

DOI:10.22144/ctu.jos.2024.345

ỨNG DỤNG MẠNG NEURAL NHÂN TẠO – MÔ HÌNH DENSENET TRONG DỰ ĐOÁN ĐẶC TÍNH ĐIỆN TỬ CỦA VẬT LIỆU

Hà Thu Hoàng¹, Đặng Thị Hồng Nhan² và Nguyễn Thành Tiên^{1*}

¹Khoa Khoa học Tự nhiên, Trường Đại học Cần Thơ

²Khoa Sau Đại học, Trường Đại học Cần Thơ

*Tác giả liên hệ (Corresponding author): nttien@ctu.edu.vn

Thông tin chung (Article Information)

Nhận bài (Received): 19/04/2024

Sửa bài (Revised): 28/06/2024

Duyệt đăng (Accepted): 04/08/2024

Title: Application of DenseNet Neural Network in Predicting Electronic Properties of Materials

Author(s): Ha Thu Hoang¹, Dang Thi Hong Nhan² and Nguyen Thanh Tien^{1*}

Affiliation(s): Can Tho University

TÓM TẮT

Nghiên cứu này trình bày một quy trình làm việc để khám phá khoa học vật liệu dựa trên dữ liệu thông qua thuật toán học máy (ML) - mạng Neural nhân tạo (ANN). Trong đó, tập trung vào việc dự đoán năng lượng vùng cấm (E_{gap}) của vật liệu, một tính chất điện tử quan trọng trong vật lý chất rắn. Bằng cách sử dụng các kỹ thuật học máy có giám sát và tập dữ liệu lớn, mô hình DenseNet được tối ưu hóa để dự đoán chính xác giá trị E_{gap} . Nghiên cứu đã chứng minh tính hiệu quả của mô hình DenseNet thông qua các chỉ số đánh giá như hệ số xác định (R^2), sai số tuyệt đối trung bình (MAE) và sai số căn quân phương (RMSE). Kết quả cho thấy mô hình đạt hiệu suất tốt nhất với R^2 là 0.7924 trên tập huấn luyện và 0.6682 trên tập kiểm định. Nghiên cứu này không chỉ đóng góp vào việc phát triển các phương pháp tính toán hiệu quả cho khoa học vật liệu mà còn mở ra những hướng nghiên cứu mới trong việc khám phá và thiết kế vật liệu mới.

Từ khóa: Năng lượng vùng cấm, máy học, mạng Neural nhân tạo

ABSTRACT

In this study, we present a workflow for data-driven materials science discovery using a machine learning (ML) algorithm - an artificial neural network (ANN). The focus is on predicting the energy band gap (E_{gap}) of materials, a key electronic property in solid-state physics. By employing supervised machine learning techniques and a large dataset, the DenseNet model is optimized to accurately predict E_{gap} values. The effectiveness of the DenseNet model is demonstrated through evaluation metrics such as the coefficient of determination (R^2), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). The results show that the model achieves its best performance with an R^2 of 0.7924 on the training set and 0.6682 on the validation set. This research not only contributes to the development of efficient computational methods for materials science but also opens up new avenues in the discovery and design of novel materials.

Keywords: Energy band gap, machine learning, Artificial neural network

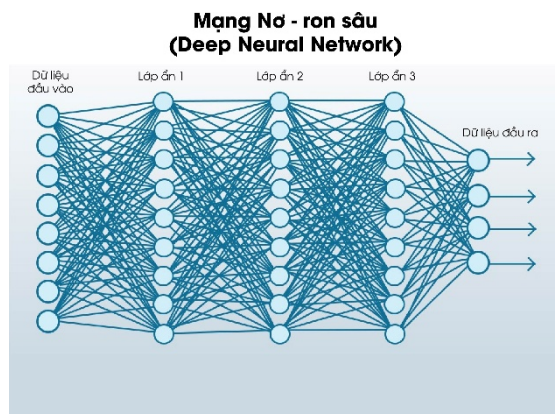
1. ĐẶT VẤN ĐỀ

Năng lượng vùng cấm (E_{gap}) là một đặc tính điện tử cơ bản của vật liệu, quyết định khả năng ứng dụng của chúng trong các thiết bị điện tử và quang điện tử. Việc xác định E_{gap} của vật liệu bằng các phương pháp khác nhau, chẳng hạn như tính toán lý thuyết phiếm hàm mật độ (DFT) và mô hình học máy (ML) đã được sử dụng để dự đoán E_{gap} với độ chính xác cao; đóng vai trò then chốt trong việc thiết kế và phát triển các thiết bị tiên tiến, đặc biệt là trong lĩnh vực năng lượng tái tạo và công nghệ bán dẫn (Mondal et al., 2023; Lin et al., 2023). Tuy nhiên, việc đo đạc E_{gap} bằng thực nghiệm thường tốn kém, phức tạp và mất nhiều thời gian (Manzotti et al., 2022). Nghiên cứu về các siêu mạng quang điện tử cho thấy vùng cấm truyền dẫn rộng hơn trong các cấu trúc phức tạp, mở ra tiềm năng cho các bộ lọc năng lượng chính xác và các thiết bị quang điện tử (Wang et al., 2022). Mặt khác, các phương pháp tính toán lý thuyết, mặc dù có độ chính xác cao nhưng lại đòi hỏi nguồn lực tính toán lớn và thường chỉ giới hạn trong các hệ vật liệu đơn giản (Chowdhury et al., 2022).

