

UNIVERSITY OF THESSALY
SCHOOL OF ENGINEERING
DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

**Machine Learning Methods for Photovoltaic Generation
Prediction: A Comparative Analysis using Operational and
Meteorological Data**

Diploma Thesis

Refenes Alexandros

Supervisor: Stamoulis Georgios

Volos 2025



UNIVERSITY OF THESSALY

SCHOOL OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

**Machine Learning Methods for Photovoltaic Generation
Prediction: A Comparative Analysis using Operational and
Meteorological Data**

Diploma Thesis

Refenes Alexandros

Supervisor: Stamoulis Georgios

Volos 2025



ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

**Μέθοδοι Μηχανικής Μάθησης για την Πρόβλεψη
Φωτοβολταϊκής Παραγωγής: Συγκριτική Ανάλυση με
χρήση Λειτουργικών και Μετεωρολογικών Δεδομένων**

Διπλωματική Εργασία

Ρεφενές Αλέξανδρος

Επιβλέπων/πουσα: Σταμούλης Γεώργιος

Βόλος 2025

Approved by the Examination Committee:

Supervisor **Stamoulis Georgios**

Professor, Department of Electrical and Computer Engineering, University of Thessaly

Member **Kolomvatsos Konstantinos**

Assistant Professor, Department of Informatics and Telecommunications, University of Thessaly

Member **Loukopoulos Athanasios**

Associate Professor, Department of Computer Science and Biomedical Informatics, University of Thessaly

Acknowledgements

I would like to express my heartfelt gratitude to my supervisor, Professor Stamoulis Georgios, for his invaluable guidance and support throughout my research. His expertise and encouragement have been instrumental in shaping my work.

I also want to extend my thanks to my family for their unwavering support and understanding during this journey. Their love and encouragement have been my greatest motivation. Thank you all for your contributions to my success.

DISCLAIMER ON ACADEMIC ETHICS AND INTELLECTUAL PROPERTY RIGHTS

«Being fully aware of the implications of copyright laws, I expressly state that this diploma thesis, as well as the electronic files and source codes developed or modified in the course of this thesis, are solely the product of my personal work and do not infringe any rights of intellectual property, personality and personal data of third parties, do not contain work / contributions of third parties for which the permission of the authors / beneficiaries is required and are not a product of partial or complete plagiarism, while the sources used are limited to the bibliographic references only and meet the rules of scientific citing. The points where I have used ideas, text, files and / or sources of other authors are clearly mentioned in the text with the appropriate citation and the relevant complete reference is included in the bibliographic references section. I also declare that the results of the work have not been used to obtain another degree. I fully, individually and personally undertake all legal and administrative consequences that may arise in the event that it is proven, in the course of time, that this thesis or part of it does not belong to me because it is a product of plagiarism».

The declarant

Refenes Alexandros

Diploma Thesis

Machine Learning Methods for Photovoltaic Generation Prediction: A Comparative Analysis using Operational and Meteorological Data

Refenes Alexandros

Abstract

This thesis investigates machine learning approaches for photovoltaic power prediction, combining operational data from solar installations with meteorological parameters from the ERA5 reanalysis dataset. The work addresses challenges in solar forecasting through systematic data integration and model evaluation.

The methodology encompasses several key components. First, we develop techniques for combining high-frequency solar generation data with ERA5 meteorological parameters, addressing temporal alignment and spatial resolution differences. Next, we implement and evaluate multiple machine learning approaches, including tree-based ensemble methods (Random Forest, XGBoost, LightGBM, CatBoost) and deep learning models (Temporal Fusion Transformer). A comprehensive validation framework assesses both temporal prediction accuracy and generalization to different system configurations.

The analysis reveals distinct characteristics of different modeling approaches. Tree-based models demonstrate robust performance with efficient computation, while the Temporal Fusion Transformer shows advantages in capturing complex temporal patterns. The work provides insights into feature importance patterns and model behavior across different prediction scenarios, contributing to the understanding of machine learning applications in solar power forecasting.

Keywords:

photovoltaic power prediction, machine learning, weather data integration, temporal modeling, solar forecasting

Διπλωματική Εργασία

Μέθοδοι Μηχανικής Μάθησης για την Πρόβλεψη Φωτοβολταϊκής

Παραγωγής: Συγκριτική Ανάλυση με χρήση Λειτουργικών και

Μετεωρολογικών Δεδομένων

Ρεφενές Αλέξανδρος

Περίληψη

Η παρούσα διπλωματική εργασία διερευνά μεθόδους μηχανικής μάθησης για την πρόβλεψη φωτοβολταϊκής ισχύος, συνδυάζοντας λειτουργικά δεδομένα από ηλιακές εγκαταστάσεις με μετεωρολογικές παραμέτρους από το σύνολο δεδομένων ERA5. Η εργασία αντιμετωπίζει προκλήσεις στην ηλιακή πρόβλεψη μέσω συστηματικής ενοποίησης δεδομένων και αξιολόγησης μοντέλων.

Η μεθοδολογία περιλαμβάνει διάφορα βασικά στοιχεία. Αρχικά, αναπτύσσουμε τεχνικές για το συνδυασμό δεδομένων ηλιακής παραγωγής υψηλής συχνότητας με μετεωρολογικές παραμέτρους ERA5, αντιμετωπίζοντας ζητήματα χρονικής ευθυγράμμισης και διαφορών χωρικής ανάλυσης. Στη συνέχεια, υλοποιούμε και αξιολογούμε πολλαπλές προσεγγίσεις μηχανικής μάθησης, συμπεριλαμβανομένων μεθόδων δενδρικών συνόλων (Random Forest, XGBoost, LightGBM, CatBoost) και μοντέλων βαθιάς μάθησης (Temporal Fusion Transformer). Ένα ολοκληρωμένο πλαίσιο επικύρωσης αξιολογεί τόσο τη χρονική ακρίβεια πρόβλεψης όσο και τη γενίκευση σε διαφορετικές διαμορφώσεις συστημάτων.

Η ανάλυση αποκαλύπτει διακριτά χαρακτηριστικά διαφορετικών προσεγγίσεων μοντελοποίησης. Τα δενδρικά μοντέλα επιδεικνύουν στιβαρή απόδοση με αποδοτικό υπολογισμό, ενώ ο Temporal Fusion Transformer παρουσιάζει πλεονεκτήματα στην καταγραφή πολύπλοκων χρονικών μοτίβων. Η εργασία παρέχει γνώσεις σχετικά με τα πρότυπα σημαντικότητας χαρακτηριστικών και τη συμπεριφορά μοντέλων σε διαφορετικά σενάρια πρόβλεψης, συμβάλλοντας στην κατανόηση των εφαρμογών μηχανικής μάθησης στην πρόβλεψη ηλιακής ενέργειας.

Λέξεις-κλειδιά:

πρόβλεψη φωτοβολταϊκής ισχύος, μηχανική μάθηση, ενοποίηση μετεωρολογικών δεδομένων, χρονική μοντελοποίηση, ηλιακή πρόβλεψη

Table of contents

Acknowledgements	ix
Abstract	xii
Περίληψη	xiii
Table of contents	xv
List of figures	xix
List of tables	xxiii
Abbreviations	xxv
1 Introduction	1
1.1 Thesis Objective	1
1.1.1 Contributions	2
1.2 Thesis Organization	2
2 Literature Review	3
2.1 Introduction to Solar Energy Forecasting	3
2.2 Overview of Numerical Weather/Climatological Data for Solar Forecasting	4
2.2.1 Numerical Weather Data in Renewable Energy Applications	4
2.2.2 Reanalysis Data vs. Satellite Data	4
2.2.3 ERA5 for Solar Energy Forecasting	4
2.3 Key Weather Variables Impacting Solar Energy Production	5
2.3.1 Solar Radiation	5
2.3.2 Temperature and Humidity	5

2.3.3	Cloud Cover and Precipitation	6
2.3.4	Wind and Other Environmental Factors	6
2.4	Machine Learning Models in Solar Energy Forecasting	6
2.4.1	Traditional and Advanced Machine Learning Approaches	7
2.4.2	Temporal Fusion Transformer (TFT)	7
2.4.3	Comparative Studies	8
2.5	Overview of Current Research	8
2.5.1	Data Integration	8
2.5.2	Model Evaluation	8
2.5.3	Feature Analysis	9
2.6	Conclusion	9
3	Dataset Construction and Preparation	11
3.1	Introduction	11
3.2	Solar Farm Data from PVOutput.org	12
3.2.1	Overview of PVOutput Platform	12
3.2.2	Installation Characteristics and Selection Criteria	12
3.2.3	Data Structure and Quality Assessment	13
3.2.4	Temporal Resolution and Aggregation Process	15
3.3	ERA5 Meteorological Data	18
3.3.1	ERA5 Reanalysis Dataset Overview	18
3.3.2	Selected Weather Parameters	19
3.3.3	Spatial and Temporal Characteristics	21
3.3.4	Data Retrieval and Processing	22
3.4	Data Integration and Preprocessing	23
3.4.1	Temporal and Spatial Alignment	23
3.4.2	Data Quality Assurance	24
3.5	Final Dataset Structure and Analysis	25
3.5.1	Dataset Composition	25
3.5.2	Variable Analysis and Relationships	27
3.5.3	Temporal Characteristics	31
3.6	Dataset Considerations	33
3.6.1	Quality and Coverage Assessment	33

3.6.2	Usage Implications	35
4	Methodology	37
4.1	Introduction	37
4.2	Feature Engineering	38
4.2.1	Feature Categories	38
4.2.2	Feature Transformation and Organization	40
4.3	Model Development	42
4.3.1	Baseline Model	42
4.3.2	Tree-based Ensemble Models	43
4.3.3	Deep Learning Approach	52
4.4	Validation Framework	54
4.4.1	Validation Strategies	54
4.4.2	Performance Metrics	56
4.4.3	Model Comparison Framework	58
4.5	Summary	60
5	Results and Analysis	61
5.1	Baseline Model Results	61
5.2	Tree-based Models Performance	62
5.2.1	Random Forest Results	62
5.2.2	XGBoost Results	73
5.2.3	LightGBM Results	83
5.2.4	CatBoost Results	93
5.3	Deep Learning Model Performance	102
5.3.1	Temporal Fusion Transformer Results	102
5.4	Comparative Analysis	115
5.4.1	Tree-based Models Comparison	115
5.4.2	Comparing Tree-based and Deep Learning Approaches	116
5.4.3	Comparative Analysis Conclusions	117
6	Conclusion	119
6.1	Summary and Conclusions	119
6.2	Future Work	120

Bibliography	121
---------------------	------------

List of figures

3.1	Raw 5-Minute Generation Data	15
3.2	Energy Measurement Comparison	17
3.3	Dataset Characteristics Overview	25
3.4	Feature Importance Analysis	28
3.5	Variable Correlation Analysis	29
3.6	Weather Pattern Generation Profiles	30
3.7	Time Series Decomposition Analysis	32
3.8	Dataset Quality Metrics	34
5.1	RF Temporal Efficiency Comparison	63
5.2	RF Absolute Power Generation	64
5.3	RF Feature Importance Analysis	65
5.4	RF Efficiency Trend Analysis	66
5.5	RF Generation Trend Analysis	67
5.6	RF Efficiency Error Distribution	68
5.7	RF Generation Error Distribution	69
5.8	RF System-based Efficiency Prediction	70
5.9	RF System-based Feature Importance	71
5.10	RF System-based Trend Analysis	72
5.11	RF System-based Error Analysis	73
5.12	XGBoost Temporal Efficiency Analysis	74
5.13	XGBoost Power Generation Prediction	75
5.14	XGBoost Feature Importance Rankings	76
5.15	XGBoost Efficiency Trend Analysis	77
5.16	XGBoost Generation Trend Analysis	77

5.17	XGBoost Efficiency Error Distribution	78
5.18	XGBoost Generation Error Distribution	79
5.19	XGBoost System-based Prediction	80
5.20	XGBoost System-based Feature Analysis	81
5.21	XGBoost System-based Trend Analysis	82
5.22	XGBoost System-based Error Analysis	83
5.23	LGBM Temporal Efficiency Analysis	84
5.24	LGBM Power Generation Prediction	85
5.25	LGBM Feature Importance Analysis	86
5.26	LGBM Efficiency Trend Analysis	87
5.27	LGBM Generation Trend Analysis	87
5.28	LGBM Efficiency Error Distribution	88
5.29	LGBM Generation Error Distribution	89
5.30	LGBM System-based Prediction	90
5.31	LGBM System-based Feature Analysis	91
5.32	LGBM System-based Trend Analysis	92
5.33	LGBM System-based Error Analysis	93
5.34	CatBoost Temporal Efficiency Analysis	94
5.35	CatBoost Power Generation Prediction	94
5.36	CatBoost Feature Importance Analysis	95
5.37	CatBoost Efficiency Trend Analysis	96
5.38	CatBoost Generation Trend Analysis	97
5.39	CatBoost Efficiency Error Distribution	98
5.40	CatBoost Generation Error Analysis	98
5.41	CatBoost System-based Prediction	99
5.42	CatBoost System-based Feature Analysis	100
5.43	CatBoost System-based Trend Analysis	101
5.44	CatBoost System-based Error Analysis	102
5.45	TFT Temporal Efficiency Analysis	103
5.46	TFT Power Generation Prediction	104
5.47	TFT Feature Importance Analysis	105
5.48	TFT Trend Analysis	107

5.49	TFT Error Distribution	108
5.50	TFT Residual Analysis	109
5.51	TFT System-based Prediction	110
5.52	TFT System-based Feature Analysis	111
5.53	TFT System-based Trend Analysis	113
5.54	TFT System-based Error Distribution	114
5.55	TFT System-based Residual Analysis	115

List of tables

3.1	Range of Static System Parameters in Dataset	14
3.2	Key statistical characteristics of primary dataset variables	27
4.1	Static System Features	38
4.2	Temporal Feature Components	39
4.3	ERA5 Meteorological Features	40
4.4	Feature Organization by Type	41
4.5	Random Forest Hyperparameter Ranges and Descriptions	44
4.6	XGBoost Hyperparameter Configuration	46
4.7	LightGBM Hyperparameter Configuration	48
4.8	CatBoost Hyperparameter Configuration	50

Abbreviations

blh	Boundary Layer Height
d2m	2-meter Dewpoint Temperature
ECMWF	European Centre for Medium-Range Weather Forecasts
ERA5	ECMWF Reanalysis v5
hcc	High Cloud Cover
kW	Kilowatt
kWh	Kilowatt-hour
lcc	Low Cloud Cover
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
mcc	Medium Cloud Cover
PV	Photovoltaic
R ²	Coefficient of determination
RMSE	Root Mean Square Error
skt	Skin Temperature
sp	Surface Pressure
SSR	Surface Solar Radiation
SSRD	Surface Solar Radiation Downwards
SSRDC	Surface Solar Radiation Downwards Clear-sky
t2m	2-meter Temperature
tcc	Total Cloud Cover
tco3	Total Column Ozone
tcwv	Total Column Water Vapour
TFT	Temporal Fusion Transformer

Chapter 1

Introduction

Solar power generation has been gaining importance in modern electrical grids in pursuit of renewable sources of energy. Photovoltaic systems promise clean and renewable power. However, the output is influenced greatly by weather conditions, seasonal patterns, and atmospheric phenomena. This variation creates challenges for grid integration, making accurate power prediction important for both research and practical uses.

There are a few factors that make predictions in photovoltaic power complex. To begin with, the short-term generation is influenced by weather conditions that depend heavily on cloud cover and atmospheric clarity. In addition, seasonal changes determine the duration and intensity of sunlight availability. Finally, local environmental conditions hamper the effective conversion of the panels. The available methods of prediction suffer either due to limited sources of data or from employing simplified statistical methods, leading to issues with accuracy during periods of rapid variations.

1.1 Thesis Objective

This study looks at predicting solar power generation using operational data from solar farms and meteorological elements. It combines the data from both sources and compares the accuracy of various machine learning methods.

Key technical problems include matching time data from different sources dealing with different spatial resolutions and using models that learn well from multiple different system configurations. The goal is to understand how different models work in this area while remaining practical.

1.1.1 Contributions

This work makes several key contributions:

1. Techniques for merging operational PV data with ERA5 reanalysis parameters
2. Execution and evaluation of various machine learning methods for power forecasting
3. Creation of validation strategies to evaluate model generalization
4. Comparative examination of modeling techniques and their attributes
5. Examination of important meteorological and operational factors

1.2 Thesis Organization

Chapter 2 reviews related work and background concepts in solar forecasting. Chapter 3 describes the construction and preprocessing of the dataset. Chapter 4 presents the modeling approaches and validation methodology. Chapter 5 provides results and analysis, while Chapter 6 offers conclusions and potential future work.

Chapter 2

Literature Review

2.1 Introduction to Solar Energy Forecasting

Transitioning to renewable energy, especially solar, is extremely important for sustainable growth and combating climate change. Solar energy forecasting is thus vital for grid stability by integrating renewables into current power systems. As the world's share of solar power continuously increases, the need for reliable prediction models becomes increasingly critical for managing supply and demand effectively [1].

Weather variability leads to significant fluctuations in solar irradiance, making solar forecasting a difficult endeavor. For example, some parameters, such as cloud coverage, air moisture, and seasonal variance, alter solar power, making forecasts very hard [2]. Due to this challenge, previous methods have failed to measure up; therefore, new methods should be designed based on improved data and machine-learning techniques.

Recent progress in numerical weather data offers new opportunities for better forecasts. Datasets like ERA5 give detailed reanalysis data that captures key atmospheric details with consistent spatial and temporal resolution. By mixing this data with machine learning algorithms, researchers aim to improve prediction accuracy and trustworthiness while fixing significant gaps that were found before. This literature review will look at the current state of solar forecasting, focusing on using the reanalysis of climate data and advanced machine learning models to predict photovoltaic output well.

2.2 Overview of Numerical Weather/Climatological Data for Solar Forecasting

Accurate solar energy forecasting relies heavily on the availability and quality of numerical weather data. This data plays a crucial role in understanding atmospheric conditions that affect solar irradiance and, consequently, photovoltaic (PV) energy production.

2.2.1 Numerical Weather Data in Renewable Energy Applications

Numerical weather data consists of information obtained from mathematical models that replicate atmospheric behavior via physical principles. This data offers crucial insights into meteorological factors such as temperature, humidity, wind speed, and solar radiation, which are essential for forecasting solar energy production [3]. Common datasets for solar forecasting encompass ERA5, NARR, and MERRA-2; each brings distinct features suited to diverse forecasting requirements. For example, ERA5 delivers high-resolution reanalysis data with detailed atmospheric variables recorded hourly, which is especially beneficial for solar energy applications [4].

2.2.2 Reanalysis Data vs. Satellite Data

Reanalysis data differs from satellite data primarily in its methodology and application. While satellite data offers real-time observations of atmospheric conditions, it can be affected by noise and limitations in temporal resolution. In contrast, reanalysis datasets like ERA5 combine historical observations with model outputs to create a consistent spatial-temporal representation of weather variables [5]. This consistency is crucial for solar forecasting, allowing for more reliable predictions over extended periods.

2.2.3 ERA5 for Solar Energy Forecasting

ERA5 has become a favored dataset for predicting solar energy because of its high spatial resolution (31 km) and broad range of atmospheric variables. It encompasses essential factors like surface solar radiation, temperature profiles, and humidity levels, which greatly affect PV output [6]. Combining ERA5 data with machine learning algorithms has produced encouraging outcomes in improving the precision of solar energy forecasts, effectively tackling

some issues linked to conventional forecasting techniques.

2.3 Key Weather Variables Impacting Solar Energy Production

Various weather variables significantly influence the production of solar energy. Understanding these variables is essential for accurately forecasting photovoltaic (PV) output. This section discusses the key weather factors that affect solar energy generation.

2.3.1 Solar Radiation

Solar radiation is the most important variable that affects the production of photovoltaic energy. Key metrics include:

- **Surface Solar Radiation (SSR):** Represents the total solar energy received per unit area at the Earth's surface.
- **Surface Solar Radiation Downwards (SSRD):** Measures the solar radiation reaching the surface.
- **Clear-Sky Irradiance:** Denotes the amount of solar radiation reaching the surface under clear sky conditions.

Research has indicated a strong link between solar radiation levels and PV output, suggesting that precise solar radiation forecasting can enhance energy generation predictions [2].

2.3.2 Temperature and Humidity

Temperature and humidity also play significant roles in PV performance:

- **Air Temperature:** Higher temperatures can reduce the efficiency of PV panels due to increased resistance in electrical components. In contrast, lower temperatures can improve efficiency.
- **Humidity:** Increased humidity can lead to condensation on panels, which may affect performance and energy output.

Research indicates that both temperature and humidity have direct impacts on the efficiency and output of solar panels [7].

2.3.3 Cloud Cover and Precipitation

Cloud cover is a critical factor influencing solar irradiance:

- **Total Cloud Cover (tcc), High Cloud Cover (hcc), and Low Cloud Cover (lcc):** These metrics quantify the sunlight reaching the surface, with increased cloud cover generally leading to reduced solar radiation.
- **Precipitation:** Rainfall can temporarily diminish solar output by blocking sunlight but can also assist in cleaning panels, potentially enhancing performance afterward.

Grasping cloud dynamics and precipitation trends is vital for improving the reliability of solar forecasting [8].

2.3.4 Wind and Other Environmental Factors

Wind speed and other environmental factors also affect PV systems:

- **Wind Speed:** Although wind can cool PV panels and improve efficiency, high winds can cause physical damage or soiling.
- **Other Atmospheric Features:** Variables such as dust accumulation and temperature fluctuations can influence local weather patterns, further affecting solar irradiance.

Incorporating these factors into forecasting models can improve their predictive capabilities [9].

2.4 Machine Learning Models in Solar Energy Forecasting

Applying machine learning (ML) models has revolutionized solar energy forecasting by providing advanced techniques for complex dataset analysis. This section reviews widely used traditional and advanced machine learning approaches, emphasizing their relevance in predicting photovoltaic (PV) output.

2.4.1 Traditional and Advanced Machine Learning Approaches

Traditional machine learning models such as Linear Regression, Random Forest, and Support Vector Machines (SVM) have usually been employed in solar energy forecasting. These models effectively capture non-linear relationships between input features and output predictions.

- **Random Forest:** This ensemble learning method constructs multiple decision trees and merges them to improve accuracy. It is beneficial for handling large datasets with numerous input variables, making it a popular choice for solar forecasting [10].
- **XGBoost:** Known for its speed and performance, XGBoost implements gradient boosted decision trees designed for efficiency. It has been shown to outperform many traditional models in various forecasting tasks, including solar energy prediction [11].
- **Support Vector Machines (SVM):** SVMs are adequate for regression tasks where the relationship between variables is complex. They work well with high-dimensional data, making them suitable for solar forecasting applications.

2.4.2 Temporal Fusion Transformer (TFT)

The Temporal Fusion Transformer (TFT) is an advanced deep learning model designed specifically for time series forecasting. It effectively handles static and temporal features by combining the strengths of recurrent neural networks (RNNs) and attention mechanisms.

Key advantages of TFT include:

- **Handling Static and Temporal Features:** TFT can incorporate both static covariates (e.g., geographical information) and dynamic inputs (e.g., weather data), allowing for more comprehensive modeling of solar energy generation.
- **Interpretability:** The attention mechanism within TFT provides transparency into the most influential features in making predictions [12].

Recent studies have demonstrated the effectiveness of TFT in solar energy forecasting, showing improved accuracy over traditional models when applied to large datasets with complex temporal patterns [13].

2.4.3 Comparative Studies

Comparative studies evaluating the performance of various machine learning models in predicting PV output have highlighted the advantages of using advanced techniques like TFT. Metrics such as Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) are commonly used to assess model performance. Research indicates that deep learning models, particularly those using attention mechanisms, tend to outperform traditional approaches in terms of prediction accuracy and robustness [14].

2.5 Overview of Current Research

This study investigates various facets of photovoltaic (PV) power prediction, emphasizing the integration of numerical weather data with machine learning techniques. It builds upon earlier solar forecasting research to discover methods for merging different data sources and modeling strategies to enhance prediction trends' comprehension.

2.5.1 Data Integration

The research delves into combining diverse data sources, tackling challenges highlighted in prior sections:

- **Data Integration:** Investigating techniques to merge PV output data with ERA5 re-analysis information.
- **System Analysis:** Analyzing prediction trends across multiple PV systems with varying capacities.
- **Temporal Coverage:** Examining data over several seasons to grasp prediction trends.

2.5.2 Model Evaluation

The research assesses multiple machine learning strategies to evaluate their efficacy in PV power prediction:

- **Tree-based Models:** Evaluation of ensemble methods such as Random Forest, XGBoost, LightGBM, and CatBoost.

- **Deep Learning:** Investigation of the Temporal Fusion Transformer for time series forecasting.
- **Model Validation:** Application of both temporal and system-based validation methods.

2.5.3 Feature Analysis

The study examines various input features and their connections to PV power generation:

- **Weather Parameters:** Evaluation of ERA5 meteorological variables essential for PV prediction.
- **Time Features:** Analysis of temporal patterns within generation data.
- **System Features:** Study of characteristics specific to the systems.

Subsequent chapters explore these elements deeply, enhancing our understanding of PV power prediction through numerical weather data and machine learning techniques. The analysis aims to glean insights into the effectiveness of various methodologies while considering their practical implementations in solar forecasting.

2.6 Conclusion

This literature review has discussed all the components of photovoltaic power prediction, however, giving particular attention to the combination of numerical weather data and machine learning methods. It has outlined a number of themes in the current work.

Accurate meteorological data is vital for PV forecasting. Advanced reanalysis datasets like ERA5 provide extensive coverage of key weather variables, along with reliable spatial and temporal resolutions. These features facilitate a more precise examination of how weather fluctuations affect photovoltaic system efficiency.

The review has also considered the most important meteorological parameters that play a role in solar power generation, ranging from direct (solar radiation) to indirect (temperature, cloud cover, wind speed) factors. Understanding these relationships is important in creating effective prediction models.

Looking at various approaches, traditional machine learning and deep learning both demonstrate strong potential in photovoltaic (PV) prediction tasks. Tree-based ensemble methods provide solid performance and clarity, whereas intricate architectures such as the Temporal Fusion Transformer excel at managing complex temporal patterns.

The literature reveals ongoing challenges in several areas:

- Integration of various temporal resolutions in weather and PV data
- Development of models which can generalize across a range of system sizes
- Effective combination of multiple weather parameters for prediction

These challenges have informed the methodology developed in the following chapters, as different approaches to PV power prediction will be examined through systematic analysis and evaluation.

Chapter 3

Dataset Construction and Preparation

3.1 Introduction

One of the most fundamental steps in any machine learning project involves identifying, creating, and curating a reliable dataset. This turns out to be an important step in the development of solar power prediction models, as it bears directly on forecasting accuracy. In this respect, this chapter discusses how a dataset has been created by fusing high-resolution operational data from Greek PV installations with ERA5 reanalysis meteorological data. It is focused on the direct prediction of power generation, taking temporal patterns and environmental factors to create accurate forecasting models.

It includes two main components: operational data from PVOutput.org, providing 5-minute interval measurements from selected Greek installations, and ERA5 reanalysis data, offering comprehensive hourly meteorological parameters.

The integration challenge presented peculiar problems related to temporal alignment and data harmonization, requiring the most careful consideration in the aggregation methodologies not to lose the essential characteristics while adapting to the common temporal resolution. It also called for sophisticated considerations of time zones, daylight saving time adjustments, and conversion of high-frequency measurements into hourly values. The preprocessing phase included cleaning and transformation procedures, namely handling of missing values and normalization for numerical features.

Feature engineering created a number of derived variables that captured complex relationships in solar power generation, focusing on temporal patterns and interactions between weather systems. Care was taken to clearly separate static features, which represent system

specifications, from dynamic features, which include weather variables and temporal indicators, and process them appropriately based on their unique characteristics. This chapter on constructing datasets lays the groundwork for subsequent chapters that discuss modelling approaches and results. Carefulness regarding data quality, integration, and feature engineering has a direct impact on the effectiveness of machine learning models and the reliability of their predictions.

3.2 Solar Farm Data from PVOutput.org

3.2.1 Overview of PVOutput Platform

PVOutput.org is a community-driven platform that allows system owners of solar power systems to publicly access their generation data on-the-fly[15]. The platform features high temporal resolution along with extensive system metadata, making it ideal for research. In this study, we utilized data from PVOutput.org, which provided solar installation metrics from Greece, including power generation recorded at 5-minute intervals.

The platform's common reporting format guarantees a uniform reporting structure for all installations, with essential measurements such as instantaneous power output, total daily energy generation, and system efficiency metrics. All data points are accurately timestamped and authenticated to a uniform format and process. The platform programmatic interface enables efficient retrieval of historical data which is of utmost importance for the training of machine learning models.

3.2.2 Installation Characteristics and Selection Criteria

This study focuses mainly on seven solar farms in northern Greece, specifically Macedonia. These installations range quite dramatically in size - from relatively modest 99.9 kW systems to impressive 1010.88 kW facilities. They're all located fairly close together, between 40.88°N and 41.08°N latitude and 23.57°E and 23.92°E longitude. This proximity means they experience similar weather conditions, though with enough spatial variation to make things interesting.

The diversity in these installations is remarkable. Our largest site boasts 2,919 solar panels, while the smallest makes do with 555. The panels themselves vary in power output from

100W to 550W. On average, the installations are about 548.93 kW in size (with a standard deviation of 432.54 kW), giving us a good spread of different capacities to study. Six of these sites have been faithfully reporting data from January 2021 through June 2024, while our seventh site (System 90854) joined the party in July 2022. The data quality is exceptional - we're getting readings every 5 minutes with almost 100% reliability, which lets us see exactly how these systems respond to quick changes in weather conditions.

3.2.3 Data Structure and Quality Assessment

The initial data structure from PVOutput.org presents power generation measurements in a cumulative daily format, with readings captured at 5-minute intervals throughout each day. Each measurement point consists of the total energy generated since midnight, creating a stepwise progression that reflects both the instantaneous system performance and the accumulated daily generation. This raw data structure provides the foundation for subsequent temporal processing and integration with meteorological parameters.

Each data record contains several essential components that characterize both the generation measurements and the system configuration. The primary measurement components include timestamped observations with 5-minute precision, cumulative daily energy generation measured in watt-hours, instantaneous power output in watts, and when available, operating voltage and temperature measurements.

The fact that the energy measurements are cumulative lends itself to data validation. A typical day sequence shows several defining features:

- Readings initialize at zero at midnight
- Values remain at zero through pre-dawn hours
- Show monotonic increase during daylight hours
- Reach their final daily value at sunset

This pattern provides natural constraints that aid in data validation, as any decrease in cumulative values during daylight hours would indicate measurement errors.

Quality assessment of the raw data structure reveals several notable characteristics:

1. **Measurement Consistency:** The data maintains exceptional regularity with consistent 5-minute intervals across all installations.

Parameter	Minimum	Maximum
System Capacity (kW)	99.96	1,010.88
Panel Count	408	2,919
Panel Rating (W)	100	550
Latitude (°N)	40.26	41.08
Longitude (°E)	23.28	23.92

Table 3.1: Range of Static System Parameters in Dataset

2. **Temporal Coverage:** Each installation provides complete measurement coverage without significant gaps or interruptions throughout their respective monitoring periods.

3. **Physical Constraints:** All measurements adhere to fundamental physical limitations:

- Non-negative cumulative values
- Generation within system capacity limits
- Zero generation during astronomical night
- Physically plausible rates of change between consecutive measurements

The raw data structure provides measurements at a temporal resolution significantly finer than typical meteorological observations. This high-frequency sampling records quick changes in power output, like those from passing clouds or brief shading, that could be smoothed over in lower temporal resolutions. Nonetheless, this timing inconsistency with meteorological data requires thoughtful attention in later processing stages.

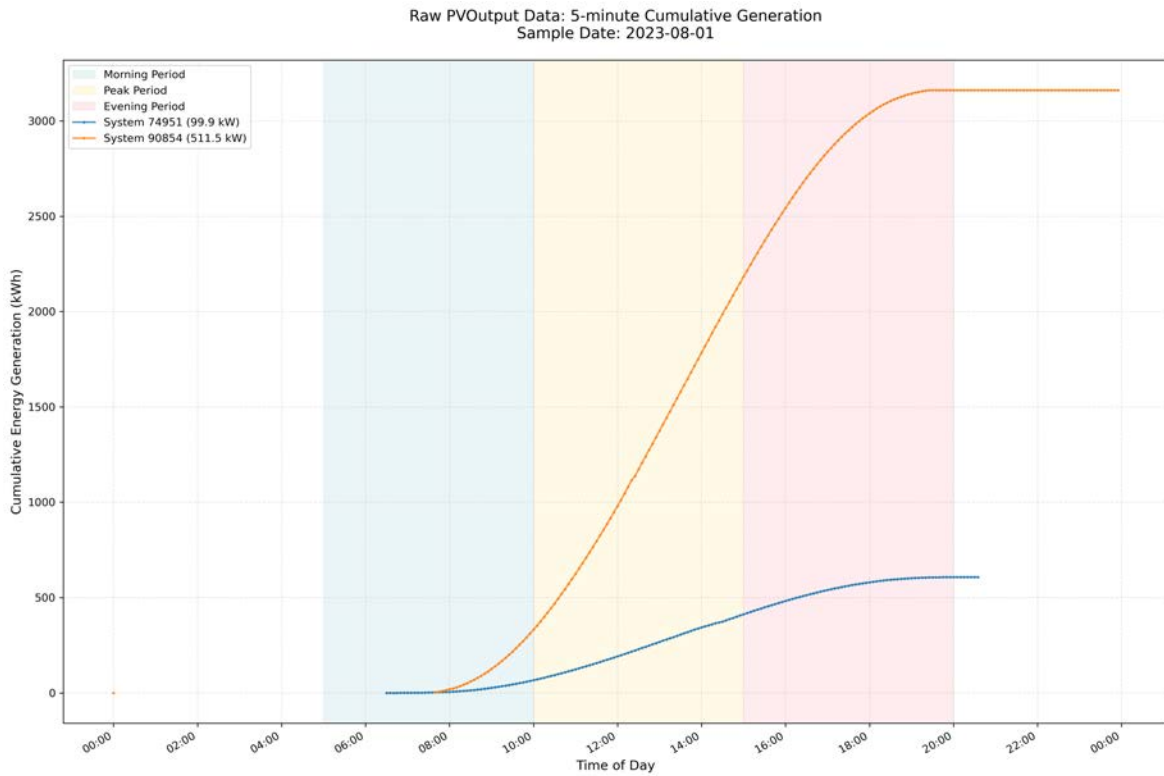


Figure 3.1: Example of Raw 5-Minute Cumulative Generation Data for a Single Day

Cross-comparison between installations reveals consistent patterns in the raw data structure. While absolute generation values differ according to system capacity, the normalized generation profiles show strong correlation between geographically proximate installations during clear sky conditions. Variations in these patterns usually relate to specific weather occurrences, offering another way to validate data quality.

This raw data structure, featuring high temporal resolution and a cumulative daily format, serves as the basis for the following processing steps. The thoroughness and quality of these measurements reduce the necessity for intricate data cleaning or imputation efforts. However, the shift from 5-minute cumulated values to hourly generation values, needed to match the hourly meteorological parameters, deserves special care and the techniques used for temporal aggregation will be elaborated upon in Section 3.2.4.

3.2.4 Temporal Resolution and Aggregation Process

Transforming cumulative daily energy measurements into distinct hourly production values is a vital preprocessing step in constructing the dataset. This method focuses on calculating the energy produced each hour instead of aggregating more frequent measurements.

Initial Data Structure

The raw data from PVOutput presents energy values as a cumulative daily total, where each 5-minute measurement represents the total energy generated since midnight. For any given time t on a particular day:

$$E_{cumulative}(t) = \sum_{t'=00:00}^t E_{generated}(t') \quad (3.1)$$

where $E_{cumulative}(t)$ represents the total energy generated from midnight until time t , and $E_{generated}(t')$ represents the actual energy generated at each time increment.

Hourly Production Calculation

The energy generated during a specific hour h can be determined by calculating the difference between cumulative values at the hour boundaries:

$$E_{hour}(h) = E_{cumulative}(h + 1 : 00) - E_{cumulative}(h : 00) \quad (3.2)$$

For example, the energy generated between 10:00 and 11:00 is calculated as:

$$E_{10-11} = E_{cumulative}(11 : 00) - E_{cumulative}(10 : 00) \quad (3.3)$$

This differential approach provides the actual energy produced within each hour, rather than a cumulative or aggregated value.

