BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC MỞ THÀNH PHỐ HỒ CHÍ MINH

TRƯƠNG THÁI BẢO

XÂY DỰNG MÔ HÌNH DEEP LEARNING DỰ ĐOÁN Ô NHIỄM KHÔNG KHÍ DỰA TRÊN CNN VÀ BI-LSTM

ĐỒ ÁN NGÀNH
NGÀNH CÔNG NGHỆ THÔNG TIN

TP. HỒ CHÍ MINH, 2025

BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC MỞ THÀNH PHỐ HỒ CHÍ MINH

TRƯƠNG THÁI BẢO

XÂY DỰNG MÔ HÌNH DEEP LEARNING DỰ ĐOÁN Ô NHIỄM KHÔNG KHÍ DỰA TRÊN CNN VÀ BI-LSTM

Mã số sinh viên: 2251050008

ĐỒ ÁN NGÀNH
NGÀNH CÔNG NGHỆ THÔNG TIN

Giảng viên hướng dẫn: Th.S DƯƠNG THÁI BẢO

TP. HỒ CHÍ MINH, 2025

LỜI CẢM ƠN

Lời đầu tiên, em xin gửi lời tri ân sâu sắc đến quý thầy, cô Khoa Công nghệ Thông tin, Trường Đại học Mở Thành phố Hồ Chí Minh. Thông qua các môn học, thầy cô đã truyền đạt cho em những kiến thức quý báu, là nền tảng để em có thể thực hiện và hoàn thành đồ án này.

Trong quá trình thực hiện, em đã nhận được nhiều sự giúp đỡ, động viên và góp ý từ thầy cô cũng như các bạn xung quanh. Đặc biệt, em xin bày tỏ lòng biết ơn chân thành đến thầy Dương Thái Bảo - giảng viên hướng dẫn, đã tận tình theo dõi, định hướng và góp ý cho em trong suốt quá trình thực hiện, giúp em hoàn thành tốt nhiệm vụ được giao.

Do kiến thức chuyên ngành khai phá dữ liệu và trí tuệ nhân tạo còn rộng lớn, chắc chắn trong quá trình làm không tránh khỏi thiếu sót. Em rất mong tiếp tục nhận được sự góp ý của quý thầy cô để đồ án được hoàn thiện hơn, đồng thời giúp em tích lũy thêm kiến thức và kinh nghiệm cho những dự án sau này.

Em xin chân thành cảm ơn quý thầy cô!

NHẬN XÉT CỦA GIẢNG VIÊN

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

TÓM TẮT ĐỒ ÁN NGÀNH

Trong bối cảnh ô nhiễm không khí ngày càng nghiêm trọng tại các đô thị lớn, việc dự báo chính xác nồng độ bụi mịn PM2.5 được xem là yếu tố quan trọng trong công tác bảo vệ sức khỏe cộng đồng. Đồ án này dựa vào kiến trúc của bài báo ***A Hybrid Deep Learning Approach for PM2.5 Concentration Prediction in Smart Environmental Monitoring*** được đăng trên tạp chí **Tech Science Press(https://www.techscience.com/iasc/v36n3/51909),** tập trung vào việc xây dựng mô hình deep learning lai, kết hợp giữa Convolutional Neural Network (CNN) và Bidirectional Long Short-Term Memory (Bi-LSTM), nhằm dự báo nồng độ PM2.5 tại Thành phố Hồ Chí Minh. Nhờ vào đó cải tiến các tham số và nâng độ chính xác của mô hình hơn nữa.

Bộ dữ liệu Air Quality HCMC dataset được sử dụng, bao gồm các biến đo lường như nhiệt độ, độ ẩm, tốc độ gió, áp suất, sương và PM2.5. Dữ liệu được tiền xử lý nhằm loại bỏ giá trị khuyết và chuẩn hóa bằng Min-Max Scaler để tránh khác biệt về thang đo giữa các biến.

Mô hình được thiết kế với hai giai đoạn chính: CNN để trích xuất đặc trưng và Bi-LSTM để khai thác quan hệ thời gian hai chiều. Sau cùng, một fully connected layer được sử dụng để dự báo giá trị PM2.5.

Kết quả thực nghiệm được đánh giá bằng các thước đo MSE, RMSE, MAE và MAPE, cho thấy mô hình đề xuất vượt trội hơn các mô hình so sánh như LSTM, Bi-LSTM, CNN, CNN-LSTM và cả PM25-CBL trên cùng tập dữ liệu.

Đồ án góp phần cung cấp một giải pháp hiệu quả trong việc dự báo và cảnh báo mức ô nhiễm không khí, hỗ trợ đề xuất các chiến lược ngắn hạn bảo vệ sức khỏe người dân và kế hoạch dài hạn trong quản lý môi trường.

MỤC LỤC

DANH MỤC HÌNH ẢNH

DANH MỤC BẢNG BIỂU

DANH MỤC TỪ VIẾT TẮT

Từ viết tắt	Giải thích
HCMC	Thành phố Hồ Chí Minh
PM2.5	Nồng độ bụi mịn 2,5
AQI	Chỉ số chất lượng không khí
WHO	Tổ chức Y tế Thế giới

Chương 1. GIỚI THIỆU

1.1. Giới thiệu đề tài

Trong những năm gần đây, hai thành phố lớn của Việt Nam là Hà Nội và Thành phố Hồ Chí Minh luôn lọt "top" những thành phố ô nhiễm không khí nhất thế giới. Thậm chí có thời điểm, chỉ số chất lượng không khí (AQI) của Hà Nội là 246, xếp hạng 1 trong số các thành phố ô nhiễm nhất thế giới. Trong khi đó, Thành phố Hồ Chí Minh cũng không kém cạnh khi xếp thứ 3 với chỉ số AQI là 193, đó là một mức độ "không lành mạnh" cho sức khỏe con người [1]. Điều này cho thấy vấn đề ô nhiễm không khí tại các thành phố lớn là điều đáng cảnh báo. Theo TS. Angela Pratt, Trưởng Đại diện Tổ chức Y tế Thế giới (WHO) tại Việt Nam, mỗi năm Việt Nam có ít nhất 70.000 người tử vong vì ô nhiễm không khí, trung bình cứ mỗi 7,5 phút lại có một người Việt Nam tử vong vì một căn bệnh do tiếp xúc với không khí bị ô nhiễm [2].

Trong các tác nhân gây hại, PM2.5 được xem là một trong những nguyên nhân cốt lõi để cảnh báo và hoạch định chính sách do khả năng xâm nhập sâu vào hệ hô hấp [3]. Từ bối cảnh đó, nghiên cứu này tập trung vào dự báo nồng độ PM2.5 tại Thành phố Hồ Chí Minh bằng cách khai thác mô hình học sâu lai (kết hợp khả năng trích đặc trưng cục bộ của CNN với năng lực nắm bắt phụ thuộc chuỗi hai chiều của Bi-LSTM). Mục tiêu là nâng độ chính xác dự báo ngắn hạn/một bước trước, làm đầu vào cho cơ chế cảnh báo sớm và tham chiếu chính sách địa phương. [4]

1.2. Nội dung thực hiện

Trong đồ án này, nội dung thực hiện tập trung vào việc xây dựng một mô hình học sâu lai cải tiến nhằm dự báo nồng độ bụi mịn PM2.5 tại Thành phố Hồ Chí Minh dựa vào bài báo [4]. Trước hết, tiến hành nghiên cứu cơ sở lý thuyết về chất lượng không khí, các yếu tố ảnh hưởng đến chỉ số ô nhiễm và các mô hình dự báo hiện có. Tiếp theo, dữ liệu từ bộ Air Quality HCMC 2020 [5] được thu thập, phân tích, xử lý và trực quan hóa nhằm làm rõ đặc điểm và mối quan hệ giữa các biến. Dựa trên nền tảng đó, thiết kế và huấn luyện mô hình dự báo kết hợp CNN và Bi-LSTM. Mô hình sau đó được so sánh với các phương pháp truyền thống và các biến thể học sâu khác để đánh giá hiệu quả.

Cuối cùng, tiến hành phân tích kết quả, rút ra nhận xét và đề xuất khả năng ứng dụng mô hình trong hệ thống cảnh báo môi trường. Trong quá trình thực hiện, có tham khảo cấu trúc và định hướng trình bày từ một số bài báo khoa học [4]. Tuy nhiên, nội dung được xây dựng lại theo đặc thù của đề tài và được diễn đạt bằng ngôn ngữ riêng, đảm bảo tính độc lập và phù hợp với mục tiêu môn học. Các ý tưởng, mô hình hoặc hình ảnh được sử dụng từ nguồn khác đều có trích dẫn và chú thích rõ ràng.

1.3. Giới hạn đề tài

Đề tài được giới hạn trong phạm vi dự báo nồng độ bụi mịn PM2.5 tại Thành phố Hồ Chí Minh, dựa trên dữ liệu thu thập từ các trạm quan trắc không khí trong năm 2020. Các chất ô nhiễm khác như PM10, SO2, NO2, CO hay O3 không nằm trong phạm vi phân tích chi tiết của nghiên cứu này. Ngoài ra, đề tài chỉ dừng lại ở việc xây dựng và đánh giá mô hình dự báo trong môi trường thử nghiệm, chưa triển khai thành ứng dụng thực tế hoặc hệ thống giám sát ngoài đời sống.

