DEVELOPING A HYBRID MODEL OF CNN-LSTM FOR SHORT-TERM PV POWER OUTPUT FORECASTING AT NHI HA SOLAR FARM

Nguyen Thi Hoai Thu*, Pham Nang Van, Tran Quang Khai
School of Electrical and Electronic Engineering - Hanoi University of Science and Technology

ARTICLE INFO

Received: 03/12/2024
Revised: 03/01/2025
Published: 04/01/2025

KEYWORDS

Forecasting
Short-term
Solar power
Hybrid model
CNN-LSTM

ABSTRACT

Integrating solar power into the grid brings not only significant economic and environmental benefits but also challenges for the power system. Accurate short-term solar power forecasting is one of the key solutions to address these challenges. This study proposes a hybrid model combining Convolutional Neural Networks and Long Short-Term Memory networks for solar power forecasting. The dataset used for model evaluation was collected from the Nhi Ha solar power plant. The study also examined 20 scenarios for training set/test set ratio and the lookback parameter to select the best one. The results indicate that the 80/20 training/test set split with a lookback window of 24 gains the best forecasting performance. To demonstrate the effectiveness of the proposed model, its forecasting performance was compared with other common machine learning and deep learning models, including Artificial Neural Networks, Convolutional Neural Networks, and Long Short-Term Memory. Additionally, typical cases, such as high/low power and strong/weak fluctuation days, are considered and evaluated. The proposed model outperforms the other three models with the lowest error values, demonstrating superior accuracy.

XÂY DỰNG MÔ HÌNH LAI CNN-LSTM DỰ BÁO NGẮN HẠN CÔNG SUẤT PHÁT CHO NHÀ MÁY ĐIỆN MẶT TRỜI NHỊ HÀ

Nguyễn Thị Hoài Thu*, Phạm Năng Văn, Trần Quang Khải
Trường Điện - Điện tử, Đại học Bách khoa Hà Nội

THÔNG TIN BÀI BÁO

Ngày nhận bài: 03/12/2024
Ngày hoàn thiện: 03/01/2025
Ngày đăng: 04/01/2025

TỪ KHÓA

Dự báo
Ngắn hạn
Công suất điện mặt trời
Mô hình lai
CNN-LSTM

TÓM TẮT

Việc tích hợp điện mặt trời vào lưới điện không chỉ mang lại lợi ích to lớn về kinh tế và môi trường mà còn đem đến những thách thức đối với hệ thống điện. Dự báo ngắn hạn công suất điện mặt trời một cách chính xác là một trong những giải pháp để khắc phục những khó khăn này. Nghiên cứu này đề xuất một mô hình lai kết hợp mạng nơ-ron tích chập – mạng có bộ nhớ dài-ngắn hạn (CNN-LSTM) để dự báo công suất điện mặt trời. Bộ dữ liệu công suất để đánh giá mô hình được thu thập từ nhà máy điện mặt trời Nhị Hà. Nghiên cứu đã đưa ra các phương án về việc chia tỉ lệ tập huấn luyện/tập kiểm tra cũng như kích thước dữ liệu đầu vào phổ biến để lựa chọn phương án tốt nhất trong đó. Kết quả cho thấy phương án chia tỷ lệ tập huấn luyện/tập dữ liệu là 80/20 với số lookback là 24 đem lại hiệu quả dự báo tốt nhất. Để chứng minh hiệu quả của mô hình đề xuất, kết quả dự báo được so sánh với các mô hình học máy và học sâu bao gồm mạng nơ ron nhân tạo, mạng nơ-ron tích chập, mạng bộ nhớ dài-ngắn hạn. Ngoài ra một số ngày điển hình như công suất cao/thấp, mức độ dao động mạnh/yếu cũng được xem xét và đánh giá. Với giá trị sai số nhỏ nhất, mô hình đề xuất có độ chính xác vượt trội khi so sánh với 3 mô hình còn lại.

DOI: https://doi.org/10.34238/tnu-jst.11648

* Corresponding author. *Email: thu.nguyenthihoai@hust.edu.vn*

1. Đặt vấn đề

Do có đặc tính sạch, không gây ô nhiễm và khả năng tái tạo, năng lượng mặt trời đã trở thành một trong những nguồn tài nguyên có triển vọng nhất và ngày càng thu hút sự chú ý trên toàn thế giới. Tuy nhiên, do tính bất định của bức xạ mặt trời nên việc sản xuất điện thường không ổn định, điều này có thể gây cản trở sự phát triển của hệ thống điện mặt trời (Photovoltaic system – PV). Một trong những phương pháp có thể giảm thiểu đáng kể vấn đề này là đưa ra dự báo ngắn hạn công suất điện mặt trời với độ chính xác và độ tin cậy cao [1].

