

A TWO-STAGE HYBRID STRATEGY FRAMEWORK FOR SOLAR
POWER FORECASTING WITH MISSING DATA CORRECTION AND
CURTAILMENT AWARENESS: A CASE STUDY OF THE SUT CAMPUS
FLOATING SOLAR SYSTEM



A Thesis Submitted in Partial Fulfillment of the Requirements for
the Degree of Master of Engineering in Electrical Engineering
Suranaree University of Technology
Academic Year 2025

กรอบการพยากรณ์กำลังไฟฟ้าพลังงานแสงอาทิตย์แบบผสมผสานสองขั้นตอน
โดยคำนึงถึงการแก้ไขข้อมูลสูญหายและการจำกัดกำลังผลิต:
กรณีศึกษาระบบผลิตไฟฟ้าพลังงานแสงอาทิตย์ลอยน้ำ
มหาวิทยาลัยเทคโนโลยีสุรนารี



นางสาววิภาดา ทองวิไล

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมไฟฟ้า
มหาวิทยาลัยเทคโนโลยีสุรนารี
ปีการศึกษา 2568

A TWO-STAGE HYBRID STRATEGY FRAMEWORK FOR SOLAR POWER
FORECASTING WITH MISSING DATA CORRECTION AND CURTAILMENT
AWARENESS: A CASE STUDY OF THE SUT CAMPUS
FLOATING SOLAR SYSTEM

Suranaree University of Technology has approved this thesis submitted in partial fulfillment of the requirements for a Master's Degree.

Thesis Examining Committee

Somchat Jiriwibhakorn

(Assoc. Prof. Dr. Somchat Jiriwibhakorn)

Chairperson

Keerati Chayakulkheeree

(Assoc. Prof. Dr. Keerati Chayakulkheeree)

Member (Thesis Advisor)

Nitikorn Junhuathon

(Asst. Prof. Dr. Nitikorn Junhuathon)

Member (Thesis Co-Advisor)

Thanatchai Kulworawanichpong

(Prof. Dr. Thanatchai Kulworawanichpong)

Member

Tosaphol Ratniyomchai

(Asst. Prof. Dr. Tosaphol Ratniyomchai)

Member

Yupaporn Ruksakulpiwat

(Assoc. Prof. Dr. Yupaporn Ruksakulpiwat)

Acting Vice Rector for Academic Affairs
and Quality Assurance

Pornsiri Jongkol

(Assoc. Prof. Dr. Pornsiri Jongkol)

Dean of Institute of Engineering

วิภาดา ทองวิไล : กรอบการพยากรณ์กำลังไฟฟ้าพลังงานแสงอาทิตย์แบบผสมผสานสอง
ขั้นตอน โดยคำนึงถึงการแก้ไขข้อมูลสูญหายและการจำกัดกำลังผลิต: กรณีศึกษาระบบผลิต
ไฟฟ้าพลังงานแสงอาทิตย์ลอยน้ำ มหาวิทยาลัยเทคโนโลยีสุรนารี (A TWO-STAGE HYBRID
STRATEGY FRAMEWORK FOR SOLAR POWER FORECASTING WITH MISSING DATA
CORRECTION AND CURTAILMENT AWARENESS: A CASE STUDY OF THE SUT
CAMPUS FLOATING SOLAR SYSTEM)

อาจารย์ที่ปรึกษา : รองศาสตราจารย์ ดร. กীরติ ชยะกุลศิริ, 166 หน้า.

คำสำคัญ : การพยากรณ์กำลังผลิตไฟฟ้าพลังงานแสงอาทิตย์ระยะสั้น / LSTM / Bi-LSTM / GRU /
RF / TFT-lite / SHAP / ปัญญาประดิษฐ์ที่อธิบายได้

วิทยานิพนธ์นี้นำเสนอแนวทางการพยากรณ์กำลังผลิตไฟฟ้าจากระบบโซลาร์เซลล์ลอยน้ำ
(Floating Photovoltaic: FPV) ของมหาวิทยาลัยเทคโนโลยีสุรนารี ภายใต้เงื่อนไขของข้อมูลไม่
สมบูรณ์ ความผันผวนตามฤดูกาล และข้อจำกัดด้านการจำกัดกำลังผลิตไฟฟ้า (Curtailment) โดย
พัฒนาแนวทางการพยากรณ์แบบสองขั้นตอนซึ่งเริ่มจากการจัดการคุณภาพข้อมูลและการพยากรณ์
ปัจจัยสิ่งแวดล้อม ก่อนนำผลลัพธ์ไปใช้เป็นข้อมูลนำเข้าสำหรับการพยากรณ์กำลังผลิตไฟฟ้าด้วย
แบบจำลองการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึกหลายรูปแบบ

จากผลการศึกษาแสดงให้เห็นว่ากรอบแนวคิดที่เสนอสามารถเพิ่มความแม่นยำและความ
เสถียรของการพยากรณ์กำลังผลิตไฟฟ้าได้อย่างมีนัยสำคัญ โดยเฉพาะในกรณีที่ข้อมูลมีการขาดหาย
และมีข้อจำกัดกำลังผลิต ทั้งนี้ การวิเคราะห์เชิงอธิบายช่วยยืนยันบทบาทสำคัญของตัวแปร
สิ่งแวดล้อมต่อกำลังผลิตไฟฟ้า และสนับสนุนความเหมาะสมของแนวทางที่นำเสนอสำหรับการ
ประยุกต์ใช้งานจริงในระบบพลังงานแสงอาทิตย์ลอยน้ำ อย่างไรก็ตาม การประยุกต์ใช้ในพื้นที่อื่นอาจ
ต้องปรับขั้นตอนการเตรียมข้อมูลและการฝึกแบบจำลองให้เหมาะสมกับลักษณะระบบที่แตกต่างกัน

สาขาวิชาวิศวกรรมไฟฟ้า

ปีการศึกษา 2568

ลายมือชื่อนักศึกษาวิภาดา ทองวิไล.....
ลายมือชื่ออาจารย์ที่ปรึกษา.....กীরติ ชยะกุลศิริ.....
ลายมือชื่ออาจารย์ที่ปรึกษาร่วม.....

WIPADA THONGWILY: A TWO-STAGE HYBRID STRATEGY FRAMEWORK FOR SOLAR
POWER FORECASTING WITH MISSING DATA CORRECTION AND CURTAILMENT
AWARENESS: A CASE STUDY OF THE SUT CAMPUS FLOATING SOLAR SYSTEM
THESIS ADVISOR: ASSOC. PROF. KEERATI CHAYAKULKHEEREE, Ph.D., 166 PP.

Keyword: short-term PV forecasting/ LSTM/ Bi-LSTM/ GRU/ RF/ TFT-lite/ SHAP/ XAI

This thesis presents a methodology for predicting the power output of the floating photovoltaic (FPV) system at Suranaree University of Technology under conditions of incomplete data, seasonal fluctuations, and power curtailment limitations. It develops a two-stage prediction approach that begins with data quality management and environmental factor forecasting, followed by using the results as input for power output prediction using various machine learning and deep learning models.

The simulation results show that the proposed conceptual framework can significantly improve the accuracy and stability of power generation forecasting, especially in cases where data is missing and there are capacity limitations. Descriptive analysis helps confirm the important role of environmental variables on power generation and supports the suitability of the proposed approach for real-world applications in floating solar power systems. However, application in other areas may require adjusting the data preparation and model training steps to suit different system characteristics.

School of Electrical Engineering
Academic Year 2025

Student's Signature อัมภดา ทองวิลย์

Advisor's Signature ศาสตราจารย์ ดร. เกียรติ ชัยกุลเขeree

Co-Advisor's Signature ศาสตราจารย์ ดร. ชัยกุลเขeree

ACKNOWLEDGEMENT

The authors are particularly grateful to the Graduate Scholarship at Suranaree University of Technology, for the financial support that enabled this research to be successfully completed.

The author would like to express her sincere appreciation to her advisor, Assoc. Prof. Dr. Keerati Chayakulkheeree, for his invaluable guidance and continuous support throughout the research process. His insightful suggestions and constructive feedback significantly enhanced the quality of this thesis. Beyond academic supervision, his mentorship fostered systematic thinking, discipline, perseverance, and a positive attitude, which greatly contributed to both the author's research development and personal growth.

In addition, the author sincerely thanks her co-advisor, Asst. Prof. Dr. Nitikorn Junhuathon, for his valuable guidance and academic support, which enhanced the analytical direction and overall reliability of this thesis.

The author would also like to extend her sincere thanks to all members of the thesis examination committee, namely Assoc. Prof. Dr. Somchat Jiriwibhakorn, Prof. Dr. Thanatchai Kulworawanichpong, and Asst. Prof. Dr. Tosaphol Ratniyomchai, for their valuable comments, suggestions, and support throughout the thesis evaluation process.

The author would like to express her deepest gratitude to her parents for their unwavering emotional and financial support. Their trust and encouragement have been a vital foundation in the successful completion of her academic journey.

Finally, the author would like to thank the faculty of the School of Electrical Engineering for their academic guidance throughout her studies. She also extends her appreciation to her friends and seniors for their cooperation, advice, and encouragement, which contributed to the successful completion of this research.

Wipada Thongwily

TABLE OF CONTENTS

	Page
ABSTRACT (THAI)	I
ABSTRACT (ENGLISH)	II
ACKNOWLEDGEMENT	III
TABLE OF CONTENTS	IV
LIST OF TABLES	VII
LIST OF FIGURES	VIII
LIST OF ABBREVIATIONS	XIII
LIST OF NOMENCLATURES	XIII
CHAPTER	
I INTRODUCTION.....	1
1.1 General Introduction.....	1
1.2 Problem Statement	2
1.3 Research Objectives.....	2
1.4 Research hypothesis	3
1.5 Scope	3
1.6 Research Benefits	3
1.7 Structure of thesis.....	4
II LITERATURE REVIEW.....	6
2.1 Introduction	6
2.2 Literature Overview.....	7
2.3 Review of Data Management in Solar Power Forecasting	14
2.3.1 Handling Missing Data and Cleaning	14
2.3.2 Feature Selection and Engineering	15
2.3.3 Data Grouping.....	18
2.4 Review of Forecasting Models	19

TABLE OF CONTENTS (Continued)

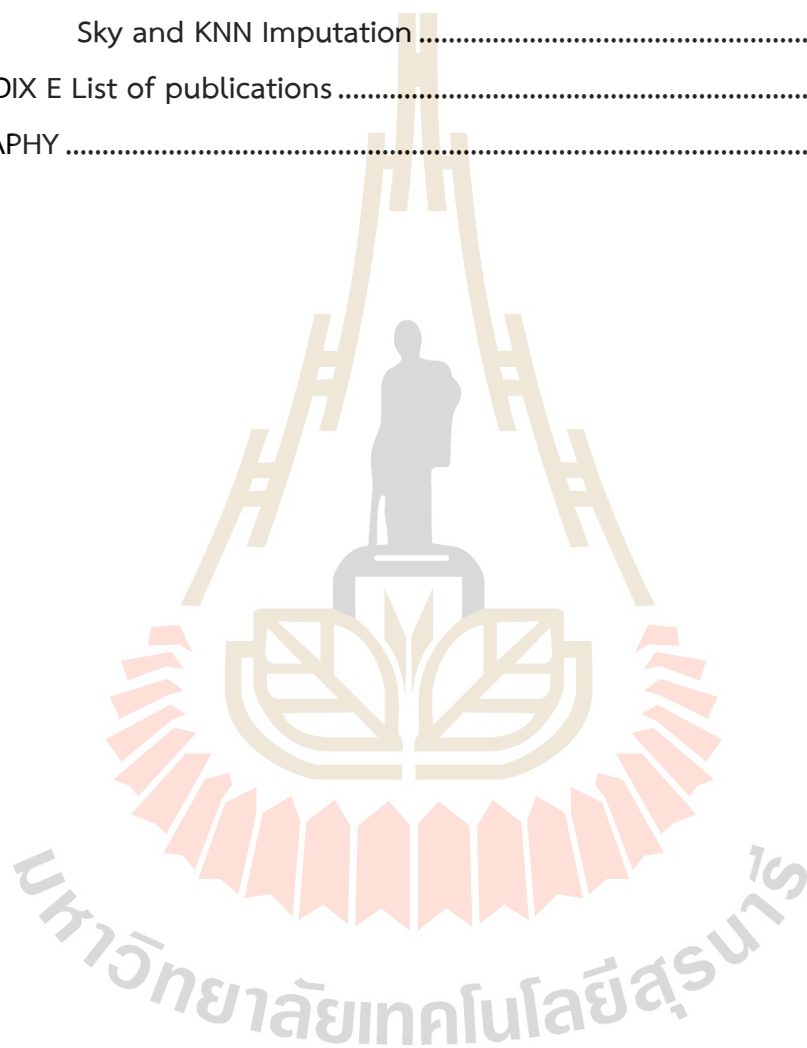
	Page
2.4.1 Data Grouping Statistical Machine Learning Methods.....	19
2.4.2 Deep Learning Models	21
2.4.3 Comparative Model Studies.....	23
2.5 Model Evaluation and Explainability	24
2.5.1 Deep Learning Models	24
2.6 Research Gaps.....	27
III TWO-STAGE DATA-DRIVEN MODELING FOR DAY-AHEAD SOLAR POWER FORECASTING: A CASE STUDY OF THE SUT CAMPUS FLOATING SOLAR SYSTEM.....	30
3.1 Introduction	30
3.2 Data preparation.....	31
3.2.1 Data collection.....	31
3.2.2 Data cleaning.....	33
3.2.3 Feature Selection and Correlation Analysis.....	37
3.2.3 Feature Selection and Correlation Analysis.....	40
3.2.4 Data Splitting	46
3.3 Two-Step Forecasting Framework.....	40
3.3.1 Environmental Factors Pre-Forecasting	40
3.3.2 PV-Forecast.....	42
3.4 Evaluation metrics.....	45
3.5 Simulation Results.....	45
3.5.1 Forecasting Results of Environmental Factors	46
3.5.2 Forecasting Results of Solar Power Generation	48
3.6 Performance Metric Comparison.....	51
3.7 Conclusion	53
IV SHORT-TERM SOLAR POWER FORECASTING UNDER INCOMPLETE DATA AND CURTAILMENT CONDITIONS	54

TABLE OF CONTENTS (Continued)

	Page
4.1 Introduction	54
4.2 Data preparation.....	55
4.2.1 Data collection.....	55
4.2.2 Data cleaning.....	55
4.2.3 Feature Selection and Exploratory Statistical Analysis.....	64
4.2.4 Data Splitting	67
4.3 Forecasting Model Development and Evaluation	70
4.4 Results and Discussion	71
4.5 Conclusion	77
V Data-Driven Solar Power Forecasting Using Pre-Forecasting and Clear-Sky/KNN Data Refinements.....	78
5.1 Introduction	78
5.2 Case 1: Augmented Data-Driven Solar Power Forecasting Using Pre-Forecasting and Clear-Sky/KNN Data Refinements.....	78
5.2.1 Data collection.....	79
5.2.2 Model development.....	80
5.2.3 Results and discussion	86
5.3 Case 2: Base case (Delete Missing Data).....	103
5.4 Conclusion.....	113
VI CONCLUSION	115
6.1 Conclusion	115
6.2 Suggestions	116
REFERENCES	118
APPENDIX A Solar cell datasheet.....	124
APPENDIX B Environmental Correlation with Solar Power Output Based on Linear Interpolation.....	126
APPENDIX C Evaluation Results for KNN Imputation Parameters.....	129

TABLE OF CONTENTS (Continued)

	Page
APPENDIX D Environmental Correlation with Solar Power Output Based on Clear-Sky and KNN Imputation	131
APPENDIX E List of publications	135
BIOGRAPHY	166



LIST OF TABLES

Table	Page
2.1 Overview of key data management aspects in solar power forecasting studies.....	8
2.2 Overview of modeling and interpretation approaches in solar power forecasting studies.....	11
2.3 Overview of XAI applications in solar power forecasting studies.....	26
2.4 Overview of key components and research gaps in solar power forecasting studies.....	27
3.1 Main specifications of Jinko Solar module (JKM540M-72HL4-V)	33
3.2 Hyperparameter configuration for model training	44
3.3 Forecasting results	52
3.4 Processing time comparison of forecasting models	52
4.1 Comparison of KNN imputation performance across k values	60
4.2 Quantitative metrics	63
4.3 Hyperparameters of environmental factors pre-forecasting models	70
4.4 Forecasting results	76
4.5 Comparison of forecasting results between before and after datasets across seasons.....	76
5.1 Hyperparameters of environmental factors pre-forecasting models	83
5.2 Hyperparameters of PV-forecasting models	83
5.3 Forecasting results with dataset 1 from LSTM&TFT pre-forecasting models	95
5.4 Forecasting results with dataset 2 from LSTM&LSTM pre-forecasting models	96

LIST OF TABLES (Continued)

Table	Page
5.5 Accuracy and stability of LSTM&TFT pre-forecasting models with dataset 2 under difference seasons.....	98
5.6 Accuracy and stability of LSTM&LSTM pre-forecasting models with dataset 1 under difference seasons.....	99
5.7 Best forecasting models of year 2023 data	100
5.8 Runtime of PV forecasting models	100
5.9 Forecasting results with dataset 1 from LSTM&TFT pre-forecasting models	110
5.10 Forecasting results with dataset 2 from LSTM&LSTM pre-forecasting models	111
5.11 Runtime of PV forecasting models	113
C.1 Comparison of KNN Imputation Performance Across K Values	129

LIST OF FIGURES

Figure	Page
3.1 The framework for solar power forecasting	30
3.2 The framework for data management process	31
3.3 Floating PV power plant at SUT.....	31
3.4 Weather data source interface (Open-meteo API)	32
3.5 Solar power output data interface (Exep platform)	32
3.6 Sample of weather data	33
3.7 Sample of power output data.....	34
3.8 Resulting dataset after data preparation	34
3.9 Power output data for the year 2023.....	34
3.10 Power output data for the year 2023 after using linear interpolation.....	35
3.11 Surface plot of PV power generation throughout the year 2023.....	36
3.12 The correlation matrix.....	37
3.13 Humidity (%) time series used as model input	38
3.14 Wind speed (m/s) time series used as model input.....	38
3.15 GTI (W/m ²) time series used as model input	39
3.16 Predicted temperature time series from LSTM model.....	39
3.17 Predicted irradiance time series from LSTM model.....	39
3.18 Architectural diagram of LSTM.....	41
3.19 Architectural diagram of Bi- LSTM	42
3.20 Architectural diagram of GTU	43
3.21 Architectural diagram of RF	44
3.22 Environmental correlation with solar power output - April 2023.....	47
3.23 Comparison of forecasted vs. actual power output using LSTM.....	49

LIST OF FIGURES (Continued)

Figure	Page
3.24 Comparison of forecasted vs. actual power output using Bi-LSTM.....	49
3.25 Comparison of forecasted vs. actual power output using GRU.....	49
3.26 Comparison of forecasted vs. actual power output using RF.....	49
3.27 Day-ahead solar power forecasting for the summer season	50
3.28 Day-ahead solar power forecasting for the rainy season	50
3.29 Day-ahead solar power forecasting for the winter season	51
4.1 The framework for solar power forecasting.....	54
4.2 The framework for data management process	55
4.3 Power output data for the year 2023 after using Clear Sky model	58
4.4 Power output data for the year 2023 after using KNN	62
4.5 Three-dimensional visualization of solar power output in 2023 after ClearSky with KNN data processing.....	63
4.6 The correlation matrix.....	64
4.7 Taylor diagram of temperature and irradiance.....	65
4.8 SHAP values impact on model output.....	66
4.9 Distribution of PV power to actual load ratio.....	68
4.10 Solar power output under curtailment conditions	68
4.11 Curtailment gap analysis by hour	69
4.12 Correlation of irradiance and temperature with solar power.....	71
4.13 MAE of environmental factor forecasts	72
4.14 Seasonal comparison of day-ahead solar power forecasting	74
5.1 The framework for case 1.....	79
5.2 The framework for data management process	79

LIST OF FIGURES (Continued)

Figure	Page
5.3 Architectural diagram of TFT-lite.....	81
5.4 MAE of environmental factor forecasts using the LSTM	87
5.5 MAE of environmental factor forecasts using the TFT-lite	87
5.6 The parameters correlation matrix after SHAP process	88
5.7 SHAP values impact on model output	89
5.8 Comparison of day-ahead solar power forecasting of the LSTM&TFT model	91
5.9 Comparison of day-ahead solar power forecasting of the LSTM&LSTM model	92
5.10 Power output data for the year 2023 after delete missing data	104
5.11 Three-dimensional visualization of solar power output in 2023 after delete missing data	104
5.12 MAE of environmental factor forecasts using the LSTM	105
5.13 MAE of environmental factor forecasts using the TFT-lite	106
5.14 The parameters correlation matrix after SHAP process	107
5.15 SHAP values impact on model output	108
A.1 Solar panel specification (data sheet) used in this study	124
B.1 Environmental correlation with solar power output - January 2023.....	126
B.2 Environmental correlation with solar power output - February 2023.....	126
B.3 Environmental correlation with solar power output - March 2023.....	126
B.4 Environmental correlation with solar power output - April 2023.....	126
B.5 Environmental correlation with solar power output - May 2023.....	127
B.6 Environmental correlation with solar power output - June 2023.....	127

LIST OF FIGURES (Continued)

Figure	Page
B.7 Environmental correlation with solar power output - July 2023.....	127
B.8 Environmental correlation with solar power output - August 2023.....	127
B.9 Environmental correlation with solar power output - September 2023.....	128
B.10 Environmental correlation with solar power output - October 2023.....	128
B.11 Environmental correlation with solar power output - November 2023.....	128
B.12 Environmental correlation with solar power output - December 2023.....	128
D.1 Environmental correlation with solar power output - January 2023.....	131
D.2 Environmental correlation with solar power output - February 2023.....	131
D.3 Environmental correlation with solar power output - March 2023.....	131
D.4 Environmental correlation with solar power output - April 2023.....	132
D.5 Environmental correlation with solar power output - May 2023.....	132
D.6 Environmental correlation with solar power output - June 2023.....	132
D.7 Environmental correlation with solar power output - July 2023.....	132
D.8 Environmental correlation with solar power output - August 2023.....	133
D.9 Environmental correlation with solar power output - September 2023.....	133
D.10 Environmental correlation with solar power output - October 2023.....	133
D.11 Environmental correlation with solar power output - November 2023.....	133
D.12 Environmental correlation with solar power output - December 2023.....	134

LIST OF ABBREVIATIONS

AI	=	Artificial Intelligence
ANN	=	Artificial Neural Network
ARIMA	=	Auto Regressive Integrated Moving Average
Bi-LSTM	=	Bi-directional LSTM
CI	=	Confidence Interval
CNN	=	Convolutional Neural Network
DL	=	Deep learning
DNN	=	Deep neural network
EDA	=	Exploratory Data Analysis
EG	=	Expected Gradients
FNN	=	Feedforward Neural Network
FPV	=	Floating Photovoltaic
GAIN	=	Generative Adversarial Imputation Nets
GHG	=	Greenhouse Gas
GRN	=	Gated Residual Network
GRU	=	Gated Recurrent Unit
IG	=	Integrated Gradients
IQR	=	Interquartile range
KNN	=	K-Nearest Neighbors
LSTM	=	Long Short-Term Memory
LSTM&LSTM	=	Forecasting model using LSTM for both temperature and irradiance
LSTM&TFT	=	Forecasting model using LSTM for temperature and TFT-lite for irradiance
MAE	=	Mean Absolute Error

LIST OF ABBREVIATIONS (Continued)

MAPE	=	Mean Absolute Percentage Error
ME	=	Mean Error
Mean	=	Arithmetic Mean
MGU	=	Minimal Gated Unit
Median	=	Median Value
ML	=	Machine Learning
MLP	=	Multi-Layer Perceptron
Min	=	Minimum Value
Max	=	Maximum Value
NGBoost	=	Natural Gradient Boosting
PV	=	Photovoltaic
PSO	=	Particle swarm optimization
R ²	=	Coefficient of Determination or R-squared
RF	=	Random Forest
RESs	=	Renewable Energy Sources
RMSE	=	Root Mean Square Error
RNN	=	Recurrent Neural Network
SD	=	Standard Deviation
SDM	=	Single Diode Model
SHAP	=	SHapley Additive Explanations
SNN	=	Spiking neural networks
SOM	=	Self-Organizing Map
SVR	=	Support Vector Regression
TCN	=	Temporal Convolutional Network
TFT	=	Temporal Fusion Transformer
VSN	=	Variable Selection Network
WMAPE	=	Weighted Mean Absolute Percentage Error

LIST OF ABBREVIATIONS (Continued)

XAI	=	Explainable Artificial Intelligence
XCFL	=	Explainable Clustered Federated Learning
XGBoost	=	Extreme Gradient Boosting



LIST OF NOMENCLATURES

A	=	Total area of the solar panels (m ²)
ClearSky with KNN	=	Missing value estimation method using Clear Sky model followed by KNN
DNI	=	Direct Normal Irradiance (W/m ²)
GHI	=	Global Horizontal Irradiance (W/m ²)
GTI	=	Global Tilted Irradiance (W/m ²)
$K_c(t)$	=	Clear-Sky Index at time t
KNN-all	=	Missing value estimation using KNN over the entire dataset
n	=	Number of samples
$P_{clear}(t)$	=	Power generation under clear sky conditions (kW)
$P_{real, no\ curtail}(t)$	=	Actual Power Output without Curtailment at time t (kW)
$P_G(t)$	=	The electrical power drawn from the utility grid (metering load)
$P_L(t)$	=	The actual electrical load (actual load)
$P_{PV}(t)$	=	The power generated by the PV system
$P_{filled}(t)$	=	Imputed Power Output at time t (kW)
Pred.irr_LSTM	=	Predicted irradiance from the LSTM model
Pred.irr_TFT	=	Predicted irradiance from the Temporal Fusion Transformer-lite model
Pred.temp_LSTM	=	Predicted temperature from the LSTM model
Pred.temp_TFT	=	Predicted temperature from the Temporal Fusion Transformer-lite model

LIST OF NOMENCLATURES (Continued)

$R(t)$	=	The ratio reflecting the relationship between PV power generation and the actual electrical load at time
$x_{missing}$	=	The missing power generation value
$y_{actual,i}$	=	Actual power generation at sample i (kW)
$y_{forecast,i}$	=	Predicted power generation at sample i (kW)
η	=	Efficiency of the photovoltaic system
α	=	Threshold defined to indicate conditions under which the PV power generation is comparable to or exceeds the actual load
α_1		Scaling Factor (Robust Performance Ratio)
β	=	Lower threshold of grid power consumption, used to indicate that PV generation predominantly supplies the load demand

CHAPTER I

INTRODUCTION

1.1 General Introduction

Currently, global energy demand continues to rise steadily due to population growth and economic development. As a result, reliance on fossil fuel sources such as oil, coal, natural gas, oil shale, and oil sands remains high. However, these energy sources are limited in quantity and have significant environmental impacts, particularly the emission of GHGs such as carbon dioxide. Therefore, reducing dependency on fossil fuels are essential. Renewable energy sources such as solar, wind, hydro, biomass, and geothermal energy offer a more sustainable alternative. Among these, solar energy has gained increasing attention as a key alternative energy source. In Thailand, the government has set a target to achieve net zero carbon emissions by the year 2050 (NXPO, 2022). This has driven the adoption of Solar PV systems in various sectors. One rapidly growing form of solar energy utilization is the FPV system, which is installed on water surfaces such as reservoirs and ponds. FPV systems offer several advantages over ground-mounted systems, including land use reduction, improved heat dissipation due to the cooling effect of water, and decreased water evaporation.

However, electricity generation from solar energy still faces significant challenges due to natural variability and the uncertainty of energy output at different times. These challenges are primarily caused by weather-related factors such as solar irradiance, temperature, humidity, and cloud cover, all of which directly affect energy generation efficiency. Such uncertainties are major obstacles to the widespread integration of renewable energy, potentially undermining the reliability, stability, and efficiency of power systems. This is especially true when integrating renewable energy sources into modern smart grid systems. Therefore, accurate forecasting of solar energy generation is essential to manage variability and uncertainty effectively while maintaining the overall balance and stability of the power grid.

Consequently, the development of precise solar power forecasting techniques is critically important. In recent years, the use of AI and ML techniques for solar power forecasting has gained significant traction due to their effectiveness in analyzing time-series data. Several models have been applied, including LSTM, Bi-LSTM, GRU, and RF, to improve forecasting accuracy. Furthermore, data quality issues, such as missing or anomalous data, also significantly impact model performance. This research aims to develop a two-stage data-driven framework that focuses on managing missing data using the Clear Sky Model and KNN technique, along with developing predictive models tailored to the system's specific constraints. The FPV system at Suranaree University of Technology (SUT) is selected as the case study for this research.

1.2 Problem Statement

Generally, research on forecasting solar power generation uses real-world weather data as input for forecasting models. However, in practice, weather data often has quality limitations, such as missing data and measurement errors, which directly affect the performance of forecasting models. Furthermore, floating solar systems face seasonal fluctuations and curtailment limitations, which further complicate power generation forecasting. Therefore, it is necessary to develop forecasting approaches that can systematically manage incomplete weather data and reflect actual operating conditions to increase the accuracy, stability, and applicability of forecasting models. This reflects the fact that simply improving the model may not be enough without proper data quality management and a suitable forecasting structure.

1.3 Research Objectives

This The main objective of this research is to forecast solar power generation.

The specific objectives of this research include.

- 1) To apply linear interpolation, the Clear Sky Model and KNN are used for missing data management.
- 2) To develop and investigate the model for solar power forecasting, focusing on the use of the Two-Step Forecasting technique.

3) To compare and analyze the performance of LSTM, Bi-LSTM, GRU, RF and TFT-lite models in forecasting electricity from the Floating PV system at SUT.

1.4 Research hypothesis

1) Handling missing and anomalous data using the Clear Sky Model and KNN is hypothesized to enhance the forecasting performance by mitigating the negative impacts of incomplete or corrupted data, which may occur in floating solar power systems.

2) The application of the Two-Step Forecasting technique for solar power prediction is hypothesized to improve forecasting accuracy. In the first step, external influencing factors such as solar irradiance and temperature are forecasted. These predicted variables are then used in the second step to forecast the actual power generation.

1.5 Scope

1) The data used in this research is obtained from the floating solar power system at Suranaree University of Technology and is assumed to be accurate and reliable.

2) Missing or anomalous data will be preprocessed using methods such as Linear Interpolation, the Clear Sky Model, and KNN before being utilized in analysis.

3) The performance of the forecasting models will be evaluated using standard metrics, namely RMSE, MAE, MAPE, and R^2

4) The proposed forecasting approach will be compared with single-step forecasting methods and commonly used baseline models, including the Persistence model, LSTM, Bi-LSTM, GRU, RF, and TFT-lite

1.6 Research Benefits

This thesis is expected to provide the following academic and practical benefits:

1) To develop an accurate solar power forecasting model that can be effectively applied in renewable energy systems and smart grid applications.

2) To compare the performance of various forecasting models, providing guidelines for selecting appropriate models based on data characteristics and operational conditions.

3) To promote systematic handling of missing data, which improves data quality, enhances model reliability, and supports more accurate and stable energy management.

4) To propose a forecasting framework specifically for FPV systems, which can serve as a reference for model development and performance evaluation in both academic research and industrial applications.

5) To support energy planners, power system operators, and relevant organizations in decision-making processes related to power generation planning and energy management.

6) To contribute to the advancement of clean energy utilization and support Thailand's national Net Zero Emission policy through improved forecasting and management of solar power generation.

1.7 Structure of thesis

This thesis is organized as follows. Chapter II presents a comprehensive literature review related to solar power forecasting, machine learning and deep learning techniques, data imputation methods, and relevant methodologies applied in previous studies, with the aim of establishing a theoretical foundation and identifying existing research gaps. Chapter III describes the research methodology employed to develop and evaluate a data-driven forecasting model for solar power generation, focusing on data preparation procedures and missing data handling using linear interpolation. Chapter IV presents the short-term solar power forecasting results using a two-stage forecasting framework, emphasizing data refinement techniques aligned with real operating conditions of the system, including Clear Sky Model and KNN imputation, consideration of curtailment effects, time-segmented forecasting in three

intra-day periods, daily result aggregation, and performance evaluation supported by XAI analysis. Chapter V extends the analyses and findings from Chapters III and IV by presenting an integrated framework for day-ahead forecasting of floating photovoltaic systems and evaluating its overall performance. Finally, Chapter VI concludes the thesis by summarizing the key findings, discussing the main contributions and limitations of the study, and providing recommendations and directions for future research.



CHAPTER II

LITERATURE REVIEW

2.1 Introduction

In recent years, electricity generation from renewable energy particularly PV systems have expanded rapidly and plays a vital role in the modern power generation landscape. However, PV systems are inherently intermittent and highly variable due to external factors such as solar irradiance, cloud cover, humidity, temperature, and seasonal changes. These fluctuations directly impact the stability of the power grid and the efficiency of long-term energy management. As a result, accurate forecasting of electricity generation has become essential, especially within the framework of smart grids that rely on real-time data for effective energy distribution. Despite advancements, solar power forecasting continues to face challenges due to the non-linear nature of the data and uncertainties introduced by weather and seasonal variability. Over the past decade, ML and DL techniques have been widely adopted to improve forecasting performance. Among these, models under the RNNs family such as LSTM, Bi-LSTM, GRU have shown strong capabilities in capturing temporal dependencies, while traditional methods such as RF are also used for performance comparison. The effectiveness of these models depends heavily on proper data management, including handling missing values, selecting relevant features, and data grouping. Data imputation methods such as linear interpolation, KNN, and comparisons with clear-sky model outputs help improve data quality for training. Furthermore, error metrics such as MAE, MAPE, and RMSE are widely used to evaluate and compare model performance. Recent trends also highlight the growing importance of model interpretability, with XAI techniques particularly SHAP providing insights into the influence of input features on prediction results. This not only enhances transparency but also builds trust in model outputs for real-world applications. In this context, Chapter II presents a comprehensive literature review on solar power forecasting for

floating photovoltaic systems, structured into five key themes: (1) overview of related studies, (2) data management for forecasting, (3) model selection and development, (4) model evaluation, and (5) model explainability using XAI techniques. The aim is to identify current research gaps and provide a foundation for model development in subsequent chapters.

2.2 Literature Overview

To gain an understanding of the trends and key components in solar power generation forecasting, this study collects, analyzes, and compares relevant literature. The reviewed studies are categorized into two main groups based on their research approaches and development focuses:

(1) studies that focus on data management, such as missing data imputation, temporal data grouping, and feature engineering to improve the learning performance of forecasting models; and

(2) studies that concentrate on model development and result interpretation, such as the application of deep learning techniques for forecasting and the implementation of XAI tools to interpret model behaviors.

To support the design of an appropriate research methodology in the subsequent chapter, this study presents an overview of the key components and research gaps identified in existing literature. Table 2.1 focuses on data management aspects, while Table 2.2 summarizes the essential components and gaps in solar power forecasting studies, particularly in relation to model selection and development, performance evaluation, and explainability of model outputs.

Table 2.1 Overview of key data management aspects in solar power forecasting studies

Ref.	Missing Data	Data Grouping	Feature Engineering	Feature Selection
Roy et al. (2022)	-	-	/	Correlation Analysis
Zhou et al. (2019)	-	-	/	-
Jailani et al. (2023)	-	-	/	-
Gao et al. (2023)	-	-	/	-
Boucetta et al. (2023)	-	-	/	Correlation Analysis
Ait Mansour et al. (2024)	-	-	/	Correlation Analysis
Sharadga et al. (2020)	-	-	/	-
Khan et al. (2024)	-	-	/	-
Huynh & Nguyen (2024)	-	-	/	-
Kuzlu et al. (2020)	-	-	-	Feature Importance (Post-hoc XAI)
Chang et al. (2020)	-	TS-SOM	/	SHAP

Table 2.1 Overview of key data management aspects in solar power forecasting studies
(Continued)

Ref.	Missing Data	Data Grouping	Feature Engineering	Feature Selection
Sarp et al. (2021)	-	-	/	ELI5
Lu et al. (2021)	-	-	/	Feature Importance (Post-hoc XAI)
Mitrentsis & Lens (2022)	-	-	/	Permutation Importance
Ali et al. (2025)	-	K-Means	/	SHAP
Chen et al. (2019)	-	K-Means	/	-
Zhou et al. (2018)	-	K-Means	/	-
Wang et al. (2024)	-	K-Means	/	Correlation Analysis
Sedliarov & Klen (2024)	-	K-Means	-	-
Junhua thon & Chayakulkheeree (2021)	-	-	/	-
Junhua thon & Chayakulkheeree (2022)	-	SOM	/	SOM

Table 2.1 Overview of key data management aspects in solar power forecasting studies
(Continued)

Ref.	Missing Data	Data Grouping	Feature Engineering	Feature Selection
Lee & Son (2024)	KNN, Gain, Linear Interpolation	-	/	-

Based on the analysis of previous studies summarized in Table 2.1, it was found that most research on solar power generation forecasting emphasizes the process of feature engineering to enhance model performance. These studies mainly focus on transforming and designing features that are aligned with the data context. However, the issue of missing data handling has not received adequate attention.

Although solar energy datasets often contain missing values due to technical issues such as sensor failures or data transmission errors, only a limited number of studies such as (Bianchini et al., 2019) have presented approaches for developing predictive models of solar power generation under data constraints. These models utilize a clear-sky solar radiation physical model as a reference signal to estimate environmental variables when measurement data is limited or incomplete. (Lee & Son, 2024) have addressed techniques to handle missing data. These techniques include KNN, information gain, and linear interpolation for data imputation. Nevertheless, most existing research lacks a systematic approach to address this issue, which may compromise the accuracy and reliability of the forecasting models.

In terms of feature selection, most studies continue to rely on linear relationship analyses, such as correlation analysis. Only a few have begun to apply model explainability techniques such as SHAP, ELI5, and Permutation Importance to understand the contribution of each feature to the model's output. However, these techniques have been applied independently of the feature selection step and have not yet been widely adopted.

In addition, certain studies for instance, (Junhua Thon & Chayakulkheeree, 2022) and (Chen et al., 2019) have employed data grouping techniques, such as K-Means and SOM, to analyze preliminary data characteristics.

Synthesizing the key points above, the critical research gaps in prior studies lie in the absence of a systematic and efficient missing data handling process and the lack of integration between data grouping and model training to enable models to learn specific patterns from each data cluster. Therefore, this research aims to develop a robust data preprocessing framework prior to model development, focusing on managing missing data using techniques such as linear interpolation, KNN imputation, and the clear sky model to restore data completeness. Additionally, K-means clustering will be employed to group temporal data, allowing for model training tailored to the characteristics of each group.

Table 2.2 Overview of modeling and interpretation approaches in solar power forecasting studies

Ref.	Model	Evaluation Metrics			Explainability (XAI) Techniques
		RMSE	MAE	MAPE	
Roy et al. (2022)	LSTM, RF	/	/	/	-
Zhou et al. (2019)	LSTM	/	/	/	-
Jailani et al. (2023)	LSTM, CNN, CNN-LSTM	/	/	/	-
Gao et al. (2023)	LSTM, GRU	/	/	/	-

Table 2.2 Overview of modeling and interpretation approaches in solar power forecasting studies (Continued)

Ref.	Model	Evaluation Metrics			Explainability (XAI) Techniques
		RMSE	MAE	MAPE	
Boucetta et al. (2023)	LSTM, GRU and MLP	/	/	-	-
Ait Mansour et al. (2024)	Bi-LSTM, GRU and 1D-CNN	/	/	-	-
Sharadga et al. (2020)	LSTM, Bi- LSTM, GRU	/	/	/	-
Khan et al. (2024)	LSTM, Bi- LSTM, GRU, CNN-LSTM, ARIMA	/	/	/	-
Huynh & Nguyen (2024)	LSTM, GRU	-	/	-	-
Kuzlu et al. (2020)	RF	/	-	-	SHAP, LIME, ELI5
Chang et al. (2020)	XGBoost, SVR, GRNN, LSTM	/	/	-	SHAP
Sarp et al. (2021)	XGBoost	/	-	-	ELI5
Lu et al. (2021)	NN, BNN	/	-	-	SHAP, IG, EG, DeepLIFT
Mitrentsis & Lens (2022)	NGBoost	/	/	-	SHAP

Table 2.2 Overview of modeling and interpretation approaches in solar power forecasting studies (Continued)

Ref.	Model	Evaluation Metrics			Explainability (XAI) Techniques
		RMSE	MAE	MAPE	
Ali et al. (2025)	XGBoost, RF, SVR	/	/	-	SHAP
Chen et al. (2019)	SNN, BP, SVM	/	-	/	-
Zhou et al. (2018)	Stacking-SVM	/	/	/	-
Wang et al. (2024)	GRU	-	/	-	-
Sedliarov & Klen (2024)	SARIMA, LSTM, Linear Regression, KNN, KNN+K- Means	/	/	/	-
Junhua thon & Chayakulkheeree (2021)	FNN, LSTM, NARX	/	/	/	-
Junhuathon & Chayakulkheeree (2022)	FNN, LSTM, FNN-SOM, LSTM-SOM	/	/	/	-
Lee & Son (2024)	CNN-GRU	/	/	-	-

Based on the analysis of the studies presented in Table 2.2, it is evident that most research related to solar power generation forecasting focuses on the development and comparison of ML and DL models. Widely adopted models include LSTM, GRU, Bi-LSTM, and RF, reflecting the potential of sequential models in handling

time-series data, as well as the effectiveness of non-sequential models such as RF. The popularity of RF, especially in contexts involving large-scale and structurally simple datasets, is evident in studies such as (Roy et al., 2022), (Kuzlu et al., 2020), and (Ali et al., 2025). Most research evaluates model performance using standard accuracy metrics, including RMSE, MAE, and MAPE.

However, the aspect of model explainability or the application of XAI techniques remains underexplored in the existing body of work. Only a few studies have applied interpretability techniques such as SHAP, LIME, ELI5, IG, EG, and DeepLIFT to tree-based models (e.g., RF, XGBoost) or neural network models to interpret model decisions. Most of these efforts have focused on post-hoc explainability rather than systematically supporting the feature selection process. This is despite the fact that many XAI methods can effectively reduce bias and enhance model transparency. Although previous research has achieved high predictive accuracy, the lack of clarity in model interpretation remains a significant limitation particularly in energy systems, where reliability, interpretability, and traceability are critical at every stage.

Therefore, this study proposes the integration of XAI techniques to enhance not only the accuracy and transparency but also the interpretability of solar power forecasting models in practical applications.

2.3 Review of Data Management in Solar Power Forecasting

Data management is a critical step in the development of solar energy forecasting systems, as high-quality and complete data significantly enhance the accuracy and reliability of prediction models. In relevant studies, data management typically encompasses several key aspects, including missing data handling, data cleaning, feature selection and engineering, and data grouping based on specific characteristics, as outlined below.

2.3.1 Handling Missing Data and Cleaning

Solar energy datasets frequently encounter issues of missing or incomplete data due to various causes such as sensor malfunctions or recording errors. Addressing missing values is therefore essential to ensure data quality. Common

imputation methods include mean, median, KNN, and the Clear Sky Model, which estimates reference irradiance based on geographic location, time, and season to fill in missing values or identify anomalies.

For detecting and removing outliers such as negative or unrealistically high-power value statistical techniques are typically employed, including Z-score and IQR methods. High-quality data preprocessing, including imputation and outlier removal, reduces model bias and significantly enhances forecasting accuracy.

(Lee & Son, 2024) compared missing data imputation methods, including KNN, GAIN, and linear interpolation, on photovoltaic power datasets using a CNN-GRU forecasting model. The results demonstrated that KNN and GAIN outperformed the linear interpolation method, evaluated by metrics such as RMSE, MAE, MRD, RMSD, and R^2 . This was especially evident when data loss exceeded 10%, highlighting the critical impact of appropriate imputation methods on forecasting accuracy.

2.3.2 Feature Selection and Engineering

The selection of important and appropriate features directly affects the performance of forecasting models. Several studies have applied variable importance analysis techniques, such as correlation analysis and SHAP-based methods, to identify features that significantly impact model outcomes. Moreover, creating new features from existing data (feature engineering), such as adding environmental features and time-based features, can significantly enhance model performance.

From the review in Table 2.1, many studies still emphasize the use of correlation analysis to analyze the linear relationship between input variables and outcomes. For instance, (Roy et al., 2022) employed LSTM and RF for day-ahead forecasting using meteorological data and historical power generation data. Although correlation analysis was conducted to select variables, there was no mention of feature creation or handling missing data.

In contrast, (Zhou et al., 2019) integrated an attention mechanism with an LSTM model to capture deep temporal characteristics, which can be considered an advanced form of feature engineering. However, explicit feature selection was not discussed.

(Gao et al., 2023) applied LSTM and GRU models with parameter tuning via Gaussian Process and designed a diverse input dataset. While no direct feature creation was mentioned, their approach reflects implicit feature importance analysis.

The study by (Ait Mansour et al., 2024) utilized 5-minute interval data from a photovoltaic system in Morocco and applied lag and smoothing techniques to capture temporal characteristics for new feature construction. However, feature selection was not explicitly addressed.

(Sharadga et al., 2020) compared statistical models with neural network models, emphasizing feature construction from historical data, such as lag features and time-of-day features. Nevertheless, systematic feature selection was not mentioned.

(Khan et al., 2024) focused on designing model architectures suitable for renewable energy data by employing a correlation matrix to select important features from meteorological and temporal data, along with normalization to stabilize learning. Their findings showed that feature selection and joint parameter tuning significantly improved model accuracy.

Similarly, (Huynh and Nguyen, 2024) used a correlation heatmap to analyze variable relationships and found high correlations between IRRADIATION and MODULE TEMPERATURE with DC POWER and AC POWER. However, their feature selection approach lacked a systematic framework.

(Sarp et al., 2021) employed an XGBoost model combined with the XAI technique ELI5 to analyze feature impacts. They experimented by removing some variables from the dataset to observe changes in outcomes, serving as a clear example of systematic feature selection.

(Chang et al., 2020) used SHAP to analyze feature importance in the context of edge computing. Although feature selection was not explicitly mentioned, the use of SHAP enabled clearer interpretation of feature importance rankings.

(Lu et al., 2021) applied XAI methods such as Integrated Gradients, Expected Gradients, and DeepLIFT to explain feature impacts at both global and local levels, offering deeper insight into data characteristics. The focus was more on feature analysis than direct selection or reduction.

(Mitrentsis and Lens, 2022) proposed an NGBoost model with strength in probabilistic forecasting using lag features and weather data. SHAP was used to explain the model, while feature importance was evaluated using permutation importance. Although missing data handling was not clearly stated, performance metrics such as CRPS, RMSE, and MAE indicated the model's robustness and reliability.

(Ali et al., 2025) presented the Explainable Clustered Federated Learning (XCFL) approach, combining Pearson Correlation for feature selection with clustering prior to decentralized learning. Models including XGBoost, RF, and SVR were applied to data from 21 stations in Germany. This reflects the integration of both feature selection and clustering to improve real-world model performance.

In the study by (Wang et al., 2024), K-Means++ was used to cluster weather data into three main groups, and key features were selected using Pearson Correlation. The GRU model was then applied in combination with the "similar day" technique an approach to feature engineering that leverages deep domain knowledge to improve forecasting performance.

In Thailand, (Junhuathon & Chayakulkheeree, 2021) investigated short-term photovoltaic power forecasting by comparing FNN, LSTM, and NARX models on the same dataset. Normalization was included as part of the feature creation process to enhance model performance.

In their subsequent study, (Junhuathon & Chayakulkheeree, 2022) introduced the integration of SOM with LSTM and FNN models to improve forecasting accuracy. SOM was used to cluster data prior to model training, representing an application of both feature engineering and feature selection via clustering. Their data preprocessing included removing periods without solar radiation and scaling the data, effectively mitigating non-linearity and seasonal effects.

Overall, many studies have demonstrated a growing emphasis on feature construction and analysis through various techniques, both explicitly and implicitly. While some research still lacks a systematic feature selection process, the combination of well-designed feature selection and feature engineering remains critical in developing more accurate and effective forecasting models.

2.3.3 Data Grouping

Data grouping is a crucial process in improving the efficiency of forecasting models. It focuses on dividing data into subgroups with similar characteristics such as climate patterns, seasons, or installation locations to reduce data heterogeneity and enable the model to effectively learn specific patterns within each group. Commonly used techniques include K-means, hierarchical clustering, and SOM.

(Chang et al., 2020) proposed a short-term solar power forecasting framework suitable for edge computing systems by applying clustering techniques in combination with tree-based algorithms such as XGBoost and Ridge Regression to enhance the learning efficiency of the model.

(Ali et al., 2025) introduced the XCFL method, which combines data grouping based on key features (selected using Pearson correlation) with federated learning in solar energy forecasting systems. They employed XGBoost, Random Forest, and SVR models on data from 21 solar power stations in Germany. The division into 3–5 clusters improved both prediction accuracy and model interpretability using SHAP. However, the study did not explicitly compare other clustering techniques.

(Chen et al., 2019) applied K-means clustering to group data based on electricity production characteristics and meteorological factors such as solar radiation, temperature, and humidity reflecting seasonal patterns. They then trained SNN separately for each group and employed PSO for parameter tuning to enhance accuracy and model stability. Results indicated that this method outperformed BP-NN and SVM. However, the study did not address missing data handling, feature selection, or result explanation through XAI.

(Zhou et al., 2018) presented a short-term solar power forecasting model using Stacking-SVM, where training data were grouped using K-Means clustering. The meta-learner was trained on the same group of data as the target forecast, which enhanced accuracy. The proposed model outperformed standalone SVM, evaluated using RMSE, MAE, and MAPE.

(Wang et al., 2024) utilized K-Means++ clustering to divide data into three major weather groups: clear, cloudy, and rainy/snowy. This approach, combined

with an enhanced similar-day technique and Pearson correlation-based feature selection, significantly improved the performance of GRU-based power forecasting.

In addition, (Sedliarov & Klen, 2024) implemented data clustering using K-Means alongside a KNN model. Data were grouped before being input into the model. The combined KNN+K-Means model showed higher performance compared to traditional methods such as SARIMA, LSTM, and linear regression, particularly for medium- to long-term forecasts (e.g., 3, 7, and 70 days ahead).

(Junhuathon & Chayakulkheeree, 2022) employed Self-Organizing Maps (SOM), an unsupervised clustering technique, prior to inputting data into LSTM and FNN models. The experimental results showed that SOM significantly reduced RMSE, MAE, and MAPE, especially in datasets exhibiting strong seasonal variability.

In summary, data grouping prior to the forecasting process whether through K-Means or SOM plays a key role in distinguishing datasets with similar characteristics, mitigating the impact of data fluctuations, and significantly enhancing model accuracy. This approach is thus highly suitable for adoption in the development of solar energy forecasting models, as proposed in this research.

2.4 Review of Forecasting Models

The selection of an appropriate forecasting model is critically important for the accuracy and performance of solar energy prediction systems. These models can be broadly categorized into three main groups: statistical models, machine learning models, and deep learning models. Each group possesses distinct advantages, limitations, and application contexts, as outlined below.

2.4.1 Data Grouping Statistical Machine Learning Methods

Statistical models such as ARIMA and linear regression offer the advantage of easy interpretability; however, they face limitations in handling nonlinear data. On the other hand, machine learning models like Random Forest RF and SVR demonstrate superior capability in managing complex and nonlinear datasets. Consequently, these models have been widely adopted for solar energy forecasting to enhance prediction accuracy.

Recent research in 2024 has compared statistical models with deep learning approaches. For instance, (Khan et al., 2024) studied and developed deep neural network (DNN) architectures for forecasting renewable energy generation, particularly solar and wind power. Their comparison among LSTM, Bi-LSTM, GRU, and CNN-LSTM models against ARIMA revealed that LSTM with hyperparameters optimized via RandomizedSearchCV achieved the best performance across multiple metrics, including MAE, MSE, RMSE, R^2 , and MAPE.

Furthermore, (Sedliarov and Klen, 2024) proposed integrating K-Nearest Neighbor (KNN) with K-Means clustering to significantly improve solar power output forecasting accuracy. Their KNN+K-Means model outperformed SARIMA, LSTM, and linear regression, especially for three-day-ahead forecasts, while maintaining strong performance for medium- and long-term horizons (7 and 70 days). This algorithm is therefore well-suited for implementation in solar power station control and energy management systems.

Regarding machine learning models, (Zhou et al., 2018) introduced a short-term solar power forecasting model using Stacking Support Vector Machine (Stacking-SVM), which combines multiple SVM models to improve accuracy. They applied K-means clustering for training data grouping and trained a meta-learner using data from the same cluster as the forecasted samples. Tests on real data from a 20-kW solar plant demonstrated that this approach outperformed single SVM models under both normal and varying meteorological conditions. In the following year, (Chen et al., 2019) proposed a short-term solar power forecasting method based on SNN combined with K-means clustering. Separate SNN models were trained per cluster, and PSO was employed to enhance model performance. Experimental results indicated superior accuracy compared to baseline methods such as BP-NN and SVM.

(Kuzlu et al., 2020) explored the use of Random Forest for solar energy forecasting complemented with XAI techniques, including SHAP, LIME, and ELI5, to facilitate interpretable and transparent model outputs. Their study utilized the well-known GEFCOM 2014 dataset and highlighted the importance of integrating XAI with machine learning models to enhance practical applicability.

In 2021, (Sarp et al., 2021) applied the boosting technique XGBoost for solar power forecasting and used ELI5 to explain feature importance. Their findings showed that appropriate feature selection effectively reduced RMSE while improving model interpretability, addressing the practical needs of real-world deployment. Similarly, (Lu et al., 2021) proposed a Bayesian Neural Network model for solar energy forecasting, incorporating XAI methods such as Integrated Gradients, Expected Gradients, and DeepLIFT to explain feature significance at both global and local levels. This work emphasized handling uncertainty and analyzing critical time intervals, making it highly relevant for weather-dependent energy data.

(Mitrentsis & Lens, 2022) introduced a short-term solar forecasting approach based on NGBoost, capable of generating probabilistic forecasts and quantifying prediction uncertainty. They further utilized SHAP and Permutation Importance for explaining key features and created lag features to incorporate temporal information. This research demonstrated an effective strategy for producing reliable and interpretable machine learning-based forecasts.

Most recently, (Ali et al., 2025) presented a XCFL framework for solar energy forecasting, combining data clustering with federated learning. Feature selection was conducted using Pearson correlation, and XAI techniques were employed to enhance model result interpretability. This study represents a cutting-edge approach that addresses future demands in intelligent energy systems.

2.4.2 Deep Learning Models

Deep learning models such as LSTM, GRU, and Bi-LSTM have demonstrated outstanding capabilities in handling time series data and nonlinear relationships. These models effectively learn long-term dependencies and complex patterns within data. Several studies indicate that GRU offers comparable performance to LSTM but with a simpler architecture and faster training speed. Meanwhile, Bi-LSTM processes data bidirectionally, making it suitable for datasets exhibiting symmetric behavioral patterns. Additionally, hybrid models like CNN-LSTM or CNN-GRU have gained popularity in solar energy forecasting due to their ability to combine the feature extraction strength of CNN with the sequence modeling capacity of RNN.

(Zhou et al., 2019) proposed a short-term solar energy forecasting approach using an LSTM model integrated with an attention mechanism to enhance temporal sequence learning. Experimental results showed that the LSTM model with attention outperformed traditional methods in accuracy, highlighting the potential of hybrid techniques for forecasting complex sequential data.

(Sharadga et al., 2020) compared statistical models such as ARMA, ARIMA, and SARIMA with neural network-based models including LSTM, Bi-LSTM, and GRU for hourly solar power output forecasting without additional meteorological inputs. Their findings demonstrated that deep learning models surpassed statistical approaches in terms of RMSE, MAE, and MAPE, confirming the suitability of deep architectures under limited data conditions.

(Junhuathon, N., & Chayakulkheeree, K., 2021) conducted a comparative study on short-term PV power forecasting using three model types: FNN, LSTM, and Nonlinear Autoregressive Network with Exogenous Inputs (NARX), employing MATLAB software. The dataset consisted of 8,760 samples, including time, solar panel temperature, solar irradiance, and power output. Results indicated that LSTM achieved the highest accuracy with an RMSE of 680 kW, an MAE of 429 kW, and an MAPE of 30.06%, followed by FNN and NARX, respectively. Notably, NARX performed better at high power levels due to its capacity to approximate peak values more accurately. Subsequently, (Junhuathon & Chayakulkheeree, 2022) introduced a hybrid approach combining deep learning with SOM clustering, comparing FNN, LSTM, FNN-SOM, and LSTM-SOM on real hourly data for an entire year. Their results demonstrated that models incorporating SOM significantly improved forecasting accuracy by reducing RMSE, MAE, and MAPE.

(Jailani et al., 2023) analyzed the potential of LSTM compared to hybrid models such as CNN-LSTM and LSTM-CNN for solar energy forecasting. Their findings revealed that hybrid models delivered higher accuracy but required longer training times. Although deep learning models show great promise for solar power forecasting, selecting an appropriate model should be aligned with the specific objectives of each case study.

(Gao et al., 2023) examined day-ahead forecasting by comparing LSTM and GRU models, integrating Gaussian Process for parameter optimization. The study found that LSTM performed well with limited data, whereas GRU was better suited for diverse historical datasets, particularly when incorporating meteorological variables.

(Boucetta et al., 2023) conducted a comparative study of LSTM, GRU, and MLP models for short-term solar power forecasting in Adrar, Algeria. The results indicated that LSTM consistently outperformed other models across various timeframes and seasons, achieving the lowest RMSE and MAE, thus demonstrating its effectiveness in handling sequential data.

(Ait Mansour et al., 2024) investigated Bi-LSTM, GRU, and 1D-CNN models for forecasting photovoltaic panel performance in semi-arid desert regions. The study showed that Bi-LSTM and GRU delivered comparable performance and outperformed 1D-CNN, especially in daytime forecasting, highlighting the suitability of bidirectional RNNs in challenging environments.

(Lee & Son, 2024) focused on methods for imputing missing data in solar energy datasets by comparing KNN, GAIN, and Linear Interpolation before applying a CNN-GRU model. Their results demonstrated that the choice of imputation method significantly affects forecasting accuracy, particularly when data gaps exceed 10%.

