Convolutional Neural Network             |
| CPU      | Central Processing Unit                  |
| CSV      | Comma Separated Values                   |
| EFB      | Exclusive Feature Bundling               |
| GB       | Gigabyte                                 |
| GBDT     | Gradient-Boosted Decision Trees          |
| GOSS     | Gradient-based One-Side Sampling         |
| GPU      | Graphics Processing Unit                 |
| GRNN     | Generalized Regression Neural Network    |
| TP.HCM   | Thành Phố Hồ Chí Minh                    |
| HTML     | Hypertext Markup Language                |
| IDW      | Inverse Distance Weighting               |
| IOA      | Indicator of Attack                      |
| IoT      | Internet of Things                       |
| LightGBM | Light Gradient Boosting Machine          |
| LSTM     | Long-Short Term Memory                   |

|         |                                |
|---------|--------------------------------|
| MAE     | Mean Absolute Error            |
| MAPE    | Mean Absolute Percentage Error |
| MB      | Megabyte                       |
| MLOps   | Machine Learning Operations    |
| MSE     | Mean Squared Error             |
| RAM     | Random-access memory           |
| RF      | Random Forest                  |
| RMSE    | Root Mean Square Error         |
| RNN     | Recurrent Neural Network       |
| RTRL    | Real-Time Recurrent Learning   |
| SGD     | Stochastic Gradient Descent    |
| SVR     | Support Vector Regression      |
| US      | United States of America       |
| USG     | Unhealthy for Sensitive Groups |
| WAQI    | World Air Quality Index        |
| WHO     | World Health Organization      |
| XGBoost | Extreme Gradient Boosting      |



## **DANH MỤC CÁC BẢNG**

|   |    |
|---|----|
| Bảng 1: Cấu hình phần cứng của máy tính cá nhân và google colab:.....   | 43 |
| Bảng 2: Số liệu về phần cứng và tổng thời training theo từng trạm ..... | 55 |
| Bảng 3: Kết quả của phương pháp IDW. ....                               | 62 |

## DANH MỤC HÌNH ẢNH, BIỂU ĐỒ, SƠ ĐỒ







|  |    |
|--|----|
| Hình 1: Thang đo mức độ nguy hiểm của chất lượng không khí (nguồn: IQAir). ....  | 1  |
| Hình 2: Bảng xếp hạng ô nhiễm không khí của các thành phố trên thế giới vào ngày 24/04/2025 theo IQAir. ....   | 2  |
| Hình 3: Kiến trúc mô hình 1D CNN-LSTM với cài đặt siêu tham số. ....   | 6  |
| Hình 4: Quá trình phát triển mô hình dự báo .....  | 8  |
| Hình 5: Workflow của mô hình dự đoán PM2.5 .....   | 9  |
| Hình 6: Flowchart phân tích các điều kiện và đặc điểm khí tượng. ....  | 10 |
| Hình 7: Framework tổng thể của mô hình được đề xuất. Đầu vào của mô hình là dữ liệu chất lượng không khí với các mức độ chi tiết khác nhau. Đầu ra của mô hình là dữ liệu chất lượng không khí. (a) Cấu trúc cụ thể của khối residual de-redundant. (b) Cấu trúc cụ thể của khối spatiotemporal attention. (c) Cấu trúc cụ thể của khối dynamic fusion. .... | 12 |
| Hình 8: Input và output của bài toán. ....   | 14 |
| Hình 9: Minh họa tìm kiếm những điểm lân cận.....  | 19 |
| Hình 10: Sơ đồ chi tiết của khối LSTM được sử dụng trong các hidden layer của recurrent neural network.....  | 21 |
| Hình 11: Kiến trúc khối học song song. Mỗi cột trong khối được sắp xếp theo giá trị tính năng tương ứng. ....  | 23 |
| Hình 12: Pipeline mô hình dự đoán chất lượng không khí ở TP.HCM theo thời gian thực .....  | 25 |
| Hình 13: Quy trình thu thập và tiền xử lý dữ liệu. ....  | 27 |
| Hình 14: Quy trình huấn luyện mô hình dự đoán chất lượng không khí .....   | 29 |
| Hình 15: Quy trình triển khai hệ thống lên nền tảng web app. ....  | 32 |
| Hình 16: Pipeline mô hình dự đoán chất lượng không khí tại những quận/huyện chưa có trạm ở TP.HCM. ....  | 35 |
| Hình 17: Giao diện biểu đồ chỉ số AQI lọc theo ngày và theo trạm. ....   | 38 |

|   |    |
|---|----|
| Hình 18: Giao diện dự đoán chất lượng không khí trong 24h tiếp theo tại trạm Quận 3.<br>.....   | 39 |
| Hình 19: Giao diện biểu đồ chất lượng không khí tại TP.HCM. ....  | 41 |
| Hình 20: Phân bố dữ liệu trong bộ dataset. Biểu đồ đường thể hiện dữ liệu qua các mốc<br>thời gian, biểu đồ cột thể hiện phân bố dữ liệu theo mức độ chất lượng không khí. .... | 46 |
| Hình 21: Giá trị RMSE của mô hình dự đoán chất lượng không khí theo từng trạm. ...  | 52 |
| Hình 22: Giá trị MAE của mô hình dự đoán chất lượng không khí theo từng trạm. ....  | 53 |
| Hình 23: Mẫu file ghi nhận kết quả dự đoán của mô hình và đánh giá mô hình dựa trên<br>dữ liệu thực. ....   | 58 |
| Hình 24: Đánh giá mô hình dự đoán chất lượng không khí dựa trên dữ liệu thực với từng<br>trạm.....  | 59 |
| Hình 25: Mẫu file ghi nhận kết quả dự đoán của phương pháp IDW và đánh giá phương<br>pháp dựa trên dữ liệu thực. ....   | 61 |

## MỞ ĐẦU

### 1. Lý do chọn đề tài

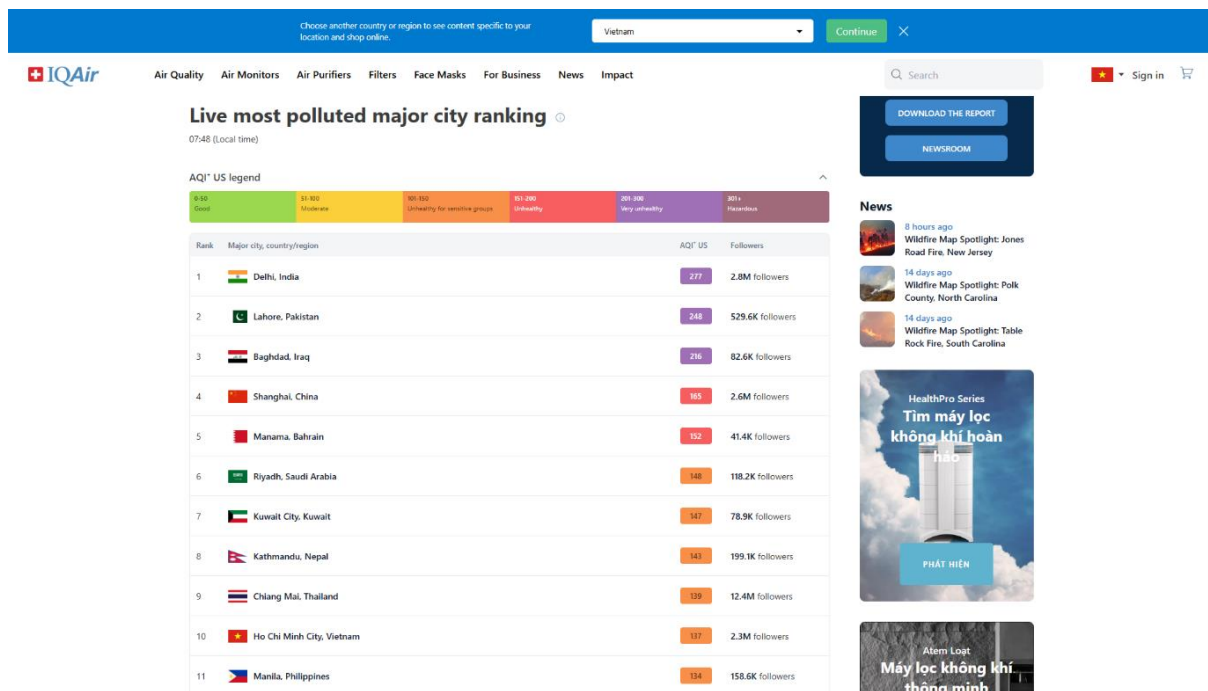
Chất lượng không khí là một trong những yếu tố then chốt quyết định đến sức khỏe cộng đồng, chất lượng sống và sự phát triển bền vững của đô thị. Theo định nghĩa của Tổ chức Y tế Thế giới (WHO), chất lượng không khí được xác định dựa trên nồng độ của các chất ô nhiễm như bụi mịn  $PM_{2.5}$ ,  $PM_{10}$ ,  $NO_2$ ,  $CO$ ,  $SO_2$ , và  $O_3$  trong không khí. Trong bài nghiên cứu này, thang đo chất lượng không khí được sử dụng là Chỉ số chất lượng không khí Hoa Kỳ (US AQI) – một chỉ số chuẩn hóa, có giá trị từ 0 đến 500, trong đó càng cao thì mức độ ô nhiễm càng nghiêm trọng. Thang đo này chia thành 6 mức độ cảnh báo từ “Tốt” đến “Nguy hại”, đi kèm với màu sắc cảnh báo và khuyến nghị ứng xử cho từng mức độ như trong hình 1.

|   | US AQI Level                           | PM <sub>2.5</sub><br>( $\mu g/m^3$ ) | Health Recommendation<br>(for 24 hour exposure)   |
|---|--|--------------------------------------|---|
|  | Good 0-50                              | 0-12.0                               | Air quality is satisfactory and poses little or no risk.  |
|  | Moderate 51-100                        | 12.1-35.4                            | Sensitive individuals should avoid outdoor activity as they may experience respiratory symptoms.                            |
|  | Unhealthy for Sensitive Groups 101-150 | 35.5-55.4                            | General public and sensitive individuals in particular are at risk to experience irritation and respiratory problems.       |
|  | Unhealthy 151-200                      | 55.5-150.4                           | Increased likelihood of adverse effects and aggravation to the heart and lungs among general public.                        |
|  | Very Unhealthy 201-300                 | 150.5-250.4                          | General public will be noticeably affected. Sensitive groups should restrict outdoor activities.                            |
|  | Hazardous 301+                         | 250.5+                               | General public at high risk of experiencing strong irritations and adverse health effects. Should avoid outdoor activities. |

Hình 1: Thang đo mức độ nguy hiểm của chất lượng không khí (nguồn: IQAir).

Sự quan tâm đến chỉ số chất lượng không khí ngày càng gia tăng trên toàn cầu khi các nghiên cứu khoa học đã chỉ ra mối liên hệ chặt chẽ giữa ô nhiễm không khí và nhiều bệnh lý nghiêm trọng như viêm đường hô hấp, tim mạch, ung thư phổi, và cả suy giảm chức năng nhận thức. Đặc biệt, những nhóm dân cư dễ bị tổn thương như trẻ em, người cao tuổi, và người có bệnh nền lại càng nhạy cảm với các tác nhân ô nhiễm trong không khí. Vì thế, việc theo dõi và dự đoán chất lượng không khí không chỉ mang ý nghĩa cảnh báo sức khỏe cá nhân mà còn là một công cụ thiết yếu trong việc hoạch định chính sách, quy hoạch đô thị và phản ứng y tế cộng đồng.

Tại Việt Nam, thực trạng chất lượng không khí ở các đô thị lớn đang có xu hướng xấu đi trong những năm gần đây. TP.HCM – trung tâm kinh tế, văn hóa, xã hội lớn nhất cả nước – đang phải đối mặt với mức độ ô nhiễm không khí đáng báo động. Theo dữ liệu từ IQAir, vào lúc **7:47 sáng ngày 24/04/2025**, TP.HCM xếp hạng **thứ 10** và Hà Nội xếp **thứ 4** trong danh sách các thành phố ô nhiễm nhất thế giới – một thứ hạng rất đáng quan ngại đối với một quốc gia đang phát triển. Số liệu vào ngày 24 tháng 04 năm 2025 được biểu hiện ở hình 2.



Hình 2: Bảng xếp hạng ô nhiễm không khí của các thành phố trên thế giới vào ngày 24/04/2025 theo IQAir.

Các ngày ô nhiễm khác trong năm 2025 cũng cho thấy xu hướng tăng cao đột biến vào giờ cao điểm hoặc những ngày trời yên, ít gió. Ví dụ, ngày 17/03/2025, AQI trung bình tại TP.HCM vượt ngưỡng 170 – mức không tốt cho sức khỏe của bất kỳ nhóm dân số nào. Thực tế này cho thấy nhu cầu cấp thiết về một hệ thống cảnh báo sớm và dự đoán chính xác tình trạng không khí trong ngắn hạn và theo thời gian thực.

Ô nhiễm không khí là một vấn đề môi trường nghiêm trọng với những tác động sâu rộng đến đời sống xã hội. Không chỉ gây ảnh hưởng trực tiếp đến sức khỏe cộng đồng thông qua việc làm gia tăng tỷ lệ mắc các bệnh lý hô hấp, tim mạch và ung thư, ô nhiễm không khí còn gián tiếp làm suy giảm năng suất lao động, gây gián đoạn trong hoạt động giáo dục, ảnh hưởng đến ngành du lịch và hạn chế các hoạt động ngoài trời. Trước thực trạng đó, nhiều nền tảng giám sát chất lượng không khí đã được phát triển nhằm cung cấp dữ liệu thời gian thực và dự báo chỉ số AQI (Air Quality Index), phục vụ công tác theo dõi, cảnh báo và ra quyết định trong cộng đồng.

Tiêu biểu trong số đó là nền tảng **IQAir AirVisual**, hiện đang theo dõi chất lượng không khí tại hơn 100 quốc gia, với dữ liệu thu thập từ hơn 80.000 trạm quan trắc bao gồm cả hệ thống công lập và cảm biến tư nhân. AirVisual sử dụng trí tuệ nhân tạo để phân tích dữ liệu lịch sử, xu hướng thời tiết và mức độ ô nhiễm hiện tại nhằm đưa ra các dự báo AQI trong ngắn hạn. Một nền tảng toàn cầu khác là **World Air Quality Index (WAQI)**, cung cấp bản đồ chất lượng không khí theo thời gian thực và miễn phí thông qua giao diện lập trình ứng dụng (Application Programming Interface - API) mở, với mạng lưới thu thập dữ liệu từ hơn 130 quốc gia. WAQI được biết đến với tính tương tác cao, cho phép người dùng tra cứu nhanh chóng chất lượng không khí tại địa phương hoặc bất kỳ khu vực nào trên thế giới. Ngoài ra, **OpenAQ** là một nền tảng dữ liệu mở quy mô toàn cầu, chuyên tập hợp và chuẩn hóa dữ liệu chất lượng không khí từ hơn 140 quốc gia và vùng lãnh thổ. Dữ liệu từ OpenAQ được sử dụng rộng rãi trong nghiên cứu học thuật, phân tích chính sách và các ứng dụng công nghệ, với sự tham gia tích cực của hàng nghìn nhà khoa học, kỹ sư và nhà hoạt động môi trường.

Một minh chứng tiêu biểu cho sự ứng dụng công nghệ trong giám sát chất lượng không khí toàn cầu là Dự án Copernicus Atmosphere Monitoring Service (CAMS) của Liên minh châu Âu. Đây là một trong những dự án lớn nhất thế giới, sử dụng dữ liệu vệ tinh kết hợp mô hình dự báo khí quyển để cung cấp các chỉ số chất lượng không khí cho toàn cầu, với độ phân giải không gian cao. Các công cụ này được tích hợp trên những nền tảng công nghệ lớn như Microsoft MSN Weather, Apple Weather, Google Maps, giúp người dùng có thể theo dõi chỉ số AQI theo từng khu vực, theo thời gian thực.

Trong bối cảnh công nghệ số phát triển mạnh mẽ và xu hướng “chuyển đổi số” lan rộng khắp thế giới, trí tuệ nhân tạo (AI) ngày càng đóng vai trò quan trọng trong việc xử lý dữ liệu lớn, học máy và ra quyết định dựa trên dữ liệu. Lĩnh vực dự đoán chất lượng không khí cũng không nằm ngoài xu thế này. Việc ứng dụng **mô hình học trực tuyến (online learning)** kết hợp với **nội suy không gian (spatial interpolation)** cho phép hệ thống vừa học từ dữ liệu mới theo thời gian thực, vừa mở rộng khả năng dự đoán cho các khu vực chưa có cảm biến. Đây chính là giải pháp thông minh, thích ứng với đặc điểm của các đô thị đang phát triển – nơi cảm biến đo lường vẫn còn hạn chế và dữ liệu thường xuyên bị thiếu hụt.

Không chỉ áp dụng cho TP.HCM – nơi có mật độ dân cư cao, nhiều hoạt động giao thông và công nghiệp, mà hệ thống này còn có thể nhân rộng ra các tỉnh, thành phố nhỏ hơn trên cả nước. Điều này góp phần hình thành một mạng lưới giám sát và dự đoán chất lượng không khí quy mô quốc gia, phục vụ cộng đồng và hỗ trợ cơ quan chức năng trong việc xây dựng các chính sách bảo vệ môi trường.

Với những lý do trên, đề tài **“Phát triển mô hình học trực tuyến kết hợp nội suy không gian để dự đoán chất lượng không khí tại TP.HCM”** được lựa chọn để tiếp tục phát triển, mở rộng từ các nghiên cứu trước đây của chính tác giả. Cụ thể, trong năm học 2024–2025, tác giả đã thực hiện hai đề tài cấp sinh viên là:

- **Ứng dụng mô hình học sâu trong việc dự đoán chất lượng không khí ở TP.HCM** – tham dự Giải thưởng Sinh viên Nghiên cứu Khoa học SIU lần thứ 17 và đạt **giải Nhất cấp trường**.
- **Ứng dụng mô hình học máy trực tuyến trong việc dự đoán chất lượng không khí ở TP.HCM theo thời gian thực** – tham dự **Giải thưởng Khoa học và Công nghệ dành cho sinh viên trong các cơ sở giáo dục đại học năm 2025**.

Các đề tài trước tập trung vào việc thu thập dữ liệu theo thời gian thực và huấn luyện mô hình học sâu để dự đoán chất lượng không khí tại TP.HCM trong vòng 24 giờ tiếp theo. Mặc dù các mô hình đã đạt kết quả tương đối tốt, nhưng vẫn còn nhiều thách thức về độ chính xác tại các khu vực không có cảm biến, cũng như khả năng thích ứng nhanh với dữ liệu mới. Vì vậy, đề tài lần này sẽ hướng đến việc **kết hợp mô hình học trực tuyến với thuật toán nội suy không gian** để cải thiện độ bao phủ và nâng cao độ chính xác của mô hình. Ngoài ra, hệ thống cũng sẽ được thiết kế để cập nhật dữ liệu AQI theo thời gian thực từ các nền tảng quốc tế, từ đó đưa ra dự báo chất lượng không khí một cách linh hoạt, chính xác và có khả năng mở rộng.

## 2. Các nghiên cứu liên quan

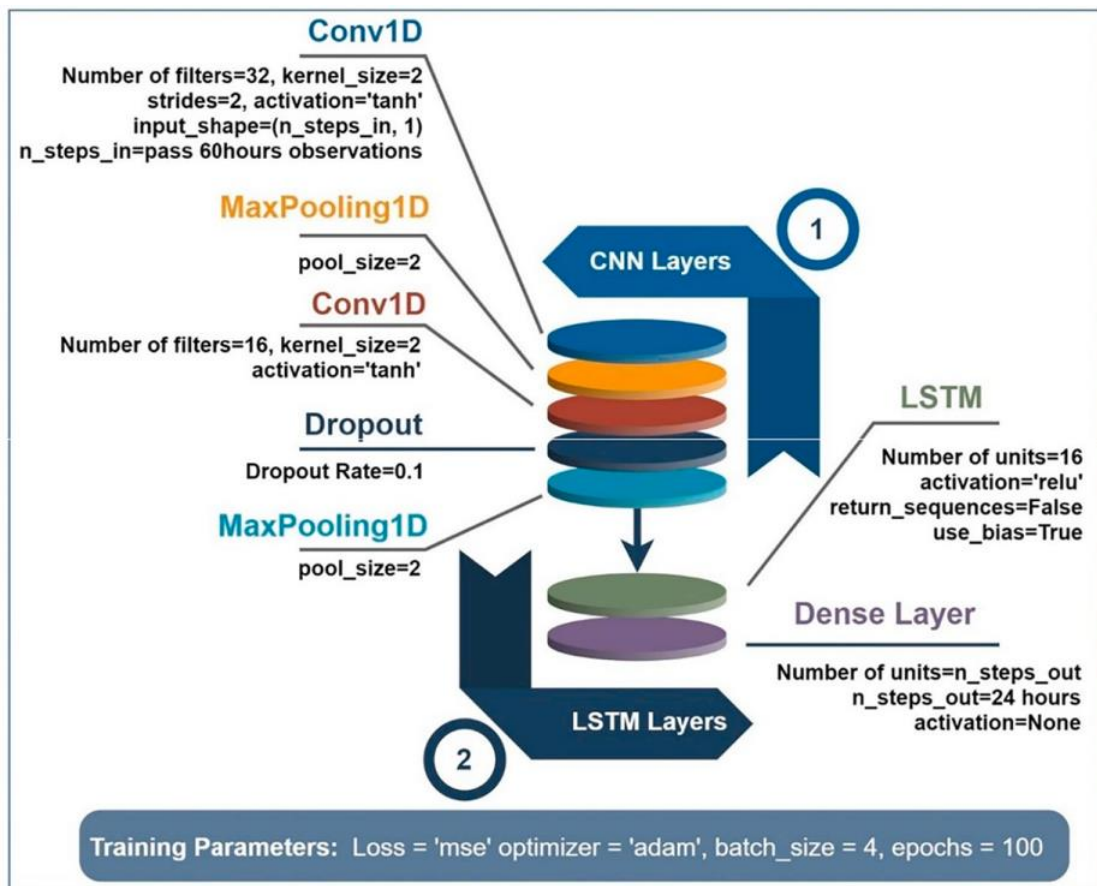
Ô nhiễm không khí là vấn đề đã tồn tại từ lâu, và giải pháp dự đoán chất lượng không khí từ sớm đã trở thành hướng nghiên cứu thu hút sự quan tâm của nhiều nhà khoa học. Phần lớn các nghiên cứu này tập trung vào việc xây dựng và tối ưu hóa các mô hình học sâu nhằm cải thiện tốc độ huấn luyện, khả năng truy xuất và tiết kiệm tài nguyên tính toán. Bên cạnh đó, một số nghiên cứu cũng hướng đến việc phát triển các ứng dụng dự báo chất lượng không khí phục vụ nhu cầu theo dõi và cảnh báo trong đời sống người dân. Tuy nhiên, các nghiên cứu trên gặp khó khăn trong việc dự đoán chất lượng không khí một cách chính xác trong thời gian thực. Cụ thể, dữ liệu huấn luyện thường được thu thập theo phương pháp truyền thống, tức là sử dụng dữ liệu lịch sử cố định và không được cập nhật theo thời gian thực. Điều này làm giảm độ chính xác của mô hình khi dự đoán chất lượng không khí trong bối cảnh hiện tại, bởi vì mức độ ô nhiễm tại TP.HCM luôn biến động liên tục theo thời gian và chịu ảnh hưởng từ nhiều yếu tố thời tiết, giao thông, công nghiệp và các yếu tố môi trường khác.

Dưới đây là những công trình nghiên cứu nổi bật liên quan:



- **Các nghiên cứu trong nước:**

Rakholia và cộng sự [1] đã phát triển ứng dụng HealthyAir trên nền tảng di động với chức năng chính là dự báo chất lượng không khí tại TP.HCM một cách chính xác. Nhóm nghiên cứu đã xây dựng mô hình học máy dựa trên các thuật toán như Stochastic Gradient Descent Regressor, mô hình kết hợp 1D CNN-LSTM, eXtreme Gradient Boosting Regressor, và Prophet, nhằm khai thác và học từ bộ dữ liệu mà họ tự thu thập. Mục tiêu chính của nghiên cứu là cải tiến so với các ứng dụng theo dõi thời tiết truyền thống vốn chỉ mang tính chất giám sát, chưa đủ khả năng đưa ra các dự báo để hỗ trợ người dùng lên kế hoạch ứng phó trong tương lai gần. Kiến trúc của mô hình dự đoán được minh họa ở hình 3.



