CẢI TIẾN MÔ HÌNH HỌC SÂU DỰA TRÊN CNN VÀ BI-LSTM TRONG VẤN ĐỀ GIẢM SỐ LƯỢNG THAM SỐ VÀ XỬ LÝ DỮ LIỆU MẤT CÂN BẰNG

Nguyễn Thị Phương Trang*, Nguyễn Đức Cường

Khoa Công nghệ thông tin, Trường Đại học Ngoại ngữ - Tin học Thành phố Hồ Chí Minh

trangntp@huflit.edu.vn, cuongnd@huflit.edu.vn

TÓM TẮT—Bài báo trình bày các cải tiến trong mô hình học sâu kết hợp CNN và Bi-LSTM, nhằm giải quyết hai vấn đề quan trọng: dữ liệu mất cân bằng và độ phức tạp tính toán. Để đối phó với dữ liệu mất cân bằng, các kỹ thuật như: kỹ thuật tăng cường mẫu thiểu số (SMOTE), kỹ thuật giảm mẫu dư thừa (Undersampling), và điều chỉnh trọng số lớp được áp dụng, giúp nâng cao độ chính xác cho các lớp thiểu số trong bộ dữ liệu. Kết quả thực nghiệm trên bộ dữ liệu UCI Student Performance cho thấy hiệu quả của mô hình trong việc dự đoán hiệu suất học tập của học sinh. Trong khi đó, để giảm độ phức tạp tính toán, bài báo áp dụng kỹ thuật tích chập phân tách theo chiều sâu (Depthwise Separable Convolutions) nhằm giảm số lượng tham số của mô hình. Kết quả được trình bày thông qua bài toán dự báo chất lượng không khí PM2.5 tại Thành phố Hồ Chí Minh, chứng minh tính hiệu quả trong việc tiết kiệm tài nguyên tính toán mà không làm giảm hiệu suất dự đoán.

Từ khóa—Bi-LSTM, CNN, Deep learning, Depthwise Separable Convolutions

I. GIỚI THIỆU

Trong bối cảnh của Cuộc cách mạng công nghiệp lần thứ tư, kỷ nguyên bùng nổ của Công nghiệp 4.0, các lĩnh vực học máy và học sâu đã nổi lên như những điểm trọng tâm thu hút sự chú ý của các nhà nghiên cứu. Một trong những lợi ích tốt nhất của học sâu là nó có thể học các đặc trưng từ dữ liệu gốc. Việc kết hợp tối ưu hóa với học máy như một hình thức học sâu dựa trên mô hình sẽ giúp quá trình học có kết quả tốt hơn [1, 2]. Mỗi mô hình học sâu sử dụng các kỹ thuật máy học khác nhau sẽ cho ra kết quả với độ chính xác khác nhau [3]. Với khả năng hỗ trợ mạnh mẽ của mô hình học sâu trong việc học các biểu diễn tính năng phức tạp từ một lượng lớn dữ liệu, xu hướng áp dụng các mô hình học sâu vào các bài toán dự đoán thuộc các lĩnh vực như kinh tế [4, 5], giáo dục [6], y tế và chăm sóc sức khỏe [7, 8], môi trường [9], quản lý tiêu thụ năng lượng [10, 11], xử lý ngôn ngữ tự nhiên [12, 13] và dự đoán chuỗi thời gian ngày càng tăng [14, 15]. Các thuật toán dựa trên học máy và học sâu là những cách tiếp cận khá phổ biến trong việc giải quyết các vấn đề dự đoán theo chuỗi thời gian. Các kỹ thuật học máy được sử dụng nhiều nhất trong các nghiên cứu trên như: RNN, CNN, LSTM, BiLSTM. Những kỹ thuật này đã được chứng minh là tạo ra kết quả chính xác hơn so với các mô hình dựa trên hồi quy thông thường. The Recurrent Neural Network (RNN) được biết đến như một trong những cách tiếp cận hiệu quả để giải quyết các bài toán về dự báo chuỗi thời gian [16]. Một số nghiên cứu khác sử dụng một số biến thể của RNN như Long Short-Term Memory (LSTM) và Bidirectional Long Short-Term Memory (Bi-LSTM) và thu được nhiều kết quả khả quan trong một số ứng dụng. Nghiên cứu của Sahoo và các đồng sự trong [17] chỉ ra rằng LSTM-RNN rất hữu ích trong quá trình xử lý chuỗi thời gian liên tục. Mô hình LSTM-RNN cho kết quả tốt hơn mô hình chỉ sử dụng RNN trong bài toán dự báo mật độ lượng nước thấp tại lưu vực sông Mahanadi (Ấn Độ) theo mục đích dự báo trước một bước theo các giá trị hàng tháng, phương pháp dựa trên tất cả các giá trị hàng tháng của năm trước. Mô hình bao gồm một lớp input, một lớp hidden là lớp LSTM với các khối nhớ và một lớp output. Kiến trúc LSTM giúp ghi nhớ các chuỗi dữ liệu đầu vào dài hơn. Một kiến trúc BiLSTM thường chứa 2 mạng LSTM đơn được sử dụng đồng thời và độc lập để mô hình hoá chuỗi đầu vào theo 2 hướng: từ trái sang phải (forward LSTM) và từ phải sang trái (backward LSTM). Trong [18] nhóm tác giả đề xuất mô hình BOP-BL dựa trên Bi-LSTM để dự đoán giá dầu. Kiến trúc mô hình gồm 2 khối, khối thứ nhất sử dụng ba lớp Bi-LSTM, khối thứ hai có một lớp kết nối đầy đủ để dự đoán giá dầu. Trong [19] nhóm tác giả chỉ ra rằng các mô hình Bi-LSTM vượt trội hơn đáng kể so với các mô hình LSTM thông thường. Cụ thể, khi sử dụng hai mô hình trên để dự báo dữ liệu chứng khoán, mô hình Bi-LSTM vượt trội hơn LSTM với tỷ lệ lỗi giảm 37,78%, tuy nhiên mô hình Bi-LSTM đạt đến trạng thái cân bằng chậm hơn so với các LSTM. Tương tự, với nghiên cứu trong [20], nhóm tác giả sử dụng mô hình Bi-LSTM dựa trên học sâu (DLBL-WQA) để dự báo các yếu tố chất lượng nước của sông Yamuna (Ấn Độ). Kết quả thực nghiệm cho thấy giá trị dự báo của mô hình tốt hơn mô hình LSTM. Dữ liệu chuỗi thời gian có các mối quan hệ bất biến theo thời gian, điều này làm cho các mô hình dựa trên CNN phù hợp với các tác vụ chuỗi thời gian. Một số nghiên cứu đã kết hợp CNN và BiLSTM trong dự đoán chuỗi thời gian. CNN giúp trích xuất và giảm kích thước tính năng và sau đó Bi-LSTM được giải trình tự để dự đoán các giá trị cụ thể. Cách tiếp cận này giúp cải thiện hiệu suất và thời gian của các mô hình dự đoán. Trong [21] các tác giã đã đề xuất phương pháp sử dụng CNN và Bi-LSTM để dự đoán mức tiêu thụ năng lượng điện. Kết quả cho thấy mô hình cải thiện hiệu suất của mô hình dự đoán. Nghiên cứu trên cho thấy các mô hình CNN và Bi-LSTM có hiệu suất tốt về phát hiện dữ liệu chuỗi thời gian và có thể tìm ra các vấn đề gây ra theo thời gian, có thể trích xuất các tính năng cục bộ một cách hiệu quả, nhưng cấu trúc mạng tương đối phức tạp và độ phức tạp tính toán cao, đòi hỏi nhiều tài nguyên tính toán trong quá trình đào tạo và dự đoán. Vì vậy sự kết hợp dựa trên CNN và Bi-LSTM không thể cải thiện hiệu suất trong một số trường hợp bởi vì các mô hình dựa trên CNN có thể mất các tính năng thiết yếu do các đặc điểm hạn chế của dữ liệu. Nhiều cấu trúc được đề xuất để giảm độ phức tạp tính toán và số lượng tham số của các mô hình CNN. Trong đó, Depthwise Separable Convolution (Tích chập chiều sâu

* Coressponding Author

tách biệt) được sử dụng rộng rãi [22-24]. Depthwise Separable Convolution là một cải tiến trong MobileNet, kiến trúc này không áp dụng bộ lọc lên toàn bộ độ sâu của layer trước đó mà chỉ được áp dụng trên một điểm đơn lẻ, giúp giảm kích thước và số lượng tham số của mô hình. Trong [24] các tác giả đề xuất mô hình chuẩn đoán lỗi của các ổ trục sử dụng CNN dựa trên Depthwise Separable Convolution giúp cải thiện hiệu quả, tăng độ chính xác của quá trình nhận dạng và đảm bảo tốc độ tính toán nhanh, không gian tham số nhỏ, mô hình có khả năng học thay đổi nhanh và hiệu suất chống nhiễu.