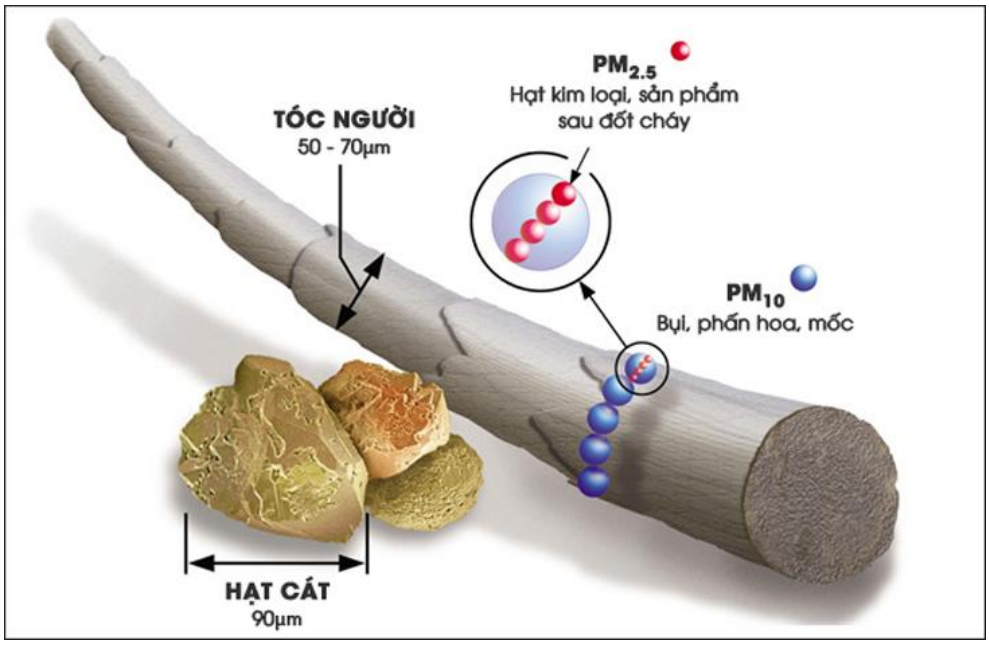
1.4. Bố cục báo cáo

Báo cáo được cấu trúc theo 4 chương chính. Chương 1 chủ yếu chỉ giới thiệu sơ lược về đề tài, nội dung thực hiện, giới hạn đề tài và bố cục của bài báo. Tiếp theo, chương 2 tổng hợp cơ sở lý thuyết liên quan đến ô nhiễm không khí, các công trình nghiên cứu về PM2.5 với phương pháp học máy, học sâu bằng mô hình dự báo chuỗi thời gian nghiên cứu trước đây. Chương 3 mô tả chi tiết phương pháp nghiên cứu, bao gồm dữ liệu sử dụng, quy trình xử lý và các mô hình được áp dụng, quá trình thực nghiệm, kết quả thu được, so sánh hiệu suất giữa các mô hình. Cuối cùng, chương 4 và đưa ra kết luận tổng quát.

Chương 2. CƠ SỞ LÝ THUYẾT

2.1. Các khái niệm cơ bản

2.1.1. Bụi mịn PM2.5



Hình 1. *Kích thước tượng trưng của một số loại bụi so với tóc người và hạt cát. [6]*

PM2.5 là tên gọi của nhóm hạt bụi siêu nhỏ trong không khí, có kích thước không vượt quá 2,5 micromet. Với kích thước chỉ bằng một phần rất nhỏ so với đường kính sợi tóc, các hạt bụi này có thể lơ lửng trong khí quyển trong thời gian dài và dễ dàng đi sâu vào phổi khi con người hít thở. Đặc điểm này khiến PM2.5 trở thành một trong những thành phần gây ô nhiễm khó xử lý và có nhiều tác động tiêu cực. Nguồn hình thành PM2.5 có thể đến từ các hiện tượng tự nhiên như cháy rừng, núi lửa hay gió bụi. Tuy vậy, phần lớn lượng bụi mịn xuất phát từ hoạt động của con người, đặc biệt là khí thải từ phương tiện giao thông, công nghiệp, xây dựng và việc đốt nhiên liệu hóa thạch. [7]

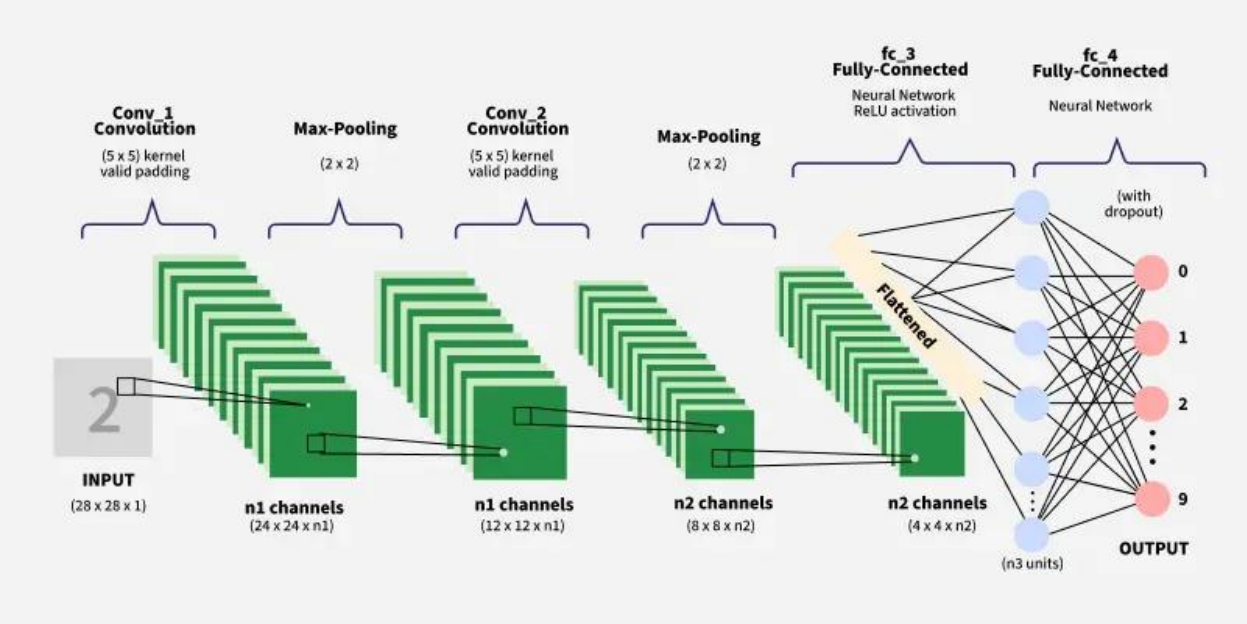
Không chỉ ảnh hưởng đến tầm nhìn và gây hiện tượng sương mù quang hóa, PM2.5 còn được ghi nhận có liên hệ chặt chẽ với các vấn đề sức khỏe nghiêm trọng. Nhiều nghiên cứu chỉ ra rằng tiếp xúc thường xuyên với PM2.5 làm gia tăng nguy cơ mắc các bệnh về hô hấp, tim mạch và ung thư phổi. PM2.5 là một trong những nguyên

nhân ô nhiễm không khí phổ biến và nguy hiểm nhất hiện nay, đòi hỏi cần có biện pháp giám sát và quản lý hiệu quả. [7]

2.1.2. Mô hình Convolutional Neural Network (CNN)

Mạng nơ-ron tích chập (Convolutional Neural Network – CNN) là một loại mô hình học sâu được thiết kế để xử lý dữ liệu có cấu trúc dạng lưới, chẳng hạn như hình ảnh, âm thanh hoặc chuỗi thời gian. CNN sử dụng các phép tích chập (convolution) thông qua các bộ lọc di chuyển trên dữ liệu đầu vào nhằm phát hiện những đặc trưng cục bộ quan trọng. Nhờ cơ chế này, CNN có khả năng học các mẫu đặc trưng từ đơn giản (như cạnh, góc) đến phức tạp (như hình dạng hoặc xu hướng biến đổi theo thời gian).

Một mô hình CNN thường bao gồm ba thành phần chính: các lớp tích chập (convolution) để trích xuất đặc trưng, các lớp pooling (down-sampling) để giảm số chiều và số lượng tham số, và các lớp fully connected để đưa ra kết quả dự báo hoặc phân loại. Ngoài ra, các hàm kích hoạt phi tuyến như ReLU được sử dụng để tăng khả năng biểu diễn của mạng.



Hình 2. *Sơ đồ biểu diễn của mạng nơ-ron tích chập với hai lớp ẩn (nguồn: [8])*

So với mạng nơ-ron truyền thống, CNN có lợi thế ở việc giảm đáng kể số lượng tham số nhờ chia sẻ trọng số, đồng thời khai thác hiệu quả tính cục bộ của dữ liệu. Chính

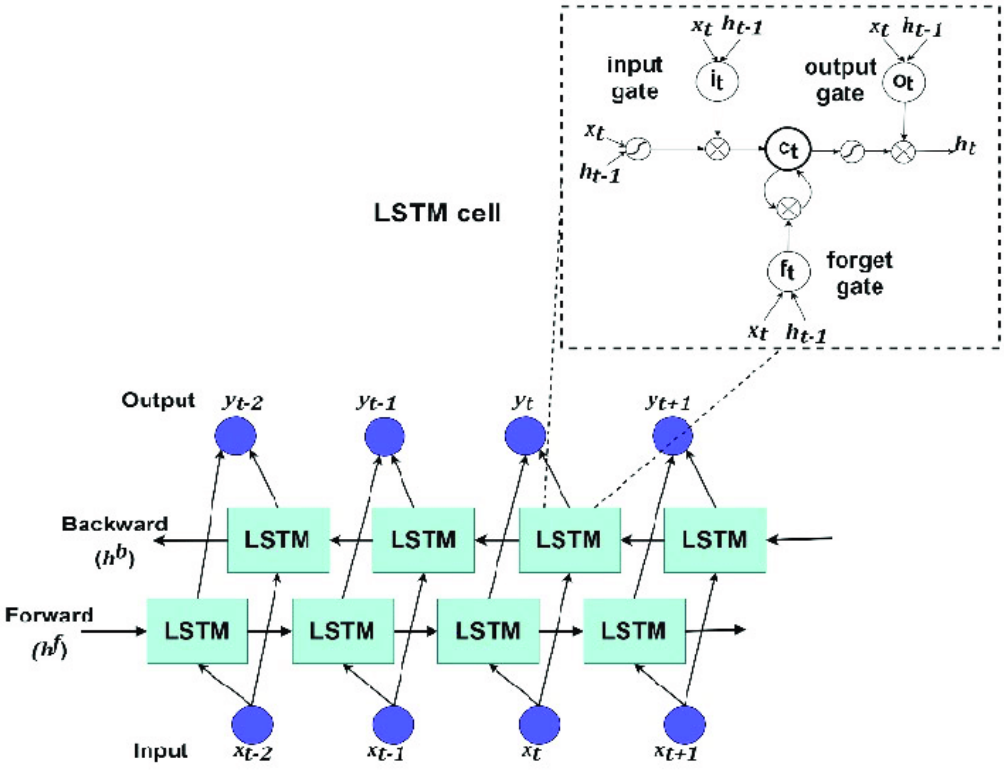
vì vậy, CNN đã trở thành một trong những công cụ chủ đạo trong thị giác máy tính, xử lý ngôn ngữ tự nhiên và ngày càng được ứng dụng nhiều trong phân tích chuỗi thời gian, đặc biệt là dự báo chất lượng không khí và nồng độ PM2.5.