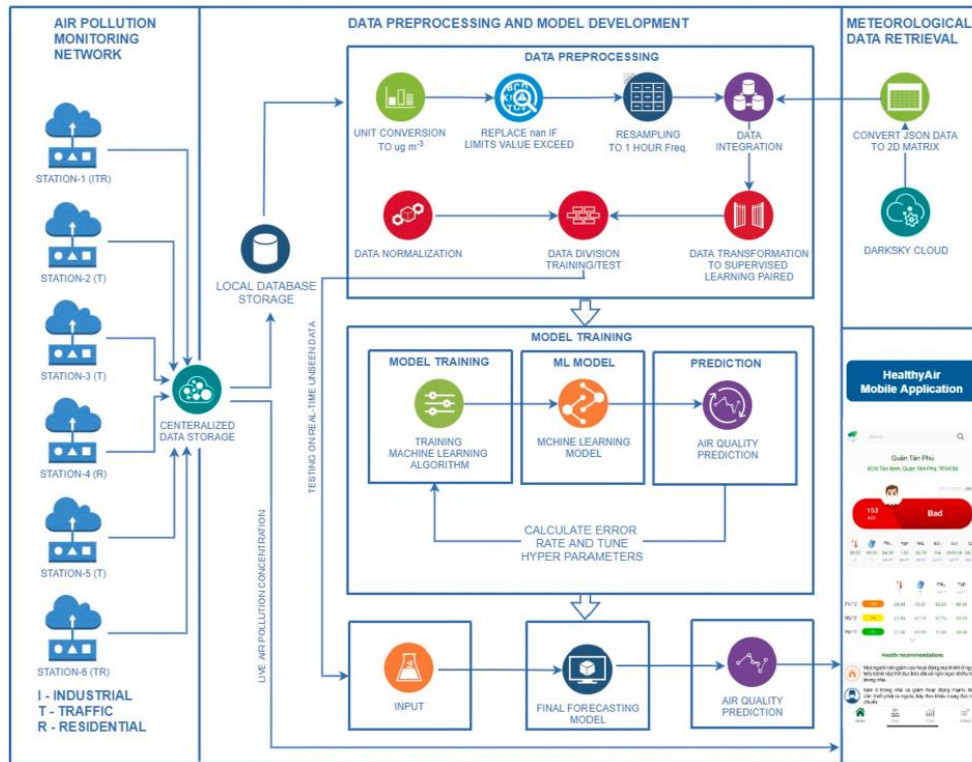
Hình 3: Kiến trúc mô hình 1D CNN-LSTM với cài đặt siêu tham số.

Bộ dữ liệu được sử dụng trong nghiên cứu này được thu thập từ một mạng lưới gồm 6 trạm quan trắc chất lượng không khí đặt tại nhiều khu vực khác nhau trong TP.HCM. Nghiên cứu này sử dụng dữ liệu trong khoảng thời gian từ tháng 2 năm 2021 đến tháng 12 năm 2021.

Tuy nhiên, nghiên cứu vẫn còn một số hạn chế nhất định. Trước hết, dữ liệu được sử dụng trong mô hình không được cập nhật liên tục theo thời gian thực, điều này khiến mô hình khó nắm bắt và phản ánh chính xác xu hướng chất lượng không khí tại thời điểm hiện tại – vốn có tính biến động cao do chịu ảnh hưởng từ nhiều yếu tố môi trường. Bên cạnh đó, việc đánh giá hiệu suất mô hình chỉ dừng lại ở các chỉ số sai số phổ biến như RMSE và MAE, trong khi chưa đề cập đến thời gian huấn luyện và tài nguyên phần cứng được sử dụng. Những yếu tố này đóng vai trò quan trọng trong việc đánh giá mức độ khả thi và hiệu quả khi triển khai mô hình trong môi trường thực tế. Cuối cùng, nghiên cứu chưa phát triển thành một ứng dụng cụ thể phục vụ người dùng đầu cuối, khiến cho tính ứng dụng trong đời sống thực tế còn bị hạn chế.

Trong nỗ lực giải quyết thách thức liên quan đến việc cập nhật dữ liệu theo thời gian thực, Rakholia cùng nhóm nghiên cứu [2] đã áp dụng mô hình mạng nơ-ron N-BEATS – một kiến trúc học sâu hiện đại có khả năng xử lý dữ liệu chuỗi thời gian hiệu quả. Với đặc thù của bài toán dự báo chất lượng không khí và mục tiêu phục vụ nhu cầu thực tiễn của người dân TP.HCM, việc đảm bảo mô hình có thể học từ dữ liệu mới cập nhật hàng ngày là một yêu cầu then chốt, ảnh hưởng trực tiếp đến độ chính xác và độ tin cậy của kết quả dự báo. Ngoài ra, để nâng cấp ứng dụng HealthyAir, nhóm nghiên cứu còn tích hợp API từ Darksky nhằm thu thập dữ liệu vi khí hậu như độ ẩm, điều kiện thời tiết,... với tần suất theo từng phút, góp phần nâng cao độ chi tiết và tính cập nhật của dữ liệu đầu vào.

Nghiên cứu sử dụng cùng bộ dữ liệu HealthyAir với khoảng thời gian từ tháng 2 năm 2021 đến tháng 8 năm 2022. Ngoài ra, dữ liệu còn thu thập dữ liệu những tác nhân ảnh hưởng đến chất lượng không khí như góc nhìn, tốc độ gió, áp suất, chỉ số tia cực tím và tầm nhìn từ Apple's Darksky weather API. Kết quả cho thấy mô hình N-BEATS đạt độ chính xác dự báo vượt trội so với các mô hình truyền thống trước đó tại TP.HCM, đặc biệt là các hệ thống dự báo sử dụng TAPM-CTM kết hợp với kiểm kê phát thải. Một ưu điểm nổi bật của N-BEATS là khả năng dự báo đồng thời nhiều chất ô nhiễm, qua đó giúp giảm đáng kể công sức trong việc phát triển và duy trì các mô hình riêng biệt cho từng loại chất ô nhiễm. Hình 4 minh họa quy trình xây dựng và triển khai mô hình dự báo này.



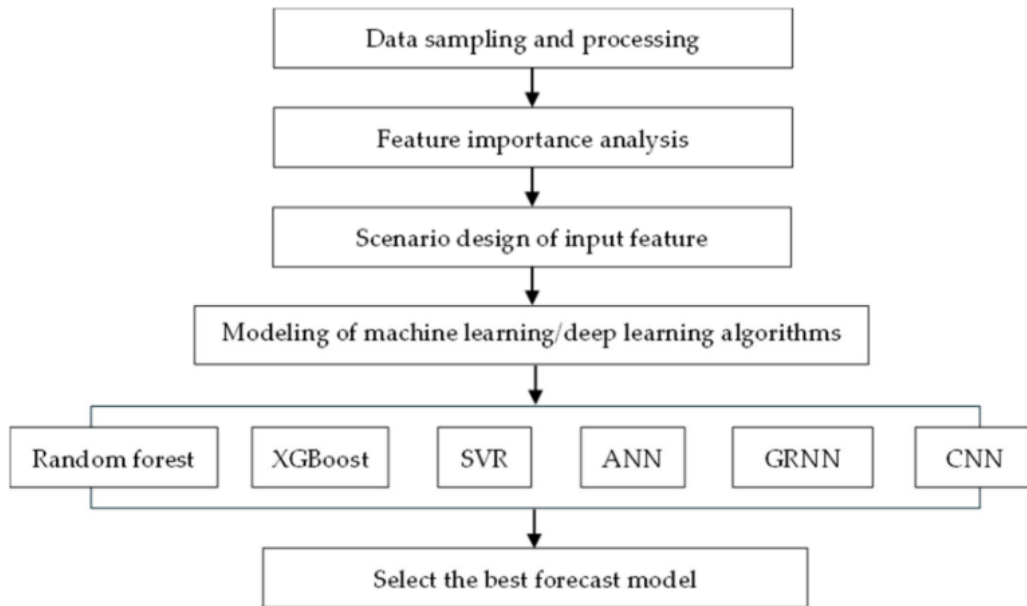
Hình 4: Quá trình phát triển mô hình dự báo

Tuy nhiên, nghiên cứu vẫn tồn tại một số hạn chế đáng chú ý. Mô hình chỉ có thể dự báo khi có đầy đủ dữ liệu lịch sử trong 48 giờ gần nhất, bao gồm cả biến mục tiêu và các đặc trưng đồng biến, dẫn đến việc hệ thống không hoạt động nếu thiếu dữ liệu và chỉ hiển thị thông báo “không tìm thấy dữ liệu”. Ngoài ra, dữ liệu AQI được sử dụng chủ yếu từ một số trạm đo cố định và chưa phủ khắp các quận, huyện tại TP.HCM, gây hạn chế trong việc phản ánh chi tiết tình trạng ô nhiễm không khí tại từng khu vực cụ thể.

Nguyen, P. H. và cộng sự [3] đã thực hiện nghiên cứu nhằm xây dựng và đánh giá hiệu suất của các mô hình học máy và học sâu trong dự báo nồng độ bụi mịn PM<sub>2.5</sub> tại Thành phố Hồ Chí Minh. Trước thực trạng ô nhiễm không khí ngày càng nghiêm trọng, đặc biệt là từ bụi PM<sub>2.5</sub>, việc xây dựng một hệ thống dự báo chính xác có ý nghĩa quan trọng trong việc đưa ra các giải pháp can thiệp kịp thời để bảo vệ sức khỏe cộng đồng và môi trường. Trong khuôn khổ nghiên cứu, nhóm tác giả đã thử nghiệm sáu thuật toán gồm: rừng ngẫu nhiên (RF), cây tăng cường độ dốc (XGB), hồi quy véc-tơ hỗ trợ (SVR), mạng nơ-ron nhân tạo (ANN), mạng nơ-ron hồi quy tổng quát (Generalized regression neural network - GRNN) và mạng nơ-ron tích chập (CNN).

Nguồn dữ liệu sử dụng bao gồm nồng độ bụi PM2.5 và các yếu tố khí tượng được thu thập liên tục trong 911 ngày, từ 1/1/2021 đến 30/6/2023. Dữ liệu khí tượng lấy từ trạm Tân Sơn Hòa, bao gồm: nhiệt độ, độ ẩm tương đối, tốc độ gió, lượng mưa, thời gian nắng và mức độ bay hơi. Việc kết hợp các yếu tố khí tượng với dữ liệu PM2.5 giúp nâng cao khả năng dự đoán, vì điều kiện thời tiết có ảnh hưởng lớn đến sự biến động của bụi mịn.

Trong quá trình thực hiện, các nhà nghiên cứu đã tích hợp nhiều mô hình để huấn luyện và so sánh hiệu suất của chúng. Kết quả cho thấy ANN là mô hình hoạt động hiệu quả nhất với chỉ số phù hợp (Indicator of Attack - IOA) đạt 0.736 và sai số dự báo thấp nhất trong giai đoạn kiểm thử, cho thấy khả năng mô hình hóa tốt các mối quan hệ phi tuyến giữa các biến đầu vào và nồng độ PM2.5. Ngoài ra, mô hình CNN cũng cho kết quả khả quan nhờ khả năng xử lý dữ liệu lớn và độ chính xác cao. Ngược lại, các mô hình truyền thống như SVR và XGB có hiệu suất kém hơn trong trường hợp này. Quy trình hoạt động chi tiết của hệ thống được minh họa trong hình 5.



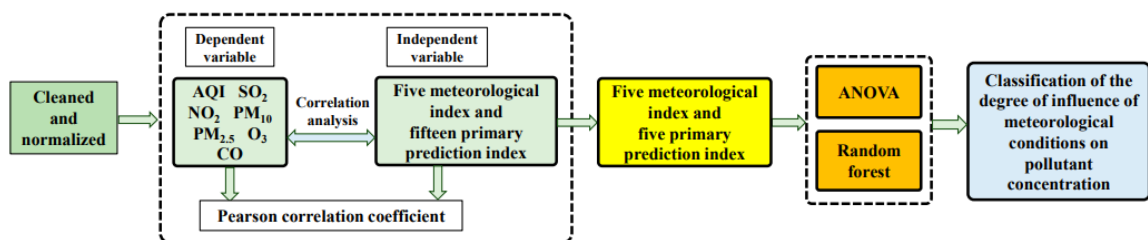
Hình 5: Workflow của mô hình dự đoán PM2.5

Mặc dù nghiên cứu đạt được nhiều kết quả tích cực, vẫn tồn tại một số hạn chế. Thứ nhất, dữ liệu được thu thập từ một trạm duy nhất có thể chưa phản ánh đầy đủ đặc điểm khí hậu và các nguồn phát thải trên toàn TP.HCM. Thứ hai, các mô hình dù có độ chính xác cao nhưng vẫn dễ bị ảnh hưởng bởi tính biến động và không chắc chắn của dữ liệu khí tượng. Cuối cùng, nghiên cứu mới chỉ tập trung vào dự báo ngắn hạn, chưa xét đến yếu tố dài hạn như xu hướng phát thải hoặc tác động của biến đổi khí hậu. Do đó, trong tương lai, việc mở rộng phạm vi dữ liệu và kết hợp các phương pháp tích hợp đa mô hình sẽ giúp nâng cao hơn nữa độ chính xác và khả năng ứng dụng thực tiễn của hệ thống dự báo.

- **Các nghiên cứu quốc tế:**

Nghiên cứu của Liu, Q. cùng nhóm nghiên cứu [4] đã đề xuất phương pháp ứng dụng các kỹ thuật học máy hiện đại để dự đoán chất lượng không khí tại Jinan, Trung Quốc. Trong đó, mô hình Light Gradient Boosting Machine (LightGBM) đạt độ chính xác lên tới 97.5%, trong khi mô hình Long Short-Term Memory (LSTM) cho kết quả dự báo chỉ số chất lượng không khí (AQI) với hệ số xác định ( $R^2$ ) đạt 91.37%. Các phân tích cũng chỉ ra rằng các yếu tố khí tượng như nhiệt độ, độ ẩm và áp suất không khí có ảnh hưởng đáng kể đến nồng độ của các chất ô nhiễm chủ yếu. Một điểm nổi bật của nghiên cứu là việc ứng dụng hai thuật toán tiên tiến – LightGBM và LSTM – để nâng cao hiệu suất dự đoán và hiểu rõ hơn vai trò của các yếu tố khí tượng trong ô nhiễm không khí.

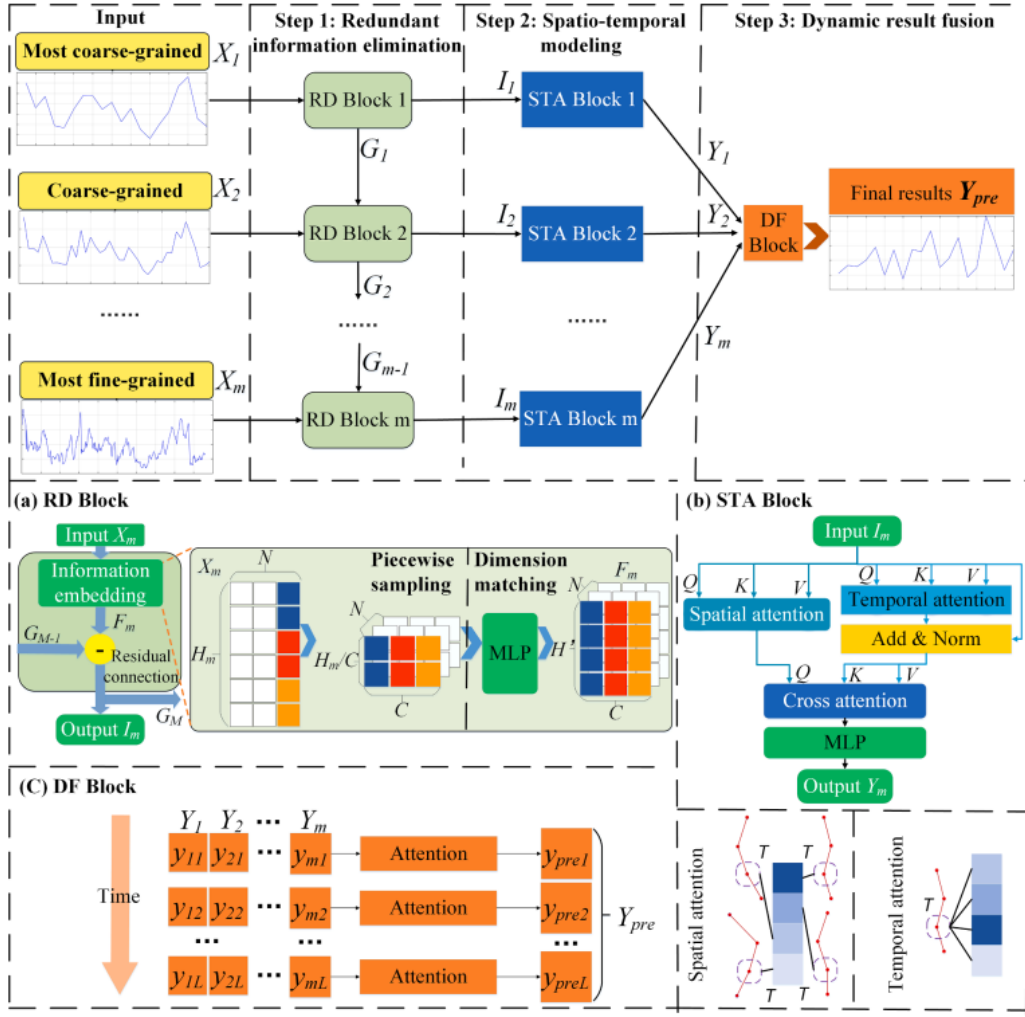
Các kết quả trên được rút ra từ bộ dữ liệu thu thập tại các trạm quan trắc không khí ở Jinan, Trung Quốc, trong khoảng thời gian từ ngày 23 tháng 6 năm 2020 đến ngày 13 tháng 7 năm 2021. Bộ dữ liệu thu thập gồm những chất liên quan đến chất lượng không khí như  $SO_2$ ,  $NO_2$ ,  $PM_{2.5}$ ,  $PM_{10}$ ,  $O_3$ ,  $CO$ . Quy trình hoạt động của hệ thống mô hình được trình bày trong sơ đồ phân tích tại hình 6.



Hình 6: Flowchart phân tích các điều kiện và đặc điểm khí tượng.

Tuy nhiên, nghiên cứu trên vẫn tồn tại một số hạn chế. Trước hết, dữ liệu được sử dụng chỉ giới hạn trong một khoảng thời gian quá khứ cố định và không có tính cập nhật theo thời gian thực, dẫn đến việc mô hình khó thích ứng và phản ánh kịp thời những biến động hiện tại của chất lượng không khí. Bên cạnh đó, dữ liệu đầu ra của mô hình không được dự báo theo từng giờ, trong khi thực tế cho thấy chỉ số chất lượng không khí có thể thay đổi đáng kể trong từng giờ. Điều này gây khó khăn trong việc nhận diện kịp thời các biến động bất thường trong ngắn hạn, từ đó làm giảm hiệu quả của các giải pháp ứng phó được đề xuất.

Yu, C. cùng nhóm nghiên cứu [5] đề xuất mô hình MGSFormer, một kiến trúc Transformer mới nhằm cải thiện độ chính xác trong dự đoán chất lượng không khí. MGSFormer tận dụng hai đặc trưng quan trọng của dữ liệu môi trường là tính đa phân giải và tương quan không gian-thời gian, thông qua ba khối chức năng: khối loại bỏ thông tin dư thừa giữa các mức phân giải (Residual De-redundant Block), khối chú ý không gian - thời gian (Spatiotemporal Attention Block), và khối hợp nhất động (Dynamic Fusion Block). Kết quả thực nghiệm trên ba bộ dữ liệu thực tế cho thấy MGSFormer vượt trội so với 11 mô hình tiên tiến khác như Airformer, DSformer và TimeMixer, với độ chính xác tăng trung bình khoảng 5% theo các chỉ số MAE, MSE và CORR. Phương pháp thực hiện của mô hình này được mô tả trong sơ đồ ở hình 7.



Hình 7: Framework tổng thể của mô hình được đề xuất. Đầu vào của mô hình là dữ liệu chất lượng không khí với các mức độ chi tiết khác nhau. Đầu ra của mô hình là dữ liệu chất lượng không khí. (a) Cấu trúc cụ thể của khối residual de-redundant. (b) Cấu trúc cụ thể của khối spatiotemporal attention. (c) Cấu trúc cụ thể của khối dynamic fusion.

Các thực nghiệm được thực hiện trên ba tập dữ liệu lớn gồm: dữ liệu từ 35 trạm đo ở Bắc Kinh, dữ liệu từ 350 thành phố và dữ liệu từ 1200 trạm tại Trung Quốc, với chuỗi thời gian kéo dài từ năm 2015 đến 2021. Các tập dữ liệu này có độ phân giải khác nhau từ 1 ngày đến 1 giờ, được thiết kế nhằm phản ánh đặc điểm đa phân giải và quy mô không gian rộng lớn. Chỉ số AQI là chỉ số chính được sử dụng, kèm theo các chất ô nhiễm quan trọng như PM2.5, PM10, SO<sub>2</sub> và NO<sub>2</sub>.

Việc cho phép dự đoán đầu ra với số giờ linh hoạt giúp mô hình trở nên thuận tiện hơn cho người sử dụng và các nhà quản lý trong việc lựa chọn các giải pháp ứng phó phù hợp dựa trên thông tin dự báo cụ thể mà họ nhận được. Tuy nhiên, mô hình được huấn luyện chung cho toàn bộ các trạm quan trắc, điều này dẫn đến hạn chế trong khả năng dự đoán đặc thù tại từng khu vực cụ thể, trong khi chất lượng không khí thường bị chi phối bởi các yếu tố mang tính địa phương như hoạt động dân cư, giao thông hay công nghiệp. Thêm vào đó, do dữ liệu không được cập nhật theo thời gian thực, mô hình gặp nhiều khó khăn trong việc phản ánh kịp thời các biến động nhanh chóng và bất thường trong chất lượng không khí, làm giảm hiệu quả trong ứng dụng thực tiễn.

### 3. Mục tiêu nghiên cứu

Đề tài **“Phát triển mô hình học trực tuyến kết hợp nội suy không gian để dự đoán chất lượng không khí tại TP.HCM”** được thực hiện với mục tiêu phát triển một hệ thống có khả năng dự báo chất lượng không khí trong vòng 24 giờ tiếp theo tại từng khu vực cụ thể trên địa bàn thành phố. Hệ thống này được kỳ vọng sẽ cung cấp thông tin nhanh chóng và chính xác cho người dân và các cơ quan chức năng, từ đó hỗ trợ việc đưa ra các biện pháp ứng phó kịp thời nhằm bảo vệ sức khỏe cộng đồng và nâng cao hiệu quả quản lý môi trường đô thị.

Mô hình học máy trực tuyến sẽ được tích hợp vào nền tảng website do tôi phát triển, cho phép hệ thống cập nhật dữ liệu mới và điều chỉnh dự báo một cách linh hoạt theo thời gian thực. Để triển khai hệ thống này, tôi sẽ tiến hành thu thập dữ liệu chất lượng không khí tại TP.HCM, thiết lập quy trình huấn luyện mô hình học trực tuyến, đồng thời đánh giá và hiệu chỉnh mô hình nhằm tối ưu hóa độ chính xác dự báo.

Trên cơ sở đó, đề tài đặt ra các mục tiêu nghiên cứu chính như sau:

- Phát triển và huấn luyện mô hình học máy trực tuyến để dự đoán chất lượng không khí trong thời gian thực.
- Xây dựng hệ thống để dự đoán chất lượng không khí tại những địa điểm lân cận không có trạm.
- Triển khai ứng dụng web để dự đoán chất lượng không khí tại TP.HCM theo thời gian thực.



- Cung cấp khuyến nghị và tư vấn dựa trên kết quả dự báo, hỗ trợ người dùng và nhà quản lý trong việc cải thiện chất lượng không khí và nâng cao chất lượng cuộc sống.

#### 4. Phát biểu bài toán

Để dự đoán chất lượng không khí tại Thành phố Hồ Chí Minh, ta tiến hành dự đoán chất lượng không khí tại từng trạm quan trắc trên địa bàn. Dựa trên kết quả dự đoán từ các trạm này, có thể suy ra mức độ ô nhiễm không khí tại các khu vực khác trong thành phố.

Bài toán được phát biểu như sau: **Cho biết chỉ số chất lượng không khí (AQI) trong 72 giờ gần nhất tại một trạm đo cụ thể, hãy dự đoán chỉ số AQI trong 24 giờ tiếp theo.** Hình 8 minh họa dữ liệu đầu vào và đầu ra của hệ thống.



Hình 8: Input và output của bài toán.

Sau khi thu thập và dự đoán chỉ số AQI tại các trạm, ta sử dụng các giá trị này làm cơ sở để ước lượng chất lượng không khí tại những khu vực không có trạm đo, chẳng hạn như các quận, huyện khác trong thành phố.

- **Input:** Chỉ số AQI của những trạm đã biết
- **Output:** Chỉ số AQI của những quận huyện khác trong Thành phố Hồ Chí Minh.

## 5. Đối tượng và phạm vi nghiên cứu

Nghiên cứu này hướng đến việc phát triển một mô hình học máy trực tuyến nhằm dự đoán chất lượng không khí một cách nhanh chóng và chính xác, với mục tiêu đồng thời tối ưu hóa độ chính xác dự báo và hiệu suất xử lý trong môi trường thời gian thực. Trên cơ sở các kết quả dự đoán, hệ thống có khả năng đánh giá tức thời mức độ ô nhiễm không khí, hỗ trợ phát hiện sớm các tình huống bất thường, đồng thời đưa ra các khuyến nghị phù hợp nhằm phục vụ công tác quản lý và bảo vệ sức khỏe cộng đồng. Hiện tại, mô hình đang được triển khai để dự đoán chất lượng không khí tại TP.HCM trong giai đoạn từ ngày 13 tháng 05 đến ngày 24 tháng 06 năm 2025.