Trong bối cảnh đó, sự tiến bộ của khoa học dữ liệu và học máy (ML) đã thực sự cách mạng hóa việc dự đoán các đặc tính vật liệu như năng lượng vùng cấm (E_{gap}) bằng cách tận dụng mạng Neural nhân tạo (ANN) để phân tích các tập dữ liệu lớn và cấu trúc vật liệu phức tạp (Askanazi & Grinberg, 2023). Các mô hình ML, đặc biệt là ANN, vượt trội trong việc nắm bắt các mối quan hệ phức tạp giữa cấu trúc và tính chất của vật liệu, cho phép dự đoán E_{gap} chính xác và hiệu quả mà không cần các thí nghiệm tốn kém hoặc các tính toán lý thuyết phức tạp (Smith., 2023). ANN là một mô hình máy học (ML), được xây dựng mô phỏng theo nguyên lý sinh học của bộ não con người (Cole, 2020). Chính vì vậy thuật toán ANN có khả năng xử lý song song với tốc độ xử lý nhanh, có khả năng dạy học thích nghi, nó thích ứng trong quá trình tự điều chỉnh trong điều chỉnh tự động. Đặc biệt, ANN có thể học hỏi từ dữ liệu để xử lý các bài toán phi tuyến. Mặc dù có những thách thức do tính không thể giải thích (*the black-box nature*) của các mô hình ML, nghiên cứu gần đây nhấn mạnh tầm quan trọng của việc hiểu độ tin cậy của các dự đoán ML, đặc biệt là đối với các tập dữ liệu nhỏ chiếm ưu thế hơn so với các tập dữ liệu lớn, độ tin cậy và độ chính xác của các mô hình dự đoán cho các tập dữ liệu lớn có thể được tăng cường bằng cách xây dựng bao lồi (*constructing convex hulls*) trong không gian đặc trưng (Askanazi et al., 2023). Tiến bộ này nhấn mạnh tiềm năng của ML trong việc thúc đẩy khám phá và đổi mới vật

liệu, đặc biệt là trong các lĩnh vực như lưu trữ năng lượng, xúc tác, điện tử. Điều này cho phép dự đoán E_{gap} của các vật liệu mới mà không cần thực hiện các thí nghiệm tốn kém hay tính toán lý thuyết phức tạp.

Trong các thuật toán ANN, mô hình DenseNet được biết đến với các kết nối dày đặc tạo điều kiện cho việc học các đặc trưng ở các mức độ trừu tượng khác nhau, đã cho thấy tiềm năng trong các lĩnh vực khác nhau như cảm biến và nhận dạng hoạt động (Liu et al., 2023). Tiềm năng của nó mở rộng đến việc dự đoán E_{gap} của vật liệu, được chứng minh bằng hiệu quả cao và vượt trội hơn các mô hình khác như ResNet (Irawan et al., 2022; Hu et al., 2023). Với khả năng này, DenseNet có tiềm năng trở thành một công cụ mạnh mẽ trong việc dự đoán E_{gap} của vật liệu.

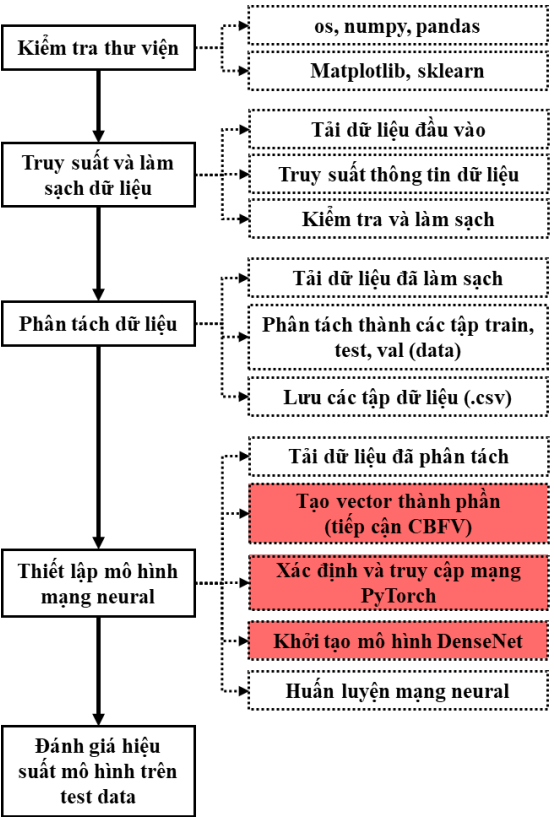


Hình 1. Trực quan hóa thuật toán mạng Neural nhân tạo

Mặc dù đã có nhiều nghiên cứu về ứng dụng học máy trong khoa học vật liệu, tuy nhiên việc tiếp cận phương pháp học sâu – mạng Neural nhân tạo (Deep Neural Network – Hình 1) để dự đoán E_{gap} vẫn còn là một cách tiếp cận mới và đầy hứa hẹn. Do đó, nghiên cứu này nhằm mục đích khám phá tiềm năng của mô hình DenseNet trong phương pháp học sâu nhằm đánh giá khả năng dự đoán E_{gap} của vật liệu trên mô hình đã đào tạo. Mô hình DenseNet sẽ được huấn luyện và tối ưu hóa trên một tập dữ liệu lớn về vật liệu, sau đó mô hình được đánh giá hiệu suất thông qua các chỉ số thống kê như hệ số xác định (R^2), sai số tuyệt đối trung bình (MAE) và sai số căn quân phương (RMSE). Nghiên cứu này không chỉ góp phần mở rộng ứng dụng của DenseNet trong lĩnh vực khoa học vật liệu mà còn cung cấp một công cụ hữu ích cho việc thiết kế và khám phá vật liệu mới.

