| LOGO DHCNTT -hinh.jpg | ĐẠI HỌC QUỐC GIA TP. HCM  **TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN** | Ngày nhận hồ sơ |  |
| --- | --- | --- | --- |
| *(Do CQ quản lý ghi)* | |

**THUYẾT MINH**

ĐỀ TÀI KHOA HỌC VÀ CÔNG NGHỆ CẤP SINH VIÊN 2025

# THÔNG TIN CHUNG

## A1. Tên đề tài

* Tên tiếng Việt (IN HOA): DỰ ĐOÁN CHỈ SỐ CHẤT LƯỢNG KHÔNG KHÍ BẰNG CÁCH SỬ DỤNG PHƯƠNG PHÁP HỌC SÂU
* Tên tiếng Anh (IN HOA): PREDICTION OF AIR QUALITY INDEX USING DEEP LEARNING

## A2. Thời gian thực hiện

**06** tháng (kể từ khi được duyệt).

## A3. Tổng kinh phí

Tổng kinh phí: **06** triệu đồng,gồm

* Kinh phí từ Trường Đại học Công nghệ Thông tin: **06** triệu đồng

## A4. Chủ nhiệm

Họ và tên: **Phan Nguyễn Lâm Hà.** **.**

Ngày, tháng, năm sinh: 271/11/2003 . Giới tính (Nam/Nữ): Nam.

Số CMND: 075203010658 ; Ngày cấp: 05/06/2021 ; Nơi cấp: Cục Cảnh sát QLHC TTXH .

Mã số sinh viên: 21522030 **.**

Số điện thoại liên lạc: 0385098714 **.**

Đơn vị (Khoa): Hệ thống thông tin **.**

Số tài khoản: 0385098714 Ngân hàng: MB BANK

## A5. Thành viên đề tài

| **TT** | **Họ tên** | **MSSV** | **Khoa** |
| --- | --- | --- | --- |
| 1 | Phan Nguyễn Lâm Hà | 21522030 | Hệ thống thông tin |
| 2 | Đinh Nhật Huy | 21522136 | Hệ thống thông tin |

# 

# MÔ TẢ NGHIÊN CỨU

## B1. Giới thiệu về đề tài

## Ô nhiễm không khí xung quanh đã trở thành "rủi ro sức khỏe môi trường lớn nhất thế giới", gây ra hàng triệu ca tử vong sớm mỗi năm và thiệt hại kinh tế không thể kể xiết. Các thành phố lớn như Seoul là hiện thân của thách thức này: công nghiệp hóa nhanh chóng, mật độ giao thông cao và địa hình giống như lưu vực bẫy một hỗn hợp các chất ô nhiễm chính (SO₂, NO₂, CO) và thứ cấp (O₃, PM₁₀, PM₂․₅) thường xuyên đẩy Chỉ số chất lượng không khí (Air Quality Index - AQI) vào vùng không lành mạnh. Tổ chức Y tế Thế giới hiện phân loại mức PM₂․₅ trung bình hàng năm của Seoul ( ≈ 46µgm⁻³) là cực kỳ kém, mức độ liên quan đến bệnh hen suyễn gia tăng, bệnh tim phổi và tử vong quá mức. Các nhà hoạch định chính sách đã ứng phó bằng các tiêu chuẩn khí thải chặt chẽ hơn, các sáng kiến ​​vận tải ít carbon và mạng lưới giám sát dày đặc, nhưng họ vẫn thiếu các dự báo có độ trung thực cao, thời gian thực có thể đưa ra các biện pháp can thiệp kịp thời và cảnh báo công chúng. Vì vậy việc xây dựng một mô hình dự báo bằng các mô hình, thuật toán Machine Learning, Deep Learning là điều cần thiết để đưa ra các tư vấn, hỗ trợ cho các nhà hoạch định, người dân một cách nhanh chóng.