Trong khuôn khổ nghiên cứu này, mô hình không đi sâu phân tích các yếu tố vật lý, hóa học hoặc xã hội có thể tác động đến chất lượng không khí như giao thông, công nghiệp, điều kiện khí tượng hay hoạt động dân cư. Thay vào đó, nghiên cứu tập trung vào việc khai thác đặc điểm chuỗi thời gian của chỉ số chất lượng không khí (AQI) để nhận diện và dự đoán xu hướng biến động theo thời gian, qua đó phục vụ mục tiêu cảnh báo sớm và hỗ trợ ra quyết định trong quản lý môi trường.

## 6. Phương pháp nghiên cứu

Trong khuôn khổ nghiên cứu này, tôi áp dụng kết hợp giữa các phương pháp lý thuyết và thực tiễn nhằm đảm bảo tính chính xác và khả năng triển khai của mô hình dự đoán chất lượng không khí. Cụ thể, các phương pháp được sử dụng bao gồm:

### **Phương pháp lý thuyết:**

- *Mô hình hóa*: Thiết kế và xây dựng mô hình học sâu (deep learning) để dự đoán chỉ số chất lượng không khí dựa trên dữ liệu thời gian thực.
- *Thu thập dữ liệu*: Tiến hành thu thập dữ liệu từ các nguồn đáng tin cậy, bao gồm hệ thống cảm biến đo lường và các cơ sở dữ liệu công khai hiện có.

### **Phương pháp thực tiễn:**

- *Thực nghiệm*: Triển khai quá trình huấn luyện và kiểm thử mô hình trên tập dữ liệu đã thu thập, từ đó đánh giá mức độ chính xác trong dự đoán.

- *Đánh giá*: Thực hiện phân tích và so sánh kết quả dự báo của mô hình với dữ liệu thực tế nhằm hiệu chỉnh mô hình và nâng cao hiệu suất cũng như tính ứng dụng trong điều kiện thực tiễn.

## 7. Cấu trúc khóa luận

Khóa luận được chia thành 4 chương với những nội dung chính như sau:

**Chương 1: Cơ sở lý thuyết:** Giới thiệu các phương pháp dự đoán và thuật toán máy học sử dụng cho việc dự đoán chất lượng không khí tại TP.HCM trong báo cáo khóa luận này.

**Chương 2: Thiết kế hệ thống:** Thiết kế pipeline bài toán và trình bày chi tiết quy trình hoạt động của hệ thống, các công cụ, thuật toán, phương pháp sử dụng cho bài toán này và cho từng giai đoạn trong bài toán. Ngoài ra, thiết kế cấu trúc, giao diện website để thực hiện, kiểm thử những chức năng của hệ thống.

**Chương 3: Thực nghiệm và đánh giá:** Trình bày quá trình, môi trường cài đặt, xây dựng hệ thống, sau đó tiến hành thực nghiệm từng chức năng của hệ thống. Từ đó đưa ra đánh giá hệ thống và so sánh với những nghiên cứu mới nhất.

**Chương 4: Kết luận và hướng phát triển:** Tổng kết đề tài, những điểm mạnh, hạn chế của đề tài và hướng phát triển trong tương lai, đồng thời nêu bật tác động tích cực và tiêu cực của đề tài đến xã hội.

## Chương 1. CƠ SỞ LÝ THUYẾT

### 1.1 Các thuật toán sử dụng:

#### 1.1.1 Kỹ thuật học máy trực tuyến

Học máy trực tuyến (Online Machine Learning) là một nhánh của học máy (Machine Learning) tập trung vào khả năng cập nhật mô hình liên tục khi dữ liệu mới xuất hiện, thay vì huấn luyện toàn bộ mô hình trên một tập dữ liệu cố định như trong học máy truyền thống (batch learning). Phương pháp này đặc biệt phù hợp với các ứng dụng yêu cầu xử lý dữ liệu theo thời gian thực hoặc dữ liệu có tốc độ phát sinh cao, chẳng hạn như giám sát chất lượng không khí, giao dịch tài chính, hoặc hệ thống khuyến nghị trực tuyến.

Trong học máy trực tuyến, mô hình được cập nhật liên tục từng bước một, mỗi khi có một điểm dữ liệu mới được đưa vào. Sau mỗi bước, mô hình tiến hành điều chỉnh các tham số dựa trên sai số giữa dự đoán hiện tại và giá trị thực tế. Quá trình này giúp mô hình thích nghi linh hoạt với các thay đổi trong dữ liệu, đồng thời giảm thiểu nhu cầu lưu trữ toàn bộ tập dữ liệu huấn luyện – một yếu tố quan trọng trong các hệ thống có giới hạn về bộ nhớ và thời gian tính toán. Khác với học tăng cường (reinforcement learning), học máy trực tuyến không cần phải tối ưu hóa chính sách hành động qua từng bước tương tác với môi trường, mà chỉ tập trung vào việc cải thiện hiệu suất dự đoán dựa trên luồng dữ liệu đầu vào liên tục.

Một số thuật toán phổ biến hỗ trợ học trực tuyến bao gồm:

- **Stochastic Gradient Descent (SGD)**: là nền tảng cơ bản cho nhiều mô hình học trực tuyến, cho phép cập nhật trọng số sau mỗi lần quan sát dữ liệu mới.
- **Hoeffding Trees**: là dạng cây quyết định có thể được xây dựng và cập nhật trực tuyến, phù hợp cho phân loại dữ liệu theo luồng.
- **Online versions của các mô hình học sâu**, tiêu biểu là các biến thể trực tuyến của Long Short-Term Memory (Online LSTM), cho phép học từ chuỗi thời gian trong môi trường thay đổi liên tục.

- **Adaptive algorithms** như Adaptive Random Forest hoặc Online Bagging, kết hợp các mô hình con để cải thiện độ chính xác và tính ổn định.

Học máy trực tuyến mang lại nhiều lợi ích trong các hệ thống dự đoán theo thời gian thực:

- **Tính thích ứng cao:** Mô hình có khả năng phản ứng linh hoạt với sự thay đổi trong phân phối dữ liệu (concept drift).
- **Tiết kiệm tài nguyên:** Không yêu cầu lưu trữ hoặc xử lý toàn bộ tập dữ liệu.
- **Khả năng mở rộng:** Dễ dàng tích hợp trong các hệ thống IoT hoặc cơ sở dữ liệu luồng.

Tuy nhiên, học trực tuyến cũng đặt ra một số thách thức, đặc biệt trong việc duy trì sự ổn định của mô hình trong điều kiện dữ liệu nhiễu, không cân bằng, hoặc thay đổi đột ngột. Ngoài ra, việc lựa chọn siêu tham số (hyperparameters) và cơ chế dừng cập nhật hiệu quả vẫn là những vấn đề mở cần được nghiên cứu thêm.

### 1.1.2 Phương pháp Inverse Distance Weighting (IDW)

Inverse Distance Weighting (IDW) là một phương pháp nội suy không gian được sử dụng rộng rãi trong lĩnh vực địa lý, khí tượng và môi trường nhằm ước lượng giá trị tại một điểm chưa biết dựa trên các giá trị đã biết từ các điểm lân cận. Nguyên lý cơ bản của IDW dựa trên giả định rằng những điểm gần nhau trong không gian có xu hướng có giá trị tương tự nhau – tức là ảnh hưởng của một điểm dữ liệu sẽ giảm dần khi khoảng cách từ điểm đó đến vị trí cần nội suy tăng lên.

Trong phương pháp IDW, giá trị tại điểm cần xác định  $x_0$  được tính bằng trung bình có trọng số các giá trị tại các điểm quan trắc lân cận  $x_i$ . Trọng số  $w_i$  được xác định dựa trên khoảng cách từ  $x_0$  đến  $x_i$  theo công thức:

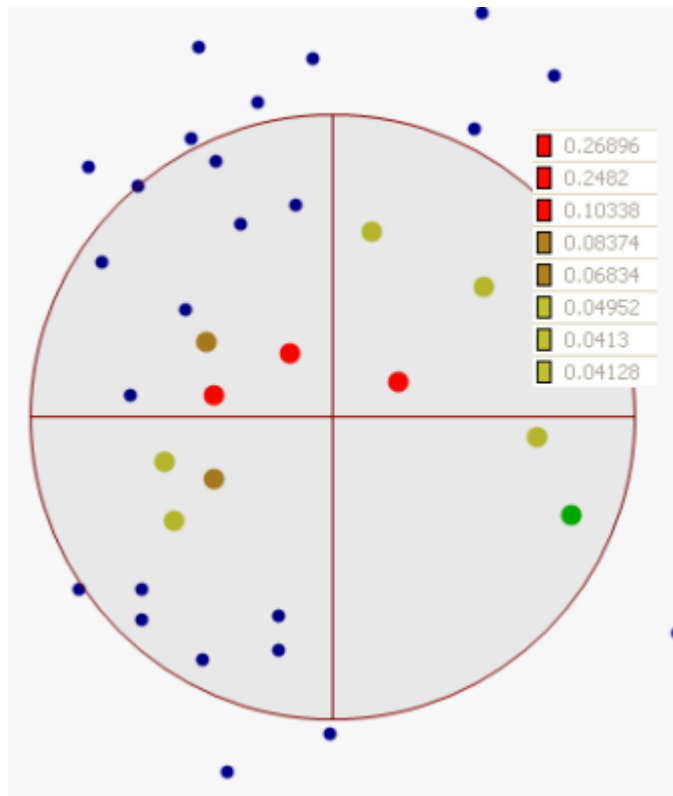
$$\hat{f}(x_0) = \frac{\sum_{i=1}^N w_i f(x_i)}{\sum_{i=1}^N w_i}, \text{ trong đó } w_i = \frac{1}{d(x_0, x_i)^p} \quad (1)$$

Trong đó:

- $\hat{f}(x_0)$  là giá trị được nội suy tại điểm cần dự đoán,
- $f_i$  là giá trị quan trắc tại điểm  $x_i$ .
- $d(x_0, x_i)$  là khoảng cách giữa điểm cần dự đoán và điểm quan trắc,
- $p$  là hệ số ảnh hưởng (thường được chọn trong khoảng từ 1 đến 3), điều chỉnh mức độ ảnh hưởng của khoảng cách.

Giá trị của số mũ  $p$  đóng vai trò kiểm soát mức độ suy giảm ảnh hưởng theo khoảng cách. Khi  $p$  tăng, những điểm gần điểm cần nội suy sẽ có ảnh hưởng lớn hơn so với những điểm xa hơn.

Hình 9 trong báo cáo minh họa cách thuật toán IDW xác định các điểm gần nhất với điểm cần nội suy, tính toán khoảng cách, và sau đó gán trọng số tương ứng để thực hiện phép nội suy. Trong thực tế, các thuật toán tìm kiếm lân cận như **k-nearest neighbors** thường được tích hợp để tối ưu quá trình chọn điểm và tăng hiệu suất xử lý.



Hình 9: Minh họa tìm kiếm những điểm lân cận.

Phương pháp IDW được đánh giá cao nhờ tính đơn giản, dễ triển khai và hiệu quả trong các tình huống có mật độ điểm quan trắc dày đặc và phân bố đều. Nhờ khả năng phản ánh trực tiếp mối quan hệ khoảng cách trong không gian, IDW cho phép nội suy nhanh chóng mà không đòi hỏi mô hình thống kê phức tạp hay giả định về phân phối dữ liệu.

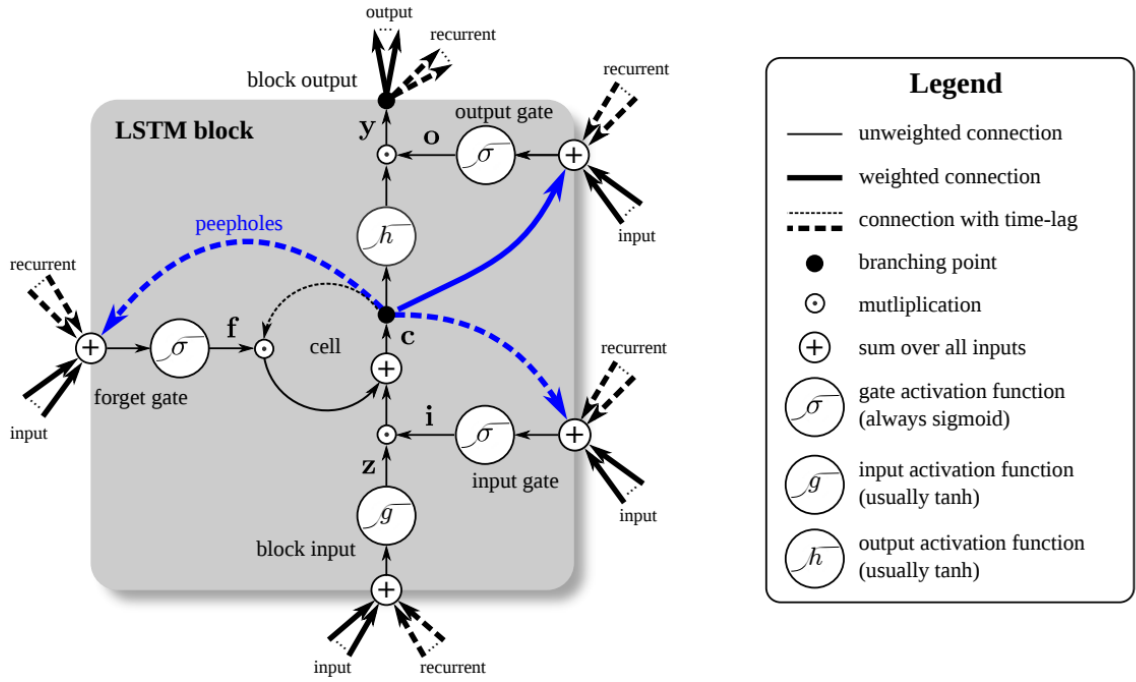
Tuy nhiên, so với các phương pháp nội suy nâng cao như Kriging – vốn xem xét cả cấu trúc không gian và phương sai của dữ liệu – thì IDW có hạn chế khi xử lý dữ liệu phân bố không đều hoặc khi hiện tượng cần dự đoán có tính biến động cao. Ngoài ra, IDW không tự động hiệu chỉnh sai số hay đưa ra ước lượng độ tin cậy cho kết quả nội suy, điều mà Kriging có thể cung cấp.

## 1.2 Các thuật toán học máy:

### 1.2.1 Long-Short Term Memory (LSTM) [6]

Long Short-Term Memory (LSTM) là một kiến trúc mạng nơ-ron hồi tiếp được thiết kế nhằm khắc phục hạn chế lớn nhất của RNN truyền thống: vấn đề tiêu biến hoặc bùng nổ gradient. Hiện tượng này khiến RNN khó học được các phụ thuộc dài hạn trong chuỗi dữ liệu. LSTM được giới thiệu như một giải pháp hiệu quả bằng cách đưa vào một khối bộ nhớ chuyên biệt gọi là “constant error carousel” (CEC), giúp duy trì dòng chảy gradient không đổi theo thời gian. CEC giữ vai trò lưu trữ trạng thái nội tại của các đơn vị nhớ (memory cells), nhờ đó bảo tồn thông tin trong thời gian dài mà không bị suy giảm tín hiệu.

Kiến trúc của mô hình LSTM được nhắc đến trong hình 10. LSTM mở rộng kiến trúc RNN bằng cách bổ sung các cổng điều khiển—bao gồm cổng vào (input gate), cổng ra (output gate) và cổng quên (forget gate). Các cổng này kiểm soát luồng thông tin vào, ra và duy trì trong ô nhớ thông qua cơ chế học dựa trên các hàm kích hoạt sigmoid. Đặc biệt, cổng quên đóng vai trò thiết yếu trong việc điều chỉnh khả năng “quên” thông tin đã lưu trữ, từ đó bảo vệ bộ nhớ khỏi việc tích lũy thông tin không còn hữu ích, một khả năng quan trọng trong xử lý dữ liệu tuần tự liên tục như giọng nói hoặc văn bản.



Hình 10: Sơ đồ chi tiết của khối LSTM được sử dụng trong các hidden layer của recurrent neural network.

Quá trình huấn luyện LSTM sử dụng phương pháp lai giữa lan truyền ngược theo thời gian (Backpropagation Through Time - BPTT) và học hồi tiếp thời gian thực (Real-Time Recurrent Learning - RTRL). BPTT chủ yếu được áp dụng cho các thành phần phía sau ô nhớ, trong khi RTRL điều chỉnh các trọng số phía trước, đặc biệt là trong và trước các ô nhớ. Cách kết hợp này cho phép mạng học được cả khi không có nhãn ở mọi bước thời gian.

Với những cải tiến đáng kể, LSTM đã chứng minh được hiệu quả vượt trội trong các tác vụ như nhận dạng giọng nói, dịch máy, và phân tích chuỗi thời gian, trở thành nền tảng cho nhiều biến thể hiện đại như BiLSTM, Grid LSTM và GRU. Tuy nhiên, LSTM vẫn đối mặt với một số hạn chế như không thể thay đổi động cấu trúc bộ nhớ, và việc mở rộng quy mô mạng có thể làm tăng độ phức tạp tính toán đáng kể (Staudemeyer & Morris, 2019).



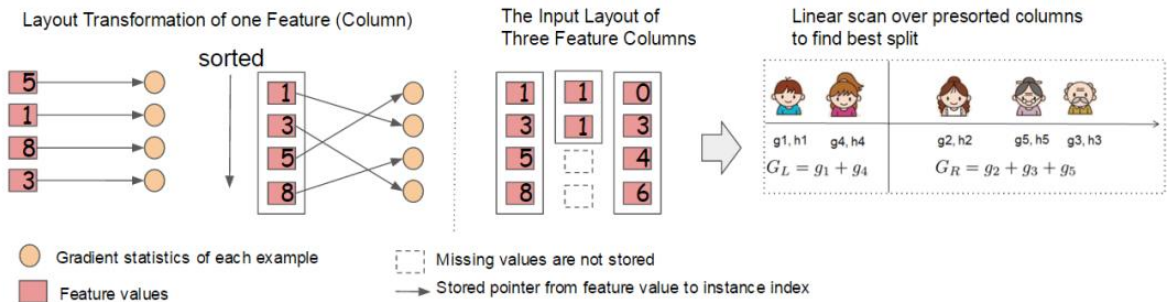
### 1.2.2 *LightGBM* [7]

*LightGBM* [7] là một thuật toán tăng cường cây quyết định gradient (Gradient-Boosted Decision Trees - GBDT) được phát triển nhằm tối ưu hiệu quả xử lý và khả năng mở rộng khi làm việc với dữ liệu có kích thước lớn và độ chiều cao đặc trưng cao. Trọng tâm của *LightGBM* là hai kỹ thuật cốt lõi: Gradient-based One-Side Sampling (GOSS) và Exclusive Feature Bundling (EFB), nhằm giải quyết giới hạn về thời gian huấn luyện và tài nguyên tính toán của các triển khai GBDT truyền thống. GOSS khai thác thông tin gradient để chọn lọc những mẫu dữ liệu quan trọng nhất trong quá trình huấn luyện, từ đó giảm đáng kể số lượng mẫu mà không làm mất tính chính xác. Cụ thể, các mẫu có gradient lớn — tức là các điểm dữ liệu có lỗi dự đoán cao — được giữ lại toàn bộ, trong khi các mẫu có gradient nhỏ được lấy mẫu ngẫu nhiên với tỷ lệ nhất định. Để đảm bảo phân phối dữ liệu không bị thay đổi, các mẫu được lấy ngẫu nhiên này sẽ được nhân với một hệ số bù khi tính toán thông tin chia tách.

Bên cạnh đó, EFB là một kỹ thuật giảm số lượng đặc trưng bằng cách gộp các đặc trưng rời rạc, hiếm khi cùng xuất hiện (tức là tương hỗ loại trừ) thành một đặc trưng duy nhất. Quá trình này được thực hiện thông qua một thuật toán tham lam dựa trên bài toán tô màu đồ thị, trong đó mỗi đặc trưng là một đỉnh và các cạnh biểu diễn sự xung đột không thể gộp. Nhờ đó, *LightGBM* có thể giảm số lượng đặc trưng hiệu quả mà vẫn giữ nguyên thông tin cần thiết để đưa ra quyết định phân nhánh chính xác trong cây quyết định. Ngoài ra, *LightGBM* sử dụng thuật toán dựa trên histogram giúp cải thiện tốc độ xây dựng cây quyết định và tối ưu bộ nhớ bằng cách lượng hóa các giá trị đặc trưng liên tục thành các bin rời rạc. Thử nghiệm thực nghiệm trên nhiều tập dữ liệu thực tế cho thấy *LightGBM* có thể tăng tốc huấn luyện gấp hơn 20 lần so với các phương pháp GBDT truyền thống như XGBoost, đồng thời duy trì được độ chính xác cao.

### 1.2.3 XGBoost [8]

XGBoost (Extreme Gradient Boosting) là một thuật toán học máy tiên tiến, thuộc nhóm các phương pháp tăng cường (boosting), nổi bật với khả năng mở rộng tốt và hiệu quả trong xử lý dữ liệu lớn cũng như dữ liệu thưa. XGBoost dựa trên nguyên lý của boosting cây quyết định, cụ thể là gradient boosting, trong đó mô hình được xây dựng một cách tuần tự bằng cách tối thiểu hóa hàm mất mát với sự trợ giúp của đạo hàm bậc nhất và bậc hai. Điều này cho phép thuật toán học được các hàm phi tuyến tính phức tạp với tốc độ hội tụ nhanh và độ chính xác cao.



Hình 11: Kiến trúc khối học song song. Mỗi cột trong khối được sắp xếp theo giá trị tính năng tương ứng.

Kiến trúc của mô hình XGBoost được mô tả trong hình 11. Một trong những đóng góp chính của XGBoost là việc đề xuất hàm mục tiêu có điều chuẩn, giúp kiểm soát độ phức tạp của mô hình và giảm nguy cơ overfitting. Cấu trúc mô hình của XGBoost bao gồm một tổ hợp cộng của các cây hồi quy (regression trees), trong đó mỗi cây được học để sửa lỗi còn lại của các cây trước đó. Ngoài ra, XGBoost còn sử dụng kỹ thuật shrinkage (giảm trọng số của mỗi cây mới), kết hợp với việc chọn ngẫu nhiên một tập con của các đặc trưng trong mỗi vòng lặp huấn luyện (column subsampling), qua đó tăng cường khả năng tổng quát hóa mô hình.

XGBoost còn nổi bật nhờ các cải tiến về mặt hệ thống như thuật toán tìm điểm chia có ý thức về độ thưa (sparsity-aware split finding), cho phép xử lý hiệu quả dữ liệu thiếu và thưa phổ biến trong các bài toán thực tế. Bên cạnh đó, nó sử dụng cấu trúc dữ liệu đặc biệt giúp truy cập bộ nhớ đệm hiệu quả (cache-aware), từ đó rút ngắn thời gian huấn luyện, đặc biệt khi làm việc với dữ liệu lớn. Đáng chú ý, XGBoost còn hỗ trợ tính toán ngoài bộ nhớ (out-of-core computing), giúp xử lý tập dữ liệu quy mô hàng tỷ dòng ngay cả trên máy tính cá nhân.

Một số ứng dụng thành công của XGBoost có thể kể đến như dự đoán tỉ lệ nhấp chuột trong quảng cáo, phân loại sự kiện vật lý năng lượng cao, dự đoán hành vi khách hàng, và nó thường xuyên xuất hiện trong các giải pháp chiến thắng tại các cuộc thi trên Kaggle. Chính khả năng mở rộng vượt trội và độ chính xác cao đã khiến XGBoost trở thành một trong những công cụ tiêu chuẩn trong kho vũ khí của các nhà khoa học dữ liệu hiện đại.

