

**BỘ CÔNG THƯƠNG
TRƯỜNG ĐẠI HỌC ĐIỆN LỰC**



NGUYỄN TUẤN ANH

LUẬN ÁN TIẾN SĨ

**NGHIÊN CỨU MÔ HÌNH HỌC MÁY HỖN HỢP DỰ
BÁO CÔNG SUẤT PHÁT ĐIỆN MẶT TRỜI TRONG
ĐIỀU KIỆN DỮ LIỆU THỰC TẾ TẠI VIỆT NAM**

Chuyên ngành	: Kỹ thuật năng lượng
Mã số chuyên ngành	: Thí điểm
Mã NCS	: 20TS0000001
Khóa học	: 2020 - 2023
Người hướng dẫn khoa học 1	: TS. Phạm Mạnh Hải
Người hướng dẫn khoa học 2	: TS. Vũ Minh Pháp

HÀ NỘI, 2026

BỘ CÔNG THƯƠNG
TRƯỜNG ĐẠI HỌC ĐIỆN LỰC

NGUYỄN TUẤN ANH

LUẬN ÁN TIẾN SĨ

NGHIÊN CỨU MÔ HÌNH HỌC MÁY HỖN HỢP DỰ
BÁO CÔNG SUẤT PHÁT ĐIỆN MẶT TRỜI TRONG
ĐIỀU KIỆN DỮ LIỆU THỰC TẾ TẠI VIỆT NAM

XÁC NHẬN CỦA CBHD VÀ CHỦ TỊCH HỘI ĐỒNG

1. XÁC NHẬN CỦA CBHD:

2. XÁC NHẬN CỦA CHỦ TỊCH HỘI ĐỒNG:

HÀ NỘI, 2026

LỜI CAM ĐOAN

Nghiên cứu sinh xin cam đoan đây là công trình nghiên cứu của riêng nghiên cứu sinh. Các thông tin trích dẫn trong luận án đã được liệt kê đầy đủ và được chỉ rõ nguồn gốc. Các số liệu, kết quả nêu trong luận án là trung thực và chưa từng được công bố trong bất kỳ công trình khoa học nào khác.

Hà Nội, ngày ... tháng ... năm 2026

Nghiên cứu sinh

Nguyễn Tuấn Anh

LỜI CẢM ƠN

Để hoàn thành luận án tiến sĩ, cùng với những nỗ lực của bản thân, nghiên cứu sinh đã nhận được sự động viên, giúp đỡ của rất nhiều cá nhân và tập thể. Nghiên cứu sinh xin trân trọng cảm ơn Trường Đại học Điện lực nơi nghiên cứu sinh thực hiện đề tài đã tạo điều kiện thuận lợi về cơ sở vật chất, học thuật và môi trường nghiên cứu nghiêm túc.

Nghiên cứu sinh cũng trân trọng gửi lời cảm ơn đến hai thầy hướng dẫn là Tiến sĩ Phạm Mạnh Hải (Trường Đại học Điện lực) và Tiến sĩ Vũ Minh Pháp (Viện Khoa học Công nghệ Năng lượng và Môi trường, Viện Hàn lâm Khoa học và Công nghệ Việt Nam) đã tận tình chỉ bảo, định hướng và hỗ trợ nghiên cứu sinh trong suốt quá trình học tập, nghiên cứu và hoàn thành luận án.

Nghiên cứu sinh cũng xin gửi lời cảm ơn chân thành tới các thầy cô khoa Kỹ thuật điện và khoa Năng lượng mới đã luôn tận tình động viên và hỗ trợ Nghiên cứu sinh trong quá trình triển khai đề tài nghiên cứu.

Nghiên cứu sinh xin trân trọng cảm ơn các nhà khoa học, các đồng nghiệp đã phản biện, đóng góp các ý kiến quý báu để xây dựng và trao đổi các vấn đề lý thuyết cũng như thực tiễn giúp hoàn thiện luận án.

Đặc biệt, nghiên cứu sinh xin được bày tỏ lời cảm ơn sâu sắc nhất đến gia đình đã luôn chia sẻ những khó khăn, luôn hỗ trợ, động viên tinh thần giúp nghiên cứu sinh có được chỗ dựa vững chắc để hoàn thành luận án này.

Hà Nội, ngày ... tháng ... năm 2026

Nghiên cứu sinh

Nguyễn Tuấn Anh

MỤC LỤC

TRANG BÌA PHỤ	i
LỜI CAM ĐOAN	i
LỜI CẢM ƠN	ii
MỤC LỤC	iii
DANH MỤC CÁC BẢNG BIỂU.....	vii
DANH MỤC HÌNH VẼ.....	ix
DANH MỤC CÁC TỪ VIẾT TẮT	xi
MỞ ĐẦU.....	1
CHƯƠNG 1. TỔNG QUAN NGHIÊN CỨU	5
1.1. Tiềm năng phát triển điện mặt trời tại Việt Nam	6
1.2. Nhu cầu và hiện trạng hệ thống dự báo công suất phát điện mặt trời tại Việt Nam	10
1.2.1. Nhu cầu và hiện trạng hệ thống dự báo công suất phát của các nhà máy điện mặt trời tại Việt Nam.....	11
1.2.2. Các khó khăn, thách thức trong dự báo công suất phát điện mặt trời tại Việt Nam	12
1.2.3. Các khung thời gian dự báo công suất phát điện mặt trời.....	13
1.3. Tổng quan các mô hình dự báo công suất phát điện mặt trời.....	14
1.3.1. Các sai số để đánh giá mô hình dự báo	14
1.3.2. Tổng quan các nghiên cứu về dự báo công suất phát điện mặt trời	15
1.4. Khoảng trống trong nghiên cứu.....	18
1.5. Kết luận chương 1	20
CHƯƠNG 2. CÁC MÔ HÌNH DỰ BÁO BỨC XẠ MẶT TRỜI VÀ CÔNG SUẤT PHÁT ĐIỆN MẶT TRỜI TRONG NGẮN HẠN	22
2.1. Bộ dữ liệu thực tế và tiền xử lý dữ liệu	22
2.1.1. Giới thiệu và phân tích bộ dữ liệu thực tế	22
2.1.2. Phương pháp tiền xử lý dữ liệu	23
2.1.3. Các bộ dữ liệu bổ sung	25
2.2. Các mô hình dự báo bức xạ mặt trời	26
2.2.1. Cơ sở lý thuyết và tầm quan trọng của dự báo bức xạ mặt trời	26
2.2.2. Các mô hình dự báo bức xạ mặt trời	28

2.2.3. Nguồn dữ liệu và hệ thống thu thập dữ liệu	29
2.3. Các mô hình dự báo công suất phát điện mặt trời	30
2.3.1. Các mô hình được nghiên cứu tại nước ngoài.....	30
2.3.2. Các mô hình được nghiên cứu trong nước	51
2.3.3. Đánh giá các mô hình dự báo công suất phát điện mặt trời	52
2.4. Kết luận chương 2	54
CHƯƠNG 3. PHÂN TÍCH ĐÁNH GIÁ CÁC MÔ HÌNH DỰ BÁO BỨC XẠ VÀ CÔNG SUẤT PHÁT ĐIỆN MẶT TRỜI.....	55
3.1. Xây dựng, huấn luyện và so sánh các mô hình dự báo bức xạ mặt trời.	55
3.1.1. Xây dựng và huấn luyện các mô hình dự báo	56
3.1.2. Đánh giá hiệu năng dự báo bức xạ mặt trời của các mô hình	59
3.1.3. Kết luận.....	62
3.2. Xây dựng, huấn luyện và so sánh đánh giá các mô hình dự báo công suất phát điện mặt trời đơn lẻ.....	62
3.2.1. Xây dựng các mô hình cây quyết định	63
3.2.2. Xây dựng các mô hình chuỗi thời gian.....	69
3.2.3. So sánh kết quả huấn luyện và dự báo của các mô hình	70
3.3. Kết hợp giữa mô hình cây quyết định và mô hình chuỗi thời gian để cải thiện hiệu năng dự báo	75
3.3.1. Phân tích ưu, nhược điểm của các mô hình trong điều kiện dữ liệu không liên tục	76
3.3.2. Ứng dụng LightGBM để tái cấu trúc chuỗi đầu vào LSTM trong điều kiện dữ liệu không hoàn chỉnh.....	77
3.3.3. Kết luận.....	86
3.4. Kết luận chương 3	87
CHƯƠNG 4. ĐỀ XUẤT MÔ HÌNH DỰ BÁO CHO MỘT SỐ NHÀ MÁY ĐIỆN MẶT TRỜI TẠI VIỆT NAM.....	90
4.1. Dự báo công suất phát cho một nhà máy	91
4.1.1. Xây dựng mô hình dự báo SE-XGB-LGBM-RF	93

4.1.2. Xây dựng mô hình SE-XGB-LGBM-BiGRU, SE-XGB-LGBM-GRU, SE-XGB-LGBM-LSTM.....	93
4.1.3. Kết quả so sánh các mô hình SE-XGB-LGBM-RF, SE-XGB-LGBM-BiGRU, SE-XGB-LGBM-GRU, SE-XGB-LGBM-LSTM.....	94
4.1.4. Xây dựng mô hình dự báo SE-XGB-LGBM-RF với bộ trọng số tối ưu (OW) ..	98
4.2. Xây dựng mô hình dự báo công suất phát điện mặt trời cho nhiều nhà máy	111
4.2.1. Triển khai mô hình Selector-Model cho XGB, LGBM, và RF	112
4.2.2. Triển khai huấn luyện mô hình SE-XGB-LGBM-RF trên từng tập dữ liệu nhà máy riêng biệt	115
4.2.3. Kết quả thực nghiệm.....	116
4.2.4. Đánh giá kết quả các mô hình dự báo	118
4.3. Đánh giá chi phí tính toán và hiệu quả sử dụng tài nguyên hệ thống	120
4.3.1. Đánh giá chi phí tính toán của SE-XGB-LGBM-RF-OW và các mô hình so sánh	121
4.3.2. Đánh giá chi phí tính toán của mô hình Selector-Model.....	122
4.4. Kết luận chương 4	124
KẾT LUẬN VÀ KHUYẾN NGHỊ	116
DANH MỤC CÁC CÔNG TRÌNH ĐÃ CÔNG BỐ	119
DANH MỤC CÁC TÀI LIỆU THAM KHẢO	120
PHỤ LỤC	128

DANH MỤC BẢNG BIỂU

Bảng 1.1. Tổng số giờ nắng tại các vùng khí hậu	6
Bảng 1.2. Các khung thời gian dự báo công suất phát điện mặt trời	13
Bảng 1.3. Các sai số được áp dụng để đánh giá hiệu suất mô hình dự báo	14
Bảng 2.1: Bảng thống kê các đặc trưng của tập dữ liệu đầu vào	23
Bảng 2.2: Các đặc trưng của ba bộ dữ liệu bổ sung tại nhà máy Quảng Trị, Đắc-Lắc và Thanh Hóa	25
Bảng 2.3. Ưu, nhược điểm các mô hình dự báo công suất phát điện mặt trời	52
Bảng 3.1: Thiết lập tham số cấu hình cho các mô hình nghiên cứu.....	57
Bảng 3.2. Tổng hợp các chỉ số đánh giá hiệu năng và sai số thực nghiệm của ba mô hình dự báo trên các kịch bản thực nghiệm.	59
Bảng 3.3. So sánh đặc tính kỹ thuật và lý do lựa chọn các thuật toán thành phần	64
Bảng 3.2. Các tham số mặc định và tối ưu của mô hình XGBoost.....	66
Bảng 3.3. So sánh giữa tham số mặc định và tham số tối ưu của mô hình XGBoost...	66
Bảng 3.4. Các tham số mặc định và tối ưu của mô hình LightGBM	67
Bảng 3.5. So sánh giữa tham số mặc định và tham số tối ưu của mô hình LightGBM	67
Bảng 3.6. Các tham số mặc định và tối ưu của mô hình Random Forest	68
Bảng 3.7. So sánh giữa tham số mặc định và tham số tối ưu của mô hình Random Forest	68
Bảng 3.10. So sánh các kết quả huấn luyện của từng mô hình với tập dữ liệu lịch sử từ 01/01/2022 đến 31/12/2022 (70% tập dữ liệu được dùng cho đào tạo và 30% cho test)	70
Bảng 3.11. So sánh sai số dự báo công suất phát điện mặt trời giữa các mô hình đơn lẻ trong giai đoạn 01-03/03/2021 trong hai trường hợp sử dụng bức xạ dự báo từ mô hình LightGBM tại mục 3.1 và sử dụng bức xạ thực tế.	72
Bảng 3.10. So sánh sai số dự báo (RMSE, NRMSE, MAPE, NMAPE) giữa mô hình đơn và LightGBM-LSTM theo tỷ lệ thiếu dữ liệu công suất (0–100%) trong các ngày 01–03/03/2021, dưới hai kịch bản bức xạ: thực tế và dự báo (LightGBM).....	79
Bảng 4.1. Kết quả so sánh dự báo công suất phát điện mặt trời của các mô hình SE-XGB-LGBM-RF, SE-XGB-LGBM-BiGRU, SE-XGB-LGBM-GRU, SE-XGB-LGBM-LSTM trên tập dữ liệu dự báo	94
Bảng 4.2. Các ngày diễn hình trong một mùa	100

Bảng 4.3. So sánh các kết quả huấn luyện của từng mô hình với tập dữ liệu lịch sử đã được chia theo mùa trong năm	101
Bảng 4.4. Kết quả dự báo của mô hình kết hợp SE-XGB-LiGBM-RF-OW với trọng số mặc định bằng nhau trong ngày diễn hình của mùa theo Bảng 4.2.....	103
Bảng 4.5. So sánh kết quả dự báo của mô hình SE-XGB-LiGBM-RF-OW với các mô hình XGBoost, LightGBM và Random Forest.....	103
Bảng 4.6. Trọng số tối ưu cho từng mô hình.....	105
Bảng 4.7. Kết quả dự báo trong các ngày diễn hình của một mùa theo Bảng 4.2 với trọng số tối ưu cho mô hình SE-XGB-LiGBM-RF-OW	106
Bảng 4.8. Kết quả của mô hình dự báo SE-XGB-LiGBM-RF-OW với thông số tối ưu khi bỏ đi mô hình Xgboost	109
Bảng 4.9. Kết quả dự báo từ mô hình Selector-Model và mô hình SE-XGB-LGBM-RF (khi được huấn luyện riêng biệt) cho từng nhà máy.....	116
Bảng 4.10. Tổng hợp hiệu quả dự báo và khả năng thích nghi của mô hình Selector-Model và mô hình SE-XGB-LGBM-RF	118
Bảng 4.11. Cấu hình phần cứng và phần mềm sử dụng trong nghiên cứu.....	120
Bảng 4.12. Tổng hợp hiệu quả chi phí tính toán của mô hình SE-XGB-LGBM-RF-OW và các mô hình đơn lẻ.....	121
Bảng 4.13. Chi phí tính toán của mô hình Selector-Model.....	122

DANH MỤC HÌNH VẼ

Hình 1.1. Phân bố bức xạ mặt trời tại Việt Nam	7
Hình 1.2. Tăng trưởng công suất điện mặt trời tại Việt Nam (2017-2022).....	8
Hình 1.3. Hệ thống dự báo công suất của nhà máy điện mặt trời	11
Hình 2.1. Các mô hình dự báo công suất phát điện mặt trời	30
Hình 2.2. Cấu trúc cơ bản của ANN.....	34
Hình 2.3. Mạng nơron nhiều lớp (MLPNN)	34
Hình 2.4: Cấu trúc của mô hình cây quyết định	35
Hình 2.5. Cấu trúc cơ bản của RNN.....	42
Hình 2.6. Cấu trúc của mô hình LSTM	43
Hình 2.7. Cấu trúc của mô hình GRU	45
Hình 2.8. Cấu trúc của mô hình BiGRU một lớp ẩn	46
Hình 2.9. Cấu trúc mô hình BiGRU hai lớp ẩn	47
Hình 2.10. Cấu trúc mô hình Stacking Ensemble	50
Hình 3.1. Lưu đồ quy trình thực nghiệm và đánh giá hiệu năng các thuật toán dự báo bức xạ.....	58
Hình 3.2. Đồ thị minh họa giá trị dự báo của ba mô hình LightGBM, LSTM, GRU so với giá trị bức xạ thực tế ngày 15/02/2021.....	61
Hình 3.3. Đồ thị minh họa giá trị dự báo của ba mô hình LightGBM, LSTM, GRU so với giá trị bức xạ thực tế ngày 01-03/03/2021.	61
Hình 3.4. Sơ đồ thuật toán thực hiện so sánh các mô hình dự báo các mô hình cây quyết định và các mô hình chuỗi thời gian.....	63
Hình 3.5. Biểu đồ so sánh hiệu suất dự báo công suất phát theo từng bước thời gian giữa các mô hình cây quyết định và chuỗi thời gian trong điều kiện thiếu dữ liệu quá khứ, dự báo theo bức xạ thực tế và bức xạ dự báo	74
Hình 3.6. Sơ đồ thuật toán kết hợp đầu vào của mô hình LightGBM-LSTM.....	77
Hình 3.7. So sánh dự báo công suất phát của mô hình LightGBM-LSTM và LSTM trong điều kiện dữ liệu chuỗi quá khứ đầy đủ và chuỗi dữ liệu quá khứ bị thiếu theo tỷ lệ 50% (a) và 100% (b).	83
Hình 3.8. So sánh các sai số RMSE, NRMSE, MAPE, NMAPE của các mô hình dự báo trong hai kịch bản sử dụng bức xạ thực tế và bức xạ dự báo.	84

Hình 4.1. Sơ đồ tích hợp các mô hình dự báo công suất phát điện mặt trời sử dụng Stacking Ensemble	92
Hình 4.2. So sánh công suất phát điện thực tế và dự báo của bốn mô hình tổ hợp a. SE-XGB-LGBM-GRU, b. SE-XGB-LGBM-LSTM, c. SE-XGB-LGBM-RF và d. SE-XGB-LGBM-BiGRU trong tập dữ liệu (01-03/03/2021)	97
Hình 4.3. Sơ đồ thuật toán mô hình dự báo đề xuất SE-XGB-LiGBM-RF-OW	99
Hình 4.4. Kết quả so sánh công suất dự báo và công suất thực tế trong các ngày điển hình của mùa xuân.....	107
Hình 4.5. Kết quả so sánh công suất dự báo và công suất thực tế trong các ngày điển hình của mùa hạ.....	108
Hình 4.6. Kết quả so sánh công suất dự báo và công suất thực tế trong các ngày điển hình của mùa thu.....	108
Hình 4.7. Kết quả so sánh công suất dự báo và công suất thực tế trong các ngày điển hình của mùa đông.....	109
Hình 4.8. Sơ đồ thuật toán mô hình thiết lập mô hình Selector-Model	113
Hình 4.9. Sơ đồ thuật toán dự báo cho nhiều nhà máy sử dụng mô hình SE-XGB-LGBM-RF	115

DANH MỤC CÁC TỪ VIẾT TẮT

Từ viết tắt	Từ tiếng anh	Nguyên nghĩa
AE	Auto Encoder	Bộ mã hóa tự động
AHAP	Adaptive Historical Analog Prediction	Dự báo tương tự lịch sử thích ứng
AI	Artificial Intelligence	Trí tuệ nhân tạo
ANN	Artificial Neural Network	Mạng noron nhân tạo
AR	Autoregressive	Mô hình tự hồi quy
ARIMA	Autoregressive Integrated Moving Average	Trung bình dịch chuyển tự động tích hợp
ARMA	Autoregressive Moving Average	Trung bình dịch chuyển tự động
BPNN	Back Propagation Neural Network	Mạng noron lan truyền ngược
BiGRU	Bidirectional Gated Recurrent Unit	Nút có cổng hồi tiếp thông tin hai chiều
BiLSTM	Bidirectional Long Short-Term Memory	Bộ nhớ ngắn-dài hạn hai chiều
CEEMDAM	Complete Ensemble Empirical Mode Decomposition	Phân tích phân thức chế độ kinh nghiệm hoàn chỉnh với nhiễu thích ứng

	with Adaptive Noise	
CFS	Climate Forecast System	Hệ thống dự báo khí hậu toàn cầu
CMV	Cloud Motion Vector	Dự báo chuyển động đám mây
CNN	Convolutional Neural Network	Mạng noron tích chập
CSI	Clear Sky Index	Chỉ số bầu trời quang đãng
CSP	Concentrated Solar Power	Hệ thống năng lượng mặt trời tập trung
DBN	Deep Belief Network	Mạng có độ tin cậy cao
DCNN	Deep Convolutional Neural Network	Mạng noron tích chập sâu
DILATE	DIstance-based LAbeling of Time sEries	Bộ biến đổi hợp nhất theo thời gian
DL	Deep Learning	Mạng học máy sâu
DSE	Dynamic State Estimation	Ước lượng trạng thái động
ECMWF	European Centre for Medium-Range Weather Forecasts	Trung tâm dự báo thời tiết Châu Âu
EEMD	Ensemble Empirical Mode Decomposition	Phân tích phân thức chế độ kinh nghiệm tập hợp

EWMA	Exponentially Weighted Moving Average	Dịch chuyển theo cấp số nhân trung bình
FBNN	Feedback Neural Network	Mạng noron phản hồi
FFNN	Feedforward Neural Network	Mạng noron truyền thẳng
GA	Genetic Algorithm	Thuật toán di truyền
GDAS	Global Data Assimilation System	Hệ thống đồng hóa dữ liệu toàn cầu
GHI	Global Horizontal Irradiance	Chỉ số bức xạ ngang toàn cầu
GRU	Gated Recurrent Unit	Nút có cổng hồi tiếp thông tin
IPSO	Improved Particle Swarm Optimization	Thuật toán tối ưu hóa bầy đàn cải tiến
ISSA	Improved Sparrow Search Algorithm	Thuật toán tìm kiếm chim sẻ cải tiến
KPCA	Kernel Principal Component Analysis	Phân tích thành phần chính sử dụng hàm Kernel
LightGBM	Light Gradient Boosting Machine	Mô hình tăng cường độ dốc
LSTM	Long Short-Term Memory	Bộ nhớ dài-ngắn hạn

MA	Moving Average	Mô hình trung bình dịch chuyển
MAE	Mean Absolute Error	Sai số trung bình tuyệt đối
MAPE	Mean Absolute Percentage Error	Sai số phần trăm trung bình tuyệt đối
ML	Machine Learning	Học máy
MLPNN	Multilayer Perceptron Neural Network	Mạng nơ-ron tri thức nhiều lớp
MRWE	Multiresolution Wavelet Encoding	Mã hóa wavelet nhiều độ phân giải
NARX	Nonlinear Autoregressive with Exogenous Inputs	Mạng tự hồi phi tuyến với đầu vào ngoại sinh
NLTT	Renewable Energy	Năng lượng tái tạo
NMAPE	Normalized Mean Absolute Percentage Error	Sai số phần trăm trung bình tuyệt đối chuẩn hóa
NOAA	National Oceanic and Atmospheric Administration	Dự báo thời tiết trong thời gian thực
NRMSE	Normalized Root Mean Square Error	Sai số bình phương trung bình chuẩn hóa
NSMO	National Load Dispatch Center	Công ty TNHH MTV Vận hành hệ thống điện và Thị trường điện quốc gia

NWP	Numerical Weather Prediction	Mô hình dự báo thời tiết số
PSO	Particle Swarm Optimization	Thuật toán tối ưu bầy đàn
PV	Photovoltaic	Pin quang điện
PVPF	Photovoltaic Power Forecasting	Dự báo công suất phát của nguồn PV
QR	Quantile Regression	Hồi quy lượng tử
QRA	Quantile Regression Averaging	Kỹ thuật hồi quy phân vị trung bình
R^2	Coefficient of Determination	Hệ số xác định
RF	Random Forest	Mô hình rừng ngẫu nhiên
RBFNN	Radial Basis Function Neural Network	Mạng nơon cơ sở phản hồi
RBN	Reduction, Filtering, and Regression Algorithm	Thuật toán giảm kích thước, lọc, hồi quy...
RCN	Recurrent Convolutional Network	Mạng chập hồi tiếp
RMSE	Root Mean Square Error	Sai số trung bình bình phương

RNN	Recurrent Neural Network	Mạng nơon hồi tiếp
SCADA	Supervisory Control and Data Acquisition	Hệ thống điều khiển, giám sát và thu thập dữ liệu
SE-XGB-LIGBM-RF-OW	Stacking Ensemble with XGBoost, LightGBM, Random Forest and Optimal Weights	Mô hình Stacking Ensemble với XGBoost, LightGBM, Random Forest và trọng số tối ưu
SLFN	Single Layer Feedforward Network	Lớp chuyển tiếp cấp dữ liệu một lớp
SRNN	Simplified Recurrent Neural Network	Mạng neural hồi tiếp đơn giản hóa
SSA	Sparrow Search Algorithm	Thuật toán tìm kiếm chim sẻ
SVM	Support Vector Machine	Thuật toán máy học lựa chọn vecto dữ liệu liên quan
SVR	Support Vector Regression	Mô hình hồi quy sử dụng thuật toán SVM
TFT	Temporal Fusion Transformer	Một mô hình biến đổi transformer tích hợp thông tin theo thời gian
VAE	Variational Autoencoder	Mô hình mã hóa tự động biến đổi
VMD	Variational Mode Decomposition	Phương pháp phân tích dữ liệu thành các chế độ không trùng lặp

WT	Wavelet Transform	Biến đổi sóng thành phần (wavelet)
XGBoost	Extreme Gradient Boosting	Mô hình tăng cường độ dốc cực đại

MỞ ĐẦU

1. Lý do chọn đề tài

Trong bối cảnh chuyển dịch năng lượng toàn cầu, Việt Nam đang tích cực thúc đẩy phát triển năng lượng tái tạo nhằm đảm bảo an ninh năng lượng, giảm phụ thuộc vào nhiên liệu hóa thạch và thực hiện các cam kết giảm phát thải khí nhà kính. Trong các nguồn tái tạo, điện mặt trời nổi lên như một trụ cột nhờ tiềm năng dồi dào và khả năng mở rộng nhanh chóng.

Tuy nhiên, bản chất phụ thuộc mạnh vào điều kiện thời tiết, đặc biệt là bức xạ mặt trời khiến công suất phát điện từ các hệ thống điện mặt trời dao động liên tục theo thời gian gây khó khăn cho điều độ và vận hành hệ thống điện. Sự bất định và thiếu ổn định này đặt ra yêu cầu cấp thiết về việc xây dựng các mô hình dự báo ngắn hạn có độ chính xác cao và khả năng ứng dụng thực tế, giúp đảm bảo cân bằng cung cầu, tối ưu hóa vận hành và nâng cao hiệu suất khai thác.

Tuy nhiên, phần lớn các mô hình dự báo công suất phát hiện nay đều giả định chuỗi dữ liệu lịch sử dài, đầy đủ và liên tục, đây là điều kiện khó đáp ứng tại nhiều nhà máy điện mặt trời ở Việt Nam. Các nhà máy mới đưa vào vận hành thường chỉ có tập dữ liệu ngắn, gián đoạn hoặc không đồng nhất, trong khi sự phân hóa rõ rệt theo mùa và vùng miền càng khiến việc xây dựng mô hình dự báo gặp nhiều thách thức.

Do đó, vấn đề đặt ra là: Làm thế nào để phát triển một mô hình dự báo ngắn hạn vừa đảm bảo độ chính xác, vừa có khả năng thích ứng theo mùa và vùng khí hậu, lại vừa có thể mở rộng triển khai cho nhiều nhà máy khác nhau trong điều kiện dữ liệu thực tế không đầy đủ?

Từ những thách thức trên, nhu cầu xây dựng một khung mô hình dự báo ngắn hạn linh hoạt, đáng tin cậy, có khả năng hoạt động hiệu quả với dữ liệu thiếu hụt, và dễ dàng mở rộng triển khai đa nhà máy đã trở nên cấp thiết.

2. Mục tiêu nghiên cứu

- Xây dựng phương pháp dự báo ngắn hạn công suất phát điện mặt trời ổn định và đáng tin cậy, có khả năng hoạt động hiệu quả ngay cả khi chuỗi dữ liệu lịch sử bị gián đoạn hoặc thiếu hụt, đồng thời đảm bảo khả năng thích ứng theo mùa và vùng khí hậu.
- Đánh giá và so sánh hiệu quả của các mô hình học máy trong dự báo bức xạ mặt trời, làm cơ sở lựa chọn đầu vào cho các mô hình dự báo công suất.

- Phát triển mô hình tổ hợp Stacking Ensemble, kết hợp các mô hình nền tảng có hiệu suất tốt nhất, đồng thời tích hợp cơ chế tối ưu trọng số theo mùa (Optimal Weights - OW) để nâng cao độ chính xác trong các giai đoạn thời tiết khác nhau.
- Thiết kế và triển khai cơ chế lựa chọn mô hình thích ứng (Selector-Model), cho phép mở rộng mô hình dự báo ra nhiều nhà máy điện mặt trời cùng lúc, mà không cần huấn luyện lại toàn bộ mô hình cho từng nhà máy riêng biệt.

3. Phương pháp nghiên cứu

- **Phương pháp tổng quan lý thuyết:** Nghiên cứu tài liệu chuyên ngành trong và ngoài nước về các phương pháp dự báo công suất phát điện mặt trời, bao gồm các mô hình học máy và học sâu như từ đó phân tích ưu, nhược điểm và xác định khoảng trống nghiên cứu.
- **Phương pháp thu thập và xử lý dữ liệu:** Thu thập dữ liệu thời tiết và công suất phát điện từ ba nhà máy điện mặt trời tại Việt Nam trong khoảng thời gian một năm; áp dụng các kỹ thuật tiền xử lý như làm sạch, chuẩn hóa để đảm bảo chất lượng dữ liệu đầu vào cho mô hình.
- **Phương pháp mô phỏng và thực nghiệm:** Xây dựng và huấn luyện các mô hình dự báo công suất phát điện mặt trời trong các kịch bản có dữ liệu đầy đủ và thiếu hụt, so sánh hiệu quả các mô hình dựa trên các chỉ số đánh giá sai số phổ biến như RMSE, MAPE, NRMSE, NMAPE.
- **Phương pháp tích hợp mô hình:** Đề xuất và triển khai mô hình tổ hợp với tối ưu trọng số, nhằm nâng cao độ chính xác và khả năng thích ứng theo điều kiện khí hậu. Đánh giá khả năng tổng quát hóa của mô hình tổ hợp khi dự báo cho nhiều nhà máy.

4. Phạm vi nghiên cứu

- **Khung thời gian dự báo:** Tập trung vào bài toán dự báo công suất phát của nhà máy điện mặt trời trong ngắn hạn (khoảng 1-3 ngày tới), đặc biệt trong điều kiện thiếu hụt hoặc gián đoạn dữ liệu quá khứ.
- **Đối tượng nghiên cứu:** Ba nhà máy điện mặt trời tiêu biểu tại Quảng Trị, Thanh Hóa và Đắk Lắk, đại diện cho các vùng khí hậu khác nhau tại Việt Nam.
- **Tập dữ liệu lịch sử:** Gồm các thông số thời tiết (bức xạ mặt trời, nhiệt độ môi trường, nhiệt độ tấm pin), đặc trưng thời gian (tháng/mùa) làm đầu vào; và công suất phát của nhà máy làm đầu ra.

- **Kịch bản dữ liệu:** Xây dựng mô hình trong bối cảnh dữ liệu bị thiếu, gián đoạn và không đồng nhất, phản ánh đúng thực tế tại các nhà máy ở Việt Nam.
- **Thử nghiệm dự báo bức xạ:** Triển khai và so sánh hiệu quả của các mô hình học máy trong việc dự báo bức xạ mặt trời, làm nền tảng đầu vào cho dự báo công suất phát của nhà máy.

5. Cơ sở khoa học và thực tiễn

Nghiên cứu dựa trên:

- **Cơ sở khoa học:** Tích hợp kiến thức từ các lĩnh vực học máy như mô hình chuỗi thời gian (LSTM, GRU, BiGRU), mô hình cây quyết định (XGBoost, LightGBM, RF), và kỹ thuật Stacking Ensemble với cơ chế tối ưu trọng số theo mùa (OW).
- **Nhu cầu thực tiễn:** Hệ thống điện Việt Nam ngày càng phụ thuộc vào nguồn năng lượng tái tạo, đòi hỏi dự báo ngắn hạn chính xác, ổn định và thích ứng trong điều kiện dữ liệu không hoàn hảo. Do đó, cần thiết phải xây dựng một khung mô hình linh hoạt, hiệu quả và có khả năng triển khai mở rộng cho nhiều nhà máy điện mặt trời.

6. Đóng góp mới của luận án

- Đề xuất mô hình tổ hợp SE-XGB-LGBM-RF-OW: Áp dụng Stacking Ensemble kết hợp với cơ chế tối ưu trọng số theo mùa, giúp nâng cao độ chính xác trong điều kiện thời tiết biến động.
- Xây dựng và đánh giá các mô hình dự báo bức xạ mặt trời bằng LightGBM, LSTM và GRU nhằm làm rõ hiệu quả của từng mô hình, qua đó hỗ trợ lựa chọn nguồn dữ liệu đầu vào tối ưu cho mô hình dự báo công suất phát.
- So sánh toàn diện các mô hình học sâu (LSTM, GRU, BiGRU) với các mô hình cây quyết định (XGBoost, LightGBM, Random Forest) trong bối cảnh thiếu dữ liệu, nhằm xác định mô hình có hiệu suất ổn định hơn.
- Phát triển mô hình hybrid LightGBM-LSTM: Trong đó LightGBM đảm nhận vai trò nội suy chuỗi dữ liệu bị thiếu, cung cấp đầu vào liên tục cho LSTM giải pháp thực tế cho bài toán dữ liệu không đầy đủ.
- Thiết kế mô hình Selector-Model: Để tự động lựa chọn mô hình dự báo dựa trên đặc điểm từng nhà máy và từng thời điểm, giúp tăng khả năng tổng quát hóa và ứng dụng đồng thời cho nhiều nhà máy khác nhau.

7. Kết cấu luận án

Luận án gồm 4 chương:

- Chương 1: Tổng quan nghiên cứu.
- Chương 2: Các mô hình dự báo bức xạ mặt trời và công suất phát điện mặt trời trong ngắn hạn.
- Chương 3: Phân tích đánh giá các mô hình dự báo bức xạ và công suất phát điện mặt trời.
- Chương 4: Đề xuất mô hình dự báo cho một số nhà máy điện mặt trời tại Việt Nam.

CHƯƠNG 1. TỔNG QUAN NGHIÊN CỨU

Năng lượng điện là yếu tố cốt lõi trong chiến lược phát triển bền vững, công nghiệp hóa, đô thị hóa và nâng cao chất lượng sống [1]. Tuy nhiên, sự phụ thuộc lâu dài vào các nguồn năng lượng hóa thạch đã gây ra những hệ lụy nghiêm trọng như ô nhiễm môi trường, biến đổi khí hậu và cạn kiệt tài nguyên. Trước thực trạng đó, nhiều quốc gia trong đó có Việt Nam đã và đang thúc đẩy mạnh mẽ quá trình chuyển dịch sang năng lượng tái tạo.

Trong số các nguồn tái tạo, điện mặt trời nổi lên như một giải pháp tiềm năng nhờ tính thân thiện với môi trường, chi phí giảm nhanh và khả năng triển khai linh hoạt. Tại Việt Nam, điện mặt trời đã có những bước phát triển ấn tượng cả về quy mô lẫn tốc độ, phản ánh rõ nét trong chiến lược an ninh năng lượng quốc gia. Tuy nhiên, đặc điểm phụ thuộc cao vào thời tiết – đặc biệt là bức xạ mặt trời – khiến công suất phát biến động lớn theo thời gian và không gian. Điều này đặt ra những thách thức lớn cho công tác điều độ, lập kế hoạch vận hành và đảm bảo ổn định hệ thống điện.

Trong bối cảnh đó, dự báo công suất phát điện mặt trời ngắn hạn trở thành một yêu cầu thiết yếu, không chỉ để hỗ trợ vận hành hệ thống hiệu quả mà còn là cơ sở quan trọng trong các quy định pháp lý và chiến lược phát triển năng lượng. Tuy nhiên, việc xây dựng các mô hình dự báo phù hợp với điều kiện đặc thù của Việt Nam vẫn còn gặp nhiều hạn chế, từ vấn đề thiếu hụt dữ liệu lịch sử, khí hậu phân hóa phức tạp, đến khó khăn trong việc mở rộng mô hình cho nhiều nhà máy khác nhau.

Với mục tiêu xây dựng một giải pháp dự báo sát với thực tiễn Việt Nam, Chương 1 sẽ tập trung phân tích các khía cạnh sau:

- Tầm quan trọng của năng lượng điện và xu hướng chuyển dịch sang năng lượng tái tạo trong chiến lược phát triển bền vững;
- Bức tranh toàn cảnh về phát triển điện mặt trời tại Việt Nam, bao gồm tiềm năng, chính sách, hiện trạng và các thách thức kỹ thuật;
- Vai trò, yêu cầu và khung pháp lý liên quan đến công tác dự báo công suất phát điện mặt trời;
- Tổng quan các mô hình dự báo công suất hiện nay, đặc biệt là các phương pháp học máy và học sâu.
- Nhận diện các khoảng trống nghiên cứu còn tồn tại, làm cơ sở cho hướng tiếp cận của luận án.

Những nội dung trên sẽ làm rõ bối cảnh, định hướng và cơ sở khoa học để phát triển một khung mô hình dự báo công suất phát điện mặt trời ngắn hạn, bền vững- thích ứng- khả chuyển, phù hợp với điều kiện dữ liệu thực tế và yêu cầu vận hành tại Việt Nam.

1.1. Tiềm năng phát triển điện mặt trời tại Việt Nam

Việt Nam có vị trí địa lý và điều kiện khí hậu thuận lợi để phát triển năng lượng mặt trời. Nước ta nằm trong vùng nhiệt đới gió mùa với nhiều giờ nắng trong năm, đặc biệt là khu vực phía Nam. Theo báo cáo của Trung tâm Khí tượng Thủy văn Quốc gia, tiềm năng năng lượng mặt trời trung bình toàn quốc dao động từ 4 đến 5 kWh/m²/ngày và số giờ nắng trung bình từ 1600 đến 2600 giờ/năm [2], tạo điều kiện lý tưởng cho việc khai thác năng lượng mặt trời.

a. Phân bố số giờ nắng và bức xạ tại Việt Nam

Phân bố số giờ nắng

Số giờ nắng tại Việt Nam phân bố không đồng đều trên toàn lãnh thổ do đặc điểm địa hình và khí hậu. Để làm rõ sự phân hóa về thời lượng chiếu nắng giữa các vùng khí hậu, bảng dưới đây trình bày tổng số giờ nắng trung bình theo tháng đại diện (I, IV, VII, X) và trung bình năm tại các vùng khí hậu chính trên lãnh thổ Việt Nam. Những khác biệt này phản ánh rõ đặc điểm mùa vụ, sự phân bố năng lượng mặt trời theo vĩ độ và là cơ sở quan trọng để điều chỉnh mô hình dự báo công suất phù hợp với từng khu vực [3]:

Bảng 1.1. Tổng số giờ nắng tại các vùng khí hậu

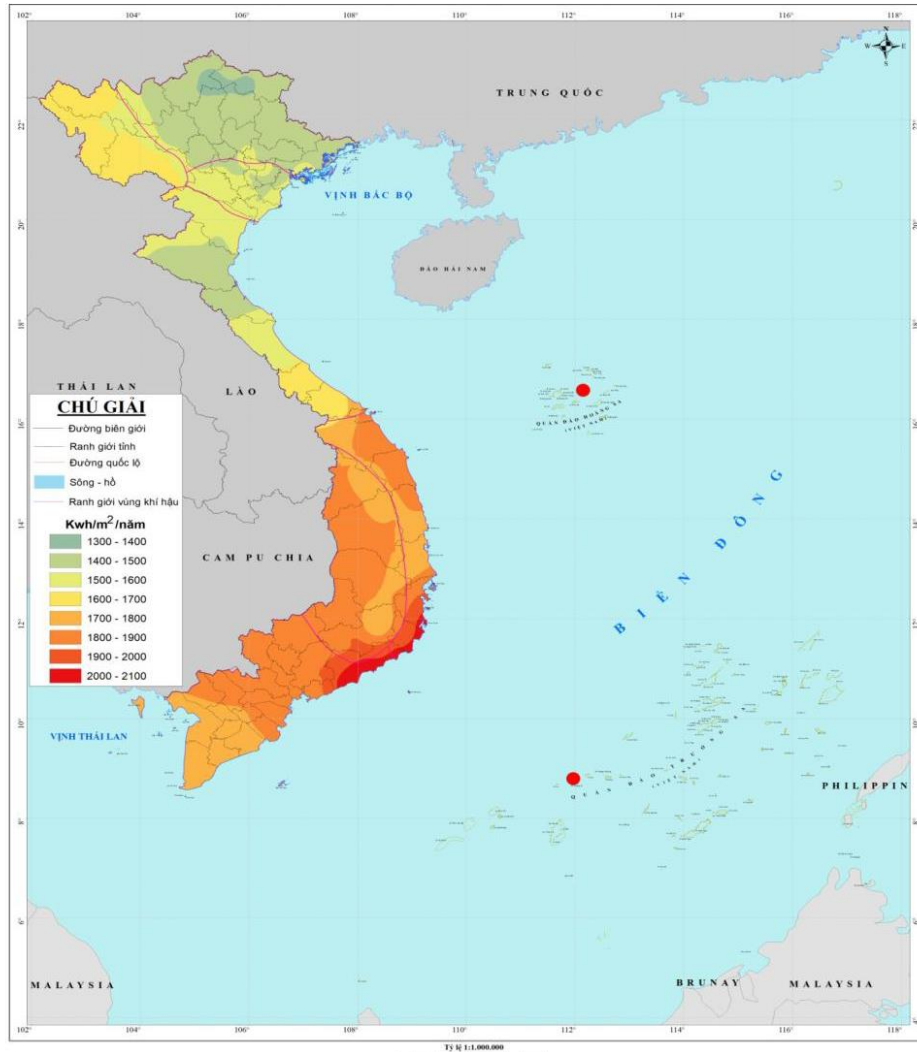
(Nguồn: Tài liệu [3])

Vùng khí hậu	Tháng 1	Tháng 4	Tháng 7	Tháng 10	Năm
Tây Bắc	128,9	171,6	145,8	156,0	1819,8
Đông Bắc	71,4	101,7	167,4	147,1	1502,4
Đồng Bằng Bắc bộ	71,6	89,7	195,1	160,8	1563,3
Bắc Trung Bộ	87,1	138,2	210,1	132,3	1653,0
Nam Trung Bộ	175,9	244,4	223,3	165,6	2364,4
Tây Nguyên	230,8	232,6	168,5	158,2	2312,4
Nam Bộ	245,8	253,9	181,1	175,9	2502,6

Phân bố bức xạ

Việt Nam có tiềm năng lớn về cường độ bức xạ mặt trời với mức trung bình khoảng 4,6 kWh/m²/ngày. Khu vực có bức xạ cao nhất là Tây Nguyên, Nam Trung Bộ và Nam Bộ, đạt từ 4,9 đến 5,7 kWh/m²/ngày, rất thuận lợi cho phát triển điện mặt trời quy mô

lớn. Trong khi đó, miền Bắc và Bắc Trung Bộ có mức cường độ bức xạ thấp hơn, từ 3,3 đến 4,6 kWh/m²/ngày, nhưng vẫn đủ đáp ứng các hệ thống điện mặt trời hộ gia đình và quy mô vừa. Sự phân bố không đồng đều này đặt ra yêu cầu quan trọng về việc điều chỉnh mô hình dự báo phù hợp với từng vùng khí hậu cụ thể. Phân bố bức xạ tại Việt Nam được thể hiện qua Hình 1.1 sau đây [3]:

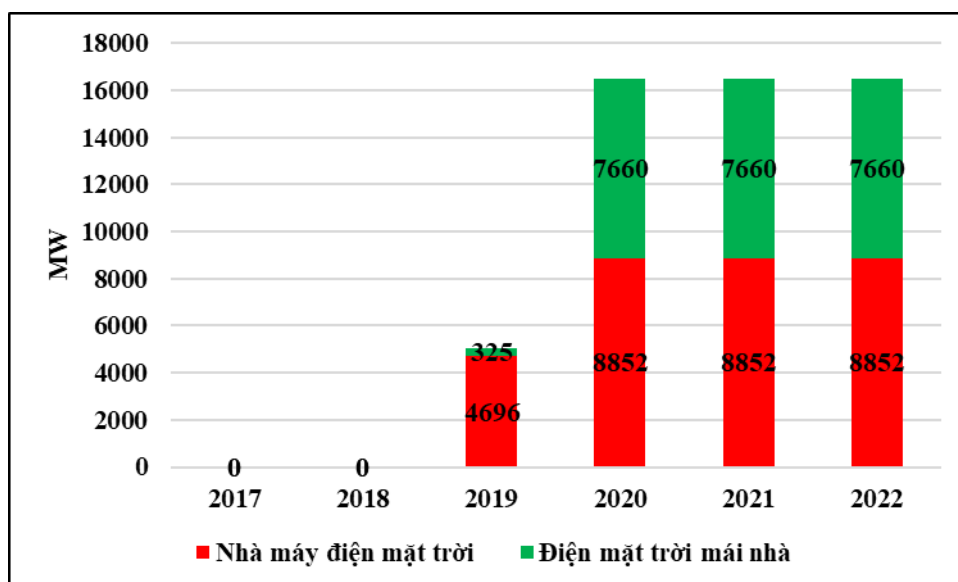


Hình 1.1. Phân bố bức xạ mặt trời tại Việt Nam

(Nguồn: Tài liệu [3])

b. Thực trạng phát triển năng lượng mặt trời

Trong những năm gần đây, công suất lắp đặt của các dự án điện mặt trời tại Việt Nam đã tăng trưởng nhanh chóng. Tính đến cuối năm 2023, tổng công suất lắp đặt của các dự án điện mặt trời tại Việt Nam đã đạt hơn 16600 MW [4]. Do vậy, Việt Nam đã trở thành một trong những quốc gia có tốc độ phát triển năng lượng mặt trời nhanh nhất thế giới. Hình 1.2 cho thấy sự tăng trưởng công suất của điện mặt trời tại Việt Nam từ 2017 đến 2022.



Hình 1.2. Tăng trưởng công suất điện mặt trời tại Việt Nam (2017-2022)

(Nguồn: Tài liệu [5])

Biểu đồ cho thấy sự tăng trưởng vượt bậc của điện mặt trời tại Việt Nam, đặc biệt trong giai đoạn 2019–2020. Công suất điện mặt trời mái nhà tăng mạnh nhờ cơ chế giá FIT. Sau khi cơ chế giá FIT kết thúc vào cuối năm 2020, tốc độ phát triển điện mặt trời tại Việt Nam có dấu hiệu chững lại, đặc biệt trong giai đoạn 2021–2023 ghi nhận sự ổn định về công suất điện mặt trời. Một số dự án điện mặt trời quy mô lớn đã được triển khai và đi vào hoạt động, tiêu biểu là các nhà máy được liệt kê trong phụ lục A.1.

Sự phát triển của thị trường tài chính xanh toàn cầu đang tạo ra cơ hội đáng kể cho đầu tư quốc tế vào lĩnh vực điện mặt trời tại Việt Nam. Với sự bùng nổ của thị trường tài chính xanh và sự hậu thuẫn từ các ngân hàng trung ương trong vai trò người cho vay cuối cùng đối với trái phiếu xanh, các dự án điện mặt trời tại Việt Nam đang có cơ hội tiếp cận nguồn vốn quốc tế với lãi suất ưu đãi và nhiều điều kiện thuận lợi [6].

c. Thách thức trong điều độ vận hành

Mặc dù có nhiều tiềm năng, phát triển năng lượng mặt trời tại Việt Nam cũng đối mặt với một số thách thức chính tác động trực tiếp đến việc triển khai các dự án điện mặt trời như sau [7]:

- Sự tăng trưởng bùng nổ của điện mặt trời, dù mang lại nhiều lợi ích, cũng tạo ra những thách thức vận hành cho hệ thống điện quốc gia. Thách thức lớn nhất đến từ chính bản chất của nguồn năng lượng này: công suất phát hoàn toàn phụ thuộc vào các điều kiện thời tiết biến động như bức xạ, mây che, gây ra sự bất định và khó lường theo từng phút, từng giờ.

- Sự biến động này gây áp lực nặng nề lên công tác điều độ, khiến việc duy trì cân bằng cung-cầu và đảm bảo ổn định lưới điện trở nên vô cùng phức tạp. Khi công suất từ các nhà máy điện mặt trời đồng loạt tăng cao vào các giờ nắng đỉnh, trong khi hạ tầng lưới truyền tải chưa theo kịp, tình trạng quá tải cục bộ và tắc nghẽn lưới điện thường xuyên xảy ra. Hệ quả trực tiếp là Trung tâm Điều độ Hệ thống điện Quốc gia (NSMO) buộc phải ra lệnh cắt giảm công suất của các nhà máy để đảm bảo an toàn gây lãng phí tài nguyên và ảnh hưởng đến hiệu quả kinh tế của các dự án.

d. Yêu cầu cấp thiết về dự báo công suất phát điện mặt trời

Bất chấp những thách thức trên, triển vọng phát triển điện mặt trời tại Việt Nam vẫn rất lớn, theo quy hoạch điện VIII điều chỉnh năng lượng tái tạo sẽ chiếm khoảng 28-36% tổng công suất lắp đặt vào năm 2030, trong đó tổng công suất các nguồn điện mặt trời (gồm điện mặt trời tập trung và điện mặt trời mái nhà, không tính đến các nguồn điện mặt trời theo khoản 5 Điều 10 Luật Điện lực số 61/2024/QH15) đạt 46.459 - 73.416 MW. Định hướng đến năm 2050 tổng công suất sẽ đạt 293.088 - 295.646 MW [8]. Tuy nhiên, để hiện thực hóa mục tiêu này và khai thác hiệu quả tiềm năng sẵn có, việc giải quyết các thách thức vận hành là yêu cầu bắt buộc.

Để hiện thực hóa mục tiêu này và khai thác hiệu quả tiềm năng sẵn có, việc dự báo chính xác công suất phát điện mặt trời trong ngắn hạn đã trở thành giải pháp kỹ thuật then chốt và cấp thiết nhất. Việc dự báo chính xác mang lại nhiều hiệu quả thiết thực trong vận hành hệ thống điện, cụ thể như sau:

- **Tối ưu hóa công tác điều độ và giảm chi phí vận hành:** Dự báo chính xác là yếu tố nền tảng cho quy trình cam kết tổ máy cho phép các nhà vận hành lập kế hoạch huy động các nguồn điện truyền thống (nhiệt điện, thủy điện) một cách hiệu quả hơn, tránh tình trạng khởi động quá nhiều tổ máy đắt tiền khi có nguồn năng lượng mặt trời giá rẻ sắp được đưa vào lưới. Điều này giúp giảm đáng kể chi phí nhiên liệu và chi phí khởi động/tắt máy, từ đó giảm tổng chi phí sản xuất điện của toàn hệ thống.
- **Giảm thiểu cắt giảm công suất và lãng phí tài nguyên:** Khi không có dự báo đáng tin cậy, hệ thống phải đối mặt với nguy cơ thừa công suất đột ngột khi trời nắng, buộc phải cắt giảm sản lượng điện mặt trời để tránh gây mất ổn định lưới điện. Các nghiên cứu đã chỉ ra rằng việc sử dụng các bản tin dự báo tốt hơn có thể giảm lượng

công suất mặt trời bị cắt giảm xuống dưới một phần ba, vì các nhà máy điện truyền thống được cam kết vận hành một cách hiệu quả hơn. Điều này giúp tối đa hóa việc khai thác nguồn năng lượng sạch và miễn phí, nâng cao hiệu quả kinh tế cho các dự án đầu tư.

- **Nâng cao độ tin cậy và an ninh lưới điện:** Việc dự báo chính xác giúp các nhà quản lý lưới điện chuẩn bị tốt hơn cho những biến động trong sản xuất điện mặt trời, từ đó duy trì cân bằng cung-cầu một cách chủ động. Nó cho phép tính toán và huy động lượng công suất dự phòng cần thiết một cách hợp lý để đối phó với các sai số dự báo, thay vì phải duy trì một lượng lớn công suất dự phòng tốn kém. Điều này đặc biệt quan trọng trong việc quản lý các sự kiện phụ tải cực đoan và đảm bảo an ninh hệ thống.
- **Hỗ trợ lập kế hoạch và ra quyết định:** Độ chính xác của các bản tin dự báo là yếu tố đầu vào quan trọng cho các quyết định mang tính chiến lược của ngành điện, từ việc thực hiện các nghĩa vụ đảm bảo độ tin cậy, quản lý các nguồn dự phòng khẩn cấp, cho đến việc lập kế hoạch phát triển điện lực tổng thể.

Tóm lại, dự báo chính xác không chỉ là một công cụ kỹ thuật mà còn là nền tảng cho việc vận hành kinh tế, an toàn và bền vững, cho phép tích hợp hiệu quả nguồn điện mặt trời vào lưới điện quốc gia.

1.2. Nhu cầu và hiện trạng hệ thống dự báo công suất phát điện mặt trời tại Việt Nam

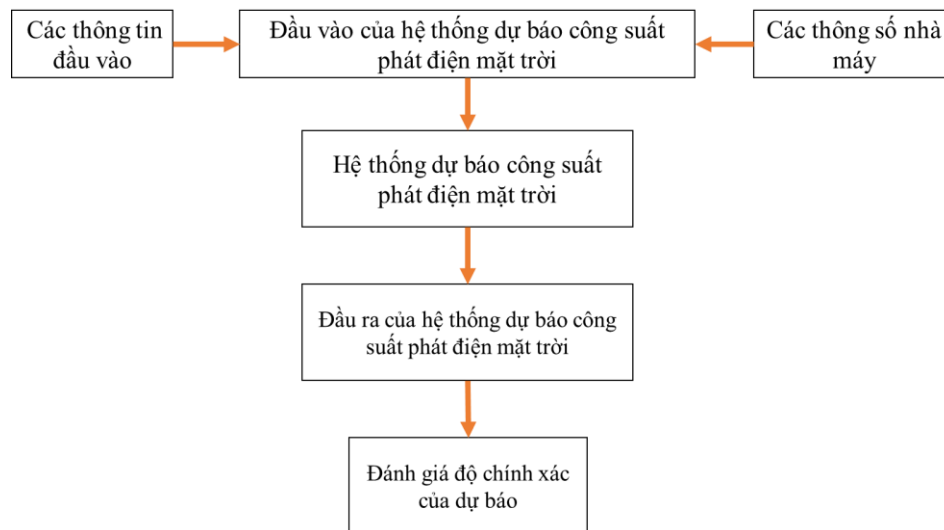
Với tiềm năng dồi dào và tốc độ phát triển nhanh chóng của điện mặt trời tại Việt Nam như đã nêu ở phần 1.1, việc dự báo chính xác công suất phát điện mặt trời đã trở thành một yêu cầu cấp thiết trong quản lý và vận hành hệ thống điện quốc gia. Đặc biệt, do đặc tính biến động mạnh theo thời tiết và chu kỳ tự nhiên, điện mặt trời là một nguồn phát khó kiểm soát và dễ gây mất cân bằng cung cầu nếu không được dự báo chính xác. Không chỉ là giải pháp kỹ thuật, hoạt động dự báo công suất còn là một nội dung được quy định cụ thể trong các văn bản pháp lý hiện hành. Do đó, mục này sẽ tập trung trình bày rõ hơn về nhu cầu, chính sách, hiện trạng và các khó khăn trong lĩnh vực dự báo công suất phát điện mặt trời tại Việt Nam hiện nay.

1.2.1. Nhu cầu và hiện trạng hệ thống dự báo công suất phát của các nhà máy điện mặt trời tại Việt Nam

Sự gia tăng nhanh chóng của các nhà máy điện mặt trời tại Việt Nam đã đặt ra yêu cầu cấp thiết về dự báo chính xác công suất phát để đảm bảo vận hành ổn định hệ thống điện. Trước thực tế này, Bộ Công Thương đã ban hành Quyết định 67/QĐ-ĐTĐL (2021) [9] và Quy định 4608/ĐĐQG-NLTT (2023) [10] nhằm quy định chi tiết quy trình và yêu cầu dự báo công suất phát của các nguồn điện năng lượng tái tạo.

Theo các văn bản nêu trên, hệ thống dự báo cần đáp ứng độ phân giải 15 phút cho vận hành trong ngày (cập nhật mỗi 30 phút), 30 phút cho dự báo ngày tới (D+1, D+2, cập nhật hai lần/ngày) và dự báo tuần tới (W+1, cập nhật mỗi tuần một lần). Yêu cầu độ chính xác của các khung thời gian dự báo cũng được công bố hằng năm, phù hợp với kế hoạch phát triển nguồn và yêu cầu bảo đảm an ninh năng lượng.

Hệ thống dự báo công suất phát có khả năng ước tính công suất phát điện tương lai của các trang trại điện mặt trời dựa trên thông số vận hành và dữ liệu khí tượng. Các hệ thống này sử dụng mô hình học máy tiên tiến, dữ liệu lịch sử, thông tin vận hành theo thời gian thực và dự báo khí tượng để tạo ra các dự báo chính xác về công suất phát điện của các nhà máy [11]. Một hệ thống dự báo công suất phát điện mặt trời được mô tả trong Hình 1.3 dưới đây:



Hình 1.3. Hệ thống dự báo công suất của nhà máy điện mặt trời

Trong sơ đồ trên, hệ thống dự báo công suất phát điện mặt trời hoạt động theo một chuỗi quy trình gồm bốn bước chính. Trước hết, các thông tin đầu vào bao gồm: dữ liệu dự báo thời tiết, dữ liệu vận hành trong quá khứ và các thông số kỹ thuật của nhà máy như vị trí địa lý, công suất đặt, cấu hình và đặc tính inverter, chuỗi tấm pin cũng như

công suất thiết kế. Ngoài ra, công suất khả dụng cũng được đưa vào hệ thống, được ước tính từ kế hoạch bảo trì, vận hành hoặc cắt giảm đột xuất của nhà máy. Sau khi tiếp nhận toàn bộ các đầu vào này, hệ thống dự báo sử dụng các mô hình tính toán tiên tiến để ước lượng công suất phát điện và sản lượng đầu ra dự kiến theo từng khung thời gian, với độ phân giải tùy thuộc yêu cầu điều độ và thị trường. Cuối cùng, độ chính xác của dự báo được đánh giá thông qua các chỉ số sai số, phổ biến nhất là sai số tuyệt đối trung bình (MAPE) giữa công suất phát dự báo với công suất phát thực tế và công suất phát dự báo với công suất lắp đặt của nhà máy, áp dụng trên tất cả các khung thời gian dự báo khác nhau.

Cơ quan thực hiện công tác dự báo là Công ty TNHH MTV Vận hành hệ thống điện và Thị trường điện quốc gia (NSMO – trước đây là A0). Việc xác định giá trị công suất dự báo được thực hiện dựa trên hai nguồn chính: dự báo của từng nhà máy và công suất dự báo tổng hợp từ nhiều nguồn dữ liệu. Trong trường hợp độ chênh lệch giữa hai giá trị không vượt quá $\pm 20\%$, NSMO chấp thuận dự báo của nhà máy; nếu vượt quá, kết quả dự báo tổng hợp sẽ được ưu tiên [12]. Hệ thống tổng hợp sử dụng bốn nguồn dữ liệu chính: (i) hai đơn vị dự báo quốc tế (qua đấu thầu), (ii) dự báo nội bộ của NSMO, (iii) dữ liệu thời tiết từ một bên thứ ba độc lập, và (iv) dữ liệu dự báo từ nhà máy. Mỗi nguồn được gán trọng số theo độ tin cậy trong quá khứ nhằm tối ưu hóa độ chính xác.