Các mô hình dự báo công suất điện mặt trời thường được phân loại theo mô hình vật lý, thống kê, học máy, và mô hình lai [2]. Các mô hình vật lý thường dựa trên các mô hình dự báo thời tiết số (NWP) và có thể dự báo với độ chính xác chấp nhận được khi điều kiện thời tiết không thay đổi, nhưng không thể đảm bảo sai số nếu thời tiết thay đổi mạnh. So với mô hình vật lý, các mô hình thống kê dựa trên mối quan hệ toán học của một hoặc nhiều biến đầu ra với các biến khác. Một số mô hình thống kê phổ biến và được sử dụng nhiều nhất là mô hình đường trung bình động tự hồi quy (ARMA) [3], bộ lọc Kalman [4], và các biến thể cải tiến khác của nó (SARIMA, ARIMA, và ARMAX) [5]. Các mô hình này có nhược điểm là không thể dự báo chính xác dữ liệu phi tuyến. Để tăng độ chính xác, các mô hình học máy đang được nghiên cứu và áp dụng trong thời gian gần đây [6]. Các mô hình học máy được sử dụng rộng rãi trong dự báo điện mặt trời bao gồm mạng nơron nhân tạo (ANN) [7], mô hình máy học cực trị (ELM) [6]. Mô hình ANN được ứng dụng nhiều trong dự báo công suất điện mặt trời vì khả năng xử lý dữ liệu nhiễu và khả năng học tập của chúng với các tập dữ liệu lớn. Tuy nhiên, các mạng ANN truyền thống có hiệu quả thấp trong việc tiếp nhận và trích xuất thông tin từ dữ liệu lớn, ảnh hưởng đến tốc độ hội tụ và kết quả thu được không phải là tối ưu toàn cục. Mặt khác, khi số lượng lớp ẩn và kích thước mẫu huấn luyện tăng lên, việc tối ưu hóa các tham số của ANN trở nên khó khăn. Do đó, các nghiên cứu gần đây thường sử dụng mạng nơron học sâu (Deep neural network - DNN). Một số nghiên cứu sử dụng mang nơ-ron tích chập (Convolution Neural Network - CNN) [8], mạng nơ-ron hồi quy (RNN) [9], hoặc mạng long-short term memory (LSTM) [6], và các kết quả nghiên cứu này đã chứng minh tính hiệu quả trong dự báo ngắn hạn công suất điện mặt trời.

Tuy nhiên, các mô hình đơn lẻ dự báo công suất điện mặt trời vẫn không đem lại độ chính xác cao trong một số trường hợp. Vì vậy, việc kết hợp hai hoặc nhiều mô hình là một giải pháp có thể đem lại kết quả dự báo tốt hơn. Điểm mạnh chính của các mô hình lai là kết hợp các cấu trúc mạng khác nhau để tăng cường hiệu quả bằng cách tận dụng ưu điểm của từng cấu trúc [2]. Trong [1], [2], các tác giả đề xuất mô hình lai CNN-LSTM và so sánh với mô hình đơn LSTM để dự báo công suất điện mặt trời. Nghiên cứu có kết luận là sử dụng mô hình đề xuất đem lại hiệu quả dự báo tốt hơn mô hình đơn, dự báo dùng số liệu đầu vào đơn biến cũng tốt hơn với đa biến. Tuy nhiên, các nghiên cứu này chưa xem xét ảnh hưởng của một số yếu tố khi huấn luyện mô hình đến kết quả dự báo, ví dụ như lựa chọn số dữ liệu quá khứ đầu vào (lookback) để đưa ra dự báo trong tương lai, cũng như tỉ lệ tập dữ liệu huấn luyện và kiểm tra, v.v.

Dựa trên các phân tích trên, nghiên cứu này đề xuất xây dựng và sử dụng mô hình lai CNN-LSTM để dự báo công suất điện mặt trời và áp dụng cho số liệu công suất từ nhà máy điện mặt trời Nhị Hà, huyện Thuận Nam, tỉnh Ninh Thuận, với tổng công suất 50 MWp [10]. Dữ liệu công suất điện mặt trời được thu thập từ nhà máy, qua bước xử lý dữ liệu trước khi đưa vào mô hình lai CNN-LSTM. Dữ liệu được chia ra thành các tập huấn luyện/kiểm tra với tỉ lệ khác nhau. Mô hình sau khi huấn luyện được dùng để dự báo công suất và được kiểm tra, đánh giá độ chính xác của dự báo thông qua tập dữ liệu kiểm tra.

Cấu trúc của bài báo như sau: sau phần đặt vấn đề là phần phương pháp luận trình bày các phương pháp xử lý dữ liệu, mô hình liên quan, mô hình đề xuất và các chỉ số đánh giá mô hình dự báo. Phần tiếp theo là kết quả và thảo luận bao gồm các trường hợp tính toán xem xét tỉ lệ dữ liệu tập huấn luyện/kiểm tra, số lượng đầu vào và so sánh kết quả dự báo với các mô hình khác. Phần cuối cùng kết luận về các kết quả thu được, đóng góp của nghiên cứu và đề xuất hướng nghiên cứu tiếp theo.

2. Phương pháp luận

2.1. Tiền xử lý dữ liệu

Các mô hình dự báo, học sâu thường làm việc với kích thước dữ liệu lớn nên nguồn dữ liệu ban đầu cần phải được chuẩn hóa, xử lý để đưa nó về định dạng phù hợp. Giai đoạn tiền xử lý dữ liệu có thể bao gồm: xử lý dữ liệu bị thiếu; xử lý dữ liệu ngoại lai; chuẩn hóa dữ liệu; chuyển đổi dữ liệu; chia dữ liệu; giảm chiều dữ liệu; định dạng lại dữ liệu.

Tiền xử lý dữ liệu nhằm mục đích tăng tính đồng nhất và cải thiện chất lượng của dữ liệu, đảm bảo cho các mô hình dự báo, học máy hoạt động hiệu quả và đạt được độ chính xác cao.

2.1.1. Xử lý dữ liệu khuyết thiếu

Với dữ liệu bị mất trong ngày, có thể sử dụng phương pháp lặp để tìm 2 thời điểm gần nhất với giá trị khuyết thiếu. Sau đó, áp dụng thuật toán Imputation để tính giá trị trung bình của 2 dữ liệu này và điền vào giá trị bị khuyết thiếu.

Trong trường hợp dữ liệu bị mất nhiều ngày liên tiếp, phương pháp được chọn là tính trung bình của các dữ liệu tại cùng thời điểm với giá trị bị thiếu mà khác "0" ở 2 ngày gần nhất để thay vào vị trí dữ liệu khuyết thiếu.