## KẾT LUẬN CHƯƠNG

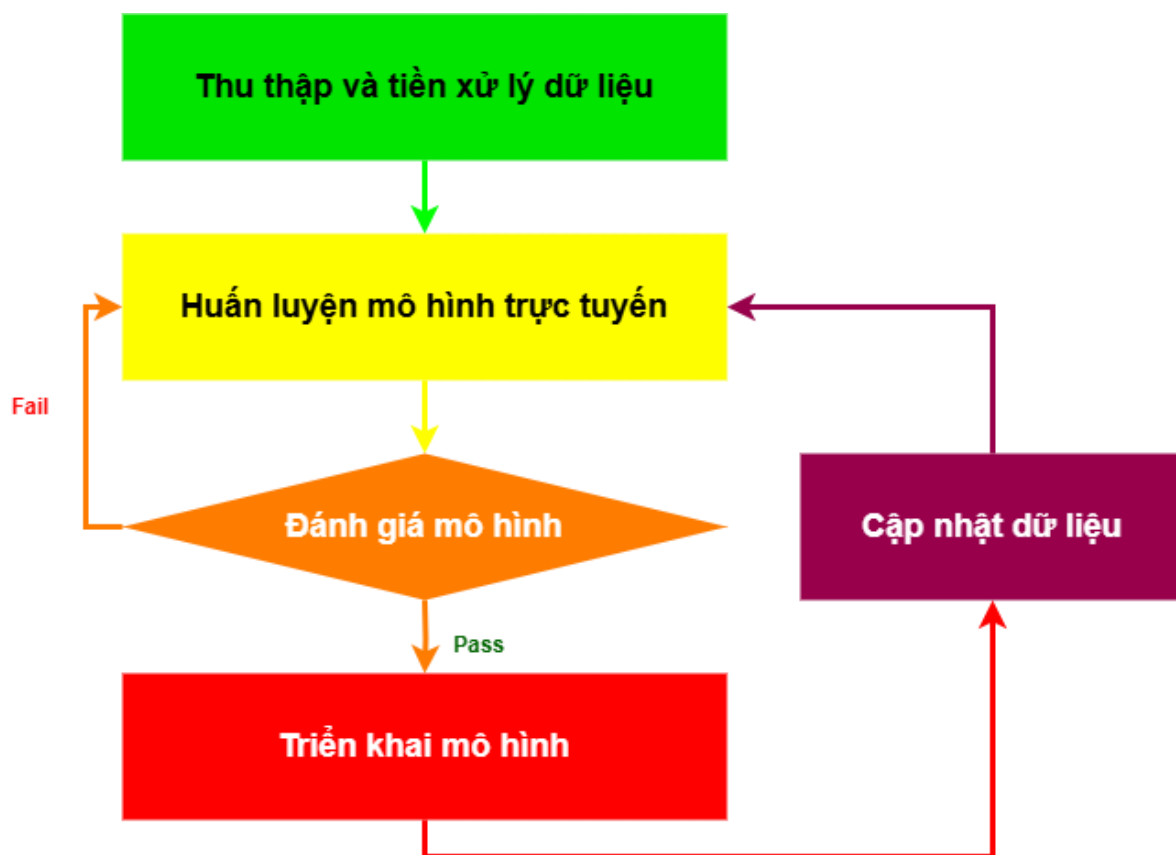
Trong chương này, các khái niệm nền tảng và thuật toán quan trọng đã được trình bày nhằm xây dựng cơ sở lý thuyết cho bài toán dự đoán chất lượng không khí theo thời gian thực. Đầu tiên, kỹ thuật **học máy trực tuyến** được giới thiệu như một giải pháp phù hợp để xử lý dữ liệu thay đổi liên tục, đặc biệt trong bối cảnh dữ liệu môi trường luôn biến động. Phương pháp **IDW** được khai thác để nội suy giá trị AQI tại những khu vực không có cảm biến đo lường, hỗ trợ cải thiện độ phủ không gian của hệ thống dự đoán. Các thuật toán học máy mạnh như **LSTM**, **LightGBM** và **XGBoost** lần lượt được phân tích chi tiết về nguyên lý hoạt động, ưu điểm, và khả năng áp dụng vào bài toán chuỗi thời gian với dữ liệu lớn.

Những kiến thức lý thuyết này không chỉ giúp hiểu rõ bản chất của từng phương pháp mà còn đóng vai trò then chốt trong việc lựa chọn mô hình, thiết kế quy trình huấn luyện, và tối ưu hệ thống trong các chương sau. Trong chương tiếp theo, các nội dung về thiết kế hệ thống, tập dữ liệu sử dụng, cũng như quy trình xây dựng và đánh giá mô hình sẽ được trình bày cụ thể hơn để hiện thực hóa các lý thuyết đã nêu.

## Chương 2 THIẾT KẾ HỆ THỐNG

### 2.1 Thiết kế mô hình học trực tuyến

Mô hình dự đoán chất lượng không khí ở TP.HCM được thiết kế với input là tập dữ liệu về chất lượng không khí ở TP.HCM được cập nhật liên tục theo mỗi giờ, và output là mô hình dự đoán chất lượng không khí trong 24 giờ tới và được chia theo từng trạm. Pipeline ở hình 12 mô tả quy trình hoạt động của hệ thống.



Hình 12: Pipeline mô hình dự đoán chất lượng không khí ở TP.HCM theo thời gian thực

Hệ thống được thiết kế để thu thập dữ liệu chất lượng không khí theo chu kỳ hàng giờ. Sau khi dữ liệu được thu thập, bước xử lý sơ bộ sẽ được thực hiện nhằm loại bỏ nhiễu và chuẩn hóa đầu vào trước khi đưa vào quy trình huấn luyện mô hình. Mô hình được học theo phương pháp huấn luyện lại mô hình định kỳ dựa trên toàn bộ dữ liệu mới được cập nhật. Sau mỗi chu kỳ huấn luyện, mô hình được đánh giá theo các tiêu chí định sẵn; nếu không đạt yêu cầu, quá trình huấn luyện sẽ được lặp lại cho đến khi đạt hiệu suất mong muốn.

Khi mô hình đạt ngưỡng hiệu suất cần thiết, nó sẽ được tích hợp vào hệ thống website để thực hiện dự đoán chất lượng không khí theo thời gian thực. Dù không sử dụng cơ chế học liên tục, hệ thống vẫn đảm bảo khả năng thích ứng với môi trường bằng cách tái huấn luyện mô hình định kỳ trên dữ liệu mới. Ngoài ra, hiệu suất dự báo của mô hình cũng được đánh giá thường xuyên nhằm đảm bảo tính ổn định; nếu mô hình không còn đáp ứng yêu cầu, một mô hình mới sẽ được huấn luyện và triển khai thay thế để duy trì độ chính xác và tin cậy của hệ thống.

### *2.1.1 Thu thập và tiền xử lý dữ liệu:*

Trong nghiên cứu này dữ liệu chất lượng không khí được thu thập từ tổ chức IQAir<sup>1</sup>. IQAir là một tổ chức quốc tế tiên phong trong lĩnh vực giám sát và cải thiện chất lượng không khí, có trụ sở tại Thụy Sĩ. Với mục tiêu nâng cao nhận thức cộng đồng và hỗ trợ chính sách môi trường, IQAir kết hợp công nghệ cảm biến tiên tiến với dữ liệu thời gian thực để cung cấp thông tin toàn diện về mức độ ô nhiễm không khí tại hàng nghìn điểm đo trên toàn thế giới. Nền tảng của IQAir cho phép người dùng truy cập nhanh chóng vào chỉ số chất lượng không khí (AQI), nhận cảnh báo về ô nhiễm, đồng thời theo dõi các phân tích chuyên sâu về hiện trạng môi trường không khí.

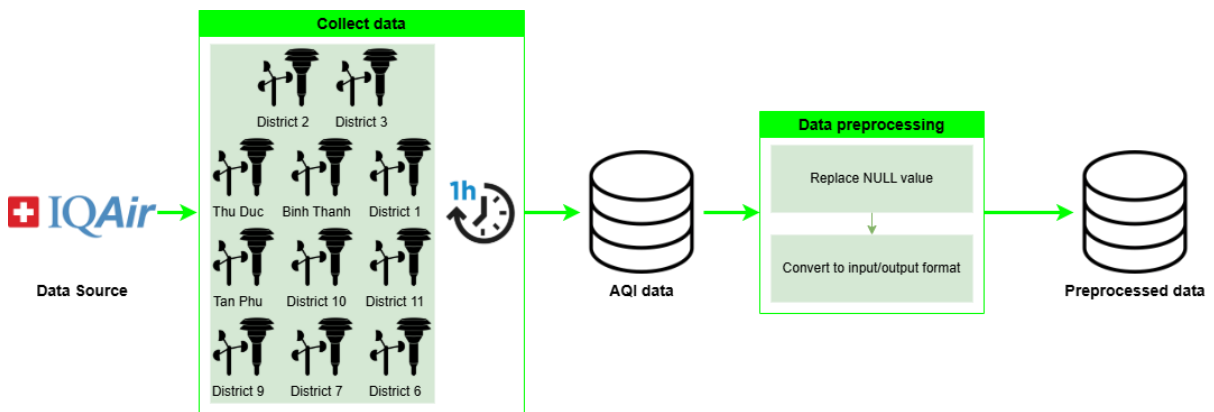
Hệ thống thu thập dữ liệu của IQAir hoạt động dựa trên mạng lưới cảm biến đa tầng, bao gồm: dữ liệu từ các trạm quan trắc chính thức của chính phủ, thiết bị cảm biến được IQAir triển khai độc lập, và các cảm biến cộng đồng do người dùng đóng góp. Nhờ sự phân bố rộng khắp tại nhiều khu vực, từ thành thị đến nông thôn, hệ thống này mang lại độ phủ dữ liệu rộng và tính đại diện cao cho từng vùng địa lý cụ thể.

---

<sup>1</sup> Tổ chức IQAir: [IQAir.com](https://www.iqair.com)

Dữ liệu được cập nhật liên tục theo thời gian thực và bao gồm các thông số chính như PM<sub>2.5</sub>, PM<sub>10</sub>, NO<sub>2</sub>, SO<sub>2</sub>, CO, O<sub>3</sub>, nhiệt độ và độ ẩm. Trong đó, chỉ số PM<sub>2.5</sub> – đại diện cho bụi mịn có đường kính nhỏ hơn 2.5 micromet – được xem là chỉ tiêu trọng yếu do ảnh hưởng nghiêm trọng đến sức khỏe hô hấp và tim mạch. Tất cả dữ liệu được xử lý, hiệu chỉnh và tổng hợp thông qua các hệ thống điện toán đám mây sử dụng thuật toán nội bộ của IQAir. Đồng thời, tổ chức này cũng áp dụng các kỹ thuật trí tuệ nhân tạo và học máy để phát hiện bất thường, cải thiện độ chính xác của dữ liệu và dự báo xu hướng chất lượng không khí trong tương lai.

Tại TP.HCM có rất nhiều trạm đo chất lượng không khí được phân bố ở hầu hết các quận huyện trên địa bàn thành phố. Dữ liệu đã được thu thập từ 11 trạm khác nhau ở những khu vực khác nhau.. Quy trình thu thập và tiền xử lý dữ liệu được đề cập ở hình 13.



Hình 13: Quy trình thu thập và tiền xử lý dữ liệu.

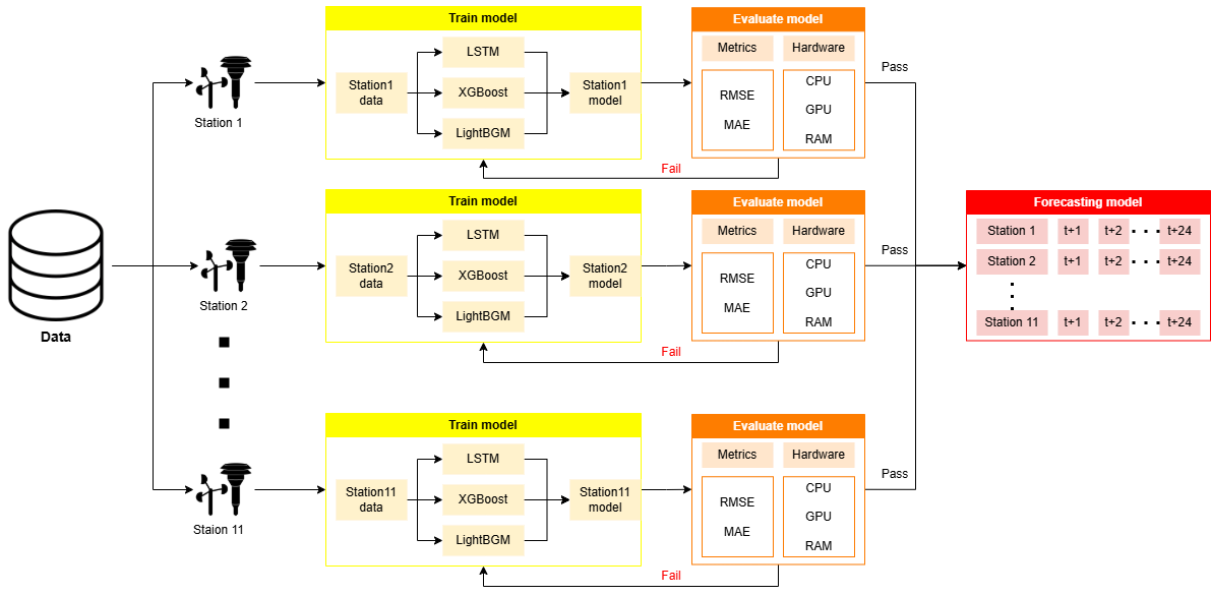
**Thu thập dữ liệu:** Dữ liệu chất lượng không khí được thu thập từ 11 trạm quan trắc được bố trí tại các vị trí đại diện trên địa bàn Thành phố Hồ Chí Minh, bao gồm các quận: Quận 1, Quận 2, Quận 3, Quận 6, Quận 7, Quận 9, Quận 10, Quận 11, Quận 12, Quận Thủ Đức, Tân Phú và Bình Thạnh. Việc thu thập được thực hiện với tần suất mỗi giờ nhằm đảm bảo độ phân giải cao theo thời gian, đồng thời phản ánh chi tiết biến động chất lượng không khí theo từng khu vực cụ thể. Dữ liệu đo lường từ các trạm được tổng hợp và lưu trữ dưới định dạng CSV, tạo điều kiện thuận lợi cho quá trình xử lý, phân tích và trực quan hóa trong các bước nghiên cứu tiếp theo.

**Tiền xử lý dữ liệu:** Dữ liệu sau khi thu thập được tổ chức và lưu trữ theo định dạng chuỗi thời gian, trong đó mỗi mốc thời gian tương ứng với thời điểm đo chỉ số chất lượng không khí (AQI). Tuy nhiên, trong quá trình thu thập, đã xuất hiện các khoảng trống dữ liệu do sự cố kỹ thuật từ thiết bị quan trắc cũng như gián đoạn trong quá trình truyền tải dữ liệu. Để xử lý vấn đề này, nghiên cứu áp dụng phương pháp trung bình động (moving average), trong đó các giá trị AQI bị thiếu được thay thế bằng giá trị trung bình của năm thời điểm gần nhất trước đó. Đối với các khoảng thiếu dữ liệu kéo dài hoặc không thể khắc phục bằng phương pháp trung bình động, kỹ thuật nội suy tuyến tính (linear interpolation) được sử dụng nhằm ước lượng và bổ sung giá trị còn thiếu. Quy trình này giúp đảm bảo tính liên tục và đồng nhất của tập dữ liệu trước khi đưa vào các giai đoạn phân tích và huấn luyện mô hình.

Sau khi xử lý dữ liệu đầu vào, chuỗi AQI được định dạng thành tập dữ liệu có cấu trúc đầu vào–đầu ra phù hợp với mô hình học máy dự báo theo thời gian. Cụ thể, mỗi mẫu dữ liệu đầu vào bao gồm chuỗi giá trị AQI trong 72 giờ liên tiếp ( $t-71$  đến  $t_0$ ), trong khi đầu ra tương ứng là chuỗi AQI dự báo cho 24 giờ kế tiếp ( $t+1$  đến  $t+24$ ). Phương pháp xây dựng dữ liệu sử dụng kỹ thuật trượt cửa sổ (sliding window) trên tập dữ liệu thời gian nhằm tạo thành các cặp đặc trưng–nhãn phục vụ cho quá trình huấn luyện và đánh giá mô hình học trực tuyến.

### 2.1.2 Huấn luyện mô hình trực tuyến:

Để đảm bảo độ phù hợp và khả năng thích ứng với điều kiện môi trường đặc thù tại từng khu vực, dữ liệu chất lượng không khí được sử dụng để huấn luyện riêng biệt cho từng trạm quan trắc. Đối với mỗi trạm, hệ thống triển khai song song ba mô hình học máy trực tuyến khác nhau, bao gồm Long Short-Term Memory (LSTM), Light Gradient Boosting Machine (LightGBM) và Extreme Gradient Boosting (XGBoost). Mỗi mô hình được huấn luyện độc lập trên dữ liệu của từng trạm, tạo thành ba lần huấn luyện riêng biệt nhằm đánh giá và so sánh hiệu suất giữa các mô hình. Từ đó, mô hình có độ chính xác và khả năng dự báo cao nhất sẽ được lựa chọn để triển khai trong hệ thống thực tế. Quy trình chi tiết của quá trình huấn luyện được minh họa trong Hình 14.



Hình 14: Quy trình huấn luyện mô hình dự đoán chất lượng không khí

Việc lựa chọn ba mô hình LSTM, LightGBM và XGBoost để đưa vào so sánh không chỉ xuất phát từ mức độ phổ biến rộng rãi của chúng trong các nghiên cứu liên quan đến dự báo chuỗi thời gian và môi trường, mà còn dựa trên những ưu điểm kỹ thuật đặc thù của từng mô hình trong bối cảnh xử lý dữ liệu chất lượng không khí theo thời gian thực. Cụ thể, mô hình LSTM (Long Short-Term Memory) là một biến thể của mạng nơ-ron hồi tiếp (RNN), được thiết kế đặc biệt để xử lý dữ liệu chuỗi có độ phụ thuộc dài hạn, cho phép mô hình học được các mối quan hệ tiềm ẩn trong chuỗi AQI liên tục và phát hiện các xu hướng phức tạp theo thời gian.

Trong khi đó, LightGBM và XGBoost là hai thuật toán tăng cường (boosting) dựa trên cây quyết định, nổi bật với hiệu suất tính toán cao và khả năng xử lý dữ liệu quy mô lớn với tốc độ huấn luyện nhanh. Những đặc điểm này đặc biệt phù hợp khi dữ liệu đầu vào bao gồm cả yếu tố không gian và thời gian, như trong bài toán dự đoán AQI theo từng khu vực cụ thể.



Mục tiêu của việc so sánh ba mô hình là nhằm xác định phương pháp tối ưu không chỉ về mặt độ chính xác dự đoán, mà còn về khả năng sử dụng hiệu quả tài nguyên tính toán – một yếu tố quan trọng khi mở rộng hệ thống ra các địa phương khác có điều kiện hạ tầng công nghệ hạn chế hơn so với TP.HCM. Bên cạnh đó, tất cả các mô hình đều được tinh chỉnh tham số một cách cẩn trọng nhằm đảm bảo chất lượng dự báo cao, đáp ứng yêu cầu triển khai trong môi trường thực tế và hỗ trợ hiệu quả cho công tác giám sát, cảnh báo ô nhiễm không khí theo thời gian thực.

Trong quá trình huấn luyện, hệ thống theo dõi và ghi nhận các chỉ số đánh giá độ chính xác của mô hình bao gồm:

- **RMSE** (Root Mean Square Error) – sai số bình phương trung bình,
- **MAE** (Mean Absolute Error) – sai số tuyệt đối trung bình,
- **Loss function** – hàm mất mát tương ứng với từng mô hình.

Bên cạnh các chỉ số đánh giá mô hình, hệ thống cũng giám sát các chỉ số liên quan đến **hiệu suất phần cứng**, bao gồm:

- Mức sử dụng RAM,
- CPU và GPU trong quá trình huấn luyện,
- Thời gian huấn luyện cho mỗi mô hình tại từng trạm.

Việc giám sát đồng thời cả độ chính xác của mô hình và mức độ sử dụng tài nguyên hệ thống đóng vai trò quan trọng trong việc đánh giá toàn diện hiệu quả triển khai trong điều kiện thực tế. Một mô hình dự đoán không chỉ cần đạt được kết quả chính xác, mà còn phải đảm bảo khả năng vận hành nhẹ, nhanh và ổn định nhằm đáp ứng yêu cầu cập nhật dữ liệu liên tục và đưa ra dự báo trong thời gian thực. Tính cân bằng giữa độ chính xác và hiệu suất xử lý là yếu tố then chốt để đảm bảo mô hình có thể ứng dụng bền vững trong môi trường có hạ tầng tính toán hạn chế, đồng thời duy trì tính khả thi khi mở rộng hệ thống trên quy mô lớn.

### 2.1.3 Đánh giá mô hình:

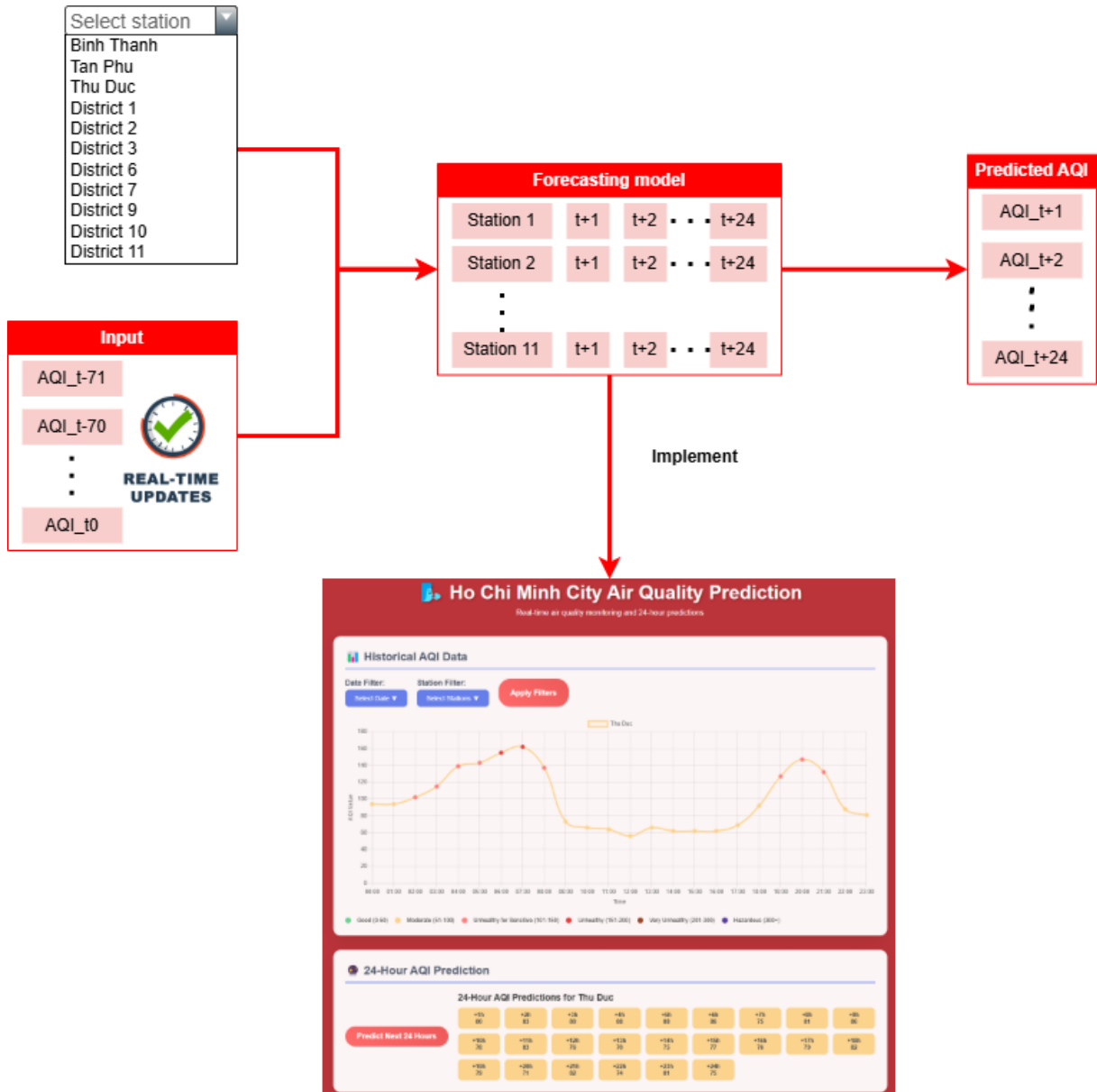
Mô hình được đánh giá dựa trên hai chỉ số chính là RMSE (Root Mean Square Error) và MAE (Mean Absolute Error), nhằm đo lường mức độ sai lệch giữa giá trị dự báo và giá trị thực tế. Trong đó, RMSE có khả năng phản ánh rõ hơn các sai số lớn do đặc điểm bình phương của công thức tính, còn MAE cung cấp một ước lượng trung bình tuyến tính về độ lệch, giúp mô tả sai số một cách trực quan và dễ hiểu. Việc sử dụng đồng thời cả hai chỉ số này giúp tạo ra một cái nhìn toàn diện về hiệu suất dự đoán của mô hình, từ đó đánh giá được mức độ tin cậy trong nhiều tình huống khác nhau.

Ngoài các chỉ số đo lường độ chính xác, quá trình huấn luyện còn được giám sát thông qua hàm mất mát (loss function), với mục tiêu đảm bảo mô hình học được các đặc trưng dữ liệu một cách hiệu quả mà không dẫn đến hiện tượng quá khớp (overfitting). Đồng thời, các thông số liên quan đến hiệu năng hệ thống như mức sử dụng bộ nhớ RAM, CPU và GPU cũng được theo dõi liên tục trong suốt quá trình huấn luyện và dự đoán. Thời gian xử lý toàn bộ quy trình từ dữ liệu đầu vào đến đầu ra dự báo cũng được ghi nhận nhằm đánh giá mức độ phù hợp khi triển khai trong môi trường yêu cầu cập nhật liên tục theo thời gian thực.

Sau khi hoàn tất quá trình đánh giá, kết quả mô hình được so sánh với các nghiên cứu trước đó trong cùng lĩnh vực, nhằm xác định vị thế tương đối của mô hình hiện tại về mặt hiệu quả và độ chính xác. Thông qua quá trình đối chiếu này, có thể nhận diện được những ưu điểm và hạn chế của phương pháp đang được áp dụng, từ đó rút ra các bài học kinh nghiệm và đề xuất các hướng cải tiến nhằm tối ưu hóa mô hình trong các giai đoạn triển khai tiếp theo.