2.4.3 Comparative Model Studies

Based on the literature review, the comparative studies of forecasting models can be categorized into two main approaches as follows. Several studies have compared the performance of deep learning models with traditional models such as ARIMA, SVM, and random forest. These studies found that deep learning models are more capable of capturing complex and nonlinear data patterns.

(Khan et al., 2024) proposed a method for developing and restructuring DNNs to forecast renewable energy production, focusing on solar and wind power. Their comparative study of advanced models, including LSTM, Bi-LSTM, GRU, and CNN-LSTM against ARIMA, demonstrated that the LSTM model optimized by RandomizedSearchCV outperformed the other models.

(Sedliarov & Klen, 2024) presented the use of KNN combined with K-Means clustering to improve the accuracy of solar power forecasting. Experimental

results on real data showed that the KNN+K-Means model yielded lower prediction errors compared to conventional forecasting systems such as SARIMA, LSTM, and linear regression, particularly for 3-day-ahead forecasts. The model also exhibited strong performance for medium- and long-term forecasts (7 days and 70 days), making this algorithm suitable for energy management and control systems at solar power plants.

Regarding studies comparing deep learning and machine learning models, (Roy et al., 2022) introduced a day-ahead solar power forecasting method using LSTM and RF models tailored for the northeastern region of India. The input features included meteorological data such as temperature, humidity, and solar radiation, along with historical power generation data. Their results indicated that LSTM provided superior forecasting accuracy compared to RF in most cases.

The review of solar power generation forecasting models reveals that although statistical models are straightforward and interpretable, machine learning and deep learning models demonstrate superior performance in handling nonlinear and complex data. Particularly, models such as LSTM, GRU, Bi-LSTM, and hybrid models that combine multiple techniques can significantly enhance forecasting accuracy. Furthermore, clustering techniques combined with KNN+K-Means present a promising approach for efficient solar power prediction.

2.5 Model Evaluation and Explainability

The evaluation and performance measurement of forecasting models are critical steps in model development, as they provide insight into the model's accuracy and stability. Moreover, these steps facilitate the comparison of performance across multiple models. Therefore, selecting appropriate criteria and evaluation methods is essential in research related to solar energy forecasting and various time series data.

2.5.1 Deep Learning Models

Currently, the development of forecasting models using Deep Learning techniques has achieved high accuracy but lacks the ability to clearly explain the internal decision-making processes of the models. Consequently, the concept of "XAI" has emerged, focusing on the development of tools or techniques that can interpret

the model's workings in an understandable manner without compromising model performance.

One of the widely adopted and standard techniques for model explanation is SHAP, which is developed based on the concept from Cooperative Game Theory. It relies on Shapley values to calculate the impact of each feature on the model's output. SHAP can explain both at the global level (overall model trends) and the local level (individual predictions). The working principle of SHAP involves assessing how much each feature contributes to the change in the prediction value by calculating the average of changes caused by adding or removing features from all possible subsets, thereby producing fair and interpretable values. Notably, (Lundberg & Lee, 2017) introduced the SHAP method, which utilizes game theory principles to explain complex model outcomes by fairly and accurately measuring the influence of each feature on predictions at both global and local levels.

In 2020, (Chang et al., 2020) proposed a short-term solar energy forecasting framework suitable for edge computing systems by combining data clustering with TS-SOM and XGBoost models. They applied SHAP to explain feature importance, enhancing model transparency and supporting practical deployment. In the same year, (Kuzlu et al., 2020) studied solar photovoltaic power forecasting using Random Forest alongside XAI tools such as SHAP, LIME, and ELI5 to analyze feature importance. Their experiments demonstrated that XAI improves understanding of the model's functioning and enhances accuracy, particularly in the context of smart grid systems that require high transparency and reliability.

The following year, (Sarp et al., 2021) presented solar power production forecasting using the XGBoost model and employed ELI5 to explain influential features. Their experimental results, comparing datasets with and without certain features, showed that the complete dataset yielded better performance and provided clearer explanation of results, thereby improving model interpretability. In the work of (Lu et al., 2021), classification and regression models for solar energy forecasting were developed with XAI techniques including SHAP, Integrated Gradients, Expected Gradients, and DeepLIFT for feature importance analysis. They found that load and solar energy values from preceding hours had the greatest impact. This study also

incorporated uncertainty analysis using Bayesian neural networks to predict time intervals with a higher risk of forecasting errors.

(Mitrentsis & Lens, 2022) proposed a short-term solar energy forecasting model using NGBoost, which is capable of uncertainty quantification. They utilized SHAP for feature importance analysis together with permutation importance to evaluate each feature's influence, employing various performance metrics such as CRPS, RMSE, and MAE.

(Ali et al., 2025) introduced the concept of XCFL, which integrates feature clustering based on Pearson correlation with decentralized Federated Learning. Using XGBoost, RF, and SVR models along with SHAP, they explained each model's outcomes, enabling clearer understanding and interpretation, while also improving forecasting accuracy.

The related research details are summarized in Table 2.3.

Table 2.3 Overview of XAI applications in solar power forecasting studies

Ref.	AI model	XAI approach	Scope
Chang et al. (2020)	XGBoost, SVR, GRNN, LSTM	Feature attribution from weight, gain and cover & SHAP	Global & local
Kuzlu et al. (2020)	RF	SHAP, LIME, ELI5	Global & local
Sarp et al. (2021)	XGBoost	ELI5	Global
Lu et al. (2021)	NN, BNN	SHAP, Integrated and Expected Gradients & DeepLIFT	Global & local
Mitrentsis & Lens (2022)	NGBoost	SHAP	Global & local
Ali et al. (2025)	XGBoost, RF, SVR	SHAP	Global & local

2.6 Research Gaps

Table 2.4 illustrates the research gaps between the proposed study and the existing literature based on the aforementioned literature survey

Table 2.4 Overview of key components and research gaps in solar power forecasting studies

Ref.	Data Management	Stage 1	Stage 2	Evaluation Metrics			XAI
	Missing Data			RMSE	MAE	MAPE	
Roy et al. (2022)	×	×	LSTM	✓	✓	✓	×
Zhou et al. (2019)	×	×	LSTM	✓	✓	✓	×
Jailani et al. (2023)	×	×	LSTM, CNN, CNN-LSTM	✓	✓	✓	×
Gao et al. (2023)	×	×	LSTM, GRU	✓	✓	✓	×
Boucetta et al. (2023)	×	×	LSTM	✓	✓	×	×
Ait Mansour et al. (2024)	×	×	Bi-LSTM, GRU	✓	✓	×	×
Sharadga et al. (2020)	×	×	LSTM, Bi-LSTM, GRU	✓	✓	✓	×
Khan et al. (2024)	×	×	LSTM	✓	✓	✓	×

Table 2.4 Overview of key components and research gaps in solar power forecasting studies (Continued)

Ref.	Data	Stage 1	Stage 2	Evaluation Metrics			XAI
	Management						
	Missing Data			RMSE	MAE	MAPE	
Huynh & Nguyen (2024)	×	×	GRU	×	✓	×	×
Kuzlu et al. (2020)	×	×	RF	✓	×	×	✓
Chang et al. (2020)	×	×	XGBoost	✓	✓	×	✓
Sarp et al. (2021)	×	×	XGBoost	✓	×	×	✓
Lu et al. (2021)	×	×	NN, BNN	✓	×	×	✓
Mitrentsis & Lens (2022)	×	×	NGBoost	✓	✓	×	✓
Ali et al. (2025)	×	×	XGBoost, RF, SVR	✓	✓	×	✓
Chen et al. (2019)	×	×	SNN	✓	×	✓	×
Zhou et al. (2018)	×	×	Stacking-SVM	✓	✓	✓	×
Wang et al. (2024)	×	×	GRU	×	✓	×	×
Sedliarov & Klen (2024)	×	×	KNN+ K-Means	✓	✓	✓	×

Table 2.4 Overview of key components and research gaps in solar power forecasting studies (Continued)

Ref.	Data Management	Stage 1	Stage 2	Evaluation Metrics			XAI
	Missing Data			RMSE	MAE	MAPE	
Junhua thon & Chayakul kheeree (2021)	×	×	LSTM	✓	✓	✓	×
Junhua thon & Chayakul kheeree (2022)	×	×	FNN-SOM	✓	✓	✓	×
Lee & Son (2024)	✓	×	CNN-GRU	✓	✓	×	×
Proposed Framework	✓	✓	LSTM, Bi-LSTM, GRU, RF, TFT-lite	✓	✓	✓	✓

CHAPTER III

TWO-STAGE DATA-DRIVEN MODELING FOR DAY-AHEAD SOLAR POWER FORECASTING: A CASE STUDY OF THE SUT CAMPUS FLOATING SOLAR SYSTEM¹

3.1 Introduction

This section presents the research methodology employed in the development and evaluation of short-term (day-ahead) solar power forecasting models for a FPV system located within the Suranaree University of Technology campus. The overall workflow is divided into four main stages: data preparation, two-step forecasting framework, model development, evaluation metric, and conclusion. An overview of the research framework is illustrated in Figure 3.1.

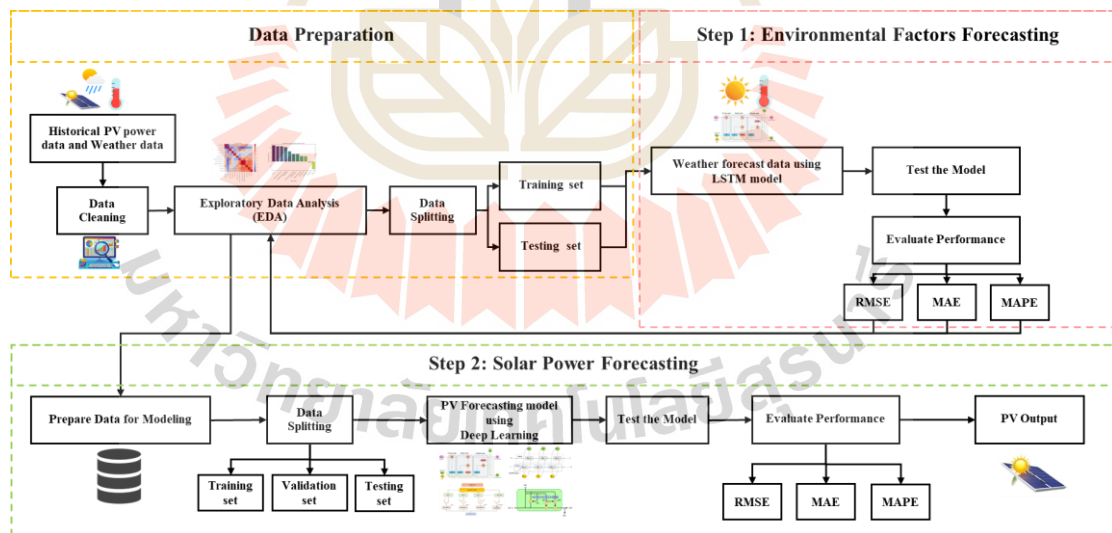


Figure 3.1 The framework for solar power forecasting

¹ Part of this chapter had been presented at the 13th International Electrical Engineering Congress (IEEECON 2025), “Carbon Neutrality: Challenges and Solution Based on Sustainable Power of Nature”, Hua Hin, Thailand, March 5–7, 2025.

3.2 Data preparation

The data management process consists of data collection, data cleaning, feature selection and correlation analysis, and data splitting, as illustrated in Figure 3.2.

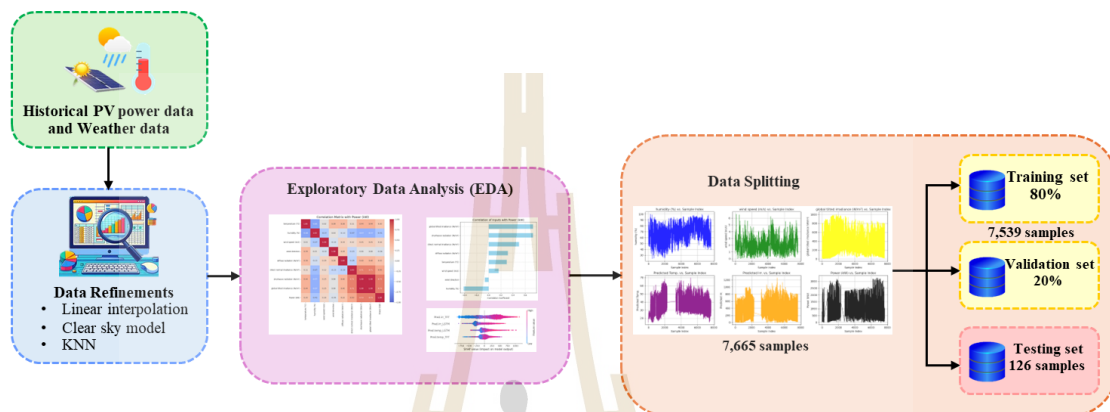


Figure 3.2 The framework for data management process

3.2.1 Data collection

The dataset used in this study is from a floating PV power plant located at 111 University Road, Suranaree Sub-district, Mueang Nakhon Ratchasima District, Nakhon Ratchasima Province, Thailand. The installed capacity of the floating PV power plant is 4.27 MW/unit, and the power output data is recorded during the date range from 01/01/2023 to 12/31/2023. During this date range, the PV power generation is consistently zero between 17:00 and 07:00. Therefore, we only consider the power between 07:00 and 17:00 in the simulation, and the power data is recorded every 30 minutes, as shown in Figure 3.3



Figure 3.3 Floating PV power plant at SUT

The datasets used in this study are divided into two main groups: weather data and solar cell power generation data. The weather data consists of variables such as temperature, relative humidity, wind speed, GHI, and GTI, which can be accessed via the website <https://open-meteo.com/en/docs/historical-weather-api>. For the solar cell power generation data, from the system installed with 16 inverters and Jinko Solar solar panels model JKM540M-72HL4-V, a total of 7,920 panels, the data can be accessed via the website <https://app-portal.exep.io/login>. Figures 3.4 and 3.5 show the websites of both sets, respectively. The data of the solar cell data sheet are presented in Table 3.1, with the corresponding figures provided in Appendix A.

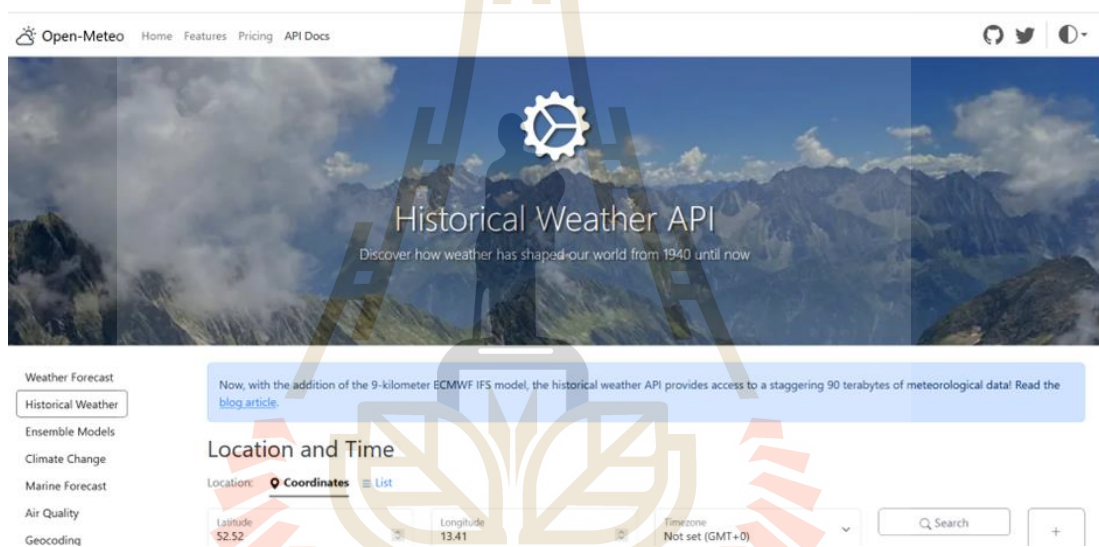


Figure 3.4 Weather data source interface (Open-meteo API)

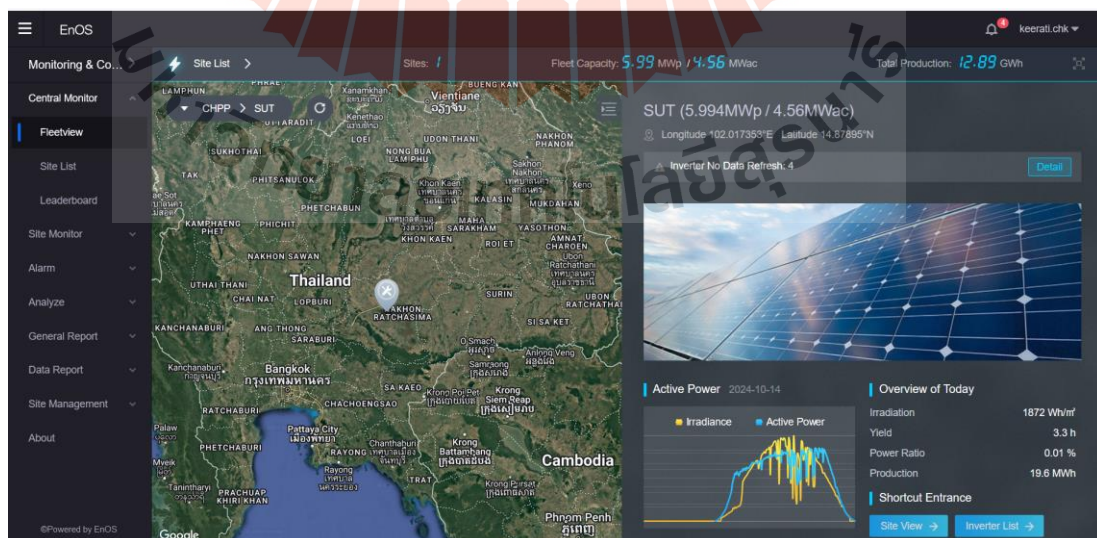


Figure 3.5 Solar power output data interface (Exep platform)

Table 3.1 Main specifications of Jinko Solar module (JKM540M-72HL4-V)

Item	Value
Module Type	JKM540M-72HL4-V
Maximum Power (P_{max} , STC)	540 Wp
Number of Modules	7,920
Dimensions	2230 × 1134 × 35 mm (87.80 × 44.65 × 1.38 in)
Module Area	2.53 m ²
Total Installed Capacity (P_{total})	4,276.8 kWp

3.2.2 Data cleaning

The data cleaning process is a crucial step in preparing data for analysis or model training. This involves detecting and handling missing and anomalous values using techniques such as linear interpolation, the clear sky model, and KNN imputation.

After downloading the weather data and solar power output data from online sources, the two datasets were merged into a single dataset. Only the time period between 07:00 and 17:00 was retained, as this is when solar energy production typically occurs. Figures 3.6 and 3.7 present examples of the weather and power data, respectively.

For the power data, the outputs from all 16 inverters were aggregated into a single column to facilitate subsequent analysis and processing. The result of this data preparation step is illustrated in Figure 3.8

date	month	time	timestep	temperature (°C)	humidity (%)	wind speed (m/s)	wind direction	diffuse radiation (W/m ²)	direct normal irradiance (W/m ²)	shortwave radiation (W/m ²)	global tilted irradiance (W/m ²)
1/1/2023	1	7:00:00	1	15	92	0.78	50	10	122.1	19	19
1/1/2023	1	7:30:00	2	15.6	88	0.92	41	45	419.8	111	111
1/1/2023	1	8:00:00	3	16.6	83	1.28	39	80	548.7	226	226
1/1/2023	1	8:30:00	4	17.6	77	1.98	49	104	610.3	329	329
1/1/2023	1	9:00:00	5	18.6	72	2.72	54	138	613.1	422	422
1/1/2023	1	9:30:00	6	19.5	68	3.02	56	184	563.7	493	493
1/1/2023	1	10:00:00	7	20.2	65	3.24	56	251	469.4	543	543
1/1/2023	1	10:30:00	8	21	63	3.38	56	327	359.9	573	573
1/1/2023	1	11:00:00	9	21.6	61	3.44	54	356	339	604	604
1/1/2023	1	11:30:00	10	22.4	59	3.5	53	341	389.3	639	639
1/1/2023	1	12:00:00	11	23.2	56	3.48	51	281	546.8	710	710
1/1/2023	1	12:30:00	12	24.2	53	3.34	51	195	757.1	792	792
1/1/2023	1	13:00:00	13	25.1	51	3.2	51	152	842.7	807	807
1/1/2023	1	13:30:00	14	25.7	49	3.26	50	134	856.2	777	777
1/1/2023	1	14:00:00	15	26	48	3.41	50	116	864.6	730	730
1/1/2023	1	14:30:00	16	26.2	47	3.48	51	103	854.4	663	663
1/1/2023	1	15:00:00	17	26.4	46	3.48	51	100	816.6	580	580
1/1/2023	1	15:30:00	18	26.3	46	3.48	51	98	757.2	483	483
1/1/2023	1	16:00:00	19	26.1	47	3.48	51	88	697.5	380	380
1/1/2023	1	16:30:00	20	25.8	48	3.48	51	72	618.8	270	270
1/1/2023	1	17:00:00	21	25.2	49	3.26	50	51	462.4	150	150

Figure. 3.6 Sample of weather data

date	mont	time	INV-01	INV-02	INV-03	INV-04	INV-05	INV-06	INV-07	INV-08	INV-09	INV-10	INV-11	INV-12	INV-13	INV-14	INV-15	INV-16	Power (kW)
1/1/2023	1	07:00:00	2.6475	2.351	2.290333	0.106	2.7036	0	2.656167	2.6792	2.708	0	2.4	2.652167	2.7584	2.6328	2.557167	2.363667	33.506
1/1/2023	1	07:30:00	6.7152	6.598	6.231	0.5572	6.645333	0	7.098	7.002667	7.089667	0	5.9756	7.0978	7.161667	6.9705	7.0498	6.6592	88.8516333
1/1/2023	1	08:00:00	21.9006	23.123	21.6538	2.3104	21.128	0	25.1232	24.7436	25.1372	0	16.3066	25.1386	25.4518	25.0996	25.294	23.7588	306.1292
1/1/2023	1	08:30:00	55.0966	52.8718	54.9388	7.5806	53.2654	0	52.5918	52.5516	55.0524	0	46.4958	52.9954	55.0986	53.467	53.6498	53.2884	699.144
1/1/2023	1	09:00:00	56.41017	56.84067	56.51233	11.94683	57.32683	0	56.59083	56.3395	56.19517	0	56.4635	56.78517	56.77083	56.48783	56.428	56.80317	747.900833
1/1/2023	1	09:30:00	50.2366	50.3246	50.5582	15.822	50.5876	0	50.562	50.9412	50.6596	0	51.052	50.6474	50.6188	51.154	50.825	50.9646	674.9536
1/1/2023	1	10:00:00	50.3918	50.6474	50.8212	18.1548	50.653	0	50.5346	50.8662	50.4678	0	50.765	50.8116	50.7554	50.657	50.7886	50.4524	676.7668
1/1/2023	1	10:30:00	48.73717	48.7725	49.4446	19.315	49.4596	0	48.995	49.3636	49.5716	0	49.057	49.19867	49.3788	49.2998	48.849	49.09617	658.5385
1/1/2023	1	11:00:00	49.213	49.2474	49.5555	20.72683	49.94067	0	50.2064	49.948	50.0385	0	50.0935	49.8546	49.917	50.38783	49.5104	49.9068	668.546433
1/1/2023	1	11:30:00	52.0062	52.1296	51.5142	21.0668	51.4916	0	52.0138	51.8256	51.4618	0	52.0394	51.7456	51.271	51.9742	51.689	51.8608	694.0896
1/1/2023	1	12:00:00	53.4052	53.5458	53.8306	21.3144	53.3762	0	53.3416	53.2156	53.2854	0	53.2416	53.7968	53.0342	53.2798	53.6276	53.3524	715.6472
1/1/2023	1	12:30:00	53.74167	53.62533	53.4885	21.01733	53.63883	0	53.41183	53.28817	53.34683	0	53.6935	54.02117	53.64683	53.25567	53.38967	53.159	716.724333
1/1/2023	1	13:00:00	54.162	54.4678	53.7588	20.1766	53.9334	0	54.2724	54.132	53.6568	0	54.0386	54.7316	53.4914	53.781	54.3586	54.3952	723.3562
1/1/2023	1	13:30:00	53.3478	53.4556	53.1948	19.2034	53.1954	0	53.3066	53.0744	53.2372	0	53.4006	53.7374	53.344	53.0958	53.3006	53.6022	712.4958
1/1/2023	1	14:00:00	55.9072	55.2625	55.3182	17.851	54.8856	0	55.1162	54.9204	55.566	0	55.7294	55.77617	55.1666	54.8588	55.0882	54.9908	726.437067
1/1/2023	1	14:30:00	56.53317	56.1772	56.7655	16.021	56.88683	0	56.572	56.75083	56.24683	0	57.20717	56.3776	56.44717	56.72117	56.7605	56.61317	752.080133
1/1/2023	1	15:00:00	59.2198	58.7448	59.0734	13.5492	59.8206	0	58.7836	59.7068	59.7336	0	59.5484	58.6146	58.6652	59.5656	59.0876	59.206	783.3192
1/1/2023	1	15:30:00	57.5908	57.724	57.781	10.9872	59.2296	0	57.879	59.3516	59.3836	0	57.9228	57.5488	59.517	59.3088	57.9	57.5454	769.6696
1/1/2023	1	16:00:00	59.397	59.22633	59.0865	8.116167	59.48467	0	58.87033	59.17617	59.32333	0	59.59483	59.59217	58.97133	59.64417	59.16267	59.0565	778.702167
1/1/2023	1	16:30:00	47.1356	46.7598	43.7598	5.1172	50.8732	0	50.3766	50.6622	50.643	0	49.8496	50.8944	50.3516	50.4656	49.9732	49.7352	646.597
1/1/2023	1	17:00:00	24.9518	24.4024	23.113	2.4438	26.8698	0	26.6344	26.5544	26.8226	0	26.3712	26.8896	26.4358	26.6474	26.273	26.3172	340.7264

Figure. 3.7 Sample of power output data

date	month	time	timestep	temperature (°C)	humidity (%)	wind speed (m/s)	wind direction	diffuse radiation (W/m²)	direct normal irradiance (W/m²)	shortwave radiation (W/m²)	global tilted irradiance (W/m²)	Power (kW)
1/1/2023	1	7:00:00	1	15	92	0.78	50	10	122.1	19	19	33.506
1/1/2023	1	7:30:00	2	15.6	88	0.92	41	45	419.8	111	111	88.85163
1/1/2023	1	8:00:00	3	16.6	83	1.28	39	80	548.7	226	226	306.1292
1/1/2023	1	8:30:00	4	17.6	77	1.98	49	104	610.3	329	329	699.144
1/1/2023	1	9:00:00	5	18.6	72	2.72	54	138	613.1	422	422	747.9008
1/1/2023	1	9:30:00	6	19.5	68	3.02	56	184	563.7	493	493	674.9536
1/1/2023	1	10:00:00	7	20.2	65	3.24	56	251	469.4	543	543	676.7668
1/1/2023	1	10:30:00	8	21	63	3.38	56	327	359.9	573	573	658.5385
1/1/2023	1	11:00:00	9	21.6	61	3.44	54	356	339	604	604	668.5464
1/1/2023	1	11:30:00	10	22.4	59	3.5	53	341	389.3	639	639	694.0896
1/1/2023	1	12:00:00	11	23.2	56	3.48	51	281	546.8	710	710	715.6472
1/1/2023	1	12:30:00	12	24.2	53	3.34	51	195	757.1	792	792	716.7243
1/1/2023	1	13:00:00	13	25.1	51	3.2	51	152	842.7	807	807	723.3562
1/1/2023	1	13:30:00	14	25.7	49	3.26	50	134	856.2	777	777	712.4958
1/1/2023	1	14:00:00	15	26	48	3.41	50	116	864.6	730	730	736.4371
1/1/2023	1	14:30:00	16	26.2	47	3.48	51	103	854.4	663	663	752.0801
1/1/2023	1	15:00:00	17	26.4	46	3.48	51	100	816.6	580	580	783.3192
1/1/2023	1	15:30:00	18	26.3	46	3.48	51	98	757.2	483	483	769.6696
1/1/2023	1	16:00:00	19	26.1	47	3.48	51	88	697.5	380	380	778.7022
1/1/2023	1	16:30:00	20	25.8	48	3.48	51	72	618.8	270	270	646.597
1/1/2023	1	17:00:00	21	25.2	49	3.26	50	51	462.4	150	150	340.7264

Figure. 3.8 Resulting dataset after data preparation

For the power generation data, it was found that the amount of data in each time period was not equal; that is, there were some time periods where data was missing, resulting in the inability to combine the power data set with the weather data, which shows the missing data set for 2023 as shown in Figure 3.9

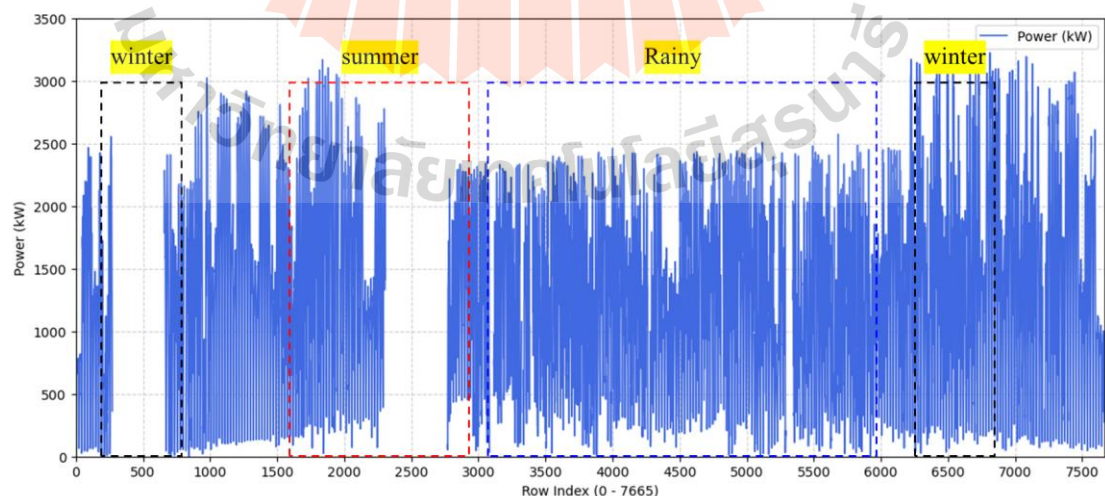


Figure. 3.9 Power output data for the year 2023

Figure 3.9 shows the missing power generation data range for all 3 seasons of Thailand in 2023. The missing data range is 1069 cells of the total data set of 7,665 cells, which is 13.95% of the total missing data consisting of 472 rows for summer, 178 rows for the rainy season, and 419 rows for winter. Therefore, the data was formatted to have the same number of rows for all inverters, and the missing data values were filled in (Missing Data Imputation) using the Linear Interpolation method, with the details as follows:

3.2.2.1 Linear Interpolation

This method is suitable for cases where data is missing in a short period of time. It estimates the missing values based on the previous and next values in the time series. Linear interpolation is particularly suitable for continuous time series data, as it preserves the overall trend and does not distort the distribution of the data. In this study, this method was used to calculate the missing power generation for the whole year 2023. Figure 3.10 shows the data after calculation using linear interpolation. Figure 3.11 presents a surface plot of the solar PV power generation capacity in 2023, with the X-axis representing the timestep within day, the Y-axis representing the day of year, and the Z-axis representing the power generation in kilowatts (kW). The data set was filled in with the missing values using linear interpolation to obtain a continuous and complete dataset before analysis.

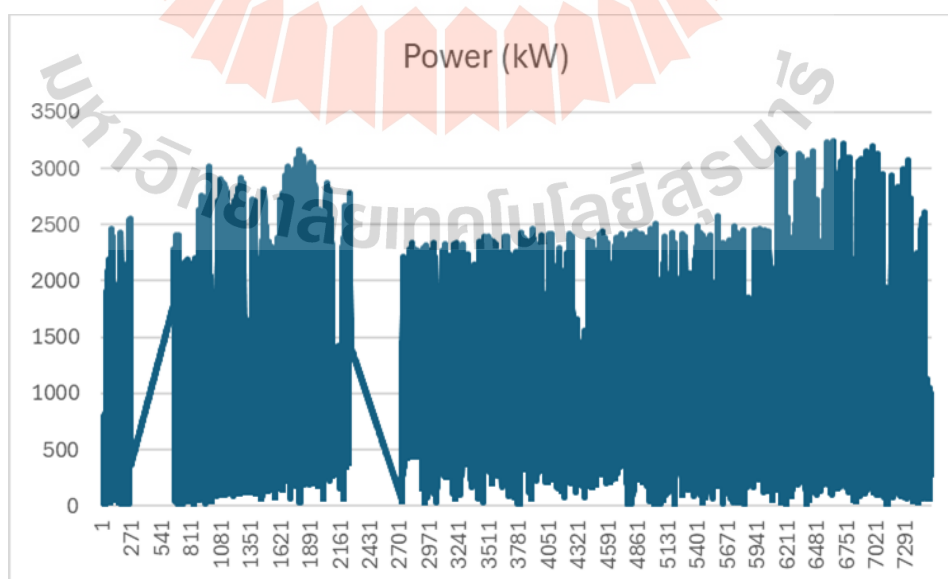


Figure. 3.10 Power output data for the year 2023 after using linear interpolation

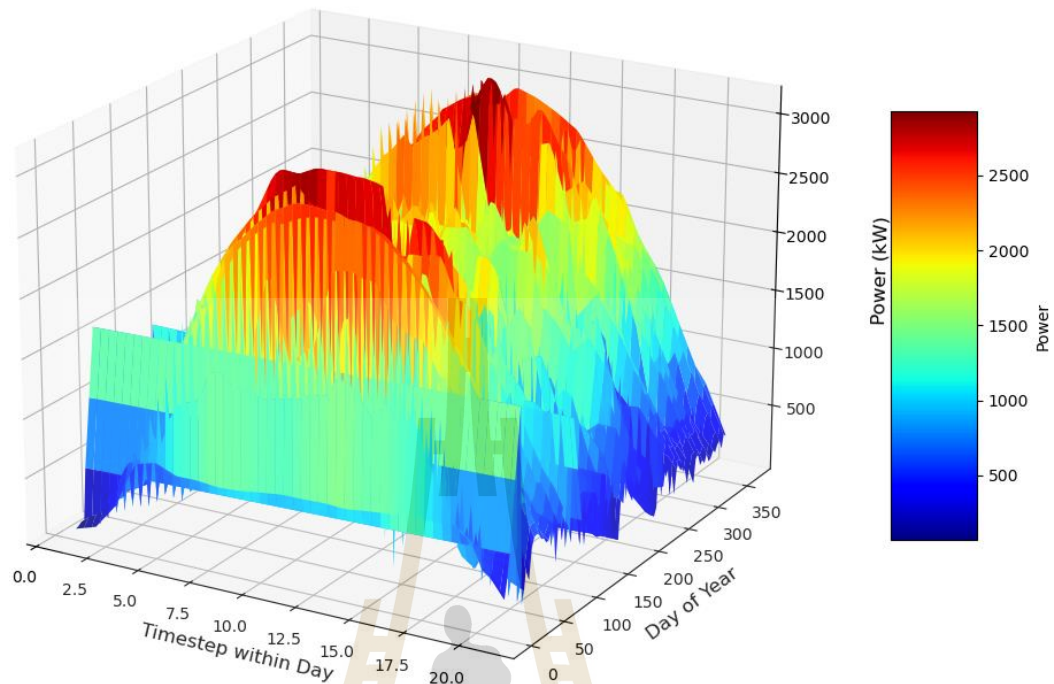


Figure. 3.11 Surface plot of PV power generation throughout the year 2023

The figure clearly shows the variation of the power output on a daily basis, which is consistent with the solar power generation behavior. In the morning, the power generation is relatively low and begins to increase when sunlight is present, reaching a maximum during the day. It then decreases in the afternoon and evening to nearly zero at night, reflecting the rising and setting of the sun. The graph's shape is an "inverted bell curve," clearly reflecting the system's diurnal power generation pattern. When considered in terms of day-to-day dimensions, the power output tends to fluctuate seasonally, with peak generation occurring in the middle of the year (approximately April–July, or approximately 100–250 days). The graph surface appears yellow to red, indicating high power generation, consistent with the peak summer light intensity. Conversely, in the early and late periods (<100 and >300 days), the power output continuously decreases. The graph surface is low and blue or green, indicating decreased power generation due to shorter sunlight exposure or increased cloud cover. Furthermore, some graph surfaces appear smooth or have sharp edges, a direct result of linear interpolation, which calculates missing values from a straight line between the previous and subsequent actual data. This results in continuous data but

cannot fully reflect the actual fluctuations of the system, especially in cases where data is missing over a long period of time. Connecting values with a straight line can oversmooth the resulting power output and may reduce peak power or daily variability from the actual values.

However, linear interpolation is still appropriate for cases where data is missing for a short period and the system is subject to continuous change, as it can maintain the main pattern of the data trend well, although it has the limitation of reducing some of the data fluctuations from the actual values.

In summary, this surface plot demonstrates the solar power generation pattern throughout 2023, covering both daily and seasonal dimensions. It also demonstrates the impact of linear interpolation, which ensures continuous data availability for subsequent forecasting steps.

3.2.3 Feature Selection and Correlation Analysis

After addressing missing and anomalous data using linear interpolation, a statistical correlation analysis was conducted employing a correlation heatmap to identify factors influencing power output. The identified variables used as inputs and outputs include temperature, humidity, wind speed, wind direction, diffuse radiation, DNI, GHI, predicted temperature, predicted irradiance, and power (kW), as shown in Figure 3.12

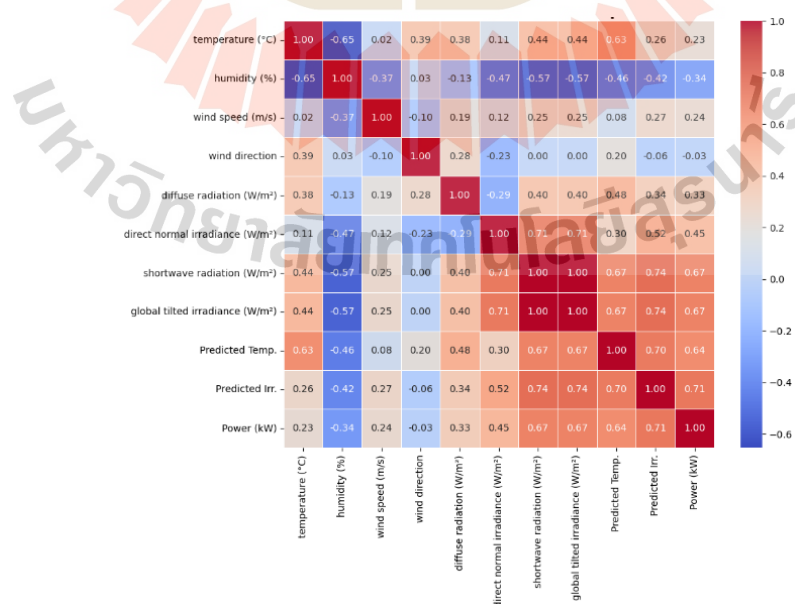


Figure. 3.12 The correlation matrix

From the correlation heatmap analysis above, it can be observed that the correlation coefficient between temperature and power (kW) is 0.23, while that between irradiance and power (kW) is 0.67. These are key factors affecting power generation. Therefore, temperature and irradiance were forecasted using the LSTM model. The forecasted results were then analyzed with a correlation heatmap, revealing increased correlation coefficients between temperature and irradiance with power (kW). Consequently, the predicted temperature and irradiance values were used to forecast power output. The input data comprised humidity, wind speed, GTI, predicted temperature, and predicted irradiance, as illustrated in Figures 3.13 -3.17, respectively.

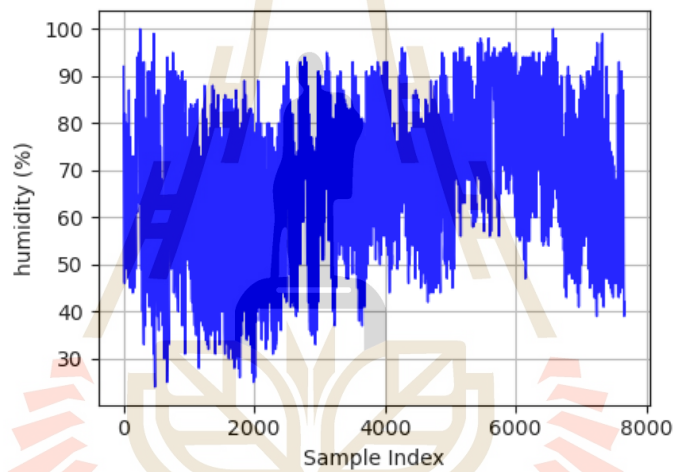


Figure 3.13 Humidity (%) time series used as model input

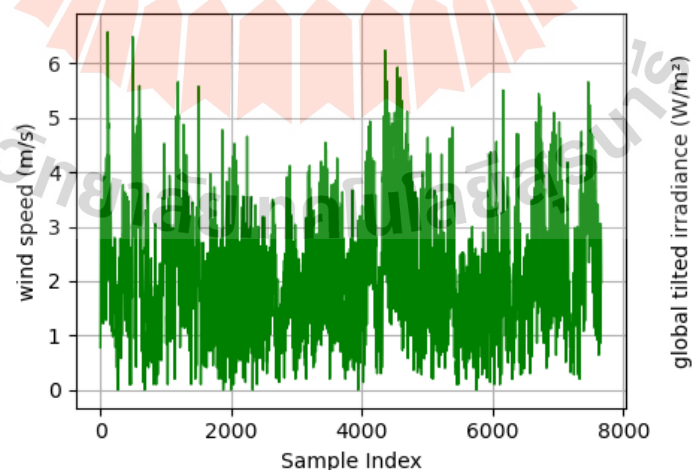


Figure 3.14 Wind speed (m/s) time series used as model input

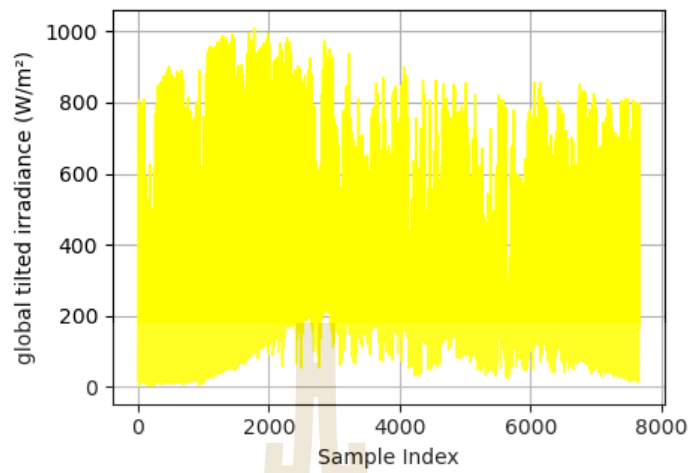


Figure 3.15 GTI (W/m²) time series used as model input

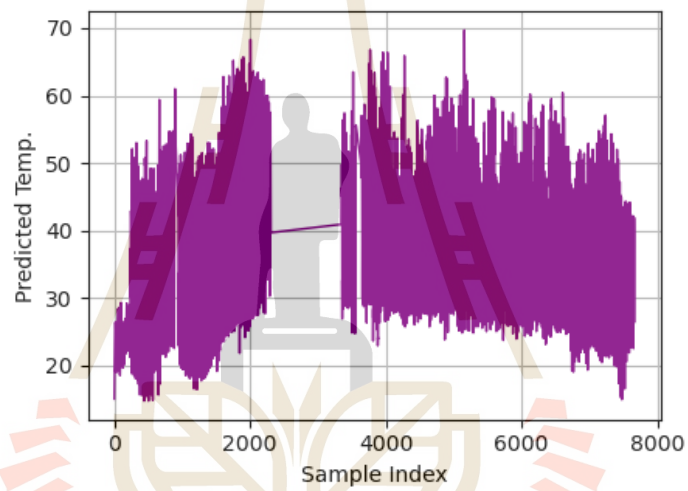


Figure 3.16 Predicted temperature time series from LSTM model

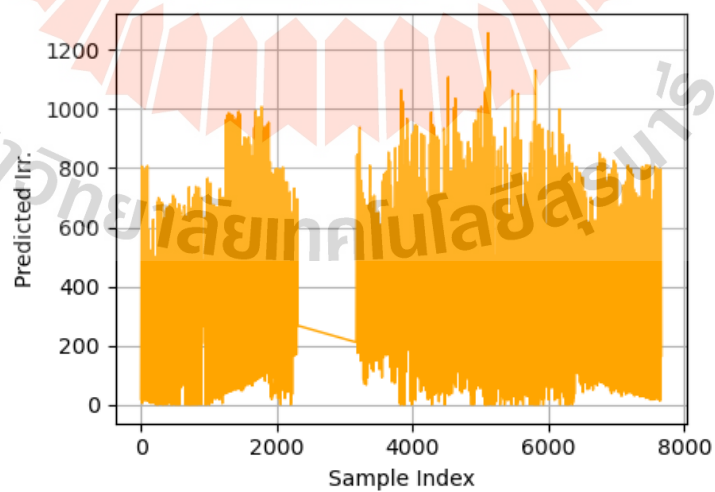


Figure 3.17 Predicted irradiance time series from LSTM model

3.2.4 Data Splitting

This study focuses on day-ahead PV power forecasting using time-series data with a 30-minute resolution covering an entire year. After conducting EDA through correlation analysis, a total of 7,665 data samples were collected and divided into three subsets: 7,539 samples were allocated to the training and validation sets (split 80% and 20%, respectively), and 126 samples were reserved for the test set.

Following data preparation, which included merging data from multiple sources, formatting it appropriately, selecting time periods that exhibit seasonal patterns relevant to power generation, and handling missing values, the resulting dataset was complete and ready to be used for developing and training the forecasting models.

To enhance the performance of the deep learning models and facilitate faster and more stable parameter optimization during training, all input features were normalized to a common scale. This normalization process plays a crucial role in accelerating model convergence and reducing issues caused by differing feature scales.

Such a systematic approach to data preparation is a critical foundation, as data quality directly impacts forecasting accuracy, especially in the context of time-series data for solar power generation systems.

3.3 Two-Step Forecasting Framework

The forecasting process in this study was designed using a two-step forecasting approach to enhance the efficiency and accuracy of solar PV power generation prediction. The details of each stage are as follows:

3.3.1 Environmental Factors Pre-Forecasting

After statistically analyzing the relationship using a correlation heatmap to identify the factors influencing the power, namely temperature and light intensity, the temperature and light intensity were forecast using the LSTM model, a deep learning model suitable for time-series data, as detailed below.

3.3.1.1 Long Short-Term Memory

LSTM is a type of RNN first introduced by Hochreiter and Schmidhuber in 1997. A standard LSTM network consists of a memory cell and three types of gates: the input gate, output gate, and forget gate. These gates regulate the flow of information into and out of the memory cell. The cell itself retains values over long periods of time, allowing the model to learn long-term dependencies in time-series data.

The forget gate determines which parts of the previous cell state should be discarded. It does so by transforming the previous state and the current input into values between 0 and 1. A value close to 1 (rounded) implies that the information is retained, while a value close to 0 indicates that the information is discarded. The input gate decides which new information should be added to the current cell state, using a mechanism similar to that of the forget gate. Finally, the output gate controls which part of the current cell state should be passed to the next layer by assigning a weight between 0 and 1, based on both the current and previous states. The structure of the LSTM architecture is illustrated in Figure 3.18

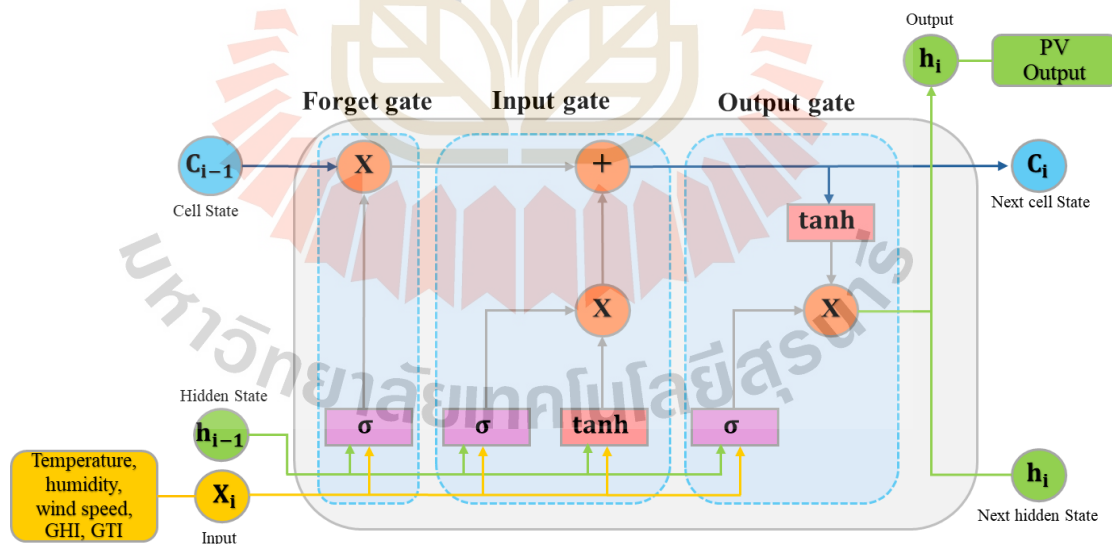


Figure 3.18 Architectural diagram of LSTM

After obtaining the results of the environmental forecast, the results were analyzed for relationships using a correlation heatmap to compare values before and after the environmental forecast. After the analysis, the features that

influenced power generation were used as inputs for the next step of the power generation forecast. The details of the input data are as shown in the content of 3.2.3 above.

3.3.2 PV-Forecast

After the data preparation process, all input data are normalized to the same standard before being fed into the model training process for forecasting power generation with LSTM, Bi-LSTM, GRU, and RF models, with details as follows:

3.3.2.1 Bidirectional LSTM

Bi-LSTM as illustrated in Figure 3.19 is developed using two temporal directions. Bi-LSTM is capable of learning more accurately than a standard unidirectional LSTM by processing information from both the past and the future. It achieves this by connecting neurons from two separate hidden layers one processing forward in time and the other backward to a single output. Input information is simultaneously propagated to all hidden layers from both directions; however, the neurons in each layer are not interconnected with each other. The output of the Bi-LSTM, denoted as y_i , is calculated by combining the results from both the forward and backward directions.

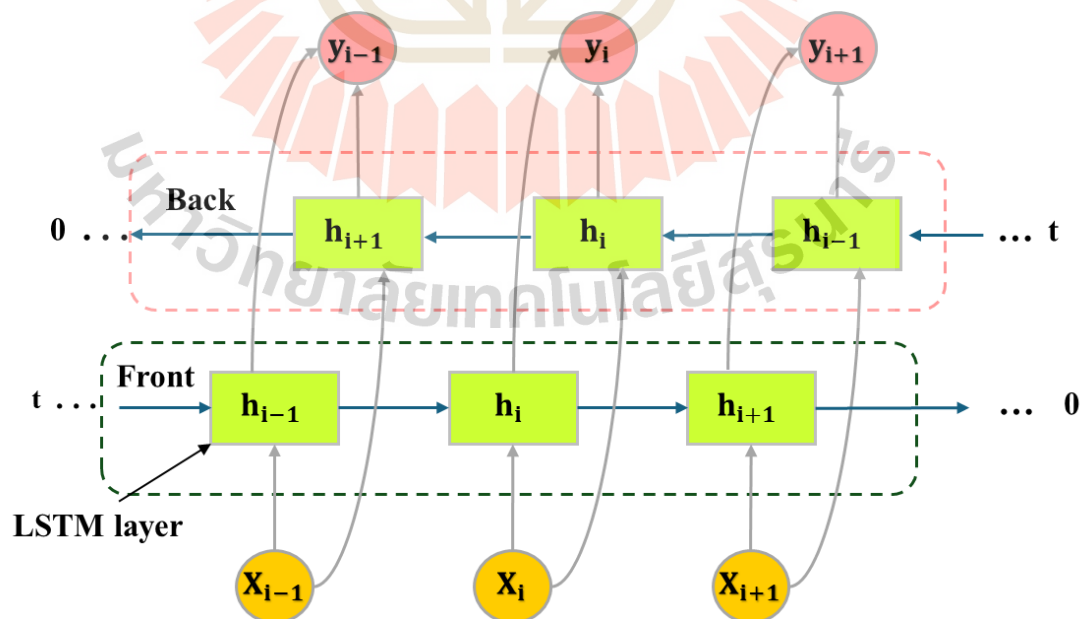


Figure 3.19 Architectural diagram of Bi- LSTM

3.3.2.2 Gated Recurrent Units

The GRU is a variant of the Recurrent Neural Network (RNN) designed to effectively capture long-term dependencies by addressing the gradient vanishing problem encountered in traditional RNN architectures. While it shares conceptual similarities with the LSTM model particularly in handling recurrent processing errors GRU has a more compact architecture with fewer parameters. This structural simplicity enables faster training compared to LSTM.

However, GRU may be less effective at modeling complex spatial relationships present in meteorological data. This limitation arises from its sensitivity to scaling and distributional characteristics of the input data, necessitating more extensive preprocessing to ensure model performance.

Figure 3.20 illustrates the architecture of the GRU model. During each training step, the input data are dynamically modified via two key components: the update gate and the reset gate. The reset gate determines the extent to which the previous hidden state should be forgotten, while the update gate controls how much of the past information should be retained and passed forward to influence the current output.

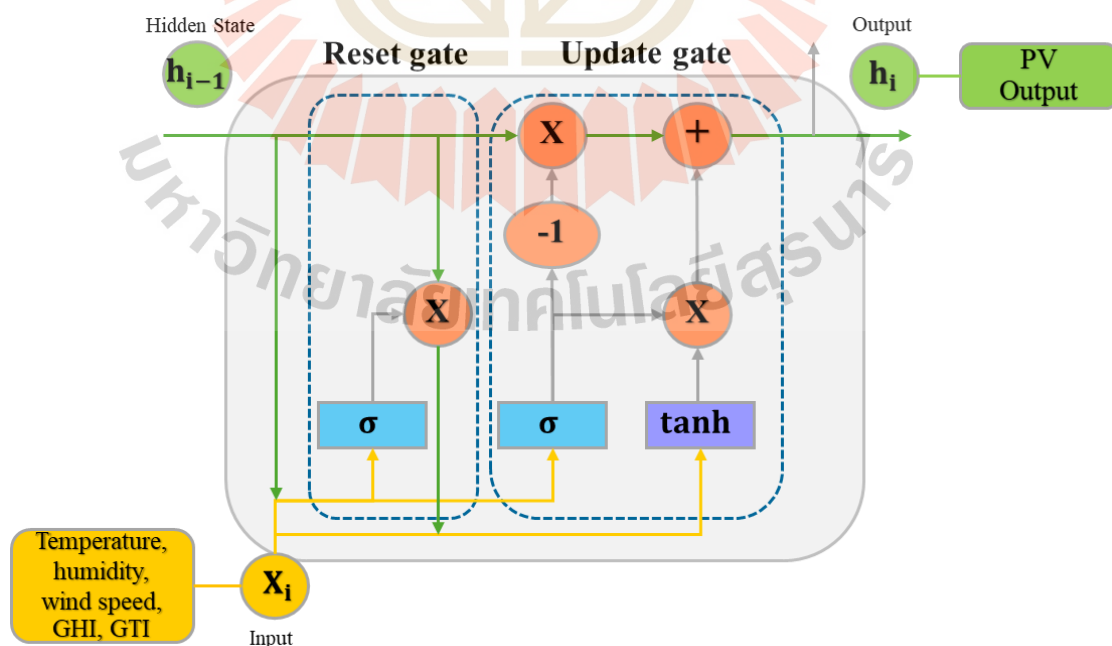


Figure 3.20 Architectural diagram of GRU

3.3.2.3 Random forest

RF is an ensemble technique that employs both bagging and feature randomness to construct a collection of uncorrelated decision trees. The final output is determined by taking the average of the predictions from all individual trees. By aggregating predictions in this manner, Random Forest significantly reduces variance and enhances the accuracy of forecasting results. The fundamental architecture of the Random Forest model is illustrated in Figure 3.21

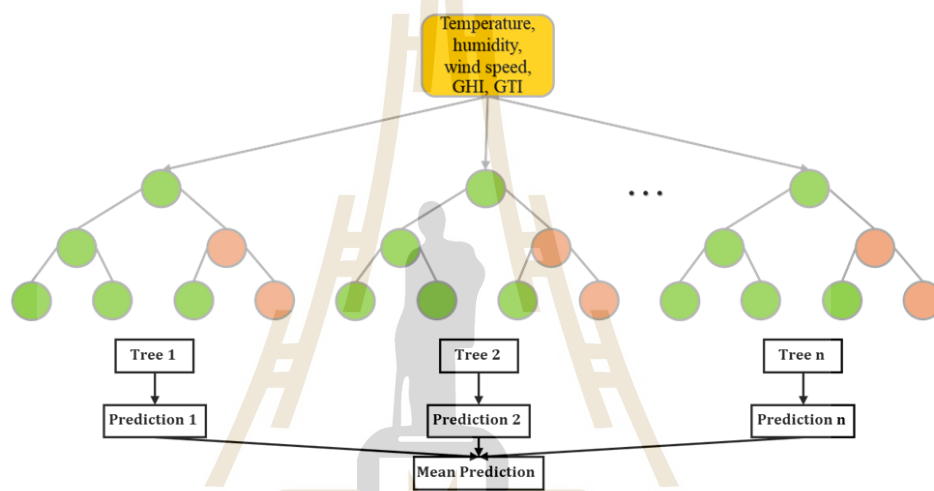


Figure 3.21 Architectural diagram of RF

For the parameter configuration, the structural parameters of all three models are designed the same for the training process, unlike the RF, which has different structures, the details of which are shown in Table 3.2.

