

ĐÁNH GIÁ ẢNH HƯỞNG CỦA TỶ LỆ TẬP DỮ LIỆU HUẤN LUYỆN/ KIỂM TRA ĐẾN ĐỘ CHÍNH XÁC CỦA DỰ BÁO CÔNG SUẤT PHÁT ĐIỆN MẶT TRỜI

EVALUATING THE IMPACT OF TRAINING/TEST DATA RATIOS ON THE ACCURACY OF SOLAR POWER FORECASTING

Nguyễn Tuấn Anh¹, Phạm Mạnh Hải^{1,*}, Lê Thành Doanh¹,
Nguyễn Ngọc Trung¹, Vũ Thị Anh Thơ¹, Trần Anh Tùng¹,
Vũ Minh Pháp², Nguyễn Hữu Nguyễn³

DOI: <http://doi.org/10.57001/huih5804.2025.110>

TÓM TẮT

Trong bài toán dự báo công suất phát của điện mặt trời, các thuật toán máy học dựa trên việc huấn luyện các tập dữ liệu quá khứ đã trở nên rất phổ biến. Một trong những yếu tố ảnh hưởng đáng kể đến hiệu suất của các mô hình dự báo là tỷ lệ phân chia tập dữ liệu huấn luyện và kiểm tra. Tỷ lệ không phù hợp có thể dẫn đến hiện tượng underfitting hoặc overfitting, làm suy giảm khả năng tổng quát hóa của mô hình. Trong nghiên cứu này, nhóm tác giả xây dựng mô hình Random Forest để dự báo công suất phát điện mặt trời và đánh giá hiệu suất mô hình với các tỷ lệ huấn luyện/kiểm tra khác nhau, từ 50/50 đến 90/10. Hai tập dữ liệu kiểm tra được thiết kế: một tập gồm các ngày có bức xạ biến động mạnh nhất, và một tập kết hợp cả các ngày có bức xạ ổn định nhất. Kết quả cho thấy, hiệu suất mô hình cải thiện rõ rệt khi tăng tỷ lệ dữ liệu huấn luyện. Tại tỷ lệ 90/10, mô hình đạt sai số MAPE thấp nhất (19,92%) trên tập kiểm tra tổng hợp. Ngoài ra, nghiên cứu cũng chỉ ra rằng MAPE là chỉ số nhạy cảm với dữ liệu công suất nhỏ, nên cần được đánh giá song song với RMSE và NMAPE để có đánh giá toàn diện. Nghiên cứu cung cấp cơ sở thực nghiệm quan trọng cho việc lựa chọn tỷ lệ huấn luyện phù hợp trong các bài toán dự báo công suất điện mặt trời.

Từ khóa: Dự báo bức xạ mặt trời, Random forest, huấn luyện/kiểm tra.

ABSTRACT

In the context of solar power forecasting, machine learning algorithms based on training historical datasets have become increasingly popular. One significant factor affecting forecasting model performance is the training and testing data ratio. An inappropriate ratio may lead to underfitting or overfitting, reducing the model's generalization ability. In this study, the authors developed a Random Forest model to forecast solar power output and evaluated its performance across various training/testing splits, ranging from 50/50 to 90/10. Two testing datasets were designed: one with days with the highest radiation variability and another combining days with stable and unstable radiation. The results show that model performance significantly improves as the proportion of training data increases. At a 90/10 split, the model achieved the lowest MAPE (19.92%) on the combined test set. Additionally, the study reveals that MAPE is highly sensitive to small power values, therefore, be interpreted alongside RMSE and NMAPE for a comprehensive performance evaluation. This research provides practical insights and empirical evidence for selecting appropriate training/testing ratios in solar power forecasting tasks.

Keywords: Solar power forecasting, Random forest, train/test.

¹Trường Đại học Điện lực

²Viện Khoa học công nghệ Năng lượng và Môi trường, Viện Hàn lâm Khoa học và Công nghệ Việt Nam

³Trường Đại học Công nghệ Đông Á

*Email: haipm@epu.edu.vn

Ngày nhận bài: 25/02/2025

Ngày nhận bài sửa sau phản biện: 05/5/2025

Ngày chấp nhận đăng: 28/5/2025

1. GIỚI THIỆU

Trong bối cảnh chuyển đổi năng lượng và nhu cầu sử dụng năng lượng tái tạo ngày càng tăng, việc dự báo chính xác công suất phát điện mặt trời đóng vai trò quan trọng trong quy hoạch lưới điện và tối ưu hóa vận hành hệ thống. Tuy nhiên, công suất phát điện mặt trời chịu ảnh hưởng mạnh từ các yếu tố môi trường như bức xạ mặt trời, nhiệt độ, mùa vụ,... [1], khiến bài toán dự báo trở nên phức tạp.

Việc phát triển các mô hình dự báo hiệu quả không chỉ mang ý nghĩa về mặt kỹ thuật mà còn có giá trị lớn về kinh tế và môi trường. Một hệ thống có khả năng dự báo chính xác sẽ giúp tối ưu hóa quá trình sản xuất và tiêu thụ điện, giảm phụ thuộc vào các nguồn điện truyền thống và thúc đẩy quá trình chuyển đổi năng lượng bền vững [2].

Hiệu suất của mô hình dự báo phụ thuộc vào nhiều yếu tố như chất lượng dữ liệu đầu vào, thuật toán học máy và đặc biệt là tỷ lệ tập huấn luyện/tập kiểm tra. Nếu tỷ lệ huấn luyện quá nhỏ, mô hình có thể không học đủ thông tin (underfitting), ngược lại nếu tập huấn luyện quá lớn, mô hình có thể ghi nhớ dữ liệu thay vì học được quy luật tổng quát (overfitting). Việc xác định tỷ lệ hợp lý là yếu tố quan trọng giúp mô hình có khả năng tổng quát hóa tốt nhất mà vẫn đạt được độ chính xác cao [3, 4]. Bảng 1 tổng hợp một số nghiên cứu tiêu biểu trong lĩnh vực dự báo công suất phát điện mặt trời có đề cập đến việc sử dụng và đánh giá các tỷ lệ huấn luyện/kiểm tra khác nhau.