#### 2.1.4 Triển khai mô hình:

Sau khi hoàn tất quá trình huấn luyện và đánh giá, mô hình học máy được lựa chọn – là mô hình có độ chính xác cao nhất trong số ba mô hình đã thử nghiệm gồm LSTM, LightGBM và XGBoost – sẽ được triển khai vào hệ thống web app nhằm phục vụ dự đoán chất lượng không khí trong môi trường thực tế. Mô hình được triển khai theo kiến trúc online, trong đó dữ liệu được cập nhật một lần mỗi giờ và các yêu cầu dự đoán được xử lý gần như theo thời gian thực. Cách thức mô hình được áp dụng được mô tả trong sơ đồ ở hình 15.



Hình 15: Quy trình triển khai hệ thống lên nền tảng web app.

Mục tiêu cụ thể của hệ thống là dự đoán chỉ số chất lượng không khí (AQI) trong vòng 72 giờ tiếp theo, dựa trên chuỗi AQI của 72 giờ gần nhất. Đây là một bài toán dự báo chuỗi thời gian phức tạp, đòi hỏi mô hình phải học được cả xu hướng dài hạn và biến động ngắn hạn trong dữ liệu môi trường đô thị, vốn thường xuyên thay đổi do ảnh hưởng của thời tiết, giao thông, và các nguồn ô nhiễm cục bộ.

Mô hình được triển khai theo kiến trúc phân tán theo trạm đo, nghĩa là mỗi trạm quan trắc chất lượng không khí trong hệ thống được gắn với một mô hình dự báo riêng biệt. Khi người dùng yêu cầu dự đoán tại một trạm cụ thể, hệ thống sẽ gọi mô hình tương ứng với trạm đó để sinh ra dự đoán AQI cho 24 giờ tới. Cách tiếp cận này giúp tăng độ chính xác cục bộ, cho phép mô hình khai thác các đặc trưng riêng của từng khu vực địa lý, thay vì sử dụng một mô hình tổng quát cho toàn thành phố.

Hệ thống web được xây dựng trên nền tảng Flask, cho phép giao tiếp giữa giao diện người dùng và mô hình một cách linh hoạt. Hiện tại, mô hình được triển khai trên môi trường máy cá nhân. Việc cập nhật mô hình mới vẫn được thực hiện thủ công sau mỗi lần huấn luyện lại, đảm bảo sự kiểm soát và quan sát chặt chẽ trong giai đoạn phát triển ban đầu. Mặc dù vậy, cấu trúc của hệ thống đã được thiết kế sẵn để hướng tới khả năng mở rộng trong tương lai, cho phép tích hợp các thành phần MLOps khi cần thiết.

#### *2.1.5 Quy trình cập nhật dữ liệu*

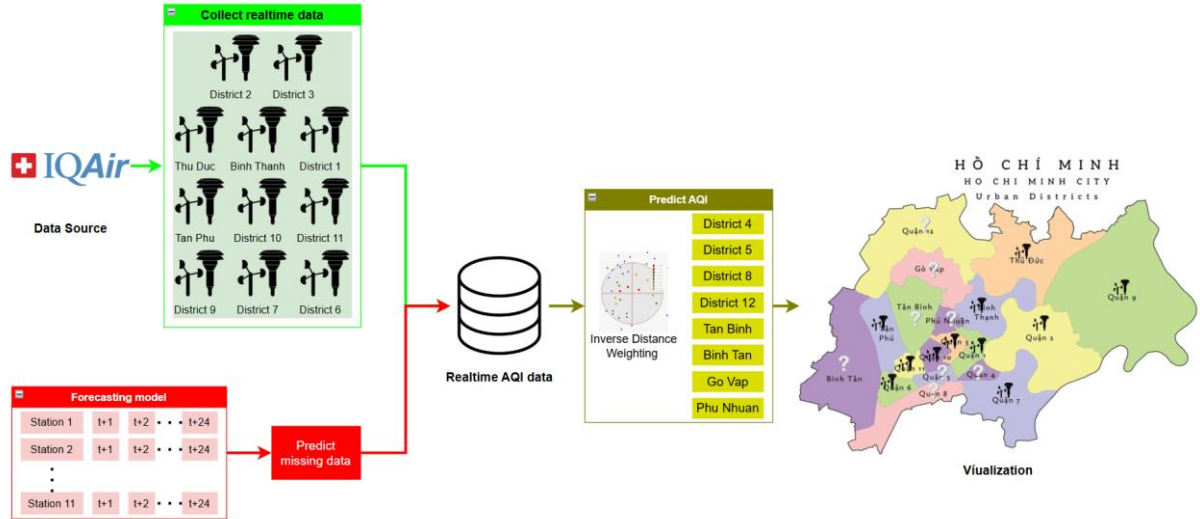
Sau khi được triển khai, mô hình vẫn tiếp tục duy trì khả năng thích ứng với những biến đổi của môi trường thông qua cơ chế cập nhật liên tục dựa trên dữ liệu mới. Cụ thể, dữ liệu chất lượng không khí tại từng trạm quan trắc được thu thập và lưu trữ định kỳ mỗi giờ. Định kỳ hai ngày một lần, hệ thống sẽ tiến hành huấn luyện lại mô hình trên tập dữ liệu mới nhằm bắt kịp các xu hướng và biến động gần nhất trong môi trường thực tế.

Quá trình cập nhật này được hỗ trợ thông qua pipeline lập trình sẵn và lưu trữ trên GitHub, giúp tự động hóa khâu thu thập, xử lý dữ liệu và kích hoạt mô hình học trực tuyến mà không cần can thiệp thủ công. Trong giai đoạn cập nhật, dữ liệu mới sẽ được bổ sung vào kho lưu trữ, đồng thời mô hình sẽ được huấn luyện lại với dữ liệu cập nhật này. Mặc dù hiện tại hệ thống ghi đè mô hình cũ bằng mô hình mới, các thông số vận hành như thời gian huấn luyện, mức sử dụng tài nguyên hệ thống (RAM, CPU, GPU), cũng như các thước đo đánh giá mô hình như RMSE và MAE đều được ghi nhận để phục vụ theo dõi hiệu năng và độ ổn định của hệ thống.

Bên cạnh đó, một bước kiểm tra thủ công được thực hiện sau mỗi lần cập nhật mô hình, trong đó kết quả dự đoán trước khi cập nhật sẽ được đối chiếu với dữ liệu thực tế mới thu thập, nhằm đánh giá khả năng khái quát hóa và mức độ sai lệch của mô hình. Dù quy trình này chưa được tự động hóa hoàn toàn, nó đóng vai trò quan trọng trong việc đảm bảo rằng mô hình duy trì độ chính xác cao theo thời gian, đặc biệt trong bối cảnh chất lượng không khí có thể biến động mạnh do các yếu tố thời tiết, giao thông hoặc ô nhiễm đột xuất.

## **2.2 Thiết kế mô hình dự đoán chất lượng không khí tại những nơi không có trạm.**

Nhằm cung cấp một cái nhìn toàn diện hơn về tình trạng chất lượng không khí trên toàn địa bàn TP.HCM, mô hình đã được thiết kế để dự đoán chỉ số AQI tại các khu vực chưa được lắp đặt trạm quan trắc. Phương pháp này cho phép mở rộng phạm vi bao phủ thông tin mà không cần triển khai thêm cảm biến vật lý. Dữ liệu đầu vào của mô hình bao gồm các chỉ số chất lượng không khí thu thập từ các trạm quan trắc hiện có tại thời điểm hiện tại; từ đó, mô hình thực hiện nội suy để ước lượng giá trị AQI tại các vị trí chưa có trạm. Quy trình hoạt động tổng thể của mô hình được minh họa trong hình 16.



Hình 16: Pipeline mô hình dự đoán chất lượng không khí tại những quận/huyện chưa có trạm ở TP.HCM.

**Thu thập dữ liệu thời gian thực:** Hệ thống được thiết lập để kết nối trực tiếp với nền tảng dữ liệu của IQAir, từ đó thu thập chỉ số chất lượng không khí (AQI) tại các trạm quan trắc đặt tại nhiều khu vực trọng điểm của TP.HCM, bao gồm các quận trung tâm và vùng ven như Quận 1, Quận 2, Quận 3, Quận 6, Quận 7, Quận 9, Quận 10, Quận 11, Tân Phú, Thủ Đức và Bình Thạnh. Dữ liệu được thu thập theo thời gian thực với tần suất cao, nhằm phản ánh chính xác tình trạng không khí tại từng khu vực và thời điểm cụ thể.

Tuy nhiên, do đặc thù biến động và gián đoạn thường gặp trong dữ liệu môi trường, một số trạm quan trắc có thể không cung cấp thông tin tại thời điểm truy xuất. Để xử lý các trường hợp thiếu dữ liệu, hệ thống triển khai mô hình học máy trực tuyến (online learning) nhằm ước lượng các giá trị bị thiếu. Mô hình này được huấn luyện liên tục trên chuỗi dữ liệu thời gian gần nhất, giúp duy trì khả năng thích ứng với điều kiện môi trường thay đổi và nâng cao độ chính xác trong dự đoán.

**Dự đoán số liệu tại những địa điểm không có trạm:** Sau khi dữ liệu từ các trạm quan trắc được chuẩn hóa và xử lý đầy đủ, hệ thống tiếp tục thực hiện bước nội suy nhằm ước lượng chỉ số AQI tại các khu vực chưa được trang bị trạm đo, bao gồm Quận 4, Quận 5, Quận 8, Quận 12, Tân Bình, Bình Tân, Phú Nhuận và Gò Vấp. Phương pháp được lựa chọn cho giai đoạn này là kỹ thuật nội suy khoảng cách nghịch đảo (Inverse Distance Weighting – IDW), một phương pháp thống kê không tham số phổ biến trong các bài toán không gian. Cơ chế của IDW dựa trên giả định rằng các điểm gần nhau trong không gian có xu hướng mang giá trị tương tự, theo đó giá trị tại vị trí cần ước lượng được tính toán từ các điểm lân cận có dữ liệu, với trọng số tỷ lệ nghịch với bình phương khoảng cách đến điểm đang xét.

Việc lựa chọn IDW thay vì các kỹ thuật nội suy khác như Kriging, spline hay các mô hình học máy được cân nhắc dựa trên các đặc điểm thực tế của bài toán. IDW có ưu điểm vượt trội về mặt đơn giản hóa tính toán, không yêu cầu giả định về phân phối xác suất của dữ liệu, đồng thời không cần lượng dữ liệu lịch sử lớn để huấn luyện mô hình — điều này đặc biệt phù hợp với điều kiện dữ liệu không gian–thời gian còn rời rạc tại TP.HCM. Ngoài ra, phương pháp này tận dụng hiệu quả tính cục bộ trong phân bố dữ liệu môi trường, phù hợp với quan sát thực tế rằng chất lượng không khí tại các khu vực liền kề thường có xu hướng tương đồng. Với định hướng xây dựng một hệ thống dự báo nhẹ, linh hoạt và dễ triển khai trong môi trường thời gian thực, IDW là một lựa chọn hợp lý và mang lại hiệu quả cao về mặt chi phí tính toán lẫn độ tin cậy kết quả.

Toàn bộ kết quả dự đoán — bao gồm dữ liệu đo thực tế từ các trạm, dữ liệu được mô hình học máy ước lượng cho các trạm bị thiếu thông tin, và dữ liệu nội suy từ IDW tại các khu vực không có trạm — được tổng hợp và trực quan hóa dưới dạng bản đồ số. Bản đồ này được tích hợp vào ứng dụng web, cho phép người dùng tương tác và theo dõi chất lượng không khí một cách trực quan và tức thời trên từng khu vực cụ thể của TP.HCM.

## 2.3 Xây dựng ứng dụng web

Nền tảng ứng dụng web được phát triển không chỉ nhằm kết nối mô hình dự đoán với người dùng cuối mà còn đóng vai trò trung tâm trong việc trực quan hóa dữ liệu và đảm bảo quá trình cập nhật thông tin diễn ra liên tục theo thời gian thực. Việc triển khai hệ thống trên môi trường web góp phần mở rộng phạm vi tiếp cận, cho phép cả cộng đồng dân cư lẫn cơ quan quản lý dễ dàng truy cập, giám sát tình hình chất lượng không khí và đưa ra các quyết định ứng phó kịp thời nhằm bảo vệ sức khỏe cộng đồng và nâng cao hiệu quả quản lý môi trường đô thị.

Các chức năng chính được thực hiện trong dự án bao gồm:

- Hiển thị dữ liệu lịch sử chất lượng không khí.
- Dự đoán chất lượng không khí trong 24h.
- Hiển thị chất lượng không khí tại các trạm trên bản đồ Thành phố Hồ Chí Minh.

### 2.3.1 *Hiển thị dữ liệu lịch sử chất lượng không khí.*

Hệ thống web được thiết kế để cung cấp tính năng truy xuất và hiển thị dữ liệu lịch sử về chất lượng không khí, nhằm hỗ trợ người dùng theo dõi xu hướng biến động của các chỉ số ô nhiễm theo thời gian. Dữ liệu chỉ số AQI được ghi nhận định kỳ theo từng giờ tại các trạm quan trắc và được lưu trữ để phục vụ mục đích tra cứu và phân tích. Người dùng có thể lựa chọn khoảng thời gian tùy chỉnh để quan sát diễn biến chất lượng không khí theo từng ngày, cũng như so sánh sự thay đổi giữa các mốc thời gian khác nhau.

Giao diện trực quan của tính năng này được minh họa tại hình 17. Để hiển thị biểu đồ dữ liệu lịch sử, người dùng cần lựa chọn ngày và trạm quan trắc tương ứng, sau đó nhấn nút "Apply filter" để hệ thống tiến hành xử lý và hiển thị kết quả. Biểu đồ được dựng bằng thư viện Chart.js — một thư viện JavaScript phổ biến và hiệu năng cao chuyên dùng cho việc trực quan hóa dữ liệu thời gian thực, giúp biểu diễn các xu hướng một cách rõ ràng và dễ hiểu.





Hình 17: Giao diện biểu đồ chỉ số AQI lọc theo ngày và theo trạm.

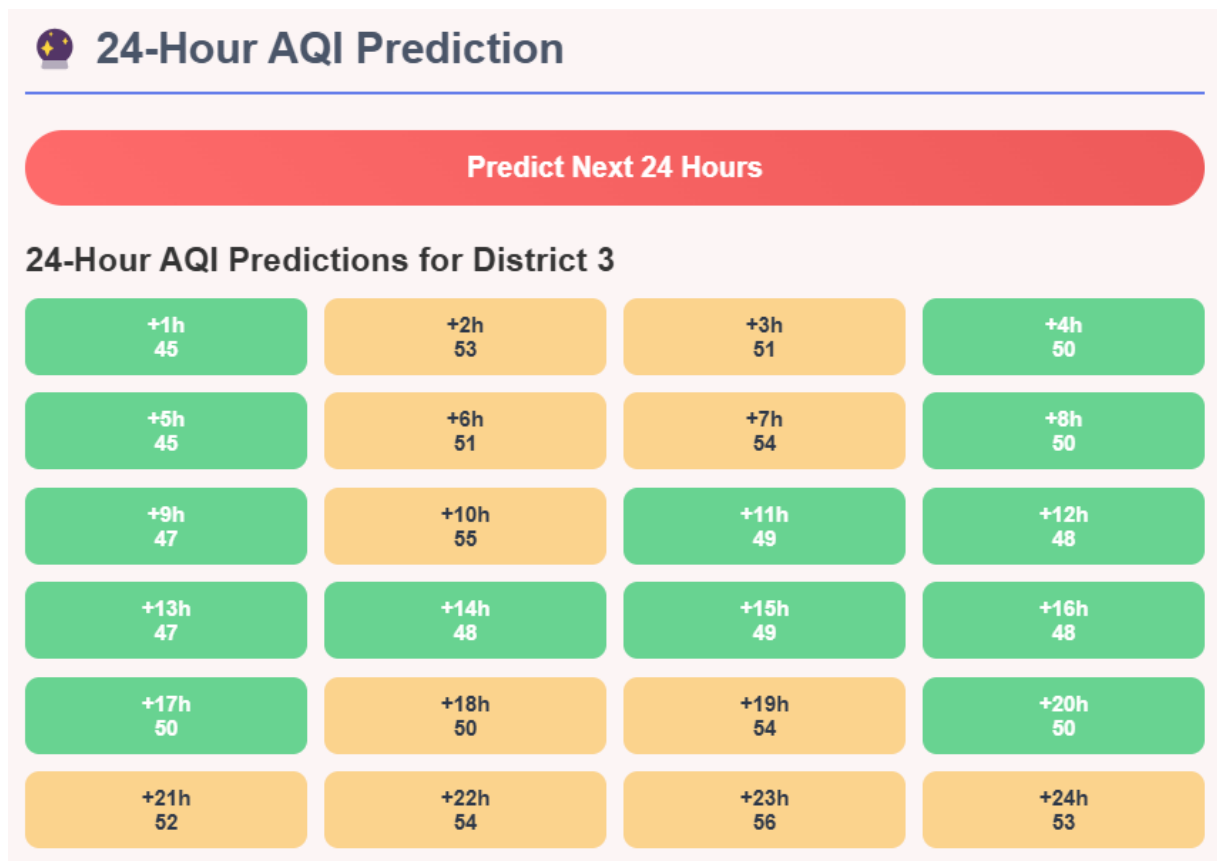
Chỉ số AQI được trình bày dưới dạng biểu đồ đường nhằm trực quan hóa xu hướng biến động của chất lượng không khí theo thời gian trong ngày. Dữ liệu đầu vào cho biểu đồ được trích xuất từ tập tin .csv đã được thu thập và lưu trữ trước đó. Hệ thống web sử dụng các bộ lọc theo ngày và theo trạm quan trắc do người dùng chọn để truy xuất các giá trị phù hợp từ tệp dữ liệu. Biểu đồ hiển thị chỉ số AQI theo từng khung giờ từ 00:00 đến 23:00 trong ngày đã chọn, qua đó giúp người dùng dễ dàng quan sát các thay đổi theo chuỗi thời gian ngắn hạn.

Mỗi điểm dữ liệu trên biểu đồ được mã hóa màu sắc tương ứng với các ngưỡng cảnh báo AQI theo tiêu chuẩn quốc tế, giúp người dùng dễ dàng nhận diện mức độ nguy hại của không khí tại từng thời điểm. Bên cạnh đó, biểu đồ hỗ trợ tính năng tương tác nâng cao: khi người dùng di chuyển con trỏ chuột đến một điểm cụ thể trên biểu đồ, hệ thống sẽ hiển thị thông tin chi tiết bao gồm tên trạm quan trắc, thời gian đo và giá trị AQI tương ứng. Trong trường hợp dữ liệu đầu vào bị thiếu hoặc không hợp lệ, hệ thống sẽ thông báo lỗi để đảm bảo tính minh bạch và độ tin cậy của thông tin hiển thị.

### 2.3.2 Dự đoán chất lượng không khí trong 24 giờ.

Tính năng dự đoán chất lượng không khí trong vòng 24 giờ tại Thành phố Hồ Chí Minh được phát triển với mục tiêu cung cấp thông tin dự báo ngắn hạn một cách chính xác và kịp thời về chỉ số AQI. Tính năng này đóng vai trò quan trọng trong việc hỗ trợ người dân chủ động phòng ngừa các rủi ro sức khỏe liên quan đến ô nhiễm không khí, đặc biệt trong bối cảnh đô thị hóa diễn ra mạnh mẽ và mức độ ô nhiễm ngày càng có xu hướng gia tăng.

Người dùng có thể dễ dàng sử dụng tính năng này bằng cách nhấn nút "Dự đoán" sau khi lựa chọn trạm quan trắc từ bộ lọc có sẵn trong phần biểu đồ. Hệ thống sau đó sẽ xử lý và trả về kết quả dự đoán chỉ số AQI trong 24 giờ tiếp theo đối với trạm được chọn. Minh họa về kết quả dự báo tại trạm Quận 3 được thể hiện trong Hình 18.



Hình 18: Giao diện dự đoán chất lượng không khí trong 24h tiếp theo tại trạm Quận 3.

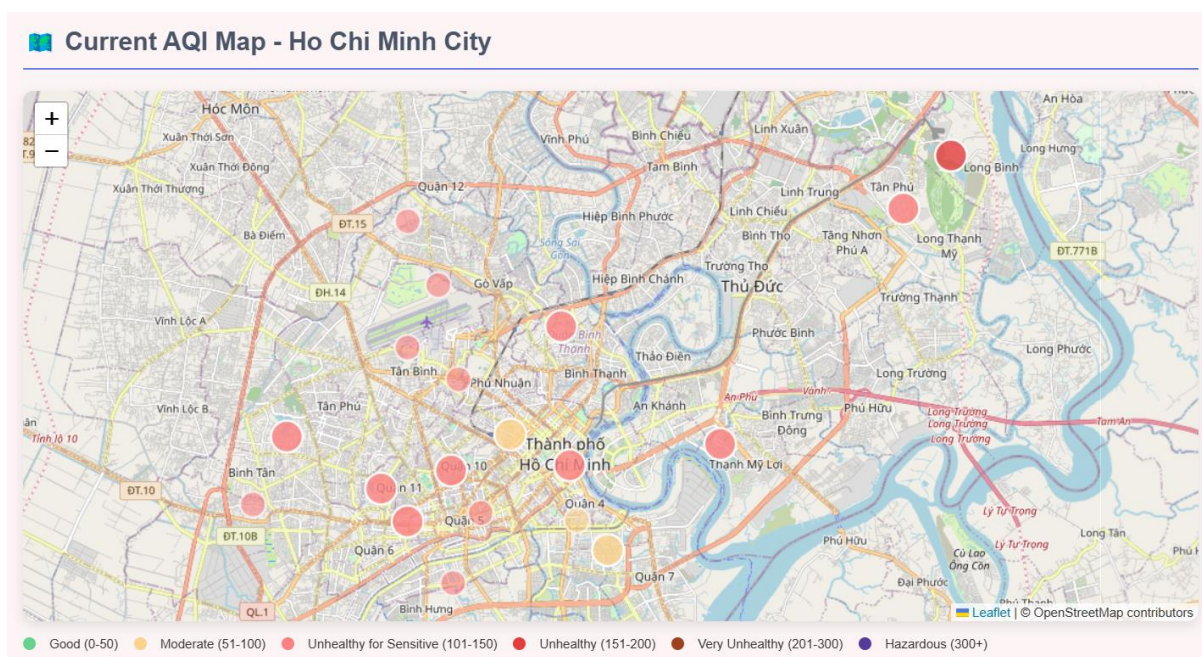
Hệ thống được tích hợp mô hình dự báo chất lượng không khí với đầu vào là chuỗi thời gian gồm 72 giờ gần nhất của chỉ số AQI tại từng trạm, và đầu ra là dự báo chi tiết cho 24 giờ kế tiếp. Mỗi giờ dự báo được trình bày cùng với màu nền tương ứng phản ánh mức độ nguy hại theo chuẩn phân loại AQI, giúp người dùng dễ dàng nắm bắt mức độ ô nhiễm theo thời gian trong ngày.

Khác với các phương pháp hiển thị dữ liệu trung bình theo giờ hoặc phân đoạn theo khung 12–24 giờ, phương pháp dự báo rời rạc theo từng giờ cho phép nắm bắt rõ ràng các biến động ngắn hạn của môi trường không khí đô thị. Việc tính trung bình có thể che lấp các đỉnh ô nhiễm ngắn nhưng nghiêm trọng – vốn có thể ảnh hưởng đáng kể đến sức khỏe cộng đồng. Trong khi đó, dự báo chi tiết từng giờ cung cấp khả năng phát hiện sớm các khung thời gian có nguy cơ cao, từ đó hỗ trợ người dân, nhà quản lý và cơ quan chức năng đưa ra các biện pháp ứng phó kịp thời và hiệu quả, góp phần bảo vệ sức khỏe cộng đồng và nâng cao hiệu quả điều hành trong lĩnh vực môi trường.

### 2.3.3 *Hiển thị chất lượng không khí tại các trạm trên bản đồ*

Tính năng “Hiển thị chất lượng không khí tại các trạm trên bản đồ” được xây dựng nhằm mang đến cho người dùng một cái nhìn trực quan và toàn diện về mức độ ô nhiễm không khí tại từng khu vực trong địa bàn TP.HCM. Khác với hình thức trình bày dữ liệu AQI bằng bảng số liệu hay biểu đồ, việc trực quan hóa trên bản đồ giúp người dùng dễ dàng nhận diện các khu vực có chỉ số ô nhiễm cao, từ đó hỗ trợ họ đưa ra các quyết định phù hợp trong việc đi lại, sinh hoạt và bảo vệ sức khỏe cá nhân.

