

DOI:10.22144/ctujos.2025.075

ỨNG DỤNG MÔ HÌNH HỌC MÁY DỰ BÁO CÔNG SUẤT PHÁT ĐIỆN GIÓ

Huỳnh Văn Vạn¹, Vũ Phan Tú^{2, 3*} và Nguyễn Văn Minh Nhứt¹

¹Khoa Điện - Điện tử, Trường Đại học Tôn Đức Thắng, Việt Nam

²Bộ môn Hệ thống điện, Trường Đại học Bách Khoa Thành phố Hồ Chí Minh, Việt Nam

³Đại học Quốc gia Thành phố Hồ Chí Minh, Việt Nam

*Tác giả liên hệ (Corresponding author): vptu@hcmut.edu.vn

Thông tin chung (Article Information)

Nhận bài (Received): 17/02/2025

Sửa bài (Revised): 15/04/2025

Duyệt đăng (Accepted): 30/05/2025

Title: Forecasting the power generation of wind turbine output using machine learning

Author: Huynh Van Van¹, Vu Phan Tu^{2,3*} and Nguyen Van Minh Nhut¹

Affiliation(s): ¹Faculty of Electrical and Electronics Engineering, Ton Duc Thang University, Viet Nam;

²Department of Electrical Systems, Ho Chi Minh City University of Technology, Viet Nam; ³Vietnam National University Ho Chi Minh City, Viet Nam

TÓM TẮT

Ngành điện gió được phát triển nhằm đáp ứng nhu cầu điện năng và giảm phụ thuộc vào năng lượng hóa thạch. Tuy nhiên, sự biến động nguồn năng lượng gió gây khó khăn cho vận hành và điều phối lưới điện dẫn đến tình trạng phát điện dư thừa hoặc thiếu hụt công suất vào từng thời điểm khác nhau. Nghiên cứu này được thực hiện nhằm nâng cao độ chính xác trong dự báo công suất phát điện gió, giúp tối ưu hóa vận hành và tích hợp năng lượng gió vào hệ thống điện. Phương pháp được sử dụng là mô hình mạng hồi quy Perceptron đa tầng (MLP Regression) và được lựa chọn để thực hiện nhiệm vụ dự báo công suất phát điện gió. Tập dữ liệu được sử dụng trong nghiên cứu bao gồm các biến đầu vào như nhiệt độ môi trường, tốc độ và hướng gió; đầu ra là công suất phát của tuabin gió. Toàn bộ quá trình xây dựng và đánh giá mô hình được triển khai bằng ngôn ngữ lập trình Python. Kết quả ghiên cứu cho thấy tiềm năng của trí tuệ nhân tạo (AI - Artificial Intelligence) và học máy trong tối ưu hóa vận hành điện gió, góp phần khai thác hiệu quả tài nguyên gió, hỗ trợ điều phối lưới điện hiệu quả và thúc đẩy hệ thống điện bền vững.

Từ khóa: Dự báo công suất điện gió, điện gió, mô hình học máy, năng lượng tái tạo

ABSTRACT

The development of the wind power sector aims to meet energy demands and reduce dependency on fossil fuels. However, the variability of wind energy sources poses challenges for grid operation and dispatch, leading to periods of surplus generation or insufficient power supply at different times. This study aims to enhance the accuracy of wind energy forecasts to optimize operations and integrate wind energy into the power grid. The applied method is the Multilayer Perceptron (MLP) regression model, which was selected for the task of wind power forecasting. The dataset used in the study includes input variables such as ambient temperature, wind speed, and wind direction, while the output corresponds to the power generated by the wind turbine. The model development and evaluation process is implemented using the Python programming language. The research highlights the potential of AI (Artificial Intelligence) and ML in optimizing wind power operations, efficiently utilizing wind resources, supporting effective grid dispatch, and promoting a sustainable power system.

Keywords: Machine learning, renewable energy, wind power, wind power forecasting

1. GIỚI THIỆU

Với nguồn tài nguyên gió dồi dào quanh năm, điện gió trở thành một giải pháp hiệu quả, đầy tiềm năng cho hệ thống năng lượng quốc gia. Việc phát triển các trang trại điện gió, tận dụng điều kiện gió thuận lợi, không chỉ giúp giảm phát thải khí nhà kính mà còn góp phần giảm sự phụ thuộc vào nhiên liệu hóa thạch truyền thống. Mặc dù sở hữu nhiều lợi thế so với các nguồn năng lượng hóa thạch, điện gió vẫn gặp phải những thách thức lớn trong khai thác hiệu quả và tích hợp vào lưới điện do tính biến động, không liên tục, khó dự đoán và những hạn chế trong việc công suất điện gió thay đổi liên tục. Do đó, việc phát triển một mô hình dự báo công suất phát điện gió có độ chính xác cao là điều cần thiết nhằm nâng cao độ tin cậy và tính ổn định của hệ thống điện gió.

Hiện nay, các phương pháp dự báo công suất gió chủ yếu được chia thành bốn nhóm chính: mô hình dựa trên vật lý, mô hình thống kê, mô hình trí tuệ nhân tạo (AI - Artificial Intelligence) và mô hình lai (Yang et al., 2023; Hao et al., 2024). Tuy nhiên, mỗi phương pháp, cách tiếp cận đều có những ưu điểm

và hạn chế riêng trong ứng dụng thực tế được trình bày trong Bảng 1. Công suất phát điện gió phụ thuộc vào các yếu tố thời tiết như: vận tốc gió, hướng gió và các điều kiện khí hậu, môi trường khác nhau. Sự biến động của các yếu tố này dẫn đến sự biến động công suất phát điện gió, gây khó khăn cho vận hành và điều phối lưới điện. Điều này nhấn mạnh sự cần thiết của dự báo chính xác để đảm bảo tối ưu vận hành lưới điện cũng như vận hành hệ thống điện ổn định. Ở nhiều nghiên cứu, việc đánh giá tác động của các yếu tố môi trường đã được thực hiện, đặc biệt là vận tốc gió đến hiệu suất tuabin trong quá trình phát điện, từ đó đề xuất các giải pháp, các mô hình dự báo nhằm giảm thiểu ảnh hưởng của những biến động này (Azimi et al., 2016; Pandit et al., 2023). Bên cạnh đó, việc phân tích và so sánh các phương pháp trích xuất dữ liệu, ước lượng vận tốc gió giúp tiếp cận hiệu quả đến giám sát số kết quả dự báo (Naik et al., 2018; Wang et al., 2021). Tuy nhiên, một thách thức lớn trong dự báo công suất gió là nguồn dữ liệu thu thập từ hệ thống SCADA (Supervisory Control and Data Acquisition) thường không đầy đủ do lỗi cảm biến, gián đoạn tín hiệu truyền hoặc điều kiện vận hành không ổn định.

Bảng 1. Đánh giá các phương pháp dự báo năng lượng gió

Mô hình	Phương pháp	Ưu điểm	Nhược điểm
Mô hình dựa trên vật lý	Mô hình Dự báo Thời tiết Số (NWP - Numerical Weather Prediction) (Zhang et al., 2020)	Độ tin cậy cao, hiệu quả trong dự báo dài hạn từ vài giờ đến vài tuần.	Khó triển khai do yêu cầu tính toán phức tạp, phụ thuộc nhiều vào chất lượng dữ liệu đầu vào.
Mô hình thống kê	ARMA (AutoRegressive Moving Average) (Tian et al., 2018)	Thiết kế đơn giản, hiệu quả tính toán cao, tận dụng tốt dữ liệu lịch sử để nhận diện quy luật biến động.	Phụ thuộc vào bộ dữ liệu ổn định, không phù hợp với dữ liệu phi tuyến tính, độ chính xác giảm khi tăng khoảng thời gian dự báo.
Mô hình AI	Học máy (ML - Machine Learning) (Mohd et al., 2023) Học sâu (DL - Deep Learning) (Lin et al., 2020)	Khả năng học tập các quy luật phi tuyến tính giúp cải thiện độ chính xác dự báo, tự động hóa quy trình, phù hợp với dữ liệu lớn.	Triển khai phức tạp, yêu cầu tài nguyên tính toán lớn, khó khăn với hệ thống có năng lực xử lý hạn chế.
Mô hình lai	Hybrid Models (Ozkan et al., 2015; Kalpana et al., 2023)	Kết hợp các ưu điểm của từng phương pháp, cải thiện độ chính xác dự báo, khả năng thích ứng với nhiều loại dữ liệu khác nhau.	Đòi hỏi kỹ thuật triển khai phức tạp, yêu cầu cao về tài nguyên tính toán và dữ liệu huấn luyện chất lượng cao.