Table 3.2 Hyperparameter configuration for model training

Model	Hyperparameters
LSTM	Input: 5, Hidden nodes: 64,32,50, Epochs: 100, Batch Size: 32, Learning Rate: 0.001, Optimizer: Adam, Early Stopping: 20
Bi-LSTM	Input: 5, Hidden nodes: 64,32,50, Epochs: 100, Batch Size: 32, Learning Rate: 0.001, Optimizer: Adam, Early Stopping: 20
GRU	Input: 5, Hidden nodes: 64,32,50, Epochs: 100, Batch Size: 32, Learning Rate: 0.001, Optimizer: Adam, Early Stopping: 20
RF	Input: 5, Number of Trees: 100, Max Depth: 10, Bootstrap: True, Criterion: MSE, Max Features: auto

3.4 Evaluation metrics

In the final step, the evaluation metrics of the appropriate model of the PV power forecast are selected to verify the prediction and efficiency of PV generation, which makes the prediction results more reliable and applicable. In this work, three fitness functions, namely RMSE, MAE, and MAPE, are used, which can be called using the Sklearn library in Python. The first comparative index considered is the RMSE. It can be measured by averaging the squares of the errors and extracting the square root as Equation 3.1

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_{actual,i} - y_{forecast,i})^2}{n}} \quad (3.1)$$

The next index used is MAE, which can be calculated by averaging the absolute values of the errors (absolute error) as in Equation 3.2

$$MAE = \frac{\sum_{i=1}^n |y_{actual,i} - y_{forecast,i}|}{n} \quad (3.2)$$

MAPE can be found by averaging the percentage errors as shown in Equation 3.3

$$MAPE = \sum_{i=1}^n \frac{|y_{actual,i} - y_{forecast,i}|}{y_{real,i}} \times 100\% \quad (3.3)$$

3.5 Simulation Results

The experimental results are divided into two main parts. The first part involves forecasting environmental factors that affect electricity production, such as temperature and solar irradiance, using the LSTM model. The second part involves using the predicted values from the first step, along with other meteorological data, as inputs to forecast electricity generation using the deep learning and ensemble models discussed above.

Additionally, a comparative analysis of each model's performance across three seasons summer, rainy, and winter was conducted using evaluation metrics including RMSE, MAE, and MAPE. These quantitative indicators were used to assess the accuracy of the forecasts. The results of these experiments not only validate the effectiveness of the two-step forecasting approach but also provide insights into the suitability of different models under varying climatic conditions.

3.5.1 Forecasting Results of Environmental Factors

In this study, a dataset from the year 2023 was used, consisting of 7,539 samples (divided into 80% for training and 20% for validation) and 126 samples for the test set. This division ensures the diversity of data for robust model validation. Due to the presence of missing data which could negatively impact the accuracy and reliability of forecasting results linear interpolation was applied to fill the gaps. Moreover, to enhance the forecasting accuracy of solar power generation, two key environmental factors solar irradiance and temperature were predicted using the LSTM model, as they directly influence electricity production.

In the first stage of the two-step forecasting approach, the LSTM model was employed to forecast temperature and solar irradiance in advance. LSTM, a deep learning model well-suited for time-series data, was selected due to its capacity to capture long-term dependencies and dynamic patterns effectively.

After forecasting the environmental factors, a heatmap was constructed to analyze the correlation coefficients between the forecasted values and the actual electricity generation. The results illustrate the relationships between various environmental factors (solar irradiance and ambient temperature) and the power output of the solar panels during the year 2023. These relationships are illustrated in Figure 3.22 for April, with the complete set of results for all twelve months presented in Appendix B.

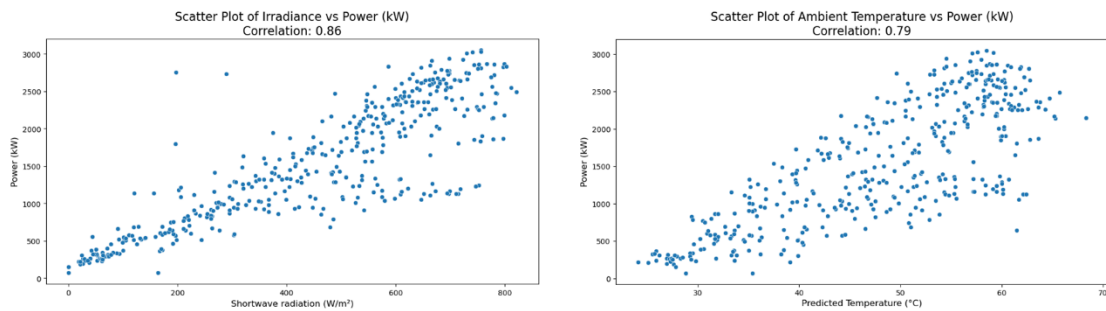


Figure 3.22 Environmental correlation with solar power output - April 2023

An analysis of the correlation between solar irradiance and ambient temperature with the electricity generation of the floating solar power system throughout 2023 revealed that April exhibited the highest correlation values. Specifically, irradiance had a correlation coefficient of 0.86 and temperature had 0.79, indicating a strong positive relationship between these two factors and the amount of electricity produced. This highlights April as the most suitable month for solar energy generation due to consistently high irradiance and temperature levels. During February to March, the correlation values for both factors remained relatively high (ranging from 0.70 to 0.80), reflecting favorable weather conditions for electricity generation. In contrast, May showed a negative correlation for solar irradiance (-0.37), possibly due to atmospheric instability, such as cloud cover or fog, which resulted in inconsistent solar radiation. However, temperature still had a moderate correlation of 0.37.

In the rainy season, from June to October, the correlation values for both factors ranged from moderate to high. Notably, in August and September, both irradiance and temperature had correlation coefficients above 0.75. This suggests that despite frequent rainfall, sufficient solar radiation was still available during certain periods, enabling continuous electricity production. For the winter season, from November to January, the correlation gradually declined especially in January, where the correlation between temperature and power generation dropped to just 0.15, indicating a reduced influence of temperature on electricity production.

In summary, the correlation between environmental factors and power output varied across seasons. The period from April to August demonstrated the highest correlation and was most suitable for model-based forecasting. In contrast,

May and the winter months exhibited lower correlations, which could affect forecasting accuracy if proper data handling techniques are not applied.

From the correlation analysis using a correlation heatmap, it was found that, prior to forecasting, the correlation coefficient between ambient temperature and power generation (kW) was 0.23, while that between solar irradiance and power generation was 0.67. These values indicate a moderate relationship between environmental factors and generated energy.

However, after forecasting environmental variables using the LSTM model, the correlation between the forecasted variables and electricity generation increased significantly. Specifically, the correlation coefficient between predicted temperature and power (kW) rose to 0.64, and the coefficient between predicted irradiance and power (kW) increased to 0.71. These results demonstrate that using forecasted environmental data can better capture the dynamic characteristics of environmental factors and clearly enhance their relationship with electricity output.

These results support the idea that the two-step forecasting approach presented above can significantly improve the model performance in the power generation forecasting stage. It also shows how important it is to handle environmental data correctly before using it in the final forecasting process.

3.5.2 Forecasting Results of Solar Power Generation

In the second step of the process, the predicted values of ambient temperature and solar irradiance were combined with additional meteorological inputs, including GTI, relative humidity, and wind speed, to forecast the next-day electricity output of the FPV system. The forecasting was carried out using four models: LSTM, Bi-LSTM, GRU, and Random Forest, respectively. Since the objective was to perform day-ahead forecasting of PV energy generation, the results were analyzed and presented according to three seasonal periods: summer (mid-February to mid-May), rainy season (mid-May to mid-October), and winter (mid-October to mid-February).

The simulation results, based on a validation dataset consisting of 1,508 samples, are illustrated in Figures 3.23, 3.24, 3.25, and 3.26, respectively. These figures provide a comparative visualization of the actual power output versus the forecasted power output for each of the four models, allowing for a clear assessment of their

performance. The results indicate that all models achieved relatively high accuracy, particularly during periods when the power output was below 2,500 kilowatts a range characterized by stable trends and low variability in the data.

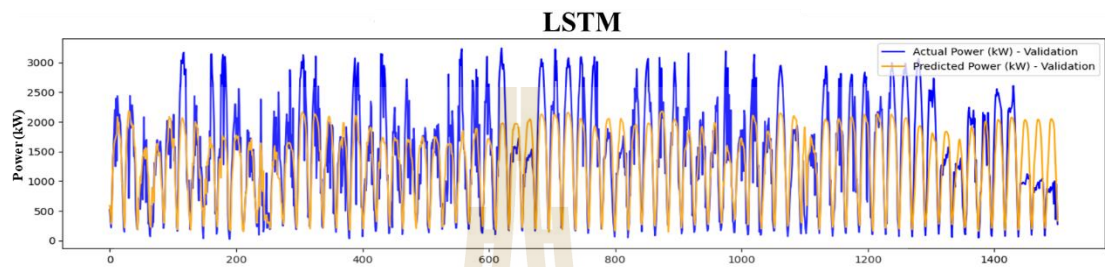


Figure 3.23 Comparison of forecasted vs. actual power output using LSTM

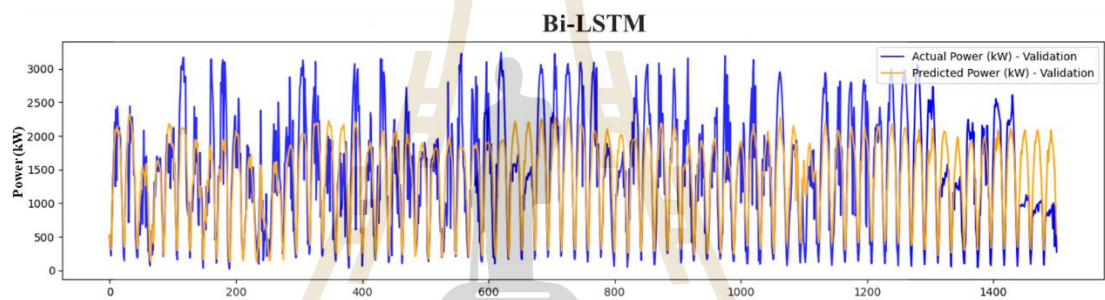


Figure 3.24 Comparison of forecasted vs. actual power output using Bi-LSTM

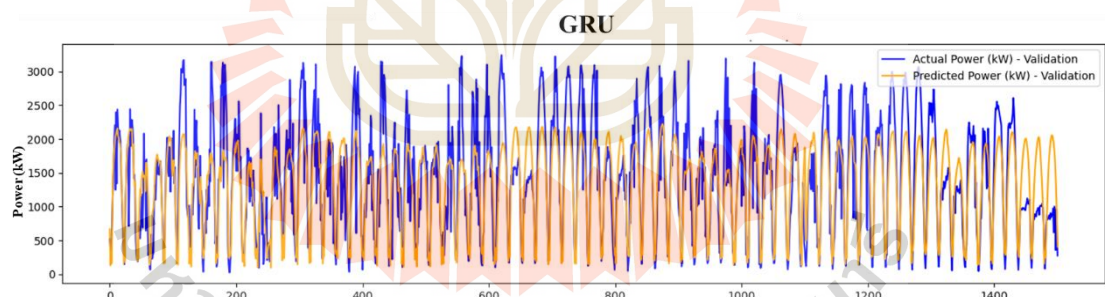


Figure 3.25 Comparison of forecasted vs. actual power output using GRU

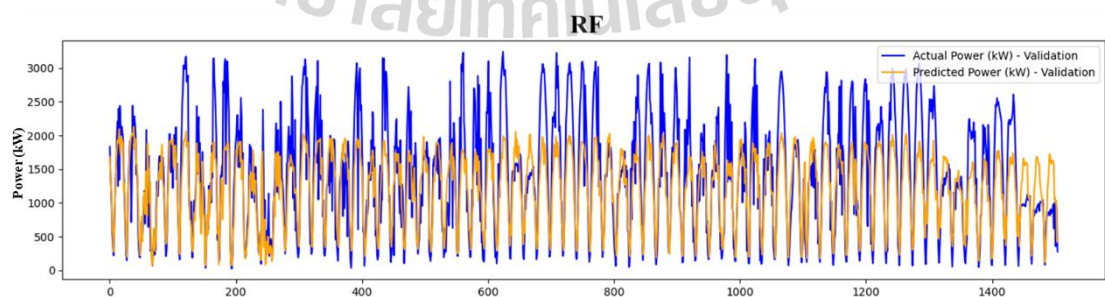


Figure 3.26 Comparison of forecasted vs. actual power output using RF

Figures 3.27-3.29 present a comparison between the actual power output and the forecasted values generated by the models across the three seasons: summer, rainy season, and winter.

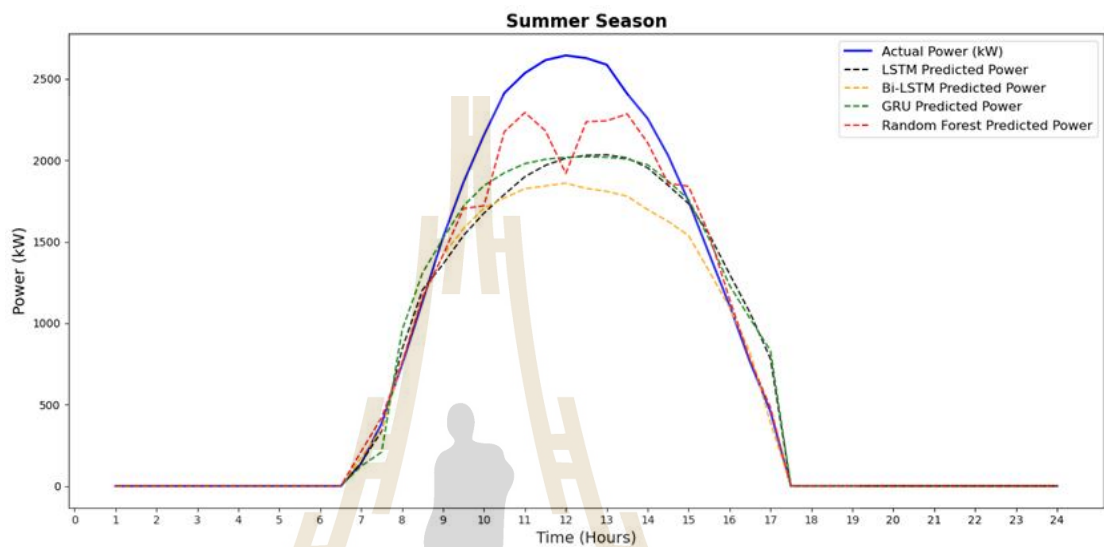


Figure 3.27 Day-ahead solar power forecasting for the summer season

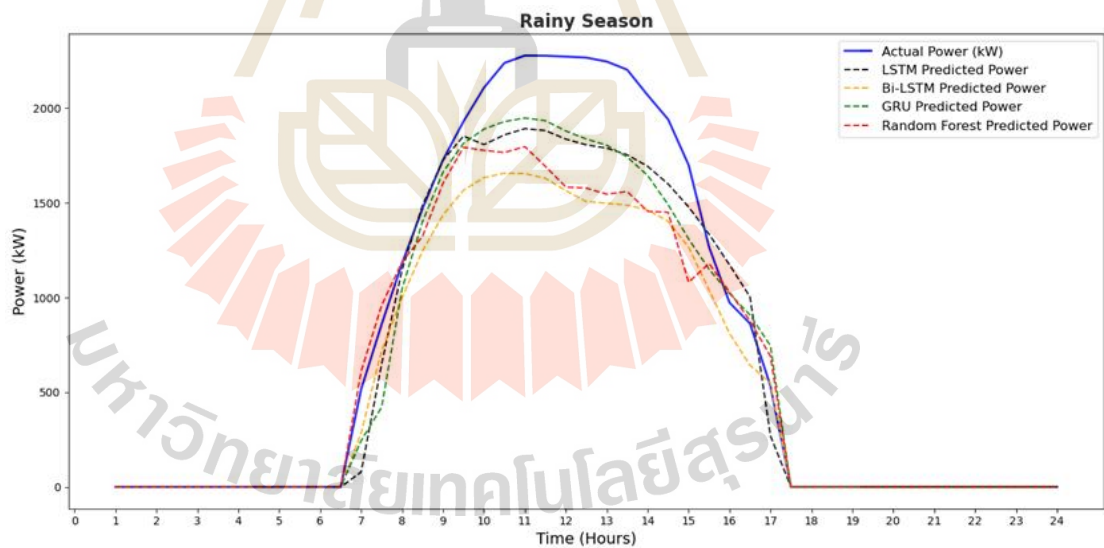


Figure 3.28 Day-ahead solar power forecasting for the rainy season

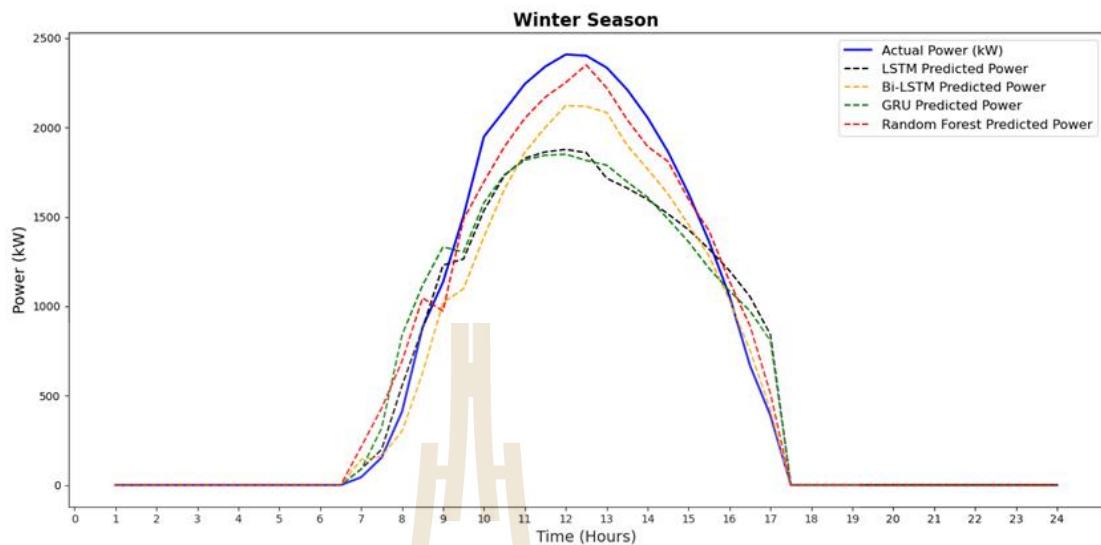


Figure 3.29 Day-ahead solar power forecasting for the winter season

3.6 Performance Metric Comparison

To evaluate the performance of the one-day-ahead electricity generation forecasting models, key metrics including RMSE, MAE, and MAPE were compared. The evaluation was conducted separately by season to assess the forecasting capability of each model: LSTM, Bi-LSTM, GRU, and RF.

After calculating the validation metrics based on the predicted and actual power output, the results are presented in Table 3.3. For the summer season, the RF model achieved the best performance among the four models with RMSE, MAE, and MAPE values of 260 kW, 188 kW, and 11.35%, respectively, indicating the most accurate forecasting. However, during the rainy season, the GRU model demonstrated the best performance with values of 310 kW, 272 kW, and 18.37%, respectively. Conversely, in the winter season, the Bi-LSTM model yielded the most accurate results with RMSE, MAE, and MAPE values of 271 kW, 227 kW, and 24.65%, respectively.

The LSTM model in the summer provided slightly less accurate results compared to the RF model but outperformed the GRU model, with RMSE, MAE, and MAPE values of 389 kW, 319 kW, and 19.26%, respectively. In the rainy season, LSTM was more accurate than Bi-LSTM, with values of 311 kW, 269 kW, and 19.20%, respectively. On the other hand, its performance in the winter season was the lowest among all three seasons.

Furthermore, the models were also compared in terms of processing time. Ranked from the shortest to the longest processing time, the results were as follows: RF (12 seconds), GRU (61 seconds), LSTM (92 seconds), and Bi-LSTM (137 seconds), as detailed in Table 3.4. These results align with the theoretical structural complexity of each model, with simpler models requiring less computational time. This consistency reinforces the expectation that model complexity directly impacts processing efficiency, highlighting the trade-off between accuracy and computational cost in forecasting applications.

Table 3.3 Forecasting results

Model	Summer			Rainy			Winter		
	RMSE (kW)	MAE (kW)	MAPE (%)	RMSE (kW)	MAE (kW)	MAPE (%)	RMSE (kW)	MAE (kW)	MAPE (%)
LSTM	389	319	19.26	311	269	19.20	365	310	29.14
BI-LSTM	468	357	17.73	481	423	24.17	271	227	24.65
GRU	360	295	20.45	310	272	18.37	376	339	36.32
RF	260	188	11.35	429	344	18.78	166	148	38.97

Table 3.4 Processing time comparison of forecasting models

Model	Processing Time (seconds)
LSTM	92
BI-LSTM	137
GRU	61
RF	12

The simulation results indicate that the RF model performs well for forecasting electricity generation during the summer season. This is because, in summer, weather conditions tend to be stable with consistent power output, resulting in continuous and

less volatile data used for model training. Consequently, the model can effectively learn the relationships among different features. In contrast, the GRU model is more suitable for forecasting power output during the rainy season, where data is characterized by high uncertainty and rapidly changing power output, causing significant fluctuations. The GRU model can adapt well to such parameter changes. For the winter season, the best forecasting method is Bi-LSTM. This is due to the fact that winter electricity generation is influenced by multiple factors such as shorter sunlight hours, lower temperatures, and higher humidity. These conditions lead to rapid changes and complex data patterns, which the Bi-LSTM model can efficiently capture by processing sequential data bidirectionally and learning the intricate relationships among various factors. Additionally, forecasting errors may arise from missing data used in the prediction process, as well as differences in weather conditions such as temperature, wind speed, or humidity, which can directly affect the efficiency of solar panels.

3.7 Conclusion

This chapter systematically describes the research approach, encompassing data collection and preparation, feature selection and correlation analysis, and the development of solar power forecasting models using deep learning techniques. This also includes model evaluation using accuracy indicators. Experimental results are presented, demonstrating that a two-step forecasting approach, which begins with environmental factors and then follows with power forecasting, significantly improves forecast accuracy. Analysis of seasonal results also revealed that no single model is superior across all seasons. Each model's suitability differs depending on the data characteristics and climate conditions at each time. The RF model performs best in the summer, when data is relatively stable and fluctuates, while the GRU model performs best in the rainy season, when data uncertainty is high. Bi-LSTM is most appropriate in winter, when data patterns are more complex. This suggests that model selection for solar power forecasting should take into account the data characteristics, climate conditions, and seasonality at each time period to ensure a more accurate and reliable solar power forecasting system in the future.

CHAPTER IV

SHORT-TERM SOLAR POWER FORECASTING UNDER INCOMPLETE DATA AND CURTAILMENT CONDITIONS²

4.1 Introduction

Chapter 3, which presented solar power generation forecasting using a two-stage forecasting framework and filling in missing data with linear interpolation, builds upon this by focusing on data management to align with the actual operating behavior of the system. The research process includes data preparation using the Clear Sky Model and KNN, along with consideration of power generation curtailment. The two-stage forecasting is performed in three time periods before combining the results to obtain daily forecasts. Finally, the results are evaluated and analyzed using XAI to understand the role of external factors affecting power generation. The conceptual framework of the research is shown in Figure 4.1.

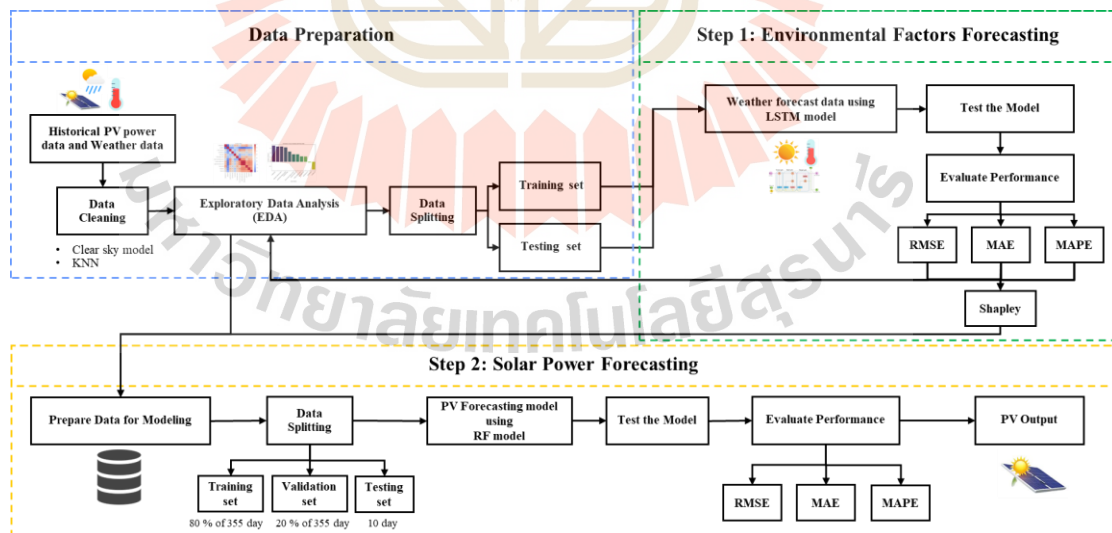


Figure 4.1 The framework for solar power forecasting

² Part of this chapter had been presented at the 14th International Electrical Engineering Congress (IEEECON 2026), “Innovating the future of engineering: fostering resilience and adaptability in complex systems”, Pattaya, Thailand, March 4–6, 2026.

4.2 Data preparation

The data management process consists of data collection, data cleaning, feature selection and correlation analysis, and data splitting, as illustrated in Figure 4.2.

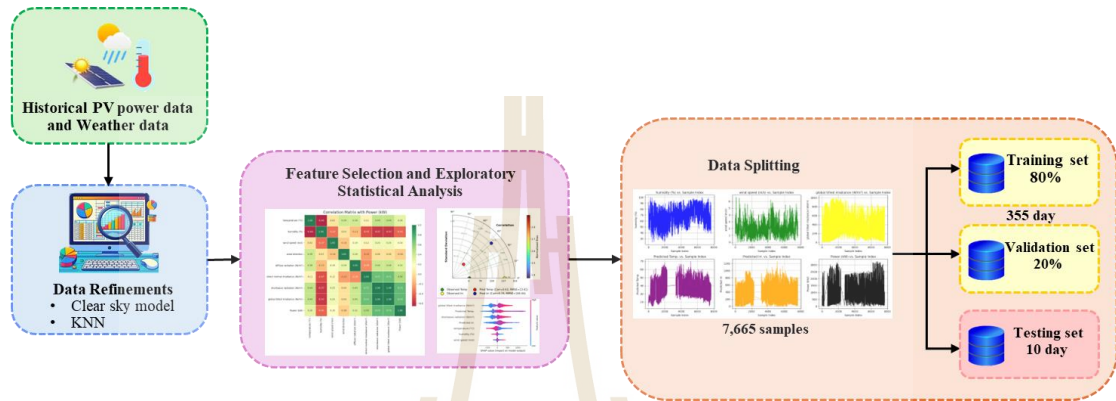


Figure 4.2 The framework for data management process

4.2.1 Data collection

The data used in this study are from a floating PV power plant located at 111 University Road, Suranaree Sub-district, Mueang Nakhon Ratchasima District, Nakhon Ratchasima Province, Thailand, and data from websites. Only energy consumption between 7:00 AM and 5:00 PM is considered, with a 30-minute simulation period, as described in Chapter 3.

For the power generation dataset used in this study, 13.95% of the data was found to be missing, as shown in Figure 3.9 This figure displays the missing power generation data for all three seasons in Thailand in 2023. Furthermore, additional data analysis revealed abnormal power generation behavior. This curtailment data is different from the missing data and may affect the learning process of the forecasting model if not managed properly.

4.2.2 Data cleaning

The selection of missing data imputation methods has a direct impact on the accuracy of solar energy forecasting models, particularly in the context of time-series data such as PV power output.

For the imputation of missing values in the power output dataset, seasonal variations in Thailand must be carefully considered, as the characteristics of solar radiation differ significantly across seasons. Thailand experiences three main seasons: summer (March–May), characterized by intense sunlight and the highest levels of solar irradiance; the rainy season (June–October), which has frequent cloud cover and rainfall, resulting in highly variable and often lower-than-expected irradiance under clear-sky conditions; and winter (November–February), when solar energy production is generally lower than in summer, though many days remain clear. Nevertheless, the potential impact of decreased temperature on solar panel efficiency should also be taken into account.

Therefore, the choice of imputation method for each season should align with solar radiation characteristics and prevailing weather conditions to ensure reasonable estimates and enhance forecasting model accuracy. For summer and winter, the Clear Sky Model was applied. In summer, skies are generally clear, with minimal cloud cover and stable weather, aligning well with the assumptions of the clear-sky model, resulting in high accuracy of estimated values. Although winter offers lower solar output than summer, clear skies are still common, making the Clear Sky Model applicable in this season as well. However, the effect of lower temperatures on panel performance should still be considered. In contrast, the rainy season presents distinct challenges for solar forecasting due to highly unpredictable weather patterns such as thick clouds, frequent rain, and overcast conditions which reduce the reliability of clear-sky estimations, as they tend to overestimate solar irradiance. Consequently, this study employs the KNN Imputation method for handling missing data during the rainy season. This approach compares the meteorological variables of the missing data periods with similar periods in the dataset that contain complete data, using the closest match to estimate the missing power output values. This method helps reduce imputation errors and contributes to improved model accuracy in subsequent forecasting steps.

4.2.2.1 Clear Sky Model

In cases where data were missing for extended periods or when missing values significantly impacted the structure of the dataset, the Clear Sky Model

was applied. This model estimates solar irradiance under clear-sky conditions by considering geographic coordinates, date, time, solar panel efficiency, and GTI. The model was implemented using the Python library pvlib, which automatically calculates clear-sky irradiance based on these parameters. The estimated irradiance was then used to approximate the expected solar power output.

The calculation of solar panel efficiency is described as follows:

$$\eta = \frac{P_{\max}}{\text{Irradiance} \times \text{Area}} \quad (4.1)$$

Based on the panel dimensions of 2230 mm × 1134 mm, the conversion from millimeters to meters yields

$$\text{Dimensions} = 2.230 \times 1.134 \text{ m} \quad (4.2)$$

Therefore, the area of 1 panel

$$\text{Area}_{1\text{panel}} = 2.230 \times 1.134 = 2.529 \text{ m}^2 \quad (4.3)$$

Accordingly, efficiency is calculated as follows:

$$\eta = \frac{P_{\max}}{\text{Irradiance} \times \text{Area}}$$

$$\eta = \frac{540}{1000 \times 2.529}$$

$$\eta = \frac{540}{2529} = 0.2136$$

Consequently, the calculated panel efficiency is 21.36%.

Subsequently, the total surface area of all solar panels (in m²) is determined as follows:

$$\text{Total}_{\text{area}} = \text{Area}_{1\text{panel}} \times \text{Number}_{\text{panels}} \quad (4.4)$$

$$\text{Total}_{\text{area}} = 2.529 \times 7920 = 20022.73 \text{ m}^2$$

Therefore, the total area is approximately 20,022.73 m².

Following this, the solar power output imputed using the Clear Sky Model can be mathematically formulated as follows:

$$P_{clear} = \frac{GTI \times \eta \times A}{1000} \quad (4.5)$$

Note: The result is divided by 1,000 to convert the power from watts (W) to kilowatts (kW).

Due to meteorological factors and system electrical losses, this research applies the K_c to adjust the values used for data filling to align with the actual behavior of the system, which is defined as follows:

$$K_c(t) = \frac{P_{real, no\ curtail}(t)}{P_{clear}(t)} \quad (4.6)$$

By choosing the median of K_c as a robust representative value resistant to outliers and setting the α_1 as follows:

$$\alpha_1 = median(K_c) \quad (4.7)$$

And calculating the value used to fill in the power generation data that corresponds to the actual behavior of the system, as shown in Equation 4.8.

$$P_{filled}(t) = \alpha_1 \times P_{clear}(t) \quad (4.8)$$

The power output data imputed using the Clear Sky Model is presented in Figure 4.3

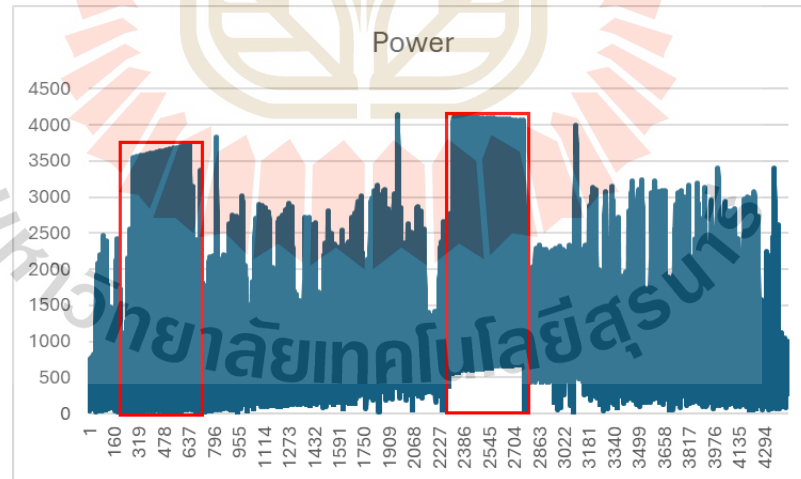


Figure 4.3 Power output data for the year 2023 after using Clear Sky model

4.2.2.2 KNN Imputation

The KNN Imputation method is a technique for filling in missing data values, suitable for cases where the missing data is not continuous or the data structure is complex. This method is based on the fundamental assumption that data

points with similar characteristics of the explanatory variable tend to have similar values of the target variable as well. Therefore, KNN imputation can maintain the relationships between variables within the dataset better than linear addition or representation by means alone.

In this research, the filling in of missing values in the power generation dataset uses meteorological variables, namely temperature, humidity, wind speed, and solar radiation values, as co-variables to calculate data similarity. The KNN imputation procedure is detailed as follows:

The first step is to define the parameter K , which represents the number of nearest neighbors used to estimate the missing data. The value of K directly affects the characteristics of the filled data. To select an appropriate parameter K , the performance of KNN Imputation was evaluated by randomly generating a model of 10% of the missing data rows in the power generation dataset. Then, the data was filled with different K values ranging from $K=1$ to 20 and larger K values: 30, 40, 50, 60, 70, 80, 90, and 100. The filled values were then compared to the actual temporarily missing values to evaluate the accuracy of data filling at each k value.

The accuracy of data filling was evaluated using error metrics: MSE, RMSE, MAE, and MAPE. Each metric reflects the characteristics of the error from a different perspective. MSE measures the average of the squares of the error between the actual and filled values. These are metrics that emphasize outliers. While RMSE and MSE, which are in the same units as the original data, facilitate easier interpretation of errors, MAE measures the mean absolute error and is less sensitive to outliers than MSE, providing a better overall picture of accuracy. MAPE shows the error in proportion to the true value. Therefore, evaluating the model's performance should consider these metrics together, rather than using only one metric.

To achieve a balanced evaluation and consider the performance of KNN Imputation in multiple dimensions, all metrics were adjusted to the same range using min-max normalization and calculated as a composite score. Each metric was assigned an equal weight to be used as a criterion for selecting the K

parameter that provides the best overall performance. Some of the K evaluation results are shown in Table 4.1, and all experimental results are shown in Appendix C.

Table 4.1 Comparison of KNN imputation performance across k values

KNN Imputer	Error Metrics				
	MSE (kW ²)	RMSE (kW)	MAE (kW)	MAPE (%)	
1	247137.63	497.13	356.22	48.15	
5	193440.98	439.82	318.37	45.86	
10	180856.18	425.27	313.28	50.21	
20	185949.87	431.22	320.46	55.24	
40	188538.09	434.21	320.73	56.47	
60	192038.22	438.22	323.56	58.92	
80	190994.01	437.03	322.64	61.52	
100	190967.32	437.00	323.45	63.41	
KNN Imputer	Normalized Metrics				Composite Score
	MSE	RMSE	MAE	MAPE	
	(kW ²)	(kW)	(kW)	(%)	
1	1.000	1.000	1.000	0.131	0.783
5	0.190	0.202	0.119	0.000	0.128
10	0.000	0.000	0.000	0.248	0.062
20	0.077	0.083	0.167	0.534	0.215
40	0.116	0.124	0.174	0.604	0.255
60	0.169	0.180	0.240	0.744	0.333
80	0.153	0.164	0.218	0.892	0.357
100	0.153	0.163	0.237	1.000	0.388

The evaluation results showed that K=10 yielded the lowest overall error score. Therefore, this K value was chosen as the parameter for filling in missing power generation data in the actual dataset.

After determining the value of K for filling in the missing data, the distance between the missing data points and the observed data points was calculated using the Euclidean distance, which is calculated from meteorological variables such as temperature, humidity, GTI, or a combination of these variables over the same or nearby time period. The number of p -factors is explained in Equation 4.9

$$d_i = \sqrt{\sum_{j=1}^p (z_{missing,j} - z_{i,j})^2} \quad (4.9)$$

Once the distances between each data point are determined, the K data points with the shortest distances are selected. The power generation values of neighboring data points are then used to estimate the missing values. In this research, an unweighted mean of neighboring data points is used, which can be written as Equation 4.10 as follows:

$$x_{missing} = \frac{1}{K} \sum_{i=1}^K x_i \quad (4.10)$$

The result of the power generation data point filled in using the KNN imputation method is shown in Figure 4.4.

มหาวิทยาลัยเทคโนโลยีสุรนารี

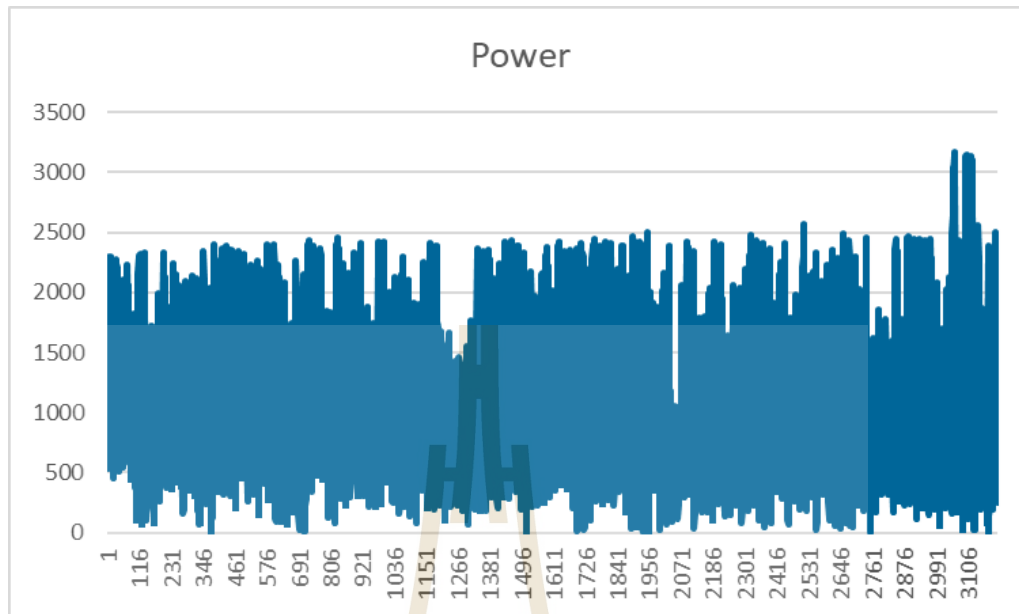


Figure 4.4 Power output data for the year 2023 after using KNN

Subsequently, the missing data points were filled in using the Clear Sky Model for summer and winter seasons and the KNN Imputation for the rainy season. These were then combined into a single dataset covering a one-year period. This data, filled in to reflect the characteristics of each season, will be used to prepare new data for the forecasting model development process in the next step.

To quantitatively evaluate the effectiveness of the two data augmentation methods, a comparison was made of baseline statistics, including mean, median, standard deviation, and interquartile range of the augmented power generation values, categorized by augmentation method. It was found that augmentation using ClearSky with KNN yielded higher baseline statistics than KNN-all. This reflects that referencing the Clear Sky model for augmentation improves the power generation to better reflect actual sunlight conditions, while maintaining clearer seasonal distribution and fluctuations. Overall, ClearSky with KNN maintained the seasonal pattern and actual power generation range better than KNN-all, as shown in Table 4.2 Therefore, the augmented data is more suitable for subsequent forecasting.

Table 4.2 Quantitative metrics

Metric	KNN-all	ClearSky+KNN
Mean	1279.71	1451.68
Median	1128.33	1356.49
SD	706.65	905.86
Min	0	0
Max	3242.40	4136.49
IQR	996.22	1304.81

To illustrate the characteristics of the prepared dataset, Figure 4.5 presents a three-dimensional surface plot of solar power generation capacity in 2023. The X-axis represents the timestep within a day, the Y-axis represents the day of the year, and the Z-axis represents the power output in kilowatts (kW), which is used as the target data for forecasting.

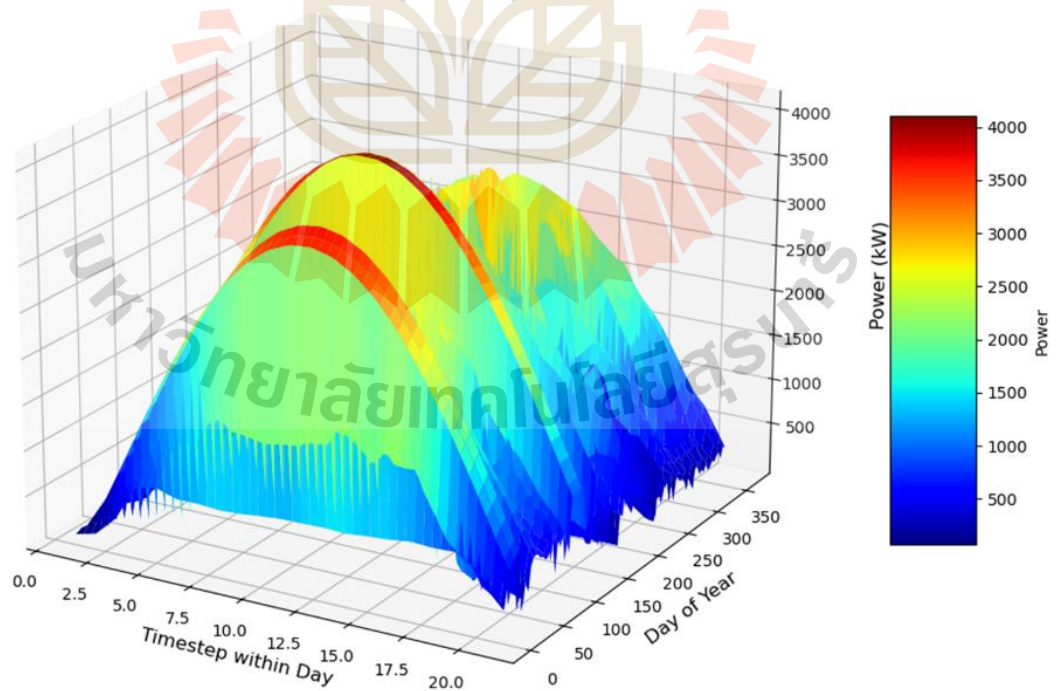


Figure 4.5 Three-dimensional visualization of solar power output in 2023 after ClearSky with KNN data processing

From the plot, a dome-shaped curve can be observed in the daily power output, with low initial power output in the morning, peaking at midday, and declining in the evening, reflecting the sunrise and sunset patterns. During the middle of the year (approximately 100–250 days), the surface plot appears yellow to red, indicating high power output, corresponding to the summer months with the highest sunlight intensity. Conversely, at the beginning and end of the year (< 100 and > 300 days), the surface plot is low and appears blue or green, indicating decreased power generation due to shorter daylight hours or increased cloud cover. Furthermore, some parts of the surface plot appear jagged or wavy, which may reflect the impact of erratic weather conditions such as clouds, shadows, or rain on certain days. The insights gained from analyzing this plot confirm the quality of the prepared data. It can also be used as a ready-to-use database for the next steps in developing and comparing the performance of forecasting models.

4.2.3 Feature Selection and Exploratory Statistical Analysis

Next, an EDA was performed using correlation analysis to identify factors influencing power output. These factors can be used to determine the input and output variables for the forecasting model. The results of the correlation heatmap are shown in Figure 4.6.

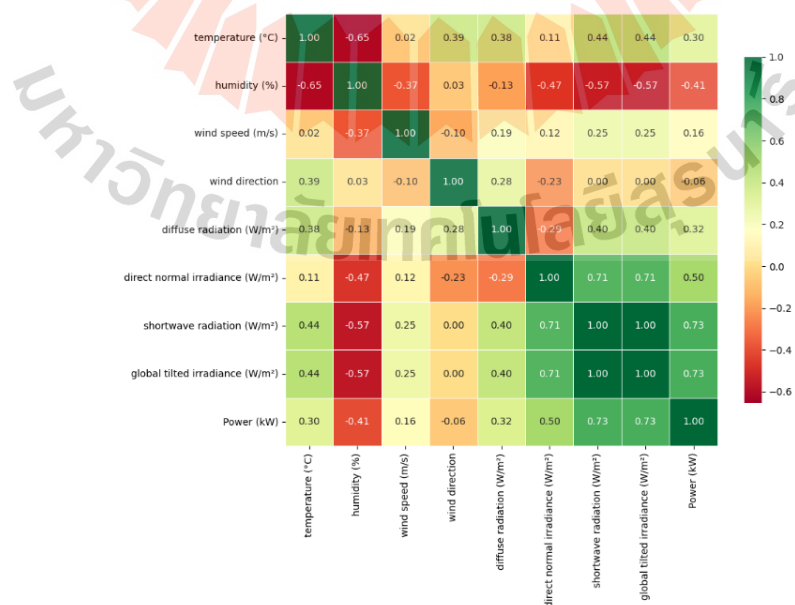


Figure 4.6 The correlation matrix

From the correlation heatmap analysis above, it can be seen that the correlation coefficient between light intensity and power (kW) is 0.73, reflecting that light intensity is the main factor determining the level of power production of the solar cell system. Meanwhile, the correlation coefficient between temperature and power (kW) is 0.30, which is an important factor affecting power output. Since the data from the station and the data loaded from the satellite web are not the same, temperature and light intensity were forecasted using the LSTM model to be used as inputs for the power output forecasting model in the next step.

To assess the quality of the input data after forecasting, the results were analyzed using a Taylor Diagram as shown in Fig. 4.7 This diagram compares the actual and predicted values of temperature and solar radiation, considering correlation, standard deviation, RMSE, and normalized bias. The results show a clear improvement in data quality after forecasting compared to the actual values. Temperature had a correlation coefficient of 0.63, while solar radiation had a high correlation coefficient of 0.74, with RMSE decreasing to 13.41°C and 184.44 W/m², respectively. This indicates a reduction in overall error and a closer alignment of the data distribution with the actual values.

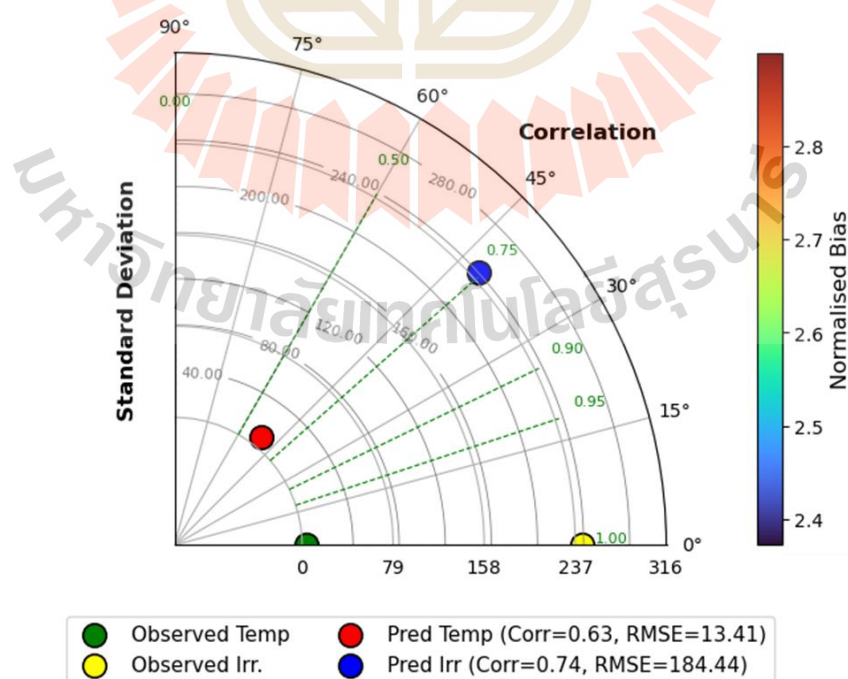


Figure 4.7 Taylor diagram of temperature and irradiance

However, the Taylor Diagram only reflects that the quality of the input data after forecasting is better compared to the actual values, but it is only an assessment of the consistency between the forecasted input and the actual input. Therefore, it cannot be directly concluded that the power generation forecast will be more accurate. The forecasted results were analyzed using SHAP, as shown in Fig. 4.8, to evaluate the factors affecting power generation. It was found that GTI remains the main factor affecting power generation the most, while the forecasted values, such as predicted temp. and predicted irr., demonstrate positive impacts consistent with the behavior of the solar cell system, reflecting the quality of post-forecast data for input in the power generation forecasting stage. Additionally, SHAP reveals non-linear relationships such as the temperature effect changing direction across value ranges and the saturation characteristics of solar radiation, which are not visible from basic statistics. This leads to a clearer understanding of the role of each variable in power generation. The input data includes predicted Irr., GTI (W/m^2), predicted temp., humidity, and wind speed (m/s).

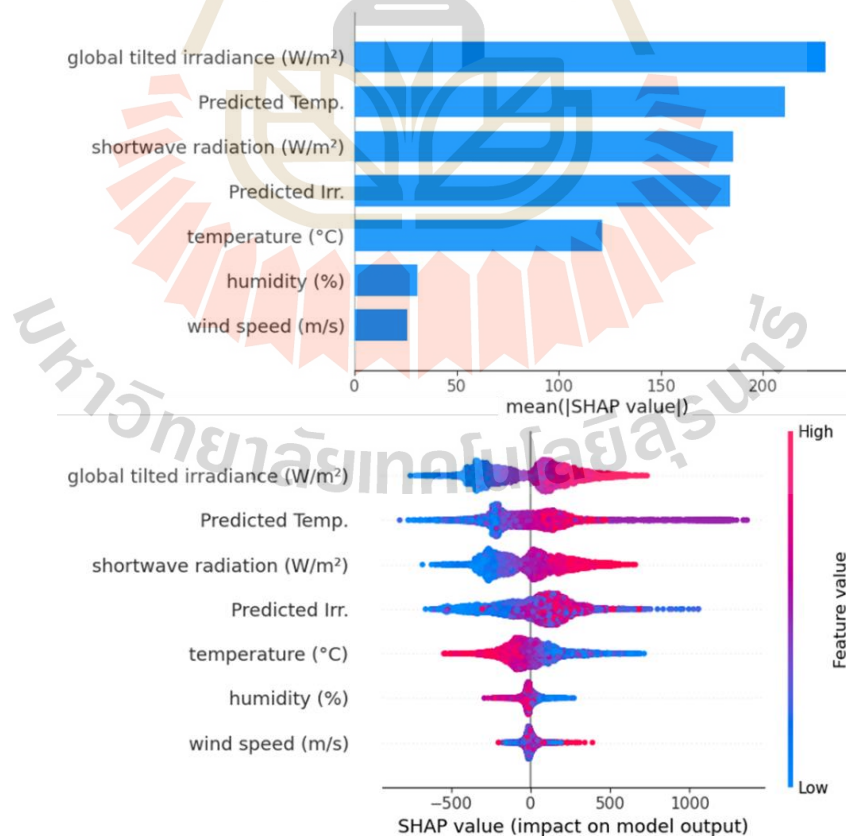


Figure 4.8 SHAP values impact on model output

4.2.4 Data Splitting

However, behavioral analysis of power generation data combined with internal system electrical load also revealed periods where solar cell system power generation was curtailed, meaning the output value did not reflect the system's true production potential even when weather conditions were favorable for power generation. Curtailment differs from missing data because it results from structural limitations of the electrical system, not data loss or measurement errors.

To systematically identify the periods of curtailment, this study considers the relationship between PV power generation, actual load, and metering load. The power-to-load ratio is defined by Equation 4.11, and the curtailment state is determined by conditional equations as shown in Equation 4.12, which considers both the condition where the power-to-load ratio is higher than the threshold obtained from statistical analysis of data throughout the year, and the condition where the metering load decreases significantly, in order to best reflect the actual operating behavior of the system. Examples of data behavior on days of curtailment and the distribution of curtailment behavior throughout the year are shown in Figures 4.9 and 4.10, respectively.

$$R(t) = \frac{P_{PV}(t)}{P_L(t)} \quad (4.11)$$

$$Curtailment(t) = \begin{cases} 1, & \text{if } R(t) \geq \alpha \text{ and } P_G(t) \leq \beta \\ 0, & \text{otherwise} \end{cases} \quad (4.12)$$

Where Both conditions must be satisfied simultaneously for the system to be classified as operating under a curtailment state.

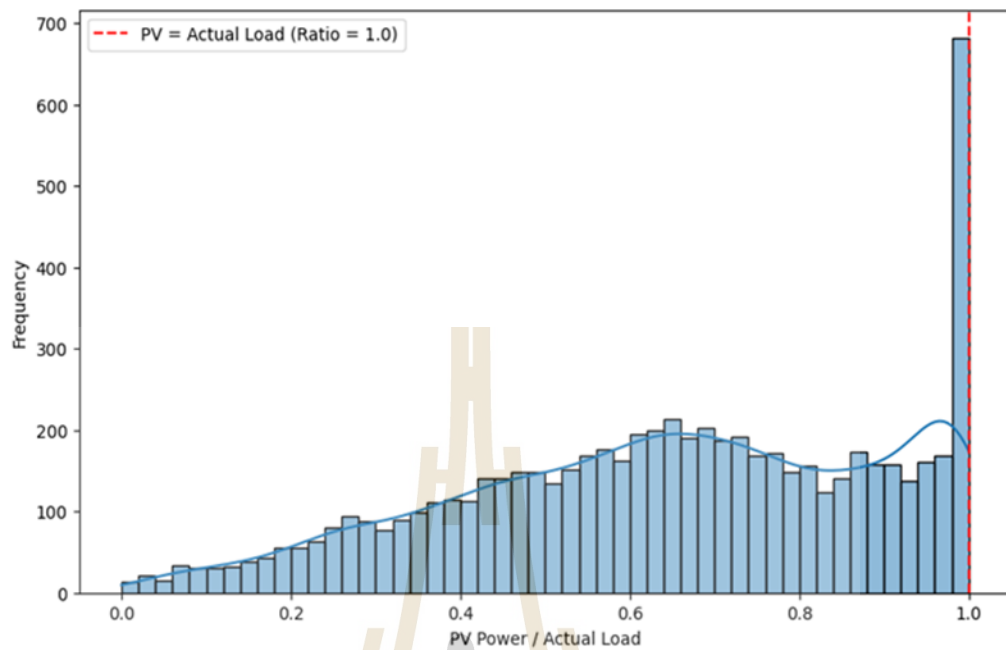


Figure 4.9 Distribution of PV power to actual load ratio

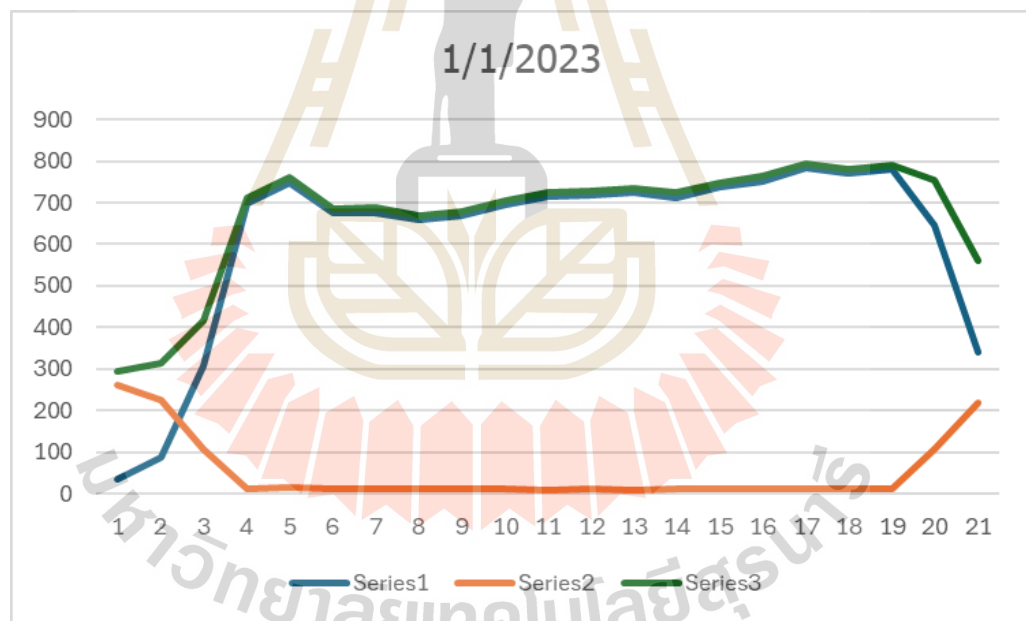


Figure 4.10 Solar power output under curtailment conditions

Figure 4.9 shows that the PV power output is close to the actual load, while the grid load drops close to zero throughout the day. This is consistent with the concentration of the PV power to actual load ratio near 1.0 throughout the day, as shown in Figure 4.10, reflecting a systematic limitation of PV power output.