Special Cases and Boundary Conditions

The hourly production calculation requires careful handling of several special cases:

1. **Midnight Transition:** At the start of each day ($t = 00 : 00$):

$$E_{cumulative}(00 : 00) = 0 \quad (3.4)$$

2. **Missing Values:** When intermediate measurements are missing, the cumulative nature of the data maintains accuracy:

$$E_{hour}(h) = \max(0, E_{cumulative}(h + 1 : 00) - E_{cumulative}(h : 00)) \quad (3.5)$$

3. **Data Quality:** Non-physical values (such as negative differences) are constrained:

$$E_{hour}(h) \geq 0 \quad \forall h \in [0, 23] \quad (3.6)$$

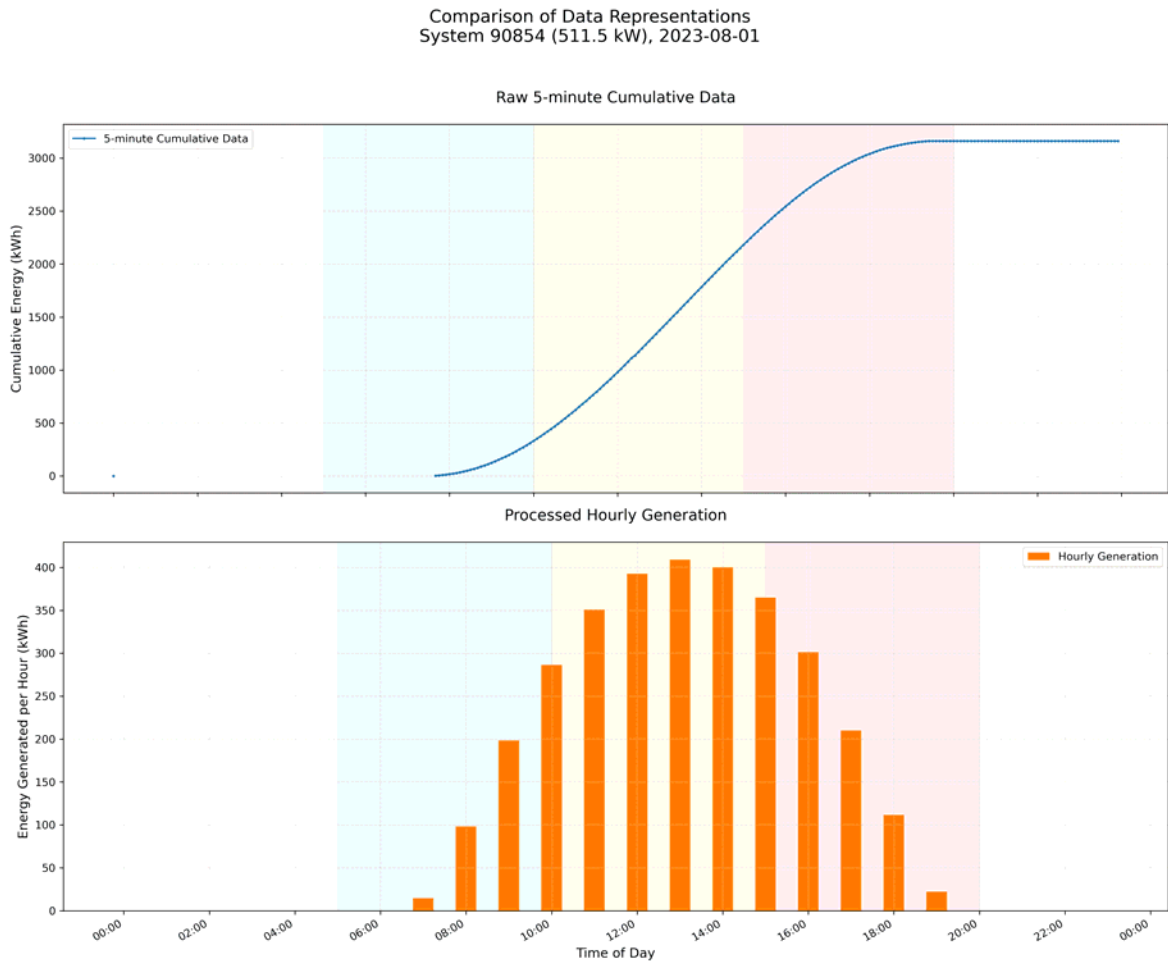


Figure 3.2: Comparison of Cumulative Energy Measurement and Derived Hourly Production

Validation Framework

The conversion process incorporates several validation mechanisms:

$$\sum_{h=0}^{23} E_{hour}(h) = E_{cumulative}(23 : 59) \quad (3.7)$$

This ensures that the sum of hourly production values matches the final cumulative value for each day. Additional validation includes:

- Verification of zero production during nighttime hours
- Confirmation that hourly production values fall within system capacity limits
- Cross-validation of production patterns between neighboring installations
- Assessment of production profiles against clear-sky models

The resulting hourly production dataset offers an accurate depiction of system performance and aligns temporally with meteorological data. This establishes a basis for further analysis regarding the effects of weather on photovoltaic generation.

3.3 ERA5 Meteorological Data

3.3.1 ERA5 Reanalysis Dataset Overview

The ERA5 dataset, created by the European Centre for Medium-Range Weather Forecasts (ECMWF), offers a thorough method for grasping historical weather conditions via reanalysis [16]. This reanalysis merges observational data with numerical weather prediction models, correcting the shortcomings of both: traditional observations deliver precise yet spatially limited data, whereas numerical models provide full coverage but may diverge from actual conditions without observational support.

These observations include a vast array of data sources:

- Satellite measurements of atmospheric temperature, humidity, and cloud cover
- Ground-based weather station measurements of temperature, pressure, and wind
- Weather balloon (radiosonde) data providing vertical profiles of the atmosphere
- Aircraft measurements during flights
- Ocean buoy data for sea surface conditions

Through 4D-Var (four-dimensional variational data assimilation), ERA5 creates optimal estimates of atmospheric conditions by weighing observation reliability against model forecasts. Satellite data plays a crucial role, providing essential measurements of cloud cover and atmospheric radiation that directly influence solar panel performance, particularly valuable in areas with sparse ground-based measurements.

A key strength of ERA5 lies in its consistent processing methodology using the IFS Cycle 41r2 forecasting system, ensuring that dataset changes reflect actual weather patterns rather than processing variations. The resulting dataset provides hourly estimates of atmospheric parameters at 0.25 degrees spatial resolution (approximately 31 kilometers), effectively capturing meteorological conditions relevant to solar power generation while maintaining computational feasibility.

3.3.2 Selected Weather Parameters

The analysis incorporates twenty-six distinct meteorological parameters from ERA5, each selected for its direct or indirect influence on photovoltaic power generation. These parameters characterize atmospheric conditions affecting solar power production through various physical mechanisms, from direct solar resource availability to system operating conditions.

Solar Radiation Parameters

Surface solar radiation downwards (SSRD), measured in J m^{-2} and accumulated hourly, represents the total incoming solar radiation reaching a horizontal plane at the Earth's surface, comprising both direct and diffuse components. For PV modeling, these values are converted to instantaneous power density (W m^{-2}).

Clear-sky radiation parameters provide baseline values under cloudless conditions:

- Surface solar radiation downward, clear sky (SSRDC): The total downward solar radiation that would reach the surface in the absence of clouds, comprising both direct and diffuse components
- Clear-sky direct solar radiation at surface: The unscattered component of solar radiation under cloudless conditions, representing radiation passing directly through the atmosphere
- Surface net solar radiation, clear sky: The difference between incoming and reflected solar radiation assuming clear-sky conditions, accounting for surface albedo effects

Cloud Coverage Parameters

ERA5 provides cloud distribution through four parameters, expressed as dimensionless fractions (0-1):

- High cloud cover: Clouds occurring at pressure levels below 0.45 times the surface pressure (typically above 6 km), often composed of ice crystals
- Medium cloud cover: Clouds in the intermediate layer between 0.45 and 0.8 times the surface pressure (2-6 km), typically containing mixed-phase hydrometeors
- Low cloud cover: Clouds at pressure levels above 0.8 times the surface pressure (below 2 km), predominantly composed of liquid water droplets

- Total cloud cover: The integrated cloud fraction throughout the atmospheric column, accounting for vertical overlap between layers

Temperature Parameters

ERA5 provides temperature parameters measured in Kelvin (K):

- 2m temperature: The air temperature at two meters above the surface, interpolated using boundary layer similarity theory
- Maximum and minimum 2m temperature since previous post-processing: Capturing the diurnal temperature range
- Skin temperature: The theoretical temperature of the Earth's surface layer
- 2m dewpoint temperature: The temperature at which air becomes saturated

Panel efficiency typically decreases by 0.3-0.5% per Kelvin above standard test conditions (25°C).

Wind Conditions

- 10m u-component and v-component of wind: Eastward and northward wind components at ten meters above the surface (m s^{-1})
- 10m wind gust since previous post-processing: Maximum three-second wind speed, important for mechanical stress and cooling effects

Atmospheric State Variables

- Surface pressure: Atmospheric pressure at the surface (Pascal)
- Total column water vapour: Integrated water vapor content (kg m^{-2})
- Total column ozone: Integrated ozone content (kg m^{-2})
- Boundary layer height: Depth of the well-mixed lower atmosphere (meters)

Surface Conditions

- Total precipitation: Accumulated liquid and frozen precipitation in meters of water equivalent
- Snow depth: Snow accumulation in meters of water equivalent
- Snow evaporation: Water loss from snow cover through evaporation
- Snowfall: Accumulated solid precipitation

The comprehensive nature of these parameters enables detailed modeling of the physical processes affecting PV power generation. Their hourly time resolution and physical consistency guarantee the effective capture of both quick changes and extended patterns in system performance.

3.3.3 Spatial and Temporal Characteristics

Grasping how ERA5 defines time and space is essential for our research. The dataset offers measurements every hour, which perfectly fits our requirements for predicting PV power. This hourly pattern is especially useful for converting our 5-minute solar measurements into hourly values, establishing a cohesive time connection across our data sources (outlined in Section 3.2.4).

ERA5 provides a global grid, with each cell measuring 0.25 degrees (approximately 31 kilometers at our study's latitude). Though contemporary satellites can observe the Earth in finer detail and regional weather models may provide higher resolution for specific areas, ERA5 maintains a careful balance. It gathers sufficient detail to comprehend weather patterns while ensuring that the data remains manageable for analysis.

What makes ERA5 particularly elegant is its consistency. In contrast to other data sources, which can provide high resolution for certain quantities and coarser resolutions for others, ERA5 preserves the same spatial and temporal resolution for all features. This homogeneity guarantees that when we look at variations in weather conditions from which solar power generation can be affected, we are seeing them all from the same point of view and at the same time. For our study area in northern Greece, this consistent view captures the regional weather patterns that matter for PV generation while keeping our computational tasks feasible.

3.3.4 Data Retrieval and Processing

The endeavor to obtain ERA5 data guided us to the Climate Data Store (CDS) Application Programming Interface (API), which provides programmatic access to the extensive meteorological archives of the European Centre for Medium-Range Weather Forecasts (ECMWF). Through meticulous evaluation of both efficiency and comprehensiveness, we devised a systematic methodology for the collection of the requisite parameters throughout the designated study period and locations.

We implemented the retrieval process through the CDS API's Python client interface, accessing ERA5's single-level reanalysis product. Given the great volume of data needed, we developed a mindful segmentation process and submitted our requests in six-month segments: January-June and July-December. This approach allowed us to strike a difficult balance between the size of the request and efficiency in processing while guaranteeing data continuity.

For each data segment, we crafted careful request structures encompassing all twenty-six meteorological parameters simultaneously:

- Temporal parameters: Hourly intervals for the specified date range
- Spatial boundaries: Specific coordinates for each PV installation
- Output format: NetCDF (Network Common Data Form) format for efficient multidimensional data storage
- Product type: Reanalysis data specification

To address the inherent latency in retrieving such large volumes of data, we developed an asynchronous processing framework. This method allows simultaneous data requests, significantly lowering retrieval times in comparison to sequential processing, all while ensuring strong error handling and request management during prolonged retrieval sessions.

We organized the retrieved data in a structured directory hierarchy, with each NetCDF file containing six months of hourly data for a specific location. This organization preserves the intricate relationships between temporal and spatial components while facilitating efficient access and management. Each file maintains ERA5's complete metadata structure, including parameter descriptions, units, and quality indicators.

Our post-retrieval processing implements automated validation of data completeness and consistency checks across all parameters, leveraging NetCDF's efficient data manipulation

capabilities. This structured method creates a strong basis for our ongoing analysis and model creation.

3.4 Data Integration and Preprocessing

Integrating PV generation data with meteorological parameters is a key aspect of building our dataset. This complex procedure demands careful focus on temporal alignment, spatial correspondence, and data quality assurance to create a dataset that meets the needs of advanced machine learning applications.

3.4.1 Temporal and Spatial Alignment

A great technical challenge was posed by the synchronization of two correlated data streams: that of high-frequency PV generation measurements and that of hourly meteorological parameters. After converting PV data to hourly resolution (see Section 3.2.4), the next step was to align the datasets temporally.

Our ERA5 data is formatted based on the NetCDF specifications, delivering meteorological parameters at hourly time-slices over six-month periods and at a system location and temporal range. By validating the correspondence in timestamps, we ensured total temporal alignment of 30,648 unique timestamps, leading to 100% hourly coverage for the entire study period.

Our approach of integration is a systematic process with two steps. First, we have kept the datetime system ID-based, wherein specific installation parameters such as physical parameters and geographical coordinates are not altered. Then, for each system group, we merged records by exact datetime matching to create a unified timeline of meteorological conditions and power generation at each installation. This allowed for a record to include information for a single hour and system with all meteorological parameters aligned in time with generation data.

For spatial integration the coordinates of each PV installation (ranging between 40.88°N to 41.08°N latitude and 23.57°E to 23.92°E longitude) were used to extract the respective ERA5 parameters. Thanks to the close geographical density of the installations, it was possible to extract parameters directly without the need for interpolation as the stations were contained within a single grid cell of ERA5, which has a spatial resolution of 0.25 degrees.

Using this process, a unified hourly dataset was created where we have generation data relevant to each installation as well as relevant meteorological parameters. This temporal resolution was sufficient to capture both meteorological variability and associated changes in PV generation, while remaining computationally practicable. Moreover, this determination is consistent with normal conditions for operation of the grid and participation in energy markets.

3.4.2 Data Quality Assurance

The quality control framework enforced comprehensive actions to 201,432 hourly observations from seven PV installations. Initial verification focused on the completeness of ERA5 parameters, each meteorological record being composed of 28 meteorological parameters, such as temperature, radiation components, cloudiness, as well as atmospheric variables. The structure of NetCDF itself guaranteed parameter consistency, eliminating missing value imputation in meteorological data.

Temporal coverage analysis revealed distinct observation periods across installations. Six systems provided continuous data from January 2021 through June 2024, while system 90854 initiated data collection in July 2022. This time distinction was maintained because these specific installation periods denoted legitimate operational timeframes backed by verified data quality. The integration process preserved only timestamps that contained valid observations from both meteorological and PV generation data.

Physical validation procedures confirmed measurement reasonableness. Temperature records (t_{2m}) ranged from 259.4K to 313.2K (-13.8°C to 40°C), consistent with expected seasonal variations in the geographical region. Solar radiation measurements ($ssrd$) exhibited proper diurnal patterns, with maximum values reaching $3.46 \times 10^6 \text{ J/m}^2$ during peak daylight hours. These ranges underwent cross-validation with historical meteorological records. Similarly, hourly energy generation values, with a maximum generation of 896 kWh, aligned with theoretical capacity limits for installations ranging from 100 kW to 1.01 MW, while demonstrating expected diurnal and seasonal variations.

3.5 Final Dataset Structure and Analysis

3.5.1 Dataset Composition

Our final integrated dataset is a complete set of 201,432 hourly observations of seven PV plants, merging generation data with ERA5 meteorological parameters. We achieved complete coverage for all 34 variables, thereby guaranteeing strong analytical ability for temporal and meteorological aspects.

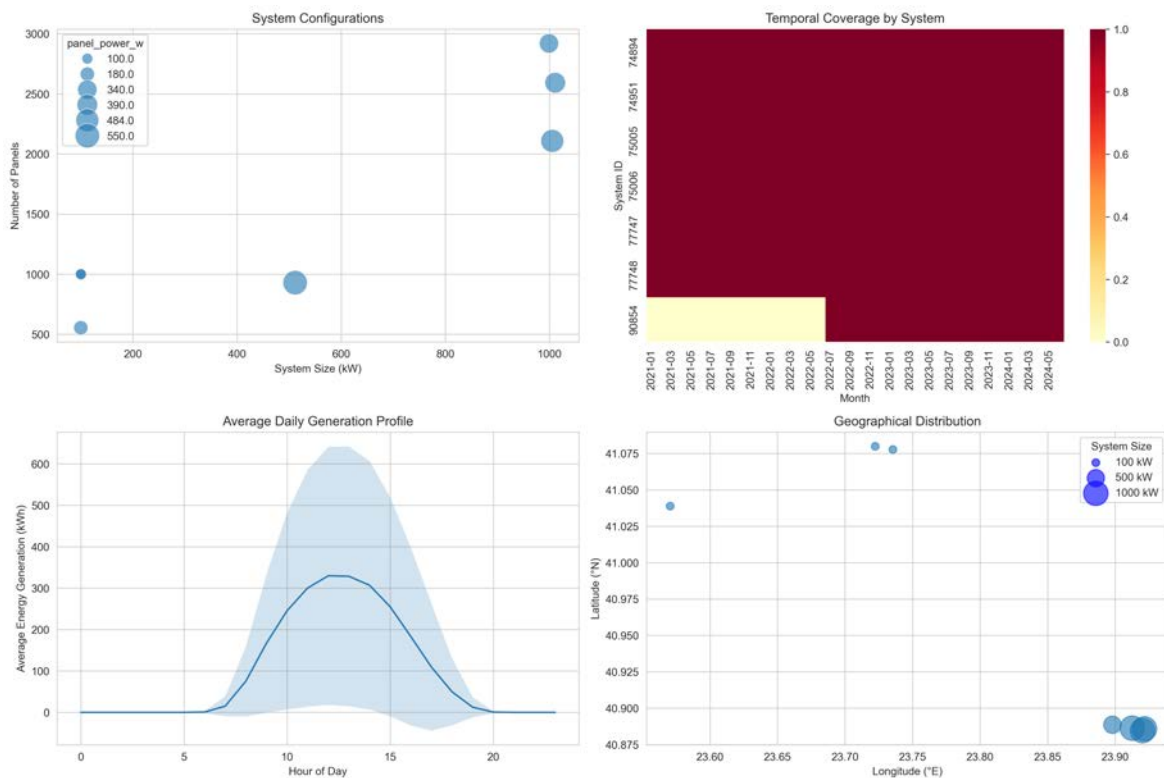


Figure 3.3: Comprehensive characteristics of the dataset include: (a) System configurations that depict the correlation among installation capacities, panel numbers, and power ratings, (b) Temporal coverage indicating data availability throughout the study period, (c) An average daily generation profile with uncertainty bounds that reflect the variability in system performance, and (d) The geographical distribution of installations, with marker sizes corresponding to system capacity.

Our dataset reveals a remarkable diversity in system configuration: the installation capacity ranges between 99.9 kW and 1,010.88 kW, with an average of 549 kW (Figure 3.3a). The panel configurations of the systems are very different: while the number of panels varies between 555 and 2,919 (on average: 1,629 panels), their power rating varies between 100W

and 550W (on average: 290W). This technical diversity provides valuable insights into performance differences under similar environmental conditions.

The temporal coverage analysis (Figure 3.3b) shows reliable data availability between January 2021 and June 2024 for the six installations and a data availability from July 2022 for System 90854. Using the dataset mean timestamp in October 2022, we observe a well-spread temporal representation. The homogeneous dark color in the coverage plot illustrates that the data is complete over the entire measurement period.

Figure 3.3c shows the typical series of daily generation with a very clear bell-shaped profile peaking in the middle of the day. Shaded area shows one standard deviation from the average and it represents the natural variability in generation profiles due to meteorological and seasonal effects. This variability is particularly accentuated during peak generation hours, thereby underscoring the complex interaction between system performance and environmental conditions.

Our installations occupy a relatively compact geographical area in northern Greece (Figure 3.3d), with latitudes between 40.88°N and 41.08°N and longitudes between 23.57°E and 23.92°E. This distribution ensures similar climatic conditions while maintaining sufficient separation to capture local meteorological variations. The marker sizes reflect system capacities, illustrating the spatial distribution of generation capabilities.

Each record contains three categories of variables:

- **System Specifications:**

- Installation capacity: ranging from 99.9 kW to 1,010.88 kW (mean: 549 kW)
- Panel configurations: varying from 555 to 2,919 panels (mean: 1,629 panels)
- Panel power ratings: spanning 100W to 550W (mean: 290W)
- Geographical coordinates: latitude 40.88°N to 41.08°N, longitude 23.57°E to 23.92°E

- **Generation Data:** Hourly energy production in kilowatt-hours, ranging from 0 to 895.95 kWh

- **Meteorological Parameters:** 28 ERA5 variables measuring atmospheric conditions

Our preliminary correlation analysis uncovers strong connections between energy generation and meteorological factors. Surface solar radiation (ssr) exhibits the highest correla-

Variable	Mean	Std Dev	Minimum	Maximum
Energy Generation (kWh)	99.18	202.14	0.00	895.95
Temperature (K)	282.23	6.60	256.67	297.81
Total Cloud Cover	0.47	0.38	0.00	1.00

Table 3.2: Key statistical characteristics of primary dataset variables

tion (0.571) with hourly generation, closely followed by downward surface solar radiation (ssrd, 0.559) and their clear-sky equivalents (ssrdc, 0.541). Moreover, the boundary layer height (blh) shows a significant correlation (0.476), highlighting how atmospheric mixing boosts generation efficiency. Combined with our comprehensive temporal and spatial coverage, these findings provide a strong foundation for investigating the impact of weather on photovoltaic generation.

3.5.2 Variable Analysis and Relationships

Our investigation into the relationships among variables in the integrated dataset uncovers intriguing patterns regarding how various factors affect photovoltaic generation. By employing correlation analysis, assessing feature importance, and clustering weather patterns, we unveil complementary insights that clarify both linear and nonlinear interactions between variables.

Our feature importance analysis, performed with a Random Forest model, highlights a distinct hierarchy of impact. Surface solar radiation (ssr) is identified as the primary factor influencing energy generation, succeeded by geographical location, particularly latitude. System characteristics, especially the number of panels, follow as the third most significant factor, with boundary layer height (blh) in fourth place. This order illustrates how energy generation patterns are shaped by the interaction of environmental factors, geographical elements, and system configurations.

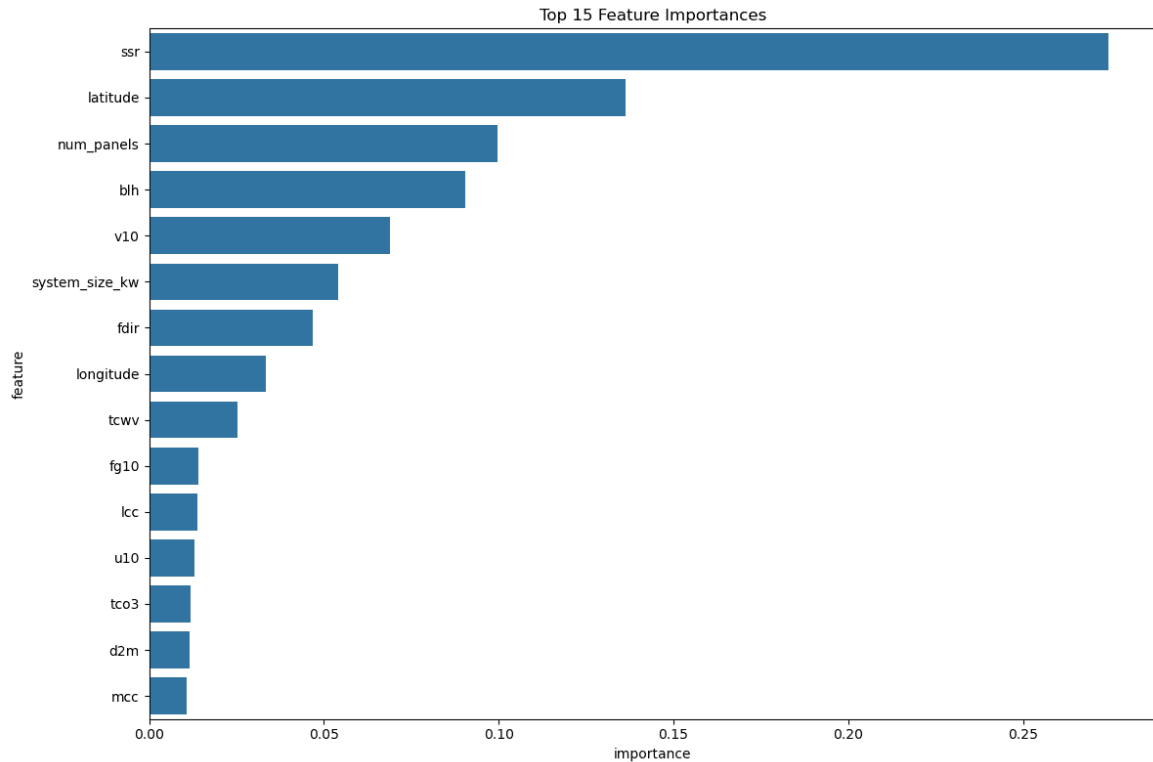


Figure 3.4: Feature importance rankings from Random Forest analysis, demonstrating the dominant influence of solar radiation (ssr) followed by geographical factors (latitude) and system characteristics (num_panels). The analysis reveals the hierarchical nature of influences on PV generation, from environmental conditions to system specifications.

Examining the correlation heatmap unveils complex interactions, especially among meteorological parameters. We observe strong positive correlations between radiation parameters (ssr, ssr_c, ssr_{dc}, ssr_d), highlighting their fundamental role in solar resource characterization. Temperature parameters (t_{2m}, d_{2m}) show moderate correlations with generation, reflecting their influence on panel efficiency, while system specifications demonstrate consistent correlation patterns.

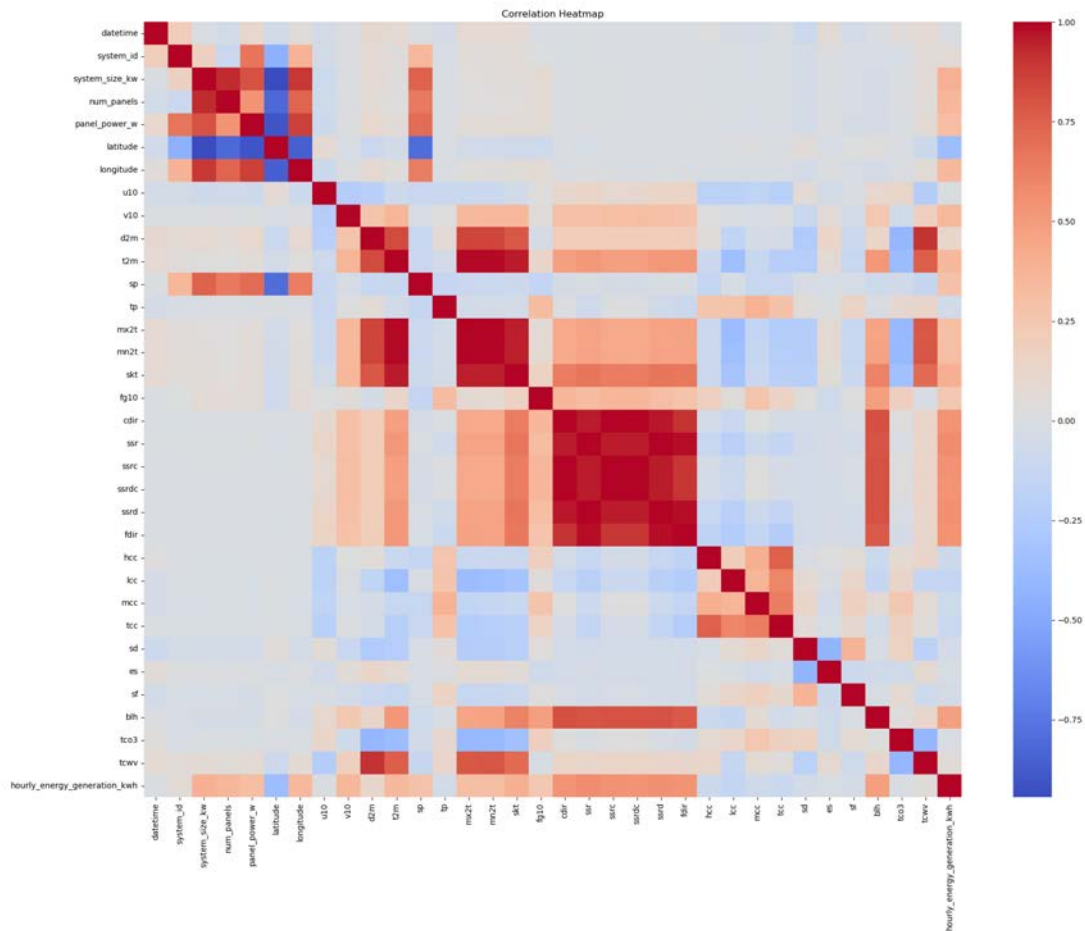


Figure 3.5: Correlation heatmap demonstrating the relationships between dataset variables. Notable patterns include strong correlations among radiation parameters, temperature measurements, and system specifications, revealing the interconnected nature of factors affecting PV generation.

Our weather pattern clustering analysis identified five distinct meteorological regimes, each with unique combinations of environmental conditions:

- Cluster 0: Early morning conditions with high cloud cover (tcc: 0.88), appearing predominantly in winter (41.8%) and fall (36.6%) during pre-dawn hours (02:00-04:00). Shows minimal solar radiation (mean SSRD: 1.04×10^5 J/m²) and lower temperatures.
- Cluster 1: Clear night conditions with very low cloud cover (tcc: 0.18), mainly in winter (59.1%) during early morning hours (02:00-04:00). Despite clear skies, timing and seasonal distribution result in minimal generation.
- Cluster 2: Evening transition periods, dominant in summer (48.8%) and spring (36.8%)

during late hours (17:00-19:00). Shows moderate temperatures (293.62 K) with reduced radiation due to solar angle.

- Cluster 3: Intermediate generation conditions during peak hours (10:00-12:00), featuring moderate cloud cover (0.57) and substantial solar radiation (mean SSRD: 1.48×10^6 J/m²). Winter dominates (43.3%) followed by spring (31.5%).
- Cluster 4: Optimal generation conditions during peak hours (10:00-12:00), characterized by low cloud cover (0.33) and maximum solar radiation (mean SSRD: 2.24×10^6 J/m²). Appears predominantly in summer (51.1%) and spring (43.2%).

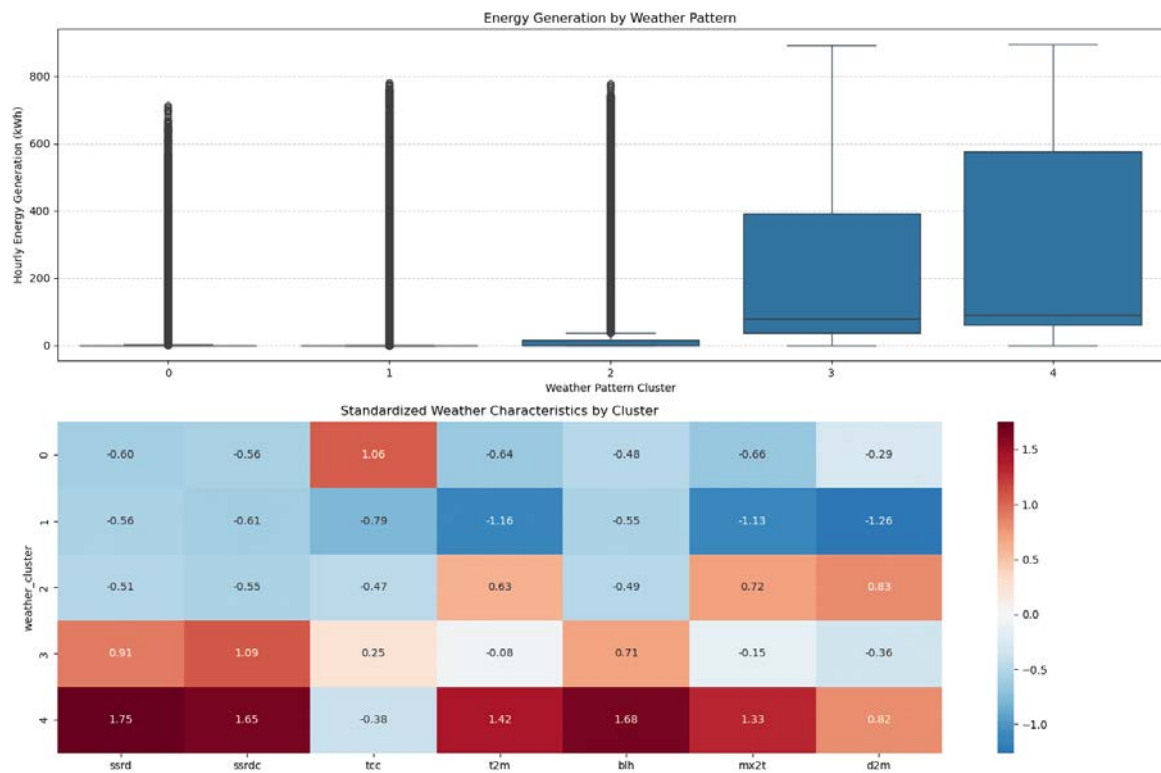


Figure 3.6: Energy generation distribution across identified weather patterns, showing distinct generation profiles for each meteorological regime. The lower panel displays standardized weather characteristics by cluster, demonstrating how combinations of meteorological parameters create these distinct generation patterns.

These findings reveal that generation patterns emerge not from individual weather parameters, but from specific combinations of meteorological conditions varying systematically with time of day and season. This understanding provides valuable context for generation forecasting and system performance optimization.

3.5.3 Temporal Characteristics

Our temporal analysis of the integrated dataset reveals fascinating yet distinct patterns across multiple timescales. These patterns demonstrate the interplay between fundamental solar physics and the modulating effects of atmospheric conditions throughout the year.

Our analysis of daily generation profiles shows a distinct bell-shaped curve influenced by weather conditions and system specifications. The highest generation occurs between 10:00 and 12:00 local time, as indicated by the weather cluster 4 distribution. Notably, we see an interesting asymmetry in generation transitions: morning increases are more pronounced than the declines in the evening, which can be linked to atmospheric factors like boundary layer development and temperature gradients.

Seasonal variations emerge through several mechanisms. Changes in solar geometry drive the most pronounced effects, yielding longer days and enhanced generation potential during summer months. This pattern manifests clearly in our weather cluster distribution, with optimal generation conditions (cluster 4) occurring predominantly in summer (51.1%) and spring (43.2%). In contrast, winter months exhibit higher frequencies of clusters 0 and 1, characterized by reduced generation potential even during daylight hours, with notable winter occurrence rates of 41.8% and 59.1%, respectively.

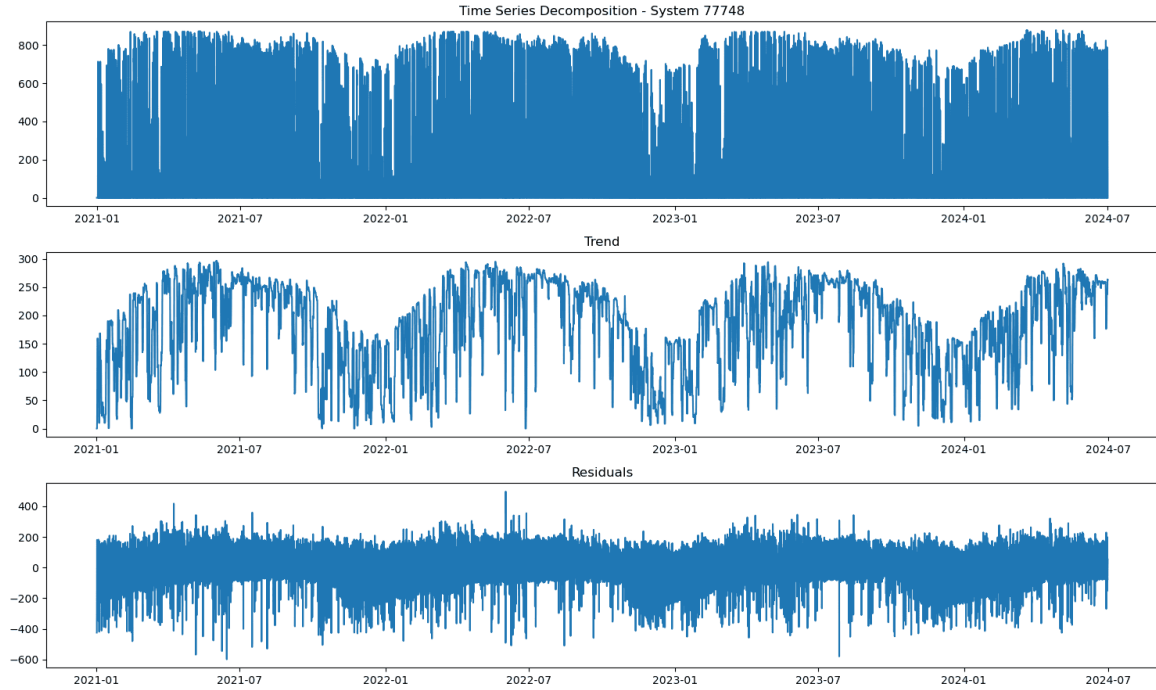


Figure 3.7: Time series decomposition for System 77748, showing (a) observed generation patterns, (b) underlying trend component capturing seasonal and long-term variations, and (c) residual variations highlighting short-term weather effects. The decomposition reveals the multi-scale temporal structure of PV generation patterns.