1.2.2. Các khó khăn, thách thức trong dự báo công suất phát điện mặt trời tại Việt Nam

Thực tế việc triển khai các mô hình dự báo công suất phát tại Việt Nam còn gặp nhiều thách thức. NSMO đang áp dụng hai hình thức: dự báo trung tâm (quy mô lớn, chi phí thấp, nhưng phức tạp và có thể thiếu dữ liệu chi tiết) và dự báo phân tán tại từng nhà máy (có độ chính xác cao hơn, nhưng tốn kém chi phí triển khai). Tuy nhiên, quy trình này bộc lộ một số bất cập:

- Thiếu quy định chi tiết về sai số: Mức sai số 15% được áp dụng chung [10], chưa có tiêu chuẩn riêng cho từng khung thời gian dự báo, trong khi độ chính xác thực tế giảm dần khi dự báo trong tương lai xa hơn.
- Một vấn đề đáng chú ý là quy định độ lệch $\pm 20\%$ giữa dự báo của nhà máy và dự báo tổng hợp [13], [14]. Hệ quả của sai số này bao gồm: (1) giảm độ chính xác dự báo chung; (2) yêu cầu bổ sung nguồn phát khác để bù đắp công suất khi dự báo không chính xác; (3) khó khăn trong việc tối ưu hóa lưu trữ điện năng.

- Hệ thống cập nhật dữ liệu còn thủ công: Việc gửi và nhận bản tin dự báo phần lớn vẫn thực hiện qua các kênh thủ công, dễ gây ra sai sót và chậm trễ, không đáp ứng yêu cầu vận hành thời gian thực.

Những khó khăn này cho thấy nhu cầu cấp bách về việc nghiên cứu và phát triển các mô hình dự báo tiên tiến, có khả năng thích ứng tốt hơn với điều kiện thực tế tại Việt Nam.

1.2.3. Các khung thời gian dự báo công suất phát điện mặt trời

Tùy theo mục tiêu vận hành và yêu cầu kỹ thuật, dự báo công suất điện mặt trời được chia thành nhiều khung thời gian khác nhau. Các khung này được trình bày trong Bảng 1.2 sau đây:

Bảng 1.2. Các khung thời gian dự báo công suất phát điện mặt trời
(Nguồn: Tài liệu tham khảo [15])

Loại dự báo	Khung dự báo	Ứng dụng
Dự báo rất ngắn hạn	Vài giây cho tới dưới 30 phút [16]	Điều khiển thời gian thực của hệ thống điện, pin lưu trữ, chào giá điện trong ngày [15], [17]
Dự báo ngắn hạn	Từ 30 phút đến 6h [16] cho tới 1 tuần [17]	Lập kế hoạch vận hành và lập kế hoạch giao dịch [18]
Dự báo trung hạn	Từ 6-24h [16] hoặc 1 tuần đến 1 tháng hoặc 1 năm [15]	Lên kế hoạch bảo trì định kỳ, tối ưu vận hành lưới điện [17]
Dự báo dài hạn	Hơn 24h [16] hoặc 1 năm đến 10 năm [17]	Quy hoạch công suất, đầu tư, phát triển hạ tầng điện [19]

Trong bối cảnh tại Việt Nam, luận án tập trung vào việc dự báo công suất điện mặt trời trong khoảng thời gian từ 1 đến 3 ngày. Theo phân loại của Utpal Kumar Das và các cộng sự đây là dạng dự báo ngắn hạn [17]. Tuy nhiên, một số nghiên cứu như R. Ahmed và các cộng sự [15] cho rằng khoảng thời gian từ 1 đến 3 ngày có thể được xem là trung hạn trong các hệ thống bảo trì quy mô lớn hoặc thị trường điện kéo dài. Trong phạm vi tại Việt Nam, nghiên cứu sinh lựa chọn khoảng thời gian từ 1 đến 3 ngày là dự báo ngắn hạn dài ngày nhằm thống nhất với các ứng dụng kỹ thuật phổ biến tại Việt Nam và phù hợp với mục tiêu điều độ, vận hành ngắn hạn của nhà máy.

1.3. Tổng quan các mô hình dự báo công suất phát điện mặt trời

Trước những khó khăn và thách thức nêu trên, việc lựa chọn và áp dụng mô hình dự báo công suất phát phù hợp đóng vai trò then chốt trong việc nâng cao độ chính xác và ổn định cho hệ thống điện. Trong phần này, luận án sẽ trình bày tổng quan về các loại mô hình dự báo công suất điện mặt trời, từ các mô hình truyền thống đến các mô hình hiện đại dựa trên trí tuệ nhân tạo, nhằm phân tích ưu nhược điểm và khả năng ứng dụng của từng phương pháp trong điều kiện thực tế.

1.3.1. Các sai số để đánh giá mô hình dự báo

Để đánh giá hiệu suất mô hình dự báo, nghiên cứu sinh sử dụng các sai số: RMSE, NRMSE, MAPE, NMAPE. Các sai số này được thể hiện qua Bảng 1.3 dưới đây:

Bảng 1.3. Các sai số được áp dụng để đánh giá hiệu suất mô hình dự báo

Sai số	Công thức tính	Giải thích các ký hiệu
RMSE [20]: Là sai số trung bình bình phương giữa các giá trị dự báo và giá trị thực tế	$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2} \quad (1.1)$	$\hat{y}_i$ là công suất dự báo (kW), y_i là công suất thực tế (kW), n là số lượng điểm dữ liệu.
NRMSE [21]: Là sai số trung bình bình phương gốc đã chuẩn hóa. Tương tự như sai số RMSE nhưng sai số này đã được chuẩn hóa để so sánh hiệu suất trên các tập dữ liệu có đơn vị đo khác nhau	$NRMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n \frac{(\hat{y}_i - y_i)^2}{Capacity}} \quad (1.2)$	$\hat{y}_i$ là công suất dự báo (kW), y_i là công suất thực tế (kW), n là số lượng điểm dữ liệu, Capacity là tổng công suất lắp đặt của nhà máy (kW)
MAPE [22]: Là sai số tuyệt đối trung bình phần trăm. Sai số MAPE cho biết tỷ lệ phần trăm trung	$MAPE = \frac{100}{n} \sum_{i=1}^n \frac{ P_i^{db} - P_i^{tt} }{P_i^{tt}} \quad (1.3)$	P_i^{db} giá trị công suất phát của dự báo thứ i (MW), là P_i^{tt} giá trị công suất trong thực tế của tín hiệu

bình của sai số tuyệt đối so với giá trị thực tế		thứ i (MW), n là số tín hiệu dự báo được đánh giá trong khung thời gian của loại dự báo
NMAPE [9]: Là sai số tuyệt đối phần trăm trung bình đã được chuẩn hóa	$NMAPE = \frac{100}{n} \sum_{i=1}^n \frac{ P_i^{db} - P_i^{tt} }{P_{dm}} \quad (1.4)$	P_i^{db} giá trị công suất phát của dự báo thứ i (MW), là 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)
MAE [23]: Là sai số tuyệt đối trung bình giữa giá trị dự báo và giá trị thực tế. Công thức tính MAE như sau	$MAE = \frac{1}{n} \sum_{i=1}^n \hat{y}_i - y_i \quad (1.5)$	$\hat{y}_i$ là bức xạ dự báo (W/m^2), y_i là bức xạ thực tế (W/m^2), n là số lượng điểm dữ liệu.

1.3.2. Tổng quan các nghiên cứu về dự báo công suất phát điện mặt trời

Việc tổng hợp và phân tích các công trình nghiên cứu nổi bật trong giai đoạn 2018–2024 cho thấy dự báo công suất điện mặt trời đã trở thành một chủ đề nghiên cứu sôi động, với nhiều công trình quốc tế và trong nước tập trung cải thiện độ chính xác cũng như khả năng ứng dụng trong thực tế vận hành hệ thống điện. Các hướng phát triển nổi bật có thể khái quát thành ba nhóm chính: (i) mở rộng và cải tiến các kiến trúc học sâu, (ii) khai thác hiệu quả các mô hình học máy dựa trên cây quyết định, và (iii) hình thành các mô hình kết hợp (hybrid/ensemble) nhằm tận dụng ưu điểm của nhiều phương pháp.

a. Sự phát triển của các mô hình học sâu và mô hình kết hợp (Hybrid Model)

Các mô hình học sâu, nhờ khả năng mô tả các mối quan hệ phi tuyến phức tạp trong dữ liệu chuỗi thời gian, tiếp tục giữ vai trò trung tâm trong lĩnh vực dự báo công suất phát.

Các mô hình học sâu nền tảng: Các mô hình hồi quy nơ-ron như LSTM và GRU được xem là những mô hình tiêu biểu. Nhiều nghiên cứu trong nước đã chứng minh hiệu quả của chúng trong các bối cảnh cụ thể. Nhóm tác giả Nguyễn Quang Ninh và cộng sự

đã áp dụng LSTM cho nhà máy Phong Điền [24], kết quả mô hình dự báo LSTM đạt sai số MAPE thấp nhất chỉ 1,5% khi sử dụng dữ liệu lịch sử của nhà máy, sai số này tăng lên 4,432% khi sử dụng dữ liệu được cung cấp từ một bên thứ ba. Trong nghiên cứu của mình [25] nghiên cứu sinh cũng khẳng định mô hình GRU có thể hoạt động đáng tin cậy khi bộ dữ liệu đầu vào đầy đủ.

Các cải tiến của kiến trúc LSTM: Nhận thấy tiềm năng của LSTM, nhiều nhóm nghiên cứu đã đề xuất các biến thể phức tạp hơn. Tiêu biểu là nghiên cứu của M. Chai (2019) và các cộng sự [26], nhóm nghiên cứu đã so sánh nhiều biến thể của LSTM và cho thấy kiến trúc AHAP-LSTM đã giảm sai số trung bình δ MAPE xuống 5,79%, vượt trội so với 19,82% của LSTM gốc.

Các mô hình kết hợp (Hybrid Models): Xu hướng kết hợp nhiều kiến trúc để khai thác thế mạnh riêng của từng loại trở nên ngày càng phổ biến. V. Suresh và nhóm nghiên cứu (2020) [27] đã áp dụng mô hình CNN-LSTM cho dự báo công suất phát điện mặt trời, kết quả cho thấy hiệu suất ổn định của mô hình dự báo trong khoảng thời gian từ 1 giờ ($RMSE = 0,053$) tới 1 tuần ($RMSE = 0,045$). Mohammed Sabri và các cộng sự (2022) [28] đã xây dựng mô hình kết hợp GRU-CNN, mô hình đạt sai số MAE trung bình là 0,081. Bên cạnh đó, nhóm nghiên cứu của Pengyun Jia (2020) [29] đã chứng minh rằng việc kết hợp phân rã tín hiệu VMD với tối ưu hóa ISSA trong mô hình VMD-ISSA-GRU có thể đưa sai số MAE xuống mức 1,012 và R^2 đạt tới 0,999. Tại Việt Nam, Nguyễn Thị Hoài Thu và cộng sự đã xây dựng mô hình kết hợp CEEMDAN-BiLSTM-PSO [30], kết quả cho thấy sai số RMSE, s-MAPE, n-RMSE, MAE của mô hình đề xuất thấp hơn hẳn các mô hình đơn lẻ ANN, LSTM, BiLSTM.

Sự phát triển của mô hình Transformer: Gần đây, kiến trúc Transformer với cơ chế tự chú ý (self-attention) được áp dụng vào dự báo công suất. Jihoon Moon và nhóm nghiên cứu (2024) [31] đã kết hợp 1D-CNN và Transformer, kết quả của mô hình kết hợp có $MAE = 1,243$ kWh, vượt trội so với CNN thuần túy. Fatma Mazen Ali Mazen và các cộng sự (2023) [32] cũng đã đề xuất mô hình GRU-DILATE-TFT (Temporal Fusion Transformer), kết quả cho thấy tiềm năng lớn của mô hình TFT khi mô hình được đề xuất có hiệu suất tốt hơn nhiều trong phép thử Diebold-Mariano so với các mô hình XGBoost, NHiTS và N-BEATS, thể hiện qua giá trị p nhỏ hơn 0,05 và các sai số $MAE=1,19$; $RMSE=1,44$; $MSE=2,08$ đều ở mức thấp.

b. Sự phát triển của các mô hình dựa trên cây quyết định

Song song với các mô hình học sâu, các mô hình Ensemble dựa trên cây quyết định như Random Forest (RF), XGBoost, và LightGBM vẫn chứng tỏ tính cạnh tranh nhờ hiệu suất ổn định, thời gian huấn luyện nhanh và khả năng diễn giải tốt hơn.

Các mô hình cây quyết định đơn lẻ: Alexandra Khalyasmaa cùng nhóm nghiên cứu (2019) [22] đã sử dụng Random Forest Regressor để dự báo công suất phát cho một nhà máy điện mặt trời có công suất 15MW tại miền nam nước Nga, kết quả cho thấy mô hình đạt độ chính xác trung bình khoảng 93%. Waqas Khan (2022) [33] đã xây dựng mô hình XGBoost kết hợp với thuật toán tập hợp xếp chồng (DSE), kết quả mô hình giúp cải thiện R^2 thêm 10–12% so với các mô hình ANN, LSTM và bagging, điều này chứng minh tính ổn định vượt trội của mô hình đề xuất trong điều kiện thời tiết biến động.

Các mô hình cây quyết định kết hợp: Zhen Wang và các cộng sự (2023) [34] đã chứng minh sự hiệu quả của mô hình LightGBM-LSTM khi sai số dự báo MAPE của nó chỉ là 3,70%, thấp hơn nhiều so với các mô hình độc lập XGBoost (4,18%), LightGBM (4,88%) và LSTM (6,94%). Jiandong Ye và các cộng sự (2023) [35] đã kết hợp LightGBM-XGBoost, kết quả dự báo của mô hình liên tục đạt RMSE và MAE thấp nhất khi so sánh với các mô hình LSTM và SVR. Safia Babikir cùng nhóm nghiên cứu (2024) [36] đã đề xuất mô hình XGBoost-CNN kết quả thu được với $R^2 = 0,9962$, mô hình kết hợp cũng duy trì hiệu suất cao ngay cả trong điều kiện thời tiết động. Tại Việt Nam, Phan Quốc Thắng và nhóm nghiên cứu (2021) [37] đã xây dựng mô hình kết hợp KPCA-XGBoost-NWP. Khi so sánh sai số RMSE của mô hình KPCA-XGBoost-NWP với PCA-XGBoost và XGBoost tại 5 nhà máy điện mặt trời tại Đài Loan thì mô hình KPCA-XGBoost-NWP có sai số RMSE thấp nhất.

c. Giới hạn phạm vi ứng dụng của các mô hình dự báo

Mặc dù các nghiên cứu đã phát triển nhiều mô hình tinh vi và đạt độ chính xác cao, nhưng một hạn chế nổi bật vẫn tồn tại: phần lớn các mô hình dự báo đều được triển khai trong phạm vi một nhà máy hoặc một địa điểm duy nhất.

Các nghiên cứu quốc tế của các nhóm như của M. Chai, V. Suresh hay Pengyun Jia đều chứng minh mô hình đề xuất cho kết quả dự báo tốt hơn các mô hình tham chiếu, nhưng chỉ áp dụng trên một bộ dữ liệu từ một nhà máy.

Tại Việt Nam, nhiều nghiên cứu cũng đi theo hướng riêng biệt. Nguyễn Quang Ninh và cộng sự lần lượt áp dụng LSTM cho Phong Điền, Thành Thành Công 1, hay cho bài

toán cắt giảm công suất [38] đều tiến hành dự báo cho một nhà máy. Các công trình khác về dự báo cho các hệ thống điện mặt trời mái nhà [39], [40] cũng chỉ dừng lại ở việc dự báo cho một hệ thống duy nhất.

Cách tiếp cận mỗi nhà máy có một mô hình dự báo giúp tối ưu hóa tại từng địa điểm, nhưng lại khó áp dụng trên quy mô hệ thống điện quốc gia, nơi tồn tại hàng trăm nhà máy với điều kiện khí hậu và dữ liệu đa dạng.

Có thể thấy, các nghiên cứu quốc tế lẫn trong nước đều tập trung vào việc áp dụng các mô hình học sâu và học máy để cải thiện độ chính xác dự báo công suất phát điện mặt trời. Tuy nhiên, điểm chung là hầu hết các nghiên cứu đều giả định dữ liệu lịch sử của nhà máy là đầy đủ và liên tục, trong khi thực tế vận hành tại Việt Nam và nhiều quốc gia khác cho thấy tình trạng thiếu hụt dữ liệu quá khứ do lỗi hệ thống hoặc gián đoạn thu thập là rất phổ biến.

Bên cạnh đó, đa số công trình chỉ thực nghiệm trên một nhà máy duy nhất, chưa xem xét khả năng tổng quát hóa và mở rộng mô hình cho nhiều nhà máy khác nhau vốn có đặc điểm khí hậu, quy mô và điều kiện vận hành khác biệt. Điều này dẫn tới khoảng trống quan trọng: chưa có một giải pháp dự báo công suất phát điện mặt trời nào vừa xử lý tốt dữ liệu thiếu hụt, vừa thích ứng khi triển khai đồng thời trên nhiều nhà máy đặt tại các khu vực địa lý khác nhau (như miền Trung, miền Bắc và Tây Nguyên) nơi có sự khác biệt rõ rệt về bức xạ, khí hậu và điều kiện vận hành.

1.4. Khoảng trống trong nghiên cứu

Từ phân tích tại các phần 1.1 đến 1.3 có thể thấy điện mặt trời ở Việt Nam đang phát triển nhanh chóng, tuy nhiên hệ thống dữ liệu thực địa còn tồn tại nhiều hạn chế như thiếu hụt, gián đoạn và không đồng nhất. Bên cạnh đó, điều kiện khí hậu phân hóa mạnh theo mùa và vùng địa lý, trong khi yêu cầu vận hành hệ thống ngày càng đòi hỏi các mô hình dự báo ngắn hạn phải đảm bảo tính chính xác, ổn định và kịp thời.

+) Khoảng trống trong nghiên cứu:

Mặc dù có nhiều tiến bộ trong lĩnh vực dự báo công suất phát điện mặt trời, vẫn còn tồn tại những khoảng trống quan trọng cần được lấp đầy để đáp ứng yêu cầu thực tiễn của hệ thống điện tại Việt Nam.

Thứ nhất, phần lớn các nghiên cứu hiện nay vẫn chưa giải quyết hiệu quả bài toán dự báo trong điều kiện chuỗi dữ liệu quá khứ bị đứt gãy hoặc thiếu hụt. Việc xử lý dữ liệu bị thiếu chủ yếu sử dụng các phương pháp nội suy đơn giản như giữ nguyên giá trị

trước đó (forward fill) hoặc trung bình trượt (moving average), các phương pháp này vốn có độ tin cậy thấp và dễ lan truyền sai số vào quá trình huấn luyện mô hình, làm giảm đáng kể độ chính xác của kết quả dự báo.

Thứ hai, có rất ít nghiên cứu đánh giá định lượng mức độ ảnh hưởng của việc sử dụng dữ liệu bức xạ dự báo thay cho dữ liệu bức xạ thực đo trong bài toán dự báo công suất ngắn hạn. Sự chênh lệch này không chỉ phụ thuộc vào chất lượng mô hình bức xạ mà còn bị ảnh hưởng bởi yếu tố mùa vụ và vùng khí hậu cụ thể, điều mà nhiều nghiên cứu trước đây còn bỏ ngỏ.

Thứ ba, các mô hình dự báo thường được đánh giá riêng lẻ mà thiếu đi những phân tích so sánh có hệ thống giữa các nhóm mô hình khác nhau, đặc biệt trong điều kiện dữ liệu đầu vào không hoàn hảo. Điều này làm hạn chế khả năng chọn lựa mô hình phù hợp nhất cho các bối cảnh khác nhau về dữ liệu và khí hậu.

Ngoài ra, một số mô hình kết hợp (Hybrid) hiện nay chủ yếu dừng lại ở việc kết hợp dữ liệu đầu vào, trong khi tiềm năng của việc kết hợp đầu ra giữa các mô hình thông qua các kỹ thuật tổ hợp như Stacking Ensemble lại chưa được khai thác một cách bài bản. Việc tận dụng sức mạnh tổng hợp của các mô hình khác nhau ở cấp độ đầu ra có thể giúp cải thiện rõ rệt độ chính xác và độ ổn định của dự báo.

Cuối cùng, việc triển khai mô hình dự báo cho nhiều nhà máy điện mặt trời vẫn còn hạn chế về khả năng tổng quát hóa, do phần lớn các nghiên cứu chỉ huấn luyện mô hình riêng biệt cho từng nhà máy. Chưa có một khung mô hình nào có khả năng thích ứng linh hoạt và khả chuyển giữa các nhà máy nằm ở những vùng khí hậu khác nhau mà không cần phải huấn luyện lại hoàn toàn từ đầu. Điều này gây khó khăn cho việc mở rộng mô hình ra quy mô lớn trên toàn quốc.

+) Hướng tiếp cận của luận án:

Để giải quyết bài toán dự báo công suất phát điện mặt trời trong điều kiện dữ liệu đầu vào không hoàn hảo, luận án đề xuất một hướng tiếp cận theo tiến trình nhiều lớp, mang tính hệ thống và có khả năng thích ứng với điều kiện vận hành tại Việt Nam.

Trước tiên, luận án tập trung vào việc xây dựng và so sánh ba mô hình học máy hiện đại là LightGBM, LSTM và GRU trong bài toán dự báo bức xạ mặt trời. Mục tiêu của bước này không phải là sử dụng trực tiếp kết quả dự báo bức xạ cho mô hình công suất, mà là để đánh giá mức độ tin cậy và tính khả dụng của từng loại dữ liệu đầu vào từ đó đưa ra khuyến nghị về thời điểm nên sử dụng dữ liệu thực đo tại trạm, dữ liệu dự báo từ

tổ chức khí tượng chuyên nghiệp, hay dữ liệu do mô hình nội bộ sinh ra. Đây là nền tảng quan trọng giúp nâng cao chất lượng đầu vào của chuỗi dự báo công suất trong các bước tiếp theo.

Tiếp theo, luận án thực hiện so sánh một cách có hệ thống hai nhóm mô hình dự báo công suất chính: nhóm mô hình cây quyết định (XGBoost, LightGBM, Random Forest) và nhóm mô hình chuỗi thời gian (LSTM, GRU, BiGRU). Việc so sánh được triển khai trong các kịch bản thiếu hụt chuỗi dữ liệu quá khứ nhằm phản ánh điều kiện dữ liệu thực tế tại nhiều nhà máy điện mặt trời ở Việt Nam. Các kết quả đánh giá sẽ giúp xác định mô hình nào có độ tin cậy cao hơn trong từng điều kiện cụ thể.

Từ kết quả trên, luận án đề xuất một cấu trúc mô hình kết hợp (hybrid) giữa LightGBM và LSTM, trong đó LightGBM đóng vai trò như một bộ lấp đầy các chuỗi dữ liệu bị thiếu để đầu ra sau cùng được đưa vào mô hình LSTM thực hiện dự báo công suất. Hướng tiếp cận này vừa tận dụng được năng lực học phi tuyến của LightGBM, vừa phát huy thế mạnh ghi nhớ chuỗi dài hạn của LSTM.

Sau khi đánh giá hiệu năng từng mô hình đơn lẻ, luận án tiếp tục tiến hành tổ hợp đầu ra các mô hình có độ chính xác cao bằng phương pháp Stacking Ensemble, trong đó trọng số kết hợp giữa các mô hình được tối ưu hóa theo từng mùa khí hậu thông qua thuật toán OW (Optimized Weights). Cách tiếp cận này không chỉ giúp bù trừ sai số của từng mô hình mà còn nâng cao độ ổn định và khả năng thích ứng theo mùa vụ.

Cuối cùng, để mở rộng khả năng ứng dụng mô hình ở quy mô lớn, luận án xây dựng một mô hình Selector-Model có khả năng tự động lựa chọn mô hình phù hợp nhất cho từng nhà máy tại từng thời điểm cụ thể. Mô hình này được huấn luyện trên dữ liệu tổng hợp của nhiều nhà máy, cho phép thực hiện dự báo đồng thời cho nhiều nhà máy có đặc điểm khí hậu và điều kiện vận hành khác biệt, mà không cần huấn luyện lại toàn bộ hệ thống. Đây là bước đột phá quan trọng nhằm hướng tới mục tiêu triển khai dự báo công suất điện mặt trời đa nhà máy- đa vùng trong thực tế, đặc biệt ở các quốc gia đang phát triển như Việt Nam.

1.5. Kết luận chương 1

Chương 1 đã cung cấp cái nhìn tổng quan về bối cảnh phát triển điện mặt trời tại Việt Nam, làm rõ tiềm năng, xu thế tăng trưởng, đồng thời nhấn mạnh các thách thức thực tiễn trong vận hành và dự báo công suất. Mặc dù sở hữu nguồn tài nguyên bức xạ dồi dào và tốc độ phát triển nhanh, hệ thống điện mặt trời Việt Nam đang đối mặt với

nhiều vấn đề đáng chú ý như: sự phụ thuộc vào điều kiện thời tiết khiến công suất biến động mạnh; các ràng buộc từ hệ thống điều độ; và đặc biệt là sự thiếu hụt, gián đoạn, không đồng nhất của dữ liệu vận hành tại các nhà máy.

Trong khi đó, yêu cầu từ cơ quan điều độ ngày càng đặt nặng tính chính xác của dự báo ngắn hạn, nhất là ở các khu vực tập trung nhiều nguồn năng lượng tái tạo. Điều này đặt ra yêu cầu cấp thiết phải phát triển các mô hình dự báo không chỉ chính xác mà còn bền vững trong điều kiện dữ liệu thiếu hụt.

Tổng quan các nghiên cứu hiện tại cho thấy, hầu hết các mô hình dự báo công suất vẫn chủ yếu được xây dựng trên dữ liệu lịch sử đầy đủ, liên tục và chỉ áp dụng cho từng nhà máy riêng biệt hoặc một mùa cụ thể. Các phương pháp tổ hợp mô hình (Stacking Ensemble) tuy được quan tâm, nhưng ít mô hình chú trọng đến khả năng thích ứng theo thời tiết hoặc khả năng mở rộng dự báo cho nhiều nhà máy cùng lúc. Do đó, luận án đã xác định được những khoảng trống nghiên cứu chính cần giải quyết.

Để giải quyết các vấn đề trên, luận án đề xuất một mô hình tổ hợp thích ứng, được phát triển theo hướng tiếp cận có hệ thống.

Mô hình được đề xuất không chỉ hướng đến tăng độ chính xác mà còn bảo đảm tính linh hoạt, khả năng thích ứng và khả năng tổng quát hóa những yếu tố then chốt đối với các hệ thống điện mặt trời tại Việt Nam.

Chương tiếp theo sẽ trình bày chi tiết về phương pháp luận, dữ liệu sử dụng và thiết kế thực nghiệm để hiện thực hóa khung dự báo tổ hợp này.

CHƯƠNG 2. CÁC MÔ HÌNH DỰ BÁO BỨC XẠ MẶT TRỜI VÀ CÔNG SUẤT PHÁT ĐIỆN MẶT TRỜI TRONG NGẮN HẠN

Chương 1 đã phác họa rõ nét bối cảnh phát triển điện mặt trời tại Việt Nam, những yêu cầu thực tiễn trong công tác điều độ vận hành và khoảng trống nghiên cứu trong dự báo công suất phát điện mặt trời. Trên nền tảng đó, chương 2 sẽ đi sâu vào ba nội dung trọng tâm, đặt nền móng cho việc xây dựng và đánh giá các mô hình dự báo:

Thứ nhất, chương giới thiệu chi tiết các tập dữ liệu được sử dụng trong nghiên cứu, bao gồm cả dữ liệu chính (từ nhà máy điện mặt trời tại Quảng Trị) và các tập dữ liệu bổ sung từ Thanh Hóa và Đắk Lắk. Việc trình bày nguồn gốc, cấu trúc, giai đoạn thời gian và các đặc trưng của dữ liệu nhằm đảm bảo tính minh bạch, đồng thời làm rõ các vấn đề liên quan đến dữ liệu thiếu, đứt gãy và phân hóa vùng miền.

Thứ hai, chương tiến hành xây dựng và đánh giá các mô hình dự báo bức xạ mặt trời ngắn hạn. Bức xạ là biến đầu vào có ảnh hưởng trực tiếp đến sản lượng điện phát ra, do đó việc lựa chọn mô hình bức xạ tối ưu là một bước quyết định đến chất lượng dự báo công suất trong các chương tiếp theo.

Thứ ba, chương trình bày việc phân tích, đánh giá các mô hình dự báo công suất phát điện mặt trời, bao gồm hai nhóm chính: Các mô hình truyền thống (mô hình quán tính, vật lý, thống kê) và nhóm mô hình học máy hiện đại như các mô hình cây quyết định (XGBoost, LightGBM, Random Forest), các mô hình học sâu chuỗi thời gian (LSTM, GRU, BiGRU) và các mô hình kết hợp, mô hình Stackinh Ensemble. Phân tích này không chỉ xác định mô hình có hiệu suất tốt mà còn mở ra cơ sở cho việc kết hợp các mô hình trong chiến lược tổ hợp ở các chương tiếp theo.

Qua ba nội dung trên, chương 2 đóng vai trò là nền tảng kỹ thuật quan trọng, vừa cung cấp dữ liệu thực nghiệm, vừa phân tích đánh giá các mô hình dự báo cốt lõi, từ đó chuẩn bị cho việc xây dựng mô hình dự báo tổ hợp phù hợp với bối cảnh vận hành tại Việt Nam.

2.1. Bộ dữ liệu thực tế và tiền xử lý dữ liệu

2.1.1. Giới thiệu và phân tích bộ dữ liệu thực tế

a. Nguồn gốc và Bối cảnh:

- **Nhà máy:** Nhà máy điện mặt trời tại tỉnh Quảng Trị.
- **Công suất lắp đặt:** 49,5 MW.

- **Thời gian thu thập:** Dữ liệu được ghi nhận liên tục trong khoảng thời gian từ ngày 01/01/2022 đến 31/12/2022.
- **Tần suất lấy mẫu:** 5 phút/lần, cung cấp một chuỗi thời gian có độ phân giải cao với tổng cộng 105.120 điểm dữ liệu.

b. Các đặc trưng dữ liệu:

Bộ dữ liệu bao gồm các biến số quan trọng ảnh hưởng trực tiếp đến quá trình phát điện của nhà máy, cụ thể:

- Bức xạ mặt trời (GHI): Cường độ bức xạ ngang toàn phần (W/m^2).
- Nhiệt độ môi trường: Nhiệt độ không khí xung quanh ($^{\circ}\text{C}$).
- Nhiệt độ tấm pin: Nhiệt độ bề mặt của tấm pin quang điện ($^{\circ}\text{C}$).
- Công suất phát thực tế: Sản lượng điện thực tế của nhà máy (kW).

c. Phân tích thống kê mô tả:

Để hiểu rõ hơn về đặc tính của bộ dữ liệu, các thông số thống kê cơ bản được tổng hợp trong Bảng 2.1 dưới đây:

Bảng 2.1: Bảng thống kê các đặc trưng của tập dữ liệu đầu vào

Đặc trưng	Trung bình	Tối thiểu	Tối đa	Độ lệch chuẩn
Bức xạ mặt trời (W/m^2)	162,38	0	1257	260,94
Nhiệt độ môi trường ($^{\circ}\text{C}$)	25,37	10,25	39,25	4,78
Nhiệt độ tấm pin ($^{\circ}\text{C}$)	27,77	11,5	61,7	9,21
Công suất phát (kW)	7105,56	0	41864,09	11423,7

Phân tích tương quan cho thấy Bức xạ mặt trời có mối tương quan mạnh nhất với công suất phát $r = 0,987$ sau đó là nhiệt độ tấm pin với $r = 0,893$.

Thông số chi tiết của tập dữ liệu tại Quảng Trị được trình bày trong bảng phụ lục A2.

2.1.2. Phương pháp tiền xử lý dữ liệu

Để đảm bảo chất lượng và tính nhất quán của dữ liệu đầu vào cho các mô hình học máy, một quy trình tiền xử lý chặt chẽ đã được áp dụng, bao gồm các bước sau:

Bước 1: Làm sạch và xử lý giá trị bất thường

Quy trình làm sạch dữ liệu được thực hiện theo hai cấp độ, nhằm đảm bảo tính khách quan và độ tin cậy vật lý cho quá trình huấn luyện mô hình.

Thứ nhất, loại bỏ các lỗi cảm biến cơ bản: Các điểm dữ liệu vi phạm các ràng buộc vật lý hiển nhiên, điển hình là trường hợp bức xạ bằng 0 nhưng vẫn ghi nhận công suất phát khác 0, được xác định là hệ quả của lỗi thiết bị đo hoặc nhiễu trong quá trình truyền dữ liệu SCADA. Thông qua quá trình rà soát, NCS đã loại bỏ 149 điểm dữ liệu thuộc nhóm sai lệch này.

Thứ hai, phân biệt và loại bỏ các nhiễu vận hành nâng cao: Để làm sạch sâu tập dữ liệu, NCS xây dựng tiêu chí phân biệt giữa biến động công suất do yếu tố thời tiết và sai lệch do vận hành, dựa trên nguyên lý đồng biến vật lý giữa bức xạ mặt trời và công suất phát.

- **Biến động do thời tiết** được xác định khi bức xạ và công suất biến thiên cùng chiều. Các trường hợp công suất giảm nhẹ trong khi bức xạ tăng, do ảnh hưởng của nhiệt độ tấm pin làm suy giảm hiệu suất, vẫn được giữ lại nhằm phản ánh đúng hành vi thực tế của hệ thống.
- **Sai lệch do vận hành** được nhận diện khi xuất hiện các nghịch lý vật lý rõ rệt, không thể giải thích bằng các cơ chế khí tượng thông thường. Dựa trên tiêu chí này, NCS tiếp tục loại bỏ 116 điểm dữ liệu bất thường, bao gồm:
 - Bức xạ tăng trong khi công suất sụt giảm mạnh từ 30% đến 50% (23 điểm).
 - Bức xạ tiếp tục gia tăng nhưng công suất duy trì ở mức cố định do ràng buộc vận hành (87 điểm).
 - Bức xạ lớn ($>100 \text{ W/m}^2$) nhưng công suất bằng 0 do bảo trì (6 điểm).

Tổng kết, sau quá trình tiền xử lý, tổng cộng 265 điểm dữ liệu bất thường (chiếm khoảng 0,25% tổng số mẫu) đã được loại bỏ. Việc loại trừ các nhiễu vận hành này giúp mô hình tránh bị chi phối bởi các yếu tố chủ quan như cắt giảm công suất hoặc sự cố thiết bị, từ đó tập trung học đúng quy luật phản ứng vật lý của nhà máy trước biến động khí tượng, nhằm nâng cao độ tin cậy của kết quả dự báo công suất phát.

Bước 2: Chuẩn hóa dữ liệu:

Toàn bộ các đặc trưng đầu vào được chuẩn hóa về cùng một thang đo trong khoảng $[0, 1]$ bằng phương pháp Min-Max scaling. Các tham số chuẩn hóa (giá trị min, max) được lưu lại để áp dụng nhất quán cho cả tập kiểm tra và dữ liệu dự báo sau này.

Quy trình xử lý này là tiêu chuẩn áp dụng cho việc xây dựng bộ dữ liệu nền tảng. Các kịch bản đặc thù nhằm giải quyết thách thức khi dữ liệu bị thiếu sẽ được trình bày riêng trong các phần thực nghiệm của luận án.

2.1.3. Các bộ dữ liệu bổ sung

Để phục vụ cho mục tiêu cuối cùng của luận án là xây dựng một phương pháp dự báo có khả năng tổng quát hóa và mở rộng cho nhiều nhà máy trên các vùng miền khác nhau, bên cạnh bộ dữ liệu chính tại Quảng Trị, nghiên cứu cũng thu thập và sử dụng hai bộ dữ liệu bổ sung từ các nhà máy tại tỉnh Đắk-Lắk (Tây Nguyên) và Thanh Hóa (Bắc Trung Bộ) và một bộ dữ liệu mới được thu thập từ 01/01/2024 đến 31/12/2024 của nhà máy tại Quảng Trị.

Việc giới thiệu sớm các bộ dữ liệu này nhằm cung cấp một cái nhìn toàn cảnh về nguồn tài nguyên dữ liệu thực tế sẽ được sử dụng để kiểm chứng hiệu quả của mô hình tổ hợp và mô hình Selector-Model ở Chương 4. Mặc dù chúng sẽ không được dùng trong các phân tích nền tảng ban đầu, việc mô tả chúng ở đây giúp làm rõ bối cảnh và phạm vi kiểm chứng của luận án.

Ba bộ dữ liệu bổ sung này, mặc dù có cùng tần suất lấy mẫu và các đặc trưng thời tiết cơ bản, nhưng có một khác biệt quan trọng so với bộ dữ liệu chính: chúng không có dữ liệu về Nhiệt độ tấm pin. Sự thiếu hụt này chính là lý do kỹ thuật then chốt để lựa chọn bộ dữ liệu Quảng Trị thu thập từ 01/01/2022 đến 31/12/2022 làm nền tảng cho việc phân tích sâu và so sánh các mô hình cơ sở ở các chương tiếp theo.

Thông tin chi tiết về hai bộ dữ liệu bổ sung được trình bày trong Bảng 2.2 dưới đây:

Bảng 2.2: Các đặc trưng của ba bộ dữ liệu bổ sung tại nhà máy Quảng Trị, Đắk-Lắk và Thanh Hóa

Các đặc điểm	Nhà máy Quảng Trị	Nhà máy Đắk-Lắk	Nhà máy Thanh Hóa
Vùng khí hậu	Trung Trung Bộ	Tây Nguyên	Bắc Trung Bộ
Công suất lắp đặt (MW)	49,5	50	30
Thời gian thu thập	01/01/2024-31/12/2024	01/01/2024-31/12/2024	01/01/2024-31/12/2024

Các đặc trưng	GHI, Nhiệt độ môi trường, Công suất	GHI, Nhiệt độ môi trường, Công suất	GHI, Nhiệt độ môi trường, Công suất
---------------	-------------------------------------	-------------------------------------	-------------------------------------

Các bộ dữ liệu này cũng sẽ được áp dụng quy trình tiền xử lý (làm sạch, xử lý dữ liệu thiếu, chuẩn hóa) tương tự như bộ dữ liệu chính để đảm bảo tính nhất quán khi được đưa vào sử dụng ở Chương 4.

Việc lựa chọn 3 nhà máy tại Thanh Hóa, Quảng Trị và Đắk Lắk nhằm đảm bảo tính đại diện cho các đới khí hậu và đặc điểm bức xạ đặc trưng tại Việt Nam.

- **Khu vực Thanh Hóa** đại diện cho vùng khí hậu Bắc Trung Bộ với chế độ bức xạ biến đổi mạnh theo mùa, đặc biệt là hiện tượng mây mù và mưa phùn vào mùa Đông - Xuân làm gia tăng tính bất định của dữ liệu đầu vào, gây thách thức lớn cho việc duy trì độ chính xác của các thuật toán học máy.
- **Khu vực Quảng Trị** đại diện cho vùng chuyển tiếp khí hậu, chịu ảnh hưởng trực tiếp của bão và hiệu ứng gió Phơn gây nhiệt độ cực đoan, tác động lớn đến hiệu suất thiết bị.
- **Khu vực Đắk Lắk** đại diện cho vùng Tây Nguyên có tiềm năng bức xạ lớn nhưng phân hóa mùa mưa - khô sâu sắc, với các cơn mưa rào bất chợt gây sụt giảm công suất phát đột ngột trong ngày.

Sự đa dạng về điều kiện tự nhiên giữa các địa điểm này cho phép đánh giá toàn diện khả năng thích nghi và tính khái quát hóa của mô hình dự báo hỗn hợp trong các kịch bản thời tiết thực tế khác nhau trên cả nước.

2.2. Các mô hình dự báo bức xạ mặt trời

Để xây dựng một mô hình dự báo bức xạ mặt trời hiệu quả, trước hết cần nắm vững các nền tảng lý thuyết và các phương pháp tiếp cận hiện có. Mục này sẽ trình bày tổng quan các yếu tố vật lý ảnh hưởng đến bức xạ mặt trời, các phương pháp dự báo từ truyền thống đến hiện đại, và hệ thống thu thập dữ liệu cần thiết, từ đó làm cơ sở cho việc lựa chọn và xây dựng mô hình thực nghiệm trong các phần sau.

2.2.1. Cơ sở lý thuyết và tầm quan trọng của dự báo bức xạ mặt trời

Cường độ bức xạ mặt trời tại một địa điểm chịu ảnh hưởng bởi sự tương tác phức tạp của nhiều yếu tố, bao gồm vị trí địa lý (vĩ độ, độ cao), thời gian (thời điểm trong ngày, mùa trong năm), và thành phần khí quyển (mây, hơi nước, sol khí) [41]. Việc hiểu rõ các yếu tố này là nền tảng cơ bản để xây dựng các mô hình dự báo bức xạ mặt trời

chính xác, từ đó hỗ trợ hiệu quả cho việc quy hoạch và vận hành các hệ thống năng lượng mặt trời.

Để mô hình hóa các yếu tố này, các mô hình toán học đã được phát triển. Mô hình đơn giản nhất ước tính BXM dựa trên góc thiên đỉnh và hệ số truyền qua khí quyển [42]:

$$I = I_0 \times \cos(\theta) \times T_t \quad (2.1)$$

trong đó: I là cường độ bức xạ tại bề mặt Trái Đất, I_0 là hằng số mặt trời (khoảng 1361 W/m^2), θ là góc thiên đỉnh, và T_t là hệ số truyền qua khí quyển. Mô hình này cung cấp một ước tính nhanh về bức xạ mặt trời, nhưng không tính đến nhiều yếu tố phức tạp của khí quyển và bề mặt Trái Đất.

Các mô hình phức tạp hơn như ASHRAE Clear Sky [43]. Mô hình này tính đến sự thay đổi theo mùa của khoảng cách Trái Đất-Mặt Trời và độ đục của khí quyển. Nó sử dụng các hệ số A, B, và C được xác định thực nghiệm để điều chỉnh cho các điều kiện khí quyển khác nhau. Công thức cơ bản của mô hình ASHRAE là [43]:

$$I_{th} = I_{DN}(\cos z + C) = |A \exp(-Bmr)|(\cos z + C) \quad (2.2)$$

trong đó: I_{DN} là cường độ bức xạ trực tiếp bình thường dưới bầu trời quang đãng, z là góc thiên đỉnh của mặt trời, C là hệ số bức xạ khuếch tán, A là cường độ bức xạ trực tiếp bình thường dự kiến tại khối khí quyển quang học bằng không, B là hệ số suy giảm khí quyển, m là khối lượng không khí quang học ở mực nước biển, được tính theo công thức của Rodgers ($m = 35/(1224\cos^2 z + 1)^{0.5}$), r là hệ số điều chỉnh theo độ cao, được giới thiệu bởi Powell ($r = (1 - \frac{h}{44308})^{5.257}$, h là độ cao của trạm đo so với mực nước biển).

Dự báo bức xạ mặt trời đóng vai trò then chốt trong việc quyết định trực tiếp đến công suất phát điện của nhà máy. Về cơ bản, công suất phát tỷ lệ trực tiếp với cường độ bức xạ theo công thức sau [44]:

$$P = \overline{H_T} \cdot A_s \cdot \eta \quad (2.3)$$

Trong đó: $\overline{H_T}$ là giá trị cường độ bức xạ mặt trời trung bình ngày (W/m^2), A_s là diện tích lắp đặt pin mặt trời (m^2), và η là hiệu suất chuyển đổi của hệ thống PV. Công thức này giả định điều kiện ổn định, không xét ảnh hưởng của nhiệt độ, suy hao truyền dẫn, hoặc các yếu tố bất lợi khác, nhưng vẫn cung cấp ước lượng sơ bộ hợp lý để kiểm tra hoặc khởi tạo mô hình dự báo.

Việc dự báo bức xạ mặt trời tốt là nền tảng cho việc dự báo công suất phát có độ chính xác cao, qua đó giúp tối ưu hóa vận hành và quản lý lưới điện, lập kế hoạch đầu tư và giảm thiểu rủi ro tài chính cho các dự án năng lượng mặt trời.

2.2.2. Các mô hình dự báo bức xạ mặt trời

a. Mô hình dự báo dựa trên dữ liệu vệ tinh và chuyển động của đám mây (CMV)

Các mô hình này là trọng tâm của dự báo rất ngắn hạn (vài phút đến vài giờ). Bằng cách sử dụng chuỗi hình ảnh liên tiếp từ các vệ tinh địa tĩnh (như GOES, Himawari), các thuật toán có thể tính toán và ngoại suy vector chuyển động của mây để dự đoán vị trí của chúng trong tương lai gần [45]. Các kỹ thuật phổ biến bao gồm tương quan chéo và dòng quang học [46], [47]. Ưu điểm của các mô hình này là khả năng nắm bắt biến động tức thời trên một khu vực rộng. Tuy nhiên, độ chính xác giảm nhanh theo thời gian và gặp khó khăn trong việc dự báo sự hình thành hay tan rã đột ngột của mây [48].

b. Dự báo thời tiết số (Numerical Weather Prediction - NWP)

Đây là mô hình tiêu chuẩn cho dự báo ngắn và trung hạn (từ 6 giờ đến vài ngày). Các mô hình NWP (như GFS, ECMWF) sử dụng các phương trình vật lý và khí động học phức tạp để mô phỏng trạng thái khí quyển trên quy mô toàn cầu hoặc khu vực, từ đó đưa ra dự báo về các biến số khí tượng, bao gồm cả bức xạ mặt trời [49]. Điểm mạnh của NWP là nền tảng vật lý vững chắc và khả năng dự báo dài hạn. Hạn chế chính là yêu cầu tài nguyên tính toán khổng lồ, độ phân giải không gian-thời gian thấp, và có thể chứa sai số hệ thống cần được hiệu chỉnh ở cấp độ địa phương [50], [51].

c. Các mô hình dự báo bức xạ mặt trời sử dụng trí tuệ nhân tạo

Các mô hình dự báo sử dụng trí tuệ nhân tạo tập trung vào việc cung cấp các dự báo dài hạn và trung hạn, thường kéo dài từ vài ngày đến vài tháng. Mục tiêu chính là đảm bảo rằng các dự báo bức xạ từ các mô hình này hoạt động ổn định và đáng tin cậy trong một khoảng thời gian dài [52]. Điều này đặc biệt quan trọng trong các lĩnh vực như quản lý năng lượng, nông nghiệp và quy hoạch đô thị, nơi các quyết định cần dựa trên những dự báo chính xác và liên tục về điều kiện thời tiết và bức xạ mặt trời.

Để đạt được mục tiêu dự báo trung và dài hạn, các phương pháp học máy hiện đại được sử dụng rộng rãi. Mỗi phương pháp có những ưu và nhược điểm riêng, phù hợp với các loại dữ liệu và yêu cầu dự báo khác nhau.

Trong số các phương pháp học máy được ứng dụng rộng rãi trong dự báo bức xạ mặt trời, các mô hình như mạng nơron nhân tạo (ANN, LSTM, GRU, BiGRU...), mô hình SVM, mô hình cây quyết (Random Forest) và các thuật toán tăng cường gradient (XGBoost, LightGBM) đã cho thấy hiệu quả cao trong việc xử lý dữ liệu phức tạp và cải thiện độ chính xác dự báo [23], [53]–[56]:

- **Mạng nơ-ron nhân tạo:** Mạng nơ-ron nhân tạo là các mô hình học máy mạnh mẽ, mô phỏng hoạt động của não bộ con người để xử lý thông tin và học từ dữ liệu. Các mạng nơ-ron đặc biệt là các mô hình học sâu có thể xử lý dữ liệu phức tạp và phi tuyến tính, giúp cải thiện độ chính xác của dự báo bức xạ mặt trời.
- **Mô hình máy học sử dụng vector hỗ trợ (SVM):** SVM là một thuật toán học máy giám sát, được sử dụng để phân loại và hồi quy. SVM có thể tạo ra các mô hình dự báo mạnh mẽ bằng cách tìm kiếm mặt phẳng phân tách tối ưu giữa các lớp dữ liệu.
- **Random Forest:** Đây là một phương pháp học máy dựa trên cây quyết định, sử dụng nhiều cây quyết định để đưa ra dự báo tổng hợp. Random Forest giúp giảm thiểu sai số và cải thiện độ chính xác của dự báo bằng cách kết hợp nhiều mô hình dự báo.
- **Gradient Boosting (XGBoost, LightGBM):** Các thuật toán này là các phương pháp tăng cường gradient, sử dụng nhiều mô hình yếu kết hợp lại để tạo thành một mô hình mạnh mẽ. XGBoost và LightGBM đã chứng tỏ hiệu quả cao trong nhiều cuộc thi học máy và ứng dụng thực tế, giúp cải thiện đáng kể độ chính xác của dự báo bức xạ mặt trời.

2.2.3. Nguồn dữ liệu và hệ thống thu thập dữ liệu

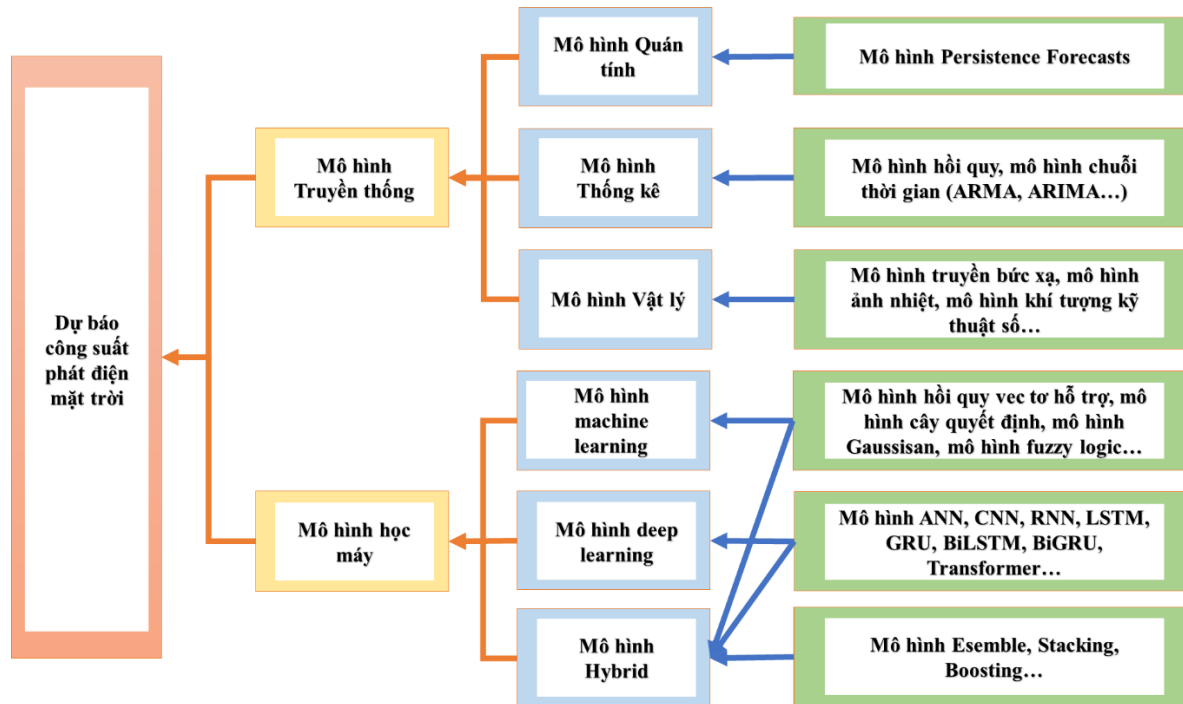
Một hệ thống dự báo BXMT hiệu quả đòi hỏi sự tích hợp của nhiều nguồn dữ liệu và một hạ tầng thu thập đáng tin cậy.

- **Nguồn dữ liệu:** Bao gồm dữ liệu khí tượng cơ bản (nhiệt độ, độ ẩm...) từ các trạm quan trắc mặt đất, dữ liệu BXMT trực tiếp từ các cảm biến chuyên dụng (pyranometer, pyrliometer), và dữ liệu viễn thám từ vệ tinh và radar [15], [23].
- **Hệ thống thu thập:** Tại các nhà máy điện mặt trời hiện đại, hệ thống SCADA đóng vai trò trung tâm trong việc thu thập, giám sát và truyền dữ liệu từ các trạm thời tiết tích hợp và các cảm biến về trung tâm điều khiển [57]. Dữ liệu này sau đó được đưa vào các hệ thống dự báo tích hợp, nơi các mô hình (NWP, AI...) được triển khai để đưa ra dự báo BXMT và công suất phát cuối cùng [58].

Trong khuôn khổ luận án, nhận thấy tiềm năng và tính linh hoạt của các phương pháp dựa trên trí tuệ nhân tạo, các thực nghiệm tiếp theo sẽ tập trung vào việc xây dựng và đánh giá hiệu suất của các mô hình học máy để giải quyết bài toán dự báo bức xạ mặt trời.

2.3. Các mô hình dự báo công suất phát điện mặt trời

Trong lĩnh vực dự báo công suất phát điện mặt trời, một số mô hình đã được sử dụng như: Mô hình quán tính, thống kê, trí tuệ nhân tạo AI (bao gồm cả mạng nơron nhân tạo ANN, mô hình hybrid để dự báo đã được sử dụng cho dự báo đầu ra PV). Các nhà nghiên cứu cũng đã so sánh các mô hình dự báo khác nhau và đưa ra các kết quả để so sánh sai số của các mô hình. Hình 2.1 thống kê một số mô hình dự báo công suất phát điện mặt trời.



Hình 2.1. Các mô hình dự báo công suất phát điện mặt trời

Việc lựa chọn mô hình dự báo phụ thuộc chặt chẽ vào đặc tính dữ liệu, miền dự báo (forecast horizon), và yêu cầu vận hành. Dưới đây là phân loại các nhóm mô hình chính, từ cơ bản đến nâng cao, cùng với phân tích vai trò, ưu nhược điểm và điều kiện áp dụng của chúng.

2.3.1. Các mô hình được nghiên cứu tại nước ngoài

2.3.1.1. Các mô hình truyền thống

a. Mô hình Quán tính (Persistence Technique)

Mô hình Quán tính phổ biến cho dự báo rất ngắn hạn và ngắn hạn. Kỹ thuật này thông qua khái niệm “ngày nay bằng ngày mai” giả định công suất phát tại thời điểm kế tiếp bằng công suất hiện tại [59]. Đây là mô hình đơn giản nhất, thường được dùng làm đường cơ sở (baseline) để so sánh với các mô hình phức tạp hơn. Phương trình của mô hình Quán tính được mô tả như sau [59]:

$$P(t + k|t) = \frac{1}{T} \sum_{i=0}^{n-1} P(t - i\Delta t) \quad (2.4)$$

Trong đó: k là khoảng thời gian dự báo, $P(t + k|t)$: Dự báo công suất cho thời gian k_t tại thời điểm tức thời t , T là khoảng dự báo, n là số phép đo trong tập dữ liệu, $P(t - i\Delta t)$ là công suất thực được đo trong thời gian t và các bước thời gian i trong T , Δt là chênh lệch thời gian của chuỗi thời gian đo.

Mô hình Quán tính có hiệu suất của dự báo tương đối chính xác ở thời điểm cực ngắn hạn và ngắn hạn (từ vài giây cho tới 1 giờ). Tuy nhiên, với sự tăng dần theo thời gian, độ chính xác của mô hình này giảm đi rất nhiều.

b. Mô hình vật lý

Trong hệ thống dự báo công suất phát điện mặt trời, các mô hình vật lý thường không được sử dụng trực tiếp để ước lượng công suất phát mà đóng vai trò trung gian, thông qua việc mô phỏng bức xạ mặt trời tới bề mặt các tấm pin mặt trời. Do đó, trong nghiên cứu này, mô hình vật lý sẽ được phân tích trong chương 2 như một thành phần quan trọng trong chuỗi dự báo bức xạ thay vì xem là mô hình độc lập trong hệ thống dự báo công suất.

c. Mô hình thống kê

Mô hình thống kê sử dụng chuỗi thời gian trong quá khứ để rút trích các mối quan hệ và khuynh hướng nhằm dự báo giá trị trong tương lai. Không giống như mô hình Quán tính hay học máy, phương pháp này chỉ sử dụng dữ liệu lịch sử và các công thức toán học thuần túy để ước lượng công suất phát, đòi hỏi nhiều dữ liệu đầu vào và phù hợp cho dự báo ngắn hạn [60].

+) Kỹ thuật dự báo dựa trên chuỗi thời gian

Chuỗi thời gian gồm các quan sát định lượng được ghi nhận tại các khoảng thời gian đều đặn (giờ, ngày, tháng...) [61]. Dựa trên chuỗi này, mô hình dự báo giá trị tương lai gần bằng cách đánh giá mối quan hệ nội tại trong quá khứ. Cornaro cùng các cộng sự [62] đã sử dụng phương pháp thống kê để thiết lập mối tương quan giữa dữ liệu quá khứ khí tượng và giờ chiếu xạ mặt trời. Các kỹ thuật được thiết lập là: cấp số nhân trung bình (EWMA), trung bình di chuyển tự động (ARMA) và trung bình trượt tích hợp tự hồi quy (ARIMA).

+) Phương pháp dịch chuyển theo cấp số nhân trung bình (EWMA)

Phương pháp dịch chuyển theo cấp số nhân trung bình (EWMA) là một kỹ thuật thống kê áp dụng trong dự báo chuỗi thời gian, trong đó các quan sát gần hiện tại được gán trọng số cao hơn so với các quan sát xa hơn trong quá khứ. Phương pháp này giúp phản ánh nhanh các biến động gần nhất của dữ liệu mà vẫn duy trì sự ổn định tổng thể. Kỹ thuật EWMA được bắt đầu từ nghiên cứu của Brown và cộng sự [63] sau đó được phát triển mở rộng bởi Holt (1957) và Winter (1960) từ đó hoàn thiện thành phương pháp Holt-Winter [64] có phương trình như sau:

$$F_{t+1} = F_t + \alpha(Y_t - F_t) \quad (2.5)$$

Trong đó: Y_t là giá trị quan sát trong hiện tại, F_{t+1} là giá trị dự báo, F_t là giá trị dự báo trước đó, α là hằng số làm mịn duy trì trong khoảng (0,1)

Từ đó, phương trình dự báo xuất giá trị dự báo tại $t+1$ bằng tổng của giá trị dự báo cuối cùng và có hệ số điều chỉnh dự báo là: $\alpha \times (Y_t - F_t)$.