Bảng 1. Một số nghiên cứu trong lĩnh vực dự báo công suất phát điện mặt trời có đề cập đến việc sử dụng và đánh giá các tỷ lệ huấn luyện/kiểm tra

Các nghiên cứu	Mô hình sử dụng	Tỷ lệ tập huấn luyện/tập kiểm tra	Kết quả chính
Enhancing solar power forecasting with machine learning using principal component analysis and diverse statistical indicators [5]	PCA + ML (LR, SVR, DTR, KNN, RF, XGBoost)	70/30	XGBoost có RMSE thấp nhất (12,7), chứng minh hiệu quả PCA và học máy.
Multivariate solar power time series forecasting using multilevel data fusion and deep neural networks [6]	MF-NBEA (multilevel fusion + N-BEATS + residual learner)	67/33	MF-NBEA vượt trội hơn các mô hình học sâu hiện tại, cung cấp dự báo dễ diễn giải và chính xác hơn.

Forecasting and Prediction of Solar Energy Generation using Machine Learning Techniques [7]	MLR, KNN, SVR, ANN	80/20	ANN đạt độ chính xác cao nhất với RMSE thấp và R ² cao trong dự báo sản lượng điện mặt trời.
Dự báo ngắn hạn sản lượng điện năng điện mặt trời mái nhà sử dụng mạng neuron nhân tạo [8]	MLFF, NARX, LSTM, NARX-LSTM	60/20/20 (60% huấn luyện, 20% xác thực, 20% kiểm tra)	MLFF có kết quả tốt nhất trong dự báo ngắn hạn 5 phút và 1 giờ.
Enhanced Solar Power Prediction Models with Integrating Meteorological Data Toward Sustainable Energy Forecasting [9]	Mô hình kết hợp giữa XGBoost - Extra Trees - Gradient Boosting	70/30	Mô hình lai đạt RMSE = 0,1087 và R ² = 0,9962 - kết quả tốt nhất trong các mô hình so sánh.
Transfer learning strategies for solar power forecasting under data scarcity [10]	Stacked LSTM + Transfer Learning	80/20	Transfer Learning giúp cải thiện RMSE ~12,6% và Skill Index ~16,3% so với mô hình LSTM huấn luyện từ đầu.

Trong nghiên cứu này, nhóm tác giả tiến hành phân tích chi tiết ảnh hưởng của tỷ lệ giữa tập huấn luyện/tập kiểm tra trên mô hình Random Forest trong dự báo công suất phát điện mặt trời. Random Forest (RF) là một thuật toán học máy dựa trên cấu trúc cây quyết định (decision tree) được đề xuất bởi Leo Breiman vào năm 2001 [11]. Đây là một phương pháp học tập tổng hợp (ensemble learning) mạnh mẽ, kết hợp nhiều cây quyết định nhỏ để tạo thành một mô hình dự đoán chính xác hơn. Random Forest có khả năng xử lý dữ liệu có nhiều biến đầu vào, đồng thời tránh được vấn đề overfitting (quá khớp), do đó rất phù hợp cho các bài toán dự báo phức tạp như dự báo công suất phát điện mặt trời. Alexandra Khalyasmaa và các cộng sự [12] đã sử dụng mô hình Random Forest để dự báo công suất phát điện mặt trời trong ngày tiếp theo. Kết quả dự báo cho một trạm điện mặt trời tại phía nam Liên Bang Nga thì độ chính xác của mô hình đạt được là 93%, các tác giả đánh giá mô hình dự báo đủ tốt cho các ứng dụng thực tế. Ming Meng và các cộng sự [13] cũng sử dụng mô hình RF để dự báo công suất phát điện mặt trời trong các ngày mùa đông tại Bắc Trung Quốc. Họ chia các ngày mùa đông

thành ba loại là ngày nhiều mây (cloudy days), ngày quang đãng (clear days), ngày mưa hoặc có tuyết rơi (rainy/snowy days) và xây dựng ba mô hình RF để dự báo cho ba loại ngày này. Kết quả thử nghiệm so sánh giữa ba mô hình dự báo khác nhau, bao gồm hồi quy vector hỗ trợ (Support Vector Regression-SVR), Lưới Đàn Hồi (Elastic Net- EN), và Cây Quyết Định Tăng Cường (Gradient Boosting Decision Tree- GBDT) cho thấy do khả năng giảm nguy cơ quá khớp (overfitting) của các cây quyết định thông qua việc cân bằng, mô hình RF để xuất đạt được các lỗi MAPE là 2,83% và 3,89% cho các ngày trời quang đãng và mây mù. Đối với các ngày có điều kiện thời tiết bất thường, sai số dự báo tương đối lớn. Trong các trường hợp như vậy, việc mở rộng mẫu huấn luyện, thực hiện phân loại và can thiệp thủ công có thể cải thiện độ chính xác của dự báo. Kết luận của họ là mô hình để xuất vượt trội hơn so với ba phương pháp khác đối với hầu hết các chỉ số đánh giá sai số trong mỗi loại.

Ảnh hưởng của tỷ lệ tập huấn luyện/tập kiểm tra đến độ chính xác của mô hình RF, sẽ được kiểm chứng bằng cách thử nghiệm tăng dần 5% tỉ lệ tập huấn luyện từ 50/50 cho đến 90/10. Kết quả sẽ giúp xác định tỷ lệ tối ưu nhằm nâng cao độ chính xác của dự báo công suất điện mặt trời. Bằng cách thực hiện các thử nghiệm trên nhiều tập dữ liệu khác nhau, nghiên cứu này sẽ cung cấp cái nhìn sâu hơn về cách chọn tỷ lệ huấn luyện hợp lý nhằm đạt hiệu suất tối ưu trong bài toán dự báo công suất điện mặt trời.

Điểm mới của nghiên cứu này là việc thiết kế hai tập kiểm tra đặc thù một tập gồm các ngày có mức dao động bức xạ lớn nhất và một tập dữ liệu mở rộng gồm các ngày có mức bức xạ thay đổi lớn nhất và nhỏ nhất. Cách tiếp cận này giúp kiểm tra khả năng tổng quát hóa của mô hình trong cả điều kiện bất lợi và thuận lợi điều chưa được thực hiện trong các nghiên cứu trước.