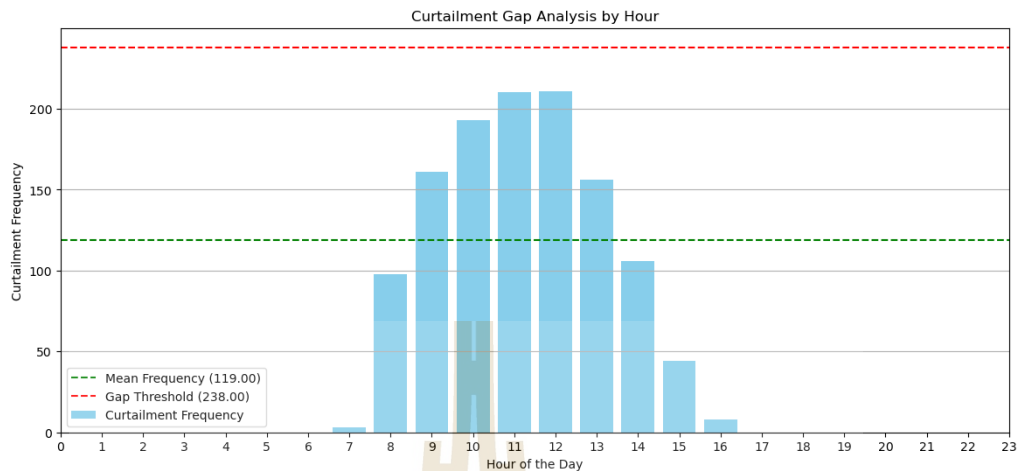


Figure 4.11 Curtailment gap analysis by hour

Figure 4.11 shows that temporal analysis reveals that curtailment begins around 8:00 AM and peaks between 10:00 AM and 1:00 PM, coinciding with the period of peak solar power output. The frequency then decreases below the average after approximately 2:00 PM and continues to decline until it is almost nonexistent after 3:00 PM–4:00 PM. These results demonstrate that curtailment is not random but follows a clear temporal pattern directly related to solar power generation potential.

Due to the specific characteristics of the curtailment data, all data was divided into three sets: data on days with curtailment, data on days without curtailment, and data for the entire year, to study the impact of curtailment on forecasting model performance. Each dataset is divided for training and validation in an 80/20 ratio. Five days of test data are used for both curtailment-existing and non-curtailment-existing sets, and ten days for the entire year's dataset, using the same dates, to ensure a fair comparison of model results.

Furthermore, each set is divided into three time periods of the day: the upward trend (07:00–08:30), the peak or curtailment period (08:30–14:30), and the downward trend (14:30–17:00) to clearly reflect the characteristics of curtailment.

To speed up the model training process and increase stability, all inputs are scaled to the same level using a normalization process. This allows the model to learn the relationships between variables stably and reduces problems arising from differences in scale between input features.

4.3 Forecasting Model Development and Evaluation

This section presents a two-step forecasting framework, a research approach used to develop and evaluate day-ahead forecasting models for solar power generation. The research process is divided into three steps: data preparation, two-step forecasting, and evaluation, detailed below.

This study utilizes LSTM and RF models for forecasting environmental factors and solar power generation. The data is time-series, nonlinear, and time-variable. LSTM's ability to learn both short-term and long-term temporal relationships makes it suitable for forecasting weather variables. Random Forest, an ensemble learning technique that utilizes bagging and feature randomness to construct multiple decision trees, reduces model variance, increases noise robustness, and efficiently learns nonlinear relationships between input variables and power generation. Therefore, it is used for forecasting power generation in the next step.

For configuring the structural parameter values of both models to suit the characteristics of the data used in the study, the details are shown in Table 4.3

Table 4.3 Hyperparameters of environmental factors pre-forecasting models

Model	Hyperparameters
LSTM	Input: 3, Hidden nodes: 128,64,32, Epochs: 200, Batch Size: 32, Learning Rate: 0.0001, Optimizer: Adam, Early Stopping: 10
RF	Input: 5, Number of Trees: 100, Max Depth: 10, Bootstrap: True, Criterion: MSE, Max Features: auto

To evaluate the performance of the forecasting models on each dataset, this study uses basic statistical indicators, including RMSE, MAE, and MAPE. These indicators are calculated using the Scikit-learn library in Python.

4.4 Results and Discussion

This section presents one-day FPV power generation forecast results from Suranaree University of Technology using a two-stage forecasting framework. This consists of forecasting environmental factors in the first stage and power generation forecasting in the next stage. The analysis results are discussed comparatively to evaluate the model's performance under different data types, as detailed below.

To improve the accuracy of solar power generation forecasting, this study uses an LSTM model to forecast environmental factors directly affecting power generation, namely solar irradiance and temperature. The relationship between various environmental factors (ambient radiation and temperature) and solar panel power generation in August is shown in Figure 4.12, and details for other months are shown in Appendix D.

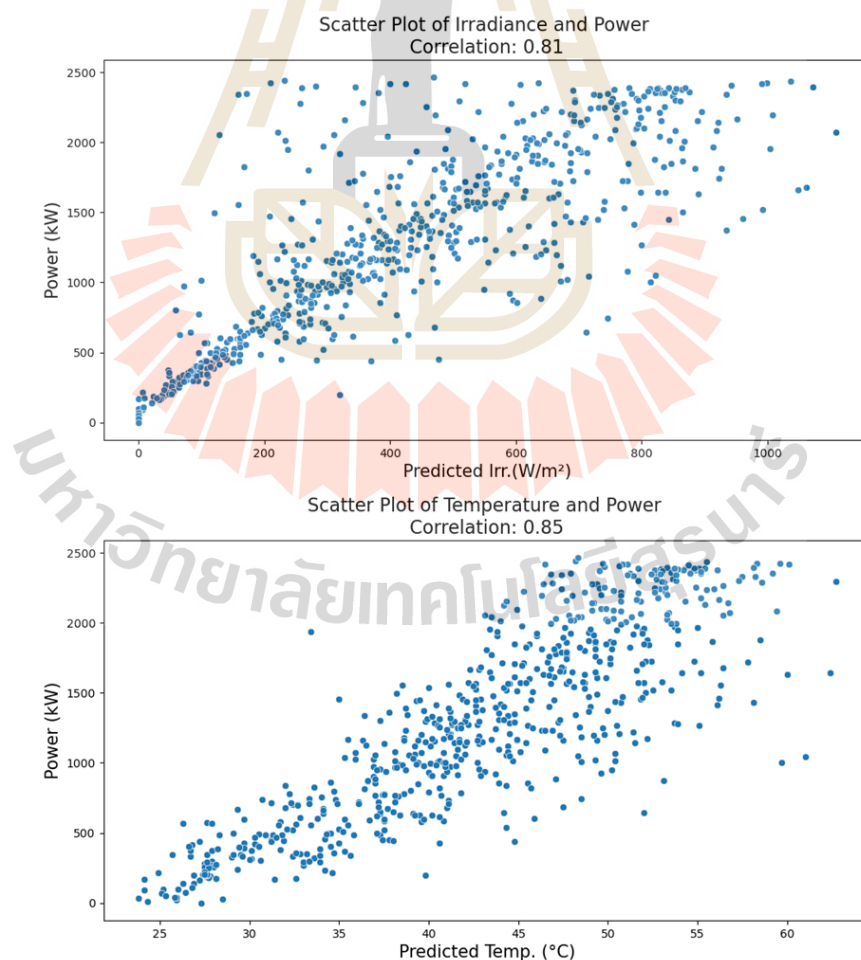


Figure 4.12 Correlation of irradiance and temperature with solar power

An analysis of the relationship between environmental factors and solar panel power generation in August revealed a correlation coefficient of 0.81 for solar intensity, reflecting a high positive correlation. This means that as light intensity increases, power generation tends to increase as well. However, some data dispersion was observed, possibly due to the influence of clouds, rain, and weather variability during the rainy season. Ambient temperature had a higher correlation coefficient of 0.85, indicating that temperature is another significant factor affecting power generation. In summary, the high correlation coefficients of both factors, approaching 1, suggest a clear positive relationship between environmental factors and power generation, supporting their suitability as input variables for forecasting power generation.

Furthermore, the MAE of the monthly forecasts for light intensity and temperature using the LSTM model are shown in Figure 4.13, presenting the results in two parts: light intensity and temperature, respectively.

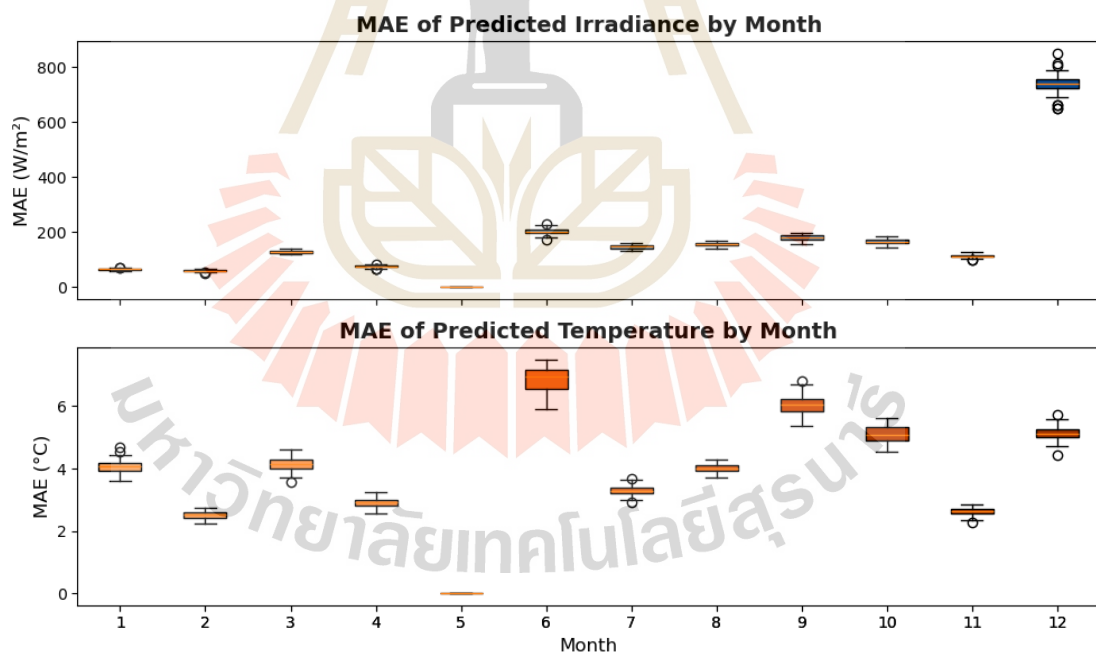


Figure 4.13 MAE of environmental factor forecasts

From the results of the monthly error analysis, it was found that the forecasts for solar irradiance and temperature showed clear differences. The MAE values for irradiance forecasts were mostly low and relatively constant throughout the year, but

in December, the MAE and dispersion were found to be higher, reflecting weather fluctuations. In May, the MAE was close to zero due to missing and incomplete data.

For temperature forecasting, it was found that most MAE values were relatively low and consistent, ranging from 2 to 5°C. However, in June and September, the MAE values were higher, approximately 6–7°C, due to sudden temperature changes during those months, which the model was not yet able to handle effectively.

However, the overall results show that the LSTM model can predict temperature more accurately and stably than solar irradiance and support the suitability of using environmental factor prediction results from LSTM as the first step in a two-stage prediction framework for FPV system power generation.

For the One-day PV energy forecast, the results are presented by 3 seasons: summer (mid-February to mid-May), rainy season (mid-May to mid-October), and winter (mid-October to mid-February). The actual power output is compared with the forecast power output for the entire year's dataset, divided into the three time periods of the day as mentioned above: the rising period, the peak period, or the curtailment period, and the falling period, respectively. The forecast for each time period is made separately for each day. The forecast results are then stacked to obtain the daily forecast power generation, making the difference between actual and forecast power generation clearly visible, as shown in Fig. 4.14

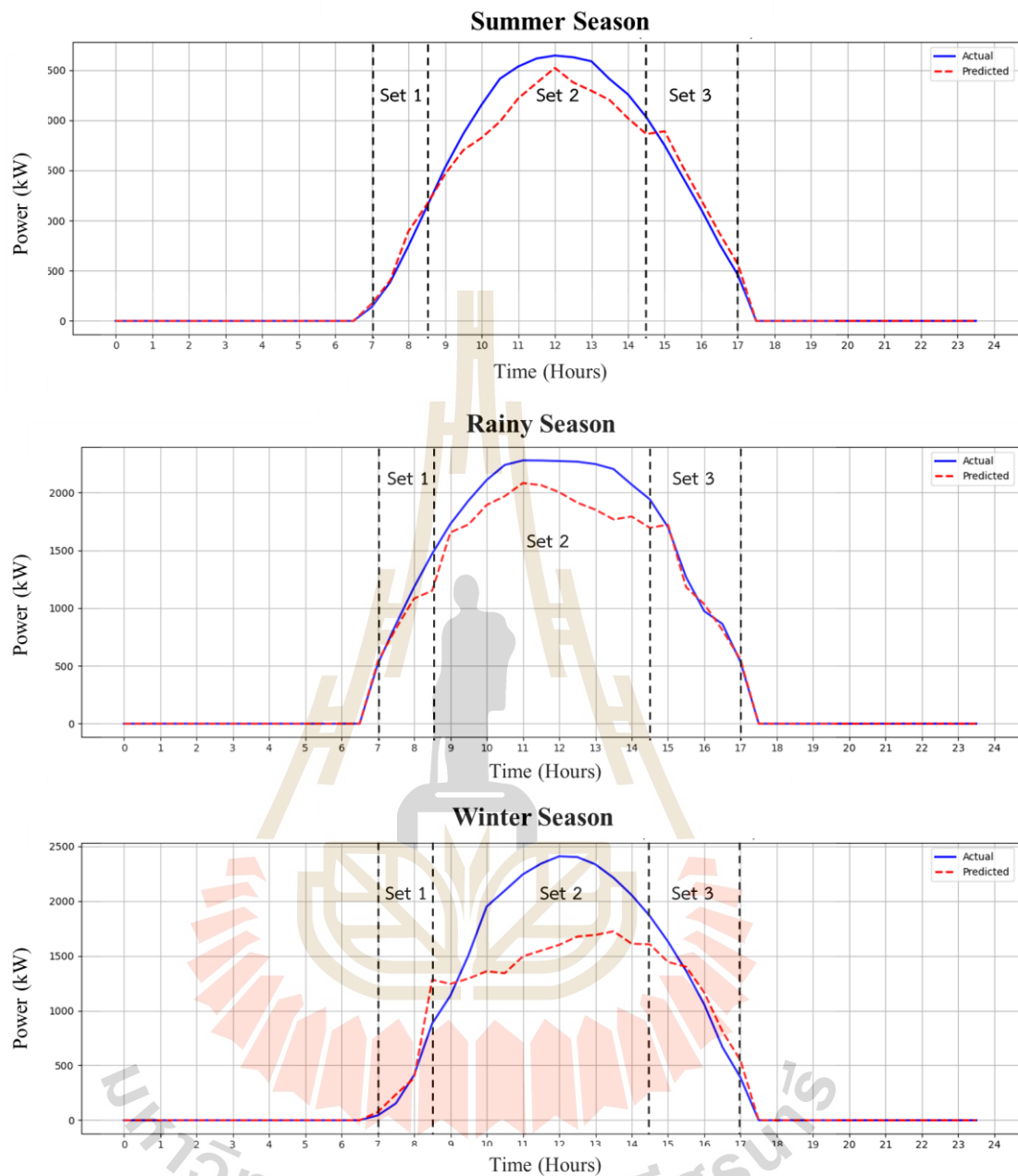


Figure 4.14 Seasonal comparison of day-ahead solar power forecasting

After calculating the power generation forecasting performance, the results are shown in Table III, displaying the RMSE, MAE, and MAPE values for each dataset in each season. It was found that the RF model could clearly learn different power generation behaviors depending on the dataset. In the summer, the Curtail and Entire datasets showed similar error values of 196 kW, 167 kW, and 11.63%, respectively, reflecting that the Curtail period can accurately represent the all-day production behavior

because the power generation is stable and the data is continuous. In contrast, the No Curtail dataset had values of 316 kW, 264 kW, and 16.87%, respectively, showing higher errors than the all-day dataset, which had values of 223 kW, 182 kW, and 10.16%, respectively. This indicates that learning only from periods without power limitations cannot adequately adapt to data fluctuations. Combining the all-day datasets reduced the error and allowed the model to better learn the relationship between Curtail and No Curtail periods. Similarly, in winter, the No Curtail period had values of 270 kW, 264 kW, and 37.57%, respectively, due to low power generation and complex data patterns resulting in high relative error. However, combining data throughout the day helped reduce the MAPE to 24.18%.

Furthermore, forecasting by dividing the data into three time periods each day and stacking the results to estimate daily energy consumption allows the model to efficiently learn the specific electricity production patterns of each time period. This clearly reflects the system's behavior during both curtail and non-curtail periods. Compared to the linear interpolation method used in the previous chapter, which employed an RF model to forecast daily energy without time period separation or curtailment as shown in Table 4.4, the After approach exhibits lower overall errors and consistently better result stability across all seasons. Stacking the results reduces time-specific errors and structural bias, resulting in higher accuracy daily energy forecasts and better generalization to varying weather and seasonal conditions.

In summary, electricity generation forecasting performance does not depend solely on the capabilities of the RF model but also significantly on the structure and timing of the data used to train the model. Using data that does not comprehensively cover the 24-hour power-limiting behavior tends to lead to errors and biases, especially in highly volatile conditions. A data management approach that considers both curtail and no-curtail periods, combined with time segmentation and data stacking, allows for more accurate, stable, and realistic daily energy forecasts that align with the actual usage of solar power systems.

Table 4.4 Forecasting results

Dataset	Summer		
	RMSE (kW)	MAE (kW)	MAPE (%)
Curtail	196	167	11.63
No Curtail	-	-	-
Entire	196	167	11.63
Dataset	Rainy		
	RMSE (kW)	MAE (kW)	MAPE (%)
Curtail	-	-	-
No Curtail	316	264	16.87
Entire	223	182	10.16
Dataset	Winter		
	RMSE (kW)	MAE (kW)	MAPE (%)
Curtail	-	-	-
No Curtail	270	264	37.57
Entire	434	326	24.18

Table 4.5 Comparison of forecasting results between before and after datasets across seasons

Model	Summer			Rainy			Winter		
	RMSE (kW)	MAE (kW)	MAPE (%)	RMSE (kW)	MAE (kW)	MAPE (%)	RMSE (kW)	MAE (kW)	MAPE (%)
RF*	260	188	11.35	429	344	18.78	166	148	38.97
RF**	196	167	11.63	223	182	10.16	434	326	24.18

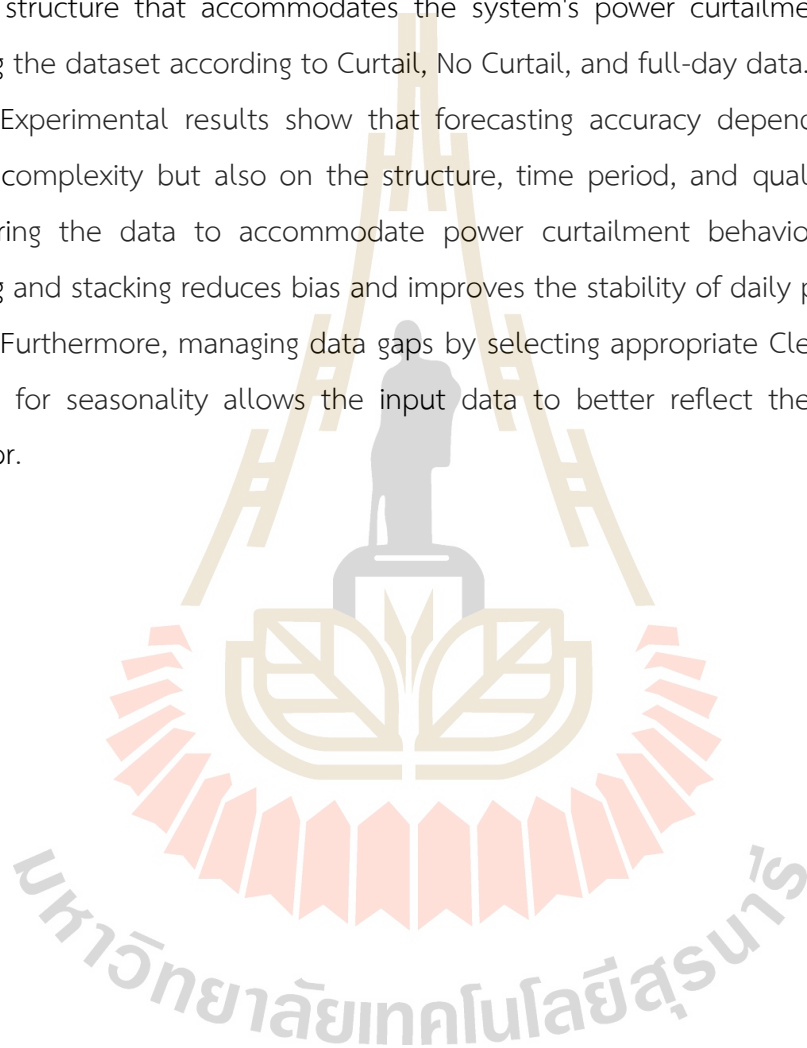
Note: *Chapter 3, **Chapter 4

4.5 Conclusion

This chapter presents a one-day forecasting framework for solar power generation using a two-step forecasting approach. The first step forecasts light intensity and temperature using an LSTM model, and the results are then used as input to a RF model to forecast power generation. The proposed framework emphasizes designing a data structure that accommodates the system's power curtailment behavior by splitting the dataset according to Curtail, No Curtail, and full-day data.

Experimental results show that forecasting accuracy depends not only on model complexity but also on the structure, time period, and quality of the data. Structuring the data to accommodate power curtailment behavior through data splitting and stacking reduces bias and improves the stability of daily power forecasts.

Furthermore, managing data gaps by selecting appropriate Clear-Sky and KNN models for seasonality allows the input data to better reflect the actual system behavior.



CHAPTER V

Data-Driven Solar Power Forecasting Using Pre-Forecasting and Clear-Sky/KNN Data Refinements³

5.1 Introduction

This chapter presents a research methodology for developing and evaluating a short-term (day-ahead) forecasting framework for electricity generation from floating solar power systems. It builds upon the analysis and findings of the previous chapters and includes the main topics: Case 1 and Case 2

5.2 Case 1: Augmented Data-Driven Solar Power Forecasting Using Pre-Forecasting and Clear-Sky/KNN Data Refinements

In this case, data quality management is presented through the Clear Sky and KNN models, and interpretability is enhanced through XAI. The first step uses LSTM and TFT-lite to forecast light intensity and temperature before the results are combined with Bi-LSTM, GRU, and RF to forecast power generation. This builds upon the research from Chapters 3 and 4. The research process is divided into three stages: data preparation, two-stage forecasting, and evaluation. The conceptual framework of the research is shown in Figure 5.1.

³ Part of this chapter has been accepted for publication in the GMSARN International Journal.

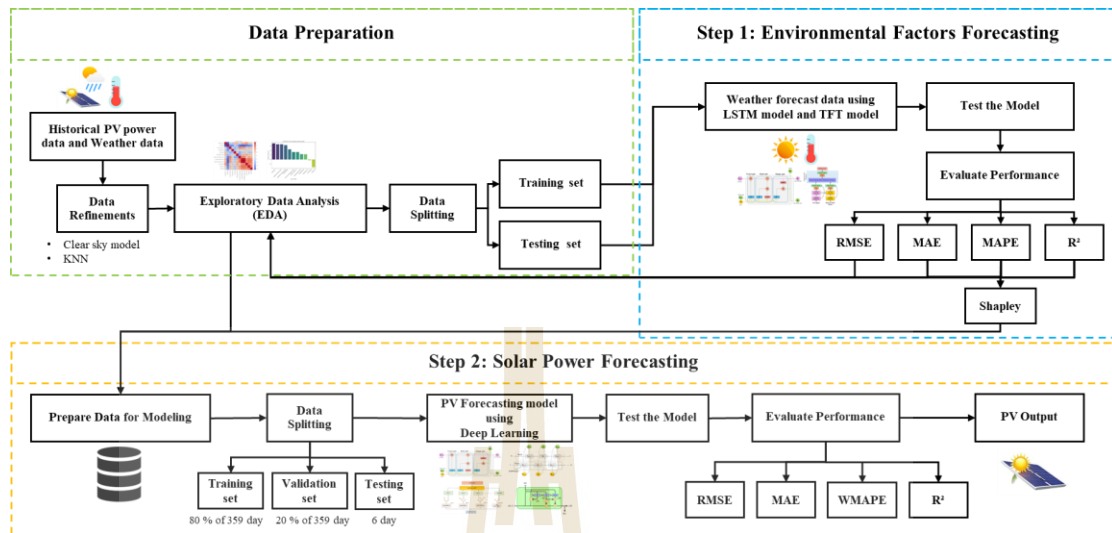


Figure 5.1 The framework for case 1

5.2.1 Data collection

The data management process diagram, as shown in Figure 5.2, comprises data collection, data refinements, EDA, and data splitting. The details are as follows:

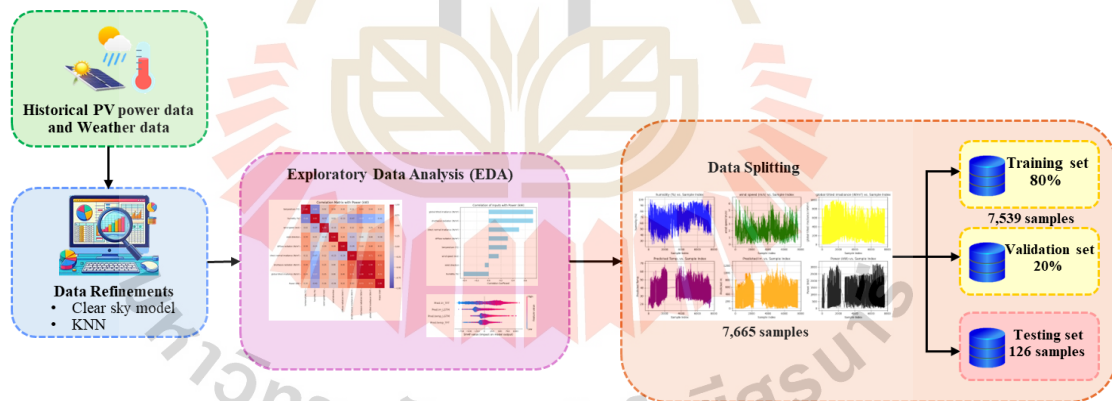


Figure 5.2 The framework for data management process

The dataset used in this study is the same as that in Chapters 3 and 4. After downloading the data, it was updated, a crucial step in preparing data for analysis and model development. The focus was on detecting and managing missing and anomaly values using the Clear Sky Model and KNN methods, as described in Chapter 4.

After analyzing the missing and anomaly data using the Clear Sky Model and KNN, statistical correlation analysis was performed using correlation heat maps to identify factors influencing power generation. The identified variables used as input and output data included temperature, humidity, wind speed, wind direction, diffuse radiation, DNI, GHI, predicted temperature, predicted radiation dose, and power (kW), as shown in Figure 5.3 in Chapter 4.

Based on the above correlation heat map analysis, temperature and light intensity were predicted using the LSTM and TFT-lite models. Analysis of the prediction results from the correlation heat maps showed that the correlation coefficient between temperature, light intensity, and power increased. Therefore, the prediction results were used as input data for the next step in forecasting power generation.

The input data consisted of the predicted correlation coefficient, GTI, predicted temperature, humidity, and wind speed. Therefore, the dataset was divided into three subsets: a sample of 7,539 samples was allocated to the training and validation sets (80% and 20%, respectively), and 126 samples were reserved for the test set.

After data preparation, a complete dataset was obtained, and all input features were standardized to a general scale, ready for use in developing and training predictive models in the next step.

5.2.2 Model development

The forecasting process is divided into two main steps (Two-Step Forecasting Approach) to improve the efficiency and accuracy of forecasting electricity production from solar cell systems. The details of each step are as follows:

5.2.2.1 Environmental Factors Pre-Forecasting

The solar radiation and temperature forecasting process is performed using LSTM and TFT-lite models. Although both models are widely recognized for time series forecasting, the selection for this research was based on their suitability for the data characteristics and limitations of the actual dataset. This is

because the LSTM model has the advantage of handling long-term relationships in time series data and its stability in learning highly continuous sequences of data, which are commonly found in weather data, such as temperature or seasonal trends in radiation. This technique allows for stable learning of gradual trends.

While TFT-lite models have a better ability to handle highly complex multivariate data, they also have the ability to automatically select important variables at each time point and learn both short-term and long-term relationships through attention mechanisms, making them suitable for highly volatile and multifactorial weather data.

A novel addition to this research is the application of the TFT-lite architecture. This not only simplifies the architecture from a full TFT but is also designed to be suitable for the real-world dataset of floating solar PV systems, which are smaller and have higher volatility than land-based solar PV systems. While maintaining the ability to capture temporal patterns and fully explain the importance of variables. The architecture and diagrams are shown in Figure 5.3.

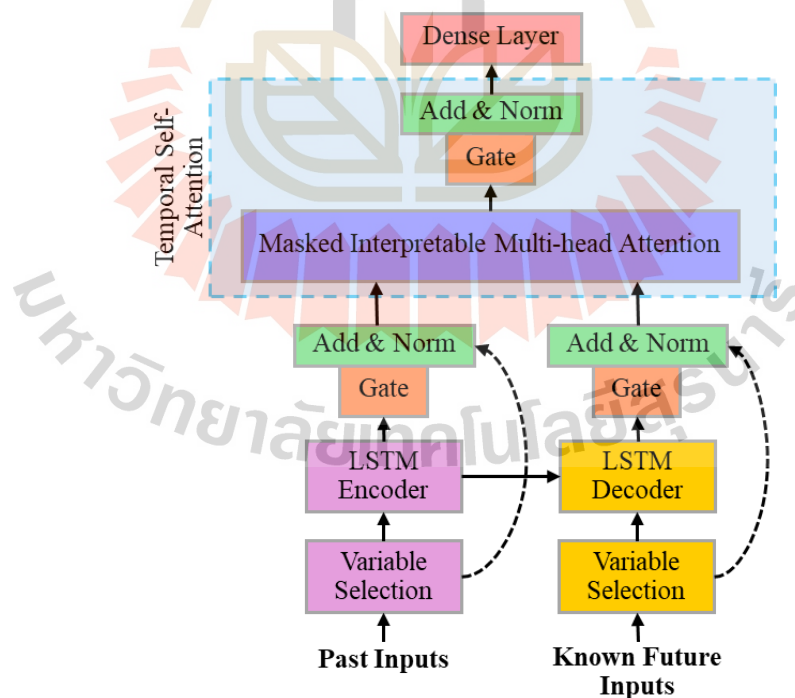


Figure 5.3. Architectural diagram of TFT-lite

5.2.2.2 PV-Forecast

This second step uses the forecasted solar irradiance and temperature from the pre-forecasting stage, together with GTI, relative humidity, and wind speed, as inputs to predict power generation using LSTM, Bi-LSTM, GRU, RF, and TFT-lite models. While these five models are standard models for power generation forecasting, their application in the context of floating solar data sets presents a unique role. Floating datasets exhibit higher intraday volatility than land-based solar, influenced by water surface humidity and temperature, as well as unstable cloud patterns. This gives LSTM and Bi-LSTM advantages in learning long-term relationships in erratic data. GRU is suitable for data with high volatility and requires a shorter training time. RF examines nonlinear relationships and variable significance, helping to clarify the structure of weather relationships. TFT-lite enhances the ability to select important variables over time and handle the overlapping multivariate data, a characteristic of real floating data, as mentioned above.

Therefore, the novelty of this work lies not in the creation of new models but in the evaluation of multiple model architectures with different strengths on real floating data, enabling concrete insights into which models best respond to which data types in this specific context.

Due to the unstable nature of the data and seasonal variations. For each model, a range of possible parameters was defined, such as the range of units in the hidden layer, the number of layers, and the range of learning rate values. The models were then run with multiple sets of different values and evaluated using the same metric to determine the one with the lowest forecast error. This approach enables the search for parameters that best fit the actual data behavior without relying on a single initial structure. It also reflects the model's learning characteristics and its response to seasonal data characteristics.

For the configuration of structural parameters for temperature and irradiance forecasting, both models are detailed in Table 5.1 For PV forecasting, the structural parameters of the LSTM, Bi-LSTM, and GRU models were kept consistent during training, whereas the RF and TFT-lite models used different configurations. The detailed hyperparameters for all PV-forecasting models are summarized in Table 5.2

Table 5.1 Hyperparameters of environmental factors pre-forecasting models

Model	Hyperparameters range	Optimum hyperparameters
LSTM	Input: [1 5], Hidden nodes: [32 128]x3 Epochs: [100 200], Batch Size: [16 32], Learning Rate: [0.0001 0.001], Early Stopping: [10 20]	Input: 3, Hidden nodes: 128,64,32, Epochs: 200, Batch Size: 32, Learning Rate: 0.0001, Early Stopping: 10
TFT-lite	Input: [1 5], Hidden nodes: [32 128]x3 Epochs: [100 200], Batch Size: [16 32], Learning Rate: [0.0001 0.001], Early Stopping: [10 20]	Input: 3, Hidden nodes: 64, 64, 32, Epochs: 200, Batch Size: 32, Learning Rate: 0.0001, Early Stopping: 10

Table 5.2 Hyperparameters of PV-forecasting models

Model	Hyperparameters range	Optimum hyperparameters
LSTM	Input: [1 8], Hidden nodes: [32 128]x3 Epochs: [100 200], Batch Size: [16 32], Learning Rate: [0.0001 0.001], Early Stopping: [10 20]	Input: 5, Hidden nodes: 64,32,50, Epochs: 100, Batch Size: 32, Learning Rate: 0.001, Early Stopping: 20

Table 5.2 Hyperparameters of PV-forecasting models (Continued)

Model	Hyperparameters range	Optimum hyperparameters
Bi-LSTM	Input: [1 8], Hidden nodes: [32 128]x3 Epochs: [100 200], Batch Size: [16 32], Learning Rate: [0.0001 0.001], Early Stopping: [10 20]	Input: 5, Hidden nodes: 64,32,50, Epochs: 100, Batch Size: 32, Learning Rate: 0.001, Early Stopping: 20
GRU	Input: [1 8], Hidden nodes: [32 128]x3 Epochs: [100 200], Batch Size: [16 32], Learning Rate: [0.0001 0.001], Early Stopping: [10 20]	Input: 5, Hidden nodes: 64,32,50, Epochs: 100, Batch Size: 32, Learning Rate: 0.001, Early Stopping: 20
RF	Input: [1 8], Number of Trees: [50 100] Max Depth: [10 20]	Input: 5, Number of Trees: 100, Max Depth: 10
TFT-lite	Input: [1 8], Hidden nodes: [32 128]x4 Epochs: [100 200], Batch Size: [16 32], Learning Rate: [0.0001 0.001], Early Stopping: [10 20]	Input: 5, Hidden nodes: 128,128,64,128, Epochs: 100, Batch Size: 16, Learning Rate: 0.0001, Early Stopping: 30

In the final step, metrics are selected to evaluate the performance of the model in forecasting solar power generation and to measure the accuracy and reliability of the forecasts, enabling the results to be used effectively. These metrics can be accessed using the Sklearn library in Python. In this study, four performance functions were used: RMSE, MAE, WMAPE, and R^2 . The equations for RMSE and MAE are shown in the previous chapter, and the next index considered is WMAPE, a measure of the accuracy of the forecasting model that adjusts the weights according to the actual values of the target variable, with the aim of reducing the sensitivity to low or zero values, which is a limitation of MAPE, shown in Equation 5.1.

$$WMAPE = \frac{\sum_{i=1}^n |y_{actual,i} - y_{forecast,i}|}{\sum_{i=1}^n y_{actual,i}} \times 100\% \quad (5.1)$$

The last index is R^2 , which measures the model's ability to explain the variance of the target variable, as shown in Equation 5.2.

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_{actual,i} - y_{forecast,i})^2}{\sum_{i=1}^n (y_{actual,i} - \bar{y}_{actual})^2} \quad (5.2)$$

In addition, the model performance was further evaluated by using the ME, SD, and 95% CI. Using these metrics, we can reflect both the “level of model error” through the mean error and the “stability of the forecast results” through the spread of the SD and the width of the CI interval, as shown in Equations 5.3-5.5, respectively.

$$ME = \frac{1}{n} \sum_{i=1}^n y_{forecast,i} - y_{actual,i} \quad (5.3)$$

$$SD = \sqrt{\frac{1}{n-1} \sum_{i=1}^n ((y_{forecast,i} - y_{actual,i}) - ME)^2} \quad (5.4)$$

$$CI_{95\%} = ME \pm 1.96 \times \frac{SD}{\sqrt{n}} \quad (5.5)$$

5.2.3 Results and discussion

The experimental results were presented in three main parts: forecasting results of environmental factors, forecasting results of PV power generation, and discussion. The details are as follows.

5.2.3.1 Forecasting Results of Environmental Factors

From the forecasting results of light intensity and temperature using the LSTM and TFT-lite models, the MAE of the monthly light intensity and temperature forecasts can be shown in Figures 5.4 and 5.5, respectively. The results are presented separately in two parts: light intensity and temperature, respectively.

Monthly error analysis revealed significant differences in irradiance and temperature forecasts. MAE for irradiance forecasts was mostly low and relatively stable throughout the year, but in December, the MAE and its distribution increased, reflecting weather volatility. In May, the MAE was near zero due to missing and incomplete data. Considering the results from the TFT-lite model, a deep model capable of handling temporal relationships and external variables, a similar trend in monthly error was observed, particularly the increase in MAE in the middle of the year. This reflects that weather uncertainty is a structural constraint on solar irradiance forecasting rather than a limitation of any particular model.

For temperature forecasts, the MAE was mostly relatively low and consistent, ranging from 2 to 5°C. However, in June and September, the MAE increased to approximately 6–7°C, consistent with the TFT-lite model's results showing higher errors during months with sudden temperature changes. This comparison indicates that, despite the potential of advanced models like TFT-lite to learn complex data structures, temperature forecasting remains more stable than solar irradiance forecasting.

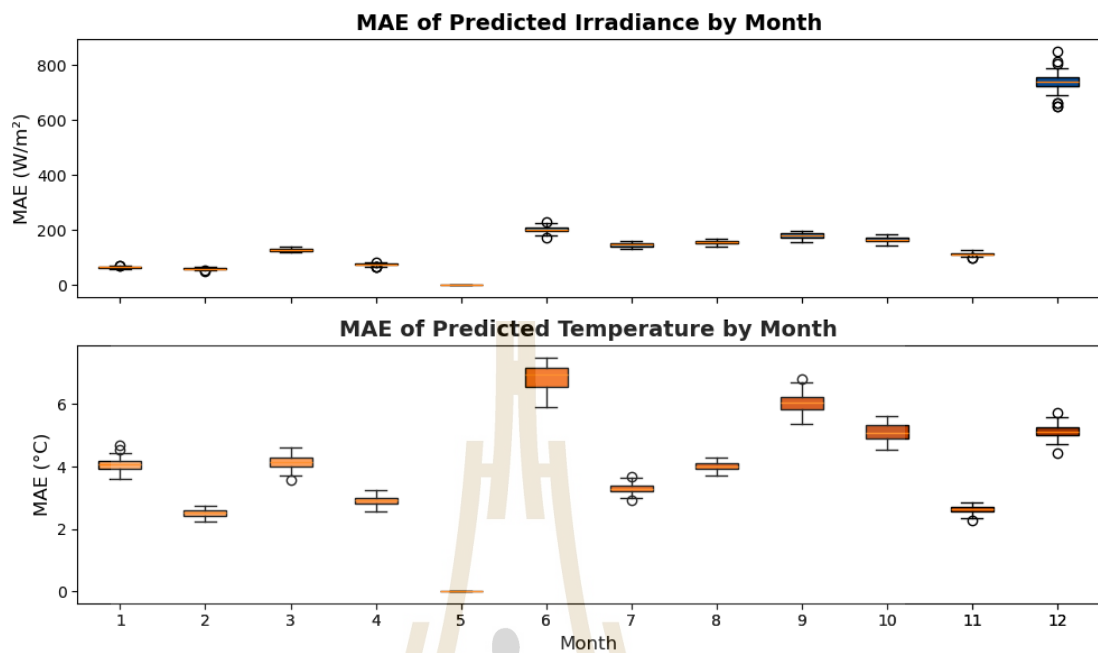


Figure 5.4 MAE of environmental factor forecasts using the LSTM

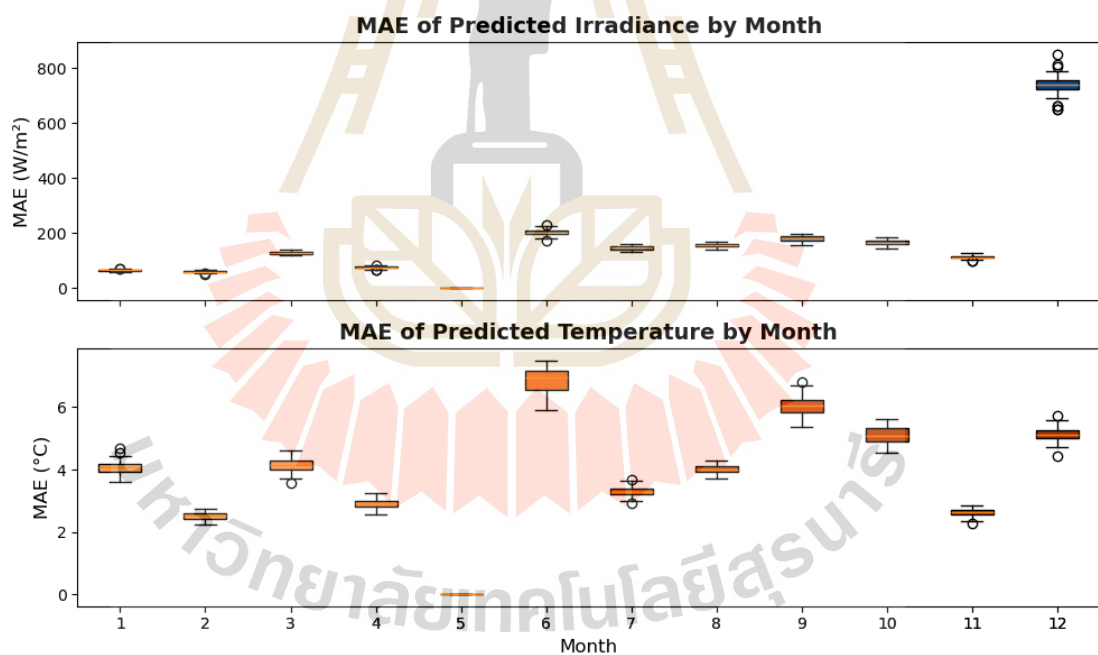


Figure 5.5 MAE of environmental factor forecasts using the TFT-lite

After forecasting solar irradiance and temperature using the LSTM and TFT-lite models, the prediction results were analyzed through a correlation heatmap to identify factors influencing power generation. The analysis indicated that

both models produced similar forecasts, with correlations to power output of 0.63 and 0.62 for temperature and 0.56 and 0.50 for solar irradiance, respectively.

Nevertheless, correlation analysis alone may not fully capture the direct influence of forecasted variables on power generation due to the nonlinear nature of real-world phenomena. To address this, SHAP analysis was conducted alongside the correlation heatmap to evaluate feature importance based on SHAP values using a nonlinear reference model, as illustrated in Fig. 5.6 and Fig. 5.7

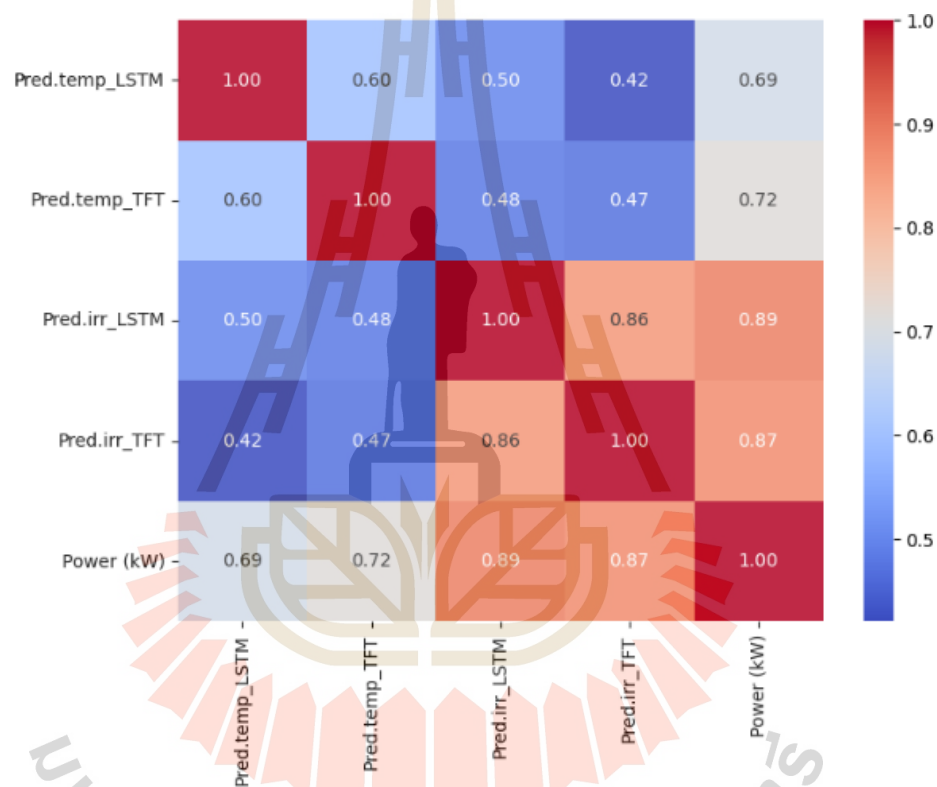


Figure 5.6 The parameters correlation matrix after SHAP process

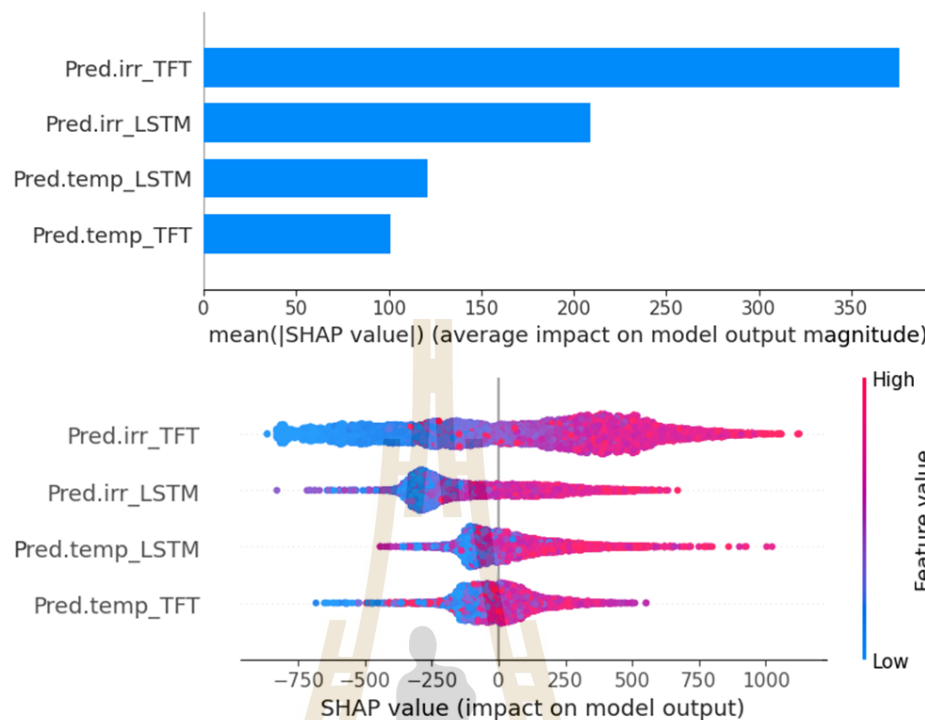


Figure 5.7 SHAP values impact on model output

Analysis of the correlation heatmaps shows that the predicted irradiance from both the LSTM and TFT-lite models is highly correlated with power, with values of 0.89 and 0.87, respectively, while the predicted temperature has a moderate correlation of approximately 0.69 and 0.72, respectively, which is consistent with the physical mechanism of the solar cell system that directly depends on the light intensity. Meanwhile, temperature still has a secondary effect on the power generation, which is consistent with the behavior of decreasing the efficiency of the panel when the temperature increases.

When compared to the correlation heatmap between the forecast data and power generation directly, the values from the SHAP analysis combined with the correlation heatmap are higher, indicating that these variables have a greater influence on power generation capacity than correlation alone can reflect, which helps to understand the impact of the variables in detail and covers both linear and non-linear aspects.

Considering the absolute mean of SHAP, the predicted irradiance from the TFT-lite model is the highest, followed by the predicted irradiance from the

LSTM model. The mean SHAP temperature of the LSTM model is slightly higher than the TFT-lite model. This result shows that the irradiance-related input has the highest influence in a reasonable manner consistent with the physics of the system and also reflects that the values obtained from TFT-lite have higher quality in the angle where the model is actually used, not just a linear statistical relationship.

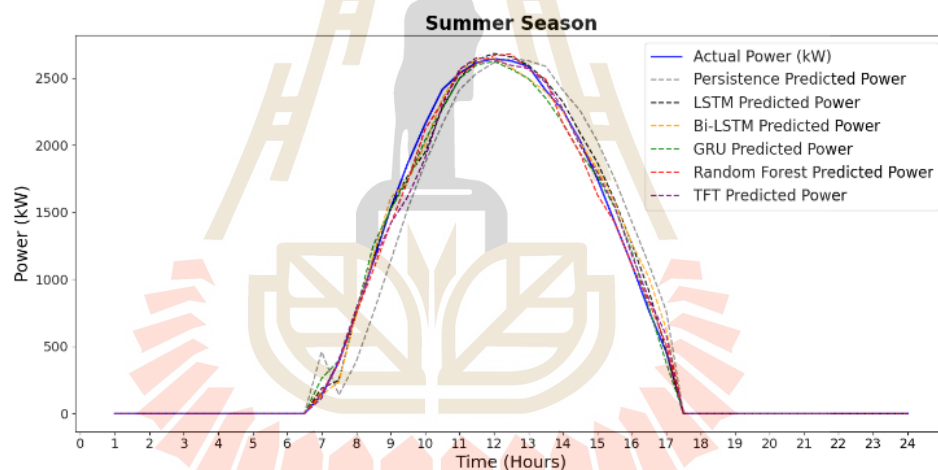
Furthermore, the SHAP summary plot results indicate that high light intensity (red) consistently induces positive SHAP, leading to increased power generation with increasing light intensity. Meanwhile, low light intensity (blue) induces negative SHAP, reflecting the direct physical behavior of solar cells. Similar to temperature, which has both positive and negative effects depending on the temperature, this suggests that high temperatures may reduce panel efficiency, but moderate temperatures can support power generation. The distribution of SHAP effects is broader, though not as pronounced as solar radiation. Furthermore, the SHAP distribution of Pred.irr_TFT is more consistent and has a higher influence on the model than that of Pred.irr_LSTM. This suggests that using TFT-lite predicted irradiance values can accurately and continuously capture solar radiation patterns, enabling power generation forecasting models to better understand causal relationships, not just statistical ones. Using SHAP summary plots not only demonstrates numerical significance but also provides a systematic and detailed understanding of the mechanisms and physical effects of input variables on power generation.

In summary, the analysis confirms that solar irradiance is the primary factor driving the fluctuations in power generation from solar panels. Although temperature plays a secondary role, it remains an important variable for model performance. These results can support feature selection to enhance both the efficiency and interpretability of the forecasting models. Furthermore, the comparison between the correlation heatmap using predicted irradiance and predicted temperature from the LSTM model as inputs and the SHAP analysis combined with the correlation heatmap using predictions from both the LSTM and TFT-lite models as inputs indicates that both input sets are suitable for power generation forecasting. Consequently, both datasets were utilized as inputs for subsequent forecasting, and their results will be compared in the following step.

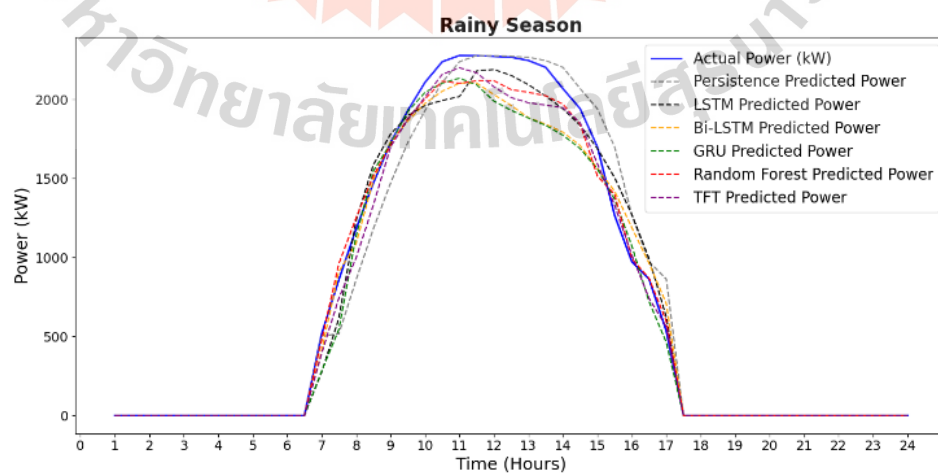
5.2.3.2 Forecasting Results of PV Power Generator

Since this is a one-day PV energy forecast, the results are presented in three seasons: summer, rainy season, and winter. A comparison is made between the actual power output and the forecast power output of the five models mentioned above. Furthermore, the results are compared with the persistence model, a basic model that uses previous power generation values as a forecast baseline to assess efficiency and clearly demonstrate the model's ability to adapt to seasonal variation.

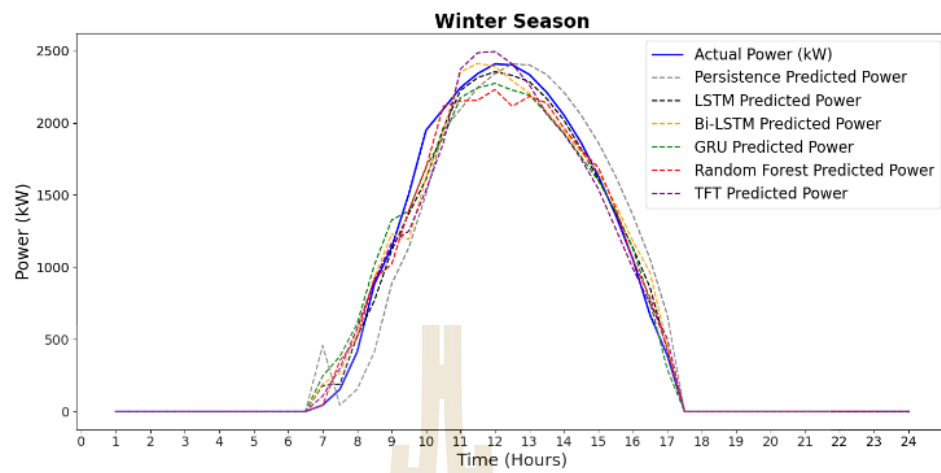
Figures 5.8 and 5.9 show the comparison between actual and forecasted power outputs for the five models, based on datasets containing predicted irradiance and temperature from the LSTM&TFT and LSTM&LSTM models, respectively. The results are presented for the three seasons: summer (a), rainy (b), and winter (c).



(a)

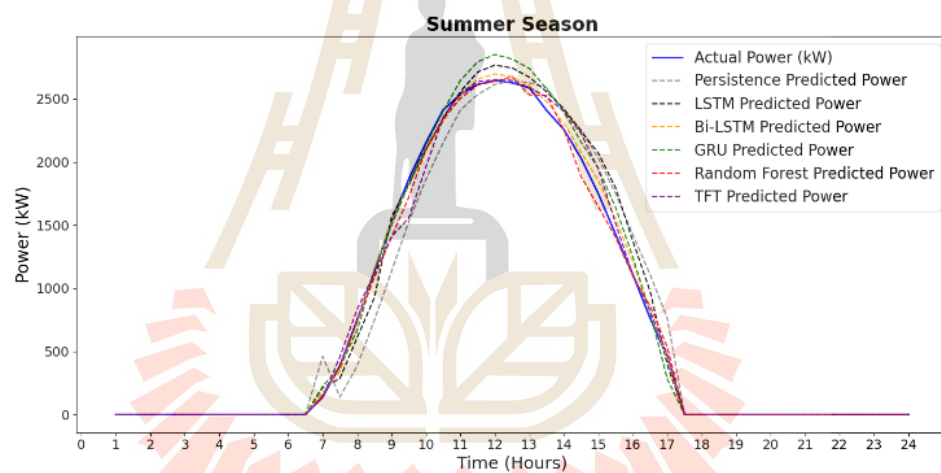


(b)

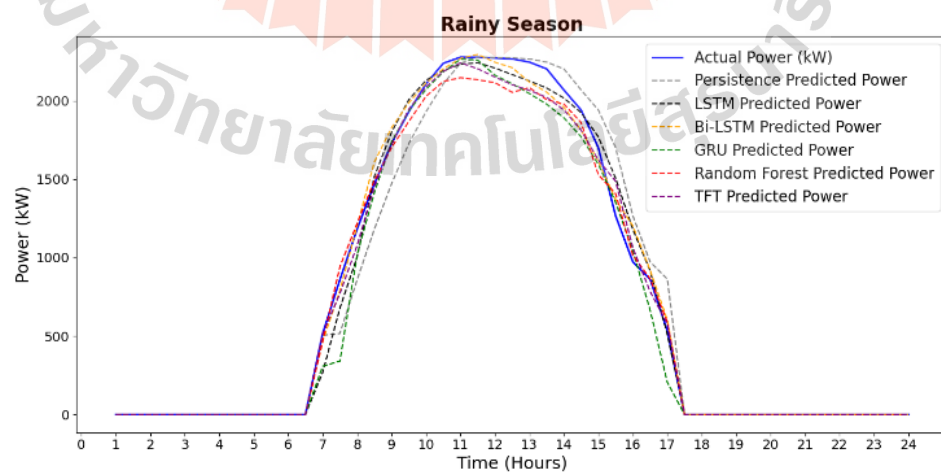


(c)

Figure 5.8 Comparison of day-ahead solar power forecasting of the LSTM&TFT models: (a) summer season (b) rainy season and (c) winter season



(a)



(b)

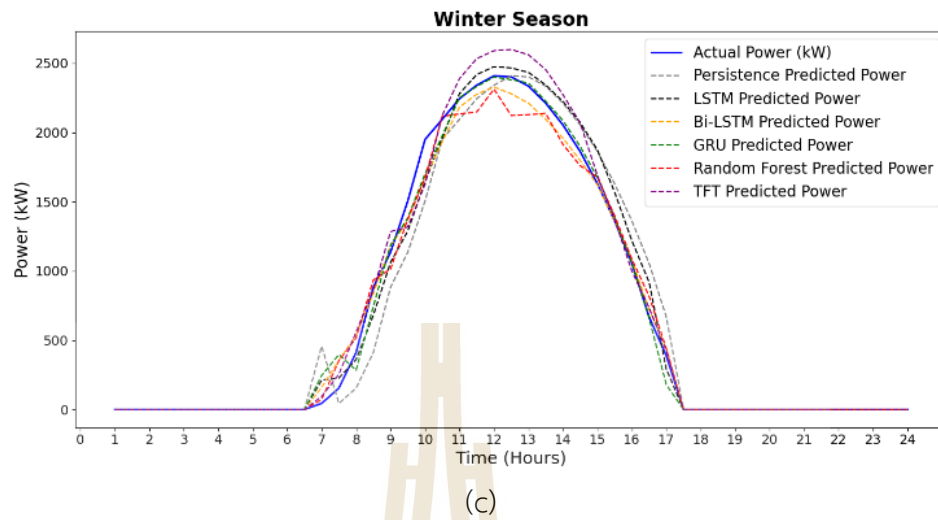


Figure 5.9 Comparison of day-ahead solar power forecasting of the LSTM&LSTM models: (a) summer season (b) rainy season and (c) winter season

Based on the validation metrics obtained from comparing the predicted and actual power outputs, the results of both test sets are shown in Table 5.3 and Table 5.4, respectively, showing the RMSE, MAE, WMAPE, and R^2 values. From the analysis of the results in Table 5.3, the RF method is 67.61 kW, 52.49 kW, 3.10%, and 0.99, respectively, which is the best forecasting method among the five methods for summer considering the four validation factors. In contrast, in the LSTM&LSTM dataset from Table 5.4, the Bi-LSTM method is the best forecasting method among the five methods, with the values of 59.13 kW, 49.79 kW, 2.94%, and 0.99, respectively. The same is true in the rainy season, with the values of 95.55 kW, 78.48 kW, 4.72%, and 0.98, respectively. However, for the dataset in Table 5.3, the RF forecasting results are the best in the rainy season, which are 123.01 kW, 106.02 kW, 6.38%, and 0.96, respectively. In contrast, for the first winter dataset, the best prediction method was LSTM with values of 110.94 kW, 81.53 kW, 5.50%, and 0.98, respectively, and for the next dataset, the GRU method had relatively accurate values of 114.99 kW, 80.61 kW, 5.44%, and 0.98, respectively.