Khi một hoặc nhiều biến đầu vào có giá trị khuyết, mô hình dự báo không thể nắm bắt đầy đủ mối quan hệ giữa các yếu tố này ảnh hưởng đến công suất, làm giảm độ chính xác của mô hình dự báo (Fan et al., 2024). Ngoài ra, dữ liệu sai lệch làm gia tăng sai số, gây khó khăn trong quản lý và vận hành hệ thống điện. Điều này đặc biệt quan trọng trong các hệ thống điện có tỷ trọng cao từ năng lượng tái tạo như điện gió, nơi mà độ chính xác của dự báo đóng vai trò quyết định trong việc đảm bảo ổn định và hiệu quả kinh tế. Do đó, việc xử lý dữ liệu ngoại lai bằng các phương pháp xử lý dữ liệu hoặc sử dụng các mô hình dự báo mạnh mẽ là yếu tố quan trọng giúp giảm thiểu rủi ro và tối ưu hóa việc tích hợp điện gió vào lưới điện (Liang et al., 2023).

Nghiên cứu này được thực hiện nhằm hướng đến việc phát triển một mô hình dự báo công suất phát điện gió kết hợp các kỹ thuật tiền xử lý dữ liệu tiên tiến nhằm giải quyết những nhược điểm và thách thức hiện tại (Tuncar et al., 2024; Yang et al., 2025). Kết quả các nghiên cứu về học sâu đã chỉ ra rằng phương pháp hồi quy Perceptron đa tầng (MLP - Multi-Layer Perceptron) đặc biệt hiệu quả trong việc nhận diện các mối quan hệ phi tuyến giữa biến đầu vào và đầu ra, phù hợp với tính chất không ổn định của tốc độ gió (Sireesha et al., 2024). Để nâng cao hiệu suất dự báo, đảm bảo độ tin cậy và độ chính xác cao hơn, mô hình lai CNN-Transformer-MLP (Bashir et al., 2025) đã được đề xuất trong nghiên cứu.

Trong bài nghiên cứu, mô hình dự báo ngắn hạn công suất điện gió kết hợp các phương pháp xử lý dữ liệu hiện đại đã được đề xuất nhằm cải thiện độ chính xác kết quả dự báo. Mô hình tận dụng khả năng xử lý dữ liệu nâng cao của phương pháp hồi quy đa tầng MLP, hướng đến việc xây dựng một mô hình dự báo có độ chính xác cao hơn. Những đóng góp của bài báo như sau:

- Mô hình học máy trong dự báo ngắn hạn công suất phát điện gió được áp dụng thành công.
- Kết quả dự báo công suất gió đạt sai số dưới 5% so với dữ liệu thực tế thu thập được từ các dự án điện gió.
- Phương pháp MLP tuy đã được áp dụng trong các nghiên cứu trước, nghiên cứu này sẽ đưa ra một cách tiếp cận mới, kết hợp các kỹ thuật tiền xử lý dữ liệu và tối ưu hóa, giúp nâng cao độ chính xác trong dự báo công suất điện gió. Cách tiếp cận này không chỉ cải thiện hiệu quả của mô hình mà còn hỗ trợ tích hợp điện gió vào lưới điện một cách hiệu quả hơn, mang lại kết quả dự báo có độ tin cậy cao hơn.

Quá trình phát triển và triển khai mô hình được thực hiện trên nền tảng Google Colaboratory và lập trình bằng ngôn ngữ Python. Điều này cho phép sử dụng các thư viện và công cụ mạnh mẽ trong phân tích dữ liệu và học máy.

2. PHƯƠNG PHÁP NGHIÊN CỨU

2.1. Tổng quan mô hình dự báo

Trong lĩnh vực dự báo công suất điện gió, trí tuệ nhân tạo (AI), đặc biệt là học máy (ML) đang nhận được sự quan tâm ngày càng lớn và được ứng dụng rộng rãi. Mô hình học máy cho phép các mô hình dự báo học hỏi và thích ứng thông qua phân tích dữ liệu thay vì chỉ dựa trên các quy tắc được xác định trước. Cách tiếp cận này giúp khám phá các quy luật phức tạp trong tập dữ liệu lớn, bao gồm dữ liệu thời tiết và dữ liệu SCADA từ tuabin gió, từ đó nâng cao độ chính xác của dự báo. Quy trình triển khai thường bao gồm nhiều bước, bao gồm: xử lý dữ liệu, trích xuất tính năng và phân loại dữ liệu như được trình bày ở Hình 1. Mục tiêu được xác định là phát triển mô hình dự báo với độ chính xác và độ tin cậy cao.



Hình 1. Sơ đồ tổng quát về dự báo phát điện gió

Dữ liệu được thu thập từ hệ thống SCADA của tuabin gió thường chứa nhiễu, giá trị khuyết và sai lệch do lỗi cảm biến hoặc điều chỉnh công suất khi dư thừa điện gió. Do đó, việc tiền xử lý dữ liệu đóng vai trò quan trọng nhằm nâng cao chất lượng đầu vào và tối ưu hóa hiệu suất của mô hình dự báo.

Tiền xử lý dữ liệu (Preprocessing): Đây là bước quan trọng giúp cải thiện chất lượng dữ liệu đầu vào:

- Xử lý dữ liệu khuyết: Các giá trị bị khuyết trong tập dữ liệu từ hệ thống SCADA được giải quyết bằng cách tính toán để xác định giá trị hợp lý.
- Loại bỏ dữ liệu sai lệch: Các giá trị sai lệch (do lỗi cảm biến hoặc điều chỉnh công suất khi dư thừa điện gió) được loại bỏ nhằm giảm nhiễu và nâng cao độ chính xác dự báo.

Lựa chọn đặc trưng: Hệ thống SCADA ghi nhận nhiều thông số dư thừa, ảnh hưởng đến hiệu quả học máy. Việc lựa chọn đặc trưng có tương quan cao giúp giảm chiều dữ liệu, tăng tốc độ tính toán và nâng cao khả năng tổng quát hóa của mô hình.

Phân cụm dữ liệu: Nhóm các mẫu dữ liệu có đặc điểm tương đồng giúp xây dựng các mô hình dự báo chuyên biệt, tối ưu hóa hiệu suất và đảm bảo độ chính xác cao hơn cho từng điều kiện vận hành của tuabin gió.

Chia dữ liệu: Để xây dựng một mô hình có khả năng tổng quát hóa tốt, tập dữ liệu đầy đủ thường được chia thành ba phần chính:

- Tập huấn luyện: dùng để huấn luyện mô hình, giúp nó học quy luật và mối quan hệ trong dữ liệu.
- Tập kiểm định: dùng để tinh chỉnh siêu tham số, đánh giá mô hình trong quá trình huấn luyện mà không sử dụng tập kiểm tra.
- Tập kiểm tra: dùng để đánh giá cuối cùng, đảm bảo mô hình có khả năng tổng quát tốt với dữ liệu chưa thấy trước.

Phân loại dữ liệu: Sau khi phân cụm, mô hình phân loại sẽ gán nhãn cho các điểm dữ liệu mới, giúp cải thiện độ chính xác của dự báo.

Lựa chọn và huấn luyện mô hình: Các thuật toán khác nhau được đánh giá để lựa chọn mô hình tối ưu, kết hợp với tinh chỉnh siêu tham số để đảm bảo hiệu suất cao nhất.

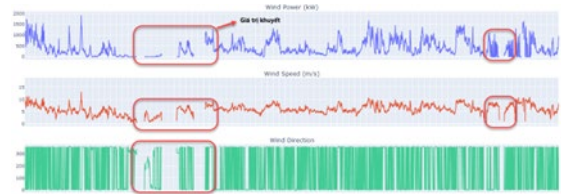
Kiểm tra và đánh giá mô hình: Hiệu suất của mô hình được kiểm tra trên tập dữ liệu kiểm tra. Đồng thời, việc thêm một số mức nhiễu vào dữ liệu

giúp đánh giá khả năng dự đoán của mô hình với các dữ liệu không hoàn hảo.