2.1.2. Xử lý dữ liệu ngoại lai

Trong bài báo này, phương pháp "Quy tắc 3-sigma" được sử dụng để xác định giá trị ngoại lai, khi đó các giá trị ngoại lai trong tập dữ liệu sẽ được xác định theo phương trình dưới đây:

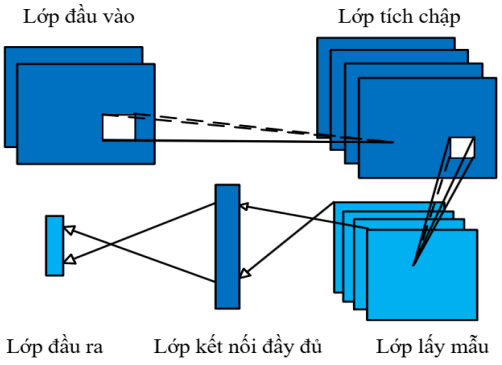
$$\mathrm{Z} = \frac{x-\mu}{\sigma}$$

$$\mathrm{x} = \begin{cases} x_i, & \text{nếu } |Z(x_i)| \le 3 \\ 0, & \text{nếu } |Z(x_i)| > 3 \end{cases} \quad (1)$$

Trong đó, μ là giá trị trung bình, độ lệch chuẩn $\sigma(.)$ cho từng khoảng thời gian.

2.2. Mô hình CNN

Mô hình CNN là một mô hình học sâu, được ứng dụng chủ yếu trong nhận dạng hình ảnh, thị giác máy tính, ngoài ra nó còn được sử dụng để phân tích dữ liệu chuỗi thời gian. Mô hình này được xây dựng từ một chuỗi các lớp tích chập liên tiếp, các nơ-ron trong các lớp được sử dụng các hàm kích hoạt phi tuyến để xử lý dữ liệu [6]. Thuật toán CNN thường bao gồm các lớp sau: lớp đầu vào, lớp tích chập, lớp lấy mẫu, lớp kết nối đầy đủ, và cuối cùng là lớp đầu ra như được mô tả trong Hình 1.



Hình 1. *Cấu trúc của mô hình CNN*

Mỗi lớp tích chập có chứa nhiều bộ lọc tích chập được biểu diễn về mặt toán học thông qua phương trình (2). Lớp tích chập sẽ trích xuất các đặc trưng của dữ liệu, tuy nhiên, các đặc trưng này có kích thước rất lớn. Do đó, lớp lấy mẫu được thêm vào ngay sau lớp tích chập giúp giảm kích thước của các đặc trưng đồng thời giảm chi phí huấn luyện mạng [8].

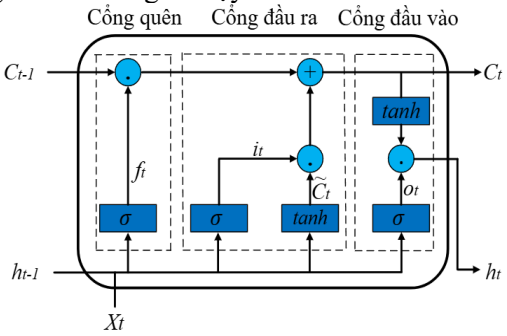
$$l_t = \tanh\left(x_t \times k_t + b_t\right) \quad (2)$$

Trong đó: l_t là giá trị đầu ra sau phép toán tích chập; x_t đại diện vector đầu vào; k_t là trọng số của bộ lọc tích chập; b_t là độ lệch của bộ lọc tích chập; tanh là hàm kích hoạt hyperbolic tangent.

2.3. Mô hình LSTM

Mạng LSTM là một mạng nơ-ron được phát triển từ mạng nơ-ron hồi tiếp RNN [6], [7]. Mô hình này được đề xuất nhằm giải quyết vấn đề gia tăng hoặc mất dần gradient. Ngoài ra mạng LSTM có thể cải thiện khả năng xử lý các phụ thuộc dài hạn trong dữ liệu chuỗi thời gian, điều mà mạng RNN gặp khó khăn khi xử lý, giúp cho mạng trở nên đáng tin cậy hơn.

Trong Hình 2, mô hình LSTM bao gồm một lớp đầu vào, một lớp ẩn, một lớp đầu ra, cổng quên, cổng đầu vào và cổng đầu ra. Quá trình hoạt động của mô hình LSTM được chia thành ba giai đoạn chính. Trong giai đoạn đầu tiên, thông tin quan trọng từ nút trước được lưu trữ, trong khi những thông tin không cần thiết sẽ bị loại bỏ trong giai đoạn quên. Đến giai đoạn trí nhớ chọn lọc, các thông tin quan trọng có thể được ghi nhớ trong các đầu vào. Cuối cùng, giai đoạn đầu ra sẽ xác định những thông tin nào được coi là đầu ra của trạng thái hiện tại, đảm bảo rằng chỉ có những thông tin hữu ích nhất được giữ lại cho bước tiếp theo.



Hình 2. *Cấu trúc của mô hình LSTM*

Mạng LSTM được mô tả thông qua các phương trình sau [6]:

Cổng đầu vào:

$$i_t = \sigma.\left(\mathrm{W}_i.\left[h_{t-1}, x_t\right] + b_i\right) \tag{3}$$

Cổng quên:

$$f_t = \sigma.\left(\mathrm{W}_f.\left[h_{t-1}, x_t\right] + b_f\right) \tag{4}$$

Cổng đầu ra:

$$o_t = \sigma.\left(\mathrm{W}_o.\left[h_{t-1}, x_t\right] + b_o\right) \tag{5}$$

Cổng cập nhật:

$$u_t = \tanh.\left(\mathrm{W}_c.\left[h_{t-1}, x_t\right] + b_c\right) \tag{6}$$