Furthermore, experimental results show that the persistence model exhibits significantly higher errors across all seasons, while the proposed model consistently yields lower RMSE and WMAPE values and higher R^2 values. This demonstrates superior performance compared to traditional forecasting methods.

Comparing the two-step framework with the one-step framework for forecasting power generation using an LSTM model, the error of the one-step LSTM is higher than that of the two-step LSTM for both datasets. This indicates that the two-step forecasting results are more accurate, clearly demonstrating the improvement achieved by the proposed two-step method.



Table 5.3 Forecasting results with dataset 1 from LSTM&TFT pre-forecasting models

Model	Summer season			
	RMSE (kW)	MAE (kW)	WMAPE (%)	R ²
LSTM*	191.33	172.66	10.19	0.95
Persistence**	259.32	230	13.70	0.90
LSTM**	97.27	75.53	4.46	0.99
Bi-LSTM**	104.25	78.66	4.64	0.98
GRU**	82.69	69.22	4.09	0.99
RF**	67.61	52.49	3.10	0.99
TFT-lite**	82.46	63.47	3.75	0.99
Model	Rainy season			
	RMSE (kW)	MAE (kW)	WMAPE (%)	R ²
LSTM*	230.67	205.99	12.39	0.86
Persistence**	374.85	260.04	16.21	0.63
LSTM**	164.91	143.82	8.65	0.93
Bi-LSTM**	196.72	169.94	10.22	0.90
GRU**	208.77	174.59	10.50	0.88
RF**	123.01	106.02	6.38	0.96
TFT-lite**	180.13	154.99	9.32	0.91
Model	Winter season			
	RMSE (kW)	MAE (kW)	WMAPE (%)	R ²
LSTM*	224.83	206.08	13.90	0.92
Persistence**	254.5842	214.44	14.82	0.88
LSTM**	110.94	81.53	5.50	0.98
Bi-LSTM**	152.21	127.26	8.58	0.96
GRU**	147.21	134.27	9.06	0.96
RF**	130.98	106.02	7.15	0.97
TFT-lite**	146.39	129.51	8.73	0.97

Note: * Without pre-forecasting, **With pre- forecasting

Table 5.4 Forecasting results with dataset 2 from LSTM&LSTM pre-forecasting models

Model	Summer season			
	RMSE (kW)	MAE (kW)	WMAPE (%)	R ²
LSTM*	191.33	172.66	10.19	0.95
Persistence**	259.32	230	13.70	0.90
LSTM**	167.35	139.01	8.21	0.96
Bi-LSTM**	59.13	49.79	2.94	0.99
GRU**	132.27	111.28	6.57	0.97
RF**	73.95	56.99	3.36	0.99
TFT-lite**	116.02	81.89	4.83	0.98
Model	Rainy season			
	RMSE (kW)	MAE (kW)	WMAPE (%)	R ²
LSTM*	230.67	205.99	12.39	0.86
Persistence**	374.85	260.04	16.21	0.63
LSTM**	119.43	95.97	5.77	0.96
Bi-LSTM**	95.55	78.48	4.72	0.98
GRU**	183.19	138.67	8.34	0.91
RF**	114.82	98.02	5.90	0.96
TFT-lite**	106.90	87.27	5.25	0.97
Model	Winter season			
	RMSE (kW)	MAE (kW)	WMAPE (%)	R ²
LSTM*	224.83	206.08	13.90	0.92
Persistence**	254.5842	214.44	14.82	0.88
LSTM**	163.07	144.89	9.77	0.96
Bi-LSTM**	114.42	94.08	6.34	0.98
GRU**	114.99	80.61	5.44	0.98
RF**	139.10	116.60	7.86	0.97
TFT-lite**	160.38	135.48	9.14	0.96

Note: * Without pre-forecasting, **With pre- forecasting

In addition, the evaluation results using mean error, SD, and CI showed that model performance differed significantly across seasons, reflecting both model accuracy and stability. Considering the width of the CI range as statistical evidence, these results can be summarized in Tables 5.5 and 5.6. The analysis of the results showed that in summer, the first dataset showed the most stable RF with narrow SD and CI values, while the second dataset, Bi-LSTM, had the lowest error and the narrowest CI range among all models. This is consistent with the fact that summer has a relatively uniform solar radiation pattern, allowing the Bi-LSTM temporal model that learns both past and future directions to perform best. In the rainy season, both datasets exhibited the widest CI, reflecting the fluctuations in cloud and rainfall that result in high data uncertainty. The RF model showed good stability on the first dataset due to its tree-based structure that copes well with nonlinearity from sudden changes. However, as data quality improved on the second dataset, Bi-LSTM emerged as the most stable model, indicating that its bidirectional temporal learning capability significantly mitigates the impact of intra-rainy season variability. While in winter, SD and CI values are moderate compared to other seasons, indicating that the data fluctuate less but are not as consistent as in summer. The first dataset shows that several models have similar performance, with the LSTM model having the lowest mean error and narrower CI than the other models. Meanwhile, in the second dataset, the GRU model has a lower mean error and narrower CI than the other models, reflecting that the gated recurrent model can handle data patterns with mixed seasonality and moderate variation more efficiently. The results for each season can be summarized as in Table 5.7

Table 5.5 Accuracy and stability of LSTM&TFT pre-forecasting models with dataset 2 under difference seasons

Model	Summer season				
	ME (kW)	SD (kW)	CI Lower (kW)	CI Upper (kW)	CI Width (kW)
LSTM	33.75	56.18	17.69	49.81	32.12
Bi-LSTM	35.15	60.82	17.76	52.53	34.78
GRU	30.93	46.30	17.69	44.17	26.47
RF	23.45	39.05	12.29	34.62	22.33
TFT-lite	31.78	58.93	14.93	48.62	33.69
Model	Rainy season				
	ME (kW)	SD (kW)	CI Lower (kW)	CI Upper (kW)	CI Width (kW)
LSTM	64.26	90.53	38.38	90.14	51.77
Bi-LSTM	75.93	108.52	44.91	106.96	62.05
GRU	78.01	116.96	44.57	111.45	66.88
RF	47.37	67.93	27.95	66.79	38.84
TFT-lite	53.87	76.44	32.01	75.72	43.71
Model	Winter season				
	ME (kW)	SD (kW)	CI Lower (kW)	CI Upper (kW)	CI Width (kW)
LSTM	36.43	65.29	17.76	55.09	37.33
Bi-LSTM	56.86	85.28	32.48	81.24	48.76
GRU	59.99	78.84	37.45	82.53	45.08
RF	47.37	74.43	26.09	68.65	42.56
TFT-lite	55.50	86.63	30.73	80.27	49.53

Table 5.6 Accuracy and stability of LSTM&LSTM pre-forecasting models with dataset 1 under difference seasons

Model	Summer season				
	ME (kW)	SD (kW)	CI Lower (kW)	CI Upper (kW)	CI Width (kW)
LSTM	62.11	94.06	35.22	89.00	53.78
Bi-LSTM	22.25	33.02	12.81	31.69	18.88
GRU	49.72	73.90	28.59	70.85	42.25
RF	25.46	42.82	13.22	37.71	24.49
TFT-lite	36.59	69.12	16.83	56.35	39.52
Model	Rainy season				
	ME (kW)	SD (kW)	CI Lower (kW)	CI Upper (kW)	CI Width (kW)
LSTM	42.88	68.06	23.42	62.34	38.92
Bi-LSTM	35.07	53.96	19.64	50.49	30.85
GRU	61.96	106.76	31.44	92.48	61.05
RF	43.80	63.71	25.58	62.01	36.43
TFT-lite	38.99	60.52	21.69	56.30	34.61
Model	Winter season				
	ME (kW)	SD (kW)	CI Lower (kW)	CI Upper (kW)	CI Width (kW)
LSTM	64.74	88.64	39.40	90.08	50.69
Bi-LSTM	42.03	64.58	23.57	60.50	36.93
GRU	36.02	68.63	16.40	55.64	39.24
RF	52.10	77.84	29.84	74.35	44.51
TFT-lite	60.53	89.44	34.96	86.10	51.14

Table 5.7 Best forecasting models of year 2023 data

Models	Best Model	
	Pre-forecast data 1	Pre-forecast data 2
	(LSTM&TFT)	(LSTM&LSTM)
Summer	RF	Bi-LSTM
Rainy	RF	Bi-LSTM
Winter	LSTM	GRU

Table 5.8 Runtime of PV forecasting models

Model	Time (s)	
	Pre-forecast data 1	Pre-forecast data 2
	(LSTM&TFT)	(LSTM&LSTM)
LSTM	134.07	163.57
Bi-LSTM	182.18	100.68
GRU	113.73	80.17
RF	6.27	5.94
TFT-lite	602.94	620.59

However, the processing time ranking, as shown in Table 5.8, indicates that for the first dataset, the models with the shortest computation times are RF, GRU, LSTM, Bi-LSTM, and TFT-lite, with times of 6.27 s, 113.73 s, 134.07 s, 182.18 s, and 602.94 s, respectively. For the second dataset, the models are ranked according to increasing processing time as RF, GRU, Bi-LSTM, LSTM, and TFT-lite, with times of 5.94 s, 80.17 s, 100.68 s, 163.67 s, and 620.59 s, respectively. These results are consistent with the theoretical expectations described above.

The simulation results indicate that each model has different strengths depending on the data characteristics at each time period. RF and Bi-LSTM perform well for summer power generation forecasting because during summer, weather changes lead to consistent power generation capacity. This results in

continuous and less fluctuating data for model training. The model can efficiently learn the relationships between these features.

Similarly, during the rainy season, RF and Bi-LSTM perform best. Due to the uncertain data characteristics and rapid power generation capacity changes, RF can effectively handle the uncertainty and capture non-hierarchical patterns. Bi-LSTM, on the other hand, utilizes historical and forward data, enabling it to learn temporal correlation patterns and reduce errors. In contrast, for winter, the most suitable models are LSTM and GRU. This is because power generation is influenced by many factors, such as shorter daylight hours, low temperature, and high humidity, leading to complex and rapidly changing data characteristics. LSTM and GRU models, which are hierarchical neural networks, can efficiently capture the hierarchy of data and learn the relationships between multiple variables. GRU, with its simpler structure, is therefore faster to train. When comparing the error values between the two data sets, the pre-forecast data set 2 (LSTM and LSTM) provides more accurate forecasting results. This observation clearly indicates that the quality of irradiance and temperature data forecasted in the first step directly affects the accuracy of the power generation forecast, both in terms of error level and consistency of the forecast over time. Furthermore, the causes of power generation forecast errors may arise from missing data for some of the forecasts and environmental factors that may directly affect the efficiency of the solar panels.

5.2.3.3 Discussion

Overall, the analysis results reflect the fundamental characteristic of FPV systems: power generation is still primarily dependent on solar radiation intensity. While temperature plays a role in panel efficiency, it is not the most significant predictor of fluctuations. This trend is evident in both the correlation heatmap and the SHAP absolute average, which are consistent, even though the two approaches approach the data from different perspectives. A heatmap describes a linear relationship, while SHAP reflects the non-linear influence of inputs on model decisions, further validating the link between weather conditions and power generation. Considering both approaches together, there is a strong consensus that

irradiance significantly influences power generation. Even when the correlation is lower than SHAP, this suggests that the relationship may be non-linear, such as short-term extreme highs or lows caused by cloud movement. This is consistent with the actual behavior of FPV.

Although the SHAP analysis in the first step indicated that the irradiance forecast from TFT-lite exhibits a signal quality that is consistent with physical laws and more consistent than LSTM, experimental results show that using LSTM irradiance forecasts results in more accurate power generation forecasts in the second step, across multiple time periods and seasons. This point reflects the specific nature of two-stage forecasting systems, where the outcome depends not only on the accuracy of the first-stage model but also on the error structure and the compatibility between the errors of the two models.

Considering the seasonal results, the model accuracy clearly reflects the data characteristics. Summer has a constant irradiance, which makes RF and Bi-LSTM perform well due to the relatively uniform weather conditions, making it easy for the model to learn the relationship between variables. Meanwhile, the rainy season has high volatility and uncertainty. Therefore, RF, with its flexible structure, still has advantages in handling uncertainty and non-order patterns. Bi-LSTM can learn the temporal order and the backward-forward pattern, thus reducing the error better than other models. Winter is the most complex case because it is affected by many factors simultaneously. Therefore, sequential models such as LSTM and GRU can well capture the temporal order and the relationship between multiple variables. GRU trains faster and is suitable for small data sets. From the results, it can be technically interpreted that RNN models, including LSTM, Bi-LSTM, and GRU, are naturally suitable for data with a temporal structure. Bi-LSTM, in particular, can use information in both backward and forward directions, which supports volatile data and reduces the error better than other models. Meanwhile, RF, which does not rely on a temporal structure, can handle highly volatile or irregular data well, thus performing highly well in both summer and rainy seasons, although it cannot learn the temporal association like RNN-based models. TFT-lite, although highly accurate, comes at the cost of time and computing

resources. This makes it suitable for applications requiring high accuracy and sufficient resources.

A comparison of two pre-forecast datasets confirms that the quality of the first-stage irradiance and temperature forecasts directly affects the second-stage power generation, both in terms of error magnitude and signal continuity. Errors often arise from missing data or meteorological errors, particularly the May data outage, which resulted in an unusually low MAE. This suggests that model estimates may be distorted by data quality rather than the actual model capability if preprocessing is not carefully examined.

In summary, the results of this discussion clearly indicate that power generation forecasting cannot be solely based on model accuracy. It must also be considered systematically, encompassing the quality of the forecast inputs, internal model mechanisms, seasonal variations, and the spatial and environmental constraints of the FPV system. However, since all data are from a single site, this approach may require parameter adjustments or additional variable selection when applied to other areas with varying climates, cloud patterns, and radiation angles. These findings contribute to a more comprehensive physical and data-driven understanding of the system, which is beneficial for smart power system planning, such as load management, real-time power mix control, and backup power management in dynamic real-world environments.

5.3 Case 2: Base case (Delete Missing Data)

This case demonstrates one method of data management by removing missing values from the dataset and considering only the periods with complete data, which are defined as the baseline reference case. This results in 5,700 complete cells, as shown in Figure 5.10.

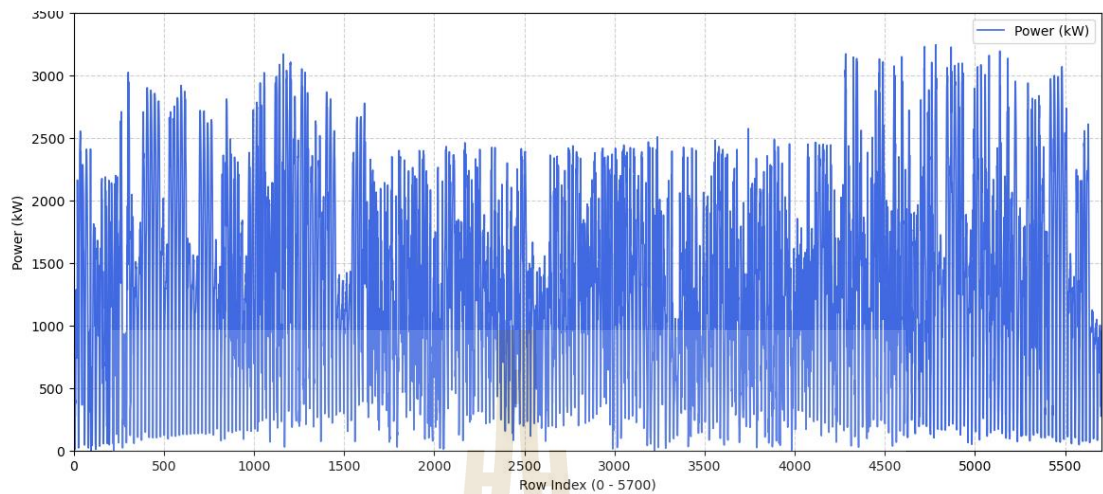


Figure 5.10 Power output data for the year 2023 after delete missing data

To illustrate the characteristics of the prepared dataset, Figure 5.11 presents a three-dimensional surface plot of solar power generation capacity in 2023. The X-axis represents the time step within a day, the Y-axis represents the day of the year, and the Z-axis represents the power output in kilowatts (kW), which is used as the target data for forecasting.

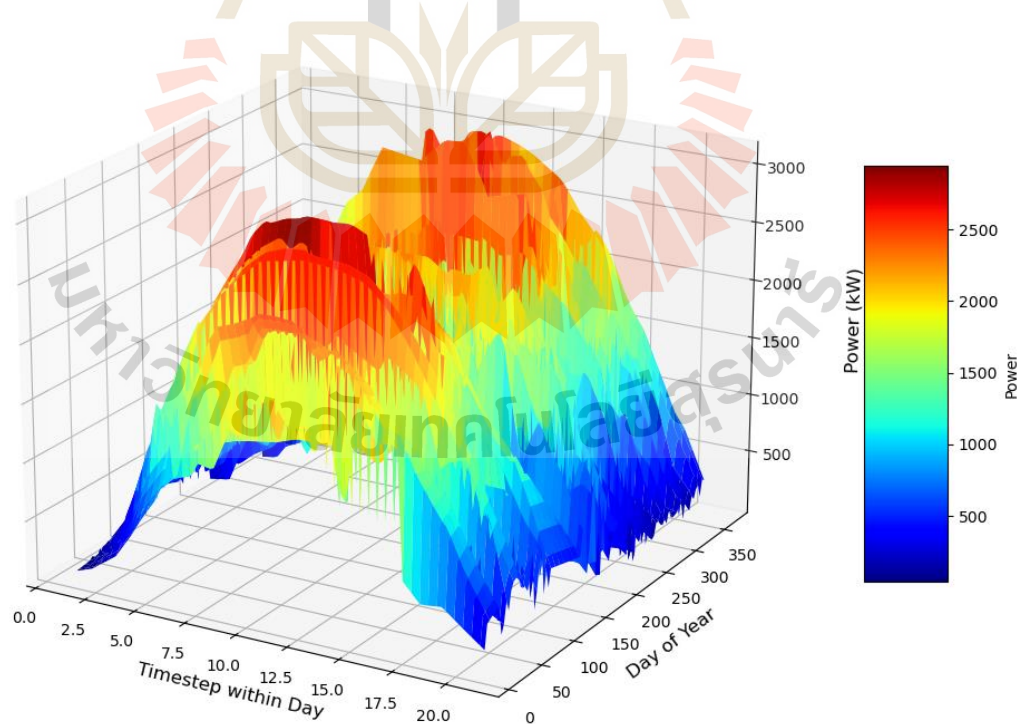


Figure 5.11 Three-dimensional visualization of solar power output in 2023 after delete missing data

Figure 5.11 shows the electricity production pattern of solar cell systems throughout 2023. Some data was processed using the Deleted Data method (removing missing data). Overall, the pattern of increasing production in the morning (peaking at midday) and decreasing in the afternoon-evening, as well as seasonal changes with peaks concentrated in the middle of the year, is still visible. However, data deletion creates temporal discontinuities, appearing as gaps or breaks in the graph surface, which reduces the completeness of the data sequence and causes the observation of short-term fluctuations (such as cloud cover) to be lost or incomplete.

After data cleaning, the data was analyzed using the methods previously described in the previous chapter, and light intensity and temperature were forecasted using LSTM and TFT-lite models, respectively. MAE of the monthly light intensity and temperature forecasts are shown in Figures 5.12 and 5.13, respectively, displaying the results in two parts: light intensity and temperature, respectively.

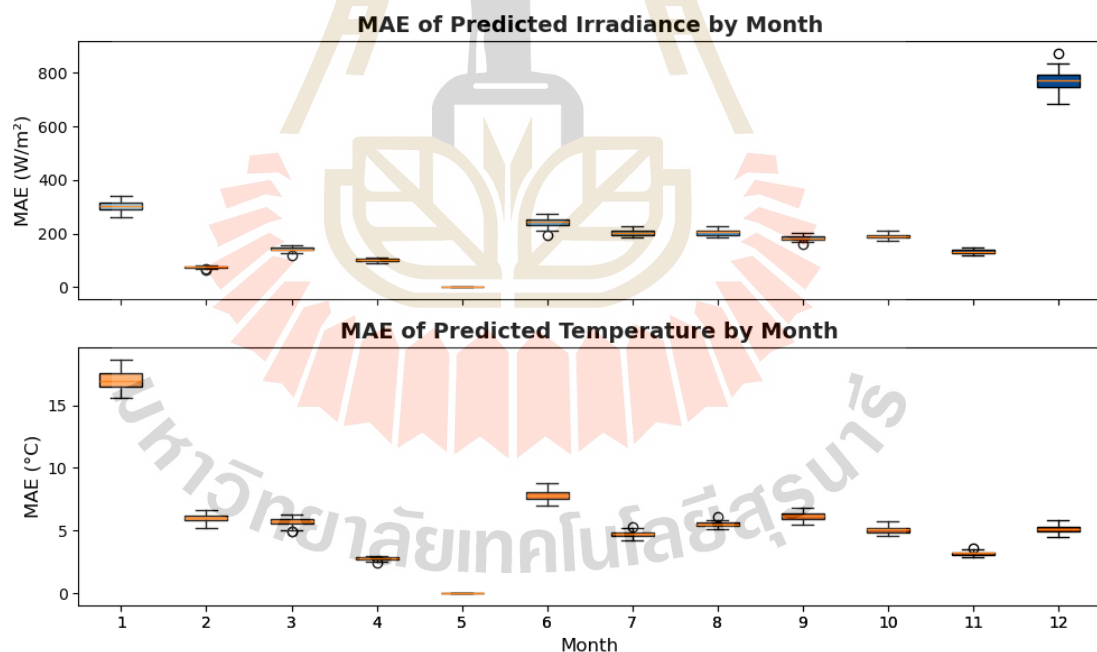


Figure 5.12 MAE of environmental factor forecasts using the LSTM

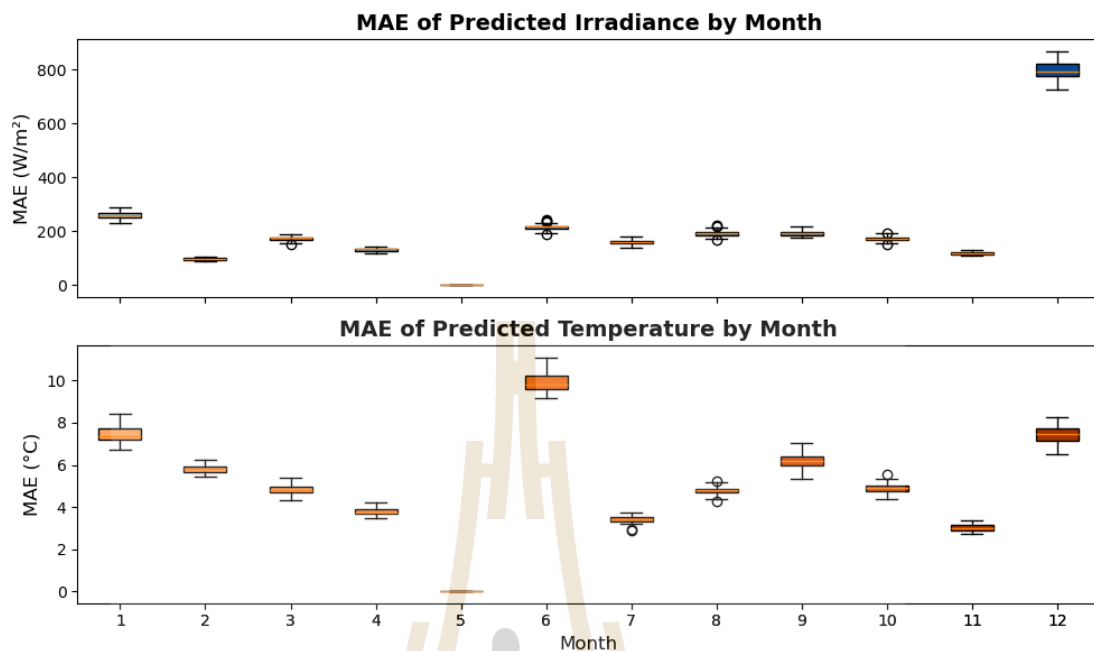


Figure 5.13 MAE of environmental factor forecasts using the TFT-lite

Analysis of the LSTM model results revealed that the MAE for solar irradiance forecast showed significant monthly variability. Some months at the beginning and end of the year exhibited high MAE values, particularly December, which had the highest MAE and a wider distribution of errors, reflecting the uncertainty of winter weather. May showed MAE values near zero, likely due to missing data rather than the model's actual performance. This trend mirrored the TFT-lite model, but with a narrower distribution of errors over several months, reflecting TFT-lite's ability to manage complex temporal relationships and its focus on relevant variables. However, December still showed a significantly high MAE, similar to the LSTM, indicating that weather volatility is a natural limitation of the data rather than a limitation of the model structure.

For temperature forecasting, the LSTM model mostly yielded MAE values in the range of approximately 3–6°C and exhibited significantly higher uniformity than the irradiance forecast. However, January and June showed significantly higher MAE values, reflecting the model's limitations in handling sudden temperature changes. Similarly, the TFT-lite model showed MAE values in the range of approximately 3–8°C with uniformity over several months. However, significantly higher MAE values were found

in June and December compared to other periods, consistent with periods of rapid temperature change. Nevertheless, the TFT-lite model demonstrated a greater ability to reduce error distribution in certain months compared to the LSTM. The consistency of results from both models supports the suitability of using LSTM or TFT-lite for forecasting environmental factors as the first step of a two-stage forecasting framework for FPV system power generation.

Before using the solar irradiance and temperature forecasts from the LSTM and TFT-lite models as input variables for power generation forecasting, statistical correlation and significance analyses of the variables were performed using correlation analysis and SHAP to evaluate the suitability of the variables used in the two-stage forecasting framework, as shown in Figures 5.14 and 5.15, respectively.

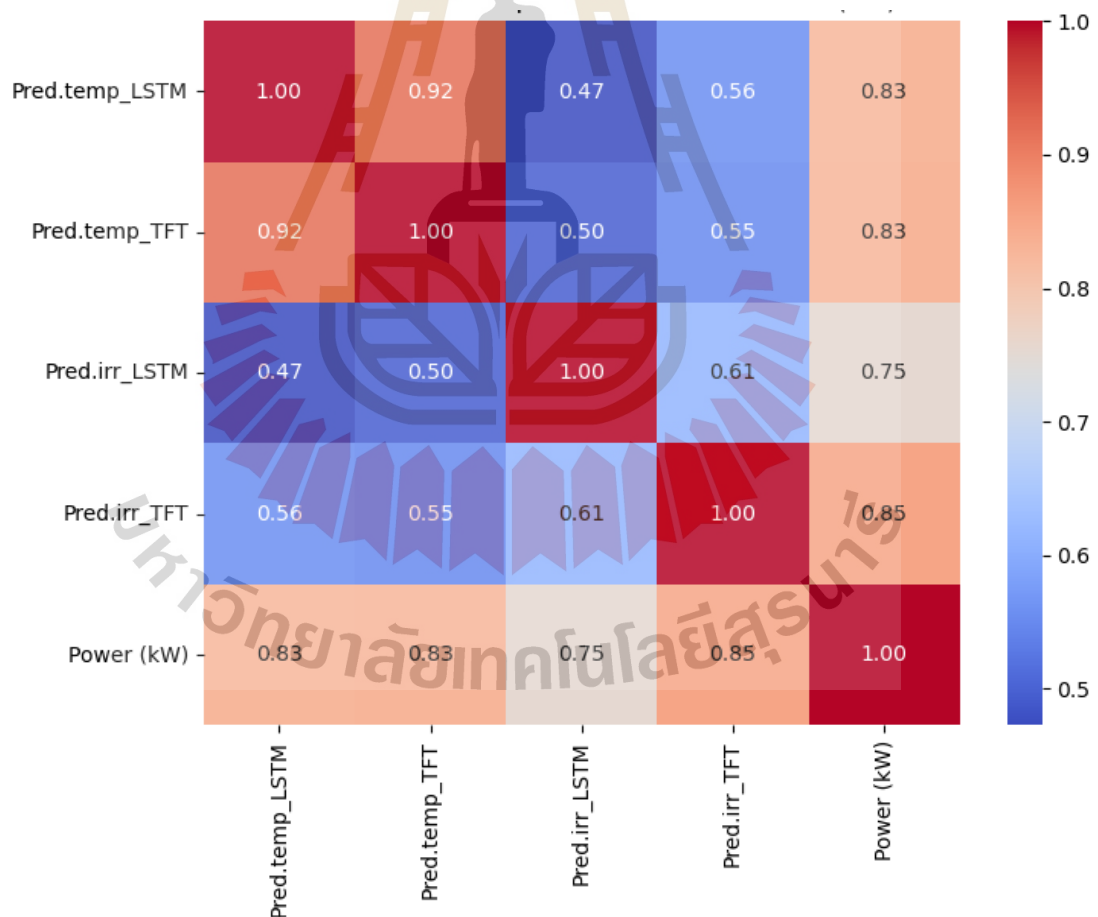


Figure 5.14 The parameters correlation matrix after SHAP process

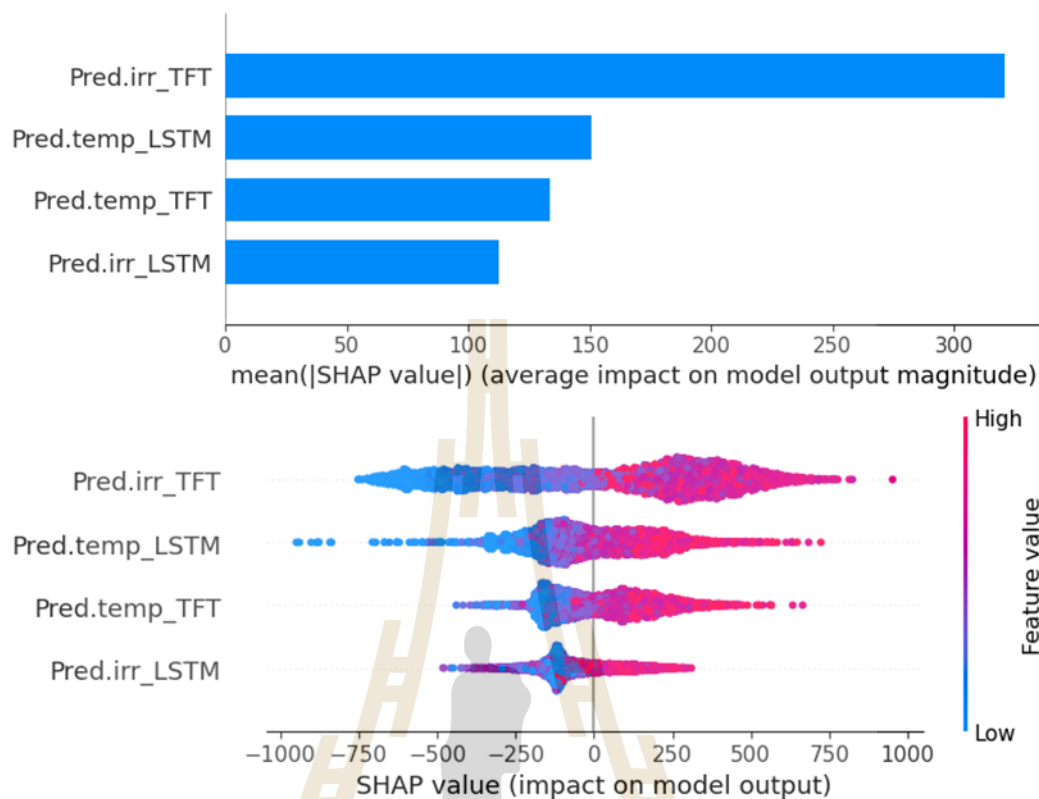


Figure 5.15 SHAP values impact on model output

The correlation analysis revealed that the variable Pred.irr_TFT had the highest correlation coefficient ($r \approx 0.85$), followed by Pred.temp_LSTM and Pred.temp_TFT ($r \approx 0.83$) and Pred.irr_LSTM ($r \approx 0.75$). These results indicate that both temperature and solar irradiance, as predicted, are directly related to power generation, especially the irradiance from the TFT-lite model, which statistically reflects the behavior of the power generation system most clearly.

The SHAP analysis further enhances the descriptive understanding of the model. From the mean (|SHAP value|), Pred.irr_TFT was found to be the variable with the greatest influence on the power generation forecast, followed by Pred.temp_LSTM and Pred.temp_TFT, respectively. Pred.irr_LSTM had less influence compared to other variables, which is consistent with the previous correlation analysis results.

Furthermore, the distribution of SHAP values shows that... High solar irradiance tends to have a clearly positive impact on power generation, while temperature has both positive and negative effects depending on the temperature range. This reflects

the actual behavior of solar cell systems, where efficiency decreases when the temperature exceeds the optimal range.

In summary, the correlation and SHAP analysis results confirm that the variables obtained from forecasting with LSTM and TFT-lite are significantly correlated with power generation and influence the model results at different levels. This analysis helps reduce the risk of selecting inappropriate variables and supports the accuracy of using the forecasted environmental factors as input variables in the next step.

For power generation forecasting, the same model as above was used, employing four validation factors: RMSE, MAE, WMAPE, and R-squared, under seasonal data divisions: summer, rainy season, and winter. The results of both test sets are shown in Tables 5.9 and 5.10, respectively.

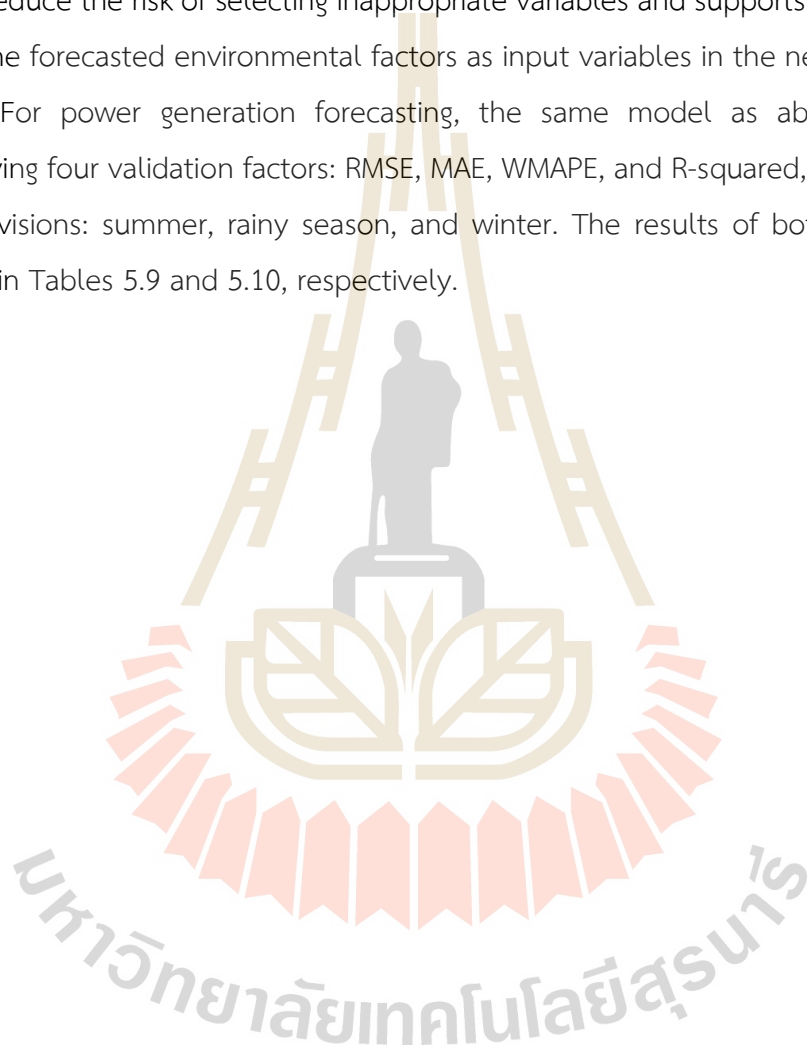


Table 5.9 Forecasting results with dataset 1 from LSTM&TFT pre-forecasting models

Model	Summer season			
	RMSE (kW)	MAE (kW)	WMAPE (%)	R ²
LSTM	135.28	118.04	6.97	0.97
Bi-LSTM	122.68	109.34	6.45	0.98
GRU	142.25	116.93	6.90	0.97
RF	55.61	42.53	2.51	1.00
TFT-lite	254.08	190.18	11.23	0.90
Model	Rainy season			
	RMSE (kW)	MAE (kW)	WMAPE (%)	R ²
LSTM	206.86	181.30	10.90	0.88
Bi-LSTM	173.15	146.79	8.83	0.92
GRU	197.94	179.49	10.79	0.89
RF	115.21	96.68	5.81	0.96
TFT-lite	226.61	173.99	10.46	0.86
Model	Winter season			
	RMSE (kW)	MAE (kW)	WMAPE (%)	R ²
LSTM	153.84	125.08	8.44	0.96
Bi-LSTM	185.27	151.50	10.22	0.94
GRU	150.24	119.58	8.06	0.96
RF	86.38	70.28	4.74	0.99
TFT-lite	158.06	137.51	9.27	0.96

Table 5.10 Forecasting results with dataset 2 from LSTM&LSTM pre-forecasting models

Model	Summer season			
	RMSE (kW)	MAE (kW)	WMAPE (%)	R ²
LSTM	103.96	77.49	4.57%	0.98
Bi-LSTM	91.85	74.92	4.42%	0.99
GRU	112.55	96.25	5.68%	0.98
RF	49.89	38.61	2.28%	1.00
TFT-lite	159.97	106.03	6.26%	0.96
Model	Rainy season			
	RMSE (kW)	MAE (kW)	WMAPE (%)	R ²
LSTM	109.89	86.01	5.17%	0.97
Bi-LSTM	119.51	98.74	5.94%	0.96
GRU	190.20	172.38	10.37%	0.90
RF	136.63	110.15	6.62%	0.95
TFT-lite	155.45	131.72	7.92%	0.93
Model	Winter season			
	RMSE (kW)	MAE (kW)	WMAPE (%)	R ²
LSTM	139.10	120.24	8.11%	0.97
Bi-LSTM	144.21	103.92	7.01%	0.97
GRU	183.47	167.51	11.30%	0.95
RF	87.70	68.05	4.59%	0.99
TFT-lite	119.50	90.56	6.11%	0.98

For power generation forecasting, the same model as above was used, employing four validation factors: RMSE, MAE, WMAPE, and R², under seasonal data divisions: summer, rainy season, and winter. The results of both test sets are shown in Tables 5.9 and 5.10, respectively.

For the summer season, Table 5.9 shows that the RF model provides the best forecasting results among the five methods, with RMSE, MAE, WMAPE, and R-squared values of 55.61 kW, 42.53 kW, 2.51%, and 1.00, respectively. This demonstrates RF's ability to handle data with low volatility and high continuity. Table 5.10, which uses

pre-forecast data from LSTM for both temperature and solar irradiance, shows that RF still provides the best results, with values of 49.89 kW, 38.61 kW, 2.28%, and 1.00, respectively. This reflects the significant impact of the quality of the initial forecasting data on error reduction.

For the rainy season, which is a period with high uncertainty and volatility in the data, Table 5.9 shows that RF provides the best forecasting results, with values of 115.21 kW, 96.68 kW, 5.81%, and 0.96, respectively. The Bi-LSTM model yields the second-best results, reflecting the fact that using environmental factors from the original data source directly results in RF's superior performance. Time-associated statistical models may be more robust to seasonal variations. However, considering the forecasting results from Table 5.10, the LSTM model showed the highest accuracy during the rainy season, with values of 109.89 kW, 86.01 kW, 5.17%, and 0.97, respectively. This is significantly better than the results from Table I, demonstrating that time-concordance (LSTM–LSTM) of solar irradiance and temperature data from the first step helps to mitigate the effects of seasonal variations. And it significantly improved the forecasting efficiency of electricity generation.

In the winter season, Table 5.9 showed that RF was the best-performing model, with values of 86.38 kW, 70.28 kW, 4.74%, and 0.99, respectively, while LSTM and GRU gave similar results. However, Table 5.10 showed that RF still provided the best forecasting results, with values of 87.70 kW, 68.05 kW, 4.59%, and 0.99, respectively. This reflects that in winter, when the data is influenced by many environmental factors, RF can efficiently manage nonlinear relationships.

However, the ranking of processing time from Table 5.11 shows that RF is the model that takes the least time among the two datasets, at only about 5–6 seconds. This is followed by LSTM, Bi-LSTM, and GRU, respectively, while TFT-lite takes the longest time, more than 500 seconds in both cases. This is consistent with the theory that transformer models have a complex structure and require higher computational resources.

Table 5.11 Runtime of PV forecasting models

Model	Time (s)	
	Pre-forecast data 1	Pre-forecast data 2
	(LSTM&TFT)	(LSTM&LSTM)
LSTM	83.84	86.64
Bi-LSTM	86.54	181.26
GRU	113.47	90.34
RF	5	5.08
TFT-lite	503.99	564.54

Furthermore, the outstanding forecasting performance of RF in multiple seasons is due to the characteristics of the data used in this case. This involves managing missing data by removing incomplete information, which can cause temporal discontinuities in the dataset. This results in more stable forecasting for time-sequence-independent models such as RF, while time-sequence models like LSTM and Bi-LSTM may be affected by such data continuity loss, particularly during the highly volatile rainy season. This reflects the fact that managing missing data plays a crucial role in the performance of power generation forecasting.

5.4 Conclusion

This chapter presents an evaluation of the performance of solar power generation forecasting models, building upon the analysis and results conducted in the previous chapter. The model assesses the use of missing data from different real-world systems, comparing both accuracy and processing time. The analysis shows that power generation forecasting cannot be considered solely based on model accuracy but requires a systemic approach encompassing data management, the quality of forecasted inputs, the model's internal mechanisms, seasonal variability, and spatial and environmental constraints specific to floating PV systems. However, since all data comes from a single site, this method may require parameter adjustments or the selection of additional variables when applied to different areas with varying climates, cloud patterns, and radiation angles.

In summary, this chapter demonstrates that data quality management combined with model selection tailored to the data structure plays a crucial role in accuracy, efficiency, and application suitability, forming a fundamental basis for the conclusions and recommendations in the next chapter.



CHAPTER VI

CONCLUSION

6.1 Conclusion

This thesis presents an in-depth study of forecasting electricity generation from FPV systems at Suranaree University of Technology under conditions of incomplete data and seasonal fluctuations. The main objective is to develop and evaluate a two-stage forecasting framework for a case study of a university-level electricity generation system with campus-based load characteristics and site-specific usage conditions.

In the first stage, the thesis focused on managing data quality and forecasting environmental factors, namely solar irradiance and temperature, using time-sequence models such as LSTM and TFT-lite. Experimental results showed that LSTM provided more accurate and stable forecasting results, especially in the case of temperature, which has lower temporal continuity and fluctuations than solar irradiance. Therefore, the results from the first stage are suitable for use as input in the next stage.

For managing real-world system data, missing data and curtailment were observed. This thesis therefore employed data management methods considering the filling of missing values and the removal of incomplete data from the dataset, which reduced the amount of data and temporal discontinuity, as well as considering curtailment data. However, the analysis results indicate that this approach helps reduce errors caused by estimates and makes model learning clearer in some cases.

In the second step, the light intensity and temperature forecasts from the first step were used as input for power generation forecasting. The performances of LSTM, Bi-LSTM, GRU, RF, and TFT-lite models, as well as the Persistence basic model, was compared as a benchmark. Experimental results show that power generation forecasting cannot be considered solely based on model accuracy but must be viewed systemically, encompassing data management, the quality of the forecasted inputs,

the model's internal mechanisms, seasonal variance, and the actual operating conditions of the system. This case study primarily reflects a campus-level application.

Furthermore, descriptive analysis using correlation analysis and SHAP confirms the crucial role of solar intensity and temperature in determining power generation and enhances the interpretability of the model results, which is a key issue for practical application in real energy systems.

However, since the data used in this study comes from a single-site production source and is a case-specific testing system, it cannot be directly generalized to the utility level. The application to other areas or systems should consider additional factors such as load structure at the electrical system level, feeder conditions, abnormal events, or disasters that may affect production and power supply, as well as adjusting parameters and model structures to fit the context of the target system before actual implementation.

Overall, this research achieves its objectives by demonstrating that data quality management combined with model selection appropriate to the data structure and seasonal conditions significantly impacts the accuracy and efficiency of campus-level FPV power generation forecasting. This can serve as a starting point for further development in other system contexts.

6.2 Suggestions

This thesis demonstrates the potential benefits of using a two-stage forecasting framework for application to floating solar power systems in Thailand, a country experiencing continuous growth in renewable energy. However, some limitations should be considered in further study to improve accuracy, stability, and practical applicability in the future. In particular, there are issues related to managing missing and limited data sets that may affect temporal continuity and the overall data structure. Furthermore, model recalibration is necessary when applied to areas with varying climates, cloud patterns, or installation conditions.

Additionally, when applied to utility-scale power systems, the model should be improved to reflect real-world system conditions, particularly by incorporating variables or features related to load behavior, power system structure, feeder

conditions, system anomalies, and environmental or disaster factors that may impact power generation and distribution.

This research utilizes single-site data, which may limit the ability to draw conclusions and apply to floating solar power systems in different spatial contexts. Nevertheless, the results of this study can serve as a guideline to support energy planners, power system operators, and power plant managers in making practical decisions, real-time power generation monitoring, and future load management.



REFERENCES

- Ait Mansour, A., Tilioua, A., & Touzani, M. (2024, 2024/03/01/). Bi-LSTM, GRU and 1D-CNN models for short-term photovoltaic panel efficiency forecasting case amorphous silicon grid-connected PV system. *Results in Engineering*, 21, 101886. <https://doi.org/https://doi.org/10.1016/j.rineng.2024.101886>
- Ali, S. S., Ali, M., Bhatti, D. M., & Choi, B. J. (2025). Explainable Clustered Federated Learning for Solar Energy Forecasting. *Energies*, 18(9).
- Bianchini, G., Pepe, D., & Vicino, A. (2020, 2020/03/01/). Estimation of photovoltaic generation forecasting models using limited information. *Automatica*, 113, 108688. <https://doi.org/https://doi.org/10.1016/j.automatica.2019.108688>
- Boucetta, L. N., Amrane, Y., & Arezki, S. (2023, 5-7 Nov. 2023). Comparative Analysis of LSTM, GRU, and MLP Neural Networks for Short-Term Solar Power Forecasting. 2023 International Conference on Electrical Engineering and Advanced Technology (ICEEAT),
- Chang, X., Li, W., Ma, J., Yang, T., & Zomaya, A. Y. (2020, 4-8 May 2020). Interpretable Machine Learning In Sustainable Edge Computing: A Case Study of Short-Term Photovoltaic Power Output Prediction. ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),
- Chen, B., Li, H., & Han, K. (2019, 24-25 Oct. 2019). Short-term photovoltaic power forecasting method based on K-means algorithm and spiking neural networks. 8th Renewable Power Generation Conference (RPG 2019),

Gao, Y., Qi, S., & Ponoćko, J. (2023, 25-29 June 2023). Assessing Critical Data Types for Deep Learning-Based PV Generation Forecasting. 2023 IEEE Belgrade PowerTech,

Huynh, A. D. N., Thanh Khuong. (2024, Received 10 Aug 2024; Revised 26 Sep 2024; Accepted 28 Sep 2024; Published Jan 2024). The Comparison of GRU and LSTM in Solar Power Generation Forecasting Application [Peer-reviewed Research Article]. *International Journal of Science and Research Archive*, 13(01), 1360–1370. <https://doi.org/10.30574/ijrsra.2024.13.1.1831> (IJSRA 2024, Vol13(1):1360–1370)

Jailani, N. L., Dhanasegaran, J. K., Alkawsi, G., Alkahtani, A. A., Phing, C. C., Baashar, Y., Capretz, L. F., Al-Shetwi, A. Q., & Tiong, S. K. (2023). Investigating the Power of LSTM-Based Models in Solar Energy Forecasting. *Processes*, 11(5).