2. KHUNG QUY TRÌNH HỌC MÁY MẠNG NEURAL NHÂN TẠO DỰA TRÊN THUẬT NGỮ HỌC TẬP SÂU

Trong đề tài này, một khung thuật toán cho mạng lưới thần kinh đơn giản được đề xuất và triển khai thành công để mô hình hóa dữ liệu năng lượng vùng cấm (E_{gap}) bằng phương pháp học sâu có giám sát.



Hình 2.1. Khung thuật toán mạng Neural nhân tạo đề xuất đào tạo dữ liệu E_{gap} bằng phương pháp học sâu có giám sát

Khung quy trình thuật toán của mô hình DenseNet được đề xuất nhằm đạt hiệu suất tốt hơn ANN truyền thống trong các tác vụ học, ở các đặc điểm:

- **Tạo vector thành phần (tiếp cận CBFV):** bước đặc trưng của quy trình DenseNet, không có trong quy trình ANN truyền thống. CBFV được tiếp cận để chuyển đổi thông tin về thành phần hóa học của vật liệu thành một vector đặc trưng, làm đầu vào cho mô hình.
- **Xác định và truy cập mạng PyTorch:** khung quy trình đề xuất chỉ rõ việc sử dụng PyTorch - một framework học sâu phổ biến hiện nay, để xây dựng và huấn luyện mô hình DenseNet.

- **Khởi tạo mô hình DenseNet:** thay vì khởi tạo một mô hình ANN truyền thống, quy trình này khởi tạo một mô hình DenseNet.

Quá trình xây dựng chương trình thuật toán chia làm 5 bước cơ bản:

- Bước 1: Kiểm tra thư viện
- Bước 2: Truy xuất và xử lý dữ liệu
- Bước 3: Phân tách dữ liệu
- Bước 4: Thiết lập mô hình mạng Neural nhân tạo
- Bước 5: Đánh giá lại hiệu suất mô hình trên tập dữ liệu thử nghiệm

2.1. Kiểm tra thư viện

Trước tiên, quá trình xây dựng mô hình ML trong ngôn ngữ lập trình Python cần một số thư viện chính, chính những nguồn thư viện này giúp hỗ trợ tốt hơn cho thực thi mô hình. Quá trình kiểm tra thư viện nhằm đảm bảo trong quá trình hoạt động thư viện đã được cài đặt trước, giúp cho mô hình hoạt động tốt hơn. Một lưu ý khi cài đặt và sử dụng thư viện cho các dự án ML là một phân vùng môi trường riêng cần được tạo cho dự án đó. Dưới đây là kết quả của các thư viện cho quá trình cài đặt và thông báo trước khi tiến hành xây dựng mô hình ML tiên đoán E_{gap} :

- [OK] numpy version 1.18.5
- [OK] pandas version 1.0.5
- [OK] matplotlib version 3.2.2
- [OK] seaborn version 0.10.1
- [OK] sklearn version 0.22.2.post1
- [OK] scipy version 1.7.3
- [OK] tqdm version 4.66.1
- [OK] jupyter_client version 7.4.9
- [OK] ipywidgets version 8.1.1
- [OK] torch version 1.3.1

2.2. Truy xuất và xử lý dữ liệu

Bước đầu tiên của một việc xây dựng mô hình học máy là chọn tập dữ liệu ta sẽ thực hiện đào tạo. Có nhiều kho lưu trữ dữ liệu dành riêng cho khoa học vật liệu. Sử dụng gói thư viện Pandas với thao tác `pd.read_csv()` để đọc dữ liệu đầu vào từ tệp excel vào khung dữ liệu. Dữ liệu đầu vào có 19039 mẫu được chia ra thành 2 cột tương ứng là: ['formula', 'target']. Cột formula là cột công thức hoá học của

các mẫu nghiên cứu, cột target là cột tương ứng với giá trị E_{gap} .