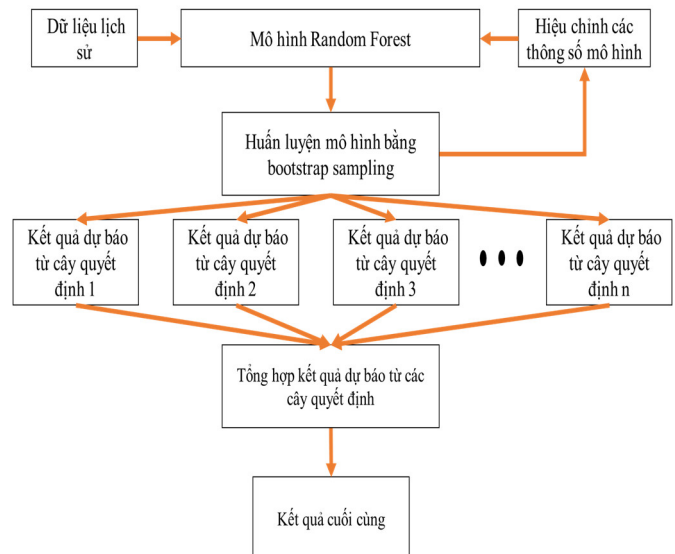
Nghiên cứu được chia thành 4 phần: phần 1 là giới thiệu chung về sự ảnh hưởng của quá trình huấn luyện mô hình tới dự báo, phần 2 trình bày về xây dựng mô hình dự báo và các tiêu chí đánh giá, phần 3 trình bày các kết quả và thảo luận, phần 4 là kết luận chung.

2. XÂY DỰNG MÔ HÌNH DỰ BÁO

Trong nghiên cứu này, nhóm tác giả chỉ sử dụng một thuật toán Random Forest làm nền tảng cho việc xây dựng mô hình dự báo, nhằm đảm bảo tính ổn định trong quá trình phân tích ảnh hưởng của tỷ lệ huấn luyện/kiểm tra. Sự nhất quán về mô hình cho phép đánh giá rõ ràng hơn hiệu ứng của thay đổi dữ liệu huấn luyện mà không bị chi phối bởi các yếu tố thuật toán khác.

2.1. Mô hình Random Forest

Mô hình RF hoạt động dựa trên việc kết hợp nhiều cây quyết định (Decision Trees) để tạo ra dự báo cuối cùng. Cấu trúc của mô hình RF được thể hiện trong hình 1.

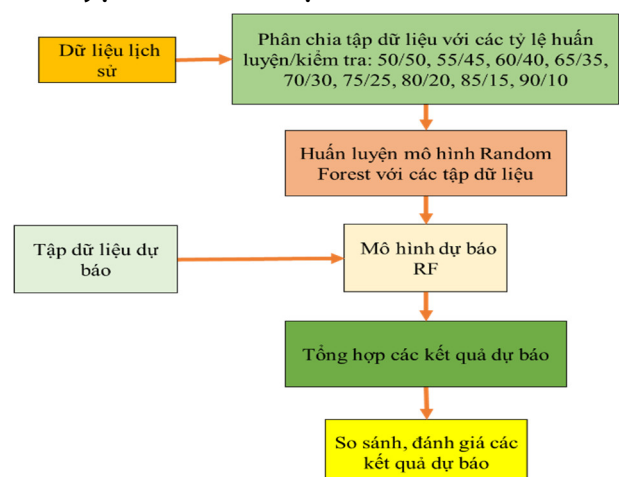


Hình 1. Cấu trúc của một mô hình Random Forest

Quá trình này được thực hiện theo các bước sau [14]:

1. Tạo tập dữ liệu con: Dữ liệu gốc được chia thành nhiều tập con bằng phương pháp bootstrap sampling.
2. Xây dựng các cây quyết định: Mỗi tập con được sử dụng để huấn luyện một cây quyết định riêng biệt.
3. Tạo dự báo từ từng cây: Mỗi cây quyết định đưa ra dự báo riêng lẻ.
4. Tổng hợp kết quả: Dự báo cuối cùng được tính bằng trung bình cộng (đối với hồi quy) hoặc bỏ phiếu đa số (đối với phân loại).

2.2. Sơ đồ thuật toán đánh giá ảnh hưởng của tỷ lệ tập huấn luyện tới mô hình dự báo

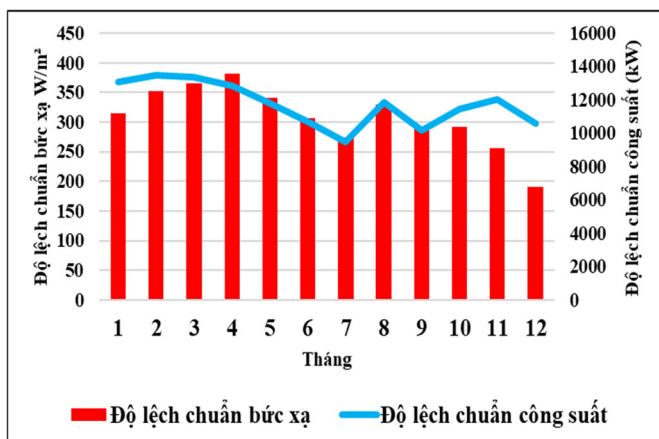


Hình 2. Sơ đồ thuật toán so sánh đánh giá ảnh hưởng của tập huấn luyện tới mô hình dự báo công suất phát điện mặt trời RF

Nghiên cứu này áp dụng một phương pháp tiếp cận có hệ thống để so sánh và đánh giá hiệu suất dự báo công suất phát điện mặt trời của mô hình RF dựa vào tỷ lệ tập huấn luyện/kiểm tra. Quy trình nghiên cứu được thiết kế để đảm bảo tính nhất quán và công bằng trong việc đánh giá các mô hình, bao gồm các bước chính được thể hiện trong hình 2.