Bài báo tập trung vào việc cải tiến mô hình học sâu bằng sử dụng Convolutional Neural Networks (CNN) và Bidirectional Long Short-Term Memory (Bi-LSTM) với việc giải quyết hai thách thức chính:

- Giảm số lượng tham số trong mô hình: Một trong những vấn đề nổi bật khi phát triển các mô hình học sâu là số lượng tham số lớn, dẫn đến chi phí tính toán cao và khó khăn trong việc triển khai trên các hệ thống tài nguyên hạn chế. Việc giảm số lượng tham số mà không làm giảm hiệu suất của mô hình là một bài toán quan trọng. Bằng cách kết hợp Depthwise Separable Convolutions (DSC) với mô hình sẽ giúp tối ưu hóa cấu trúc mô hình sao cho giảm thiểu tham số mà vẫn duy trì khả năng học tốt từ dữ liệu.
- Xử lý dữ liệu mất cân bằng (Imbalanced Data): Dữ liệu mất cân bằng, trong đó một số lớp có số lượng mẫu quá ít so với các lớp khác, là một thách thức lớn trong việc huấn luyện mô hình học sâu. Các mô hình học sâu có xu hướng thiên về các lớp có tần suất cao, dẫn đến việc dự đoán không chính xác cho các lớp ít phổ biến. Bài báo sẽ nghiên cứu các kỹ thuật như điều chỉnh trọng số, tái cân bằng dữ liệu và các phương pháp học không giám sát để cải thiện hiệu suất mô hình đối với dữ liệu mất cân bằng. Cụ thể, bài báo đi sâu vào ứng dụng của kỹ thuật tăng cường mẫu thiểu số (Synthetic Minority Over - sampling Technique – SMOTE), biến thể của SMOTE là Borderline-SMOTE, kỹ thuật lấy mẫu tổng hợp thích nghi (Adaptive Synthetic Sampling - ADASYN) và các kỹ thuật tương tự trong phân tích dữ liệu bảng, cung cấp thông tin chi tiết về cách giải quyết hiệu quả các thách thức mất cân bằng dữ liệu

Những đóng góp chính của nghiên cứu này được tóm tắt như sau:

- Xây dựng mô hình tích hợp CNN và Bi-LSTM (gọi là mô hình CLS) kết hợp SMOTE để đối phó với dữ liệu mất cân bằng trong bài toán dự đoán kết quả học tập của học sinh trên bộ dữ liệu UCI Student Performance. Thực nghiệm trên 4 mô hình khác nhau: CNN-LSTM, CNN-Bi-LSTM và Naive Predictor, XGBoost. Kết quả cho thấy mô hình của đề xuất hoạt động tốt hơn các phương pháp thực nghiệm khác.
- Xây dựng mô hình tích hợp CNN và Bi-LSTM, trong đó CNN được kết hợp với Depthwise Separable Convolutions để tăng hiệu suất mô hình (gọi là mô hình CDL). Phần thực nghiệm tiến hành đánh giá khả năng dự báo của phương pháp đề xuất và so sánh kết quả với các mô hình LSTM, Bi-LSTM, CNN-LSTM, ARIMA và PM25-CBL trong bài toán dự báo nồng độ bụi mịn PM2.5 trong bộ dữ liệu Air Quality HCMC tại Vietnam. Kết quả chỉ ra rằng mô hình CDL vượt trội so với các phương pháp thử nghiệm khác đối với tập dữ liệu Air Quality HCMC về các chỉ số MSE, RMSE, MAE và MAPE.

Phần còn lại của bài viết này được tổ chức như sau. Tóm tắt các nghiên cứu liên quan được trình bày trong Phần 2. Các mô hình đề xuất được trình bày trong Phần 3. Kết quả thực nghiệm ở Phần 4. Cuối cùng, Phần 5 đưa ra kết luận của nghiên cứu này.

II. CÁC CÔNG TRÌNH LIÊN QUAN

Trong những năm gần đây, nhiều nghiên cứu đã chỉ ra rằng việc xử lý hiệu quả dữ liệu mất cân bằng có thể cải thiện đáng kể hiệu suất của mô hình học sâu. Các phương pháp như SMOTE (Synthetic Minority Over-sampling Technique), ADASYN (Adaptive Synthetic Sampling) đã được áp dụng rộng rãi để giải quyết vấn đề này. Nghiên cứu của Kannan và đồng sự trong [25] áp dụng SMOTE và ADASYN để tăng cường mẫu cho lớp thiểu số trong các bài toán phân loại y tế. Kết quả cho thấy việc sử dụng các kỹ thuật này giúp cải thiện độ chính xác của mô hình, với độ chính xác đạt 99.2% trên bộ dữ liệu COVID-19, 99.4% trên bộ dữ liệu Kidney, và 99.5% trên bộ dữ liệu Dengue. Họ cũng sử dụng TabNet, một mô hình học sâu chuyên biệt cho dữ liệu bảng, để xử lý dữ liệu mất cân bằng và đạt được hiệu suất cao trong các bài toán phân loại y tế. Ngoài ra, nghiên cứu của Joloudari và cộng sự [26] đã kết hợp SMOTE với mạng CNN để xử lý dữ liệu mất cân bằng trong các bài toán phân loại. Kết quả cho thấy mô hình kết hợp này đạt độ chính xác lên cao trên 24 bộ dữ liệu mất cân bằng, vượt trội hơn so với các phương pháp khác như chỉ sử dụng CNN hoặc SMOTE riêng lẻ. Một nghiên cứu khác của Liu và cộng sự [27] đã đề xuất phương pháp Deep Attention SMOTE, kết hợp giữa SMOTE và cơ chế chú ý trong học sâu để cải thiện việc tạo mẫu cho lớp thiểu số. Kết quả cho thấy phương pháp này giúp cải thiện khả năng phân loại lớp thiểu số và giảm thiểu sự thiên lệch của mô hình. Những nghiên cứu này cho thấy rằng việc áp dụng các phương pháp xử lý dữ liệu mất cân bằng như SMOTE, ADASYN có thể giúp cải thiện hiệu suất của mô hình học sâu trong các bài toán phân loại, đặc biệt là khi đối mặt với dữ liệu mất cân bằng. Trong phân tích dữ liệu bảng, việc giải quyết tình trạng mất cân bằng dữ liệu là rất quan trọng và các phương pháp lấy mẫu quá mức như SMOTE và ADASYN đóng vai trò quan trọng trong việc cân bằng phân phối lớp và ngăn ngừa sai lệch khi đào tạo. SMOTE tạo ra các lập trường tổng hợp cho lớp thiểu số, trong khi ADASYN điều chỉnh quy trình để tập trung vào các trường hợp đầy thách thức. Các phương pháp này không chỉ cân bằng các tập dữ liệu mà còn tăng cường khả năng phân biệt mô hình trong các lớp thiểu số, thúc đẩy

độ chính xác dự đoán. Bài báo này đi sâu vào ứng dụng của SMOTE, ADASYN và các kỹ thuật tương tự trong phân tích dữ liệu bảng, cung cấp thông tin chi tiết về cách giải quyết hiệu quả các thách thức mất cân bằng dữ liệu.