Bảng 2.1. Dữ liệu đầu vào từ tập dữ liệu E_{gap}

STT	formula	target	STT	formula	target
0	N ₁ Nd ₁	0.0000	3998	Cu ₁ In ₁ Mo ₂ O ₈	0.7201
1	Dy ₁ Rh ₂	0.0000	...	...	...
2	Ga ₁ S ₁	1.7100	7997	As ₂ O ₇ Sr ₂	3.3287
3	Tb ₁ Zn ₁	0.0000	...	...	...
...	...	...	11999	Ca ₁ O ₁ S ₁ Zn ₁	3.1552
291	Bi ₂ Fe ₄ O ₉	1.6087	...	...	...
...	...	...	...	...	...
2999	Li ₅ O ₅ Sb ₁	2.9349	19038	Cs ₃ F ₇ Ge ₁	4.3375

Quá trình thu thập dữ liệu từ các nguồn khác nhau với dữ liệu thu thập đều là dữ liệu tiền xử lý (*dữ liệu chưa qua làm sạch của các tính toán mô phỏng hoặc lý thuyết*), do đó, tập dữ liệu có thể có những sai sót hoặc lỗi trong quá trình nhập liệu hoặc tính toán dẫn đến các giá trị E_{gap} hoặc công thức hóa học không hợp lệ. Việc loại bỏ các giá trị không phù hợp này trong quá trình làm sạch dữ liệu là cần thiết để đảm bảo tính chính xác và độ tin cậy của kết quả nghiên cứu.

Các cấu trúc hóa học có giá trị E_{gap} âm, không phải số (NaN) hoặc trùng lặp được thực hiện loại bỏ (làm sạch). Sau quá trình xử lý dữ liệu, dữ liệu được tiến hành lưu dưới dạng tập .csv cho bước tiếp theo. Thư viện Pandas được sử dụng cho phép lưu dữ liệu dưới dạng tệp .csv thông qua cú pháp: *df.to_csv()*. Ban đầu dữ liệu là 19039, sau quá trình xử lý dữ liệu, thu được dữ liệu với 9609 mẫu tương ứng với 2 cột 'formula' và 'target'.

2.3. Phân tách dữ liệu

Từ dữ liệu đã làm sạch, thông thường, dữ liệu chỉ cần được phân chia bằng hàm sklearn. Hàm scikit-learn *train_test_split* chia ngẫu nhiên một tập dữ liệu thành tập dữ liệu huấn luyện và tập dữ liệu kiểm tra.

Bảng 2.2. Thông số phân tách tập dữ liệu E_{gap} đã làm sạch

	train split	validation split	test split
Tỷ lệ phân tách tập dữ liệu (%)	70	20	10
Lượng data tương ứng (data)	6726	1922	961

Ở đây chương trình *train_test_split* được sử dụng để phân chia dữ liệu của mình thành các tập dữ liệu "đào tạo" và "kiểm tra" trước tiên, sau đó hàm

này được sử dụng lại để phân chia tiếp tục dữ liệu "đào tạo" thành các tập dữ liệu "đào tạo" và "xác thực".

2.4. Thiết lập mô hình mạng Neural nhân tạo

Khi tiến hành xây dựng mô hình học sâu, các thư viện cần thiết cần được khai báo, sau đó là các tập dữ liệu đầu vào train-val-test đã được phân tách được đọc. Tiếp theo, mô hình học sâu được xây dựng theo dữ liệu:

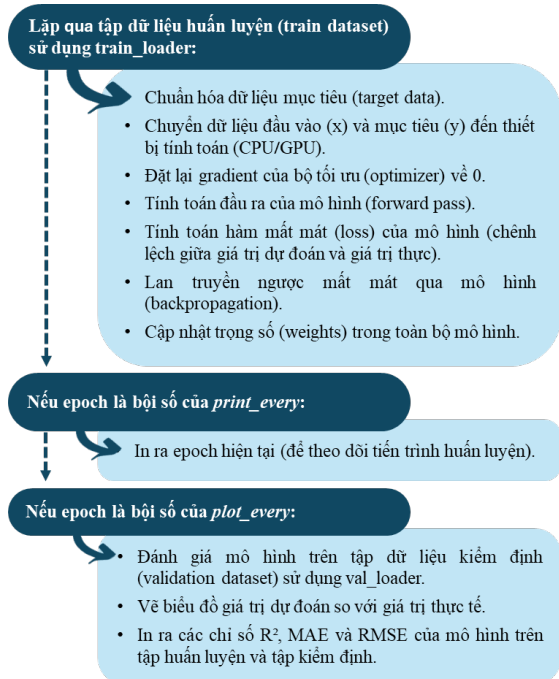
- Train data: bao gồm các dữ liệu được đào tạo trực tiếp để xây dựng mô hình.
- Test data: gồm các dữ liệu được dùng để đánh giá hiệu quả mô hình. Dữ liệu thử nghiệm là điều kiện cần để một mô hình hiệu quả là kết quả đánh giá trên cả train và test đều cao.
- Val data: được sử dụng trong việc lựa chọn các siêu tham số mô hình. Một mô hình hoạt động hiệu quả trên dữ liệu train chưa chắc đã hoạt động hiệu quả trên dữ liệu test. Để tăng hiệu quả của mô hình trên dữ liệu test, một tập dữ liệu nữa được gọi là tập dữ liệu xác thực (val) cần được dùng thêm.

Xây dựng mô hình mạng Neural nhân tạo bằng phương pháp học sâu có giám sát. Việc lựa chọn mô hình ML rất quan trọng, các mô hình sẽ quyết định đến xu hướng phát triển và dự đoán dữ liệu cũng như tính hiệu quả của mô hình cho việc dự đoán.