2.2.1. Các bước thực hiện thuật toán

*) **Tập dữ liệu dùng để huấn luyện:** tập dữ liệu lịch sử bao gồm 98.208 điểm dữ liệu, được thu thập liên tục với chu kỳ lấy mẫu 5 phút trong suốt năm 2024 tại một nhà máy điện mặt trời 50MW ở khu vực Tây Nguyên. Các thông số trong tập dữ liệu lịch sử bao gồm: Bức xạ mặt trời, nhiệt độ môi trường, công suất phát và các tháng trong năm. Hình 3 minh họa mức độ biến động theo tháng của bức xạ và công suất thông qua chỉ số độ lệch chuẩn. Quá trình tiền xử lý bao gồm việc làm sạch dữ liệu, xử lý giá trị thiếu, và chuẩn hóa các đặc trưng. Dữ liệu sau đó được chia thành các tỷ lệ tập huấn luyện/tập kiểm tra: 50/50, 55/45, 60/40, 65/35, 70/30, 75/25, 80/20, 85/15, 90/10. Ngoài ra, một phần dữ liệu được tách ra để làm các tập dự báo độc lập (validation set), phục vụ mục tiêu đánh giá khả năng tổng quát hóa của mô hình trong các điều kiện thời tiết đặc trưng.



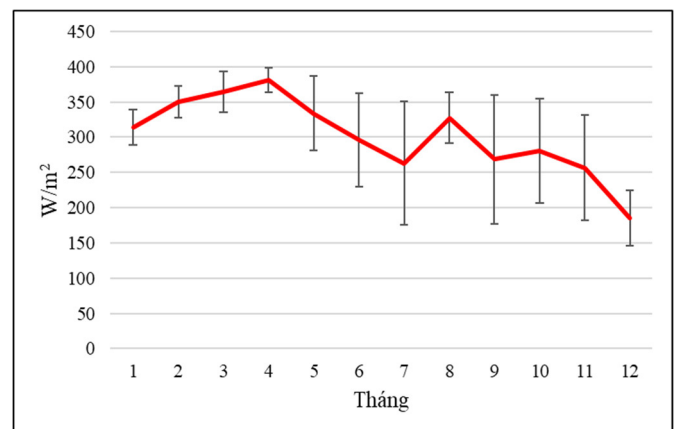
Hình 3. Độ lệch chuẩn theo tháng của bức xạ (W/m^2) và công suất phát (kW)

Hình 3 thể hiện độ lệch chuẩn theo tháng của bức xạ mặt trời (trục trái - W/m^2) và công suất phát điện (trục phải - kW) trong năm 2024. Có thể thấy rõ rằng trong các tháng 7 đến 10, độ biến động của cả hai đại lượng đều tăng cao, phản ánh ảnh hưởng của điều kiện thời tiết mùa mưa tại khu vực Tây Nguyên. Ngược lại, các tháng đầu và cuối năm có xu hướng ổn định hơn. Sự phân tích này đóng vai trò quan trọng trong việc lựa chọn các ngày đại diện cho tập dữ liệu dự báo, nhằm đánh giá hiệu quả mô hình trong các điều kiện thời tiết đặc trưng khác nhau.

*) Tập dữ liệu dự báo:

Để đảm bảo tính khách quan và đánh giá khả năng tổng quát hóa của các mô hình, nhóm tác giả đã sử dụng tập dữ liệu độc lập không thuộc các tập dữ liệu huấn luyện và kiểm tra để kiểm tra hiệu suất dự báo. Tập dữ liệu dự báo gồm hai tập dữ liệu thời tiết như sau:

Tập thứ nhất bao gồm các ngày có mức biến động bức xạ mặt trời lớn nhất trong từng tháng của năm 2024, cụ thể là các ngày: 30/01, 26/02, 26/03, 06/04, 10/05, 14/06, 25/07, 25/08, 02/09, 14/10, 11/11 và 22/12. Những ngày này được tách ra từ tập dữ liệu lịch sử. Phương pháp xác định dựa trên việc tính độ lệch chuẩn bức xạ theo từng ngày trong tháng, sau đó chọn ra ngày có độ lệch chuẩn cao nhất đại diện cho điều kiện thời tiết biến động mạnh, ảnh hưởng rõ rệt đến công suất phát điện. Hình 4 minh họa độ lệch chuẩn bức xạ của các ngày được chọn, kèm theo thanh sai số thể hiện mức chênh lệch so với độ lệch chuẩn trung bình của tháng. Có thể thấy trong các tháng như 7, 9 và 10, mức dao động cao hơn rõ rệt so với trung bình, khẳng định rằng những ngày này có thể tạo ra thách thức lớn cho mô hình dự báo.



Hình 4. Độ lệch chuẩn bức xạ của các ngày có biến động bức xạ lớn nhất

Tập dữ liệu dự báo thứ hai là tập dữ liệu mở rộng của tập dữ liệu thứ nhất, được kết hợp từ các ngày có bức xạ biến động mạnh nhất (các ngày của tập dữ liệu 1) và yếu nhất trong năm 2024 bao gồm các ngày: 06/01, 27/02, 29/03, 19/04, 29/05, 23/06, 18/07, 30/08, 18/09, 28/10, 08/11, 13/12 (những ngày này đã được tách khỏi tập dữ liệu lịch sử). Việc đưa vào cả những ngày có điều kiện bức xạ ổn định giúp tạo ra một tập dữ liệu kiểm tra đa dạng hơn, phản ánh đầy đủ cả hai kịch bản thời tiết đối lập. Qua đó, mô hình không chỉ được đánh giá trong điều kiện bất lợi mà còn trong cả điều kiện thuận lợi khi bức xạ ổn định và ít biến động.

Hai tập dự báo độc lập giúp đánh giá hiệu quả mô hình trong điều kiện dữ liệu bất lợi và ổn định, đồng thời

kiểm tra khả năng tổng quát hóa và hiện tượng overfitting hoặc underfitting.

***) Xây dựng và huấn luyện mô hình:** Với mỗi tập huấn luyện:

- Áp dụng Bootstrap Sampling để tạo các tập con ngẫu nhiên.
- Xây dựng n cây quyết định với các tập con khác nhau.
- Huấn luyện mô hình trên tập huấn luyện.

***) Đánh giá mô hình:** Các mô hình được huấn luyện trên tập dữ liệu huấn luyện và đánh giá ban đầu dựa trên tập dữ liệu kiểm tra. Quá trình đánh giá tập trung vào yếu tố là độ chính xác của dự báo được thể hiện qua các chỉ số đánh giá RMSE, NRMSE, MAPE, NMAPE.