Junhuathon, N., & Chayakulkheeree, K. (2021, 20-22 Oct. 2021). Comparative Study of Short-Term Photovoltaic Power Generation Forecasting Methods. 2021 International Conference on Power, Energy and Innovations (ICPEI),

Junhuathon, N., & Chayakulkheeree, K. (2022, 07/01). Deep-learning-based short-term photovoltaic power generation forecasting using improved self-organization map neural network. *Journal of Renewable and Sustainable Energy*, 14, 043702. <https://doi.org/10.1063/5.0091454>

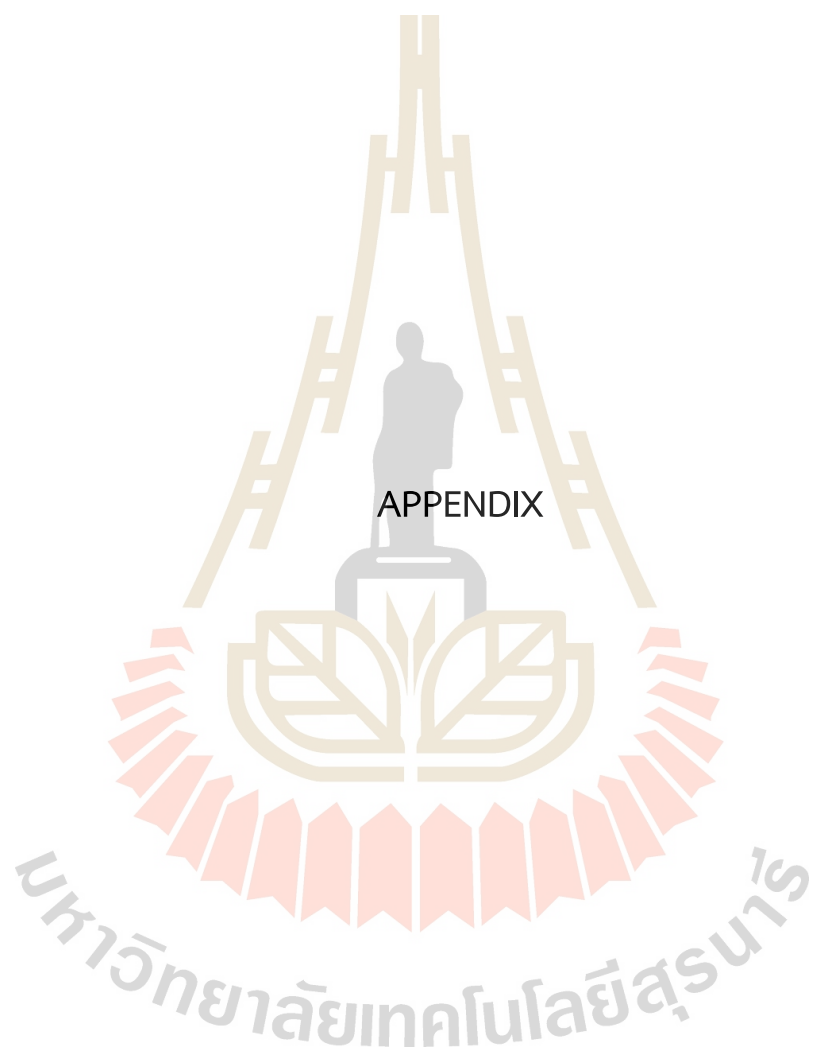
khan, S., Mazhar, T., Shahzad, T., Waheed, W., Waheed, A., Saeed, M. M., & Hamam, H. (2024, 2024/11/12). Optimizing deep neural network architectures for renewable energy forecasting. *Discover Sustainability*, 5(1), 394. <https://doi.org/10.1007/s43621-024-00615-6>

- Kuzlu, M., Cali, U., Sharma, V., & Ö, G. (2020). Gaining Insight Into Solar Photovoltaic Power Generation Forecasting Utilizing Explainable Artificial Intelligence Tools. *IEEE Access*, 8, 187814-187823. <https://doi.org/10.1109/ACCESS.2020.3031477>
- Lee, D. S., & Son, S. Y. (2024). PV Forecasting Model Development and Impact Assessment via Imputation of Missing PV Power Data. *IEEE Access*, 12, 12843-12852. <https://doi.org/10.1109/ACCESS.2024.3352038>
- Lu, Y., Murzakhanov, I., & Chatzivasileiadis, S. (2021, 25-28 Oct. 2021). Neural network interpretability for forecasting of aggregated renewable generation. 2021 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm),
- Lundberg, S. M. a. L., Su-In. (2017, December 2017). A unified approach to interpreting model predictions [Peer-reviewed Conference Paper]. *Advances in Neural Information Processing Systems*, 30, 4765–4774. <https://doi.org/10.48550/arXiv.1705.07874> (NeurIPS 2017)
- Mitrentsis, G., & Lens, H. (2022, 2022/03/01/). An interpretable probabilistic model for short-term solar power forecasting using natural gradient boosting. *Applied Energy*, 309, 118473. <https://doi.org/https://doi.org/10.1016/j.apenergy.2021.118473>
- NXPO. (2022). Thailand's development and use of clean energy for a sustainable future. Office of National Higher Education Science Research and Innovation Policy Council (NXPO). <https://www.nxpo.or.th/th/9651/>
- Roy, N., Tripathy, P., De, S. C., Swargiary, B., Kumar, S., Das, S., & Pathak, N. (2022, 17-19 Dec. 2022). Day-ahead Solar Power Generation Forecasting using LSTM and Random Forest Methods for North Eastern Region of India. 2022 22nd National Power Systems Conference (NPSC),

- Sarp, S., Kuzlu, M., Cali, U., Elma, O., & Guler, O. (2021, 16-18 Feb. 2021). An Interpretable Solar Photovoltaic Power Generation Forecasting Approach Using An Explainable Artificial Intelligence Tool. 2021 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT),
- Sedliarov, Y., & Klen, K. (2024, 24-26 Sept. 2024). Predictive Control Algorithm for Solar Station based on K-Nearest Neighbor and K-Means. 2024 IEEE 7th International Conference on Smart Technologies in Power Engineering and Electronics (STEE),
- Sharadga, H., Hajimirza, S., & Balog, R. S. (2020, 2020/05/01/). Time series forecasting of solar power generation for large-scale photovoltaic plants. *Renewable Energy*, 150, 797-807.
<https://doi.org/https://doi.org/10.1016/j.renene.2019.12.131>
- Wang, J., Wang, S., Han, J., Ma, Z., Li, H., & Zhang, J. (2024, 25-27 Oct. 2024). An Approach to Forecasting Short-Term Photovoltaic Power Generation. 2024 8th International Conference on Electrical, Mechanical and Computer Engineering (ICEMCE),
- W. Thongwily, N. Junhuathon, and K. Chayakulkheeree (2025), Two-Stage Data-Driven Modeling for Day-Ahead Solar Power Forecasting: A Case Study of the SUT Campus Floating Solar System, 2025 13th International Electrical Engineering Congress (iEECON2025), Huahin, Thailand.
- Zhou, H., Zhang, Y., Yang, L., & Liu, Q. (2018, 19-21 Oct. 2018). Short-Term Photovoltaic Power Forecasting Based on Stacking-SVM. 2018 9th International Conference on Information Technology in Medicine and Education (ITME),

Zhou, H., Zhang, Y., Yang, L., Liu, Q., Yan, K., & Du, Y. (2019). Short-Term Photovoltaic Power Forecasting Based on Long Short Term Memory Neural Network and Attention Mechanism. *IEEE Access*, 7, 78063-78074.
<https://doi.org/10.1109/ACCESS.2019.2923006>

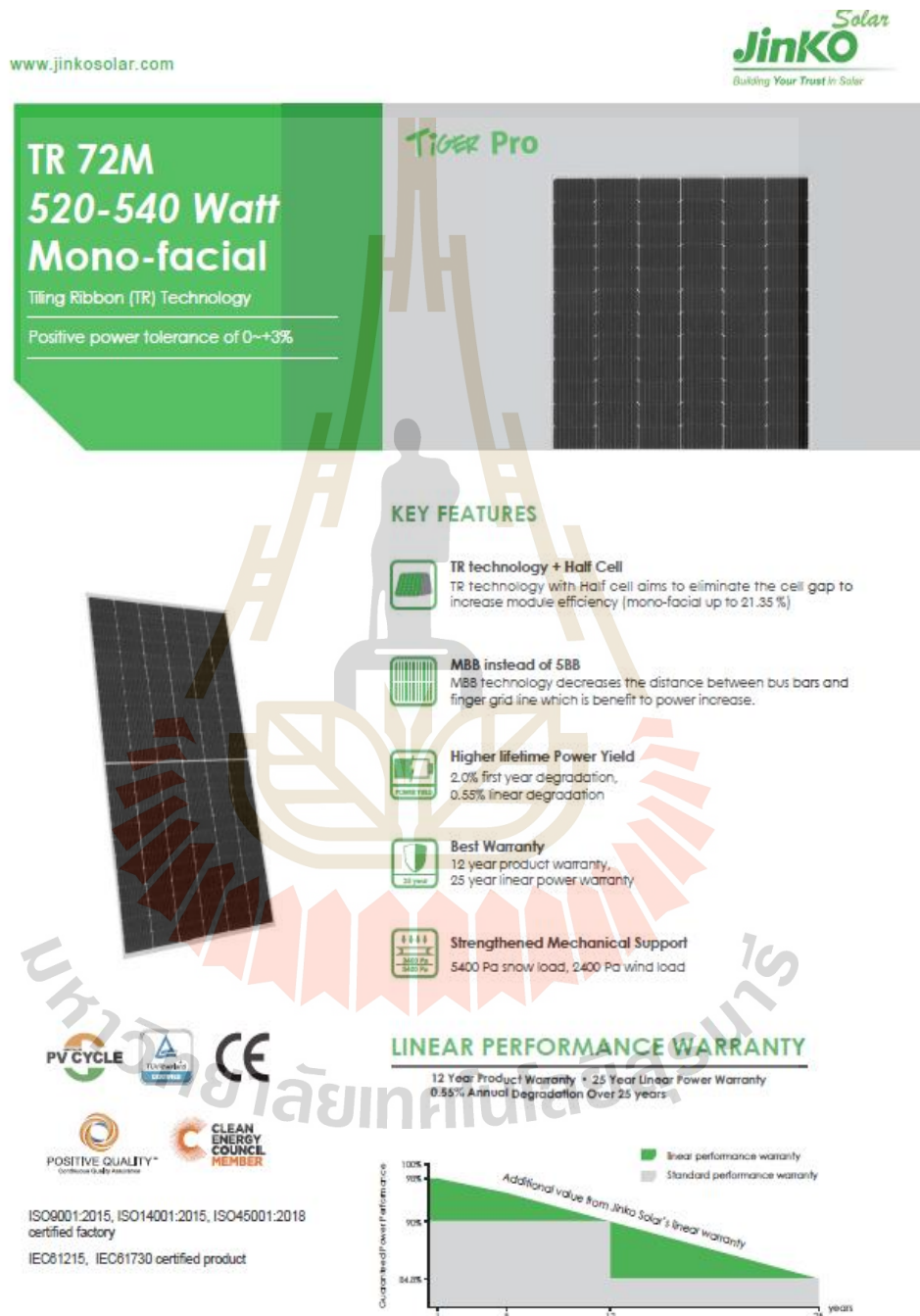




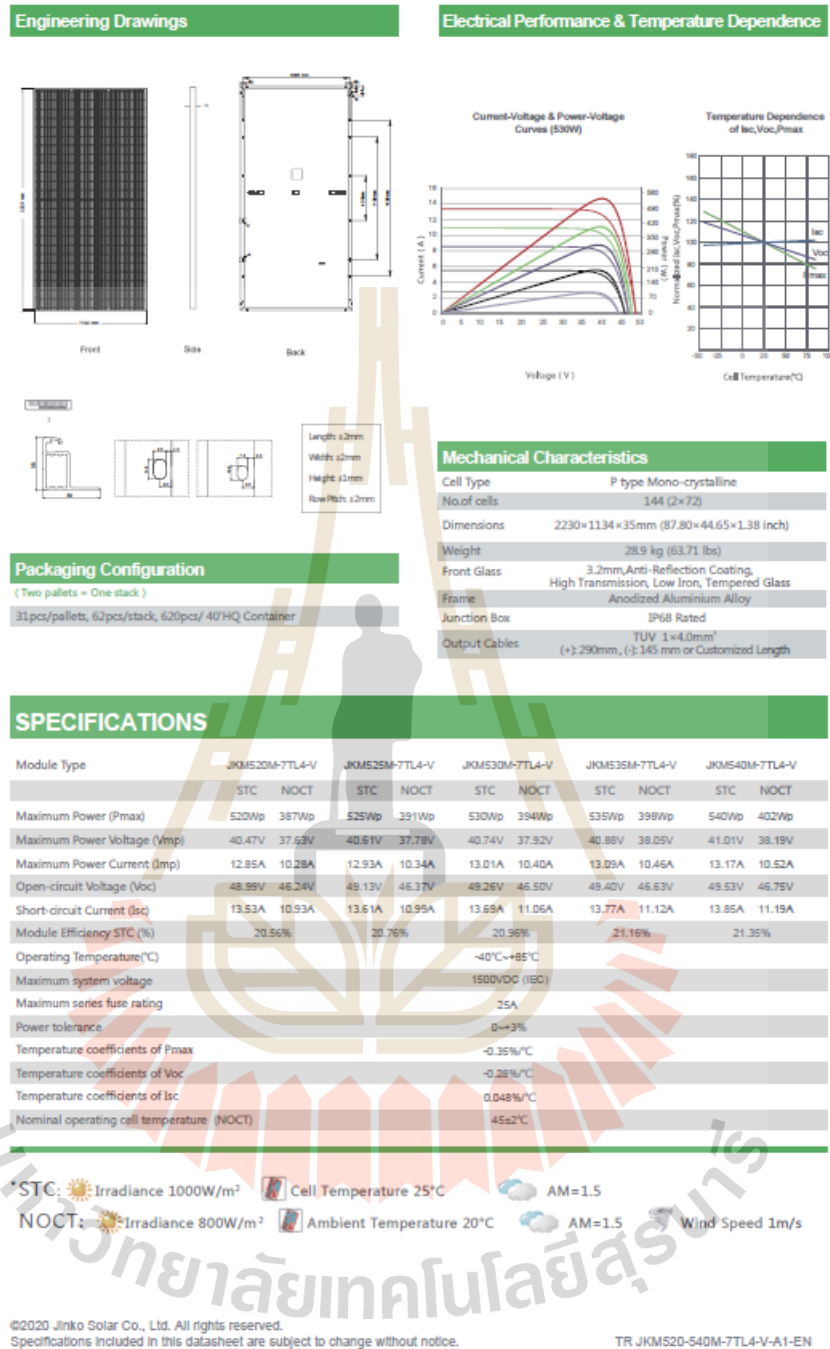
APPENDIX

APPENDIX A

Solar cell datasheet



(a)



(b)

Figure A.1 Solar panel specification (data sheet) used in this study

APPENDIX B

Environmental Correlation with Solar Power Output Based on Linear Interpolation

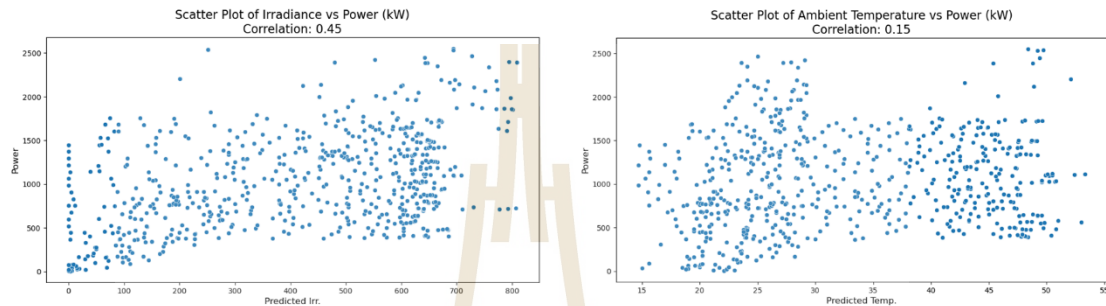


Figure B.1 Environmental correlation with solar power output - January 2023

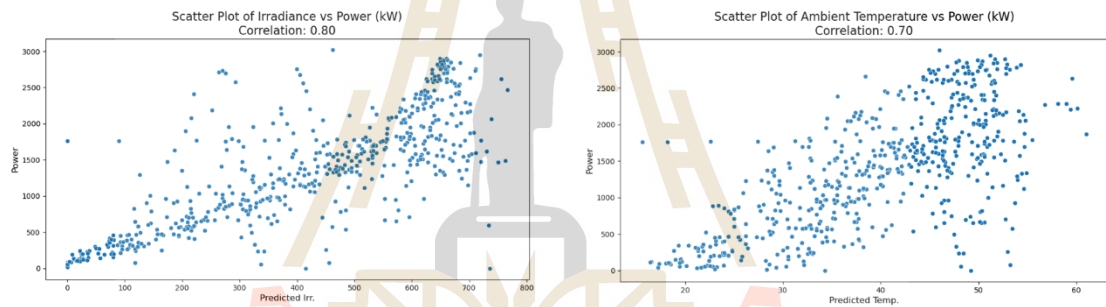


Figure B.2 Environmental correlation with solar power output - February 2023

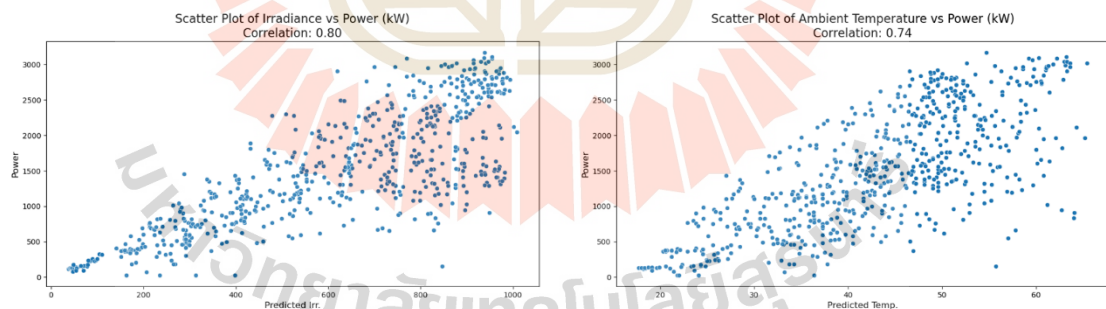


Figure B.3 Environmental correlation with solar power output - March 2023

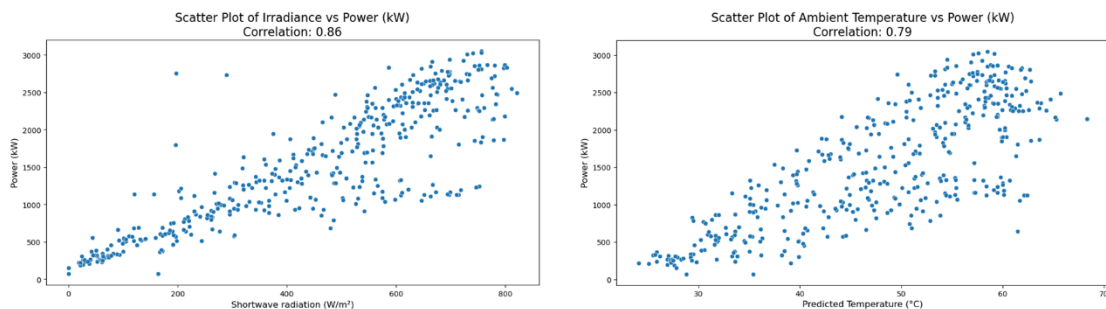


Figure B.4 Environmental correlation with solar power output - April 2023

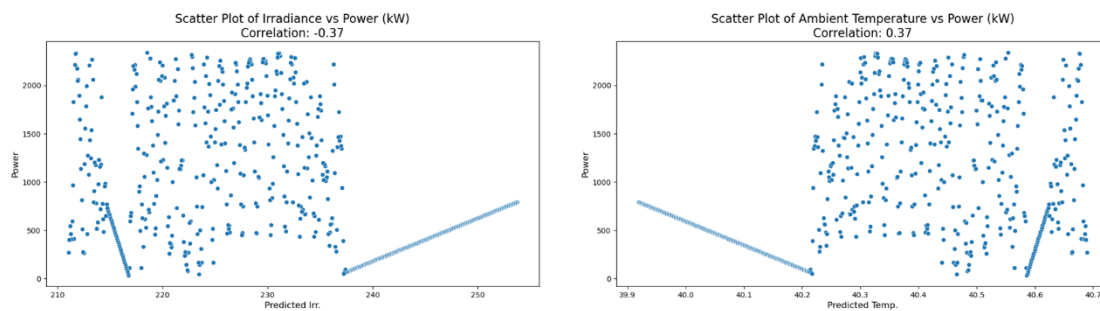


Figure B.5 Environmental correlation with solar power output - May 2023

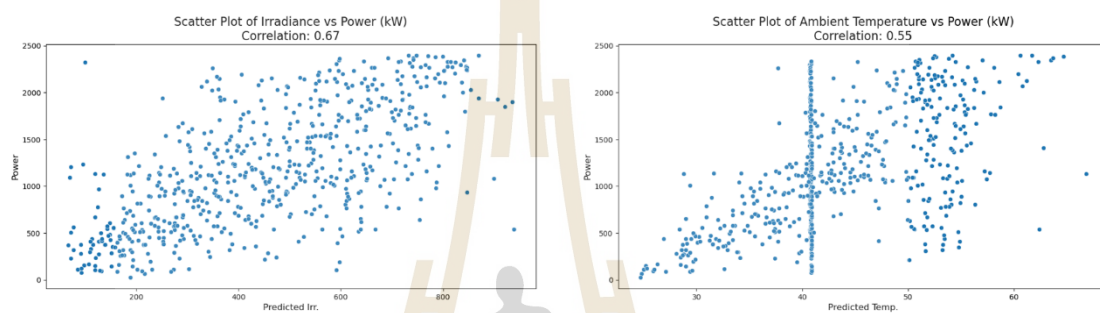


Figure B.6 Environmental correlation with solar power output - June 2023

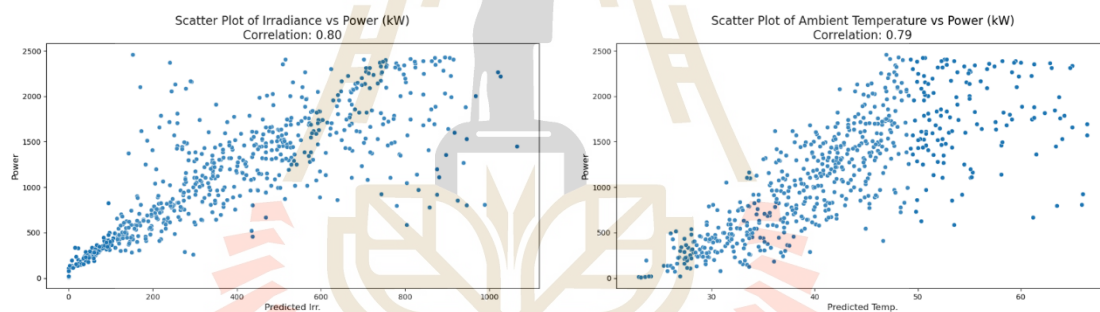


Figure B.7 Environmental correlation with solar power output - July 2023

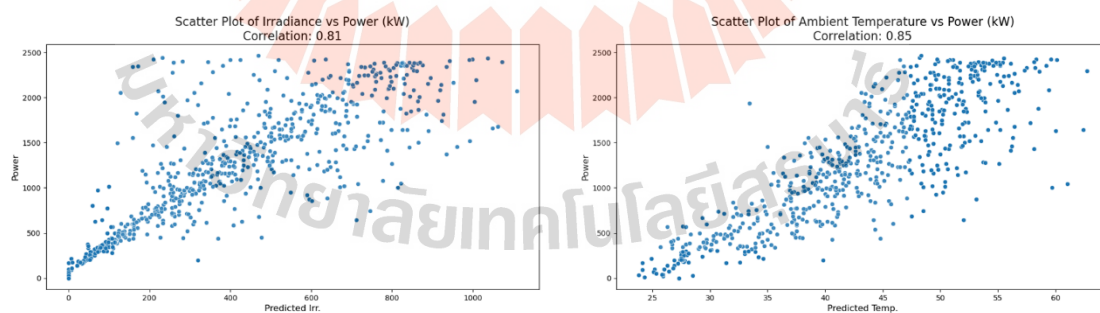


Figure B.8 Environmental correlation with solar power output - August 2023

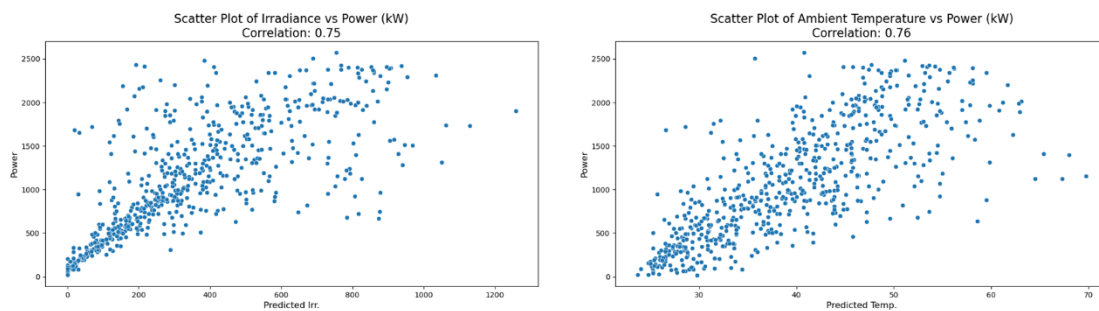


Figure B.9 Environmental correlation with solar power output - September 2023

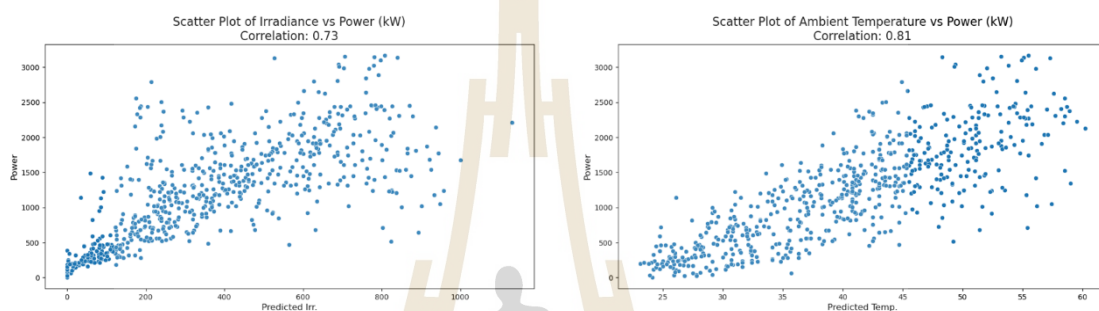


Figure B.10 Environmental correlation with solar power output – October 2023

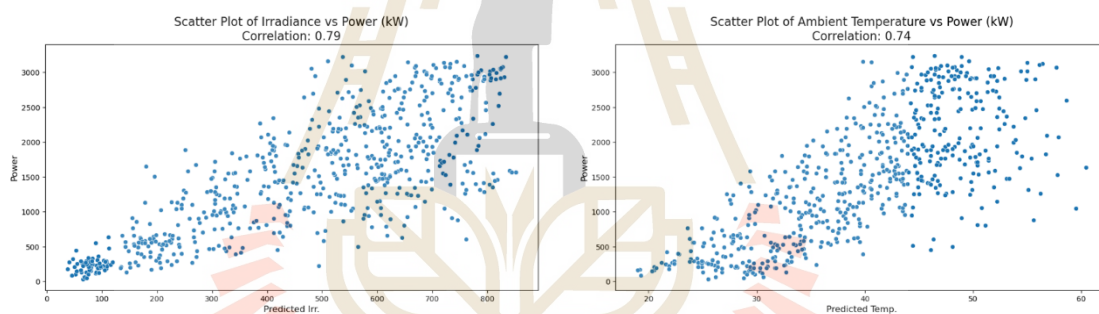


Figure B.11 Environmental correlation with solar power output – November 2023

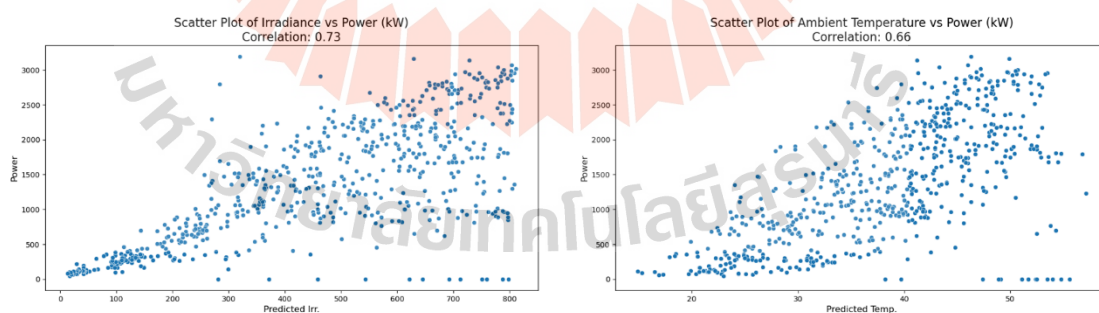


Figure B.12 Environmental correlation with solar power output – December 2023

APPENDIX C

Evaluation Results for KNN Imputation Parameters

Table C.1 Comparison of KNN Imputation Performance Across K Values

KNN Imputer	Error Metrics				Normalized Metrics				Composite Score
	MSE	RMSE	MAE	MAPE	MSE	RMSE	MAE	MAPE	
	(kW ²)	(kW)	(kW)	(%)	(kW ²)	(kW)	(kW)	(%)	
1	247137.63	497.13	356.22	48.15	1.000	1.000	1.000	0.131	0.783
2	213703.47	462.28	329.29	47.92	0.496	0.515	0.373	0.117	0.375
3	205756.83	453.60	325.99	47.41	0.376	0.394	0.296	0.088	0.289
4	192733.91	439.01	317.29	46.93	0.179	0.191	0.093	0.061	0.131
5	193440.98	439.82	318.37	45.86	0.190	0.202	0.119	0.000	0.128
6	191176.07	437.24	319.30	47.50	0.156	0.167	0.140	0.093	0.139
7	185050.12	430.17	316.61	48.80	0.063	0.068	0.078	0.168	0.094
8	182136.27	426.77	313.37	49.78	0.019	0.021	0.002	0.223	0.066
9	182489.25	427.19	313.28	49.51	0.025	0.027	0.000	0.208	0.065
10	180856.18	425.27	313.28	50.21	0.000	0.000	0.000	0.248	0.062
11	182676.52	427.41	314.16	50.95	0.027	0.030	0.020	0.290	0.092
12	183632.75	428.52	313.50	51.73	0.042	0.045	0.005	0.334	0.107
13	184262.34	429.26	315.46	51.97	0.051	0.055	0.051	0.348	0.126
14	188142.06	433.75	319.44	52.40	0.110	0.118	0.144	0.372	0.186
15	186245.86	431.56	318.21	52.66	0.081	0.088	0.115	0.387	0.168
16	184762.38	429.84	316.78	53.09	0.059	0.064	0.082	0.412	0.154
17	187352.20	432.84	319.73	54.15	0.098	0.105	0.150	0.473	0.207
18	187602.19	433.13	321.60	54.95	0.102	0.109	0.194	0.518	0.231
19	186504.11	431.86	320.51	55.01	0.085	0.092	0.168	0.521	0.217
20	185949.87	431.22	320.46	55.24	0.077	0.083	0.167	0.534	0.215
30	190235.39	436.16	323.37	56.17	0.142	0.152	0.235	0.588	0.279
40	188538.09	434.21	320.73	56.47	0.116	0.124	0.174	0.604	0.255
50	189285.68	435.07	320.60	57.41	0.127	0.136	0.170	0.658	0.273
60	192038.22	438.22	323.56	58.92	0.169	0.180	0.240	0.744	0.333
70	191911.58	438.08	323.40	60.26	0.167	0.178	0.236	0.820	0.350
80	190994.01	437.03	322.64	61.52	0.153	0.164	0.218	0.892	0.357

Table C.1 Comparison of KNN Imputation Performance Across K Values

KNN Imputer	Error Metrics				Normalized Metrics				Composite Score
	MSE	RMSE	MAE	MAPE	MSE	RMSE	MAE	MAPE	
	(kW ²)	(kW)	(kW)	(%)	(kW ²)	(kW)	(kW)	(%)	
90	191542.96	437.66	323.12	62.26	0.161	0.172	0.229	0.934	0.374
100	190967.32	437.00	323.45	63.41	0.153	0.163	0.237	1.000	0.388



APPENDIX D

Environmental Correlation with Solar Power Output Based on Clear-Sky and KNN Imputation

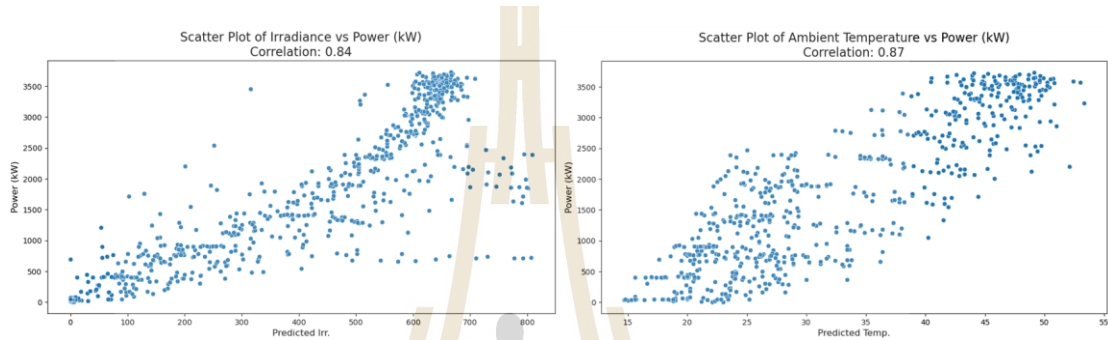


Figure D.1 Environmental correlation with solar power output - January 2023

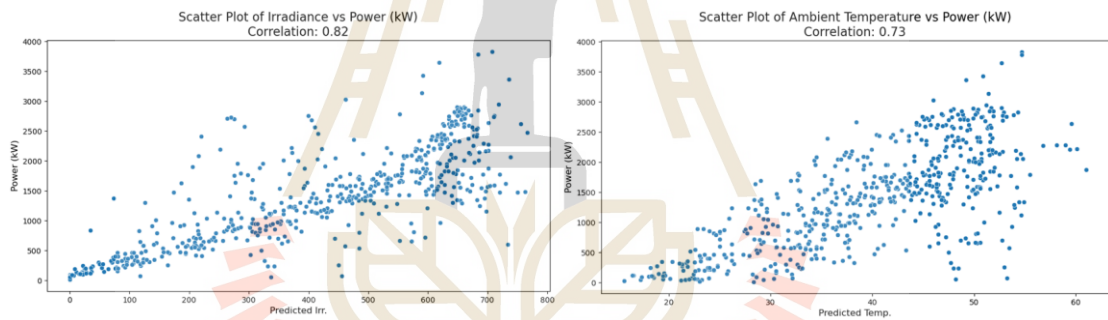


Figure D.2 Environmental correlation with solar power output - February 2023

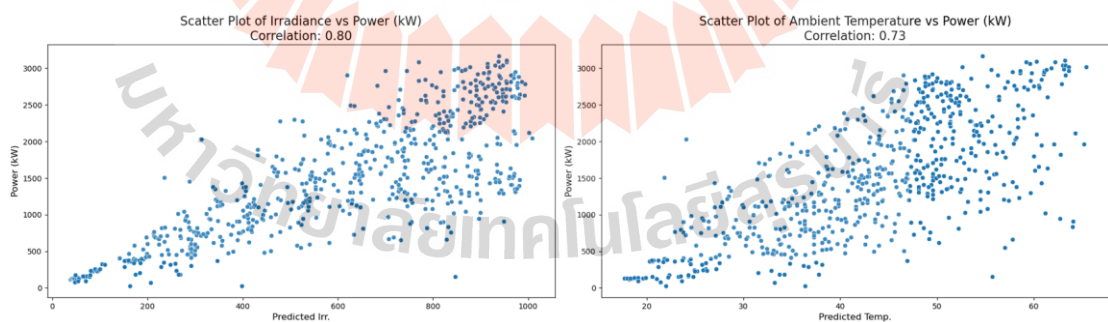


Figure D.3 Environmental correlation with solar power output - March 2023

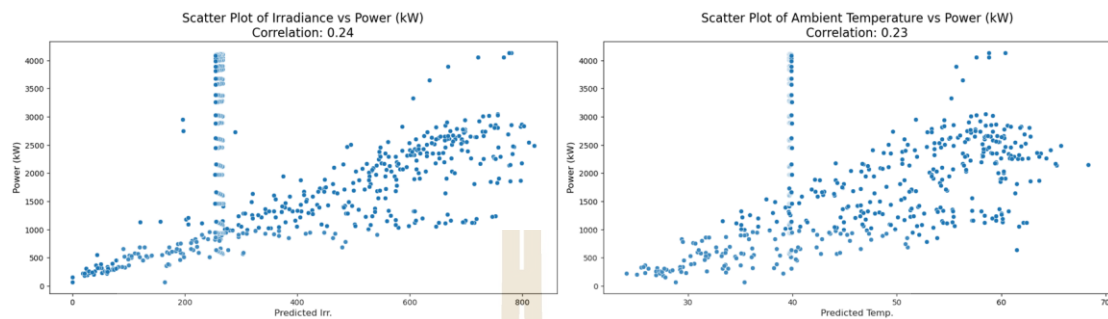


Figure D.4 Environmental correlation with solar power output - April 2023

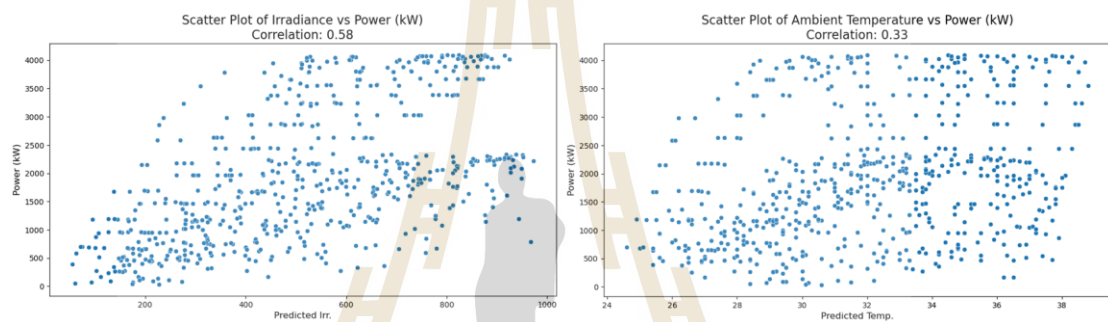


Figure D.5 Environmental correlation with solar power output - May 2023

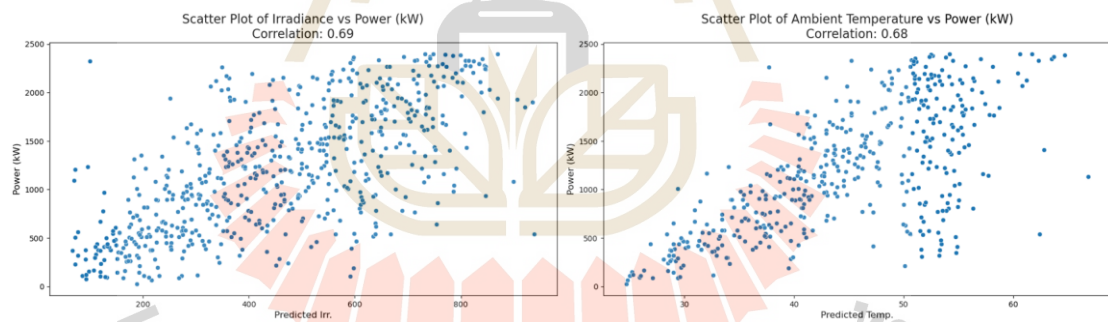


Figure D.6 Environmental correlation with solar power output - June 2023

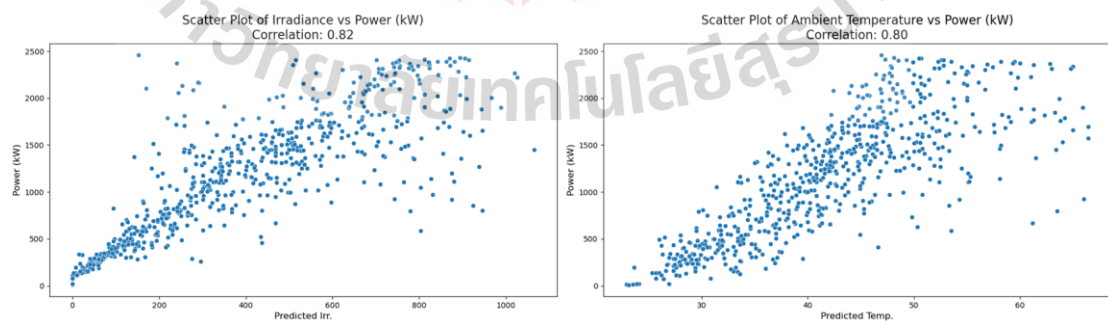


Figure D.7 Environmental correlation with solar power output - July 2023

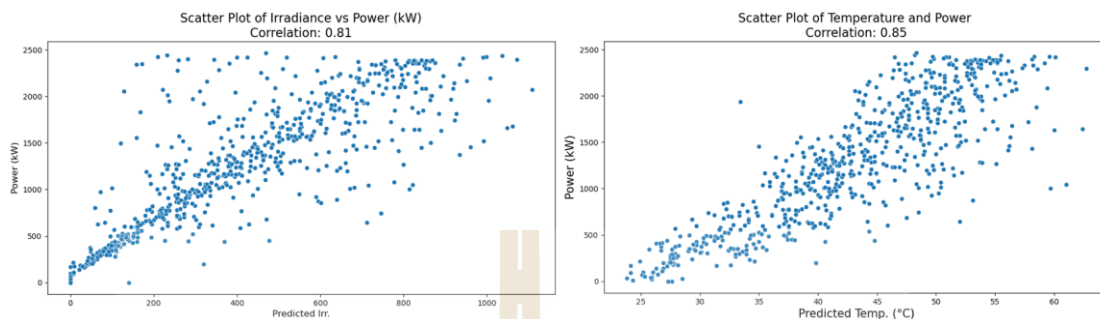


Figure D.8 Environmental correlation with solar power output - August 2023

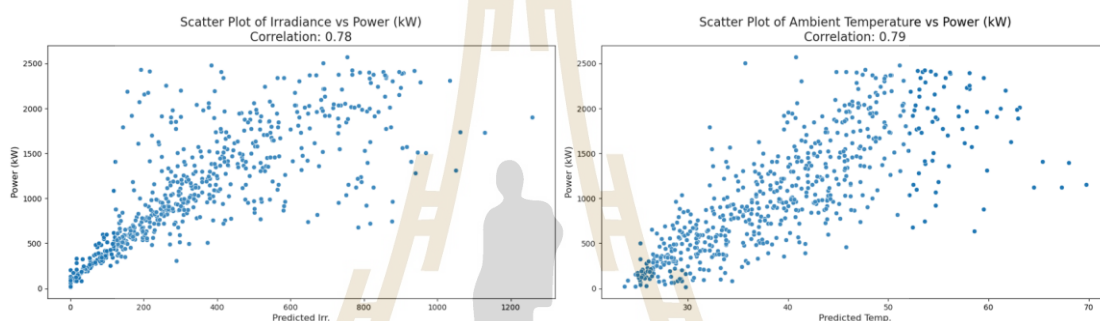


Figure D.9 Environmental correlation with solar power output - September 2023

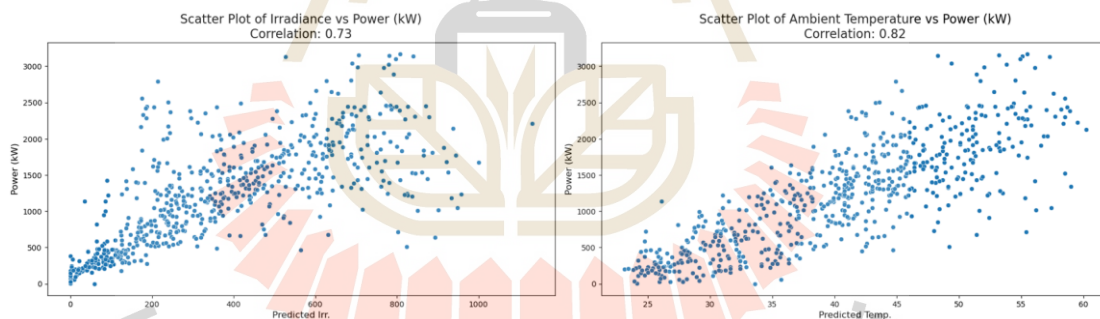


Figure D.10 Environmental correlation with solar power output – October 2023

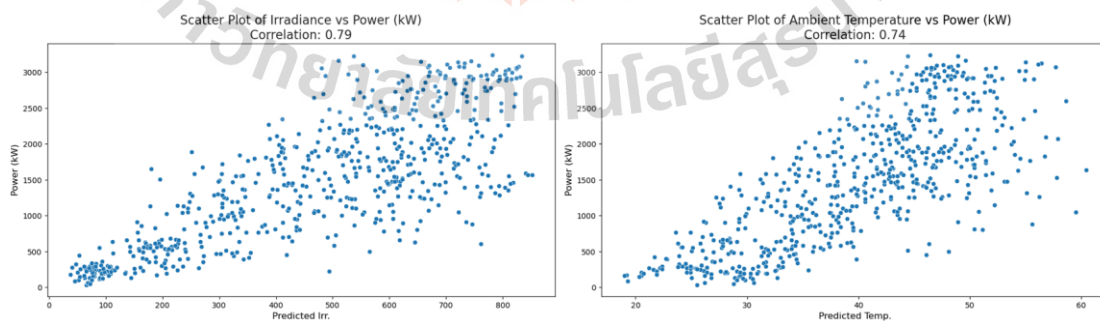


Figure D.11 Environmental correlation with solar power output – November 2023

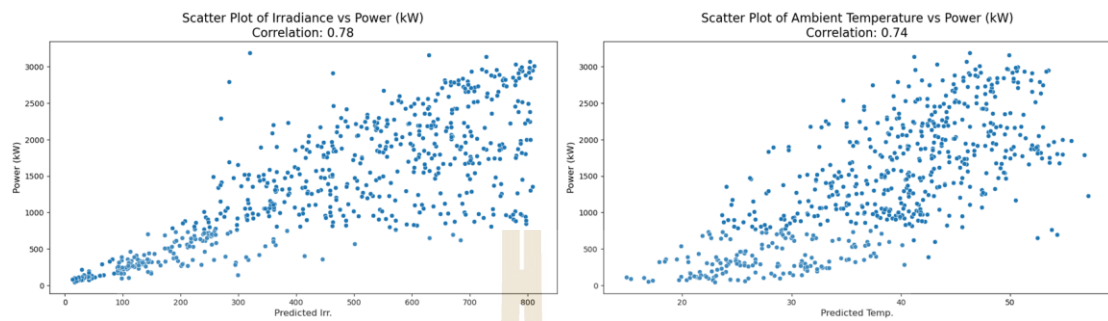
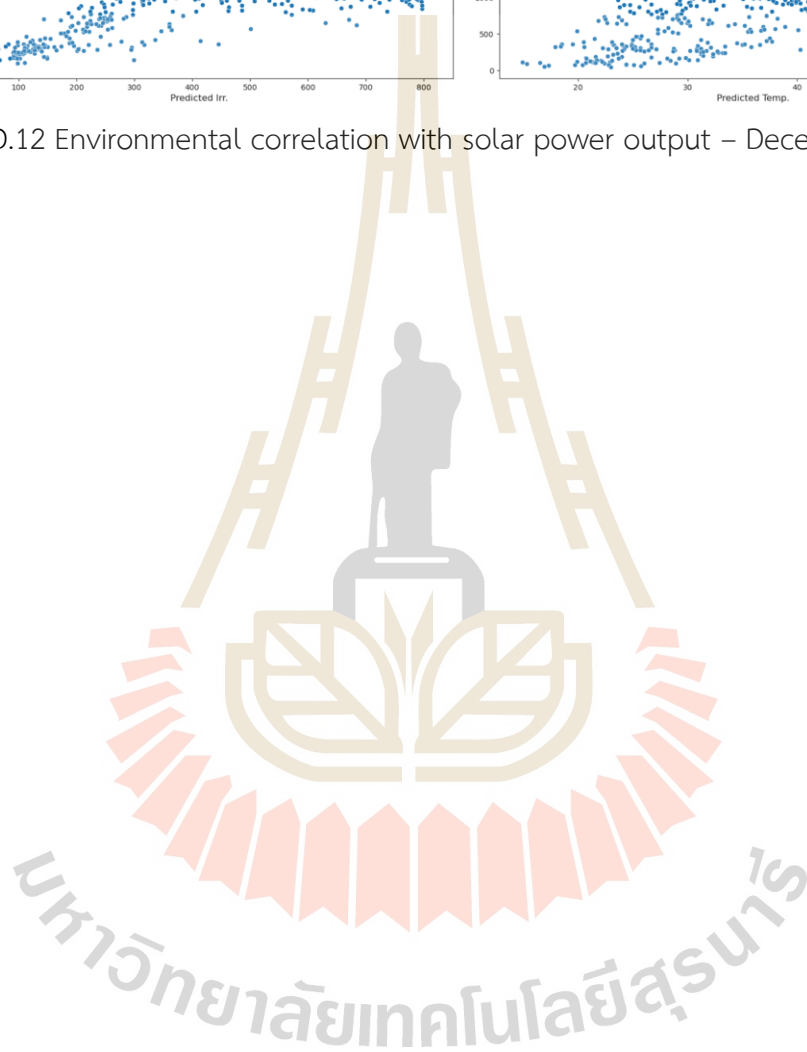


Figure D.12 Environmental correlation with solar power output – December 2023



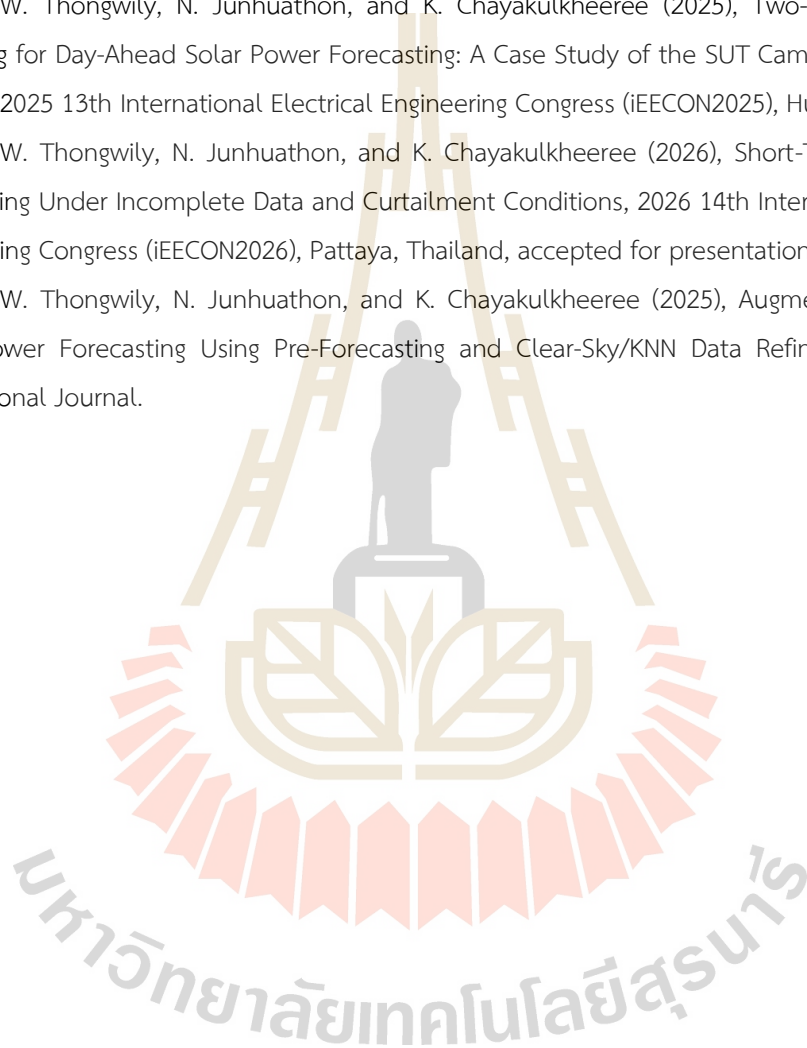
APPENDIX E

List of publications

W. Thongwily, N. Junhuathon, and K. Chayakulkheeree (2025), Two-Stage Data-Driven Modeling for Day-Ahead Solar Power Forecasting: A Case Study of the SUT Campus Floating Solar System, 2025 13th International Electrical Engineering Congress (iEECON2025), Huahin, Thailand.

W. Thongwily, N. Junhuathon, and K. Chayakulkheeree (2026), Short-Term Solar Power Forecasting Under Incomplete Data and Curtailment Conditions, 2026 14th International Electrical Engineering Congress (iEECON2026), Pattaya, Thailand, accepted for presentation.

W. Thongwily, N. Junhuathon, and K. Chayakulkheeree (2025), Augmented Data-Driven Solar Power Forecasting Using Pre-Forecasting and Clear-Sky/KNN Data Refinements, GMSARN International Journal.



Two-Stage Data-Driven Modeling for Day-Ahead Solar Power Forecasting: A Case Study of the SUT Campus Floating Solar System

Wipada Thongwily
School of electrical engineering
Institute of Engineering
Suranaree University of Technology
Nakhon Ratchasima, Thailand
wipada.thon@gmail.com

Nitikorn Junhuathon
Department of electrical engineering
Faculty of engineering
Rajamangala University of
Technology Thanyaburi
Pathum Thani, Thailand
nitikorn_j@rmutt.ac.th

Keerati Chayakulkheeree*
School of electrical engineering
Institute of Engineering
Suranaree University of Technology
Nakhon Ratchasima, Thailand
keerati.ch@sut.ac.th

Abstract— This paper presents a comparative study of forecasting methods for solar PV power generation, using a two-step forecasting approach. The data includes solar irradiance, global tilted irradiance, temperature, humidity, and wind speed, analyzed with LSTM, Bi-LSTM, GRU, and RF models, 7,665 samples of 30-minute recordings from 2023. In the first part of the study, the important factors affecting power generation, irradiance and temperature, are forecasted using the LSTM method to consider the relationship between external factors and power generation. In the second part of the study, the results from the aforementioned forecasting are used to predict power generation. The forecast results obtained by LSTM, Bi-LSTM, GRU, and RF, were investigated and compared. This study not only helps to better understand the relationship between external factors and power generation but also significantly improves the overall forecasting efficiency.

Keywords—short-term PV forecasting, LSTM, Bi-LSTM, GRU, RF

I. INTRODUCTION

The world's population has been rapidly increasing, led to a rise in the need for energy and power worldwide. In addition to being finite, fossil fuels pollute the environment by releasing greenhouse gases (GHGs) and carbon dioxide. Reducing reliance on these energy sources is therefore essential in light of these findings. As a result, renewable energy sources (RESs), which are sustainable, clean, and non-polluting energy sources, are becoming a significant alternative to polluting energy resources, such as coal, oil, and natural gas [1]. It also helps reduce GHG emissions that cause global warming, supporting the Net Zero Emission target as a policy of the Thai government by 2050 [2,3]. It also reduces the import of fuels from abroad, promotes community participation in electricity production, and reduces the cost of electricity production [4,5].

Nevertheless, the uncertainty of renewable energy (RE) generation are major challenges for the widespread utilization of RE [6]. Among the types of RE, photovoltaic (PV) energy is a clean energy that has been widely utilized in various fields. However, solar energy production depends on various factors; weather conditions, solar irradiance, ambient temperature, module temperature, cloud cover, and relative humidity received by solar panels, resulting in high fluctuations and uncertainties in the amount of energy produced at each time, which may affect the reliability and stability and difficult to integrate REs into the electricity grid effectively [7]. Therefore, accurate solar energy production forecasting is required to deal with uncertainty and variability while maintaining the balance of the power system. Solar energy

production forecasting is one of the main pillars of Thailand's smart grid plan, to promote the development of a modern electricity grid system that is efficient and environmentally friendly [8].

Basically, the forecasting of energy production from PV systems is divided into four types according to the period of data collection and training which are, very short-term, short-term, medium-term, and long-term, as shown in Fig. 1. [9]. Each period has different purposes and methods. Very Short-Term Forecasting focuses the period from one second to one hour, suitable for real-time power control or response to sudden fluctuations in sunlight. Meanwhile, short-term forecasting is implemented for the period of one day, which is used for load scheduling and daily energy balancing. Medium-term forecasting forecasts in the period from one week to one month for energy management planning. Next, long-term forecasting uses the period of annual basis for strategic planning, such as PV system design, energy grid investment, or climate change impact assessment.

In [10], it is proposed that PV power forecasting studies can be categorized into five types: persistence models, physical models, statistical models, artificial intelligence (AI)-based models, and hybrid models. In the past, persistence models, physical models, and statistical models were used to predict PV power output. Persistence models have been used since the beginning of PV forecasting because they are the easiest to use and use historical data to predict the future, assuming that weather conditions or electricity costs will not change in the next period. They are suitable for situations where weather conditions do not change much. Physical models have played an important role in the past, especially in the era when data for model training was limited. They use physics principles and system parameters such as temperature, solar irradiance, and solar panel angles, so they have good accuracy when the physical data is correct. When time series data was introduced, it was developed into statistical models using statistical techniques such as ARIMA and linear regression, which are suitable for time series data. However, the above-mentioned techniques often provide low-accuracy forecasts and cannot work properly with non-linear data. Currently, due to the advancement of AI and the collection of high-resolution weather data, it has been developed into AI-based models, such as artificial neural networks (ANN), long short-term memory (LSTM), and deep learning (DL) models, due to their high accuracy for complex data and their ability to manage complex and non-linear data. Meanwhile, hybrid models, which combine physical models and AI-based models for flexible and highly accurate are proposed by many researchers to reduce errors of single models.

*Corresponding author

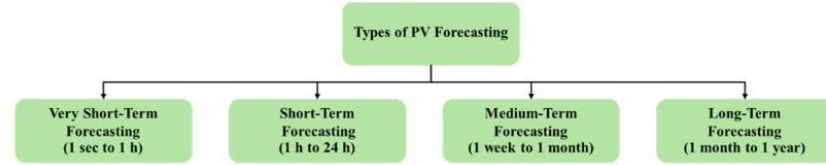


Fig. 1. Classification Methods for PV Power Prediction

In solar power forecasting, LSTM is a very popular technique due to its suitability for time-series data and its ability to learn complex features well. This can be seen from the research [17, 18]. Later, in [19], proposed a method to compare the one-day solar power forecasting of LSTM and GRU by using a Gaussian process for hyperparameters optimization. GRU outperforms LSTM for longer training data, especially with comprehensive weather forecasting data. However, the authors in [20] predicted the PV power generation using LSTM, GRU, and MLP. From the results, the LSTM outperforms both GRU and MLP models. Later, in 2024, the authors in [21] forecasted the PV output power using two years of 5-min collected datasets from a fixed rooftop grid-connected PV system using a Bidirectional LSTM (Bi-LSTM), One-Dimensional Convolutional Neural Network (1D-CNN), and a GRU models. The results demonstrate the accuracy of GRU and BiLSTM. In [16], the PV power generation forecasting of the Rowta Solar Power Plant was presented using an LSTM model and an RF model. The results indicate that the LSTM method is suitable for future solar power generation forecasting.

However, the forecasting of solar cell power generation has been a topic of widespread interest in recent years. Most research mainly focuses on improving the forecasting accuracy and model training efficiency, which may overlook the factors that influence power generation, such as light intensity and temperature. Therefore, the motivation of this paper is to propose a two-step forecasting method that focuses on forecasting the key factors affecting power generation, namely light intensity and temperature, in the first step and then using the forecasting results to forecast power generation in the next step. This method provides a deeper analysis of the relationship between external factors and power generation and improves the overall forecasting accuracy.

This paper addresses a comparative investigation of methods for solar power generation forecasting, focusing on two-step forecasting. The comparative study on LSTM, bi-LSTM, GRU, and RF were investigated with the data from floating PV system of Suranaree University of Technology (SUT) campus.

The rest of the paper is arranged as follows: Section II illustrates the Data Preparation, Section III describes the Model Development and Evaluation Process, Section IV describes the simulation result, and the final section is the conclusion.

II. DATA PREPARATION

The dataset used in this paper is obtained from a floating PV power plant located at latitude 14.8801, longitude 102.0153, within Suranaree University of Technology, Nakhon Ratchasima Province, Thailand. The installed capacity is 1.5 MW/unit, and The power output data was recorded for the entire year of 2023. The PV power generation

is zero between 17:00 and 07:00. Therefore, we only consider the power between 07:00 and 17:00 in the simulation, and the power data is recorded every 30 minutes.

In particular, this work aims at short-term PV power forecasting with 30-min resolution data for DL model. The one-day PV power forecast is divided into: summer, rainy season, and winter. The input data for DL model are PV power generation, solar irradiance, temperature, wind speed, humidity, and GTI. The dataset consists of 7,665 samples, divided into 3 sets: 7,539 samples of the training set and the validation set, which are 80% and 20%, respectively, and 126 samples of the test set.

The data cleaning process is an important step in preparing data for analysis or model training by checking and dealing with missing data and abnormal data by filling data with the linear interpolation method. Then, perform exploratory data analysis (EDA) by using correlation analysis to examine which factors affect electricity generation and can be used to determine input and output values from the evaluation of the relationship from the correlation heatmap, which is shown in Fig. 2.

The variables with high correlation coefficients with power are predicted irradiance, global tilted irradiance, and predicted temperature, which show a tendency to increase the electrical power, unlike the humidity variable, which has a negative correlation coefficient with the electrical power, which shows that the electrical power tends to decrease. Meanwhile, wind speed has a moderate relationship, which may have some effect on the electrical power, but not as high as the first group. Therefore, the variables suitable for use as inputs in forecasting solar cell electrical power are predicted irradiance, global tilted irradiance, predicted temperature, humidity, and wind speed.

From the above correlation heatmap analysis, the correlation coefficient between temperature and power is 0.23, and the value between light intensity and power is 0.67, which are important factors affecting electricity generation. Therefore, the temperature and light intensity are forecasted using the LSTM model, and the forecast results are analyzed from the correlation heatmap. It was found that the correlation coefficient between temperature and light intensity and power has increased. Therefore, the values obtained from the temperature and light intensity forecast are used to forecast electricity generation. The input data includes predicted irradiance, global tilted irradiance, predicted temperature, humidity, and wind speed, which can be explained in the work diagram from Fig. 3.

III. MODEL DEVELOPMENT AND EVALUATION PROCESS

For the models used in forecasting the power generation of solar cells, they consist of LSTM, Bi-LSTM, GRU, and RF, which are detailed as follows:

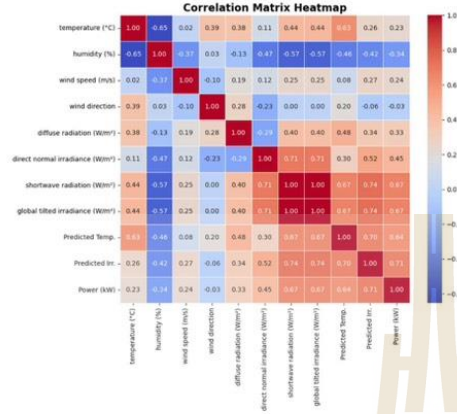


Fig. 2. The correlation matrix.

LSTM, introduced by Hochreiter and Schmidhuber in 1997 [11]. It consists of a cell and three gates: input, output, and forget gates [12-13]. These gates control the flow of data into and out of the cell. The forget gate decides which data to discard by assigning a value between 0 and 1, where 1 means data is retained and 0 means discarded. The input gate determines which new data to store, while the output gate controls the output of the current cell state. The structure of LSTM is shown in Fig. 4.

Bi-LSTM is developed using two temporal directions [14]. Bi-LSTM can learn more accurately than unidirectional LSTM. Bi-LSTM takes information from both past and future as input by connecting neurons from two separate hidden layers to a single output. Information is sent to all hidden layers from both the front and back layers, but the neurons in each layer are not in contact with each other. The output of Bi-LSTM, y_i , is computed by combining the results from both directions. The structure of Bi-LSTM is shown in Fig. 5.

GRUs are RNN models developed to efficiently track long-term dependencies. In handling gradient errors in the iterative processing unit, they are similar to LSTM models but have fewer parameters than LSTMs, resulting in higher training speeds. Fig. 6. shows the architecture of a GRU. At each step of training, the model modifies the incoming data through the update gate and the outgoing data through the reset gate, where the reset gate determines whether to use the previous state, while the update gate determines what data to recall from the previous step.

RF is an ensemble method using bagging and feature randomness to create uncorrelated decision trees [15]. The final result is the average of the predictions from all the trees. Since it averages predictions from multiple trees, this method results in low-variance predictions and increases the prediction accuracy. The structure of RF is shown in Fig. 7 [16].

For the parameter configuration, the structural parameters of all three models are designed the same for the training process, unlike the RF, which has different structures, the details of which are shown in Table I.