Bên cạnh những ưu điểm đó, CNN tồn tại một nhược điểm chí mạng đó là chỉ học được mối quan hệ cục bộ ngắn hạn, khó nắm bắt phụ thuộc dài hạn theo thời gian. Điều này sẽ làm giảm đi độ chính xác, nên ta sẽ kết hợp nó với LSTM do LSTM có thể học được quan hệ dài hạn.

2.1.3. Mô hình Long Short-Term Memory (LSTM) và Bidirectional LSTM (Bi-LSTM)

Mạng bộ nhớ ngắn hạn dài (Long Short-Term Memory – LSTM) là một biến thể cải tiến của mạng nơ-ron hồi tiếp (Recurrent Neural Network – RNN), được Hochreiter và Schmidhuber (1997) [9] đề xuất nhằm giải quyết vấn đề mất dần hoặc bùng nổ gradient trong RNN truyền thống. LSTM sử dụng các "cổng" (gates) đặc biệt để kiểm soát việc lưu giữ, quên hoặc bổ sung thông tin trong chuỗi, bao gồm cổng quên, cổng đầu vào và cổng đầu ra. Nhờ cấu trúc này, LSTM có khả năng học được mối quan hệ phụ thuộc dài hạn trong dữ liệu chuỗi, vốn rất khó đối với RNN thông thường.

Bidirectional LSTM (Bi-LSTM) là sự mở rộng của LSTM, trong đó dữ liệu được xử lý đồng thời theo cả hai chiều: từ quá khứ đến hiện tại (forward) và từ tương lai quay về quá khứ (backward). Cách tiếp cận này giúp mạng khai thác đầy đủ ngữ cảnh của chuỗi, nhờ đó cải thiện đáng kể độ chính xác trong các bài toán dự báo. Trong lĩnh vực dự báo chất lượng không khí, Bi-LSTM đặc biệt hữu ích vì nồng độ bụi mịn PM2.5 thường chịu ảnh hưởng từ cả xu hướng trước đó và các biến động kế tiếp của chuỗi thời gian. [10]



Hình 3. *Kiến trúc lớp Bidirectional LSTM (nguồn: [11])*

2.2. Các công trình nghiên cứu liên quan

2.2.1. A deep learning-based hybrid method for PM2.5 prediction in central and western China

Mục tiêu nghiên cứu là cải thiện độ chính xác dự báo PM2.5 thông qua một mô hình lai học sâu mới, tích hợp mạng LSTM và Transformer với thuật toán PSO để tối ưu tham số. Dữ liệu đầu vào gồm quan trắc PM2.5 cùng nồng độ các chất ô nhiễm liên quan (CO, NO2, O3, PM10, SO2) và nhiệt độ không khí tại hai thành phố Nanchang và Wuhan giai đoạn 2018–2022, thu thập từ Viện Khoa học Địa lý và Tài nguyên thuộc Viện Hàn lâm Khoa học Trung Quốc. Kết quả cho thấy mô hình đề xuất dự báo PM2.5 chính xác hơn so với LSTM và PSO-LSTM, thể hiện R^2 cao hơn cùng các sai số MAE, MBE, RMSE, MAPE thấp hơn. Đồng thời, mô hình duy trì hiệu suất ổn định trên nhiều thành phố và giai đoạn. [12]

2.2.2. PM2.5 forecasting for an urban area based on deep learning and decomposition method

Nghiên cứu này đề xuất một mô hình lai kết hợp Phân rã EEMD (Ensemble Empirical Mode Decomposition) và LSTM để dự báo nồng độ PM2.5 hàng giờ tại các trạm quan trắc không khí đô thị ở Malaysia. Dữ liệu đầu vào gồm các thông số ô nhiễm như PM2.5, PM10, SO_2, NO_2, O_3, CO từ hai trạm ở Kuala Lumpur (Cheras và Batu Muda) trong khoảng thời gian từ 2018 đến 2019. Trước khi đưa vào mô hình, chuỗi thời gian được phân rã thành các thành phần nhỏ hơn (IMFs và phần dư) bằng phương pháp EEMD; sau đó mỗi phân đoạn được dự báo riêng với mạng LSTM xếp chồng. Các kết quả dự báo sau đó được tổng hợp để tạo ra giá trị PM2.5 chung. Mô hình EEMD-LSTM vượt trội so với các mô hình đơn lẻ (LSTM, Bi-LSTM, GRU, CNN-LSTM, EMD-LSTM, etc.) về các chỉ số sai số như RMSE, MAE, MAPE và hệ số xác định (R^2). Cụ thể, ở trạm Cheras có RMSE ~4.2083 $\mu g/m^3$, MAE ~2.8190 $\mu g/m^3$; ở trạm Batu Muda RMSE khoảng 4.8949 $\mu g/m^3$, MAE ~2.7724 $\mu g/m^3$. [13]

2.2.3. A hybrid model for enhanced forecasting of PM2.5 spatiotemporal concentrations with high resolution and accuracy

Bài nghiên cứu của Feng và cộng sự (2024) trình bày một mô hình lai nhằm nâng cao khả năng dự báo ngắn hạn nồng độ PM2.5 bằng cách kết hợp mô hình vận chuyển hóa học khí quyển (atmospheric chemistry transport model) với thuật toán học sâu để khai thác thông tin môi trường và khí tượng. Mô hình được thử nghiệm với dữ liệu quan trắc không khí tại vùng trung tâm và miền tây Trung Quốc, và cho kết quả chính là độ chính xác dự báo được cải thiện rõ rệt so với các phương pháp thuần học sâu hoặc mô hình hóa khí tượng truyền thống. Kết quả đo bằng các chỉ số sai số (như RMSE, MAE) và hệ số tương quan cho thấy mô hình lai này có hiệu suất tốt hơn đáng kể. Mô hình cũng chứng tỏ tính ổn định khi áp dụng cho nhiều khu vực có điều kiện khí hậu và địa hình khác nhau trong vùng nghiên cứu. [14]

2.2.4. Advancing Air Quality Monitoring: Deep Learning-Based CNN-RNN Hybrid Model for PM2.5 Forecasting

Nghiên cứu này đề xuất một mô hình lai kết hợp **CNN** và **RNN** để dự báo nồng độ PM2.5 giờ tiếp theo tại ba thành phố: India, Milan và Frankfurt. Dữ liệu sử dụng bao

gồm các quan trắc giờ PM2.5 từ ba bộ dữ liệu thực tế lấy từ India, Milan và Frankfurt (sensor / IoT/open-source) với các bước xử lý như loại bỏ giá trị bất thường, khuyết dữ liệu, chuẩn hóa. Mô hình CNN được dùng để trích đặc trưng cục bộ từ chuỗi thời gian, sau đó RNN tiếp nhận để dự báo tiếp theo. Kết quả so sánh cho thấy mô hình CNN-RNN vượt các phương pháp khác bao gồm các mạng học máy truyền thống (như K-NN, RF, SVM, DT, LR), các mạng học sâu đơn lẻ (CNN, RNN, LSTM), về các chỉ số sai số (RMSE, MAE) và hệ số xác định (R^2); đặc biệt ở bộ dữ liệu của India và Frankfurt đạt R^2 rất cao. [15]

2.2.5. Air-pollution prediction in smart city, deep learning approach

Nghiên cứu của Abdellatif Bekkar và cộng sự (2021) nhằm phát triển một giải pháp học sâu để dự báo PM2.5 hàng giờ tại Bắc Kinh, Trung Quốc, sử dụng mô hình lai CNN-LSTM với thông tin không gian-thời gian (spatio-temporal features). Dữ liệu đầu vào gồm nồng độ các chất ô nhiễm (PM2.5 cùng các pollutant khác), dữ liệu khí tượng, và dữ liệu từ các trạm lân cận. Các mô hình được so sánh bao gồm LSTM, Bi-LSTM, GRU, Bi-GRU, CNN, và CNN-LSTM lai đa biến. Kết quả thực nghiệm cho thấy mô hình đa biến hybrid CNN-LSTM cho kết quả tốt hơn tất cả các mô hình còn lại về khả năng dự báo, với sai số thấp hơn và độ chính xác cao hơn. [16]

Chương 3. XÂY DỰNG MÔ HÌNH DEEP LEARNING DỰ ĐOÁN Ô NHIỄM KHÔNG KHÍ DỰA TRÊN CNN VÀ BI-LSTM

3.1. Tổng quan phương pháp hiện thực

Trong nghiên cứu này, phương pháp được áp dụng là dựa trên mô hình PM25-CBL đã được nghiên cứu trong bài báo [4], một mô hình lai kết hợp giữa mạng tích chập (CNN) và mạng bộ nhớ ngắn hạn hai chiều (Bi-LSTM). CNN đảm nhận vai trò trích xuất đặc trưng cục bộ từ dữ liệu chuỗi thời gian nhiều biến (temperature, dew, humidity, pressure, wind speed), trong khi Bi-LSTM giúp học các quan hệ phụ thuộc dài hạn theo cả hai chiều thời gian. Sau đó, đầu ra được đưa vào lớp Fully Connected để dự báo giá trị PM2.5. Cách tiếp cận này tận dụng được thế mạnh của cả hai kiến trúc: CNN xử lý

đặc trưng cục bộ nhanh và hiệu quả, Bi-LSTM khai thác quan hệ theo trình tự, từ đó tăng độ chính xác so với các mô hình đơn lẻ như LSTM, CNN. Đề tài này được lấy ý tưởng từ bài báo [4] và dựa vào đó cải tiến mô hình hơn nữa với ít tham số hơn, giúp giảm thời gian học nhưng đồng thời cũng tăng được độ chính xác .