2.2.2. Các chỉ số đánh giá

Các chỉ số đánh giá độ chính xác của các mô hình dự báo thường được các nhóm nghiên cứu sử dụng trong dự báo bức xạ mặt trời được thể hiện bằng các chỉ số sau đây:

a) Sai số RMSE

RMSE là sai số trung bình bình phương gốc giữa giá trị dự báo và giá trị thực tế. Công thức tính RMSE như sau [15]:

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

Trong đó: $\hat{y}_i$ là công suất dự báo (kW), y_i là công suất thực tế (kW), n là số lượng điểm dữ liệu.

RMSE đo lường độ lớn trung bình của lỗi dự báo và nhạy cảm với các lỗi lớn do việc bình phương sai số. Đơn vị của RMSE giống với đơn vị của biến được dự báo, giúp dễ dàng diễn giải. RMSE thấp hơn chỉ ra dự báo chính xác hơn.

b) Sai số NRMSE

NRMSE [16] là sai số trung bình bình phương đã được chuẩn hóa. Công thức tính NRMSE như sau:

$$NRMSE = \frac{\sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2}}{y_{\text{công suất lắp đặt}}} \quad (2)$$

Trong đó: $\hat{y}_i$ là công suất dự báo (kW), y_i là công suất thực tế (kW), n là số lượng điểm dữ liệu, $y_{\text{công suất lắp đặt}}$ là tổng công suất lắp đặt của nhà máy (MW).

c) Sai số MAPE

MAPE [12] là sai số tuyệt đối trung bình phần trăm. Sai số MAPE cho biết tỷ lệ phần trăm trung bình của sai số tuyệt đối so với giá trị thực tế. công thức tính MAPE như sau:

$$MAPE = \frac{100}{n} \sum_{i=1}^n \frac{|P_i^{db} - P_i^{tt}|}{P_i^{tt}} \quad (3)$$

Trong đó: P_i^{db} giá trị công suất phát của dự báo thứ i (MW), P_i^{tt} giá trị công suất trong thực tế của tín hiệu thứ i (MW), n là số tín hiệu dự báo được đánh giá trong khung thời gian của loại dự báo.

d) Sai số NMAPE

NMAPE [17] là sai số tuyệt đối phần trăm trung bình đã được chuẩn hóa. Công thức tính sai số NMAPE như sau:

$$NMAPE = \frac{100}{n} \sum_{i=1}^n \frac{|P_i^{db} - P_i^{tt}|}{P_{dm}} \quad (4)$$

Trong đó: P_i^{db} giá trị công suất phát của dự báo thứ i (MW), P_i^{tt} giá trị công suất trong thực tế của tín hiệu thứ i (MW), P_{dm} là công suất lắp đặt định mức của nhà máy (MW).

3. KẾT QUẢ VÀ THẢO LUẬN

3.1. Kết quả

Nhóm nghiên cứu tiến hành quá trình đánh giá mô hình qua các kết quả dự báo của các mô hình RF theo các tập dữ liệu dự báo. Việc thực hiện hai kịch bản dự báo này nhằm đánh giá sự tác động của miền dự báo tới hiệu suất làm việc của các mô hình. Kết quả dự báo cho các ngày có lượng bức xạ thay đổi lớn nhất được trình bày chi tiết trong bảng 2.

Bảng 2. Kết quả dự báo công suất phát của mô hình RF cho các ngày có lượng bức xạ thay đổi lớn nhất

Mô hình với các tập huấn luyện theo tỷ lệ tập huấn luyện/ kiểm tra	Sai số RMSE (kW)	Sai số NRMSE (%)	Sai số MAPE (%)	Sai số NMAPE (%)
50% dữ liệu huấn luyện - 50% kiểm tra	1754,53	3,51	20,1	1,58
55% dữ liệu huấn luyện - 45% kiểm tra	1658,17	3,37	19,2	1,49
60% dữ liệu huấn luyện - 40% kiểm tra	1554,45	3,11	17,64	1,37
65% dữ liệu huấn luyện - 35% kiểm tra	1449,28	2,98	18,26	1,31
70% dữ liệu huấn luyện - 30% kiểm tra	1390,2	2,78	14,76	1,22
75% dữ liệu huấn luyện - 25% kiểm tra	1284,74	2,57	13,77	1,13
80% dữ liệu huấn luyện - 20% kiểm tra	1193,1	2,39	18,9	1,07

85% dữ liệu huấn luyện - 15% kiểm tra	1129,94	2,25	12,91	1,00
90% dữ liệu huấn luyện - 10% kiểm tra	1055,84	2,11	12,35	0,93

Bảng 2 thể hiện sự thay đổi của các chỉ số sai số (RMSE, NRMSE, MAPE, NMAPE) theo các tỷ lệ chia tập huấn luyện và kiểm tra khác nhau. Có thể thấy rằng, khi tỷ lệ dữ liệu huấn luyện tăng lên từ 50% đến 90%, tất cả các chỉ số sai số đều có xu hướng giảm, phản ánh rõ sự cải thiện hiệu suất của mô hình dự báo công suất phát điện mặt trời. Cụ thể, chỉ số RMSE giảm đều từ 1754,53kW xuống còn 1055,84kW, đồng thời NRMSE cũng giảm từ 3,509 xuống 2,11, cho thấy mô hình học tốt hơn khi được cung cấp nhiều dữ liệu hơn. Các chỉ số MAPE và NMAPE cũng giảm mạnh, từ 20,99% xuống 12,35% và từ 1,58% xuống 0,93% tương ứng. Tuy nhiên, tại tỷ lệ 80/20, MAPE có tăng nhẹ lên 18,9%, có thể do ảnh hưởng của dữ liệu kiểm tra nhỏ hoặc một số ngoại lệ trong tập dữ liệu. Tổng thể, mô hình hoạt động tốt hơn rõ rệt khi tỷ lệ dữ liệu huấn luyện cao hơn, đặc biệt từ tỷ lệ 75/25 trở lên. Do đó, các tỷ lệ như 85/15 hoặc 90/10 có thể sẽ là lựa chọn tối ưu nếu mục tiêu là đạt độ chính xác cao nhất. Trong khi đó, nếu cần đảm bảo độ đa dạng của tập kiểm tra để đánh giá toàn diện hơn, tỷ lệ 75/25 có thể là lựa chọn cân bằng giữa hiệu suất và khả năng kiểm chứng. Tuy nhiên, việc lựa chọn một tỉ lệ lớn cho tập huấn luyện có thể dẫn đến khả năng overfitting tùy thuộc tập dữ liệu quá khứ và bản chất của đối tượng dự báo.