Giao diện của tính năng này được thể hiện như hình 19. Ngay khi trang web được tải, bản đồ khu vực TP.HCM sẽ được hiển thị bằng thư viện Leaflet, sử dụng dữ liệu không gian định dạng GeoJSON kết hợp với lớp nền bản đồ từ OpenStreetMap. Các trạm đo chất lượng không khí sẽ được đánh dấu bằng các vòng tròn màu sắc tại đúng vị trí địa lý trên bản đồ. Màu của mỗi vòng tròn thể hiện mức AQI hiện tại tại từng trạm, giúp người dùng nhanh chóng đánh giá tình trạng không khí tại các khu vực khác nhau. Ngoài ra, người dùng có thể sử dụng các nút điều khiển (+/-) ở góc trên bên trái để phóng to hoặc thu nhỏ bản đồ, hỗ trợ quan sát chi tiết hơn theo từng khu vực cụ thể.



Hình 19: Giao diện biểu đồ chất lượng không khí tại TP.HCM.

Khi hệ thống được khởi chạy, dữ liệu chất lượng không khí theo thời gian thực sẽ được tự động thu thập từ nền tảng IQAir. Tuy nhiên, do mạng lưới các trạm quan trắc hiện tại vẫn còn phân bố không đồng đều và chưa bao phủ toàn bộ thành phố, nhiều khu vực trên bản đồ không có dữ liệu đo trực tiếp để hiển thị. Để khắc phục điểm hạn chế này và đảm bảo tính liên tục về mặt không gian trong việc hiển thị chất lượng không khí, hệ thống đã tích hợp thuật toán Inverse Distance Weighting (IDW) — một phương pháp nội suy không gian phổ biến. Thuật toán này cho phép ước lượng chỉ số AQI tại các vị trí chưa có trạm bằng cách tính trung bình trọng số từ các trạm lân cận, trong đó trọng số giảm dần theo khoảng cách.

Cần lưu ý rằng chất lượng không khí tại TP.HCM không phân bố đồng nhất — từng khu vực có thể bị ảnh hưởng bởi các yếu tố đặc thù như mật độ giao thông, mức độ đô thị hóa hay hoạt động công nghiệp. Vì vậy, việc chỉ dựa trên dữ liệu từ các trạm hiện có là chưa đủ để phản ánh toàn diện tình hình. Để khắc phục điều này, chức năng **dự đoán AQI cho các khu vực chưa có trạm quan trắc** được phát triển nhằm cung cấp một cái nhìn đầy đủ hơn cho người dân khi đưa ra các lựa chọn về sinh hoạt, di chuyển, cũng như hỗ trợ cơ quan quản lý xây dựng các chính sách môi trường phù hợp với đặc thù từng khu vực, thay vì áp dụng giải pháp chung cho toàn thành phố.

## KẾT LUẬN CHƯƠNG

Trong chương này, quy trình thiết kế hệ thống dự đoán chất lượng không khí theo thời gian thực tại TP. Hồ Chí Minh đã được trình bày chi tiết với trọng tâm là việc áp dụng các mô hình học máy trực tuyến. Hệ thống được xây dựng gồm ba thành phần chính: thu thập và tiền xử lý dữ liệu, huấn luyện mô hình học trực tuyến cho từng trạm quan trắc, và đánh giá mô hình ở cả hai giai đoạn: sau huấn luyện và trong quá trình dự đoán thực tế.

Thông qua việc sử dụng dữ liệu AQI từ 11 trạm quan trắc khác nhau trên địa bàn thành phố, kết hợp với ba mô hình LSTM, LightGBM và XGBoost, hệ thống được thiết kế nhằm đạt được sự cân bằng giữa độ chính xác dự đoán và khả năng triển khai thực tế. Việc theo dõi đồng thời cả hiệu năng mô hình lẫn tài nguyên tính toán giúp đảm bảo rằng hệ thống không chỉ chính xác mà còn hiệu quả và ổn định trong môi trường hoạt động liên tục.

Chương tiếp theo sẽ trình bày quá trình **thực nghiệm và đánh giá kết quả** từ hệ thống này, qua đó kiểm định khả năng áp dụng của mô hình trong bối cảnh thực tiễn cũng như làm rõ tiềm năng cải tiến trong tương lai.

## Chương 3 THỰC NGHIỆM VÀ ĐÁNH GIÁ

### 3.1 Môi trường thực nghiệm

#### 3.1.1 Môi trường thực nghiệm cho mô hình học máy trực tuyến

Trong quá trình thực nghiệm, mô hình LSTM được triển khai và huấn luyện trên máy tính cá nhân, trong khi hai mô hình XGBoost và LightGBM được đào tạo trong môi trường Google Colab. Bảng 1 trình bày chi tiết cấu hình phần cứng của cả máy tính cá nhân và nền tảng Google Colab được sử dụng trong quá trình nghiên cứu.:

*Bảng 1: Cấu hình phần cứng của máy tính cá nhân và google colab:*

|            | <b>Máy tính cá nhân</b>   | <b>Google Colab</b> |
|------------|---------------------------|---------------------|
| <b>CPU</b> | Intel Core i5 8365U       | Intel Xeon          |
| <b>GPU</b> | Intel(R) UHD Graphics 620 | NVIDIA Tesla T4     |
| <b>RAM</b> | 16GB                      | 13GB                |

Việc phân bổ môi trường huấn luyện cho các mô hình được lựa chọn dựa trên đặc điểm kỹ thuật và yêu cầu tính toán riêng biệt của từng thuật toán. Cụ thể, mô hình LSTM với cấu trúc mạng vừa phải có thể hoạt động ổn định trên CPU mà không cần đến tài nguyên phần cứng cao cấp. Trong khi đó, các mô hình XGBoost và LightGBM lại phát huy hiệu quả vượt trội khi được huấn luyện trên GPU nhờ khả năng khai thác tính toán song song, từ đó rút ngắn đáng kể thời gian xử lý đối với tập dữ liệu lớn. Chiến lược lựa chọn môi trường như vậy giúp tối ưu hóa hiệu suất thực nghiệm đồng thời tiết kiệm chi phí tài nguyên.

#### 3.1.2 Môi trường thực nghiệm cho việc xây dựng ứng dụng:

Về mặt phần mềm, Python 3.9 cùng với framework PyTorch đã được sử dụng để xây dựng và huấn luyện mô hình học sâu. Để triển khai kết quả dự đoán lên nền tảng web, Flask được sử dụng cho phần backend, trong khi phần frontend được đảm nhận bởi HTML và JavaScript.

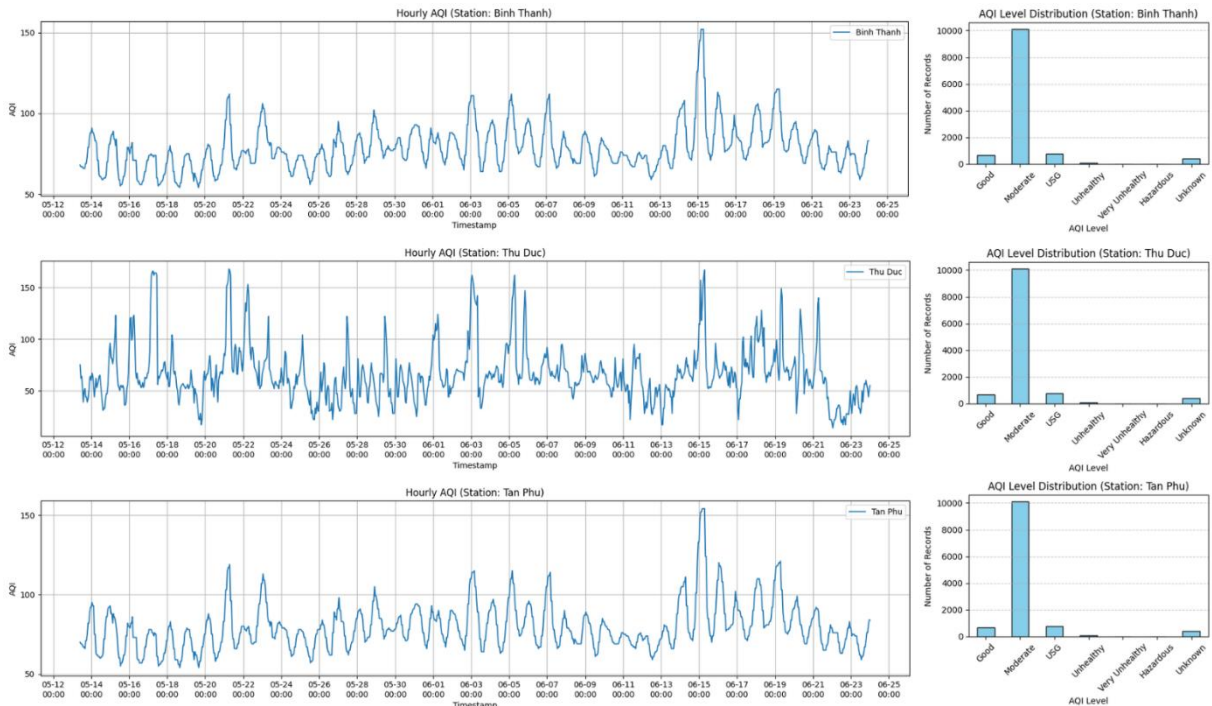
### 3.2 Tập dữ liệu

Dữ liệu chất lượng không khí đã được thu thập theo thời gian thực với tần suất mỗi giờ từ nền tảng IQAir. Trong khuôn khổ khóa luận tốt nghiệp này, dữ liệu phục vụ cho quá trình thực nghiệm được giới hạn trong khoảng thời gian từ ngày 13 tháng 5 năm 2025 đến ngày 24 tháng 6 năm 2025, tập trung tại khu vực Thành phố Hồ Chí Minh — nơi đang phải đối mặt với tình trạng ô nhiễm không khí nghiêm trọng do tốc độ đô thị hóa cao và mật độ giao thông dày đặc.

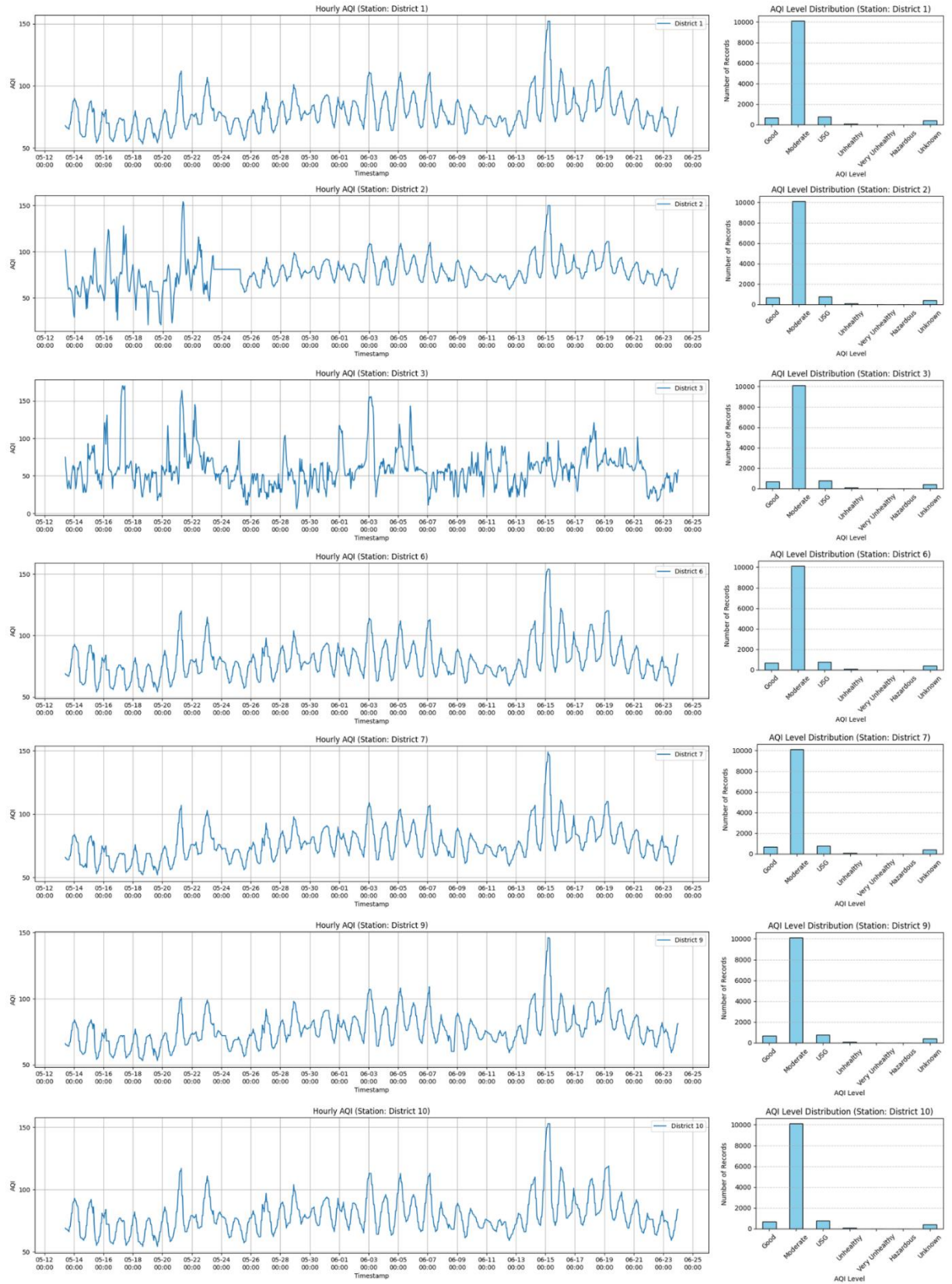
Tập dữ liệu thu thập được bao gồm 11.088 bản ghi, tương ứng với số lần quan trắc theo giờ tại các trạm đo trong toàn thành phố. Mỗi bản ghi chứa các thông tin sau:

- **Thời điểm:** Thời gian cụ thể khi dữ liệu được ghi nhận.
- **Tên trạm:** Tên trạm quan trắc ghi nhận dữ liệu.
- **Chỉ số AQI:** Air Quality Index (chỉ số chất lượng không khí), phản ánh mức độ ô nhiễm tại thời điểm quan trắc, được đo theo tiêu chuẩn US-AQI.

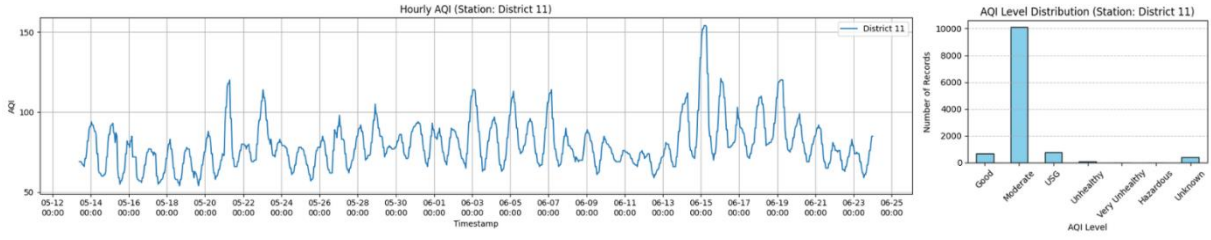
Tỷ lệ dữ liệu bị thiếu chiếm khoảng 2%, tuy nhiên điều này không gây trở ngại đáng kể cho quá trình tiền xử lý. Hình 20 minh họa sự phân bố dữ liệu sau khi đã được xử lý tại các trạm quan trắc trên địa bàn Thành phố Hồ Chí Minh.











Hình 20: Phân bố dữ liệu trong bộ dataset. Biểu đồ đường thể hiện dữ liệu qua các mốc thời gian, biểu đồ cột thể hiện phân bố dữ liệu theo mức độ chất lượng không khí.

Quan sát các biểu đồ thời gian, có thể nhận thấy rằng AQI tại đa số trạm dao động chủ yếu trong khoảng từ 50 đến 120, trong đó đa số dữ liệu nằm ở mức độ “Moderate”. Một số trạm như Quận 2, Quận 3 và Thủ Đức có xu hướng biến động mạnh hơn với các đỉnh AQI cao đột ngột, phản ánh khả năng xảy ra các đợt ô nhiễm cục bộ hoặc ảnh hưởng từ hoạt động giao thông, công nghiệp. Trong khi đó, các trạm như Quận 6, Quận 11 và Tân Phú thể hiện mức độ biến động thấp hơn, với các đường biểu diễn AQI có dạng chu kỳ đều đặn, nhiều khả năng phản ánh ảnh hưởng theo thời gian trong ngày. Các mô hình dao động tuần hoàn tại nhiều trạm gợi ý về tác động của yếu tố khí tượng và hoạt động sinh hoạt hàng ngày.

Phân tích biểu đồ phân bố mức độ AQI cho thấy phần lớn giá trị AQI nằm trong ngưỡng “Moderate” (Vừa phải), chiếm ưu thế tại tất cả các trạm. Một lượng nhỏ các quan trắc rơi vào nhóm “Unhealthy for Sensitive Groups” (Không lành mạnh cho nhóm nhạy cảm), trong khi các mức cao hơn như “Unhealthy”, “Very Unhealthy” hoặc thấp hơn như “Good” (Tốt) xuất hiện với tần suất rất thấp hoặc gần như không đáng kể. Điều này cho thấy môi trường không khí tại TP.HCM trong giai đoạn khảo sát nhìn chung không đạt ngưỡng “tốt”, nhưng vẫn chưa đến mức báo động nghiêm trọng trên diện rộng.

So sánh giữa các khu vực, có thể thấy rằng các quận trung tâm như Quận 1, Quận 3 và Bình Thạnh thường có biên độ dao động AQI lớn hơn so với các quận ngoại vi như Quận 6, Quận 11 hoặc Tân Phú. Điều này cho thấy ảnh hưởng của các yếu tố như mật độ dân cư, hoạt động kinh tế - giao thông có thể là nguyên nhân dẫn đến sự khác biệt trong mức độ ô nhiễm không khí giữa các khu vực.

### 3.3 Thang đo và các chỉ số đánh giá

Trong quá trình đánh giá hiệu suất mô hình dự đoán chất lượng không khí, RMSE (Root Mean Squared Error) và MAE (Mean Absolute Error) được lựa chọn làm hai thang đo chính nhờ khả năng phản ánh trực tiếp sai số giữa giá trị dự đoán và giá trị thực tế. MAE cung cấp cái nhìn tổng quan về mức độ sai lệch trung bình, dễ diễn giải và không bị ảnh hưởng bởi các giá trị ngoại lệ. Trong khi đó, RMSE có ưu thế trong việc nhấn mạnh các sai số lớn, từ đó hỗ trợ nhận diện và cải thiện các điểm dự đoán sai lệch nghiêm trọng. Việc kết hợp cả hai thang đo này giúp đánh giá mô hình một cách toàn diện hơn so với các chỉ số thống kê khác như  $R^2$  hay MAPE (Mean Absolute Percentage Error), vốn có thể kém chính xác trong bối cảnh dữ liệu AQI dao động không đều và có sự xuất hiện của cực trị.

Để đánh giá độ hiệu quả của phương pháp Inverse Distance Weighting (IDW) trong việc dự đoán chất lượng không khí tại các khu vực không có trạm quan trắc, hai chỉ số được sử dụng để đánh giá là precision và recall. Việc lựa chọn hai chỉ số này xuất phát từ mục tiêu cốt lõi của hệ thống: không chỉ dự đoán giá trị gần đúng mà còn phân loại chính xác các mức độ ô nhiễm không khí theo các ngưỡng cảnh báo (như "tốt", "trung bình", "kém", "nguy hại"), nhằm phục vụ cho việc ra quyết định và cảnh báo cộng đồng. Cụ thể, precision phản ánh mức độ chính xác trong các dự đoán nguy cơ ô nhiễm – tức là trong số các điểm mà hệ thống dự đoán là ô nhiễm, có bao nhiêu điểm thực sự bị ô nhiễm, còn recall đo lường khả năng phát hiện đầy đủ các điểm thực sự bị ô nhiễm.

#### 3.3.1 Thang đo Root Mean Square Error (RMSE)

**Root Mean Squared Error (RMSE)** là một thang đo phổ biến được sử dụng để đánh giá hiệu suất của các mô hình dự đoán, đặc biệt là trong các bài toán hồi quy. RMSE đo lường độ lệch trung bình của các giá trị dự đoán so với giá trị thực tế, với trọng số cao hơn dành cho những sai số lớn.

Công thức tính RMSE được biểu diễn như sau:

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

Trong đó:

- $n$  là số lượng mẫu trong tập dữ liệu đánh giá.
- $\hat{y}_i$  là giá trị dự đoán của mô hình tại thời điểm thứ  $i$ .
- $y_i$  là giá trị thực tế tại thời điểm thứ  $i$ .
- $(\hat{y}_i - y_i)^2$  là bình phương sai số tại mỗi điểm dữ liệu.

Trong bài toán dự đoán chất lượng không khí 24 giờ, RMSE được tính riêng biệt cho từng giờ dự báo. Thay vì gộp trung bình tất cả sai số, hệ thống giữ nguyên 24 giá trị RMSE ứng với 24 giờ. Cách tiếp cận này cho phép đánh giá chi tiết hiệu suất mô hình theo từng thời điểm trong ngày, từ đó phát hiện các khung giờ mô hình hoạt động kém hiệu quả hoặc dễ bị nhiễu. Đây là cách đo trực quan và phù hợp trong các ứng dụng yêu cầu phân tích sai số theo thời gian.

RMSE là một chỉ số đánh giá hiệu suất dự báo hữu ích khi cần nhấn mạnh các sai số lớn, từ đó giúp mô hình tập trung cải thiện các điểm dự đoán sai lệch nghiêm trọng. Với đơn vị giống giá trị gốc (ví dụ: AQI), RMSE dễ diễn giải và phù hợp trong các bài toán dự báo biến động chất lượng không khí theo giờ. Tuy nhiên, RMSE lại nhạy với ngoại lệ: chỉ một vài điểm sai số cao có thể khiến chỉ số này tăng mạnh, làm lu mờ hiệu suất tổng thể. Vì vậy, nên kết hợp RMSE với các chỉ số khác như MAE để đánh giá toàn diện hơn.

### 3.3.2 Thang đo Mean Absolute Error (MAE)

**Mean Absolute Error (MAE)** là một thang đo phổ biến dùng để đánh giá độ chính xác của mô hình dự báo. MAE đo lường sai số trung bình tuyệt đối giữa giá trị dự đoán và giá trị thực tế. Khác với RMSE, MAE xử lý sai số theo tuyến tính, không bình phương, nên phản ánh sai số một cách ổn định hơn, đặc biệt khi dữ liệu chứa nhiễu hoặc ngoại lệ. Đây là chỉ số dễ hiểu, trực quan, phù hợp cho cả người dùng kỹ thuật và phi kỹ thuật. MAE thể hiện mức độ sai lệch trung bình mà không làm sai số lớn bị khuếch đại như RMSE.

MAE được tính theo công thức:

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i| \quad (3)$$

Trong đó:

- $n$  là số lượng mẫu trong tập dữ liệu đánh giá.
- $\hat{y}_i$  là giá trị dự đoán của mô hình tại thời điểm thứ  $i$ .
- $y_i$  là giá trị thực tế tại thời điểm thứ  $i$ .
- $(\hat{y}_i - y_i)^2$  là bình phương sai số tại mỗi điểm dữ liệu.

Tương tự RMSE, MAE trong bài toán dự đoán AQI được tính riêng cho từng giờ trong 24 giờ sắp tới. Hệ thống không lấy trung bình mà giữ nguyên 24 chỉ số MAE riêng biệt, giúp đánh giá chính xác sai số tại từng thời điểm. Việc giữ từng thang đo riêng biệt giúp mô hình dễ được hiệu chỉnh ở những giờ dự báo kém hiệu quả hoặc thường xuyên gặp sai lệch do nhiễu thời gian thực.

MAE có ưu điểm là dễ diễn giải, ổn định và không bị ảnh hưởng bởi sai số lớn một cách cực đoan như RMSE. Chỉ số này phản ánh chính xác sai số trung bình chung, phù hợp trong các bối cảnh dữ liệu chứa nhiễu hoặc giá trị ngoại lệ. Tuy nhiên, điểm yếu của MAE là nó không nhấn mạnh các sai số lớn, do đó có thể bỏ sót các điểm dự báo nguy hiểm mà mô hình cần cải thiện đặc biệt trong các ứng dụng như cảnh báo sớm chất lượng không khí.

### 3.3.3 Chỉ số đánh giá precision

Precision là một chỉ số đánh giá quan trọng trong các bài toán phân loại, phản ánh mức độ chính xác của mô hình khi dự đoán một lớp cụ thể – trong trường hợp này là các mức độ ô nhiễm không khí. Về mặt định nghĩa, precision được tính bằng tỉ lệ giữa số lượng dự đoán đúng thuộc lớp dương (True Positives) và tổng số dự đoán là dương (gồm cả đúng và sai – True Positives + False Positives).