3.2. Công cụ hiện thực

Quá trình hiện thực mô hình được thực hiện bằng Python 3.x trên môi trường Google Colab.

Các thư viện chính bao gồm:

- **NumPy, Pandas**: xử lý và thao tác dữ liệu.
- **Matplotlib, Seaborn**: trực quan hóa dữ liệu và kết quả.
- **Scikit-learn**: chuẩn hóa dữ liệu, chia tập train/test và tính toán các chỉ số đánh giá.
- **TensorFlow/Keras**: xây dựng và huấn luyện mô hình CNN, Bi-LSTM và Fully Connected.
- Các thư viện liên quan khác

3.3. Tập dữ liệu

Dữ liệu sử dụng là **Air Quality HCMC dataset** được cung cấp bởi Open Development Mekong [5]. Bộ dữ liệu gồm 7 biến: ngày (*Date*) được dùng làm index trong bộ dữ liệu , nhiệt độ (*Temperature*), độ ẩm (*Humidity*), tốc độ gió (*Wind Speed*), *PM2.5* (giá trị mục tiêu), sương (*Dew*) và áp suất (*Pressure*). Trong đó, *PM2.5* là biến cần dự báo, các biến còn lại được dùng làm đầu vào.

Bảng 1. *Các biến của tập dữ liệu Air Quality HCMC*

#No	Biến	Mô tả	Min	Max	Median	Đơn vị
1	Date	Ngày	30-12-2019	20-01-2021	10-07-2020	Ngày
2	Temperature	Nhiệt độ trung bình	23	31	27.5	ºC
3	Humidity	Độ ẩm trung bình	47	100	76.5	%
4	Wind speed	Tốc độ gió trung bình	0.5	5.4	2.3	m/s
5	Dew	Sương trung bình	14.5	26.5	24	ºC
6	Pressure	Áp suất trung bình	1003	1014	1009	hPa
7	PM25	Nồng độ PM2.5 trung bình	5	171	68	µg/m³

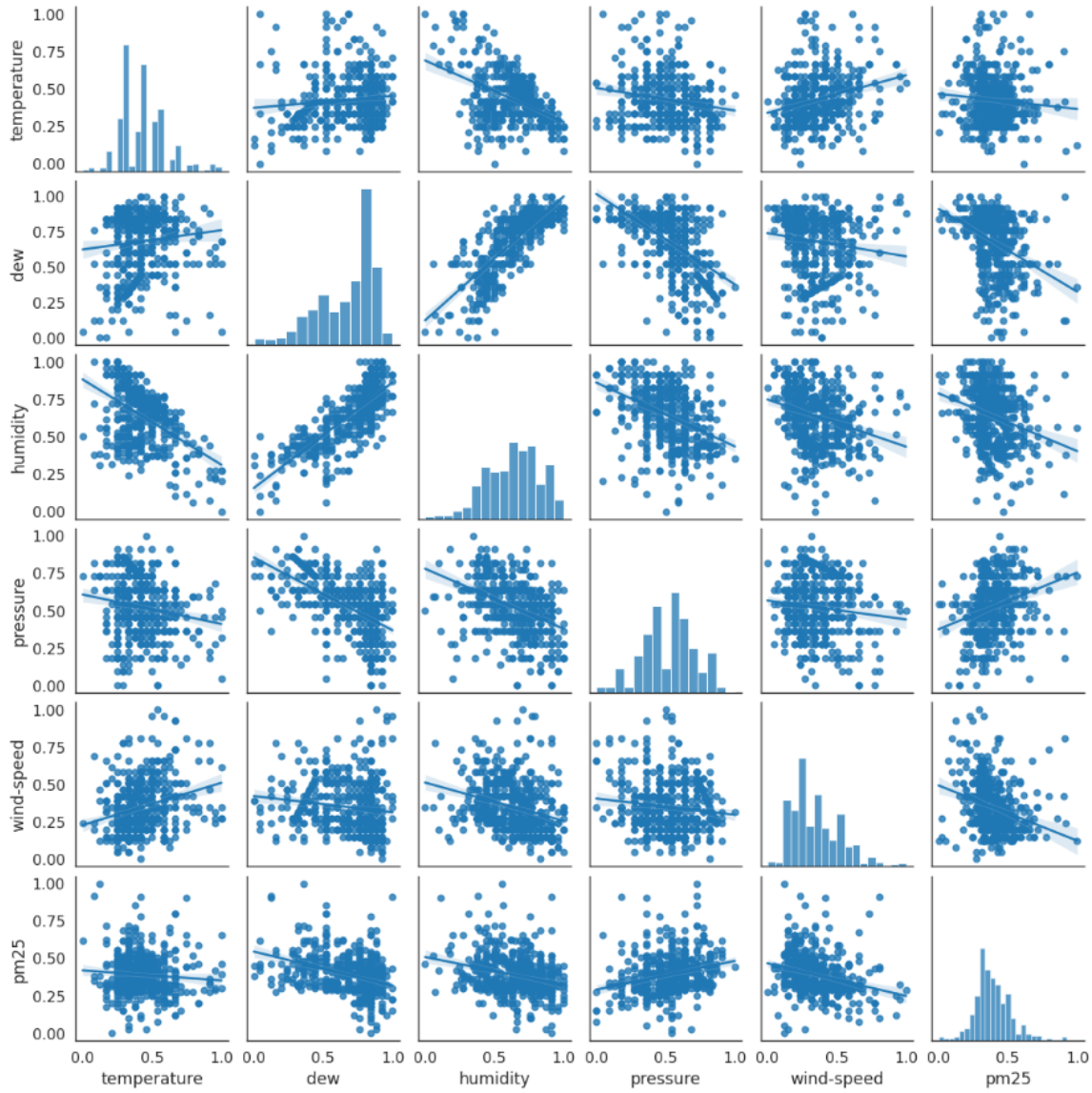
3.4. Tiền xử lý dữ liệu

Bộ dữ liệu **Air Quality HCMC dataset** ghi nhận các chỉ số chất lượng không khí hằng ngày tại một số thành phố của Việt Nam. Tuy nhiên, trong nghiên cứu này, phạm vi được thu hẹp và chỉ lựa chọn dữ liệu từ Thành phố Hồ Chí Minh để phục vụ cho mục tiêu dự báo nồng độ PM2.5.

Nhằm hạn chế sự khác biệt về thang đo giữa các biến đầu vào, sáu biến gồm temperature, humidity, wind speed, pm25, dew và pressure được chuẩn hóa về cùng một khoảng giá trị bằng phương pháp Min–Max Scaler, với công thức tính như sau:

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Mối quan hệ giữa các biến temperature, humidity, wind speed, dew và pressure với nồng độ PM2.5 trong **Air Quality HCMC dataset** được minh họa trong ***Hình 4***. Quan sát từ cột cuối của ma trận phân tán cho thấy temperature, dew, humidity và wind speed có xu hướng tương quan âm nhẹ với nồng độ PM2.5, trong khi pressure thể hiện mối tương quan dương nhẹ. Vì vậy, nghiên cứu này lựa chọn sử dụng toàn bộ các biến temperature, dew, humidity, wind speed và pressure trong mô hình đề xuất.



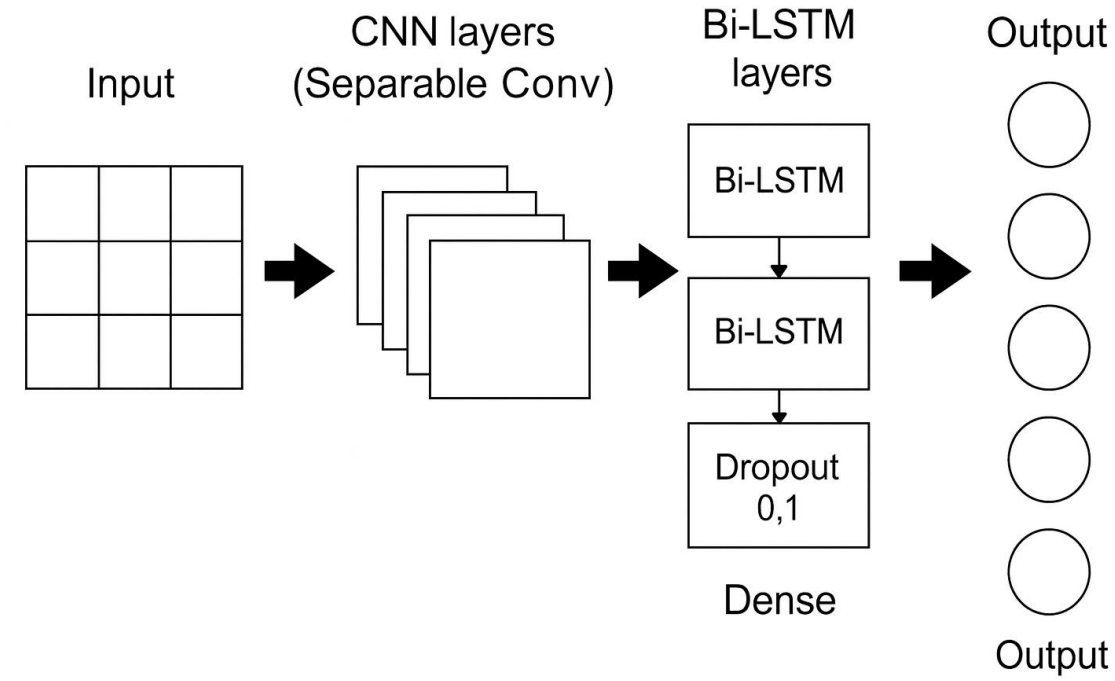
Hình 4. *Ma trận phân tán của các biến với biến pm2.5*