+) Mô hình trung bình trượt kết hợp tự hồi quy (ARIMA)

Có hai biến thể của mô hình ARIMA: Không theo mùa và theo mùa. Nếu trong dữ liệu chuỗi thời gian có tính thời vụ thì mô hình ARIMA theo mùa được sử dụng. Mặt khác mô hình ARIMA không theo mùa được sử dụng cho các trường hợp chung. ARIMA không theo mùa được mô hình hóa theo cách sau:

$$y'_t = \mu + \phi_1 y_{t-1} + \dots + \phi_p y_{t-p} - \theta_1 e_{t-1} - \dots - \theta_q e_{t-q} \quad (2.6)$$

Trong đó: y'_t là biến thứ d^{th} của chuỗi thời gian không cố định Y ; p , d và q tương ứng là độ trễ tự động hồi phục, độ trễ sai lệch và độ trễ trung bình động, ϕ và θ là các tham số tự động hồi phục và trung bình động, μ là hằng số.

Tùy thuộc vào các giá trị của p , d và q , một quy trình ARIMA có thể thực hiện hình thức trung bình trượt (MA) hoặc tự hồi quy (AR) hoặc trung bình trượt tự động (ARMA).

Một mô hình kết hợp bao gồm cả mô hình không theo mùa và theo mùa, được sử dụng để đại diện cho mô hình ARIMA theo mùa. ARIMA theo mùa thường được thể hiện theo cách sau:

$$\text{ARIMA}(p; d; q) \times (P; D; Q)_s \quad (2.7)$$

Với: P là thứ tự AR theo mùa, Q là thứ tự MA theo mùa, D là thứ tự khác biệt theo mùa, s là khoảng thời gian của mẫu theo mùa.

Một mô hình rõ hơn của mô hình ARIMA theo mùa là:

$$(1 - \phi_1 B - \dots - \phi_p B^p)(1 - \Phi_1 B^s - \dots - \Phi_p B^{ps})(1 - B)^d \quad (2.8)$$

$$(1 - B^s)^D y_t = (1 + \theta_1 B + \dots + \theta_1 B^q)(1 + \lambda_1 B^s + \lambda_Q B^{qs})_{et}$$

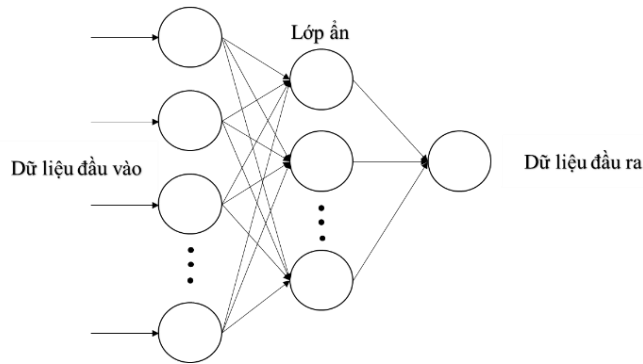
Trong đó, B là toán tử trễ được định nghĩa theo công thức: $B^m y_t = y_{t-m}$ nghĩa là dịch chuỗi dữ liệu y_t về phía sau n bước thời gian.

2.3.1.2. Mô hình học máy (Machine Learning)

Hiện nay, các mô hình dự báo công suất phát điện mặt trời đã khai thác những tiến bộ trong học máy, đây là một cách tiếp cận dựa trên trí tuệ nhân tạo (AI). Phương pháp này dựa vào khả năng AI học hỏi kinh nghiệm với dữ liệu lịch sử và tiếp tục trau dồi khả năng dự báo của nó thông qua việc chạy các chương trình. Các máy tính cấu hình cao được yêu cầu chạy nhiều lần lặp trước khi có thể đưa dự báo cuối cùng. Nó có khả năng tự học hỏi theo những thuật toán được định hình cấu trúc với điều kiện dừng học cho trước. Các ứng dụng của nó rất nhiều như nhận dạng mẫu, khai thác dữ liệu, các vấn đề phân loại, lọc và dự báo... Các kỹ thuật chính của học máy là mạng nơron nhân tạo (ANN), mạng nơron tri thức nhiều lớp (MLPNN), cây quyết định (Decision Tree).

a. Mạng nơron nhân tạo (ANN)

ANN áp dụng cơ chế xử lý thông tin của bộ não con người [61]. Nó có một khả năng độc đáo để các hàm phi tuyến có độ chính xác cao, nó được sử dụng trong các lĩnh vực đa dạng như dự báo khí tượng, tài chính, vật lý, kỹ thuật, y học.... Hình 2.2 là một mô hình cơ bản của mạng. Kiến trúc cơ bản của mạng nơron nhân tạo (ANN) bao gồm ba thành phần chính: lớp đầu vào, lớp ẩn và lớp đầu ra. Mỗi lớp được cấu thành từ các nơron nhân tạo, chúng được kết nối với nhau thông qua các trọng số. Trong mạng ANN, mỗi nơron được xem như một nút kích hoạt nơi diễn ra quá trình xử lý thông tin và ra quyết định. Một nơron tại lớp bất kỳ sẽ nhận tín hiệu đầu vào từ các nơron của lớp trước đó, sau đó kết hợp các đầu vào này theo các trọng số đã học để tạo thành tổng trọng số. Kết quả này được đưa qua một hàm kích hoạt nhằm xác định đầu ra của nơron. Thông tin đầu ra sẽ tiếp tục được truyền đến lớp kế tiếp, cho đến khi quá trình lan truyền hoàn tất và cho ra kết quả dự báo cuối cùng tại lớp đầu ra. Các hàm kích hoạt phổ biến trong ANN bao gồm: Sigmoid, tanh (Sigmoid tiếp tuyến hyperbolic), hàm cơ sở Gaussian, các hàm bước tuyến tính đơn cực và lưỡng cực, cũng như các hàm tuyến tính đơn cực và lưỡng cực. Các kỹ thuật dự báo sử dụng ANN đã được phát triển và điều chỉnh linh hoạt để phù hợp với các bài toán có cấu trúc đầu vào và đầu ra khác nhau, từ đơn biến đến đa biến và từ tuyến tính đến phi tuyến.

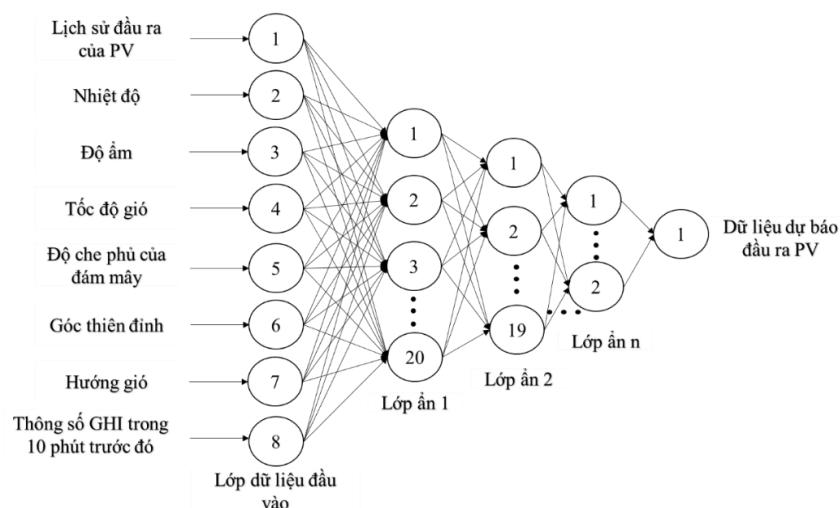


Hình 2.2. Cấu trúc cơ bản của ANN

(Nguồn: Tài liệu [61])

b. Mạng nơ-ron nhiều lớp (MLPNN)

Mô hình MLPNN là một kỹ thuật cơ bản và hiệu quả trong các bài toán dự báo và mô hình hóa phi tuyến [65]. Nhờ cấu trúc linh hoạt và khả năng học sâu, MLPNN có thể xử lý các vấn đề phức tạp mà các mạng nơ-ron đơn tầng thông thường không thể giải quyết hiệu quả [66]. Về cấu trúc, MLPNN bao gồm ba lớp trở lên, trong đó mỗi lớp bao gồm các nơ-ron phi tuyến được kích hoạt độc lập. Mỗi nút ở một lớp nhất định được kết nối với một số nút trong lớp tiếp theo, tạo thành một mạng liên kết chặt chẽ giữa đầu vào và đầu ra. Nhờ đó, MLPNN có khả năng học và khái quát mối quan hệ phi tuyến giữa các biến đầu vào và kết quả đầu ra thông qua cơ chế huấn luyện bằng dữ liệu. Mối tương quan giữa số lượng nút và số lớp ẩn là rất cần thiết [67], có những khó khăn xảy ra khi tăng số lượng nút [68]. Hình 2.3 cho thấy cấu trúc đơn giản hóa của một MLPNN với tám đầu vào để dự báo năng lượng mặt trời.

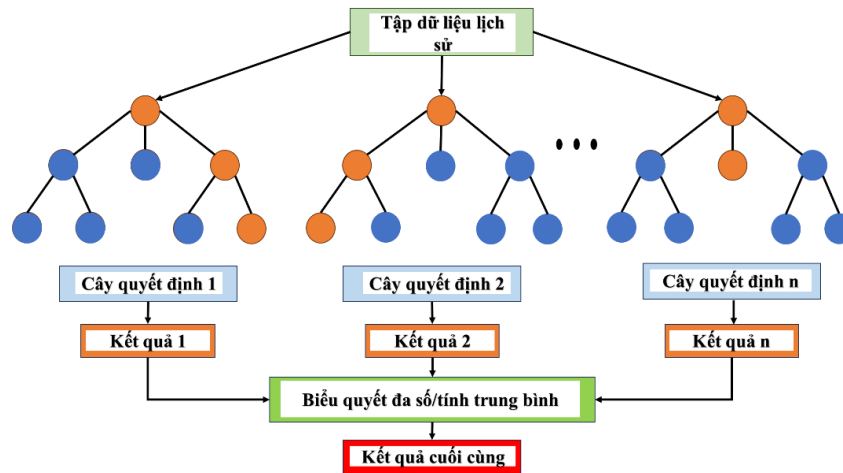


Hình 2.3. Mạng nơ-ron nhiều lớp (MLPNN)

(Nguồn: Tài liệu [67])

c. Mô hình cây quyết định (Decision Tree)

Cây quyết định là một mô hình học máy phi tham số được sử dụng rộng rãi cho cả bài toán phân loại và hồi quy. Nó có cấu trúc giống như một cây, trong đó mỗi nút trong biểu diễn một quyết định dựa trên một đặc trưng, mỗi nhánh thể hiện kết quả của quyết định đó, và mỗi lá nút đại diện cho một dự đoán.



Hình 2.4: Cấu trúc của mô hình cây quyết định

(Nguồn: Tài liệu [69])

Dựa trên ý tưởng của cây quyết định, một số thuật toán ensemble learning đã được phát triển và cho thấy hiệu quả cao trong nhiều bài toán dự báo, trong đó có dự báo công suất phát điện mặt trời.

Ưu nhược điểm của các mô hình cây quyết định so với các mô hình ANN truyền thống:

Ưu điểm:

- Các mô hình này đều cho kết quả dự báo với độ chính xác cao, thể hiện qua các chỉ số như RMSE, MAE, MAPE thấp hơn so với các mô hình truyền thống khác như ANN, SVR [34], [35], [70]–[74].
- Khả năng xử lý dữ liệu nhiễu và tránh hiện tượng overfitting (quá khớp) tốt hơn so với mạng nơ-ron [20].
- Cấu trúc đơn giản, dễ hiểu và dễ triển khai, thời gian xử lý nhanh [73].
- Linh hoạt trong cấu trúc và xử lý đa biến [74].
- Khả năng dự báo tức thời, ít bị phụ thuộc vào dữ liệu quá khứ.

Nhược điểm:

- Kết quả dự báo phụ thuộc nhiều vào quá trình tiền xử lý các đặc trưng và tối ưu tham số của mô hình [35], [72].

- Có nguy cơ mắc kẹt vào điểm cực tiểu cục bộ làm giảm khả năng tổng quát hóa và dự báo chính xác trên dữ liệu mới khi cây quyết định được xây dựng quá sâu và phức tạp [74].
- Với điều kiện khí hậu bất thường, các mô hình này có thể gặp khó khăn dẫn đến sai số lớn hơn [20].

Các mô hình cây quyết định tiêu biểu

a. Random Forest (RF):

Random Forest là một thuật toán học máy dựa trên cấu trúc cây quyết định (decision tree) được đề xuất bởi Leo Breiman vào năm 2001 [75]. 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, do đó rất phù hợp cho các bài toán dự báo phức tạp như dự báo công suất phát điện mặt trời. Quá trình huấn luyện một mô hình Random Forest bao gồm các bước sau:

Bước 1: Tạo tập dữ liệu con bằng cách lấy mẫu có hoàn lại (Bootstrap Sampling)

- Từ tập dữ liệu gốc, Random Forest sẽ tạo ra nhiều tập dữ liệu con (subsets) bằng cách lấy mẫu có hoàn lại.
- Mỗi tập dữ liệu con này sẽ được sử dụng để huấn luyện một cây quyết định riêng biệt.

Bước 2: Xây dựng cây quyết định từng cây một cách ngẫu nhiên

- Tại mỗi nút của cây quyết định, thay vì xem xét tất cả các biến đầu vào, Random Forest chỉ chọn ngẫu nhiên một tập con các biến đầu vào.
- Từ tập con này, thuật toán sẽ tìm biến và điều kiện phân chia tối ưu để tạo nút con mới.
- Quá trình này được lặp lại cho đến khi cây quyết định đạt được điều kiện dừng (ví dụ: độ sâu tối đa hoặc số lượng mẫu tối thiểu tại nút lá).

Bước 3: Dự báo bằng cách kết hợp các cây quyết định

- Sau khi huấn luyện xong tất cả các cây quyết định, mô hình Random Forest sẽ sử dụng tất cả chúng để dự báo kết quả mới.
- Trong bài toán phân loại, dự báo cuối cùng là nhãn lớp được đa số các cây bầu chọn.

- Trong bài toán hồi quy (như dự báo công suất phát điện mặt trời), dự đoán cuối cùng là trung bình của các giá trị dự đoán từ tất cả các cây.

Bằng cách sử dụng nhiều cây quyết định được huấn luyện trên các tập dữ liệu con khác nhau, Random Forest giảm thiểu được hiện tượng overfitting và cải thiện đáng kể tính chính xác và đồng nhất của mô hình. Các cây quyết định trong Random Forest hoạt động độc lập và đa dạng, giúp mô hình có khả năng tổng quát hóa tốt hơn so với một cây quyết định đơn lẻ.

b. Mô hình XGBoost

XGBoost là một thuật toán dựa trên cây quyết định và sử dụng phương pháp Boosting, được đề xuất bởi Tianqi Chen vào năm 2016 [76]. Boosting là phương pháp thay vì cố gắng xây dựng một mô hình tốt duy nhất, nó sẽ tạo ra một loạt các mô hình yếu, học bổ sung lẫn nhau [77]. Nói cách khác, trong Boosting, các mô hình sau sẽ cố gắng học để hạn chế lỗi lầm của các model trước. Tuy nhiên, phương pháp boosting có nhược điểm là tốn thời gian và có nguy cơ overfitting.

Thuật toán XGBoost đã phát triển hiệu suất của mô hình Boosting thông thường bằng cách sử dụng tính năng thu nhỏ theo tỷ lệ của các nút lá (cắt tỉa cây), song song hóa và các thuật ngữ regularization. Giá trị dự báo của mô hình XGBoost được tính bằng Công thức (1) sau [76].

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^K f_k(x_i), \quad f_k \in \mathcal{F} \quad (2.9)$$

Trong đó: $\hat{y}_i$ là giá trị công suất dự báo, $f_k()$ là mô hình cây thứ k trong tổ hợp các cây quyết định, x_i là đặc trưng đầu vào tại thời điểm t, K là số lượng cây trong tổ hợp, $\mathcal{F}$ là không gian chức năng (functional space) chứa tất cả các cây trong tổ hợp. Công thức này thể hiện cách giá trị dự đoán của mô hình XGBoost là tổng của các đầu ra của từng cây quyết định, mỗi cây đóng góp vào giá trị dự báo cuối cùng. Hàm mục tiêu trong bộ hồi quy XGBoost bao gồm thuật ngữ chính quy hóa và được xác định bởi Phương trình (2) [76], như sau:

$$\mathcal{L}(\phi) = \sum_{t=1}^n l(y_t, \hat{y}_t) + \sum_{k=1}^K \Omega(f_k) \quad (2.10)$$

Trong đó: $\mathcal{L}(\phi)$ là hàm mục tiêu cần được tối ưu hóa, n là chiều dài của chuỗi dữ liệu thời gian, y_t là giá trị thực tế tại thời điểm t, $\hat{y}_t$ là giá trị dự báo tại thời điểm t, K là số lượng lá của mỗi cây, f_k là cây quyết định thứ k trong tổ hợp các cây quyết định,

$l(y_i, \hat{y}_i)$ là hàm mất mát đo lường sự chênh lệch giữa giá trị thực tế và giá trị dự báo, $\Omega(f_k)$ là hàm kiểm soát regularization của cây quyết định thứ k. Công thức này thể hiện rằng mục tiêu của mô hình XGBoost là cực tiểu hóa tổng mất mát trên tập huấn luyện và tổng regularization của các cây quyết định. Điều này giúp kiểm soát overfitting và tối ưu hóa mô hình để dự đoán chính xác trên dữ liệu mới. Hàm regularization Ω trong mô hình XGBoost được biểu diễn bởi công thức sau [76]:

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \|\omega\|^2 \quad (2.11)$$

Trong đó: f là cây quyết định cần được kiểm soát, γ là tham số kiểm soát tổng số lượng lá (T) của cây, λ là tham số kiểm soát độ lớn của các nút lá cây, T là tổng số lượng lá cây trong cây quyết định f , ω là giá trị của nút lá trong cây. Hàm regularization $\Omega(f)$ có hai thành phần chính. Phần γT kiểm soát tổng số lượng lá của cây và phần $\frac{1}{2} \lambda \|\omega\|^2$ kiểm soát độ lớn của các lá cây thông qua việc cộng tổng bình phương của giá trị của các lá ω nhân với tham số regularization λ . Mục tiêu của regularization là ngăn chặn mô hình quá mức phức tạp và giúp mô hình tổ hợp trở nên ổn định và khả năng tổng quát hóa tốt hơn trên dữ liệu mới. Các tham số γ và λ thường được điều chỉnh để đạt được sự cân bằng giữa hiệu suất trên tập huấn luyện và khả năng tổng quát hóa của mô hình.

c. Mô hình LightGBM:

LightGBM (light gradient boosting machine) là một thuật toán Gradient Boosting Decision Tree (GBDT) được nhóm nghiên cứu của Guolin Ke tại Microsoft giới thiệu vào năm 2017 [78]. LightGBM được thiết kế nhằm tối ưu hóa hiệu suất huấn luyện và sử dụng bộ nhớ trong quá trình xây dựng mô hình cây tăng cường. Hai cải tiến kỹ thuật then chốt giúp LightGBM đạt được hiệu quả vượt trội là GOSS (Gradient-based One-Side Sampling) và EFB (Exclusive Feature Bundling). Các kỹ thuật này cho phép thuật toán tăng tốc đáng kể quá trình huấn luyện mà vẫn duy trì độ chính xác cao, đặc biệt khi xử lý các tập dữ liệu có kích thước lớn và phân bố rời rạc về đặc trưng [78].

GOSS [78] giữ lại toàn bộ các điểm dữ liệu có gradient lớn (đóng góp cao cho cập nhật mô hình), và chỉ lấy mẫu ngẫu nhiên từ phần còn lại. Để bù lại việc mất cân bằng dữ liệu, các điểm có gradient nhỏ được gán trọng số nhân $\frac{1-a}{b}$ khi tính toán mức tăng thông tin. Bằng cách đó, chúng tập trung nhiều hơn vào các phiên bản chưa được đào tạo mà không làm thay đổi nhiều về phân phối dữ liệu gốc.

Cho O là bộ dữ liệu huấn luyện trên một nút cố định của cây quyết định. Độ chệch phương sai của việc chia đặc trưng j tại điểm d cho nút này được định nghĩa là [78]:

$$V_{j|O}(d) = \frac{1}{n_O} \left(\frac{\left(\sum_{\{x_i \in O: x_{ij} \leq d\}} g_i \right)^2}{n_{l|O}^j(d)} + \frac{\left(\sum_{\{x_i \in O: x_{ij} > d\}} g_i \right)^2}{n_{r|O}^j(d)} \right) \quad (2.12)$$

Trong đó: $n_O = \sum I[x_i \in O]$, $n_l^j(d) = \sum I[x_i \in O: x_{ij} \leq d]$ và $n_{r|O}^j(d) = \sum I[x_i \in O: x_{ij} > d]$. n_O là số lượng trường hợp trong tập dữ liệu huấn luyện tại nút O . $I[\dots]$ là hàm chỉ số (indicator function) đánh dấu xác định điều kiện. x_{ij} là phần tử thứ j của vector đặc trưng x_i . g_i là đạo hàm âm của hàm mất mát đối với đầu ra của mô hình tại điểm x_i . $n_{jO}(d)$ là số lượng trường hợp trong tập O mà thỏa mãn $x_{ij} \leq d$. $n_{jro}(d)$ là số lượng trường hợp trong tập O mà thỏa mãn $x_{ij} > d$. Công thức này giúp đo lường sự gia tăng của phương sai sau khi chia tại một nút, giúp GBDT chọn ra sự chia tốt nhất dựa trên việc giảm thiểu phương sai.

Trong thuật toán cây quyết định đối với đặc trưng j , thuật toán chọn $d_j^* = \text{argmax}_j V_j(d)$ và tính toán độ chệch lớn nhất $V_j(d_j^*)$. Sau đó, dữ liệu được chia theo đặc trưng j^* tại điểm d_j thành hai nút con trái và phải.

Trong phương pháp GOSS được đề xuất, đầu tiên thuật toán sắp xếp các trường hợp huấn luyện theo giá trị tuyệt đối của độ dốc của chúng theo thứ tự giảm dần. tiếp theo, thuật toán giữ lại $a \times 100\%$ trường hợp với độ dốc lớn hơn và tạo ra một tập con A . Sau đó, đối với tập còn lại A^c bao gồm $(1 - a) \times 100\%$ trường hợp với độ dốc nhỏ hơn, tiếp tục chọn một tập con B với kích thước $b \times |A^c|$. Cuối cùng, thuật toán chia các trường hợp dựa trên độ chệch phương sai ước lượng $\tilde{V}_j(d)$ trên tập hợp $A \cup B$ tức là:

$$\tilde{V}_j(d) = \frac{1}{n} \left(\frac{\left(\sum_{x_i \in A_l} g_i + \frac{1-a}{b} \sum_{x_i \in B_l} g_i \right)^2}{n_l^j(d)} + \frac{\left(\sum_{x_i \in A_r} g_i + \frac{1-a}{b} \sum_{x_i \in B_r} g_i \right)^2}{n_r^j(d)} \right) \quad (2.13)$$

Trong đó: $A_l = \{x_i \in A: x_{ij} \leq d\}$, $A_r = \{x_i \in A: x_{ij} > d\}$, $B_l = \{x_i \in B: x_{ij} \leq d\}$, $B_r = \{x_i \in B: x_{ij} > d\}$ và hệ số $\frac{1-a}{b}$ được sử dụng để chuẩn hóa tổng các gradient trên B trở lại kích thước của A^c . Do đó, trong GOSS, chúng ta sử dụng $\tilde{V}_j(d)$ ước lượng

trên một tập con nhỏ hơn các mẫu thay vì $V_j(d)$ chính xác trên toàn bộ tập dữ liệu để xác định điểm chia, từ đó giảm đáng kể chi phí tính toán.

Quan trọng hơn, định lý sau đây chỉ ra rằng GOSS sẽ không mất nhiều độ chính xác trong quá trình huấn luyện và sẽ vượt trội hơn so với việc lấy mẫu ngẫu nhiên.

Định lý: Sai số xấp xỉ $\varepsilon(d) = |\tilde{V}_j(d) - V_j(d)|$ và $\bar{g}_l^j(d) = \frac{\sum_{x_i \in (A \cup A^c)_l} |g_i|}{n_l^j(d)}$,

$$\bar{g}_r^j(d) = \frac{\sum_{x_i \in (A \cup A^c)_r} |g_i|}{n_r^j(d)}$$

Với xác suất ít nhất là $1 - \delta$, ta có:

$$\varepsilon(d) \leq C_{a,b}^2 \ln 1/\delta \cdot \max \left\{ \frac{1}{n_l^j(d)}, \frac{1}{n_r^j(d)} \right\} + 2DC_{a,b} \sqrt{\frac{\ln 1/\delta}{n}} \quad (2.14)$$

Trong đó $C_{a,b} = \frac{1-a}{\sqrt{b}} \max_{x_i \in A^c} |g_i|$, và $D = \max(\bar{g}_l^j(d), \bar{g}_r^j(d))$

Từ định lý trên ta có tỉ lệ xấp xỉ theo hình thức dạng xấp xỉ của GOSS được biểu diễn là $O\left(\frac{1}{n_l^j(d)} + \frac{1}{n_r^j(d)} + \frac{1}{\sqrt{n}}\right)$ trong đó $n_l^j(d)$ và $n_r^j(d)$ đại diện cho số lượng mẫu trong các nút trái và phải sau khi chia, và p là tổng số mẫu. Thảo luận nhấn mạnh rằng khi tỷ số không quá mất cân bằng ($n_l^j(d) \geq O\sqrt{n}, n_r^j(d) \geq O\sqrt{n}$) thì sai số xấp xỉ sẽ chủ yếu bị ảnh hưởng bởi mệnh đề thứ hai trong bất đẳng thức, và nó có xu hướng tiến về 0 khi n tiến về vô cùng. Điều này gợi ý rằng sai số xấp xỉ là khá chính xác khi xử lý một lượng lớn dữ liệu. Lấy mẫu ngẫu nhiên được xem xét như một trường hợp đặc biệt của GOSS với $a=0$. Trong nhiều trường hợp GOSS có thể vượt trội so với việc lấy mẫu ngẫu nhiên, điều này xảy ra dưới điều kiện $C_{0,\beta} > C_{\alpha,\beta-\alpha}$, tương đương với $\frac{\alpha a}{\sqrt{\beta}} > \frac{1-a}{\sqrt{\beta-a}}$ với $a_\alpha = \max_{x_i \in A \cup A^c} |g_i| / \max_{x_i \in A^c} |g_i|$.

Hiệu suất của GOSS được đánh giá dựa trên sai số tổng quát $E_{GEN}^{GOSS}(d) = |\tilde{V}_j(d) - V_j(d)|$. Sai số tổng quát được giới hạn bởi tổng của sai số xấp xỉ và sai số tổng quát thực tế $E_{GEN}^{GOSS}(d) \leq |\tilde{V}_j(d) - V_j(d)| + |V_j(d) - V_*(d)| \triangleq E_{GOSS}(d) + E_{GEN}(d)$.

Phương pháp Exclusive Feature Bundling (EFB) [78] được đề xuất nhằm giảm độ phức tạp tính toán khi huấn luyện mô hình Gradient Boosting Decision Trees (GBDT) bằng cách đóng gói các đặc trưng loại trừ lẫn nhau vào các bundle. Cách EFB thực hiện: thứ nhất xây dựng đồ thị xung đột giữa các đặc trưng. Thứ hai, áp dụng thuật toán tham

lam để phân nhóm vào các bundle. Cuối cùng gán offset khác nhau cho các đặc trưng trong cùng bundle để không trùng lặp giá trị.

Bằng cách đóng gói các đặc trưng loại trừ lẫn nhau, EFB giúp giảm số đặc trưng mà không làm mất thông tin ban đầu, tối ưu cả tốc độ và độ chính xác.

2.3.1.3. Các mô hình sử dụng học máy sâu (Deep Learning)

Hiện nay, các kỹ thuật học máy tiên tiến như học sâu đã đạt nhiều thành tựu nổi bật trong xử lý dữ liệu lớn, nhờ khả năng tự động học từ dữ liệu thông qua mạng nơron nhiều lớp (Deep Neural Networks – DNN). Các mô hình DNN có khả năng khai thác cấu trúc phi tuyến phức tạp của dữ liệu, sử dụng các thuật toán tối ưu hóa hiện đại và nhiều tầng ẩn để trích xuất đặc trưng hiệu quả. DL bao gồm nhiều hướng tiếp cận như Restricted Boltzmann Machines (RBM), mạng tin cậy sâu (DBN), bộ mã hóa tự động (Autoencoder), và đặc biệt là mạng nơron tích chập sâu (DCNN) vốn rất phổ biến trong xử lý ảnh, nhận diện khuôn mặt, giọng nói, hành vi, và phân loại hình ảnh. DCNN vận hành thông qua các khối tích chập, kích hoạt và gộp, giúp giảm số tham số huấn luyện nhờ chia sẻ trọng số và khai thác hiệu quả đặc trưng cục bộ [79].

Tuy nhiên, do đặc trưng thiết kế hướng đến dữ liệu hình ảnh có cấu trúc lưới, DCNN không thường được áp dụng trực tiếp trong phân loại thời tiết hay dự báo công suất phát điện mặt trời [80]. Thay vào đó, các mô hình học sâu có khả năng xử lý chuỗi thời gian như RNN, LSTM hoặc BiGRU thường phù hợp hơn trong các bài toán dự báo năng lượng tái tạo.

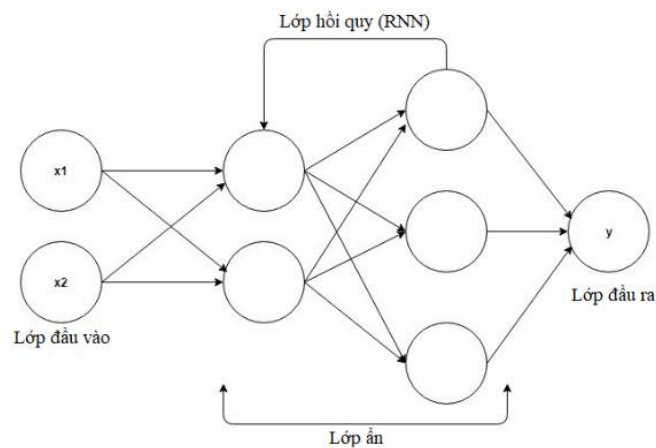
a. Mạng nơron hồi tiếp (RNN)

Trong những năm gần đây, với sự phát triển nhanh chóng của các thuật toán trí tuệ nhân tạo, các thuật toán học sâu đã phá vỡ những hạn chế của các mô hình học máy nông. Mô hình học sâu có nhiều ưu điểm như tự học các tính năng và biểu diễn chúng theo cách phân cấp ở các cấp độ khác nhau, học sâu thích hợp xử lý lượng dữ liệu lớn hơn và phức tạp hơn mô hình học máy truyền thống [81]. Các mô hình học sâu được phổ biến là mạng nơron tích chập (CNN) [27], [82] và gần đây nhất mạng nơron hồi tiếp (Recurrent Neural Network – RNN), trong nhiều tài liệu tại Việt Nam còn gọi là mạng nơron hồi quy [26], [83], [84]. Chúng đã được sử dụng rộng rãi trong lĩnh vực dự báo công suất PV ngắn hạn. RNN chủ yếu được sử dụng để xử lý dữ liệu thông tin dạng chuỗi, tuy nhiên có một thách thức trong việc huấn luyện RNN đó là vanishing gradient và exploding gradient [85].

RNN là một mạng ANN nổi bật có thể học tập và xử lý nhiều mối quan hệ phức tạp khác nhau. Phần cốt lõi của RNN chính là một vòng hồi tiếp (feedback loop), nơi kết quả của bước tính toán trước (thường là trạng thái ẩn) được đưa ngược lại làm một phần đầu vào cho bước tính toán hiện tại. Cơ chế hồi tiếp thông tin này là nền tảng cho khả năng 'ghi nhớ' và xử lý hiệu quả các loại dữ liệu có tính chuỗi (sequential data) của mạng. Về cơ bản ứng dụng mô hình RNN rất linh hoạt từ nhận dạng giọng nói đến xe không người lái...[15]. Mô hình RNN được làm nổi bật bởi nghiên cứu của Elman và các cộng sự [86], họ đã xây dựng mô hình RNN có một vòng phản hồi có thể giao tiếp với lớp ẩn của mạng và lớp đầu vào. Trong RNN, mỗi nơ-ron hoạt động được kết nối với tất cả các nơ-ron xử lý khác và với chính nó [15]. Vì thế, kết quả đầu ra của RNN nghiêng về tín hiệu phản hồi ở bước thời gian trước đó và tín hiệu đầu vào. Hàm kích hoạt của RNN là tổng trọng số của tín hiệu đầu vào và tín hiệu phản hồi. Do đó, phương trình hàm kích hoạt là [87]:

$$S_k(t) = \sum_{i \in U} W_{ki} y_i(t) + \sum_{i \in I} W_{ki} x_i(t) = \sum_{i \in U \cup I} W_{ki} z_i(t) \quad (2.15)$$

Trong đó: $S_k(t)$ là tổng đầu vào thuần (net input) của nơ-ron k tại thời điểm t , U là tập các đơn vị ẩn và đầu ra (hidden + output units), i là tập các đầu vào (input units), W_{ki} là trọng số kết nối từ $i \rightarrow k$, $y_i(t)$ là đầu ra của đơn vị i ở t (nếu $i \in U$), $x_i(t)$ là đầu ra của đơn vị i ở t (nếu $i \in I$), $z_i(t)$ là tổng hợp cả $y_i(t)$ và $x_i(t)$.



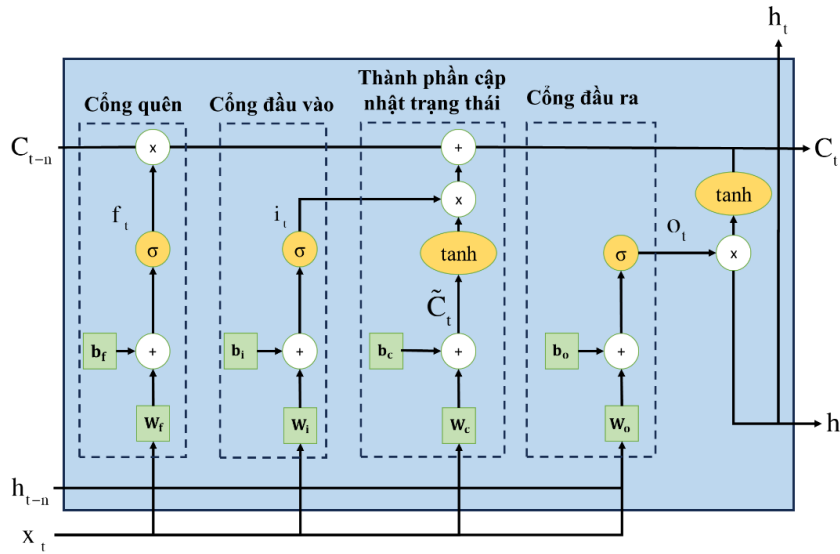
Hình 2.5. Cấu trúc cơ bản của RNN

(Nguồn: Tài liệu [15])

b. Mô hình dự báo LSTM

Vào năm 1997, Hochreiter và các cộng sự [88] đã đề xuất mô hình mạng LSTM. LSTM là một cấu trúc mở rộng của RNN, LSTM được thiết kế để giải quyết các bài

toán về phụ thuộc xa (long-term dependencies) trong mạng RNN do bị ảnh hưởng của vanishing gradient.



Hình 2.6. Cấu trúc của mô hình LSTM

(Nguồn: Tài liệu [89])

Trọng tâm của kiến trúc LSTM là các cổng, đóng vai trò quan trọng trong việc kiểm soát luồng thông tin. Có ba cổng chính trong LSTM: cổng forget, cổng input và cổng output. Các cổng này được xác định bởi các công thức sau [90]:

+) **Cổng quên (forget gate) f_t** : Cổng forget đóng vai trò như một bộ lọc, quyết định những thông tin nào từ trạng thái tế bào trước đó cần được giữ lại hoặc loại bỏ. Nó hoạt động bằng cách tính toán một vector có giá trị từ 0 đến 1 cho mỗi số trong trạng thái tế bào. Giá trị gần 1 chỉ ra rằng thông tin đó nên được giữ lại, trong khi giá trị gần 0 nghĩa là thông tin đó có thể bị quên đi. Cơ chế này giúp mạng tránh được tình trạng quá tải thông tin và tập trung vào những dữ liệu thực sự quan trọng.

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \quad (2.16)$$

+) **Cổng vào (input gate) i_t** : Cổng input, như tên gọi của nó, quyết định những thông tin mới nào sẽ được lưu trữ vào trạng thái tế bào. Nó hoạt động thông qua hai bước: đầu tiên, một lớp sigmoid quyết định những giá trị nào sẽ được cập nhật; sau đó, một lớp tanh tạo ra một vector các giá trị mới có thể được thêm vào trạng thái. Hai phần này được kết hợp để tạo ra một bản cập nhật cho trạng thái tế bào. Cơ chế này cho phép LSTM liên tục cập nhật kiến thức của nó, thích ứng với thông tin mới mà nó nhận được.

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \quad (2.17)$$

+) Thành phần cập nhật trạng thái C_t : Đây là giá trị mới được tính toán dựa trên thông tin hiện tại và các trạng thái trước đó, và nó sẽ đóng góp vào việc cập nhật trạng thái ô hiện tại C_t sau khi đi qua các cổng của LSTM.

$$\begin{aligned}\tilde{C}_t &= \tanh(W_c[h_{t-1}, x_t] + b_c) \\ C_t &= f_t \odot C_{t-1} \oplus i_t \odot \tilde{C}_t\end{aligned}\tag{2.18}$$

+) Cổng ra (output gate) h_t : Cuối cùng, cổng output quyết định phần nào của trạng thái tế bào hiện tại sẽ được xuất ra làm đầu ra. Nó sử dụng một lớp sigmoid để quyết định những phần nào của trạng thái tế bào sẽ được xuất ra, sau đó nhân kết quả này với một phiên bản đã được xử lý bởi hàm kích hoạt tanh của trạng thái tế bào. Điều này cho phép mạng kiểm soát chặt chẽ thông tin nào được truyền đến trạng thái ẩn tiếp theo và cuối cùng là đầu ra của mạng.

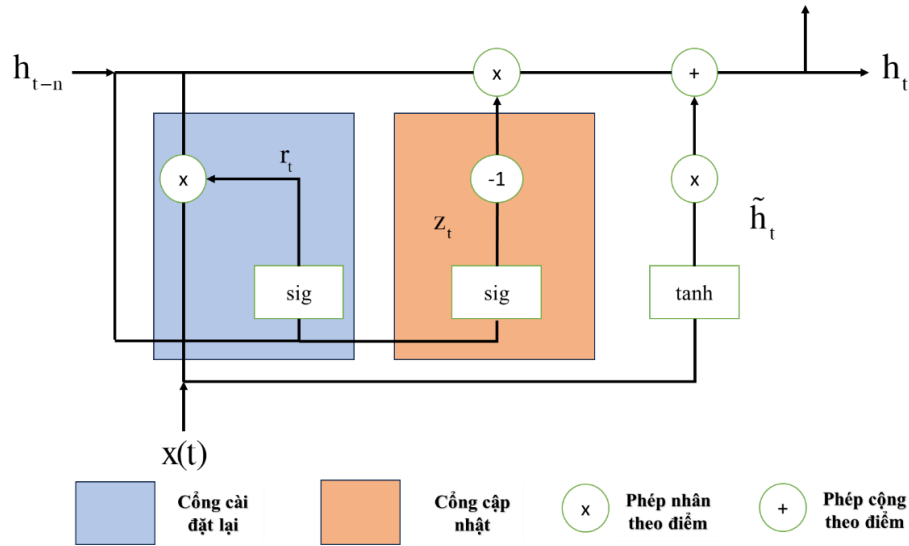
$$\begin{aligned}o_t &= \sigma(W_o[h_{t-1}, x_t] + b_o) \\ h_t &= o_t \odot \tanh(C_t)\end{aligned}\tag{2.19}$$

Trong các công thức (14), (15), (16), (17) trên: σ là hàm sigmoid, $\tanh$ là hàm hyperbolic tangent, W_f , W_i , W_c , W_o và b_f , b_i , b_c , b_o là các tham số của mô hình, x_t là đầu vào tại thời điểm t , h_{t-1} là trạng thái ẩn từ thời điểm trước đó, C_t là trạng thái của tế bào nơ-ron tại thời điểm t , $\tilde{C}_t$ là giá trị cập nhật trạng thái của tế bào nơ-ron, $\oplus$ là phép cộng hadamard, $\odot$ là phép nhân hadamard.

Sự kết hợp của ba cổng này tạo nên sức mạnh của LSTM. Chúng cho phép mạng duy trì thông tin quan trọng trong thời gian dài, đồng thời vẫn có khả năng cập nhật và thích ứng với thông tin mới. Cơ chế này giúp LSTM đặc biệt hiệu quả trong việc xử lý các chuỗi dài, nơi mà các mối quan hệ giữa các phần tử có thể kéo dài qua nhiều bước thời gian. Nhờ đó, LSTM đã trở thành một công cụ mạnh mẽ trong nhiều ứng dụng như dịch máy, nhận dạng giọng nói, và dự báo chuỗi thời gian.

c. Mô hình dự báo GRU

Để cải tiến mạng RNN đơn giản, năm 2014, Cho và các cộng sự [91] đã đề xuất kiến trúc mạng mới được gọi là GRU (gate recurrent unit). Nhóm nghiên cứu đầu tiên về GRU này đã chỉ ra ưu điểm của GRU so với LSTM là có cấu trúc đơn giản hơn, tham số huấn luyện ít hơn [92].



Hình 2.7. Cấu trúc của mô hình GRU

(Nguồn: Tài liệu [89])

+) Cổng cập nhật (update gate) z_j [91]: Vai trò của cổng này là quyết định bao nhiêu thông tin từ trạng thái ẩn trước đó h_{t-1} sẽ được mang vào trạng thái ẩn hiện tại h_t . Cổng này giúp mô hình quyết định có nên lưu giữ thông tin từ quá khứ hay không. Công thức tính:

$$z_j = \sigma \left([W_z X]_j + [U_z h_{(t-1)}]_j \right) \quad (2.20)$$

trong đó σ là hàm sigmoid logistic và $[\cdot]_j$ biểu thị phần tử thứ j của một vector. x và $h_{(t-1)}$ lần lượt là đầu vào và trạng thái ẩn trước đó. W_z và U_z là các ma trận trọng số đã học.

Khi x được đưa vào mạng, nó được nhân với trọng số W_z . $h_{(t-1)}$ chứa thông tin của các đơn vị tại thời điểm $t-1$ trước đó và nhân với trọng số U_z . Cả hai kết quả được cộng với nhau và qua hàm sigmoid để nén kết quả trong khoảng $[0,1]$. Cổng cập nhật giúp mô hình xác định được lượng thông tin trong quá khứ (thông tin ở bước $t-1$) cần chuyển đến tương lai (bước t). Điều này mang tới ưu điểm là mô hình có thể quyết định copy tất cả thông tin từ quá khứ và loại bỏ nguy cơ vanishing gradient.

+) Cổng cài đặt lại (reset gate) r_j [91]: Quyết định bao nhiêu thông tin từ trạng thái ẩn trước đó sẽ được quên đi khi tính toán nội dung nhớ hiện tại. Công thức tính

$$r_j = \sigma \left([W_r X]_j + [U_r h_{(t-1)}]_j \right) \quad (2.21)$$

Công thức này giống như là công thức ở cổng update đã nêu trên. Sự khác biệt chỉ là ở trọng số và mức sử dụng của cổng này. Về cơ bản, cổng này được sử dụng từ mô hình để quyết định lượng thông tin trong quá khứ được quên đi.

+) **Nội dung nhớ hiện tại $\tilde{h}_j^{(t)}$** : thông tin nội dung của bộ nhớ mới sẽ sử dụng cổng reset để lưu trữ thông tin có liên quan đến quá khứ được tính như sau [91]:

$$\tilde{h}_j^{(t)} = f([WX]_j + r_j[Uh_{(t-1)}]) \quad (2.22)$$

Các bước thực hiện của công thức trên:

- Phép nhân x với hàm trọng số W và $h_{(t-1)}$ với hàm trọng số U
- Phép nhân giữa cổng reset r_j và $Uh_{(t-1)}$. Phép nhân này sẽ xác định xem những gì cần phải xóa ở thời điểm trước đó.
- Cộng kết quả bước 1 và bước 2.
- Sử dụng hàm kích hoạt $f(\tanh, \dots)$.

+) **Bộ nhớ tại thời điểm hiện tại $h_j^{(t)}$** : ở bước cuối cùng, đầu ra của mạng cần tính là $h_j^{(t)}$ là vector chứa toàn bộ thông tin ở tại thời điểm t và truyền nó đi. Để thực hiện điều này, cần có cổng update. Nó xác định nội dung thu thập từ bộ nhớ hiện tại $\tilde{h}_j^{(t)}$ và những gì từ các bước trước đó $h_j^{(t)}$ [91]:

$$h_j^{(t)} = z_j h_j^{(t-1)} + (1 - z_j) \tilde{h}_j^{(t)} \quad (2.23)$$

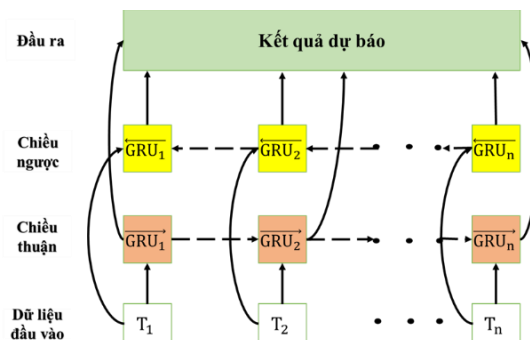
Các bước thực hiện của công thức trên:

- Nhân cổng update z_j và $h_j^{(t-1)}$
- Nhân $(1-z_j)$ với $\tilde{h}_j^{(t)}$
- Cộng kết quả của hai bước trên.

d. Mô hình dự báo BiGRU

BiGRU là sự kết hợp của hai lớp GRU đơn hướng độc lập. Một lớp duy trì các trạng thái ẩn phía trước, trong khi lớp kia duy trì các trạng thái ẩn phía sau. Cấu trúc của một mô hình BiGRU như sau [93]:

+) **Mô hình BiGRU một lớp ẩn**



Hình 2.8. Cấu trúc của mô hình BiGRU một lớp ẩn

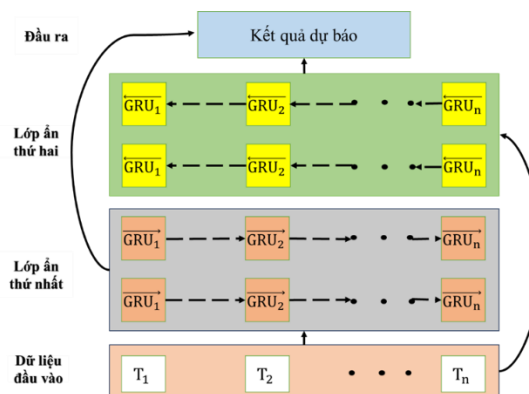
(Nguồn: Tài liệu [93])

Mô hình BiGRU với một lớp ẩn sử dụng cả hai chiều (thuận và nghịch) của mạng nơ-ron GRU để xử lý dữ liệu đầu vào. Lớp GRU thuận của mô hình xử lý thông tin theo chiều thuận từ thời điểm trước đến thời điểm hiện tại (thông tin đầu vào theo các bước thời gian $T_1, T_2, \dots, T_n$), trong khi lớp GRU nghịch xử lý theo chiều nghịch từ thời điểm hiện tại đến thời điểm trước đó ($T_n, \dots, T_2, T_1$). Mỗi lớp GRU bao gồm các cổng cập nhật và cổng khôi phục, giúp điều chỉnh thông tin được truyền đi và quên thông tin không cần thiết.

Lớp GRU thuận và lớp GRU nghịch sử dụng các tham số (trọng số và độ lệch) để tính toán đầu ra và trạng thái ẩn cho mỗi thời điểm. Kết quả từ hai chiều được tích hợp bằng cách sử dụng các phép toán như cộng, trung bình, hoặc các kỹ thuật kết hợp khác tùy thuộc vào mục tiêu của mô hình. Quá trình này giúp mô hình BiGRU hiểu được ngữ cảnh từ cả hai phía của mỗi thời điểm, làm tăng khả năng học và mô hình hóa các mối quan hệ phức tạp trong dữ liệu đầu vào.

+) Mô hình BiGRU hai lớp ẩn

Mô hình BiGRU với hai lớp ẩn có các đặc điểm giống với mô hình một lớp ẩn, tuy nhiên số lớp ẩn trong mô hình tăng thêm một lớp nữa. Cấu trúc của mô hình BiGRU hai lớp ẩn như Hình 2.9 dưới đây [93]:



Hình 2.9. Cấu trúc mô hình BiGRU hai lớp ẩn

(Nguồn: Tài liệu [93])

Lớp ẩn thứ nhất (Forward GRU): Lớp này chịu trách nhiệm xử lý dữ liệu theo chiều từ trái sang phải ($T_1, T_2, \dots, T_n$). Lớp này nhận đầu vào từ bước thời gian hiện tại và trạng thái ẩn của lớp trước đó. Quá trình này giúp lớp ẩn thứ nhất nắm bắt thông tin quan trọng từ quá khứ và tiếp tục truyền thông tin đó qua chuỗi.

Lớp ẩn thứ hai (Backward GRU): Lớp này xử lý dữ liệu theo chiều từ phải sang trái ($T_n, \dots, T_2, T_1$). Tương tự như lớp ẩn thứ nhất, lớp này cũng nhận đầu vào từ bước thời gian hiện tại và trạng thái ẩn của lớp trước đó. Quá trình này giúp lớp ẩn thứ hai nắm bắt thông tin tương lai và truyền ngược lại qua chuỗi.

Kết hợp thông tin từ cả hai lớp ẩn:

Tổng hợp trạng thái: trạng thái ẩn của mỗi lớp được tổng hợp lại để tạo ra trạng thái ẩn cuối cùng cho mỗi bước thời gian.

Thông tin đầu ra: là quá trình kết hợp các trạng thái ẩn cuối cùng của một hay nhiều lớp ẩn để tạo ra kết quả dự báo.

e. Ưu nhược điểm của các mô hình chuỗi thời gian:

Ưu điểm:

- Nắm bắt mối liên hệ theo thời gian: các mô hình như LSTM và GRU được thiết kế để ghi nhớ các mối liên hệ trong chuỗi dữ liệu, đặc biệt hữu ích trong dự báo công suất điện mặt trời vốn có tính chu kỳ và phụ thuộc mạnh vào quá khứ gần
- Khả năng học xu hướng phi tuyến phức tạp: Không bị ràng buộc bởi giả định tuyến tính như các mô hình thống kê, các mô hình học sâu chuỗi thời gian có thể phát hiện các mẫu phức tạp ẩn trong dữ liệu như đỉnh- đáy, biên độ dao động, hoặc độ trễ do mây che nắng.
- Tùy biến linh hoạt với dữ liệu nhiều chiều: Các mô hình này dễ dàng tích hợp nhiều đặc trưng đầu vào như nhiệt độ, bức xạ, tháng, v.v, và học các tương tác giữa các yếu tố này qua thời gian.

Nhược điểm:

- Phụ thuộc mạnh vào chuỗi dữ liệu đầy đủ và liên tục: Nếu thiếu dữ liệu ở các điểm gần thời gian dự báo, hoặc chuỗi bị gián đoạn, mô hình sẽ mất ngữ cảnh và dẫn đến kết quả sai lệch (underfitting hoặc trơn hóa đường dự báo).
- Thời gian huấn luyện dài, yêu cầu tài nguyên tính toán cao: Do cấu trúc mạng phức tạp và phụ thuộc thời gian, các mô hình LSTM/GRU cần nhiều thời gian huấn luyện hơn so với các mô hình đơn giản như cây quyết định.
- Không chống chịu tốt với nhiễu hoặc sai số từ đầu vào: Nếu đặc trưng đầu vào (ví dụ: bức xạ mặt trời) bị sai do cảm biến hoặc nội suy, mô hình dễ bị sai lệch toàn bộ chuỗi dự báo vì tính tích lũy theo thời gian.

2.3.1.4. Các mô hình kết hợp

Trong phần này, nghiên cứu sinh tập trung vào ba nhóm mô hình hybrid chính cho dự báo công suất phát điện mặt trời: các mô hình kết hợp dựa trên các mô hình chuỗi thời gian (LSTM, GRU, Transformer), các mô hình cây quyết định và các phương pháp Stacking Ensemble. Lựa chọn này dựa trên khả năng vượt trội của các mô hình này trong việc xử lý dữ liệu chuỗi thời gian phi tuyến và phức tạp đặc trưng của hệ thống điện mặt trời. LSTM và GRU, với khả năng học các phụ thuộc dài hạn, đặc biệt phù hợp cho việc nắm bắt các mẫu thời tiết và chu kỳ năng lượng mặt trời. Mô hình Transformer [94] với kiến trúc dựa trên cơ chế Attention, cho phép học phụ thuộc không tuần tự giữa các thời điểm, giúp mô hình linh hoạt hơn trong việc nhận diện các biến động không đều của dữ liệu. Trong khi đó, Stacking Ensemble cho phép kết hợp hiệu quả nhiều mô hình để tăng cường độ chính xác dự báo. Mặc dù các kết hợp khác như ARIMA-ANN, WT-SVM, ANN-SVM... cũng đã được nghiên cứu nhưng ba nhóm mô hình này thể hiện tiềm năng lớn hơn trong việc cải thiện độ chính xác dự báo và khả năng thích ứng với đặc tính biến động cao của dữ liệu năng lượng mặt trời.

a. Các mô hình kết hợp chuỗi thời gian

Các mô hình kết hợp dựa trên kiến trúc LSTM hoặc GRU thường được sử dụng để khai thác đặc trưng chuỗi thời gian và quan hệ phụ thuộc dài hạn giữa các yếu tố khí tượng và công suất phát. Nhiều nghiên cứu đề xuất tích hợp LSTM với CNN để trích xuất đặc trưng không gian, hoặc với các lớp Attention để tăng tính linh hoạt theo mùa và điều kiện thời tiết. Một số hướng khác kết hợp LSTM với các thuật toán học máy truyền thống (như SVR, RF) để cải thiện khả năng khái quát trên dữ liệu nhiễu. Gần đây, mô hình Transformer với cơ chế Attention tự thích ứng cũng được áp dụng kết hợp với CNN hoặc các phương pháp tiền xử lý đặc trưng để tăng hiệu quả dự báo nhiều bước trong các hệ thống năng lượng tái tạo, đặc biệt trong điều kiện dữ liệu biến động theo mùa và độ phân giải thời gian cao.

Danh sách mô hình cụ thể, dữ liệu sử dụng và kết quả so sánh đã được trình bày chi tiết trong phụ lục A.3.

b. Các mô hình kết hợp cây quyết định

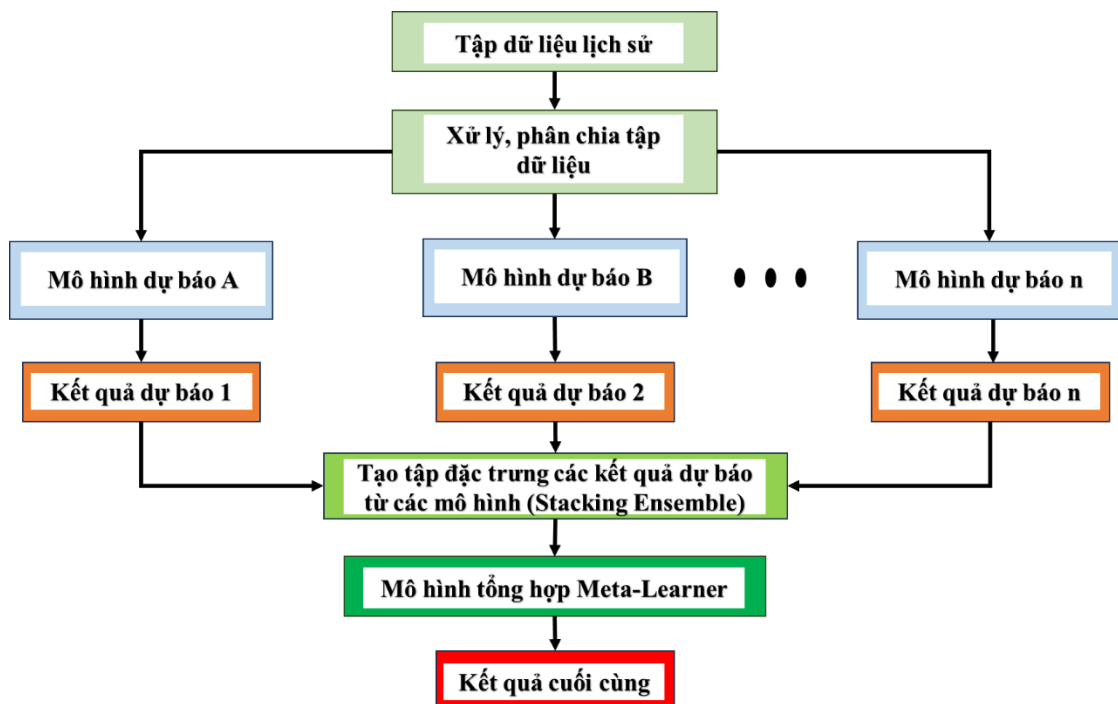
Nhóm mô hình này tập trung vào việc khai thác sức mạnh phân loại của các thuật toán cây quyết định tăng cường (boosting/bagging), vốn có khả năng xử lý tốt dữ liệu thiếu hụt và không yêu cầu chuẩn hóa phức tạp. Một số nghiên cứu kết hợp XGBoost

với các bước trích đặc trưng đầu vào bằng PCA, EMD hoặc CNN; một số khác sử dụng cây quyết định làm tầng cuối trong mô hình kết hợp.

Do đặc tính dễ triển khai, tốc độ huấn luyện nhanh và độ chính xác cao trong điều kiện dữ liệu không lý tưởng, nhóm mô hình này rất phù hợp với thực tiễn Việt Nam. Chi tiết các nghiên cứu đã được tổng hợp tại phụ lục A.4.

c. Mô hình kết hợp Stacking-Ensemble

Xu hướng xây dựng các mô hình dự báo công suất phát điện mặt trời ngày nay là các mô hình kết hợp. Một trong các cách kết hợp đó chính là mô hình Stacking Ensemble. Khác với kiểu kết hợp dạng Hybrid khi mà kết quả đầu ra của mô hình này sẽ làm dữ liệu đầu vào cho mô hình kia (kết hợp nối tiếp-Sequential Hybrid), Stacking Ensemble (hay Stacked Generalization) [95] là kỹ thuật huấn luyện nhiều mô hình học máy (gọi là mô hình cơ sở) trên cùng tập dữ liệu (kết hợp song song). Sau đó, một mô hình khác gọi là Meta-Learner được huấn luyện trên đầu ra của các mô hình cơ sở để tổng hợp và kết hợp chúng lại. Kỹ thuật này khai thác lợi thế của mỗi mô hình riêng lẻ, giúp tạo ra một mô hình tổng hợp có hiệu suất dự báo cao hơn so với từng mô hình đơn lẻ. Sơ đồ mô tả cấu trúc của mô hình Stacking Ensemble được thể hiện dưới Hình 2.10. Sơ đồ này được tạo ra dựa trên tổng hợp kiến thức từ các nguồn [95]–[97] nhằm minh họa trực quan quy trình hoạt động của phương pháp Stacking Ensemble.