Toán học hóa, precision được biểu diễn dưới dạng:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (4)$$

### 3.3.4 Chỉ số đánh giá recall

Recall là một chỉ số đánh giá hiệu suất của mô hình phân loại, đặc biệt quan trọng trong các tình huống mà việc phát hiện đầy đủ các trường hợp thuộc lớp dương là ưu tiên hàng đầu. Trong ngữ cảnh này, recall đo lường khả năng của mô hình trong việc nhận diện đúng các khu vực thực sự đang có chất lượng không khí kém hoặc nguy hại. Công thức tính recall là tỉ lệ giữa số lượng dự đoán đúng thuộc lớp dương (True Positives) và tổng số trường hợp thực tế thuộc lớp dương (True Positives + False Negatives):

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (5)$$

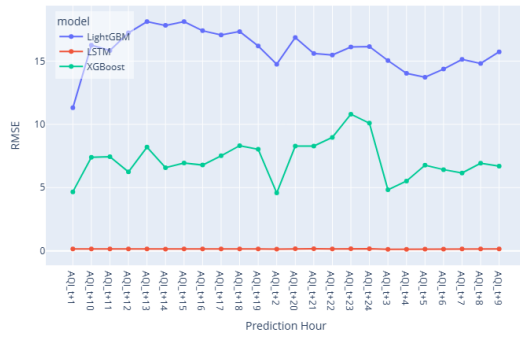
## 3.4 Kết quả thực nghiệm

### 3.4.1 Kết quả huấn luyện mô hình học trực tuyến

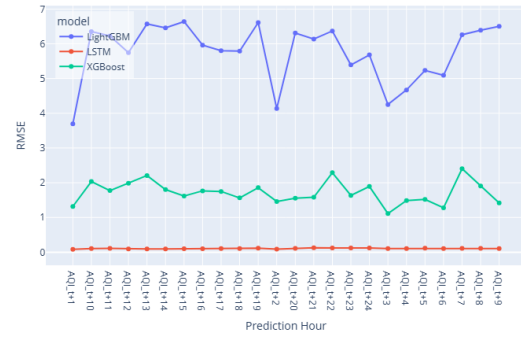
Trong quá trình thực nghiệm, ba mô hình học máy gồm LSTM, LightGBM và XGBoost đã được huấn luyện riêng biệt cho từng trạm quan trắc. Để tăng độ tin cậy cho kết quả, mỗi mô hình tại mỗi trạm được huấn luyện lặp lại ba lần. Trong suốt các giai đoạn huấn luyện, xác thực và kiểm thử, các chỉ số đánh giá hiệu suất bao gồm MAE (Mean Absolute Error) và RMSE (Root Mean Square Error) đã được ghi nhận. Nhằm đánh giá mức độ ổn định của từng mô hình tại mỗi trạm, giá trị trung bình và độ lệch chuẩn của hai thang đo trên được tính toán từ ba lần huấn luyện lặp lại. Kết quả sau cùng được trực quan hóa thông qua biểu đồ: Hình 21 minh họa giá trị RMSE theo từng trạm quan trắc, trong khi Hình 22 trình bày các giá trị MAE tương ứng.



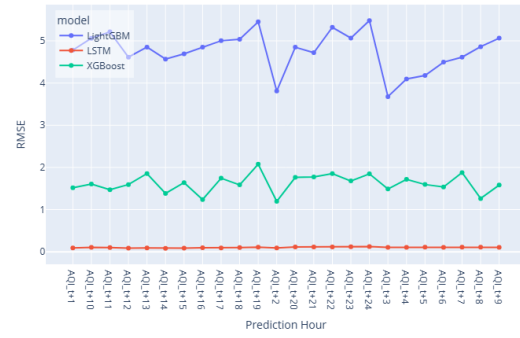
RMSE (Station: District 3)



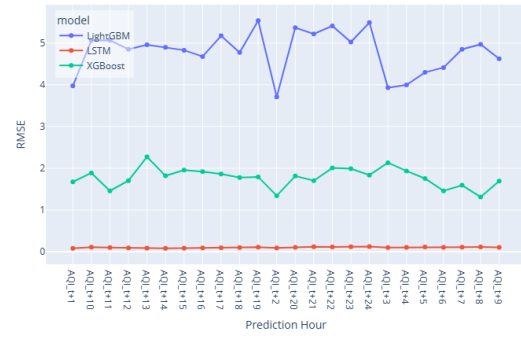
RMSE (Station: District 6)



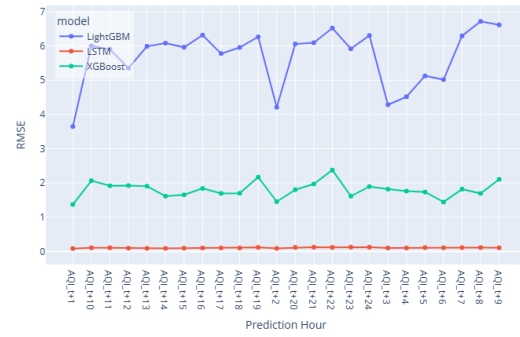
RMSE (Station: District 7)



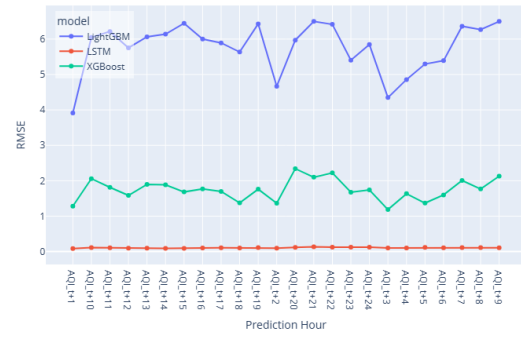
RMSE (Station: District 9)



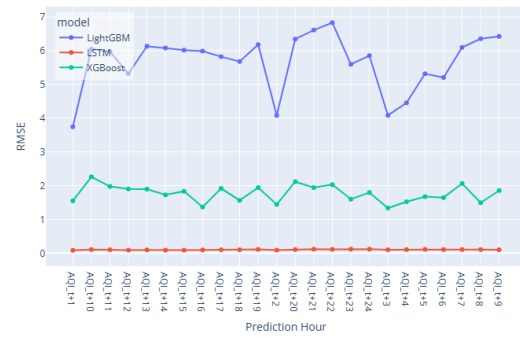
RMSE (Station: District 10)



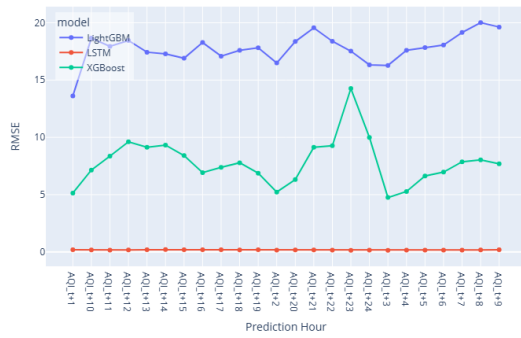
RMSE (Station: District 11)

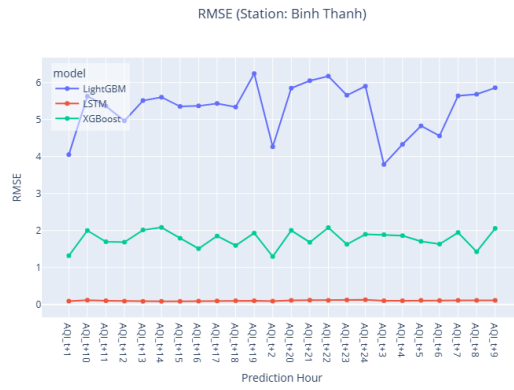


RMSE (Station: Tan Phu)



RMSE (Station: Thu Duc)





Hình 21: Giá trị RMSE của mô hình dự đoán chất lượng không khí theo từng trạm.





Hình 22: Giá trị MAE của mô hình dự đoán chất lượng không khí theo từng trạm.

Biểu đồ trực quan hóa kết quả cho thấy sự khác biệt rõ ràng về độ chính xác giữa các mô hình được khảo sát. Mô hình **LSTM** nổi bật với sai số tuyệt đối trung bình (MAE) duy trì ổn định ở mức rất thấp, dao động quanh ngưỡng **0.1** trong suốt 24 giờ dự đoán. Kết quả này phản ánh rõ khả năng học và biểu diễn hiệu quả các đặc trưng tuần hoàn cũng như các động thái phi tuyến tính trong chuỗi thời gian – thế mạnh vốn có của kiến trúc mạng nơ-ron tuần hoàn như LSTM.



Trái ngược, mô hình **LightGBM** cho thấy mức sai số khá cao, với MAE dao động từ khoảng **2.5 đến 4.0** ở hầu hết các trạm, đồng thời thể hiện sự dao động mạnh theo từng khung giờ dự đoán, cho thấy mô hình gặp khó khăn trong việc nắm bắt sự biến đổi liên tục theo thời gian. Mô hình **XGBoost** đạt kết quả trung gian với MAE nằm trong khoảng **0.5 đến 2**. Mặc dù XGBoost cải thiện đáng kể so với LightGBM, nhưng vẫn chưa đạt được độ ổn định cao như LSTM trong suốt dải thời gian dự đoán.

Kết quả về sai số bình phương trung bình căn (RMSE) cũng củng cố thêm kết luận trên. LSTM tiếp tục giữ ưu thế với RMSE ổn định quanh mức **0.1**, cho thấy không chỉ sai số trung bình thấp mà còn kiểm soát tốt các sai số lớn – một yếu tố quan trọng trong các ứng dụng dự báo theo thời gian thực. Trong khi đó, **LightGBM** có RMSE dao động từ **3.5 đến 7**, cao nhất trong ba mô hình, và **XGBoost** có RMSE dao động từ **1 đến 3**, cho thấy xu hướng sai số biến đổi theo từng khung giờ dự báo.

Đáng chú ý, khi phân tích chi tiết tại hai trạm quan trắc Quận 3 và Thủ Đức, có thể nhận thấy mức sai số dự báo tại trạm Thủ Đức cao hơn đáng kể so với Quận 3 ở cả ba mô hình. Cụ thể, LightGBM tại Thủ Đức có RMSE lên đến gần 20 và MAE đạt 14, trong khi tại Quận 3, các giá trị tương ứng chỉ dao động trong khoảng 15–18 và 12–13. XGBoost tại Thủ Đức cũng ghi nhận RMSE vượt 13 và MAE trên 10 ở một số giờ dự đoán, trong khi tại Quận 3 các chỉ số này thường dưới 10 và 7. Ngược lại, LSTM duy trì độ chính xác ổn định với RMSE và MAE gần như không đổi quanh mức rất thấp ở cả hai trạm. Những chênh lệch này cho thấy dữ liệu tại Thủ Đức có thể biến động mạnh hoặc chứa nhiều nhiễu hơn, gây khó khăn cho các mô hình không chuyên xử lý chuỗi thời gian như các mô hình boosting.

Những kết quả này cho thấy rằng mặc dù các mô hình boosting như XGBoost và LightGBM thường thể hiện tốt trong các bài toán dữ liệu dạng bảng, nhưng lại chưa thực sự phù hợp để mô hình hóa chuỗi thời gian có tính liên tục và phụ thuộc lẫn nhau như dữ liệu chất lượng không khí (AQI). Trong bối cảnh này, LSTM tỏ ra là phương pháp ưu việt hơn, đặc biệt khi mục tiêu là đạt được độ chính xác ổn định và kiểm soát sai số trong các hệ thống dự báo thời gian thực.

Nhìn chung, mô hình LSTM đã thể hiện hiệu quả vượt trội trong nhiệm vụ dự đoán chất lượng không khí theo thời gian thực, đặc biệt nổi bật tại trạm quan trắc Bình Thạnh. Với độ sai số thấp và ổn định trong suốt khoảng thời gian dự đoán, mô hình cho thấy tiềm năng ứng dụng cao trong các hệ thống giám sát môi trường. Kết quả đạt được từ nghiên cứu này có thể so sánh với độ chính xác của những công trình gần đây trong lĩnh vực dự báo chất lượng không khí sử dụng học sâu, qua đó củng cố vai trò của các mô hình học trực tuyến trong việc xây dựng hệ thống cảnh báo và giám sát môi trường đô thị hiện đại.

Ngoài việc trực quan hóa các chỉ số đánh giá mô hình, hiệu năng phần cứng của thiết bị dùng để huấn luyện cũng được theo dõi và phân tích. Trong quá trình này, các thông số bao gồm tỷ lệ sử dụng CPU, dung lượng bộ nhớ RAM RSS, RAM VMS và thời gian hoàn thành mỗi epoch đã được ghi nhận. Bảng [X] tổng hợp các thông số trên, thể hiện giá trị trung bình, tối đa và tối thiểu của từng chỉ số, được thống kê theo từng trạm trong mỗi lần huấn luyện.

*Bảng 2: Số liệu về phần cứng và tổng thời training theo từng trạm*

| Station           | Mô hình         | Tổng số giây | Phần trăm CPU | RAM RSS (MB) | RAM VMS (MB) |
|-------------------|-----------------|--------------|---------------|--------------|--------------|
| <b>Bình Thạnh</b> | <b>LSTM</b>     | 17.67        | 9.70          | 320.81       | 759.82       |
|                   | <b>XGBoost</b>  | 13.01        | 91.63         | 475.96       | 7070.82      |
|                   | <b>LightGBM</b> | 29.69        | 49.82         | 1486.92      | 12977.95     |
| <b>Thu Duc</b>    | <b>LSTM</b>     | 21.50        | 9.50          | 321.18       | 759.30       |
|                   | <b>XGBoost</b>  | 21.50        | 91.17         | 477.44       | 7071.50      |
|                   | <b>LightGBM</b> | 35.61        | 49.82         | 1475.93      | 12977.95     |
| <b>Tan Phu</b>    | <b>LSTM</b>     | 17.48        | 11.20         | 320.51       | 758.38       |
|                   | <b>XGBoost</b>  | 18.45        | 92.19         | 477.51       | 7071.50      |
|                   | <b>LightGBM</b> | 32.11        | 51.92         | 1470.45      | 12977.95     |
| <b>Quận 1</b>     | <b>LSTM</b>     | 17.42        | 8.90          | 320.54       | 758.70       |
|                   | <b>XGBoost</b>  | 16.81        | 91.61         | 474.81       | 7070.82      |
|                   | <b>LightGBM</b> | 30.96        | 49.89         | 1438.90      | 12977.95     |
| <b>Quận 2</b>     | <b>LSTM</b>     | 16.58        | 13.40         | 321.50       | 759.67       |
|                   | <b>XGBoost</b>  | 16.88        | 91.09         | 474.55       | 7069.09      |

|                |                 |       |       |         |          |
|----------------|-----------------|-------|-------|---------|----------|
|                | <b>LightGBM</b> | 29.72 | 49.80 | 1431.40 | 12977.95 |
| <b>Quận 3</b>  | <b>LSTM</b>     | 17.32 | 9.60  | 320.80  | 759.21   |
|                | <b>XGBoost</b>  | 17.42 | 91.64 | 386.89  | 6713.30  |
|                | <b>LightGBM</b> | 43.36 | 52.70 | 1213.84 | 12977.95 |
| <b>Quận 6</b>  | <b>LSTM</b>     | 17.42 | 6.30  | 320.39  | 758.42   |
|                | <b>XGBoost</b>  | 14.78 | 91.35 | 482.15  | 7081.88  |
|                | <b>LightGBM</b> | 32.60 | 50.61 | 1529.75 | 12977.95 |
| <b>Quận 7</b>  | <b>LSTM</b>     | 17.56 | 12.00 | 320.72  | 758.90   |
|                | <b>XGBoost</b>  | 12.26 | 91.46 | 482.15  | 7081.88  |
|                | <b>LightGBM</b> | 29.20 | 49.75 | 1526.05 | 12977.95 |
| <b>Quận 9</b>  | <b>LSTM</b>     | 15.82 | 10.10 | 320.49  | 758.39   |
|                | <b>XGBoost</b>  | 11.63 | 91.06 | 482.15  | 7081.88  |
|                | <b>LightGBM</b> | 32.04 | 49.77 | 1501.19 | 12977.95 |
| <b>Quận 10</b> | <b>LSTM</b>     | 18.05 | 10.00 | 321.08  | 759.48   |
|                | <b>XGBoost</b>  | 13.27 | 91.02 | 482.15  | 7081.88  |
|                | <b>LightGBM</b> | 32.58 | 49.82 | 1469.00 | 12977.95 |
| <b>Quận 11</b> | <b>LSTM</b>     | 17.40 | 10.00 | 320.77  | 759.06   |
|                | <b>XGBoost</b>  | 15.08 | 90.54 | 482.15  | 7081.88  |
|                | <b>LightGBM</b> | 31.37 | 49.90 | 1507.60 | 12977.95 |

Kết quả thực nghiệm cho thấy sự khác biệt đáng kể về hiệu suất phân cứng giữa ba mô hình học máy – LSTM, XGBoost và LightGBM – trong bài toán dự đoán chất lượng không khí tại TP.HCM. Đáng chú ý, mô hình LSTM được triển khai và huấn luyện trên máy tính cá nhân với cấu hình trung bình, trong khi hai mô hình còn lại, XGBoost và LightGBM, được huấn luyện trên nền tảng Google Colab. Sự phân bổ môi trường huấn luyện được lựa chọn có chủ đích nhằm phù hợp với đặc điểm kỹ thuật và yêu cầu tính toán riêng của từng mô hình.

Mô hình LSTM cho thấy khả năng hoạt động ổn định trong môi trường tính toán thông thường, với thời gian huấn luyện tương đối thấp (trung bình 16–21 giây), mức sử dụng CPU dao động dưới 14%, và bộ nhớ RAM (RSS) duy trì ở mức xấp xỉ 320 MB. Điều này phản ánh ưu điểm của LSTM trong việc tiết kiệm tài nguyên, đặc biệt phù hợp với hệ thống không có GPU hỗ trợ.

Trong khi đó, các mô hình XGBoost và LightGBM, được huấn luyện trên Google Colab, thể hiện khả năng tận dụng tốt tài nguyên xử lý. XGBoost đạt hiệu suất huấn luyện nhanh, thường dưới 20 giây, nhưng sử dụng CPU ở mức rất cao (trên 90%), cùng với bộ nhớ ảo (VMS) lên tới khoảng 7 GB, cho thấy khả năng xử lý song song mạnh mẽ nhưng cũng tiềm ẩn nguy cơ quá tải hệ thống nếu triển khai trên môi trường giới hạn. Ngược lại, LightGBM dù có mức sử dụng CPU thấp hơn (khoảng 50%) nhưng lại tiêu tốn thời gian huấn luyện nhiều hơn (dao động từ 29 đến hơn 43 giây) và bộ nhớ RAM cao nhất trong ba mô hình (trên 1400 MB), với VMS gần 13 GB cho mọi trạm.

Từ những số liệu trên, có thể thấy rằng việc phân bổ mô hình LSTM cho môi trường tính toán cơ bản là hoàn toàn hợp lý, đảm bảo tính khả thi mà không làm ảnh hưởng đến tiến trình huấn luyện. Đồng thời, việc tận dụng nền tảng tính toán cao hơn như Google Colab cho XGBoost và LightGBM giúp phát huy được sức mạnh xử lý của hai mô hình này, tối ưu thời gian huấn luyện đối với tập dữ liệu lớn. Tuy nhiên, do Google Colab không cung cấp được thông tin chi tiết về GPU trong quá trình thực nghiệm, nên hiệu suất thực sự của các mô hình GPU-based chưa được đánh giá đầy đủ trong nghiên cứu này.

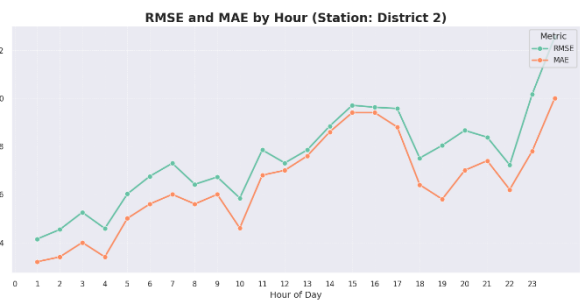
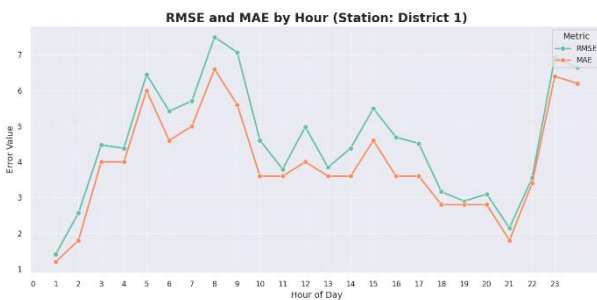
### 3.4.2 Kết quả dự đoán dữ liệu chất lượng không khí trong 24 giờ trên thời gian thực

Sau khi mô hình hoàn tất quá trình dự đoán chất lượng không khí, kết quả dự báo tại từng trạm quan trắc được ghi lại. Khi dữ liệu thực tế mới được cập nhật, các giá trị dự đoán trước đó được so sánh với dữ liệu quan trắc nhằm đánh giá độ chính xác của mô hình. Quy trình này được lặp lại 5 lần cho mỗi trạm để đảm bảo tính ổn định và nhất quán của kết quả dự báo. Hai chỉ số đánh giá sai số phổ biến là RMSE và MAE đã được sử dụng làm cơ sở đánh giá hiệu suất mô hình. Hình 23 minh họa một mẫu tệp được sử dụng để ghi nhận kết quả dự đoán và thực hiện so sánh với dữ liệu thực tế.

|    | A    | B         | C         | D          | E   | F         | G         | H          | I   | J         | K         | L          | M   | N         | O         | P          | Q   | R         | S         | T          | U   | V           | W    |
|----|------|-----------|-----------|------------|-----|-----------|-----------|------------|-----|-----------|-----------|------------|-----|-----------|-----------|------------|-----|-----------|-----------|------------|-----|-------------|------|
| 1  | hour | predicted | real data | $\Delta^2$ | Abs | predicted | real data | $\Delta^2$ | Abs | predicted | real data | $\Delta^2$ | Abs | predicted | real data | $\Delta^2$ | Abs | predicted | real data | $\Delta^2$ | Abs | RMSE        | MAE  |
| 2  | 1    | 85        | 87        | 4          | 2   | 85        | 89        | 16         | 4   | 83        | 79        | 16         | 4   | 90        | 94        | 16         | 4   | 86        | 85        | 1          | 1   | 3.255764119 | 3    |
| 3  | 2    | 86        | 83        | 9          | 3   | 88        | 92        | 16         | 4   | 80        | 75        | 25         | 5   | 91        | 97        | 36         | 6   | 89        | 80        | 81         | 9   | 5.779273311 | 5.4  |
| 4  | 3    | 84        | 82        | 4          | 2   | 87        | 94        | 49         | 7   | 77        | 76        | 1          | 1   | 91        | 95        | 16         | 4   | 88        | 79        | 81         | 9   | 5.495452666 | 4.6  |
| 5  | 4    | 84        | 82        | 4          | 2   | 88        | 94        | 36         | 6   | 77        | 76        | 1          | 1   | 93        | 95        | 4          | 2   | 90        | 79        | 121        | 11  | 5.761944116 | 4.4  |
| 6  | 5    | 83        | 81        | 4          | 2   | 84        | 96        | 144        | 12  | 80        | 77        | 9          | 3   | 88        | 92        | 16         | 4   | 86        | 78        | 64         | 8   | 6.884765791 | 5.8  |
| 7  | 6    | 82        | 85        | 9          | 3   | 85        | 93        | 64         | 8   | 82        | 79        | 9          | 3   | 90        | 86        | 16         | 4   | 86        | 77        | 81         | 9   | 5.983310121 | 5.4  |
| 8  | 7    | 82        | 85        | 9          | 3   | 84        | 93        | 81         | 9   | 81        | 79        | 4          | 2   | 88        | 86        | 4          | 2   | 84        | 77        | 49         | 7   | 5.422176685 | 4.6  |
| 9  | 8    | 81        | 88        | 49         | 7   | 81        | 90        | 81         | 9   | 85        | 82        | 9          | 3   | 86        | 80        | 36         | 6   | 82        | 75        | 49         | 7   | 6.693280212 | 6.4  |
| 10 | 9    | 80        | 84        | 16         | 4   | 83        | 81        | 4          | 2   | 85        | 87        | 4          | 2   | 86        | 75        | 121        | 11  | 83        | 72        | 121        | 11  | 7.293833012 | 6    |
| 11 | 10   | 81        | 84        | 9          | 3   | 82        | 81        | 1          | 1   | 85        | 87        | 4          | 2   | 85        | 75        | 100        | 10  | 80        | 72        | 64         | 8   | 5.966573556 | 4.8  |
| 12 | 11   | 79        | 79        | 0          | 0   | 78        | 71        | 49         | 7   | 86        | 92        | 36         | 6   | 82        | 70        | 144        | 12  | 78        | 69        | 81         | 9   | 7.874007874 | 6.8  |
| 13 | 12   | 79        | 76        | 9          | 3   | 80        | 68        | 144        | 12  | 88        | 94        | 36         | 6   | 83        | 68        | 225        | 15  | 80        | 72        | 64         | 8   | 9.777525249 | 8.8  |
| 14 | 13   | 76        | 76        | 0          | 0   | 74        | 68        | 36         | 6   | 92        | 94        | 4          | 2   | 75        | 68        | 49         | 7   | 73        | 71        | 4          | 2   | 4.312771731 | 3.4  |
| 15 | 14   | 79        | 72        | 49         | 7   | 78        | 64        | 196        | 14  | 90        | 97        | 49         | 7   | 79        | 67        | 144        | 12  | 77        | 70        | 49         | 7   | 9.869143833 | 9.4  |
| 16 | 15   | 75        | 70        | 25         | 5   | 72        | 64        | 64         | 8   | 94        | 95        | 1          | 1   | 71        | 67        | 16         | 4   | 70        | 69        | 1          | 1   | 4.626013402 | 3.8  |
| 17 | 16   | 78        | 70        | 64         | 8   | 77        | 64        | 169        | 13  | 90        | 95        | 25         | 5   | 76        | 67        | 81         | 9   | 74        | 69        | 25         | 5   | 8.532291603 | 8    |
| 18 | 17   | 75        | 68        | 49         | 7   | 81        | 65        | 256        | 16  | 87        | 92        | 25         | 5   | 83        | 67        | 256        | 16  | 82        | 69        | 169        | 13  | 12.28820573 | 11.4 |
| 19 | 18   | 79        | 72        | 49         | 7   | 78        | 70        | 64         | 8   | 87        | 86        | 1          | 1   | 79        | 73        | 36         | 6   | 78        | 69        | 81         | 9   | 6.797058187 | 6.2  |
| 20 | 19   | 80        | 72        | 64         | 8   | 78        | 70        | 64         | 8   | 90        | 86        | 16         | 4   | 78        | 73        | 25         | 5   | 78        | 69        | 81         | 9   | 7.071067812 | 6.8  |
| 21 | 20   | 80        | 77        | 9          | 3   | 80        | 75        | 25         | 5   | 86        | 80        | 36         | 6   | 81        | 80        | 1          | 1   | 80        | 69        | 121        | 11  | 6.196773354 | 5.2  |
| 22 | 21   | 80        | 83        | 9          | 3   | 78        | 83        | 25         | 5   | 88        | 75        | 169        | 13  | 78        | 87        | 81         | 9   | 79        | 74        | 25         | 5   | 7.861297603 | 7    |
| 23 | 22   | 82        | 83        | 1          | 1   | 83        | 83        | 0          | 0   | 84        | 75        | 81         | 9   | 85        | 87        | 4          | 2   | 85        | 79        | 36         | 6   | 4.939635614 | 3.6  |
| 24 | 23   | 81        | 88        | 49         | 7   | 81        | 91        | 100        | 10  | 84        | 70        | 196        | 14  | 84        | 95        | 121        | 11  | 83        | 85        | 4          | 2   | 9.695359715 | 8.8  |
| 25 | 24   | 84        | 88        | 16         | 4   | 86        | 97        | 121        | 11  | 80        | 68        | 144        | 12  | 89        | 100       | 121        | 11  | 86        | 87        | 1          | 1   | 8.977750275 | 7.8  |

Hình 23: Mẫu file ghi nhận kết quả dự đoán của mô hình và đánh giá mô hình dựa trên dữ liệu thực.