## Các phương pháp tiếp cận thế hệ đầu tiên như tự hồi quy (autoregressive - AR), trung bình động tích hợp tự hồi quy (Autoregressive Integrated Moving Average - ARIMA [1]), hồi quy tuyến tính bội và mô hình xám mang lại khả năng diễn giải nhưng lại không ổn định khi đối mặt với tính không ổn định và sự thay đổi chế độ đột ngột. Ngay cả với sự khác biệt theo mùa hoặc phần mở rộng logic mờ, hiệu suất của chúng vẫn giảm khi hành vi của chất ô nhiễm trở nên phi tuyến tính cao. Hồi quy vectơ hỗ trợ, rừng ngẫu nhiên và XGBoost [2] đã cải thiện độ chính xác bằng cách nắm bắt các tính phi tuyến tính và tương tác của các tính năng, và đã được triển khai để dự đoán AQI hàng ngày hoặc hàng giờ tại Trung Quốc, Ấn Độ và Châu Âu. Tuy nhiên, chúng coi thời gian chỉ là một tính năng khác và do đó gặp khó khăn với các phụ thuộc thời gian dài, đặc biệt là trong điều kiện dữ liệu thưa thớt hoặc trôi dạt khái niệm.

## Mạng nơ-ron hồi quy (Recurrent neural networks - RNN) [3] và các biến thể có cổng của chúng LSTM [4] và GRU [5] học các mẫu tuần tự trực tiếp từ các luồng cảm biến thô, vượt trội hơn các đường cơ sở ML trong các thiết lập nhiều chất ô nhiễm. Các lớp tích chập (CNN) [6] nắm bắt các họa tiết thời gian cục bộ và các vùng lân cận không gian; việc xếp chồng CNN với RNN giúp nâng cao hiệu suất hơn nữa. Tuy nhiên, vẫn còn ba nút thắt về mặt phương pháp như sự chú ý biến mất đối với các sự kiện cực đoan. Các LSTM tiêu chuẩn có thể làm loãng các đột biến hiếm gặp nhưng quan trọng đối với sức khỏe khi các gradient biến mất trên các đường chân trời dài. Chiều dài cửa sổ, các đơn vị ẩn và tốc độ học phải được điều chỉnh cho từng chất ô nhiễm và trạm, thường thông qua tìm kiếm lưới bằng vũ lực. DL vượt trội ở các giá trị dư phi tuyến tính nhưng có xu hướng "lãng phí năng lực" đối với tính theo mùa tuyến tính mà các mô hình đơn giản hơn có thể xử lý.

## Do đó, công trình gần đây hợp nhất các mô-đun bổ sung: đường ống ARIMA→LSTM để phân tích tuyến tính - phi tuyến tính, các ngăn xếp CNN ‑ GRU để ghép nối không gian thời gian hoặc các trình học tổng hợp như SD ‑ SE ‑ LSTM ‑ BA ‑ LSSVM​. Để tự động điều chỉnh, các nhà nghiên cứu nhập các siêu thuật toán PSO, Harris‑Hawks, Cuckoo Search, Multi‑Verse Optimiser. Tuy nhiên, PSO cổ điển có thể trì trệ ở các cực tiểu cục bộ; biến thể lượng tử của nó (QPSO) giới thiệu các bản cập nhật vị trí xác suất mở rộng bán kính tìm kiếm chỉ với một tham số điều khiển, mang lại sự hội tụ nhanh hơn và khám phá toàn cầu tốt hơn​.

## Tuy nhiên bài toán sẽ có nhiều vấn đề bất cập. Chất lượng dữ liệu là một trong những vấn đề quan trọng, bởi tính đồng nhất và chất lượng dữ liệu, đòi hỏi các chiến lược làm sạch và xử lý trước khi xây dựng mô hình. Các thuật toán học sâu như cơ chế tự chú ý (attention) làm tăng bộ nhớ và thời gian chạy, đặc biệt là khi mô hình hóa sáu chất gây ô nhiễm trong nhiều năm, nhiều tập dữ liệu trạm. Bên cạnh đó khả năng chuyển giao cũng là một vấn đề quan trọng, một mô hình được điều chỉnh theo hồ sơ phát thải của Seoul phải được thử nghiệm ứng suất trên các vùng khí hậu khác để đánh giá khả năng khái quát hóa. Bằng cách đối mặt với những rào cản này, nghiên cứu này nhằm mục đích cung cấp một công cụ dự báo có thể triển khai, diễn giải và có độ chính xác cao, có thể cung cấp thông tin cho việc giảm thiểu theo thời gian thực, dài hạn.