Mô hình được đào tạo thông qua việc tạo các vector đặc trưng thành phần bằng tiếp cận CBFV (Composition-Based Feature Vectorization). CBFV là một công cụ mạnh mẽ trong khoa học vật liệu tính toán, đặc biệt là trong các ứng dụng học máy. Nó cho phép chuyển đổi thông tin về thành phần hóa học của vật liệu thành một vector đặc trưng (*feature vector*) đại diện cho vật liệu đó (Yadav et al., 2024). Vector đặc trưng này chứa thông tin về các thuộc tính của các nguyên tố cấu thành nên vật liệu, chẳng hạn như số hiệu nguyên tử, khối lượng nguyên tử, độ âm điện, bán kính nguyên tử, và các thông tin khác. Các đặc trưng này được tính toán dựa trên các tính chất của các nguyên tố trong bảng tuần hoàn và có thể được sử dụng làm đầu vào cho các mô hình học máy để dự đoán các tính chất của vật liệu. Cụ thể trong nghiên cứu này, vector đặc trưng CBFV được sử dụng làm đầu vào cho mô hình DenseNet để dự đoán E_{gap} .

Trong lĩnh vực học sâu, epoch là thông số đại diện cho một lần duyệt qua toàn bộ tập dữ liệu huấn luyện trong quá trình đào tạo mô hình. Nó đóng vai trò quan trọng trong việc xác định hiệu suất của mô hình bằng cách giúp ngăn ngừa tình trạng quá khớp

(*overfitting*) hoặc kém khớp (*underfitting*) dữ liệu. Số lượng epoch được chọn để huấn luyện là một siêu tham số ảnh hưởng đến quá trình học và khả năng khái quát hóa của mô hình (Niharika et al., 2013; Cui et al., 2022).



Hình 2.2. Quy trình huấn luyện (training procedure) cho một mạng Neural nhân tạo (neural network)

Các nhà nghiên cứu đã khám phá những phương pháp sáng tạo như “Học tập theo quy trình Gaussian Tiến hóa Epoch (GPGL)” của Jiabao Cui et al. (2022), các tác giả đề xuất một phương pháp học máy mới, trong đó mô hình GP (*Gaussian Process*) được sử dụng để hướng dẫn quá trình cập nhật trọng số của mạng Neural nhân tạo. Mô hình GP này không cố định mà “tiến hóa” qua từng epoch, nghĩa là nó liên tục cập nhật và điều chỉnh để phù hợp hơn với dữ liệu huấn luyện (Cui et al., 2022). Hiểu rõ và tối ưu hóa tham số epoch là điều cần thiết để đạt được hiệu suất và độ chính xác tối ưu trong các nhiệm vụ học sâu (Saahil et al., 2020). Trong đề tài này, quá trình huấn luyện được thực hiện với tối đa là 500 epoch.

Đồng thời, hiệu chỉnh mô hình bằng cách sử dụng tập dữ liệu kiểm định (*validation set*) là một kỹ thuật phổ biến trong học máy để tránh *overfitting* (quá khớp) và đảm bảo khả năng khái quát hóa của mô hình. Các nghiên cứu trước đây đã chứng minh tính hiệu quả của phương pháp này trong nhiều lĩnh vực, bao gồm cả khoa học vật liệu (Faber et al., 2017).

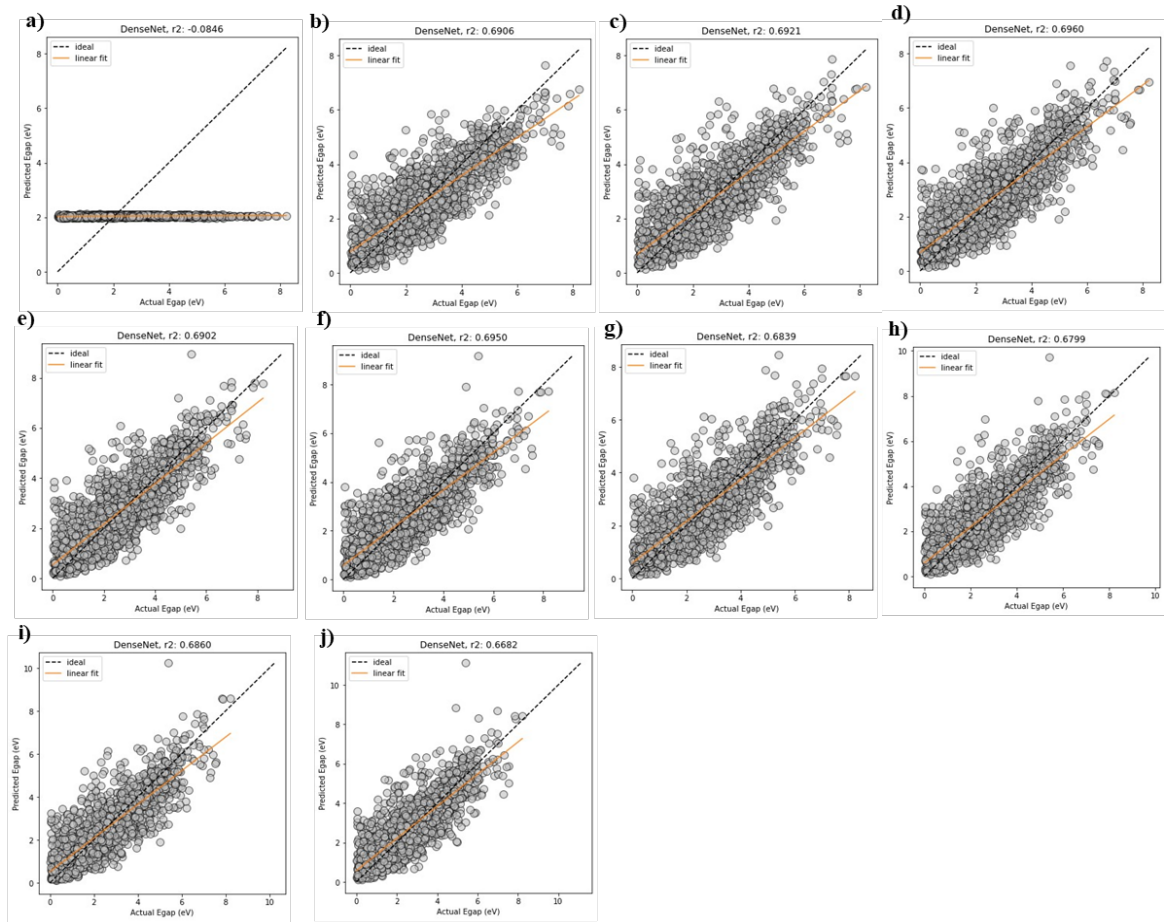
2.5. Đánh giá lại hiệu suất mô hình trên tập dữ liệu thử nghiệm