Hiện nay, vấn đề về ô nhiễm không khí đang là mối quan tâm của thế giới và Việt Nam, bởi tác động nghiêm trọng của nó đến môi trường và sức khỏe con người. Không khí xung quanh chứa các chất gây ô nhiễm bao gồm: các chất ô nhiễm dạng bụi (particulate matter, aerosol) và các chất ô nhiễm dạng khí. Trong đó PM2.5 là chất ô nhiễm dạng bụi được xem là kẻ giết người thầm lặng nguy hiểm nhất hiện nay. Có nhiều nghiên cứu sử dụng máy học và học sâu để dự đoán PM2.5. Pak và cộng sự trong [28] đã sử dụng Học sâu để dự đoán nồng độ PM2.5 ở một số thành phố ở Trung Quốc. Mô hình sử dụng CNN-LSTM có hiệu suất dự đoán tốt và ổn định hơn so với các mô hình perceptron (MLP). Vo và các cộng sự trong [29] phát triển mô hình học sâu PM25-CBL tích hợp CNN và Bi LSTM để dự đoán nồng độ PM2.5 trong bộ dữ liệu Air Quality HCMC. Kết quả thực nghiệm xác nhận rằng PM25-CBL Kết quả chỉ ra rằng mô hình PM25-CBL vượt trội so với các phương pháp đoán chuỗi thời gian bao gồm LSTM, Bi-LSTM, CNN, CNN-LSTM, CNN-Bi-LSTM và ARIMA trên bộ dữ liệu Air Quality HCMC về các chỉ số MSE, RMSE, MAE và MAPE. Các tác giả trong [30] xây dựng mô hình sử dụng CNN và LSTM để dự đoán nồng độ PM2.5 hàng giờ trong 24 giờ sắp tới trên sáu địa điểm tại Thành phố Hồ Chí Minh (HCMC), với tập dữ liệu được lấy từ 6 trạm đo không khí theo thời gian thực. Nghiên cứu chỉ ra kết quả mô hình dự đoán tốt hơn khi so sánh với các phương pháp SGDRegressor, XGBoost. Các mô hình trên đưa ra dự đoán PM2.5 với độ chính xác cao, tuy nhiên vẫn còn hạn chế về thời gian huấn luyện mô hình hoặc một số mô hình không thể áp dụng cho việc dự đoán tại các địa điểm khác có tập dữ liệu với nhiều tính năng khác nhau. Trong nghiên cứu này, chúng tôi phân tích sự ảnh hưởng của các biến số đến nồng độ PM2.5 trong bộ dữ liệu Air Quality HCMC tại Vietnam. Tiếp đến xây dựng mô hình PM25-CBL tích hợp CNN và Bi-LSTM, trong đó CNN được kết hợp với Depthwise Separable Convolutions để tăng hiệu suất mô hình. Phần thực nghiệm tiến hành đánh giá khả năng dự báo của phương pháp đề xuất và so sánh kết quả với các mô hình LSTM, Bi-LSTM, CNN-LSTM, ARIMA và PM25-CBL. Kết quả chỉ ra rằng mô hình PM25-CDSCBL vượt trội so với các phương pháp thử nghiệm khác đối với tập dữ liệu Air Quality HCMC về các chỉ số MSE, RMSE, MAE và MAPE.

III. CÁC MÔ HÌNH ĐỀ XUẤT

A. MÔ HÌNH DỰ ĐOÁN KẾT QUẢ HỌC TẬP SINH VIÊN

1. CÁC BỘ DỮ LIỆU

Trong bài báo này, chúng tôi sử dụng ba tập dữ liệu: tập dữ liệu Student Performance in Mathematics [30], tập dữ liệu Student Performance in Portuguese language [30] và tập dữ liệu Students' Academic Performance (xAPI) [31]. Thu thập từ UCI ML Repository, tập dữ liệu Student Performance in Mathematics có 395 bộ với 33 thuộc tính, trong đó thuộc tính 'G3' được chọn làm đối tượng dự đoán. Với tập dữ liệu Student Performance in Portuguese language, có tổng cộng 649 bộ, bao gồm 33 thuộc tính, trong đó thuộc tính 'G3' được chọn làm đối tượng dự đoán. Cuối cùng, tập dữ liệu Students' Academic Performance bao gồm 480 bộ và 16 thuộc tính, trong đó các nhãn dự đoán được chia thành 3 lớp.

2. MÔ HÌNH ĐỀ XUẤT

Trong phần này, chúng tôi giới thiệu kiến trúc tổng thể bao gồm tiền xử lý, phân chia tập dữ liệu, lấy mẫu xử lý mất cân bằng, phân loại và đánh giá. Kiến trúc chung được trình bày trong Hình 1.

a) Bước 1: Tiền xử lý dữ liệu (Data preprocessing):

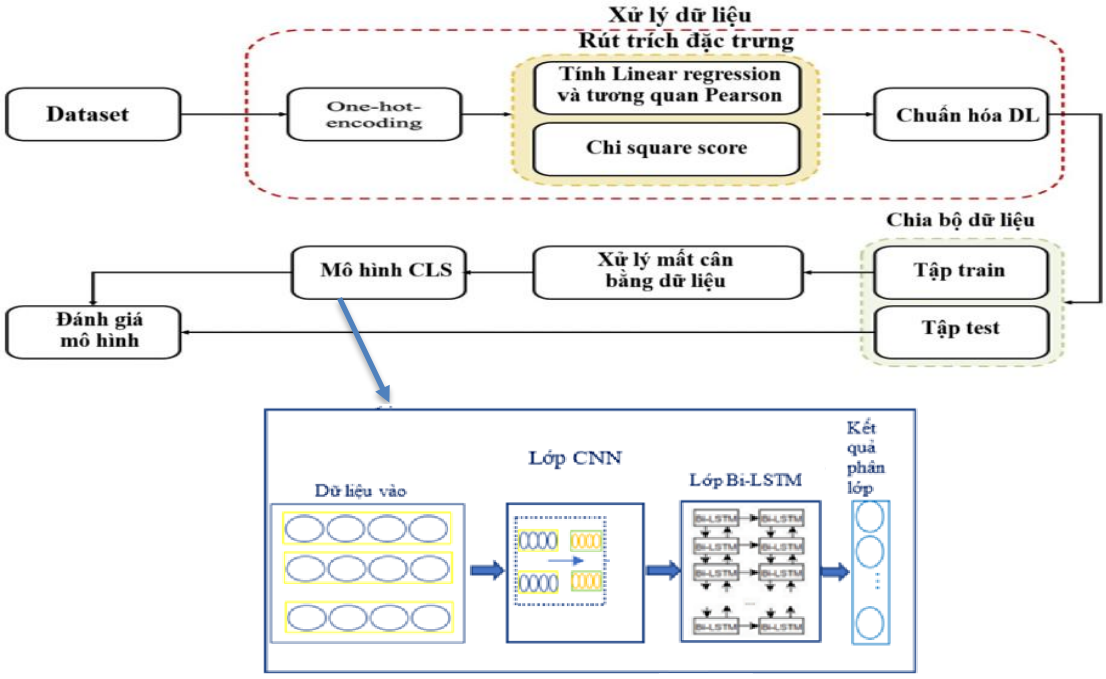
Trước khi đào tạo mô hình, để đảm bảo mô hình xử lý hiệu quả, chúng tôi đã áp dụng mã hóa one-hot trong quá trình xử lý dữ liệu trước. Điều này chuyển đổi các biến phân loại thành dạng nhị phân, giúp mô hình học dễ dàng hơn với dữ liệu phân loại. Bước chọn đặc trưng (Feature selection) sử dụng các phương pháp như Pearson Correlation và Linear Regression, cùng với Chi square score để chọn các đặc trưng quan trọng nhất cho mô hình. Điều này giúp giảm thiểu độ phức tạp và cải thiện hiệu suất mô hình. Đầu tiên, mô hình sẽ tính Pearson Correlation. Thuộc tính nào có sigps >= 0,6 hoặc sigps <= 0,6 trong Pearson Correlation sẽ được giữ, nếu abs (sigps) <= 0,6, mô hình sẽ xóa thuộc tính này. Sau khi xóa các thuộc tính được xác định là không tương quan, chúng tôi chuẩn hóa dữ liệu bằng phương pháp chia tỷ lệ Min-max cho các thuộc tính (trừ các tính năng được dự đoán). Bước chuẩn hóa dữ liệu (Normalize data) đảm bảo tất cả các đặc trưng có cùng thang đo, giúp mô hình hoạt động hiệu quả hơn.

b) Bước 2: Chia dữ liệu

Tiếp theo, chia tập dữ liệu thành tập huấn luyện và tập kiểm tra bằng cách sử dụng Random sampling và Fractional Sampling Floor. Vì bộ dữ liệu có sự chênh lệch về số lượng phần tử giữa các lớp dự đoán, nên mô hình sử dụng các phương pháp đánh trọng số và tăng kích thước mẫu lớp thiểu số để đối phó với bộ dữ liệu mất cân bằng.

c) Bước 3: Huấn luyện và đánh giá mô hình

Tập huấn luyện sẽ được đưa vào để huấn luyện với các thuật toán học máy và thu được mô hình dự đoán. Mô hình dự đoán này sẽ được kiểm tra dựa vào tập với các chỉ số Accuracy, Precision, Recall, F1-Score. Trước khi huấn luyện, chúng tôi thiết lập ba trường hợp: Repeated Cross-validation, oversampling với SMOTE, oversampling với ADASYN, và oversampling với Borderline SMOTE.