3.5. Xây dựng mô hình học sâu lai CNN–Bi-LSTM cải tiến tối giản tham số cho dự báo PM2.5

Trong phạm vi đồ án, mô hình đề xuất được ký hiệu là **PM25-CBLo** (*PM2.5–CNN–BiLSTM optimized*). Ký hiệu "o" trong CBLo thể hiện đây là phiên bản **tối ưu hóa và tối giản tham số** so với mô hình gốc PM25-CBL trong bài báo tham khảo [4], đồng thời vẫn đạt được độ chính xác cao hơn. Kiến trúc của mô hình học sâu lai PM25-CBLo được xây dựng nhằm dự đoán nồng độ PM2.5 trong bộ dữ liệu **Air**

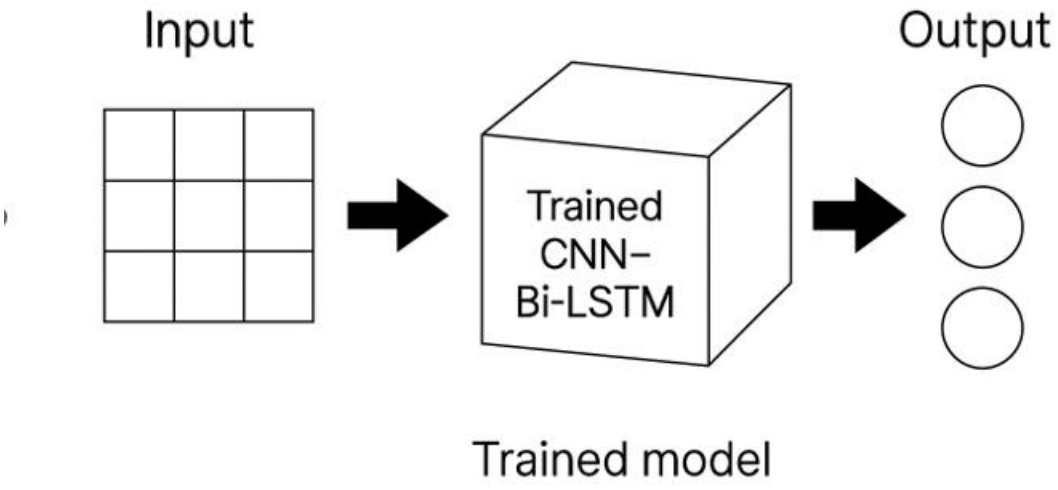
Quality HCMC . Mô hình này, được minh họa ở ***Hình 5***, là sự kết hợp giữa CNN (Separable Convolution) và Bi-LSTM, nghiên cứu giải quyết dự báo một bước trước (one-step ahead) từ dữ liệu t−t suy ra PM2.5 ở thời điểm t để giải quyết bài toán dự báo chuỗi thời gian.

Mô hình hoạt động qua hai giai đoạn. Ở giai đoạn huấn luyện, sáu biến đầu vào từ bộ dữ liệu **Air Quality HCMC** ở thời điểm t-1 gồm $PM2.5_{t-1}, temperature_{t-1}, humidity_{t-1}, wind\ speed_{t-1}, dew_{t-1}$ và $pressure_{t-1}$ được đưa vào khối CNN để trích xuất đặc trưng. Các đặc trưng thu được sau đó được chuyển tiếp đến khối Bi-LSTM. Cuối cùng, chỉ có một lớp kết nối đầy đủ (Fully Connected) (Dense(1)) mà không có thêm các lớp ẩn khác, đảm nhiệm việc dự đoán nồng độ PM2.5. Đó cũng là sự khác biệt so với mô hình PM25-CBL.



Hình 5. *Kiến trúc mô hình PM25-CBLo giai đoạn huấn luyện*

Trong giai đoạn kiểm thử, sử dụng đúng sáu biến gồm temperature, humidity, wind speed, dew và pressure,pm25 ở thời điểm t-1 đưa vào mô hình đã huấn luyện để dự báo giá trị PM2.5 tại thời điểm t.



Hình 6. *Kiến trúc mô hình PM25-CBLo giai đoạn kiểm thử*

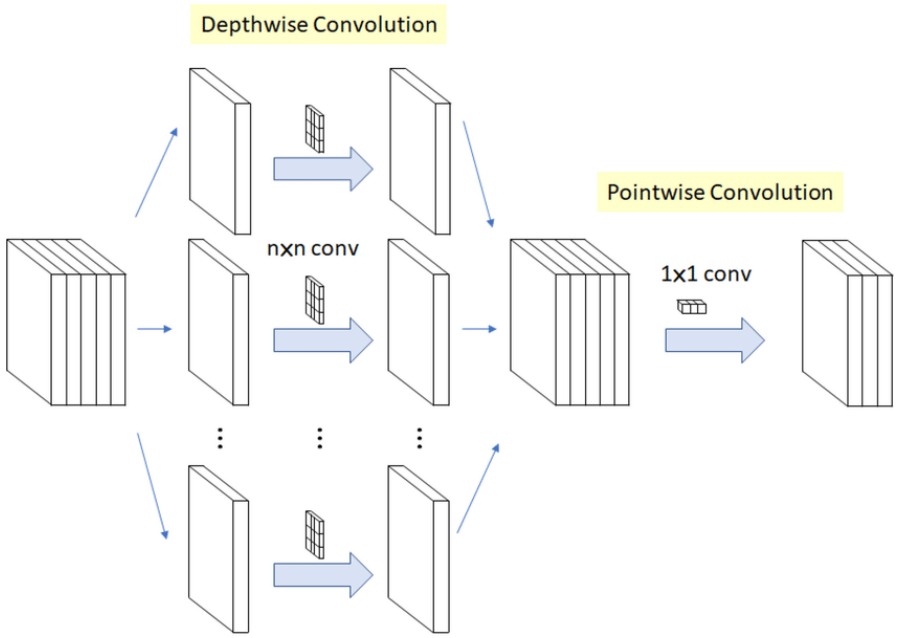
3.5.1. Mô hình CNN (Depthwise Separable Convolution)

Trong các mô hình học sâu truyền thống, lớp tích chập (convolution) thường thực hiện phép lọc trên toàn bộ các kênh đầu vào cùng lúc. Cách tiếp cận này giúp mô hình học được nhiều đặc trưng phức tạp, nhưng lại kéo theo số lượng tham số rất lớn và chi phí tính toán cao. Để khắc phục hạn chế này, kỹ thuật **Depthwise Separable Convolution (SeparableConv)** được giới thiệu nhằm tách riêng quá trình tích chập thành hai bước độc lập: Depthwise Convolution và Pointwise Convolution.

Depthwise Convolution: thay vì áp dụng một bộ lọc trên tất cả các kênh, mỗi kênh đầu vào sẽ được xử lý bởi một bộ lọc riêng biệt. Điều này cho phép trích xuất đặc trưng cục bộ của từng kênh một cách hiệu quả, đồng thời làm giảm đáng kể số phép toán cần thực hiện.

Pointwise Convolution (1x1 Convolution): sau khi các đặc trưng được tách riêng ở bước depthwise, lớp pointwise thực hiện tích chập với kernel kích thước 1x1 để kết hợp thông tin giữa các kênh. Nhờ vậy, mô hình vẫn có thể học được mối quan

hệ phức tạp giữa các kênh mà không cần sử dụng một lượng lớn tham số như convolution truyền thống.



Hình 7. *Kiến trúc Depthwise Separable Convolution (nguồn: [17])*

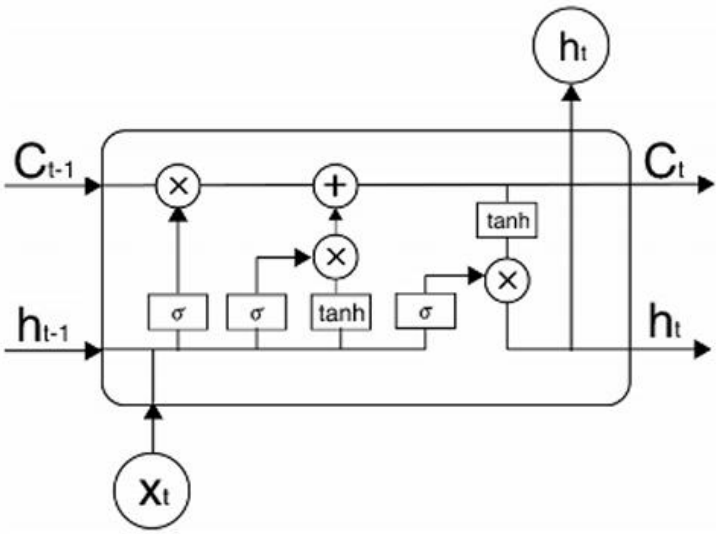
Lợi ích của Separable Convolution so với Convolution thông thường

1. **Giảm số tham số và chi phí tính toán**: So với convolution chuẩn có số tham số tỉ lệ theo k × k × M × N, Depthwise Separable Convolution chỉ cần k × k × M + M × N, giúp giảm hàng chục đến hàng trăm lần số lượng tham số khi mô hình có nhiều kênh đầu vào và đầu ra.
2. **Tăng tốc độ huấn luyện và suy luận**: Nhờ giảm mạnh số phép nhân–cộng, mô hình sử dụng SeparableConv có thể được huấn luyện nhanh hơn và triển khai dự báo với độ trễ thấp hơn.
3. **Hạn chế overfitting**: Với ít tham số hơn, mô hình gọn nhẹ hơn và giảm nguy cơ khớp quá mức với dữ liệu huấn luyện. Điều này đặc biệt hữu ích trong các bài toán mà dữ liệu không quá lớn.
4. **Giữ nguyên khả năng học đặc trưng**: Dù tối ưu về tính toán, SeparableConv vẫn đảm bảo mô hình có thể học được cả đặc trưng cục bộ (qua depthwise) và đặc trưng tổng hợp (qua pointwise).