Sau khi mô hình DenseNet được huấn luyện xong, hiệu suất của mô hình trên tập dữ liệu thử nghiệm (*test dataset*) được đánh giá để đảm bảo rằng mô hình có khả năng khái quát hóa tốt cho dữ liệu mới, chưa từng được nhìn thấy trong quá trình huấn luyện và quá trình hiệu chỉnh.

3. KẾT QUẢ VÀ THẢO LUẬN

Các mô hình trong Hình 3.1 được huấn luyện trên dữ liệu E_{gap} đầu vào, với 70% dữ liệu được sử dụng để huấn luyện, 20% để kiểm tra và 10% để xác thực. Trục tung và trục hoành lần lượt biểu diễn giá trị E_{gap} dự đoán và giá trị E_{gap} thực tế. Đường đứt nét màu đen là đường lý tưởng của mô hình, đường nét liền màu cam là đường khớp dữ liệu của mô hình. Sự khớp giữa hai đường này càng cao, mô hình càng chính xác.

Số liệu của các mô hình có thể được quan sát ở Bảng 3.1. Mô hình DenseNet thể hiện khả năng dự đoán tốt với R^2 trên tập huấn luyện (R^2_{train}) và tập kiểm định (R^2_{val}) đều tăng dần theo số epoch. Điều này cho thấy mô hình đang học và cải thiện khả năng dự đoán qua quá trình huấn luyện. Giá trị R^2 đạt giá trị cao nhất là 0.7924 trên tập huấn luyện và 0.6682 trên tập kiểm định ở epoch 420-440, cho thấy mô hình có khả năng khái quát hóa tốt. MAE trên cả hai tập dữ liệu giảm dần, đặc biệt là trong giai đoạn đầu của quá trình huấn luyện (epoch 0-100). Điều này cho thấy mô hình đang giảm thiểu sai số dự đoán trung bình. Tương tự MAE, RMSE cũng giảm dần, chỉ ra rằng mô hình đang giảm thiểu độ lệch giữa giá trị dự đoán và giá trị thực. Trong quá trình huấn luyện, mô hình dường như hội tụ sau khoảng 200 epoch, khi các chỉ số R^2 , MAE và RMSE đã bão hòa và hầu như không cải thiện đáng kể sau đó.



Hình 3.1. Mô hình DenseNet dự đoán cho giá trị E_{gap}

Bảng 3.1. Số liệu mô hình đã được đào tạo trong quá trình huấn luyện

Model Training	epoch	R^2_{train}	MAE train	RMSE train	R^2_{val}	MAE val	RMSE val
a	0	-0.0944	1.2884	1.6556	-0.0846	1.2419	1.6028
b	20 - 40	0.7438	0.5653	0.8011	0.6906	0.6441	0.8560
c	60 - 80 - 100	0.7616	0.5340	0.7727	0.6921	0.6318	0.8541
d	120 - 140	0.7724	0.5172	0.7550	0.6960	0.6279	0.8486
e	160 - 180 - 200	0.7766	0.5068	0.7480	0.6902	0.6297	0.8566
f	220 - 240	0.7856	0.4992	0.7328	0.6950	0.6266	0.8499
g	260 - 280 - 300	0.7875	0.4944	0.7295	0.6839	0.6413	0.8652
h	320 - 340	0.7854	0.4942	0.7332	0.6799	0.6406	0.8707
i	360 - 380 - 400	0.7856	0.4982	0.7328	0.6860	0.6378	0.8625
j	420 - 440	0.7924	0.4777	0.7211	0.6682	0.6469	0.8865