Để đánh giá khả năng dự báo của mô hình trong các điều kiện thời tiết khác nhau, nhóm nghiên cứu đã lựa chọn các ngày có độ biến động bức xạ mặt trời cao nhất và thấp nhất trong từng tháng. Các ngày này đại diện cho các tình huống bất ổn và ổn định về bức xạ, qua đó cho phép kiểm tra hiệu suất mô hình trong cả điều kiện phức tạp lẫn lý tưởng.

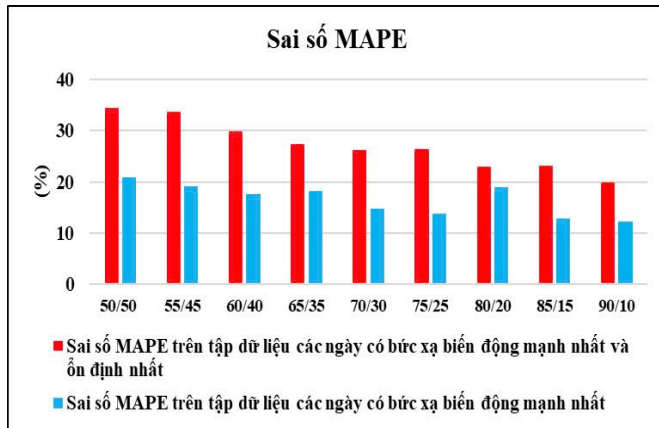
Bảng 3. Kết quả dự báo công suất phát của mô hình RF với tập dữ liệu dự báo các ngày có lượng bức xạ thay đổi lớn nhất và nhỏ nhất

Mô hình với các tập huấn luyện theo tỷ lệ tập huấn luyện/kiểm tra	Sai số RMSE (kW)	Sai số NRMSE (%)	Sai số MAPE (%)	Sai số NMAPE (%)
50% dữ liệu huấn luyện - 50% kiểm tra	2536,7	5,07	34,52	2,21
55% dữ liệu huấn luyện - 45% kiểm tra	2467,91	4,94	33,73	2,15
60% dữ liệu huấn luyện - 40% kiểm tra	2369,51	4,74	29,86	2,02

65% dữ liệu huấn luyện - 35% kiểm tra	2269,32	4,54	27,38	1,93
70% dữ liệu huấn luyện - 30% kiểm tra	2185,17	4,37	26,18	1,83
75% dữ liệu huấn luyện - 25% kiểm tra	2064,66	4,13	26,52	1,72
80% dữ liệu huấn luyện - 20% kiểm tra	1947,42	3,89	23	1,61
85% dữ liệu huấn luyện - 15% kiểm tra	1775,61	3,55	23,23	1,49
90% dữ liệu huấn luyện - 10% kiểm tra	1593,14	3,19	19,92	1,37

Bảng 3 cho thấy hiệu suất mô hình khi áp dụng dự báo trên tập dữ liệu gồm các ngày có bức xạ mặt trời biến động mạnh nhất và yếu nhất trong năm 2024 với nhiều tỷ lệ huấn luyện/kiểm tra khác nhau. Nhìn chung, các chỉ số sai số đều có xu hướng giảm ổn định khi tăng tỷ lệ huấn luyện. Cụ thể, RMSE giảm từ 2536,7kW (với tỷ lệ 50/50) xuống còn 1593,14kW (với tỷ lệ 90/10), và NRMSE giảm từ 5,07 xuống 3,19 phản ánh khả năng học tốt hơn của mô hình khi được cung cấp nhiều dữ liệu huấn luyện hơn. Chỉ số MAPE cũng giảm từ 34,52% xuống 19,92%, tuy nhiên có một điểm bất thường tại tỷ lệ 80/20, khi MAPE tăng lên 23,23%. Nguyên nhân có thể do sự xuất hiện của các giá trị công suất thực tế thấp trong tập kiểm tra, khiến sai số phần trăm bị ảnh hưởng lớn mặc dù sai số tuyệt đối không lớn. Ngược lại, chỉ số NMAPE giảm đều từ 2,21% xuống còn 1,37%, thể hiện độ chính xác tổng thể ổn định và đáng tin cậy hơn so với MAPE, đặc biệt trong điều kiện dữ liệu đa dạng. Việc kết hợp các ngày có mức biến động bức xạ đối lập trong tập kiểm tra đã tạo ra kịch bản dữ liệu có tính đại diện cao, bao gồm cả điều kiện thuận lợi và bất lợi. Kết quả này cho thấy mô hình vẫn duy trì được hiệu suất tốt và có khả năng khái quát hóa trong các tình huống thực tế phức tạp.

Trong số các chỉ số sai số được sử dụng, nhóm nghiên cứu lựa chọn MAPE để trực quan hóa vì đây là chỉ số nhạy cảm với chênh lệch tương đối và thường được sử dụng phổ biến trong các bài toán dự báo có đơn vị đầu ra biến động lớn như công suất điện mặt trời. MAPE cho thấy rõ ràng tác động của các tình huống cực đoan, đặc biệt là khi công suất thực tế thấp. Do đó, biểu đồ MAPE là công cụ trực quan phù hợp để làm nổi bật sự khác biệt giữa hai tập kiểm tra, cũng như phản ánh ảnh hưởng của tỷ lệ huấn luyện đến khả năng tổng quát hóa của mô hình. Hình 5 thể hiện sai số MAPE của mô hình theo từng tỷ lệ huấn luyện/kiểm tra, áp dụng trên hai tập kiểm tra nói trên.