Individual systems exhibit characteristic temporal behaviors influenced by their specific configurations and local conditions. Our time series decomposition reveals gradual changes in system performance that may be attributed to factors such as panel degradation or systematic environmental changes. These long-term patterns overlay the more pronounced seasonal and daily cycles, creating system-specific temporal signatures.

The residual component of our decomposition captures short-term variations primarily driven by weather conditions. These deviations demonstrate strong correlation with the identified weather clusters, particularly during primary generation hours. Notably, our analysis reveals that the largest deviations occur during optimal generation conditions (cluster 4), suggesting that clear-sky periods paradoxically exhibit the highest variability between actual and expected generation.

The boundary layer height (blh), which emerged as significant in our feature analysis, demonstrates notable daily cycles aligning with generation patterns during clear conditions. This alignment reveals the complex coupling between atmospheric dynamics and PV system performance, where atmospheric mixing influences both radiation transmission and panel

cooling efficiency.

While our dataset's temporal span (January 2021 through June 2024) constrains multi-year trend analysis, the consistent patterns observed across multiple annual cycles suggest robust underlying relationships between meteorological conditions and generation performance.

3.6 Dataset Considerations

3.6.1 Quality and Coverage Assessment

Our integration of high-frequency PV generation data with ERA5 meteorological parameters demands a thorough examination of data quality. Our comprehensive analysis reveals exceptional completeness and consistency across multiple dimensions.

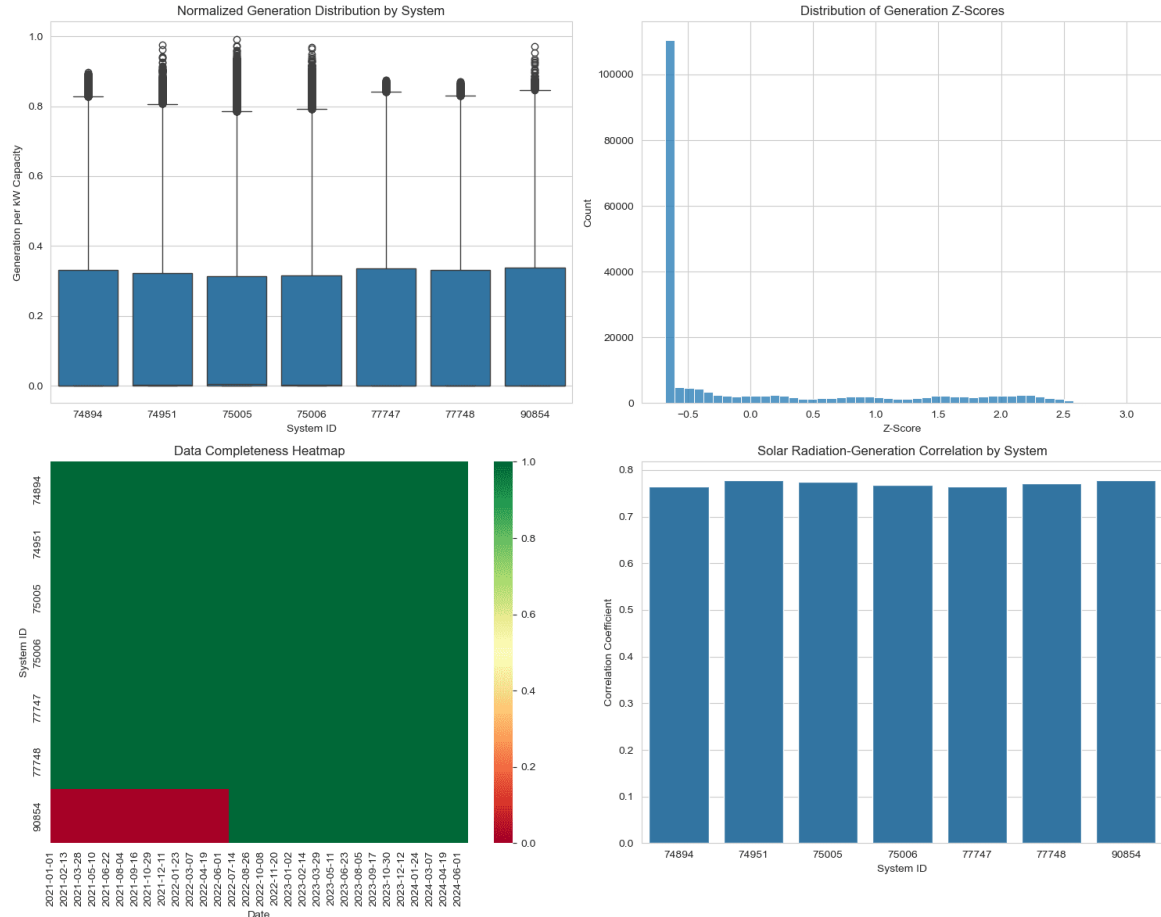


Figure 3.8: Dataset quality assessment showing: (a) Normalized generation distribution demonstrating consistent performance patterns across systems, (b) Generation z-score distribution indicating robust outlier characteristics, (c) Temporal completeness heatmap confirming data availability, and (d) System-specific solar radiation correlations demonstrating consistent weather response patterns.

Our system-level analysis reveals a remarkable consistency in normalized generation patterns across installations (Figure 3.8a). When accounting for system capacities, we observe uniform median values and interquartile ranges, suggesting robust performance characteristics across diverse configurations.

Through z-score analysis, we discovered an exceptionally clean dataset with notable outliers in only two systems. Systems 74951 and 75005 exhibited interesting anomalies during midday hours (13:00-14:00) in spring months. System 75005 showed three significant cases in April 2021 (z-scores > 3.0 , generation 97-99 kWh) under varying solar radiation (2.07×10^6 to 2.57×10^6 J/m²). System 74951 displayed two similar deviations. Intriguingly, these outliers represent positive deviations during high radiation periods, suggesting

exceptional performance rather than data errors.

Our temporal completeness analysis revealed fascinating intricacies in handling daylight saving time transitions. Most systems show a completeness ratio of 1.0008, reflecting our careful preservation of temporal information during DST transitions. This slight excess above unity stems from our explicit handling of fall transitions, where we maintain both timestamps during clock setbacks. System 90854, beginning in July 2022, shows a slightly higher ratio (1.0014) due to proportionally more DST transitions in its shorter operational period.

The relationship between solar radiation and generation demonstrates remarkable consistency, with correlation coefficients tightly clustered between 0.764 and 0.778. Peak hour performance stability remains consistent, with coefficients of variation ranging from 0.426 to 0.465.

3.6.2 Usage Implications

Our quality assessment reveals several crucial considerations for researchers:

- System-size normalization effectively addresses installation capacity variations, though slight differences in peak hour coefficients (0.426 to 0.465) warrant attention in prediction modeling
- The consistent radiation correlations (0.764-0.778) suggest reliable weather-based modeling across installations
- Outlier characteristics deserve special consideration:
 - Their concentration in specific temporal periods suggests blanket removal strategies may be inappropriate
 - Association with high radiation periods indicates potential exceptional performance
 - Their rarity (5 instances in 201,432 observations, 0.0025%) supports overall data reliability
- Methodological implications include:
 - Standard statistical techniques suffice given low outlier rates
 - No need for missing data imputation

- Weather-based models should transfer effectively between systems
- Different modeling approaches may benefit peak generation periods

These findings establish a robust foundation for diverse analytical approaches while highlighting important considerations for research design. The demonstrated consistency across quality dimensions supports both system-specific and comparative analyses, though researchers should remain mindful of peak generation stability variations.

Chapter 4

Methodology

4.1 Introduction

This chapter introduces a general methodology for the prediction and optimization of the production of photovoltaic farms based on the data set presented in Chapter 3. In this approach, Greek photovoltaic plant high-frequency operational data is blended with ERA5 reanalysis meteorological parameters for constructing accurate prediction models.

The dataset of 201,432 hourly observations from seven photovoltaic installations presents several key methodological challenges:

- Effectively aligning 5-minute PV generation data (aggregated hourly) with ERA5 meteorological parameters
- Integrating 26 ERA5 meteorological parameters with system-specific characteristics
- Developing features that capture weather patterns and seasonal variations
- Creating models that generalize across systems ranging from 100 kW to 1.01 MW

Building on patterns identified in Chapter 3, the methodology is structured around four components: feature engineering, model development, validation strategy, and performance evaluation. The feature engineering expands the initial dataset integration by capturing temporal and meteorological patterns. The model development implements machine learning approaches explicitly selected to handle the complex relationships revealed in the correlation analysis.

The methodology follows an efficiency-based approach that predicts system efficiency (kWh/kW) rather than absolute power output. This maintains consistency with observed patterns while enabling better generalization across diverse system sizes.

4.2 Feature Engineering

The feature engineering process builds upon the integrated dataset described in Chapter 3, creating additional features and organizing them into logical groups to support effective model development. This section examines the existing features preserved from the original dataset and the new features engineered to capture temporal patterns and system performance characteristics.

4.2.1 Feature Categories

Static System Features

These fundamental features, preserved from the original dataset, represent the unchanging characteristics of each installation:

Table 4.1: Static System Features

Feature	Description
system_id	Unique identifier for each installation
system_size_kw	Installation capacity in kilowatts
num_panels	Total number of panels in the installation
panel_power_w	Power rating of individual panels in watts
latitude	Installation latitude coordinate
longitude	Installation longitude coordinate

Temporal Features

The datetime information yields new features that capture temporal patterns, incorporating basic temporal components and their cyclical transformations.

The cyclical nature of temporal features necessitates these trigonometric transformations:

$$feature_{sin} = \sin(2\pi \frac{value}{period}) \quad (4.1)$$

$$feature_{cos} = \cos(2\pi \frac{value}{period}) \quad (4.2)$$

Table 4.2: Temporal Feature Components

Category	Feature	Range/Description
Basic	hour	0-23
	month	1-12
	year	Observation year
	day_of_week	0-6
	day_of_year	1-365
Cyclical	hour_sin, hour_cos	period = 24
	month_sin, month_cos	period = 12
	day_of_year_sin, day_of_year_cos	period = 365
Seasonal	season	0: Winter, 1: Spring, 2: Summer, 3: Fall

Meteorological Features

The ERA5 dataset described in Chapter 3 provides these meteorological features, ensuring a comprehensive environmental context:

Table 4.3: ERA5 Meteorological Features

Category	Parameter	Description
Radiation	ssrd	Surface solar radiation downwards
	ssrdc	Surface solar radiation downwards clear-sky
	fdir	Total sky direct solar radiation at surface
	cdir	Clear-sky direct solar radiation at surface
	ssr	Surface net solar radiation
	ssrc	Surface net clear-sky solar radiation
Temperature	t2m	2-meter temperature
	mx2t	Maximum 2-meter temperature
	mn2t	Minimum 2-meter temperature
	d2m	2-meter dewpoint temperature
	skt	Skin temperature
Wind	u10	10-meter U wind component
	v10	10-meter V wind component
	fg10	10-meter wind gust
Cloud Coverage	tcc	Total cloud cover
	hcc	High cloud cover
	mcc	Medium cloud cover
	lcc	Low cloud cover
Atmospheric	sp	Surface pressure
	blh	Boundary layer height
	tcwv	Total column water vapour
	tco3	Total column ozone

4.2.2 Feature Transformation and Organization

Feature Grouping

Features are organized into four logical groups to support model development:

Table 4.4: Feature Organization by Type

Category	Features
static_categoricals	system_id
static_reals	system_size_kw, num_panels, panel_power_w, latitude, longitude
time_varying_known_reals	hour, month, day_of_year, season, day_of_week, hour_sin, hour_cos, month_sin, month_cos, day_of_year_sin, day_of_year_cos
time_varying_unknown_reals	All ERA5 parameters

This grouping structure supports both traditional machine learning models and the specialized requirements of the Temporal Fusion Transformer architecture.

Standardization

Numerical features undergo standardization via the Standard Scaler:

$$x_{scaled} = \frac{x - \mu}{\sigma} \quad (4.3)$$

Here, μ and σ are determined from the training data to avoid data leakage. This standardization is applied to:

- All ERA5 meteorological parameters
- System specifications (system_size_kw, panel_power_w)
- Generated cyclical features

Target Variable

The main target variable for model development is system efficiency, which is calculated as follows:

$$efficiency = \frac{hourly_energy_generation_kwh}{system_size_kw} \quad (4.4)$$

This efficiency-focused method normalizes energy generation across varying system sizes, ensuring interpretability and allowing absolute generation predictions through post-processing.

4.3 Model Development

4.3.1 Baseline Model

The implementation uses a seasonal naive model, taking advantage of a basic performance benchmark that reflects the daily periodicity in photovoltaic generation patterns. This foundational model offers a context for assessing more advanced methods while still recognizing essential temporal trends in the data.

Model Definition

The seasonal naive model bases predictions on the principle that efficiency at any given hour matches the efficiency observed at the same hour on the previous day:

$$\hat{y}_{h,d} = y_{h,d-1} \quad (4.5)$$

where $\hat{y}_{h,d}$ represents the predicted efficiency for hour h on day d , and $y_{h,d-1}$ represents the observed efficiency for the same hour on the previous day.

Implementation

The baseline implementation follows three sequential steps:

1. Data Preparation:

- Concatenation of training and test data ensures temporal continuity
- Chronological sorting by datetime maintains temporal order

2. Prediction Generation:

- Grouping observations by hour of day (0-23)
- Applying 24-hour shifts within each hourly group

3. Post-processing:

- Filling missing values with zeros
- Enforcing non-negative constraints:

$$\hat{y}_{final} = \max(0, \hat{y}) \quad (4.6)$$

Justification and Limitations

Several characteristics of photovoltaic generation justify this baseline approach:

- Strong daily cyclical patterns in solar generation
- Minimal computational requirements
- Interpretable predictions
- Direct efficiency comparisons across different system sizes

The approach carries inherent limitations:

- No consideration of weather variations between consecutive days
- Inability to capture seasonal trends
- No incorporation of system-specific characteristics
- Limited adaptability to irregular patterns

These limitations establish clear opportunities for more sophisticated modeling approaches. However, these approaches must demonstrate significant performance gains to justify their additional complexity.

4.3.2 Tree-based Ensemble Models

Random Forest Implementation

Random Forest is an ensemble learning technique that works by creating various decision trees and aggregating their predictions [17, 18]. Essentially, the algorithm generates each tree from a unique bootstrap sample of the training data, a method referred to as bagging (bootstrap aggregating). Combined with random feature selection at each split, this sampling technique introduces randomness that helps prevent overfitting and improves the model's generalization capability.

Every decision tree within the forest learns to divide the feature space into distinct regions, forming a hierarchy of choices based on feature values. These choices may include splits on factors like solar radiation levels, temperature readings, or time-of-day variables, specifically

for predicting photovoltaic efficiency. The ensemble's final prediction results from averaging the outputs of all the trees.

$$\hat{y} = \frac{1}{T} \sum_{t=1}^T f_t(x) \quad (4.7)$$

where T is the total number of trees, $f_t(x)$ represents the prediction of the t -th tree, and $\hat{y}$ is the ensemble's final prediction.

Our implementation utilizes scikit-learn's RandomForestRegressor with Bayesian optimization for hyperparameter tuning. The hyperparameter space encompasses four key parameters:

Table 4.5: Random Forest Hyperparameter Ranges and Descriptions

Parameter	Range	Description
n_estimators	[100, 1000]	Controls the number of trees in the forest. A larger number of trees generally improves model robustness by reducing variance, though with diminishing returns beyond a certain point.
max_depth	[3, 20]	Limits the maximum depth of individual trees. Deeper trees can capture more complex patterns but risk overfitting to noise in the training data.
min_samples_split	[2, 10]	Specifies the minimum number of samples required to split an internal node. Larger values prevent the creation of nodes with too few samples, reducing overfitting.
min_samples_leaf	[1, 10]	Defines the minimum number of samples required at each leaf node. This constraint ensures that predictions are based on sufficient training examples.

The Bayesian optimization method effectively investigates the parameter space through a series of sequential trials. Beginning with five randomly selected configurations, the method performs five iterations of optimization to identify the best parameter combination. This procedure aims to reduce the Root Mean Square Error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (4.8)$$

where y_i represents the actual efficiency values and $\hat{y}_i$ represents the predicted values. This metric provides a balanced prediction accuracy assessment by penalizing large and small deviations from actual values.

The implementation includes several key technical considerations. Feature scaling through `StandardScaler` ensures that all input features contribute appropriately to the splitting decisions. Predictions are maintained under non-negative constraints to ensure physically meaningful results, and the model preserves the capability to convert between efficiency values and absolute power generation through system size scaling.

XGBoost Implementation

XGBoost (eXtreme Gradient Boosting) extends traditional gradient boosting by incorporating second-order gradients and advanced regularization techniques [19]. Unlike Random Forest's parallel tree construction, XGBoost builds trees sequentially, with each tree learning to correct the errors of the previous ensemble.

XGBoost, standing for eXtreme Gradient Boosting, takes the concept of traditional gradient boosting to the next level by adding in second-order gradients and some smart regularization techniques [19]. Instead of building trees in parallel like Random Forest does, XGBoost creates them one after the other, where each new tree is designed to mend the mistakes made by its predecessor.

The model constructs each tree by optimizing the objective:

$$\mathcal{L} = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^K \Omega(f_k) \quad (4.9)$$

where l represents the loss function, Ω is the regularization term, and f_k represents the k -th tree in the ensemble.

Our implementation explores a comprehensive hyperparameter space:

Table 4.6: XGBoost Hyperparameter Configuration

Category	Parameter	Range & Description
Core Learning	max_depth	[3, 10] Controls tree growth and model complexity
	learning_rate	[0.01, 0.3] Step size shrinkage to prevent overfitting
	n_estimators	[100, 1000] Number of boosting rounds
Sampling	colsample_bytree	[0.5, 1] Fraction of features to consider for each tree
	subsample	[0.5, 1] Fraction of samples used for tree building
Regularization	min_child_weight	[1, 10] Minimum sum of instance weight in child node
	gamma	[0, 5] Minimum loss reduction for node splitting
	max_delta_step	[0, 10] Maximum delta step for tree weight estimation
	reg_lambda	[0.1, 10] L2 regularization term on weights
	reg_alpha	[0, 10] L1 regularization term for feature selection

LightGBM Implementation

LightGBM introduces a novel approach to gradient boosting through its leaf-wise tree growth strategy [20]. Unlike traditional algorithms such as XGBoost that construct trees

level-by-level, LightGBM determines and splits the leaf that produces the greatest reduction in loss, resulting in asymmetric trees. This strategy generally leads to quicker training and improved management of sparse features, which is especially beneficial for our temporal dataset, where certain patterns might be infrequent over time.

The algorithm optimizes the following objective function during tree construction:

$$\text{Obj} = \sum_{i=1}^n L(y_i, \hat{y}_i) + \sum_{j=1}^T \Omega(f_j) \quad (4.10)$$

where L represents the training loss for each instance, Ω penalizes model complexity, and f_j represents the j -th tree. The leaf-wise growth strategy specifically targets the split that maximizes the gain:

$$\text{Gain} = \frac{1}{2} \left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right] \quad (4.11)$$

where G and H represent first and second order gradients respectively, and subscripts L and R denote left and right child nodes.

Our implementation defines a comprehensive hyperparameter space:

Table 4.7: LightGBM Hyperparameter Configuration

Category	Parameter	Range & Description
Core	max_depth	[3, 10] Maximum tree depth, though less critical due to leaf-wise growth
	learning_rate	[0.01, 0.3] Controls the contribution of each tree
	num_leaves	[20, 100] Maximum number of leaves per tree, key for model complexity
Sampling & Feature	feature_fraction	[0.5, 1] Controls feature sampling for each tree
	bagging_fraction	[0.5, 1] Fraction of data to use for each iteration
Regularization	min_child_samples	[1, 20] Minimum number of samples per leaf
	min_child_weight	[0.001, 10] Minimum sum of instance weight per leaf
	min_split_gain	[0, 1] Minimum gain needed for a split
	reg_alpha	[0, 10] L1 regularization
	reg_lambda	[0, 10] L2 regularization

The optimization process employs the same Bayesian approach used in previous models but with specific consideration for LightGBM's unique characteristics. The objective function remains the minimization of RMSE, but the parameter space exploration is mainly focused on the interplay between num_leaves and other regularization parameters, as this balance is crucial for LightGBM's leaf-wise growth strategy.

Feature preprocessing follows our established pipeline with StandardScaler while maintaining non-negative prediction constraints. The model preserves the ability to convert be-

tween efficiency values and absolute power generation, ensuring consistency across our modeling framework.

CatBoost Implementation

CatBoost introduces several improvements to gradient boosting, the most significant one of which being its ordered boosting procedure, which minimizes the prediction drift by using different permutations of the training samples [21]. This method solves a core defect in gradient boosting, where the targets for different observations are calculated using various models, resulting in statistical bias that accumulates during training.

The algorithm implements a unique gradient estimation process that can be expressed as:

$$\hat{g}_i = \frac{\partial L(y_i, s)}{\partial s} \Big|_{s=\hat{F}^{i-1}(x_i)} \quad (4.12)$$

where $\hat{g}_i$ represents the gradient estimate for the i -th example, L is the loss function, and $\hat{F}^{i-1}$ is the model constructed using only the examples that come before the i -th example in a random permutation.

For the photovoltaic efficiency prediction task, the hyperparameter space is defined with careful consideration of CatBoost's ordered boosting characteristics:

Table 4.8: CatBoost Hyperparameter Configuration

Category	Parameter	Range & Description
Core	iterations	[100, 1000] Number of boosting iterations, balanced for model performance and training time
	learning_rate	[0.01, 0.3] Step size for gradient updates, crucial for convergence quality
	depth	[3, 10] Maximum tree depth, controlling model complexity while maintaining interpretability
Learning Control	rsm	[0.5, 1] Random subspace method, controlling feature sampling for each split
	random_strength	[0, 1] Amount of randomness in split scoring, helping prevent overfitting
Regularization	l2_leaf_reg	[1, 10] L2 regularization coefficient for leaf values
	bagging_temperature	[0, 1] Controls, the intensity of bagging, affecting model robustness

The training process incorporates several key technical considerations specific to CatBoost. The model employs an oblivious tree structure where the same splitting criterion is used across an entire tree level, which helps prevent overfitting and enables faster inference. Additionally, the implementation uses CPU-based computation with the RMSE loss function:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (4.13)$$

The optimization utilizes Bayesian methods, starting with 10 initial points and 50 iterations to effectively explore the parameter space while optimizing computational efficiency. An early stopping criterion is applied, with a patience of 10 rounds, tracking the evaluation metric on a validation set to avoid overfitting.

As with our other implementations, the model maintains consistent preprocessing through `StandardScaler` and preserves non-negative prediction constraints. This ensures compatibility with our broader modeling framework while taking advantage of CatBoost's unique strengths in handling ordered boosting and preventing prediction shifts.

Model Training and Optimization

The training and optimization algorithm maintains a uniform framework among all ensemble models their respective properties. This single framework allows for an equitable comparison of models and makes it possible for each model to maximize its own strengths.

Bayesian optimization forms the core of the optimization strategy, which can effectively search the high-dimensional hyperparameter spaces. This approach models the relationship between hyperparameters and model performance using a Gaussian process, selecting new configurations that balance exploring unknown regions with exploiting known suitable parameters. The acquisition function employs Expected Improvement (EI), expressed as:

$$EI(x) = \mathbb{E} [\max(f(x) - f^+ - \xi, 0)] \quad (4.14)$$

where f^+ represents the best observed value, and ξ controls the exploration-exploitation trade-off.

Each model's optimization process begins with an initial exploration phase using 10 randomly selected configurations, followed by 50 sequential optimization steps. This balance ensures a thorough exploration of the parameter space while maintaining practical computational constraints. The objective function targets the minimization of Root Mean Square Error:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (4.15)$$

Early stopping with a patience of 10 epochs prevents overfitting during training. This mechanism monitors validation set performance and halts training when no improvement occurs, preserving the best-performing model configuration. The validation strategy splits training data chronologically, maintaining temporal coherence in the evaluation process.

Feature preprocessing employs StandardScaler to normalize input features:

$$x_{\text{scaled}} = \frac{x - \mu}{\sigma} \quad (4.16)$$

where μ and σ are calculated from training data to prevent data leakage, this scaling ensures proportional feature contribution to model decisions, particularly important for tree-based models where split decisions could otherwise be biased by feature scales.

The training process incorporates several key technical considerations:

- Non-negative Constraints:

All models enforce non-negative predictions, maintaining physical meaningfulness:

$$\hat{y}_{\text{final}} = \max(0, \hat{y}_{\text{predicted}}) \quad (4.17)$$

- Reproducibility:

Fixed random seeds (42) ensure reproducible results across training runs

- Computational Efficiency:

Parallel processing utilization where applicable, particularly in Random Forest and during hyperparameter optimization

Model predictions are carefully handled to maintain consistency with our efficiency-based approach. The primary output of each model is the predicted efficiency, which can be converted to absolute power generation through system size scaling when required:

$$P_{\text{predicted}} = \eta_{\text{predicted}} \cdot S_{\text{capacity}} \quad (4.18)$$

where $P_{\text{predicted}}$ represents the predicted power output in kilowatt-hours, $\eta_{\text{predicted}}$ is the predicted efficiency, and S_{capacity} is the system capacity in kilowatts.

4.3.3 Deep Learning Approach

Temporal Fusion Transformer

The Temporal Fusion Transformer (TFT) marks a notable shift from tree-based models, featuring an advanced architecture tailored for temporal sequence modeling [22, 23, 24]. Unlike traditional approaches that treat time points in isolation, TFT explicitly addresses temporal dependencies using recurrent layers combined with self-attention mechanisms. This

design is particularly effective for predicting photovoltaic generation, as it captures not only brief weather influences but also long-term seasonal trends while factoring in the static characteristics of the system.

TFT's key innovation is found in its variable selection networks coupled with an interpretable multi-head attention mechanism. These components enable the flexible adjustment of the importance of input variables over different prediction horizons, which is crucial for understanding the intricate connections between weather factors and the performance of photovoltaic systems. With 1.1 million trainable parameters, the model achieves a balance between the complexity required to capture detailed temporal patterns and computational effectiveness.

Architecture and Implementation

The TFT architecture implements a hierarchical processing pipeline beginning with specialized input-handling networks. A MultiEmbedding layer processes categorical variables, while three variable selection networks manage distinct aspects of input data. The static network (26.2K parameters) processes system characteristics, while encoder and decoder networks (151K and 147K parameters) manage temporal features, enabling dynamic feature utilization across prediction horizons.

Static information flows through four identical Gated Residual Networks (GRN), each containing 66.3K parameters, serving context variable selection, LSTM state initialization, and static enrichment. The GRN architecture implements gated skip connections:

$$\text{GRN}(x) = \text{LayerNorm}(x + \text{GLU}(W_2\sigma(W_1x + b_1) + b_2)) \quad (4.19)$$

The temporal processing core consists of dual LSTM networks, each containing 132K parameters, which capture sequential patterns in the data. These recurrent networks are augmented by post-processing components, including a Gated Linear Unit (33K parameters) and normalization layers (256 parameters). The temporal processing implements the standard LSTM update equations:

$$h_t, c_t = \text{LSTM}(x_t, h_{t-1}, c_{t-1}) \quad (4.20)$$

An Interpretable Multi-Head Attention module (37.1K parameters) processes temporal relationships through 8 parallel attention heads:

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V \quad (4.21)$$

Training Configuration

The training configuration balances model performance with computational efficiency. The architecture employs 128-unit hidden dimensions and 8 attention heads for temporal processing. Dropout (0.1), learning rate (1e-3), and weight decay (1e-4) maintain stable optimization dynamics.

PyTorch Lightning enables distributed GPU training, processing 256-sample batches with 4-step gradient accumulation. Training spans a maximum of 10 epochs with 3-epoch patience early stopping. Gradient clipping (1.0) ensures stable updates.

The temporal processing considers 168-hour context windows for 24-hour predictions, incorporating 24, 48, and 168-hour target lags. Feature processing treats system identifiers as categorical variables, while specifications and coordinates form static features. Temporal features divide into known (cyclical time encodings) and unknown (meteorological) parameters, with group normalization applied per system.

Mean Absolute Error is the training objective chosen for outlier robustness and interpretability. AdamW optimization with AMSGrad provides adaptive learning rates with effective regularization, navigating the loss landscape while preventing overfitting.

4.4 Validation Framework

4.4.1 Validation Strategies

Validating machine learning models for predicting photovoltaic power necessitates thorough examination of temporal dependencies and unique system features. Two complementary validation strategies tackle various dimensions of this prediction challenge, offering a thorough evaluation across a range of real-world conditions.

Temporal Validation Design

The temporal validation strategy employs a forward-chronological split, reflecting practical deployment scenarios. Designating 2024 data as the test set while training on previous

years creates a framework mirroring real-world applications in which models predict future performance based solely on historical data.

The temporal validation process follows a strict chronological separation defined mathematically as:

$$\mathcal{D}_{train} = \{(x_t, y_t) | t < t_{split}\} \quad (4.22)$$

$$\mathcal{D}_{test} = \{(x_t, y_t) | t \geq t_{split}\} \quad (4.23)$$

where t_{split} represents January 1, 2024, x_t comprises the feature vector at time t , and y_t represents the target efficiency value. This temporal segregation avoids data leakage and maintains temporal continuity, it retains seasonal patterns and long-term trends that are likely to be broken by the randomized split.

This method implements a stringent performance evaluation that necessitates authentic forecasting skills instead of simply interpolating between established data points. Models are required to adapt to new timeframes, facing unfamiliar weather patterns and seasonal impacts that were not present in the training data.

System-based Validation Design

System-based validation evaluates model generalization across different photovoltaic installations by holding out complete systems for testing while training on the remaining installations. This validation approach is formally expressed as:

$$\mathcal{D}_{train} = \{(x_i, y_i) | s_i \in \mathcal{S}_{train}\} \quad (4.24)$$

$$\mathcal{D}_{test} = \{(x_i, y_i) | s_i \in \mathcal{S}_{test}\} \quad (4.25)$$

where s_i represents system identifier, and $\mathcal{S}_{train}$ and $\mathcal{S}_{test}$ are disjoint sets of system identifiers. This design tests generalization across different configurations, locations, and installation characteristics.

The challenge is adapting to entirely new system characteristics not seen during training. Success indicates that models learn generalizable relationships between inputs and performance rather than memorizing system-specific patterns, which is crucial for predicting new installations or systems with limited historical data.

Together, temporal and system-based validation provide a comprehensive assessment. While temporal validation ensures future prediction capability, system-based validation confirms generalization across installations. This combined framework offers strong performance guarantees under diverse conditions and identifies robust and reliable prediction approaches.

4.4.2 Performance Metrics

A complete set of performance metrics needs to be considered to properly evaluate PV power prediction models, accounting for the precision of efficiency predictions and the ability to transform them into actual power generation values. Our evaluation framework uses various complementary metrics to comprehensively evaluate model performance when considered across different aspects of the prediction task.

Efficiency Metrics

The primary evaluation focuses on the models' ability to predict system efficiency: the ratio of energy generation to system capacity. Root Mean Square Error (RMSE) serves as our primary efficiency metric, providing a quadratic penalty to prediction errors:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (\eta_i - \hat{\eta}_i)^2} \quad (4.26)$$

where η_i represents the actual efficiency and $\hat{\eta}_i$ the predicted efficiency for observation i . This metric is particularly valuable as it emphasizes larger errors, especially important in power generation forecasting.

RMSE is complemented with Mean Absolute Error (MAE) to provide a linear scale assessment of prediction accuracy:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |\eta_i - \hat{\eta}_i| \quad (4.27)$$

The combination of RMSE and MAE shows the distribution of prediction errors. The relationship between the two metrics reveals whether errors are uniform (when RMSE and MAE are comparable) or tend to include large deviations (when RMSE is much bigger than MAE).

Energy Generation Metrics

Energy generation metrics evaluate predictions in absolute kilowatt-hours, converting from efficiency through:

$$P_i = \eta_i \cdot S_c \quad (4.28)$$

where P_i represents power generation and S_c is the system capacity. Here is the calculation of RMSE and MAE for absolute generation:

$$\text{RMSE}_{kWh} = \sqrt{\frac{1}{n} \sum_{i=1}^n (P_i - \hat{P}_i)^2} \quad (4.29)$$

$$\text{MAE}_{kWh} = \frac{1}{n} \sum_{i=1}^n |P_i - \hat{P}_i| \quad (4.30)$$

Additionally, the coefficient of determination (R^2) assists in assessing the proportion of variance explained by the models:

$$R^2 = 1 - \frac{\sum_{i=1}^n (P_i - \hat{P}_i)^2}{\sum_{i=1}^n (P_i - \bar{P})^2} \quad (4.31)$$

where $\bar{P}$ represents the mean observed power generation.

Error Analysis Methods

Our error analysis framework uses both quantitative measures and visualization tools to gain a deep understanding of model performance. The most important error computation compares predicted and actual values:

$$e_t = y_t - \hat{y}_t \quad (4.32)$$

where e_t represents the prediction error at time t , y_t is the actual value, and $\hat{y}_t$ is the predicted value.

Multiple approaches analyze these errors: Temporal visualization is sensitive to systematic deviations and biases, while residual analysis via scatter plots detects possible heteroscedasticity. Histogram plots analyze residual distributions, highlighting outliers and anomalous behavior. This analysis spans both prediction methodologies, thus achieving performance understanding at both measurement scales.

4.4.3 Model Comparison Framework

A comprehensive framework that enables fair and insightful evaluation across diverse architectures is required to systematically compare different modeling approaches. Our framework addresses the challenges of comparing traditional tree-based ensemble methods with deep learning approaches while maintaining consistent evaluation criteria across all models.

Cross-model Evaluation

Cross-model evaluation at the same time poses interesting challenges for comparing models based on fundamentally different architectures that range from Random Forest all the way to Temporal Fusion Transformer. Through standardized preprocessing and evaluation procedures, our approach guarantees fair comparison. All models work on the same feature sets and are evaluated from the same temporal/systematic train-test splits.

The seasonal naive approach is used to establish a baseline, providing a fundamental reference point for model performance assessment. The relative improvement over this baseline is calculated for each model as follows:

$$\text{Improvement}_{\text{RMSE}} = \frac{\text{RMSE}_{\text{baseline}} - \text{RMSE}_{\text{model}}}{\text{RMSE}_{\text{baseline}}} \times 100\% \quad (4.33)$$

This standardized comparison reveals not only absolute performance differences but also the relative strengths of each modeling approach. The tree-based models (Random Forest, XGBoost, LightGBM, and CatBoost) share standard evaluation metrics, while the Temporal Fusion Transformer requires additional consideration of its temporal learning capabilities.

Feature Importance Analysis

Feature importance analysis provides crucial insights into model behavior and decision-making processes. Each modeling approach implements feature importance calculation differently, requiring careful interpretation across models. For tree-based models, feature importance is extracted through gain-based metrics:

$$\text{Importance}(f) = \sum_{t \in \text{trees}} \text{gain}(f, t) \quad (4.34)$$

where $\text{gain}(f, t)$ represents the improvement in accuracy contributed by feature f in tree t .

The Temporal Fusion Transformer provides feature importance through its variable selection

networks and attention mechanisms, offering temporal context-dependent importance scores.