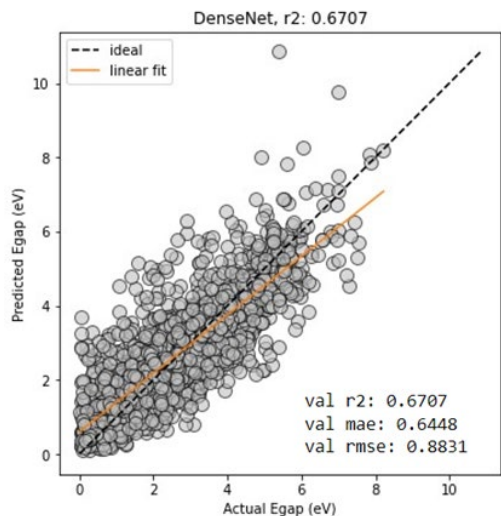
Giờ đây, với mạng Neural nhân tạo đã được huấn luyện, hiệu suất của mô hình (khi kết thúc giai đoạn huấn luyện) có thể được đánh giá trên tập dữ liệu xác thực (Hình 3.2). Dựa trên các dữ liệu đã xử lý, bởi chương trình cụ thể:

```
target_val, pred_val = predict(model, val_loader)
scores = evaluate(target_val, pred_val)

print_scores(scores, label='val')

plot = plot_pred_act(target_val, pred_val, model, label='')
```

Mô hình DenseNet cuối cùng (Hình 3.2) được đưa ra với hiệu suất đạt xấp xỉ 67% bằng việc đào tạo các dữ liệu đầu vào của E_{gap} với 70% cho dữ liệu đào tạo, 20% cho dữ liệu thử nghiệm và 10% cho dữ liệu xác thực. Các điểm dữ liệu có xu hướng phân tán đồng đều và đối xứng quanh đường tuyến tính lý tưởng.



Hình 3.2. Mô hình DenseNet cuối cùng được đánh giá trên tập dữ liệu xác thực (val data)

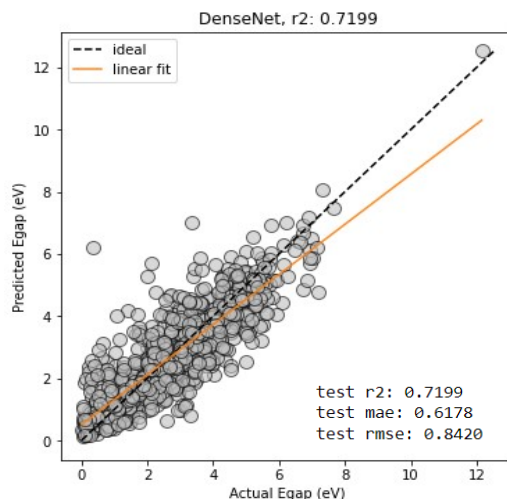
Cuối cùng, mô hình lên tập dữ liệu thử nghiệm (Hình 3.3) được đánh giá hiệu suất, cụ thể bằng chương trình:

```
target_test, pred_test = predict(model, test_loader)
scores = evaluate(target_test, pred_test)

print_scores(scores, label='test')

plot = plot_pred_act(target_test, pred_test, model, label=f' $E_{\text{gap}}$ '
                    i(C)_{\mathrm{p}}$  $E_{\text{gap}}$ )
```

Thông qua việc đánh giá hiệu năng bằng dữ liệu thử nghiệm (20%) cho thấy hiệu năng của mô hình tăng cao hơn so với dữ liệu xác thực (10%). Hiệu suất lúc này đạt xấp xỉ 72%, khẳng định rằng mô hình phù hợp với dữ liệu E_{gap} và khả năng dự đoán tốt của mô hình với dữ liệu train-test-val (70%-20%-10%) và tối đa là 500 epoch. Nghiên cứu đã đưa ra mô hình dự đoán tốt cho E_{gap} bằng phương pháp học sâu – mạng Neural nhân tạo đồng thời đề xuất khung quy trình thuật toán cụ thể.



Hình 3.3. Mô hình đánh giá hiệu suất trên tập dữ liệu thử nghiệm (test data)

4. KẾT LUẬN

Nghiên cứu này đã trình bày tổng quan khoa học vật liệu, phương pháp học máy cũng như các kỹ thuật xây dựng mô hình để tiên đoán giá trị E_{gap} cho vật liệu, đồng thời đã đạt được một số kết quả khả quan sau.

- Thứ nhất, mô hình ML được dựng lại gồm các bước: kiểm tra các thư viện trong chương trình tính toán, truy vấn, xử lý và tổ chức hiển thị mô hình với hiệu suất làm việc của từng mô hình cũng như là sai số tuyệt đối trung bình và sai số căn quân phương.
- Thứ hai, mô hình ML mạng Neuron nhân tạo được xây dựng thành công để dự đoán giá trị E_{gap} .
- Thứ ba, mô hình đạt hiệu suất tốt nhất với R^2 là 0.7924 trên tập huấn luyện và 0.6682 trên tập kiểm định (Hình 3.1-j).

Với các kết quả đạt được, nghiên cứu này giúp mở ra thêm một tài liệu tham khảo ML khi nghiên cứu các vật liệu khối, xây dựng quy trình học cũng như các kỹ thuật đặc trưng cho vật liệu. Bên cạnh đó, kết quả cũng mở ra một hướng nghiên cứu năng lượng vùng cấm bằng phương pháp học sâu triển vọng; từ đó giúp tiên đoán các trị năng lượng vùng cấm, đánh giá chung dữ liệu, phân tích giá trị năng lượng và sử dụng giá trị đó cho lần học dữ liệu tiếp theo.