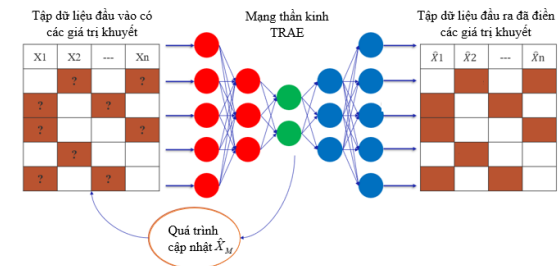
2.2. Xử lý dữ liệu khuyết

Sự xuất hiện của giá trị khuyết trong tập dữ liệu đầu vào của mô hình dự báo công suất điện gió làm tăng độ phức tạp của quá trình khai thác dữ liệu. Để giải quyết vấn đề này một cách hiệu quả, việc ứng dụng mô hình mạng thần kinh nhân tạo (Track-Removed AutoEncoder (TRAЕ)) đã được thực hiện. Phương pháp này tái cấu trúc động kiến trúc của các nơ-ron ẩn dựa trên nền tảng AutoEncoder (AE) truyền thống (Xiaochen et al., 2019).

Cụ thể, các giá trị dữ liệu khuyết (Hình 2) được đưa vào quá trình huấn luyện mạng để tối ưu hóa các tham số của mô hình. Trong quá trình tối ưu hóa các giá trị bị khuyết được cập nhật thông qua hàm tổn thất của TRAЕ (Hình 3). Mô hình giúp sử dụng dữ liệu một cách toàn diện hơn thay vì chỉ dựa trên các giá trị đã biết. Khi quá trình tối ưu tiến dần đến hội tụ, sai số ước lượng của giá trị khuyết giảm dần, giúp cải thiện độ chính xác của mô hình và chất lượng nội suy giá trị.



Hình 2. Giá trị khuyết trong tập dữ liệu



Hình 3. Quy trình xác định giá trị khuyết với mạng thần kinh TRAЕ và thuật toán SGD

Giả sử ta có một tập dữ liệu ban đầu $X \in R^{m \times n}$ trong đó:

- m : là số lượng dữ liệu đầu vào (ví dụ: số giờ đo gió),
- n : là số đặc trưng của dữ liệu (ví dụ: tốc độ gió, công suất phát, nhiệt độ, áp suất, v.v.).

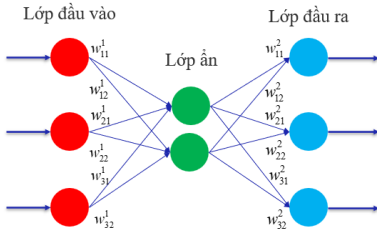
Định nghĩa hai tập dữ liệu con X_{obs} và X_{miss} như sau (Xiaochen et al., 2019):

$$X_{obs} = \{(x_{ij}) \in X \mid x_{ij} \neq NaN\} \quad (1)$$

$$X_{miss} = \{(x_{ij}) \in X \mid x_{ij} = NaN\} \quad (2)$$

Bước 1: Khởi tạo tham số mô hình TRAE

Tham số mô hình TRAE bao gồm: trọng số $W^{(1)}$ (trọng số liên kết lớp đầu vào với lớp ẩn), $W^{(2)}$ (trọng số liên kết lớp ẩn với lớp đầu ra), $b^{(1)}$ ngưỡng kích hoạt lớp ẩn và $b^{(2)}$ ngưỡng kích hoạt của lớp đầu ra (Xiaochen et al., 2019; Tian et al., 2025) (Hình 4).



Hình 4. Cấu trúc mạng thần kinh TRAE

Hàm giá trị lớp ẩn (Xiaochen et al., 2019):

$$h = f(w^{(1)}x^{(t)} + b^{(1)}) \quad (3)$$

Hàm giá trị đầu ra (Xiaochen et al., 2019):

$$y^{(t)} = g(w^{(2)}h + b^{(2)}) \quad (4)$$

Trong đó: $f(\bullet)$, $g(\bullet)$ là hàm kích hoạt (ví dụ: ReLU, Sigmoid), $y^{(t)}$ là giá trị đầu ra mô hình TRAE.

Bước 2: Xây dựng vector đầu vào

Trong quá trình huấn luyện mô hình TRAE, tập dữ liệu đầu vào bao gồm cả các giá trị quan sát được và các giá trị bị khuyết. Tại vòng lặp t , vector đầu vào được xây dựng như sau (Xiaochen et al., 2019):

$$X^{(t)} = \begin{cases} x_i, & i \in X_{obs} \\ \hat{x}_m^{(t-1)}, & m \in X_{miss} \end{cases} \quad (5)$$

Trong đó: X_{obs} là tập hợp các giá trị quan sát được, giữ nguyên trong suốt quá trình huấn luyện. X_{miss} là tập hợp các giá trị khuyết, được thay thế bằng giá trị dự đoán từ vòng lặp trước đó $\hat{x}_m^{(t-1)}$.

Ban đầu, các giá trị khuyết trong tập dữ liệu được khởi tạo bằng giá trị trung bình của các quan sát sẵn có. Giá trị khuyết $\hat{x}_m^{(0)}$ được tính theo công thức (Xiaochen et al., 2019):

$$\hat{x}_m^{(0)} = \frac{1}{|X_{obs,m}|} \sum_{i \in X_{obs,m}} x_i \quad (6)$$

Trong đó: $X_{obs,m}$ là tập hợp các giá trị quan sát thứ m , $|X_{obs,m}|$ là số lượng phần tử trong tập $X_{obs,m}$, và x_i là các giá trị quan sát trong tập dữ liệu huấn luyện.

Cách tiếp cận này giúp đảm bảo rằng mô hình TRAE có thể học cách nội suy giá trị khuyết một cách hiệu quả thông qua quá trình tối ưu hóa lặp dần, thay vì chỉ dựa vào các phương pháp điền khuyết truyền thống.

Bước 3: Cập nhật tham số mô hình

Hàm tổn thất $L^{(t)}$ đo lường sai số giữa giá trị ước lượng mô phỏng y_i và giá trị thực tế x_i trong dữ liệu đầu vào. Hàm tổn thất này bao gồm hai phần: một phần cho các giá trị đã biết (X_{obs}) và một phần cho các giá trị khuyết (X_{miss}) (Yang et al., 2024):

$$L^{(t)} = \frac{1}{2} \sum_{i \in X_{obs}} (x_i - y_i^{(t)})^2 + \frac{1}{2} \sum_{i \in X_{miss}} (\hat{x}_m^{(t-1)} - y_m^{(t)})^2 \quad (7)$$

Đặt $\theta = \{W^{(1)}, W^{(2)}, b^{(1)}, b^{(2)}, \hat{x}_m^{(t)}\}$ là vector chứa tất cả các tham số cần tối ưu, bao gồm cả các giá trị khuyết. Quá trình tối ưu sử dụng phương pháp Stochastic Gradient Descent (SGD) để cập nhật tham số theo đạo hàm của hàm tổn thất $L^{(t)}$ (Hu et al., 2016):

$$\theta^{(t)} = \theta^{(t-1)} - \eta \frac{\partial L^{(t)}}{\partial \theta} \quad (8)$$

Trong đó: $\theta^{(t)}$ là vector chứa tất cả tham số cần tối ưu tại lần lặp t , $\theta^{(t-1)}$ là giá trị của tham số ở lần lặp $t-1$ và $\frac{\partial L^{(t)}}{\partial \theta}$ là đạo hàm của hàm tổn thất $L^{(t)}$ đối với các tham số trong vector θ tại lần lặp t .

Bước 4: Cập nhật giá trị khuyết từ giá trị mô hình TRAE

Các giá trị được cập nhật từ quá trình tối ưu SGD sẽ thay thế các giá trị bị khuyết trong tập dữ liệu bằng các ước lượng mới từ bước 3. Tập dữ liệu sau khi cập nhật sẽ được sử dụng cho lần lặp tiếp theo. (Hu et al., 2016)

$$\hat{x}_m^{(t)} = \hat{x}_m^{(t-1)} - \eta \frac{\partial L^{(t)}}{\partial \hat{x}_m^{(t-1)}} \quad (9)$$

Trong đó: $\hat{x}_m^{(t)}$ là giá trị được xác định của giá trị bị khuyết tại lần lặp t , $\hat{x}_m^{(t-1)}$ là giá trị được xác định của giá trị bị khuyết tại lần lặp $t - 1$ và $\frac{\partial L^{(t)}}{\partial \hat{x}_m^{(t-1)}}$ là đạo hàm của hàm tổn thất $L^{(t)}$ đối với giá trị $\hat{x}_m^{(t-1)}$ tại lần lặp $t - 1$.