Hình 1. *Mô hình CLS dự đoán kết quả sinh viên*

B. MÔ HÌNH DỰ ĐOÁN NỒNG ĐỘ BỤI MỊN PM2.5

1. BỘ DỮ LIỆU

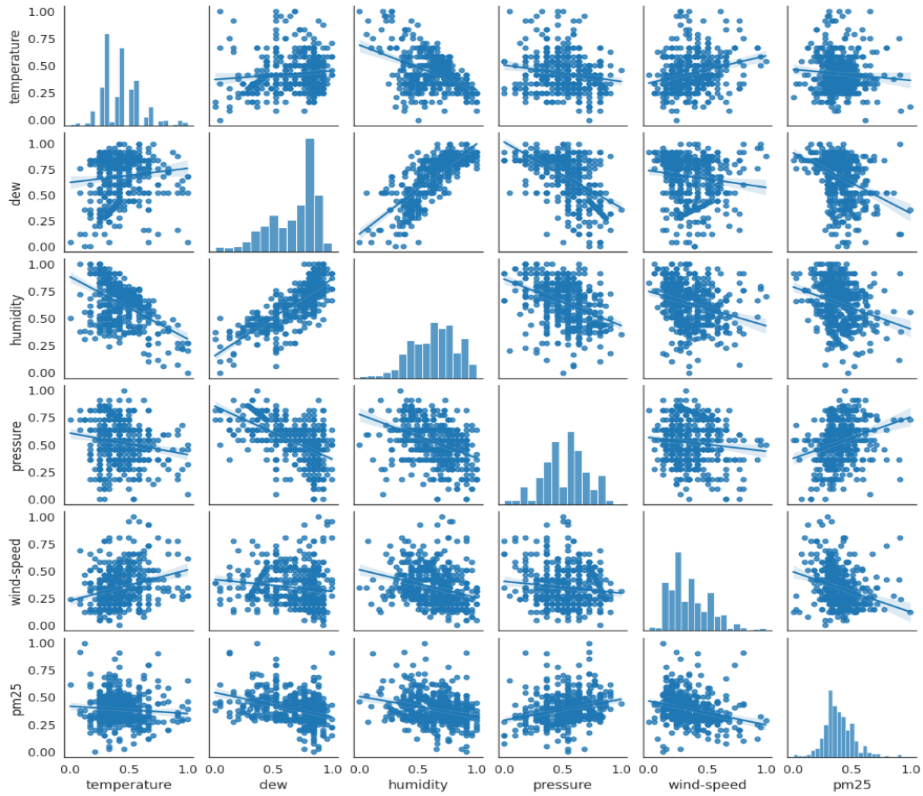
Dữ liệu sử dụng là Air Quality HCMC dataset được cung cấp bởi Open Development Mekong [32]. Bộ dữ liệu gồm 7 biến: ngày (Date) được dùng làm index trong bộ dữ liệu, nhiệt độ (Temperature), độ ẩm (Humidity), tốc độ gió (Wind Speed), PM2.5 (giá trị mục tiêu), sương (Dew) và áp suất (Pressure) (Bảng 1). Trong đó, PM2.5 là biến cần dự báo, các biến còn lại được dùng làm đầu vào. 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 2. 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.

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

TT	Biến	Min	Max	Median	Đơn vị
1	Ngày	30-12-2019	20-01-2021	10-07-2020	Ngày
2	Nhiệt độ trung bình	23	31	27.5	^oC
3	Độ ẩm trung bình	47	100	76.5	%
4	Tốc độ gió trung bình	0.5	5.4	2.3	m/s
5	Sương trung bình	14.5	26.5	24	^oC
6	Áp suất trung bình	1003	1014	1009	hPa
7	Nồng độ PM2.5 trung bình	5	171	68	$\mu g/m^3$

2. MÔ HÌNH ĐỀ XUẤT

Mô hình CDL được phát triển nhằm giải quyết bài toán dự đoán nồng độ PM2.5 trong không khí, một trong những yếu tố quan trọng trong việc bảo vệ sức khỏe cộng đồng và môi trường. Mô hình CDL kết hợp các kỹ thuật học sâu hiện đại như Convolutional Neural Networks (CNN), Bidirectional Long Short-Term Memory (Bi-LSTM) và Depthwise Separable Convolution (DSC) để đạt được hiệu quả cao trong việc dự đoán nồng độ PM2.5. Kiến trúc mô hình CDL được trình bày trong hình 3 với hai giai đoạn chính: Huấn luyện và Thử nghiệm.



Hình 2. *Ma trận phân tán của các biến với biến pm2.5*

a) Lớp Convolutional Neural Networks (CNN) với Depthwise Separable Convolution (DSC)

Lớp CNN kết hợp với DSC giúp giảm thiểu số lượng tham số trong mô hình mà vẫn giữ được hiệu suất cao trong việc trích xuất đặc trưng không gian. DSC là một kỹ thuật giúp phân tách các phép toán convolution thành hai phần nhỏ hơn, từ đó giảm thiểu tính toán và số lượng tham số. Lớp CNN này đóng vai trò quan trọng trong việc học các đặc trưng không gian của dữ liệu, chẳng hạn như các xu hướng thời gian và sự thay đổi trong nồng độ PM2.5. Việc kết hợp CNN với DSC giúp giảm số lượng tham số trong mô hình, đồng thời duy trì khả năng học đặc trưng mạnh mẽ từ dữ liệu.

Depthwise Separable Convolution có thể được chia thành hai bước chính: Depthwise Convolution: Thực hiện phép tích chập riêng biệt cho từng đặc trưng và Pointwise Convolution để kết hợp các đặc trưng.

Công thức Depthwise Convolution:

$$y_{i,c} = \sum_{k=0}^{K-1} x_{i+k,c} * w_{k,c}$$

Trong đó:

$y_{i,c}$ là giá trị đầu ra tại vị trí i của đặc trưng c

$x_{i+k,c}$ là giá trị mẫu i+k trong đặc trưng c

$w_{k,c}$ là trọng số bộ lọc với chiều dài K

Công thức Pointwise Convolution:

$$y_{i,c'} = \sum_{c=0}^{A} x_{i,c} * w_{c,c'}$$

Trong đó:

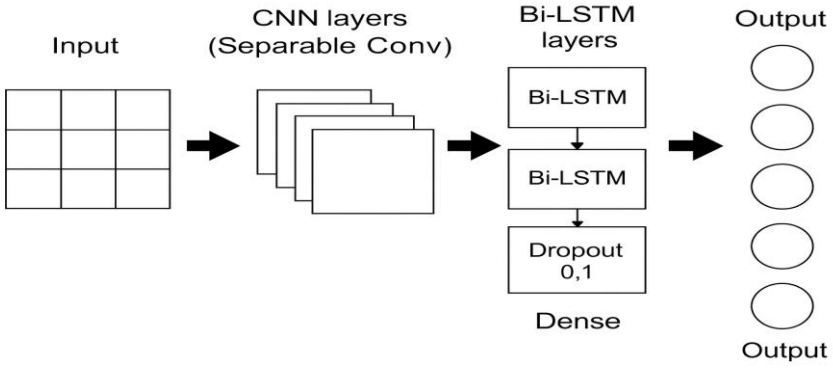
$y_{i,c'}$ là giá trị đầu ra tại vị trí i của đặc trưng c'

$x_{i,c}$ là giá trị đầu ra tại vị trí i của đặc trưng c'

$w_{c,c'}$ là trọng số bộ lọc kết hợp các đặc trưng để tạo ra đặc trưng đầu ra c' trong bảng dữ liệu với A đặc trưng

b) Lớp Bidirectional Long Short-Term Memory (Bi-LSTM)

Bi-LSTM có khả năng xử lý thông tin theo cả hai chiều thời gian (từ quá khứ đến hiện tại và từ hiện tại đến quá khứ). Điều này giúp mô hình hiểu rõ hơn về các mối quan hệ giữa các sự kiện trong chuỗi thời gian, đặc biệt là trong các bài toán như dự đoán nồng độ PM2.5, nơi thông tin từ quá khứ và tương lai đều có thể ảnh hưởng đến dự đoán.