To enable cross-model comparison of feature importance, their importance scores are normalized within each model:

$$\text{Normalized_Importance}(f) = \frac{\text{Importance}(f)}{\sum_{i \in \text{features}} \text{Importance}(i)} \quad (4.35)$$

This normalization allows the consistent identification of essential features across different models, providing insights into the most crucial predictors of photovoltaic system performance.

Visualization Methods

Our visualization framework designs a structured approach to analyze and compare model performance using several complementary techniques. It is intended to provide a broad view of general prediction accuracy and specific performance characteristics.

The temporal comparison of forecasted vs. actual values provides the backbone for our performance analysis. Full time series plots show overall generation patterns and prediction accuracy over longer time horizons. This type of visualization is very effective in clearly demonstrating how various models capture long-generation cycles.

Error analysis visualizations complement basic comparisons to investigate prediction accuracy patterns further. Our framework sets up complementary error visualization approaches to create a holistic model performance picture. Temporal error distribution plots track the deviations of predictions over time. A reference zero line helps identify any possible bias in systematic predictions. These visualizations explain variations in prediction accuracy with operating conditions and seasonal patterns by pointing out particularly challenging scenarios for models.

Long-term patterns are analyzed using STL decomposition-based visualizations to separate the trend components between actual and forecasted values and compare them. Visualizations at a two-week periodicity show how models capture changes in the underlying system behavior and suppress short-term variations. The trend comparison plots stress model adaptation to the systems' seasonal changes and systematic evolution in performance, which allow particular insight into the long-term prediction reliability.

Feature importance visualizations, using normalized importance rankings, allow direct comparison of the relevance of predictive factors. Such visualizations will show which fea-

tures drive prediction accuracy. This can be especially helpful in gaining insight into how environmental conditions relate to prediction performance and, again, giving guidance on how models may be improved by better feature selection or feature engineering.

Combining these visualization techniques gives comprehensive insight into the model's behavioral characteristics and patterns of prediction accuracy. The framework sustains constant styling and scale conventions across all visualizations, allowing direct comparisons of different performance aspects without losing clarity of patterns.

4.5 Summary

This chapter outlined a detailed methodological framework for predicting the performance of photovoltaic systems using machine learning. It showcased feature engineering, a range of modeling techniques from tree-based ensembles to deep learning architectures, as well as robust validation strategies. Our approach guarantees that the generalization capabilities of the models are thoroughly assessed over time and across different system configurations, facilitated by both temporal and system-based validation, along with comprehensive performance analysis. It allowed us to systematically compare different modeling techniques while providing insight into some of the most critical drivers of prediction accuracy. Therefore, it forms a methodological basis that would definitely support reliable photovoltaic power prediction and further advance our understanding of machine learning applications in renewable energy forecasting.

Chapter 5

Results and Analysis

Our analysis begins with establishing fundamental performance benchmarks through the baseline model - a simple 24-hour persistence approach that predicts each hour's generation will match the same hour from the previous day. This naive approach provides context for evaluating more sophisticated models, revealing both the inherent patterns and challenges in photovoltaic power prediction.

5.1 Baseline Model Results

The baseline model, implementing 24-hour shift prediction, demonstrates surprisingly robust performance despite its simplicity. Under temporal validation, where we predict future performance based on historical data, the model achieves an efficiency RMSE of 0.1459 and MAE of 0.0649, with an R^2 value of 0.7028. When translated to absolute generation, this corresponds to an RMSE of 98.39 kWh and MAE of 34.81 kWh.

Interestingly, under system-based validation, where we predict an entirely unseen system, the baseline shows comparable performance with an efficiency RMSE of 0.1405 and MAE of 0.0593 (R^2 of 0.7245). However, the absolute metrics reveal a different story, with RMSE increasing to 140.31 kWh and MAE to 59.21 kWh, highlighting how system capacity differences amplify prediction errors.

These results demonstrate the strong daily periodicity in solar generation and the fundamental limitations of such a simple approach. The relatively high R^2 values (around 0.72-0.75) indicate that simply assuming tomorrow will match today captures a significant portion of generation patterns. However, this baseline approach cannot:

- Account for weather changes between consecutive days
- Adapt to seasonal variations in generation patterns
- Consider system-specific characteristics beyond basic persistence
- Respond to irregular patterns or unexpected conditions

The baseline's performance provides a critical reference point for evaluating more sophisticated approaches. Any advanced model should demonstrate substantial improvement over these metrics to justify its additional complexity. The baseline's ability to achieve R^2 values above 0.70 sets a non-trivial threshold for improvement, while its clear limitations outline specific areas where more sophisticated models need to excel.

5.2 Tree-based Models Performance

Following our preliminary analysis, we explore the performance of tree-based ensemble algorithms, each of which have a different story to tell about the prediction problem. These models show a substantial improvement over baseline, by learning rich relationships between meteorological features and power production patterns.

5.2.1 Random Forest Results

The Random Forest model substantially improves its performance over the baseline with its parallel tree construction and feature bagging approach. We analyze its capabilities across both validation scenarios to understand its generalization characteristics.

Temporal Validation Performance

Random Forest has performed remarkably well in the temporal validation, where the model was trained on data from 2021 to 2023 and predicts generation patterns for 2024. The model had an efficiency RMSE of 0.0795 and MAE of 0.0364, while the R^2 is calculated at 0.9117. These are the metrics that indicate an improvement of about 45.50% over the baseline approach. This goes to mean that indeed, the model captures much more about the pattern than a simple day-ahead persistence. Translated into absolute generation values, this gives RMSE of 53.01 kWh and MAE of 19.50 kWh, while R^2 improves to 0.9301.

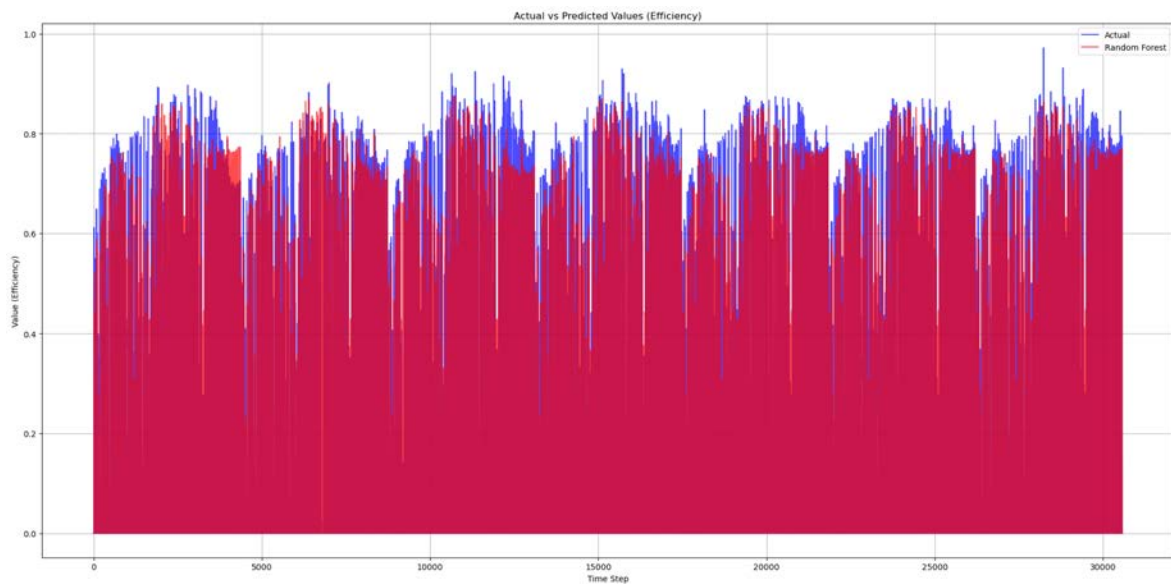


Figure 5.1: Comparison of actual and predicted efficiency values using Random Forest under temporal validation. The close alignment between actual (blue) and predicted (red) values demonstrates the model's ability to capture complex generation patterns.

The time series visualization of predictions reveals several important characteristics of the model's performance. As shown in Figure 5.1, the Random Forest accurately tracks complex daily generation patterns. The overlay of actual and predicted values shows powerful performance in capturing:

- Daily generation peaks with high precision
- Seasonal variations in generation patterns
- Rapid transitions during sunrise and sunset periods

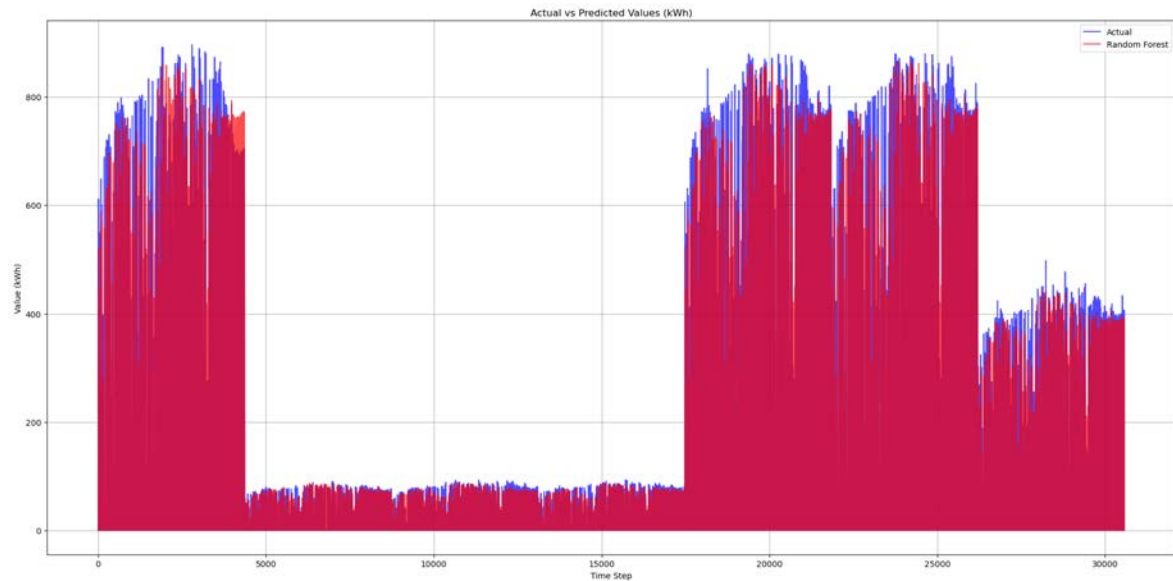


Figure 5.2: Actual versus predicted absolute power generation in kWh. The visualization demonstrates the model's ability to adapt to different generation scales throughout the prediction period.

The absolute generation predictions shown in Figure 5.2 reflect mainly the model's ability to adaptation to different generation scales throughout the year. In a quantitative way, this increased predictive capability is reflected in the improved R^2 , from 0.7591 in the baseline to 0.9301 in absolute generation terms. The close alignment between predicted and actual values across high-generation summer and lower-generation winter periods indicates good performance under variable seasonal conditions. Simultaneously, minor deviations in peak predictions at extreme generation events point to an inability to perfectly predict the most exceptional conditions.

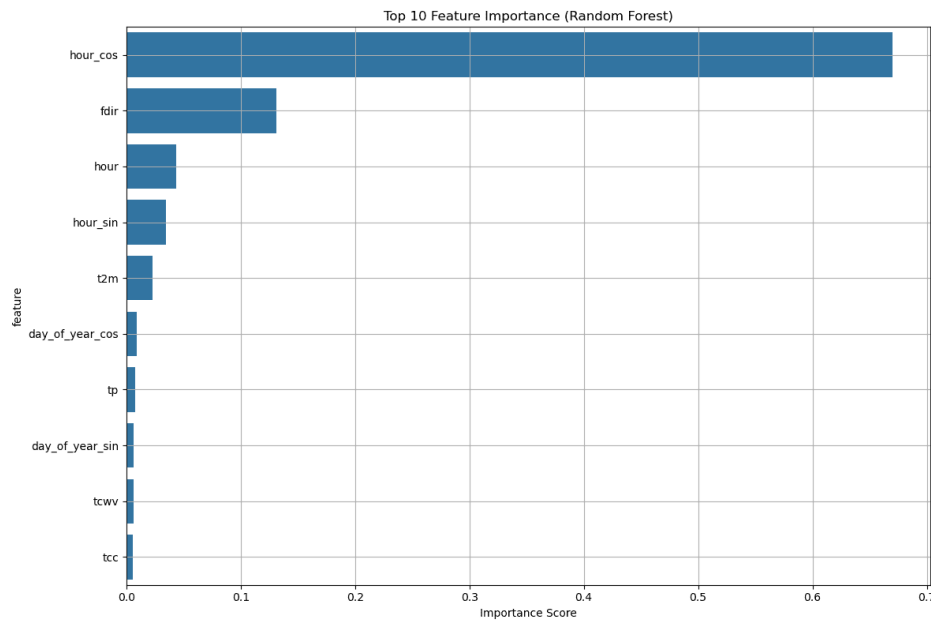


Figure 5.3: Top 10 most important features in the Random Forest model. The dominance of temporal features, particularly `hour_cos`, highlights the fundamental importance of daily solar patterns in generation prediction.

Feature importance analysis (Figure 5.3) provides insight into how the model achieves this performance. The cyclical encoding of hour (`hour_cos`) emerges as the dominant predictive feature, highlighting the fundamental importance of daily solar patterns in generation prediction. This is followed by the total sky direct solar radiation (`fdir`), which captures immediate weather effects. The strong influence of temporal features (`hour`, `hour_sin`) combined with meteorological parameters (`t2m`) demonstrates how the model balances structural and environmental factors in its predictions.

Further insight into the model's performance can be gained through trend analysis using Seasonal-Trend decomposition (STL). The trend comparisons reveal the model's ability to capture short-term fluctuations and long-term patterns in generation efficiency. As shown in Figure 5.4, the predicted efficiency trend closely mirrors the actual trend throughout the prediction period, demonstrating the model's robust capability to capture underlying generation patterns. This close alignment is particularly notable given that the model predicts future behavior based solely on historical training data.

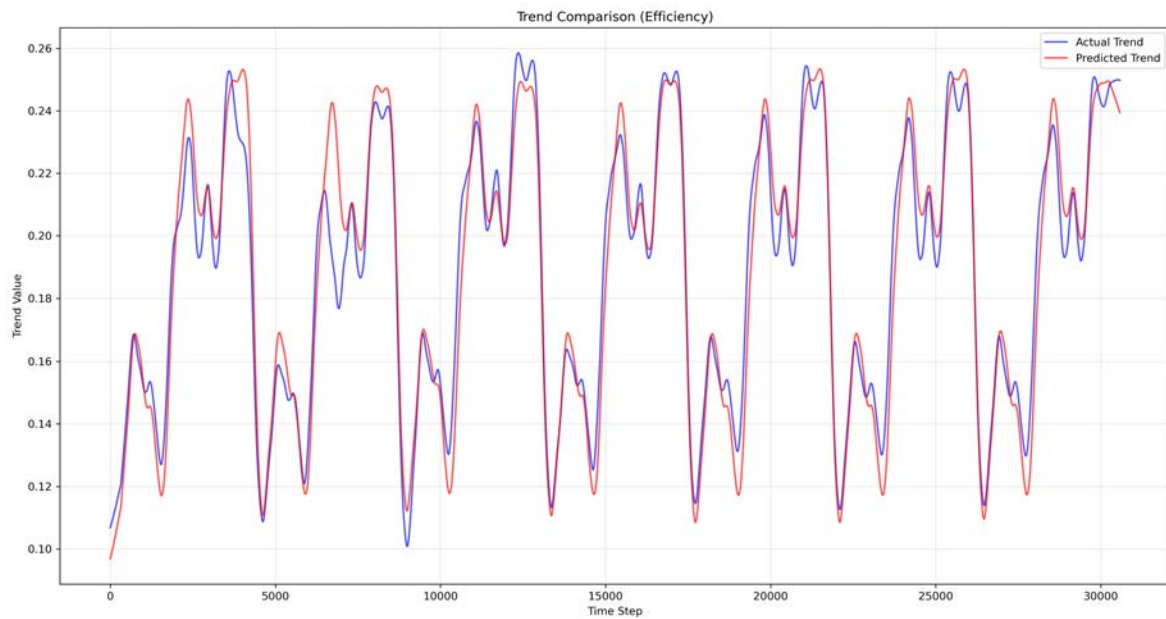


Figure 5.4: Trend comparison of actual versus predicted efficiency values, highlighting the model's ability to capture short-term and seasonal fluctuations. The close alignment between actual (blue) and predicted (red) trends demonstrates robust temporal pattern recognition.

When examining the absolute generation trends (Figure 5.5), the model demonstrates remarkable consistency in tracking generation patterns across different scales. The decomposed trends show strong alignment between predicted and actual values throughout the test period, with the model successfully capturing both gradual seasonal variations and shorter-term generation patterns. This consistent trend-following behavior across different temporal scales suggests that the model has effectively learned the underlying relationships between our input features and generation patterns.

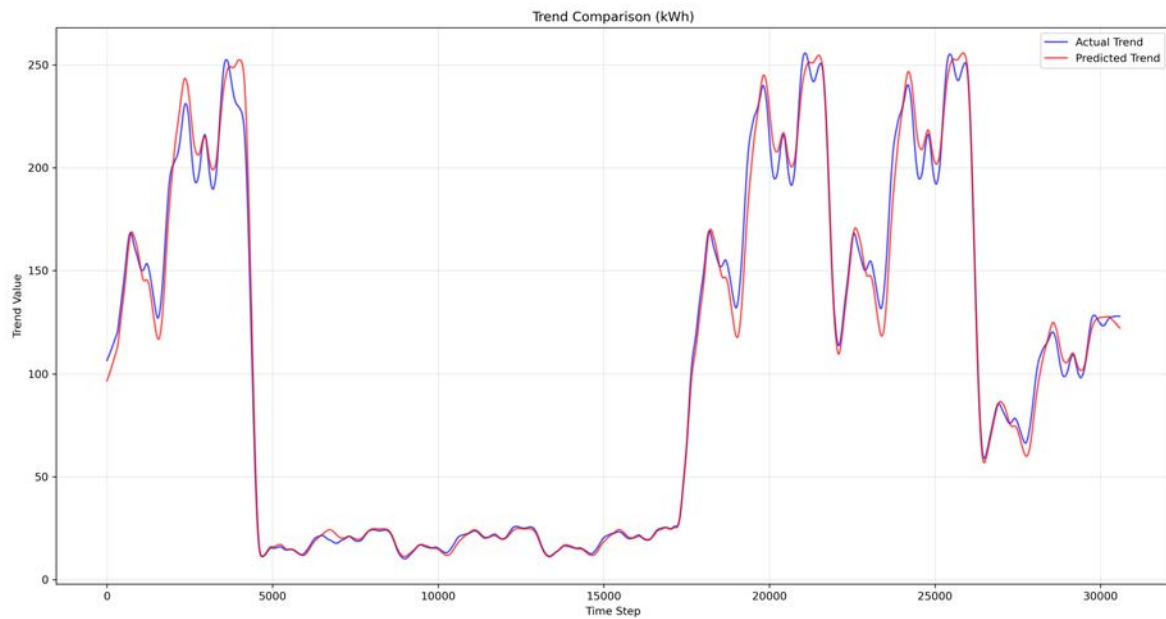


Figure 5.5: Trend comparison of actual versus predicted absolute generation in kWh, demonstrating the model's ability to adapt to different generation scales throughout the year. The analysis reveals consistent performance across both high and low generation periods.

Observing prediction errors over time provides additional insights into the model's performance characteristics. The residual analysis for efficiency predictions (Figure 5.6) reveals a symmetric distribution of errors around zero, with most deviations contained within ± 0.2 efficiency units. This pattern suggests unbiased predictions, though occasional more significant deviations up to ± 0.6 are observed during particularly challenging conditions, often corresponding to rapid weather changes or extreme generation events.

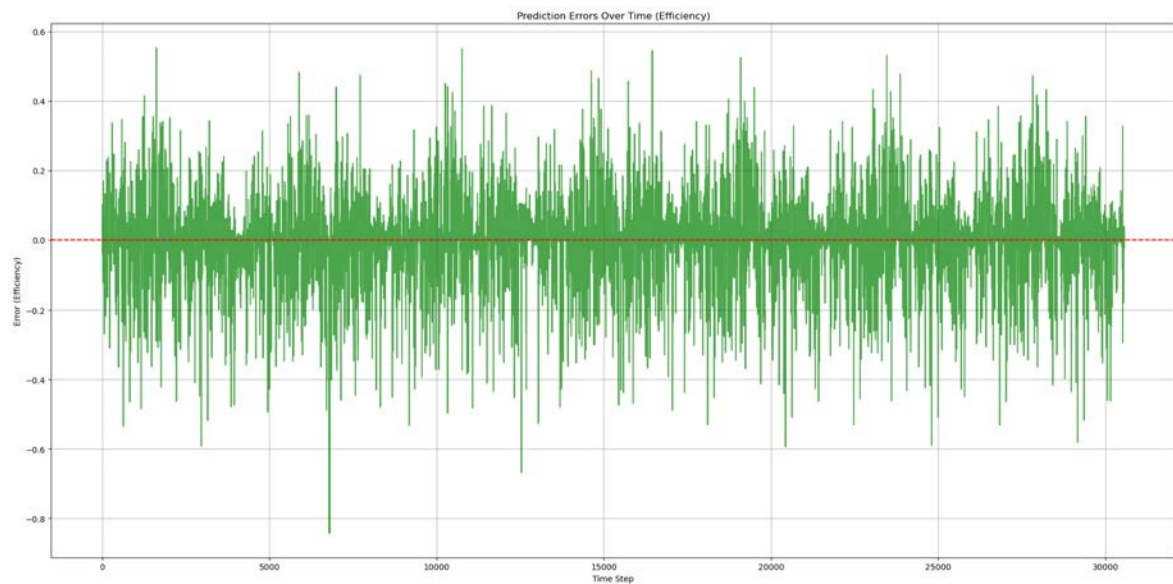


Figure 5.6: Temporal distribution of prediction errors for efficiency values. The symmetric pattern around zero indicates unbiased predictions, while the consistent spread demonstrates stable model performance across the prediction period.

The absolute generation error patterns (Figure 5.7) reflect our dataset's diverse range of system sizes, spanning from 99.9 kW to 1010.88 kW installations. This variation in system capacity naturally leads to more significant absolute error ranges compared to the normalized efficiency metrics. The heteroscedastic pattern in the absolute errors primarily reflects these underlying system size differences rather than model performance limitations, which is why the efficiency metrics provide a more standardized view of prediction accuracy across all installations.

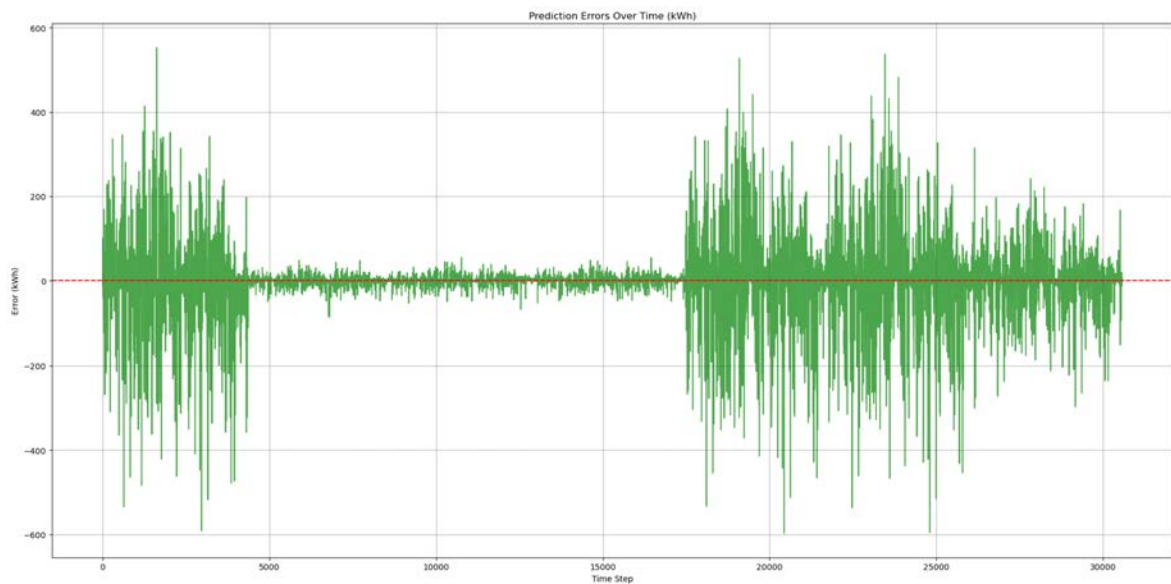


Figure 5.7: Temporal distribution of prediction errors for absolute generation in kWh. The larger error ranges reflect the diverse system sizes in the dataset (99.9-1010.88 kW), with larger absolute errors naturally corresponding to predictions for higher-capacity installations.

Of particular note is the model's handling of daily transitions. Unlike the baseline approach, which relies solely on the previous day's values, the Random Forest successfully captures variations in generation patterns caused by changing weather conditions. This is evidenced by the smooth tracking of generation curves even during high variability, though some minor discrepancies remain during rapid weather changes. The reduction in RMSE from 98.39 kWh in the baseline to 53.01 kWh demonstrates this improved ability to handle transitions and varying conditions.

System-based Validation Performance

Under system-based validation, the Random Forest is trained on data from six installations and predicts the generation patterns of a completely unseen system (System 74894, 998.955 kW). The model demonstrates exceptional predictive capabilities. It achieves an efficiency RMSE of 0.0281 and MAE of 0.0108, with an R^2 value of 0.9890. These efficiency metrics for this particular system translate to absolute generation errors of 28.04 kWh RMSE and 10.82 kWh MAE, representing an 80.02% improvement over the baseline approach.

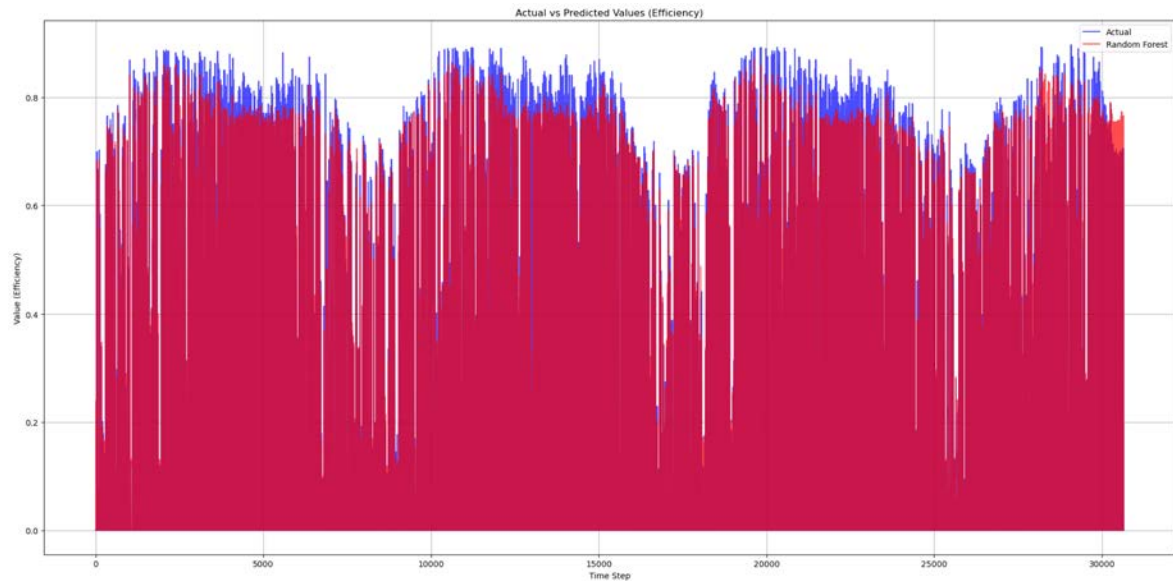


Figure 5.8: Comparison of actual and predicted efficiency values for System 74894. The close alignment between actual (blue) and predicted (red) values demonstrates the model's ability to generalize to an unseen system configuration.

The time series visualization reveals remarkable prediction accuracy across different generation conditions. As shown in Figure 5.8, the model successfully captures the overall generation patterns and fine-grained variations, with efficiency values ranging from near-zero during non-generation hours to peaks around 0.85 during optimal conditions. This accurate tracking across the full range of generation conditions is particularly impressive given that the model had no exposure to this specific system during training.

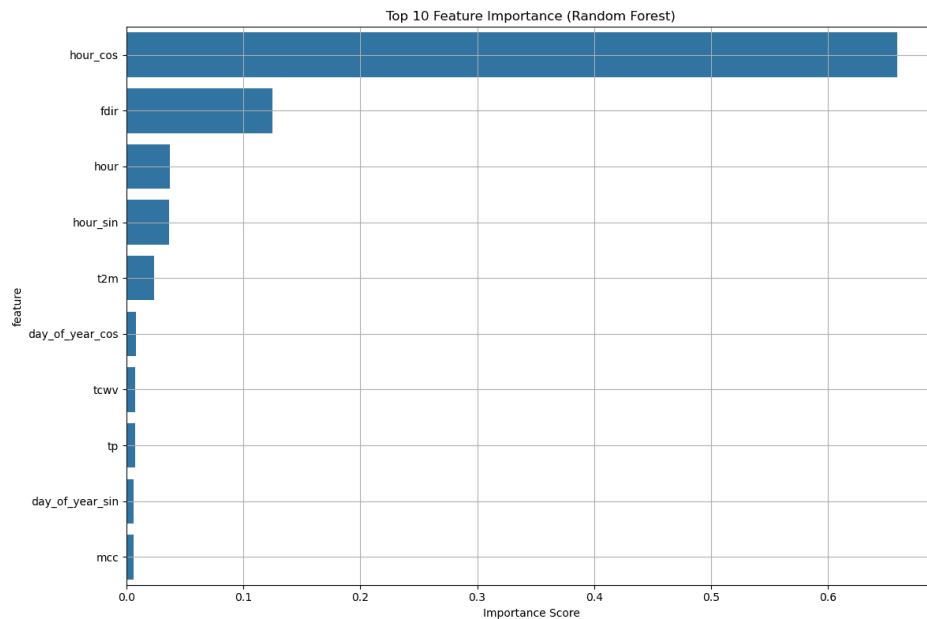


Figure 5.9: Top 10 most important features in the Random Forest model under system-based validation. The dominant influence of temporal features (*hour_cos*) and solar radiation (*fdir*) remains consistent with the temporal validation results.

Feature importance analysis (Figure 5.9) reveals patterns similar to those observed in temporal validation, with *hour_cos* remaining the dominant predictive feature followed by total sky direct solar radiation (*fdir*). This consistency in feature importance rankings between temporal and system-based validation suggests that the model has learned generalizable relationships between environmental conditions and generation patterns that extend beyond system-specific characteristics. The strong influence of cyclical temporal features (*hour_cos*, *hour_sin*) combined with meteorological parameters demonstrates how the model effectively combines temporal and environmental factors to achieve high prediction accuracy.

The trend analysis through Seasonal-Trend decomposition (STL) provides deeper insight into the model's exceptional generalization capabilities. As shown in Figure 5.10, the predicted efficiency trend demonstrates remarkable alignment with the actual trend throughout the prediction period. This close tracking is particularly notable in the smooth transitions between different generation regimes, with efficiency trends ranging from 0.05 to 0.25 across seasonal variations. The model's ability to capture these long-term patterns for an unseen system demonstrates its robust understanding of the underlying relationships between environmental conditions and photovoltaic generation.

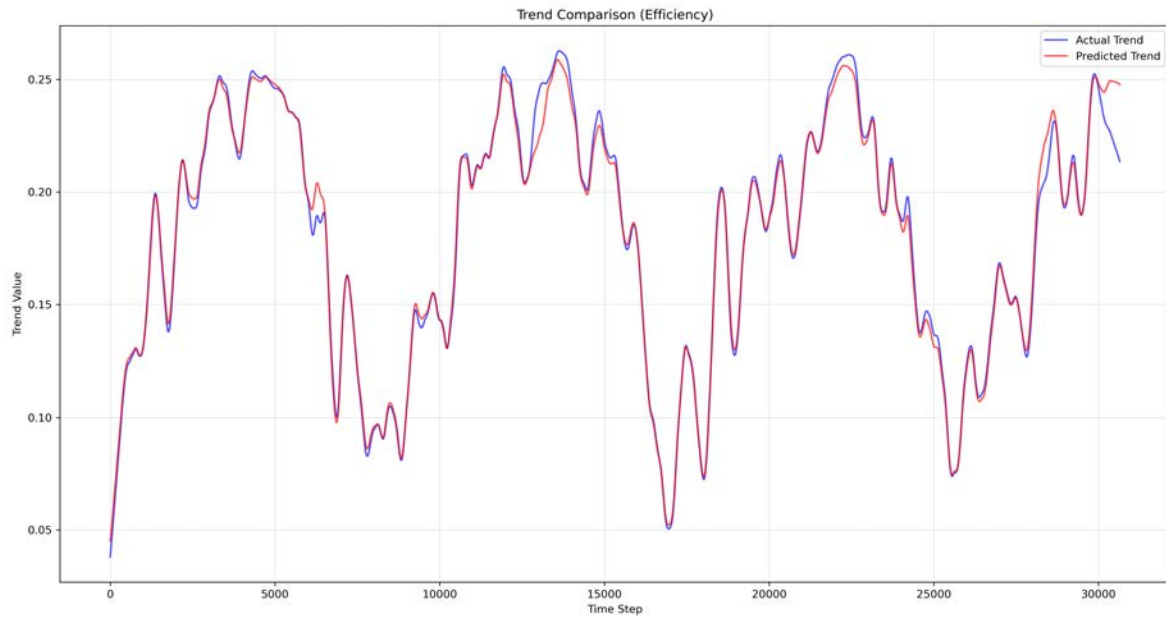


Figure 5.10: Trend comparison of actual versus predicted efficiency values for System 74894, highlighting the model's ability to capture gradual transitions and seasonal patterns. The near-perfect alignment between actual (blue) and predicted (red) trends demonstrates exceptional generalization to an unseen system.

The residual error analysis (Figure 5.11) further reinforces the model's outstanding performance. The prediction errors show a remarkably tight distribution around zero, with most deviations within ± 0.1 efficiency units. This consistent accuracy is maintained across the entire prediction period, with only occasional more significant deviations reaching up to ± 0.4 during particularly challenging conditions. The symmetrical nature of these residuals around zero indicates unbiased predictions, while their small magnitude - reflected in the MAE of just 0.0108 - demonstrates the model's consistent accuracy.

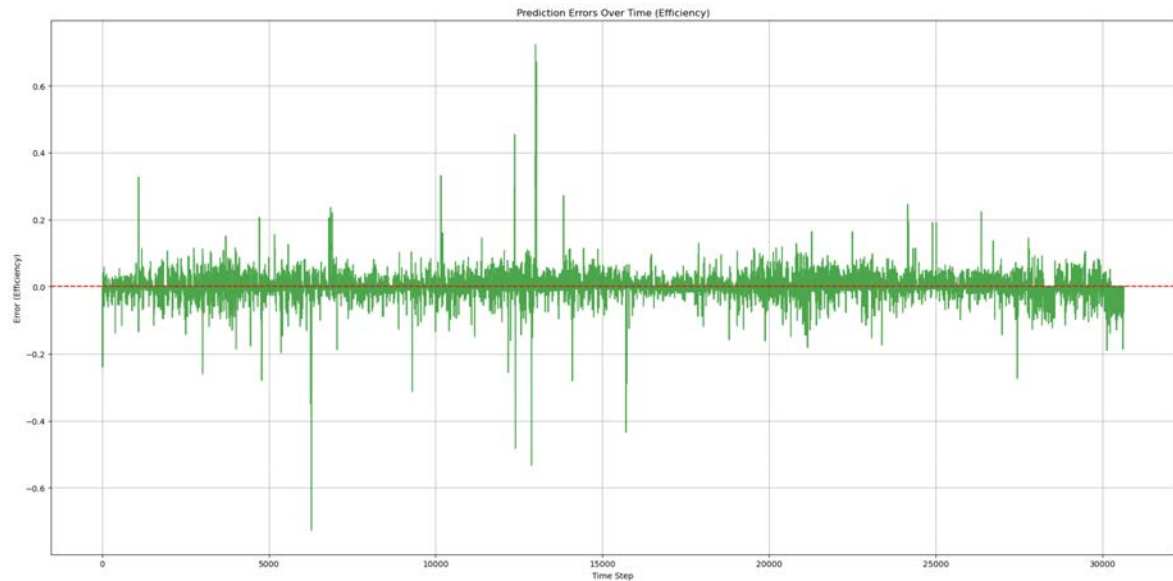


Figure 5.11: Temporal distribution of prediction errors for System 74894. The tight clustering around zero and limited range of deviations demonstrate the model's consistent accuracy in predicting an unseen system's behavior.