Bước 5: Kiểm tra điều kiện hội tụ và kết thúc

Khi số lần lặp tối đa hoặc khi sự thay đổi trong hàm tổn thất nhỏ hơn ngưỡng được xác định trước, quá trình đào tạo kết thúc và các giá trị ước tính được coi là kết quả cuối cùng.

Hàm kiểm tra điều kiện hội tụ trong quá trình đào tạo mô hình hoặc tối ưu hóa được định nghĩa như sau:

$$|L^{(t)} - L^{(t-1)}| < \varepsilon \quad (10)$$

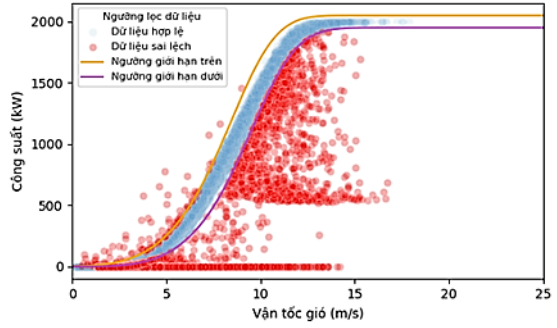
Trong đó: $L^{(t)}$ là giá trị của hàm tổn thất tại lần lặp t , là giá trị của hàm tổn thất tại lần lặp trước đó ($t - 1$) và ε là ngưỡng, thường được đặt trong phạm vi từ 10^{-4} đến 10^{-6} .

2.3. Loại bỏ giá trị sai lệch

Nghiên cứu sau khi thực hiện bước lựa chọn đặc trưng đã cho ra kết quả vận tốc gió là yếu tố có ảnh hưởng lớn nhất đến công suất phát của tuabin. Do đó, tập dữ liệu được xét sẽ tập trung vào hai thành phần chính: vận tốc gió (giá trị đầu vào) và công suất tuabin (giá trị đầu ra).

Để xây dựng mô hình dự báo sản lượng điện của tuabin gió, cần loại trừ các tác động bên ngoài ảnh hưởng đến dữ liệu điện gió như: cắt giảm lưới điện, ngừng hoạt động hoặc giảm sản lượng điện. Các giá trị ngoại lệ phải được phát hiện và loại bỏ trước khi lập mô hình. Tuy nhiên, việc lọc dữ liệu quá mức có thể dẫn đến quá khớp, điều này không mong muốn, đặc biệt đối với các ứng dụng như dự báo bằng dữ liệu chưa biết (Hình 5).

Một cách tiếp cận để quản lý các giá trị ngoại lệ liên quan đến việc loại trừ các điểm dữ liệu nằm ngoài phạm vi được xác định bởi $Q1 - 1,5 \times IQR$ và $Q3 + 1,5 \times IQR$, trong đó $Q1$ và $Q3$ lần lượt biểu thị tứ phân vị thứ nhất và thứ ba, và IQR (Phạm vi liên tứ phân vị) được tính là $Q3 - Q1$. Hệ số nhân 1,5 bắt nguồn từ định lý giới hạn trung tâm (Central Limit Theorem), kết quả chỉ ra rằng trong phân phối chuẩn, các điểm dữ liệu vượt quá 1,5 lần IQR được coi là giá trị ngoại lệ.



Hình 5. Xác định ngưỡng giới hạn của dữ liệu

Bước 1: Thu thập dữ liệu từ hệ thống SCADA từ tuabin gió:

$$X = \{(v_i, P_i) | i = 1, 2, \dots, N\} \quad (11)$$

Trong đó: v_i là vận tốc gió (m/s), P_i là công suất của tuabin (kW) tại thời điểm i .

Bước 2: Xây dựng mô hình đường cong công suất

Mô hình đường cong công suất được thiết kế để mô tả mối quan hệ phi tuyến giữa công suất và tốc độ gió. Sau khi thử nghiệm nhiều hàm sigmoid (logistic, tanh) trên dữ liệu thực tế, hàm hyperbolic tangent (tanh) kết hợp với đa thức bậc 4 của tốc độ gió được lựa chọn nhờ độ phù hợp cao nhất. Phương trình mô hình đường cong công suất được biểu diễn như sau:

$$\hat{P}(v) = P_{\max} \times \tanh((a \times v + b)^4) \quad (12)$$

Trong đó: $\hat{P}(v)$ là công suất dự đoán tương ứng với tốc độ gió v , $P_{\max}$ là công suất tối đa của tuabin; v là tốc độ gió (m/s) và a, b là các hệ số cần tối ưu hóa.

Bước 3: Tối ưu hóa tham số bằng hồi quy phân vị (Quantile Regression)

Để xác định các đường cong biên $Q1$ (25%) và $Q3$ (75%), phương pháp hồi quy phân vị (Quantile Regression) được sử dụng. Phương pháp này giảm thiểu các lỗi tuyệt đối có trọng số dựa trên phân vị, khác với hồi quy tuyến tính truyền thống, dựa trên bình phương tối thiểu thông thường, do đó giảm ảnh hưởng của các giá trị cực trị lên đường cong được lắp. Hàm tổn thất được định nghĩa (He et al., 2018):

$$\text{Nếu: } P_i - \hat{P}(v_i) \geq 0$$

$$L(P_i, \hat{P}(v_i)) = \sum_{i=1}^N \tau \times (P_i - \hat{P}(v_i)) \quad (13)$$

Nếu: $P_i - \hat{P}(v_i) < 0$

$$L(P_i, \hat{P}(v_i)) = \sum_{i=1}^N (\tau - 1) \times (P_i - \hat{P}(v_i)) \quad (14)$$

Trong đó: τ là mức phân vị ($\tau = 0,25$ cho Q1 và $\tau = 0,75$ cho Q3)

Vì hàm tổn thất là hàm không tuyến tính (kết hợp hàm tan hyperbolic và đa thức bậc bốn), phương pháp Gradient Descent được sử dụng để tối ưu hóa các tham số dựa trên đạo hàm của hàm tổn thất, với hàm mục tiêu (He et al., 2018):

$$(a^*, b^*) = \arg \min_{a,b} \sum_{i=1}^n \rho_{\tau}(P_i - \hat{P}(v_i)) \quad (15)$$

Phương pháp Gradient Descent được áp dụng để tối ưu tham số a, b. Các tham số được cập nhật theo công thức (He et al., 2018):

$$a^{(t+1)} = a^{(t)} - \eta \frac{\partial L^{(t)}}{\partial a}, \quad b^{(t+1)} = b^{(t)} - \eta \frac{\partial L^{(t)}}{\partial b} \quad (16)$$

Trong đó: $a^{(t)}, b^{(t)}$ là giá trị tham số a, b tại lần lặp t; $a^{(t+1)}, b^{(t+1)}$ là giá trị tham số a tại lần lặp t+1; $\frac{\partial L^{(t)}}{\partial a}, \frac{\partial L^{(t)}}{\partial b}$ là đạo hàm của hàm tổn thất L đối với tham số a, b.

Bước 4: Xây dựng đường cong dự đoán các phân vị:

Đối với Q1(v) (phân vị thứ 25, $\tau = 0,25$):

$$Q1(v) = f_{0,25}(x) = P_{max} \times \tanh((a_{0,25}v + b_{0,25})^4) \quad (17)$$

Đối với Q3(v) (phân vị thứ 75, $\tau = 0,75$):

$$Q3(v) = f_{0,75}(x) = P_{max} \times \tanh((a_{0,75}v + b_{0,75})^4) \quad (18)$$

Trong đó, các tham số $a_{0,25}, b_{0,25}$ và $a_{0,75}, b_{0,75}$ được tìm ra bằng quá trình tối ưu hóa hàm tổn thất của hồi quy phân vị ở bước 3.