Hình 3. *Kiến trúc mô hình CDL giai đoạn huấn luyện*

Trong giai đoạn kiểm thử, chỉ năm biến gồm temperature, humidity, wind speed, dew và pressure được đưa vào mô hình đã huấn luyện để dự báo giá trị PM2.5.

IV. KẾT QUẢ THỰC NGHIỆM

A. THIẾT LẬP THỰC NGHIỆM

Kết quả thực nghiệm được triển khai trên framework Keras với máy tính Ubuntu Intel Core i7-4790K (4.0 GHz x 8 cores), RAM 32 GB, và GeForce GTX 1080 Ti.

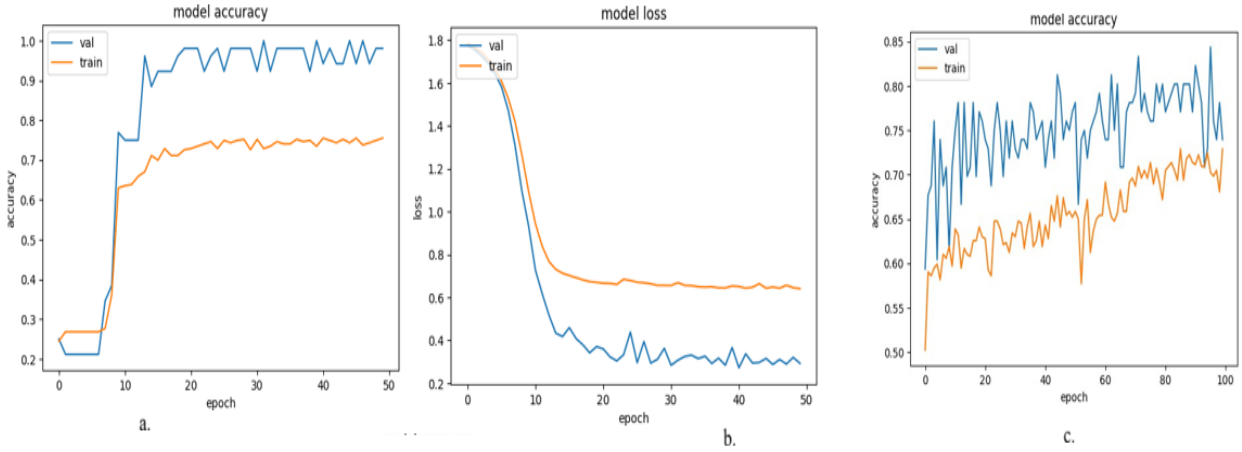
B. KẾT QUẢ MÔ HÌNH DỰ ĐOÁN KẾT QUẢ HỌC TẬP SINH VIÊN

Các thông số cấu hình của cả hai mô hình được trình bày trong Bảng 2. Mô hình CLS bao gồm ba lớp tích chập 1D (1D Convolution) với 64 bộ lọc mỗi lớp. Bên cạnh đó, mô hình cũng sử dụng các lớp Max Pooling và Bi-LSTM để học các đặc trưng phức tạp.

Kết quả thực nghiệm sau khi sử dụng SMOTE kết hợp CNN với LSTM và Bi-LSTM, kết quả của mô hình CLS khi kết hợp CNN với Bi-LSTM có sự cải thiện cao, quá trình đào tạo được thể hiện ở hình 4 và lần lượt thể hiện các tham số độ chính xác và mất mát (accuracy và loss). Như chúng ta có thể thấy, độ chính xác của cả tập huấn luyện và tập kiểm tra đều tăng nhanh trong 20 epochs đầu và đạt chỉ số chính xác cao cho tập kiểm tra. Bên cạnh đó, lớp Dropout được thêm vào để giảm hiện tượng overfitting, do đó chúng ta có thể thấy chỉ số accuracy của tập huấn luyện nhỏ đạt kết quả tốt. Ngược lại, chỉ số loss có xu hướng giảm đáng kể trong 20 kỷ nguyên đầu tiên. Các kết quả thực nghiệm của mô hình được trình bày và so sánh trong bảng 3 cho thấy hiệu quả vượt trội giữa kết quả của mô hình.

Bảng 2. *Các thông số cấu hình của mô hình CLS*

STT	Loại lớp	Số nơ-ron (Neurons)	Số tham số (Parameters)
1	Tích chập 1 chiều (1D Convolution)	(None, None, 5, 64)	192
2	Gộp cực đại 1 chiều (1D Max Pooling)	(None, None, 5, 64)	0
3	Tích chập 1 chiều (1D Convolution)	(None, None, 4, 64)	8,256
4	Gộp cực đại 1 chiều (1D Max Pooling)	(None, None, 4, 64)	0
5	Flatten	(None, None, 256)	0
6	LSTM hai chiều (Bi-LSTM)	(None, None, 128)	164,352
7	Dropout	(None, 128)	0
8	LSTM hai chiều (Bi-LSTM)	(None, 64)	41,216



Hình 4. *Độ chính xác khi huấn luyện và kiểm tra dữ liệu với CNN và Bi-LSTM trong ba bộ dữ liệu Student Performance in Mathematics Dataset (a) Portuguese language dataset (b) và xAPI dataset (c)*

Bảng 3. *Dự đoán kết quả trong Student Performance in Mathematics Dataset và Portuguese language dataset*

	CNN với LSTM (của bài báo)	**CNN với Bi-LSTM (của bài báo - CLS)**	**Random forest**	**XGBoost**
Accuracy	89.23%	92.31%	76.9%	78.9%

Đối với dữ liệu xAPI, chúng tôi sử dụng lấy mẫu quá mức với phương pháp ADASYN khá hiệu quả khi kết hợp với các thuật toán học máy, tạo ra kết quả khá tích cực khi chỉ số độ chính xác vượt trội so với các phương pháp lấy mẫu quá mức khác. Với CNN kết hợp với LSTM và Bi- LSTM (CLS), chúng tôi tăng thời gian huấn luyện lên 100 epochs và đạt kết quả tốt khi chỉ số chính xác đạt tới 84,38%. Bảng 4 so sánh kết quả giữa CNN với LSTM, CLS (CNN với BiLSTM) và kết quả thực nghiệm trước đó của Elaf Abu Amrieh và đồng sự [30]. Kết quả của chúng tôi cao hơn 10% so với kết quả của Elaf Abu Amrieh và cho thấy xAPI kết hợp với BiLSTM tạo ra kết quả thực sự hiệu quả.