3.5.2. Mô hình Bi-LSTM

Để hiểu rõ về Bi-LSTM, trước tiên cần xem qua mạng LSTM cơ bản. LSTM được Graves và Schmidhuber [18] đề xuất nhằm khắc phục hạn chế của mạng nơ-ron hồi tiếp truyền thống (RNN). Thay vì sử dụng các nơ-ron ẩn như trong RNN, LSTM

áp dụng một tập hợp đặc biệt gọi là "ô nhớ" (memory cells). Các ô nhớ này có cơ chế cổng (gate) để lọc, duy trì và cập nhật trạng thái thông tin. Ba loại cổng chính trong mỗi ô nhớ là: cổng nhập (input gate), cổng quên (forget gate) và cổng xuất (output gate). Bên trong mỗi ô nhớ còn sử dụng hai hàm kích hoạt phi tuyến phổ biến là *sigmoid* và *tanh*.



Hình 8. *Minh họa cấu trúc của một ô nhớ tại thời điểm t [4]*

Trong đó, cổng quên quyết định phần thông tin nào từ trạng thái trước sẽ bị loại bỏ. Cổng này nhận đầu vào là trạng thái ẩn ở bước t−1 (h_{t-1}) và đầu vào hiện tại (x_t). Hai thành phần này được kết hợp thành một vector và xử lý qua hàm sigmoid để tạo ra hệ số quên f_t, được biểu diễn như trong công thức:

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

Trong đó, W_f và b_f lần lượt là ma trận trọng số và vector bias của cổng quên. Hệ số f_t xác định mức độ thông tin từ trạng thái ô trước (C_{t-1}) được giữ lại và chuyển tiếp sang trạng thái hiện tại (C_t). Nhờ cơ chế này, LSTM nói chung và Bi-LSTM nói riêng có khả năng ghi nhớ thông tin dài hạn cũng như khai thác quan hệ theo cả hai chiều thời gian.

Tiếp theo, cổng nhập kiểm soát lượng thông tin mới từ đầu vào x_t được lưu vào ô nhớ. Nó bao gồm hai thành phần: một vector i_t được tính bằng hàm sigmoid để

quyết định vị trí cập nhật, và một vector ứng viên C_t' được tính qua hàm tanh nhằm bổ sung thông tin mới.

Trạng thái ô nhớ được cập nhật bằng công thức:

$$C_t = f_t \cdot C_{t-1} + i_t \cdot C_t'$$

Cổng xuất quyết định phần nào của trạng thái ô nhớ hiện tại sẽ được đưa ra ngoài. Đầu tiên, một vector o_t được tính bằng hàm sigmoid. Sau đó, trạng thái ẩn h_t được tính bằng cách nhân o_t với tanh(C_t), cho phép mô hình tạo ra đầu ra phù hợp tại thời điểm t.

Tuy nhiên, LSTM chỉ xử lý dữ liệu theo một chiều (từ quá khứ đến hiện tại), điều này có thể làm mất đi các mẫu quan trọng. Để khắc phục, Bi-LSTM được phát triển nhằm khai thác cả hai chiều: tiến (forward) và lùi (backward). Cấu trúc này sử dụng hai lớp LSTM song song, một lớp đọc chuỗi dữ liệu từ đầu đến cuối, trong khi lớp còn lại đọc theo chiều ngược lại. Bằng cách này, Bi-LSTM có khả năng nắm bắt thông tin theo cả hai hướng thời gian, từ đó cải thiện khả năng nhận diện và dự báo sự thay đổi nồng độ PM2.5 trong cả quá khứ lẫn tương lai (đã minh họa ở ***Hình 2)***.

3.5.3. Mô hình kết nối hoàn toàn (Fully connected layer)

Trong phiên bản gốc PM25-CBL, đầu ra từ các lớp Bi-LSTM thường được đưa qua một hoặc nhiều lớp ẩn trước khi đi vào lớp kết nối đầy đủ (fully connected layer). Tuy nhiên, ở mô hình cải tiến PM25-CBLo, cấu trúc này đã được tinh giản: thay vì sử dụng thêm các lớp ẩn trung gian, đầu ra từ Bi-LSTM được đưa trực tiếp vào lớp fully connected cuối cùng.

Cách tiếp cận này có hai ưu điểm rõ rệt. Thứ nhất, số lượng tham số cần huấn luyện được giảm xuống đáng kể, giúp mô hình gọn nhẹ hơn và tiết kiệm thời gian tính toán. Thứ hai, việc loại bỏ các tầng ẩn trung gian cũng góp phần hạn chế nguy cơ overfitting, đặc biệt trong bối cảnh dữ liệu huấn luyện không quá lớn.

Mặc dù được đơn giản hóa, lớp fully connected cuối cùng vẫn đảm nhận nhiệm vụ quan trọng: sinh ra giá trị dự đoán nồng độ PM2.5. Nhờ đó, mô hình PM25-CBLo

vừa duy trì được độ chính xác, vừa đạt được sự tối ưu về mặt hiệu quả tính toán. Cấu hình chi tiết của mô hình sẽ được trình bày trong ***Bảng 2.***

3.5.4. Thông số cấu hình cải tiến

Trong mô hình PM25-CBLo, cấu trúc mạng được điều chỉnh để giảm đáng kể số lượng tham số so với mô hình gốc PM25-CBL. Sự thay đổi quan trọng nằm ở việc sử dụng Separable Convolution thay cho convolution thông thường và giảm số lượng neuron trong các lớp Bi-LSTM. Cách tối ưu này không chỉ giúp mô hình trở nên gọn nhẹ hơn, mà còn rút ngắn thời gian huấn luyện và hạn chế hiện tượng overfitting.

Cụ thể, số kênh ở các lớp convolution được giảm từ 64 xuống 32, đồng thời các lớp Bi-LSTM được thiết kế với số lượng neuron ít hơn. Việc tinh giản này cũng kéo theo lớp Flatten có kích thước nhỏ hơn, từ đó giảm thêm số tham số cần học.

Bảng 2. *Thông số cấu hình của mô hình PM25-CBLo*

#No	Layer type	Neurons	Parameters
1	1D Separable Convolution	(None, None, 5, 32)	66
2	1D Max Pooling	(None, None, 5, 32)	0
3	1D Separable Convolution	(None, None, 4, 32)	1120
4	1D Max Pooling	(None, None, 4, 32)	0
5	Flatten	(None, None, 128)	0
6	Bi-LSTM	(None, None, 64)	41,216
7	Dropout	(None, None, 64)	0
8	Bi-LSTM	(None, 64)	24,832
9	Fully connected layer	(None, 1)	65

Tổng tham số = 66 + 1120 + 41216 + 24832 + 65 = 67299 (tham số)

Tổng tham số của mô hình PM25-CBLo chỉ còn 67.299 trong khi ở mô hình gốc PM25-CBL con số này lên đến 214.081 (nhiều hơn ~**3.18 lần)**.

#No	Layer type	Neurons	Parameters
1	1D Convolution	(None, None, 5, 64)	192
2	1D Max Pooling	(None, None, 5, 64)	0
3	1D Convolution	(None, None, 4, 64)	8256
4	1D Max Pooling	(None, None, 4, 64)	0
5	Flatten	(None, None, 256)	0
6	Bi-LSTM	(None, None, 128)	164,352
7	Dropout	(None, 128)	0
8	Bi-LSTM	(None, 64)	41,216
9	Fully connected layer	(None, 1)	65

Hình 9. *Các thông số cấu hình của PM25-CBL trong bài báo [4]*

Sự khác biệt hơn **3 lần** về số lượng tham số cho thấy mô hình cải tiến đã đạt được mức độ tối ưu hóa cao, tiết kiệm tài nguyên tính toán mà vẫn duy trì hiệu quả dự báo. Thậm chí, trong một số thử nghiệm, độ chính xác của mô hình cải tiến còn vượt trội hơn so với phiên bản gốc, minh chứng rằng việc giảm tham số không đồng nghĩa với suy giảm chất lượng dự báo. Đây chính là ưu điểm nổi bật của PM25-CBLo trong việc dự đoán nồng độ PM2.5 trên dữ liệu chuỗi thời gian.

3.6. Phương pháp đánh giá

Trong phần này, nghiên cứu tiến hành đánh giá mô hình đề xuất và so sánh với một số phương pháp tiên tiến trong dự đoán chuỗi thời gian như CNN, LSTM, Bi-LSTM, CNN-LSTM và PM25-CBL trên bộ dữ liệu **Air Quality HCMC.** Để kiểm chứng hiệu quả của mô hình PM25-CBLo, bộ dữ liệu **Air Quality HCMC** được tách theo trục thời gian 80% giai đoạn đầu cho huấn luyện, 20% giai đoạn cuối cho kiểm thử (không xáo trộn). Sử dụng 4 chỉ số đánh giá hiệu suất phổ biến:

- **MSE (Mean Squared Error):** Trung bình bình phương sai số giữa giá trị dự đoán và giá trị thực tế.

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

- **RMSE (Root Mean Squared Error):** Căn bậc hai của MSE, thể hiện mức sai số trung bình tính theo cùng đơn vị với dữ liệu gốc.

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

- **MAE (Mean Absolute Error):** Trung bình giá trị tuyệt đối của sai số, không xét đến hướng sai lệch.