Bước 5: Xác định ngưỡng giới hạn:

Ngưỡng giới hạn được tính bằng quy tắc $1,5 \times IQR$, trong đó $IQR = Q3 - Q1$, đồng thời điều chỉnh bởi một hệ số dung sai để tránh overfitting, đặc biệt khi huấn luyện mô hình dự báo trên dữ liệu chưa thấy trước đó. Các ngưỡng được xác định như sau:

- Ngưỡng giới hạn trên (Upper limit):

$$(1 - tol) \times (Q3(v) - 1,5 \times IQR(v)) \quad (19)$$

- Ngưỡng giới hạn dưới (Lower limit):

$$(1 - tol) \times (Q1(v) - 1,5 \times IQR(v)) \quad (20)$$

Trong đó, hệ số dung sai $\pm 1,5\%$ được áp dụng một cách tùy ý để giữ lại các điểm dữ liệu có độ lệch nhỏ, đặc biệt trong các vùng mà công suất đầu ra đạt mức tối đa của tuabin gió.

Bước 6: Lọc dữ liệu sai lệch

Đối với mỗi điểm dữ liệu (v_i, P_i) trong tập dữ liệu $X = \{(v_i, P_i) | i = 1, 2, \dots, N\}$, nếu giá trị công suất P_i nằm trong giới hạn ngưỡng, nó được coi là giá trị bình thường:

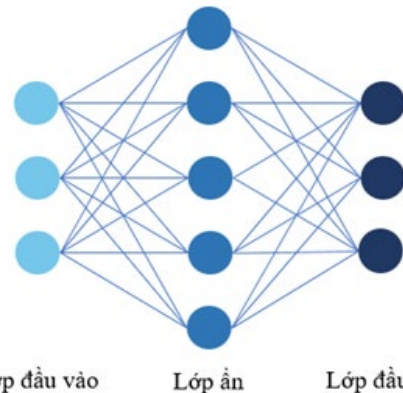
$$Lower\ Limit \leq P_i \leq Upper\ Limit \quad (21)$$

Các giá trị công suất P_i vượt quá ngưỡng trên sẽ được xác định là ngoại lai và bị loại khỏi tập dữ liệu nhằm đảm bảo chất lượng dữ liệu trước khi thực hiện mô hình hóa hoặc dự báo.

2.4. Mô hình dự báo

Mô hình mạng hồi quy Perceptron đa tầng (MLP Regression) (Hình 6) là một kỹ thuật học máy (Machine Learning) tận dụng kiến trúc mạng nơ-ron nhiều lớp, lấy cảm hứng từ cấu trúc thần kinh của não bộ con người. Trong MLP, các nơ-ron được tổ chức thành các lớp kết nối với nhau, trong đó mỗi nơ-ron ở một lớp sẽ liên kết với các nơ-ron ở lớp tiếp theo thông qua các trọng số có thể điều chỉnh. Trong quá trình huấn luyện, các trọng số này được tối ưu hóa liên tục để nâng cao khả năng của mô hình trong việc phân tích và nhận diện các quy luật ẩn trong dữ liệu.

Thế mạnh chính của MLP nằm ở khả năng mô hình hóa và học các mối quan hệ phi tuyến giữa các đặc trưng đầu vào và đầu ra mục tiêu, giúp dự đoán hoặc phân loại chính xác hơn.



Hình 6. Sơ đồ mạng nơ-ron hồi quy Multi-Layer Perceptron – MLP

Cấu trúc của MLP bao gồm ba loại lớp chính:

Lớp đầu vào: nhận dữ liệu đầu vào thô, đóng vai trò là điểm khởi đầu để mạng thực hiện tính toán.

Lớp ẩn: bao gồm một hoặc nhiều lớp nơ-ron, có nhiệm vụ xử lý dữ liệu đầu vào và trích xuất các đặc trưng quan trọng. Các lớp ẩn sử dụng hàm kích hoạt phi tuyến để biến đổi dữ liệu, giúp mạng học được các mẫu phức tạp.

Lớp đầu ra: tạo ra kết quả đầu ra của mô hình dựa trên các đặc trưng đã học từ các lớp ẩn. Đối với bài toán hồi quy, đầu ra thường là một giá trị liên tục (ví dụ: công suất điện gió dự báo).

Cơ chế hoạt động của MLP bắt đầu từ lớp đầu vào, nơi dữ liệu được đưa vào mạng. Mỗi nơ-ron tính tổng có trọng số của các đặc trưng đầu vào, áp dụng một hàm kích hoạt, rồi truyền giá trị tính được đến lớp ẩn đầu tiên. Quá trình này tiếp tục qua các lớp ẩn cho đến khi dữ liệu đi đến lớp đầu ra, nơi dự báo cuối cùng được tạo ra.

Việc xác định số lượng lớp ẩn tối ưu trong mô hình MLP đòi hỏi một quá trình thử nghiệm lặp đi lặp lại. Các cấu hình mạng khác nhau được thử nghiệm và đánh giá dựa trên các chỉ số sai số để xác định cấu trúc mang lại độ chính xác cao nhất. Cách tiếp cận này giúp phát triển một mô hình đáng tin cậy, có khả năng xử lý tốt sự phức tạp của dữ liệu.

3. KẾT QUẢ VÀ THẢO LUẬN

3.1. Dữ liệu

Dữ liệu trong bài báo này được thu thập từ các dự án hợp tác khoa học tại Đài Loan, liên quan đến phát điện gió. Tập dữ liệu bao gồm hai bộ dữ liệu riêng biệt như sau:

Tập dữ liệu huấn luyện (Training Data):

- Khoảng thời gian giữa các điểm dữ liệu: 10 phút
- Thời gian thu thập: 1 năm.
- Cấu trúc: 4 cột (3 biến đầu vào, 1 biến đầu ra), khoảng 52.560 dòng (tính theo: 365 ngày $\times$ 24 giờ $\times$ 6 khoảng thời gian/giờ).
- Dữ liệu đầu vào: nhiệt độ môi trường, tốc độ gió, hướng gió.

Tập dữ liệu này được sử dụng để huấn luyện mô hình dự báo bằng cách phân cụm dữ liệu thành các nhóm khác nhau, giúp mô hình học được các quy luật đặc trưng của sản xuất điện gió.

Tập dữ liệu kiểm tra (Testing Data):

- Khoảng thời gian giữa các điểm dữ liệu: 10 phút.
- Thời gian thu thập: 7 tháng.
- Cấu trúc: 4 cột (3 biến đầu vào, 1 biến đầu ra), khoảng 30.660 dòng (tính theo: 213 ngày $\times$ 24 giờ $\times$ 6 khoảng thời gian/giờ).
- Dữ liệu đầu vào: nhiệt độ môi trường, tốc độ gió, hướng gió.

Tập dữ liệu này được sử dụng để dự báo công suất điện gió và đánh giá tính ổn định của mô hình khi xử lý dữ liệu có sai số, kiểm tra độ tin cậy và khả năng thích ứng của mô hình.

Tập dữ liệu huấn luyện (training data) được sử dụng để đánh giá khả năng dự báo công suất phát điện gió của mô hình, bao gồm các bước: xử lý dữ liệu, kỹ thuật lọc dữ liệu và phương pháp xây dựng mô hình dự báo công suất phát điện gió dựa trên phân cụm. Việc tham khảo một số nghiên cứu trước thường có giá trị sai số khá cao, mức sai số đã được đặt ra khoảng 5% hoặc thấp hơn cho việc xác nhận rằng mô hình đáp ứng yêu cầu, vận hành ổn định và có độ tin cậy cao (Lin & Liu, 2020; Zhang et al., 2020).

Sau khi hoàn thiện mô hình dự báo, bộ dữ liệu kiểm tra sẽ được sử dụng để xác minh các bước thực hiện và các phương pháp được chọn khi áp dụng cho dữ liệu mới. Quá trình này giúp đánh giá khả năng ứng dụng của mô hình trong điều kiện thực tế và khả năng mở rộng cho các loại dữ liệu khác nhau, đảm bảo tính hiệu quả của mô hình dự báo điện gió trong thực tế.

3.2. Sai số của các mô hình dự báo công suất phát điện gió

Trong bài báo này, hai chỉ số sai số được lựa chọn làm tiêu chí đánh giá để xác định số lượng lớp ẩn phù hợp trong các bước xây dựng mô hình dự báo phát điện gió, bao gồm sai số bình phương trung bình gốc (RMSE - Root Mean Squared Error) và sai số tương đối trung bình (MRE - Mean Relative Error).