Hình 2.10. Cấu trúc mô hình Stacking Ensemble

(Nguồn: Tài liệu [95])

Mô hình Stacking Ensemble đã được áp dụng thành công trong nhiều lĩnh vực, đặc biệt là trong dự báo năng lượng tái tạo. Nghiên cứu gần đây của Abdallah Abdellatif [98] và các cộng sự đã đề xuất một thuật toán Stacking Ensemble (Stack-ETR) để dự báo công suất đầu ra của các hệ thống điện mặt trời một ngày trước đó, kết hợp ba mô hình học máy là Random Forest, XGBoost và AdaBoost làm mô hình cơ sở. Ngoài ra, một mô hình Extra Trees Regressor được sử dụng làm Meta-Learner để tích hợp các dự báo từ các mô hình cơ sở, nhằm nâng cao độ chính xác dự báo. Mô hình đề xuất này đã được kiểm chứng trên ba hệ thống điện mặt trời thực tế với bốn năm dữ liệu khí tượng, và cho thấy hiệu suất vượt trội so với các mô hình Ensemble khác về chỉ số RMSE và MAE. Cụ thể, mô hình Stack-ETR được đề xuất đã làm giảm RMSE xuống 24,49%, 40,2% và 27,95% và MAE xuống 28,88%, 47,2% và 40,88% so với mô hình cơ sở ETR khác sử dụng các loại vật liệu làm pin như pin mặt trời màng mỏng (Thin Film-TF), pin quang điện đơn tinh thể (monocrystalline- MC) và đa tinh thể (polycrystalline-PC). Ví dụ này minh họa rõ ràng cho tiềm năng của các mô hình Stacking Ensemble trong việc cải thiện chất lượng dự báo công suất điện mặt trời. Như vậy, việc lựa chọn phương pháp Stacking Ensemble để phục vụ dự báo công suất phát điện mặt trời cho phép kết hợp hiệu quả các mô hình học máy khác nhau, giúp bù đắp những nhược điểm của từng mô hình riêng lẻ và khai thác tối đa ưu điểm của chúng. Hơn nữa, các nghiên cứu trước đây [33], [99] đã chứng minh Stacking Ensemble thường cho kết quả dự báo tốt hơn so với các mô hình đơn lẻ trong nhiều lĩnh vực ứng dụng khác nhau. Với các ưu điểm đã được liệt kê ở trên, Nghiên cứu sinh lựa chọn xây dựng mô hình dự báo dựa trên mô hình Stacking Ensemble.

2.3.2. Các mô hình được nghiên cứu trong nước

Mặc dù hoạt động dự báo công suất phát điện mặt trời đã được quy định rõ ràng trong các văn bản pháp lý và đang dần trở thành yêu cầu bắt buộc đối với các đơn vị vận hành, nhưng hệ thống nghiên cứu khoa học trong nước về lĩnh vực này vẫn còn khá khiêm tốn. Đặc biệt, các công trình tập trung vào việc ứng dụng các mô hình học máy, học sâu, hoặc tổ hợp mô hình để nâng cao độ chính xác dự báo vẫn chưa nhiều, và chủ yếu dừng lại ở quy mô thử nghiệm nhỏ hoặc mô hình đơn lẻ.

Một số nghiên cứu đã bước đầu áp dụng các kỹ thuật thống kê và mô hình cây quyết định (như Random Forest, XGBoost), tuy nhiên chưa có sự so sánh hệ thống giữa các nhóm mô hình, cũng như chưa khai thác triệt để tính linh hoạt của các phương pháp tổ

hợp hoặc khả năng thích nghi trong điều kiện thiếu dữ liệu lịch sử. Tổng quan các nghiên cứu hiện có về dự báo công suất phát điện mặt trời tại Việt Nam được trình bày trong Phụ lục A.5.

2.3.3. Đánh giá các mô hình dự báo công suất phát điện mặt trời

Các nghiên cứu trong và ngoài nước đã chỉ ra rằng việc sử dụng kết hợp các mô hình dự báo để tăng độ chính xác của kết quả đầu ra đang ngày càng phổ biến. Ưu nhược điểm của từng mô hình dự báo được phân tích trong Bảng 2.3 dưới đây.

Bảng 2.3. Ưu, nhược điểm các mô hình dự báo công suất phát điện mặt trời

Nhóm mô hình dự báo	Các Mô hình đại diện	Ưu điểm	Nhược điểm	Khoảng thời gian dự báo hiệu quả
Mô hình truyền thống	Mô hình Quán tính, mô hình thống kê và mô hình vật lý [59] [60]	- Nhanh, đơn giản, chi phí tính toán thấp. - Có thể làm baseline tốt cho các so sánh.	- Mô hình quán tính và mô hình ARIMA có sai số lớn khi thời tiết biến động - Mô hình NWP cần hiệu chỉnh bias.	Phổ biến cho dự báo rất ngắn hạn và ngắn hạn.
Mô hình cây quyết định [69]	XGBoost, LightGBM, Random Forest	- Dễ hiểu và dễ triển khai. - Xử lý tốt dữ liệu phi tuyến, rất bền vững với dữ liệu bị khuyết thiếu. - Thời gian huấn luyện và dự báo nhanh.	- Khó nắm bắt được quan hệ theo chuỗi thời gian.	Hàng giờ tới hàng ngày

Mô hình ANN đơn giản (có 1 lớp ẩn) và mô hình MLPN nhiều lớp [61] [67]		- Mạnh với dữ liệu phi tuyến, tính linh hoạt cao	- Dễ bị quá khớp, không tận dụng cấu trúc chuỗi thời gian tốt như RNN.	1 ngày
Mô hình học sâu chuỗi thời gian [26], [83], [84] [88] [91] [31]	RNN, LSTM, GRU, BiGRU, BiLSTM, Transformer	- Xử lý dữ liệu thông tin dạng chuỗi tốt - Transformer có thể học quan hệ xa và xử lý song song.	- Có một thách thức trong việc huấn luyện đó là vanishing gradient và exploding gradient - Rất nhạy cảm với dữ liệu trong chuỗi quá khứ bị khuyết thiếu. - Chi phí tính toán cao, thời gian huấn luyện và dự báo chậm.	Hàng ngày
Các mô hình tổ hợp Ensemble [95]	Stacking Ensemble.	- Tăng độ ổn định và tính chính xác, bù trừ sai số của các mô hình riêng lẻ	- Quy trình huấn luyện phức tạp, yêu cầu lượng dữ liệu đủ lớn để huấn luyện - Cần giám sát và tái định kỳ các tham số.	Hàng giờ tới hàng ngày.

2.4. Kết luận chương 2

Chương 2 đã trình bày chi tiết về các bộ dữ liệu được sử dụng trong nghiên cứu, bao gồm cả dữ liệu huấn luyện mô hình và dữ liệu kiểm định khả năng mở rộng. Cụ thể, bộ dữ liệu chính được thu thập từ nhà máy điện mặt trời tại tỉnh Quảng Trị trong giai đoạn từ 01/01/2022 đến 31/12/2022, bao gồm các thông số vận hành và điều kiện bức xạ thực đo theo thời gian thực. Bộ dữ liệu này đóng vai trò trung tâm trong việc xây dựng và đánh giá hiệu quả các mô hình dự báo trong điều kiện dữ liệu thực tế tại Việt Nam. Bên cạnh đó, ba bộ dữ liệu bổ sung từ các nhà máy khác tại Quảng Trị, Thanh Hóa và Đắk Lắk được sử dụng nhằm kiểm định khả năng tổng quát hóa và mở rộng của mô hình trong bối cảnh đa nhà máy một yêu cầu quan trọng trong ứng dụng thực tiễn. Ngoài ra, chương cũng mô tả đầy đủ và hệ thống các quy trình tiền xử lý dữ liệu nhằm đảm bảo chất lượng và tính toàn vẹn của đầu vào.

Chương 2 cũng mô tả tổng quan về các phương pháp dự báo bức xạ, đây là nền tảng cho việc dự báo công suất phát có độ chính xác cao.

Đặc biệt, chương đã phân tích những thách thức đặc thù của bài toán dự báo công suất phát trong bối cảnh dữ liệu thiếu hụt và không đồng đều giữa các nhà máy, đây cũng chính là một trong những khoảng trống nghiên cứu được luận án đề xuất giải quyết.

Những cơ sở dữ liệu và tiền xử lý nêu trong chương này đóng vai trò then chốt cho các chương tiếp theo, nơi các mô hình dự báo được xây dựng, hiệu chỉnh và đánh giá. Việc chuẩn bị và hiểu rõ đặc điểm dữ liệu không chỉ đảm bảo tính chính xác trong huấn luyện mô hình mà còn góp phần định hình chiến lược xây dựng mô hình phù hợp với điều kiện vận hành thực tế của các nhà máy điện mặt trời tại Việt Nam.

CHƯƠNG 3. PHÂN TÍCH ĐÁNH GIÁ CÁC MÔ HÌNH DỰ BÁO BỨC XẠ VÀ CÔNG SUẤT PHÁT ĐIỆN MẶT TRỜI

Sau khi đã thiết lập cơ sở dữ liệu và quy trình tiền xử lý trong chương 2, chương này tiến hành xây dựng và phân tích đánh giá dựa trên thực nghiệm các mô hình dự báo công suất phát điện mặt trời, nhằm xác định những mô hình dự báo hiệu quả và phù hợp với điều kiện dữ liệu thực tế tại Việt Nam. Trước khi đi sâu vào dự báo công suất, chương cũng xem xét hiệu quả của một số mô hình học máy trong bài toán dự báo bức xạ mặt trời, nhằm chuẩn bị dữ liệu đầu vào cho các mô hình công suất phát.

Trọng tâm của chương là đánh giá hai nhóm mô hình chính:

- **Nhóm mô hình chuỗi thời gian** (LSTM, GRU, BiGRU) có khả năng học được tính phụ thuộc thời gian sâu sắc trong dữ liệu;
- **Nhóm mô hình cây quyết định** (XGBoost, LightGBM, Random Forest) có ưu điểm nổi bật trong xử lý dữ liệu gián đoạn và dễ triển khai trong thực tế.

Việc lựa chọn và huấn luyện các mô hình này được thực hiện dựa trên bộ dữ liệu thực đo tại nhà máy điện mặt trời đã giới thiệu ở chương trước. Kết quả dự báo sẽ được so sánh dựa trên các chỉ số sai số phổ biến để:

- Làm rõ điểm mạnh, yếu của từng nhóm mô hình;
- Đánh giá khả năng ứng dụng trong điều kiện dữ liệu còn thiếu hụt hoặc không đồng nhất. Từ đó, lựa chọn ra các mô hình nền tảng tốt nhất để làm đầu vào cho mô hình tổ hợp Stacking Ensemble được trình bày trong Chương 4.

Thông qua các đánh giá này, chương 3 đóng vai trò thiết lập nền tảng thực nghiệm vững chắc cho việc xây dựng một mô hình dự báo tổ hợp thích ứng với điều kiện vận hành tại Việt Nam.

3.1. Xây dựng, huấn luyện và so sánh các mô hình dự báo bức xạ mặt trời.

Trên thế giới hiện nay, các hướng tiếp cận dự báo bức xạ mặt trời chủ yếu được phân loại thành ba nhóm chính. Đầu tiên là nhóm mô hình vật lý, với hai đại diện tiêu biểu là phương pháp hình ảnh bầu trời (SIM) [49] và dự báo thời tiết số (NWP) [100]. Nguyên lý hoạt động của nhóm này dựa trên việc giải các hệ phương trình vật lý mô phỏng sự truyền dẫn bức xạ qua tầng khí quyển, kết hợp với các tham số đầu vào như nhiệt độ, độ ẩm và thành phần không khí. Mặc dù phương pháp này có ưu điểm vượt trội về cơ sở khoa học và độ tin cậy trong dự báo dài hạn, nhưng rào cản lớn nhất nằm

ở yêu cầu khắt khe về tài nguyên tính toán (thường cần siêu máy tính) và sự phức tạp của dữ liệu đầu vào [100].

Song song với đó, nhóm mô hình thống kê (như ARIMA, hồi quy tuyến tính) lại được ưa chuộng nhờ sự đơn giản và hiệu quả trong các kịch bản dự báo ngắn hạn. Tuy nhiên, hạn chế cốt lõi của phương pháp thống kê truyền thống là khả năng thích ứng kém đối với các mối quan hệ phi tuyến phức tạp của dữ liệu khí tượng [101].

Trước bối cảnh đó, sự phát triển của trí tuệ nhân tạo (AI) trong những năm gần đây đã mở ra hướng đi mới đầy tiềm năng. Nhiều công trình nghiên cứu [23], [56], [102]–[105] đã chứng minh tính hiệu quả của các thuật toán học máy và học sâu, đặc biệt là ba mô hình: LightGBM, LSTM và GRU. Cụ thể:

LightGBM [78]: Là một kiến trúc Gradient Boosting hiệu năng cao, thuật toán này nổi bật với tốc độ huấn luyện nhanh và khả năng tự động trích xuất các đặc trưng phi tuyến. Đây là công cụ đắc lực để giải quyết bài toán tương tác đa biến giữa các yếu tố khí tượng.

LSTM [88]: Được thiết kế chuyên biệt cho dữ liệu chuỗi thời gian, mạng nơ-ron bộ nhớ dài-ngắn (LSTM) khắc phục được vấn đề nhớ lại thông tin trong quá khứ xa. Đặc tính này biến LSTM trở thành lựa chọn tối ưu để nắm bắt các xu hướng biến thiên dài hạn của bức xạ mặt trời.

GRU [91]: Là phiên bản tinh gọn của mạng nơ-ron hồi quy (RNN), GRU sở hữu khả năng học các phụ thuộc thời gian tương tự như LSTM nhưng với cấu trúc đơn giản hơn. Ưu điểm này giúp mô hình giảm thiểu đáng kể chi phí tính toán mà vẫn duy trì hiệu suất dự báo ổn định.

Dựa trên các phân tích nêu trên, nghiên cứu sinh tiến hành xây dựng và so sánh thực nghiệm ba mô hình LightGBM, LSTM và GRU trên cùng một tập dữ liệu quan trắc thực tế. Mục tiêu của thử nghiệm này là đánh giá toàn diện khả năng học xu hướng và tốc độ xử lý của từng thuật toán, từ đó làm cơ sở khoa học để đề xuất mô hình tối ưu cho bài toán dự báo bức xạ và công suất phát điện mặt trời trong các chương tiếp theo.

3.1.1. Xây dựng và huấn luyện các mô hình dự báo

a. Thông số các mô hình dự báo.

Chi tiết các siêu tham số tối ưu được thiết lập cho quá trình huấn luyện và kiểm chứng hiệu năng của ba mô hình dự báo được tổng hợp tại Bảng 3.1 dưới đây. Trong đó, hai mô hình LSTM và GRU được xây dựng dựa trên kiến trúc nền tảng đã trình bày

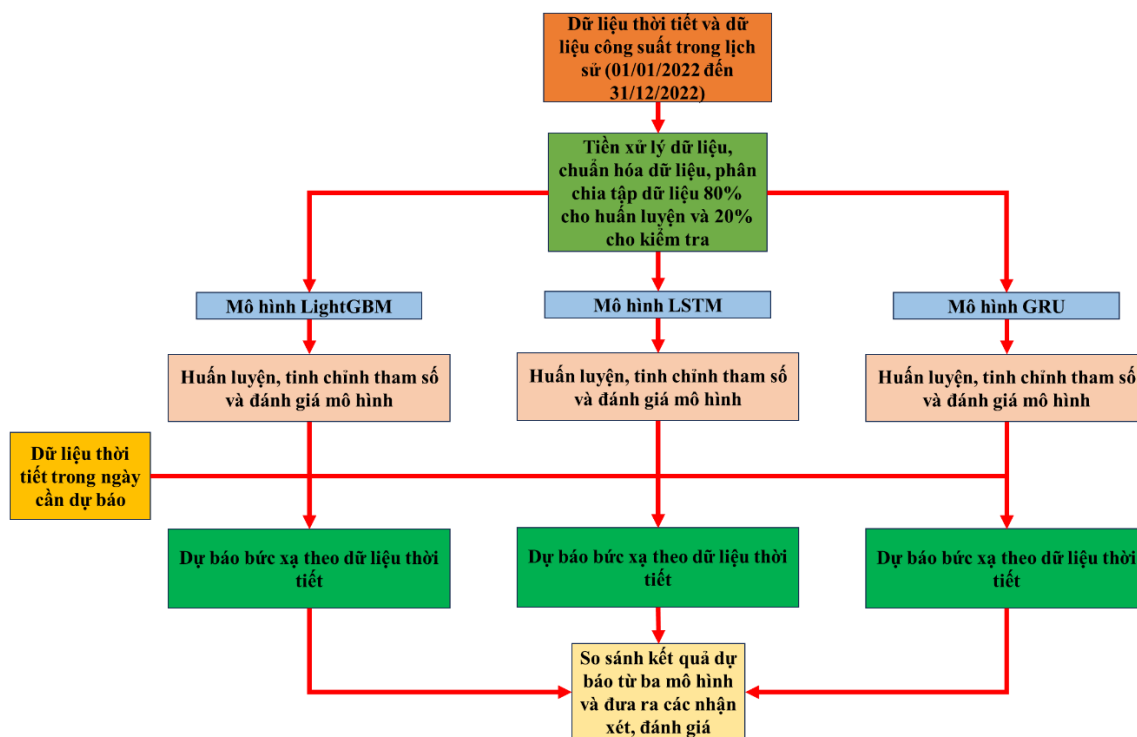
tại mục 2.3.1.3, đồng thời được cấu hình hoàn toàn đồng nhất về bộ tham số vận hành nhằm đảm bảo tính khách quan khi so sánh thực nghiệm.

Bảng 3.1: Thiết lập tham số cấu hình cho các mô hình nghiên cứu

Mô hình	Tham số	Giá trị	Vai trò
LightGBM	Learning_rate	0,1	Kiểm soát tốc độ hội tụ
	Max_depth	8	Giới hạn độ sâu cây để kiểm soát độ phức tạp mô hình
	Num_leaves	31	Tăng độ chính xác phân nhánh, tránh hiện tượng overfitting.
	Min_child_samples	20	Đảm bảo số lượng dữ liệu tối thiểu tại mỗi nút lá.
	colsample_bytree	1	Tỷ lệ chọn mẫu đặc trưng ngẫu nhiên cho từng cây.
LSTM/GRU	Hàm kích hoạt	tanh	Chuẩn hóa đầu ra dữ liệu trong biên độ [-1, 1]
	Hàm sai số	MSE	Tăng độ nhạy của mô hình với các sai số dự báo nhỏ
	Số vòng lặp	30	Để đảm bảo khai thác dữ liệu toàn cục
	Batch size	32	Tối ưu hóa tốc độ tính toán và cập nhật trọng số
	Look-back window	24 giờ	Cung cấp dữ liệu từ ngữ cảnh quá khứ đầy đủ (24 giờ tương đương với 288 điểm dữ liệu trong quá khứ)

b. Sơ đồ thuật toán thực hiện dự báo bức xạ mặt trời

Các bước chính thực hiện dự báo bức xạ mặt trời được thể hiện dưới sơ đồ Hình 3.1 sau đây:



Hình 3.1. Lưu đồ quy trình thực nghiệm và đánh giá hiệu năng các thuật toán dự báo bức xạ.

Chi tiết các bước triển khai trong quy trình thực nghiệm:

Bước 1: Chuẩn bị dữ liệu huấn luyện (Training Data Set)

Nguồn dữ liệu đầu vào được thu thập từ cơ sở dữ liệu quan trắc tại nhà máy điện mặt trời Quảng Trị (như đã mô tả tại mục 2.1.1), sau khi đã lược bỏ trường thông tin về công suất phát thực tế. Trước khi đưa vào mô hình, dữ liệu thô trải qua quy trình tiền xử lý nghiêm ngặt bao gồm: làm sạch nhiễu, khôi phục giá trị khuyết thiếu và chuẩn hóa đặc trưng. Tập dữ liệu sau xử lý được phân chia theo tỷ lệ tiêu chuẩn: 80% dành cho huấn luyện và 20% dành cho kiểm tra nội bộ

Bước 2: Thiết lập kịch bản dự báo độc lập

Để đảm bảo tính khách quan và kiểm chứng khả năng tổng quát hóa của mô hình, nghiên cứu sinh thiết lập hai kịch bản thử nghiệm sử dụng các tập dữ liệu hoàn toàn tách biệt (chưa từng xuất hiện trong quá trình huấn luyện):

- Kịch bản 1 (Ngắn hạn trong ngày): Sử dụng chuỗi số liệu ngày 15/02/2021. Mục tiêu là đánh giá độ nhạy của mô hình trong điều kiện vận hành hằng ngày.
- Kịch bản 2 (Ngắn hạn dài ngày): Sử dụng dữ liệu chuỗi 03 ngày liên tiếp (từ 01/03 đến 03/03/2021). Kịch bản này mô phỏng bài toán lập kế hoạch điều độ và bảo trì hệ thống, yêu cầu sự ổn định của mô hình qua nhiều chu kỳ ngày đêm.

Bước 3: Cấu hình chuỗi quá khứ (Look-back Window)

Đối với các mô hình học sâu the chuỗi thời gian như LSTM và GRU, việc cung cấp ngữ cảnh quá khứ là yếu tố tiên quyết để mô hình nắm bắt được các phụ thuộc thời gian. Chuỗi quá khứ quan sát được trong các kịch bản thử nghiệm được ấn định là 24 giờ (tương đương 288 điểm dữ liệu quá khứ). Cụ thể:

- Tại kịch bản 1: Dữ liệu ngày 14/02/2021 được nạp làm đầu vào để dự báo cho ngày kế tiếp.
- Tại kịch bản 2: Dữ liệu ngày 28/02/2021 đóng vai trò là ngữ cảnh khởi tạo cho chuỗi dự báo kéo dài 3 ngày sau đó. Việc đảm bảo tính liên tục của chuỗi dữ liệu quá khứ giúp tránh hiện tượng "học thiếu" (underfitting) và cho phép các thuật toán khai thác tối đa đặc trưng động học của bức xạ mặt trời..

Bước 4: Huấn luyện và Tối ưu hóa

Ba mô hình (LightGBM, LSTM, GRU) được huấn luyện song song trên cùng một không gian dữ liệu chuẩn hóa. Quá trình này đi kèm với việc tinh chỉnh các siêu tham số (hyperparameters) để tối ưu hóa hiệu suất của từng mô hình.

Bước 5: Đánh giá hiệu năng đa chiều

Chiến lược đánh giá được phân kỳ thành hai giai đoạn:

- **Giai đoạn 1 (Kiểm chứng nội bộ):** Đánh giá sơ bộ dựa trên các chỉ số sai số thu được từ tập kiểm tra (20% dữ liệu ban đầu).
- **Giai đoạn 2 (Kiểm chứng thực tế):** Áp dụng mô hình vào hai kịch bản dự báo độc lập nêu trên để đo lường độ chính xác, tốc độ tính toán và khả năng thích nghi theo thời gian. Kết quả tổng hợp từ các giai đoạn này sẽ là cơ sở để phân tích ưu/nhược điểm và định hướng lựa chọn giải pháp tối ưu. Các thang đo định lượng cụ thể được trình bày chi tiết tại mục 1.3.1.

3.1.2. Đánh giá hiệu năng dự báo bức xạ mặt trời của các mô hình

Kết quả dự báo bức xạ mặt trời của ba mô hình qua các kịch bản thử nghiệm được tổng hợp chi tiết trong Bảng 3.2.

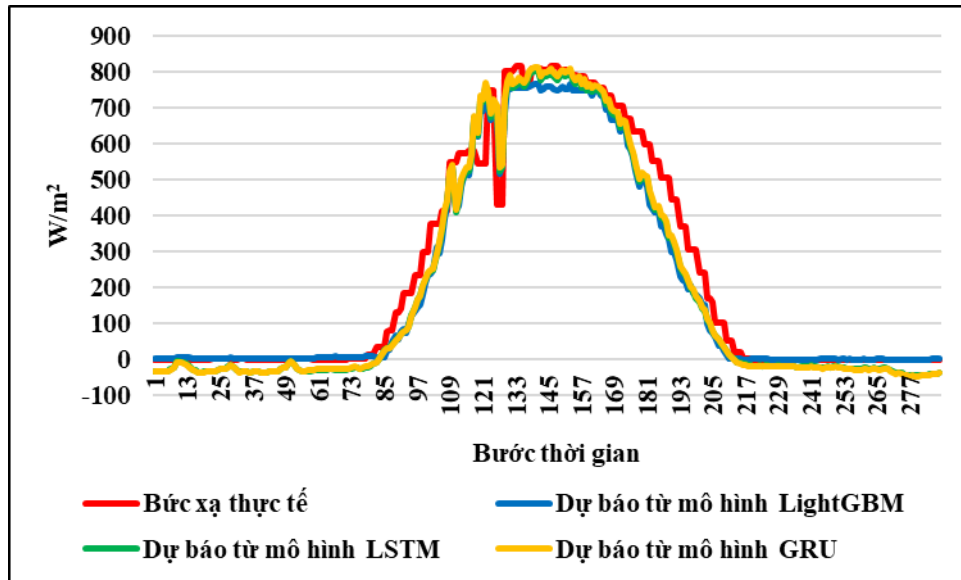
Bảng 3.2. Tổng hợp các chỉ số đánh giá hiệu năng và sai số thực nghiệm của ba mô hình dự báo trên các kịch bản thực nghiệm.

Các chỉ số		Mô hình LightGBM	Mô hình LSTM	Mô hình GRU
Tập kiểm tra	Thời gian thực hiện (s)	0,53	456,57	397,20

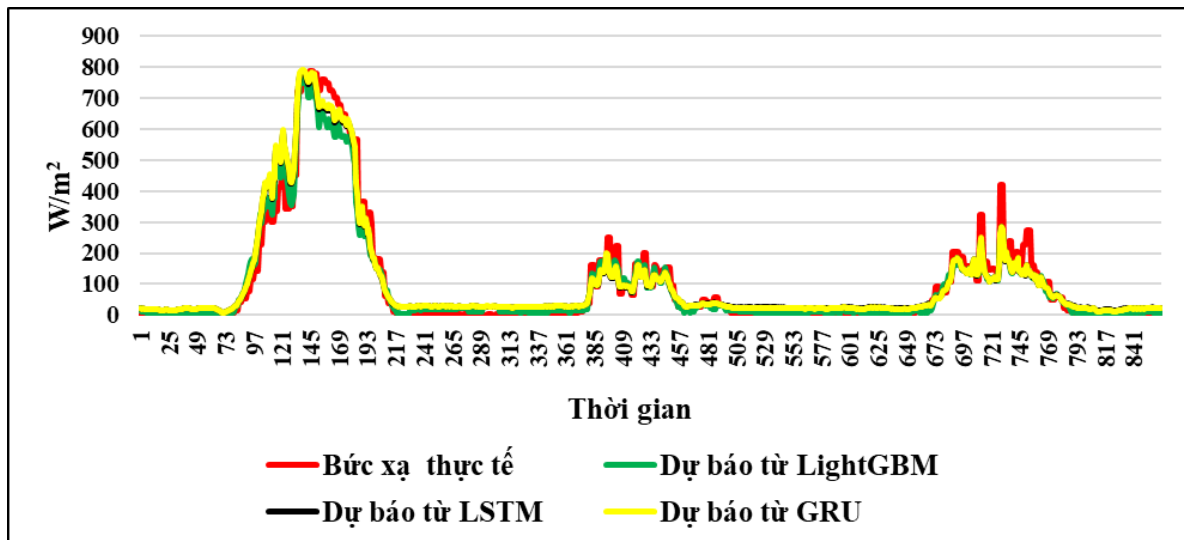
	RMSE (W/m²)	54,82	59,16	59,33
	MAE (W/m²)	27,60	34,79	34,75
Tập dữ liệu thời tiết ngày 15-02-2021	Thời gian thực hiện (s)	0,12	1,45	1,76
	RMSE (W/m²)	60,01	58,84	57,45
	MAE (W/m²)	36,22	44,45	42,75
Tập dữ liệu thời tiết các ngày từ 01-03/03/2021	Thời gian thực hiện (s)	0,21	2,48	1,57
	RMSE (W/m²)	36,27	38,24	38,14
	MAE (W/m²)	21,05	27,48	26,97

Dựa trên dữ liệu thực nghiệm ở Bảng 3.2, kết quả cho thấy mô hình Gradient Boosting có hiệu suất tốt hơn so với nhóm mô hình học sâu, đặc biệt là trên thử nghiệm với tập dữ liệu kiểm tra. Cụ thể, LightGBM đạt độ chính xác cao hơn LSTM và GRU với sai số RMSE chỉ ở mức 54,817 W/m² và sai số (MAE là 27,599 W/m². Trong khi đó, hai mô hình mạng nơ-ron hồi quy ghi nhận mức sai số cao hơn đáng kể: LSTM có RMSE là 59,164 W/m² và GRU là 59,331 W/m². Ưu thế này tiếp tục được khẳng định qua tốc độ xử lý khi LightGBM có thời gian thực thi chỉ 0,534 giây, nhanh hơn gấp hàng trăm lần so với 456,571 giây của LSTM và 397,197 giây của GRU. Xu hướng tương tự cũng được quan sát thấy trong các kịch bản dự báo thực tế ngắn hạn và dài ngày. Mặc dù chênh lệch về chỉ số RMSE giữa ba mô hình là tương đương nhau (dao động trong từ 57 đến 60 W/m² cho ngày 15/02/2021 và khoảng 36 đến 38 W/m² với kịch bản dự báo từ 01 đến 03/03/2021), nhưng sự phân hóa về độ ổn định lại thể hiện rõ qua chỉ số MAE, nơi LightGBM luôn duy trì mức sai số thấp nhất (đạt 36,219 W/m² trong ngày 15/02/2021 và giảm sâu xuống còn 21,048 W/m² trong chuỗi ngày 01-03/03/2021), đồng thời đáp ứng rất tốt với yêu cầu thời gian thực với thời gian thực hiện dự báo không đáng kể.

Mức độ tương quan giữa giá trị dự báo và dữ liệu thực tế được minh họa trực quan tại Hình 3.2 và Hình 3.3; trên không gian biểu diễn với trục hoành là chuỗi thời gian (độ phân giải 5 phút) và trục tung là cường độ bức xạ, các đường đặc tính dự báo của cả ba mô hình đều bám sát xu hướng biến thiên của bức xạ thực tế, đặc biệt thể hiện khả năng thích nghi tốt tại các thời điểm xuất hiện sụt giảm bức xạ đột ngột dù biên độ dao động tại các điểm cực trị này vẫn còn tồn tại sai lệch nhất định.



Hình 3.2. Đồ thị minh họa giá trị dự báo của ba mô hình LightGBM, LSTM, GRU so với giá trị bức xạ thực tế ngày 15/02/2021.



Hình 3.3. Đồ thị minh họa giá trị dự báo của ba mô hình LightGBM, LSTM, GRU so với giá trị bức xạ thực tế ngày 01-03/03/2021.

Kết quả minh họa trực quan tại Hình 3.2 và Hình 3.3 cho thấy mức độ tương quan chặt chẽ giữa giá trị dự báo của ba mô hình và chuỗi dữ liệu quan trắc thực tế. Các mô hình đều thể hiện khả năng khái quát hóa tốt, phản ánh chính xác động thái biến thiên của cường độ bức xạ trong chu kỳ ngày. Đặc biệt, trong các giai đoạn xuất hiện nhiễu động thời tiết (gây sụt giảm mạnh bức xạ cục bộ), các mô hình đã chứng minh được độ chính xác khi nhận diện kịp thời xu hướng thay đổi đột ngột của đồ thị. Tuy nhiên, một hạn chế tồn tại là sự sai lệch nhất định về biên độ tại các điểm cực trị. Kết quả này phản ánh thách thức cố hữu đối với các mô hình học máy thuần túy dựa trên dữ liệu lịch sử

trong việc định lượng chính xác các giá trị dị biệt (outliers) hoặc các biến động cực đoan của môi trường.

3.1.3. Kết luận

Từ các kết quả thực nghiệm và phân tích nêu trên, có thể rút ra một số nhận định tổng quan về đặc tính của các phương pháp dự báo:

Thứ nhất, về hiệu suất dự báo: Kết quả cho thấy mô hình LightGBM đạt được độ chính xác khả quan hơn so với hai mô hình học sâu trong hầu hết các kịch bản thử nghiệm. Mặc dù sự chênh lệch về sai số (RMSE, MAE) giữa các mô hình là không quá lớn, nhưng LightGBM cho thấy sự ổn định tốt hơn, đặc biệt là trong các bài toán dự báo ngắn hạn dài ngày. Ngược lại, LSTM và GRU tuy có chỉ số sai số cao hơn một chút nhưng vẫn thể hiện năng lực tốt trong việc nắm bắt các quy luật biến thiên của chuỗi dữ liệu, chứng tỏ tiềm năng ứng dụng nếu được tinh chỉnh sâu hơn về kiến trúc [89].

Thứ hai, về tài nguyên tính toán: Ưu điểm lớn nhất của LightGBM nằm ở tốc độ xử lý và yêu cầu phần cứng thấp. So với các mô hình mạng nơ-ron hồi quy vốn đòi hỏi thời gian huấn luyện đáng kể, LightGBM cho phép triển khai nhanh chóng và phù hợp hơn với các hệ thống yêu cầu phản hồi tức thời. Tuy nhiên, LSTM và GRU vẫn là các công cụ mạnh mẽ nếu hệ thống có đủ tài nguyên tính toán để khai thác các đặc trưng phụ thuộc chuỗi phức tạp [89].

Thứ ba, định hướng cải thiện: Kết quả nghiên cứu cũng chỉ ra rằng, việc chỉ dựa vào dữ liệu lịch sử tại chỗ có những giới hạn nhất định về độ chính xác, chưa thể so sánh ngang bằng với các dịch vụ khí tượng chuyên nghiệp. Do đó, để nâng cao chất lượng dự báo, hướng tiếp cận tiềm năng là kết hợp các mô hình học máy này với nguồn dữ liệu dự báo thời tiết số (NWP) từ các đơn vị chuyên ngành, giúp bổ sung thông tin đầu vào và giảm thiểu sai số trong các điều kiện thời tiết biến động [89].

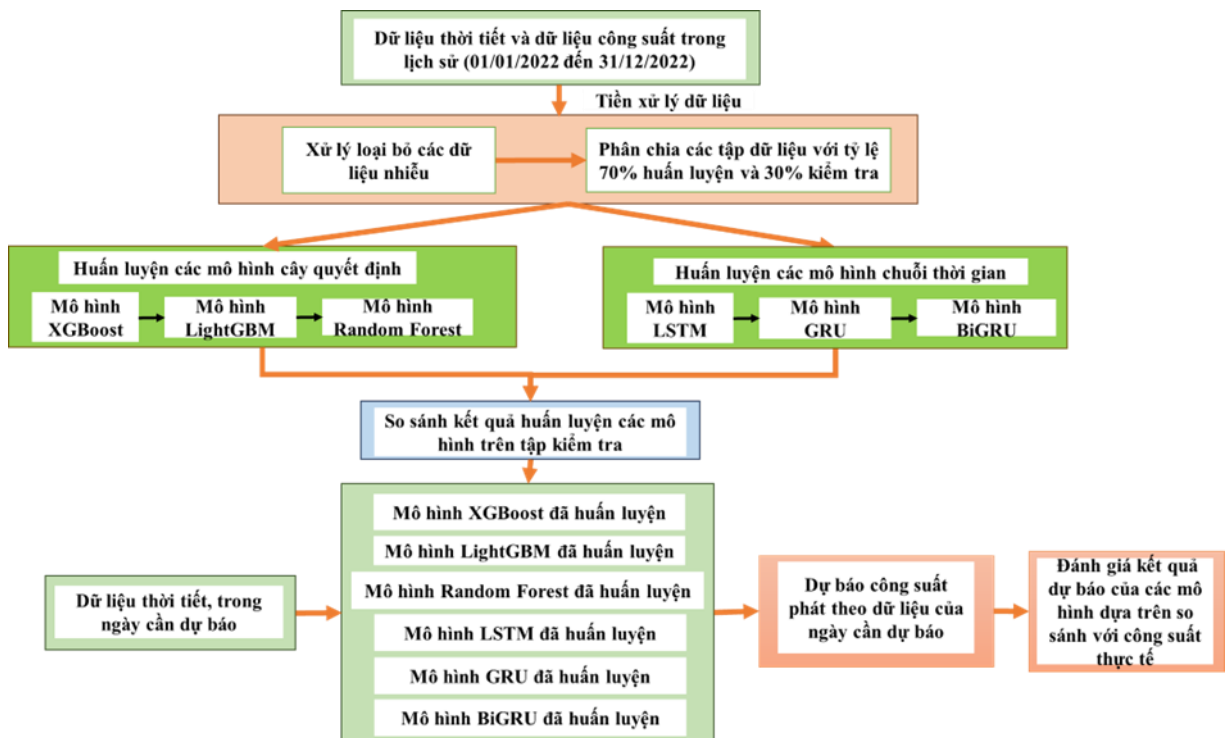
3.2. Xây dựng, huấn luyện và so sánh đánh giá các mô hình dự báo công suất phát điện mặt trời đơn lẻ

Phần này sẽ trình bày quy trình thiết lập và huấn luyện các mô hình dự báo công suất phát điện mặt trời đơn lẻ do nghiên cứu sinh thực hiện. Các mô hình được lựa chọn thuộc hai nhóm chính:

- Nhóm mô hình chuỗi thời gian: bao gồm LSTM, GRU và BiGRU là các mạng nơron hồi tiếp có khả năng học mối liên hệ phụ thuộc thời gian giữa các điểm dữ liệu.

- Nhóm mô hình cây quyết định (phi chuỗi thời gian): bao gồm XGBoost, LightGBM và Random Forest là các thuật toán cây quyết định tăng cường, có khả năng khai thác mạnh các mối quan hệ phi tuyến mà không yêu cầu dữ liệu quá khứ liên tục.
- Việc lựa chọn các mô hình này nhằm mục tiêu so sánh hiệu quả dự báo giữa hai nhóm kỹ thuật, từ đó đánh giá khả năng thích nghi của từng nhóm trong điều kiện dữ liệu thực tế tại Việt Nam nơi thường xuyên xảy ra hiện tượng thiếu hụt chuỗi dữ liệu công suất.

Bộ dữ liệu dùng để huấn luyện các mô hình dự báo là bộ dữ liệu đã được nêu trong phần 2.1.1 có sử dụng công suất phát trong lịch sử. tập dữ liệu được phân chia 70% cho huấn luyện và 30% cho kiểm tra. Sơ đồ thuật toán thực hiện so sánh các mô hình dự báo đơn lẻ được trình bày trong Hình 3.4 dưới đây:



Hình 3.4. Sơ đồ thuật toán thực hiện so sánh các mô hình dự báo các mô hình cây quyết định và các mô hình chuỗi thời gian.

3.2.1. Xây dựng các mô hình cây quyết định

Trong bài toán dự báo công suất phát điện mặt trời, các biến số khí tượng như bức xạ và nhiệt độ thường mang tính chất phi tuyến cao, biến động phức tạp và phụ thuộc mạnh mẽ vào các điều kiện ngoại cảnh đặc thù tại Việt Nam. Các mô hình hồi quy tuyến tính truyền thống thường bộc lộ hạn chế trong việc mô hình hóa các mối quan hệ đa chiều này, dẫn đến sai số lớn khi thời tiết thay đổi đột ngột [106]. Do đó, việc ứng dụng các

mô hình dựa trên cấu trúc cây quyết định (Decision Trees) trở thành một nhu cầu tất yếu nhờ khả năng phân tách không gian dữ liệu phức tạp, không yêu cầu giả định về phân phối dữ liệu và đặc biệt là tính bền vững (robustness) trước các nhiễu đo lường từ hệ thống SCADA [75].

Tuy nhiên, sự phong phú của các thuật toán thuộc họ cây quyết định đòi hỏi một sự phân tích kỹ lưỡng về cơ chế học và khả năng hội tụ để lựa chọn ra các mô hình thành phần tối ưu. Nhằm làm rõ tiêu chí lựa chọn và thiết lập cơ sở khoa học cho việc loại bỏ các thuật toán không phù hợp trong phạm vi nghiên cứu này các đặc tính kỹ thuật cốt lõi của các mô hình được tổng hợp và đối chiếu trong Bảng 3.3.

Bảng 3.3. So sánh đặc tính kỹ thuật và lý do lựa chọn các thuật toán thành phần

So sánh	Random Forest	XGBoost	LightGBM	Catboost
Cơ chế Ensemble	Bagging	Boosting	Boosting	Boosting
Cấu trúc cây	Xây dựng cây đầy đủ độc lập	Xây dựng cây theo tầng	Xây dựng cây theo chiến lược mở rộng nút lá	Cây quyết định có cấu trúc đối xứng
Xử lý dữ liệu số	Tính bền vững cao	Tối ưu nhất	Nhanh nhất	Trung bình
Thế mạnh chính	Kháng nhiễu SCADA	Chống (overfitting) tốt	Hiệu suất cao trên dữ liệu lớn	Xử lý biến định danh

Dựa trên kết quả so sánh tại Bảng 3.3, mặc dù CatBoost là một thuật toán Boosting mạnh mẽ, nhưng ưu thế cốt lõi của nó nằm ở việc xử lý các biến định danh (categorical features). Đối với tập dữ liệu dự báo công suất trong nghiên cứu này vốn hoàn toàn là dữ liệu số liên tục thì CatBoost không đem lại sự cải thiện đáng kể về độ chính xác so với XGBoost hay LightGBM [107], đồng thời lại đòi hỏi tài nguyên tính toán cao hơn do đặc thù của cấu trúc cây đối xứng. Do đó, để đảm bảo tính tinh gọn và hiệu quả tính toán, NCS không lựa chọn xây dựng mô hình CatBoost.

Bên cạnh các mô hình dựa trên cây quyết định, NCS cũng xem xét mô hình SVR (Support Vector Regression) một thuật toán học máy dựa trên hàm hạt nhân (kernel-based) và nguyên lý tối ưu hóa siêu phẳng. Tuy nhiên, mô hình SVR tồn tại một số hạn chế đáng kể về hiệu suất tính toán và độ bền vững khi áp dụng trên dữ liệu thực tế. Cụ thể, độ phức tạp tính toán của SVR tỷ lệ thuận với bình phương số lượng mẫu huấn luyện ($O(n^2)$) gây áp lực lớn lên tài nguyên tính toán khi xử lý các tập dữ liệu SCADA có mật độ cao, từ đó làm giảm khả năng đáp ứng các yêu cầu nghiêm ngặt của vận hành thời gian thực [108]. Hơn nữa, khác với cơ chế phân tách không gian dữ liệu mang tính linh hoạt và cục bộ của các thuật toán dựa trên cây quyết định, SVR tập trung vào việc ước lượng một hàm hồi quy trơn (smooth function) trên toàn bộ tập huấn luyện. Đặc điểm này khiến mô hình trở nên nhạy cảm hơn đối với các điểm dữ liệu nhiễu vốn thường xuyên phát sinh do các biến động thời tiết đột ngột và cực đoan tại Việt Nam từ đó làm suy giảm tính ổn định của mô hình và gia tăng sai số dự báo [108].

Dựa trên các phân tích đã trình bày, NCS lựa chọn xây dựng bộ ba mô hình bao gồm XGBoost, LightGBM và Random Forest. Sự lựa chọn này nhằm đảm bảo tính bao quát đối với hai cơ chế học quan trọng là Bagging và Boosting, đồng thời khai thác hiệu quả các ưu thế nổi bật của nhóm thuật toán dựa trên cây quyết định, bao gồm khả năng chống nhiễu tốt, năng lực mô hình hóa các quan hệ phi tuyến phức tạp và hiệu quả cao khi áp dụng trên dữ liệu số trong các bài toán thực tiễn.

a. Xây dựng mô hình XGBoost

+) Các tham số của mô hình XGBoost:

Nghiên cứu sinh sử dụng mô hình Optuna để tìm thông số tối ưu của mô hình XGBoost thông qua tập kiểm tra khi huấn luyện mô hình với tập dữ liệu được giới thiệu trong phần 2.1.1 Optuna [109] là một khung tối ưu hóa siêu tham số hiện đại, được thiết kế để tự động tìm kiếm các tham số tối ưu cho các mô hình học máy. Điểm mạnh của Optuna nằm ở khả năng sử dụng các phương pháp tối ưu hóa tiên tiến như TPE và CMA-ES, cho phép nhanh chóng và hiệu quả tìm ra giá trị siêu tham số tối ưu, ngay cả trong các không gian tham số phức tạp và có kích thước lớn.

Quy trình tối ưu hóa với Optuna bắt đầu bằng việc xác định không gian tìm kiếm tham số, sau đó Optuna tự động tiến hành các thử nghiệm và điều chỉnh tham số để giảm thiểu sai số dự báo (ví dụ như RMSE hoặc MAPE). Với các chiến lược tìm kiếm thích

ứng, Optuna tối ưu hóa hiệu quả hiệu suất của mô hình mà không cần phụ thuộc vào các phương pháp điều chỉnh tham số thủ công, vốn tốn nhiều thời gian và công sức.

Bảng 3.4. Các tham số mặc định và tối ưu của mô hình XGBoost

Các thông số	Thông số mặc định	Thông số tối ưu sau khi tinh chỉnh
learning_rate	0,1	0,022
max_depth	8	15
n_estimators	150	262
Min_child_weight	1	10
Gamma	0	0,745
subsample	1	0,813
colsample_bytree	1	0,819

Trong đó tham số learning_rate kiểm soát kích thước của các bước mà mô hình thực hiện khi tối ưu hóa hàm mất mát. Tỷ lệ học càng nhỏ, mô hình càng cần nhiều cây hơn nhưng có khả năng dẫn đến khả năng khái quát hóa tốt hơn. Max_depth xác định độ sâu tối đa của mỗi cây; cây sâu hơn có thể nắm bắt các mẫu phức tạp hơn nhưng cũng dễ bị quá khớp. N_estimators chỉ định số lượng cây trong mô hình. Nhiều cây hơn thường giúp tăng độ chính xác, tuy nhiên cũng có thể gây ra quá khớp và tăng nhu cầu tính toán. Min_child_weight đặt tổng trọng số tối thiểu cho một nút con; các giá trị cao hơn giúp ngăn mô hình học các mẫu quá chi tiết và giảm nguy cơ quá khớp. Gamma kiểm soát mức giảm tối thiểu trong hàm mất mát để phân tách thêm tại một nút lá; giá trị cao hơn dẫn đến mô hình đơn giản hơn và bảo thủ hơn. Subsample chỉ phần trăm dữ liệu đào tạo được chọn ngẫu nhiên cho mỗi cây; các giá trị thấp hơn có thể giảm quá khớp, nhưng nếu quá thấp sẽ gây ra hiện tượng không khớp. Cuối cùng, colsample_bytree xác định phần trăm các đặc trưng được chọn ngẫu nhiên cho mỗi cây; các giá trị thấp hơn giúp tạo ra sự đa dạng giữa các cây, giảm nguy cơ quá khớp.

Các kết quả với tập dữ liệu huấn luyện khi sử dụng thông số mặc định và thông số tối ưu được thể hiện tại bảng dưới đây

Bảng 3.5. So sánh giữa tham số mặc định và tham số tối ưu của mô hình XGBoost

Các so sánh	Thông số mặc định	Thông số tối ưu
Thời gian hoàn thành	2,22	2,17
RMSE (kW)	1397,277	1397,277

NRMSE (%)	2,823	2,823
NMAPE (%)	1,012	1,012

Như vậy khi sử dụng thông số tối ưu cho mô hình XGboost thì thời gian hoàn thành và sai số RMSE và NRMSE là tương đương nhau. Từ đó nghiên cứu sinh lựa chọn tham số tối ưu cho mô hình XGBoost.

b. Xây dựng mô hình LightGBM

Mô hình LightGBM được xây dựng giống phần 3.1.1 nhưng bổ sung thêm tham số tối ưu.

+) Các tham số của mô hình LightGBM:

Nghiên cứu sinh sử dụng mô hình Optuna để tìm thông số tối ưu của mô hình LightGBM thông qua tập test khi huấn luyện mô hình với tập dữ liệu được đề cập trong phần 2.1.1.

Bảng 3.6. Các tham số mặc định và tối ưu của mô hình LightGBM

Các thông số	Thông số mặc định	Thông số tối ưu sau khi tinh chỉnh
Learning_rate	0,1	0,106
Max_depth	8	9
Num_leaves	31	111
Min_child_samples	20	14
reg_alpha	0	0,949
reg_lambda	0	0,553
Colsample_bytree	1	0,756

Các kết quả với tập dữ liệu khi sử dụng thông số mặc định và thông số tối ưu được thể hiện tại bảng dưới đây.

Bảng 3.7. So sánh giữa tham số mặc định và tham số tối ưu của mô hình LightGBM

Các so sánh	Thông số mặc định	Thông số tối ưu
Thời gian hoàn thành	0,737	0,607
RMSE (kW)	1373,272	1373,272
NRMSE (%)	2,774	2,774
NMAPE (%)	1,004	1,004

Từ bảng kết quả trên ta thấy thời gian huấn luyện mô hình và các sai số RMSE, NMAPE và NRMSE khi sử dụng thông số tối ưu và thông số mặc định của mô hình LightGBM gần như tương đương nhau. Do đó nghiên cứu sinh lựa chọn tham số tối ưu cho mô hình LightGBM.

c. Mô hình Random Forest

+ Các tham số của mô hình Random Forest:

Nghiên cứu sinh sử dụng mô hình Optuna để tìm thông số tối ưu của mô hình Random Forest thông qua tập test khi huấn luyện mô hình với tập dữ liệu được đề cập trong phần 2.1.1.

Bảng 3.8. Các tham số mặc định và tối ưu của mô hình Random Forest

Các thông số	Thông số mặc định	Thông số tối ưu
n_estimators	100	123
Max_depth	10	30
min_samples_split	2	2
min_samples_leaf	1	3

Trong đó: Tham số n_estimators trong Random Forest xác định số lượng cây trong rừng. Nhiều cây hơn thường cải thiện hiệu suất và độ ổn định của mô hình, tuy nhiên sẽ làm tăng nhu cầu về tài nguyên tính toán. Max_depth đặt độ sâu tối đa của mỗi cây; cây càng sâu có thể nắm bắt các mẫu phức tạp hơn nhưng cũng dễ bị quá khớp hơn. Tham số min_samples_split xác định số lượng mẫu tối thiểu cần thiết để chia một nút, giúp kiểm soát độ phát triển của cây. Các giá trị cao hơn có thể ngăn cây trở nên quá phức tạp và giảm nguy cơ quá khớp. Min_samples_leaf quy định số lượng mẫu tối thiểu phải có trong một nút lá, giúp đảm bảo cây không trở nên quá tinh chỉnh với dữ liệu đào tạo, giảm thiểu nguy cơ quá khớp. Những tham số này kết hợp để cân bằng độ phức tạp và khả năng khái quát hóa của mô hình, giúp Random Forest đạt hiệu suất tốt trên dữ liệu chưa thấy.

Các kết quả với tập dữ liệu thời tiết khi sử dụng thông số mặc định và thông số tối ưu được thể hiện tại bảng dưới đây

Bảng 3.9. So sánh giữa tham số mặc định và tham số tối ưu của mô hình Random Forest

Các so sánh	Thông số mặc định	Thông số tối ưu
-------------	-------------------	-----------------

Thời gian hoàn thành	16,58	55,402
RMSE (kW)	1431,01	1398,010
NRMSE (%)	2,888	2,888
NMAPE (%)	1,033	1,033

Từ bảng trên có thể thấy rằng việc điều chỉnh các tham số của mô hình Random Forest để tối ưu hiệu suất đã dẫn đến cải thiện trong việc giảm RMSE và NRMSE. Tuy nhiên, điều này đi kèm với tăng thời gian hoàn thành của quá trình huấn luyện mô hình. Trong quá trình cân nhắc giữa việc giảm sai số và việc tăng thời gian tính toán, nghiên cứu sinh đã quyết định sử dụng các tham số mặc định cho mô hình Random Forest.

3.2.2. Xây dựng các mô hình chuỗi thời gian

Các mô hình chuỗi thời gian LSTM, GRU và BiGRU có cấu trúc như được trình bày trong phần 2.3.1.3. Để tiết kiệm thời gian huấn luyện nghiên cứu sinh sử dụng các mô hình BiGRU, GRU và LSTM với một lớp ẩn có 64 nơron có cấu hình huấn luyện như sau:

+) **Số look-back window:** 288 điểm dữ liệu quá khứ tương ứng với 24 giờ quá khứ (tần suất lấy mẫu là 5 phút/lần)

Chức năng: là thông số dữ liệu quá khứ được sử dụng để tích hợp vào dự báo chuỗi thời gian:

+) **Thuật toán tối ưu hóa:** Adam

Chức năng: thuật toán tối ưu Adam được sử dụng để điều chỉnh trọng số của mô hình dựa trên độ lớn của gradient. Adam giúp mô hình hội tụ nhanh hơn so với các thuật toán tối ưu truyền thống và giảm nguy cơ rơi vào các điểm cực tiểu cục bộ.

+) **Hàm kích hoạt:** tanh

Chức năng: hàm kích hoạt tanh được sử dụng để giới hạn đầu ra của mô hình trong khoảng $[-1, 1]$. Giới hạn đầu ra giúp mô hình học được mối quan hệ phức tạp và chủ động trong việc xử lý dữ liệu đầu vào.

+) **Hàm sai số (Loss function):** MSE

Chức năng: MSE được sử dụng để đo lường độ lớn của sai số giữa dự báo của mô hình và giá trị thực tế. MSE là một hàm phổ biến trong các nhiệm vụ dự báo, giúp mô hình học được các thay đổi nhỏ và đồng thời giảm độ lớn của sai số.

+) **Số lần huấn luyện (epochs):** 50

Chức năng: epochs là số lần huấn luyện trên toàn bộ dữ liệu. Số lượng epochs lớn giúp mô hình có cơ hội học được nhiều hơn từ dữ liệu, tăng khả năng hiểu biết và dự báo.

+) Kích thước dữ liệu cập nhật trọng số (Batch size): 32

Chức Năng: batch size là số lượng mẫu dữ liệu được sử dụng trong mỗi lần cập nhật trọng số (iteration). Việc sử dụng batch size giúp tăng tốc độ huấn luyện và giảm áp lực tính toán, đồng thời duy trì tính đại diện của dữ liệu trong quá trình cập nhật trọng số.

3.2.3. So sánh kết quả huấn luyện và dự báo của các mô hình

Bảng 3.10 dưới đây trình bày so sánh các kết quả huấn luyện trên tập kiểm tra của các mô hình dự báo:

Bảng 3.10. So sánh các kết quả huấn luyện của từng mô hình với tập dữ liệu lịch sử từ 01/01/2022 đến 31/12/2022 (70% tập dữ liệu được dùng cho đào tạo và 30% cho test)

Các so sánh	Mô hình XGBoost	Mô hình LightGBM	Mô hình Random Forest	Mô hình LSTM	Mô hình GRU	Mô hình BiGRU
Thời gian hoàn thành (s)	2,12	0,607	16,58	594,512	609,591	562,68
RMSE (kW)	1397,277	1373,272	1431,003	1670,691	1717,464	1554,82
NRMSE (%)	2,823	2,774	2,888	3,387	3,469	3,141
NMAPE (%)	1,012	1,004	1,033	1,35	1,521	1,18
MAPE (%)	Không xác định	Không xác định	Không xác định	Không xác định	Không xác định	Không xác định

Kết quả huấn luyện các mô hình XGBoost, LightGBM, Random Forest, LSTM, GRU và BiGRU được trình bày trong Bảng 3.10 cho thấy sự khác biệt rõ rệt về cả hiệu suất huấn luyện và độ chính xác dự báo. Trong số các mô hình học máy, LightGBM nổi bật với thời gian huấn luyện nhanh nhất (0,607 giây) và sai số thấp nhất (RMSE = 1373,272 kW, NMAPE = 1,004%), thể hiện sự phù hợp với các ứng dụng cần tính phản hồi nhanh. XGBoost có hiệu quả gần tương đương về độ chính xác, nhưng chậm hơn đôi chút trong thời gian huấn luyện. Ngược lại, Random Forest tuy có khả năng tổng quát hóa tốt, song thời gian huấn luyện dài (16,58 giây) khiến nó kém hiệu quả hơn khi xử lý lượng lớn dữ liệu hoặc khi cần huấn luyện lại thường xuyên. Đối với nhóm mô hình học sâu, BiGRU cho kết quả khá hơn so với GRU và LSTM với RMSE = 1554,82

kW và $NMAPE = 1,18\%$, tuy nhiên tất cả đều yêu cầu thời gian huấn luyện rất lớn (trên 560 giây), trong khi sai số lại không thấp hơn so với các mô hình học máy thông thường. Điều này cho thấy rằng, trong điều kiện dữ liệu đầy đủ cả năm và không yêu cầu xử lý chuỗi quá phức tạp, các mô hình cây quyết định như LightGBM, XGBoost và RF vẫn là lựa chọn ưu tiên nhờ thời gian huấn luyện, hiệu suất và độ ổn định cao. Khi tính toán MAPE dựa trên công thức (1.3), có xuất hiện các giá trị công suất bằng 0 điều này làm cho kết quả tính toán trở nên không xác định. Để giải quyết vấn đề này, nghiên cứu sinh đã quyết định loại bỏ các điểm dữ liệu có công suất bằng 0 khi sử dụng các tập dữ liệu tính toán với sai số MAPE.

Để đánh giá khả năng dự báo tức thời của các mô hình trong điều kiện thực tế, nghiên cứu sinh lựa chọn ba ngày liên tiếp từ 01 đến 03/03/2021 (tập dữ liệu được sử dụng trong dự báo bức xạ tại chương 2) làm tập dữ liệu thử nghiệm chính. Với mục tiêu kiểm tra khả năng dự báo tức thời không phụ thuộc vào chuỗi dữ liệu quá khứ. Đồng thời, nghiên cứu sinh cũng thực hiện đánh giá ảnh hưởng của dữ liệu bức xạ được dự báo bằng mô hình LightGBM (trình bày tại phần 3.1) so với dữ liệu bức xạ thực tế, đối với kết quả dự báo công suất phát của các mô hình đơn lẻ. Việc này xuất phát từ thực tế rằng mô hình LightGBM đạt độ chính xác ổn định hơn hai mô hình LSTM và GRU khi dự báo bức xạ mặt trời với tập dữ liệu từ ngày 01 đến 03/03/2021. Các mô hình được áp dụng để đưa ra dự báo công suất phát điện trong từng khoảng thời gian trong ba ngày, dựa hoàn toàn trên thông số thời tiết tại thời điểm hiện tại và không có dữ liệu công suất quá khứ làm đầu vào.

Cách thiết kế thử nghiệm này nhằm làm rõ sự khác biệt giữa hai nhóm mô hình:

- Các mô hình chuỗi thời gian (như LSTM, GRU, BiGRU), vốn yêu cầu chuỗi dữ liệu lịch sử liên tục để duy trì độ chính xác;
- Và các mô hình cây quyết định (XGBoost, LightGBM, Random Forest), có khả năng dự báo tức thời một cách linh hoạt và không phụ thuộc vào dữ liệu quá khứ.