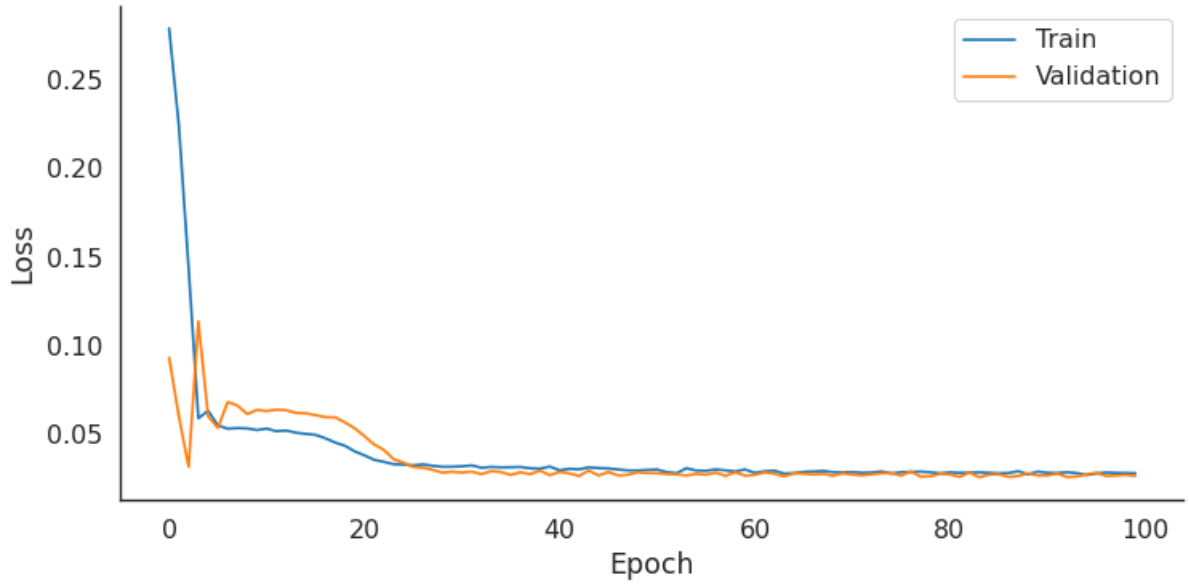
$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y_i}|$$

- **MAPE (Mean Absolute Percentage Error):** Tỉ lệ phần trăm sai số tuyệt đối trung bình, cho biết độ chính xác dự đoán dưới dạng phần trăm.

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y_i}}{y_i}\right|$$

3.7. Kết quả thực nghiệm

Trong quá trình thực nghiệm, tôi đã theo dõi giá trị loss ở cả giai đoạn huấn luyện và kiểm thử (xem ***Hình 10***). Kết quả cho thấy đường loss ở hai giai đoạn gần như ổn định sau khoảng 100 epoch, vì vậy mô hình được ấn định huấn luyện trong 100 epoch. Ngoài ra, nghiên cứu sử dụng batch size = 30 và bộ tối ưu Adam cho mô hình PM25-CBLo, với learning rate khởi tạo là 0,001.



Hình 10. *Model loss*

Tiếp theo, để cho minh chứng cho sự vượt trội hơn so với các mô hình còn lại. Phần này sẽ báo cáo hiệu suất của các phương pháp thực nghiệm trên, bao gồm LSTM, Bi-LSTM, CNN, CNN-LSTM, PM25-CBL và cuối cùng là PM25-CBLo sử dụng bộ dữ liệu **Air Quality HCMC** với các chỉ số MSE, RMSE, MAE, MAPE. Tổng quan, PM25-CBLo đạt kết quả tốt nhất trên cả bốn thước đo, với MSE = 0,907, RMSE = 0,952, MAE = 0,762 và MAPE = 2,909. So với đường cơ sở CNN, mô hình đề xuất giảm MSE ~62,7% (2,433 → 0,907), RMSE ~38,9% (1,560 → 0,952), MAE ~38,3% (1,235 → 0,762) và MAPE ~39,1% (4,772 → 2,909). Điều này cho thấy kiến trúc kết hợp tích chập – chuỗi thời gian hai chiều cùng lớp fully-connected tinh gọn giúp mô hình hóa quan hệ phi tuyến và phụ thuộc theo thời gian tốt hơn. Cũng như với

việc giảm tham số, độ chính xác của PM25-CBLo cải tiến có độ chính xác cao hơn so với PM25-CBL rất nhiều. Việc này cho thấy với cách cấu hình trong bài báo [4] vẫn còn có thể cải tiến tốt hơn nữa.

Xếp hạng tiếp theo là CNN-LSTM, với các chỉ số đứng thứ hai (MSE 1,110; RMSE 1,053; MAE 0,838; MAPE 3,207). PM25-CBL vẫn vượt CNN-LSTM lần lượt khoảng 18,3% (MSE), 9,6% (RMSE), 9,1% (MAE) và 9,3% (MAPE), phản ánh lợi ích của hướng đọc hai chiều và tối ưu số tham số.

Bi-LSTM và LSTM cho kết quả trung bình, kém hơn nhóm kết hợp CNN do thiếu tầng trích chọn đặc trưng không gian cục bộ trước khi đưa vào mô hình chuỗi. CNN thuần cho hiệu năng thấp nhất trên mọi chỉ số, chứng tỏ chỉ dựa vào đặc trưng cục bộ mà bỏ qua phụ thuộc dài hạn theo thời gian là chưa đủ.

Bảng 3. *Kết quả của các phương pháp thực nghiệm*

#No	Mô hình	MSE	RMSE	MAE	MAPE
1	CNN	2.433	1.560	1.235	4.772
2	LSTM	1.218	1.104	0.880	3.384
3	Bi-LSTM	1.295	1.138	0.906	3.486
4	CNN-LSTM	1.110	1.053	0.838	3.207
5	PM25-CBLo	**0.907**	**0.952**	**0.762**	**2.909**
6	PM25-CBL	1.368	1.170	0.944	3.165

Tiếp theo, phần này sẽ đánh giá thời gian xử lí của các phương pháp thực nghiệm với bộ dữ liệu Air Quality HCMC. Kết quả ở ***bảng 4*** cho thấy sự khác biệt đáng kể về thời gian xử lý giữa các mô hình. CNN có tốc độ nhanh nhất trong cả hai giai đoạn, với thời gian huấn luyện chỉ 23.71 và kiểm thử 0.31. Tuy nhiên, hiệu năng dự báo của CNN lại thấp hơn nhiều so với các mô hình khác (như đã phân tích ở ***bảng 3***), cho thấy ưu điểm tốc độ đi kèm với hạn chế về độ chính xác.

Ngược lại, Bi-LSTM yêu cầu thời gian huấn luyện cao nhất (62.86) và kiểm thử chậm (2.62), phản ánh độ phức tạp do sử dụng hai chiều xử lý chuỗi. LSTM đơn thuần cũng có thời gian huấn luyện đáng kể (38.57) và kiểm thử 1.36, cho thấy mô hình chuỗi nhìn chung tốn kém hơn so với CNN.

CNN-LSTM đạt sự cân bằng hơn khi thời gian huấn luyện 30.54 và kiểm thử 1.33, nhanh hơn so với LSTM và Bi-LSTM, đồng thời duy trì độ chính xác cao (xem lại Bảng 3).

Đối với PM25-CBL, thời gian huấn luyện là 35.71 và kiểm thử 2.62, cao hơn so với CNN-LSTM. Điều này xuất phát từ cấu trúc kết hợp nhiều tầng tích chập và Bi-LSTM, khiến mô hình phức tạp hơn trong giai đoạn dự đoán. Tuy nhiên, nhờ độ chính xác vượt trội, sự đánh đổi về thời gian xử lý của PM25-CBL vẫn được xem là chấp nhận được cho các ứng dụng thực tế, đặc biệt khi mục tiêu là dự báo chất lượng không khí chính xác và tin cậy. Và tất nhiên, PM25-CBL có nhiều tham số hơn nên việc học sẽ tốn thời gian hơn mô hình cải tiến rất nhiều.

Bảng 4. *Thời gian training và testing*

#No	Mô hình	Training phase	Testing phase
1	CNN	23.71	0.31
2	LSTM	38.57	1.36
3	Bi-LSTM	62.86	2.62
4	CNN-LSTM	30.54	1.33
5	PM25-CBLo	35.71	2.62
6	PM25-CBL	53.99	2.65

Cuối cùng, đánh giá mô hình PM25-CBLo một lần nữa bằng phương pháp Cross Validation với 5 fold.

Bảng 5. *Kết quả đánh giá bằng phương pháp cross validation*

#No	Thông số	Kết quả
1	MSE	1.46 ± 0.64
2	RMSE	1.18 ± 0.27
3	MAE	0.99 ± 0.28
4	MAPE	$3.53\% \pm 0.93\%$
5	Trung bình thời gian training/fold	11.40(giây)

Kết quả 5-fold cho PM25-CBLo đạt **MSE = 1.46 ± 0.64, RMSE = 1.18 ± 0.27, MAE = 0.99 ± 0.28, MAPE = 3.53% ± 0.93%,** thời gian huấn luyện trung bình ≈ 11.40s/fold. So với đánh giá hold-out 80/20 ở ***Bảng 3*** **(MSE 0.907; RMSE 0.952; MAE 0.762; MAPE 2.909)**, các chỉ số từ cross validation thấp hơn. Điều này là phù hợp với kỳ vọng và có thể giải thích bởi các nguyên nhân sau:

1. Kích thước tập huấn luyện theo fold nhỏ hơn. Ở các fold đầu, mô hình chỉ được học trên một phần nhỏ dữ liệu (20%–40%), nên khả năng khái quát hoá kém hơn so với huấn luyện một lần trên 80% dữ liệu.
2. Độ khó không đồng đều theo thời gian. Mỗi fold bao phủ một đoạn thời gian khác nhau; các đoạn có nhiều biến động đột ngột (spike) hoặc thời tiết bất thường khiến sai số lớn hơn. Trong khi đó, tập 20% dùng cho hold-out có thể "dễ" hơn về mặt dao động.
3. Chuẩn hoá đặc trưng theo từng fold. Khi tái chuẩn hoá (Min-Max) trong từng fold, các thống kê được ước lượng từ tập huấn luyện con thay vì toàn bộ 80% như ở kịch bản hold-out. Điều này làm cho phân phối dữ liệu ở tập kiểm tra của fold chênh lệch hơn → sai số tăng.
4. Cơ chế dừng sớm/siêu tham số cố định. EarlyStopping dừng sớm ở những fold có tập kiểm tra khó, số epoch hiệu dụng nhỏ hơn; đồng thời các siêu tham số (epochs, batch size) không được tối ưu riêng cho từng fold.
5. Cách tổng hợp nghiêm ngặt hơn. Cross validation báo cáo trung bình ± độ lệch chuẩn qua 5 kịch bản độc lập, nên phản ánh thực lực mô hình trong nhiều điều kiện hơn; vì vậy thường khắt khe hơn một lần đo duy nhất trên 20% cuối.