Sai số bình phương trung bình gốc (RMSE) được sử dụng phổ biến trong các bài toán hồi quy để đo căn bậc hai của trung bình tổng bình phương sai lệch giữa giá trị dự đoán của mô hình và giá trị thực tế. Đây là một trong những tiêu chí quan trọng để đánh giá, xác định và lựa chọn các mô hình dự báo hiệu quả vì nó cung cấp một thước đo lỗi có cùng đơn vị với dữ liệu gốc. RMSE càng thấp thì độ chính xác càng cao và càng phù hợp với dữ liệu thực tế, vì nó

trừng phạt mạnh các sai số lớn trong khi phản ánh hiệu suất tổng thể của mô hình.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (22)$$

Sai số tương đối trung bình (MRE) là trung bình của các sai số tuyệt đối giữa giá trị dự đoán của mô hình và giá trị thực tế, được biểu thị dưới dạng phần trăm của giá trị gốc. Đây là một chỉ số quan trọng để đánh giá mô hình dự báo, vì nó chuẩn hóa sai số theo độ lớn của dữ liệu thực tế, cung cấp cái nhìn sâu sắc về độ chính xác tỷ lệ của dự đoán. MRE thấp hơn thể hiện độ chính xác cao hơn và hiệu suất mô hình dự báo tốt hơn, đặc biệt trong việc phản ánh chính xác độ lệch tương đối so với quan sát thực tế.

$$MRE = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (23)$$

Trong đó: n là tổng số quan sát, y_i là giá trị thực tế tại quan sát thứ i , $\hat{y}_i$ là giá trị dự báo tại quan sát thứ i .

Tuy nhiên, sai số MRE được hiển thị dưới dạng phần trăm (%) nhằm giúp dễ dàng so sánh mức độ sai lệch giữa giá trị thực tế và giá trị dự đoán trong một phạm vi chuẩn hóa, đồng thời làm cho kết quả trở nên dễ hiểu hơn khi phân tích và đánh giá độ chính xác của mô hình.

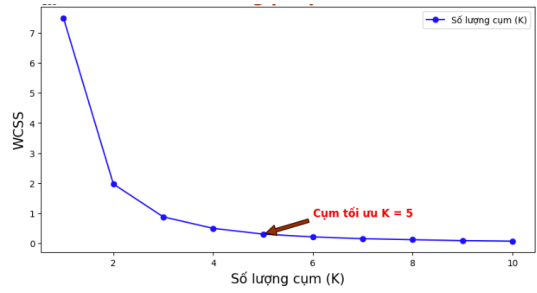
3.3. Kết quả

Dữ liệu đầu vào được xử lý theo các bước được trình bày trong lưu đồ các bước, quy trình xây dựng mô hình dự báo học máy. Sau khi hoàn tất quá trình tiền xử lý, tập dữ liệu đầu vào trở thành một tập dữ liệu hoàn chỉnh, sẵn sàng để huấn luyện mô hình dự báo.

Trong giai đoạn phân cụm dữ liệu, việc xác định số cụm tối ưu được thực hiện để đảm bảo phân nhóm hiệu quả và nâng cao độ chính xác cũng như khả năng diễn giải kết quả. Phương pháp Elbow (Hình 7) được sử dụng cho mục đích này, dựa trên tổng bình phương khoảng cách trong cụm (WCSS) để xác định số lượng cụm K . Công thức tính WCSS được định nghĩa như sau:

$$WCSS = \sum_{i=1}^K \sum_{x \in C_i} \|x - \mu_i\|^2 \quad (24)$$

Khi số cụm K tăng lên, giá trị WCSS sẽ giảm, nhưng đến một điểm nhất định, sự suy giảm này trở nên không đáng kể, tức là xuất hiện điểm gãy khuỷu tay (elbow), có nghĩa là việc tăng thêm số cụm K sẽ không cải thiện đáng kể kết quả phân cụm.



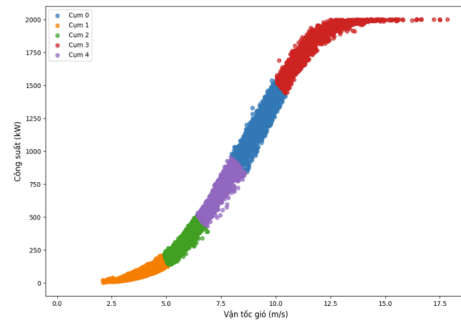
Hình 7. Đồ thị thuật toán Elbow

Kết quả được thể hiện ở Hình 7 cho thấy WCSS giảm đáng kể đến một điểm tại $K = 5$, sau đó sự suy giảm trở nên chậm lại và ổn định. Do đó, năm cụm được xác định là số lượng tối ưu để phân nhóm dữ liệu hiệu quả (Hình 8).

Thay vì tạo một mô hình duy nhất cho toàn bộ tập dữ liệu, một mô hình riêng biệt sẽ được phát triển cho từng cụm. Cách tiếp cận này được lựa chọn vì các điểm dữ liệu trong cùng một cụm thường có đặc điểm giống nhau. Do đó, mô hình của từng cụm sẽ được phân tích chi tiết và phù hợp hơn với các thuộc tính riêng của cụm đó, từ đó nâng cao độ chính xác của dự báo.

Các mô hình mạng nơ-ron nhân tạo MLP với 1, 2, 3, 4 và 5 lớp ẩn, mỗi lớp chứa 32 nơ-ron được thiết lập cho từng cụm dữ liệu. Số lượng nơ-ron, cấu trúc mạng và các tham số được xác định thông qua khảo sát thực nghiệm để cân bằng giữa độ phức tạp của mô hình và hiệu quả tính toán.

Sau khi huấn luyện và lưu trữ các tham số tối ưu, các mô hình này được so sánh để đánh giá mức độ tương đồng giữa giá trị dự báo và giá trị thực tế tại tuabin gió (Hình 9).

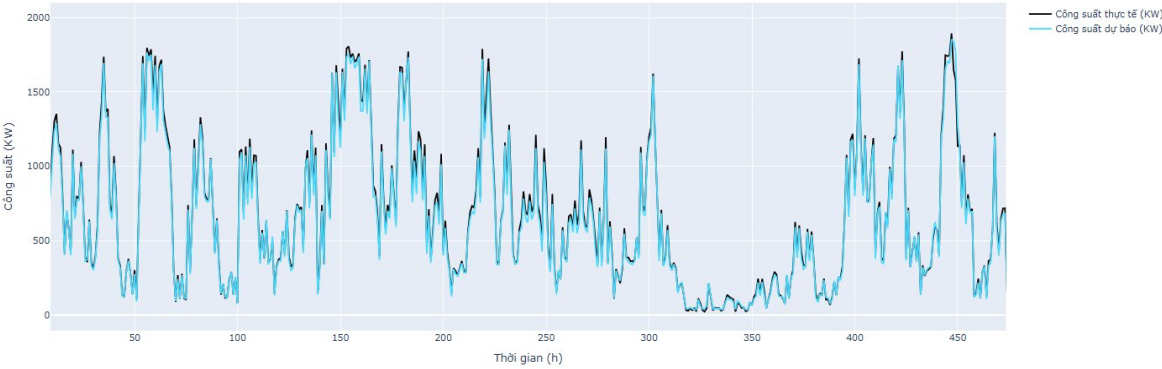


Hình 8. Đồ thị phân cụm với 5 cụm dữ liệu

Dựa trên tiêu chí chọn số lớp ẩn có sai số thấp nhất, ta có:

– Cụm 0: 3 hidden layers (RMSE = 38,06, MRE = 2,52%).

- Cụm 1: 5 hidden layers (RMSE = 12,58, MRE=1,5%).
- Cụm 2: 5 hidden layers (RMSE = 16,37, MRE = 1,6%).
- Cụm 3: 4 hidden layers (RMSE = 34,13, MRE = 1,50%).
- Cụm 4: 4 hidden layers (RMSE = 24,84, MRE = 2,67%).