Cổng tế bào:

$$C_t = f_t \times C_{t-1} + i_t \times u_t \tag{7}$$

Trạng thái lớp ẩn:

$$h_t = o_t.\tanh(C_t) \tag{8}$$

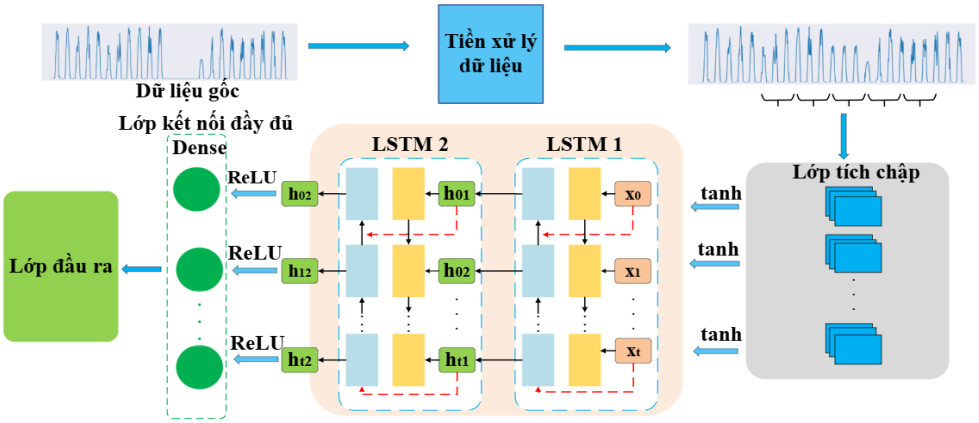
Đầu ra của kết quả dự báo cuối cùng:

$$y_t = \mathrm{W}_y.h_t + b_y \tag{9}$$

Trong đó, các ma trận trọng số của từng lớp là $\mathrm{W_i}$, $\mathrm{W_f}$, $\mathrm{W_o}$ và $\mathrm{W_c}$. Các giá trị f_t, i_t, o_t nằm trong khoảng (0,1); các độ lệch tương ứng của các cổng là b_f, b_i, b_o và độ lệch đầu ra là b_y. Các hàm σ (.) là hàm sigmoid và tanh (.) là hàm hyperbolic tangent. h_{t-1} và h_t lần lượt là các trạng thái của lớp ẩn tại các thời điểm t-1 và t, C_{t-1} và C_t là trạng thái của tế bào tương ứng ở thời điểm $t-1$ và t.

2.4. Mô hình đề xuất

Bài báo này đề xuất mô hình kết hợp CNN và LSTM để dự báo công suất điện mặt trời với sơ đồ cấu trúc thể hiện trên Hình 3. Đầu tiên, dữ liệu thô được qua khối tiền xử lý dữ liệu rồi chuyển qua mạng CNN. CNN sẽ đảm nhiệm việc xử lý liên quan đến chuỗi dữ liệu, bằng cách tự trích xuất các đặc trưng của tập dữ liệu theo thời gian công suất phát nhà máy điện mặt trời và chuyển tiếp qua các lớp LSTM. Còn lớp LSTM sẽ thực hiện dự báo thời gian, giữ lại và truyền dữ liệu trong nhiều thời điểm trong quá trình huấn luyện, giúp mô hình có thể tìm ra mối quan hệ giữa các thời điểm trong chuỗi dữ liệu. Mô hình sau khi huấn luyện được dùng để dự báo công suất điện mặt trời.



Hình 3. *Sơ đồ cấu trúc của CNN-LSTM*

2.5. Các chỉ số đánh giá mô hình

Trong nghiên cứu này, ba chỉ số sai số được sử dụng để đánh giá mô hình đề xuất. Các chỉ số này bao gồm: Sai số trung phương (RMSE), RMSE chuẩn hóa (N-RMSE), và Sai số trung bình tuyệt đối (MAE).

$$RMSE = \sqrt{\frac{1}{n} \times \sum \left(y_{true} - y_{pred} \right)^2} \tag{10}$$

$$N - RMSE = \frac{RMSE}{\max(y_{true}) - \min(y_{true})} \times 100\% \tag{11}$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} \left| y_{true} - y_{pred} \right| \tag{12}$$

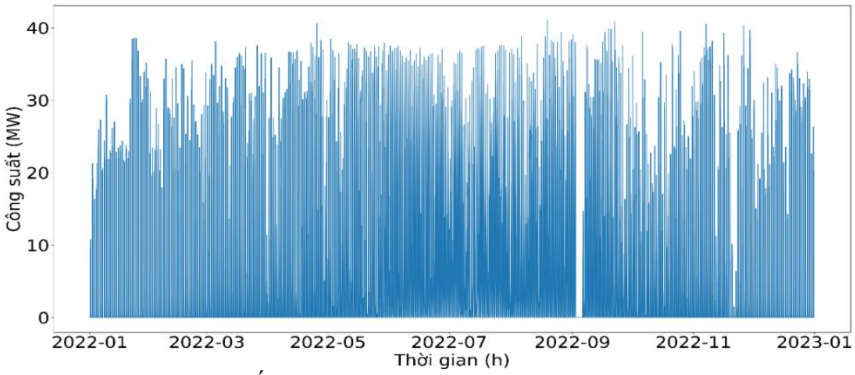
Trong đó: n là số lượng mẫu dữ liệu; y_{true} là giá trị thực tế của mẫu thứ i; y_{pred} là giá trị dự báo của mẫu thứ i.

Trong nghiên cứu này, độ chính xác của mô hình được xem là xuất sắc khi N-RMSE < 10%; tốt nếu 10% < N-RMSE < 20%; trung bình nếu 20% < N-RMSE < 30%; và kém nếu N-RMSE ≥ 30% [11].

3. Kết quả và thảo luận

3.1. Thu thập và xử lý dữ liệu

Nhà máy điện mặt trời Nhị Hà được xây dựng trên diện tích 60 ha tại xã Nhị Hà, huyện Thuận Nam, tỉnh Ninh Thuận, và đã hoàn thành giai đoạn 1 vào năm 2019. Nhà máy được trang bị hơn 150.000 tấm pin năng lượng mặt trời với tổng công suất đạt 50 MWp [10].