Mặc dù các chỉ số cao hơn so với hold-out, kết quả 5-fold vẫn nằm trong mức sai số tương đối ~3–4%, chứng tỏ PM25-CBLo giữ được độ ổn định và khả năng khái quát trên các phân đoạn thời gian khác nhau. Điều này củng cố kết luận rằng kiến trúc tối ưu (SeparableConv + Bi-LSTM, bỏ các fully-connected ẩn) là lựa chọn phù hợp cho dự báo PM2.5 trong điều kiện tài nguyên tính toán hạn chế.

Chương 4. KẾT LUẬN

4.1. Nhận xét kết quả thực nghiệm

Hiệu năng tổng quát: Mô hình PM25-CBLo đạt các chỉ số lỗi thấp (MSE 0,907; RMSE 0,952; MAE 0,762; MAPE 2,909), ổn định hơn so với các baseline (CNN, LSTM, Bi-LSTM, CNN-LSTM) và cũng nhỉnh hơn cấu hình gốc PM25-CBL. Với mức MAPE xấp xỉ 2–3%, sai số tương đối ở đa số thời điểm nằm trong ngưỡng chấp nhận cho ứng dụng giám sát và cảnh báo.

Tác động của tối ưu kiến trúc: Việc thay Conv1D bằng SeparableConv1D và loại bỏ các fully-connected ẩn giúp số tham số giảm còn ~1/3, kéo theo thời gian huấn luyện và suy luận ngắn hơn. Quan trọng hơn, độ chính xác không suy giảm mà còn cải thiện, cho thấy phần lớn khả năng biểu diễn đến từ khối trích đặc trưng và Bi-LSTM thay vì các tầng dày đặc cuối.

Hành vi sai số: Sai số có xu hướng tăng ở các thời điểm biến động đột ngột (spike) do mô hình cần thời gian để cập nhật xu thế mới; ở các khoảng ổn định, dự báo bám sát giá trị thực. Điều này phù hợp với đặc tính chuỗi thời gian PM2.5 chịu ảnh hưởng tức thời của thời tiết và hoạt động đô thị.

Tính ổn định: Trong các lần lặp huấn luyện, kết quả ít dao động khi giữ nguyên cấu hình và hạt giống khởi tạo, cho thấy quy trình tiền xử lý và cấu trúc mô hình đủ bền vững.

Ý nghĩa vận hành: Tốc độ suy luận nhanh cùng kiến trúc gọn nhẹ giúp mô hình phù hợp chạy trên máy chủ cấu hình trung bình hoặc tích hợp vào pipeline realtime/near-realtime.

4.2. Những hạn chế của đề tài

Nghiên cứu dùng dữ liệu năm 2020 của TP.HCM; sự thay đổi theo mùa/năm hoặc ảnh hưởng các sự kiện bất thường (ví dụ dịch bệnh, cháy rừng) chưa được mô hình hoá tường minh.

Đặc trưng ngoại sinh: mới khai thác các biến khí tượng cơ bản; chưa kết hợp thêm nguồn dữ liệu không gian (trạm lân cận, vệ tinh), hoạt động giao thông/công nghiệp, hay dữ liệu dự báo thời tiết tương lai.

Đánh giá theo nhiều mức mục tiêu: chưa thử nghiệm phân loại/đánh giá theo các ngưỡng PM2.5 (ví dụ chuẩn WHO), hoặc các bài toán multi-step/multi-horizon (dự báo nhiều bước tiếp theo).

4.3. Ý nghĩa thực tiễn

Kết quả cho thấy việc tối ưu kiến trúc là hướng tiếp cận hiệu quả để nâng cao độ chính xác dự báo mà không đánh đổi chi phí tính toán. Mô hình có thể tích hợp vào bảng điều khiển (dashboard) theo thời gian gần-thực, hỗ trợ cảnh báo sớm, lập kế hoạch vận hành đô thị (khuyến cáo sức khoẻ, điều tiết giao thông, lịch tưới rửa đường, v.v.) và làm đầu vào cho các mô hình chính sách môi trường.

4.4. Kết luận chung

Đề tài chứng minh rằng một kiến trúc lai CNN–BiLSTM được tối ưu bằng Separable Convolution có thể vừa nhẹ vừa chính xác trong dự báo PM2.5. Phiên bản PM25-CBLo không chỉ giảm đáng kể tham số và thời gian xử lý mà còn đạt kết quả tốt hơn các mô hình đối sánh. Những cải tiến này là cơ sở vững chắc để phát triển một hệ thống dự báo–cảnh báo chất lượng không khí có khả năng vận hành ổn định trong điều kiện tài nguyên hạn chế, đồng thời mở ra nhiều hướng nghiên cứu tiếp theo giàu tiềm năng.

TÀI LIỆU THAM KHẢO

[1] V. Điệp, "VietNamNet," Công ty Cổ phần Truyền thông VietNamNet, 16 4 2025. [Online]. Available: https://vietnamnet.vn/ha-noi-va-tphcm-chim-trong-bui-lot-top-thanh-pho-o-nhiem-toan-cau-2391614.html.

[2] Nguyễn Hoài - Thanh Hiếu, "Tienphong," 4 6 2024. [Online]. Available: https://tienphong.vn/cu-hon-7-phut-lai-co-nguoi-tu-vong-do-o-nhiem-khong-khi-post1642891.tpo.

[3] WHO global air quality guidelines, 2021.

[4] M. T. Vo, A. H. Vo, H. Bui and T. Le, "A Hybrid Deep Learning Approach for PM2.5 Concentration Prediction in Smart Environmental Monitoring," *Intelligent Automation & Soft Computing,* vol. 36, no. 3, p. 3029–3041, 2023.

[5] Open Development Mekong, "[English] Dataset on air quality in Vietnam in 2020," Open Development Mekong, 20 7 2021. [Online]. Available: https://data.opendevelopmentmekong.net/dataset/timelines-dataset-on-air-quality-in-vietnam/resource/c3712765-0d13-4d83-8695-fe803a6d9933.

[6] N. D. T. Thảo, "Điện máy xanh," Điện máy xanh, 2 6 2023. [Online]. Available: https://www.dienmayxanh.com/kinh-nghiem-hay/bui-pm-2-5-gay-o-nhiem-khong-khi-la-gi-tac-hai-va-1203859?srsltid=AfmBOoonWUW14Z77AlD7JzPlGu3Tks5fGCzegAJkd3ehp3XcyRTm57cT.

[7] WHO, "Ambient (outdoor) air pollution," WHO, 24 10 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health.

[8] geeksforgeeks, "Geeksforgeeks," GeeksforGeeks, 23 7 2025. [Online]. Available: https://www.geeksforgeeks.org/deep-learning/convolutional-neural-network-cnn-in-machine-learning/.

[9] S. Hochreiter and J. Schmidhuber, " Long Short-Term Memory," *Neural Computation,* vol. 9, no. 8, p. 1735–1780, 1997.

[10] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Networks,* vol. 18, no. 5-6, p. 602–610, 2005.

[11] Kanwal Yousaf; Tabassam Nawaz, "The architecture of the BiLSTM model," *A Deep Learning-Based Approach for Inappropriate Content Detection and Classification of YouTube Videos,* vol. 10, pp. 16283-16298, 2022.

[12] Zuhan Liu, Zihai Fang, Yuanhao Hu, "A deep learning-based hybrid method for PM2.5 prediction in central and western China," *Scientific Reports,* vol. 15, p. 10080, 2025.

[13] Zaini, N.; Ean, L. W.; Ahmed, A. N.; Abdul Malek, M.; Chow, M. F., "PM2.5 forecasting for an urban area based on deep learning and decomposition method," *Scientific Reports,* vol. 12, 2022.

[14] Xiaoxiao Feng, Xiaole Zhang, Stephan Henne, Yi-Bo Zhao, Jie Liu, Jing Wang, "A hybrid model for enhanced forecasting of PM2.5 spatiotemporal concentrations with high resolution and accuracy," *Environmental Pollution,* vol. 355, p. 124263, 2024.

[15] Anıl Utku; Umit Can; Mustafa Alpsülün; Hasan Celal Balıkçı; Azadeh Amoozegar; Abdulmuttalip Pilatin; Abdulkadir Barut, "Advancing Air Quality Monitoring: Deep Learning-Based CNN-RNN Hybrid Model for PM2.5 Forecasting," *Atmosphere,* vol. 16, no. 9, p. 1003, 2025.

[16] Bekkar, A.; Hssina, B.; Douzi, S.; Douzi, K., "Air-pollution prediction in smart city, deep learning approach," *Journal of Big Data,* vol. 8, no. 1, p. 161, 2021.

[17] Imran Junejo; Naveed Ahmed, "Depthwise Separable Convolutional Neural Networks for Pedestrian Attribute Recognition," *SN Computer Science,* 2021.

[18] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation,* vol. 9, no. 8, pp. 1735-1780, 1997.