Bảng 4. *Đánh giá mô hình trong xAPI dataset*

	CNN với Bi-LSTM (CLS)	**CNN với LSTM (của bài báo)**	**ANN [30]**
Accuracy	84.38%	80.21%	73.8%
Precision	84.26%	80.17%	73.8%
Recall	85.21%	80.59%	73.9%
F1-Score	84.47%	80.33%	73.2%

C. KẾT QUẢ MÔ HÌNH DỰ ĐOÁN NỒNG ĐỘ BỤI MỊN PM2.5

Trong phần này, để chứng minh hiệu quả của mô hình, chúng tôi cài đặt trên cùng bộ dữ liệu với các mô hình: CNN-LSTM, ARIMA, PM25-CBL trong bài báo [31] và mô hình của chúng tôi là CDL. chúng tôi so sánh hai mô hình PM25-CBL [31] và CDL về cấu trúc và hiệu quả tính toán. Các thông số cấu hình của cả hai mô hình được trình bày trong Bảng 5 và Bảng 6. Trong quá trình thực nghiệm, giá trị loss ở cả giai đoạn huấn luyện và kiểm thử được theo dõi hình (5). 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 CDL, với learning rate khởi tạo là 0,001.

Mô hình PM25-CBL (Bảng 5) bao gồm ba lớp tích chập 1D (1D Convolution) với 64 bộ lọc mỗi lớp. Bên cạnh đó, mô hình cũng sử dụng các lớp Max Pooling và Bi-LSTM để học các đặc trưng phức tạp. Tổng cộng, mô hình PM25-CBL có 214,081 tham số. Lớp Bi-LSTM đầu tiên có 164,352 tham số, chiếm phần lớn trong tổng số tham số của mô hình. Mô hình CDL (Bảng 6) thay thế các lớp tích chập thông thường bằng lớp SeparableConv, giúp giảm thiểu số lượng tham số và độ phức tạp tính toán. Cụ thể, CDL sử dụng hai lớp SeparableConv và một lớp Bi-LSTM giống như mô hình PM25-CBL. Tổng số tham số của mô hình này là 67,299, thấp hơn đáng kể so với PM25-CBL, chủ yếu do việc sử dụng lớp SeparableConv giúp giảm thiểu số lượng tham số.

Bảng 5. *Các thông số cấu hình của mô hình PM25-CBL [31]*

STT	Loại lớp	Số nơ-ron (Neurons)	Số tham số (Parameters)
1	Tích chập 1 chiều (1D Convolution)	(None, None, 5, 64)	192
2	Gộp cực đại 1 chiều (1D Max Pooling)	(None, None, 5, 64)	0
3	Tích chập 1 chiều (1D Convolution)	(None, None, 4, 64)	8,256
4	Gộp cực đại 1 chiều (1D Max Pooling)	(None, None, 4, 64)	0
5	Flatten	(None, None, 256)	0
6	LSTM hai chiều (Bi-LSTM)	(None, None, 128)	164,352
7	Dropout	(None, 128)	0
8	LSTM hai chiều (Bi-LSTM)	(None, 64)	41,216

Tổng tham số: 214,081

Bảng 6. *Các thông số cấu hình của mô hình CDL*

STT	Loại lớp	Số nơ-ron (Neurons)	Số tham số (Parameters)
1	Tích chập 1 chiều (1D Convolution)	(None, None, 5, 64)	66
2	Gộp cực đại 1 chiều (1D Max Pooling)	(None, None, 5, 64)	0
3	Tích chập 1 chiều (1D Convolution)	(None, None, 4, 64)	1,120
4	Gộp cực đại 1 chiều (1D Max Pooling)	(None, None, 4, 64)	0
5	Flatten	(None, None, 256)	0
6	LSTM hai chiều (Bi-LSTM)	(None, None, 128)	41,216
7	Dropout	(None, 128)	0
8	LSTM hai chiều (Bi-LSTM)	(None, 64)	24,832
9	Fully connected	(None, 1)	65

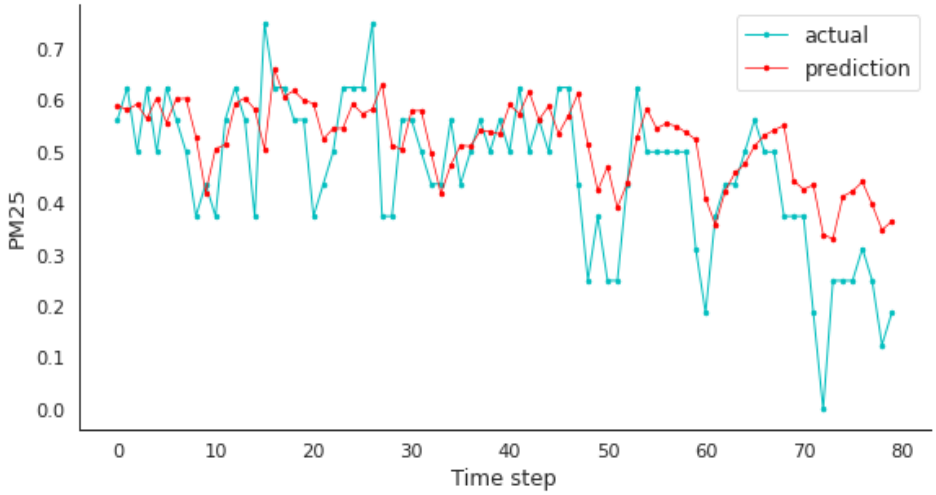
Tổng tham số: 67,299

Việc sử dụng lớp SeparableConv giúp giảm thiểu số lượng tham số, giúp mô hình có thời gian huấn luyện nhanh hơn (Bảng 7). Mô hình CNN-LSTM có thời gian huấn luyện là 30.54 giây và thời gian kiểm thử là 1.33 giây, trong khi mô hình CDL có thời gian huấn luyện là 35.71 giây và thời gian kiểm thử là 2.62 giây. Mặc dù thời gian huấn luyện và kiểm thử của CDL cao hơn một chút so với CNN-LSTM, nhưng điều quan trọng là CDL đã sử dụng Depthwise Separable Convolution (DSC), giúp giảm đáng kể số lượng tham số trong mô hình mà vẫn giữ được hiệu suất tốt.

Bảng 7. *Thời gian huấn luyện và kiểm thử.*

#No	**Mô hình**	**Huấn luyện**	**Kiểm thử**
1	CNN-LSTM	30.54	1.33
2	CDL	35.71	2.62
3	PM25-CBL	53.99	2.65

Việc áp dụng DSC giúp mô hình CDL giảm được số lượng phép toán trong các lớp chập so với các mô hình truyền thống, từ đó giúp tối ưu hóa bộ nhớ và tốc độ tính toán, CDL thực sự thể hiện hiệu quả vượt trội khi áp dụng Depthwise Separable Convolution, vì mô hình này không chỉ giảm số lượng tham số mà còn duy trì được khả năng học đặc trưng mạnh mẽ từ dữ liệu. Đặc biệt khi so với các mô hình không sử dụng DSC như PM25-CBL, vốn có thời gian huấn luyện dài hơn đáng kể (53.99 giây). Điều này chứng tỏ rằng CDL có sự cân bằng tốt giữa độ chính xác và khả năng tối ưu tài nguyên, đồng thời cung cấp một phương pháp hiệu quả trong việc giảm thiểu chi phí tính toán mà không làm giảm độ chính xác trong các nhiệm vụ dự đoán.



Hình 5. *Minh họa kết quả dự đoán của mô hình CDL*

Kết quả dự đoán được minh họa trong hình 5. Đường xanh đại diện cho giá trị thực tế của nồng độ PM2.5 qua các bước thời gian, trong khi đường đỏ thể hiện giá trị dự báo từ mô hình. Các kết quả cho thấy sự khớp khá tốt giữa giá trị thực tế và giá trị dự báo. Nhìn chung, mô hình dự báo đã theo sát được xu hướng thay đổi của dữ liệu thực tế, đặc biệt là ở những giai đoạn biến động không khí mạnh. Tuy nhiên, tại một số điểm, đặc biệt là khi nồng độ PM2.5 có sự dao động mạnh hoặc đột ngột thay đổi, sự chênh lệch giữa giá trị thực tế và giá trị dự báo vẫn tồn tại. Điều này chỉ ra rằng mô hình đã có khả năng tổng quát tốt nhưng có thể cần cải tiến thêm để nâng cao độ chính xác, đặc biệt trong các giai đoạn mà dữ liệu có sự thay đổi nhanh và mạnh. Mặc dù vậy, kết quả này vẫn cho thấy mô hình dự báo có tiềm năng ứng dụng thực tế cao trong việc theo dõi chất lượng không khí.