Kết quả dự báo cho các ngày từ 01 đến 03/03/2021 được thể hiện trong Bảng 3.11, kết quả này sẽ là cơ sở để đánh giá hiệu năng các mô hình dự báo, cũng như làm nổi bật tính khả thi trong ứng dụng thực tế của mô hình tổ hợp được đề xuất ở chương tiếp theo.

Bảng 3.11. So sánh sai số dự báo công suất phát điện mặt trời giữa các mô hình đơn lẻ trong giai đoạn 01-03/03/2021 trong hai trường hợp sử dụng bức xạ dự báo từ mô hình LightGBM tại mục 3.1 và sử dụng bức xạ thực tế.

Sử dụng bức xạ thực tế						
Các sai số	XGB	LGBM	RF	GRU	LSTM	BiGRU
RMSE (kW)	1457,02	1438,34	1483,57	9489,06	9489,09	1941,6
NRMSE (%)	2,94	2,94	3,00	19,16	19,16	3,92
MAPE (%)	59,31	59,8	59,36	99,5	99,5	124,11
NMAPE (%)	1,26	1,25	1,3	9,05	9,05	1,81
Sử dụng bức xạ dự báo từ mô hình LightGBM						
Các sai số	XGB	LGBM	RF	GRU	LSTM	BiGRU
RMSE (kW)	1591,75	1534,36	1591,03	9489,41	9489,13	1636,78
NRMSE (%)	3,21	3,1	3,21	19,171	19,17	3,31
MAPE (%)	144,23	145,63	151,3	99,51	99,6	178,8
NMAPE (%)	1,87	1,82	1,88	9,06	9,06	1,97

Bảng 3.11 cung cấp cái nhìn tổng quát về hiệu suất của các mô hình dự báo công suất trong hai tình huống phổ biến: sử dụng bức xạ thực tế và bức xạ dự báo làm đầu vào. Đồng thời, bảng này cũng phản ánh rõ hiệu quả của từng mô hình khi áp dụng trong kịch bản dự báo tức thời, tức là không có chuỗi dữ liệu công suất quá khứ. Kết quả cho thấy, nhóm mô hình cây quyết định (XGBoost, LightGBM, Random Forest) giữ được hiệu suất tốt và ổn định trong cả hai tình huống. Khi sử dụng bức xạ thực tế, các chỉ số RMSE, NRMSE và NMAPE đều đạt mức thấp (RMSE $\approx$ 1450 kW, NRMSE $\approx$ 2,9%, NMAPE $\approx$ 1,25%), cho thấy độ chính xác cao. Dù sai số tăng khi chuyển sang bức xạ dự báo (RMSE tăng lên $\sim$ 1590 kW, MAPE tăng đáng kể), nhóm mô hình này vẫn duy trì được tính ổn định và khả năng tổng quát hóa. Ngược lại, hai mô hình chuỗi thời gian là GRU và LSTM thể hiện hiệu suất kém rõ rệt trong cả hai tình huống. Việc thiếu chuỗi dữ liệu công suất quá khứ vốn là điều kiện thiết yếu để các mô hình này phát huy hiệu quả, khiến chúng gần như mất khả năng học. RMSE và MAPE đều đạt mức rất cao, với RMSE > 9400 kW và MAPE > 99%, bất kể đầu vào là bức xạ thực tế hay dự báo. Mặc dù mô hình BiGRU có cấu trúc hai chiều giúp cải thiện một số chỉ số so với hai mô hình LSTM và GRU nhưng sai số vẫn lớn hơn các mô hình cây quyết định.

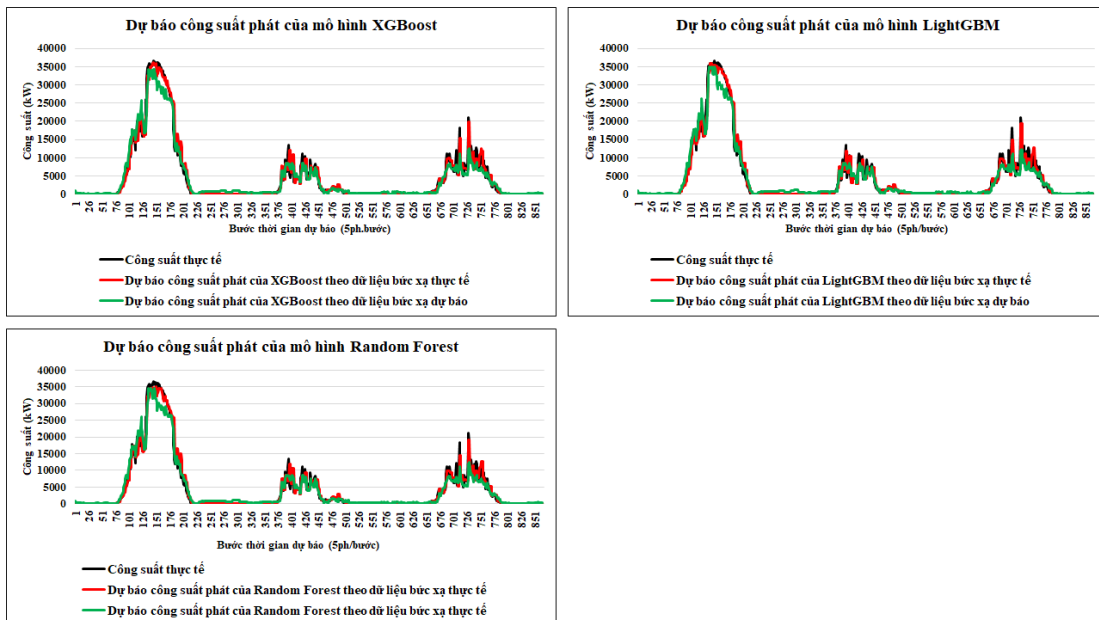
Tổng hợp lại, Bảng 3.11 đồng thời khẳng định hai điều quan trọng:

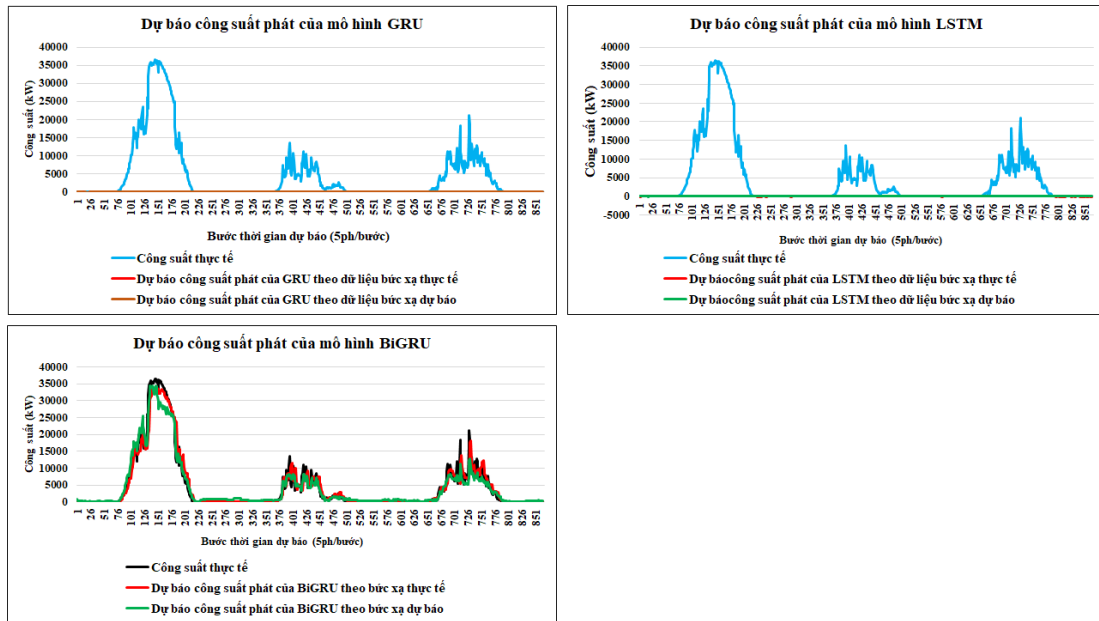
- Sai số trong dự báo công suất tăng rõ rệt khi sử dụng dữ liệu bức xạ dự báo thay vì bức xạ thực tế (đặc biệt là sai số MAPE). Điều này phản ánh mức độ ảnh hưởng của

sai số đầu vào trong chuỗi dự báo liên hoàn, đồng thời cho thấy vai trò then chốt của chất lượng dữ liệu bức xạ trong toàn bộ hệ thống dự báo công suất.

- Trong điều kiện dự báo tức thời, thiếu dữ liệu quá khứ hoặc chuỗi dữ liệu trong quá khứ bị ngắt quãng, các mô hình cây quyết định vượt trội hơn rõ rệt so với các mô hình học sâu theo chuỗi, và do đó phù hợp hơn cho triển khai thực tế tại các nhà máy điện mặt trời có dữ liệu quá khứ không phải lúc nào cũng đầy đủ.

Nhằm minh họa trực quan sự khác biệt về hiệu năng giữa các mô hình, đồng thời kiểm chứng khả năng bám sát thực tế của từng phương pháp dự báo, luận án tiến hành so sánh giá trị công suất phát điện thực tế với giá trị dự báo của các mô hình qua từng khoảng thời gian trong ngày. Các biểu đồ sau thể hiện kết quả dự báo trong ba ngày liên tiếp (từ 01 đến 03/03/2021) giai đoạn được chọn làm tập thử nghiệm điển hình, đại diện cho điều kiện vận hành tại nhà máy điện mặt trời. Trục hoành thể hiện các mốc thời gian trong ngày, trục tung là giá trị công suất (kW), mỗi đường biểu diễn thể hiện kết quả dự báo của một mô hình, kèm theo đường biểu diễn giá trị thực tế làm chuẩn tham chiếu. Việc đối chiếu trực quan này cho phép đánh giá mức độ bám sát xu hướng, khả năng bắt đỉnh và đáy công suất, cũng như độ trễ phản ứng của từng mô hình, từ đó củng cố thêm cơ sở cho các nhận định định lượng đã trình bày ở phần trước.





Hình 3.5. Biểu đồ so sánh hiệu suất dự báo công suất phát theo từng bước thời gian giữa các mô hình cây quyết định và chuỗi thời gian trong điều kiện thiếu dữ liệu quá khứ, dự báo theo bức xạ thực tế và bức xạ dự báo

Dựa trên đồ thị tại Hình 3.5, có thể thấy rõ sự khác biệt về đặc điểm và hiệu suất dự báo giữa hai nhóm mô hình trong hai điều kiện đầu vào là bức xạ thực tế và bức xạ dự báo.

Các mô hình chuỗi thời gian như GRU và LSTM thể hiện dấu hiệu đường dự báo bị phẳng, gần như không phản ánh được dao động công suất theo thời gian, cả khi sử dụng bức xạ thực tế lẫn bức xạ dự báo. Đây là hệ quả trực tiếp của việc thiếu chuỗi dữ liệu quá khứ điều kiện bắt buộc để các mô hình này học được quy luật của các đặc trưng đầu vào theo thời gian. Sai số dự báo trong cả hai kịch bản đều rất lớn, đặc biệt khi sử dụng bức xạ dự báo, đường cong hầu như không theo được nhịp biến động công suất, phản ánh sự suy giảm nghiêm trọng về khả năng tổng quát hóa trong điều kiện dữ liệu không lý tưởng.

Ngược lại, mô hình BiGRU, mặc dù vẫn thuộc nhóm học chuỗi thời gian, lại thể hiện kết quả khả quan hơn đáng kể. Đường dự báo của BiGRU cho thấy khả năng bám sát xu hướng công suất thực tế trong cả hai điều kiện bức xạ. Sự vượt trội này đến từ kiến trúc hai chiều (bidirectional), cho phép mô hình khai thác cả thông tin trong và ngoài phạm vi thời điểm hiện tại để dự báo. Tuy nhiên, sai số vẫn ở mức cao, đặc biệt là về biên độ dao động cho thấy BiGRU bắt đúng hình dạng nhưng sai về biên độ, khiến MAPE và NMAPE vẫn ở mức cao, điều này gây bất lợi cho vận hành thực tế vốn đòi hỏi độ chính xác tức thời cao.

Đối lập với nhóm học sâu, các mô hình cây quyết định như XGBoost, LightGBM và Random Forest cho thấy đường dự báo bám sát tốt đường thực tế, cả về hình dạng lẫn biên độ. Dù có sự suy giảm nhẹ khi sử dụng bức xạ dự báo thay vì thực tế, các mô hình này vẫn duy trì được đường cong mượt, phản ánh đúng đỉnh và đáy công suất trong ngày. Lý do là vì các mô hình cây không phụ thuộc vào chuỗi thời gian, mà học mối liên hệ trực tiếp giữa các đặc trưng đầu vào (nhiệt độ, thời gian, bức xạ...) và công suất đầu ra tại từng thời điểm. Do đó, chúng ít bị ảnh hưởng bởi sự ngắt đoạn của dữ liệu, đồng thời tận dụng tốt các đặc trưng tức thời vốn là dạng dữ liệu phổ biến tại nhiều nhà máy điện mặt trời ở Việt Nam.

Một điểm đáng lưu ý là dù tất cả các mô hình đều có sai số tăng khi chuyển từ bức xạ thực tế sang bức xạ dự báo, nhóm cây quyết định vẫn giữ được cấu trúc đường cong hợp lý, trong khi nhóm mô hình chuỗi gần như bị triệt tiêu phản ứng thể hiện rõ rệt trong các khung giờ đỉnh nắng và biến động mây. Điều này càng củng cố vai trò ưu việt của mô hình cây trong môi trường dữ liệu nhiễu và thiếu như thực tế hiện nay.

Từ các quan sát trên, có thể rút ra kết luận rằng:

Mô hình cây quyết định tuy không học được chu kỳ phức tạp như mô hình chuỗi thời gian, nhưng lại bù đắp bằng độ chính xác tức thời, khả năng ổn định khi thiếu dữ liệu, và tính linh hoạt trong triển khai. Trong khi đó, các mô hình học sâu chuỗi thời gian chỉ thực sự phát huy hiệu quả khi có dữ liệu liên mạch và chất lượng cao điều khó đạt được tại phần lớn các nhà máy điện mặt trời tại Việt Nam hiện nay.

Do đó, việc lựa chọn XGBoost, LightGBM và Random Forest làm nền tảng cho các mô hình tổ hợp ở các bước sau, nhằm cân bằng giữa độ chính xác, tính ổn định và khả năng ứng dụng thực tiễn, là một hướng đi phù hợp và có cơ sở khoa học rõ ràng.

3.3. Kết hợp giữa mô hình cây quyết định và mô hình chuỗi thời gian để cải thiện hiệu năng dự báo

Sau khi đánh giá hiệu năng của các mô hình đơn lẻ trong điều kiện thiếu chuỗi dữ liệu quá khứ, kết quả thực nghiệm đã chỉ ra sự phân hóa rõ rệt về mức độ thích ứng giữa hai nhóm mô hình cây quyết định và học sâu chuỗi thời gian. Trong khi nhóm mô hình cây thể hiện tính ổn định cao, sai số thấp và khả năng hoạt động tốt ngay cả khi dữ liệu không đầy đủ, thì nhóm mô hình chuỗi lại phụ thuộc mạnh mẽ vào độ liên mạch của dữ liệu để duy trì hiệu suất dự báo.

Tuy nhiên, mỗi nhóm mô hình lại sở hữu những thế mạnh riêng biệt về mặt kỹ thuật: mô hình cây quyết định phù hợp với các bài toán có tính phân mảnh dữ liệu cao, trong khi mô hình chuỗi thời gian có khả năng học được các quy luật biến thiên phức tạp theo thời gian nếu được cung cấp đủ dữ liệu lịch sử. Chính vì vậy, việc khai thác tính bổ sung lẫn nhau giữa hai hướng tiếp cận này là một hướng đi tiềm năng để cải thiện hiệu năng dự báo tổng thể trong các hệ thống thực tế.

Mục 3.3 sẽ đi sâu phân tích các khía cạnh ưu – nhược của từng nhóm mô hình, từ đó xây dựng và thử nghiệm kết hợp đó là: ứng dụng mô hình cây quyết định để lấp đầy dữ liệu thiếu nhằm cải thiện đầu vào cho mô hình chuỗi thời gian.

3.3.1. Phân tích ưu, nhược điểm của các mô hình trong điều kiện dữ liệu không liên tục

Dựa trên phân tích định lượng (sai số dự báo) và định tính (đường biểu diễn công suất) đã trình bày trong Bảng 3.10, Bảng 3.11 và Hình 3.5, có thể thấy rõ sự khác biệt về cơ chế hoạt động và mức độ thích ứng giữa các nhóm mô hình đơn lẻ trong điều kiện dữ liệu chuỗi không đầy đủ. Các mô hình cây quyết định như LightGBM, XGBoost và Random Forest sở hữu nhiều ưu điểm nổi bật khi áp dụng trong môi trường dữ liệu không liên tục, đặc biệt trong lĩnh vực dự báo công suất điện mặt trời. Do không phụ thuộc vào tính chuỗi của dữ liệu, các mô hình này có thể hoạt động hiệu quả ngay cả khi dữ liệu quá khứ bị ngắt quãng hoặc thiếu hụt. Độ nhạy của chúng với dữ liệu bị thiếu là rất thấp, trong khi khả năng khái quát và chống nhiễu lại khá cao. Đặc biệt, các mô hình cây còn có tính diễn giải tốt và dễ triển khai trong thực tế, nhờ vào cấu trúc phân nhánh rõ ràng và khả năng xử lý quan hệ phi tuyến giữa các biến đầu vào. Tuy nhiên, hạn chế lớn của nhóm mô hình này là không thể học được các chu kỳ hoặc xu hướng dài hạn trong dữ liệu yếu tố vốn rất quan trọng trong các bài toán dự báo theo thời gian.

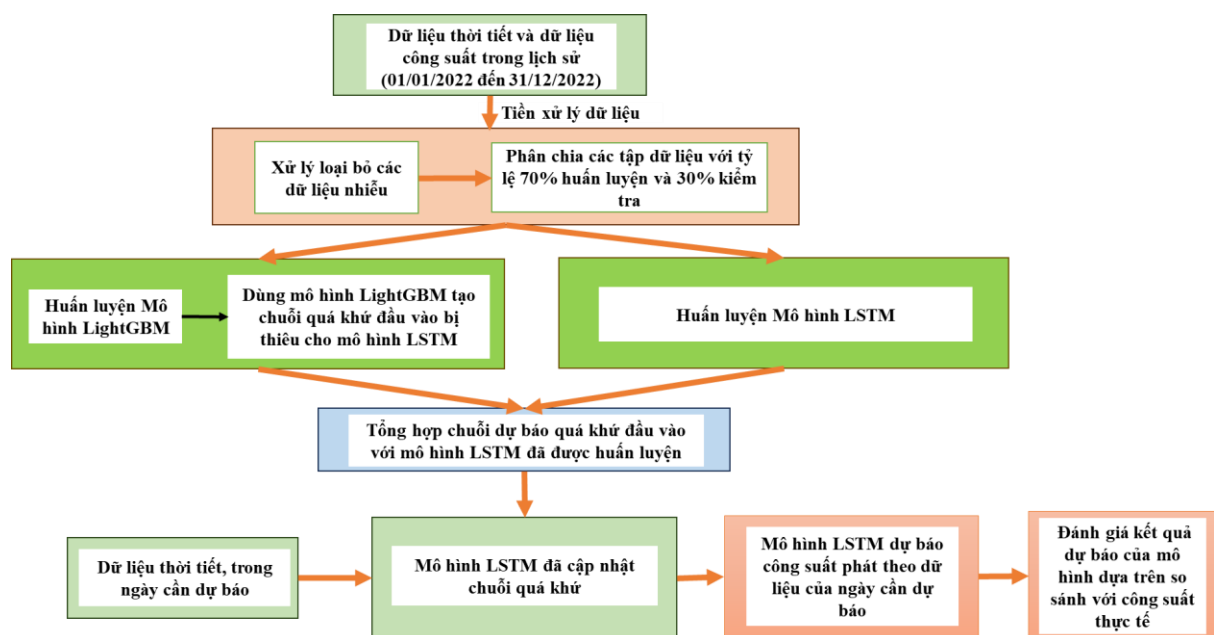
Ngược lại, các mô hình học sâu chuỗi thời gian như LSTM, GRU hay BiGRU được thiết kế để nắm bắt tốt mối liên hệ dài hạn và mô hình hóa được các chu kỳ biến động phức tạp. Chúng có khả năng học sâu hơn về cấu trúc ẩn trong chuỗi thời gian, từ đó cho kết quả dự báo sát xu hướng và ổn định trong điều kiện lý tưởng. Tuy nhiên, hạn chế nghiêm trọng nhất của nhóm mô hình này là sự phụ thuộc tuyệt đối vào tính liên tục của dữ liệu. Khi chuỗi dữ liệu bị thiếu hoặc nhiễu, mô hình dễ rơi vào trạng thái underfitting, dẫn đến sai số lớn và khả năng học kém.

Từ những phân tích trên, có thể thấy rằng: mô hình cây quyết định và mô hình chuỗi thời gian mỗi loại đều có thể mạnh riêng, và việc khai thác hợp lý sự bổ trợ lẫn nhau giữa hai hướng tiếp cận sẽ là tiền đề quan trọng cho các phương pháp lai ghép (hybrid models) nhằm tăng độ chính xác và độ tin cậy trong môi trường dữ liệu thực tế nhiều biến động.

3.3.2. Ứng dụng LightGBM để tái cấu trúc chuỗi đầu vào LSTM trong điều kiện dữ liệu không hoàn chỉnh

Để khai thác ưu điểm thời gian huấn luyện nhanh, khả năng tổng quát hóa tốt và bền với dữ liệu nhiễu của mô hình LightGBM và khả năng ghi nhớ chuỗi thời gian dài hạn của mô hình LSTM, nghiên cứu sinh tiến hành thử nghiệm theo kịch bản kết hợp mô hình LightGBM với LSTM. Trong bối cảnh dữ liệu công suất phát điện có thể bị thiếu tại một số thời điểm, nghiên cứu sinh lựa chọn sử dụng mô hình LightGBM để lấp đầy các điểm dữ liệu công suất bị thiếu, thay vì để mô hình chuỗi thời gian như LSTM tự dự báo nối tiếp. Cách tiếp cận này nhằm hạn chế hiện tượng tích lũy sai số một vấn đề phổ biến trong dự báo đa bước liên tiếp của các mô hình học sâu. Cụ thể, LightGBM được sử dụng để dự báo công suất phát cho ngày 28/02/2021 đây là thời điểm bị thiếu dữ liệu công suất, và kết quả dự báo này được đưa vào làm tập dữ liệu quá khứ (theo đúng look-back window là 24 giờ) để phục vụ mô hình LSTM trong việc dự báo công suất phát cho các ngày từ 01 đến 03/03/2021, như đã trình bày tại mục 3.1.

Kịch bản cụ thể được trình bày trong Hình 3.6 sau đây:



Hình 3.6. Sơ đồ thuật toán kết hợp đầu vào của mô hình LightGBM-LSTM

+) Các tập dữ liệu được sử dụng:

- Tập dữ liệu huấn luyện gồm tập dữ liệu từ 01/01/2022 đến 31/12/2022 đã được tiền xử lý dữ liệu ở phần 2.1.1 với các đặc trưng thời tiết gồm: bức xạ mặt trời, nhiệt độ môi trường, nhiệt độ tấm pin và công suất phát trong lịch sử. Tập dữ liệu huấn luyện sẽ được chia thành tỷ lệ 70% cho huấn luyện và 30% cho kiểm tra.
- Tập dữ liệu quá khứ ngày 28/02/2021: chỉ có dữ liệu thời tiết, thiếu toàn bộ giá trị công suất phát. Để thể hiện chi tiết hơn sự ảnh hưởng tới kết quả dự báo của các mô hình trong điều kiện thiếu dữ liệu quá khứ, nghiên cứu sinh sẽ áp dụng kịch bản lần lượt dự báo trong điều kiện tập dữ liệu quá khứ thiếu theo các tỷ lệ 10%, 20%, 30%, 50% và 100%.
- Tập dữ liệu dự báo: gồm các ngày 01–03/03/2021, chỉ có dữ liệu thời tiết bao gồm các thông số bức xạ thực tế và bức xạ dự báo theo mô hình LightGBM ở phần 3.1, nhiệt độ môi trường, nhiệt độ tấm pin và không có dữ liệu công suất.

+) Các bước thực hiện

- Xử lý dữ liệu đầu vào giống phần 3.1, phân chia các tập dữ liệu với tỷ lệ 70% cho huấn luyện và 30% cho kiểm tra.
- Huấn luyện mô hình LightGBM và LSTM trên tập huấn luyện.
- Dự báo cho ngày 28/02/2021: sử dụng mô hình LightGBM đã được huấn luyện tiến hành dự báo công suất phát theo tỷ lệ bị mất 10%, 20%, 30%, 50% và 100% dựa trên các thông số thời tiết của ngày 28/02/2021.
- Ghép chuỗi dữ liệu cho mô hình LSTM: ghép ngày 28/02/2021 vừa được dự báo từ mô hình LightGBM với tập dữ liệu dự báo từ 01-03/03/2021 để tạo thành chuỗi quá khứ với đầy đủ dữ liệu công suất, làm tập dữ liệu quá khứ cho LSTM.
- Dự báo công suất cho các ngày 01- 03/03/2021 bằng mô hình LSTM: mô hình LSTM được huấn luyện lại với chuỗi hoàn chỉnh và sử dụng đầu vào là dữ liệu thời tiết từ 01- 03/03/2021.

Kết quả thử nghiệm sử dụng mô hình LightGBM-LSTM dự báo cho các ngày 01-03/03/2021 được thể hiện trong Bảng 3.12 dưới đây. Nghiên cứu sinh tiến hành so sánh hiệu năng của mô hình kết hợp LightGBM–LSTM với ba mô hình cây quyết định (XGBoost, LightGBM và Random Forest) trong điều kiện dữ liệu đầu vào bị thiếu chuỗi quá khứ theo tỷ lệ, đồng thời đối chiếu với mô hình LSTM trong hai kịch bản: (1) không có chuỗi dữ liệu quá khứ và (2) có đầy đủ chuỗi dữ liệu quá khứ (điều kiện lý tưởng).

Mục tiêu của so sánh này là đánh giá tính ổn định và hiệu quả của mô hình dự báo kết hợp trong bối cảnh dữ liệu đầu vào bị thiếu hụt, đây là một vấn đề hay gặp trong thực tế dự báo công suất phát điện mặt trời:

Bảng 3.12. So sánh sai số dự báo (RMSE, NRMSE, MAPE, NMAPE) giữa mô hình đơn và LightGBM–LSTM theo tỷ lệ thiếu dữ liệu công suất (0–100%) trong các ngày 01–03/03/2021, dưới hai kịch bản bức xạ: thực tế và dự báo (LightGBM)

Kết quả dự báo với thông số bức xạ thực tế					
Mô hình	Tỷ lệ dữ liệu quá khứ bị thiếu (%)	RMSE (kW)	NRMSE (%)	MAPE (%)	NMAPE (%)
LSTM (đầy đủ dữ liệu quá khứ)	0	1460,52	2,95	41,04	1,40
XGBoost	0-100	1510,83	3,10	64,47	1,30
LightGBM	0-100	1431,22	2,90	57,73	1,20
Random Forest	0-100	1493,81	3,01	65,02	1,30
LSTM (không được lấp đầy dữ liệu quá khứ)	10	2268,99	4,58	42,92	1,86
	20	3188,13	6,44	48,93	2,59
	30	3872,06	7,82	55,62	3,18
	50	5225,56	10,56	62,67	4,42
	100	8940,90	18,06	97,41	8,58
LightGBM-LSTM	10	1513,61	3,06	42,18	1,45
	20	1519,97	3,07	46,55	1,46
	30	1553,51	3,14	45,37	1,48
	50	1712,14	3,46	49,06	1,6
	100	1977,93	4,06	62,62	1,85
Kết quả dự báo với thông số bức xạ dự báo từ mô hình LightGBM					
Mô hình	Tỷ lệ thiếu dữ liệu quá khứ (%)	RMSE (kW)	NRMSE (%)	MAPE (%)	NMAPE (%)

LSTM (đầy đủ dữ liệu quá khứ)	0	1447,32	2,92	115,79	1,63
XGBoost	0-100	1562,32	3,16	138,98	1,80
LightGBM	0-100	1543,65	3,12	138,04	1,81
Random Forest	0-100	1535,78	3,10	163,65	1,85
LSTM (không được lấp đầy dữ liệu quá khứ)	10	2231,18	4,51	116,49	2,07
	20	3255,04	6,58	115,47	2,68
	30	3753,65	7,58	119,56	3,25
	50	5007,9	10,12	127,86	4,45
	100	8902,82	17,99	151,08	8,60
LightGBM-LSTM	10	1491,53	3,01	121,33	1,69
	20	1561,36	3,15	130,03	1,76
	30	1650,54	3,33	132,73	1,85
	50	1666,44	3,37	143,67	1,98
	100	2064,96	4,17	192,64	2,41

Từ Bảng 3.12 cho thấy với việc lấp đầy chuỗi công suất thiếu bằng LightGBM rồi dự báo bằng LSTM giúp giải quyết nhược điểm của LSTM khi chuỗi đầu vào bị thiếu hoặc đứt đoạn, nhờ đó giữ sai số ở mức thấp trong dải thiếu thấp và trung bình, mô hình LightGBM-LSTM vượt trội so với LSTM không lấp đầy khi tỷ lệ thiếu dữ liệu quá khứ cao. Cụ thể, với bức xạ thực tế, ở mức thiếu 10–30% mô hình LightGBM–LSTM đạt NRMSE 3,06–3,14% và NMAPE 1,45–1,48%, trong khi LSTM không lấp đầy vọt lên 4,58–7,82% (NMAPE 1,86–3,18%); với bức xạ dự báo (LightGBM), bức tranh tương tự: 3,01–3,33% (NMAPE 1,69–1,85%) so với 4,51–7,58% (NMAPE 2,07–3,25%). Khi thiếu $\geq 50\%$, LightGBM–LSTM có suy giảm (NMAPE thực tế: 3,46 – 4,06%; dự báo: 3,37 – 4,17%), nhưng vẫn tốt hơn đáng kể so với LSTM không lấp đầy (NMAPE thực tế: 10,56 – 18,06%; dự báo: 10,12 – 17,99%). Mô hình LightGBM–LSTM luôn đạt sai số NMAPE thấp hơn ngưỡng yêu cầu của NSMO ($\leq 15\%$) trong tất cả các kịch bản khảo sát. Đặc biệt, ngay cả trong trường hợp thiếu hụt hoàn toàn dữ liệu quá khứ (100%), tương đương với việc mất toàn bộ look-back window 24 giờ (1 ngày quá khứ) của LSTM, mô hình vẫn duy trì độ chính xác cao với sai số 1,85% khi sử dụng bức xạ thực tế và 2,41% khi sử dụng bức xạ dự báo, cho thấy LightGBM vẫn có hiệu suất tốt và đáng tin cậy trong việc lấp đầy phần dữ liệu công suất bị thiếu dựa trên các đặc trưng khí

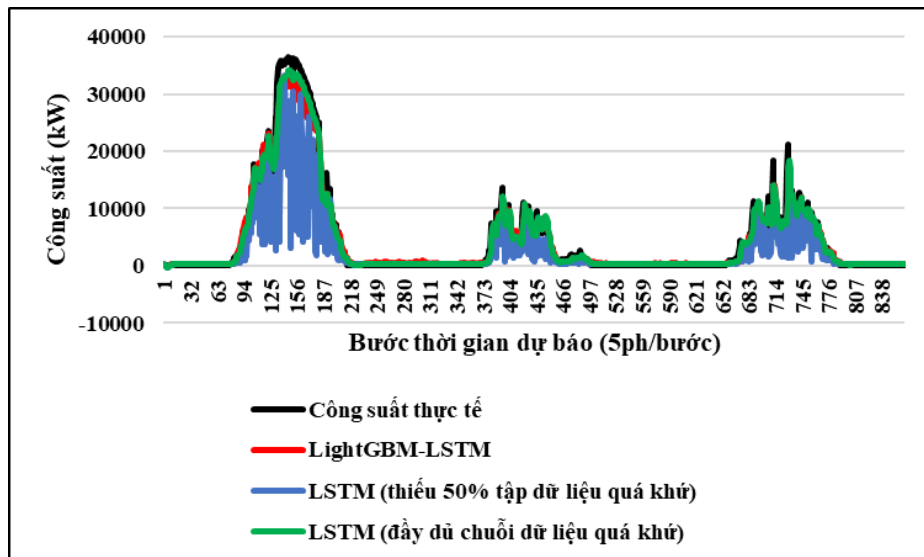
tượng hiện tại. Đồng thời, khi mức thiếu hụt dữ liệu quá khứ dưới 30%, LightGBM phát huy hiệu quả nội suy tối ưu nhất, giúp mô hình đạt trạng thái ổn định cao với sai số NMAPE dao động hẹp trong khoảng 1,45%–1,48%. Trên cơ sở đó, tỷ lệ thiếu hụt dữ liệu quá khứ ở mức 30% được xem là một ngưỡng vận hành tin cậy, tại đó mô hình LightGBM-LSTM vẫn bảo toàn được sự ổn định gần như tuyệt đối của các chỉ số sai số. Những kết quả này khẳng định vai trò trụ cột của LightGBM trong việc đảm bảo độ tin cậy của dự báo khi dữ liệu lịch sử bị gián đoạn, cũng như duy trì hiệu năng tối ưu trong các kịch bản thiếu dữ liệu ở mức vừa phải. Đồng thời với các sai số NMAPE chỉ dao động hẹp khác với mô hình LSTM không được lấp đầy chuỗi quá khứ (tăng từ 1,40 lên 8,58%-gấp 4,6 lần) cho thấy mô hình LightGBM-LSTM đã thể hiện tính bền vững vượt trội hơn so với LSTM. Tuy nhiên, nếu xét độ ổn định dài hạn khi tỷ lệ dữ liệu thiếu có thể tăng trong triển khai dự báo cho nhiều giai đoạn vận hành thì các mô hình cây quyết định (LightGBM, Random Forest, XGBoost) là lựa chọn đáng tin cậy: các chỉ số của chúng hầu như bất biến theo tỷ lệ thiếu (bức xạ thực tế: NRMSE ~2,90–3,10%, NMAPE ~1,20–1,30%; bức xạ dự báo: NRMSE ~3,10–3,16%, NMAPE ~1,80–1,85%). Từ các phân tích độ nhạy của sai số NMAPE thu được, có thể rút ra rằng cấu trúc mô hình LightGBM-LSTM được đề xuất và các mô hình cây quyết định không chỉ duy trì hiệu năng ổn định trước sự gián đoạn dữ liệu lịch sử, mà còn cho phép định lượng rõ ràng mức độ chịu lỗi của hệ thống dự báo trong điều kiện vận hành thực tế. Vì vậy, về chiến lược sử dụng các mô hình dự báo, khi có đầy đủ tập dữ liệu quá khứ có thể dùng LSTM, khi tập dữ liệu quá khứ thiếu nhẹ và vừa ưu tiên sử dụng LightGBM-LSTM để cân bằng độ chính xác và khả năng chống đứt gãy chuỗi. Còn nếu mục tiêu ổn định vận hành dài hạn trước biến động chất lượng dữ liệu, dùng các mô hình cây quyết định làm xương sống trong dự báo (có thể kết hợp các mô hình này theo dạng Ensemble), vì chúng không phụ thuộc look-back nên ít nhạy cảm với sự thiếu hụt dữ liệu quá khứ.

Việc sử dụng LightGBM để nội suy dữ liệu công suất bị thiếu được coi là một bước tiền xử lý chiến lược nhằm đảm bảo tính toàn vẹn về mặt lịch sử cho mô hình LSTM. Thay vì sử dụng các phương pháp nội suy truyền thống vốn có sai số lớn đối với dữ liệu phi tuyến biến động mạnh [110] [111], LightGBM tận dụng ưu thế về khả năng học máy phi tuyến và tốc độ tính toán (trung bình 1 giây trên mỗi chu kỳ dự báo) để ước lượng chính xác giá trị thiếu dựa trên dữ liệu khí tượng sẵn có. Hiệu quả của cơ chế này được thể hiện rõ qua thực nghiệm tại Bảng 3.12, giúp duy trì sai số NMAPE ở mức thấp

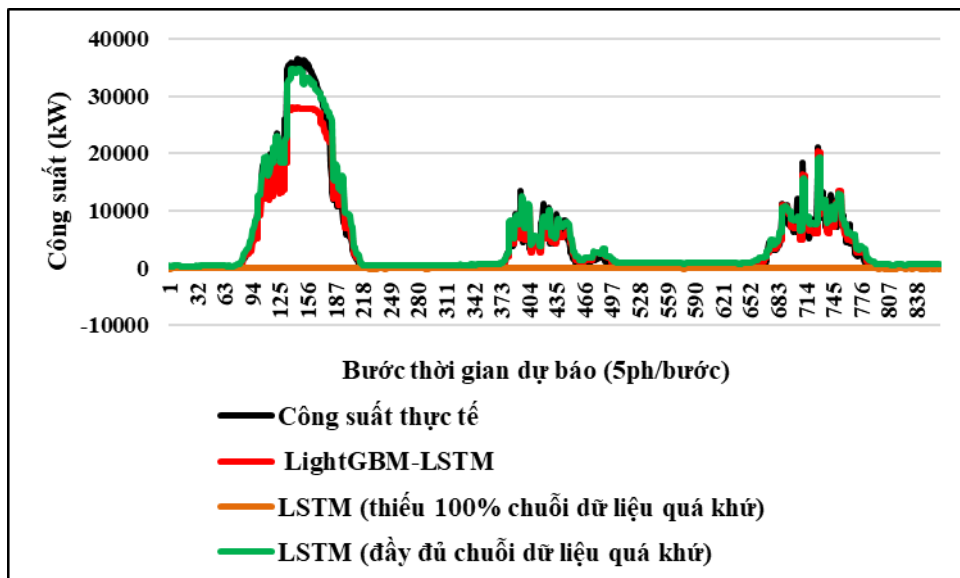
(1,85%) ngay cả trong tình huống mất dữ liệu chuỗi quá khứ kéo dài, thay vì mức sai số 8,58% nếu chỉ sử dụng LSTM thuần túy không có bước lấp đầy dữ liệu.

Để minh họa rõ hơn ảnh hưởng của chuỗi dữ liệu quá khứ đến hiệu năng của mô hình LSTM, Hình 3.7 dưới đây thể hiện so sánh giữa đường dự báo công suất phát trong hai điều kiện: (1) sử dụng mô hình LSTM có chuỗi dữ liệu quá khứ liên tục (sau khi được tái cấu trúc bằng LightGBM), (2) sử dụng mô hình LSTM trong điều kiện thiếu bộ dữ liệu công suất của ngày 28/02/2021 theo tỷ lệ 50% và 100% sử dụng bức xạ thực tế, (3) sử dụng mô hình LSTM trong điều kiện có đầy đủ bộ dữ liệu công suất quá khứ của ngày 28/02/2021.

Sự khác biệt giữa hai đường dự báo giúp làm rõ vai trò thiết yếu của chuỗi dữ liệu đầu vào đối với khả năng học và tái tạo xu hướng của mô hình LSTM.



a. So sánh các mô hình dự báo với tỷ lệ thiếu 50% chuỗi dữ liệu quá khứ

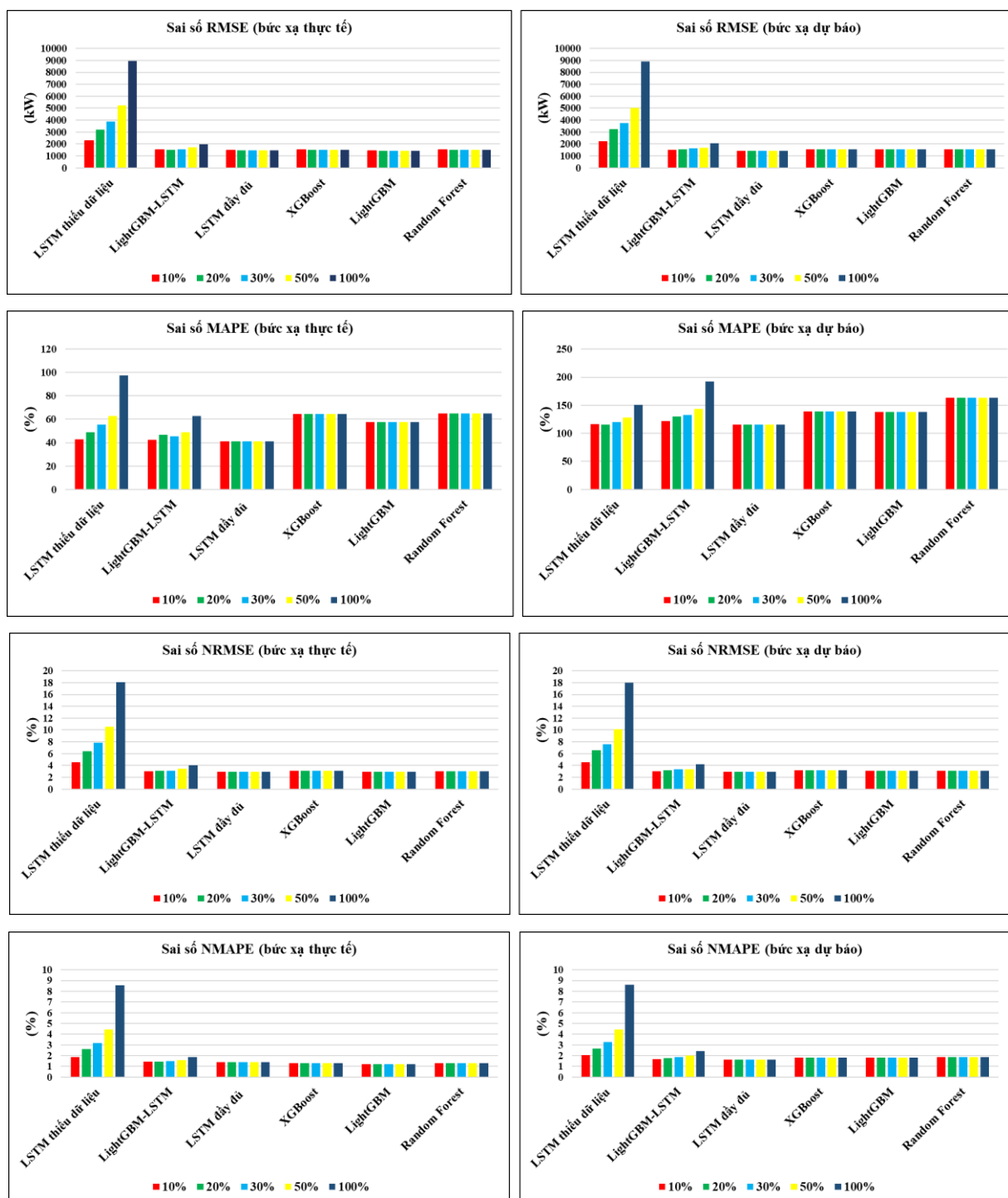


b. So sánh các mô hình dự báo với tỷ lệ thiếu dữ liệu quá khứ 100%

Hình 3.7. So sánh dự báo công suất phát của mô hình LightGBM-LSTM và LSTM trong điều kiện dữ liệu chuỗi quá khứ đầy đủ và chuỗi dữ liệu quá khứ bị thiếu theo tỷ lệ 50% (a) và 100% (b).

Biểu đồ Hình 3.7 cho thấy mô hình LSTM khi bị mất 50% chuỗi dữ liệu quá khứ (đường màu xanh da trời đậm) vẫn thể hiện được sự tăng giảm của dự báo tuy nhiên có nhiều dao động lớn cho thấy mô hình không nắm bắt đúng được các xu hướng và cường độ tăng giảm của công suất. Mô hình LSTM thiếu 100% chuỗi dữ liệu quá khứ (đường màu cam) gần như mất hoàn toàn khả năng học xu hướng, dẫn đến đường dự báo là một đường nằm ngang, không phản ứng trước các biến động thực tế. Đây là biểu hiện rõ của hiện tượng underfitting do thiếu chuỗi lịch sử, phản ánh đúng bản chất kỹ thuật của mạng LSTM là phụ thuộc chặt chẽ vào độ liền mạch của dữ liệu đầu vào để học được các quy luật ẩn theo thời gian. Ngược lại, mô hình LSTM với chuỗi dữ liệu đầy đủ (đường màu xanh lục) thể hiện rõ khả năng tái tạo hình dạng tổng thể và biên độ dao động của công suất, đặc biệt tại các đỉnh cao vào buổi trưa và chiều. Tuy nhiên, mô hình này thường dự báo vượt mức gây sai số tương đối lớn, nhất là trong các điểm công suất thấp hoặc công suất giảm nhanh. Mô hình kết hợp LightGBM-LSTM (đường màu đỏ) cho thấy khả năng bám sát xu hướng tổng thể của công suất phát khá tốt. Đặc biệt, mô hình này xử lý mượt các đoạn công suất thấp. Tuy nhiên, có thể quan sát thấy mô hình này chưa tái hiện chính xác biên độ dao động tại các điểm công suất cao, mô hình kết hợp không theo kịp về biên độ tại một số điểm công suất cao so với đường công suất thực tế và đường LSTM với đầy đủ chuỗi dữ liệu. Sự khác biệt giữa ba đường dự báo này xác nhận giá trị của phương pháp tái cấu trúc chuỗi bằng mô hình LightGBM như đã trình bày. Dù bản thân LightGBM không nắm bắt được tính tuần tự, nhưng việc sử dụng nó để lấp đầy phần dữ liệu thiếu một cách có định hướng giúp phục hồi chuỗi đầu vào đủ dài, từ đó khôi phục khả năng học xu hướng của mô hình chuỗi sâu như LSTM.

Để phân tích rõ hơn độ chính xác của các mô hình dự báo trong cả hai kịch bản là sử dụng bức xạ thực tế và bức xạ dự báo, nghiên cứu sinh tiến hành so sánh sự biến động các sai số RMSE, NRMSE, MAPE, NMAPE được thể hiện trong Hình 3.8 dưới đây.



Hình 3.8. So sánh các sai số RMSE, NRMSE, MAPE, NMAPE của các mô hình dự báo trong hai kịch bản sử dụng bức xạ thực tế và bức xạ dự báo.

Hình 3.8 là kết quả phân tích dựa trên bốn chỉ số sai số (RMSE, NRMSE, MAPE, NMAPE) trong cả hai kịch bản dữ liệu bức xạ cho thấy sự khác biệt rõ ràng về hiệu suất giữa các nhóm mô hình khi chạy với các mức độ thiếu hụt dữ liệu quá khứ khác nhau. Nhóm mô hình học máy theo mô hình cây quyết định, bao gồm XGBoost, LightGBM và Random Forest, thể hiện mức độ ổn định và khả năng thích ứng vượt trội; hiệu suất

của chúng hầu như không bị suy giảm ngay cả trong trường hợp toàn bộ dữ liệu lịch sử bị mất, qua đó khẳng định ưu thế của các mô hình này trong việc khai thác các đặc trưng tức thời mà không phụ thuộc vào tính liên tục của chuỗi thời gian. Ngược lại, các mô hình dựa trên chuỗi thời gian cho thấy sự nhạy cảm cao đối với tính toàn vẹn của dữ liệu: sai số dự báo tăng tuyến tính theo tỷ lệ dữ liệu bị mất, mặc dù trong điều kiện lý tưởng (dữ liệu đầy đủ), mô hình này vẫn đóng vai trò là một chuẩn tham chiếu mạnh mẽ. Đáng chú ý, mô hình kết hợp LightGBM- LSTM duy trì hiệu quả khi mức thiếu hụt dữ liệu ở mức vừa phải ($\leq 50\%$), nhưng khi tỷ lệ dữ liệu thiếu từ 50-100% thì độ chính xác của mô hình đã có sự suy giảm, đặc biệt khi toàn bộ dữ liệu lịch sử bị thiếu.

Những kết quả từ Bảng 3.12, Hình 3.7 và Hình 3.8 cho thấy sự hỗ trợ giữa mô hình cây quyết định và mô hình chuỗi thời gian có thể mang lại hiệu quả cao trong điều kiện dữ liệu thiếu hụt nhẹ, nhưng nếu tỷ lệ thiếu dữ liệu tăng cao thì sai số sẽ tăng dần theo dạng cộng dồn. Mô hình hybrid chỉ thực sự tối ưu khi có khả năng kiểm soát được chất lượng của chuỗi dữ liệu đã được lấp đầy.

Một điểm quan trọng rút ra từ kết quả thực nghiệm là khi sử dụng bức xạ dự báo thay cho bức xạ đo thực tế, các chỉ số sai số tuyệt đối phần trăm trung bình (MAPE) có xu hướng tăng rất cao do nhiễu và sai lệch tích lũy trong chuỗi đầu vào, đặc biệt tại những thời điểm công suất nhỏ. Cụ thể, MAPE của LSTM (đầy đủ dữ liệu) tăng từ 41,04% lên 115,79%; nhóm cây quyết định tăng từ khoảng 57,73–65,02% lên 138,04–163,65%; trong khi mô hình LightGBM–LSTM tăng từ khoảng 42,18–62,62% lên 121,33–192,64%. Điều này cho thấy việc sử dụng bức xạ dự báo có thể làm lan truyền sai số đầu vào và khuếch đại hạn chế vốn có của các mô hình nhạy cảm. Tuy vậy, cả bức xạ đo thực tế và bức xạ dự báo đều có thể được sử dụng song song trong vận hành và đánh giá hệ thống dự báo. Đối với yêu cầu độ chính xác cao tuyệt đối, bức xạ đo thực tế từ nhà máy cần được ưu tiên. Khi sử dụng bức xạ dự báo, việc đánh giá hiệu năng nên tập trung vào các chỉ số ổn định hơn như NRMSE hoặc NMAPE thay vì MAPE.

Trong điều kiện thiếu hụt chuỗi dữ liệu quá khứ, có thể lựa chọn sử dụng mô hình LightGBM-LSTM cho các mức thiếu hụt dữ liệu nhẹ đến trung bình nhằm giảm suy hao do thiếu hụt chuỗi dữ liệu quá khứ. Nhóm mô hình cây quyết định có thể được sử dụng làm nền tảng cho dự báo ổn định dài hạn hoặc là các mô hình thành phần trong các mô hình kết hợp (Stacking Ensemble) do chúng bị không phụ thuộc vào tính đầy đủ của chuỗi dữ liệu quá khứ.

3.3.3. Kết luận

Các mô hình cây quyết định như LightGBM, XGBoost và Random Forest sở hữu nhiều ưu điểm nổi bật khi áp dụng trong môi trường dữ liệu không liên tục, đặc biệt trong lĩnh vực dự báo công suất điện mặt trời. Do không phụ thuộc vào đặc tính chuỗi thời gian của dữ liệu, các mô hình này có thể hoạt động hiệu quả ngay cả khi dữ liệu quá khứ bị ngắt quãng hoặc thiếu hụt. Độ nhạy của chúng với dữ liệu bị thiếu là rất thấp, trong khi khả năng khái quát và chống nhiễu lại khá cao. Đặc biệt, các mô hình cây còn có tính diễn giải tốt và dễ triển khai trong thực tế, nhờ vào cấu trúc phân nhánh rõ ràng và khả năng xử lý quan hệ phi tuyến giữa các biến đầu vào. Tuy nhiên, hạn chế lớn của nhóm mô hình này là không thể học được các chu kỳ hoặc xu hướng dài hạn trong dữ liệu, yếu tố vốn rất quan trọng trong các bài toán dự báo theo thời gian.

Ngược lại, các mô hình học sâu chuỗi thời gian như LSTM hay GRU được thiết kế để nắm bắt tốt mối liên hệ dài hạn và mô hình hóa được các chu kỳ biến động phức tạp. Chúng có khả năng học sâu hơn về cấu trúc ẩn trong chuỗi thời gian, từ đó cho kết quả dự báo sát xu hướng và ổn định trong điều kiện lý tưởng. Tuy nhiên, hạn chế nghiêm trọng nhất của nhóm mô hình này là sự phụ thuộc tuyệt đối vào tính liên tục của dữ liệu. Khi chuỗi dữ liệu bị thiếu hoặc nhiễu, mô hình dễ rơi vào trạng thái underfitting, dẫn đến sai số lớn và khả năng học kém. Bên cạnh đó, đặc tính phức tạp của mô hình học sâu cũng khiến việc giải thích kết quả trở nên khó khăn hơn so với các mô hình truyền thống.

Từ những phân tích đối lập nhưng bổ sung giữa hai dòng mô hình cây quyết định vốn mạnh mẽ trong xử lý dữ liệu gián đoạn, và mô hình chuỗi thời gian có khả năng nắm bắt xu hướng dài hạn đã đặt nền móng cho việc phát triển các mô hình kết hợp LightGBM-LSTM trong đó sử dụng mô hình LightGBM để tái cấu trúc dữ liệu đầu vào cho mô hình LSTM cho kết quả khả quan khi LSTM đã hết tình trạng underfitting, bám sát được xu hướng biến thiên của đường công suất thực tế và độ chính xác có cải thiện nhưng vẫn thấp hơn so với các mô hình cây quyết định hoạt động đơn lẻ. Tuy nhiên, trong bối cảnh dữ liệu thực tế thường xuyên khuyết thiếu hoặc nhiễu loạn, các mô hình học sâu chuỗi thời gian vẫn có thể thiếu chính xác khi sử dụng dữ liệu dự báo thay thế dữ liệu thực tế. Vì vậy, thay vì tiếp tục cố gắng cải thiện hiệu năng bằng cách xử lý đầu vào cho các mô hình học sâu, nghiên cứu sinh lựa chọn một hướng tiếp cận hiệu quả và

khả thi hơn: kết hợp đầu ra của nhiều mô hình cây quyết định thông qua kỹ thuật Stacking Ensemble.

Ba mô hình XGBoost, LightGBM và Random Forest được lựa chọn để xây dựng nên bộ mô hình tổ hợp là kết quả của quá trình so sánh chi tiết về hiệu suất, độ ổn định và khả năng thích ứng trong điều kiện dữ liệu thiếu. Không chỉ sở hữu thời gian huấn luyện nhanh, khả năng khái quát hóa tốt, và sức đề kháng cao với nhiễu và khuyết dữ liệu, mà cả ba mô hình còn có thể dễ dàng nhân rộng và triển khai đồng thời cho nhiều nhà máy khác nhau, một yếu tố rất quan trọng trong hệ thống giám sát và dự báo quy mô lớn. Ngoài ra, sự đa dạng trong cơ chế hoạt động giữa XGB, LGBM và RF còn đảm bảo độ phân tán về lỗi, góp phần gia tăng hiệu quả tổng thể của mô hình kết hợp. Những đặc điểm này sẽ là nền tảng cho chương 4, nơi nghiên cứu sinh sẽ xây dựng mô hình tổ hợp Stacking Ensemble để khai thác triệt để tiềm năng của các mô hình cây quyết định trong điều kiện thực tiễn.

3.4. Kết luận chương 3

Trong chương 3, các mô hình học máy như LightGBM, LSTM và GRU đã được sử dụng nhằm khảo sát khả năng ứng dụng trong bài toán dự báo bức xạ mặt trời và các mô hình này đều cho hiệu suất tốt trong cả dự báo ngắn hạn một ngày và ngắn hạn dài ngày. Tuy nhiên, cần khẳng định rằng việc chỉ sử dụng các mô hình học máy huấn luyện trên tập dữ liệu lịch sử hạn chế từ một nhà máy cụ thể thường không thể đạt được độ chính xác ngang bằng với các dịch vụ dự báo chuyên nghiệp từ các tổ chức khí tượng và trung tâm bức xạ quy mô lớn, vốn sở hữu hệ thống đo đạc vệ tinh, mô hình số và năng lực tính toán chuyên sâu. Do đó, để nâng cao độ chính xác trong vận hành thực tế, việc tích hợp thông tin dự báo từ các tổ chức chuyên ngành khí tượng vào quy trình nội bộ của nhà máy là một bước đi cần thiết, giúp nâng cao độ tin cậy trong lập kế hoạch và điều độ công suất.

Chương 3 cũng tiến hành so sánh kết quả dự báo công suất phát điện mặt trời giữa các mô hình đơn lẻ XGBoost, LightGBM, Random Forest, LSTM, GRU, BiGRU trên tập dữ liệu ngắn hạn dài ngày cho thấy rõ sự khác biệt trong hiệu năng giữa hai nhóm mô hình: mô hình cây quyết định và mô hình chuỗi thời gian.

Trong các trường hợp thiếu hụt dữ liệu quá khứ, vốn là tình trạng phổ biến tại nhiều nhà máy điện mặt trời ở Việt Nam hiện nay do mất tín hiệu, lỗi SCADA hoặc hạn chế thiết bị lưu trữ, các mô hình chuỗi thời gian như LSTM, GRU và BiGRU đều gặp khó

khẩn nghiêm trọng, thể hiện qua hiện tượng sai số lớn, và kết quả dự báo thiếu ổn định. Trong khi đó, các mô hình cây quyết định như XGBoost, LightGBM và Random Forest vẫn duy trì được khả năng dự báo tức thời với sai số thấp và hiệu suất ổn định, nhờ đặc điểm không phụ thuộc vào chuỗi dữ liệu liên tục.

Tuy nhiên, khi thực hiện việc kết hợp mô hình LightGBM để tái cấu trúc dữ liệu đầu vào cho mô hình LSTM thì độ chính xác cải thiện rõ rệt, mô hình LSTM đã hết tình trạng underfitting và có các sai số RMSE, NRMSE, NMAPE gần ngang bằng các mô hình cây quyết định. Như vậy, trong điều kiện có đầy đủ và liên tục dữ liệu lịch sử, các mô hình chuỗi thời gian vẫn cho thấy khả năng dự báo tốt trong các hệ thống giám sát dài hạn, nhất là khi cần phân tích xu hướng theo chu kỳ hoặc chuỗi thời gian. Tuy nhiên khi các dữ liệu thiếu ngày càng nhiều sẽ ảnh hưởng tới độ chính xác của mô hình chuỗi thời gian. Do đó mô hình kết hợp LightGBM-LSTM chỉ được xem như một phương án ứng dụng trong các tình huống thiếu hụt nhẹ dữ liệu quá khứ.

Ngoài ra, kết quả thực nghiệm trong chương 3 cũng cho thấy mô hình dự báo công suất sử dụng đầu vào là dữ liệu bức xạ được đo tại nhà máy cho các sai số RMSE, NRMSE, NMAPE đặc biệt là sai số MAPE thấp hơn đáng kể so với trường hợp dùng dữ liệu bức xạ được dự báo từ mô hình nội bộ. Điều này phản ánh ảnh hưởng rõ rệt của chất lượng dữ liệu bức xạ đến độ chính xác của công suất dự báo. Tuy nhiên, trong trường hợp không có điều kiện tiếp cận dữ liệu dự báo chuyên sâu, các mô hình học máy như đã trình bày trong chương 3 vẫn có thể cung cấp một giải pháp thay thế khả thi, đặc biệt trong điều kiện thiếu thốn dữ liệu đo đạc tại chỗ hoặc hạ tầng hạn chế.

Từ các kết quả thực nghiệm và cân nhắc điều kiện triển khai thực tế tại Việt Nam, luận án lựa chọn ba mô hình cây quyết định XGBoost, LightGBM và Random Forest để xây dựng mô hình tổ hợp Stacking Ensemble trong chương tiếp theo. Việc lựa chọn này hướng đến giải quyết triệt để bài toán dự báo công suất phát trong điều kiện thiếu hụt dữ liệu quá khứ, đảm bảo khả năng triển khai thực tế cao và tính linh hoạt trong môi trường vận hành có nhiều biến động.