Hình 5. Biểu đồ sai số MAPE theo tỷ lệ tập huấn luyện của mô hình RF trên hai tập kiểm tra

Kết quả từ hình 5 cho thấy sự khác biệt rõ rệt về sai số MAPE giữa hai tập kiểm tra. Mô hình RF có MAPE tại tập dữ liệu của các ngày có bức xạ thay đổi lớn nhất và bé nhất (các cột màu đỏ) luôn cao hơn so với khi chỉ kiểm tra trên tập dữ liệu các ngày có bức xạ thay đổi lớn nhất (các cột màu xanh). Điều này phản ánh rằng tập dữ liệu có tính hỗn hợp không làm bài toán trở nên dễ hơn, mà ngược lại còn làm sai số tăng lên do xuất hiện nhiều giá trị công suất thực tế thấp - vốn khiến MAPE nhạy cảm và dễ bị đẩy cao.

Sự chênh lệch giữa hai tập kiểm tra là rõ ràng hơn ở các tỷ lệ huấn luyện thấp (50/50, 55/45 và 65/35), cho thấy mô hình gặp khó khăn trong việc tổng quát hóa khi lượng dữ liệu huấn luyện còn hạn chế. Tuy nhiên, khi tỷ lệ huấn luyện tăng lên từ 70/30 trở đi, sai số MAPE ở cả hai tập đều giảm đáng kể, cho thấy hiệu suất mô hình được cải thiện ổn định.

Đáng chú ý, tại tỷ lệ 85/15 và 90/10, cả hai tập đều cho sai số MAPE thấp nhất, và khoảng cách giữa hai đường biểu diễn cũng thu hẹp lại. Điều này cho thấy mô hình đã đạt mức độ học đủ tốt để thích ứng với cả những tình huống dữ liệu biến động khó lường, đồng thời cũng phản ánh rằng tỷ lệ huấn luyện từ 80/20 trở lên là phù hợp và đáng được khuyến nghị trong các bài toán dự báo công suất điện mặt trời có dữ liệu kiểm tra đa dạng.

3.2. Thảo luận

Kết quả nghiên cứu cho thấy rằng tỷ lệ tập huấn luyện/tập kiểm tra có ảnh hưởng rõ rệt đến hiệu suất dự báo của mô hình RF. Khi tỷ lệ huấn luyện quá thấp (ví dụ 50/50), mô hình không có đủ dữ liệu để học, dẫn đến sai số cao, đặc biệt là MAPE - chỉ số phản ánh độ chính xác tương đối. Tại tỷ lệ này, MAPE lên tới 34,52% trên tập dữ liệu gồm cả ngày có bức xạ biến động mạnh nhất và ổn định nhất. Trong khi đó, khi tăng dần tỷ lệ huấn luyện, sai

số giảm dần rõ rệt, thể hiện ở sự cải thiện liên tục của RMSE, NRMSE và NMAPE. Đặc biệt, từ tỷ lệ 80/20 trở lên, mô hình cho thấy độ chính xác ổn định và khả năng khái quát hóa tốt hơn.

Để đánh giá sâu hơn hiệu suất mô hình trong các điều kiện bức xạ khác nhau, nghiên cứu đã thiết kế hai tập kiểm tra độc lập: (1) tập các ngày có bức xạ thay đổi mạnh nhất trong năm và (2) tập kết hợp các ngày có bức xạ thay đổi mạnh nhất và ổn định nhất. Biểu đồ MAPE theo các tỷ lệ huấn luyện cho thấy mô hình hoạt động tốt hơn khi kiểm tra trên tập (1), với sai số luôn thấp hơn so với tập (2). Điều này phản ánh rằng việc đưa thêm các ngày có bức xạ ổn định vào tập kiểm tra không làm sai số giảm, mà ngược lại còn làm tăng MAPE do ảnh hưởng của các giá trị công suất thực tế thấp. MAPE vốn nhạy cảm với các giá trị nhỏ, nên dễ bị khuếch đại trong các khung giờ sáng sớm hoặc chiều muộn, khi công suất gần bằng 0.

Mặc dù vậy, các chỉ số khác như RMSE và NMAPE vẫn cho thấy hiệu suất dự báo ổn định và đáng tin cậy. Do đó, trong đánh giá mô hình dự báo công suất điện mặt trời, cần kết hợp nhiều chỉ số khác nhau để có cái nhìn toàn diện hơn, thay vì chỉ dựa vào MAPE.

Những kết quả thu được cho thấy rõ tác động của tỷ lệ huấn luyện đến hiệu suất dự báo. Trong các nghiên cứu tiếp theo, nhóm tác giả dự kiến sẽ mở rộng thử nghiệm với các thuật toán khác như XGBoost, LSTM hoặc Transformer nhằm kiểm chứng lại xu hướng và kết quả khi thay đổi mô hình dự báo.

4. KẾT LUẬN

Nghiên cứu này đã phân tích ảnh hưởng của tỷ lệ tập huấn luyện/tập kiểm tra đến độ chính xác của mô hình Random Forest trong dự báo công suất điện mặt trời. Kết quả chỉ ra rằng tỷ lệ huấn luyện càng cao thì hiệu suất mô hình càng được cải thiện rõ rệt, đặc biệt là từ tỷ lệ 75/25 trở lên. Tỷ lệ 90/10 cho kết quả MAPE thấp nhất ở cả hai tập dữ liệu kiểm tra, cho thấy đây có thể là một lựa chọn tối ưu nếu mục tiêu là đạt độ chính xác cao (trừ trường hợp overfitting).

Ngoài ra, việc xây dựng hai tập kiểm tra với đặc trưng bức xạ đối lập đã giúp kiểm chứng khả năng khái quát hóa của mô hình trong các điều kiện thực tế đa dạng. Tập kết hợp các ngày có bức xạ dao động mạnh và yếu đặt ra thử thách lớn hơn cho mô hình, nhưng đồng thời phản ánh khả năng thích ứng của mô hình trong môi trường thực tế. Biểu đồ MAPE đã minh họa rõ sự khác biệt này và cho thấy rằng chỉ số này, dù phổ biến, vẫn có giới hạn khi công suất thực tế thấp.