Bảng 8. *Kết quả thực nghiệm trên bộ dữ liệu Air Quality HCMC 2020.*

STT	**Mô hình**	**MSE**	**RMSE**	**MAE**	**MAPE**
1	CNN-LSTM	1.4	1.1	0.9	3.4
2	ARIMA	1.9	1.4	0.9	16.1
3	PM25-CBL [28]	1.2	1.0	0.8	3.1
4	CDL	0.9	0.9	0.7	2.9

Kết quả thực nghiệm tại bảng cho thấy mô hình CDL thể hiện hiệu suất vượt trội so với các mô hình khác như CNN-LSTM, PM25-CBL, và ARIMA trong việc dự đoán giá trị cho bộ dữ liệu Air Quality HCMC 2020. Cụ thể, khi đánh giá qua các chỉ số MSE (Mean Squared Error), RMSE (Root Mean Squared Error), MAE (Mean Absolute Error) và MAPE (Mean Absolute Percentage Error), mô hình CDL luôn đạt các giá trị thấp nhất, chứng tỏ độ chính xác cao và sai số nhỏ. Cụ thể, mô hình CDL đạt MSE = 0.9, RMSE = 0.9, MAE = 0.7, và MAPE = 2.9%. Những chỉ số này cho thấy CDL không chỉ có sai số tuyệt đối và bình phương trung bình thấp mà còn mang lại hiệu quả dự đoán chính xác với tỷ lệ phần trăm sai lệch thấp, giúp tối ưu hóa các chiến lược dự báo và phân tích. Trong khi đó, mô hình CNN-LSTM cho kết quả với các chỉ số MSE = 1.4, RMSE = 1.1, MAE = 0.9 và MAPE = 3.4%, mặc dù mô hình này có các chỉ số khá tốt, nhưng không thể vượt qua mô hình CDL về độ chính xác và hiệu quả. Mô hình PM25-CBL cũng cho thấy hiệu suất khá ổn với các chỉ số MSE = 1.2, RMSE = 1.0, MAE = 0.8, và MAPE = 3.1%. Tuy nhiên, mặc dù PM25-CBL vượt trội hơn so với CNN-LSTM và ARIMA, nhưng vẫn không thể đạt được mức độ chính xác như CDL, đặc biệt là trong các chỉ số MAE và MAPE, nơi mà CDL thể hiện sự vượt trội rõ rệt. Với những kết quả trên, nghiên cứu kết luận rằng CDL là mô hình tối ưu trong việc dự đoán các giá trị trong bộ dữ liệu Air Quality HCMC 2020, mang lại độ chính xác cao nhất và các chỉ số hiệu suất tốt nhất. Mô hình này có khả năng dự đoán chính xác hơn các mô hình truyền thống như ARIMA và các mô hình học máy kết hợp như CNN-LSTM. Điều này làm nổi bật tầm quan trọng của việc chọn lựa mô hình phù hợp khi giải quyết các bài toán dự đoán và phân tích dữ liệu. Với hiệu suất vượt trội và khả năng xử lý dữ liệu phức tạp, mô hình CDL được đề xuất là lựa chọn ưu tiên trong các ứng dụng thực tế và nghiên cứu trong tương lai, đặc biệt là trong các bài toán yêu cầu độ chính xác cao và tính ổn định trong dự đoán.

V. KẾT LUẬN

Việc cải tiến mô hình CNN và Bi-LSTM trong việc giảm số lượng tham số và xử lý dữ liệu mất cân bằng là một cách tiếp cận quan trọng nhằm tối ưu hóa mô hình học sâu, làm giảm tài nguyên tính toán và đồng thời cải thiện hiệu suất đối với dữ liệu không cân bằng. Các chiến lược như giảm kích thước đầu vào, sử dụng các lớp CNN nhẹ, áp dụng các phương pháp cân bằng lớp và điều chỉnh loss function sẽ giúp tăng hiệu quả của mô hình. Đề tài trình bày hướng giải quyết của hai bài toán dự đoán khi sử dụng mô hình CNN và Bi-LSTM. Mô hình CLS dự đoán thành tích của học sinh, là vấn đề đã thu hút sự quan tâm của nhiều người trong và ngoài ngành. Kết quả đạt được và cũng là mục tiêu của bài báo này là tìm ra các đặc trưng có tác động đến đối tượng dự đoán, đó là điểm G3 trong 3 bộ dữ liệu Portuguese Language Dataset, Student Performance in Mathematics Dataset and outputClass data sets with xAPI. Kết quả đạt được khá tốt với mô hình CNN kết hợp với BiLSTM. Đặc biệt, chúng tôi không chỉ tinh chỉnh và cải thiện thành công mô hình thuật toán dự đoán để giải quyết các bài toán đa lớp mà còn đưa ra kết quả tốt nhất so với các mô hình thuật toán dự đoán khác bằng cách sử dụng các kỹ thuật lấy mẫu quá mức với SMOTE, ADASYN và Borderline-SMOTE SVM. Bài toán thứ hai chúng tôi đề xuất mô hình CDL để cải thiện kết quả của mô hình PM25-CBL trong nghiên cứu [31] trên tập dữ liệu Air Quality HCMC. Mô hình PM25-CBL bao gồm ba mô đun: CNN, Bi-LSTM và Kết nối đầy đủ (Fully connected). Ưu điểm của mô hình là khả năng dự báo chính xác cao nhưng mô hình có khuyết điểm là thời gian huấn luyện lâu, hầu như gấp đôi so với mô hình CNN thông thường. Kiến trúc mô hình PM25-CDSCBL là sự kết hợp giữa CNN với Depthwise Separable Convolutions và Bi-LSTM. Depthwise Separable Convolutions được kết hợp với CNN để giúp giảm số lượng tham số của mô hình PM25-CBL từ 214,081 xuống 67,299 giúp cho quá trình huấn luyện mô hình đạt hiệu quả về mặt thời gian tốt hơn, nhưng vẫn giữ được hiệu suất dự đoán cao. Đồng thời mô hình áp dụng Bi-LSTM dựa trên quá trình học chuyển đổi để dự đoán chất lượng không khí trong tình huống thiếu dữ liệu. Kết quả chỉ ra CDL đạt được RMSE thấp nhất khi so sánh với các phương pháp CNN-LSTM, PM25-CBL, ARIMA khi thử nghiệm trong môi trường mới.

VI. TÀI LIỆU THAM KHẢO

[1] N. Shlezinger, Y. C. Eldar and S. P. Boyd, "Model-Based Deep Learning: On the Intersection of Deep Learning and Optimization," in IEEE Access, vol. 10, pp. 115384-115398, 2022, doi: 10.1109/ACCESS.2022.3218802.

[2] El Nahhas, Omar SM, et al. "From whole-slide image to biomarker prediction: end-to-end weakly supervised deep learning in computational pathology." *Nature Protocols* 20.1 (2025): 293-316.

[3] Razak Olu-Ajayi, Hafiz Alaka, Ismail Sulaimon, Funlade Sunmola, Saheed Ajayi (2022). Building energy consumption prediction for residential buildings using deep learning and other machine learning techniques, Journal of Building Engineering, Volume 45, 2022,103406

[4] Hajek, Petr, and Michal Munk. "Corporate financial distress prediction using the risk-related information content of annual reports." *Information Processing & Management* 61.5 (2024): 103820.

[5] Chu Zhang, Huixin Ma, Lei Hua, Wei Sun, Muhammad Shahzad Nazir, Tian Peng (2022). "An evolutionary deep learning model based on TVFEMD, improved sine cosine algorithm, CNN and BiLSTM for wind speed prediction", Energy, Volume 254, Part A, 2022,124250.

[6] A. S. Aljaloud *et al*., "A Deep Learning Model to Predict Student Learning Outcomes in LMS Using CNN and LSTM," in *IEEE Access*, vol. 10, pp. 85255-85265, 2022, doi: 10.1109/ACCESS.2022.3196784.