Importantly, the model's performance is excellent despite not being trained on this specific system. With an RMSE as low as 0.0281 and a very good R^2 of 0.9890, the Random Forest has learned general patterns without being limited to system-specific characteristics. It improves by 80.02% over baseline, which means the model has learnt some fundamental relationship between meteorological conditions and the photovoltaic generation valid under different system settings.

5.2.2 XGBoost Results

Temporal Validation Performance

In a temporal validation setting, where the model is trained with data from 2021 to 2023 to forecast generation patterns for 2024, XGBoost shows impressive predictive performance. It records an RMSE of 0.0770 and an MAE of 0.0369, along with an R^2 value of 0.9173. These efficiency metrics translate to absolute generation values of 51.29 kWh RMSE and 19.73 kWh MAE for our test period, representing a 47.24% improvement over the baseline approach.

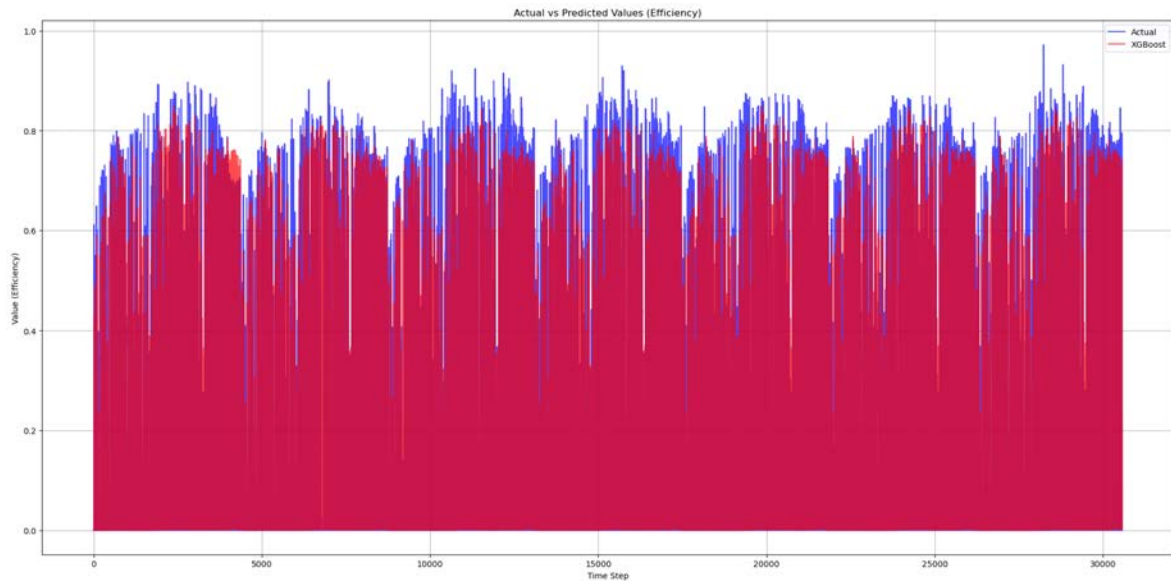


Figure 5.12: Comparison of actual and predicted efficiency values using XGBoost under temporal validation. The close alignment between actual (blue) and predicted (red) values demonstrates the model’s ability to capture complex generation patterns across the prediction horizon.

The time series visualization reveals XGBoost’s capability to track generation patterns across different conditions. As shown in Figure 5.12, the model successfully captures the overall generation profile and detailed variations, with efficiency values ranging from near-zero during non-generation hours to peaks approaching 0.85 during optimal conditions. The close alignment between predicted and actual values demonstrates the model’s robust performance across different generation regimes, though some minor deviations are observed during peak generation periods.

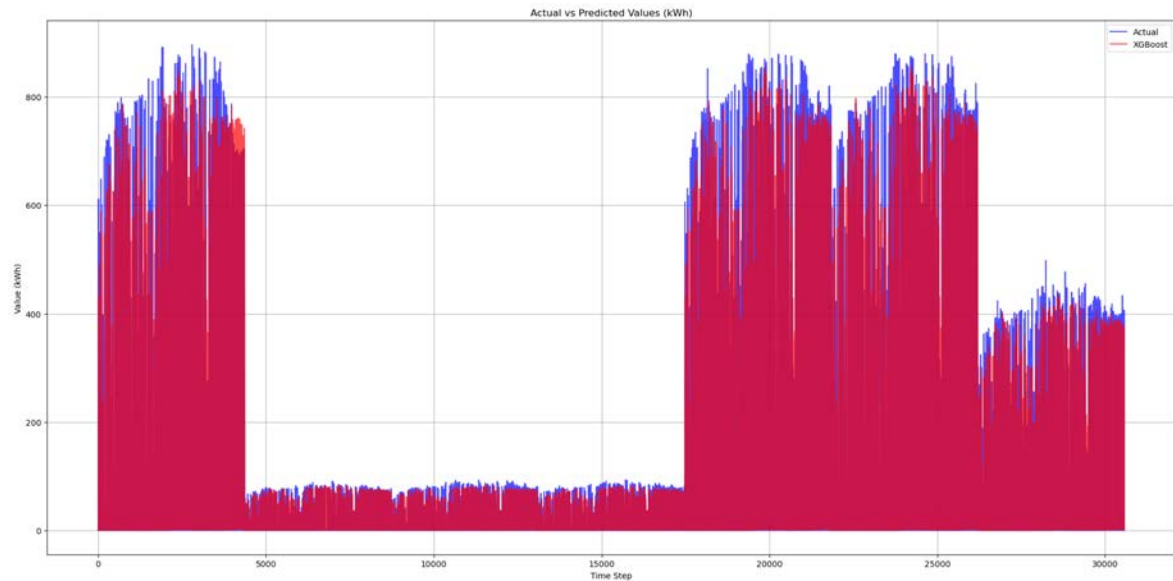


Figure 5.13: Actual versus predicted absolute power generation in kWh. The visualization demonstrates how the efficiency predictions translate to real-world generation values, with the model maintaining accuracy across systems of varying capacities.

We can observe the model's performance across our diverse set of installations when examining the absolute generation predictions (Figure 5.13). The visualization clearly shows the parallel prediction of systems with different capacities, from our 1000 kW installations generating up to 800 kWh during peak conditions to our 100 kW systems operating in their proportionally lower ranges. The model maintains consistent prediction accuracy across this wide range of system sizes, demonstrating its ability to learn and apply generation patterns regardless of installation capacity. This robust performance across different system scales contributes to the substantial R^2 value of 0.9345 in absolute generation terms.

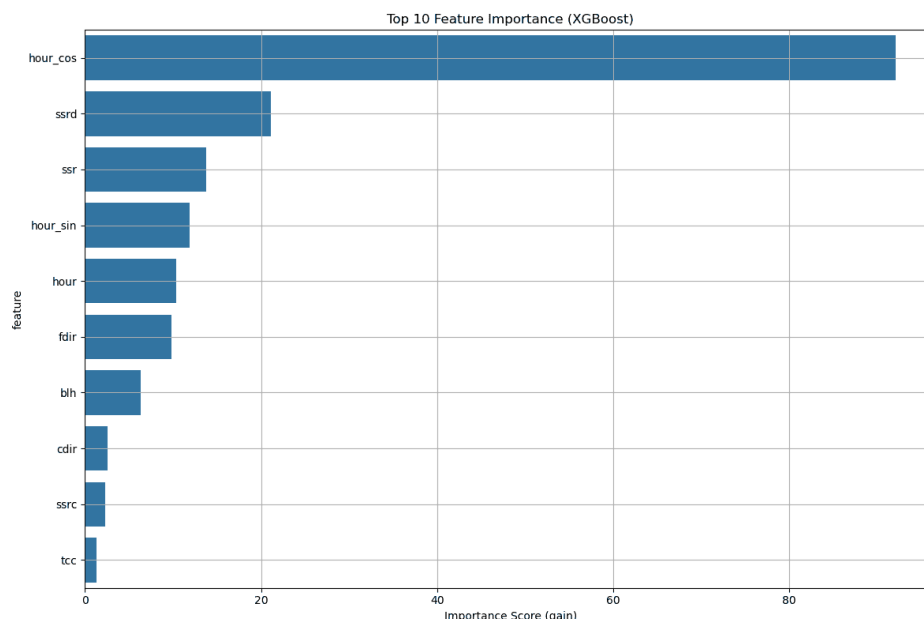


Figure 5.14: Top 10 most important features in the XGBoost model, revealing a distinctive hierarchy where solar radiation parameters (ssrd, ssr) play a more prominent role compared to traditional time-based features.

The feature importance analysis reveals the surprising aspect of XGBoost’s learning process. Despite the paramount role of `hour_cos` for predictions, the model uses a dramatically alternative approach to incorporate features. XGBoost favors measurements of direct solar radiation (ssrd and ssr), suggesting a predictive model that strongly depends on current environmental factors, even though the time series may consist only of temporal trends. This inclination towards radiation metrics signifies that XGBoost builds its predictions by placing greater emphasis on the actual solar resources available, which may render it more reactive to swift changes in weather conditions.

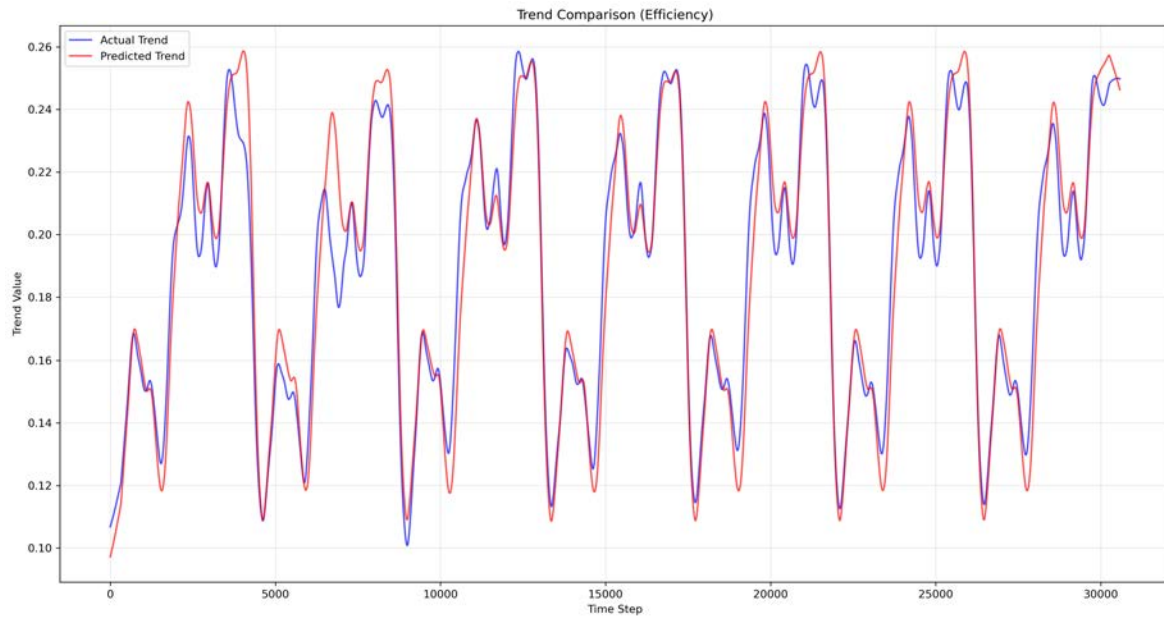


Figure 5.15: Trend comparison of actual versus predicted efficiency values. The decomposed trends reveal XGBoost’s characteristic handling of transitional periods, with powerful performance during seasonal shifts.

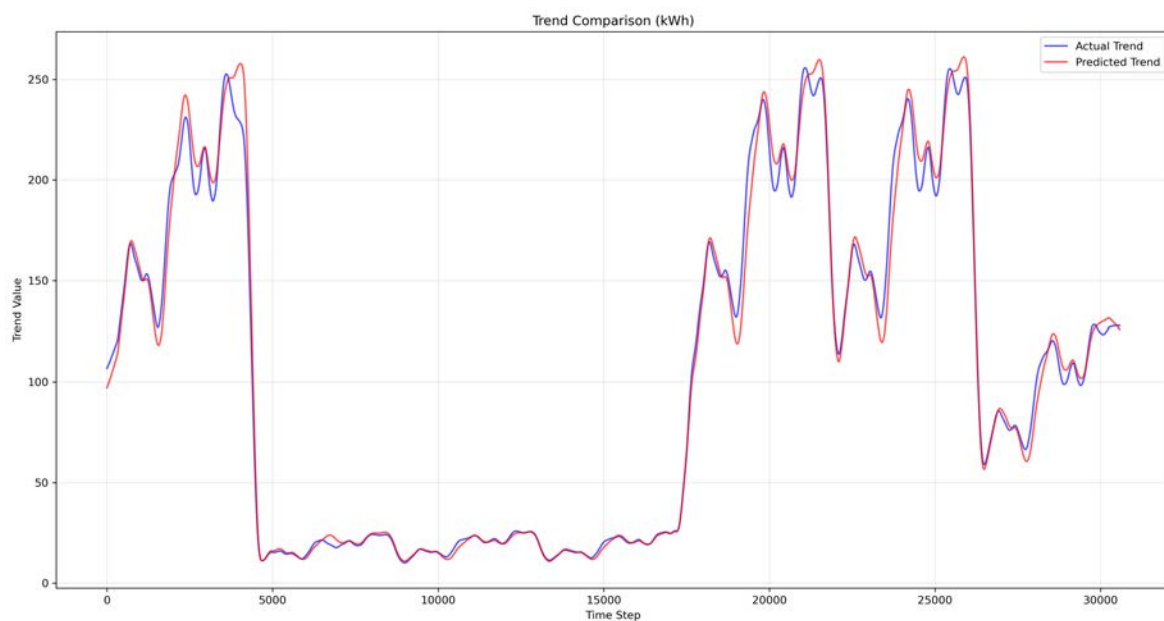


Figure 5.16: Trend comparison of actual versus predicted absolute generation values. The visualization demonstrates how efficiency predictions scale across our installation range, with notable precision in tracking major and minor generation fluctuations.

The trend analysis uncovers subtle but significant aspects of XGBoost’s prediction behavior. The efficiency trends (Figure 5.15) exhibit a distinctive pattern where the model antici-

pates changes slightly ahead of their occurrence, particularly visible in the transition points between different generation regimes. This forward-looking characteristic suggests that XGBoost's gradient-boosting approach might be especially adept at identifying early indicators of changing generation conditions.

We observe how this predictive behavior translates across our diverse installation set when examining the absolute generation trends (Figure 5.16). The model exhibits an impressive degree of consistency in maintaining prediction accuracy at all system scales, precise tracking through the intermediate generation intervals. This implies that XGBoost's boosting process naturally tunes the prediction strategies across different system sizes without degrading the prediction quality of either side of the capacity spectrum.

The residual error analysis provides insight into XGBoost's prediction behavior across our diverse installations. The efficiency error distribution (Figure 5.17) shows a characteristic pattern where most deviations remain contained within ± 0.2 efficiency units, with the model exhibiting remarkably stable performance during mid-day generation periods. While occasional more significant deviations reaching up to ± 0.4 efficiency units occur, they appear primarily during rapid transition periods or challenging weather conditions.

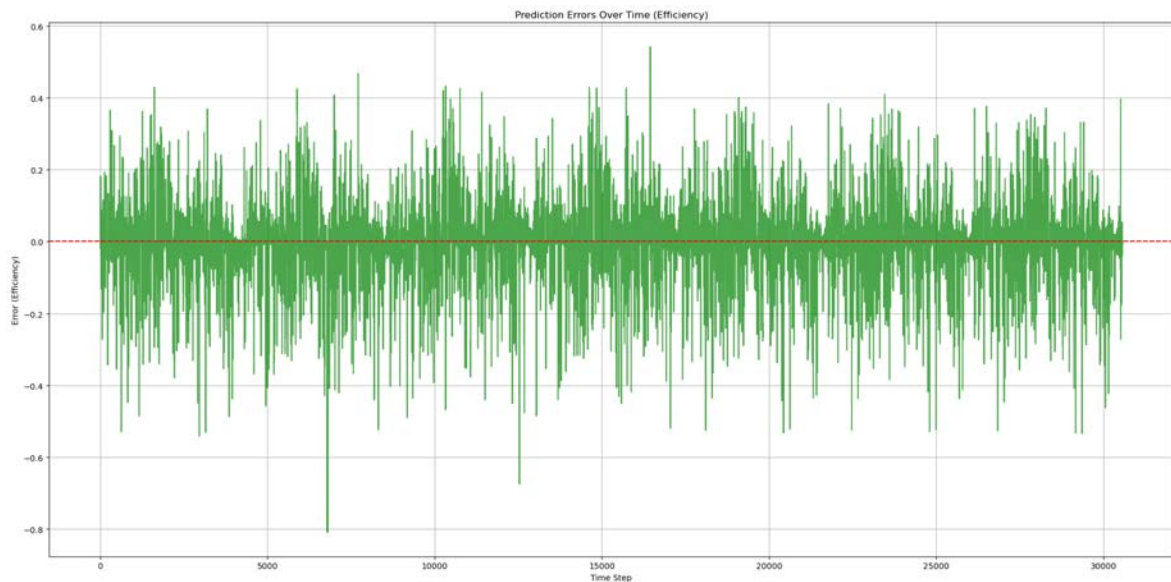


Figure 5.17: Temporal distribution of prediction errors in efficiency terms. The symmetric nature of the errors around zero, combined with their consistent spread, indicates XGBoost's balanced prediction approach across different operating conditions.

When examining the absolute generation errors (Figure 5.18), we can observe how XGBoost's prediction characteristics manifest across different system scales. The error pattern

reveals an interesting segmentation that aligns with our installation capacities - the more significant magnitude errors naturally correspond to our 1000 kW installations, while proportionally more minor deviations appear for our 100 kW systems. While maintaining consistent relative accuracy, this scaling of errors with system size demonstrates XGBoost's ability to adapt its predictions proportionally across our installation range.

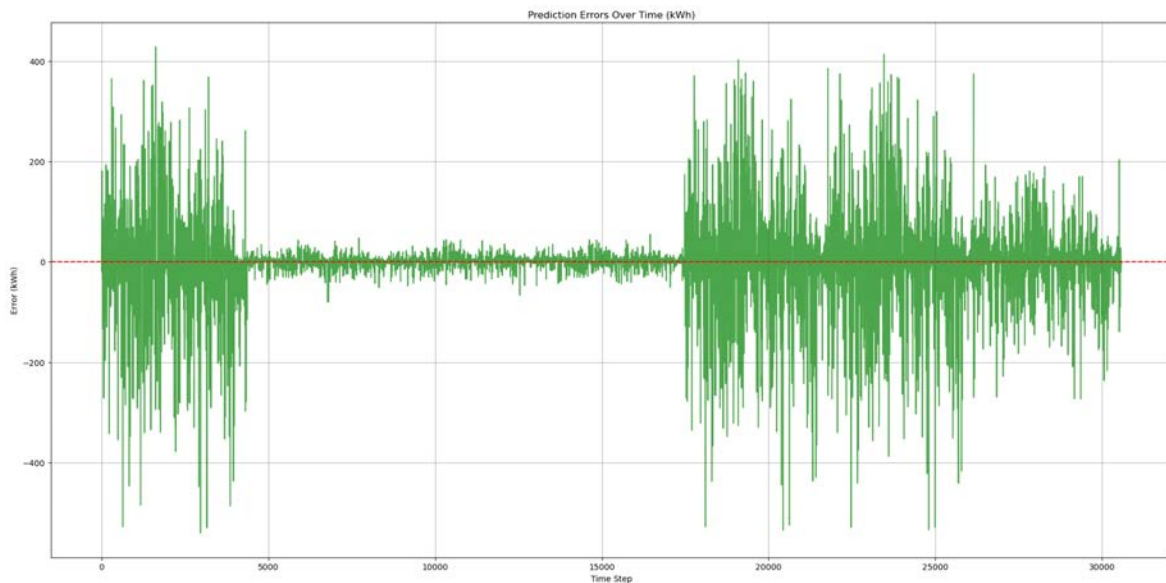


Figure 5.18: Temporal distribution of prediction errors in absolute generation terms. The varying error magnitudes reflect the diverse system sizes in our dataset, with larger deviations corresponding to predictions for higher-capacity installations.

The model's 47.24% improvement over baseline performance reflected in an efficiency RMSE of 0.0770, suggests that XGBoost's gradient boosting approach effectively captures the complex relationships between our environmental predictors and generation outcomes. The consistent error patterns and symmetric distribution around zero indicate that the model achieves this performance without introducing systematic bias, making it particularly reliable for operational forecasting applications.

The identified error patterns, along with XGBoost's high prediction accuracy and significant enhancement over the baseline, showcase the model's capability in forecasting temporal photovoltaic generation. The even distribution of residuals and stable performance across various system sizes indicate that the model has effectively grasped the essential relationships influencing solar generation, establishing it as a dependable resource for upcoming power prediction initiatives.

System-based Validation Performance

Under system-based validation, where XGBoost predicts generation patterns for an entirely unseen installation (System 74894, 998.955 kW), the model demonstrates good predictive capabilities while exhibiting some characteristic behaviors. The model achieves an efficiency RMSE of 0.0441 and MAE of 0.0206, with an R^2 value of 0.9728. These efficiency metrics for this particular system translate to absolute generation errors of 44.10 kWh RMSE and 20.59 kWh MAE, representing a 68.57% improvement over the baseline approach.

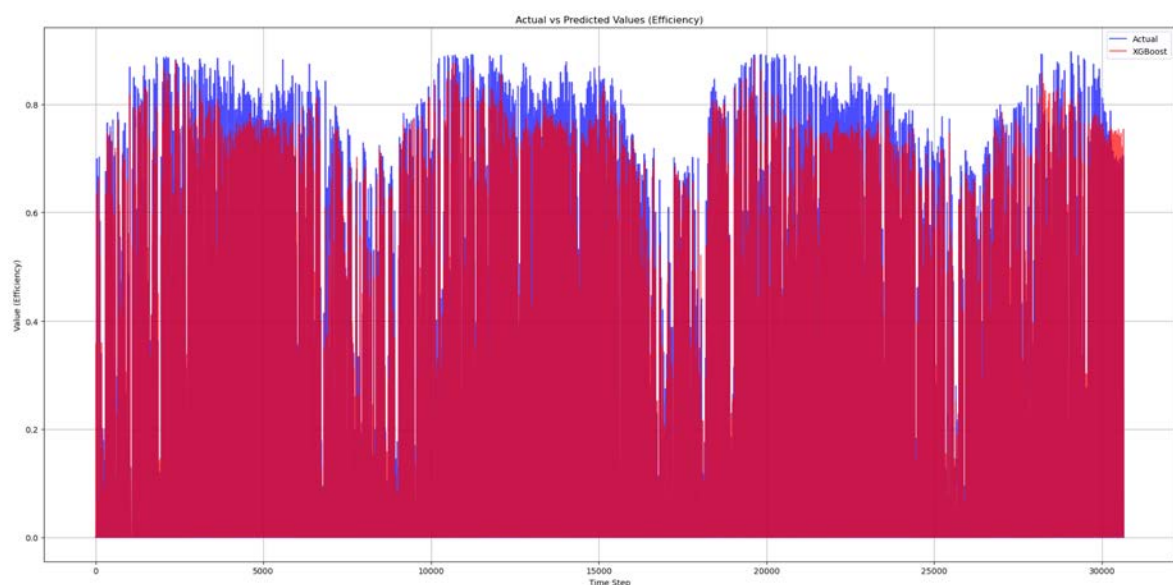


Figure 5.19: Comparison of actual and predicted efficiency values for System 74894. The visualization reveals XGBoost’s tendency toward slight underprediction, which is evident during peak generation periods, though it maintains strong overall tracking of generation patterns.

The time series visualization reveals interesting aspects of XGBoost’s generalization to unseen systems. As shown in Figure 5.19, while the model successfully captures the overall generation patterns, there is a noticeable tendency toward conservative predictions, particularly during peak generation periods where efficiency values approach 0.85. This systematic slight underprediction suggests that XGBoost’s gradient boosting approach may prioritize prediction stability over capturing extreme values when generalizing to new system configurations.

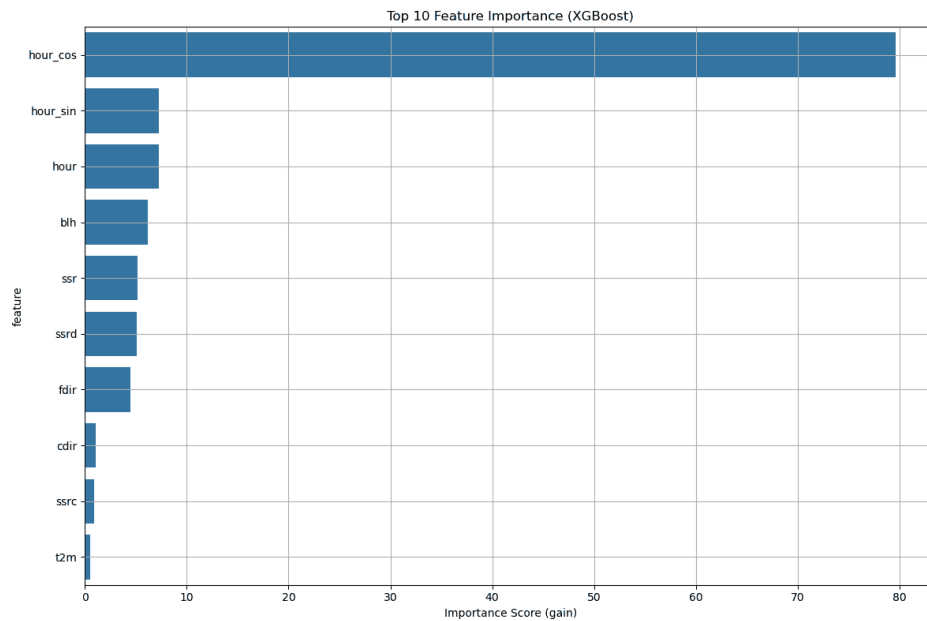


Figure 5.20: Feature importance analysis for system-based prediction reveals XGBoost’s heavy reliance on `hour_cos`, with a marked difference in utilizing other features compared to temporal validation. This suggests a shift in the model’s strategy when generalizing to unseen systems.

The feature importance analysis reveals a notable shift in XGBoost’s prediction strategy when generalizing to an unseen system. Unlike its temporal validation behavior, where radiation parameters played prominent roles, the system-based model shows an overwhelming dependence on `hour_cos`, with other features having significantly reduced influence. This dramatic reweighting suggests that when faced with an unknown system configuration, XGBoost adopts a more conservative approach centered around fundamental daily solar patterns rather than relying heavily on meteorological parameters.

The secondary features - `hour_sin`, `hour`, and boundary layer height (`blh`) - maintain modest but meaningful contributions to the predictions, while radiation parameters (`ssr`, `ssrd`, `fdir`) show surprisingly diminished importance. This reorganization of feature priorities might explain the model’s tendency toward slight underprediction, as it appears to favor stable, time-based patterns over potentially more volatile radiation measurements when generalizing to new system configurations.

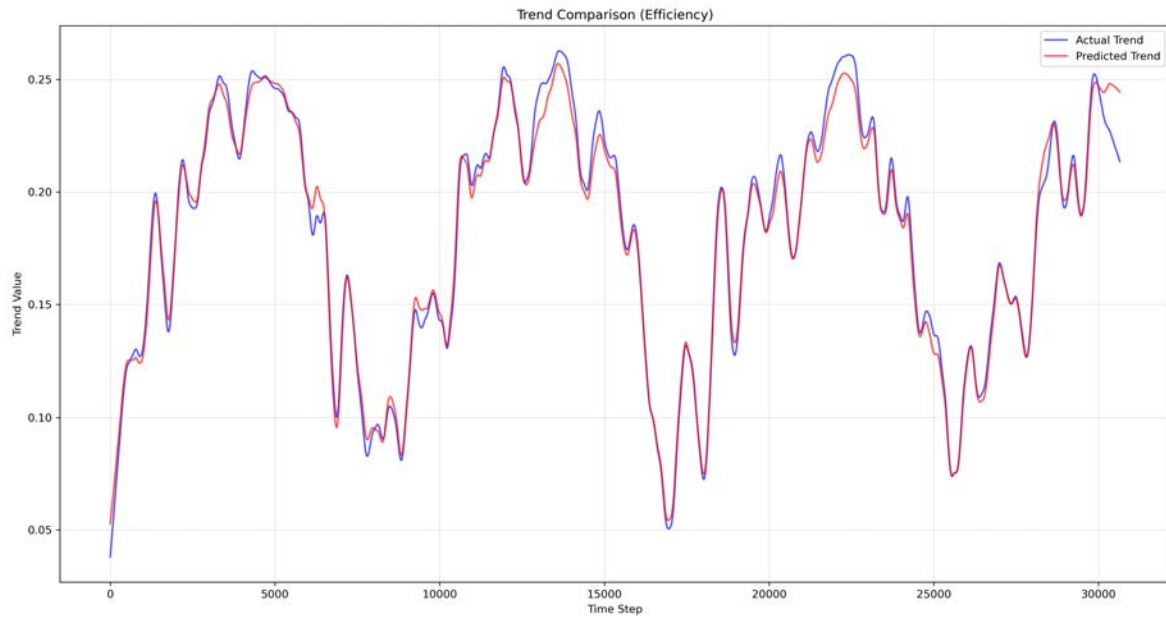


Figure 5.21: Trend comparison of actual versus predicted efficiency values for System 74894. The trend analysis reveals XGBoost’s tendency to smooth out extreme variations, which is particularly visible during peak efficiency periods.

The trend analysis through STL decomposition provides further insight into XGBoost’s system-based prediction behavior. As shown in Figure 5.21, the model’s predictions closely follow the overall generation patterns while exhibiting a characteristic smoothing effect. This behavior is particularly noticeable during peak efficiency periods, where the predicted trend line consistently runs slightly below the actual trend, aligning with the model’s observed tendency toward conservative predictions when generalizing to new systems.

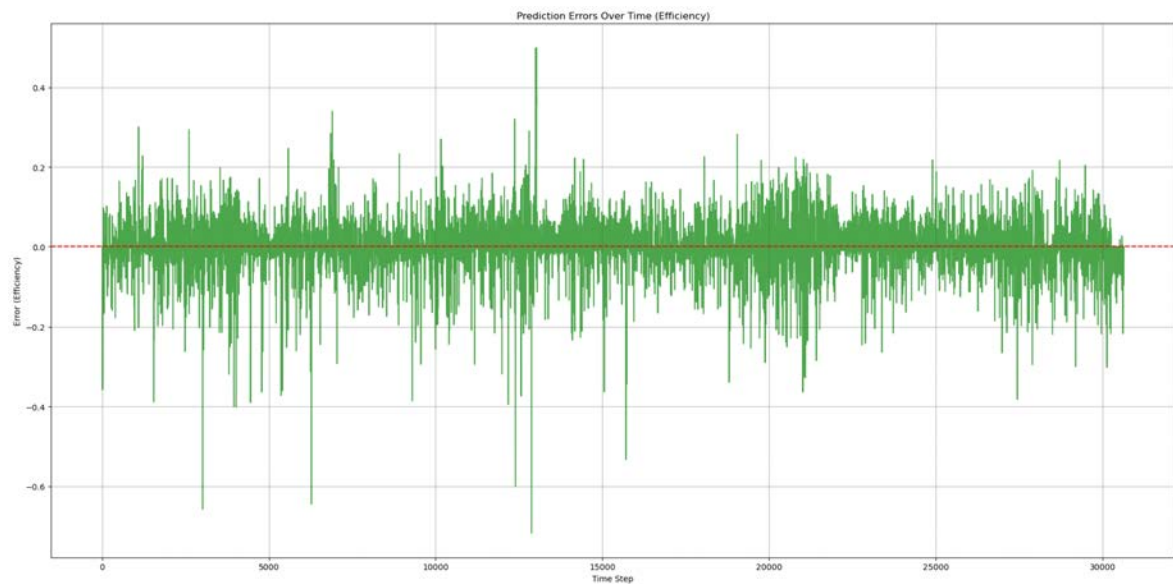


Figure 5.22: Temporal distribution of prediction errors for System 74894. The error pattern shows a balanced distribution with occasional larger deviations during challenging prediction periods.

The error analysis (Figure 5.22) reveals a generally stable prediction pattern with most errors within ± 0.2 efficiency units. While the model experiences occasional more significant deviations, reaching up to ± 0.4 during particularly challenging conditions, these represent a small fraction of the overall predictions. The error distribution maintains a relatively symmetric pattern around zero, though with a slight bias toward negative values that reflects the model's conservative prediction strategy.

The validation of XGBoost shows a notable shift in its prediction strategy with new systems. Increased importance on temporal features and conservative predictions indicate that XGBoost is cautious when generalizing to different configurations. Although performance is slightly lower than other methods, it still improves significantly over the baseline, evidenced by a 68.57% RMSE reduction. This balance of stability and accuracy makes XGBoost a dependable, albeit conservative, option for photovoltaic generation forecasting.

5.2.3 **LightGBM Results**

Temporal Validation Performance

Under temporal validation, LightGBM demonstrates effective predictive capabilities while showcasing its characteristic leaf-wise tree growth approach. The model achieves an effi-

ciency RMSE of 0.0774 and MAE of 0.0375, with an R^2 value of 0.9164. These efficiency metrics translate to absolute generation values of 51.67 kWh RMSE and 20.17 kWh MAE, representing a 46.97% improvement over the baseline approach.

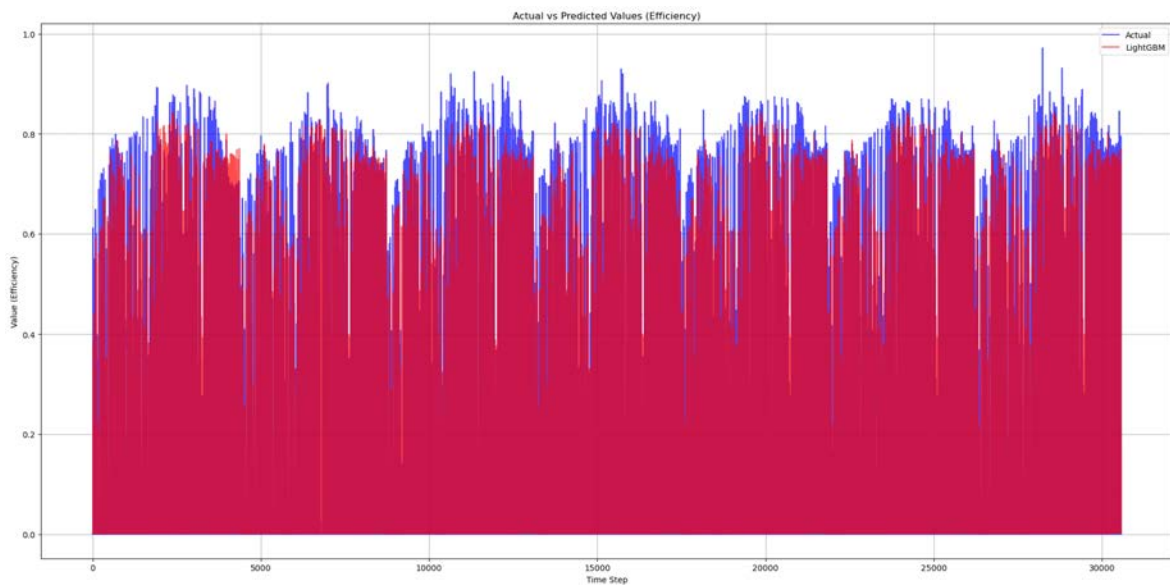


Figure 5.23: Comparison of actual and predicted efficiency values under temporal validation. LightGBM's leaf-wise growth strategy enables detailed pattern capture while maintaining computational efficiency.

The time series visualization reveals LightGBM's ability to track generation patterns with notable computational efficiency. While matching the prediction accuracy of other gradient boosting approaches, LightGBM achieves this performance through its unique tree growth strategy, which is evident in how it handles the parallel prediction of systems with different capacities.