Do đó, việc lựa chọn tỷ lệ huấn luyện hợp lý cần căn cứ vào mục tiêu sử dụng mô hình: nếu ưu tiên độ chính xác, tỷ lệ từ 80/20 trở lên là có thể được lựa chọn; nếu cần đánh giá toàn diện, tỷ lệ 75/25 đến 85/15 có thể phù hợp hơn. Trong tương lai, nghiên cứu có thể mở rộng với các mô hình tiên tiến như XGBoost hoặc mạng nơ-ron nhân tạo để tiếp tục đánh giá hiệu quả mô hình trong các điều kiện dữ liệu phức tạp hơn đối với từng bộ dữ liệu.

TÀI LIỆU THAM KHẢO

- [1]. N. Tuan Anh, P. Manh-Hai, V. Minh Phap, D. Trung Kien, "A recent invasion wave of deep learning in solar power forecasting techniques using ANN," *IEEE Int. Futur. Energy Electron. Conf.*, 2021. doi: 10.1109/IFEEEC53238.2021.9661747.
- [2]. L. J. Wang Zhen, "Short-Term Photovoltaic Power Generation Prediction Based on LightGBM-LSTM Model," *Mod. Electr. Power*, 40, 6, 899-905, 2023. doi: 10.19725/j.cnki.1007-2322.2022.0111.
- [3]. J. Gaboitaolelwe, A. M. Zungeru, A. Yahya, C. K. Lebekwe, D. N. Vinod, A. O. Salau, "Machine Learning Based Solar Photovoltaic Power Forecasting: A Review and Comparison," *IEEE Access*, 11, no. April, 40820-40845, 2023. doi: 10.1109/ACCESS.2023.3270041.
- [4]. J. Ye, B. Zhao, H. Deng, "Photovoltaic Power Prediction Model Using Pre-train and Fine-tune Paradigm Based on LightGBM and XGBoost," *Procedia Comput. Sci.*, 224, 407-412, 2023. doi: 10.1016/j.procs.2023.09.056.
- [5]. Y. Djeldjeli, et al., "Enhancing solar power forecasting with machine learning using principal component analysis and diverse statistical indicators," *Case Stud. Therm. Eng.*, 61, no. August, 104924, 2024. doi: 10.1016/j.csite.2024.104924.
- [6]. S. Almaghrabi, M. Rana, M. Hamilton, M. Saiedur Rahaman, "Multivariate solar power time series forecasting using multilevel data fusion and deep neural networks," *Inf. Fusion*, 104, no. September 2023, 102180, 2024. doi: 10.1016/j.inffus.2023.102180.
- [7]. A. J. M. Lari, *Forecasting and Prediction of Solar Energy Generation using Machine Learning Techniques*. Doctor of Philosophy, Department of Electronic and Electrical Engineering Swansea University, 2023.
- [8]. P. K. Nguyễn, L. Trần Kim, C. Hà Phú, H. Nguyễn Minh, H. Hoàng Công, "Dự báo ngắn hạn sản lượng điện năng điện mặt trời mái nhà sử dụng mạng neuron nhân tạo," *Can Tho Univ. J. Sci.*, 59, no. ETMD, 251-260, 2023. doi: 10.22144/ctu.jvn.2023.051.
- [9]. M. A. Atiea, A. M. Shaheen, A. Alassaf, I. Alsaleh, "Enhanced Solar Power Prediction Models With Integrating Meteorological Data Toward Sustainable Energy Forecasting," *International Journal of Energy Research*, 2024.

- [10]. E. Sarmas, N. Dimitropoulos, V. Marinakis, Z. Mylona, H. Doukas, "Transfer learning strategies for solar power forecasting under data scarcity," *Sci. Rep.*, 12, 1, 1-14, 2022. doi: 10.1038/s41598-022-18516-x.
- [11]. Leo Breiman, "Random Forests," *Statistics Dep. Univ. Calif. Berkeley, CA 94720*, 1-33, 2001. doi: 10.1023/A:1010950718922.
- [12]. S. K. Alexandra Khalyasmaa, Stanislav A. Eroshenko, Teja Piepur Chakravarthy, Venu Gopal Gasi, Sandeep Kumar Yadav Bollu, Raphaël Caire, Sai Kumar Reddy Atluri, "Prediction of Solar Power Generation Based on Random Forest Regressor Model Ale," in *2019 Int. Multi-Conference Eng. Comput. Inf. Sci.*, 1-10, 2019.
- [13]. M. Meng, C. Song, "Daily photovoltaic power generation forecasting model based on random forest algorithm for north china in winter," *Sustain.*, 12, 6, 2020. doi: 10.3390/su12062247.
- [14]. R. Jiang, W. Tang, X. Wu, W. Fu, "A random forest approach to the detection of epistatic interactions in case-control studies," *BMC Bioinformatics*, 10, no. SUPPL. 1, 1-12, 2009. doi: 10.1186/1471-2105-10-S1-S65.
- [15]. E. Nziyumva, R. Hu, Y. Xu, J. Niyogisubizo, Z. Li, P. C. Nshimyumukiza, "Solar Radiation Forecasting Using Ensemble-Based Hybrid LGBM-GB-MLP Model: A Novel Stacked Generalization Method," *SSRN Electron. J.*, no. October, 2022. doi: 10.2139/ssrn.4076358.
- [16]. Q. T. Phan, Y. K. Wu, Q. D. Phan, H. Y. Lo, "A Novel Forecasting Model for Solar Power Generation by a Deep Learning Framework with Data Preprocessing and Postprocessing," *IEEE Trans. Ind. Appl.*, 59, 1, 220-231, 2023. doi: 10.1109/TIA.2022.3212999.
- [17]. MOIT, *Decision 67 QĐ/ĐTĐL, Forecasting capacity and electricity generation of renewable energy power sources*. Vietnam, 2021.

AUTHORS INFORMATION

**Nguyen Tuan Anh¹, Pham Manh Hai¹, Le Thanh Doanh¹,
Nguyen Ngoc Trung¹, Vu Thi Anh Tho¹, Tran Anh Tung¹,
Vu Minh Phap², Nguyen Huu Nguyen³**

¹Electric Power University, Vietnam

²Institute of Science and Technology for Energy and Environment,
Vietnam Academy of Science and Technology, Vietnam

³East Asia University of Technology, Vietnam