Sau quá trình huấn luyện các mô hình học trực tuyến, mô hình LSTM đã được đánh giá là có độ chính xác cao nhất. Do đó, mô hình này đã được lựa chọn để áp dụng trong việc dự đoán chất lượng không khí tại các trạm quan trắc trên dữ liệu thực tế. Các kết quả dự đoán và đánh giá hiệu suất mô hình được trực quan hóa dưới dạng biểu đồ, như thể hiện trong Hình 24.





Hình 24: Đánh giá mô hình dự đoán chất lượng không khí dựa trên dữ liệu thực với từng trạm

Kết quả đánh giá mô hình trên dữ liệu thực cho thấy mức độ biến động sai số theo từng giờ tại các trạm quan trắc có sự khác biệt rõ rệt, phản ánh tính phức tạp và đặc thù của dữ liệu thực tế so với dữ liệu huấn luyện. Tại đa số trạm như Quận 1, Quận 2, Quận 7, Quận 9 và Quận 10, cả RMSE và MAE đều cho thấy xu hướng tăng dần vào ban ngày, đặc biệt là vào các khung giờ cao điểm từ 6h đến 9h và 14h đến 17h, sau đó giảm nhẹ vào buổi tối. Ví dụ, tại Quận 1, RMSE đạt đỉnh trên 7 vào khoảng 8h sáng và 23h, trong khi tại Quận 2, giá trị này lên đến gần 13 vào khoảng 15h và 23h. Điều này cho thấy những thời điểm có mật độ giao thông cao hoặc biến động khí tượng lớn có thể làm tăng sai số dự đoán.

Một số trạm như Quận 6, Quận 11 và Bình Thanh thể hiện sai số dao động nhẹ hơn, tuy vẫn xuất hiện những đỉnh cao đột biến ở một số giờ (như 13h và 17h), cho thấy ảnh hưởng của nhiễu hoặc bất thường cục bộ trong dữ liệu. Các biến động này cần được lưu ý trong các bước xử lý tiền xử lý và huấn luyện mô hình trong tương lai.

Đặc biệt, hai trạm Quận 3 và Thủ Đức tiếp tục thể hiện đặc trưng sai số cao tương tự như khi đánh giá trên dữ liệu huấn luyện, cho thấy sự nhất quán trong thách thức mà mô hình gặp phải với dữ liệu từ hai khu vực này. Tại Quận 3, RMSE dao động mạnh từ khoảng 5 đến gần 17 tùy theo giờ, với các đỉnh rõ rệt ở khung giờ 8h, 14h và 16h. Trong khi đó, Thủ Đức tiếp tục là trạm có sai số cao nhất trong toàn bộ hệ thống, với RMSE vượt ngưỡng 20 vào một số thời điểm như 23h và MAE thường xuyên trên 10, đặc biệt trong khoảng thời gian từ 4h đến 17h.

Kết quả này phản ánh đúng xu hướng sai số cao đã quan sát được trong giai đoạn huấn luyện, cho thấy các mô hình, đặc biệt là mô hình boosting như LightGBM và XGBoost, vẫn chưa thích ứng hiệu quả với tính biến động và nhiễu động dữ liệu tại các khu vực này. Trong khi đó, mặc dù không hiển thị riêng trong các biểu đồ này, LSTM trong giai đoạn huấn luyện cho thấy khả năng kiểm soát sai số tốt hơn nhờ khai thác được đặc tính chuỗi thời gian – điều mà các mô hình boosting thường bỏ qua.

### 3.4.3 Kết quả dự đoán dữ liệu sử dụng phương pháp IDW

Để đánh giá hiệu quả của phương pháp nội suy IDW được áp dụng trong nghiên cứu này, một giả định về giá trị AQI tại một trạm quan trắc đã được thiết lập, coi như kết quả cần dự đoán. Giá trị dự báo sau đó được so sánh với dữ liệu thực tế bằng cách sử dụng các chỉ số đánh giá độ chính xác gồm **precision** và **recall**. Hình 25 minh họa một mẫu tệp ghi nhận kết quả dự đoán sử dụng phương pháp IDW, cùng với dữ liệu thực tế để phục vụ quá trình đánh giá.

|    | A   | B         | C         | D               | E          | F     | G | H             | I             | J      | K     | L       |
|----|-----|-----------|-----------|-----------------|------------|-------|---|---------------|---------------|--------|-------|---------|
| 1  | No. | predicted | real data | predicted level | real level | match |   | AQI Category  | Precision     | Recall | Count | Count ✓ |
| 2  | 1   | 64        | 61        | Moderate        | Moderate   | ✓     |   | Good          | 0.83          | 1.00   | 30    | 20      |
| 3  | 2   | 66        | 67        | Moderate        | Moderate   | ✓     |   | Moderate      | 0.89          | 0.76   | 30    | 25      |
| 4  | 3   | 65        | 67        | Moderate        | Moderate   | ✓     |   | USG           | 0.80          | 0.63   | 30    | 24      |
| 5  | 4   | 63        | 58        | Moderate        | Moderate   | ✓     |   | Unhealthy     | 0.00          | -      | 10    | 0       |
| 6  | 5   | 66        | 67        | Moderate        | Moderate   | ✓     |   | VeryUnhealthy | -             | -      | 0     | 0       |
| 7  | 6   | 65        | 67        | Moderate        | Moderate   | ✓     |   | Hazardous     | -             | -      | 0     | 0       |
| 8  | 7   | 62        | 55        | Moderate        | Moderate   | ✓     |   |               |               |        |       |         |
| 9  | 8   | 63        | 64        | Moderate        | Moderate   | ✓     |   |               |               |        |       |         |
| 10 | 9   | 62        | 63        | Moderate        | Moderate   | ✓     |   |               |               |        |       |         |
| 11 | 10  | 63        | 59        | Moderate        | Moderate   | ✓     |   |               |               |        |       |         |
| 12 | 11  | 64        | 64        | Moderate        | Moderate   | ✓     |   |               |               |        |       |         |
| 13 | 12  | 64        | 64        | Moderate        | Moderate   | ✓     |   |               |               |        |       |         |
| 14 | 13  | 61        | 61        | Moderate        | Moderate   | ✓     |   |               |               |        |       |         |
| 15 | 14  | 63        | 64        | Moderate        | Moderate   | ✓     |   |               |               |        |       |         |
| 16 | 15  | 63        | 64        | Moderate        | Moderate   | ✓     |   |               |               |        |       |         |
| 17 | 16  | 99        | 114       | Moderate        | USG        | X     |   |               |               |        |       |         |
| 18 | 17  | 101       | 103       | USG             | USG        | ✓     |   |               |               |        |       |         |
| 19 | 18  | 102       | 104       | USG             | USG        | ✓     |   | Threshold     | Level         |        |       |         |
| 20 | 19  | 111       | 129       | USG             | USG        | ✓     |   | 0             | Good          |        |       |         |
| 21 | 20  | 106       | 106       | USG             | USG        | ✓     |   | 51            | Moderate      |        |       |         |
| 22 | 21  | 110       | 108       | USG             | USG        | ✓     |   | 101           | USG           |        |       |         |
| 23 | 22  | 109       | 116       | USG             | USG        | ✓     |   | 151           | Unhealthy     |        |       |         |
| 24 | 23  | 106       | 106       | USG             | USG        | ✓     |   | 201           | VeryUnhealthy |        |       |         |
| 25 | 24  | 109       | 108       | USG             | USG        | ✓     |   | 301           | Hazardous     |        |       |         |
| 26 | 25  | 109       | 92        | USG             | Moderate   | X     |   |               |               |        |       |         |
| 27 | 26  | 107       | 109       | USG             | USG        | ✓     |   |               |               |        |       |         |
| 28 | 27  | 111       | 112       | USG             | USG        | ✓     |   |               |               |        |       |         |
| 29 | 28  | 100       | 94        | Moderate        | Moderate   | ✓     |   |               |               |        |       |         |
| 30 | 29  | 100       | 103       | Moderate        | USG        | X     |   |               |               |        |       |         |
| 31 | 30  | 103       | 105       | USG             | USG        | ✓     |   |               |               |        |       |         |
| 32 | 31  | 101       | 89        | USG             | Moderate   | X     |   |               |               |        |       |         |
| 33 | 32  | 100       | 103       | Moderate        | USG        | X     |   |               |               |        |       |         |
| 34 | 33  | 103       | 105       | USG             | USG        | ✓     |   |               |               |        |       |         |

Hình 25: Mẫu file ghi nhận kết quả dự đoán của phương pháp IDW và đánh giá phương pháp dựa trên dữ liệu thực.

Một tập gồm 100 mẫu dữ liệu đại diện cho hầu hết các mức độ nguy hiểm về chất lượng không khí tại Thành phố Hồ Chí Minh đã được thu thập để phục vụ quá trình đánh giá. Bảng [X] trình bày kết quả các chỉ số **precision** và **recall** thu được từ việc áp dụng phương pháp nội suy IDW trên tập dữ liệu này.



*Bảng 3: Kết quả của phương pháp IDW.*

| <b>US AQI Level</b>            | <b>Precision</b> | <b>Recall</b> |
|--------------------------------|------------------|---------------|
| Good                           | 0.83             | 1.00          |
| Moderate                       | 0.89             | 0.76          |
| Unhealthy for Sensitive groups | 0.80             | 0.63          |
| Unhealthy                      | 0.00             | -             |
| Very unhealthy                 | -                | -             |
| Hazardous                      | -                | -             |
| <b>Total</b>                   | <b>0.82</b>      | <b>0.76</b>   |

Kết quả đánh giá phương pháp nội suy IDW (Inverse Distance Weighting) cho thấy phương pháp này đạt được mức hiệu quả đáng kể trong việc dự đoán chất lượng không khí tại các khu vực không có trạm quan trắc. Nhìn chung, độ chính xác (precision) đạt mức cao đối với các nhóm AQI phổ biến như Good (0,83), Moderate (0,89) và Unhealthy for Sensitive Groups – USG (0,80). Điều này cho thấy phương pháp IDW có khả năng nội suy hợp lý chỉ số AQI tại các vị trí chưa được giám sát trực tiếp, đặc biệt trong trường hợp các trạm lân cận cung cấp dữ liệu đầy đủ và có sự phân bố đồng đều.

Tuy nhiên, đối với nhóm Unhealthy, precision ghi nhận giá trị bằng 0 và recall không thể xác định do không có mẫu nào trong nhóm này được dự đoán chính xác. Nguyên nhân chủ yếu được cho là do trong thời điểm thử nghiệm, hầu như không có trạm nào ghi nhận giá trị AQI ở mức Unhealthy, khiến việc nội suy tại vùng này trở nên không đáng tin cậy hoặc không khả thi. Đây là một hạn chế mang tính khách quan, xuất phát từ đặc điểm phân bố không gian của dữ liệu thực tế hơn là từ bản chất của thuật toán nội suy.

Mặc dù tồn tại hạn chế nêu trên, khi xét trên toàn bộ tập dữ liệu, phương pháp IDW vẫn thể hiện hiệu quả đáng kể trong điều kiện thực tế, với 38 trên tổng số 100 mẫu thử nghiệm được dự đoán chính xác (tương đương 76%), chủ yếu thuộc các nhóm phổ biến như Moderate và USG. Phương pháp này đặc biệt phù hợp trong bối cảnh hệ thống quan trắc còn thưa thớt, đóng vai trò hỗ trợ trong việc bổ sung dữ liệu tại các khu vực thiếu thông tin, từ đó nâng cao khả năng trực quan hóa bản đồ chất lượng không khí một cách toàn diện và liên tục hơn.

#### 3.4.4 *Đánh giá hệ thống web*

Hệ thống web đã được thiết kế với định hướng lấy trải nghiệm người dùng làm trung tâm, đảm bảo giao diện trực quan, dễ thao tác, đồng thời duy trì tốc độ phản hồi nhanh và độ chính xác cao trong quá trình dự đoán.

Trong chức năng hiển thị biểu đồ, hệ thống cho phép phản hồi gần như tức thì sau khi người dùng thay đổi các bộ lọc. Các điểm dữ liệu trên biểu đồ được thiết kế có khả năng tương tác, cho phép hiển thị thông tin chi tiết tương ứng, qua đó hỗ trợ quá trình phân tích trực quan một cách linh hoạt và hiệu quả.

Đối với chức năng dự đoán chất lượng không khí trong 24 giờ tiếp theo, thời gian phản hồi được ghi nhận là rất nhanh — kết quả dự báo thường được trả về trong thời gian dưới 1 giây kể từ khi người dùng kích hoạt thao tác tìm kiếm. Tính năng này đóng vai trò thiết yếu trong việc cung cấp thông tin theo thời gian thực, phục vụ hiệu quả cho các nhu cầu giám sát môi trường và hỗ trợ ra quyết định kịp thời.

Trong chức năng dự đoán tại các khu vực không có trạm quan trắc, bản đồ Thành phố Hồ Chí Minh được hiển thị với độ phân giải cao, hỗ trợ khả năng thu phóng linh hoạt và giao diện thể hiện rõ ràng. Dữ liệu từ các trạm quan trắc được trình bày dưới dạng các điểm màu thể hiện mức độ ô nhiễm không khí, với khả năng tương tác để hiển thị thông tin chi tiết như tên trạm, thời gian đo và chỉ số AQI. Mặc dù quá trình thu thập dữ liệu từ nền tảng IQAir có thể chịu ảnh hưởng bởi tốc độ kết nối mạng, hiệu suất thu thập vẫn được duy trì ở mức chấp nhận được, đảm bảo tính cập nhật và minh bạch của thông tin hiển thị.

## KẾT LUẬN CHƯƠNG

Chương này đã trình bày chi tiết quá trình thực nghiệm và đánh giá hiệu suất của hệ thống dự đoán chất lượng không khí tại TP.HCM dựa trên mô hình học trực tuyến và kỹ thuật nội suy không gian. Thông qua việc thiết lập môi trường thử nghiệm thực tế và sử dụng dữ liệu thu thập liên tục trong 43 ngày, các mô hình như LSTM, LightGBM và XGBoost đã được triển khai, huấn luyện và so sánh hiệu quả trên nhiều trạm quan trắc khác nhau. Kết quả thực nghiệm cho thấy mỗi mô hình có những ưu thế riêng biệt, song mô hình được lựa chọn cuối cùng đạt hiệu suất vượt trội về cả độ chính xác dự báo và khả năng vận hành ổn định theo thời gian thực.

Bên cạnh việc đánh giá các mô hình dự báo AQI tại các trạm, chương này cũng phân tích hiệu quả của thuật toán Inverse Distance Weighting (IDW) trong việc nội suy chất lượng không khí tại những khu vực chưa có trạm quan trắc. Kết quả minh chứng cho khả năng mở rộng không gian của hệ thống mà không đòi hỏi hạ tầng phần cứng bổ sung. Ngoài ra, hệ thống web cũng được thử nghiệm và đánh giá về mặt hiệu năng, khả năng cập nhật mô hình định kỳ và giao diện người dùng, từ đó chứng minh được tính ứng dụng thực tiễn trong môi trường đô thị năng động.

Nhìn chung, các kết quả thu được trong chương này đã xác thực tính khả thi và hiệu quả của hệ thống được đề xuất, đồng thời làm cơ sở thực tiễn vững chắc cho những kết luận tổng thể và định hướng phát triển được trình bày trong chương tiếp theo.

## Chương 4 KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

### 4.1 Kết luận

Trong khóa luận “**Phát triển mô hình học trực tuyến kết hợp nội suy không gian để dự đoán chất lượng không khí tại TP.HCM**”, một hệ thống dự báo chất lượng không khí tại Thành phố Hồ Chí Minh đã được phát triển dựa trên mô hình học trực tuyến. Cụ thể, tập dữ liệu đã được mở rộng để tăng khả năng dự đoán của mô hình. Ngoài ra, khóa luận còn đề xuất phương pháp nội suy giúp dự đoán chất lượng không khí tại các địa điểm lân cận.

Hệ thống dự báo được tích hợp vào một ứng dụng web, cho phép người dùng tra cứu chất lượng không khí tại từng trạm, với dữ liệu được cập nhật tự động theo chu kỳ hàng giờ. Bên cạnh đó, một tính năng mở rộng cũng đã được phát triển nhằm ước lượng chất lượng không khí tại các khu vực không có trạm quan trắc, thông qua việc áp dụng các kỹ thuật nội suy không gian. Kết quả dự báo tại các khu vực này được trực quan hóa dưới dạng bản đồ và biểu đồ tương tác, góp phần nâng cao khả năng tiếp cận thông tin và hỗ trợ người dùng trong việc theo dõi chất lượng môi trường xung quanh.

Tuy nhiên, nghiên cứu vẫn còn tồn tại một số hạn chế cần được khắc phục trong các giai đoạn phát triển tiếp theo. Thứ nhất, mô hình hiện tại chủ yếu dựa trên xu hướng biến động của chỉ số chất lượng không khí theo thời gian, chưa tích hợp các yếu tố ngoại cảnh có ảnh hưởng trực tiếp đến mức độ ô nhiễm không khí như nhiệt độ, độ ẩm, tốc độ gió, áp suất khí quyển và lượng mưa. Bên cạnh đó, các yếu tố liên quan đến hoạt động của con người, đặc biệt là giao thông và công nghiệp, cũng chưa được đưa vào như là biến đầu vào trong quá trình huấn luyện mô hình, dù đây là những nguồn phát thải quan trọng trong môi trường đô thị.

Thứ hai, khả năng tự động hóa của hệ thống còn nhiều hạn chế. Cụ thể, dữ liệu đầu vào hiện đang được cập nhật thủ công nhằm đảm bảo tính chính xác và tránh các lỗi phát sinh do dữ liệu thiếu hụt tạm thời. Việc sử dụng API để thu thập dữ liệu thời gian thực tuy khả thi nhưng chưa được triển khai rộng rãi do chi phí cao. Bên cạnh đó, quá trình huấn luyện lại mô hình cũng đang được thực hiện bằng tay, khiến việc cập nhật mô hình theo thời gian thực chưa đạt được tính linh hoạt và hiệu quả như mong đợi.

## 4.2 Tác động của đề tài đến xã hội

Đề tài “Phát triển mô hình học trực tuyến kết hợp nội suy không gian để dự đoán chất lượng không khí tại TP.HCM” mang lại nhiều tác động đáng kể đến cộng đồng, đặc biệt trong bối cảnh đô thị hóa nhanh chóng và tình trạng ô nhiễm không khí ngày càng nghiêm trọng tại các thành phố lớn như Thành phố Hồ Chí Minh.

Về mặt tích cực, hệ thống dự báo chất lượng không khí theo thời gian thực có khả năng hỗ trợ người dân chủ động điều chỉnh lịch trình sinh hoạt, đặc biệt là các nhóm dễ bị tổn thương như trẻ em, người cao tuổi và người mắc bệnh hô hấp. Việc kết hợp nội suy không gian giúp mở rộng phạm vi dự báo đến những khu vực không có trạm quan trắc, qua đó cung cấp bức tranh toàn diện hơn về tình trạng ô nhiễm không khí trên toàn địa bàn thành phố. Ngoài ra, hệ thống có thể trở thành công cụ hỗ trợ ra quyết định cho các cơ quan quản lý trong việc đưa ra các chính sách can thiệp kịp thời nhằm giảm thiểu tác động của ô nhiễm không khí đến sức khỏe cộng đồng và môi trường đô thị.

Tuy nhiên, đề tài cũng có thể mang đến một số tác động tiêu cực. Việc công bố thông tin chất lượng không khí theo thời gian thực, nếu không đi kèm với hướng dẫn đầy đủ và truyền thông phù hợp, có thể gây hoang mang trong dư luận, đặc biệt khi chỉ số ô nhiễm ở mức cao. Hơn nữa, độ chính xác của mô hình vẫn phụ thuộc vào chất lượng dữ liệu đầu vào và hiệu quả của thuật toán nội suy, nên nguy cơ đưa ra dự báo sai lệch là điều không thể loại trừ, từ đó dẫn đến những hành động không phù hợp từ phía người dân hoặc nhà quản lý. Cuối cùng, việc duy trì và vận hành hệ thống dự báo theo thời gian thực đòi hỏi nguồn lực kỹ thuật và tài chính liên tục, điều này có thể là rào cản trong quá trình ứng dụng rộng rãi nếu không có sự đầu tư và hỗ trợ lâu dài.

## 4.3 Hướng phát triển

Hệ thống hiện đang trong quá trình hoàn thiện với mục tiêu trở thành một công cụ ứng dụng thiết thực cho cư dân và các nhà quản lý đô thị. Trong tương lai, định hướng phát triển của hệ thống bao gồm các mục tiêu sau:

- Nâng cao chất lượng và độ bao phủ của tập dữ liệu về chất lượng không khí thu thập tại các trạm quan trắc trên địa bàn TP.HCM nhằm đảm bảo tính chính xác và đại diện của đầu vào mô hình.

- Mở rộng phạm vi phân tích bằng cách tích hợp các yếu tố môi trường và xã hội có ảnh hưởng đến chất lượng không khí, từ đó cải thiện hiệu suất dự đoán và cung cấp các khuyến nghị đa chiều hơn cho người dân cũng như các nhà quản lý.
- Tăng cường mức độ tự động hóa trong quy trình cập nhật dữ liệu và tái huấn luyện mô hình nhằm duy trì tính thích ứng với những biến động thực tế, đồng thời giảm thiểu sự can thiệp thủ công trong vận hành hệ thống.

## TÀI LIỆU THAM KHẢO

- [1] Rakholia, R., Le, Q., Vu, K., Ho, B. Q., & Carbajo, R. S. (2022). AI-based air quality PM2. 5 forecasting models for developing countries: A case study of Ho Chi Minh City, Vietnam. *Urban Climate*, 46, 101315.
- [2] Rakholia, R., Le, Q., Ho, B. Q., Vu, K., & Carbajo, R. S. (2023). Multi-output machine learning model for regional air pollution forecasting in Ho Chi Minh City, Vietnam. *Environment international*, 173, 107848.
- [3] Nguyen, P. H., Dao, N. K., & Nguyen, L. S. P. (2024). Development of machine learning and deep learning prediction models for PM2. 5 in Ho Chi Minh City, Vietnam. *Atmosphere*, 15(10), 1163.
- [4] Liu, Q., Cui, B., & Liu, Z. (2024). Air quality class prediction using machine learning methods based on monitoring data and secondary modeling. *Atmosphere*, 15(5), 553.
- [5] Yu, C., Wang, F., Wang, Y., Shao, Z., Sun, T., Yao, D., & Xu, Y. (2025). MGSFformer: A multi-granularity spatiotemporal fusion transformer for air quality prediction. *Information Fusion*, 113, 102607.
- [6] Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., & Schmidhuber, J. (2016). LSTM: A search space odyssey. *IEEE transactions on neural networks and learning systems*, 28(10), 2222-2232.
- [7] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.
- [8] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).