Hình 9. Đồ thị tương quan công suất dự báo và thực tế

Bảng 2. Đánh giá sai số mô hình ở các cụm dữ liệu

Lớp ẩn	1		2		3		4		5	
Cụm	RMSE	MRE	RMSE	MRE	RMSE	MRE	RMSE	MRE	RMSE	MRE
0	48,67	3,12%	39,28	2,70%	38,06	2,52%	39,50	2,63%	42,08	2,85%
1	14,34	1,58%	12,68	1,51%	12,84	1,51%	12,71	1,7%	12,58	1,5%
2	19,33	1,74%	16,96	1,66%	16,94	1,66%	16,86	1,64%	16,37	1,6%
3	41,40	1,88%	46,39	1,93%	36,46	1,60%	34,13	1,50%	37,94	1,75%
4	32,62	3,67%	24,81	2,75%	24,38	3,66%	24,84	2,67%	24,66	2,69%

Trong điều kiện thực tế, sai số trong dự báo tốc độ gió có thể thay đổi do nhiều yếu tố khác nhau (địa hình, thời tiết cục đoạn, hạn chế của các mô hình dự báo thời tiết). Việc đưa sai số vào tập dữ liệu của mô hình dự báo năng lượng gió có nhiều mục đích giúp đánh giá khả năng áp dụng của mô hình trong các kịch bản khác nhau.

Thứ nhất, nó kiểm tra độ bền vững của mô hình bằng cách đánh giá khả năng xử lý dữ liệu không hoàn hảo, đảm bảo tính đáng tin cậy trong các tình huống thực tế. Thứ hai, nó xác thực độ chính xác của mô hình dự báo bằng cách mô phỏng các kịch bản thực tế, trong đó dữ liệu có thể bị nhiễu hoặc không chính xác, qua đó xác nhận hiệu quả của mô hình trong điều kiện không lý tưởng. Ngoài ra, việc thêm sai số giúp đánh giá khả năng tổng quát hóa của mô hình dự báo, xác định xem mô hình có thể hoạt động tốt, chính xác trên dữ liệu mới hay không.

Quá trình này cũng hỗ trợ tối ưu hóa các tham số của mô hình nhằm cải thiện độ chính xác của dự báo và tăng hiệu quả tính toán. Hơn nữa, việc đưa sai số vào giúp tăng sự đa dạng của dữ liệu, cho phép mô hình học được nhiều dạng mẫu khác nhau, từ đó cải thiện khả năng dự báo.

Cụ thể, trong bài báo này, sáu kịch bản đã được kiểm tra với các mức sai số ngẫu nhiên trong khoảng -10%, -20%, -30%, +10%, +20% và +30% sử dụng Việc phân phối Gaussian để mô hình hóa sai số dự báo (Xu et al., 2024). Cách tiếp cận này đảm bảo rằng mô hình dự báo năng lượng gió đạt độ chính xác cao, bền vững và đáng tin cậy trong các ứng dụng thực tiễn, qua đó hỗ trợ phát triển và sử dụng hiệu quả năng lượng gió như một nguồn năng lượng thay thế bền vững.

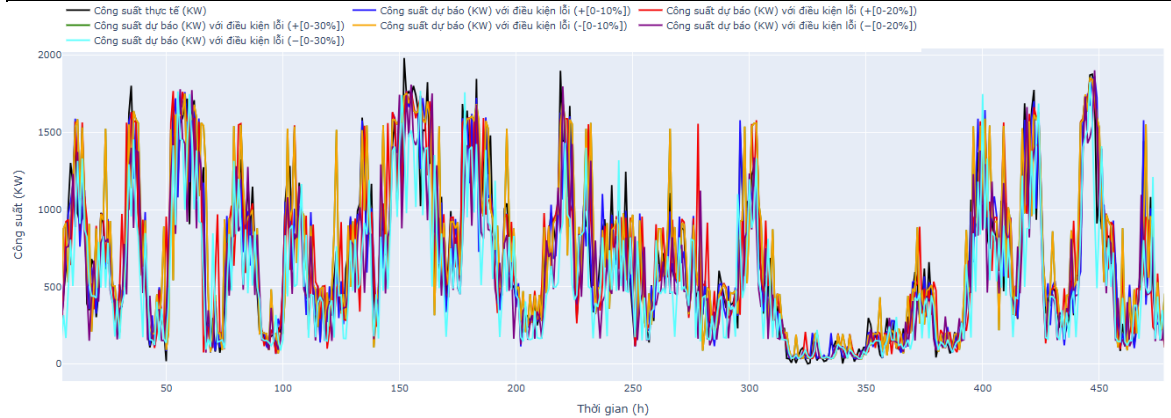
$$f_{v_{error}}(v_{error}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(v_{error} - \mu)^2}{2\sigma^2}\right) \quad (25)$$

Trong đó: $f(\bullet)$ là hàm mật độ xác suất của phân phối Gaussian, μ là giá trị trung bình hoặc kỳ vọng của phân phối và σ^2 là phương sai.

Mô hình dự báo với dữ liệu thời tiết có sai số trong sáu kịch bản: -10%, -20%, -30%, +10%, +20% và +30%. Kết quả được đánh giá bằng cách so sánh giá trị dự báo với giá trị công suất thực tế (Hình 10). Các kết quả của quá trình dự báo được tổng hợp trong Bảng 3.

Bảng 3. Đánh giá sai số mô hình với dữ liệu thời tiết có sai số

Sai số	-10%	-20%	-30%	+10%	+20%	+30%
RMSE	112,75	98,97	155,14	39,18	68,42	112,75
MRE	3,45%	3,34%	4,24%	1,30%	2,07%	3,45%



Hình 10. Đồ thị tương quan công suất dự báo và thực tế dữ liệu thời tiết có sai số

Khi so với các nghiên cứu trước (Lin & Liu, 2020; Yao et al., 2020; Zhang et al., 2020; Wang et al., 2021; Zhao & Bai, 2024; Tian et al., 2025), các chỉ số sai số như RMSE và MRE cũng được sử dụng trong các bài viết này để đánh giá hiệu quả của mô hình dự báo. Tuy nhiên, kết quả sai số từ các nghiên cứu này đều có giá trị cao hơn. Cụ thể, trong bài báo (Lin & Liu, 2020), giá trị sai số RMSE có thể dao động từ 517,33 cho đến hơn 22303.30; trong khi đó, ở bài báo (Tian et al., 2025), sai số dao động từ 101.7532 đến 110.4740, và trong bài báo (Zhang et al., 2020), sai số dao động từ 129,3 đến 175,1. So với giá trị sai số RMSE như trên, các kết quả thu được từ các cụm trong mô hình dự báo của nghiên cứu này chỉ dao động từ 38.06 trở xuống, cho thấy hiệu quả vượt trội của mô hình. Đối với chỉ số MRE, mức sai số ở các bài nghiên cứu khác thường vượt quá 5%, với sai số thấp nhất ghi nhận là 5.8942% trong nghiên cứu của Wang et al., (2021). Ở một nghiên cứu khác, Yao et al. (2020) việc so sánh nhiều phương pháp dự báo và nhận thấy mô hình LSTM (Long Short-Term Memory) đã được thực hiện và cho kết quả tốt nhất với sai số 19,24%. Bên cạnh đó, trong nghiên cứu của Zhao and Bai (2024), phương pháp dự báo trong hai điều kiện thời tiết đặc trưng – mùa hè và mùa đông đã được áp dụng và ghi nhận sai số lần lượt là 12.1573% và 21.6826%. Trong khi đó, kết quả của nghiên cứu này duy trì sai số dưới 5%. Điều này cho thấy mô hình trong bài báo này đạt hiệu quả dự báo tốt hơn về mặt sai số. Trong ba nghiên cứu của Yao et al. (2020), Wang et

al. (2021) và Zhao and Bai (2024), chỉ số sai số MAPE (Mean Absolute Percentage Error) đã được sử dụng, nhưng về bản chất, MAPE là một cách thể hiện sai số của MRE dưới dạng phần trăm. Do đó, việc so sánh kết quả sai số giữa các nghiên cứu là hợp lý và có cơ sở. Hơn nữa, các tập dữ liệu được sử dụng trong mỗi nghiên cứu là khác nhau. Vì vậy, việc so sánh với các nghiên cứu trước dựa trên sự tương đồng trong việc lựa chọn các chỉ số sai số RMSE và MRE được thực hiện, với kết quả của nghiên cứu này cho thấy sai số thấp hơn, khẳng định độ chính xác của mô hình dự báo.