Hình 4. *Dữ liệu công suất phát của nhà máy điện mặt trời Nhị Hà trong năm 2022*

Dữ liệu được sử dụng trong nghiên cứu này là dữ liệu công suất phát của nhà máy điện mặt trời Nhị Hà, được thu thập trong khoảng thời gian từ 01/01/2022 đến 31/12/2022. Bước lấy mẫu của bộ dữ liệu là 30 phút, bộ dữ liệu gồm 16984 điểm dữ liệu. Hình 4 cho thấy dữ liệu công suất phát của nhà máy điện mặt trời Nhị Hà trong năm 2022, có một số điểm khuyết thiếu. Để xử lý giá trị khuyết thiếu, chúng tôi sử dụng phương pháp tính giá trị trung bình. Ngoài ra, để loại bớt dữ liệu ngoại lai, nghiên cứu này đã sử dụng phương pháp "Quy tắc 3-sigma" như đã trình bày ở phần trước của bài báo.

Đặc điểm của tập dữ liệu trước và sau khi xử lý được thể hiện trong Bảng 1, với giá trị lớn nhất và nhỏ nhất của tập dữ liệu được giữ nguyên, bổ sung thêm giá trị khuyết thiếu nên tổng số điểm tăng lên, loại trừ các điểm ngoại lai nên giá trị trung bình và phương sai thay đổi nhỏ nhưng không đáng kể.

Bên cạnh đó, khi huấn luyện và kiểm tra mô hình dự báo công suất phát của nhà máy điện mặt trời, các giá trị dữ liệu công suất phát bằng 0 ở ban đêm được loại bỏ, tương ứng khoảng thời gian từ 6h tối đến 6h sáng. Việc loại bỏ các giá trị công suất phát bằng 0 này giúp giảm thiểu nhiễu từ dữ liệu đầu vào vì chúng không cần thiết trong việc huấn luyện dự báo. Ngoài ra, việc này còn giúp giảm bớt lượng dữ liệu cần xử lý, huấn luyện, giúp mô hình dự báo chính xác hơn.

Bảng 1. *Đặc điểm tập dữ liệu trước và sau khi xử lý dữ liệu khuyết thiếu*

Công suất phát	Trước xử lý	Sau xử lý
Count	16984	17520
Mean	8196,09	8010,20
Std	11294,70	11221,94
Min.	0	0
Max.	41039	41039

3.2. Siêu tham số của các mô hình

Trong bài báo này, thông số các mô hình cho các phương án được thể hiện trong Bảng 2. Tất cả các bước xử lý dữ liệu cũng như mô hình được lập trình sử dụng ngôn ngữ Python. Các chương trình được xây dựng và kết quả tính toán được chạy trên máy tính cá nhân AMD Ryzen 7 5700U CPU 1.8 GHz và RAM 8 GB.

Bảng 2. *Cấu hình các lớp của ANN, CNN, LSTM*

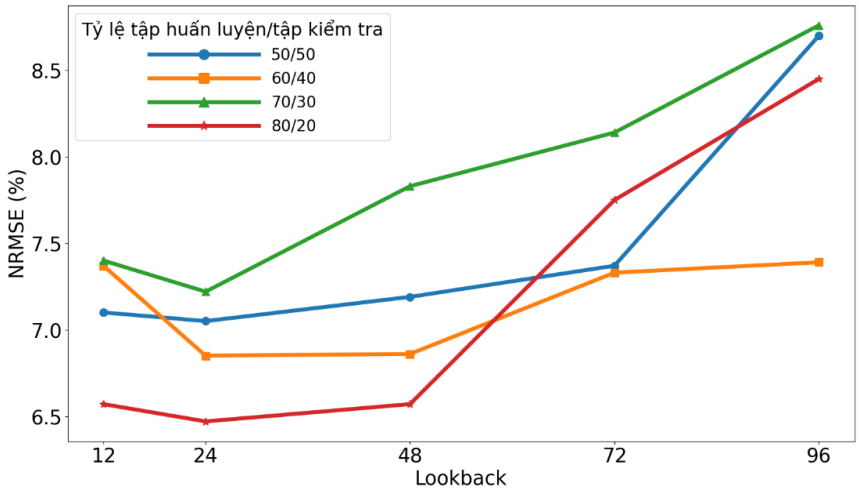
Siêu tham số của mô hình	Mô hình			
	ANN	**CNN**	**LSTM**	**CNN-LSTM**
Conv1D	-	128	-	54
MaxPooling	-	2	-	-
LSTM1	-	-	56	112
LSTM2	-	-	28	56
Dropout	0,2	0,2	0,1	0,2
Dense1	64	64	1	-
Dense2	32	-	-	-
Hàm kích hoạt	ReLU	ReLU	ReLU	ReLU
Tối ưu	Adam	Adam	Adam	Adam

3.3. Ảnh hưởng của kích thước dữ liệu đầu vào và tỉ lệ chia tập dữ liệu huấn luyện/kiểm tra đến kết quả dự báo

Trong các nghiên cứu [12] – [15], tỷ lệ tập huấn luyện/tập kiểm tra thường được sử dụng là 70/30 và 80/20. Để đánh giá ảnh hưởng của tỉ lệ này đến sai số dự báo, các tỷ lệ 50/50, 60/40, 70/30, 80/20 được xem xét để tìm ra tỷ lệ phù hợp nhất với dữ liệu đã có. Tương ứng với mỗi tỷ lệ, xét ảnh hưởng của kích thước dữ liệu đầu vào lookback là 12, 24, 48, 72 và 96 điểm. Kết quả sai số dự báo của mô hình đề xuất trong các phương án về tỷ lệ tập dữ liệu/tập kiểm tra và lookback được thể hiện trong Bảng 3.