## B2. Mục tiêu, nội dung, kế hoạch nghiên cứu

### B2.1 Mục tiêu

## Tìm hiểu các phương pháp dự đoán theo thời gian như: ARIMA, CNN, GNN, LSTM, ứng dụng trong bài toán dự báo AQI

* Thiết kế và triển khai đường ống kết hợp bộ mã hóa CNN dựa trên Attention, bộ giải mã LSTM được tối ưu hóa QPSO và bộ điều chỉnh tinh chỉnh XGBoost (ACNN‑QPSO‑LSTM‑XGBoost). Dự kiến, kết hợp này sẽ nắm bắt được cả các tính năng toàn cục-cục bộ và các phụ thuộc thời gian dài hạn.
* Nghiên cứu sử dụng QPSO để tinh chỉnh siêu tham số LSTM (kích thước cửa sổ, đơn vị ẩn, kích thước lô) để mạng hội tụ nhanh hơn, tránh sự dư thừa và học tốt hơn các mẫu bất thường có thể bị bỏ sót.
* Xác thực mô hình được đề xuất trên dữ liệu hàng giờ trong hai năm từ 25 trạm giám sát của Seoul và cho thấy những cải tiến được đánh giá trên các chỉ số MSE, MAE và R² so với các đường cơ sở thống kê, học máy và học sâu​.

### B2.2 Nội dung và phương pháp nghiên cứu

**Nội dung 1: Liệt kê nội dung, phương pháp thực hiện và dự kiến kết quả.**

* Mô tả: Chỉ số chất lượng không khí (AQI) hàng giờ của Seoul dao động dưới ảnh hưởng kết hợp của sáu chất ô nhiễm được quy định (PM₂․₅, PM₁₀, SO₂, NO₂, CO và O₃). Vì các chuỗi thời gian này thể hiện tính theo mùa mạnh, đột biến và liên kết phức tạp giữa các chất ô nhiễm, nên các mô hình thống kê cổ điển (AR, ARIMA) và thậm chí cả các mạng sâu độc lập (LSTM, CNN‑GRU) chỉ mang lại độ chính xác ở mức trung bình. Tuy nhiên, các dự báo AQI chính xác ở cấp độ trạm là không thể thiếu đối với các cảnh báo sức khỏe cộng đồng theo thời gian thực và các chính sách kiểm soát khí thải.
* Phương pháp nghiên cứu:
* Trích xuất tuyến tính: Loại bỏ tính theo mùa/xu hướng để người học sâu có thể tập trung vào cấu trúc phi tuyến tính​ bằng cách huấn luyện mô hình ARIMA(p,d,q) cho mỗi chuỗi chất ô nhiễm
* Bộ mã hóa phi tuyến tính: Truyền các giá trị dư qua Attention-CNN (ACNN) kết hợp các phép tích chập đa thang với sự tự chú ý. Ghi lại cả các xung cục bộ (thông qua hạt nhân) và các phụ thuộc giữa các chất gây ô nhiễm tầm xa (thông qua sự chú ý)​
* Bộ giải mã với điều chỉnh tự động: Đưa các tính năng ACNN vào LSTM có các siêu tham số (chiều dài cửa sổ, đơn vị ẩn, kích thước lô) được tối ưu hóa bằng Tối ưu hóa bầy hạt lượng tử (Quantum‑behaved Particle Swarm Optimisation - QPSO)
* Hợp nhất và tinh chỉnh: Nối các dự báo ARIMA và đầu ra LSTM; đào tạo một bộ hồi quy XGBoost trên bộ tính năng.
* Kết quả dự kiến:
* Đánh giá mô hình trên các độ đo như MSE, MAE và R², so sánh với các phương pháp đã sử dụng trước đây.
* Xây dựng được kiến trúc kết hợp từ ARIMA, CNN, LSTM, và XGBoost, ngoài ra kết hợp với phương pháp tối ưu hóa bằng Tối ưu hóa bầy hạt lượng tử (Quantum‑behaved Particle Swarm Optimisation - QPSO)