4. KẾT LUẬN

Machine Learning trong dự báo công suất phát điện gió được ứng dụng trong nghiên cứu này với sai số trong từng kịch bản đều dưới 5%. Các bước phân tích dữ liệu, tiền xử lý dữ liệu (điền dữ liệu khuyết, loại bỏ dữ liệu ngoại lai) cùng với cách tiếp cận có hệ thống được trình bày trong nghiên cứu đã giúp xây dựng một mô hình dự báo công suất phát điện gió hiệu quả, đạt được dự báo có tỷ lệ sai số thấp và độ chính xác cao.

Ngoài ra, các kết quả từ bài báo này cũng có thể áp dụng cho việc dự báo công suất điện gió thực tế bằng cách sử dụng các tập dữ liệu huấn luyện lớn được thu thập từ các nhà máy điện gió trong và ngoài nước. Những kết quả này tạo ra một nền tảng vững chắc cho sự phát triển tương lai của công nghệ dự báo năng lượng tái tạo và việc ứng dụng nó trong điều độ công suất phát của hệ thống điện.

TÀI LIỆU THAM KHẢO (REFERENCES)

- Azimi, R., Ghofrani, M., & Ghayekhloo, M. (2016). A hybrid wind power forecasting model based on data mining and wavelets analysis. *Energy Conversion and Management*, 127, 208- 225. <https://doi.org/10.1016/j.enconman.2016.09.002>
- Bashir, T., Wang, H., Tahir, M., & Zhang, Y. (2025). Wind and solar power forecasting based on hybrid CNN-ABiLSTM, CNN-transformer-MLP models. *Renewable Energy*, 239, 122055. <https://doi.org/10.1016/j.renene.2024.122055>
- Fan, Y., Feng, C., Wu, R., Liu, C., & Jiang, D. (2024). Multiscale-attention masked autoencoder for missing data imputation of wind turbines. *Knowledge-Based Systems*, 299, 112114. <https://doi.org/10.1016/j.knosys.2024.112114>
- Hao, Y., Wang, X., Wang, J., & Yang, W. (2024). A new perspective of wind speed forecasting: Multi-objective and model selection-based ensemble interval-valued wind speed forecasting system. *Energy Conversion and Management*, 299, 117868. <https://doi.org/10.1016/j.enconman.2023.117868>
- He, Y., & Li, H. (2018). Probability density forecasting of wind power using quantile regression neural network and kernel density estimation. *Energy Conversion and Management*, 164, 374-384. <https://doi.org/10.1016/j.enconman.2018.03.010>
- Hu, Q., Zhang, S., Yu, M., & Xie, Z. (2016). Short-term wind speed or power forecasting with heteroscedastic support vector regression. *IEEE Transactions on Sustainable Energy*, 7(1), 241-249. <https://doi.org/10.1109/TSTE.2015.2480245>
- Kalpana, R., Subburaj, V., Lokanadham, R., Amudha, K., Beena Bethel, G. N., Shukla, A. K., Kshirsagar, P. R., & Rajaram, A. (2023). Internet of Things (IoT) based machine learning techniques for wind energy harvesting. *Electric Power Components and Systems*, 1-17. <https://doi.org/10.1080/15325008.2023.2293952>
- Liang, G., Su, Y., Wu, X., Ma, J., Long, H., & Song, Z. (2023). Abnormal data cleaning for wind turbines by image segmentation based on active shape model and class uncertainty. *Renewable Energy*, 216, 118965. <https://doi.org/10.1016/j.renene.2023.118965>
- Lin, Z., & Liu, X. (2020). Wind power forecasting of an offshore wind turbine based on high-frequency SCADA data and deep learning neural network. *Energy*, 201, 117693. <https://doi.org/10.1016/j.energy.2020.117693>
- Mohd, Y., & Singh, H. (2023). Machine learning for analysis and prediction of wind energy. *International Conference on Circuit Power and Computing Technologies 2023*, (pp.1438-1445). IEEE. <https://doi.org/10.1109/ICCPCT58313.2023.10245555>
- Naik, J., Satapathy, P., & Dash, P. K. (2018). Short-term wind speed and wind power prediction using hybrid empirical mode decomposition and kernel ridge regression. *Applied Soft Computing*, 70, 1167-1188. <https://doi.org/10.1016/j.asoc.2017.12.010>
- Ozkan, M. B., & Karagoz, P. (2015). A novel wind power forecast model: Statistical hybrid wind power forecast technique (SHWIP). *IEEE Transactions on Industrial Informatics*, 11(2), 375-387. <https://doi.org/10.1109/TII.2015.2396011>
- Pandit, R., Infield, D., & Santos, M. (2023). Accounting for environmental conditions in data-driven wind turbine power models. *IEEE Transactions on Sustainable Energy*, 14(1), 168-177. <https://doi.org/10.1109/TSTE.2022.3204453>
- Sireesha, P. V., & Thotakura, S. (2024). Wind power prediction using optimized MLP-NN machine learning forecasting model. *Electrical Engineering*, 106, 7643-7666. <https://doi.org/10.1007/s00202-024-02440-6>
- Tian, C., Niu, T., & Li, T. (2025). Developing an interpretable wind power forecasting system using a transformer network and transfer learning. *Energy Conversion and Management*, 323, 119155. <https://doi.org/10.1016/j.enconman.2024.119155>
- Tian, S., Fu, Y., Ling, P., Wei, S., Liu, S., & Li, K. (2018). *Wind power forecasting based on ARIMA-LGARCH model*. International Conference on Power System Technology 2018 (pp. 1285-1289). IEEE. <https://doi.org/10.1109/POWERCON.2018.8601740>
- Tuncar, E. A., Sağlam, Ş., & Oral, B. (2024). A review of short-term wind power generation forecasting methods in recent technological trends. *Energy Reports*, 12, 197-209. <https://doi.org/10.1016/j.egyr.2024.06.006>
- Wang, Y., Zou, R., Liu, F., Zhang, L., & Liu, Q. (2021). A review of wind speed and wind power forecasting with deep neural networks. *Applied Energy*, 304, 117766. <https://doi.org/10.1016/j.apenergy.2021.117766>
- Xiaochen, L., Xia, W., Liyong, Zh., Wei, L., & Chongquan, Zh. (2019). Imputations of missing values using a tracking-removed autoencoder trained with incomplete data. *Neurocomputing*,

- 366, 54-65.
<https://doi.org/10.1016/j.neucom.2019.07.066>.
- Xu, Z. Q., Xue, T., Chen, X. Y., Feng, J., Zhang, G. W., Wang, C., Lu, C. H., Chen, H. S., & Ding, Y. H. (2024). *Wind power correction model designed by the quantitative assessment for the impacts of forecasted wind speed error*. *Advances in Climate Change Research*.
<https://doi.org/10.1016/j.accre.2024.12.006>
- Yang, M., Huang, Y., Xu, C., Liu, C., & Dai, B. (2025). Review of several key processes in wind power forecasting: Mathematical formulations, scientific problems, and logical relations. *Applied Energy*, 377, 124631.
<https://doi.org/10.1016/j.apenergy.2024.124631>
- Yang, S., Yang, W., Wang, X., & Hao, Y. (2023). A novel selective ensemble system for wind speed forecasting: From a new perspective of multiple predictors for subseries. *Energy Conversion and Management*, 294, 117590.
<https://doi.org/10.1016/j.enconman.2023.117590>
- Yang, T., Yang, Z., Li, F., & Wang, H. (2024). A short-term wind power forecasting method based on multivariate signal decomposition and variable selection. *Applied Energy*, 360, 122759.
<https://doi.org/10.1016/j.apenergy.2024.122759>
- Yao, F., Liu, W., Zhao, X., & Song, L. (2020). Integrated Machine Learning and Enhanced Statistical Approach-Based Wind Power Forecasting in Australian Tasmania Wind Farm. *Complexity*, 2020(1), 9250937.
<https://doi.org/10.1155/2020/9250937>
- Zhang, Y., Li, Y., & Zhang, G. (2020). Short-term wind power forecasting approach based on Seq2Seq model using NWP data. *Energy*, 213, 118371.
<https://doi.org/10.1016/j.energy.2020.118371>
- Zhao, Z., & Bai, J. (2024). Ultra-Short-Term Wind Power Forecasting Based on the MSADBO-LSTM Model. *Energies*, 17(22), 5689.
<https://doi.org/10.3390/en17225689>