Bên cạnh đó, luận án cũng định hướng mở rộng mô hình tổ hợp sang kịch bản dự báo công suất phát đồng thời cho nhiều nhà máy điện mặt trời. Cách tiếp cận này không chỉ tận dụng lợi thế dữ liệu được cung cấp từ nhiều nguồn khác nhau nhằm nâng cao độ tin cậy dự báo, mà còn phản ánh đúng thực tiễn vận hành hệ thống điện tại Việt Nam nơi lưới điện phải tiếp nhận đồng thời công suất từ nhiều nhà máy phân bố trên một khu

vực địa lý rộng lớn. Việc triển khai dự báo nhiều nhà máy góp phần hỗ trợ công tác điều độ nguồn năng lượng tái tạo, đảm bảo cân bằng cung cầu và tăng cường tính ổn định của hệ thống trong điều kiện biến động thời tiết phức tạp.

CHƯƠNG 4. ĐỀ XUẤT MÔ HÌNH DỰ BÁO CHO MỘT SỐ NHÀ MÁY ĐIỆN MẶT TRỜI TẠI VIỆT NAM

Trên cơ sở ba mô hình XGBoost, LightGBM và Random Forest đã được lựa chọn trong chương 3 nhờ hiệu suất cao và khả năng thích ứng tốt với điều kiện thiếu dữ liệu đo đạc tại nhà máy điện mặt trời, chương 4 tiếp tục phát triển các mô hình tổ hợp theo hướng Stacking Ensemble (SE) nhằm khai thác tối đa ưu thế tổng hợp của từng mô hình thành phần.

Sau khi đánh giá phương pháp kết hợp đầu vào (input fusion) trong chương 3 cụ thể là sử dụng LightGBM để nội suy dữ liệu thiếu và cung cấp làm đầu vào cho mô hình LSTM cho thấy kết quả dự báo đã được cải thiện trong một số trường hợp. Tuy nhiên, khi tỷ lệ dữ liệu bị thiếu vượt quá ngưỡng cho phép, việc lấp đầy trở nên kém chính xác và dẫn đến sai số tổng thể của mô hình tăng lên do hiện tượng lan truyền lỗi. Xuất phát từ hạn chế này, luận án chuyển hướng sang phương pháp kết hợp đầu ra (output fusion), nhằm tách biệt quá trình huấn luyện của từng mô hình thành phần và tối ưu hóa kết quả đầu ra bằng cách học trọng số tổ hợp dựa trên hiệu suất thực tế. Phương pháp Stacking Ensemble được sử dụng trong chương này không những giúp giảm thiểu ảnh hưởng của lỗi tích lũy mà còn tỏ ra phù hợp hơn trong bối cảnh dữ liệu thời tiết đo đạc có thể gián đoạn, không đồng nhất hoặc đến từ nhiều nguồn khác nhau.

Trong chương 3 có thể thấy các chỉ số MAPE của các mô hình đơn lẻ tăng rất cao khi sử dụng dữ liệu bức xạ dự báo. Do đó, để đảm bảo độ tin cậy và tính chính xác cho việc xây dựng mô hình dự báo công suất phát theo dạng Stacking Ensemble tại chương này, nghiên cứu sinh sẽ sử dụng tập dữ liệu bức xạ và thời tiết do nhà máy điện mặt trời cung cấp đây là nguồn dữ liệu đã qua xử lý, đồng bộ và phù hợp với điều kiện vận hành thực tế.

Bên cạnh việc xây dựng tổ hợp giữa các mô hình cây quyết định, luận án cũng tiến hành thử nghiệm với các cấu hình kết hợp giữa cây quyết định và các mô hình học sâu chuỗi thời gian như GRU, LSTM và BiGRU, nhằm đánh giá khả năng cải thiện hiệu suất dự báo khi tích hợp các cấu trúc học khác biệt. Các mô hình SE được triển khai theo hai nhóm chính:

- Nhóm mô hình tổ hợp học máy truyền thống (XGBoost, LightGBM, Random Forest);
- Nhóm mô hình kết hợp học sâu (BiGRU, GRU, LSTM) với XGBoost và LightGBM.

Mục tiêu của các thử nghiệm này là đánh giá hiệu năng tương đối giữa các cấu hình tổ hợp, dựa trên ba tiêu chí: độ chính xác, thời gian xử lý, và khả năng tổng quát hóa trong các điều kiện thời tiết khác nhau.

Tiếp theo, chương mở rộng bài toán từ phạm vi một nhà máy điện mặt trời sang nhiều nhà máy có điều kiện khí hậu khác nhau, thông qua việc tích hợp thêm dữ liệu từ hai nhà máy mới tại Đắk Lắk và Thanh Hóa. Trên cơ sở đó, hai kịch bản triển khai được thiết kế:

- Huấn luyện riêng mô hình cho từng nhà máy;
- Sử dụng kiến trúc Selector- Model cho phép một mô hình duy nhất tự động thích nghi với đặc trưng vùng khí hậu từng địa phương.

Các nội dung trong chương này đóng vai trò then chốt trong việc kiểm chứng khả năng ứng dụng thực tiễn và tiềm năng mở rộng quy mô của mô hình Stacking Ensemble khi triển khai trên các hệ thống điện mặt trời phân tán tại Việt Nam.

4.1. Dự báo công suất phát cho một nhà máy

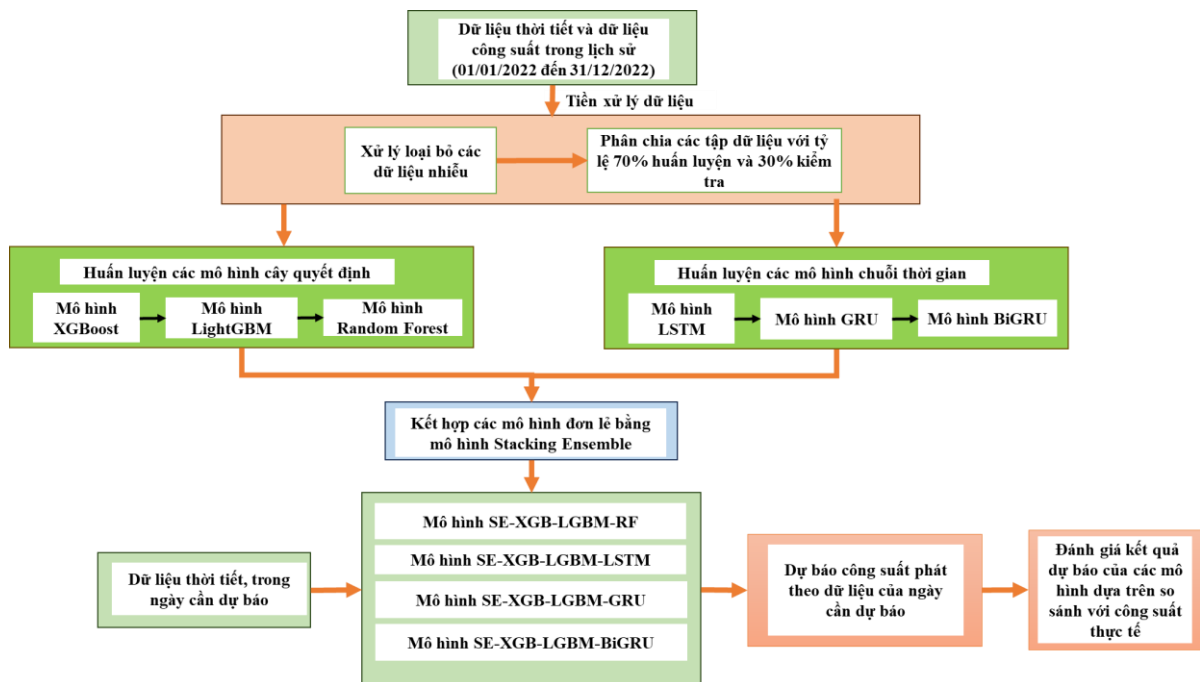
Trong bối cảnh bài toán dự báo công suất phát điện mặt trời với dữ liệu thời tiết mang đặc trưng chuỗi thời gian và phi tuyến tính mạnh, đồng thời thường xuyên gặp tình trạng thiếu hụt hoặc không liên tục về dữ liệu lịch sử tại các nhà máy, việc lựa chọn cấu trúc mô hình phù hợp trở thành yếu tố then chốt nhằm đảm bảo độ chính xác và khả năng tổng quát hóa của hệ thống dự báo. Nhằm đánh giá toàn diện hiệu quả dự báo trong nhiều điều kiện thực tế khác nhau bao gồm cả tình huống thiếu dữ liệu quá khứ, nghiên cứu sinh lựa chọn sử dụng kỹ thuật Stacking Ensemble (Stacked Generalization), một phương pháp học máy nâng cao cho phép kết hợp nhiều mô hình thành phần với đặc trưng khác nhau, từ đó xây dựng mô hình tổng hợp có hiệu suất ổn định và độ chính xác vượt trội [95].

Trong thiết kế kiến trúc Stacking Ensemble, hai mô hình thành phần đầu tiên được giữ cố định là XGBoost và LightGBM, bởi cả hai đều là các thuật toán cây quyết định mạnh mẽ, đã được kiểm chứng về hiệu quả xử lý dữ liệu phi tuyến, khả năng chống overfitting tốt và thời gian huấn luyện ngắn. Việc giữ nguyên hai mô hình này nhằm đảm bảo nền tảng vững chắc và ổn định cho tổ hợp, đồng thời tạo điều kiện kiểm soát chính xác ảnh hưởng của mô hình thứ ba khi thay đổi [34], [35].

Trên cơ sở đó, mô hình thứ ba trong tổ hợp được luân phiên thay đổi nhằm xây dựng hai cấu hình tổ hợp mang tính đối chứng:

- Cấu hình thứ nhất: nhóm mô hình thuần cây quyết định, sử dụng Random Forest làm thành phần thứ ba, tạo thành mô hình SE-XGB-LGBM-RF. Random Forest là thuật toán học máy truyền thống, có khả năng xử lý nhiễu tốt và được ứng dụng rộng rãi trong các bài toán dự báo thực tiễn.
- Cấu hình thứ hai: nhóm mô hình kết hợp, trong đó Random Forest được thay thế bằng một trong ba mạng nơ-ron hồi tiếp BiGRU, GRU hoặc LSTM, lần lượt hình thành các tổ hợp SE-XGB-LGBM-BiGRU, SE-XGB-LGBM-GRU và SE-XGB-LGBM-LSTM. Các mô hình học sâu này có khả năng khai thác hiệu quả đặc trưng thời gian trong chuỗi dữ liệu, đặc biệt trong bối cảnh công suất phát điện bị ảnh hưởng mạnh bởi biến động khí tượng theo giờ.

Kịch bản thực hiện kết hợp các mô hình sử dụng Stacking Ensemble được thể hiện trong sơ đồ Hình 4.1 dưới đây:



Hình 4.1. Sơ đồ tích hợp các mô hình dự báo công suất phát điện mặt trời sử dụng Stacking Ensemble

Việc triển khai song song hai cấu hình tổ hợp trên không chỉ cho phép đánh giá vai trò thực tiễn của các kiến trúc học sâu trong môi trường dữ liệu chuỗi thời gian, mà còn giúp kiểm chứng khả năng tổng hợp của mô hình Stacking khi kết hợp giữa học máy truyền thống và học sâu. Từ đó, việc so sánh các mô hình theo các tiêu chí như RMSE, NRMSE, MAPE, NMAPE, thời gian huấn luyện và khả năng triển khai thực tế sẽ cung cấp cơ sở khoa học vững chắc để lựa chọn mô hình ưu việt, phục vụ cho các bước tiếp

theo: tối ưu trọng số theo mùa và mở rộng triển khai theo kiến trúc selector-model cho nhiều nhà máy điện mặt trời tại các vùng khí hậu khác nhau.

4.1.1. Xây dựng mô hình dự báo SE-XGB-LGBM-RF

Trong các bài toán dự báo như công suất phát điện mặt trời, các mô hình cây quyết định như XGBoost, LightGBM, và Random Forest thường ít bị overfitting hơn so với các mô hình học sâu như LSTM và GRU. Điều này là do các mô hình cây quyết định áp dụng các kỹ thuật kiểm soát overfitting mạnh mẽ như cây con giới hạn (pruning) và bỏ mẫu ngẫu nhiên (bagging), giúp duy trì tính khái quát tốt hơn khi áp dụng trên dữ liệu mới. Ngược lại, các mô hình học sâu như LSTM và GRU, mặc dù có khả năng nắm bắt các mối quan hệ phức tạp trong dữ liệu chuỗi thời gian, lại dễ bị overfitting nếu không được tối ưu hóa và điều chỉnh siêu tham số kỹ càng, đặc biệt là khi kích thước tập dữ liệu huấn luyện hạn chế.

Theo hướng tiếp cận thứ nhất, để khai thác tối đa ưu điểm riêng biệt của từng mô hình, luận án đề xuất xây dựng tổ hợp ba mô hình cây quyết định: XGBoost, LightGBM và Random Forest. Các mô hình con được xây dựng như đã trình bày trong mục 3.2.1, và kết quả đầu ra từ ba mô hình này được đưa vào mô hình Stacking Ensemble SE-XGB-LGBM-RF nhằm tổng hợp dự báo.

Mỗi mô hình thành phần có khả năng xử lý các dạng đặc trưng khác nhau của dữ liệu phi tuyến và có nhiễu, việc kết hợp chúng trong một kiến trúc stacking cho phép giảm thiểu overfitting, tăng khả năng tổng quát hóa, và nâng cao độ chính xác trên dữ liệu mới.

Trong cấu hình cơ bản, mô hình SE-XGB-LGBM-RF sử dụng trọng số tổ hợp bằng nhau cho ba đầu ra thành phần tương đương phép trung bình đơn giản như một bước nền trước khi đánh giá các phiên bản nâng cao hơn. Mô hình này sẽ được huấn luyện trên tập dữ liệu của nhà máy điện mặt trời tại Quảng Trị như trình bày trong mục 2.1.1.

4.1.2. Xây dựng mô hình SE-XGB-LGBM-BiGRU, SE-XGB-LGBM-GRU, SE-XGB-LGBM-LSTM

Theo hướng tiếp cận thứ hai, luận án xây dựng các mô hình tổ hợp lại giữa mô hình cây quyết định và mô hình học sâu chuỗi thời gian, bao gồm: SE-XGB-LGBM-BiGRU, SE-XGB-LGBM-GRU, và SE-XGB-LGBM-LSTM. Các mô hình này kết hợp ba thành phần: XGBoost, LightGBM (trình bày trong mục 3.2.1) và BiGRU, GRU, hoặc LSTM (trình bày trong mục 3.2.2), với mục tiêu khai thác đồng thời đặc trưng phi tuyến mạnh

của mô hình cây quyết định và khả năng ghi nhớ chuỗi thời gian dài hạn của mạng học sâu.

Kết quả dự báo từ các mô hình thành phần được kết hợp tại tầng đầu ra bằng phương pháp trung bình cộng như một cấu hình baseline của mô hình tổ hợp. Sau khi xây dựng, các mô hình tổ hợp này sẽ được huấn luyện và đánh giá trên tập dữ liệu lịch sử của nhà máy điện mặt trời tại Quảng Trị như đã mô tả trong mục 2.1.1.

4.1.3. Kết quả so sánh các mô hình SE-XGB-LGBM-RF, SE-XGB-LGBM-BiGRU, SE-XGB-LGBM-GRU, SE-XGB-LGBM-LSTM

Sau khi hoàn thành việc thiết lập các mô hình Stacking Ensemble gồm SE-XGB-LGBM-RF, SE-XGB-LGBM-BiGRU, SE-XGB-LGBM-GRU và SE-XGB-LGBM-LSTM các mô hình này được huấn luyện và đánh giá trên cùng một tập dữ liệu thời tiết bị từ ngày 01 đến 03/03/2021 và vẫn bị giới hạn chuỗi quá khứ gần, tương tự như điều kiện mô tả trong chương 3. Mục tiêu của việc so sánh là đánh giá hiệu suất thực tế của từng mô hình trong bài toán dự báo công suất phát điện theo nhiều tiêu chí khác nhau bao gồm:

- Độ chính xác dự báo: thể hiện qua các chỉ số RMSE, NRMSE, MAPE, NMAPE;
- Thời gian huấn luyện mô hình, đánh giá khả năng triển khai thực tế;
- Khả năng tổng quát hóa trên tập dự báo thực tế trong điều kiện thiếu chuỗi dữ liệu.

Các kết quả được trình bày dưới dạng bảng và đồ thị để làm rõ sự khác biệt giữa hai phương pháp tiếp cận. Trên cơ sở đó, mô hình có hiệu suất tốt hơn sẽ được lựa chọn để sử dụng trong các bước tiếp theo, bao gồm xây dựng mô hình Stacking Ensemble với trọng số tối ưu và việc mở rộng dự báo cho nhiều nhà máy theo kiến trúc selector-model.

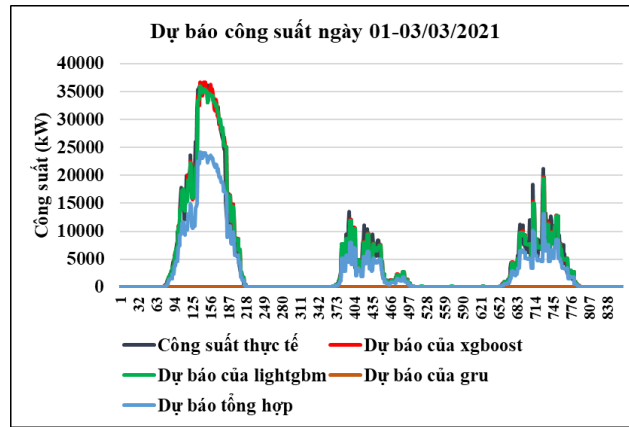
Bảng 4.1. Kết quả so sánh dự báo công suất phát điện mặt trời của các mô hình SE-XGB-LGBM-RF, SE-XGB-LGBM-BiGRU, SE-XGB-LGBM-GRU, SE-XGB-LGBM-LSTM trên tập dữ liệu dự báo

Các tiêu chí đánh giá	Mô hình SE-XGB-LGBM-RF	Mô hình SE-XGB-LGBM-BiGRU	Mô hình SE-XGB-LGBM-GRU	Mô hình SE-XGB-LGBM-LSTM
Thời gian dự báo (s)	23,28	101,57	98,39	98,94

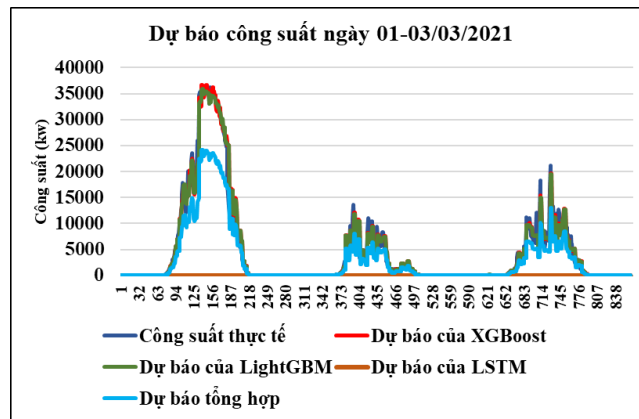
RMSE (kW)	1453,85	1497,67	3359,07	3359,08
NRMSE (%)	2,93	3,02	6,78	6,79
MAPE (%)	59,39	79,95	54,67	54,68
NMAPE (%)	1,26	1,35	3,07	3,08

Từ kết quả của Bảng 4.1 cho thấy mô hình SE-XGB-LGBM-RF khẳng định sự vượt trội về thời gian xử lý, với thời gian dự báo là 23,28 giây thấp hơn rõ rệt so với các mô hình sử dụng mạng nơron như BiGRU, GRU hay LSTM (có thời gian dự báo xấp xỉ trong khoảng 100 giây). Về độ chính xác, kết quả cho thấy mô hình SE-XGB-LGBM-RF thể hiện sự ổn định và độ chính xác cao nhất (RMSE = 1453,85 kW; NMAPE = 1,26%), vượt trội so với các mô hình còn lại. Hai mô hình SE-XGB-LGBM-LSTM và SE-XGB-LGBM-GRU có độ chính xác kém nhất với RMSE cao và sai số MAPE trung bình lớn hơn đáng kể. Từ đó có thể thấy, mô hình SE-XGB-LGBM-RF có lợi thế vượt trội về tốc độ và độ ổn định, phù hợp với các bài toán dự báo thời gian thực trong điều kiện thiếu dữ liệu quá khứ và triển khai quy mô lớn. Trong khi đó, mô hình SE-XGB-LGBM-BiGRU có thể là lựa chọn tiềm năng cho các kịch bản cần tối ưu độ chính xác ngắn hạn.

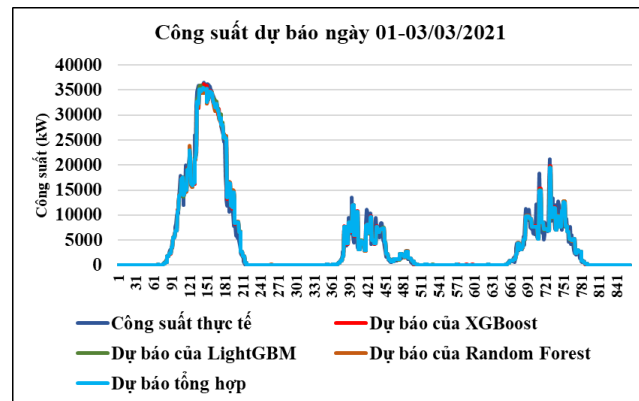
Để trực quan hóa hiệu suất dự báo của các mô hình tổ hợp SE-XGB-LGBM-RF, SE-XGB-LGBM-BiGRU, SE-XGB-LGBM-GRU và SE-XGB-LGBM-LSTM trong tập dữ liệu ngắn hạn đã được đánh giá trong Bảng 4.1, các biểu đồ sau đây minh họa đường cong công suất phát điện thực tế so sánh với dự báo của từng mô hình theo từng thời điểm trong ngày. Mỗi biểu đồ cung cấp cái nhìn rõ ràng về khả năng bám sát xu hướng, độ mượt của đường dự báo và khả năng phản ứng với biến động thực tế của từng mô hình. Đây là cơ sở trực quan giúp lý giải thêm sự chênh lệch về các chỉ số RMSE, NRMSE, MAPE, NMAPE và thời gian dự báo đã được trình bày ở Bảng 4.1.



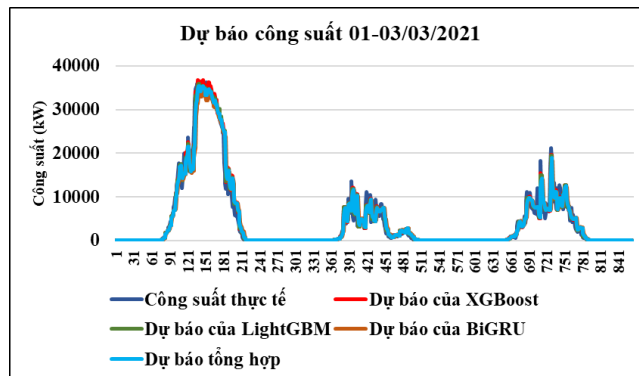
a. Dự báo công suất từ mô hình SE-XGB-LGBM-GRU



b. Dự báo công suất từ mô hình SE-XGB-LGBM-LSTM



c. Dự báo công suất từ mô hình SE-XGB-LGBM-RF



d. Dự báo công suất từ mô hình SE-XGB-LGBM-BiGRU

Hình 4.2. So sánh công suất phát điện thực tế và dự báo của bốn mô hình tổ hợp a. SE-XGB-LGBM-GRU, b. SE-XGB-LGBM-LSTM, c. SE-XGB-LGBM-RF và d. SE-XGB-LGBM-BiGRU trong tập dữ liệu (01-03/03/2021)

Từ Hình 4.2 và có thể nhận thấy hai mô hình GRU và LSTM có dấu hiệu rõ của hiện tượng underfitting do đường dự báo các mô hình GRU và LSTM thể hiện gần như nằm ngang điều này làm ảnh hưởng đến tính chính xác tổng thể của hai mô hình SE-XGB-LGBM-GRU và SE-XGB-LGBM-LSTM (đường dự báo tổng hợp của hai mô hình cách xa đáng kể so với đường công suất thực tế). Ngược lại, mô hình SE-XGB-LGBM-BiGRU cho kết quả khá cân bằng: nó tái hiện tốt hình dạng của công suất thực tế trong suốt thời gian trong ngày, tránh được hiện tượng underfitting hoặc dao động bất thường, cho thấy khả năng học được xu thế chính xác và hiệu quả của dữ liệu. Tuy nhiên, ta thấy thời gian huấn luyện và dự báo của BiGRU vẫn còn khá cao, vượt xa so với các mô hình sử dụng cây quyết định XGB, LGBM, RF (Bảng 4.1 và Bảng 3.10). Độ chính xác của mô hình SE-XGB-LGBM-BiGRU cũng kém hơn so với mô hình SE-XGB-LGBM-RF. Mô hình SE-XGB-LGBM-RF tiếp tục cho thấy độ chính xác và ổn định nhất (đường dự báo tổng hợp bám sát đường công suất thực tế).

Tổng hợp các kết quả từ Bảng 4.1 và Hình 4.2 cho thấy có thể khẳng định rằng việc chuyển hướng sang kết hợp đầu ra bằng mô hình Stacking Ensemble (thay vì kết hợp đầu vào như chương 3) là phù hợp, đặc biệt khi chọn thành phần là các mô hình cây quyết định. Việc bổ sung các mô hình học sâu chuỗi thời gian (GRU, LSTM) vào tầng đầu ra trong mô hình Stacking không mang lại cải thiện đáng kể về độ chính xác, trong khi lại làm tăng thời gian huấn luyện và độ phức tạp hệ thống. Kết quả này cũng nhất quán với nhận định từ chương 3: trong điều kiện thiếu chuỗi quá khứ liên tục, các mô hình LSTM và GRU không phát huy được tối đa khả năng học đặc trưng chuỗi, và việc kết hợp đầu ra không đủ bù đắp cho hạn chế đầu vào.

Kết hợp các kết quả trên với các phân tích định lượng và định tính từ chương 3, có thể khẳng định rằng mô hình kết hợp SE-XGB-LGBM-RF là phương án tối ưu nhất: nó vừa có độ chính xác cao, vừa đảm bảo tính ổn định, tốc độ xử lý nhanh, và dễ triển khai trên các hệ thống vận hành thực tế. Do đó, mô hình này được lựa chọn làm nền tảng cho các bước tiếp theo của luận án, bao gồm: (1) xây dựng mô hình tổ hợp có trọng số tối ưu theo mùa, và (2) triển khai dự báo mở rộng cho nhiều nhà máy.

4.1.4. Xây dựng mô hình dự báo SE-XGB-LGBM-RF với bộ trọng số tối ưu (OW)

a. Xây dựng mô hình dự báo SE-XGB-LGBM-RF-OW

Từ kết quả trên, nghiên cứu sinh sẽ chọn mô hình SE-XGB-LGBM-RF để xây dựng mô hình Stacking Ensemble với bộ trọng số tối ưu (SE-XGB-LiGBM-RF-OW)

Quá trình hoạt động của Stacking Ensemble bao gồm hai giai đoạn:

Giai đoạn 1 (Base Models): Huấn luyện các mô hình riêng lẻ (hay mô hình cơ sở) trên toàn bộ tập dữ liệu huấn luyện. Trong trường hợp này, ba mô hình XGBoost, LightGBM và Random Forest được huấn luyện riêng rẽ.

Giai đoạn 2 (Meta Learner): Tạo một tập dữ liệu mới từ các dự báo của các mô hình cơ sở trên tập dữ liệu huấn luyện và dữ liệu đầu ra thực tế. Trên tập dữ liệu mới này, một mô hình meta learner gồm các tập huấn luyện của ba mô hình XGBoost, LightGBM và Random Forest. Ba mô hình này được huấn luyện để học cách kết hợp các dự báo từ các mô hình cơ sở với một bộ trọng số tối ưu.

Quá trình tối ưu bộ trọng số được thực hiện bởi meta learner nhằm khai thác tối đa ưu điểm của từng mô hình riêng lẻ. Khi dự báo trên dữ liệu mới, dự báo cuối cùng là kết quả dự báo của từng mô hình từ ba mô hình cơ sở XGBoost, LightGBM và Random Forest nhân với tổng có trọng số của các mô hình theo công thức đề xuất như sau:

$$y_{\text{dự báo}} = w_1 \times y_{\text{xgb}} + w_2 \times y_{\text{lgbm}} + w_3 \times y_{\text{rf}} \quad (4.1)$$

Trong đó, y_{xgb} , y_{lgbm} , y_{rf} lần lượt là dự báo từ XGBoost, LightGBM và Random Forest; w_1 , w_2 , w_3 là các trọng số tối ưu tương ứng được xác định bởi meta learner. Các trọng số w_1 , w_2 , w_3 là các trọng số biểu thị mức độ đóng góp của từng mô hình vào dự báo cuối cùng. Các trọng số này được tối ưu hóa nhằm giảm thiểu sai số của mô hình dự báo tổng hợp, với các sai số thường dùng là NMAPE, MAPE, RMSE và NRMSE.

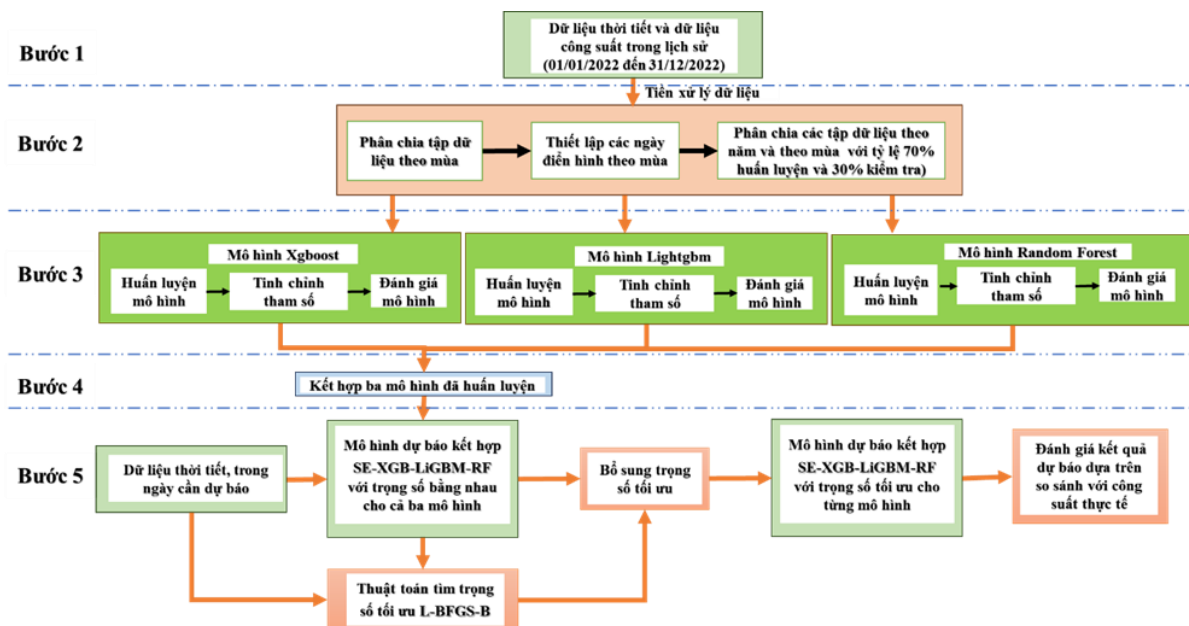
Quy trình tối ưu hóa nhằm xác định bộ trọng số tối ưu để đạt sai số nhỏ nhất có thể cho mô hình dự báo tổng hợp. Phương pháp tối ưu hóa được sử dụng trong nghiên cứu này là L-BFGS-B một thuật toán tối ưu hóa dựa trên gradient, nó đặc biệt hiệu quả cho các bài toán tối ưu hóa có ràng buộc. L-BFGS-B [112] là biến thể của phương pháp BFGS, trong đó ma trận Hessian được xấp xỉ nhằm đạt hiệu quả tối ưu hóa cao hơn so với các phương pháp giảm dần gradient thông thường.

L-BFGS-B nổi bật với khả năng sử dụng bộ nhớ hạn chế, tiết kiệm tài nguyên và phù hợp cho các bài toán quy mô lớn. Thuật toán này còn có thể xử lý các ràng buộc Box, đảm bảo các biến nằm trong khoảng giá trị xác định, cụ thể là $0 \leq w_i \leq 1$ (với $i=1,2,3$)

và $w_1 + w_2 + w_3 = 1$. Bằng cách xấp xỉ ma trận Hessian, L-BFGS-B cải thiện tốc độ hội tụ nhờ xác định các hướng tìm kiếm hiệu quả. Trong học máy, L-BFGS-B thường được dùng để tối ưu hóa các tham số mô hình có ràng buộc, giúp nâng cao hiệu suất và giảm thiểu nguy cơ quá khớp.

Bằng cách sử dụng Stacking Ensemble với bộ trọng số tối ưu, mô hình kết hợp có khả năng thu nhận điểm mạnh của từng mô hình riêng lẻ và khắc phục các nhược điểm của chúng. Điều này dẫn đến mô hình tổng hợp có hiệu suất dự báo cao hơn so với từng mô hình đơn lẻ.

Sơ đồ thuật toán cho mô hình dự báo kết hợp SE-XGB-LiGBM-RF-OW được thể hiện trong Hình 4.3.



Hình 4.3. Sơ đồ thuật toán mô hình dự báo đề xuất SE-XGB-LiGBM-RF-OW

Các bước thực hiện thuật toán như sau:

- **Tập dữ liệu lịch sử:** Tập dữ liệu lịch sử được giới thiệu trong phần 2.1.1.
- **Phân chia tập dữ liệu:** Dữ liệu lịch sử được phân chia thành các mùa trong năm: mùa xuân (từ 01/03 đến 31/05), mùa hạ (từ 01/06 đến 31/08), mùa thu (từ 01/09 đến 30/11), và mùa đông (từ 01/01 đến 28/02, cùng với ngày từ 01/12 đến 31/12). Sau khi chia thành các tập dữ liệu theo mùa, nghiên cứu sinh đã tách một số ngày đặc trưng trong từng mùa theo Bảng 4.2 ra khỏi tập dữ liệu lịch sử để thiết lập tập dữ liệu phục vụ cho công tác dự báo, bao gồm ngày có biến động công suất lớn nhất, ngày có biến động bức xạ lớn nhất, ngày có biến động công suất nhỏ nhất và ngày có biến động bức xạ nhỏ nhất. Tiếp theo, nhóm phân chia dữ liệu với tỷ lệ 70% cho huấn luyện (training) và 30% cho kiểm tra (test).

- **Dữ liệu thời tiết trong ngày cần dự báo:** tập dữ liệu theo từng mùa sẽ được tách các ngày điển hình theo Bảng 4.2 để làm dữ liệu dự báo sau này. Cách xác định ngày điển hình như sau: tính giá trị trung bình của công suất và bức xạ trong từng mùa. Sau đó sử dụng độ lệch chuẩn để so sánh với giá trị trung bình này. Ngày nào có độ lệch chuẩn của công suất và bức xạ xa giá trị trung bình nhất thì đó là ngày biến động về công suất, bức xạ lớn nhất. Ngày nào có độ lệch chuẩn của công suất và bức xạ gần giá trị trung bình nhất thì đó là ngày có biến động nhỏ nhất.

Bảng 4.2. Các ngày điển hình trong một mùa

Ngày điển hình	Ngày có công suất thay đổi lớn nhất	Ngày có công suất thay đổi nhỏ nhất	Ngày có bức xạ thay đổi lớn nhất	Ngày có bức xạ thay đổi nhỏ nhất
Mùa xuân	06/04/2022	04/03/2022	26/03/2022	05/03/2022
Mùa hạ	28/08/2022	08/08/2022	30/08/2022	08/08/2022
Mùa thu	18/09/2022	12/10/2022	18/09/2022	12/10/2022
Mùa đông	18/02/2022	18/12/2022	07/01/2022	18/12/2022

- **Trọng số cho từng mô hình:** Mỗi mô hình sẽ thực hiện dự báo công suất phát điện mặt trời tại các thời điểm trong ngày điển hình ở Bảng 4.2, sau đó kết quả dự báo của mỗi mô hình được nhân với một trọng số tương đương, với giá trị là 0,33 cho mỗi mô hình. Các kết quả dự báo từ các mô hình được tổng hợp bằng cách cộng đại số với nhau tạo ra kết quả dự báo cuối cùng.

Trong lần dự báo này, trọng số của các mô hình được đặt cố định là 0,33. Trọng số 0,33 giúp cân bằng đóng góp của từng mô hình trong kết quả cuối cùng, giảm nguy cơ mô hình tổng hợp bị ảnh hưởng quá mức bởi điểm yếu của bất kỳ mô hình nào. Cách tiếp cận này đặc biệt hữu ích khi mỗi mô hình có cả điểm mạnh và điểm yếu riêng, việc kết hợp các mô hình có thể tận dụng những điểm mạnh của chúng và bù đắp cho những điểm yếu, tạo ra dự báo tổng thể chính xác và ổn định hơn.

Tiếp theo, nghiên cứu sinh sử dụng thuật toán tối ưu L-BFGS-B với 100 lần lặp để điều chỉnh trọng số của mỗi mô hình cho mỗi mùa trong năm, với dữ liệu đầu vào là tập dữ liệu kết quả dự báo từ phân tích toán trước với trọng số đồng nhất. Kết quả từ thuật toán này được cập nhật vào mô hình dự báo để tối ưu hóa trọng số.

Sau đó, nghiên cứu sinh sử dụng mô hình dự báo với trọng số tối ưu đã được cập nhật để dự báo công suất phát điện mặt trời cho các ngày điển hình được liệt kê trong

Bảng 4.2. Kết quả từ các mô hình dự báo được tổng hợp lại với nhau theo phép cộng đại số để tạo ra kết quả dự báo cuối cùng.

- **So sánh kết quả với công suất thực tế:** kết quả dự báo công suất phát điện từ mô hình với trọng số tối ưu được so sánh với công suất thực tế và kết quả dự báo khi sử dụng trọng số đồng nhất để đánh giá hiệu suất của mô hình trong dự báo công suất mặt trời.

b. Đánh giá mô hình dự báo SE-XGB-LGBM-RF-OW

Để phân tích sâu hơn, và giảm sai số của mô hình SE-XGB-LGBM-RF-OW các mô hình dự báo riêng biệt XGBoost, LightGBM và RF được đào tạo trên các tập dữ liệu đề cập trong phần 2.1.1 nhưng được chia theo mùa như đã nói ở trên (trừ những ngày được liệt kê trong Bảng 4.2). Sau khi được huấn luyện với tập dữ liệu theo mùa, các mô hình này vẫn được so sánh về hiệu suất huấn luyện với các mô hình LSTM và GRU. Kết quả huấn luyện cho từng mùa được trình bày trong Bảng 4.3 dưới đây:

Bảng 4.3. So sánh các kết quả huấn luyện của từng mô hình với tập dữ liệu lịch sử đã được chia theo mùa trong năm

Các mùa	Các so sánh	XGBoost Model	LightGBM Model	RF Model	Mô hình LSTM	Mô hình GRU
Mùa xuân	Thời gian hoàn thành (s)	0,768	0,311	5,191	151,015	109,434
	RMSE (kW)	1085,779	1012,222	1042,355	1199,625	1179,281
	NRMSE (%)	2,193	2,044	2,105	2,423	2,382
	MAPE (%)	74,068	214,691	41,438	18285,84	5338,227
	NMAPE (%)	0,812	0,766	0,785	1,232	1,013
Mùa hạ	Thời gian hoàn thành (s)	0,781	0,318	5,179	148,341	104,264
	RMSE (kW)	1251,062	1196,135	1241,774	1396,369	1395,076
	NRMSE (%)	2,527	2,416	2,508	2,821	2,818
	MAPE (%)	246,264	53,964	23,587	1527,623	1360,625
	NMAPE (%)	1,103	1,068	1,100	1,458	1,405
Mùa thu	Thời gian hoàn thành (s)	0,745	0,324	4,794	147,146	99,923
	RMSE (kW)	2177,740	2095,553	2154,299	2305,123	2312,86
	NRMSE (%)	4,399	4,233	4,352	4,656	4,672

	MAPE	46,468	33,376	29,802	222,647	224,971
	NMAPE	1,520	1,466	1,485	1,821	1,854
Mùa đông	Thời gian hoàn thành (s)	0,695	0,333	4,356	140,802	103,623
	RMSE (kW)	1264,509	1194,748	1209,349	1308,065	1318,722
	NRMSE (%)	2,554	2,417	2,443	2,642	2,664
	MAPE (%)	94,415	120,76	14,323	2017,471	280,037
	NMAPE (%)	0,752	0,733	0,711	0,865	0,873

Từ Bảng 4.3, Các mô hình XGBoost, LightGBM và Random Forest (RF) thể hiện sự vượt trội rõ ràng so với LSTM và GRU, khiến chúng trở thành lựa chọn tối ưu để xây dựng mô hình Stacking Ensemble. Cụ thể, LightGBM cho thấy hiệu suất tuyệt vời với RMSE thấp nhất trong tất cả các mùa, như 1012,222 kW vào mùa xuân và 1194,748 kW vào mùa đông, đồng thời có thời gian hoàn thành cực kỳ ngắn, chỉ từ 0,311 đến 0,333 giây. XGBoost cũng đáng chú ý với RMSE ổn định và thấp, như 1085,779 kW vào mùa xuân và thời gian hoàn thành nhanh, chỉ từ 0,695 đến 0,781 giây. Tuy RF có thời gian hoàn thành dài hơn một chút (khoảng 4 đến 5 giây), nhưng nó lại vượt trội về độ chính xác, với MAPE đặc biệt thấp vào mùa đông, chỉ ở mức 14,323%. Ngược lại, LSTM và GRU thể hiện hiệu suất kém hơn đáng kể, với RMSE cao hơn nhiều, với 2305,123 kW và 2312,86 kW vào mùa thu, và MAPE cực kỳ cao, đặc biệt là đối với LSTM, đạt 18285,84% vào mùa xuân. Ngoài ra, thời gian huấn luyện cho LSTM và GRU rất dài, lên tới 151,015 giây và 109,434 giây vào mùa xuân, khiến chúng ít phù hợp hơn cho các nhiệm vụ đòi hỏi dự báo nhanh và chính xác. Dựa trên các phân tích này, có thể kết luận rằng XGBoost, LightGBM và RF tối ưu hơn LSTM và GRU cho nhiệm vụ này, với thời gian đào tạo nhanh hơn đáng kể, độ chính xác tốt hơn và hiệu suất ổn định hơn qua các mùa. Việc kết hợp các mô hình mạnh mẽ và ổn định như XGBoost, LightGBM và RF trong một Stacking Ensemble không chỉ giúp giảm lỗi mà còn tạo ra một mô hình có độ chính xác cao hơn và thời gian xử lý nhanh hơn so với việc sử dụng LSTM và GRU. Điều này chứng minh rõ ràng tính ưu việt của các mô hình dựa trên cây quyết định trong dự báo năng lượng mặt trời.

Với các thông số của các mô hình XGBoost, LightGBM và RF sau khi huấn luyện, mô hình dự báo kết hợp đề xuất là mô hình SE-XGB-LiGBM-RF-OW sẽ được sử dụng

để dự báo cho các ngày điển hình của mỗi mùa trong Bảng 4.2, Trọng số 0,33 được sử dụng cho bước thực hiện này, kết quả dự báo được thể hiện ở Bảng 4.4.

Bảng 4.4. Kết quả dự báo của mô hình kết hợp SE-XGB-LiGBM-RF-OW với trọng số mặc định bằng nhau trong ngày điển hình của mùa theo Bảng 4.2

Công suất		$P_{\max}$ (kW)	$P_{\min}$ (kW)	$P_{\text{trung bình}}$ mùa (kW)	RMSE (kW)	NRMSE (%)	MAPE (%)	NMAPE (%)
Mùa xuân	Thực tế	41541,56	0	7928,31	1184,44	2,39	8,62	1,10
	Dự báo	40989,01	1,547					
	$\Delta P_{\text{thực tế-dự báo}}$	552,55	1,547					
Mùa hạ	Thực tế	40287,79	0	9970,87	968,62	1,95	14,14	0,73
	Dự báo	39994,85	3,400					
	$\Delta P_{\text{thực tế-dự báo}}$	292,94	3,400					
Mùa thu	Thực tế	40710,07	0	6495,09	666,62	1,34	17,05	0,57
	Dự báo	39179,89	0,610					
	$\Delta P_{\text{thực tế-dự báo}}$	1530,18	0,610					
Mùa đông	Thực tế	39052,66	0	3780,308	1306,62	2,64	17,32	0,96
	Dự báo	40680,15	1,819					
	$\Delta P_{\text{thực tế-dự báo}}$	1627,49	1,819					

Từ Bảng 4.4 nhận thấy mô hình kết hợp SE-XGB-LiGBM-RF-OW cho sai số RMSE, NRMSE, NMAPE khi dự báo cho các ngày điển hình đều cho kết quả không vượt quá 5%, Sai số MAPE có cao hơn tại các mùa tương ứng với mùa đông cao nhất đạt 17,322%,

Để thấy được sự hỗ trợ lẫn nhau để cải thiện hiệu suất dự báo của các mô hình khi kết hợp chúng trong mô hình Stacking Ensemble, nghiên cứu sinh tiến hành so sánh kết quả dự báo của mô hình kết hợp với kết quả dự báo của từng mô hình riêng biệt, Kết quả so sánh được cho trong Bảng 4.5 dưới đây

Bảng 4.5. So sánh kết quả dự báo của mô hình SE-XGB-LiGBM-RF-OW với các mô hình XGBoost, LightGBM và Random Forest

Các mùa	Các so sánh	XGBoost	LightGBM	Random Forest	SE-XGB-LiGBM-RF-OW
Mùa xuân	RMSE (kW)	1259,072	1185,638	1209,077	1184,441
	NRMSE (%)	2,543	2,395	2,442	2,392
	MAPE (%)	8,929	9,097	8,745	8,616
	NMAPE (%)	1,120	1,118	1,140	1,103
Mùa hạ	RMSE (kW)	1046,0006	951,009	961,522	968,627
	NRMSE (%)	2,113	1,921	1,942	1,956
	MAPE (%)	20,433	13,042	11,591	14,136
	NMAPE (%)	0,777	0,732	0,729	0,731
Mùa thu	RMSE (kW)	749,514	700,547	635,233	666,626
	NRMSE (%)	1,514	1,415	1,283	1,346
	MAPE	17,200	19,315	16,581	17,056
	NMAPE	0,641	0,602	0,536	0,574
Mùa đông	RMSE (kW)	2166,008	993,404	1147,764	1306,616
	NRMSE (%)	4,375	2,006	2,318	2,639
	MAPE (%)	18,720	18,522	16,989	17,322
	NMAPE (%)	1,352	0,987	0,901	0,961

Từ Bảng 4.5 có thể nhận thấy tại các mùa, mô hình SE-XGB-LiGBM-RF-OW là mô hình cho kết quả thấp hơn hoặc tương đương với các mô hình dự báo riêng lẻ. Tuy nhiên, mô hình kết hợp không phải luôn luôn tốt trong mọi trường hợp, ví dụ mô hình LightGBM đạt sai số NMAPE thấp nhất trong mùa đông, nhưng mô hình Random Forest lại là tốt nhất trong mùa thu và mùa hạ. Như vậy, khi kết hợp các mô hình dự báo trong mô hình SE-XGB-LiGBM-RF-OW, những ưu điểm của từng mô hình riêng lẻ đã được khai thác và hỗ trợ lẫn nhau, giúp mô hình kết hợp đạt hiệu suất tốt hơn so với việc sử dụng một mô hình đơn lẻ.

Sau khi nhận được kết quả dự báo từ mô hình SE-XGB-LiGBM-RF-OW với trọng số mặc định, nghiên cứu sinh sử dụng kết quả này làm đầu vào cho mô hình tinh chỉnh thông số tối ưu sử dụng thuật toán L-BFGS-B. Kết quả của việc tìm trọng số tối ưu cho từng mô hình được trình bày dưới dạng Bảng 4.6 sau đây:

Bảng 4.6. Trọng số tối ưu cho từng mô hình

Các mùa	Trọng số	XGBoost Model	LightGBM Model	Random Forest Model
Mùa xuân	Mặc định	0,33	0,33	0,33
	Tối ưu	0,135	0,373	0,517
Mùa hạ	Mặc định	0,33	0,33	0,33
	Tối ưu	-0,452	1,04	0,403
Mùa thu	Mặc định	0,33	0,33	0,33
	Tối ưu	-0,130	0,164	0,988
Mùa đông	Mặc định	0,33	0,33	0,33
	Tối ưu	-0,313	0,828	0,503

Chiến lược tối ưu hóa trọng số theo mùa (OW) được lựa chọn thay cho các chu kỳ ngắn hạn (như theo ngày hoặc theo tháng) nhằm đạt được sự cân bằng hợp lý giữa độ chính xác theo đặc trưng khí hậu và tính ổn định vận hành của hệ thống dự báo. Việc duy trì bộ trọng số cố định trong từng mùa giúp hạn chế sai số do sự dao động mô hình và nguy cơ quá khớp đối với các nhiễu dữ liệu ngắn hạn, đồng thời giảm thiểu rủi ro mất ổn định hội tụ và gián đoạn quá trình dự báo tổng thể do tần suất tái huấn luyện quá cao. Trong bối cảnh xuất hiện các biến động thời tiết đột ngột hoặc hiện tượng cực đoan như bão, bộ trọng số w_i vẫn được giữ cố định theo đặc trưng mùa và giá trị dự báo tổng hợp là:

$$P_{\text{dự báo tổng hợp}} = \sum_{i=1}^n w_i P_i \quad (4.2)$$

Giá trị dự báo này vẫn thể hiện khả năng thích ứng linh hoạt và bám sát diễn biến thực tế. Điều này đạt được nhờ độ nhạy cao của các mô hình thành phần ở tầng dưới, nơi các giá trị P_i phản ứng tức thời trước sự thay đổi của dữ liệu khí tượng dự báo.

Bộ trọng số tối ưu theo mùa được xây dựng dựa trên một năm dữ liệu nền có chất lượng cao (năm 2022) và được thiết kế để sử dụng ổn định trong chu kỳ trung hạn từ 3 đến 5 năm. Cách tiếp cận này cho phép bảo toàn các quy luật vật lý đã học được, đồng thời hạn chế tác động tiêu cực của dữ liệu không đồng nhất trong các năm tiếp theo. Sau chu kỳ trung hạn, bộ trọng số có thể được cập nhật theo hướng học chuyển giao (Transfer Learning), trong đó các trọng số đã học được được xem như tri thức nền và chỉ được

hiệu chỉnh cục bộ trên dữ liệu mới, thay vì tái huấn luyện toàn bộ từ đầu, nhằm đảm bảo tính kế thừa, ổn định và hiệu quả tính toán của hệ thống

Sau khi đã có trọng số tối ưu cho từng mô hình, nghiên cứu sinh đã tích hợp các trọng số này vào mô hình kết hợp SE-XGB-LiGBM-RF-OW, Kết quả dự báo cho các ngày diễn hình của một mùa dựa trên Bảng 4.2 với các trọng số tối ưu, được minh họa trong Bảng 4.7 dưới đây:

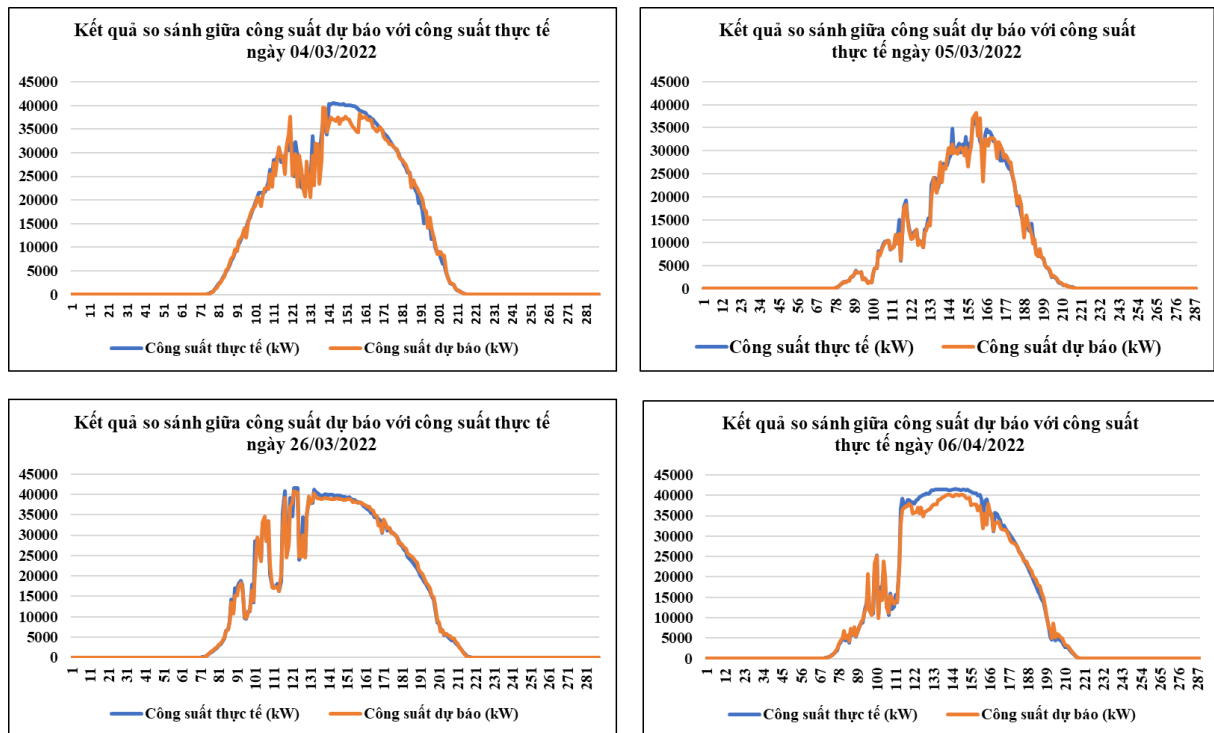
Bảng 4.7. Kết quả dự báo trong các ngày diễn hình của một mùa theo Bảng 4.2 với trọng số tối ưu cho mô hình SE-XGB-LiGBM-RF-OW

Công suất		$P_{\max}$ (kW)	$P_{\min}$ (kW)	$P_{\text{trung bình}}$ mùa (kW)	RMSE (kW)	NRMSE (%)	MAPE (%)	NMAPE (%)
Mùa xuân	Thực tế	41541,56	0	7928,31	1183,33	2,390	8,572	1,112
	Dự báo	40841,49	1,673					
	$\Delta P_{\text{thực tế-dự báo}}$	700,07	1,673					
Mùa hạ	Thực tế	40287,79	0	9970,87	939,332	1,897	12,616	0,744
	Dự báo	39869,53	0					
	$\Delta P_{\text{thực tế-dự báo}}$	418,26	0					
Mùa thu	Thực tế	40710,07	0	6495,09	635,920	1,284	16,721	0,539
	Dự báo	39194,77	0					
	$\Delta P_{\text{thực tế-dự báo}}$	1515,3	0					
Mùa đông	Thực tế	39052,66	0	3780,308	975,106	1,969	17,797	0,874
	Dự báo	38945,49	0					
	$\Delta P_{\text{thực tế-dự báo}}$	107,17	0					

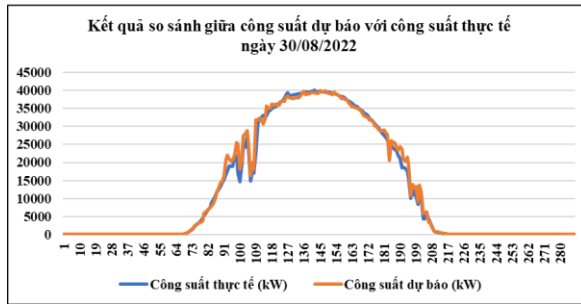
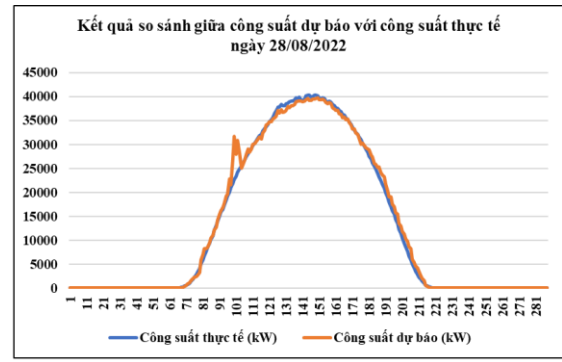
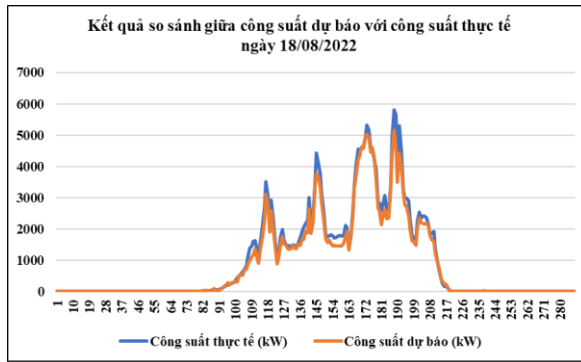
Từ Bảng 4.7, nhận thấy sử dụng trọng số tối ưu đã giảm sai số RMSE, NRMSE và MAPE một chút so với khi sử dụng trọng số mặc định, trong khi sai số NMAPE tăng không đáng kể trong mùa xuân và hạ và NMAPE giảm một chút trong mùa thu và đông, Trong các mùa, sai số NRMSE và NMAPE tỏ ra tương đối thấp, Tuy nhiên, sai số MAPE lại ở mức cao, đặc biệt là trong mùa thu và đông, Tại Việt Nam, tiêu chuẩn cho sai số dự báo công suất điện mặt trời là NMAPE không vượt quá 15% [10], vậy từ bảng 22, chúng ta có thể thấy rằng sai số NMAPE của mô hình SE-XGB-LiGBM-RF-OW hoàn

toàn đáp ứng tiêu chuẩn dự báo tại Việt Nam, Theo cách tính sai số MAPE truyền thống, sai số trong mùa xuân và hạ cũng đạt tiêu chuẩn, trong khi sai số trong mùa thu và đông cao hơn tiêu chuẩn một chút, Tóm lại, mô hình SE-XGB-LiGBM-RF-OW theo cách tính sai số MAPE và NMAPE đều có thể hoạt động hiệu quả tại Việt Nam.