[7] Liu, Tianming and Siegel, Eliot and Shen, Dinggang (2022), Deep Learning and Medical Image Analysis for COVID-19 Diagnosis and Prediction, Annual Review of Biomedical Engineering, Volumn 24, 1, 179-201

[8] T. P. T. Nguyen, N. L.-T. Nguyen, and T. H. Duong, "Deepo: An ontology-based deep learning system for disease prediction," Int. J. Intell. Inf. Database Syst., vol. 15, no. 2, pp. 166–182, 2022, doi: 10.1504/IJIIDS.2022.121897

[9] Bihter Das, Ömer Osman Dursun, Suat Toraman (2022). Prediction of air pollutants for air quality using deep learning methods in a metropolitan city, Urban Climate, Volume 46, 2022,101291

[10] Ali Agga, Ahmed Abbou, Moussa Labbadi, Yassine El Houm, Imane Hammou Ou Ali (2022), CNN-LSTM: An efficient hybrid deep learning architecture for predicting short-term photovoltaic power production, Electric Power Systems Research, Volume 208, 2022,107908

[11] Rodney Rick, Lilian Berton, Energy forecasting model based on CNN-LSTM-AE for many time series with unequal lengths, Engineering Applications of Artificial Intelligence, Volume 113, 2022, 104998,

[12] Vo, M.T., Vo, A.H., Nguyen, T., Sharma, R. and Le, T. (2021), "Dealing with the class imbalance problem in the detection of fake job descriptions", Computers, Materials and Continua, Vol. 68 No. 1, pp. 521-535.

[13] Haque, R.; Islam, N.; Islam, M.; Ahsan, M.M. A Comparative Analysis on Suicidal Ideation Detection Using NLP, Machine, and Deep Learning. *Technologies* 2022, *10*, 57. https://doi.org/10.3390/technologies10030057

[14] Tony Salloom, Okyay Kaynak, Xinbo Yu, Wei He (2022). "Proportional integral derivative booster for neural networks-based time-series prediction: Case of water demand prediction", Engineering Applications of Artificial Intelligence, Volume 108, 2022, 104570

[15] Godahewa, R., Bergmeir, C., Webb, G.I., Hyndman, R.J., Montero-Manso, P.: Monash time series forecasting archive. In: NeurIPS Track on Datasets and Benchmarks (2021).

[16] Hansika Hewamalage, Christoph Bergmeir, Kasun Bandara, “Recurrent Neural Networks for Time Series Forecasting: Current status and future directions”, International Journal of Forecasting, Volume 37, Issue 1, 2021, 388-427

[17] Sahoo, B.B., Jha, R., Singh, A. et al. Long short-term memory (LSTM) recurrent neural network for low-flow hydrological time series forecasting. Acta Geophys. 67, 1471–1481 (2019). https://doi.org/10.1007/s11600-019-00330-1

[18] H. A. Vo, T. Nguyen and T. Le, “Brent oil price prediction using Bi-LSTM network,” Intelligent Automation & Soft Computing, vol. 26, no. 6, pp. 1307–1317, 2020.

[19] S. Siami-Namini, N. Tavakoli and A. S. Namin, "The Performance of LSTM and BiLSTM in Forecasting Time Series," *2019 IEEE International Conference on Big Data (Big Data)*, Los Angeles, CA, USA, 2019, pp. 3285-3292, doi: 10.1109/BigData47090.2019.9005997.

[20] Sakshi Khullar, Nanhey Singh, “Water quality assessment of a river using deep learning Bi-LSTM methodology: forecasting and validation” (2022), Environmental Science and Pollution Research, 29, 12875-12889

[21] M. T. Vo, D. Vu, H. Nguyen, H. Bui and T. Le, “Predicting monthly household water consumption,” in Proc. of Int. Conf. on Computing and Communication Technologies, Ho Chi Minh City, Vietnam, pp. 720– 724, 2022.

[22] H. Srivastava and K. Sarawadekar, "A Depthwise Separable Convolution Architecture for CNN Accelerator," 2020 IEEE Applied Signal Processing Conference (ASPCON), Kolkata, India, 2020, pp. 1-5, doi: 10.1109/ASPCON49795.2020.9276672.

[23] Liu, B.; Zou, D.; Feng, L.; Feng, S.; Fu, P.; Li, J. An FPGA-Based CNN Accelerator Integrating Depthwise Separable Convolution. *Electronics* 2019, *8*, 281.

[24] S. Ma, W. Liu, W. Cai, Z. Shang and G. Liu, "Lightweight Deep Residual CNN for Fault Diagnosis of Rotating Machinery Based on Depthwise Separable Convolutions," in IEEE Access, vol. 7, pp. 57023-57036, 2019, doi: 10.1109/ACCESS.2019.2912072.

[25] M. Kannan, "An enhancement of machine learning model performance in disease prediction with synthetic data generation," Scientific Reports, vol. 15, no. 33482, pp. 1–12, 2025. [Online]. Available: https://www.nature.com/articles/s41598-025-15019-3

[26] J. H. Joloudari, A. Marefat, M. A. Nematollahi, S. S. Oyelere, and S. Hussain, "Effective class-imbalance learning based on SMOTE and convolutional neural networks," Applied Sciences, vol. 13, no. 6, p. 4006, 2023. [Online]. Available: https://doi.org/10.3390/app13064006

[27] D. Liu, "Deep attention SMOTE: Data augmentation with a deep attention synthetic minority over-sampling technique," Journal of Computational Science, vol. 68, p. 101622, 2023. [Online]. Available: https://doi.org/10.1016/j.jocs.2023.101622

[28] P. Y. Wong, H. Y. Lee, Y. C. Chen, Y. T. Zeng, Y. R. Chern et al., “Using a land use regression model with machine learning to estimate ground level PM2.5,” Environmental Pollution, vol. 277, pp. 116846, 2021.

[29] Vo M. T., Vo H. A., Bui H., Le T.: A Hybrid Deep Learning Approach for PM2.5 Concentration Prediction in Smart Environmental Monitoring. Intelligent Automation & Soft Computing, Volume 36 (2023) 3029-3041 http://www.techscience.com/iasc/v36n3/51909

[30] Elaf Amrieh, Thair Hamtini, and Ibrahim Aljarah. Mining educational data to predict student’s academic performance using ensemble methods. International Journal of Database Theory and Application, 9:119–136, 09 2016.17

[31] P. Cortez, "Student Performance [Dataset]," UCI Machine Learning Repository, 2014. [Online]. Available: https://archive.ics.uci.edu/dataset/320/student%2Bperformance

[32] Dataset on Air Quality in Vietnam in 2020. https://data.opendevelopmentmekong.net/dataset/timelines-dataset-on-air-quality-in-vietnam, accessed on February 20, 2025.

AN IMPROVED CNN WITH BI-LSTM MODEL FOR PARAMETER REDUCTION AND HANDLING IMBALANCED DATA

Nguyen Thi Phuong Trang, Nguyen Duc Cuong

ABSTRACT— This paper presents improvements to a deep learning model that combines CNN and Bi-LSTM to address two key issues: imbalanced data and computational complexity. To address the issue of imbalanced data, techniques such as SMOTE, undersampling, and class weight adjustments are employed, thereby enhancing the accuracy of minority classes in the dataset. Experimental results on the UCI Student Performance dataset demonstrate the model's effectiveness in predicting student academic performance. Meanwhile, to reduce computational complexity, Depthwise Separable Convolutions are employed to decrease the number of model parameters. Results are presented through the air quality prediction task for PM2.5 in Ho Chi Minh City, showing the efficiency in saving computational resources without sacrificing prediction performance.

Keywords—Bi-LSTM, CNN, Deep learning, Depthwise Separable Convolutions

Nguyễn Thị Phương Trang là thạc sĩ Công nghệ thông tin. Các hướng nghiên cứu tập trung về khai thác dữ liệu và học máy, deep learning. Các chủ đề nghiên cứu tập trung vào các mô hình khuyến nghị như Dự đoán kết quả học của sinh viên, dự đoán giá dầu, xử lý mất cân bằng dữ liệu.

Nguyễn Đức Cường là tiến sĩ Công nghệ Thông tin. Các hướng nghiên cứu tập trung vào machine learning, khai thác dữ liệu, dự báo thiên tai và xử lý dữ liệu thiếu. Các ứng dụng bao gồm nhận dạng thực thể và mô hình hóa người dùng.