TABLE I. HYPERPARAMETER CONFIGURATION FOR MODEL TRAINING

Model	Hyperparameters
LSTM	Input: 5, Hidden nodes: 64,32,50, Epochs: 100, Batch Size: 32, Learning Rate: 0.001, Optimizer: Adam, Early Stopping: 20
Bi-LSTM	Input: 5, Hidden nodes: 64,32,50, Epochs: 100, Batch Size: 32, Learning Rate: 0.001, Optimizer: Adam, Early Stopping: 20
GRU	Input: 5, Hidden nodes: 64,32,50, Epochs: 100, Batch Size: 32, Learning Rate: 0.001, Optimizer: Adam, Early Stopping: 20
RF	Input: 5, Number of Trees: 100, Max Depth: 10, Bootstrap: True, Criterion: MSE, Max Features: auto

In the final step, the evaluation metrics of the appropriate model of the PV power forecast are selected to verify the prediction and efficiency of PV generation, which makes the prediction results more reliable and applicable. In this work three fitness functions, namely RMSE, MAE, and MAPE, are used, which can be called by Sklearn library of Python. The first comparative index considered is the Root Mean Square Error (RMSE). It can be measured by averaging the squares of the errors and extracting the square root as Eq.(1)

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_{actual,i} - y_{forecast,i})^2}{n}} \quad (1)$$

The next index used is Mean Absolute Error (MAE), as in Eq.(2). Mean Absolute Percentage Error (MAPE), as shown in Eq.(3).

$$MAE = \frac{\sum_{i=1}^n |y_{actual,i} - y_{forecast,i}|}{n} \quad (2)$$

$$MAPE = \sum_{i=1}^n \frac{|y_{actual,i} - y_{forecast,i}|}{y_{real,i}} \times 100\% \quad (3)$$

Where $y_{real,i}$ is the actual power generated at sample i^{th} italic, Where $y_{forecast,i}$ is the predicted power generated at sample i^{th} , and n is sample number.

IV. SIMULATION RESULT

In this research, the performance evaluation of LSTM, Bi-LSTM, GRU, and RF models for day-ahead power generation forecasting of floating PV power plants at Suranaree University of Technology was studied. The dataset used consisted of data in 2023, which was divided into 3 datasets: 7,539 samples of training and validation sets, which were divided into 80% and 20%, respectively, and 126 samples of testing sets, which ensured the diversity of data for model validation. Since some missing data may affect the accuracy and reliability of the forecasting results, the data were filled in using linear interpolation, which increased the accuracy of solar cell power generation forecasting by forecasting solar irradiance and temperature using LSTM, which are factors that directly affect power generation. The results of the relationship between environmental factors (irradiance and ambient temperature) and solar cell power generation in April can be shown in Fig. 8. Then, the data obtained from the

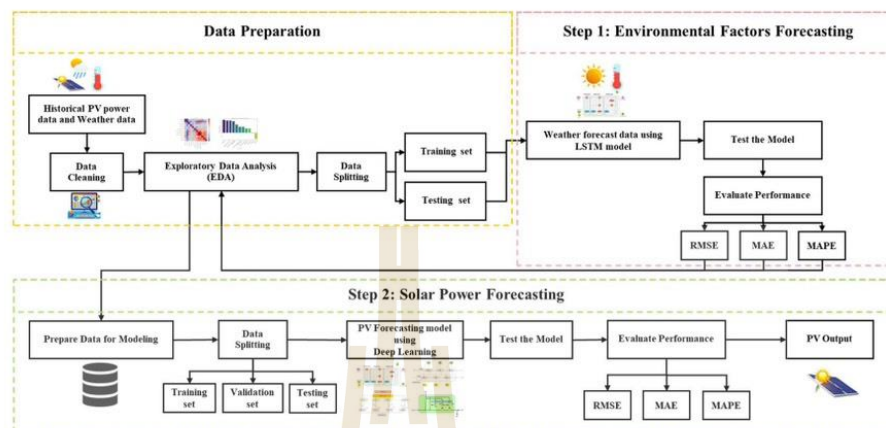


Fig. 3. The Framework for Solar Power Forecasting

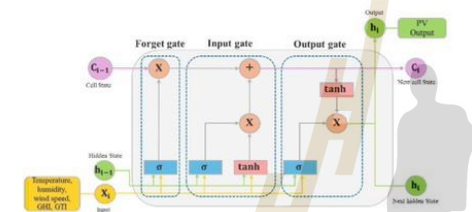


Fig. 4. Architectural Diagram of LSTM

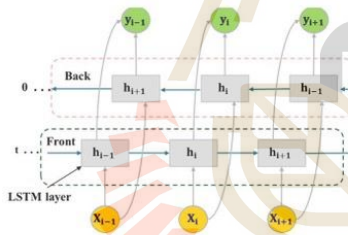


Fig. 5. Architectural Diagram of Bi-LSTM

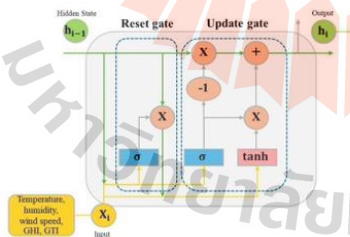


Fig. 6. Architectural Diagram of GRU

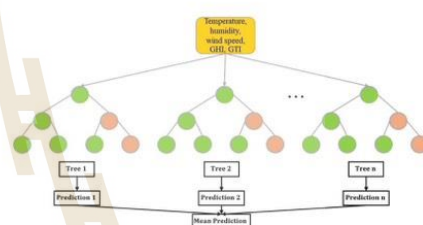


Fig. 7. Architectural Diagram of RF

forecasting were used to forecast solar cell power generation using all 4 models above, and evaluate the model accuracy using RMSE, MAE, and MAPE.

Figure 9, shows the comparison between the actual and predicted power generation of the four models, of three seasons: summer season (a), rainy season (b), and winter season (c).

The calculation results of the evaluation indicator using the predicted and actual power are shown in Table II, which shows the RMSE, MAE, and MAPE values for RF, which are 260 kW, 188 kW, and 11.35%, respectively, which are the best forecasting methods among the four methods for summer considering the three validation factors. However, in the rainy season, the prediction results of GRU are the best, which are 310 kW, 272 kW, and 18.37%, respectively. On the contrary, for the winter season, the best forecasting method is Bi-LSTM with values of 271 kW, 227 kW, and 24.65%, respectively. For the LSTM method, in summer, the accuracy is relatively lower than RF but higher than GRU, which are 389 kW, 319 kW, and 19.26%, respectively. In the rainy season, the accuracy is higher than Bi-LSTM, which are 311 kW, 269 kW, and 19.20%, respectively. In contrast, winter has the lowest accuracy among the three seasons. However, the order of processing time for the models from least to most is as follows: RF, GRU, LSTM, and Bi-LSTM, with times of 12 seconds, 61 seconds, 92 seconds, and 137 seconds,

respectively, which is consistent with the aforementioned theory.

The simulation results show that RF performs well for summer power generation forecasting. During the summer, the changes in weather conditions result in a stable power generation capacity, leading to continuous and less volatile data used for model training. This allows the model to learn the relationships between various features effectively. In contrast, GRU is suitable for forecasting power generation during the rainy season, where the data is characterized by uncertainty and rapid changes in power generation capacity, resulting in high data volatility. This enables the model to adapt well to changes in various parameters. Bi-LSTM is the best forecasting method for winter, as several factors such as shorter daylight hours, lower temperatures, and higher humidity can influence power generation. This results in complex data patterns, allowing the model to capture sequential data in both directions and learn the complex relationships between various factors effectively.

In addition, the cause of the error in power generation forecasting may occur from some missing forecast data and differences in weather conditions which may directly affect the operating condition of solar panels.

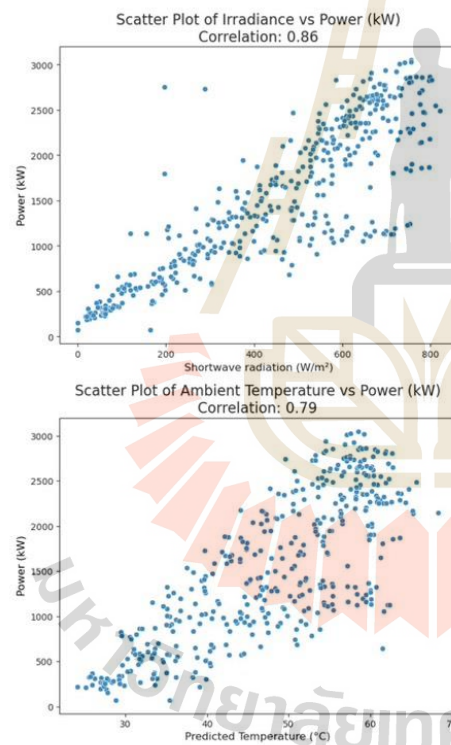


Fig. 8. Correlation of Radiation and Temperature with Solar Power

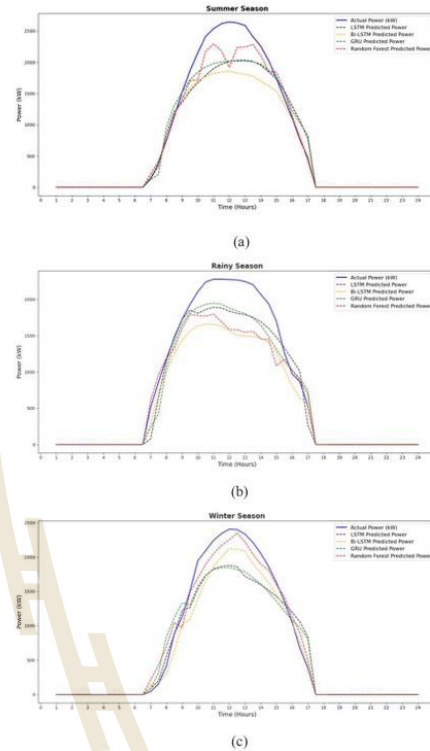


Fig. 9. Comparison of Day-Ahead Solar Power Forecasting Across Summer, Rainy, and Winter Seasons

TABLE II. FORECASTING RESULTS

Model	Summer		
	RMSE (kW)	MAE (kW)	MAPE (%)
LSTM	389	319	19.26
Bi-LSTM	468	357	17.73
GRU	360	295	20.45
RF	260	188	11.35
	Rainy		
	RMSE (kW)	MAE (kW)	MAPE (%)
LSTM	311	269	19.20
Bi-LSTM	481	423	24.17
GRU	310	272	18.37
RF	429	344	18.78
	Winter		
	RMSE (kW)	MAE (kW)	MAPE (%)
LSTM	365	310	29.14
Bi-LSTM	271	227	24.65
GRU	376	339	36.32
RF	166	148	38.97

CONCLUSION

This paper presents a comparative study of short-term PV forecasting methods in day-ahead with 30-minute details using LSTM, Bi-LSTM, GRU, and RF to improve the accuracy of power generation forecasting. This study forecasts solar irradiance and temperature in advance before using them in the actual power generation forecasting process. The

simulation results show that we should apply the RF method for continuous data with low fluctuation, such as in summer; the GRU method is suitable for data with high fluctuation during the rainy season, and the Bi-LSTM method is suitable for winter when the data is quite complex. The model processing is ranked in order of the least time as RF, GRU, LSTM, and Bi-LSTM, respectively. In addition, the appropriate management of some missing data may affect the model learning and increase the forecasting accuracy. Because the missing data characteristics in each season are different, it affects the model learning. If the data is improperly managed, the forecasting performance for that season may decrease. The accuracy could be further improved by data grouping in data management before forecasting to ensure appropriate data management and enhance parameter optimization to improve model efficiency.

ACKNOWLEDGMENT

This research was supported by the Suranaree University of Technology.

REFERENCES

- [1] EGAT. "Renewables." Electricity Generating Authority of Thailand. <https://www.egat.co.th/home/renewables/>.
- [2] NXPO. "Thailand's Development and Use of Clean Energy for a Sustainable Future." *Office of National Higher Education Science Research and Innovation Policy Council (NXPO)*. <https://www.nxpo.or.th/th/9651/>.
- [3] BOI. "Promoting Investment in Renewable Energy from Small-Scale Sources for a Sustainable Future." *Board of Investment of Thailand (BOI)*. <https://ic.or.th/th/about-us/boi-news/1275>.
- [4] Rafique MM, Bahaidarah HM, Anwar MK. Enabling private sector investment in off-grid electrification for cleaner production: Optimum designing and achievable rate of unit electricity. *J Clean Prod* 2019;206:508–23. <http://dx.doi.org/10.1016/j.jclepro.2018.09.123>.
- [5] Benti, N.E.; Mekonnen, Y.S.; Asfaw, A.A. Combining green energy technologies to electrify rural community of Wollega, Western Ethiopia. *Sci. Afr.* 2023, 19, e01467. <https://doi.org/10.1016/j.sciaf.2022.e01467>
- [6] Denholm, P.; Arent, D.J.; Baldwin, S.F.; Bilello, D.E.; Brinkman, G.L.; Cochran, J.M.; Cole, W.J.; Frew, B.; Gevorgian, V.; Heeter, J.; et al. The challenges of achieving a 100% renewable electricity system in the United States. *Joule* 2021, 5, 1331–1352. <https://doi.org/10.1016/j.joule.2021.03.028>
- [7] Imprim, S.; Nese, S.V.; Oral, B. Challenges of renewable energy penetration on power system flexibility: A survey. *Energy Strat. Rev.* 2020, 31, 100539. <https://doi.org/10.1016/j.esr.2020.100539>
- [8] Palunggan Wann. "Renewable Energy News and Updates." Palunggan Wann. <https://palungganwann.com/news/view/1217>.
- [9] Akhter, M.N.; Mekhilef, S.; Mokhlis, H.; Shah, N.M. Review on Forecasting of Photovoltaic Power Generation Based on Machine Learning and Metaheuristic Techniques. *IET Renew. Power Gener.* 2019, 13, 1009–1023. <https://doi.org/10.1049/iet-rpg.2018.5649>
- [10] <https://www.sciencedirect.com/science/article/pii/S2352484723009022#b1> <https://doi.org/10.1016/j.egy.2023.05.154>
- [11] Sepp Hochreiter; Jürgen Schmidhuber (1997). "Long short-term memory". *Neural Computation*. 9 (8): 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- [12] Hochreiter, Sepp; Schmidhuber, Jürgen (1996-12-03). "LSTM can solve hard long time lag problems". *Proceedings of the 9th International Conference on Neural Information Processing Systems. NIPS'96*. Cambridge, MA, USA: MIT Press: 473–479. <https://dl.acm.org/doi/10.5555/2998981.2999048>
- [13] Felix A. Gers; Jürgen Schmidhuber; Fred Cummins (2000). "Learning to Forget: Continual Prediction with LSTM". *Neural Computation*. 12 (10): 2451–2471. <http://dx.doi.org/10.1162/089976600300015015>
- [14] A. Graves, N. Jaitly, A.R. Mohamed, Hybrid speech recognition with deep bidirectional LSTM, in: 2013 IEEE Work. Autom. Speech Recognit. Under standing, ASRU 2013- Proc, IEEE, 2013, pp. 273e278, <https://doi.org/10.1109/ASRU.2013.6707742>.
- [15] IBM. "Random Forest: An Overview of Applications and How It Works." *IBM*. <https://www.ibm.com/topics/random-forest>.
- [16] N. Roy et al., "Day-ahead Solar Power Generation Forecasting using LSTM and Random Forest Methods for North Eastern Region of India," 2022 22nd National Power Systems Conference (NPSC), New Delhi, India, 2022, pp. 854–859, <https://doi.org/10.1109/NPSC57038.2022.10069833>
- [17] H. Zhou, Y. Zhang, L. Yang, Q. Liu, K. Yan and Y. Du, "Short-Term Photovoltaic Power Forecasting Based on Long Short Term Memory Neural Network and Attention Mechanism," in *IEEE Access*, vol. 7, pp. 78063–78074, 2019, <https://doi.org/10.1109/ACCESS.2019.2923006>
- [18] Jailani, N. L. M., Dhanasegaran, J. K., Alkaws, G., Alkahtani, A. A., Phing, C. C., Baashar, Y., Capretz, L. F., Al-Shetwi, A. Q., & Tiong, S. K. (2023). Investigating the Power of LSTM-Based Models in Solar Energy Forecasting. *Processes*, 11(5), 1382. <https://doi.org/10.3390/pr11051382>
- [19] Y. Gao, S. Qi and J. Ponočko, "Assessing Critical Data Types for Deep Learning-Based PV Generation Forecasting," 2023 *IEEE Belgrade PowerTech*, Belgrade, Serbia, 2023, pp. 1–6, <https://doi.org/10.1109/PowerTech55446.2023.10202973>
- [20] L. N. Boucetta, Y. Amrane and S. Arezki, "Comparative Analysis of LSTM, GRU, and MLP Neural Networks for Short-Term Solar Power Forecasting," 2023 International Conference on Electrical Engineering and Advanced Technology (ICEEAT), Batna, Algeria, 2023, pp. 1–6, <https://doi.org/10.1109/ICEEAT60471.2023.10426346>
- [21] Ait Mansour, A., Tilioua, A., & Touzani, M. (2024). Bi-LSTM, GRU, and 1D-CNN models for short-term photovoltaic panel efficiency forecasting case: Amorphous silicon grid-connected PV system. *Renewable and Intelligent Energy*, 101886. <https://doi.org/10.1016/j.rineng.2024.101886>

Biography



Wipada Thongwily earned her B.Eng. in EE from Suranaree University of Technology, Thailand, in 2024. She is now enrolled in M.Eng. in EE program at the School of EE, Institute of Engineering, SUT. Her research interests include deep learning technique application in power systems and related fields.



Nitikorn Junhuathon completed his B.Eng., M.Eng., and Ph.D. in EE from SUT Thailand in 2017, 2019, and 2023, respectively. Currently, he serves as a lecturer in the Department of EE at the Faculty of Engineering, Rajamangala University of Technology Thanyaburi (RMUTT), Thailand. His research focuses on machine learning, optimization techniques for power forecasting, electric vehicle applications, and related areas.



Keerati Chayakulkheeree earned his B.Eng. in EE from King Mongkut's Institute of Technology Ladkrabang (KMUTL) in 1995, followed by M.Eng. and D.Eng. in Electric Power System Management from the Asian Institute of Technology (AIT), Thailand, in 1999 and 2004, respectively. He is an associate professor at the School of EE, Institute of Engineering, SUT, Thailand. His research areas encompass power system optimization, artificial intelligence applications in power systems, smart grid technologies, and related fields.

Short-Term Solar Power Forecasting Under Incomplete Data and Curtailment Conditions

Wipada Thongwily
School of Electrical Engineering
Institute of Engineering
Suranaree University of Technology
Nakhon Ratchasima, Thailand
email: wipada.thon@gmail.com

Nitikorn Junhuathon
Department of Electrical Engineering
Faculty of Engineering
Rajamangala University of
Technology Thanyaburi
Pathum Thani, Thailand
email: nitikorn_j@rmutt.ac.th

Keerati Chayakulkheeree*
School of Electrical Engineering
Institute of Engineering
Suranaree University of Technology
Nakhon Ratchasima, Thailand
email: keerati.ch@sut.ac.th

Abstract—This paper presents a framework for day-ahead solar PV power generation forecasting for floating solar power plants, using 30-minute data from 7,665 samples in 2023. This study focuses on a two-step forecasting approach, predicting solar irradiance and temperature using an LSTM before power generation is forecasted using a Random Forest (RF) model with Explainable AI (XAI) added for interpretability. The dataset is aligned with the system's curtailment limitations. Experimental results show that considering curtailment and time intervals significantly reduces daily relative error (MAPE) by up to 33% and improves forecasting accuracy for each season. This study not only provides a deeper understanding of the impact of external factors on power generation but also highlights that the accuracy and stability of forecasts do not solely depend on the complexity of the model but crucially on proper data management. The proposed framework can be further leveraged to support effective planning and operation of PV systems under curtailment constraints.

Keywords— Power Generation Forecasting, Solar Energy, LSTM, RF, Curtailment, XAI

I. INTRODUCTION

In recent years, the global trend toward a low-carbon society has increased the demand for clean electricity [1], making renewable energy a crucial player in reducing reliance on fossil fuels and greenhouse gas emissions, especially solar power, which has high potential for sustainable electricity generation. In the future, solar power will be used more extensively and may become the primary fuel source, according to the 2024-2037 AEDP draft target that 68% of renewable electricity by 2037 will come from solar power [2]. Therefore, forecasting solar power generation capacity is important for planning the electricity system and managing the load to operate efficiently, supporting Thailand's Net Zero Emission target [3]. However, the uncertainty of solar radiation and weather fluctuations affects the accuracy of forecasts, making it necessary to develop techniques and models that can manage these uncertainties. Renewable electricity generation also faces resource limitations that affect production stability, such as the intermittent and volatile nature of solar power, which reduces production efficiency and causes uncertainty for the overall power system.

Another significant challenge in solar energy systems is curtailment, or adjusting electricity generation to match load demand. Even though PV systems can generate electricity at full capacity, it is necessary to reduce output to match actual load in order to maintain system stability and prevent energy loss. Research [4-5] has studied curtailment at the regional level using Stochastic Unit Commitment or the Perfect Forecast Metric to reduce forecasting errors and address the uncertainty of PV. Unlike this study, which forecasts curtailment at the university level, curtailed forecasting differs

from general forecasting because it must consider both maximum generation capacity and actual electrical load, not just full potential output. This type of forecasting helps plan energy use efficiently, reduce losses, and support stable electrical system management.

The prediction of electricity generation from solar cell systems has rapidly evolved with advancements in data and computational technology. The study in [6] categorizes forecasting approaches into five types: persistence models, physical models, statistical models, artificial intelligence (AI)-based models, and hybrid models. This reflects the development from basic methods to more intelligent deep learning models. Numerous research studies have applied AI and machine learning (ML) techniques for solar power generation forecasting using time series data. Deep learning models such as LSTM [7] can learn non-linear relationships from time series data well. While the Random Forest (RF) model [8] is not a direct time series model, it can handle non-linear feature variables well, providing high accuracy for multi-feature data. In addition to applying AI and ML for solar power generation forecasting, they are also used for temperature [9] and solar irradiance [10] forecasting.

However, the accuracy of the forecast heavily depends on the quality of the data, as missing and poor data can be caused by sensor errors or severe weather, which directly impacts model performance. Therefore, data quality management before the model learning process is extremely important. Only a few research studies have systematically focused on preparing data before model learning, finding that various methods such as linear interpolation [10-11] and K-Nearest Neighbors (KNN) [12] are used to handle missing data. Additionally, current research trends [13] emphasize the ability to explain model results using Explainable AI (XAI) techniques, particularly SHapley Additive exPlanations (SHAP), which helps identify the importance of independent variables to the prediction results, making it easier to understand the model's decisions and increasing its reliability for practical application.

Additionally, most research on forecasting solar cell power generation focuses primarily on improving the accuracy of single-step forecasting, which may overlook factors influencing power generation such as light intensity and temperature. Therefore, the motivation for this article is to present a two-step forecasting method that focuses on forecasting light intensity and temperature in the first step and then uses the results of this forecasting to predict power generation in the next step. This method also handles missing and anomalous data using the Clear Sky Model and KNN techniques and restructures the data to align with system curtailment conditions. Furthermore, it develops deep learning models for forecasting solar cell power generation using real

*Corresponding author

data from the FPV system at Suranaree University of Technology (SUT). This method not only helps understand the system's behavior under production constraints and data fluctuations but also improves the overall forecasting accuracy. This paper therefore presents a comparative study of methods for solar energy production forecasting, focusing on two-step forecasting.

The rest of the paper is arranged as follows: Section II illustrates the Data Preparation, Section III describes Forecasting Model Development and Evaluation, Section IV describes the simulation result, and the final section is the conclusion.

II. DATA PREPARATION

The dataset used in this paper is obtained from a floating PV (FPV) power plant located at latitude 14.8801, longitude 102.0153, within Suranaree University of Technology, Nakhon Ratchasima Province, Thailand. The installed capacity is 4.27 MW/unit, and The power output data was recorded for the entire year of 2023. The PV power generation is zero between 17:00 and 07:00. Therefore, we only consider the power between 07:00 and 17:00 in the simulation, and the power data is recorded every 30 minutes.

The data used in this study consists of weather data and electricity generation data from solar cell systems. This data can be accessed through the websites [14] and [15], respectively. The weather data includes the following variables: temperature ($^{\circ}\text{C}$), relative humidity, wind speed (m/s), global horizontal irradiance (W/m^2), and global tilted irradiance (W/m^2). For the power generation dataset used in the study, it was found that 13.95% of the total data was missing. Additionally, further data analysis revealed anomalous behavior in power generation.

The data cleaning process is a crucial step in preparing data for analysis and model training. It involves checking and addressing missing and anomalous data by filling in data using the Clear Sky model and KNN Imputation to reflect sunlight and weather conditions at the time. This ensures that missing data estimations are more reasonable and improves the accuracy of forecasting models. The Clear Sky model was used for summer and winter seasons, while KNN Imputation was used to fill in missing power generation data for the rainy season.

TABLE I. QUANTITATIVE METRICS

Metric	KNN-all	ClearSky with KNN
Mean	1279.71	1451.68
Median	1128.33	1356.49
SD	706.65	905.86
IQR	996.22	1304.81

To quantitatively evaluate the effectiveness of both methods, a comparison of baseline statistics, including mean, median, standard deviation, and interquartile range of the filled-in power generation values, was conducted, separated by method. The results showed that ClearSky with KNN Imputation yielded higher baseline statistics than KNN-all, reflecting that referencing the ClearSky model for data filling improves the accuracy of power generation to better reflect actual sunlight conditions. Furthermore, it maintains clearer seasonal distribution and volatility. Overall, ClearSky with KNN is able to preserve seasonal patterns and real production

ranges better than KNN-all, as shown in Table I., resulting in data that is more suitable for subsequent forecasting.

Then, an exploratory data analysis (EDA) was performed using correlation analysis to identify factors influencing power consumption. These factors can be used to define the input and output variables for the forecasting model. The results of the correlation heatmap are shown in Fig. 1.

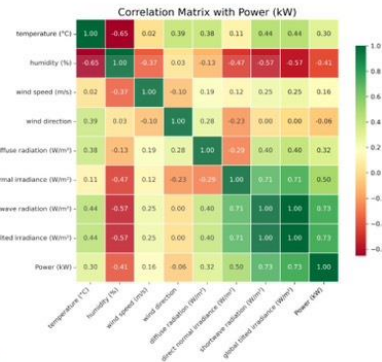


Fig. 1. The correlation matrix.

From the correlation heatmap analysis, it can be seen that the correlation coefficient between light intensity and power (kW) is 0.73, reflecting that light intensity is the main factor determining the power generation level of the solar cell system. Meanwhile, the correlation coefficient between temperature and power (kW) is 0.30, which is an important factor affecting power generation. Therefore, temperature and light intensity were predicted using the LSTM model to be used as inputs for the power generation prediction model in the next step.

To assess the quality of the input data after forecasting, the results were analyzed using a Taylor Diagram as shown in Fig. 2. This diagram compares the actual and predicted values of temperature and solar radiation, considering correlation, standard deviation, RMSE, and normalized bias. The results show a clear improvement in data quality after forecasting compared to the actual values. Temperature had a correlation coefficient of 0.63, while solar radiation had a high correlation coefficient of 0.74, with RMSE decreasing to 13.41 $^{\circ}\text{C}$ and 184.44 W/m^2 , respectively. This indicates a reduction in overall error and a closer alignment of the data distribution with the actual values.

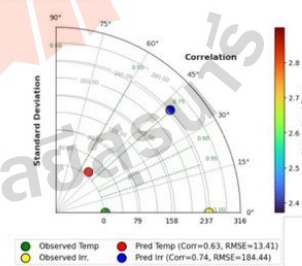


Fig. 2. Taylor diagram of temperature and irradiance.

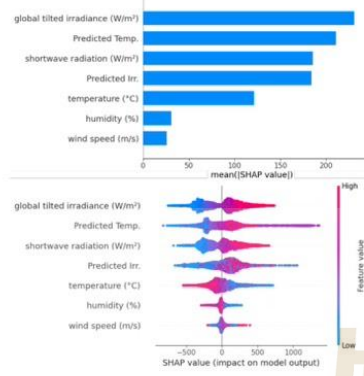


Fig. 3. SHAP Values Impact on Model Output.

However, the Taylor Diagram only reflects that the quality of the input data after forecasting is better compared to the actual values, but it is only an assessment of the consistency between the forecasted input and the actual input. Therefore, it cannot be directly concluded that the power generation forecast will be more accurate. The forecasted results were analyzed using SHAP, as shown in Fig. 3, to evaluate the factors affecting power generation. It was found that GTI remains the main factor affecting power generation the most, while the forecasted values, such as predicted temp. and predicted irr., demonstrate positive impacts consistent with the behavior of the solar cell system, reflecting the quality of post-forecast data for input in the power generation forecasting stage. Additionally, SHAP reveals non-linear relationships such as the temperature effect changing direction across value ranges and the saturation characteristics of solar radiation, which are not visible from basic statistics. This leads to a clearer understanding of the role of each variable in power generation. The input data includes predicted irr., global tilted irradiance (W/m²), predicted temp., humidity, and wind speed (m/s).

However, behavioral analysis of power generation data combined with internal system electrical load also revealed periods where solar cell system power generation was curtailed, meaning the output value did not reflect the system's true production potential even when weather conditions were favorable for power generation. Curtailment differs from missing data because it results from structural limitations of the electrical system, not data loss or measurement errors.

To systematically identify periods of curtailment, this study examines the relationship among PV power generation, actual load, and grid power consumption (metering load). The ratio between PV power generation and actual load is defined according to Eq. (1), while the curtailment status is determined using a piecewise conditional equation, as shown in Eq. (2).

$$R(t) = \frac{P_{PV}(t)}{P_L(t)} \quad (1)$$

Where $R(t)$ represents the ratio reflecting the relationship between PV power generation and the actual electrical load at time t , $P_{PV}(t)$ denotes the power generated by the PV system, $P_L(t)$ denotes the actual electrical load (actual load), and $P_G(t)$

denotes the electrical power drawn from the utility grid (metering load).

$$Curtailment(t) = \begin{cases} 1, & \text{if } R(t) \geq \alpha \text{ and } P_G(t) \leq \beta \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

The parameter α represents a threshold defined to indicate conditions under which the PV power generation is comparable to or exceeds the actual load. Meanwhile, β represents a lower threshold of grid power consumption, used to indicate that PV generation predominantly supplies the load demand. Both conditions must be satisfied simultaneously for the system to be classified as operating under a curtailment state.

The proposed criteria jointly consider the condition that the PV to load ratio exceeds a threshold derived from statistical analysis of the year-long dataset and the condition that grid power consumption decreases significantly. This combined criterion is designed to accurately reflect the actual operating behavior of the system. Representative examples of data behavior on curtailment days and the annual distribution of curtailment events are illustrated in Fig. 4.

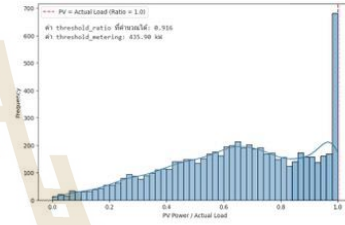


Fig. 4. Distribution of PV Power to Actual Load Ratio

In Fig. 4, it can be seen that the ratio of PV system output to actual load is concentrated near 1.0 throughout the daytime. This indicates that the output of the PV system is controlled to match the actual load, significantly reducing reliance on the power grid. It also suggests that the PV system's power output is systematically and efficiently limited.

The time series analysis results show that curtailment began around 8:00 AM. And the highest frequency is between 10:00 AM and 1:00 PM. This aligns with the peak production period for solar panel systems, before the frequency drops below average after approximately 2:00 PM, and it decreased continuously, becoming almost undetectable after 3:00–4:00 PM. These results show that curtailment is not random but has a clear temporal pattern directly related to the potential for solar power generation.

Due to the specific characteristics of curtailment data, all data was divided into three sets: the Curtailment Date Data Set, the Non-Curtailment Date Data Set, and the Year-Round Data Set, to study the impact of curtailment on the performance of the forecasting model. In each data set, the data was divided for training and validation at an 80/20 ratio, using 5 test days for the Curtailment and Non-Curtailment sets and 10 days for the Year-Round data set, which were the same days, to ensure a fair comparison of the model results.

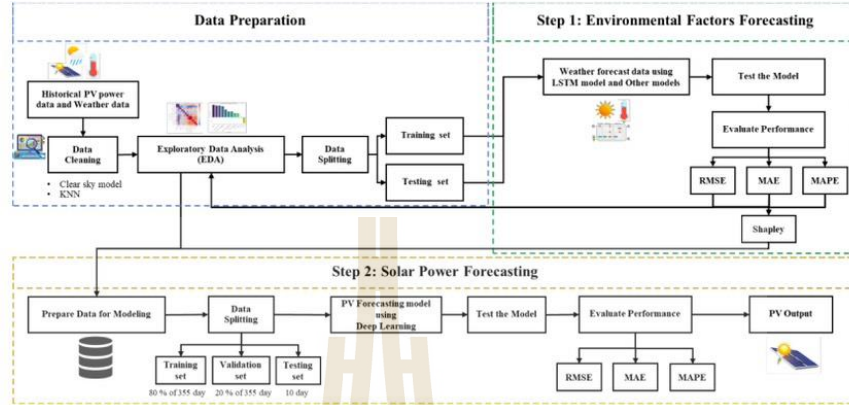


Fig. 5. The Framework for PV Power Forecasting

Additionally, each set is divided into 3 time periods of the day: the ramp-up period from 07:00 to 08:30, the peak or curtailment period from 08:30 to 14:30, and the ramp-down period from 14:30 to 17:00. To clearly reflect the nature of curtailment. To make the model training process faster and more stable, all inputs were scaled using a normalization process. This operation helps the model learn the relationships between variables stably, accelerates the convergence of the training process, and reduces problems caused by differences in scale between input features, which can be illustrated by the diagram in Fig. 5.

III. FORECASTING MODEL DEVELOPMENT AND EVALUATION

In this study, LSTM and RF models were used to predict environmental factors and electricity generation from solar cell systems. This is because the data is time series data with non-linear characteristics and time-varying variability. LSTM is a type of recurrent neural network [16] that is an artificial neural network capable of learning both short-term and long-term temporal relationships, making it suitable for predicting weather variables. RF, on the other hand, is an ensemble learning technique that relies on creating multiple decision trees using bagging and feature randomness, which helps reduce model variance, increase robustness to noise, and efficiently learn non-linear relationships between input variables and electricity generation [17]. Therefore, it was used to predict electricity generation in the next step. For configuring the structural parameter values of both models to suit the characteristics of the data used in the study, the details are shown in Table II.

TABLE II. HYPERPARAMETERS CONFIGURATION FOR MODELS TRAININGS

Model	Hyperparameters
LSTM	Input: 3, Hidden nodes: 128,64,32, Epochs: 200, Batch Size: 32, Learning Rate: 0.001, Optimizer: Adam, Early Stopping: 20
RF	Input: 5, Number of Trees: 100, Max Depth: 10, Bootstrap: True, Criterion: MSE, Max Features: auto

To evaluate the performance of the forecasting models on each dataset, this study uses basic statistical indicators,

including RMSE, MAE, and MAPE. These indicators are calculated using the Scikit-learn library in Python.

IV. SIMULATION RESULT

To improve the accuracy of solar cell power generation forecasting, this study uses an LSTM model to predict environmental factors that directly affect power generation, including solar irradiance and temperature. These factors directly impact power generation. The results showing the relationship between various environmental factors (solar radiation and ambient temperature) and the power generation of solar panels in August are shown in Fig. 6. The predicted values are then used as input for power generation forecasting in the next step.

Analysis of the relationship between environmental factors and solar panel power generation in August revealed that solar irradiance had a correlation coefficient of 0.81, indicating a high positive correlation. This means that as solar irradiance increases, power generation tends to increase accordingly. However, some data dispersion was also observed, which may be due to the influence of clouds, rain, and weather variability during the rainy season. Meanwhile, ambient temperature had a higher correlation coefficient of 0.85, indicating that temperature is another significant factor influencing power generation. In summary, the correlation coefficients for both factors were high and close to 1, indicating a clear positive relationship between environmental factors and power generation. This supports the suitability of using these factors as input variables for power generation forecasting.

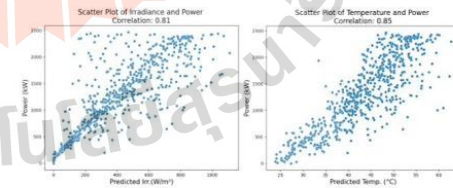


Fig. 6. Correlation of Irradiance and Temperature with Solar Power

Additionally, the monthly MAE of LSTM model for light intensity and temperature predictions are shown in Fig. 7. The results are presented in two parts: light intensity and temperature, respectively.

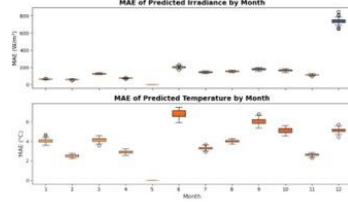


Fig. 7. MAE of Environmental Factor Forecasts

From the results of the monthly error analysis, it was found that the forecasts for solar irradiance and temperature showed clear differences. The MAE values for irradiance forecasts were mostly low and relatively constant throughout the year, but in December, the MAE and dispersion were found to be higher, reflecting weather fluctuations. In May, the MAE was close to zero due to missing and incomplete data.

For temperature forecasting, it was found that most MAE values were relatively low and consistent, ranging from 2 to 5°C. However, in June and September, the MAE values were higher, approximately 6–7°C, due to sudden temperature changes during those months, which the model was not yet able to handle effectively.

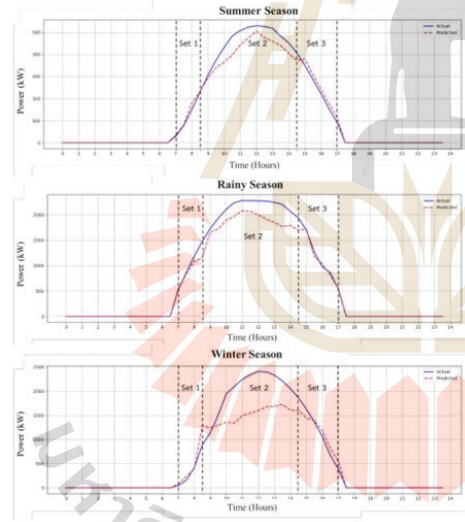


Fig. 8. Seasonal Comparison of Day-Ahead Solar Power Forecasting

However, the overall results show that the LSTM model can predict temperature more accurately and stably than solar irradiance and support the suitability of using environmental factor prediction results from LSTM as the first step in a two-stage prediction framework for FPV system power generation.

For the 1-day PV energy forecast, the results are presented by 3 seasons: summer (mid-February to mid-May), rainy season (mid-May to mid-October), and winter (mid-October to mid-February). The actual power output is compared with the forecast power output for the entire year's dataset, divided into the three time periods of the day as mentioned above: the rising period, the peak period, or the curtailment period, and the falling period, respectively. The forecast for each time period is made separately for each day. The forecast results are then stacked to obtain the daily forecast power generation, making the difference between actual and forecast power generation clearly visible, as shown in Fig. 8.

TABLE III. FORECASTING RESULTS

Dataset	Summer		
	RMSE (kW)	MAE (kW)	MAPE (%)
Curtail	196	167	11.63
No Curtail	-	-	-
Entire	196	167	11.63
Dataset	Rainy		
	RMSE (kW)	MAE (kW)	MAPE (%)
Curtail	-	-	-
No Curtail	316	264	16.87
Entire	223	182	10.16
Dataset	Winter		
	RMSE (kW)	MAE (kW)	MAPE (%)
Curtail	-	-	-
No Curtail	270	264	37.57
Entire	434	326	24.18

TABLE IV. COMPARISON OF FORECASTING RESULTS BETWEEN BEFORE AND AFTER DATASETS ACROSS SEASONS.

RF	Summer		
	RMSE (kW)	MAE (kW)	MAPE (%)
[11]	260	188	11.35
Proposed model	196	167	11.63
RF	Rainy		
	RMSE (kW)	MAE (kW)	MAPE (%)
[11]	429	344	18.78
Proposed model	223	182	10.16
RF	Winter		
	RMSE (kW)	MAE (kW)	MAPE (%)
[11]	166	148	38.97
Proposed model	434	326	24.18

After calculating the power generation forecasting performance, the results are shown in Table III, displaying the RMSE, MAE, and MAPE values for each dataset in each season. It was found that the RF model could clearly learn different power generation behaviors depending on the dataset. In the summer, the Curtail and Entire datasets showed similar error values of 196 kW, 167 kW, and 11.63%, respectively, reflecting that the Curtail period can accurately represent the all-day production behavior because the power generation is stable and the data is continuous. In contrast, the No Curtail dataset had values of 316 kW, 264 kW, and 16.87%, respectively, showing higher errors than the all-day dataset, which had values of 223 kW, 182 kW, and 10.16%, respectively. This indicates that learning only from periods without power limitations cannot adequately adapt to data fluctuations. Combining the all-day datasets reduced the error and allowed the model to better learn the relationship between Curtail and No Curtail periods. Similarly, in winter, the No Curtail period had values of 270 kW, 264 kW, and 37.57%, respectively, due to low power generation and complex data patterns resulting in high relative error. However, combining data throughout the day helped reduce the MAPE to 24.18%.

Furthermore, forecasting by dividing the data into three time periods each day and stacking the results to get daily

energy data allows the model to efficiently learn the specific electricity production patterns of each time period and clearly reflects the system's behavior both during periods with and without curtailing. Compared to the approach from previous work [11] that used an RF model to forecast daily energy without separating time periods and without considering curtailment, as shown in Table IV. The results from the proposed approach showed the lowest overall error and the most consistent stability of results across all seasons. The stacking of results reduces time-specific errors and structural bias, resulting in higher accuracy of daily energy forecasting and better generalization to different weather and seasonal conditions.

In summary, the performance of power generation forecasting does not solely depend on the capabilities of the RF model but significantly on the structure and time period of the data used to train the model. Using data that does not cover curtailment behavior throughout the day tends to lead to errors and bias, especially in highly volatile conditions. Data management approaches that consider both curtailment and no-curtailment periods, along with time segmentation and stacking results, improve the accuracy, stability, and alignment with the actual operation of solar power systems for daily energy forecasting.

V. CONCLUSION

This paper presents a framework for forecasting solar power generation one day in advance using a two-step forecasting approach. This approach first forecasts irradiance and temperature using LSTM model and then uses the results as input for the RF model to predict power generation. The proposed methodology emphasizes designing the data structure to align with the system's power curtailment behavior by dividing the dataset into Curtail, No Curtail, and All Day data.

The experimental results show that the accuracy of forecasting does not depend solely on model complexity but significantly on the structure, time period, and quality of the data. Structuring the data to align with curtailment behavior through data splitting and stacking results reduces bias and increases the stability of daily energy forecasts.

Additionally, handling missing data by selectively using the Clear-Sky Model and KNN based on the season allows the input data to better reflect the actual behavior of the system. Although this research uses data from a single site, the proposed conceptual framework can be extended to effectively support system operation planning and energy management in PV systems with power limitations. For future research, the study can be expanded to compare other power generation forecasting models besides RF, as well as develop models for forecasting meteorological variables to analyze the impact of uncertainty during the forecasting stage. Additionally, the concept of handling missing data to align with the season can be applied to PV systems in other areas. Furthermore, integrating data from energy storage systems or applying it within a digital twin framework will enhance the study's potential to effectively support the management of solar energy systems in real-world operating conditions.

ACKNOWLEDGMENT

This research was supported by the Suranaree University of Technology.

REFERENCES

- [1] "Power plant sector shows strong growth with bright profit outlook; industrial and data center sectors to increase clean energy usage by 2026" (Translated from Thai), MGR Online, Sep. 11, 2025. [Online]. Available: <https://mgronline.com/business/detail/968000086927> [Accessed: Aug. 21, 2025].
- [2] "Thailand's renewable electricity business is expected to grow in 2026 in line with increasing demand" (Translated from Thai), Kasikom Research Center, Sep. 3, 2025. [Online]. Available: <https://www.kasikomresearch.com/th/analysis/k-social-media/Pages/IAO143-Electricity-FB-2025-09-03.aspx>. [Accessed: Aug. 21, 2025].
- [3] Office of National Higher Education Science Research and Innovation Policy Council (NXPO), "Thailand's development and use of clean energy for a sustainable future," 2025. [Online]. Available: <https://www.nxpo.or.th/th/9651/>. [Accessed: Aug. 21, 2025].
- [4] J. G. da Silva Fonseca Junior, Y. Nishitsuji, K. Ogimoto, T. Oozeki, and Y. Udagawa, "Regional photovoltaic power curtailment with forecasting and unit commitment scheduling: A study on the Kanto region in Japan," in Proc. Solar World Congress, 2017, doi: 10.18086/swc.2017.21.01.
- [5] T. Masuta, D. Kobayashi, N. H. H. Viet, and H. Ohtake, "Annual evaluation of supply-demand with BESS charging/discharging schedule and UC updating based on intraday forecasted PV power outputs," in Proc. Power Syst. Comput. Conf. (PSCC), Dublin, Ireland, Jun. 2018, pp. 1–7, doi: 10.23919/PSCC.2018.8442599.
- [6] T. Nguyen Trong, H. Vu Xuan Son, H. Do Dinh, H. Takano, and T. Nguyen Duc, "Short-term PV power forecast using hybrid deep learning model and variational mode decomposition," Energy Reports, vol. 9, pp. 712–717, Oct. 2023, doi: 10.1016/j.egyr.2023.05.154.
- [7] H. Sharadga, S. Hajimirza, and R. S. Balog, "Time series forecasting of solar power generation for large-scale photovoltaic plants," Renewable Energy, vol. 150, pp. 797–807, 2020, doi: 10.1016/j.renene.2019.12.131.
- [8] N. Roy et al., "Day-ahead solar power generation forecasting using LSTM and random forest methods for North Eastern Region of India," in Proc. 22nd Natl. Power Syst. Conf. (NPSC), Gandhinagar, India, Dec. 2022, pp. 854–859, doi: 10.1109/NPSC57038.2022.10069833.
- [9] P. Li et al., "Advancing photovoltaic panel temperature forecasting: A comparative study of numerical simulation and machine learning in two types of PV power plant," Renewable Energy, vol. 237, Art. no. 121602, 2024, doi: 10.1016/j.renene.2024.121602.
- [10] D. Diaz-Bedoya, "Solar irradiance forecasting with deep learning and ensemble models," IET Smart Grid, vol. 5, Art. no. 70019, 2025, doi: 10.1049/stg2.70019.
- [11] W. Thongwily, N. Junhuathon, and K. Chayakulkheeree, "Two-stage data-driven modeling for day-ahead solar power forecasting: A case study of the SUT campus floating solar system," in Proc. 13th Int. Electr. Eng. Congr. (IEEECON), Mar. 2025, pp. 1–6, doi: 10.1109/IEEECON64081.2025.10987684.
- [12] D. S. Lee and S. Y. Son, "PV forecasting model development and impact assessment via imputation of missing PV power data," IEEE Access, vol. 12, pp. 12843–12852, 2024, doi: 10.1109/ACCESS.2024.3352038.
- [13] S. S. Ali, M. Ali, D. M. Bhatti, and B. J. Choi, "Explainable clustered federated learning for solar energy forecasting," Energies, vol. 18, no. 9, Art. no. 2380, 2025, doi: 10.3390/en18092380.
- [14] Open-Meteo, "Historical weather API documentation," 2025. [Online]. Available: <https://open-meteo.com/en/docs/historical-weather-api> [Accessed: Aug. 21, 2025].
- [15] Exep, "Monitoring platform," 2025. [Online]. Available: <https://app-portal.exep.io/login>. [Accessed: Aug. 21, 2025].
- [16] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.
- [17] IBM, "Random forest: An overview of applications and how it works," 2025. [Online]. Available: <https://www.ibm.com/topics/random-forest>. [Accessed: Aug. 21, 2025].



Augmented Data-Driven Solar Power Forecasting Using Pre-Forecasting and Clear-Sky/KNN Data Refinements

Wipada Thongwily¹, Nitikorn Junhuathon^{2,*},
and Keerati Chayakulkeeree^{1,*}

ARTICLE INFO

Article history:

Received:

Revised:

Accepted:

Keywords:

Short-term PV forecasting

Clear Sky Model

KNN imputation

Explainable AI (XAI)

Deep Learning

ABSTRACT

This paper presents a day-ahead, two-stage solar power forecast for a floating solar power plant at Suranaree University of Technology. Data quality is managed using Clear Sky and KNN models, and interpretability is enhanced using Explainable AI (XAI). In the first step, LSTM and TFT are used to forecast light intensity and temperature. The results are then combined with Bi-LSTM, GRU, and RF to forecast power generation. The proposed two-stage forecasting framework significantly enhances floating photovoltaic (FPV) power prediction accuracy by integrating Clear Sky/KNN-based data refinement and explainable AI. Simulation results demonstrate that the proposed approach reduces RMSE by approximately 30–60% compared with conventional single-stage forecasting models, while achieving a WMAPE reduction exceeding 99% under missing and noisy data conditions. These improvements enable more reliable PV power forecasting, supporting practical applications in smart grid planning, operational scheduling, and renewable energy integration.

NOMENCLATURE

Abbreviations

AI: Artificial Intelligence

ANN: Artificial Neural Network

ARIMA: Auto Regressive Integrated Moving Average

Bi-LSTM: Bi-directional LSTM

CCPI: Climate Change Performance Index

CI: Confidence Interval

CNN: Convolutional Neural Network

EDA: Exploratory Data Analysis

FNN: Feedforward Neural Network

FPV: Floating Photovoltaic

GHG: Greenhouse Gas

GRN: Gated Residual Network

GRU: Gated Recurrent Unit

IQR: Interquartile range

KNN: K-Nearest Neighbors

LSTM: Long Short-Term Memory

LSTM&LSTM: Forecasting model using LSTM for both temperature and irradiance

LSTM&TFT: Forecasting model using LSTM for temperature and TFT for irradiance

MAE: Mean Absolute Error

MAPE: Mean Absolute Percentage Error

ME: Mean Error

Mean: Arithmetic mean

MGU: Minimal Gated Unit

Median: Median value

ML: Machine Learning

MLP: Multi-Layer Perceptron

Min: Minimum value

Max: Maximum value

PV: Photovoltaic

R²: Coefficient of Determination

RF: Random Forest

RESs: Renewable Energy Sources

RMSE: Root Mean Square Error

RNN: Recurrent Neural Network

SD: Standard deviation

SDM: Single Diode Model

SHAP: SHapley Additive explanations

SOM: Self-Organizing Map

SVR: Support Vector Regression

TCN: Temporal Convolutional Network

TFT: Temporal Fusion Transformer

VSN: Variable Selection Network

WMAPE: Weighted Mean Absolute Percentage Error

XAI: Explainable Artificial Intelligence

XGBoost: Extreme Gradient Boosting

Parameters

ClearSky with KNN: Missing value estimation method using Clear Sky model followed by KNN

DNI: Direct Normal Irradiance (W/m²)

GHI: Global Horizontal Irradiance (W/m²)

¹ School of Electrical Engineering, Institute of Engineering, Suranaree University of Technology, Nakhon Ratchasima, Thailand

² Department of Electrical Engineering, Faculty of Engineering, Rajamangala University of Technology Thanyaburi, Pathum Thani, Thailand

*Corresponding author: Email: keerati.ch@sut.ac.th, nitikorn.j@rmutt.ac.th

GTI: Global Tilted Irradiance (W/m^2)

KNN-all: Missing value estimation using KNN over the entire dataset

n: Number of samples

Pred. irr_LSTM: Predicted irradiance from the LSTM model

Pred. irr_TFT: Predicted irradiance from the Temporal Fusion Transformer model

$y_{actual,i}$: Actual power generation at sample i (kW)

$y_{forecast,i}$: Predicted power generation at sample i (kW)

1. INTRODUCTION

Currently, Thailand is ranked 24th in the CCPI [1], making the use of clean RESs an important role in meeting the increasing energy demand, along with reducing GHG emissions and reducing carbon emissions by reducing dependence on fossil fuels, which supports the Net Zero Emission target of the Thai government policy [2, 3]. In addition, the use of RESs also reduces fuel imports from abroad, promotes community participation in electricity generation, and reduces the cost of electricity production [4, 5].

However, electricity generation from renewable energy sources is subject to variability and uncertainty due to natural factors. Key influencing factors include solar irradiance, ambient temperature, module temperature, cloud cover, relative humidity, and overall weather conditions. These factors directly affect the amount of electricity produced at any given time. Such fluctuations not only impact the stability of the power system but also its reliability and efficiency, particularly in the context of smart grids. Therefore, accurate solar power generation forecasting is essential to manage variability and uncertainty, while maintaining system balance, enhancing stability, and supporting efficient energy management planning. For this reason, solar power forecasting is recognized as one of the key pillars of Thailand's Medium-Term Smart Grid Development Plan (2022-2031), which promotes the development of a modern, efficient, and environmentally friendly power system.

In [6], it was proposed that solar power generation forecasting studies can be divided into five categories: persistence models, physical models, statistical models, AI-based models, and hybrid models. In the past, persistence models, physical models, and statistical models were used to predict power generation. The persistence model is the simplest method, relying on historical data and assuming that the weather conditions do not change much. It is suitable for situations where the weather conditions are relatively stable. Meanwhile, the physical model is very important when the model training data is limited. By using the principles of physics and system parameters such as temperature, solar radiation intensity, and the tilt angle of the solar panels, it achieves high accuracy when the data is correct. Later, time series data were introduced, leading to the development of statistical models using techniques such as ARIMA and linear regression, which are suitable for time

series data. However, the aforementioned techniques still have limitations in terms of accuracy and cannot handle nonlinear data efficiently. Nowadays, with the advancement of AI and the availability of high-resolution weather data, AI-based models such as ANN, LSTM, and deep learning models have been developed, which have become very popular due to their ability to handle complex and non-linear data well and produce highly accurate forecasts. In addition, hybrid models These approaches combine the advantages of physical models and AI-based models to increase flexibility and reduce errors for single models. The popularity of these approaches reflects the research and application trends in the solar industry, which are likely to continue to develop with new techniques in the future.

Over the past decade, a large number of research projects have applied AI and ML techniques for forecasting solar power generation using time series data. Deep learning models, such as LSTM [13, 14], [16]–[21], Bi-LSTM [13, 23], and GRU [13], [19]–[23], can learn non-linear relationships from time-series data well. Meanwhile, RF [17, 25], although not a direct time-series model, can well handle non-linear feature variables, providing high accuracy for multi-feature data and tolerating outliers. Furthermore, compared to other models, in 2022 [14], TFT was applied to forecast hourly solar power generation. It was found to provide higher accuracy than ARIMA, LSTM, MLP, and XGBoost models due to its ability to capture temporal relationships and handle multiple variables simultaneously. It provides better accuracy and understandability than other models, and in the following years, TFT is still used for power generation forecasting [18]. When compared with TCN and MGU [24], the analysis results show that TFT still provides the most accurate results. From the above literature review, LSTM, Bi-LSTM, GRU, RF, and TFT models are chosen for power generation forecasting because they can support temporal sequence, variable relationship, and descriptive analysis to be suitable for this data set. Besides the application of AI and ML for solar power generation forecasting, it has also been used for temperature [7, 8] and solar intensity forecasting [9]–[11].

However, the accuracy of forecasting heavily depends on data quality. Missing or poor-quality data, often caused by sensor errors or extreme weather conditions, can directly impact model performance. Therefore, managing data quality prior to model training is crucial, including handling missing values and feature selection. Although previous studies have primarily focused on developing highly accurate models, few have systematically addressed the preprocessing of data before model training. A literature review indicates that various methods have been employed to handle missing data, such as the average value of a variable [8,11,14], linear interpolation [9, 22, 24], KNN [22], and Gain [22]. Moreover, recent research trends [12, 25] emphasize the interpretability of model outputs using XAI techniques, particularly SHAP, which allow the identification of the relative importance of input variables in predictions.

Table 1. A comprehensive literature review on PV power output forecasting

Ref.	Forecasting Category	Missing Data Refinement	Feature Selection	Model	Explainability (XAI) Techniques
[12]	PV Power	-	Feature Importance (Post-hoc XAI)	RF	SHAP, LIME, ELI5
[13]	PV Power	-	-	LSTM, Bi-LSTM, and GRU	-
[14]	PV Power	Average value of variable	Backward Elimination	TFT, ARIMA, LSTM, MLP, and XGBoost	-
[15]	PV Power	-	-	SDM and Regression equations	-
[16]	PV Power	-	SOM	FNN, LSTM, FNN-SOM, and LSTM-SOM	-
[17]	PV Power	-	Correlation Analysis	LSTM and RF	-
[18]	PV Power	✓	-	TFT, LSTM, and ARIMA	-
[19]	PV Power	-	Correlation Analysis	LSTM, GRU, and MLP	-
[20]	PV Power	-	Domain-Knowledge	LSTM and GRU	-
[21]	PV Power	-	-	LSTM and GRU	-
[22]	PV Power	KNN, Gain, and Linear Interpolation	-	CNN-GRU	-
[23]	PV Power	-	Correlation Analysis	Bi-LSTM, GRU and 1D-CNN	-
[24]	PV Power	Interpolation	-	TFT, TCN and MGU	-
[25]	PV Power	-	SHAP	XGBoost, RF, and SVR	SHAP
Proposed Framework	Temperature, Irradiance and PV Power	Clear sky model and KNN	Correlation Analysis, SHAP	LSTM, Bi-LSTM, GRU, RF, and TFT	SHAP

This enhances understanding of the model's decision-making process and increases confidence in its practical application. A summary of the literature review is presented in Table 1.

Moreover, most research on photovoltaic (PV) power generation forecasting focuses on improving forecast accuracy and model training efficiency, primarily using a single-stage forecasting approach that may overlook key factors influencing power generation, such as light intensity and temperature. Therefore, the motivation of this paper is to propose a two-stage forecasting approach that focuses on first forecasting key factors influencing power generation, namely light intensity and temperature, and then using the forecast results to forecast power generation in the next step. This approach enables the model to more clearly learn the relationship between weather variables and power generation and mitigates the impact of incomplete data,

which is a major limitation of commonly used single-stage forecasting methods.

The proposed method is evaluated using real data from the FPV system at Suranaree University of Technology (SUT). This approach not only enhances understanding of the relationship between external factors and PV power generation but also improves the overall forecasting accuracy.

In summary, the main contribution of this research are as follows:

- (i) Designing a solar power generation forecasting framework capable of handling incomplete data through the application of the Clear Sky Model and KNN.
- (ii) Develop and validate deep learning models for solar power forecasting based on a two-stage prediction approach.
- (iii) Evaluate and compare the accuracy of temperature and solar irradiance forecasting models used as inputs for power generation forecasting.