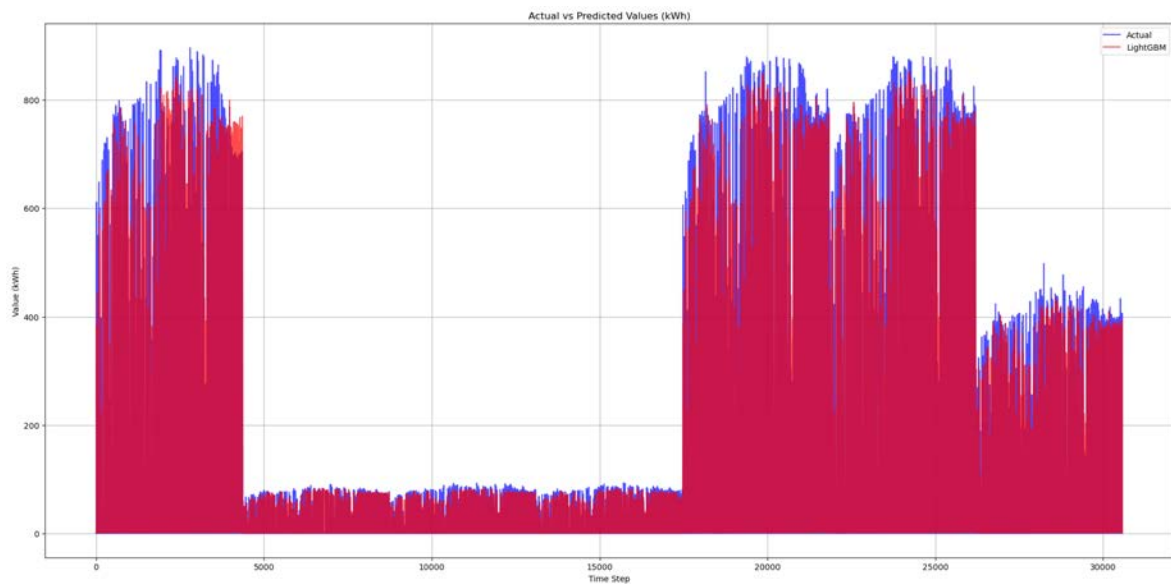


Figure 5.24: Actual versus predicted absolute power generation in kWh, demonstrating LightGBM’s ability to adapt predictions across different system scales while maintaining fast execution times.

The absolute generation predictions highlight LightGBM’s balanced handling of our diverse installation set. The model maintains consistent accuracy across different system scales, from our 100 kW installations to the larger 1000 kW systems. It showcases how its efficient leaf-wise growth strategy effectively adapts to varying generation capacities without compromising prediction quality.

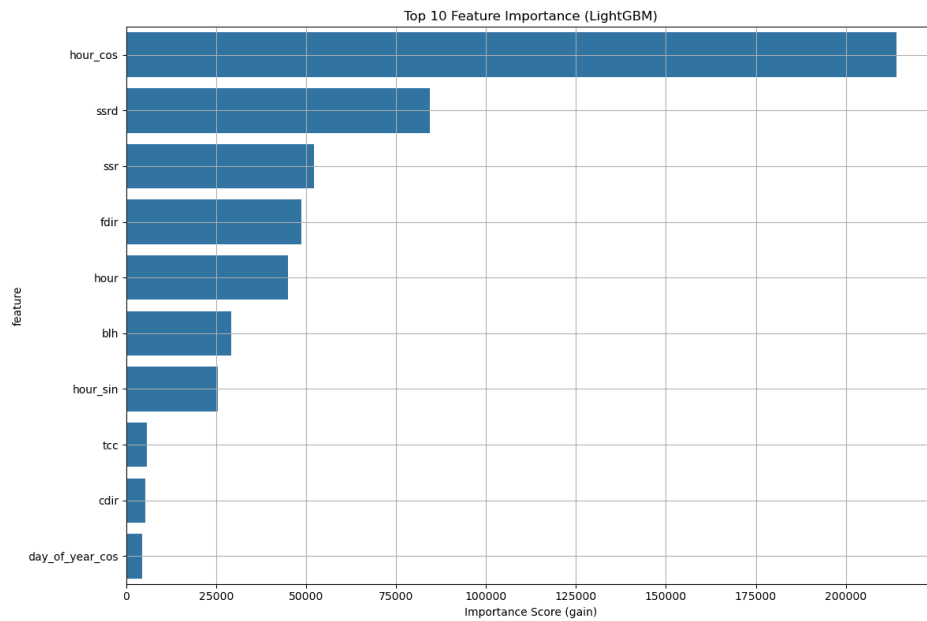


Figure 5.25: Feature importance distribution in LightGBM reveals an emphasis on `hour_cos` followed by meteorological parameters (`ssrd`, `ssr`). The model's leaf-wise growth strategy leads to a more pronounced separation between primary and secondary features.

The feature importance analysis reveals an interesting aspect of LightGBM's learning approach. While `hour_cos` maintains its primary role, the model exhibits a particularly steep drop-off in importance scores between primary and secondary features, likely due to its leaf-wise growth strategy. This hierarchical feature utilization, where `ssrd` and `ssr` follow as distant secondary factors, suggests that LightGBM's efficient tree-building approach leads to a more concentrated reliance on its most discriminative features.

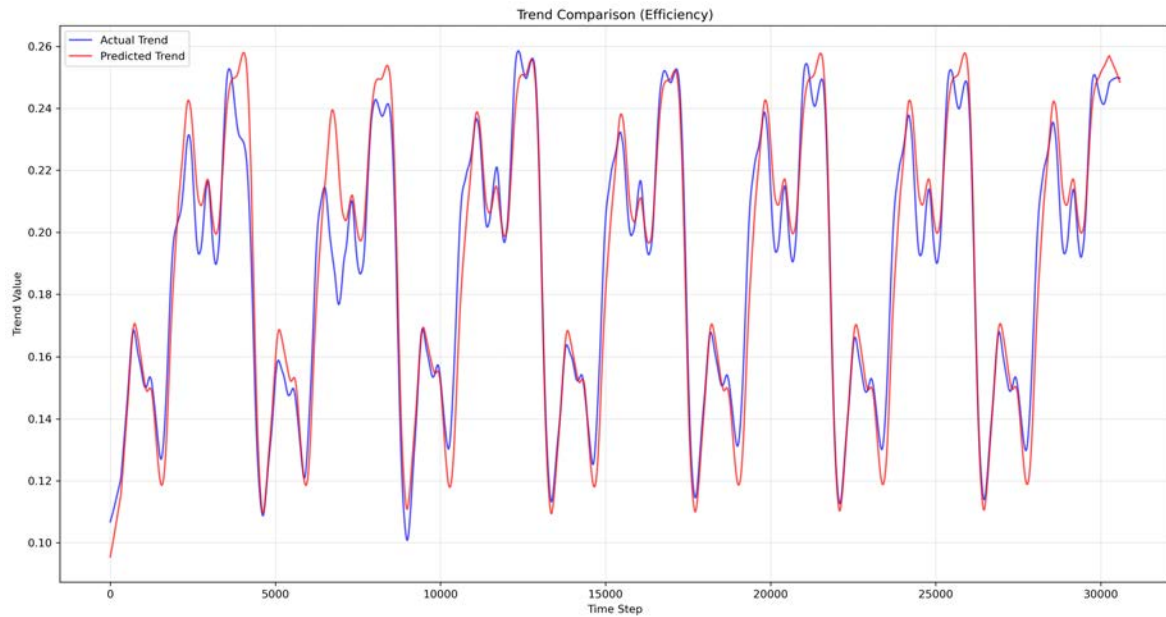


Figure 5.26: Trend comparison showing LightGBM’s characteristic handling of peak periods, with a slight tendency to overpredict maxima while maintaining accurate tracking of lower generation periods.

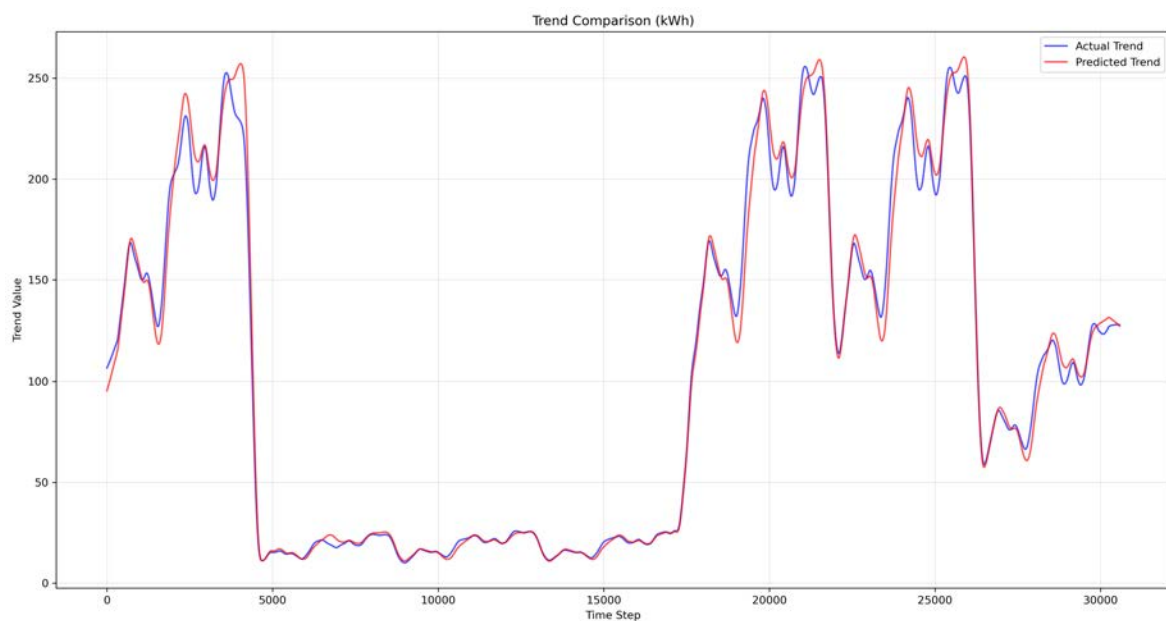


Figure 5.27: Absolute generation trend comparison demonstrating consistent prediction patterns across different system scales, with the model’s leaf-wise optimization maintaining accuracy while reducing computational overhead.

The trend analysis through STL decomposition reveals subtle characteristics of LightGBM’s prediction behavior. The model slightly tends to overpredict peak values while main-

taining precise tracking during lower generation periods. This pattern suggests its leaf-wise growth strategy may lead to more aggressive fitting at the extremes of the value range. While minor in impact, this behavior demonstrates how LightGBM's unique tree-building approach influences its prediction patterns even as it maintains overall accuracy comparable to other gradient boosting methods.

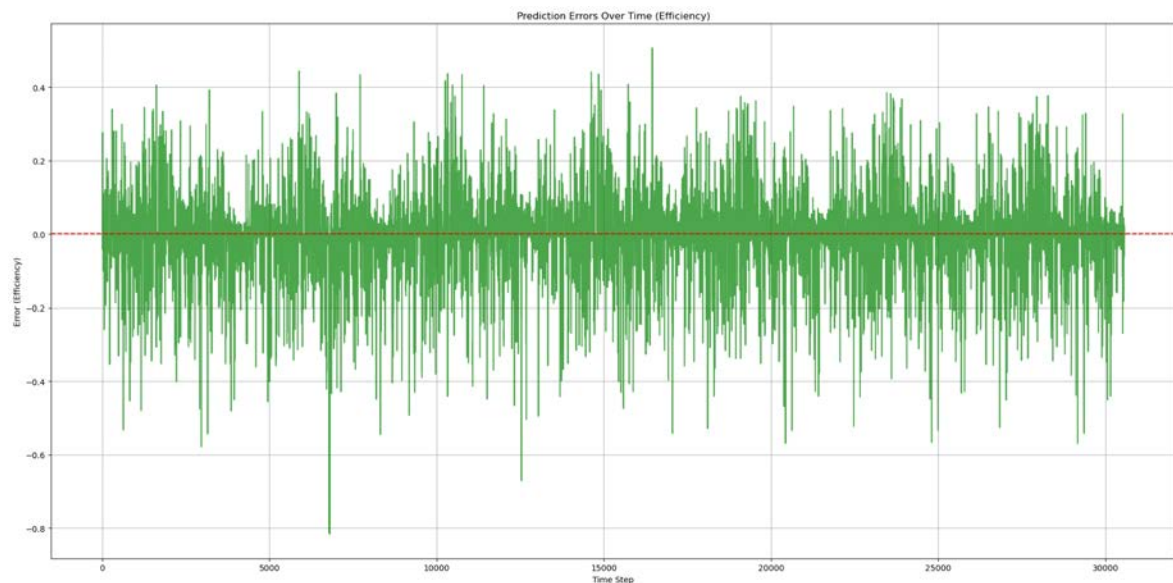


Figure 5.28: Error distribution pattern in efficiency predictions, showing LightGBM's balanced error distribution with characteristic spikes that reflect its leaf-wise growth strategy's handling of extreme events.

The error analysis reveals a consistent pattern that aligns with LightGBM's computational approach. The efficiency errors maintain a balanced distribution around zero, typically contained within ± 0.2 efficiency units, with occasional excursions reaching ± 0.4 during challenging prediction scenarios. These error patterns, obvious during transition periods, reflect how LightGBM's leaf-wise tree growth strategy handles rapid changes in generation conditions.

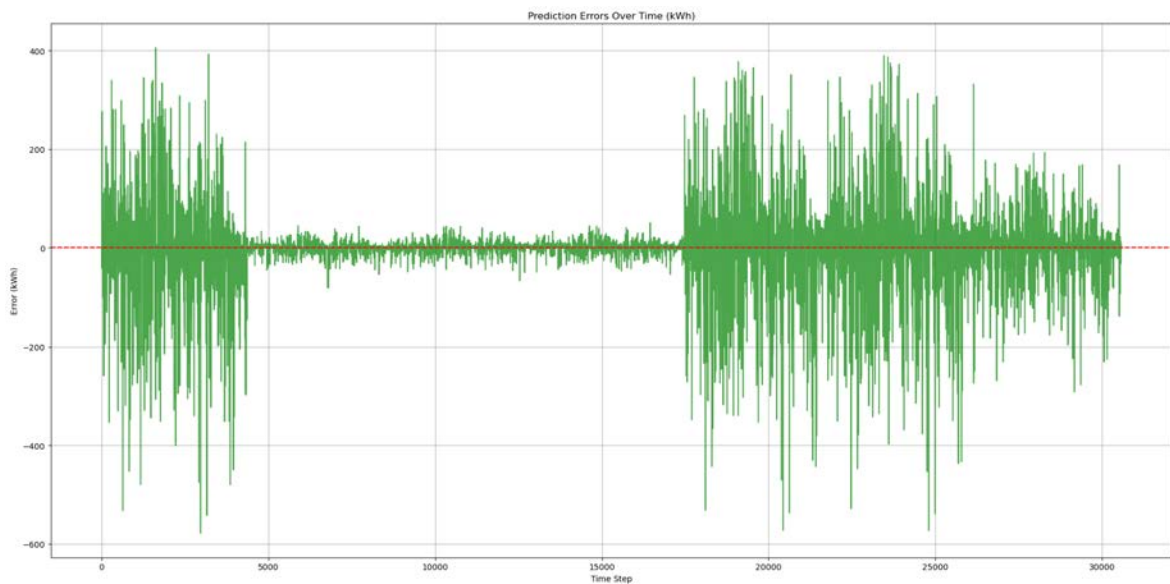


Figure 5.29: Absolute generation error patterns reflecting the diverse system sizes in our dataset, with error magnitudes naturally scaling with installation capacity while maintaining consistent relative accuracy.

LightGBM's temporal validation performance demonstrates how its efficient computational approach can maintain prediction accuracy while reducing computational overhead. The model's 46.97% improvement over baseline, achieved through its distinctive leaf-wise growth strategy and concentrated feature utilization, suggests that its streamlined approach to tree construction effectively captures the fundamental patterns driving photovoltaic generation. This balance of computational efficiency and prediction accuracy makes it particularly suitable for large-scale deployment scenarios where processing speed and prediction performance are critical considerations.

System-based Validation Performance

In system-based validation, where LightGBM predicts generation patterns for an entirely unseen installation (System 74894, 998.955 kW), the model demonstrates strong predictive capabilities while maintaining its characteristic computational efficiency. The model achieves an efficiency RMSE of 0.0446 and MAE of 0.0216, with an R^2 value of 0.9722. These efficiency metrics for this particular system translate to absolute generation errors of 44.56 kWh RMSE and 21.61 kWh MAE, representing a 68.24% improvement over the baseline approach.

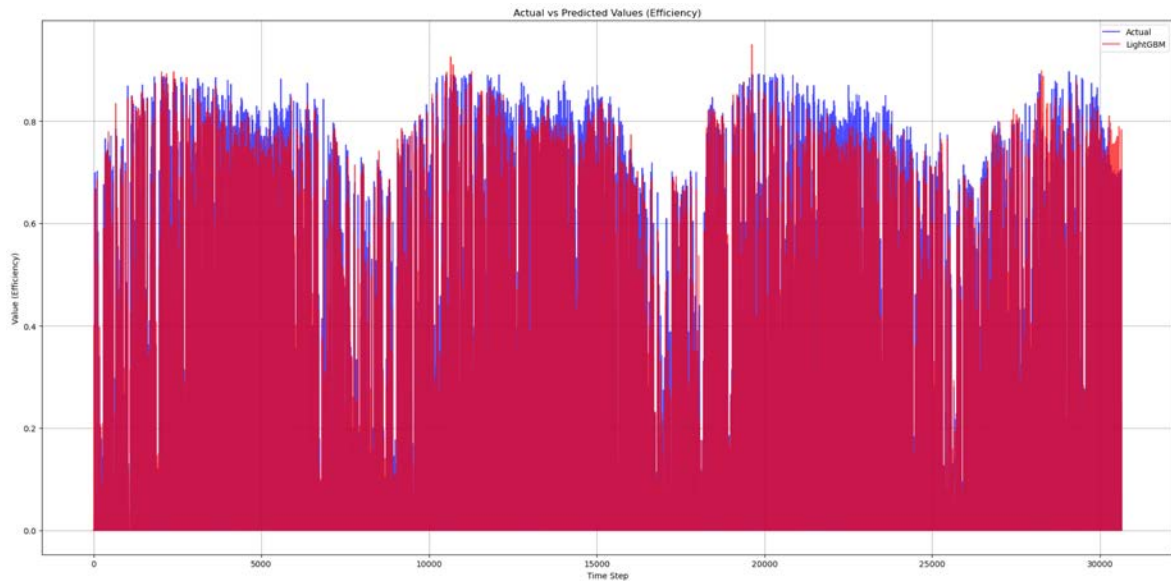


Figure 5.30: Comparison of actual and predicted efficiency values for System 74894. LightGBM's predictions show remarkably balanced tracking across the full range of generation conditions, demonstrating more neutral predictions than other gradient boosting approaches.

The time series visualization reveals an interesting aspect of LightGBM's generalization capabilities. Unlike other gradient boosting approaches that tend toward underprediction, LightGBM maintains more balanced predictions across all generation values. This balanced prediction pattern suggests that its leaf-wise growth strategy and efficient tree construction approach may help avoid systematic bias when generalizing to unseen system configurations.

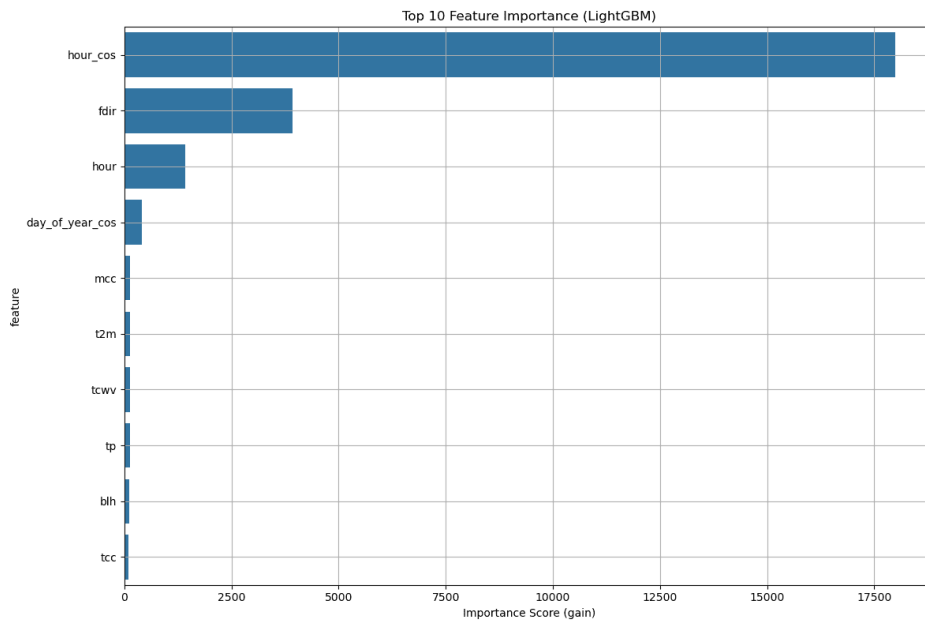


Figure 5.31: Feature importance analysis reveals LightGBM’s highly focused feature utilization strategy when generalizing to an unseen system, with `hour_cos` and `fdir` emerging as the dominant predictors while other features show minimal influence.

The feature importance analysis provides insight into how LightGBM adapts its prediction strategy for an unseen system. The model shows an even more pronounced concentration on key temporal and radiation features than in temporal validation, with `hour_cos` and `fdir` emerging as the clear dominant predictors. This extreme feature prioritization, characteristic of LightGBM’s efficient learning approach, contributes to its balanced prediction behavior while maintaining computational speed advantages.

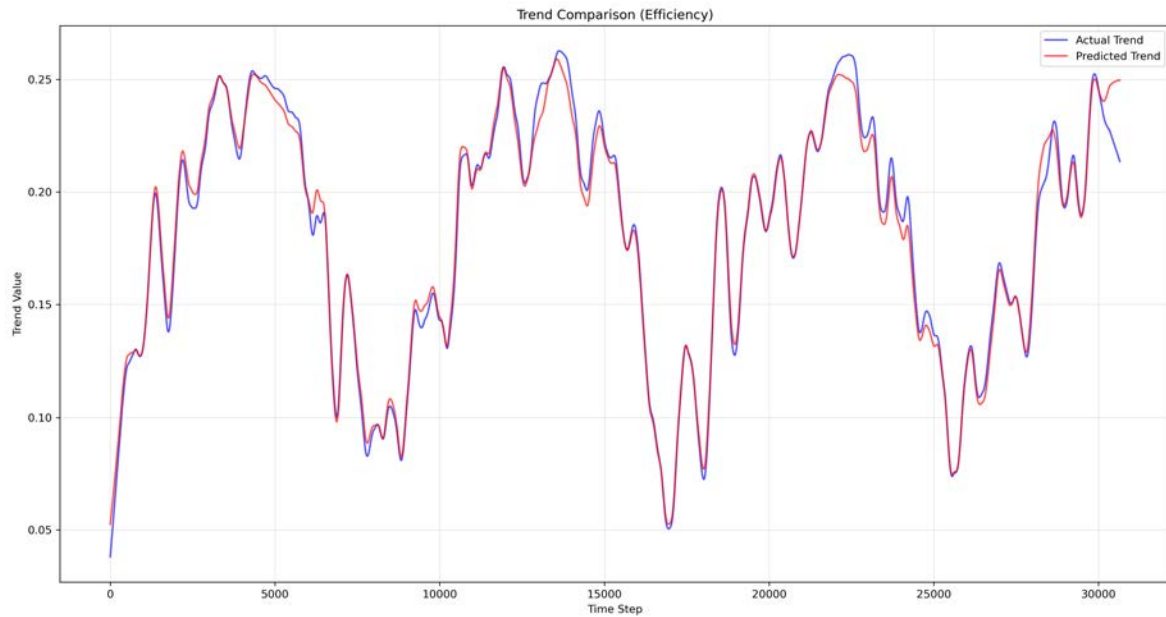


Figure 5.32: Trend comparison of actual versus predicted efficiency values for System 74894. The decomposed trends demonstrate LightGBM’s capacity to track gradual and rapid changes in generation patterns while maintaining its computational efficiency advantages.

The trend analysis through STL decomposition reveals how LightGBM’s leaf-wise growth strategy translates into long-term prediction patterns. The decomposed efficiency trends show remarkably close alignment between predicted and actual values, with the model successfully capturing both gradual seasonal variations and shorter-term generation patterns. This performance suggests that LightGBM’s focused feature utilization strategy, particularly its heavy reliance on `hour_cos` and `fdir`, provides sufficient information to reconstruct underlying generation patterns even for an unseen system.

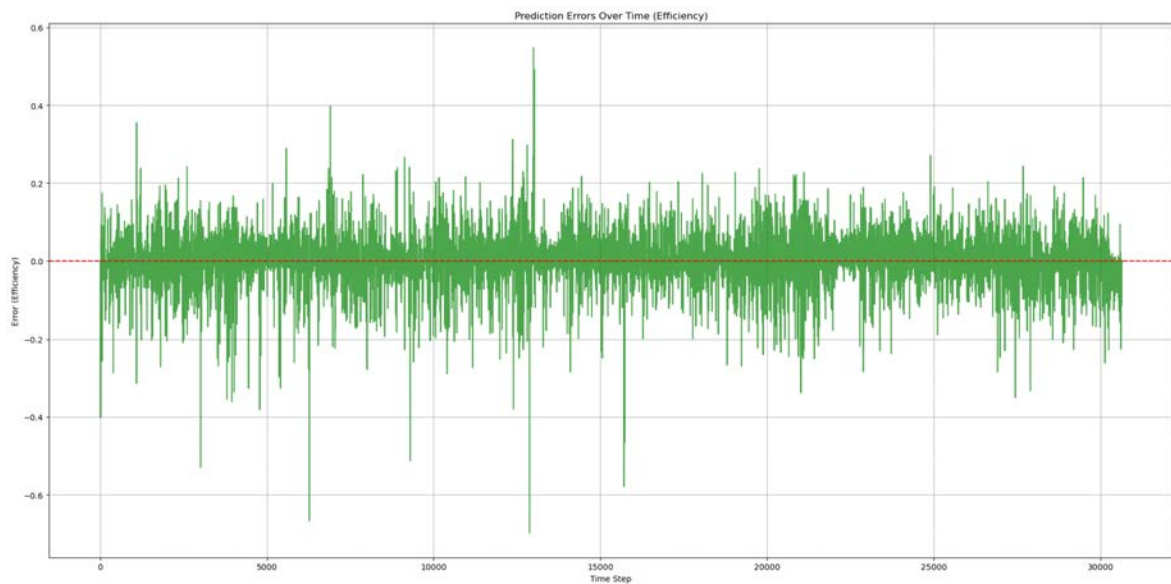


Figure 5.33: Temporal distribution of prediction errors for System 74894. The error pattern reveals LightGBM’s balanced prediction behavior, with most deviations contained within ± 0.2 efficiency units and only occasional larger excursions during challenging conditions.

Residual error analysis confirms these findings and adds to the understanding of LightGBM’s predictive behavior. Most errors occur in a small area within ± 0.2 efficiency units, where the spread is significantly balanced around zero. Although random larger variations up to ± 0.6 can occur under the most demanding conditions, they correspond to only a small percentage of the predictions. This pattern of errors, in conjunction with the model’s phenomenal 68.24% gain over baseline, shows how LightGBM’s ability to learn efficiently makes it effective at generalizing to unseen system configurations, without sacrificing its usual computational efficiency.

5.2.4 CatBoost Results

Temporal Validation Performance

On the basis of temporal validation, where CatBoost models generation patterns for 2024 using 2021-2023 training data, the model reveals interesting properties through its cascaded boosting process. The model obtains an efficiency RMSE value of 0.0778 and MAE value of 0.0376, and an R^2 value of 0.9154. These efficiency measures correspond to absolute generated values of 52.07 kWh RMSE and 20.24 kWh MAE, which is a 46.66% increase compared to baseline.

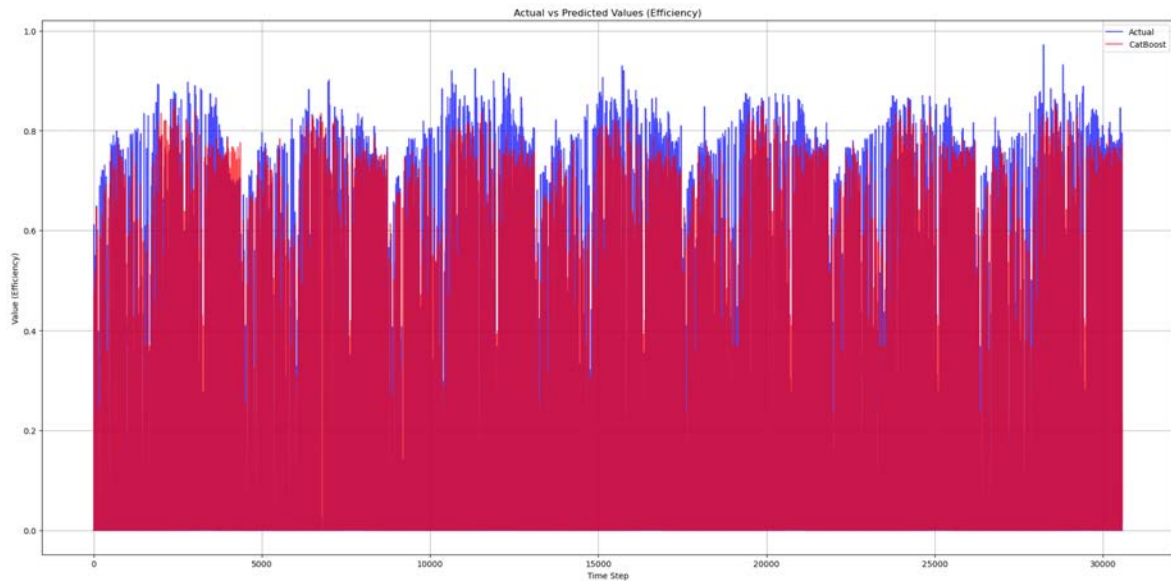


Figure 5.34: Comparison of actual and predicted efficiency values under temporal validation. CatBoost’s native handling of categorical variables and ordered boosting approach shows through in its prediction patterns.

The efficiency time series visualization reveals CatBoost’s distinctive handling of the prediction task. As shown in Figure 5.34, the model maintains consistent tracking of generation patterns while treating system identifiers as native categorical variables, a unique feature of CatBoost that influences its adaptation to different system characteristics.

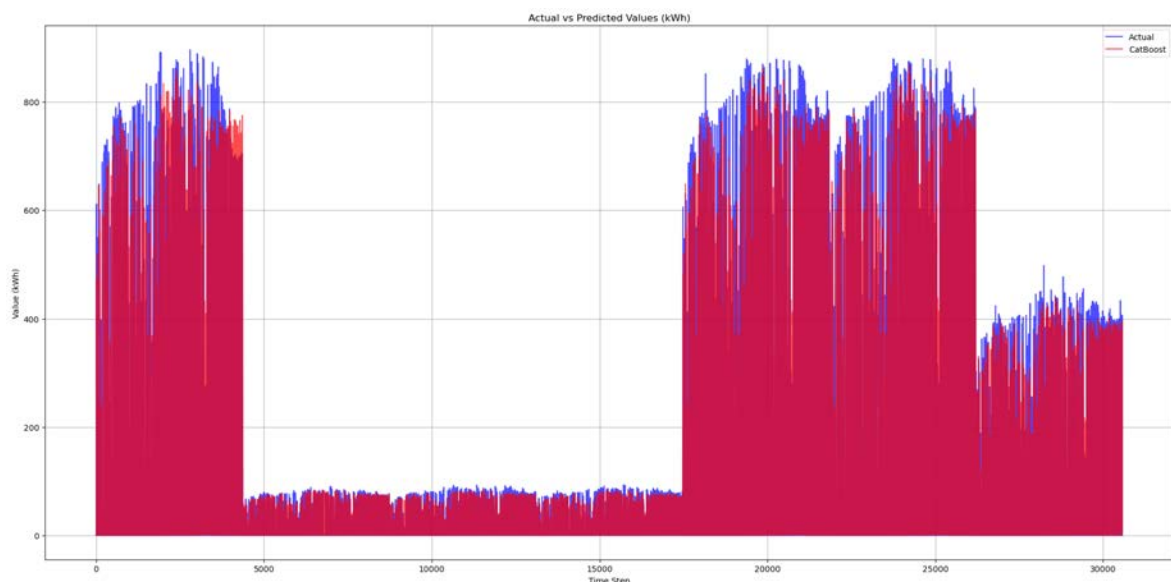


Figure 5.35: Actual versus predicted absolute power generation in kWh. The visualization demonstrates how CatBoost’s categorical handling approach translates across different system scales, from 100 kW to 1000 kW installations.

When examining the absolute generation predictions (Figure 5.35), we can observe how CatBoost’s approach scales across our diverse installations. The model maintains consistent prediction accuracy across different system sizes, from our 100 kW systems to the larger 1000 kW installations. This scaling behavior demonstrates how CatBoost’s native categorical handling enables it to learn system-specific characteristics while maintaining its ordered boosting advantages.

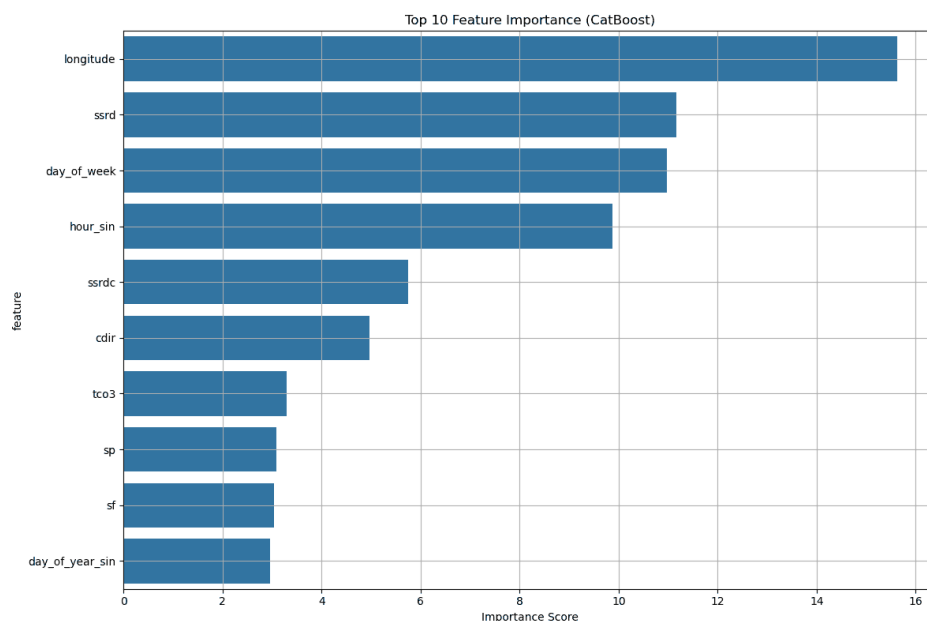


Figure 5.36: Feature importance analysis reveals CatBoost’s unique feature prioritization strategy, with geographical and radiation parameters playing more prominent roles than other gradient boosting approaches.

The most revealing side of CatBoost’s method appears in the analysis of its feature importance. In contrast to other gradient boosting techniques, which give time feature most weight, CatBoost exhibits a significantly different pattern of feature usage. Longitude emerges as the dominant predictor, followed by solar radiation parameters (ssrd) and temporal features like day_of_week and hour_sin. This distinctive feature hierarchy suggests that CatBoost’s ordered boosting mechanism and categorical handling lead it to discover different relationships in the data compared to other approaches. However, this doesn’t translate to superior temporal prediction performance.