TÀI LIỆU THAM KHẢO

- Askanazi, E., & Grinberg, I. (2023). Distance-based Analysis of Machine Learning Prediction Reliability for Datasets in Materials Science and Other Fields. *arXiv*.
<https://doi.org/10.48550/arXiv.2304.01146>
- Askanazi, E. M., Lazar, E. A., & Grinberg, I. (2023). Identification of high-reliability regions of machine learning predictions in materials science using transparent conducting oxides and perovskites as examples. *arXiv*.
<https://doi.org/10.48550/arXiv.2304.02218>
- Chowdhury, S., Hill, H. M., Rigosi, A. F., Vora, P. M., Walker, A. R. H., & Tavazza, F. (2022). Review of Theoretical and Computational Methods for 2D Materials Exhibiting Charge Density Waves. *arXiv*.
<https://doi.org/10.48550/arXiv.2201.06615>
- Cole, J. M. (2020). A Design-to-Device Pipeline for Data-Driven Materials Discovery. *Accounts of Chemical Research*, 53(3), 599–610.
<https://doi.org/10.1021/acs.accounts.9b00470>
- Cui, J., Li, X., Zhao, H., Wang, H., Li, B., & Li, X. (2022). Epoch-Evolving Gaussian Process Guided Learning for Classification. *IEEE Transactions on Neural Networks and Learning Systems*, 35(1), 326–337.
<https://doi.org/10.1109/TNNLS.2022.3174207>
- Faber, F. A., Lindmaa, A., Von Lilienfeld, O. A., & Armiento, R. (2017). Machine Learning Energies of 2 Million Elpasolite (ABC2D6) Crystals. *Physical Review Letters*, 117(13), 135502.
<https://doi.org/10.1103/PhysRevLett.117.135502>
- Hu, Z., Li, Y., Lyu, J., Gao, D., & Vasconcelos, N. (2023). Dense network expansion for class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 11858–11867).
<https://doi.org/10.48550/arXiv.2303.12696>
- Irawan, A., Putra, A. M., & Ramadhan, H. (2022, July). A DenseNet Model for Joint Activity Recognition and Indoor Localization. In *2022 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)* (pp. 61–65). IEEE.
<https://doi.org/10.1109/IAICT55358.2022.9887407>
- Lin, G., Chen, F., Zhang, Z., Zhang, A., Wang, X., & Zhou, C. (2023, May). DenseNeXt: An Efficient Backbone for Image Classification. In *2023 15th International Conference on Advanced Computational Intelligence (ICACI)* (pp. 1–6).
<https://doi.org/10.1109/ICACI58115.2023.10146197>
- Liu, H., Xu, L., Ma, Z., Li, Z., Li, H., Zhang, Y., ... & Wang, L. L. (2023). Accurate prediction of semiconductor bandgaps based on machine learning and prediction of bandgaps for two-dimensional heterojunctions. *Materials Today Communications*, 36, 106578.
<https://doi.org/10.1016/j.mtcomm.2023.106578>
- Manzotti, P., Schiavi, F., Nosenzo, F., Pitra, P., & Ballèvre, M. (2022). A journey towards the forbidden zone: a new, cold, UHP unit in the Dora-Maira Massif (Western Alps). *Contributions to Mineralogy and Petrology*, 177(6), 59.
<https://doi.org/10.1007/s00410-022-01923-8>
- Mondal, B., Kröner, M., Hepp, T., Volz, K., & Tonner-Zech, R. (2023). Accurate first principles band gap predictions in strain engineered ternary III-V semiconductors. *Physical Review B*, 108(3), 035202.
<https://doi.org/10.1103/PhysRevB.108.035202>
- Niharika, D., S., D. (2023). Epoch and accuracy based empirical study for cardiac MRI segmentation using deep learning technique. *PeerJ*, 11, e14939.
<https://doi.org/10.7717/peerj.14939>
- Saahil, A., Smitha, R. (2020). Significance Of Epochs On Training A Neural Network. *International Journal of Scientific & Technology Research*. 9(06), 485–488.
<https://www.ijstr.org/final-print/jun2020/Significance-Of-Epochs-On-Training-A-Neural-Network.pdf>
- Smith, J. (2023). Effects of Machine Learning Algorithms for Predicting and Optimizing the Properties of New Materials in the United States. *European Journal of Physical Sciences*, 6(1), 23–34.
<https://doi.org/10.47672/ejps.1444>
- Wang, S., Luo, L., & Guo, Y. (2022). Wider transmission forbidden gaps and tamm state in a complex optoelectronic superlattice based on monolayer MoS₂. *Micro and Nanostructures*, 170, 207378.
<https://doi.org/10.1016/j.micrna.2022.207378>
- Yadav, L. (2024). Atoms as words: A novel approach to deciphering material properties using NLP-inspired machine learning on crystallographic information files (CIFs). *AIP Advances*, 14(4).
<https://doi.org/10.1063/5.0187741>