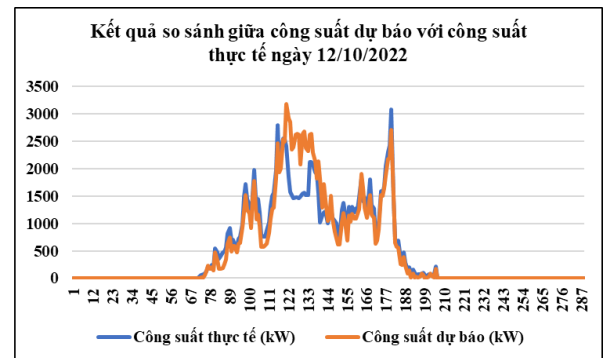
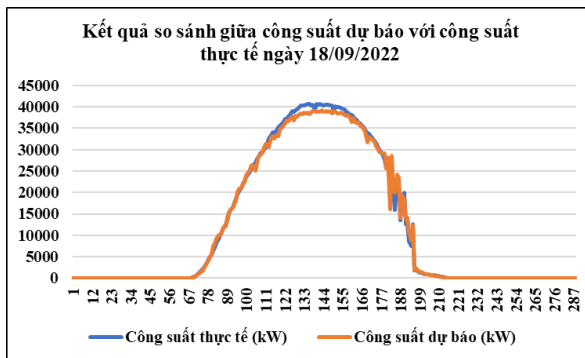
Để đánh giá sâu hơn hiệu quả của mô hình dự báo SE-XGB-LiGBM-RF-OW, Hình 4.4, Hình 4.5, Hình 4.6 và Hình 4.7 dưới đây trình bày các biểu đồ so sánh giữa giá trị dự báo và thực tế của các ngày điển hình theo từng mùa (1 ngày được chia ra làm 281 khoảng thời gian dự báo, mỗi khoảng thời gian dự báo là 5 phút), giúp minh họa trực quan mức độ phù hợp của các mô hình trong từng môi trường thời tiết khác nhau.



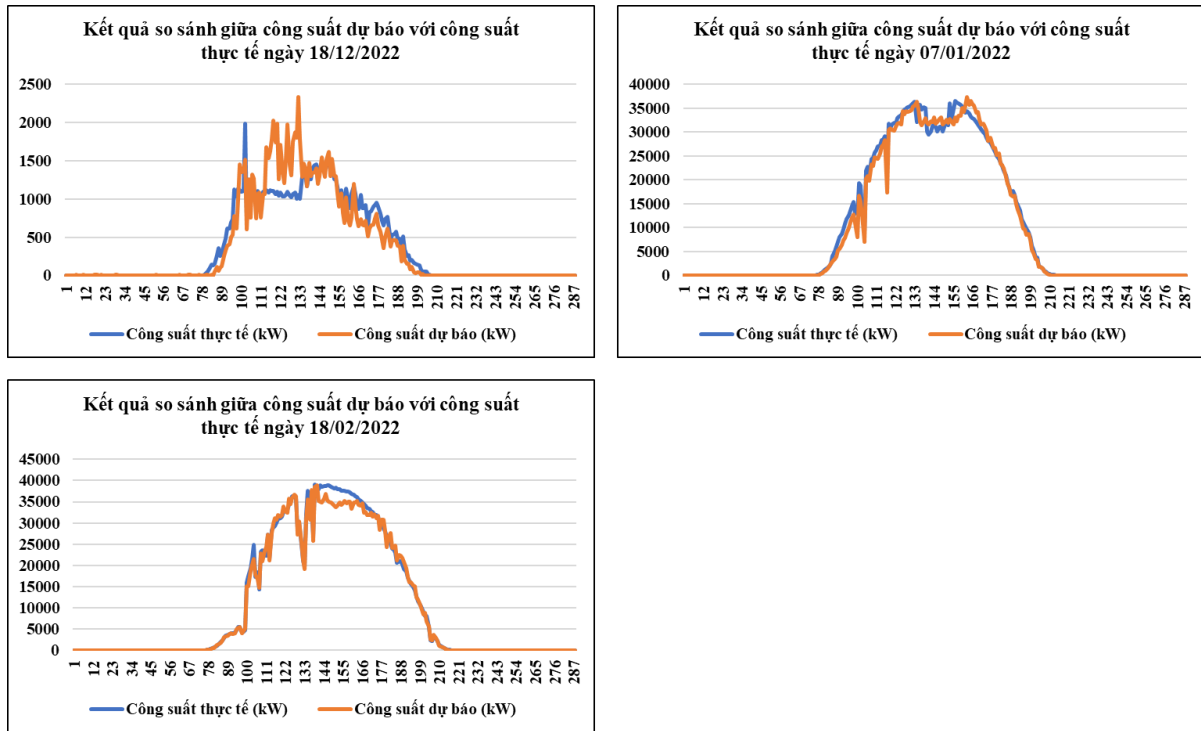
Hình 4.4. Kết quả so sánh công suất dự báo và công suất thực tế trong các ngày điển hình của mùa xuân



Hình 4.5. Kết quả so sánh công suất dự báo và công suất thực tế trong các ngày điển hình của mùa hạ



Hình 4.6. Kết quả so sánh công suất dự báo và công suất thực tế trong các ngày điển hình của mùa thu



Hình 4.7. Kết quả so sánh công suất dự báo và công suất thực tế trong các ngày điển hình của mùa đông

Kết quả so sánh giữa công suất dự báo và công suất thực tế theo từng mùa trong năm được thể hiện thông qua mô hình SE-XGB-LiGBM-RF-OW, áp dụng cho các ngày điển hình của mỗi mùa cho thấy trong phần lớn các trường hợp, đường cong dự báo bám sát xu hướng của đường cong công suất thực tế, cho thấy mô hình có độ chính xác cao trong việc phản ánh xu hướng biến động công suất. Nhìn chung, mô hình SE-XGB-LiGBM-RF-OW cho thấy hiệu quả rõ rệt trong việc nắm bắt và dự báo xu hướng chung của công suất phát điện mặt trời theo mùa.

Do phần trọng số tối ưu có xuất hiện trọng số âm của XGBoost trong ba mùa là hạ, thu và đông nên nghiên cứu sinh tiến hành kiểm tra xem trọng số này mang tính tích cực hay tiêu cực đối với mô hình dự báo bằng cách loại bỏ mô hình XGBoost này ra khỏi mô hình kết hợp. Kết quả so sánh sai số RMSE, NRMSE, MAPE, NMAPE khi loại bỏ XGBoost ra khỏi mô hình kết hợp được trình bày tại Bảng 4.8 dưới đây:

Bảng 4.8. Kết quả của mô hình dự báo SE-XGB-LiGBM-RF-OW với thông số tối ưu khi bỏ đi mô hình Xgboost

Mô hình kết hợp (với trọng số tối ưu)	Thời gian hoàn thành (s)	RMSE (kW)	NRMSE (%)	MAPE (%)	NMAPE (%)
---------------------------------------	--------------------------	-----------	-----------	----------	-----------

Mùa hạ	XGBoost+LightGBM+ Random Forest	29,386	939,332	1,897	12,616	0,921
	LightGBM+Random Forest	30,861	947,433	1,914	12,236	0,898
Mùa thu	XGBoost+LightGBM+ Random Forest	16,690	635,920	1,284	16,721	0,667
	LightGBM+Random Forest	21,728	636,545	1,285	16,679	0,666
Mùa đông	XGBoost+LightGBM+ Random Forest	22,555	975,106	1,969	17,797	1,081
	LightGBM+Random Forest	21,840	1009,547	2,039	96,688	1,073

Bảng 4.8 cho thấy rằng RMSE và NRMSE tăng lên khi loại bỏ lần lượt mô hình có trọng số âm là Xgboost khi dự báo cho các ngày điển hình trong mùa hạ, thu, đông, Với sai số MAPE thì khi bỏ mô hình Xgboost thì sai số tăng cao nhất trong mùa đông, còn 3 mùa kia có giảm đi nhưng không đáng kể, Như vậy, tuy giá trị trọng số của Xgboost là âm khi dự báo cho mùa hạ, thu, đông nhưng mô hình này cũng mang lại các đóng góp và cải thiện cho mô hình dự báo kết hợp SE-XGB-LiGBM-RF-OW. Trong mô hình tổng hợp, trong đó các dự báo từ nhiều mô hình như XGBoost, LightGBM và Random Forest được kết hợp, việc gán trọng số âm cho dự báo của mô hình XGBoost đóng vai trò quan trọng. Trọng số âm này có nghĩa là đầu ra của XGBoost bị trừ đi, thay vì được thêm vào, dự báo tổng hợp cuối cùng. Quyết định như vậy thường được đưa ra khi mô hình XGBoost có xu hướng dự báo quá mức hoặc nắm bắt các mẫu có thể dẫn đến quá khớp (over fitting) dẫn đến đưa nhiều hoặc lỗi tiềm ẩn vào dự báo cuối cùng. Bằng cách áp dụng trọng số âm, mô hình tổng hợp sẽ hiệu chỉnh hiệu quả các xu hướng này, giảm lỗi tổng thể. Phương pháp này không chỉ cải thiện tính ổn định của mô hình tổng hợp bằng cách khai thác những ưu điểm đặc thù của XGBoost, đồng thời giảm thiểu các điểm yếu của nó. Về bản chất, việc sử dụng trọng số âm trong việc tổng hợp là một phương pháp mạnh mẽ giúp giảm lỗi và tăng tính ổn định của các dự báo, khiến mô hình kết hợp trở nên đáng tin cậy hơn so với bất kỳ mô hình đơn lẻ nào.

Kết quả của quá trình dự báo sử dụng mô hình SE-XGB-LiGBM-RF-OW đã cho thấy một hiệu suất khá ổn định và đáng tin cậy, Sự kết hợp của ba mô hình XGBoost,

LightGBM và Random Forest đã mang lại một mô hình kết hợp mạnh mẽ, giúp cải thiện tính chính xác của dự báo.

Quá trình này đã minh chứng cho sự hiệu quả của phương pháp Stacking Ensemble trong việc giảm thiểu hiện tượng overfitting và tăng tính ổn định của mô hình, Bằng cách kết hợp các mô hình độc lập với nhau, chúng ta đã tạo ra một mô hình dự báo mạnh mẽ có khả năng tổng quát hóa tốt trên các tập dữ liệu mới, Tuy nhiên, cần lưu ý rằng việc đánh giá và tối ưu hóa các trọng số của mô hình vẫn là bước quan trọng trong quá trình này, Đồng thời, việc xem xét và điều chỉnh các phương pháp đo lường hiệu suất như RMSE, NRMSE, MAPE và NMAPE là rất quan trọng để đảm bảo tính khách quan và chính xác của dự báo.

c. Đánh giá kết quả dự báo của mô hình SE-XGB-LGBM-RF-OW

Kết quả của quá trình dự báo sử dụng mô hình SE-XGB-LGBM-RF-OW đã cho thấy một hiệu suất khá ổn định và đáng tin cậy. Sự kết hợp của ba mô hình XGBoost, LightGBM và Random Forest đã mang lại một mô hình kết hợp mạnh mẽ, giúp cải thiện tính chính xác của dự báo. Mô hình kết hợp SE-XGB-LGBM-RF-OW có hiệu suất tốt hơn các mô hình học sâu như LSTM và GRU, mang lại độ chính xác cao hơn và thời gian đào tạo nhanh hơn đáng kể.

Quá trình kết hợp này đã minh chứng cho sự hiệu quả của phương pháp Stacking Ensemble trong việc giảm thiểu hiện tượng overfitting và tăng tính ổn định của mô hình, bằng cách kết hợp các mô hình độc lập với nhau, chúng ta đã tạo ra một mô hình dự báo mạnh mẽ có khả năng tổng quát hóa tốt trên các tập dữ liệu mới. Tuy nhiên, cần lưu ý rằng việc đánh giá và tối ưu hóa các trọng số của mô hình vẫn là bước quan trọng trong quá trình này, Đồng thời, việc xem xét và điều chỉnh các phương pháp đo lường hiệu suất như RMSE, NRMSE, MAPE và NMAPE là rất quan trọng để đảm bảo tính khách quan và chính xác của dự báo.

Tóm lại, với sai số NMAPE $< 2\%$ tại tất cả các mùa, quá trình dự báo sử dụng mô hình SE-XGB-LGBM-RF-OW đã đem lại kết quả khả quan và là một công cụ hữu ích trong việc dự báo công suất điện mặt trời tại Việt Nam.

4.2. Xây dựng mô hình dự báo công suất phát điện mặt trời cho nhiều nhà máy

Sau khi kết hợp các mô hình XGB, LGBM và RF để xây dựng và phát triển mô hình dự báo công suất phát SE-XGB-LGBM-RF-OW cho một nhà máy điện mặt trời, bước tiếp theo nghiên cứu sinh sẽ mở rộng việc kết hợp ba mô hình trên để áp dụng cho dự

báo công suất điện mặt trời ở ba nhà máy đặt tại Thanh Hóa (miền bắc), Quảng Trị (miền trung) và Đắk-lắc (tây nguyên). Mỗi vùng có đặc trưng khí hậu và điều kiện vận hành khác nhau, ảnh hưởng đáng kể đến hiệu quả phát điện từ năng lượng mặt trời. Do đó, việc kiểm tra khả năng tổng quát hóa và độ chính xác của mô hình dự báo khi áp dụng cho nhiều vùng là rất cần thiết.

Tuy nhiên, quá trình tối ưu trọng số trong mô hình SE-XGB-LGBM-RF-OW yêu cầu rất nhiều công sức và thời gian, bao gồm việc huấn luyện theo mùa, đánh giá hiệu suất từng mô hình thành phần và tinh chỉnh tổ hợp. Vì vậy, trong phạm vi dự báo cho nhiều nhà máy, nghiên cứu sinh không tiến hành tối ưu trọng số riêng biệt cho từng trường hợp, mà thay vào đó đề xuất hai phương pháp mở rộng đơn giản và khả thi hơn, cụ thể như sau:

- **Kịch bản 1 xây dựng mô hình Selector-Model (tự động lựa chọn mô hình):** Gộp dữ liệu của cả ba nhà máy thành một tập dữ liệu chung, sau đó tiến hành huấn luyện ba mô hình thành phần XGBoost, LightGBM và Random Forest trên toàn bộ tập này. Mô hình Selector-Model sẽ được tích hợp để tự động lựa chọn mô hình con phù hợp nhất tại từng thời điểm và trong từng điều kiện thời tiết cụ thể. Cách làm này giúp mô hình thích ứng tốt với sự biến đổi khí hậu theo vùng và theo mùa mà không cần huấn luyện riêng cho từng nhà máy.
- **Kịch bản 2 (Mô hình tổ hợp SE-XGB-LGBM-RF với trọng số bằng nhau):** Mỗi mô hình thành phần (XGB, LGBM, RF) được huấn luyện riêng biệt trên tập dữ liệu của từng nhà máy, sau đó được kết hợp thành mô hình tổ hợp SE-XGB-LGBM-RF với trọng số bằng nhau giữa ba mô hình thành phần. Trọng số không được tối ưu thêm nhằm giảm thời gian xử lý và giữ sự đồng đều trong so sánh. Kết quả dự báo từ mô hình này sẽ được sử dụng để đối chiếu với kịch bản 1, nhằm đánh giá hiệu quả tổng thể và khả năng mở rộng thực tế của mô hình tổ hợp.

4.2.1. Triển khai mô hình Selector-Model cho XGB, LGBM, và RF

a. Giới thiệu về mô hình Selector-Model

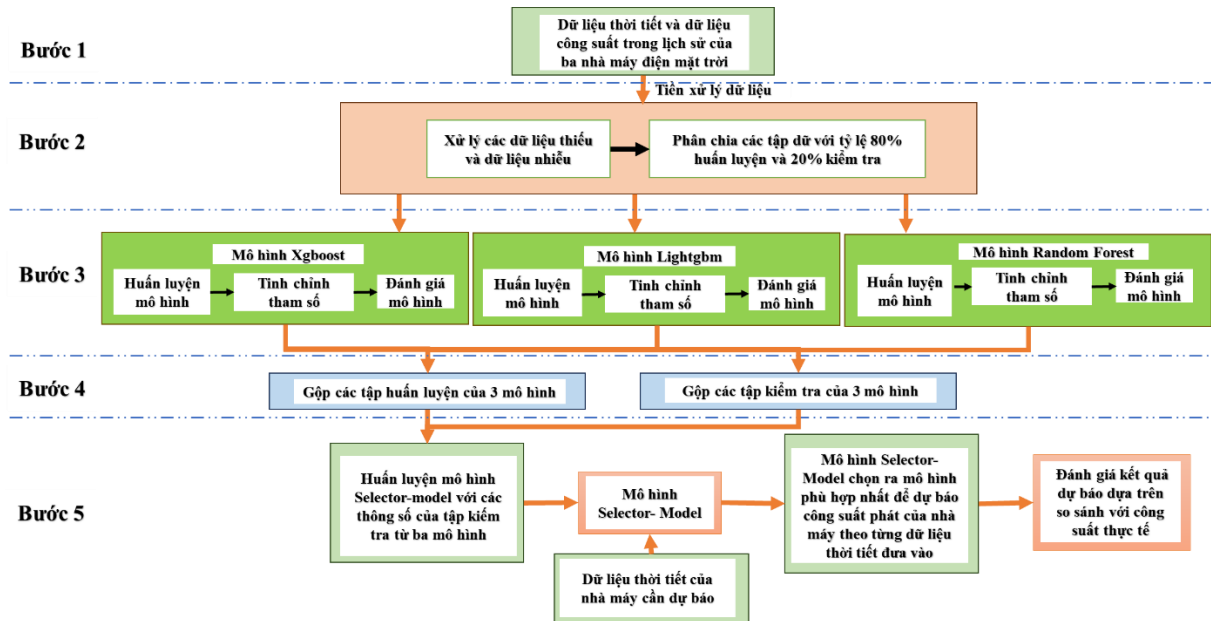
Mô hình Selector-Model được xây dựng nhằm mục đích tự động lựa chọn mô hình con phù hợp nhất tại mỗi thời điểm dựa trên đặc trưng đầu vào của dữ liệu thời tiết. Thay vì sử dụng một mô hình cố định cho tất cả điều kiện thời tiết, hệ thống này có thể nhận diện đặc điểm của dữ liệu đầu vào và lựa chọn mô hình dự báo phù hợp nhất.

Cấu trúc mô hình gồm các phần chính:

- Ba mô hình thành phần (XGBoost, LightGBM, Random Forest) được huấn luyện chung trên dữ liệu của cả ba nhà máy.
- Mô hình Selector-Model: Một mô hình phân loại được huấn luyện để lựa chọn mô hình thành phần phù hợp nhất dựa trên sai số dự báo (MAE, RMSE).

b. Quy trình triển khai Selector-Model

Quy trình triển khai mô hình Selector-Model được thể hiện trong Hình 4.8 dưới đây:



Hình 4.8. Sơ đồ thuật toán mô hình thiết lập mô hình Selector-Model

Bước 1: Chuẩn bị dữ liệu và gộp dữ liệu từ ba nhà máy

- Tập dữ liệu lịch sử được sử dụng trong phần này gồm tập dữ liệu thời tiết và công suất phát của ba nhà máy điện mặt trời đã được trình bày trong phần 2.1.3: Nhà máy đặt tại Quảng trị có công suất 49,5 MW, nhà máy tại Đắk-Lắk có công suất 50MW và nhà máy đặt tại Thanh Hóa có công suất 30MW. Các tập dữ liệu này đều được thu thập từ 01/01/2024 đến 31/12/2024 với tần suất lấy mẫu 5 phút/lần, có tổng cộng 105.120 điểm dữ liệu mỗi nhà máy. Đặc trưng đầu vào của ba tập dữ liệu gồm có: bức xạ mặt trời, nhiệt độ môi trường, thời gian các tháng trong năm, mã định danh của nhà máy và công suất phát điện mặt trời.
- Dữ liệu của ba nhà máy này được kết hợp lại thành một tập dữ liệu duy nhất.
- Tạo thêm một tập dữ liệu dùng cho dự báo bằng cách lấy ngẫu nhiên 2 ngày liên tiếp mỗi tháng (24 ngày/năm) giữ nguyên các đặc trưng thời tiết và bỏ qua giá trị công suất phát ra khỏi tập dữ liệu lịch sử của ba nhà máy (mỗi nhà máy có một tập dữ liệu dự báo riêng).
- Cách xử lý dữ liệu được thực hiện giống phần 2.1.2.

- Dữ liệu được chia theo tỷ lệ 80% huấn luyện - 20% kiểm tra để huấn luyện các mô hình dự báo.

Bước 2: Huấn luyện ba mô hình XGBoost, LightGBM, RF trên tập dữ liệu chung

- Huấn luyện ba mô hình dự báo riêng biệt (XGBoost, LightGBM, Random Forest) trên tập dữ liệu kết hợp của ba nhà máy.
- Các mô hình được xây dựng dựa trên các đặc trưng thời tiết và thời gian của các tháng trong năm, nhằm dự báo công suất phát điện tại từng thời điểm.

Bước 3: Xác định mô hình con tối ưu tại mỗi thời điểm

- Trên tập kiểm tra, tính toán sai số MAE và RMSE của từng mô hình con tại mỗi thời điểm.
- Với mỗi mẫu dữ liệu, mô hình có sai số nhỏ nhất sẽ được gán làm “nhãn phân loại”, phục vụ huấn luyện Selector-Model, đây là bước chuẩn bị dữ liệu huấn luyện cho mô hình phân loại lựa chọn mô hình con.

Bước 4: Huấn luyện Selector-Model để lựa chọn mô hình tối ưu

- Dữ liệu đầu vào để huấn luyện Selector-Model bao gồm các đặc trưng thời tiết (bức xạ mặt trời, nhiệt độ môi trường, tháng...), cùng với mã định danh nhà máy.
- Dữ liệu đầu ra (nhãn phân loại) là tên mô hình con (XGBoost, LightGBM hoặc Random Forest) có hiệu suất dự báo tốt nhất (tức có MAE hoặc RMSE thấp nhất) tại mỗi mẫu trong tập kiểm tra.
- Thuật toán Random Forest được sử dụng để xây dựng mô hình phân loại nhằm học cách xác định mô hình con phù hợp nhất với từng điều kiện thời tiết đầu vào tại từng nhà máy.
- Sau khi huấn luyện, mô hình Selector-Model có khả năng tự động lựa chọn mô hình con tối ưu tương ứng với điều kiện thời tiết thực tế tại thời điểm dự báo, giúp nâng cao độ chính xác và khả năng mở rộng hệ thống.
- Để duy trì độ chính xác bền vững, mô hình Selector-Model sẽ được tái huấn luyện định kỳ và học chuyển giao (Transfer Learning). Theo đó, bộ Selector-Model sẽ không cố định các quy luật lựa chọn và các mô hình thành phần sẽ được huấn luyện cập nhật dựa trên dữ liệu vận hành thực tế của 1-2 năm gần nhất của các nhà máy. Cơ chế này cho phép mô hình học các đặc tính mới của hệ thống để thích nghi với các thay đổi trong hệ thống như sự xuống cấp tự nhiên của các tấm pin (thường từ 0,5%–1% mỗi năm), sự thay đổi cấu hình thiết bị hoặc các tác động dài hạn từ môi

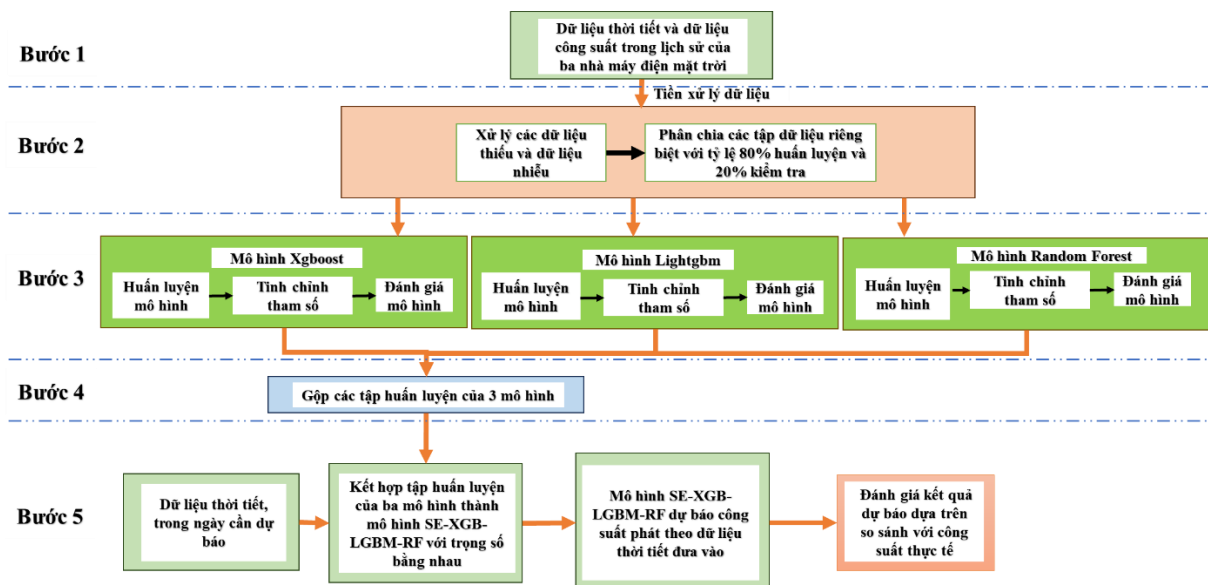
trường có thể làm dịch chuyển tương quan giữa bức xạ đầu vào và công suất đầu ra. Việc này giúp đảm bảo hiệu suất dự báo không bị suy giảm theo vòng đời của dự án.

Bước 5: Dự báo công suất phát điện với mô hình Selector-Model

- Dựa trên đặc trưng của dữ liệu đầu vào của các nhà máy tại mỗi thời điểm trong tập dự báo, mô hình Selector-Model xác định mô hình con dự báo phù hợp nhất trong số các mô hình đã được huấn luyện.
- Mô hình con được lựa chọn sau đó được sử dụng để thực hiện dự báo công suất phát điện cho nhà máy tương ứng tại thời điểm trong tập dữ liệu dùng cho dự báo.
- Đánh giá kết quả dự báo cho từng nhà máy của mô hình Selector-Model dựa trên các sai số RMSE, NRMSE, MAPE, NMAPE.

4.2.2. Triển khai huấn luyện mô hình SE-XGB-LGBM-RF trên từng tập dữ liệu nhà máy riêng biệt

Quy trình triển khai mô hình SE-XGB-LGBM-RF cho việc dự báo công suất phát với nhiều nhà máy được thể hiện trong Hình 4.9 dưới đây:



Hình 4.9. Sơ đồ thuật toán dự báo cho nhiều nhà máy sử dụng mô hình SE-XGB-LGBM-RF

Bước 1: Chuẩn bị dữ liệu riêng cho từng nhà máy

- Dữ liệu thu thập từ ba nhà máy giống phần dữ liệu lịch sử đầu vào của mô hình Selector- Model.

- Dữ liệu được xử lý giống như trong mô hình SE-XGB-LGBM-RF-OW, bao gồm các đặc trưng đầu vào: nhiệt độ môi trường, bức xạ mặt trời, thời điểm các tháng trong năm.

Bước 2: Chia tập dữ liệu huấn luyện/kiểm tra

- Chia tập dữ liệu theo tỷ lệ 80% huấn luyện - 20% kiểm tra cho từng nhà máy.
- Tạo thêm một tập dữ liệu dự báo bằng cách lấy ngẫu nhiên 2 ngày liên tiếp mỗi tháng (24 ngày/năm) ra khỏi tập huấn luyện của mỗi nhà máy.

Bước 3: Huấn luyện mô hình SE-XGB-LGBM-RF cho từng nhà máy

- Ba mô hình con (XGBoost, LightGBM, Random Forest) được huấn luyện riêng trên tập dữ liệu huấn luyện cho từng nhà máy.
- Kết hợp ba mô hình con này theo phương pháp Stacking Ensemble với trọng số bằng nhau.

Bước 4: Đánh giá mô hình

- Kiểm tra sai số trên tập test bằng các chỉ số RMSE, NRMSE MAPE, NMAPE.
- So sánh kết quả dự báo giữa mô hình huấn luyện riêng biệt và mô hình Selector-Model để đánh giá tính hiệu quả.

4.2.3. Kết quả thực nghiệm

Kết quả dự báo từ mô hình khi sử dụng mô hình Selector-Model và khi được huấn luyện riêng biệt (SE-XGB-LGBM-RF) cho từng nhà máy được thể hiện trong Bảng 4.9 dưới đây:

Bảng 4.9. Kết quả dự báo từ mô hình Selector-Model và mô hình SE-XGB-LGBM-RF (khi được huấn luyện riêng biệt) cho từng nhà máy

Mô hình dự báo sử dụng Selector-Model				
Nhà máy	RMSE (kW)	NRMSE (%)	MAPE (%)	NMAPE (%)
Thanh Hóa	2718,84	9,06	85,5	4,94
Quảng Trị	2310,55	4,66	49,2	2,07
Đắk-Lắk	3938,48	7,87	49,5	4,22
Mô hình SE-XGB-LGBM-RF dự báo riêng cho 3 nhà máy				
Nhà máy	RMSE (kW)	NRMSE (%)	MAPE (%)	NMAPE (%)
Thanh Hóa	2773,94	9,26	78,75	4,97

Quảng Trị	1897,83	3,83	8,7	1,32
Đắk-Lắk	3309,02	6,68	40,67	3,26

Bảng 4.9 cho thấy cả hai hướng tiếp cận mô hình Selector-Model (dự báo trên tập dữ liệu gộp) và mô hình huấn luyện riêng biệt cho từng nhà máy đều đạt các chỉ số sai số nằm trong giới hạn yêu cầu kỹ thuật, đặc biệt là NMAPE đều dưới 5%, đáp ứng yêu cầu của NSMO (trước đây là A0) trong đánh giá độ chính xác của hệ thống dự báo công suất điện mặt trời. Xét tổng thể trên cả bốn chỉ số đánh giá (RMSE, NRMSE, MAPE và NMAPE), mô hình huấn luyện riêng biệt cho thấy hiệu năng vượt trội hơn rõ rệt tại nhà máy Quảng Trị với MAPE giảm mạnh từ 49,2% xuống 8,7%, RMSE giảm từ 2310,55 kW xuống 1897,83 kW và NMAPE từ 2,07% xuống còn 1,32%. Kết quả này cho thấy mô hình SE-XGB-LGBM-RF dự báo riêng biệt cho các nhà máy có khả năng xử lý tốt hơn các nhiễu cục bộ trong dữ liệu, đặc biệt tại các thời điểm công suất nhỏ yếu tố thường gây sai số cao cho MAPE. Tại hai nhà máy Thanh Hóa và Đắk Lắk, sự chênh lệch giữa hai mô hình là không đáng kể, đặc biệt với chỉ số NMAPE đều dưới 5%. Điều này cho thấy mô hình Selector-Model vẫn duy trì hiệu quả tốt trong điều kiện dữ liệu gộp, và hoàn toàn có thể là lựa chọn phù hợp khi cần triển khai đồng bộ cho nhiều nhà máy cùng lúc.

Ưu điểm của mô hình Selector-Model là tiết kiệm thời gian huấn luyện, không cần xây dựng mô hình riêng biệt cho từng nhà máy, khả năng tổng quát hóa tốt, có thể mở rộng cho nhiều nhà máy khác nhau. Phù hợp cho các hệ thống có nhiều nhà máy điện mặt trời với điều kiện thời tiết tương tự nhau. Bên cạnh khả năng thích ứng với điều kiện thời tiết, hệ thống Selector-Model cũng được thiết kế để đối phó với sự thay đổi đặc tính vật lý của hệ thống điện mặt trời theo thời gian. Tuy nhiên, sai số của mô hình Selector-Model cao hơn so với mô hình huấn luyện riêng biệt, đặc biệt khi dữ liệu của các nhà máy có sự khác biệt lớn. Không phù hợp cho các nhà máy có điều kiện khí hậu thay đổi phức tạp.

Ưu điểm của mô hình huấn luyện riêng biệt (SE-XGB-LGBM-RF) cho từng nhà máy). Độ chính xác cao hơn, xử lý tốt hơn các dữ liệu nhiễu giúp giảm sai số đáng kể so với mô hình Selector-Model. Việc tối ưu hóa mô hình theo đặc trưng khí hậu từng vùng, giúp dự báo chính xác hơn. Tuy nhiên, mô hình huấn luyện riêng biệt yêu cầu nhiều công sức huấn luyện và quản lý mô hình hơn. Không linh hoạt khi mở rộng sang các nhà máy mới, cần phải huấn luyện lại từ đầu.

4.2.4. Đánh giá kết quả các mô hình dự báo

Dựa trên các kết quả thực nghiệm thu thập từ ba nhà máy điện mặt trời đại diện cho các vùng khí hậu khác nhau tại Thanh Hóa, Quảng Trị và Đắk Lắk, NCS tiến hành phân tích và thảo luận sâu về mối quan hệ giữa đặc thù khí hậu vùng miền và hiệu quả dự báo của các mô hình được đề xuất.

a. Phân tích tính thích nghi theo đặc thù khí hậu vùng miền

Kết quả thực nghiệm cho thấy sai số dự báo của mô hình có sự phân hóa rõ rệt theo các đới khí hậu, phản ánh mức độ thích nghi khác nhau của thuật toán trước các hình thái thời tiết đặc trưng tại Việt Nam Bảng 4.10. Điều này cho thấy hiệu quả dự báo không chỉ phụ thuộc vào cấu trúc mô hình, mà còn chịu ảnh hưởng đáng kể bởi điều kiện khí tượng khí hậu địa phương.

Bảng 4.10. Tổng hợp hiệu quả dự báo và khả năng thích nghi của mô hình Selector-Model và mô hình SE-XGB-LGBM-RF

Khu vực	Đặc điểm khí hậu	Sai số (NMAPE %)	Phân tích cơ chế thích nghi
Thanh Hóa	Bắc Trung Bộ: Mây mù, mưa phùn (mùa Đông-Xuân), bức xạ yếu.	4,94 - 4,97	Điều kiện bức xạ thấp và nhiều cao làm gia tăng độ bất định của công suất phát. Trong bối cảnh này, thành phần LightGBM thể hiện ưu thế nhờ chiến lược học tập trung vào các điểm gây sai số lớn, góp phần cải thiện độ ổn định của dự báo.
Quảng Trị	Nam Trung Bộ: Gió Phơn Tây Nam nhiệt độ cao, bão và áp thấp nhiệt độ.	1,32 - 2,07	Nhiệt độ môi trường cao và biến thiên mạnh ảnh hưởng trực tiếp đến hiệu suất tấm pin. Việc tích hợp biến nhiệt độ cho phép mô hình bù trừ hiệu

			quả các tổn thất do suy giảm hiệu suất nhiệt.
Đắc-Lắc	Tây Nguyên: Phân hóa mùa mưa – mùa khô, mưa rào bất chợt trong ngày.	3,26 – 4,22	Công suất phát có thể biến động đột ngột do các hiện tượng thời tiết cục bộ. Bộ Selector-Model cho phép chuyển dịch lựa chọn nhanh sang các mô hình có độ nhạy cao như XGBoost, từ đó nâng cao độ chính xác dự báo

Từ các phân tích trong Bảng 4.10, có thể khẳng định khả năng đáp ứng đa dạng của hai cấu trúc mô hình đề xuất đối với các kịch bản thời tiết thực tế. Mô hình SE-XGB-LGBM-RF cho thấy ưu thế trong việc mô hình hóa chính xác các đặc trưng vi khí hậu của từng địa điểm cụ thể. Ngược lại, mô hình Selector-Model thể hiện tính linh hoạt vượt trội thông qua cơ chế tự động chuyển đổi mô hình dựa trên đặc điểm dữ liệu tức thời. Cơ chế này giúp hệ thống không chỉ vận hành hiệu quả ở các khu vực tiềm năng lớn mà còn thích nghi tốt với các hình thái thời tiết cực đoan, mây mù hoặc bức xạ yếu, vốn là những thách thức lớn nhất đối với công tác dự báo điện mặt trời tại Việt Nam.

b. Đánh giá chiến lược lựa chọn mô hình trong vận hành thực tế

Trên cơ sở phân tích đồng thời tính thích nghi khí hậu và hiệu quả tính toán, NCS đề xuất một số định hướng áp dụng mô hình trong thực tiễn vận hành hệ thống điện như sau:

- **Ưu tiên độ chính xác tuyệt đối:** Trong kịch bản vận hành tại từng nhà máy đơn lẻ, với hạ tầng tính toán đủ mạnh và dữ liệu lịch sử đầy đủ, mô hình huấn luyện riêng biệt SE-XGB-LGBM-RF được xem là lựa chọn tối ưu. Cách tiếp cận này cho phép mô hình học sâu các đặc trưng vi khí hậu tại chỗ, từ đó giảm thiểu sai số dự báo NMAPE xuống mức thấp nhất.
- **Ưu tiên khả năng mở rộng và tính linh hoạt:** Đối với các đơn vị quản lý danh mục nhiều nhà máy hoặc các hệ thống điện mặt trời phân tán quy mô nhỏ, mô hình Selector-Model thể hiện ưu thế vượt trội. Nhờ khả năng tự động lựa chọn và điều

chỉnh mô hình con phù hợp với từng bối cảnh dữ liệu, phương pháp này giúp giảm đáng kể chi phí tính toán và nhu cầu huấn luyện lại mô hình.

- Chiến lược kết hợp đa tầng: Trong các hệ thống điều độ quy mô lớn, NCS đề xuất chiến lược kết hợp theo hai giai đoạn: sử dụng Selector-Model để khởi tạo dự báo nhanh cho các nhà máy mới hoặc trong giai đoạn thiếu dữ liệu, sau đó tiến hành tinh chỉnh (fine-tuning) các mô hình riêng biệt cho từng địa điểm khi đã tích lũy đủ dữ liệu lịch sử.

Nhìn chung, việc kết hợp giữa phân tích đặc thù khí hậu vùng miền và chiến lược lựa chọn mô hình linh hoạt đã khẳng định tính thực tiễn và khả năng ứng dụng cao của phương pháp đề xuất. Các kết quả đạt được cho thấy luận án đóng góp một công cụ dự báo tin cậy, góp phần hỗ trợ hiệu quả quá trình tích hợp năng lượng tái tạo vào hệ thống điện Việt Nam.

4.3. Đánh giá chi phí tính toán và hiệu quả sử dụng tài nguyên hệ thống

Bên cạnh việc đánh giá độ chính xác qua các chỉ số sai số, khả năng triển khai thực tế của một hệ thống dự báo còn phụ thuộc mật thiết vào chi phí tính toán. Một mô hình tối ưu không chỉ cần đảm bảo sai số thấp nhất mà còn phải đáp ứng yêu cầu về tốc độ phản hồi trong điều độ thời gian thực và khả năng vận hành trên hạ tầng phần cứng thông dụng. Phần này thực hiện phân tích chi tiết chi phí tính toán cho hai cấu trúc đề xuất: mô hình tổ hợp trọng số SE-XGB-LGBM-RF-OW và hệ thống lựa chọn tối ưu Selector-Model.

Máy tính được NCS sử dụng cho việc đánh giá chi phí tính của hai mô hình có cấu hình như Bảng 4.11 dưới đây:

Bảng 4.11. Cấu hình phần cứng và phần mềm sử dụng trong nghiên cứu

Thành phần	Thông số chi tiết
Thiết bị	Laptop Dell Precision 3520
Bộ vi xử lý (CPU)	Intel® Core™ i7-7700HQ (4 nhân, 8 luồng), xung nhịp 2,80 GHz
Bộ nhớ (RAM)	8 GB DDR4
Đồ họa (GPU)	Intel® HD Graphics 630 (Tích hợp)
Hệ điều hành	Microsoft Windows 10 Pro 64-bit (Build 18363)

Nền tảng tính toán	CPU-based (Không sử dụng tăng tốc GPU)
---------------------------	--

Nhằm kiểm chứng tính hiệu quả về mặt tài nguyên của các thuật toán đề xuất (như SE-XGB-LGBM-RF-OW và Selector-Model), quá trình huấn luyện và dự báo hoàn toàn được thực hiện dựa trên nền tảng tính toán CPU. Hệ thống đồ họa tích hợp Intel® HD Graphics 630 chỉ được sử dụng cho các tác vụ hiển thị thông thường, không tham gia vào quá trình tăng tốc tính toán học máy.

4.3.1. Đánh giá chi phí tính toán của SE-XGB-LGBM-RF-OW và các mô hình so sánh

Trong kịch bản này, các mô hình được đánh giá dựa trên tập dữ liệu huấn luyện là tập dữ liệu của Quảng Trị được trình bày trong phần 2.1, tập dữ liệu dự báo từ 01 đến 03 tháng 03 năm 2021 nhằm đảm bảo tính đồng nhất về điều kiện thử nghiệm.

Kết quả đánh giá chi phí tính toán của mô hình SE-XGB-LGBM-RF-OW được trình bày trong bảng Bảng 4.12 dưới đây:

Bảng 4.12. Tổng hợp hiệu quả chi phí tính toán của mô hình SE-XGB-LGBM-RF-OW và các mô hình đơn lẻ

Mô hình	Thời gian dự báo (s)	Thời gian huấn luyện (s)
SE-XGB-LGBM-RF-OW	22,797	19,31
LSTM	92,84	594,52
GRU	90,88	609,59
BiGRU	94,52	562,68
XGBoost	2,17	2,2
LightGBM	0,85	0,61
RF	18,37	16,58

Kết quả thực nghiệm cho thấy một sự phân hóa rõ rệt về hiệu suất tính toán giữa các nhóm thuật toán. Các mô hình dựa trên cấu trúc cây quyết định (XGBoost, LightGBM) thể hiện ưu thế vượt trội với thời gian huấn luyện cực ngắn, chỉ dưới 3 giây. Ngược lại, nhóm mô hình học sâu như LSTM, GRU và BiGRU đòi hỏi chi phí tính toán rất lớn với

thời gian huấn luyện trung bình vượt quá 580 giây, cao gấp gần 30 lần so với mô hình đề xuất. Đối với mô hình tổ hợp SE-XGB-LGBM-RF-OW, tổng thời gian huấn luyện ghi nhận được là 19,31 giây. Mặc dù thời gian dự báo (22,79 giây) có phần cao hơn các mô hình đơn lẻ do phải thực hiện cơ chế dự báo tuần tự để mô phỏng quy trình vận hành trực tuyến, nhưng con số này vẫn thấp hơn rất nhiều so với nhóm mô hình học sâu. Điều này chứng minh cấu trúc tổ hợp trọng số đề xuất đạt được sự cân bằng tối ưu giữa độ phức tạp thuật toán và tốc độ xử lý thực tế.

Mặc dù việc định lượng trực tiếp giá trị kinh tế của các mô hình dự báo tại các nhà máy điện mặt trời ở Việt Nam hiện nay gặp nhiều hạn chế đáng kể, chủ yếu do các ràng buộc nghiêm ngặt về bảo mật dữ liệu vận hành, chi phí huy động nguồn và cơ chế thị trường điện, giá trị ứng dụng của mô hình SE-XGB-LGBM-RF-OW vẫn có thể được khẳng định gián tiếp thông qua việc đối chiếu với các tiêu chuẩn kinh tế năng lượng quốc tế. Cụ thể, việc duy trì độ chính xác vượt trội ($NMAPE < 2\%$) được xem là yếu tố then chốt trong việc tối ưu hóa biểu đồ huy động nguồn và đảm bảo an ninh hệ thống. Dựa trên các thực nghiệm kinh tế năng lượng mới nhất việc đạt được độ chính xác cao ở mức này có khả năng giúp hệ thống giảm thiểu tới $\sim 5\%$ chi phí dự phòng công suất và hỗ trợ các nhà máy cắt giảm đến $\sim 82\%$ chi phí xử phạt do sai số trong lập lịch ngày trước [113]. Đồng thời, các mô hình tối ưu này còn đóng góp trực tiếp vào việc giảm khoảng 4% chỉ số chi phí điện năng quy dẫn (LCOE) [114], góp phần nâng cao hiệu quả tài chính cho toàn hệ thống. Những kết quả định lượng này tạo nền tảng khoa học vững chắc cho các nghiên cứu tiếp theo về kinh tế năng lượng chuyên sâu, dựa trên hệ thống dự báo đã được kiểm chứng trong khuôn khổ luận án.

4.3.2. Đánh giá chi phí tính toán của mô hình Selector-Model

Kịch bản này đánh giá hiệu năng của bộ chọn Selector-Model khi vận hành trên tập dữ liệu quy mô lớn của 3 nhà máy điện mặt trời (Quảng trị, Đắk-Lắk, Thanh Hóa) nhằm kiểm chứng khả năng mở rộng của hệ thống. Tập huấn luyện và dự báo được sử dụng như phần 4.2.1. Phần đánh giá chi phí tính toán của mô hình Selector-Model được thể hiện trong Bảng 4.13 sau đây:

Bảng 4.13. Chi phí tính toán của mô hình Selector-Model

Mô hình Selector-Model		
Thành phần hệ thống	Thời gian huấn luyện (s)	Thời gian dự báo (s)

Mô hình dự báo XGBoost	19,09	0,093
Mô hình dự báo LightGBM	1,59	0,102
Mô hình dự báo Random Forest	450,81	2,103
Bộ phân loại Random Forest Classifier	20,729	0,621
Tổng thời gian	492,23	2,55
Số mẫu huấn luyện	294336 mẫu	
Số lượng mẫu dự báo	21024 mẫu	

Từ bảng Bảng 4.13 có thể thấy mô hình Selector-Model được huấn luyện trên tập dữ liệu lớn với 294336 mẫu, dẫn đến tổng thời gian huấn luyện đạt 492,23 giây (khoảng 8,2 phút). Trong đó, thành phần tốn kém tài nguyên nhất là Random Forest (chiếm hơn 91% thời gian) do đặc thù xây dựng số lượng lớn các cây quyết định có độ sâu cao để bao quát kịch bản thời tiết của 3 khu vực khác nhau. Mô hình Selector-Model có ở tốc độ dự báo rất nhanh, tổng thời gian xử lý cho 21024 mẫu chỉ mất 2,55 giây, bao gồm cả thời gian chạy song song 3 mô hình dự báo và thời gian ra quyết định của bộ chọn (0,621 giây). Tốc độ đáp ứng trung bình đạt mức 0,12 ms/mẫu. Kết quả này khẳng định hệ thống có khả năng xử lý lượng dữ liệu khổng lồ trong thời gian cực ngắn, hoàn toàn đáp ứng các tiêu chuẩn khắt khe về thời gian thực của hệ thống điều độ SCADA/EMS hiện đại.

Từ bảng Bảng 4.12 và Bảng 4.13 có thể thấy rõ kết quả thực nghiệm về thời gian thực thi cho thấy sự khác biệt rõ rệt về hiệu suất tính toán giữa hai cấu trúc mô hình đề xuất. Mặc dù các trọng số tối ưu (w_i) trong mô hình SE-XGB-LGBM-RF-OW đã được xác định sẵn từ quá trình huấn luyện và không tiêu tốn thời gian tính toán lại trong quá trình dự báo, nhưng thời gian thực thi của mô hình này vẫn cao hơn đáng kể so với Selector-Model. Nguyên nhân chính nằm ở sự khác biệt trong lộ trình tính toán, mô hình SE-XGB-LGBM-RF-OW bắt buộc phải kích hoạt đồng thời toàn bộ các mô hình thành phần (XGBoost, LightGBM và Random Forest), sau đó thực hiện phép toán tổ hợp với trọng số tối ưu để đưa ra kết quả dự báo. Quá trình này tạo ra độ trễ hệ thống do phải quản lý dữ liệu đầu ra từ nhiều luồng mô hình trước khi thực hiện bước tổng hợp kết quả. Ngược lại, Selector-Model thể hiện tốc độ vượt trội nhờ cơ chế điều hướng thông

minh, chỉ kích hoạt duy nhất một mô hình con phù hợp nhất cho mỗi mẫu dữ liệu, giúp tiết giảm tối đa khối lượng phép tính và tài nguyên CPU. Tuy nhiên, cần nhấn mạnh rằng ngay cả khi phải mất thêm thời gian để tổng hợp kết quả, thời gian dự báo của mô hình SE-XGB-LGBM-RF-OW vẫn chỉ ở mức dưới 23 giây, thấp hơn nhiều lần so với các mô hình học sâu như LSTM hay GRU (hơn 90 giây, các mô hình này vốn đòi hỏi các phép toán hồi tiếp tuần tự phức tạp). Điều này khẳng định cả hai giải pháp đề xuất đều đáp ứng hoàn hảo yêu cầu vận hành thời gian thực tại các nhà máy điện mặt trời Việt Nam, ngay cả trên các hệ thống máy tính công nghiệp có cấu hình phổ thông.

4.4. Kết luận chương 4

Chương 4 đã xây dựng và đánh giá toàn diện các mô hình dự báo công suất phát điện mặt trời dựa trên dữ liệu thực tế thu thập từ nhiều nhà máy tại Việt Nam. Trong đó, trọng tâm là khai thác hiệu quả mô hình tổ hợp Stacking Ensemble (SE) để thích ứng với đặc điểm dữ liệu thực tế: phi tuyến tính cao, biến động mạnh theo thời tiết, và đặc biệt là thiếu hụt hoặc không đầy đủ chuỗi dữ liệu lịch sử một vấn đề phổ biến tại các nhà máy điện mặt trời tại Việt Nam hiện nay. Các vấn đề đã được giải quyết trong chương này như sau:

- **Vấn đề thứ nhất:** chương này đã xây dựng các tổ hợp mô hình SE khác nhau gồm: SE-XGB-LGBM-RF, SE-XGB-LGBM-BiGRU, SE-XGB-LGBM-GRU và SE-XGB-LGBM-LSTM nhằm kiểm chứng vai trò của các kiến trúc học sâu trong môi trường dữ liệu chuỗi thời gian không hoàn chỉnh. Kết quả cho thấy: mặc dù các mô hình kết hợp với mạng nơ-ron hồi tiếp có khả năng học đặc trưng chuỗi, song hiệu suất không vượt trội, trong khi thời gian huấn luyện dài và độ ổn định kém hơn. Trong khi đó, mô hình SE-XGB-LGBM-RF cho kết quả tối ưu cả về độ chính xác lẫn tốc độ, nhờ khả năng hoạt động tốt trong điều kiện dữ liệu rời rạc, phi tuyến và không liên tục. Đây là ưu điểm nổi bật trong bối cảnh thực tế tại Việt Nam.
- **Vấn đề thứ hai:** mô hình SE-XGB-LGBM-RF được tiếp tục tối ưu hóa theo mùa thông qua việc tính toán trọng số tối ưu cho từng mô hình con. Kết quả cho thấy sau tối ưu, sai số NMAPE của mô hình luôn dưới 2% ở mọi mùa trong năm, đáp ứng tốt yêu cầu vận hành thực tế của hệ thống điện mặt trời.
- **Vấn đề thứ ba:** chương 4 đã mở rộng mô hình sang nhiều nhà máy điện mặt trời đại diện cho các vùng khí hậu khác nhau (Bắc Miền Trung, Tây Nguyên), với hai phương án triển khai:

Kịch bản 1 - Selector-Model: hợp nhất dữ liệu từ ba nhà máy để huấn luyện một mô hình SE duy nhất có khả năng tự lựa chọn cấu hình phù hợp theo điều kiện đầu vào. Mô hình này phù hợp với hệ thống có yêu cầu mở rộng quy mô và tối ưu chi phí vận hành.

Kịch bản 2 - Huấn luyện riêng biệt: xây dựng mô hình riêng cho từng nhà máy, với các đặc trưng đầu vào được tùy biến theo vùng khí hậu. Kết quả cho thấy mô hình riêng có sai số thấp hơn, đặc biệt ở Quảng Trị (Bắc Trung Bộ) nơi dữ liệu thời tiết và công suất có biến động lớn và độ thiếu hụt cao.

- **Vấn đề thứ tư:** Đánh giá về chi phí tính toán và khả năng đáp ứng thời gian thực. Kết quả thực nghiệm đã khẳng định tính khả thi vượt trội của các mô hình đề xuất khi vận hành trên hạ tầng phần cứng thông dụng. Mô hình SE-XGB-LGBM-RF-OW cho thấy sự cân bằng lý tưởng với thời gian huấn luyện (~19 giây) và dự báo (~22 giây) ngắn, phù hợp cho việc triển khai độc lập tại từng nhà máy. Trong khi đó, cấu trúc Selector-Model thể hiện ưu thế đặc biệt về tốc độ phản hồi với thời gian dự báo trung bình chỉ 0,12 ms/mẫu. Mặc dù Selector-Model đòi hỏi thời gian huấn luyện ban đầu lớn hơn (~492 giây) do quy mô dữ liệu gộp từ nhiều nhà máy, nhưng tốc độ dự báo cực nhanh cho phép hệ thống xử lý đồng thời lượng lớn dữ liệu trong các hệ thống điều độ tập trung mà không gây trễ mạng. Đặc biệt, việc cả hai mô hình đều vận hành mượt mà trên CPU mà không cần các thiết bị tính toán chuyên dụng đắt tiền (GPU) là một lợi thế quan trọng trong việc ứng dụng thực tế tại Việt Nam.

+) Đánh giá tổng hợp cho thấy:

Nếu ưu tiên độ chính xác, đặc biệt ở các vùng có khí hậu khắc nghiệt và dữ liệu thiếu hụt, mô hình SE-XGB-LGBM-RF huấn luyện riêng biệt cho từng nhà máy là lựa chọn tốt nhất.

Nếu ưu tiên khả năng tổng quát hóa và tiết kiệm chi phí triển khai, Selector-Model là phương án hợp lý và khả thi.

Trong các hệ thống lớn, có thể khởi tạo bằng Selector-Model, sau đó tinh chỉnh từng nhà máy theo nhu cầu cụ thể.

Đặc biệt, thông qua việc đối chiếu với các nghiên cứu tiêu biểu được công bố tại Việt Nam (trình bày chi tiết trong Phụ lục A5), do tồn tại những hạn chế trong việc so sánh định lượng trực tiếp xuất phát từ sự khác biệt về tập dữ liệu sử dụng cũng như điều kiện khí hậu đặc thù của từng nghiên cứu, NCS tập trung phân tích và so sánh trên phương diện phương pháp luận. Kết quả cho thấy, trong khi phần lớn các nghiên cứu

trước đây, chủ yếu dựa trên các mô hình học sâu phổ biến như LSTM, GRU, thường chỉ xem xét trường hợp tập dữ liệu đầu vào đầy đủ hoặc chỉ thực hiện dự báo cho một nhà máy đơn lẻ, các giải pháp SE-XGB-LGBM-RF-OW và Selector-Model được đề xuất trong luận án có khả năng xử lý hiệu quả các tập dữ liệu bị thiếu và đồng thời thực hiện dự báo cho nhiều nhà máy. Các mô hình này vẫn duy trì được hiệu suất cao và độ ổn định tốt nhờ cấu trúc tổ hợp linh hoạt dựa trên các cây quyết định. Bên cạnh đó, thời gian dự báo của hệ thống chỉ ở mức ms/mẫu, vượt trội so với các mô hình học sâu có cấu trúc phức tạp, qua đó khẳng định tính khả thi và mức độ sẵn sàng cho việc triển khai thực tế trên các nền tảng phần cứng phổ thông tại nhà máy.

Tóm lại, chương này đã chứng minh rằng mô hình SE-XGB-LGBM-RF-OW, nhờ cấu trúc linh hoạt và khả năng xử lý tốt dữ liệu không liên tục, là một giải pháp phù hợp để ứng dụng trong bối cảnh thiếu chuỗi dữ liệu quá khứ vốn là đặc điểm thường gặp trong vận hành thực tế ở nhiều nhà máy điện mặt trời tại Việt Nam. Mô hình vừa đạt độ chính xác cao, vừa đảm bảo khả năng mở rộng và thích nghi theo vùng khí hậu, tạo tiền đề vững chắc cho việc áp dụng vào các hệ thống vận hành thực tế trên phạm vi rộng. Mô hình Selector-Model phù hợp với những nhà máy đặt tại nơi có điều kiện khí hậu tương đồng, nơi mà việc huấn luyện trên tập dữ liệu gộp không gây suy giảm đáng kể về hiệu suất. Với lợi thế về chi phí tính toán thấp và khả năng triển khai đồng loạt cho nhiều nhà máy, mô hình này đặc biệt thích hợp trong các hệ thống quản lý tập trung, nơi yêu cầu độ chính xác ở mức ổn định và tính linh hoạt trong vận hành theo cụm vùng miền.

KẾT LUẬN VÀ KIẾN NGHỊ

1. KẾT LUẬN

Luận án “Nghiên cứu mô hình học máy hỗn hợp dự báo công suất phát điện mặt trời trong điều kiện dữ liệu thực tế tại Việt Nam” đã hoàn thành toàn diện các mục tiêu đề ra, từ việc xây dựng cơ sở dữ liệu thực tế, phát triển mô hình dự báo thích ứng, đến việc đánh giá thực nghiệm trong các điều kiện đặc thù tại Việt Nam. Những đóng góp chính của luận án được tổng kết như sau:

a. Phân tích bối cảnh và xác lập vấn đề nghiên cứu

Luận án đã làm rõ đặc điểm phát triển điện mặt trời tại Việt Nam: tiềm năng lớn, tốc độ cao, nhưng tồn tại nhiều thách thức trong điều độ do công suất dao động và dữ liệu thiếu hụt. Qua tổng quan, luận án chỉ ra rằng các mô hình hiện tại thường chưa phù hợp với điều kiện dữ liệu gián đoạn. Từ đó, định hướng xây dựng mô hình tổ hợp có khả năng thích ứng và mở rộng.

b. Xây dựng và xử lý bộ dữ liệu thực nghiệm

Luận án đã thu thập, xử lý và xây dựng tập dữ liệu công suất thực đo tại nhiều nhà máy (Quảng Trị, Thanh Hóa, Đắk Lắk) và đánh giá vai trò của dữ liệu bức xạ đầu vào (thực đo và dự báo). Việc tiền xử lý dữ liệu thiếu và chuẩn hóa được thực hiện bài bản nhằm đảm bảo tính toàn vẹn đầu vào cho mô hình.

c. Đánh giá mô hình đơn và đề xuất hướng kết hợp

Các mô hình học sâu theo chuỗi thời gian (LSTM, GRU, BiGRU) và mô hình cây quyết định (LightGBM, XGBoost, RF) được xây dựng và đánh giá trong điều kiện dữ liệu thực tế. Kết quả cho thấy nhóm cây quyết định có ưu thế rõ rệt về độ ổn định và hiệu suất khi dữ liệu bị thiếu. Một mô hình kết hợp LightGBM-LSTM cũng được thử nghiệm để xử lý chuỗi bị khuyết, tuy đạt kết quả khá hơn LSTM đơn thuần nhưng vẫn chưa vượt qua các mô hình cây quyết định. Do đó, giải pháp này chỉ phù hợp với trường hợp thiếu dữ liệu nhẹ.

d. Phát triển mô hình tổ hợp thích ứng

Luận án đã xây dựng và đánh giá bốn mô hình tổ hợp dạng Stacking Ensemble. Mô hình SE-XGB-LGBM-RF cho kết quả tối ưu, và sau khi tích hợp cơ chế tối ưu trọng số theo mùa (OW), đã đạt sai số NMAPE dưới 2%, đáp ứng yêu cầu kỹ thuật thực tế. Bên cạnh đó, luận án phát triển mô hình chọn lọc tự động (Selector-Model) để lựa chọn

mô hình phù hợp với từng nhà máy dựa trên đặc trưng khí hậu và dữ liệu, góp phần mở rộng khả năng ứng dụng thực tiễn.

e. Các đóng góp mới

Đề xuất mô hình tổ hợp SE-XGB-LGBM-RF-OW phục vụ dự báo công suất phát điện mặt trời dựa trên đặc tính thời tiết của từng mùa trong năm. Kiến trúc này là sự kết hợp tối ưu giữa các thuật toán cây quyết định thông qua phương pháp Stacking Ensemble, tích hợp cơ chế tối ưu hóa trọng số theo mùa nhằm tăng cường độ chính xác và khả năng thích ứng của hệ thống dự báo trước những biến động phức tạp của thời tiết theo vùng miền và thời gian.