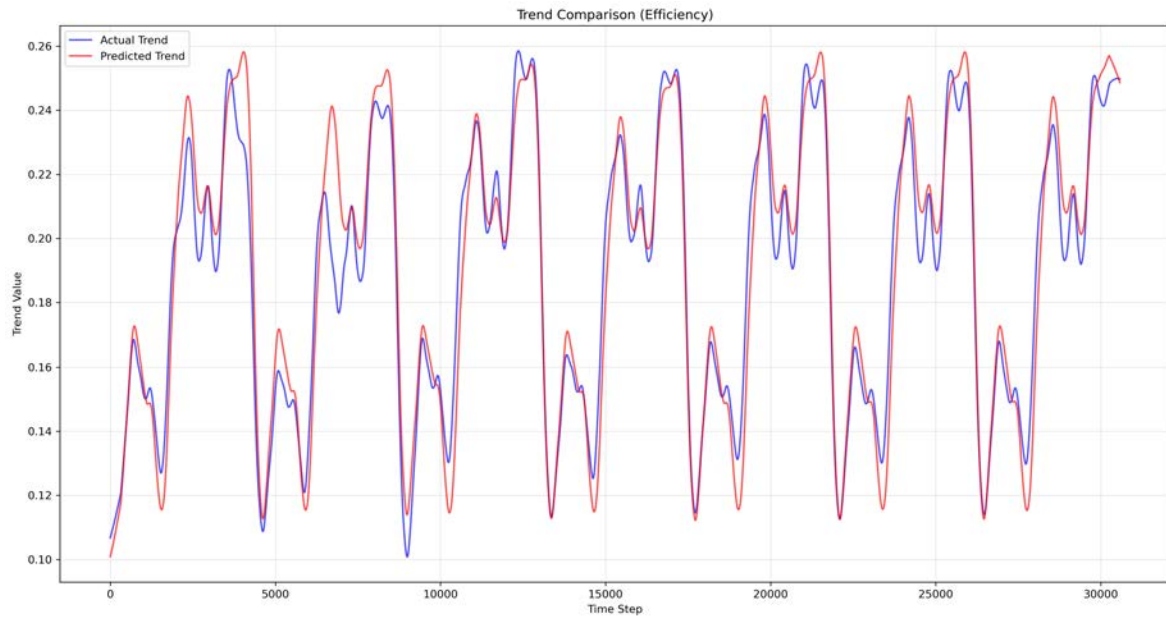


Figure 5.37: Trend comparison showing CatBoost’s characteristic tendency to amplify extremes, with slight overprediction during peaks and underprediction during troughs. This pattern reflects its ordered boosting approach to handling categorical variables.

The trend analysis via STL decomposition highlights a notable feature of CatBoost’s prediction behavior. The efficiency trends exhibit a regular pattern where the model tends to amplify extremes—slightly exceeding peak efficiency values while undervaluing during lower generation phases. This behavior is likely a result of CatBoost’s ordered boosting method and its distinctive approach to categorical variables, which enables it to construct predictions sequentially while considering system-specific traits.

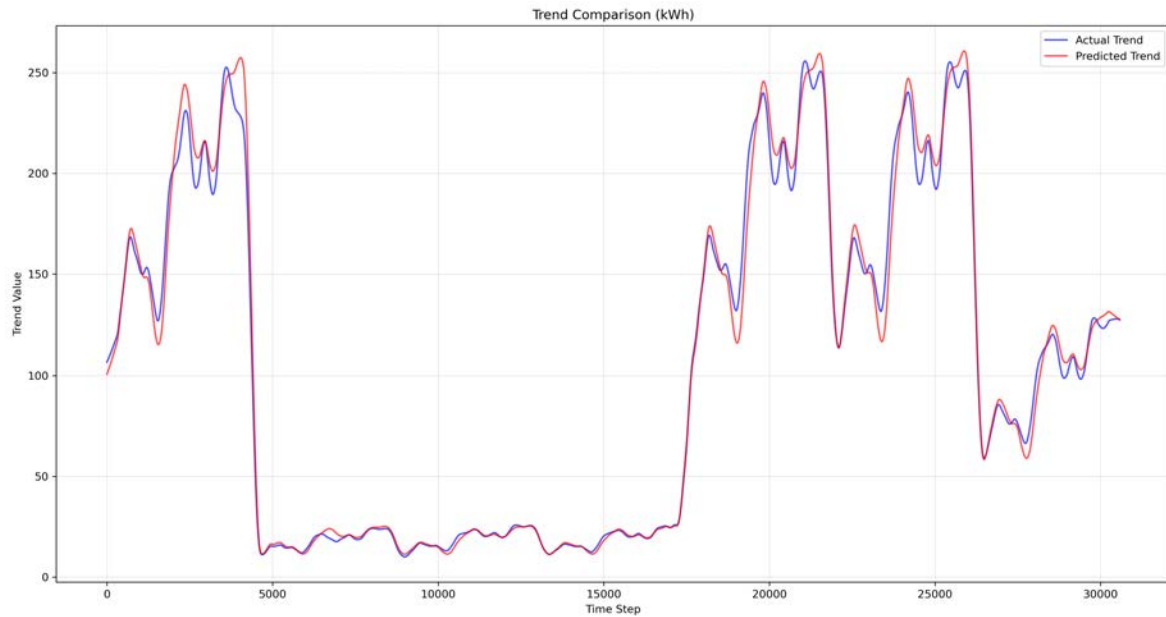


Figure 5.38: Absolute generation trend comparison demonstrating how CatBoost’s amplification of extremes scales with system size, particularly visible during peak generation periods of larger installations.

This amplification effect becomes even more apparent across different system scales when examining absolute generation trends. The model’s tendency to exaggerate extremes is particularly noticeable during high-generation periods for larger installations, though it maintains relatively accurate tracking during intermediate-generation conditions. This pattern suggests that CatBoost’s focus on geographical and radiation parameters in its feature importance may make it more sensitive to extreme conditions compared to models that prioritize temporal features.

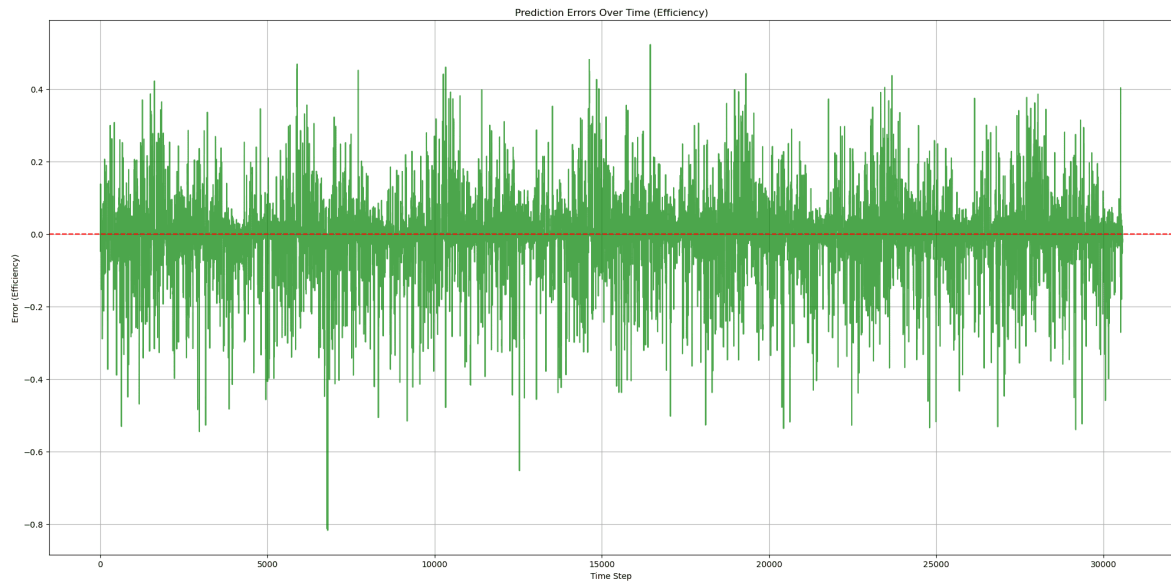


Figure 5.39: The Distribution of prediction errors in efficiency terms shows CatBoost’s characteristic error patterns.

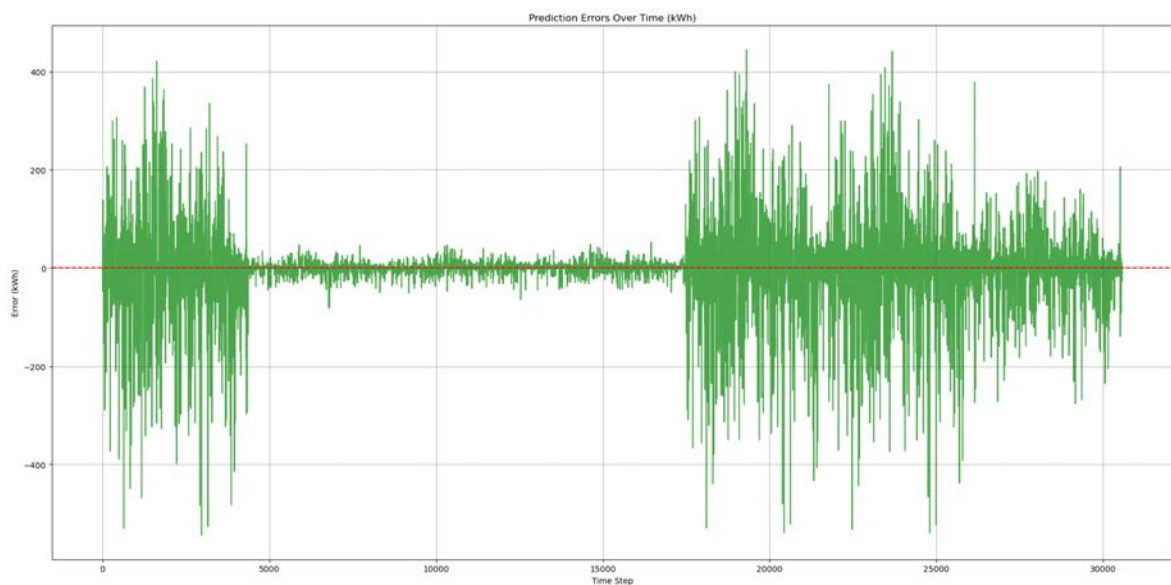


Figure 5.40: Absolute generation prediction errors demonstrating how deviations scale with system size.

The error analysis reinforces CatBoost’s distinctive prediction behavior in temporal validation. The efficiency errors maintain a relatively symmetric distribution around zero, with most deviations contained within ± 0.2 efficiency units, though showing slightly larger excursions during challenging conditions than other approaches. When examining these patterns in absolute generation terms, the error magnitudes naturally scale with system size, with more

significant deviations corresponding to our 1000 kW installations and proportionally more minor errors for 100 kW systems. This scaling of errors, combined with CatBoost's 46.66% improvement over baseline and its unique feature utilization strategy, suggests that while the model's ordered boosting approach offers an interesting alternative perspective on the prediction task, it doesn't translate to superior temporal prediction performance compared to other gradient boosting methods.

System-based Validation Performance

In system-based validation, CatBoost predicts generation patterns for a completely new installation (System 74894, 998.955 kW). The model shows exceptional predictive performance that distinguishes it from other methods. It reaches an efficiency RMSE of 0.0276, an MAE of 0.0109, and an impressive R^2 of 0.9893. For this system, these efficiency metrics translate to absolute generation errors of 27.61 kWh RMSE and 10.86 kWh MAE, representing an impressive 80.32% improvement over the baseline approach.

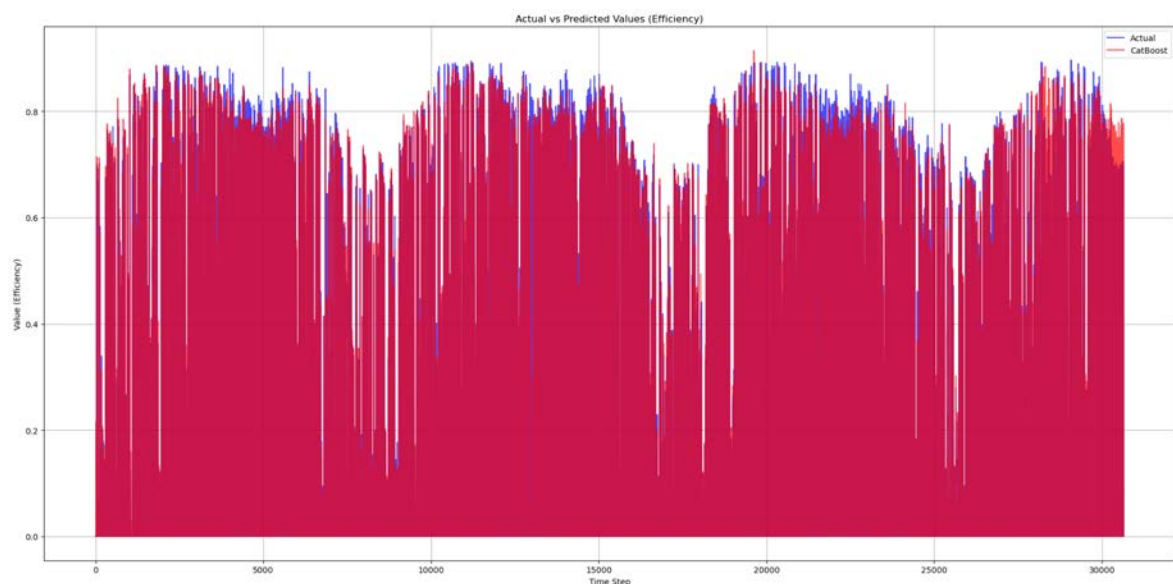


Figure 5.41: Comparison of actual and predicted efficiency values for System 74894. The remarkable correspondence between predicted and true values show how CatBoost generalizes exceptionally to unseen system configurations.

The time series visualization reveals CatBoost's ability to generalize to an unseen system. Unlike its temporal validation performance, the model demonstrates precise tracking of generation patterns, suggesting that its native handling of categorical variables particularly excels when focusing on a single, previously unseen system configuration.

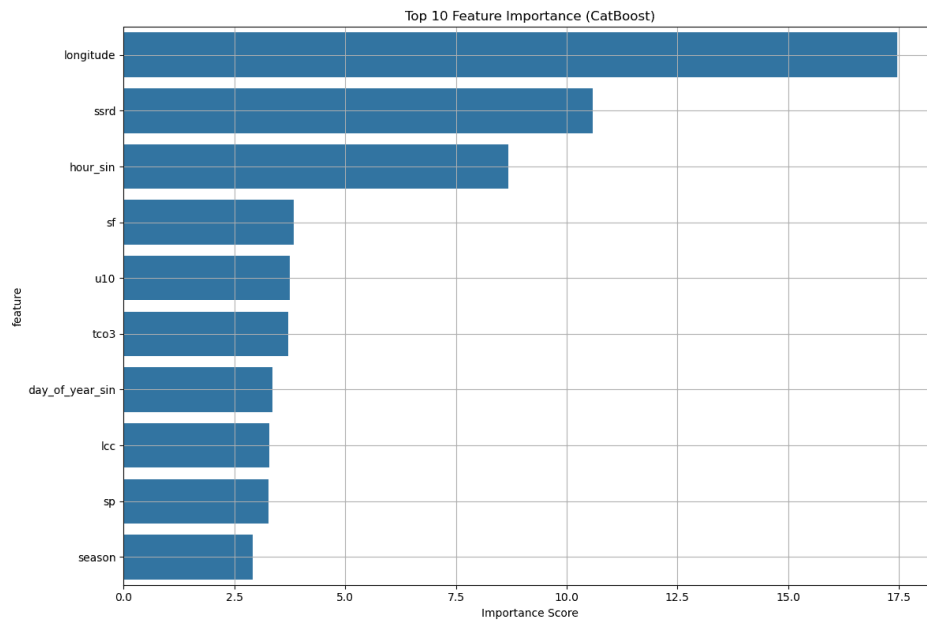


Figure 5.42: Feature importance analysis reveals CatBoost’s unique approach to system-based prediction, with geographical features playing a central role alongside radiation and temporal parameters. This distinctive feature utilization pattern helps explain its superior generalization performance.

The analysis of feature importance reveals key insights into CatBoost’s outstanding performance. In a notable contrast to other methods, longitude stands out as the primary predictor, succeeded by solar radiation (ssrd) and hour_sin. This distinctive feature hierarchy, where geographical location plays a central role alongside environmental and temporal parameters, suggests that CatBoost’s ordered boosting mechanism discovers fundamentally different relationships when generalizing to new systems. The model appears to anchor its predictions in spatial context first, then refining them through radiation and temporal features - a remarkably effective strategy for system-based prediction.

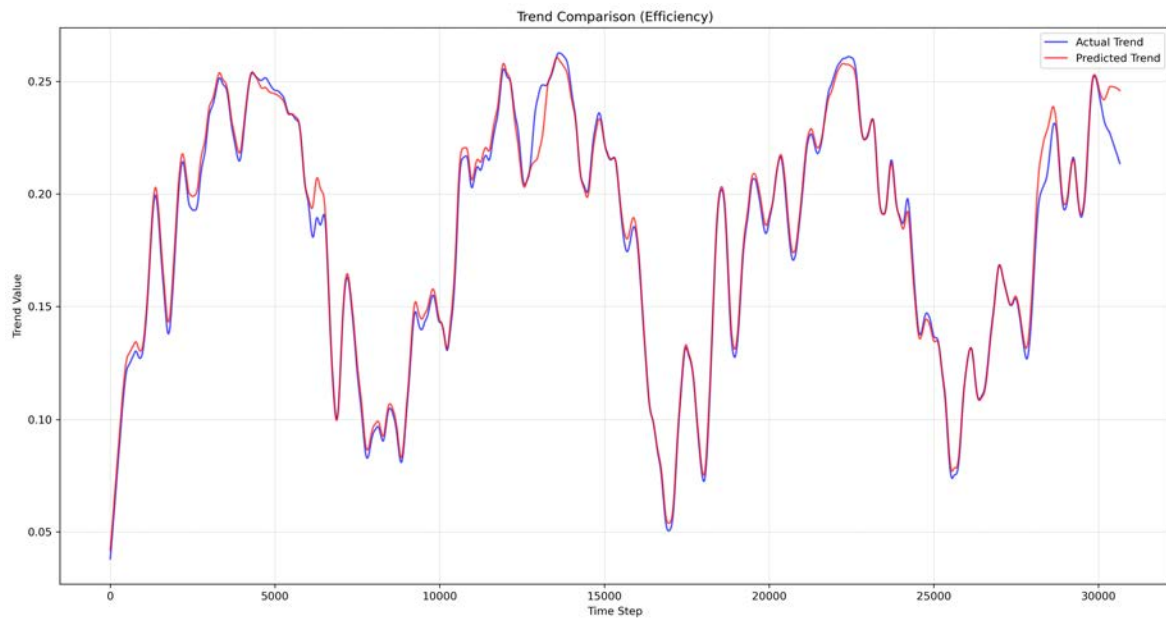


Figure 5.43: Trend comparison showing CatBoost’s precise tracking of generation patterns for System 74894. The decomposed trends demonstrate near-perfect alignment between predicted and actual values throughout the prediction period.

The trend analysis through STL decomposition further reinforces CatBoost’s outstanding performance in system-based prediction. The decomposed efficiency trends show exceptionally precise alignment between predicted and actual values, with the model successfully capturing both gradual seasonal variations and shorter-term generation patterns. This remarkable trend-following capability suggests that CatBoost’s unique approach of anchoring predictions in geographical context before refining them with radiation and temporal features creates a particularly robust foundation for generalizing to unseen systems.

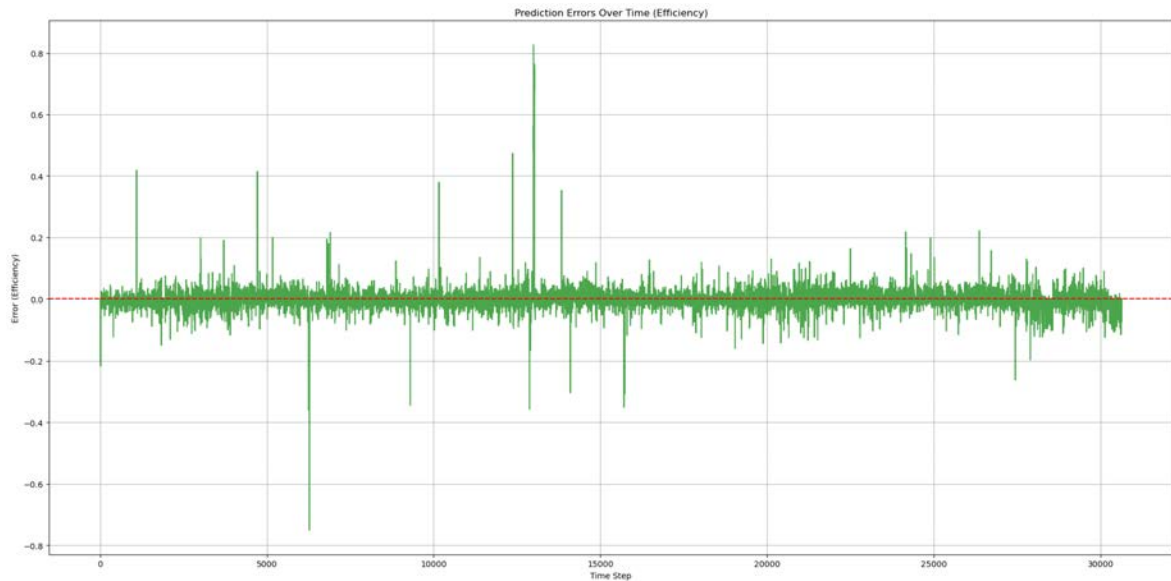


Figure 5.44: Error distribution showing CatBoost’s characteristically tight error bounds in system-based prediction, with only rare excursions beyond ± 0.2 efficiency units.

The error analysis exposes the other unique characteristic of CatBoost’s system-based performance. Residuals show a remarkably good clustering around zero with few predictions differing significantly from actual values. While occasional more extensive excursions occur, reaching up to ± 0.4 efficiency units in rare cases, these represent isolated events rather than systematic errors. This error pattern, combined with the model’s exceptional 80.32% improvement over baseline and its unique feature utilization strategy, suggests that CatBoost’s ordered boosting approach is particularly well-suited for system-based prediction tasks where geographical context plays a crucial role in understanding generation patterns.

5.3 Deep Learning Model Performance

5.3.1 Temporal Fusion Transformer Results

The Temporal Fusion Transformer (TFT) significantly differs from traditional tree-based approaches, offering a sophisticated architecture designed explicitly for temporal sequence modeling. Its multi-head attention mechanisms enable it to process variable-length sequences and capture dependencies at multiple time scales. At the same time, its variable selection networks dynamically adjust feature importance across different prediction horizons. This architecture is particularly well-suited for photovoltaic generation prediction, where both short-

term weather effects and long-term seasonal patterns influence system performance.

Temporal Validation Performance

TFT shows remarkable predictive performance under temporal validation, distinguishing it from tree-based methods. The model reaches an efficiency RMSE of 0.0568, an MAE of 0.0237, and an impressive R^2 of 0.9559. These metrics correspond to absolute generation values of 38.76 kWh RMSE and 12.92 kWh MAE (with an R^2 of 0.9635), indicating a significant 61.07% enhancement over the baseline method.

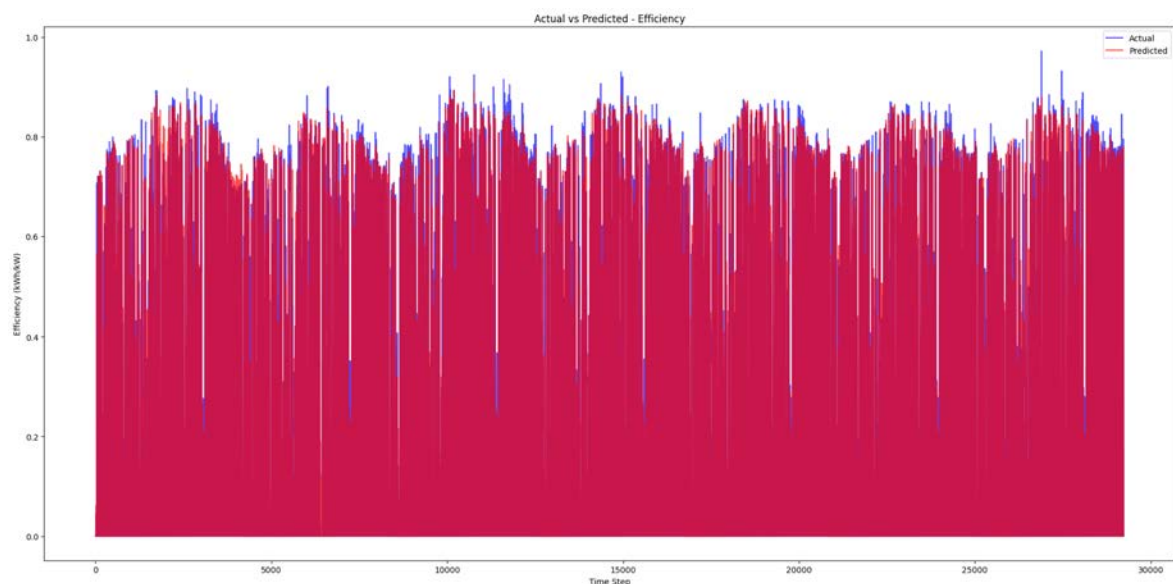


Figure 5.45: Comparison of actual and predicted efficiency values under temporal validation. TFT's attention mechanisms enable exact tracking of generation patterns, with powerful performance during transition periods and complex weather conditions.

The efficiency time series visualization reveals TFT's exceptional ability to capture complex temporal patterns. The model's attention mechanisms prove particularly effective during challenging prediction scenarios, showing remarkable consistency in tracking efficiency values from near-zero during non-generation hours to peaks approaching 0.95 during optimal conditions. This sophisticated temporal modeling is especially evident in the model's ability to accurately capture the sharp daily transitions and more subtle variations in peak generation efficiency across different weather conditions.

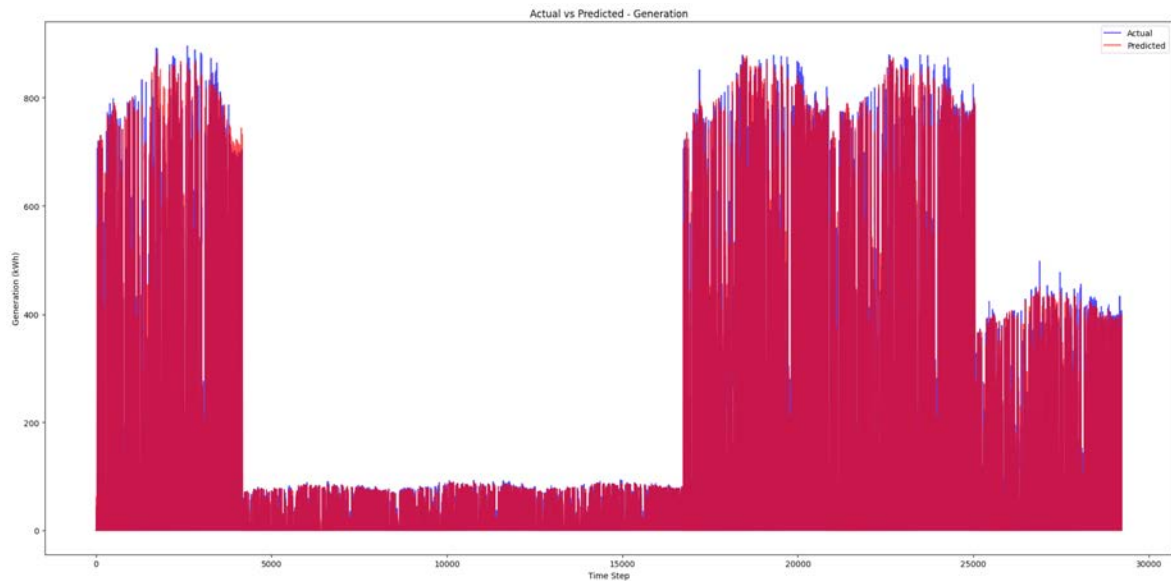


Figure 5.46: Actual versus predicted absolute power generation in kWh. The visualization demonstrates TFT's ability to maintain prediction accuracy across different system scales while capturing intricate temporal dependencies.

When examining the absolute generation predictions, TFT's sophisticated temporal modeling capabilities become even more apparent. The visualization clearly shows three distinct generation regimes: high-capacity systems generating up to 800 kWh during peak conditions, mid-range systems operating around 400 kWh, and smaller installations consistently generating under 100 kWh. The model maintains exceptional accuracy across all these scales, demonstrating its ability to learn and apply system-specific characteristics and temporal patterns simultaneously.

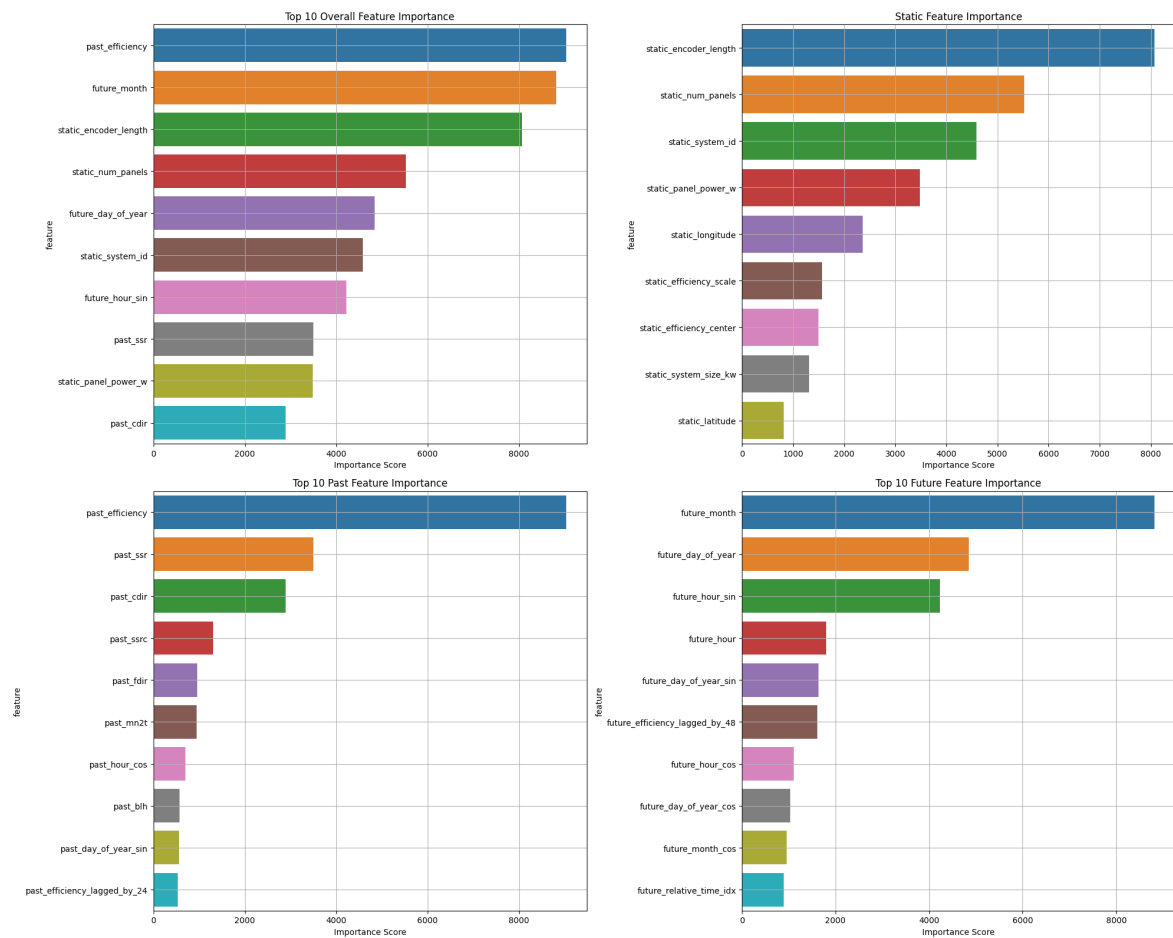


Figure 5.47: Comprehensive feature importance analysis from TFT's variable selection networks showing: (Top Left) Overall feature rankings across all categories, (Top Right) Static feature importance revealing system-specific characteristics, (Bottom Left) Encoder feature importance showing historical context utilization, and (Bottom Right) Decoder feature importance demonstrating forward-looking pattern recognition.

TFT's variable selection networks provide a multi-faceted view of how different features contribute to the prediction process. The comprehensive feature importance analysis, shown in Figure 5.47, reveals distinct patterns across four key perspectives: overall importance, static features, encoder inputs, and decoder inputs.

In the overall ranking (top left), encoder inputs (past_efficiency) emerge as the dominant predictor, followed by decoder features (future_month) and static system characteristics. This hierarchy demonstrates TFT's balanced approach to prediction, where historical context from the encoder is combined with known future information from the decoder and static system attributes.

The static feature analysis (top right) provides insight into how the model establishes the

foundational context shared by the encoder and decoder. The encoder length and number of panels are the most critical static features, followed by system identification and physical parameters like panel power. These features provide crucial context for both historical interpretation and future prediction.

Examining encoder feature importance (bottom left), we observe a clear temporal hierarchy in how historical information is processed. Recent efficiency values dominate, followed by environmental conditions (ssr, cdir). This demonstrates how TFT's attention mechanisms prioritize recent states while maintaining awareness of historical weather patterns. This encoder structure enables the model to represent the system's historical context better.

The decoder feature importance analysis (bottom right) reveals how the model utilizes known future information, with `future_month` and `future_day_of_year` emerging as key predictors. This indicates that TFT's decoder effectively incorporates known temporal patterns into its predictions, using this information to adjust forecasts based on expected cyclical variations in generation potential.

The trend analysis through STL decomposition provides further insight into TFT's prediction capabilities. As shown in Figure 5.48, the model demonstrates remarkable precision in capturing both efficiency and absolute generation trends. The efficiency trends reveal TFT's ability to track subtle variations, with values ranging from 0.125 to 0.250. In contrast, the absolute generation trends highlight how this performance scales across different system capacities, from periods of low generation around 20 kWh to peaks exceeding 250 kWh. The close alignment between predicted and actual trends in both metrics demonstrates TFT's robust temporal pattern recognition, enabled by its sophisticated encoder-decoder architecture and attention mechanisms.

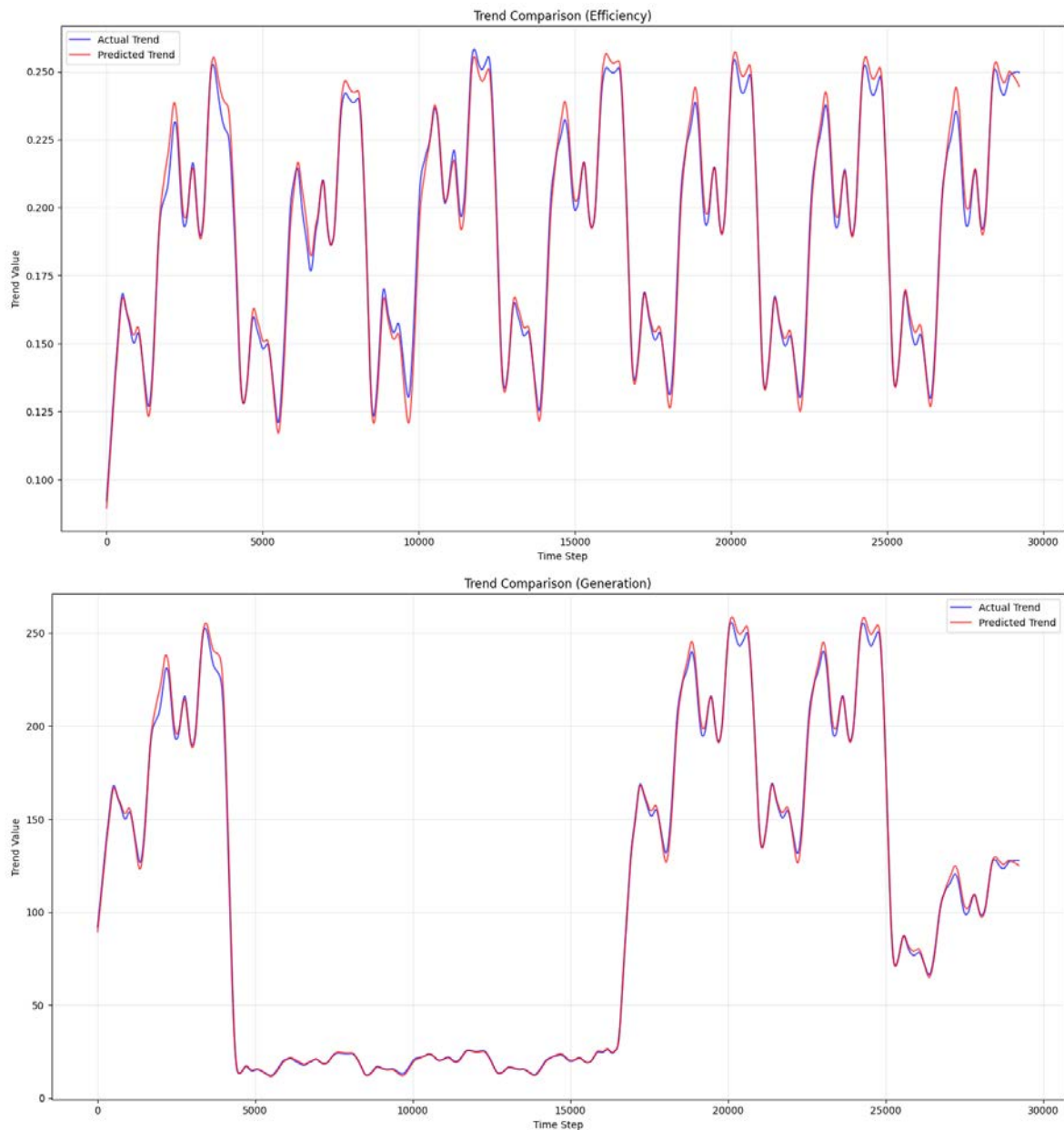


Figure 5.48: Trend comparison showing TFT's ability to capture both efficiency patterns (top) and their translation to absolute generation values (bottom). The decomposed trends demonstrate the model's consistent performance across different temporal scales and generation regimes.