Bảng 3. *Sai số dự báo của mô hình lai CNN-LSTM với các phương án tỉ lệ tập huấn luyện/kiểm tra và lookback khác nhau*

Tỉ lệ tập huấn luyện/kiểm tra	Lookback	MAE (kW)	RMSE(kW)	N-RMSE (%)
50/50	12	1703,73	2913,15	7,10
	24	1434,23	2892,85	7,05
	48	1656,84	2952,07	7,19
	72	1625,17	3022,79	7,37
	96	2043,74	3571,98	8,70
60/40	12	1582,65	3023,89	7,37
	24	1560,89	2813,04	6,85
	48	1464,55	2813,38	6,86
	72	1513,40	3006,63	7,33
	96	1613,15	3032,22	7,39
70/30	12	1608,20	3022,04	7,40
	24	1537,43	2947,55	7,22
	48	1547,40	3196,90	7,83
	72	1542,69	3323,82	8,14
	96	1578,60	3580,72	8,76
80/20	12	1372,85	2662,28	6,57
	24	1321,23	**2622,65**	**6,47**
	48	**1304,97**	2660,75	6,57
	72	1510,11	3140,24	7,75
	96	1840,29	3426,00	8,45



Hình 5. *Kết quả sai số NRMSE của mô hình CNN-LSTM trong các phương án đề xuất*

Kết quả sai số trong Bảng 3 cho thấy tỷ lệ tập huấn luyện/ kiểm tra là 80/20 với số lookback là 24 có kết quả sai số RMSE và N-RMSE nhỏ nhất trong khi sai số MAE chỉ cao hơn không đáng kể so với việc dùng lookback là 48. Kết quả sai số N-RMSE của mô hình CNN-LSTM trong các phương án đề xuất được thể hiện trong Hình 5. Các nhận xét từ Bảng 3 và Hình 5 như sau:

- Với cùng tỉ lệ tập huấn luyện/kiểm tra, sai số nhỏ nhất là khi lookback = 24, có xu hướng tăng lên khi số lookback tăng cao hơn 24 hoặc giảm nhỏ hơn 24. Điều này có thể giải thích là do đầu ra chỉ dựa trên số lượng đầu vào quá nhỏ hoặc quá lớn thì có thể sẽ không chính xác. Khi số lookback quá lớn (> 48) thì việc huấn luyện sẽ khó khăn hơn do đầu ra phụ thuộc vào quá nhiều đầu vào, mối quan hệ cần tính toán tối ưu sẽ phức tạp hơn rất nhiều, do đó độ chính xác sẽ giảm. Do đó việc chọn số lượng lookback là cần thiết.

- Với cùng số lookback, khi lookback ≤ 48, tỉ lệ huấn luyện/kiểm tra là 80/20 đem lại kết quả tốt nhất. Điều này có thể giải thích do dữ liệu huấn luyện nhiều thì mô hình học được nhiều thông

tin hơn và dự báo chính xác hơn trên đặc điểm dữ liệu này. Khi số lookback lớn thì như trên đã giải thích, sai số sẽ tăng lên và không thu được quy luật sai số theo tỉ lệ tập huấn luyện/kiểm tra như khi lookback nhỏ.

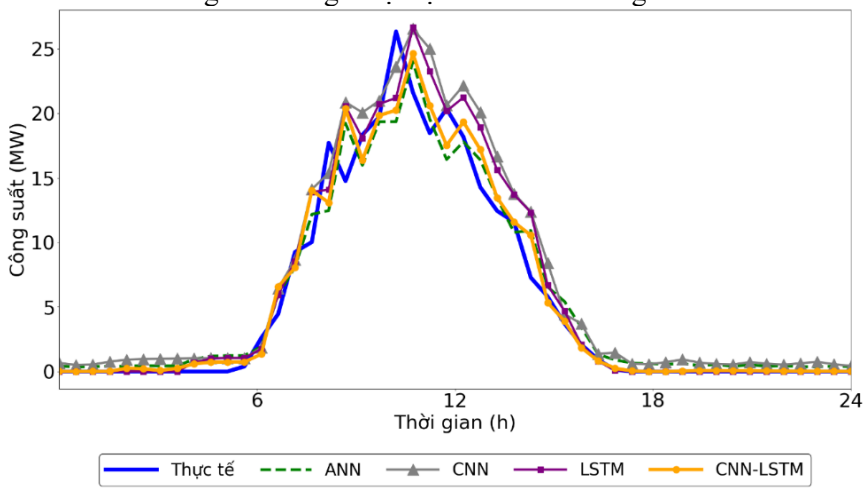
3.4. So sánh mô hình đề xuất với các mô hình khác

Với kết quả sai số nhỏ nhất khi tỉ lệ dữ liệu huấn luyện/kiểm tra là 80/20 và số lookback là 24, mô hình đề xuất được so sánh với các mô hình học máy/học sâu phổ biến trong các nghiên cứu gần đây là ANN, CNN, và LSTM. Kết quả sai số dự báo của các mô hình được thể hiện trong Bảng 4. Từ Bảng 4, có thể thấy các mô hình đề xuất và các mô hình dùng để so sánh đều cho ra kết quả dự báo rất tốt với chỉ số sai số N-RMSE <10%. Trong đó, mô hình kết hợp CNN-LSTM có khả năng dự báo tốt nhất khi so sánh với các mô hình khác với các chỉ số sai số RMSE, N-RMSE, MAE thấp nhất, lần lượt là 2622,65 kW, 6,47%, 1321,23 kW.