**Nội dung 2: Trích xuất tuyến tính bằng mô hình ARIMA**

Huấn luyện mô hình ARIMA(p,d,q) với mỗi chuỗi chất ô nhiễm. Các tham số được chọn thông qua kiểm định tính dừng ADF và giảm thiểu AIC (giới hạn p,q≤5, d≤2). Tính theo mùa và xu hướng chiếm phần lớn phương sai; việc loại bỏ chúng cho phép trình học sâu tập trung vào các mô hình phức tạp, phi tuyến tính.

Tại sao nên bắt đầu với ARIMA? Chuỗi chất ô nhiễm theo giờ thể hiện cấu trúc xác định mạnh: Xu hướng – tăng/giảm dần nồng độ nền do hoạt động kinh tế hoặc chính sách dài hạn thúc đẩy. Tính theo mùa – các đỉnh giao thông hàng ngày, chu kỳ ngày trong tuần/cuối tuần, các đợt tăng nhiệt độ vào mùa đông.

Các thành phần tuyến tính này giải thích một phần lớn phương sai. Nếu không loại bỏ chúng, mạng lưới sâu sẽ lãng phí dung lượng học lại những gì bộ lọc thống kê nhẹ có thể nắm bắt. Do đó, trước tiên các tác giả loại bỏ tín hiệu tuyến tính bằng mô hình Tự hồi quy-Tích hợp-Di chuyển-Trung bình (ARIMA) trước khi chuyển các phần dư cho khối attention-CNN-LSTM​.

**Nội dung 3: Bộ mã hóa phi tuyến tính**

Đưa phần dư vào CNN 1 chiều với nhiều kích thước hạt nhân + sự tự chú ý của nhiều đầu để tạo ra ma trận ngữ cảnh H. Lớp chú ý sẽ cân nhắc lại các bước thời gian và chất gây ô nhiễm, phát hiện ra các mối liên kết ô nhiễm chéo và tầm xa (ví dụ: CO ↔ NO₂).

Bộ mã hóa phi tuyến tính trong nghiên cứu này được thiết kế riêng để xử lý các tín hiệu dư thu được từ bước trích xuất tuyến tính ARIMA. Các dư này biểu thị các thành phần phi tuyến tính phức tạp của chuỗi thời gian ô nhiễm không khí, không thể được nắm bắt bằng các mô hình tuyến tính. Do đó, bộ mã hóa sử dụng kiến ​​trúc học sâu tinh vi kết hợp hai cơ chế mạng nơ-ron mạnh mẽ. Mạng nơ-ron tích chập (CNN) với nhiều kích thước hạt nhân. Cơ chế tự chú ý nhiều đầu.

CNN có hiệu quả trong việc xác định các mẫu cục bộ trong chuỗi bằng cách trượt các hạt tích chập theo thời gian. Với CNN 1 chiều xử lý dữ liệu tuần tự như chuỗi thời gian, bảo toàn thứ tự thời gian. Mỗi hạt trượt qua chuỗi dư để phát hiện các mẫu thời gian cục bộ, chẳng hạn như các đột biến phát xạ đột ngột, các sự kiện tồn tại trong thời gian ngắn hoặc các chu kỳ hàng ngày mà ARIMA bỏ lỡ.

Cơ chế tự chú ý nhiều đầu. Để bổ sung cho chế độ xem cục bộ của CNN, bộ mã hóa tích hợp Tự chú ý nhiều đầu từ kiến ​​trúc Transformer, được Vaswani và cộng sự phổ biến (2017). Tự chú ý đo lường mức độ liên quan giữa tất cả các vị trí trong chuỗi, cho phép mô hình tập trung vào các sự kiện và mối quan hệ quan trọng trong dữ liệu. Sự chú ý nhiều đầu áp dụng nhiều cơ chế chú ý (đầu) cùng lúc để nắm bắt các mối quan hệ khác nhau một cách độc lập, cung cấp biểu diễn phong phú hơn.