Thực hiện đánh giá có hệ thống hai nhóm mô hình dự báo là mô hình cây quyết định (XGBoost, LightGBM, Random Forest) và mô hình mạng nơ-ron học sâu chuỗi thời gian (LSTM, GRU, BiGRU) trong bối cảnh dữ liệu bị thiếu liên tục. Các phân tích so sánh chi tiết về độ chính xác, tốc độ huấn luyện và tính ổn định không chỉ làm rõ đặc tính kỹ thuật của từng nhóm mà còn cung cấp cơ sở khoa học để lựa chọn mô hình dự báo phù hợp cho các ứng dụng thực tế sau này.

Xây dựng mô hình kết hợp LightGBM-LSTM, trong đó LightGBM được sử dụng để lấp đầy dữ liệu công suất bị thiếu trong ngắn hạn, đóng vai trò trung gian đầu vào cho LSTM giúp cải thiện hiệu quả dự báo trong điều kiện dữ liệu không đầy đủ.

Thiết kế mô hình Selector-Model, có khả năng tự động lựa chọn mô hình dự báo tối ưu cho từng nhà máy tại từng thời điểm, dựa trên đặc trưng khí hậu, quy mô và lịch sử dữ liệu. Đây là bước tiến quan trọng hướng đến triển khai dự báo đồng thời trên nhiều nhà máy điện mặt trời tại các vùng khí hậu khác nhau.

Xác nhận hiệu quả của các mô hình dự báo bức xạ (LightGBM, LSTM, GRU) như một lớp trung gian hỗ trợ trong hệ thống dự báo công suất, đặc biệt hữu ích trong điều kiện thiếu dữ liệu đo đạc thực tế hoặc dữ liệu dự báo từ các trung tâm dự báo khí tượng chuyên trách. Kết quả góp phần làm rõ vai trò của lớp dữ liệu bức xạ trong toàn bộ chuỗi dự báo.

2. KIẾN NGHỊ

Đề phát huy hơn nữa giá trị của nghiên cứu và mở rộng ứng dụng của mô hình dự báo công suất nguồn điện mặt trời trong thực tế, nghiên cứu sinh đề xuất một số kiến nghị như sau:

a. Ứng dụng thực tiễn

- Các cơ quan quản lý vận hành hệ thống điện và các nhà máy điện mặt trời nên xem xét triển khai mô hình SE-XGB-LGBM-RF-OW vào thực tế, đặc biệt tại những khu vực có tỷ trọng năng lượng mặt trời cao, thời tiết biến động mạnh hoặc thiếu dữ liệu lịch sử.
- Trong các cụm nhà máy điện mặt trời, có thể áp dụng kịch bản mô hình gộp Selector-Model để triển khai ban đầu, sau đó tinh chỉnh mô hình riêng biệt cho những nhà máy có yêu cầu dự báo chính xác hơn hoặc đặc điểm khí hậu riêng biệt.

b. Về dữ liệu và hệ thống hỗ trợ

- Cần tăng cường đầu tư vào hệ thống thu thập dữ liệu tại chỗ và tích hợp dữ liệu vệ tinh, dự báo khí tượng số để làm giàu đầu vào cho mô hình dự báo.
- Nâng cấp hạ tầng tính toán (máy chủ, GPU, phần mềm hỗ trợ) để đáp ứng yêu cầu tính toán khi áp dụng các mô hình học máy và học sâu trên quy mô lớn.

c. Về hướng nghiên cứu tiếp theo

- Mở rộng mô hình dự báo cho các nguồn năng lượng tái tạo khác như điện gió, thủy điện nhỏ và nghiên cứu mô hình tổng hợp cho hệ thống hỗn hợp (multi-source forecasting).
- Tiếp tục cải tiến mô hình theo hướng tự động hóa quá trình lựa chọn mô hình con và cập nhật trọng số theo điều kiện thời tiết biến động (adaptive ensemble).
- Nghiên cứu tích hợp các mô hình học sâu tiên tiến (như Transformer, attention-based hoặc AutoML) vào khung Selector-Model và Stacking Ensemble theo kiến trúc mô-đun, nhằm nâng cao khả năng tự thích nghi, mở rộng và tính bền vững của hệ thống.

DANH MỤC CÁC CÔNG TRÌNH ĐÃ CÔNG BỐ

Hội thảo quốc tế (Scopus): 03 công bố

- [1] Tuan-Anh Nguyen, Manh-Hai Pham, Vu Minh Phap, Quang-hiep Do, Ngoc-Trung Nguyen, Dang-Toan Nguyen, Thi Nga Nguyen., “Forecasting of solar power generation in Vietnam deploying a simple GRU model,” Conf. Proc. - 2023 IEEE Asia Meet. Environ. Electr. Eng. EEE-AM 2023, pp. 1–6, 2023, doi: 10.1109/EEE-AM58328.2023.10395560.
- [2] N. Tuan Anh, P. Manh-Hai, V. Minh Phap, and 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.
- [3] Ngoc-Trung Nguyen, Phap Vu Minh, Thi Anh Tho Vu, Tuan-Anh Nguyen, Manh-Hai Pham, Trung Nguyen Thanh, Dang Toan Nguyen, Thanh-Doanh Le, Son Tran-Thanh, “An Experimental Comparison of TCN, LSTM, and GRU for Short-term Solar Power Forecasting,” IEEE-2025 12th International Conference on Power and Energy Systems Engineering (CPESE 2025). (đã được Acceptance Full PAPER).

Bài báo quốc tế (Scopus (Q2)): 01 công bố

- [1] Tuan Anh Nguyen, Manh Hai Pham, Minh Phap Vu, Ngoc Trung Nguyen, Dang Toan Nguyen, Thi Anh Tho Vu, Trong Tuan Tran, Anh Tuan Do, "A Study on Novel Solar Power Forecasting using XGB-LGBM-RF Hybrid Model and L-BFGS-B Optimization Algorithm", Engineering, Technology & Applied Science Research, vol. 15, no. 4, pp. 24516–24522, 2025.

Tạp chí trong nước: 03 công bố

- [1] Nguyễn Tuấn Anh, Phạm Mạnh hải, Vũ Minh Pháp, Nguyễn Đăng Toàn, Vũ Thị Anh Thơ, “Dự báo công suất nguồn phát điện mặt trời bằng mô hình BiGRU,” Tạp chí Khoa học và Công nghệ Năng lượng - Trường Đại học Điện lực, vol. 34, pp. 1–14, 2024.
- [2] Nguyễn Tuấn Anh, Phạm Mạnh hải, Nguyễn Đăng Toàn, Lê Thành Doanh, Nguyễn Ngọc Trung, Vũ Thị Anh Thơ, Vũ Minh Pháp, Trần Trọng Tuấn, “Đánh giá các mô hình dự báo bức xạ mặt trời: LightGBM, LSTM và GRU,” J. Sci. Technol. - HaUI, vol. 60, no. 8, pp. 3–10, Aug. 2024, doi: 10.57001/huih5804.2024.256.
- [3] Manh-Hai Pham, Nguyen Tuan Anh, Vu Minh Phap, Pham Văn Duy, Le Thanh Doanh, Vu Thi Anh Tho, Nguyen Dang Toan, “A Comparative Study on the Operational Effectiveness of Machine Learning Models in Solar Power Forecasting” ,” JST: Smart Systems and Devices *Volume 35, Issue 2, May 2025, 054-061.*

DANH MỤC CÁC TÀI LIỆU THAM KHẢO

- [1] Z. Guo, K. Zhou, C. Zhang, X. Lu, W. Chen, and S. Yang, “Residential electricity consumption behavior: Influencing factors, related theories and intervention strategies,” *Renew. Sustain. Energy Rev.*, vol. 81, no. January 2017, pp. 399–412, 2018, doi: 10.1016/j.rser.2017.07.046.
- [2] E. R. Sanseverino *et al.*, “Review of Potential and Actual Penetration of Solar Power in Vietnam,” *Energies*, vol. 13, no. 10, pp. 7–9, May 2020, doi: 10.3390/en13102529.
- [3] Bộ tài nguyên và môi trường, “Đánh giá tiềm năng năng lượng bức xạ, gió và sóng tại Việt Nam,” *báo cáo*, 2022.
- [4] ICT VietNam, “Phát triển điện mặt trời tại Việt Nam.” .
- [5] Linh Duy Bui and Quynh Pham, “Vietnam, a System in Rapid Transition - ESIG Spring Technical Workshop 2023,” 2023, [Online]. Available: <https://www.esig.energy/event/2023-spring-technical-workshop/>.
- [6] Hoàng Thị Thu Hương, “Giải pháp phát triển điện mặt trời tại Việt Nam,” *Kinh tế và dự báo*, pp. 38–41, 2021.
- [7] Đặng Hoàng Hợp, Nguyễn Đăng Cường, Nguyễn Xuân Thành, Nguyễn Văn Thành và Lê Thị Thu Hiền, “Phát triển điện mặt trời tại Việt nam: Hiện trạng và những rào cản,” *Tạp chí Khoa học Công nghệ Việt Nam*, vol. 3, pp. 23–24, 2022, doi: 10.3390/en13102529.
- [8] Chính phủ, “768/QĐ-TTg Phê duyệt Điều chỉnh Quy hoạch phát triển điện lực quốc gia thời kỳ 2021-2030, tầm nhìn đến năm 2050,” 2025.
- [9] Bộ Công Thương, “67 QĐ/quy trình dự báo công suất, điện năng phát của các nguồn điện năng lượng tái tạo,” pp. 1–22, 2021.
- [10] EVNNLDC, “4608/ĐĐQG-NLTT, yêu cầu độ chính xác đối với dự báo công suất phát các NMĐ NLTT năm 2024,” pp. 1–10, 2023.
- [11] Nguyễn Duy Minh và Nguyễn Bá Tiến, “Hiện trạng dự báo công suất phát nhà máy điện gió và mặt trời tại Việt Nam.pdf.” p. Tạp chí Công thương.
- [12] “Trung tâm Điều độ hệ thống điện quốc gia - Tập đoàn Điện lực Việt Nam (2023). Công văn số 1957/ĐĐQG-NLTT ngày 07/06/2023 về Sử dụng giá trị công suất dự báo tổng hợp trong công tác vận hành các nguồn năng lượng tái tạo.,” 2023.
- [13] “Trung tâm Điều độ hệ thống điện quốc gia - Tập đoàn Điện lực Việt Nam (2021). Công văn số 404/ ĐĐQG-NLTT ngày 02/02/2021 về Nguyên tắc sử dụng nguồn dữ liệu phục vụ tính toán phân bổ công suất các nguồn NLTT,” pp. 1–9, 2021.
- [14] “Trung tâm Điều độ hệ thống điện quốc gia - Tập đoàn Điện lực Việt Nam (2021). Công văn số 2270/ ĐĐQG-NLTT ngày 16/06/2021 về Cập nhật công suất dự báo chu kỳ tới trong tính toán giá trị công suất dự báo được lựa chọn.,” vol. 3, pp. 1–8, 2021.
- [15] R. Ahmed, V. Sreeram, Y. Mishra, and M. D. Arif, “A review and evaluation of the state-of-the-art in PV solar power forecasting: Techniques and optimization,” *Renew. Sustain. Energy Rev.*, vol. 124, no. February, p. 109792, 2020, doi: 10.1016/j.rser.2020.109792.
- [16] Y. Ren, P. N. Suganthan, and N. Srikanth, “Ensemble methods for wind and solar power forecasting - A state-of-the-art review,” *Renew. Sustain. Energy Rev.*, vol.

- 50, pp. 82–91, 2015, doi: 10.1016/j.rser.2015.04.081.
- [17] U. K. Das *et al.*, “Forecasting of photovoltaic power generation and model optimization: A review,” *Renew. Sustain. Energy Rev.*, vol. 81, pp. 912–928, Jan. 2018, doi: 10.1016/j.rser.2017.08.017.
 - [18] M. Diagne, M. David, P. Lauret, J. Boland, and N. Schmutz, “Review of solar irradiance forecasting methods and a proposition for small-scale insular grids,” *Renew. Sustain. Energy Rev.*, vol. 27, pp. 65–76, 2013, doi: 10.1016/j.rser.2013.06.042.
 - [19] M. K. Behera, I. Majumder, and N. Nayak, “Solar photovoltaic power forecasting using optimized modified extreme learning machine technique,” *Eng. Sci. Technol. an Int. J.*, vol. 21, no. 3, pp. 428–438, 2018, doi: 10.1016/j.jestch.2018.04.013.
 - [20] M. Meng and C. Song, “Daily photovoltaic power generation forecasting model based on random forest algorithm for north china in winter,” *Sustain.*, vol. 12, no. 6, 2020, doi: 10.3390/su12062247.
 - [21] Q. T. Phan, Y. K. Wu, Q. D. Phan, and 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.*, vol. 59, no. 1, pp. 220–231, 2023, doi: 10.1109/TIA.2022.3212999.
 - [22] 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,” *2019 Int. Multi-Conference Eng. Comput. Inf. Sci.*, pp. 1–10, 2019.
 - [23] E. Nziyumva, R. Hu, Y. Xu, J. Niyogisubizo, Z. Li, and 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.
 - [24] N. Quang, L. Duy, B. Van, and Q. Dinh, “Applying Artificial Intelligence in Forecasting the Output of Industrial Solar Power Plant in Vietnam,” *EAI Endorsed Trans. Energy Web*, p. 169166, 2018, doi: 10.4108/eai.29-3-2021.169166.
 - [25] T. A. Nguyen *et al.*, “Forecasting of solar power generation in Vietnam deploying a simple GRU model,” *Conf. Proc. - 2023 IEEE Asia Meet. Environ. Electr. Eng. EEE-AM 2023*, pp. 1–6, 2023, doi: 10.1109/EEE-AM58328.2023.10395560.
 - [26] M. Chai, F. Xia, S. Hao, D. Peng, C. Cui, and W. Liu, “PV Power Prediction Based on LSTM with Adaptive Hyperparameter Adjustment,” *IEEE Access*, vol. 7, pp. 115473–115486, 2019, doi: 10.1109/ACCESS.2019.2936597.
 - [27] V. Suresh, P. Janik, J. Rezmer, and Z. Leonowicz, “Forecasting solar PV output using convolutional neural networks with a sliding window algorithm,” *Energies*, vol. 13, no. 3, 2020, doi: 10.3390/en13030723.
 - [28] M. Sabri and M. El Hassouni, “A Novel Deep Learning Approach for Short Term Photovoltaic Power Forecasting Based on GRU-CNN Model,” *E3S Web Conf.*, vol. 336, 2022, doi: 10.1051/e3sconf/202233600064.
 - [29] X. G. Pengyun Jia, HAIBO ZHANG, XINMIAO LIU, “Short-term photovoltaic power forecasting based on MIE-LSTM,” *Dianli Xitong Baohu yu Kongzhi/Power Syst. Prot. Control*, vol. 48, no. 7, pp. 50–57, 2020, doi: 10.19783/j.cnki.pspc.190611.
 - [30] N. T. H. Thu, P. Q. Bao, and N. V. N. Nam, “Multiple Step Ahead Forecasting of Rooftop Solar Power Based on a Novel Hybrid Model of CEEMDAN -

- Bidirectional LSTM Network with Structure Optimized by PSO Method,” *2022 11th Int. Conf. Control. Autom. Inf. Sci. ICCAIS 2022*, pp. 522–528, 2022, doi: 10.1109/ICCAIS56082.2022.9990323.
- [31] J. Moon, “A Multi-Step-Ahead Photovoltaic Power Forecasting Approach Using One-Dimensional Convolutional Neural Networks and Transformer,” *Electron.*, vol. 13, no. 11, 2024, doi: 10.3390/electronics13112007.
 - [32] F. M. A. Mazen, Y. Shaker, and R. A. Abul Seoud, “Forecasting of Solar Power Using GRU–Temporal Fusion Transformer Model and DILATE Loss Function,” *Energies*, vol. 16, no. 24, p. 8105, Dec. 2023, doi: 10.3390/en16248105.
 - [33] W. Khan, S. Walker, and W. Zeiler, “Improved solar photovoltaic energy generation forecast using deep learning-based ensemble stacking approach,” *Energy*, vol. 240, no. December 2021, p. 122812, 2022, doi: 10.1016/j.energy.2021.122812.
 - [34] L. J. Wang Zhen, “Short-Term Photovoltaic Power Generation Prediction Based on LightGBM-LSTM Model,” *Mod. Electr. Power*, vol. 40, no. 6, pp. 899–905, 2023, doi: 10.19725/j.cnki.1007-2322.2022.0111.
 - [35] J. Ye, B. Zhao, and H. Deng, “Photovoltaic Power Prediction Model Using Pre-train and Fine-tune Paradigm Based on LightGBM and XGBoost,” *Procedia Comput. Sci.*, vol. 224, pp. 407–412, 2023, doi: 10.1016/j.procs.2023.09.056.
 - [36] S. B. Bashir, M. M. Farag, A. K. Hamid, A. A. Adam, A. G. Abo-Khalil, and R. Bansal, “A Novel Hybrid CNN-XGBoost Model for Photovoltaic System Power Forecasting,” *Proc. 2024 6th Int. Youth Conf. Radio Electron. Electr. Power Eng. REEPE 2024*, no. February, 2024, doi: 10.1109/REEPE60449.2024.10479878.
 - [37] Q. T. Phan, Y. K. Wu, and Q. D. Phan, “Short-term Solar Power Forecasting Using XGBoost with Numerical Weather Prediction,” *2021 IEEE Int. Futur. Energy Electron. Conf. IFEEEC 2021*, no. May 2022, pp. 1–2, 2021, doi: 10.1109/IFEEEC53238.2021.9661874.
 - [38] L. D. Bui, N. Q. Nguyen, B. Van Doan, and E. R. Sanseverino, “Forecasting energy output of a solar power plant in curtailment condition based on LSTM using P/GHI coefficient and validation in training process, a case study in Vietnam,” *Electr. Power Syst. Res.*, vol. 213, 2022, doi: 10.1016/j.epsr.2022.108706.
 - [39] H. Sangrody, N. Zhou, and Z. Zhang, “Similarity-Based Models for Day-Ahead Solar PV Generation Forecasting,” *IEEE Access*, vol. 8, pp. 104469–104478, 2020, doi: 10.1109/ACCESS.2020.2999903.
 - [40] N. N. V. Nhat, D. N. Huu, and T. T. H. Nguyen, “Short-term multi-step forecasting of rooftop solar power generation using a combined data decomposition and deep learning model of EEMD-GRU,” *J. Renew. Sustain. Energy*, vol. 16, no. 1, 2024, doi: 10.1063/5.0176951.
 - [41] Z. Wang *et al.*, “Investigating the all-sky surface solar radiation and its influencing factors in the Yangtze River Basin in recent four decades,” *Atmos. Environ.*, vol. 244, no. August 2020, p. 117888, 2021, doi: 10.1016/j.atmosenv.2020.117888.
 - [42] S. Liang and J. Wang, *Chapter 5 - Solar radiation*. 2020.
 - [43] S. Barbaro, G. Cannistraro, C. Giaconia, and A. Orioli, “The ASHRAE clear sky model. An evaluation in the Mediterranean Zone,” *Sol. Wind Technol.*, vol. 5, no. 1, pp. 111–116, 1988, doi: 10.1016/0741-983X(88)90097-5.
 - [44] V. Parikh, C. Desai, D. Joshi, and G. Nagababu, “Estimation of electricity

- generation potential by solar radiation on Sardar Sarovar dam,” *Energy Procedia*, vol. 158, pp. 167–172, 2019, doi: 10.1016/j.egypro.2019.01.065.
- [45] P. Kosmopoulos *et al.*, “Short-term forecasting of large-scale clouds impact on downwelling surface solar irradiation,” *Energies*, vol. 13, no. 24, 2020, doi: 10.3390/en13246555.
 - [46] B. A. Raut *et al.*, “Optimizing cloud motion estimation on the edge with phase correlation and optical flow,” *Atmos. Meas. Tech.*, vol. 16, no. 5, pp. 1195–1209, 2023, doi: 10.5194/amt-16-1195-2023.
 - [47] D. M. L. H. Dissawa, G. M. R. I. Godaliyadda, M. P. B. Ekanayake, J. B. Ekanayake, and A. P. Agalgaonkar, “Cross-correlation based cloud motion estimation for short-term solar irradiation predictions,” *2017 IEEE Int. Conf. Ind. Inf. Syst. ICIIS 2017 - Proc.*, vol. 2018-Janua, pp. 1–6, 2017, doi: 10.1109/ICIINFS.2017.8300338.
 - [48] D. Aicardi, P. Musé, and R. Alonso-Suárez, “A comparison of satellite cloud motion vectors techniques to forecast intra-day hourly solar global horizontal irradiation,” *Sol. Energy*, vol. 233, no. January, pp. 46–60, 2022, doi: 10.1016/j.solener.2021.12.066.
 - [49] E. Lorenz, J. Hurka, D. Heinemann, and H. G. Beyer, “Irradiance Forecasting for the Power Prediction of Grid-Connected Photovoltaic Systems,” *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 2, no. 1, pp. 2–10, 2009, doi: 10.1109/JSTARS.2009.2020300.
 - [50] J. Kühnert, E. Lorenz, and D. Heinemann, *Satellite-Based Irradiance and Power Forecasting for the German Energy Market*. 2013.
 - [51] E. W. Law, A. A. Prasad, M. Kay, and R. A. Taylor, “Direct normal irradiance forecasting and its application to concentrated solar thermal output forecasting - A review,” *Sol. Energy*, vol. 108, pp. 287–307, 2014, doi: 10.1016/j.solener.2014.07.008.
 - [52] D. B. Preston, “Robust Forecasting,” *Appl. Time Ser. Anal. II*, vol. 1919034, pp. 761–794, 1981, doi: 10.1016/b978-0-12-256420-8.50028-9.
 - [53] S. Ghimire, R. C. Deo, N. J. Downs, and N. Raj, “Global solar radiation prediction by ANN integrated with European Centre for medium range weather forecast fields in solar rich cities of Queensland Australia,” *J. Clean. Prod.*, vol. 216, pp. 288–310, 2019, doi: 10.1016/j.jclepro.2019.01.158.
 - [54] L. Olatomiwa, S. Mekhilef, S. Shamshirband, K. Mohammadi, D. Petković, and C. Sudheer, “A support vector machine-firefly algorithm-based model for global solar radiation prediction,” *Sol. Energy*, vol. 115, pp. 632–644, 2015, doi: 10.1016/j.solener.2015.03.015.
 - [55] C. G. Villegas-Mier, J. Rodriguez-Resendiz, J. M. Álvarez-Alvarado, H. Jiménez-Hernández, and Á. Odry, “Optimized Random Forest for Solar Radiation Prediction Using Sunshine Hours,” *Micromachines*, vol. 13, no. 9, 2022, doi: 10.3390/mi13091406.
 - [56] J. Park, J. Moon, S. Jung, and E. Hwang, “Multistep-ahead solar radiation forecasting scheme based on the light gradient boosting machine: A case study of Jeju Island,” *Remote Sens.*, vol. 12, no. 14, 2020, doi: 10.3390/rs12142271.
 - [57] N. Q. Nguyen, L. D. Bui, B. Van Doan, E. R. Sanseverino, D. Di Cara, and Q. D. Nguyen, “A new method for forecasting energy output of a large-scale solar power plant based on long short-term memory networks a case study in Vietnam,” *Electr. Power Syst. Res.*, vol. 199, no. June, p. 107427, 2021, doi:

- 10.1016/j.epsr.2021.107427.
- [58] A. T. Eseye, J. Zhang, and D. Zheng, "Short-term photovoltaic solar power forecasting using a hybrid Wavelet-PSO-SVM model based on SCADA and Meteorological information," *Renew. Energy*, vol. 118, pp. 357–367, 2018, doi: 10.1016/j.renene.2017.11.011.
 - [59] S. Dutta *et al.*, "Load and Renewable Energy Forecasting for a Microgrid using Persistence Technique," *Energy Procedia*, vol. 143, no. December, pp. 617–622, 2017, doi: 10.1016/j.egypro.2017.12.736.
 - [60] U. Firat, S. N. Engin, M. Saraclar, and A. B. Ertuzun, "Wind speed forecasting based on second order blind identification and autoregressive model," *Proc. - 9th Int. Conf. Mach. Learn. Appl. ICMLA 2010*, pp. 686–691, 2010, doi: 10.1109/ICMLA.2010.106.
 - [61] S. Sobri, S. Koohi-Kamali, and N. A. Rahim, "Solar photovoltaic generation forecasting methods: A review," *Energy Convers. Manag.*, vol. 156, pp. 459–497, Jan. 2018, doi: 10.1016/j.enconman.2017.11.019.
 - [62] C. Cornaro, M. Pierro, and F. Bucci, "Master optimization process based on neural networks ensemble for 24-h solar irradiance forecast," *Sol. Energy*, vol. 111, pp. 297–312, 2015, doi: 10.1016/j.solener.2014.10.036.
 - [63] G. Sbrana and A. Silvestrini, "Random switching exponential smoothing and inventory forecasting," *Int. J. Prod. Econ.*, vol. 156, pp. 283–294, 2014, doi: 10.1016/j.ijpe.2014.06.016.
 - [64] L. Ferbar Tratar and E. Strmčnik, "The comparison of Holt-Winters method and Multiple regression method: A case study," *Energy*, vol. 109, pp. 266–276, 2016, doi: 10.1016/j.energy.2016.04.115.
 - [65] D. Tien Bui, K. T. T. Bui, Q. T. Bui, C. Van Doan, and N. D. Hoang, *Hybrid Intelligent Model Based on Least Squares Support Vector Regression and Artificial Bee Colony Optimization for Time-Series Modeling and Forecasting Horizontal Displacement of Hydropower Dam*, 1st ed. Elsevier Inc., 2017.
 - [66] L. Sadowski, "Multi-scale evaluation of the interphase zone between the overlay and concrete substrate: Methods and descriptors," *Appl. Sci.*, vol. 7, no. 9, 2017, doi: 10.3390/app7090893.
 - [67] T. Hegazy, P. Fazio, and O. Moselhi, "Developing Practical Neural Network Applications Using Back-Propagation," *Comput. Civ. Infrastruct. Eng.*, vol. 9, no. 2, pp. 145–159, 1994, doi: 10.1111/j.1467-8667.1994.tb00369.x.
 - [68] D. Tien Bui, V. H. Nhu, and N. D. Hoang, "Prediction of soil compression coefficient for urban housing project using novel integration machine learning approach of swarm intelligence and Multi-layer Perceptron Neural Network," *Adv. Eng. Informatics*, vol. 38, no. April, pp. 593–604, 2018, doi: 10.1016/j.aei.2018.09.005.
 - [69] M. Y. Khan, A. Qayoom, M. S. Nizami, M. S. Siddiqui, S. Wasi, and S. M. K. U. R. Raazi, "Automated Prediction of Good Dictionary EXamples (GDEX): A Comprehensive Experiment with Distant Supervision, Machine Learning, and Word Embedding-Based Deep Learning Techniques," *Complexity*, vol. 2021, 2021, doi: 10.1155/2021/2553199.
 - [70] C. Saigustia and P. Pijarski, "Time Series Analysis and Forecasting of Solar Generation in Spain Using eXtreme Gradient Boosting: A Machine Learning Approach," *Energies*, vol. 16, no. 22, 2023, doi: 10.3390/en16227618.
 - [71] R. Saraswat, D. Jhanwar, and M. Gupta, "Enhanced Solar Power Forecasting

- Using XG Boost and PCA-Based Sky Image Analysis,” *Trait. du Signal*, vol. 41, no. 1, pp. 503–510, 2024, doi: 10.18280/ts.410145.
- [72] Y. Peng, S. Wang, W. Chen, J. Ma, C. Wang, and J. Chen, “LightGBM-Integrated PV Power Prediction Based on Multi-Resolution Similarity,” *Processes*, vol. 11, no. 4, 2023, doi: 10.3390/pr11041141.
 - [73] K. Reba, J. Bevc, A.- Vázquez, and M. Jankovec, “Photovoltaic Energy Production Forecasting using LightGBM,” *55th Int. Conf. Microelectron. Devices Mater. with Work. Laser Syst. Photonics*, pp. 36–39, 2019.
 - [74] J. Wang, P. Li, R. Ran, Y. Che, and Y. Zhou, “A short-term photovoltaic power prediction model based on the Gradient Boost Decision Tree,” *Appl. Sci.*, vol. 8, no. 5, 2018, doi: 10.3390/app8050689.
 - [75] Leo Breiman, “Random Forests,” *IStatistics Dep. Univ. Calif. Berkeley, CA 94720*, pp. 1–33, 2001, doi: 10.1023/A:1010950718922.
 - [76] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System Tianqi,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, vol. 42, no. 8, pp. 785–794, doi: 10.1145/2939672.2939785.
 - [77] D. J. Bae, B. S. Kwon, and K. Bin Song, “XGboost-based day-ahead load forecasting algorithm considering behind-the-meter solar PV generation,” *Energies*, vol. 15, no. 1, 2022, doi: 10.3390/en15010128.
 - [78] G. Ke and I.-Y. Meng, QiFinely, ThomasWang, TaifengChen, WeiMa, WeidongYe, QiweiTLiu, “LightGBM: A Highly Efficient Gradient Boosting Decision Tree,” *31st Conf. Neural Inf. Process. Syst. (NIPS 2017), Long Beach, CA, USA*, no. Nips, 2017.
 - [79] X. Jiang, Y. Pang, M. Sun, and X. Li, “Cascaded Subpatch Networks for Effective CNNs,” *IEEE Trans. Neural Networks Learn. Syst.*, vol. 29, no. 7, pp. 2684–2694, 2018, doi: 10.1109/TNNLS.2017.2689098.
 - [80] F. Wang *et al.*, “Generative adversarial networks and convolutional neural networks based weather classification model for day ahead short-term photovoltaic power forecasting,” *Energy Convers. Manag.*, vol. 181, pp. 443–462, Feb. 2019, doi: 10.1016/j.enconman.2018.11.074.
 - [81] N. Sharma, R. Sharma, and N. Jindal, “Machine Learning and Deep Learning Applications-A Vision ☆,” *Glob. Transitions Proc.*, vol. 2, no. 1, pp. 24–28, 2021, doi: 10.1016/j.gltp.2021.01.004.
 - [82] Y. Sun, V. Venugopal, and A. R. Brandt, “Convolutional Neural Network for Short-term Solar Panel Output Prediction,” *2018 IEEE 7th World Conf. Photovolt. Energy Conversion, WCPEC 2018 - A Jt. Conf. 45th IEEE PVSC, 28th PVSEC 34th EU PVSEC*, pp. 2357–2361, 2018, doi: 10.1109/PVSC.2018.8547400.
 - [83] M. Konstantinou, S. Peratikou, and A. G. Charalambides, “Solar photovoltaic forecasting of power output using lstm networks,” *Atmosphere (Basel)*, vol. 12, no. 1, pp. 1–17, 2021, doi: 10.3390/atmos12010124.
 - [84] H. K. Ahn and N. Park, “Deep RNN-Based Photovoltaic Power Short-Term Forecast Using Power IoT Sensors,” *Energies*, vol. 14, no. 2, p. 436, 2021, doi: 10.3390/en14020436.
 - [85] A. Dairi, F. Harrou, Y. Sun, and S. Khadraoui, “Short-term forecasting of photovoltaic solar power production using variational auto-encoder driven deep learning approach,” *Appl. Sci.*, vol. 10, no. 23, pp. 1–20, 2020, doi: 10.3390/app10238400.

- [86] J. L. Elman, "Finding structure in time," *Cogn. Sci.*, vol. 14, no. 2, pp. 179–211, 1990, doi: 10.1016/0364-0213(90)90002-E.
- [87] R. J. Williams and D. Zipser, "Experimental Analysis of the Real-time Recurrent Learning Algorithm," *Conn. Sci.*, vol. 1, no. 1, pp. 87–111, 1989, doi: 10.1080/09540098908915631.
- [88] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.
- [89] Nguyễn Tuấn Anh, Phạm Mạnh hải, Nguyễn Đăng Toàn, Lê Thành Doanh, Nguyễn Ngọc Trung, Vũ Thị Anh Thơ, Vũ Minh Pháp, Trần Trọng Tuấn, "Đánh giá các mô hình dự báo bức xạ mặt trời: LightGBM, LSTM và GRU," *J. Sci. Technol. - HaUI*, vol. 60, no. 8, pp. 3–10, Aug. 2024, doi: 10.57001/huih5804.2024.256.
- [90] M. Aslam, S. J. Lee, S. H. Khang, and S. Hong, "Two-Stage Attention over LSTM with Bayesian Optimization for Day-Ahead Solar Power Forecasting," *IEEE Access*, vol. 9, pp. 107387–107398, 2021, doi: 10.1109/ACCESS.2021.3100105.
- [91] K. Cho *et al.*, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *EMNLP 2014 - 2014 Conf. Empir. Methods Nat. Lang. Process. Proc. Conf.*, no. June, pp. 1724–1734, 2014, doi: 10.3115/v1/d14-1179.
- [92] R. A. Rajagukguk, R. A. A. Ramadhan, and H.-J. Lee, "A Review on Deep Learning Models for Forecasting Time Series Data of Solar Irradiance and Photovoltaic Power," *Energies*, vol. 13, no. 24, p. 6623, 2020, doi: 10.3390/en13246623.
- [93] L. Kumar and S. Sarkar, *Neural Search: Learning Query and Product Representations in Fashion E-commerce*, vol. 1, no. 1. Association for Computing Machinery, 2021.
- [94] A. Vaswani *et al.*, "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 2017-Decem, no. Nips, pp. 5999–6009, 2017.
- [95] D. Wolpert, "Stacked Generalization," *Neural Networks*, vol. 5, pp. 241–259, 1992.
- [96] L. Breiman, "Stacked regressions," *Mach. Learn.*, vol. 24, no. 1, pp. 49–64, 1996, doi: 10.1023/A:1018046112532.
- [97] T. G. Dietterich and Oregon, "Ensemble methods in machine learning. In: International Workshop on Multiple Classifier Models," *Oncogene*, vol. 12, no. 2, pp. 1–15, 2000.
- [98] A. Abdellatif *et al.*, "Forecasting Photovoltaic Power Generation with a Stacking Ensemble Model," *Sustain.*, vol. 14, no. 17, pp. 1–21, 2022, doi: 10.3390/su141711083.
- [99] P. Kumari and D. Toshniwal, "Extreme gradient boosting and deep neural network based ensemble learning approach to forecast hourly solar irradiance," *J. Clean. Prod.*, vol. 279, p. 123285, 2021, doi: 10.1016/j.jclepro.2020.123285.
- [100] Vincent E. Larson, *Forecasting Solar Irradiance with Numerical Weather Prediction Models*. Elsevier, 2013.
- [101] E. Chodakowska, J. Nazarko, Ł. Nazarko, H. S. Rabayah, R. M. Abendeh, and R. Alawneh, "ARIMA Models in Solar Radiation Forecasting in Different Geographic Locations," *Energies*, vol. 16, no. 13, 2023, doi: 10.3390/en16135029.
- [102] A. Teta, M. Medkour, A. Rabehi, B. Korich, and D. Bakria, "Long-term solar

- radiation forecasting based on LSTM and attention mechanism: a case study in Algeria,” *Stud. Eng. Exact Sci.*, vol. 5, no. 1, pp. 1117–1134, 2024, doi: 10.54021/seesv5n1-058.
- [103] C. N. Obiora, A. Ali, and A. N. Hasan, “Forecasting Hourly Solar Irradiance Using Long Short-Term Memory (LSTM) Network,” *11th Int. Renew. Energy Congr. IREC 2020*, no. Irec, 2020, doi: 10.1109/IREC48820.2020.9310449.
- [104] M. Aslam, K. H. Seung, S. Jae Lee, J. M. Lee, S. Hong, and E. H. Lee, “Long-term Solar Radiation Forecasting using a Deep Learning Approach-GRUs,” *APAP 2019 - 8th IEEE Int. Conf. Adv. Power Syst. Autom. Prot.*, pp. 917–920, 2019, doi: 10.1109/APAP47170.2019.9224661.
- [105] B. Gao, X. Huang, J. Shi, Y. Tai, and R. Xiao, “Predicting day-ahead solar irradiance through gated recurrent unit using weather forecasting data,” *J. Renew. Sustain. Energy*, vol. 11, no. 4, 2019, doi: 10.1063/1.5110223.
- [106] R. H. Inman, H. T. C. Pedro, and C. F. M. Coimbra, “Solar forecasting methods for renewable energy integration,” *Prog. Energy Combust. Sci.*, vol. 39, no. 6, pp. 535–576, 2013, doi: 10.1016/j.pecs.2013.06.002.
- [107] C. B. · A. C. · G. Martínez-Muñoz, “A Comparative Analysis of Gradient Boosting Algorithms,” vol. 3, 2021.
- [108] J. Zeng and W. Qiao, “Short-term solar power prediction using a support vector machine,” *Renew. Energy*, vol. 52, pp. 118–127, 2013, doi: 10.1016/j.renene.2012.10.009.
- [109] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, “Optuna: A Next-generation Hyperparameter Optimization Framework,” *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 2623–2631, 2019, doi: 10.1145/3292500.3330701.
- [110] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, and B. Mphago, *A survey on missing data in machine learning*. Springer International Publishing, 2021.
- [111] Y. Zhou, S. Aryal, and M. R. Bouadjenek, “A Comprehensive Review of Handling Missing Data: Exploring Special Missing Mechanisms,” 2024.
- [112] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal, “Algorithm 778: L-BFGS-B: Fortran Subroutines for Large-Scale Bound-Constrained Optimization,” *ACM Trans. Math. Softw.*, vol. 23, no. 4, pp. 550–560, 1997, doi: 10.1145/279232.279236.
- [113] W. Wang *et al.*, “Economics of physics-based solar forecasting in power system day-ahead scheduling,” *Renew. Sustain. Energy Rev.*, vol. 199, p. 114448, Jul. 2024, doi: 10.1016/j.rser.2024.114448.
- [114] S. Al-Dahidi *et al.*, “Techno-economic implications and cost of forecasting errors in solar PV power production using optimized deep learning models,” *Energy*, vol. 323, 2025, doi: 10.1016/j.energy.2025.135877.
- [115] Chính phủ, “Quy hoạch phát triển điện lực quốc gia thời kỳ 2021-2030, tầm nhìn đến năm 2050 (Quy hoạch điện VIII - Vietnam Eight Power Development Plan).” 2023.
- [116] G. Li, S. Xie, B. Wang, J. Xin, Y. Li, and S. Du, “Photovoltaic Power Forecasting with a Hybrid Deep Learning Approach,” *IEEE Access*, vol. 8, pp. 175871–175880, 2020, doi: 10.1109/ACCESS.2020.3025860.
- [117] F. Mei *et al.*, “Day-Ahead nonparametric probabilistic forecasting of photovoltaic power generation based on the LSTM-QRA ensemble model,” *IEEE Access*, vol. 8, pp. 166138–166149, 2020, doi: 10.1109/ACCESS.2020.3021581.
- [118] S. Liu, X. Hao, Z. Meng, J. Li, T. Cui, and L. Wei, “Application of SRNN-GRU

- in Photovoltaic power Forecasting,” *E3S Web Conf.*, vol. 256, pp. 1–7, 2021, doi: 10.1051/e3sconf/202125602001.
- [119] M. Massaoudi *et al.*, “An Effective Hybrid NARX-LSTM Model for Point and Interval PV Power Forecasting,” *IEEE Access*, vol. 9, pp. 36571–36588, 2021, doi: 10.1109/ACCESS.2021.3062776.
 - [120] M. Abumohsen, A. Y. Owda, M. Owda, and A. Abumihsan, “Hybrid machine learning model combining of CNN-LSTM-RF for time series forecasting of Solar Power Generation,” *e-Prime - Adv. Electr. Eng. Electron. Energy*, vol. 9, no. March, p. 100636, 2024, doi: 10.1016/j.prime.2024.100636.
 - [121] C. F. Yen, H. Y. Hsieh, K. W. Su, M. C. Yu, and J. S. Leu, “Solar Power Prediction via Support Vector Machine and Random Forest,” in *E3S Web of Conferences*, Nov. 2018, vol. 69, doi: 10.1051/e3sconf/20186901004.
 - [122] N. D. Tuyen, V. Xuan, S. Huu, and L. V. Thinh, “Photovoltaic Power Generation Forecasting Utilizing Long Short Term Memory,” *Meas. Control Autom.*, vol. 1, no. April, pp. 2–7, 2021.
 - [123] Nguyen Tuan Anh *et al.*, “Forecasting of solar power generation in Vietnam deploying a simple GRU model,” *Conf. Proc. - 2023 IEEE Asia Meet. Environ. Electr. Eng. EEE-AM 2023*, pp. 1–6, 2023, doi: 10.1109/EEE-AM58328.2023.10395560.
 - [124] Nguyễn Tuấn Anh, Phạm Mạnh hải, Vũ Minh Pháp, Nguyễn Đăng Toàn và Vũ Thị Anh Thơ, “Dự báo công suất nguồn phát điện mặt trời bằng mô hình BiGRU,” *Tạp chí Khoa học và Công nghệ Năng lượng - Trường Đại học Điện lực*, vol. 34, pp. 1–14, 2024.

PHỤ LỤC

Bảng A.1. Một số nhà máy điện mặt trời tại Việt Nam

(Nguồn [115])

STT	Tên nhà máy	Công suất (MWp)	Địa điểm
1	Xuân Thiện Ea Súp	600	Ea Súp, Đắk Lắk
2	Cụm nhà máy điện mặt trời Lộc Ninh 1, 2, 3	550	Lộc Ninh, Bình Phước
3	Trung Nam Thuận Bắc	450	Thuận Nam, Ninh Thuận
5	Nhà máy điện mặt trời Trung Nam Thuận Nam	450	Thuận Nam, Ninh Thuận
4	Cụm nhà máy điện mặt trời Dầu Tiếng 1, 2	420	Dương Minh Châu, Tây Ninh
6	Nhà máy điện mặt trời Phù Mỹ	330	Phù Mỹ, Bình Định
7	Nhà máy điện mặt trời BIM	330	Thuận Nam, Ninh Thuận
9	Nhà máy điện mặt trời Trung Nam	258	Ninh Thuận
8	Nhà máy điện mặt trời Hòa Hội	257	Phú Hòa, Phú Yên
12	Nhà máy điện mặt trời Hồng Phong 1A	195	Bình Thuận
10	Nhà máy điện mặt trời Sông Lũy 1	168	Bắc Bình, Bình Thuận
11	Nhà máy điện mặt trời CMX Renewable	168	Ninh Thuận
13	Nhà máy điện mặt trời Hồng Phong 1B	130	Bình Thuận
14	Nhà máy điện mặt trời Phước Hữu	65	Ninh Phước, Ninh Thuận
18	Nhà máy điện mặt trời Fujiwara Bình Định	50	Quy Nhơn, Bình Định
19	Nhà máy điện mặt trời Sre Pok 1	50	Buôn Đôn, Đắk Lắk
15	Nhà máy điện mặt trời Cát Hiệp	49,5	Phù Cát, Bình Định

17	Nhà máy điện mặt trời Phong Điền	48	Phong Điền, Thừa Thiên Huế
16	Nhà máy điện mặt trời Vĩnh Hảo	42,65	Tuy Phong, Bình Thuận

Bảng A.2: Tập dữ liệu lịch sử của nhà máy tại Quảng Trị

Thời gian	Nhiệt độ môi trường (°C)	Nhiệt độ tấm pin (°C)	Bức xạ mặt trời (W/m ²)	Công suất phát (kW)
1/1/2022 0:00	19	17,7	0	0
1/1/2022 0:05	19	17,7	0	0
1/1/2022 0:10	19	17,9	0	0
1/1/2022 0:15	19	17,8	0	0
1/1/2022 0:20	19	17,7	0	0
1/1/2022 0:25	19	17,7	0	0
1/1/2022 0:30	19	17,8	1	0
1/1/2022 0:35	19	18	0	0
1/1/2022 0:40	19	18	0	0
1/1/2022 0:45	18,9	18,1	1	0
.....				
10/6/2022 12:00	35,1	43,4	588	27212,89
10/6/2022 12:05	35,1	43,9	635	28793,98
10/6/2022 12:10	35,3	44,1	653	29028,22
10/6/2022 12:15	35,4	44,3	643	28536,05
10/6/2022 12:20	35,4	44,5	668	30020,79
10/6/2022 12:25	35,5	44,5	656	29354,99
10/6/2022 12:30	35,5	44,8	662	29648,79
10/6/2022 12:35	35,65	45	669	30136,1
10/6/2022 12:40	35,5	44,5	631	27757,42
10/6/2022 12:45	35,6	44,6	672	29976,04
10/6/2022 12:50	35,6	44,8	679	30594,15
10/6/2022 12:55	35,6	45	693	31475,51
10/6/2022 13:00	35.65	45,5	674	30613,02

.....				
31/12/2022 23:10	16,55	15,7	5	0
31/12/2022 23:15	16,5	15,6	5	0
31/12/2022 23:20	16,45	15,6	6	0
31/12/2022 23:25	16,45	15,6	6	0
31/12/2022 23:30	16,45	15,5	6	0
31/12/2022 23:35	16,45	15,5	6	0
31/12/2022 23:40	16,35	15,5	6	0
31/12/2022 23:45	16,4	15,4	8	0
31/12/2022 23:50	16,4	15,4	6	0

Bảng A.3. Các mô hình hybrid trong dự báo điện mặt trời

Năm	Tác giả	Mô hình	Độ chính xác	Khoảng thời gian dự báo hiệu quả
2019	M. Chai và các cộng sự [26]	LSTM	δ_{MAPE} (trung bình) = 19,8225%	24 giờ
		Time-LSTM	δ_{MAPE} (trung bình)= 9,3881%	24 giờ
		FAF-LSTM	δ_{MAPE} (trung bình)= 7,8261%	24 giờ
		MRWE-LSTM	δ_{MAPE} (trung bình)= 6,3397%	24 giờ
		LFA-LSTM	δ_{MAPE} (trung bình)= 6,0024%	24 giờ
		AHAP-LSTM	δ_{MAPE} (trung bình)= 5,7980%	24 giờ
2020	V. Suresh và các cộng sự [27]	CNN-LSTM	RMSE= 0,053; MAE= 0,053	1 giờ
			RMSE= 0,051; MAE= 0,035	24 giờ
			RMSE= 0,045; MAE= 0,030	168 giờ
2020	Li và các cộng sự [116]	CNN-LSTM	MAE =1,028%- 17,236% RMSE=1,448%- 33,405%	< 1 giờ đến 3 giờ
2020	Mei và các cộng sự [117]	LSTM-QRA	MAE (pu)= 34,8857; $L_{pinball-m}$ = 49,0370	24 giờ

		DNN-QRA	MAE (pu)= 41,6019; $L_{pinball-m}=76,7559$	24 giờ
		ANN-QRA	MAE (pu)= 70,4060; $L_{pinball-m}=89,4531$	24 giờ
		Persistence-QRA	MAE (pu)= 78,1367; $L_{pinball-m}=122,0268$	24 giờ
		QRNN	$L_{pinball-m}=59,9841$	24 giờ
2020	Pengyun Jia và các cộng sự [29]	IPSO-LSTM	MAE= 1,714; $R^2= 0,998$	24 giờ
		ISSA-LSTM	MAE= 1,417; $R^2= 0,998$	24 giờ
		PSO-GRU	MAE= 1,935; $R^2= 0,997$	24 giờ
		IPSO-GRU	MAE= 1,847; $R^2= 0,998$	24 giờ
		SSA-GRU	MAE= 1,676; $R^2= 0,998$	24 giờ
		ISSA-GRU	MAE= 1,402; $R^2= 0,998$	24 giờ
		VMD-IPSO-LSTM	MAE= 1,126; $R^2= 0,999$	24 giờ
		VMD-IPSO-GRU	MAE= 1,145; $R^2= 0,999$	24 giờ
		VMD-ISSA-GRU	MAE= 1,012; $R^2= 0,999$	24 giờ
2021	S. Liu và các cộng sự [118]	SRNN-GRU	SRNN-GRU có thời gian đào tạo nhanh hơn gấp 4 lần so với các mô hình RNN, LSTM và GRU thông thường (SRNN-GRU=27s, RNN= 60s, LSTM=117s, GRU=100s). sai số MAE và MSE của mô hình SRNN-GRU (MAE= 0,1102, MSE= 0,1145) cũng thấp hơn nhiều so với 3 mô hình RNN (MAE=0,5673, MSE=	< 1 giờ

			2,3149); LSTM (MAE= 0,3485; MSE= 0,6037) và GRU (MAE= 0,3813; MSE= 0,6939).	
2021	Mohamed Massaoudi và các cộng sự [119]	NARX-LSTM	Với tập dữ liệu 1: nRMSE=1,98%. Với tập dữ liệu 2: nRMSE=1,33%	1 giờ
2022	Mohammed Sabri và các cộng sự [28]	GRU-CNN	RMSE _{trung bình} =0,135; R^2 _{trung bình} =0,998; MAE _{trung bình} = 0,081; MSE _{trung bình} =0,019	1 giờ
2023	Fatma Mazen Ali Mazen và các cộng sự [32]	GRU-DILATE-TFT	MAE= 1,19; RMSE=1,44; MSE=2,08	Hàng giờ
2024	Mobarak Abumohsen và các cộng sự [120]	CNN-LSTM-RF	R^2 = 92%; RMSE=0,07kW; MAE=0,05 kW	1 giờ đến 1 ngày
2024	Jihoon Moon [31]	1D-CNN + Transformer (gộp single & multi-step)	Mô hình đề xuất cho kết quả MAE = 1.243 kWh và RMSE = 1.813 kWh trên tập DKASC-ASA-1A, vượt trội so với 1D-CNN đơn thuần và các mô hình kết hợp không có dự báo bước đơn	5 phút đến 1 giờ

Bảng A.4. Các mô hình kết hợp sử dụng cây quyết định

Năm	Tác giả	Mô hình	Độ chính xác
2020	Zhen Wang và các cộng sự [34]	LightGBM-LSTM	RMSE _{LightGBM-LSTM} = 0,1575; MAPE _{LightGBM-LSTM} = 3,70% so với RMSE _{XGB} = 0,232; MAPE _{XGB} = 4,18%; RMSE _{LightGBM} = 0,181; MAPE _{LightGBM} =

			4,88%; $RMSE_{LSTM} = 0,263$; $MAPE_{LSTM} = 6,94\%$ thì mô hình LightGBM-LSTM có độ chính xác cao nhất.
2024	Safia Babikir và các cộng sự [36]	XGBoost-CNN	Mô hình được đề xuất đã cho thấy hiệu suất vượt trội so với các mô hình đương đại, mô tả RMSE là 44,18 và R^2 là 0,9962, cung cấp dự báo chính xác về công suất DC trong điều kiện thời tiết động.
2018	Chih-Feng Yen và các cộng sự [121]	RF-SVR	Trong các kết quả dự báo, RF hoạt động tốt hơn kết quả dự đoán của SVM, trong đó RF ước tính cao hơn một chút các giá trị sản lượng thực tế và SVM đánh giá thấp chúng. Về các phép đo khác, RF hoạt động tốt hơn khoảng 37% đến 40% về RMSE, MAE và MASE.
2023	Jiandong Ye và các cộng sự [35]	LightGBM-XGBoost	Khi so sánh kết quả dự báo của mô hình kết hợp LightGBM-XGBoost với các mô hình LSTM, SVR, nuSVR, PA thì mô hình kết hợp trong 3 kịch bản đều có kết quả RMSE, MAE thấp nhất.
2021	Quoc-Thang Phan [37]	KPCA-XGBoost-NWP	Khi so sánh sai số RMSE của mô hình KPCA-XGBoost-NWP với PCA-XGBoost và XGBoost tại 5 nhà máy điện mặt trời tại Đài Loan thì mô hình KPCA-XGBoost-NWP có sai số RMSE thấp nhất.
2019	Alexandra Khalyasmaa [22]	RF -Regressor	Độ chính xác dự báo trung bình đạt được là khoảng 93%
2022	Waqas Khan [33]	DSE-XGB	Phương pháp DSE-XGB được đề xuất thể hiện sự kết hợp tốt nhất giữa tính nhất

			quán và độ ổn định trong các nghiên cứu trường hợp khác nhau bất kể sự thay đổi của thời tiết và chứng minh giá trị R^2 được cải thiện từ 10% đến 12% so với các mô hình ANN, LSTM, bagging.
--	--	--	--

Bảng A.5. Các nghiên cứu về dự báo công suất phát điện mặt trời tại Việt Nam

Số TT	Tên bài báo	Tên tác giả	Mô hình dự báo	Kết quả thu được
1	Applying Artificial Intelligence in Forecasting the Output of Industrial Solar Power Plant in Vietnam [24]	Nguyễn Quang Ninh, Bùi Duy Linh, Đoàn Văn Bình, Nguyễn Đình Quang	LSTM, áp dụng cho nhà máy Phong Điền	- Khi áp dụng với tập dữ liệu lịch sử của nhà máy thì MAPE bằng 1,5% và $MAPE_{min}$ bằng 1,085%. - Khi áp dụng tập dữ liệu của bên thứ ba thì RMSE tăng từ 0,996 MW lên 3,216 MW. $MAPE_{min}$ tăng từ 1,085% lên 4,432%
2	A new method for forecasting energy output of a large-scale solar power plant based on long short-term memory networks a case study in Vietnam [57]	Nguyễn Quang Ninh, Bùi Duy Linh, Đoàn Văn Bình, Nguyễn Đình Quang, Eleonora Riva Sanseverino, Dario Di Cara	LSTM, áp dụng cho nhà máy Thành Công 1	- Sử dụng dữ liệu thời tiết dự báo gây ra sai số MAPE lớn hơn so với việc sử dụng dữ liệu thời tiết lịch sử (10,857% so với 3,491%). - Sau khi cải thiện mô hình, bổ sung tính năng mới và xử lý sai số từ dự báo thời tiết, MAPE giảm từ 10,857% xuống còn 9,881%.
3	Forecasting energy output of a solar power plant in	Nguyễn Quang Ninh, Bùi Duy	LSTM kết hợp với kỹ thuật chia	So với mô hình cũ sai số MAPE giảm 6,059% và sai số RMSE giảm 6,710%

	curtailment condition based on LSTM using P/GHI coefficient and validation in training process, a case study in Vietnam [38]	Linh, Đoàn Văn Bình, Eleonora Riva Sanseverino	khoảng GHI và cộng hệ số P/GHI	
4	Photovoltaic Power Generation Forecasting Utilizing Long Short Term Memory [122]	Nguyễn Đức Tuyên, Vũ Xuân Sơn Hữu, Lê Viết Thịnh	GA-LSTM	- Phương án 1: dự báo công suất điện mặt trời trong tháng 9-2006 với RMSE là 4,9 W/m² và MAPE là 10,3% khi dự báo trước 1 ngày (thấp hơn mô hình tham khảo RMSE= 6,74 W/m², và MAPE= 13,12%. Trong trường hợp dự báo trước 2 ngày, RMSE và MAPE lần lượt là 4,48 W/m² và 8,78%, thấp hơn so với mô hình tham khảo (RMSE= 5,08 W/m² và MAPE= 12,05 %). - Phương án 2: dự báo công suất điện mặt trời trong tháng 12-2006, cho thấy RMSE là 4,03 W/m² và MAPE là 6,92% khi dự báo trước 1 ngày (so với mô hình tham khảo RMSE: 4,52 W/m² và MAPE= 10,51%). Đối với dự báo

				trước 2 ngày, RMSE và MAPE của phương án đề xuất lần lượt là 9,14 W/m² và 12,91%, thấp hơn so với mô hình tham khảo (RMSE= 10,7 W/m², MAPE= 13,86%). - Họ kết luật mô hình GA-LSTM hoạt động hiệu quả hơn mô hình tham khảo dựa trên so sánh kết quả RMSE và MAPE.
5	A Novel Forecasting Model for Solar Power Generation by a Deep Learning Framework With Data Preprocessing and Postprocessing [21]	Phan Quốc Thắng, Yuan-Kang Wu, Phan Quốc Dũng, Hsin-Yen Lo	KPCA-XGB-GRU	- Kết quả thực nghiệm cho thấy mô hình KPCA-XGB-GRU có sai số NMAPE và NRMSE trong phân dự báo lần lượt là 1,52% và 2,92% chính xác dự báo tốt hơn các mô hình điển hình khác (ANN, GRU, XGBoost). - Việc tiền xử lý dữ liệu và hậu xử lý dữ liệu đóng vai trò quan trọng vì thu thập các đặc trưng quan trọng nhất của tập dữ liệu và cải thiện chất lượng của tập dữ liệu để đào tạo mô hình.
6	Multiple Step Ahead Forecasting of Rooftop Solar Power Based on a	Nguyễn Thị Hoài Thu, Phan Quốc	CEEMDAM-BiLSTM-PSO	- Kết quả cho thấy mô hình đề xuất có sai số thấp nhất khi so sánh với các mô

	Novel Hybrid Model of CEEMDAN - Bidirectional LSTM Network with Structure Optimized by PSO Method [30]	bảo, Nguyễn Vũ Nhất Nam		hình khác như ANN, LSTM và BiLSTM đơn
7	Short-term multi-step forecasting of rooftop solar power generation using a combined data decomposition and deep learning model of EEMD-GRU [40]	Nguyễn Vũ Nhất Nam, Nguyễn Hữu Đức, Nguyễn Thị Hoài Thu	EEMD-GRU	- Kết quả thí nghiệm cho thấy mô hình này vượt trội hơn hẳn so với các mô hình đơn lẻ như ANN, LSTM và GRU khi đánh giá trên các chỉ số sai số MAE, RMSE và NRMSE trong cả ba kịch bản dự báo 1 bước, 2 bước và 3 bước.
8	Forecasting of solar power generation in Vietnam deploying a simple GRU model [123]		GRU	- Mô hình GRU đơn giản hoạt động khá tốt và đáng tin cậy khi được cung cấp bộ dữ liệu lịch sử hoàn chỉnh, nhưng sự hiện diện của các điểm giá trị nhỏ trong bộ dữ liệu vẫn ảnh hưởng đến độ chính xác của các dự báo.
9	Dự báo công suất nguồn phát điện mặt trời bằng mô hình BiGRU [124]		BiGRU (1 lớp ẩn và 2 lớp ẩn)	- Hai mô hình BiGRU một lớp ẩn và hai lớp ẩn đều có kết quả dự báo tốt cho bài toán dự báo công suất điện mặt trời. Việc tiêu tốn tài nguyên và chi phí cao hơn của mô hình hai lớp ẩn đối với quá trình tính toán nên

				được cân nhắc kỹ lưỡng, đặc biệt khi sai số không có sự cải thiện đáng kể so với mô hình một lớp ẩn điều này cho thấy khả năng dự báo của mô hình BiGRU một lớp ẩn không thua kém quá nhiều so với mô hình hai lớp ẩn. Nếu sự cải thiện về hiệu suất dự báo của mô hình hai lớp ẩn không đáng kể và thời gian tính toán là yếu tố quan trọng thì mô hình một lớp ẩn là lựa chọn hiệu quả hơn
--	--	--	--	---