Bảng 4. *Kết quả so sánh sai số của các mô hình*

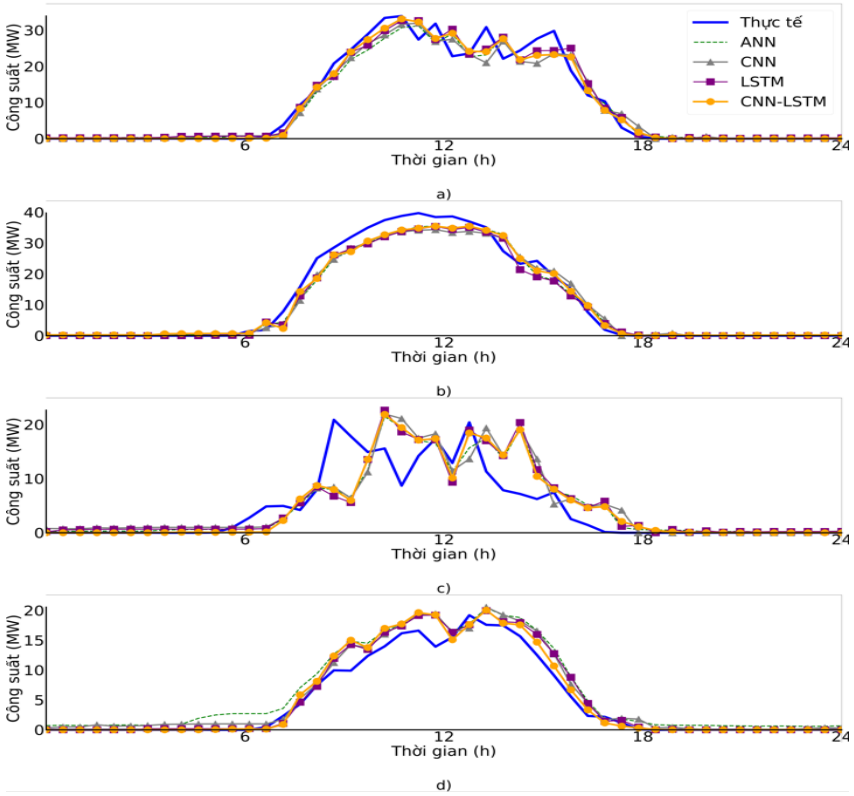
	ANN	CNN	LSTM	Mô hình đề xuất
MAE (kW)	1512,90	1542,04	1478,55	1321,23
RMSE (kW)	2698,40	2697,35	2666,37	2622,65
NRMSE (%)	6,66	6,67	6,58	6,47

Hình 6 mô tả kết quả dự báo trong 1 ngày bất kỳ sử dụng mô hình đề xuất và các mô hình so sánh trong phương án tỷ lệ tập huấn luyện/kiểm tra và lookback đã xác định ở trên. Mô hình đề xuất CNN-LSTM có khả năng bám sát giá trị thực tế tốt nhất trong các mô hình.



Hình 6. *Kết quả dự báo với các mô hình khác nhau của phương án tỷ lệ huấn luyện/ kiểm tra 80/20, lookback 24*

Việc dự báo trên toàn tập kiểm tra cho thấy hiệu quả chung của các mô hình. Tuy nhiên, để quan sát kỹ tính hiệu quả của mô hình đề xuất khi công suất thay đổi mạnh hoặc yếu, nghiên cứu có xét một số ngày điển hình là: công suất đỉnh lớn, dao động nhẹ; công suất đỉnh lớn, dao động mạnh; công suất đỉnh nhỏ, dao động nhẹ và công suất đỉnh nhỏ, dao động mạnh. Kết quả so sánh các mô hình trong một số ngày điển hình được biểu diễn trong Hình 7 và sai số của các ngày này được thể hiện trong Bảng 5. Có thể thấy các ngày có công suất đỉnh lớn dao động nhẹ có hiệu quả dự báo cao trong khi ngày có công suất đỉnh nhỏ dao động mạnh thì sai số khá lớn, cần cải thiện mô hình hơn nữa để đạt kết quả dự báo tốt hơn. Độ chính xác dự báo của mô hình CNN-LSTM trong các ngày này vẫn vượt trội hơn so với các mô hình đơn khi dự báo cho các ngày điển hình với các chỉ số sai số N-RMSE, MAE, RMSE nhỏ nhất.



Hình 7. *Kết quả dự báo cho một số ngày điển hình: (a) Công suất đỉnh lớn dao động mạnh, (b) công suất đỉnh lớn dao động nhẹ, (c) công suất đỉnh nhỏ dao động mạnh, (d) công suất đỉnh nhỏ dao động nhẹ*

Bảng 5. *Kết quả sai số của các ngày điển hình trong phương án 3*

			Ngày điển hình			
		Toàn tập test	**Đỉnh lớn dao động nhẹ**	**Đỉnh lớn dao động mạnh**	**Đỉnh nhỏ dao động nhẹ**	**Đỉnh nhỏ dao động mạnh**
ANN	**RMSE (kW)**	2698,40	2583,62	2792,11	2111,45	4124,99
	N-RMSE (%)	6,66	6,51	8,23	11,01	19,76
	MAE (kW)	1512,90	1881,87	1769,34	1706,90	2409,21
CNN	**RMSE (kW)**	2697,35	2545,39	2906,32	1720,01	4481,91
	N-RMSE (%)	6,67	6,41	8,57	8,96	21,47
	MAE (kW)	1542,04	1674,91	1811,13	1207,49	2741,01
LSTM	**RMSE (kW)**	2666,37	2508,55	2622,54	1600,98	4317,40
	N-RMSE (%)	6,58	6,32	7,73	8,34	20,68
	MAE (kW)	1478,55	1574,73	1661,51	915,28	2488,39
Mô hình đề xuất	**RMSE (kW)**	2622,65	2372,21	2441,68	1521,26	4120,92
	N-RMSE (%)	6,47	5,97	7,20	7,93	19,74
	MAE (kW)	1321,23	1537,76	1393,26	844,20	2279,99

4. Kết luận

Bài báo này đề xuất một mô hình lai CNN-LSTM dự báo ngắn hạn công suất điện mặt trời và áp dụng cho dữ liệu công suất phát của nhà máy điện mặt trời Nhị Hà. Nghiên cứu đã xem xét 20 phương án khác nhau về tỉ lệ tập huấn luyện/tập kiểm tra và số lượng dữ liệu đầu vào lookback. Từ đó, chọn được mô hình CNN-LSTM với tỉ lệ dữ liệu huấn luyện/kiểm tra là 80/20 và lookback là 24 đem lại hiệu quả dự báo cao nhất với các sai số RMSE, N-RMSE, MAE nhỏ nhất lần lượt là 2622,65 kW, 6,47%, 1321,23 kW. Kết quả dự báo được so sánh với ba mô hình ANN,