In addition, the key aspects of the proposed framework are summarized as follows:

- (i) the two-stage forecasting framework that separately forecasts environmental factors before PV power prediction,
- (ii) the integration of the Clear Sky model with KNN for robust missing data handling while preserving seasonal patterns, and
- (iii) the incorporation of XAI techniques to enhance model interpretability and provide physical insight into the influence of environmental variables on power generation.

The remainder of this paper is organized as follows: Section 2 describes the data preparation process; Section 3 discusses model development and evaluation; Section 4 presents the simulation results; and the final section provides the conclusion.

2. DATA PREPARATION

The dataset used in this study is from SUT FPV power plant located at 111 University Road, Suranaree Sub-district, Mueang Nakhon Ratchasima District, Nakhon Ratchasima Province, Thailand. The installed capacity of the floating PV power plant is 4.27 MW, and the power output data is recorded during the date range from 01/01/2023 to 12/31/2023. During this date range, the PV power generation is consistently zero between 17:00 and 07:00. Therefore, the power between 07:00 and 17:00 is considered in the simulation.

In particular, this research focuses on short-term PV power forecasting using 30-min resolution data for training and testing the deep learning model. The one-day PV power forecast is divided into 3 seasons: summer (mid-February to mid-May), rainy season (mid-May to mid-October), and winter (mid-October to mid-February). The input data for the deep learning model are PV power generation, solar irradiance, temperature, wind speed, humidity, and GTI.

The dataset consists of 7,665 samples, divided into 3 sets: 7,539 samples of the training set and the validation set, which are 80% and 20%, respectively, and 126 samples of the test set. The data management process consists of data collection, data cleaning, feature selection and correlation analysis, and data splitting. The details are as follows in Fig. 1.

2.1. Data Collection

The datasets used in this study are divided into two main groups: weather data and solar power generation data. The weather data includes variables such as temperature, relative humidity, wind speed, GHI, and GTI. These data were obtained from [26]. The solar power generation data can be accessed via [27]. The data sheet of 7,920 Jinko Solar JKM540M-72HL4-V solar panels is shown in Table 2

2.2. Data Refinement

Data refinement is an important step in data preparation for analysis and model development, focusing on detecting and managing missing values and outliers. In this study, the number of records of weather data and solar power generation data downloaded from online sources was not consistent across time periods, making it impossible to directly integrate the data. This results in missing solar power generation data, as shown in Fig. 2. Therefore, the data structure was adjusted to ensure that all inverters had the same number of rows, and two missing value estimation methods were used: the clear-sky model and KNN.

Figure 2 shows the missing power generation data for all three seasons of Thailand in 2023. The missing data range is 1,069 cells out of the total data set of 7,665 cells, which is 13.95% of the total missing data. Therefore, the data structure is adjusted to ensure that all inverters have the same number of rows, and two methods of estimating missing values are used: the clear sky and the KNN model.

To quantitatively evaluate the results of the two data addition methods, basic statistics (mean, median, SD, min, max, and IQR) of the added power generation values were compared, separated by data addition method. It was found that adding data with Clear Sky with KNN gave higher mean and median values than KNN-all, indicating that adding values according to the Clear Sky model helps to raise the power generation to reflect the actual sunlight conditions more. The SD, IQR, and Max of ClearSky with KNN were higher than KNN-all, indicating that the added values were more spread out and had more extreme values, which clearly reflected the seasonal pattern. Therefore, it can be said that adding data with ClearSky with KNN can preserve the seasonal pattern and the range of the actual power generation values better than adding data with KNN all year round, as shown in Table 3.

Table 2. The specifications of Jinko solar panels

Item	Value
Module Type	JKM540M-72HL4-V
Maximum Power (P_{max} , STC)	540 Wp
Number of Modules	7,920
Dimensions	2230 × 1134 × 35 mm (87.80 × 44.65 × 1.38 in)
Module Area	2.53 m ²
Total Installed Capacity (P_{total})	4,276.8 kWp



Fig. 1. The Framework for data management process

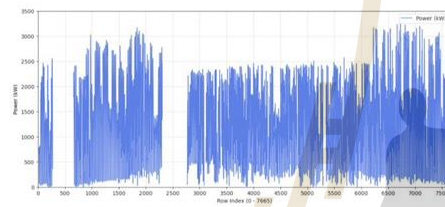


Fig. 2. PV power output of year 2023

In The choice of missing value replacement method directly impacts the accuracy of the forecast model, particularly for time series data such as solar power generation, which exhibits significant seasonal variations. Thailand has three main seasons: the summer season with the highest solar radiation; the rainy season with frequent cloudiness and rain, resulting in high solar radiation fluctuations; and the winter season, where production levels decrease despite clear weather for several days.

Table 3. Quantitative metrics

Metric	KNN-all	ClearSky with KNN
Mean	1279.71	1451.68
Median	1128.33	1356.49
SD	706.65	905.86
Min	0	0
Max	3242.40	4136.49
IQR	996.22	1304.81

To optimize the estimation for climate conditions, the clear-sky model, which is consistent with stable weather conditions, was used for the rainy season, while the KNN method was used to capture the similarity of meteorological

variables and reduce biases resulting from overestimation of clear-sky conditions. The data enriched by this method were merged into an annual dataset for use in the next model development step.

To illustrate the characteristics of the prepared dataset, Fig. 3 presents a three-dimensional surface plot of the solar power output from the photovoltaic system during the year 2023. The X-axis represents the time steps within each day, the Y-axis indicates the day of the year, and the Z-axis shows the power output values in kilowatts (kW), which serve as the target data for forecasting.

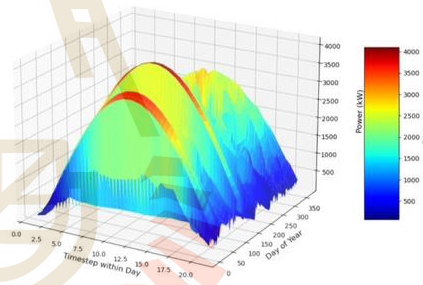


Fig. 3. PV power output of year 2023

From Fig. 3, the dome-shaped daily power output pattern can be observed, where the generation starts low in the morning, peaks around midday, and decreases in the evening, reflecting the sun's rising and setting cycle. During the middle of the year (approximately days 100–250), the surface plot displays yellow to red hues, indicating high power output consistent with the summer season, which experiences the highest solar irradiance. Conversely, at the beginning and end of the year (days < 100 and > 300), the surface is lower and shows blue to green colors, representing reduced power generation due to shorter daylight hours or increased cloud and rain coverage. Additionally, portions of

the surface plot appear rough or jagged, which may reflect the effects of variable weather conditions such as clouds, shadows, or rainfall on certain days. These insights derived from the visualization confirm the quality of the prepared data and demonstrate its readiness for use as a reliable foundation in the subsequent stages of forecasting model development and performance comparison.

2.3. Feature Selection and Correlation Analysis

After addressing missing and anomalous data using the clear-sky model and KNN, a statistical correlation analysis was conducted employing a correlation heatmap to identify factors influencing power output. The identified variables used as inputs and outputs include temperature, humidity, wind speed, wind direction, diffuse radiation, DNI, GHI, predicted temperature, predicted irradiance, and power (kW), as shown in Fig. 4.

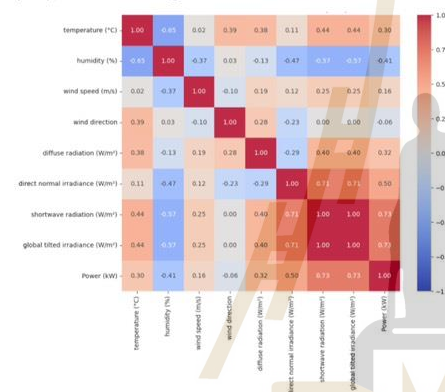


Fig. 4. The correlation matrix of all parameters

From the correlation heatmap analysis above, it can be observed that the correlation coefficient between temperature and power (kW) is 0.30, while that between irradiance and power (kW) is 0.73, which are important factors affecting power generation. Since the station data and the load data obtained from the satellite source are not identical, temperature and irradiance were forecasted using the LSTM and TFT models, and the forecast results were analyzed through the correlation heatmap. It was found that the correlation coefficients between temperature, irradiance, and power output increased. Consequently, the forecasted results were used as inputs for the next step of power generation forecasting.

The input data included predicted irradiance, GTI,

predicted temperature, humidity, and wind speed. This approach not only provides a deeper understanding of the relationships between external factors and power generation but also improves the overall forecast accuracy.

2.4. Data Splitting

This study focuses on day-ahead PV power forecasting using time-series data with a 30-minute resolution covering an entire year. After conducting EDA through correlation analysis, a total of 7,665 data samples were collected and divided into three subsets: 7,539 samples were allocated to the training and validation sets (split 80% and 20%, respectively), and 126 samples were reserved for the test set.

Following data preparation, which included merging data from multiple sources, formatting it appropriately, selecting time periods that exhibit seasonal patterns relevant to power generation, and handling missing values, the resulting dataset was complete and ready to be used for developing and training the forecasting models. To enhance the performance of the deep learning models and facilitate faster and more stable parameter optimization during training, all input features were normalized to a common scale. This normalization process plays a crucial role in accelerating model convergence and reducing issues caused by differing feature scales.

Such a systematic approach to data preparation is a critical foundation, as data quality directly impacts forecasting accuracy, especially in the context of time-series data for solar power generation systems.

3. MODEL DEVELOPMENT AND EVALUATION PROCESS

This section presents a two-step forecasting model, which is one of the research approaches employed to develop and evaluate short-term (day-ahead) PV power generation forecasting models. The research process is divided into three stages: data preparation, two-step forecasting, and evaluation. The overall research framework is illustrated in Fig. 5

The forecasting process in this study was designed as a two-step forecasting approach to enhance the accuracy and efficiency of PV power generation prediction. The steps are detailed as follows:

3.1. Environmental Factors Pre-Forecasting

The solar radiation and temperature forecasting process is performed using LSTM and TFT models. Although both models are widely recognized for time series forecasting, the selection for this research was based on their suitability for the data characteristics and limitations of the actual dataset.

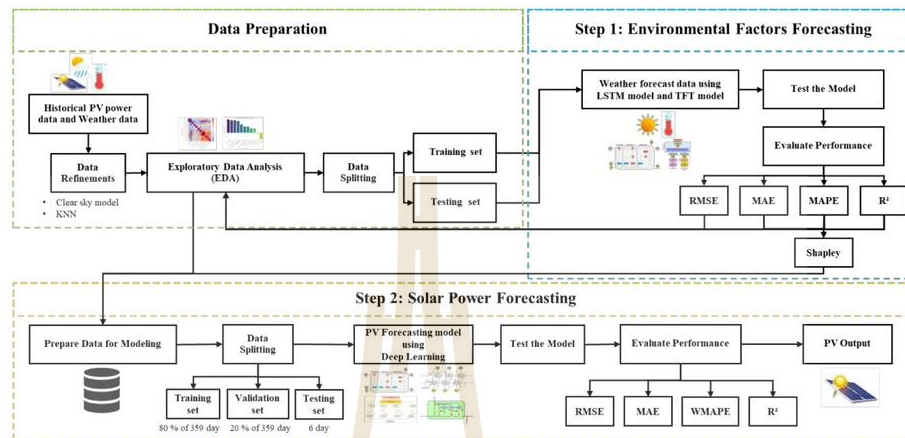


Fig. 5. The proposed framework for solar power forecasting

This is because the LSTM model has the advantage of handling long-term relationships in time series data and its stability in learning highly continuous sequences of data, which are commonly found in weather data, such as temperature or seasonal trends in radiation. This technique allows for stable learning of gradual trends.

While TFT models have a better ability to handle highly complex multivariate data, they also have the ability to automatically select important variables at each time point and learn both short-term and long-term relationships through attention mechanisms, making them suitable for highly volatile and multi-factorial weather data.

A novel addition to this research is the application of the TFT-lite architecture. This not only simplifies the architecture from a full TFT but is also designed to be suitable for the real-world dataset of floating solar PV systems, which are smaller and have higher volatility than land-based solar PV systems. While maintaining the ability to capture temporal patterns and fully explain the importance of variables, it also increases training speed. The architecture details and diagrams are provided in the appendix for clarity.

3.2. PV-Forecast

This second step use the forecasted solar irradiance and temperature from pre-forecasting stage, together with GTI, relative humidity, and wind speed, as inputs to predict power generation using LSTM, Bi-LSTM, GRU, RF, and TFT models. While these five models are standard models for power generation forecasting, their application in the context of floating solar data sets presents a unique role. Floating datasets exhibit higher intraday volatility than land-based solar, influenced by water surface humidity and

temperature, as well as unstable cloud patterns. This gives LSTM and Bi-LSTM advantages in learning long-term relationships in erratic data. GRU is suitable for data with high volatility and requires a shorter training time. RF examines nonlinear relationships and variable significance, helping to clarify the structure of weather relationships. TFT-lite enhances the ability to select important variables over time and handle the overlapping multivariate data, a characteristic of real floating data, as mentioned above.

Therefore, the novelty of this work lies not in the creation of new models but in the evaluation of multiple model architectures with different strengths on real floating data, enabling concrete insights into which models best respond to which data types in this specific context. The architecture details and diagrams are presented in the appendix.

For model structural parameterization for forecasting, an experimental approach was primarily used. Due to the unstable nature of the data and seasonal variations. For each model, a range of possible parameters was defined, such as the range of units in the hidden layer, the number of layers, and the range of learning rate values. The models were then run with multiple sets of different values and evaluated using the same metric to determine the one with the lowest forecast error. This approach enables the search for parameters that best fit the actual data behavior without relying on a single initial structure. It also reflects the model's learning characteristics and its response to seasonal data characteristics.

Regarding the selection of structural parameters, this study primarily adopts an experimental tuning approach due to the non-stationary and seasonal characteristics of the PV power data. For each model, the feasible parameter ranges are defined, and the obtained optimal hyperparameters are

presented in Tables 4 and 5, respectively.

For the configuration of structural parameters for temperature and irradiance forecasting, both models are detailed in Table 4. For PV forecasting, the structural parameters of the LSTM, Bi-LSTM, and GRU models were kept consistent during training, whereas the RF and TFT models used different configurations. The detailed hyperparameters for all PV-forecasting models are summarized in Table 5.

Table 4. Hyperparameters of environmental factors pre-forecasting models

Model	Hyperparameters range	Optimum hyperparameters
LSTM	Input: [1 5], Hidden nodes: [32 128]x3 Epochs: [100 200], Batch Size: [16 32], Learning Rate: [0.0001 0.001], Early Stopping: [10 20]	Input: 3, Hidden nodes: 128,64,32, Epochs: 200, Batch Size: 32, Learning Rate: 0.0001, Early Stopping: 10
TFT	Input: [1 5], Hidden nodes: [32 128]x3 Epochs: [100 200], Batch Size: [16 32], Learning Rate: [0.0001 0.001], Early Stopping: [10 20]	Input: 3, Hidden nodes: 64, 64, 32, Epochs: 200, Batch Size: 32, Learning Rate: 0.0001, Early Stopping: 10

In the final step, metrics are selected to evaluate the model performance in forecasting solar power generation capacity to measure the accuracy and reliability of the predictions, enabling efficient use of the results. These metrics can be accessed using the Sklearn library in Python. In this study, four fitness functions are used: RMSE, MAE, WMAPE, and R^2 . The first comparative index considered is RMSE, which can be calculated by averaging the squares of the errors and taking the square root as shown in Eq. (1).

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_{actual,i} - y_{forecast,i})^2}{n}} \quad (1)$$

The next index used is MAE, which represents the average error between the actual and forecast values, as shown in Eq. (2).

$$MAE = \frac{\sum_{i=1}^n |y_{actual,i} - y_{forecast,i}|}{n} \quad (2)$$

Eq. (3) shows WMAPE as a measure of the accuracy of the forecasting model that adjusts the weights based on the actual values of the target variable, with the aim of reducing sensitivity to low or zero values, which is a limitation of MAPE.

$$WMAPE = \frac{\sum_{i=1}^n |y_{actual,i} - y_{forecast,i}|}{\sum_{i=1}^n y_{actual,i}} \times 100\% \quad (3)$$

Table 5. Hyperparameters of PV-forecasting models

Model	Hyperparameters range	Optimum hyperparameters
LSTM	Input: [1 8], Hidden nodes: [32 128]x3 Epochs: [100 200], Batch Size: [16 32], Learning Rate: [0.0001 0.001], Early Stopping: [10 20]	Input: 5, Hidden nodes: 64,32,50, Epochs: 100, Batch Size: 32, Learning Rate: 0.001, Early Stopping: 20
Bi-LSTM	Input: [1 8], Hidden nodes: [32 128]x3 Epochs: [100 200], Batch Size: [16 32], Learning Rate: [0.0001 0.001], Early Stopping: [10 20]	Input: 5, Hidden nodes: 64,32,50, Epochs: 100, Batch Size: 32, Learning Rate: 0.001, Early Stopping: 20
GRU	Input: [1 8], Hidden nodes: [32 128]x3 Epochs: [100 200], Batch Size: [16 32], Learning Rate: [0.0001 0.001], Early Stopping: [10 20]	Input: 5, Hidden nodes: 64,32,50, Epochs: 100, Batch Size: 32, Learning Rate: 0.001, Early Stopping: 20
RF	Input: [1 8], Number of Trees: [50 100] Max Depth: [10 20]	Input: 5, Number of Trees: 100, Max Depth: 10
TFT	Input: [1 8], Hidden nodes: [32 128]x4 Epochs: [100 200], Batch Size: [16 32], Learning Rate: [0.0001 0.001], Early Stopping: [10 20]	Input: 5, Hidden nodes: 128,128,64,128, Epochs: 100, Batch Size: 16, Learning Rate: 0.0001, Early Stopping: 30

The last index is R^2 , which measures the model's ability to explain the variance of the target variable, as shown in Eq. (4).

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_{actual,i} - y_{forecast,i})^2}{\sum_{i=1}^n (y_{actual,i} - \bar{y}_{actual})^2} \quad (4)$$

In addition, the model performance was further evaluated by using the ME, SD, and 95% CI. Using these metrics, we can reflect both the “level of model error” through the mean error and the “stability of the forecast results” through the spread of the SD and the width of the CI interval, as shown in Equations 5-7, respectively.

$$ME = \frac{1}{n} \sum_{i=1}^n y_{forecast,i} - y_{actual,i} \quad (5)$$

$$SD = \sqrt{\frac{1}{n-1} \sum_{i=1}^n ((y_{forecast,i} - y_{actual,i}) - ME)^2} \quad (6)$$

$$CI_{95\%} = ME \pm 1.96 \times \frac{SD}{\sqrt{n}} \quad (7)$$

4. RESULTS AND DISCUSSION

This research investigated a two-stage forecasting approach for solar power generation. The experimental results were presented in three main parts: forecasting results of environmental, forecasting results of PV power generation, and discussion. The details are as follows.

4.1 Forecasting Results of Environmental Factors

Based on the forecasting results of solar irradiance and temperature using the LSTM and TFT models, the monthly MAE of the LSTM model is presented in Fig. 6. The figure illustrates two separate results, corresponding to solar irradiance and temperature, respectively.

The monthly error analysis reveals clear differences between irradiance and temperature forecasting. For irradiance, the MAE remains relatively low and stable throughout most of the year; however, in December, both the average error and its variability increase substantially. This indicates that the model fails to adequately capture seasonal fluctuations or irregular atmospheric conditions.

In contrast, the MAE in May is close to zero, which is attributable to missing or incomplete data rather than true predictive accuracy. For temperature forecasting, the errors are generally more consistent, with most months showing MAE values between 2 and 5°C, which is more stable than irradiance forecasting.

Nevertheless, in June and September the errors rise to around 6–7°C, reflecting sudden shifts in temperature during

those months—conditions that the model is not yet able to handle effectively.

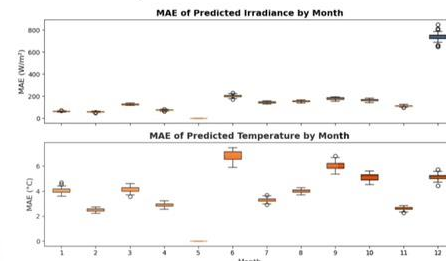


Fig. 6. MAE of environmental factors forecasting.

After forecasting solar irradiance and temperature using the LSTM and TFT models, the prediction results were analyzed through a correlation heatmap to identify factors influencing power generation. The analysis indicated that both models produced similar forecasts, with correlations to power output of 0.63 and 0.62 for temperature and 0.56 and 0.50 for solar irradiance, respectively.

Nevertheless, correlation analysis alone may not fully capture the direct influence of forecasted variables on power generation due to the nonlinear nature of real-world phenomena. To address this, SHAP analysis was conducted alongside the correlation heatmap to evaluate feature importance based on SHAP values using a nonlinear reference model, as illustrated in Fig. 7 and Fig. 8.

Analysis of the correlation heatmaps shows that the predicted irradiance from both the LSTM and TFT models is highly correlated with power, with values of 0.89 and 0.87, respectively, while the predicted temperature has a moderate correlation of approximately 0.69 and 0.72, respectively, which is consistent with the physical mechanism of the solar cell system that directly depends on the light intensity. Meanwhile, temperature still has a secondary effect on the power generation, which is consistent with the behavior of decreasing the efficiency of the panel when the temperature increases.

When compared to the correlation heatmap between the forecast data and power generation directly, the values from the SHAP analysis combined with the correlation heatmap are higher, indicating that these variables have a greater influence on power generation capacity than correlation alone can reflect, which helps to understand the impact of the variables in detail and covers both linear and non-linear aspects.

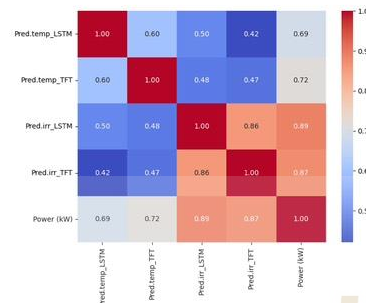


Fig. 7. The parameters correlation matrix after SHAP process

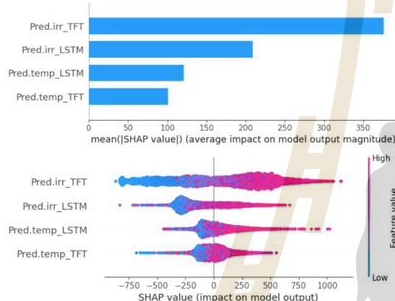


Fig. 8. SHAP values impact on model output

Considering the absolute mean of SHAP, the predicted irradiance from the TFT model is the highest, followed by the predicted irradiance from the LSTM model. The mean SHAP temperature of the LSTM model is slightly higher than the TFT model. This result shows that the irradiance-related input has the highest influence in a reasonable manner consistent with the physics of the system and also reflects that the values obtained from TFT have higher quality in the angle where the model is actually used, not just a linear statistical relationship.

Furthermore, the SHAP summary plot results indicate that high light intensity (red) consistently induces positive SHAP, leading to increased power generation with increasing light intensity. Meanwhile, low light intensity (blue) induces negative SHAP, reflecting the direct physical behavior of solar cells. Similar to temperature, which has both positive and negative effects depending on the temperature, this suggests that high temperatures may reduce panel efficiency, but moderate temperatures can support power generation. The distribution of SHAP effects

is broader, though not as pronounced as solar radiation. Furthermore, the SHAP distribution of Pred. irr_TFT is more consistent and has a higher influence on the model than that of Pred. irr_LSTM. This suggests that using TFT-predicted irradiance values can accurately and continuously capture solar radiation patterns, enabling power generation forecasting models to better understand causal relationships, not just statistical ones. Using SHAP summary plots not only demonstrates numerical significance but also provides a systematic and detailed understanding of the mechanisms and physical effects of input variables on power generation.

In summary, the analysis confirms that solar irradiance is the primary factor driving the fluctuations in power generation from solar panels. Although temperature plays a secondary role, it remains an important variable for model performance. These results can support feature selection to enhance both the efficiency and interpretability of the forecasting models. Furthermore, the comparison between the correlation heatmap using predicted irradiance and predicted temperature from the LSTM model as inputs and the SHAP analysis combined with the correlation heatmap using predictions from both the LSTM and TFT models as inputs indicates that both input sets are suitable for power generation forecasting. Consequently, both datasets were utilized as inputs for subsequent forecasting, and their results will be compared in the following step.

4.2 Forecasting Results of PV power generation

Since this study focuses on one-day-ahead PV power forecasting, the results are presented separately for three seasons: summer, rainy, and winter. A comparison was conducted between the actual power output and the forecasted power output of the five models to clearly illustrate the differences and to evaluate model accuracy using RMSE, MAE, WMAPE, and R^2 .

Figures 9 and 10 show the comparison between actual and forecasted power outputs for the five models, based on datasets containing predicted irradiance and temperature from the LSTM&TFT and LSTM&LSTM models, respectively. The results are presented for the three seasons: summer (a), rainy (b), and winter (c).

Based on the validation metrics obtained from comparing the predicted and actual power outputs, the results of both test sets are shown in Table 6 and Table 7, respectively, showing the RMSE, MAE, WMAPE, and R^2 values. From the analysis of the results in Table 6, the RF method is 67.61 kW, 52.49 kW, 3.10%, and 0.99, respectively, which is the best forecasting method among the five methods for summer considering the four validation factors. In contrast, in the LSTM&LSTM dataset from Table 7, the Bi-LSTM method

is the best forecasting method among the five methods, with the values of 59.13 kW, 49.79 kW, 2.94%, and 0.99, respectively. The same is true in the rainy season, with the values of 95.55 kW, 78.48 kW, 4.72%, and 0.98, respectively. However, for the dataset in Table 6, the RF forecasting results are the best in the rainy season, which are 123.01 kW, 106.02 kW, 6.38%, and 0.96, respectively. In contrast, for the first winter dataset, the best prediction method was LSTM with values of 110.94 kW, 81.53 kW, 5.50%, and 0.98, respectively, and for the next dataset, the GRU method had relatively accurate values of 114.99 kW, 80.61 kW, 5.44%, and 0.98, respectively.

In addition, the evaluation results using mean error, SD, and CI showed that model performance differed significantly across seasons, reflecting both model accuracy and stability.

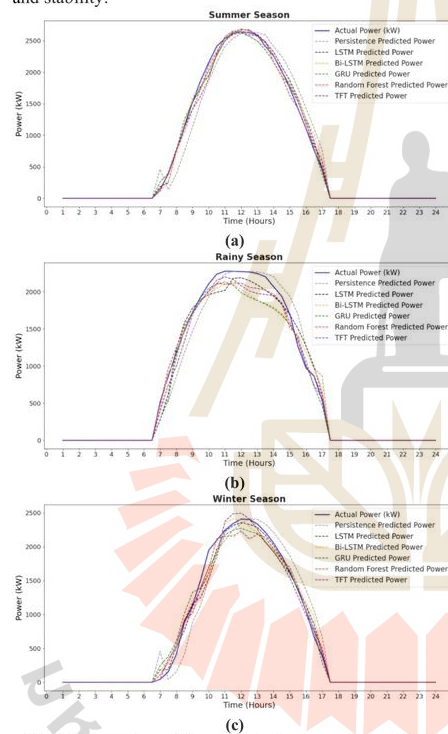


Fig. 9. Comparison of day-ahead solar power forecasting of the LSTM& TFT models: (a) summer season (b) rainy season and (c) winter season

Considering the width of the CI range as statistical evidence, these results can be summarized in Tables 8 and 9. The analysis of the results showed that in summer, the

first dataset showed the most stable RF with narrow SD and CI values, while the second dataset, Bi-LSTM, had the lowest error and the narrowest CI range among all models. This is consistent with the fact that summer has a relatively uniform solar radiation pattern, allowing the Bi-LSTM temporal model that learns both past and future directions to perform best. In the rainy season, both datasets exhibited the widest CI, reflecting the fluctuations in cloud and rainfall that result in high data uncertainty. The RF model showed good stability on the first dataset due to its tree-based structure that copes well with nonlinearity from sudden changes. However, as data quality improved on the second dataset, Bi-LSTM emerged as the most stable model, indicating that its bidirectional temporal learning capability significantly mitigates the impact of intra-rainy season variability. While in winter, SD and CI values are moderate compared to other seasons, indicating that the data fluctuate less but are not as consistent as in summer.

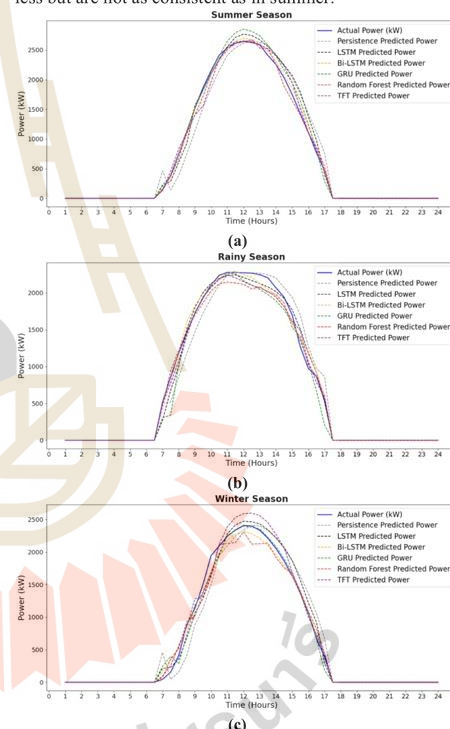


Fig. 10. Comparison of day-ahead solar power forecasting of the LSTM& LSTM models: (a) summer season (b) rainy season and (c) winter season

Table 6. Forecasting results with dataset 1 from LSTM&TFT pre-forecasting models

Model	Summer season			
	RMSE (kW)	MAE (kW)	WMAPE (%)	R ²
LSTM*	191.33	172.66	10.19	0.95
Persistence**	259.32	230	13.70	0.90
LSTM**	97.27	75.53	4.46	0.99
Bi-LSTM**	104.25	78.66	4.64	0.98
GRU**	82.69	69.22	4.09	0.99
RF**	67.61	52.49	3.10	0.99
TFT**	82.46	63.47	3.75	0.99
Model	Rainy season			
	RMSE (kW)	MAE (kW)	WMAPE (%)	R ²
LSTM*	230.67	205.99	12.39	0.86
Persistence**	374.85	260.04	16.21	0.63
LSTM**	164.91	143.82	8.65	0.93
Bi-LSTM**	196.72	169.94	10.22	0.90
GRU**	208.77	174.59	10.50	0.88
RF**	123.01	106.02	6.38	0.96
TFT**	180.13	154.99	9.32	0.91
Model	Winter season			
	RMSE (kW)	MAE (kW)	WMAPE (%)	R ²
LSTM*	224.83	206.08	13.90	0.92
Persistence**	254.5842	214.44	14.82	0.88
LSTM**	110.94	81.53	5.50	0.98
Bi-LSTM**	152.21	127.26	8.58	0.96
GRU**	147.21	134.27	9.06	0.96
RF**	130.98	106.02	7.15	0.97
TFT**	146.39	129.51	8.73	0.97

Note: * Without pre-forecasting, **With pre-forecasting

The first dataset shows that several models have similar performance, with the LSTM model having the lowest mean error and narrower CI than the other models. Meanwhile, in the second dataset, the GRU model has a lower mean error and narrower CI than the other models, reflecting that the gated recurrent model can handle data patterns with mixed seasonality and moderate variation more efficiently. The results for each season can be summarized as in Table 10.

However, the processing time ranking, as shown in Table 11, indicates that for the first dataset, the models with the shortest computation times are RF, GRU, LSTM, Bi-LSTM, and TFT, with times of 6.27 s, 113.73 s, 134.07 s, 182.18 s, and 602.94 s, respectively. For the second dataset,

the models are ranked according to increasing processing time as RF, GRU, Bi-LSTM, LSTM, and TFT, with times of 5.94 s, 80.17 s, 100.68 s, 163.67 s, and 620.59 s, respectively. These results are consistent with the theoretical expectations described above.

The simulation results indicate that each model has different strengths depending on the data characteristics at each time period. RF and Bi-LSTM perform well for summer power generation forecasting because during summer, weather changes lead to consistent power generation capacity. This results in continuous and less fluctuating data for model training. The model can efficiently learn the relationships between these features.

Table 7. Forecasting results with dataset 2 from LSTM&LSTM pre-forecasting models

Model	Summer season			
	RMSE (kW)	MAE (kW)	WMAPE (%)	R ²
LSTM*	191.33	172.66	10.19	0.95
Persistence**	259.32	230	13.70	0.90
LSTM**	167.35	139.01	8.21	0.96
Bi-LSTM**	59.13	49.79	2.94	0.99
GRU**	132.27	111.28	6.57	0.97
RF**	73.95	56.99	3.36	0.99
TFT**	116.02	81.89	4.83	0.98
Model	Rainy season			
	RMSE (kW)	MAE (kW)	WMAPE (%)	R ²
LSTM*	230.67	205.99	12.39	0.86
Persistence**	374.85	260.04	16.21	0.63
LSTM**	119.43	95.97	5.77	0.96
Bi-LSTM**	95.55	78.48	4.72	0.98
GRU**	183.19	138.67	8.34	0.91
RF**	114.82	98.02	5.90	0.96
TFT**	106.90	87.27	5.25	0.97
Model	Winter season			
	RMSE (kW)	MAE (kW)	WMAPE (%)	R ²
LSTM*	224.83	206.08	13.90	0.92
Persistence**	254.5842	214.44	14.82	0.88
LSTM**	163.07	144.89	9.77	0.96
Bi-LSTM**	114.42	94.08	6.34	0.98
GRU**	114.99	80.61	5.44	0.98
RF**	139.10	116.60	7.86	0.97
TFT**	160.38	135.48	9.14	0.96

Note: * Without pre-forecasting, **With pre-forecasting

Table 8. Accuracy and stability of LSTM&TFT pre-forecasting models with dataset 2 under difference seasons

Model	Summer season				
	ME (kW)	SD (kW)	CI Lower (kW)	CI Upper (kW)	CI Width (kW)
LSTM	33.75	56.18	17.69	49.81	32.12
Bi-LSTM	35.15	60.82	17.76	52.53	34.78
GRU	30.93	46.30	17.69	44.17	26.47
RF	23.45	39.05	12.29	34.62	22.33
TFT	31.78	58.93	14.93	48.62	33.69
Model	Rainy season				
	ME (kW)	SD (kW)	CI Lower (kW)	CI Upper (kW)	CI Width (kW)
LSTM	64.26	90.53	38.38	90.14	51.77
Bi-LSTM	75.93	108.52	44.91	106.96	62.05
GRU	78.01	116.96	44.57	111.45	66.88
RF	47.37	67.93	27.95	66.79	38.84
TFT	53.87	76.44	32.01	75.72	43.71
Model	Winter season				
	ME (kW)	SD (kW)	CI Lower (kW)	CI Upper (kW)	CI Width (kW)
LSTM	36.43	65.29	17.76	55.09	37.33
Bi-LSTM	56.86	85.28	32.48	81.24	48.76
GRU	59.99	78.84	37.45	82.53	45.08
RF	47.37	74.43	26.09	68.65	42.56
TFT	55.50	86.63	30.73	80.27	49.53

Similarly, during the rainy season, RF and Bi-LSTM perform best. Due to the uncertain data characteristics and rapid power generation capacity changes, RF can effectively handle the uncertainty and capture non-hierarchical patterns. Bi-LSTM, on the other hand, utilizes historical and forward data, enabling it to learn temporal correlation patterns and reduce errors. In contrast, for winter, the most suitable models are LSTM and GRU. This is because power generation is influenced by many factors, such as shorter daylight hours, low temperature, and high humidity, leading to complex and rapidly changing data characteristics. LSTM and GRU models, which are hierarchical neural networks, can efficiently capture the hierarchy of data and learn the relationships between multiple variables. GRU, with its simpler structure, is therefore faster to train.

When comparing the error values between the two data sets, the pre-forecast data set 2 (LSTM and LSTM) provides more accurate forecasting results.

Table 9. Accuracy and stability of LSTM&LSTM pre-forecasting models with dataset 2 under difference seasons

Model	Summer season				
	ME (kW)	SD (kW)	CI Lower (kW)	CI Upper (kW)	CI Width (kW)
LSTM	62.11	94.06	35.22	89.00	53.78
Bi-LSTM	22.25	33.02	12.81	31.69	18.88
GRU	49.72	73.90	28.59	70.85	42.25
RF	25.46	42.82	13.22	37.71	24.49
TFT	36.59	69.12	16.83	56.35	39.52
Model	Rainy season				
	ME (kW)	SD (kW)	CI Lower (kW)	CI Upper (kW)	CI Width (kW)
LSTM	42.88	68.06	23.42	62.34	38.92
Bi-LSTM	35.07	53.96	19.64	50.49	30.85
GRU	61.96	106.76	31.44	92.48	61.05
RF	43.80	63.71	25.58	62.01	36.43
TFT	38.99	60.52	21.69	56.30	34.61
Model	Winter season				
	ME (kW)	SD (kW)	CI Lower (kW)	CI Upper (kW)	CI Width (kW)
LSTM	64.74	88.64	39.40	90.08	50.69
Bi-LSTM	42.03	64.58	23.57	60.50	36.93
GRU	36.02	68.63	16.40	55.64	39.24
RF	52.10	77.84	29.84	74.35	44.51
TFT	60.53	89.44	34.96	86.10	51.14

Table 10. Best forecasting models of year 2023 data

Models	Best Model	
	Pre-forecast data 1 (LSTM&TFT)	Pre-forecast data 2 (LSTM&LSTM)
Summer	RF	Bi-LSTM
Rainy	RF	Bi-LSTM
Winter	LSTM	GRU

Table 11. Runtime of PV forecasting models

Model	Time (s)	
	Pre-forecast data 1 (LSTM&TFT)	Pre-forecast data 2 (LSTM&LSTM)
LSTM	134.07	163.57
Bi-LSTM	182.18	100.68
GRU	113.73	80.17
RF	6.27	5.94
TFT	602.94	620.59

This observation clearly indicates that the quality of irradiance and temperature data forecasted in the first step directly affects the accuracy of the power generation forecast, both in terms of error level and consistency of the forecast over time. Furthermore, the causes of power generation forecast errors may arise from missing data for some of the forecasts and environmental factors that may directly affect the efficiency of the solar panels.

4.3 Discussion

Overall, the analysis results reflect the fundamental characteristic of FPV systems: power generation is still primarily dependent on solar radiation intensity. While temperature plays a role in panel efficiency, it is not the most significant predictor of fluctuations. This trend is evident in both the correlation heatmap and the SHAP absolute average, which are consistent, even though the two approaches approach the data from different perspectives. A heatmap describes a linear relationship, while SHAP reflects the non-linear influence of inputs on model decisions, further validating the link between weather conditions and power generation. Considering both approaches together, there is a strong consensus that irradiance significantly influences power generation. Even when the correlation is lower than SHAP, this suggests that the relationship may be non-linear, such as short-term extreme highs or lows caused by cloud movement. This is consistent with the actual behavior of FPV.

Although the SHAP analysis in the first step indicated that the irradiance forecast from TFT exhibits a signal quality that is consistent with physical laws and more consistent than LSTM, experimental results show that using LSTM irradiance forecasts results in more accurate power generation forecasts in the second step, across multiple time periods and seasons. This point reflects the specific nature of two-stage forecasting systems, where the outcome depends not only on the accuracy of the first-stage model but also on the error structure and the compatibility between the errors of the two models.

Considering the seasonal results, the model accuracy clearly reflects the data characteristics. Summer has a constant irradiance, which makes RF and Bi-LSTM perform well due to the relatively uniform weather conditions, making it easy for the model to learn the relationship between variables. Meanwhile, the rainy season has high volatility and uncertainty. Therefore, RF, with its flexible structure, still has advantages in handling uncertainty and non-order patterns. Bi-LSTM can learn the temporal order and the backward-forward pattern, thus reducing the error better than other models. Winter is the most complex case because it is affected by many factors simultaneously.

Therefore, sequential models such as LSTM and GRU can well capture the temporal order and the relationship between multiple variables. GRU trains faster and is suitable for small data sets. From the results, it can be technically interpreted that RNN models, including LSTM, Bi-LSTM, and GRU, are naturally suitable for data with a temporal structure. Bi-LSTM, in particular, can use information in both backward and forward directions, which supports volatile data and reduces the error better than other models. Meanwhile, RF, which does not rely on a temporal structure, can handle highly volatile or irregular data well, thus performing highly well in both summer and rainy seasons, although it cannot learn the temporal association like RNN-based models. TFT, although highly accurate, comes at the cost of time and computing resources. This makes it suitable for applications requiring high accuracy and sufficient resources.

A comparison of two pre-forecast datasets confirms that the quality of the first-stage irradiance and temperature forecasts directly affects the second-stage power generation, both in terms of error magnitude and signal continuity. Errors often arise from missing data or meteorological errors, particularly the May data outage, which resulted in an unusually low MAE. This suggests that model estimates may be distorted by data quality rather than the actual model capability if preprocessing is not carefully examined.

In summary, the results of this discussion clearly indicate that power generation forecasting cannot be solely based on model accuracy. It must also be considered systematically, encompassing the quality of the forecast inputs, internal model mechanisms, seasonal variations, and the spatial and environmental constraints of the FPV system. However, since all data are from a single site, this approach may require parameter adjustments or additional variable selection when applied to other areas with varying climates, cloud patterns, and radiation angles. These findings contribute to a more comprehensive physical and data-driven understanding of the system, which is beneficial for smart power system planning, such as load management, real-time power mix control, and backup power management in dynamic real-world environments.

5. Conclusion

This paper presents a comparative study of short-term PV forecasting methods for the next 30 minutes, using LSTM, Bi-LSTM, GRU, RF, and TFT, with missing data handling using the Clear Sky Model and KNN in a two-stage forecasting model. The experimental results indicate that the quality of the first-stage forecast affects the accuracy of the second-stage forecast, but accuracy alone is not a true indicator of performance. The error structure and model

compatibility also play important roles. Moreover, the simulation results suggest that RF and Bi-LSTM should be preferred in the summer and rainy seasons, as RF can handle high fluctuations well, while Bi-LSTM relies on both historical and future data to learn temporal relationships and effectively reduce errors. Meanwhile, in winter, the complex and rapidly changing data types are suitable for LSTM and GRU models. GRU has the advantages of a simple structure and shorter training time. In terms of model processing, the order of time is the shortest: RF, GRU, LSTM, Bi-LSTM, and TFT, respectively. Furthermore, the proper handling of some missing data is crucial for model performance. If not handled properly, it may result in decreased forecast accuracy.

Overall, the proposed two-stage forecasting framework significantly enhances floating PV (FPV) power prediction performance, achieving RMSE reductions of approximately 30–60% and WMAPE improvements of over 99% compared with baseline single-stage methods. The integration of Clear Sky/KNN-based data refinement with explainable AI (XAI) analysis not only improves robustness under incomplete data conditions but also provides transparent insights into the influence of environmental variables on PV power generation. These features make the proposed framework highly suitable for practical smart grid applications, including PV integration planning, operational scheduling, and energy management system design.

The results of this work can support energy planners, power operators, and power plant managers in decision-making, real-time energy monitoring, and efficient load management. For future work, many areas can be developed, such as improving model accuracy, forecasting systems with multiple devices such as energy storage, and applying Digital Twin to support smart power system planning.

In addition, the proposed framework can be extended to multi-site validation by applying it to PV systems located in different climatic regions, including both floating and land-based installations. Furthermore, transfer learning techniques will be explored to adapt the pre-trained forecasting models to new sites with limited historical data by fine-tuning selected network layers. This approach is expected to enhance scalability, reduce retraining costs, and improve the practical deployment of the proposed method in real-world power system operations.

ACKNOWLEDGEMENTS

This research was supported by the Suranaree University of Technology.

REFERENCES

- [1] Germanwatch; NewClimate Institute; CAN International. 2025. Thailand – Climate Performance Ranking. Climate Change Performance Index (CCPI). Available online: <https://ccpi.org/country/tha/>; accessed August 21, 2025.
- [2] NXPO. 2025. Thailand's Development and Use of Clean Energy for a Sustainable Future. Office of National Higher Education Science Research and Innovation Policy Council (NXPO). Available online: <https://www.nxpo.or.th/th/9651/>; accessed August 21, 2025.
- [3] BOI. 2025. Promoting Investment in Renewable Energy from Small-Scale Sources for a Sustainable Future. Board of Investment of Thailand (BOI). Available online: <https://ic.or.th/th/about-us/boi-news/1275/>; accessed August 21, 2025.
- [4] Rafique, M.M.; Bahaidarah, H.M.; and Anwar, M.K. 2019. Enabling private sector investment in off-grid electrification for cleaner production: Optimum designing and achievable rate of unit electricity. *Journal of Cleaner Production* 206: 508–523. <https://doi.org/10.1016/j.jclepro.2018.09.123>
- [5] Benti, N.E.; Mekonnen, Y.S.; and Asfaw, A.A. 2023. Combining green energy technologies to electrify rural community of Wollega, Western Ethiopia. *Scientific African* 19: e01467. <https://doi.org/10.1016/j.sciaf.2022.e01467>
- [6] Nguyen, T.T.; Vu, X.S.H.; Do, D.H.; Takano, H.; and Nguyen, D.T. 2023. Short-term PV power forecast using hybrid deep learning model and Variational Mode Decomposition. *Energy Reports* 9: 712–717. <https://doi.org/10.1016/j.egy.2023.05.154>
- [7] Li, P.; et al. 2024. Advancing photovoltaic panel temperature forecasting: A comparative study of numerical simulation and machine learning in two types of PV power plant. *Renewable Energy* 237: 121602. <https://doi.org/10.1016/j.renene.2024.121602>
- [8] Keddouda, A.; et al. 2024. Photovoltaic module temperature prediction using various machine learning algorithms: Performance evaluation. *Applied Energy* 363: 123064. <https://doi.org/10.1016/j.apenergy.2024.123064>
- [9] Diaz-Bedoya, D. 2025. Solar irradiance forecasting with deep learning and ensemble models. *IET Smart Grid* 5: 70019. <https://doi.org/10.1049/stg2.70019>
- [10] Hamdaoui, J.H.; Benghoulam, E.M.; Chaibi, M.; Berrada, M.; and Hmadi, A.E. 2025. Estimating daily global solar radiation using deep learning. *Results in Engineering* 27: 106132. <https://doi.org/10.1016/j.rineng.2025.106132>
- [11] Alorf, A.; and Khan, M.U.G. 2025. Solar irradiance forecasting using temporal fusion transformers. *International Journal of Energy Research* 2025: Article 3534500. <https://doi.org/10.1155/er/3534500>
- [12] Kuzlu, M.; Cali, U.; Sharma, V.; and Ö., G. 2020. Gaining insight into solar photovoltaic power generation forecasting utilizing explainable artificial intelligence tools. *IEEE Access* 8: 187814–187823. <https://doi.org/10.1109/ACCESS.2020.3031477>
- [13] Sharadga, H.; Hajimirza, S.; and Balog, R.S. 2020. Time series forecasting of solar power generation for large-scale photovoltaic plants. *Renewable Energy* 150: 797–807. <https://doi.org/10.1016/j.renene.2019.12.131>
- [14] López Santos, M.; García-Santiago, X.; Echevarría Camarero, F.; Blázquez Gil, G.; and Carrasco Ortega, P. 2022. Application of Temporal Fusion Transformer for Day-Ahead PV Power Forecasting. *Energies* 15(14): 5232. <https://doi.org/10.3390/en15145232>
- [15] Zohal, B.; Amiry, H.; Yadir, S.; Baghaz, E.; Bendaoud, R.; El-Abidi, A.; Chanaa, F.; and Benhmda, M. 2023. Prediction of the performance of photovoltaic modules from the behavior of physical parameters according to irradiance and

- temperature. *International Journal of Intelligent Engineering and Systems* 16(1): 502–517.
<https://doi.org/10.22266/ijies2023.0228.44>
- [16] Junhuathon, N.; and Chayakulkeeree, K. 2022. Deep-learning-based short-term photovoltaic power generation forecasting using improved self-organization map neural network. *Journal of Renewable and Sustainable Energy* 14: 043702. <https://doi.org/10.1063/5.0091454>
- [17] Roy, N.; et al. 2022. Day-ahead solar power generation forecasting using LSTM and random forest methods for North Eastern Region of India. *Proceedings of the 22nd National Power Systems Conference (NPSC)*, Gandhinagar, India, Dec. 2022: 854–859.
<https://doi.org/10.1109/NPSC57038.2022.10069833>
- [18] Islam, M.; Shuvo, S.S.; Shohan, J.A.; and Faruque, O. 2023. Forecasting of PV Plant Output Using Interpretable Temporal Fusion Transformer Model. *Proceedings of the North American Power Symposium (NAPS)*, Denver, USA, Oct. 15–17, 2023: 1–6.
<https://doi.org/10.1109/NAPS58826.2023.10318698>
- [19] Boucetta, L.N.; Amrane, Y.; and Arezki, S. 2023. Comparative analysis of LSTM, GRU, and MLP neural networks for short-term solar power forecasting. *Proceedings of the International Conference on Electrical Engineering and Advanced Technology (ICEEAT)*, Nov. 5-7 2023.
- [20] Fakir, K.; Haba, J.A.; Ennawaoui, C.; and El Mouden, M. 2023. Forecasting electricity generated in a Rankine power station running in grid-connected mode. *International Journal of Intelligent Engineering and Systems* 16(6): 339–355.
<https://doi.org/10.22266/ijies2023.1231.28>
- [21] Huynh, A.D.N.; and Khuong, T. 2024. The comparison of GRU and LSTM in solar power generation forecasting application. *International Journal of Scientific Research Archive* 13(1): 1360–1370.
<https://doi.org/10.30574/ijrsra.2024.13.1.1831>
- [22] Lee, D.S.; and Son, S.Y. 2024. PV forecasting model development and impact assessment via imputation of missing PV power data. *IEEE Access* 12: 12843–12852.
<https://doi.org/10.1109/ACCESS.2024.3352038>
- [23] Ait Mansour, A.; Tilioua, A.; and Touzani, M. 2024. Bi-LSTM, GRU and 1D-CNN models for short-term photovoltaic panel efficiency forecasting case amorphous silicon grid-connected PV system. *Results in Engineering* 21: 101886. <https://doi.org/10.1016/j.rineng.2024.101886>
- [24] Al-Dahidi, S.; et al. 2025. Multistep PV power forecasting using deep learning models and the reptile search algorithm. *Results in Engineering* 27: 106265.
<https://doi.org/10.1016/j.rineng.2025.106265>
- [25] Ali, S.S.; Ali, M.; Bhatti, D.M.; and Choi, B.J. 2025. Explainable clustered federated learning for solar energy forecasting. *Energies* 18(9).
<https://doi.org/10.3390/en18092380>
- [26] Open-Meteo. 2025. Historical Weather API Documentation. Available online: <https://open-meteo.com/en/docs/historical-weather-api>; accessed August 21, 2025.
- [27] Exep. 2025. Monitoring Platform. Available online: <https://app-portal.exep.io/login>; accessed August 21, 2025.
- [28] Hochreiter, S.; and Schmidhuber, J. 1996. LSTM can solve hard long time lag problems. *Proceedings of the 10th International Conference on Neural Information Processing Systems (NIPS'96)*, Cambridge, MA, USA, Dec. 3-8 1996: 473–479.
- [29] Lim, B.; Arik, S.Ö.; Loeff, N.; and Pfister, T. 2021. Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting* 37(4): 1748–1764. <https://doi.org/10.1016/j.ijforecast.2021.03.012>
- [30] IBM. 2025. Random Forest: An Overview of Applications and How It Works. IBM online documentation. Available online: <https://www.ibm.com/topics/random-forest>; accessed August 21, 2025.

APPENDIX A: DL architectures for two-stage solar power forecasted

A.1 LSTM Architecture

An LSTM is a RNN whose standard architecture consists of a memory cell and three types of gates: an input gate, an output gate, and a forget gate [28]. These mechanisms control the flow of information into and out of the memory cells, enabling the model to retain values over long periods of time and learn long-term relationships in the time series data. Namely, a forget gate determines which parts of the previous cell state to retain or discard by calculating a value between 0 and 1 from the previous state and the current input. A value closer to 1 means the information is retained, while a value closer to 0 means it is discarded. The input gate determines which new information should be added to the current cell state, and the output gate controls which parts of the cell state to pass to the next layer, considering both the current and previous states. The LSTM architecture structure is shown in Fig. 11.

A.2 TFT Architecture

The TFT model is an advanced deep learning architecture for multi-horizon time-series forecasting. It efficiently processes both static and temporal data. TFT integrates RNN, attention mechanisms, and GRNs to learn temporal relationships and dynamically select important variables [29]. TFT's architecture consists of VSNs for selecting relevant variables at each time point, an LSTM encoder-decoder for learning sequences, a temporal fusion decoder that uses multi-head attention to capture long-term relationships, and a GRN to improve the quality of selected features. TFT also supports multi-step probabilistic forecasting via quantile regression to estimate outcome uncertainty.

In this research, the TFT-lite architecture was applied to forecast temperature, irradiance, and solar power generation. TFT-lite preserves the core components of TFT, including VSN, attention mechanisms, encoder-decoder, and GRNs, while simplifying the structure by removing static variables, reducing encoder-decoder layers, and

performing deterministic one-step-ahead forecasting. These modifications enable faster training, are suitable for small datasets and retain TFT's ability to capture temporal dependencies and select important features. The architecture is illustrated in Fig. 12

A.3 Bi-LSTM Architecture

In this stage, Bi-LSTM, as shown in Fig. 13, is a RNN that processes data in two temporal directions. By learning from both past and future data, Bi-LSTM can achieve higher accuracy than unidirectional LSTMs. Its structure consists of two separate hidden layers one processing forward in time and the other backward both connected to a single output. The input is fed to both hidden layers simultaneously, while neurons within each layer remain unconnected. The Bi-LSTM output, denoted by y_i , is obtained by combining the outputs from both directions.

A.4 GRU Architecture

The GRU is a variant of RNN designed to capture long-term dependencies and mitigate the vanishing gradient problem typical of traditional RNNs. Conceptually similar to LSTM, the GRU has a simpler architecture with fewer parameters, allowing for more efficient training. Fig. 14 illustrates the GRU architecture, which includes two types of gates: the update gate and the reset gate. The reset gate controls the degree to which information from the previous hidden state is forgotten, while the update gate determines how much past information is retained and passed forward to the current output.

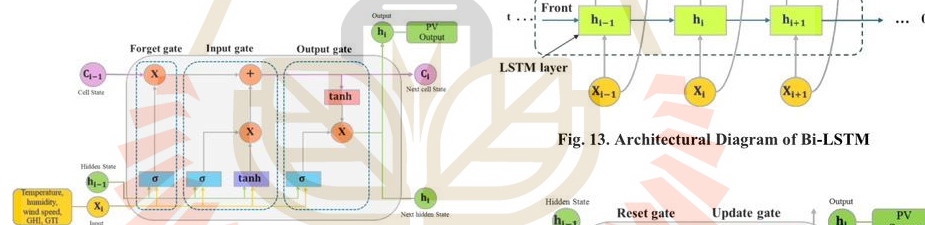


Fig. 11. Architectural Diagram of LSTM

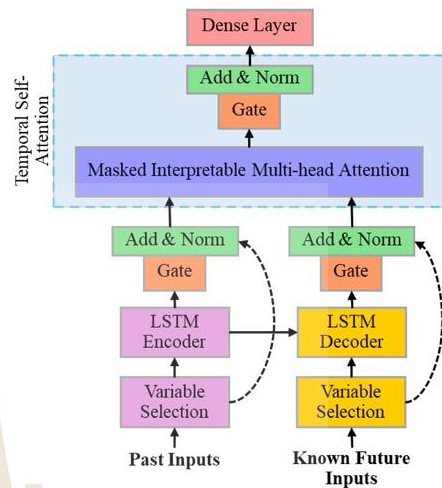


Fig. 12. Architectural diagram of TFT-lite

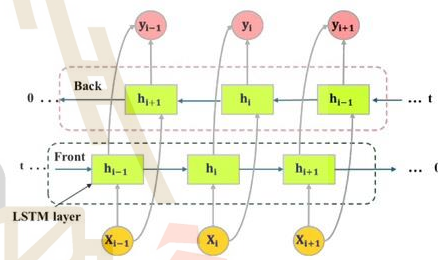


Fig. 13. Architectural Diagram of Bi-LSTM

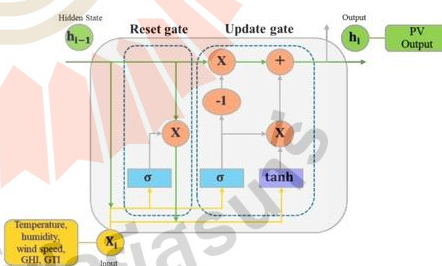


Fig. 14. Architectural Diagram of GRU

A.5 RF Architecture

In addition to RNN-based models, ensemble techniques such as RF are widely used for forecasting. RF constructs a set of uncorrelated decision trees using both bagging and feature randomness [30]. The final output is obtained by averaging the predictions of individual trees, which reduces variance and improves forecast accuracy. The basic structure of RF is shown in Fig. 15

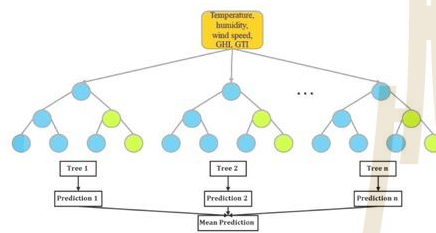


Fig. 15. Architectural Diagram of RF

BIOGRAPHY

Ms. Wipada Thongwily was born on February 9, 2001, in Nakhon Ratchasima Province, Thailand. She graduated with a Bachelor of Engineering degree in Electrical Engineering from Suranaree University of Technology in 2024 and, in the same year, pursued a Master of Engineering degree in Electrical Engineering at the School of Engineering, Suranaree University of Technology. Her academic interests include the application of deep learning techniques in power systems, time-sensitive data analysis and forecasting in electrical power, and the application of artificial intelligence and machine learning models in electrical engineering.

During her graduate studies, she served as a teaching assistant in courses related to power systems and renewable energy, which enhanced her academic, communication, and teaching support skills. Furthermore, she presented her research at the International Electrical Engineering Congress (IEEECON 2025), an experience that further developed her international academic communication abilities.

She is committed to contributing to the development and application of advanced technologies to improve the efficiency and accuracy of power system analysis, as well as forecasting and management of modern energy systems. Her goal is to pursue a career and create academic work in this field to support the sustainable development of energy infrastructure and enhance Thailand's energy security in the future.