**Nội dung 4**: **Bộ giải mã với điều chỉnh tự động**

Bộ giải mã với điều chỉnh tự động Bi-LSTM chuỗi-sang-chuỗi nhận H và đưa ra dự báo phi tuyến tính. Các siêu tham số {cửa sổ trượt, neuron₁, neuron₂, kích thước lô} được mã hóa dưới dạng hạt trong Quantum-behaved Particle Swarm Optimiser (QPSO). Không giống như PSO cổ điển, phiên bản lượng tử cho phép các hạt "đào hầm" ra khỏi các cực tiểu cục bộ, khám phá không gian tìm kiếm rộng hơn với một hệ số co-giãn duy nhất do đó điều chỉnh nhanh hơn, đáng tin cậy hơn.

Bộ giải mã với giai đoạn tự động điều chỉnh tận dụng mô hình học sâu lai tinh vi kết hợp với kỹ thuật tối ưu hóa tiên tiến để dự đoán động lực học chất lượng không khí phi tuyến tính. Giai đoạn này cụ thể liên quan đến mạng Bộ nhớ dài ngắn hạn song hướng (Bi-LSTM) có các siêu tham số được tự động điều chỉnh bằng Tối ưu hóa Bầy đàn có hành vi lượng tử (QPSO).

Tại sao cần tự động điều chỉnh ? Các mô hình học sâu như Bi-LSTM liên quan đến các siêu tham số quan trọng ảnh hưởng đáng kể đến hiệu suất dự đoán như số lượng nơ-ron (đơn vị ẩn) trong các lớp LSTM, Số lượng chuỗi được xử lý đồng thời trong quá trình đào tạo. Việc tìm ra các kết hợp tối ưu của các tham số này là một thách thức và tốn nhiều tính toán do không gian tìm kiếm siêu tham số rộng lớn. Điều chỉnh thủ công và các phương pháp tìm kiếm truyền thống (tìm kiếm lưới hoặc tìm kiếm ngẫu nhiên) thường dẫn đến các mô hình không tối ưu hoặc kém hiệu quả về mặt tính toán.

**Nội dung 5**: **Hợp nhất và tinh chỉnh**

Nối dự báo tuyến tính ARIMA và dự báo phi tuyến tính LSTM thành một vectơ đặc trưng; đào tạo một bộ hồi quy XGBoost để đưa ra giá trị AQI cuối cùng. Cây tăng cường gradient nắm bắt bất kỳ độ lệch hoặc tương tác còn lại nào mà cả ARIMA và LSTM đều không mô hình hóa, hoạt động như một bộ điều chỉnh nhẹ.

Sau khi mô hình hóa riêng biệt động lực tuyến tính và phi tuyến tính bằng ARIMA và CNN dựa trên sự chú ý kết hợp với Bi-LSTM được tối ưu hóa QPSO, bước quan trọng tiếp theo là hợp nhất hai dự đoán này để tạo ra dự báo AQI cuối cùng. Bước "Hợp nhất và tinh chỉnh" bao gồm việc tích hợp các dự đoán từ cả hai mô hình thành một vectơ đặc trưng duy nhất và tinh chỉnh chúng bằng cách sử dụng bộ hồi quy XGBoost. Bước này rất cần thiết vì:

* ARIMA nắm bắt hiệu quả các xu hướng tuyến tính và các mô hình theo mùa nhưng lại gặp khó khăn với các thay đổi đột ngột và phi tuyến tính.
* CNN-LSTM dựa trên sự chú ý mô hình hóa chính xác các động lực phi tuyến tính phức tạp nhưng vẫn có thể có sai lệch hoặc lỗi hệ thống.
* Việc tích hợp cả hai đầu ra với XGBoost sẽ giải quyết những khoảng trống này, tạo ra một đường ống dự báo mạnh mẽ và chính xác.