CNN, LSTM dựa trên ba chỉ số đánh giá RMSE, N-RMSE, MAE. Ngoài ra, nghiên cứu cũng xét thêm tính hiệu quả của các phương án và mô hình trong bốn ngày điển hình. Kết quả thu được là mô hình CNN-LSTM có khả năng dự báo chính xác vượt trội hơn các mô hình dự báo đơn trong toàn tập kiểm tra cũng như trong các ngày điển hình với sai số N-RMSE nhỏ nhất lần lượt tương ứng cho các ngày là 5,97%, 7,20%, 7,93%, 19,74%. Điều này cho thấy mô hình đề xuất có thể áp dụng để dự báo công suất điện mặt trời với độ chính xác cao. Tuy nhiên, sai số trong các ngày có dao động mạnh vẫn còn cao, do đó hướng phát triển trong tương lai là cải thiện mô hình, kết hợp thêm các phương pháp phân tách dữ liệu để nâng cao độ chính xác.

REFERENCES

[1] T. H. T. Nguyen, N. V. Pham, Q. B. Phan, V. N. N. Nguyen, H. M. Pham, and N. Q. Tran, "Short-term Forecasting of Solar Radiation Using a Hybrid Model of CNN-LSTM Integrated with EEMD," in *2022 6th International Conference on Green Technology and Sustainable Development (GTSD)*, Jul. 2022, pp. 854-859, doi: 10.1109/GTSD54989.2022.9988761.

[2] A. Agga, A. Abbou, M. Labbadi, Y. E. Houm, and I. H. Ou Ali, "CNN-LSTM: An efficient hybrid deep learning architecture for predicting short-term photovoltaic power production," *Electric Power Systems Research*, vol. 208, Jul. 2022, Art. no. 107908, doi: 10.1016/j.epsr.2022.107908.

[3] E. Erdem and J. Shi, "ARMA based approaches for forecasting the tuple of wind speed and direction," *Applied Energy*, vol. 88, no. 4, pp. 1405-1414, Apr. 2011, doi: 10.1016/j.apenergy.2010.10.031.

[4] M. Poncela, P. Poncela, and J. R. Perán, "Automatic tuning of Kalman filters by maximum likelihood methods for wind energy forecasting," *Applied Energy*, vol. 108, pp. 349-362, Aug. 2013, doi: 10.1016/j.apenergy.2013.03.041.

[5] Y. Li, Y. Su, and L. Shu, "An ARMAX model for forecasting the power output of a grid connected photovoltaic system," *Renewable Energy*, vol. 66, pp. 78-89, Jun. 2014, doi: 10.1016/j.renene.2013.11.067.

[6] T. H. T. Nguyen, N. V. Pham, V. K. Ngo, and X. B. Do, "A Comparative Study of Machine Learning–based Models for Short-Term Multi-step Forecasting of Solar Power: An Application for Nhi Ha Solar Farm," *Measurement, Control, and Automation*, vol. 5, no. 1, Apr. 2024, Art. no. 1.

[7] M. Q. Raza, M. Nadarajah, and C. Ekanayake, "On recent advances in PV output power forecast," *Solar Energy*, vol. 136, pp. 125-144, Oct. 2016, doi: 10.1016/j.solener.2016.06.073.

[8] D. T. Nguyen, D. H. Do, G. Fujita, and T. S. Tran, "Multi 2D-CNN-based model for short-term PV power forecast embedded with Laplacian Attention," *Energy Reports*, vol. 12, pp. 2086-2096, Dec. 2024, doi: 10.1016/j.egyr.2024.08.020.

[9] H. Shi, M. Xu, and R. Li, "Deep Learning for Household Load Forecasting—A Novel Pooling Deep RNN," *IEEE Transactions on Smart Grid*, vol. 9, no. 5, pp. 5271-5280, Sep. 2018, doi: 10.1109/TSG.2017.2686012.

[10] Bitexco Power, "Nhi Ha," 2023. [Online]. Available: https://bitexcopower.com.vn/projects/nhi-ha/. [Accessed Dec. 30, 2024].

[11] M.-F. Li, X.-P. Tang, W. Wu, and H.-B. Liu, "General models for estimating daily global solar radiation for different solar radiation zones in mainland China," *Energy Conversion and Management*, vol. 70, pp. 139-148, Jun. 2013, doi: 10.1016/j.enconman.2013.03.004.

[12] C. Liu, W. Li, C. Hu, T. Xie, Y. Jiang, R. Li, S. Soomro, and Y. Xu, "Research on runoff process vectorization and integration of deep learning algorithms for flood forecasting," *Journal of Environmental Management*, vol. 362, Jun. 2024, Art. no. 121260, doi: 10.1016/j.jenvman.2024.121260.

[13] V. R. Joseph, "Optimal ratio for data splitting," *Statistical Analysis*, vol. 15, no. 4, pp. 531-538, Aug. 2022, doi: 10.1002/sam.11583.

[14] Z. Wang, Y. Wei, and S. Wang, "Forecasting the carbon price of China's national carbon market: A novel dynamic interval-valued framework," *Energy Economics*, vol. 141, Jan. 2025, Art. no. 108107, doi: 10.1016/j.eneco.2024.108107.

[15] Y. Liang, D. Zhang, J. Zhang, and G. Hu, "A state-of-the-art analysis on decomposition method for short-term wind speed forecasting using LSTM and a novel hybrid deep learning model," *Energy*, vol. 313, Dec. 2024, Art. no. 133826, doi: 10.1016/j.energy.2024.133826.