**B2.3 Kế hoạch nghiên cứu**.

Kế hoạch nghiên cứu được trình bày trong bảng sau:

| **Nội dung nghiên cứu** | **Thời gian** |
| --- | --- |
| 1. Tìm hiểu bài toán, xây dựng vấn đề, mục tiêu giải quyết bài toán ô nhiễm không khí bằng mô hình học sâu | 1 tuần |
| 2. Tìm hiểu mô hình trung bình động tích hợp hồi quy tự động (ARIMA) để trích xuất tuyến tính. | 1 tuần |
| 3. Tìm hiểu mạng nơ-ron tích chập dựa trên sự chú ý (ACNN) để trích xuất tính năng sâu. | 2 tuần |
| 4. Tìm hiểu mạng bộ nhớ dài hạn ngắn hạn (LSTM) được cải tiến theo phương pháp tối ưu hóa bầy đàn (QPSO) lấy cảm hứng từ lượng tử để dự đoán phi tuyến tính. | 2 tuần |
| 5. Hồi quy XGBoost để tinh chỉnh các dự đoán AQI cuối cùng​ | 1 tuần |
| 6. Thu thập và xử lý dữ liệu trước: Dữ liệu AQI hàng giờ (O₃, CO, NO₂, SO₂, PM₁₀, PM₂.₅) |  |
| 7. Mô hình lai được đánh giá nghiêm ngặt so với các mô hình cơ sở phổ biến (ví dụ: các biến thể ARIMA, XGBoost, CNN-LSTM, BiLSTM và QPSO) | 2 tuần |
| 8. So sánh với kết quả phương pháp nghiên cứu trước đây và đưa ra những đề xuất mới | 2 tuần |

## B3. Kết quả dự kiến

* Hiểu được các mô hình trong bài toán dự báo thời gian, cụ thể hơn là bài toán dự đoán mức độ ô nhiễm không khí theo thời gian
* Hiểu được cách tối ưu hóa các mô hình học sâu như LSTM bằng cách kết hợp phương pháp tối ưu hóa bầy đàn (QPSO)
* Tích hợp XGBoost tinh chỉnh đáng kể các dự báo AQI cuối cùng bằng cách hiệu chỉnh hiệu quả các sai lệch còn lại và nắm bắt các tương tác giữa các thành phần tuyến tính (ARIMA) và phi tuyến tính (LSTM).
* Hiệu quả tính toán của khuôn khổ lai được đề xuất cho phép dự đoán AQI theo thời gian thực. Độ chính xác và độ tin cậy cao hỗ trợ quản lý môi trường chủ động, khuyến cáo sức khỏe cộng đồng hiệu quả và ra quyết định kịp thời trong các chính sách môi trường và quy hoạch đô thị.

## B4. Tài liệu tham khảo

[1] B. K. Nelson, "Time series analysis using autoregressive integrated moving average (ARIMA) models," \*Academic Emergency Medicine\*, vol. 5, no. 7, pp. 739–744, 1998.

[2] A. Ogunleye and Q. G. Wang, "XGBoost model for chronic kidney disease diagnosis," \*IEEE/ACM Transactions on Computational Biology and Bioinformatics\*, vol. 17, no. 6, pp. 2131–2140, 2019.

[3] A. Sherstinsky, "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network," \*Physica D: Nonlinear Phenomena\*, vol. 404, p. 132306, 2020.

[4] Y. Yu, X. Si, C. Hu, and J. Zhang, "A review of recurrent neural networks: LSTM cells and network architectures," \*Neural Computation\*, vol. 31, no. 7, pp. 1235–1270, 2019.

[5] R. Dey and F. M. Salem, "Gate-variants of gated recurrent unit (GRU) neural networks," in \*Proc. IEEE 60th Int. Midwest Symp. Circuits and Systems (MWSCAS)\*, Boston, MA, USA, Aug. 2017, pp. 1597–1600.

[6] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: Analysis, applications, and prospects," \*IEEE Transactions on Neural Networks and Learning Systems\*, vol. 33, no. 12, pp. 6999–7019, 2021.

| *Ngày \_\_ tháng \_\_ năm 20\_*  **Giảng viên hướng dẫn**  (Ký và ghi rõ họ tên) | *Ngày \_\_ tháng \_\_ năm 20\_*  **Chủ nhiệm đề tài**  (Ký và ghi rõ họ tên) |
| --- | --- |
|  |  |
|  |  |
|  |  |