The error analysis provides insight into TFT's prediction patterns across multiple dimensions. Examining the temporal distribution of errors (Figure 5.49), we observe that efficiency deviations remain contained mainly within ± 0.2 units, with occasional excursions reaching ± 0.4 during challenging conditions. These efficiency errors translate to absolute generation deviations that scale with system capacity, showing more considerable variations during high-

generation periods while maintaining proportionally smaller errors during low-generation periods.

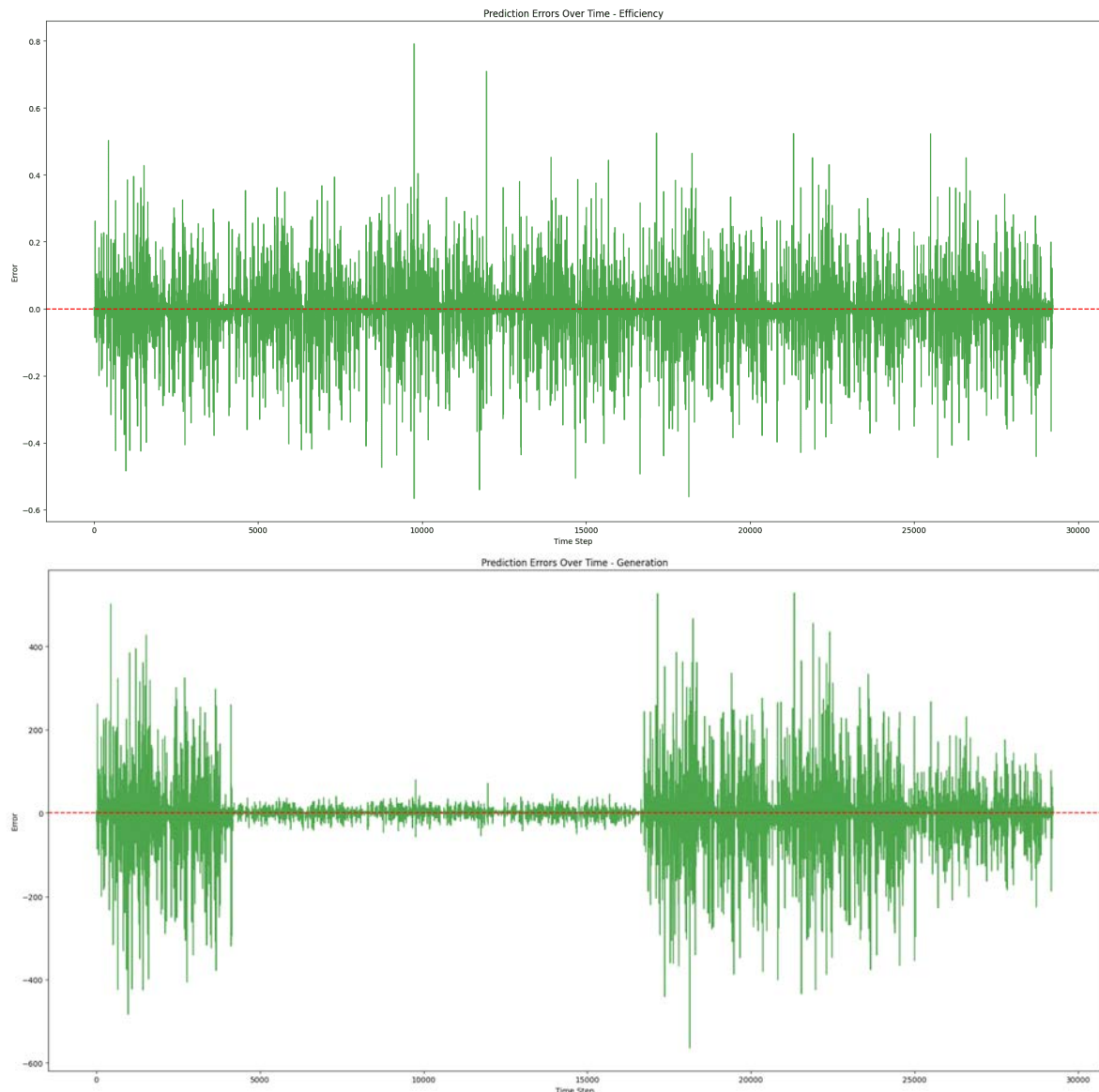


Figure 5.49: Temporal distribution of prediction errors showing: (top) efficiency errors maintaining consistent bounds throughout the prediction period, and (bottom) absolute generation errors demonstrating how deviations scale with system capacity.

The residual analysis (Figure 5.50) reveals how prediction accuracy varies with generation levels. For efficiency predictions, the residuals show a symmetric distribution around zero with slight heteroscedasticity - error variance increases moderately with predicted efficiency values. This pattern translates to the absolute generation residuals, where we observe similar symmetric behavior but with error magnitudes that naturally scale with predicted output levels. The consistent funnel shape in both plots suggests that while TFT maintains

unbiased predictions across all generation regimes, its uncertainty increases proportionally with expected values.

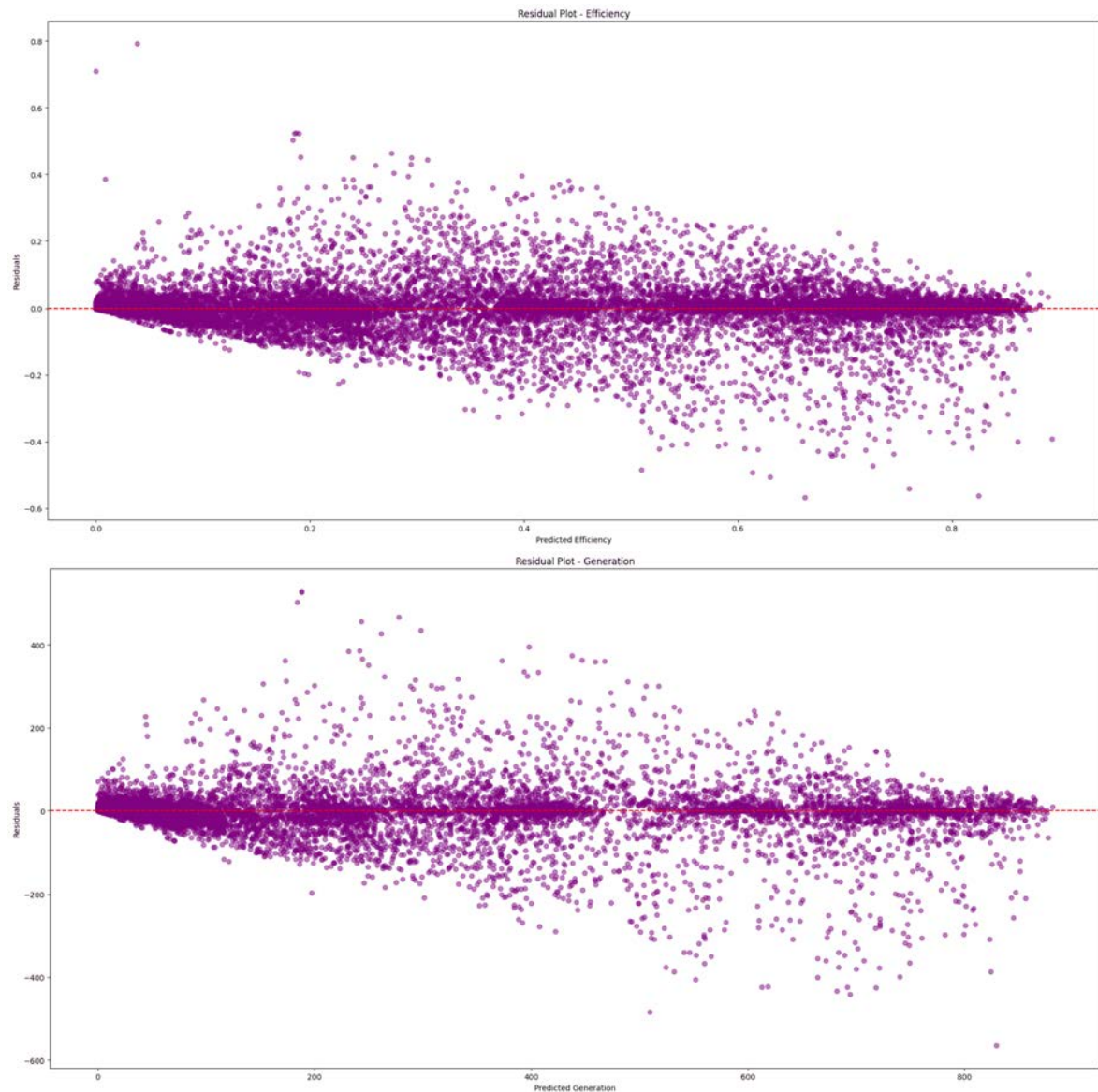


Figure 5.50: Residual analysis showing: (top) efficiency prediction errors versus predicted values demonstrating unbiased estimates with moderate heteroscedasticity, and (bottom) generation prediction errors showing similar patterns scaled by system capacity.

These comprehensive analyses demonstrate TFT's exceptional performance in temporal validation, with its 61.07% improvement over baseline reflecting the model's sophisticated handling of temporal dependencies. Encoder-decoder architecture, attention mechanisms, and dynamic feature selection enable TFT to capture short-term weather effects and long-term seasonal patterns while maintaining consistent accuracy across different system scales. This robust performance, evidenced by the stable error distributions and unbiased

residual patterns, suggests that TFT's architectural advantages translate directly into superior prediction capabilities for photovoltaic generation forecasting.

System-based Validation Performance

Under system-based validation, where TFT predicts generation patterns for an entirely unseen system (System 74894, 998.955 kW), the model maintains its strong predictive capabilities, achieving an efficiency RMSE of 0.0546 and MAE of 0.0220, with an R^2 value of 0.9585. These efficiency metrics translate to absolute generation errors of 54.54 kWh RMSE and 21.99 kWh MAE, representing a 62.58% improvement over the baseline approach. Notably, this performance is remarkably consistent with the model's temporal validation results, suggesting that TFT's encoder-decoder architecture enables robust generalization across both temporal and system-based prediction tasks.

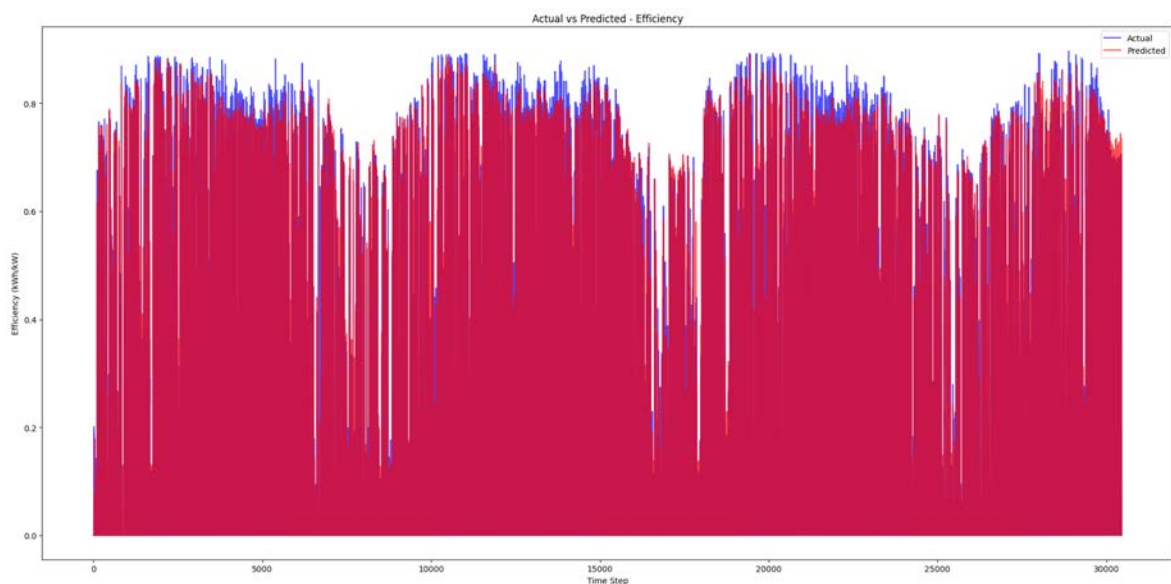


Figure 5.51: Comparison of actual and predicted efficiency values for System 74894. The close alignment between predictions and actual values demonstrates TFT's ability to maintain consistent performance when generalizing to an unseen system.

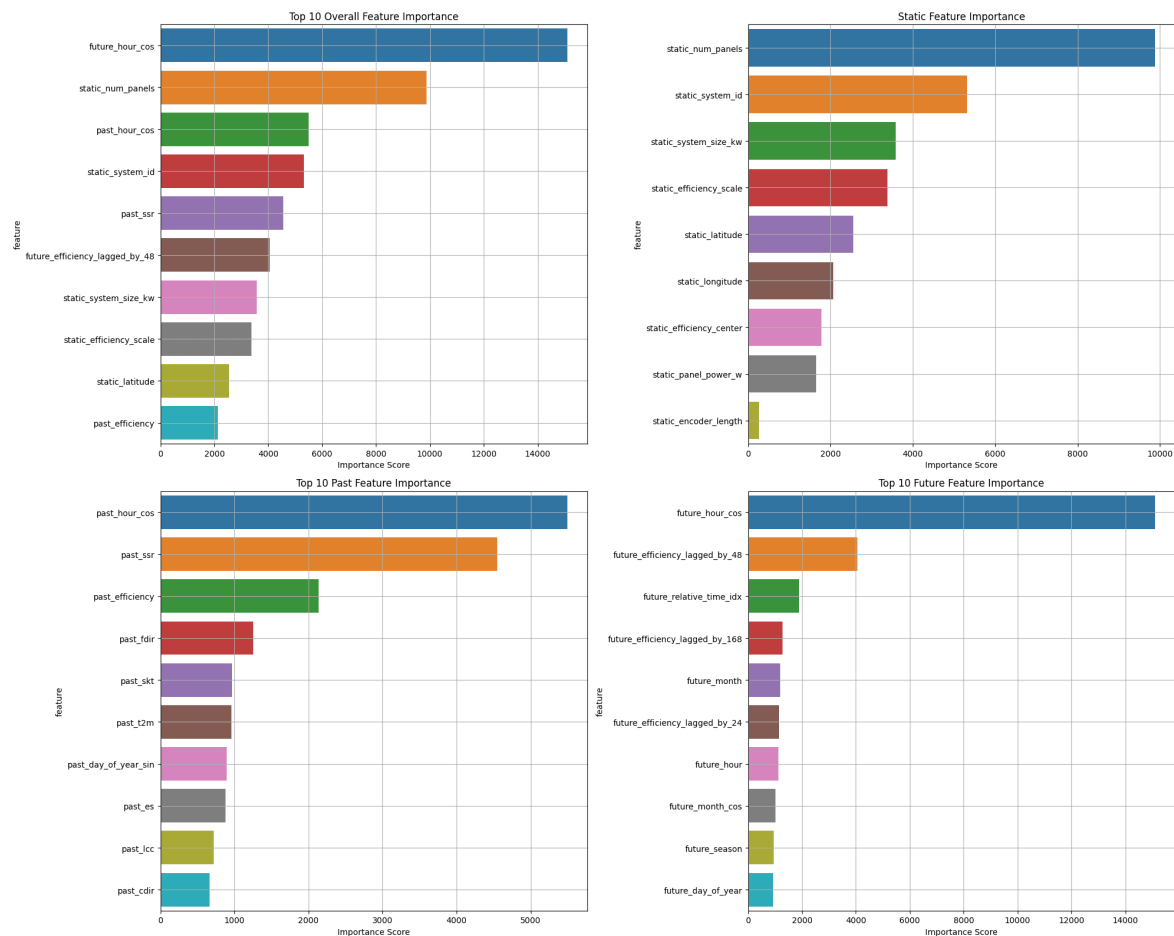


Figure 5.52: Feature importance analysis showing: (Top Left) Overall importance across feature categories, (Top Right) Static feature rankings, (Bottom Left) Encoder feature utilization, and (Bottom Right) Decoder feature importance. The balanced distribution across temporal and static features demonstrates TFT’s sophisticated approach to system-based prediction.

The feature importance analysis reveals interesting adaptations in TFT’s prediction strategy when facing an unseen system. However, these adaptations do not translate to the same level of performance improvement seen in tree-based models. While the temporal validation showed TFT outperforming tree-based approaches, we observe a different story in system-based validation - the tree-based models demonstrate remarkable improvement in their predictive accuracy. At the same time, TFT maintains performance levels similar to its temporal validation results.

Looking at the feature importance distribution in Figure 5.52, we see `future_hour_cos` emerging as the dominant predictor, followed by `static_num_panels` and `past_hour_cos`. While logical, this reliance on temporal patterns might explain why TFT doesn’t achieve the same performance gains as tree-based models when generalizing to new systems. Where tree-based

approaches appear to effectively leverage their simpler architectures to adapt to unseen system characteristics, TFT's sophisticated encoder-decoder structure maintains its temporal-focused strategy.

The static feature hierarchy and the stark dominance of `future_hour_cos` in the decoder features suggest that TFT may be somewhat constrained by its architectural complexity when adapting to new systems. While this architecture proves powerful for temporal prediction, the simpler, more direct approach of tree-based models appears more effective at capturing and generalizing system-specific patterns. This contrast in performance adaptation between model types offers valuable insights for choosing appropriate architectures based on the specific prediction context.

The trend analysis and error patterns provide deeper insights into TFT's system-based prediction behavior. Looking at the trend comparison in Figure 5.53, we observe TFT's tendency to overpredict efficiency values across different generation regimes slightly. While the model successfully captures the overall seasonal and daily patterns, maintaining the natural rhythm of photovoltaic generation, a consistent pattern of prediction values is running slightly above the actual trends. This overprediction behavior becomes particularly noticeable during transition periods between high and low generation states, where the model appears to anticipate changes slightly ahead of their actual occurrence.

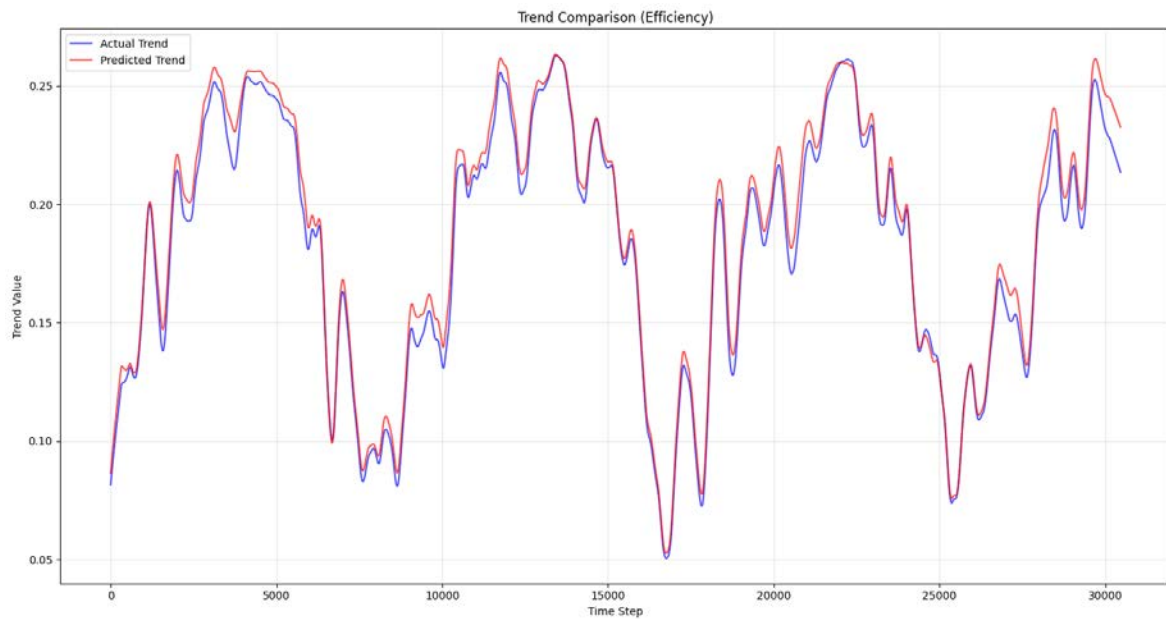


Figure 5.53: Trend comparison showing TFT’s slight overprediction tendency when generalizing to an unseen system. While the model captures major efficiency patterns, the predicted trend (red) consistently runs slightly above the actual trend (blue) during transitional periods.

The temporal error distribution, shown in Figure 5.54, reveals how prediction accuracy varies over time. Most errors remain contained within ± 0.2 efficiency units, though we observe occasional more significant deviations reaching up to ± 0.6 . What’s particularly interesting is the error pattern’s consistency throughout the prediction period. Unlike tree-based models that often show improved accuracy over time when predicting new systems, TFT maintains a relatively stable error distribution, suggesting its prediction strategy remains constant rather than adapting to system-specific patterns.

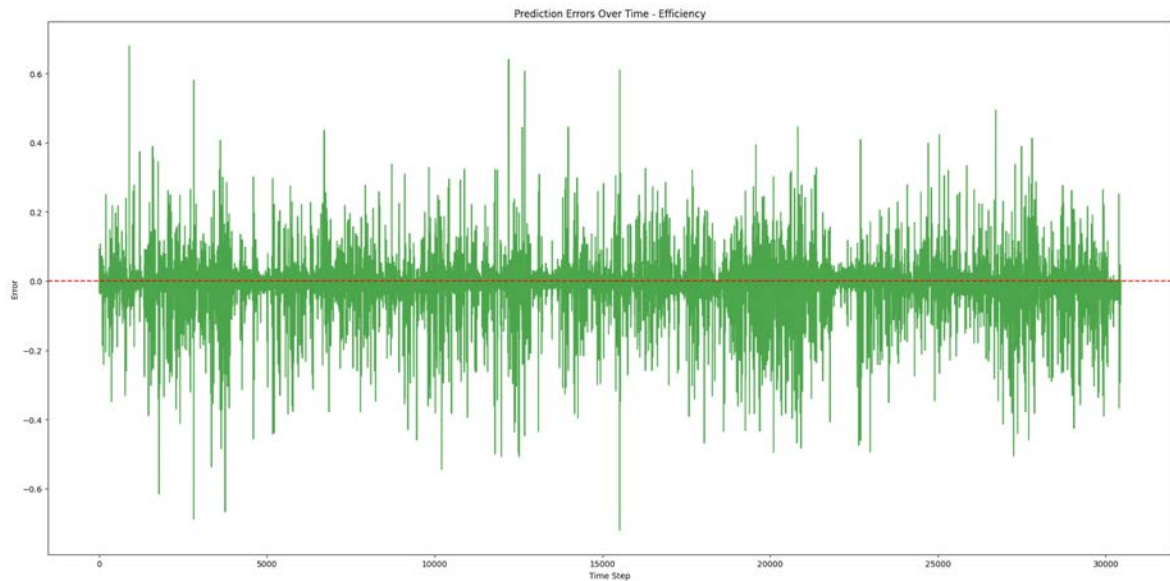


Figure 5.54: Temporal distribution of prediction errors showing consistent bounds throughout the prediction period, though with notable spikes during challenging conditions.

The residual analysis (Figure 5.55) provides perhaps the most telling insight into TFT's system-based prediction characteristics. The scatter plot reveals a distinctive pattern where error variance increases with predicted efficiency values, forming a characteristic fan shape. This heteroscedastic behavior suggests that while TFT maintains reasonable accuracy during low and moderate generation periods, its predictions become less confident during high-efficiency conditions - precisely when accurate predictions are most valuable for system operators.

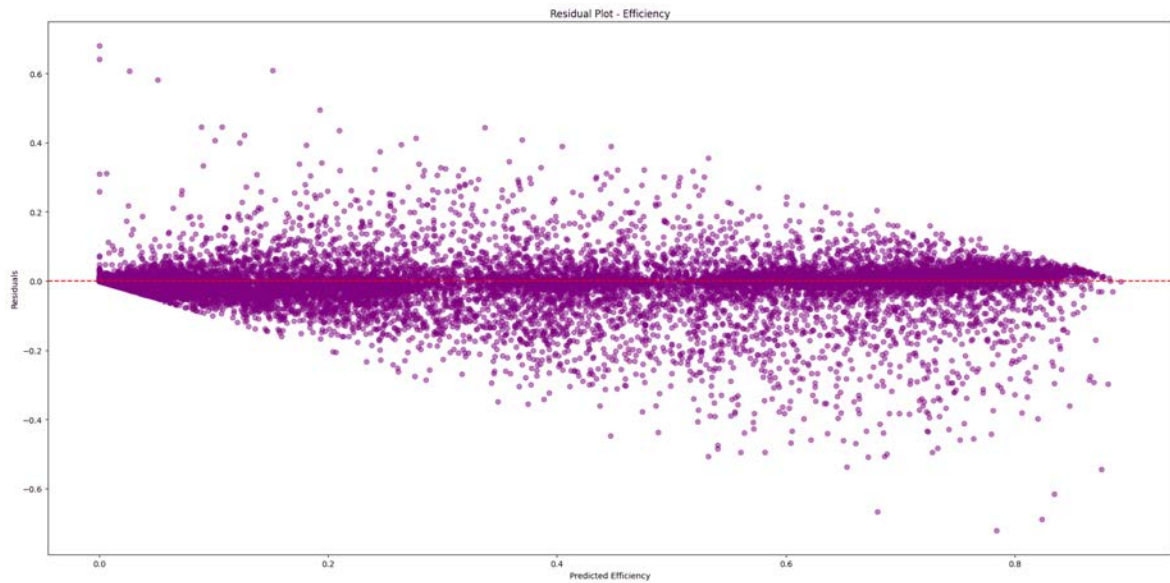


Figure 5.55: Residual analysis reveals increased prediction uncertainty during high-efficiency periods, demonstrated by the widening spread of residuals as predicted efficiency increases.

These analyses—from trend patterns to error distributions—reveal a key characteristic of TFT’s behavior: its sophisticated architecture maintains consistent performance levels across different validation scenarios. While tree-based models show marked improvement in system-based validation, TFT’s stable performance across both temporal and system-based tests suggests that its complex encoder-decoder structure learns fundamental patterns rather than adapting to specific conditions. This consistency between prediction contexts provides valuable insights for model selection based on application requirements.

5.4 Comparative Analysis

5.4.1 Tree-based Models Comparison

The four tree-based methods—Random Forest, XGBoost, LightGBM, and CatBoost—show very similar performance trends in temporal validation, with improvements over baseline between 45.50% and 47.24%. This narrow performance range indicates that, despite their different architectures, all four models adopt comparable strategies for temporal prediction. However, interesting distinctions emerge in their system-based validation results.

Random Forest and CatBoost show exceptional improvement when predicting an unseen system, achieving around 80% improvement over the baseline (80.02% and 80.32% , re-

spectively). At the same time, XGBoost and LightGBM show more modest improvements of about 68% (68.57% and 68.24%). This difference in performance is especially interesting considering their comparable results in temporal validation.

The feature importance analysis uncovers unique patterns in prediction approaches among each model. During temporal validation, all models significantly depend on temporal features, especially `hour_cos`, yet there are notable variances in their selection of secondary features. Random Forest exhibits a balanced distribution among radiation parameters (`fdir`) and additional temporal features (`hour`, `hour_sin`), indicating a solid strategy for recognizing temporal patterns. Conversely, XGBoost and LightGBM prioritize direct solar radiation measurements (`ssrd`, `ssr`), revealing a strategy that focuses more on current environmental conditions.

System-based validation uncovers even more intriguing differences. CatBoost distinguishes itself by emphasizing geographical features, such as longitude, in tandem with temporal attributes, indicating that its outstanding system-based performance may arise from a superior use of spatial data. Random Forest maintains its balanced feature hierarchy from temporal validation, while XGBoost and LightGBM show more pronounced shifts toward system-specific characteristics. This adaptation in feature utilization might explain why Random Forest and CatBoost achieve superior performance in system-based validation—their feature strategies appear better suited to generalizing across different system configurations.

The computational demands of these models show notable differences. LightGBM is highly efficient, finishing training in merely 3.23 seconds for temporal validation and 3.77 seconds for system-based validation. XGBoost shows strong performance, requiring 9.60 seconds for temporal validation and 15.81 seconds for system-based validation. In comparison, Random Forest takes much longer, with an average training time of about 200 seconds for both approaches. CatBoost demonstrates intriguing behavior; while it needs 76.59 seconds for temporal validation, its training time skyrockets to 314.48 seconds for system-based validation, indicating a more elaborate process of adaptation when learning to predict from various systems.

5.4.2 Comparing Tree-based and Deep Learning Approaches

The differences between tree-based models and TFT provide intriguing insights into forecasting photovoltaic power generation. In our main prediction task, temporal validation,

TFT's advanced architecture shows a notable performance boost of 61.07% over baseline. This greatly surpasses the performance of tree-based methods, which typically range from 46-47%. This advantage in temporal prediction demonstrates that TFT's encoder-decoder architecture and attention layers are designed to effectively capture complex, time-dependent patterns.

The patterns of feature utilization reveal these architectural distinctions. Tree-based models, which learn via recursive binary splitting, treat features as independent variables at every node, leading to distinct hierarchical importance rankings. In contrast, TFT's neural architecture facilitates intricate feature interactions through its multi-headed attention mechanism and dedicated processing streams. While tree models evaluate each feature independently for split decisions, TFT processes information through three distinct but interconnected pathways: historical patterns through its encoder network (`past_efficiency`, `past_ssr`), future temporal information through its decoder network (`future_hour_cos`), and static system characteristics through its context network. This architectural design enables TFT to understand non-linear feature interactions and capture nuanced dependencies among meteorological conditions, temporal patterns, and system characteristics that tree-based models, with their rigid binary splitting, might overlook. The attention mechanism enhances this by dynamically adjusting feature importance based on context instead of adhering to fixed importance hierarchies.

System-based validation provides additional context about each approach's characteristics. Tree-based models show remarkable improvement (reaching up to 80% over baseline), while TFT maintains its temporal validation performance level. However, for our time series prediction task, TFT's consistent performance suggests reliable temporal pattern recognition regardless of the prediction context.

The computational requirements reflect these architectural differences, with tree-based models training in seconds to minutes compared to TFT's 30+ minutes. While this increased computational cost is significant, it represents the investment required for superior temporal pattern recognition, our primary objective in photovoltaic power prediction.

5.4.3 Comparative Analysis Conclusions

Our thorough analysis of photovoltaic power prediction highlights distinctive features and strengths across various modeling approaches:

1. **TFT** shows exceptional temporal prediction abilities, improving by 61.07% over the

baseline thanks to its advanced encoder-decoder architecture. The model effectively captures intricate temporal dependencies through attention mechanisms, resulting in improved forecasting accuracy.

2. **XGBoost** provides robust temporal predictions with a 47.24% improvement while maintaining excellent computational efficiency, with a training time of 9.6 seconds. This ideal combination of performance and efficiency makes it especially suitable for real-world applications.
3. **LightGBM** also offers comparable temporal results (46.97% enhancement) and remarkable computational efficiency (3.2 seconds of training time), demonstrating how its leaf-wise growth technique effectively captures temporal patterns while minimizing computational expenses.
4. **Random Forest** and **CatBoost** demonstrate solid temporal prediction abilities (45.50% and 46.66% improvement, respectively) and show exceptional generalization to new systems in our additional validation tests.

These findings highlight an important relationship between architectural sophistication and prediction capabilities. The tree-based models demonstrate efficient, robust temporal predictions through recursive partitioning approaches. At the same time, TFT's intricate architectural design facilitates advanced temporal pattern recognition via complex feature interactions, albeit at an increased computational expense. This balance between architectural intricacy and computational efficiency offers essential insights for choosing models in photovoltaic power prediction applications.

Chapter 6

Conclusion

In this thesis, we investigated different machine learning approaches for photovoltaic power prediction. We developed and analyzed a comprehensive framework that combines operational data from Greek photovoltaic installations with ERA5 meteorological parameters. By rigorously comparing tree-based and deep learning models, we gained valuable insights into their prediction capabilities and characteristics.

6.1 Summary and Conclusions

Our analysis revealed several key findings:

1. Substantial improvements over baseline performance across all models demonstrate the effectiveness of combining high-frequency operational data with meteorological parameters for prediction.
2. The superior temporal prediction capabilities of the Temporal Fusion Transformer, achieving the most significant improvement over baseline through its sophisticated handling of temporal dependencies.
3. The impressive efficiency of tree-based models, particularly LightGBM and XGBoost, delivers robust predictions with minimal computational overhead.
4. Interesting patterns in system-based generalization, suggesting different strengths between model architectures when adapting to unseen systems.

These results validate our approach to photovoltaic power prediction while highlighting important trade-offs between architectural sophistication and computational efficiency.

6.2 Future Work

Our analysis reveals a number of promising research avenues. One important area to explore is assessing model generalization across very different system configurations, especially looking into TFT's possible benefits in these scenarios. Broadening the geographical focus beyond northern Greece would help confirm our findings in different climate zones and installation types, thus offering insights into the models' ability to adapt to a variety of environmental conditions.

From an architectural standpoint, examining hybrid methods that merge the computational efficiency of tree-based models with the exceptional temporal pattern recognition of TFT may foster innovative solutions in predicting photovoltaic power. Additionally, incorporating extra data sources, especially weather forecast information from meteorological services, emerges as a promising avenue. This integration could utilize TFT's encoder-decoder architecture to analyze both historical and projected weather patterns. Moreover, high-resolution satellite imagery and comprehensive panel degradation data could boost prediction accuracy by supplying more detailed environmental and system-specific insights.

Lastly, extending the prediction horizon would facilitate a broader range of operational planning scenarios, although this would necessitate careful consideration of how various architectures manage increasing temporal dependencies.

Bibliography

- [1] Bashar Mahmood Ali. Solar energy forecasting techniques based on machine learning: Survey. *2023 6th International Conference on Engineering Technology and its Applications (IICETA)*, pages 814–819, 2023.
- [2] Ewa Chodakowska, Joanicjusz Nazarko, Łukasz Nazarko, and Hesham S. Rabayah. Solar radiation forecasting: A systematic meta-review of current methods and emerging trends. *Energies*, 2024.
- [3] T. Rajasundrapandiyanleebanon, Kavitha Kumaresan, Sakthivel Murugan, M. S. P. Subathra, and Mahima Sivakumar. Solar energy forecasting using machine learning and deep learning techniques. *Archives of Computational Methods in Engineering*, 30:3059 – 3079, 2023.
- [4] Murugesan S, M. Mahasree, F. Kavin, and N. Bharathiraja. Solar energy forecasting with performance optimization using machine learning techniques. *Electric Power Components and Systems*, 2024.
- [5] Nataliya Shakhovska, Mykola Medykovskyi, Oleksandr Gurbych, Mykhailo Mamchur, and Mykhailo Melnyk. Enhancing solar energy production forecasting using advanced machine learning and deep learning techniques: A comprehensive study on the impact of meteorological data. *Computers, Materials & Continua*, 2024.
- [6] Rekha Kumari, Rajesh Kumhar, and Rajesh Narayan Deo. Data driven approaches to improve solar energy forecasting using machine learning algorithm. *2024 Second International Conference on Networks, Multimedia and Information Technology (NMIT-CON)*, pages 1–7, 2024.

- [7] Francesco Nicoletti and Piero Bevilacqua. Hourly photovoltaic production prediction using numerical weather data and neural networks for solar energy decision support. *Energies*, 2024.
- [8] Thomas Shering, Eduardo Alonso, and Dimitra Apostolopoulou. Investigation of load, solar and wind generation as target variables in lstm time series forecasting, using exogenous weather variables. *Energies*, 2024.
- [9] C. Vennila, Anita Titus, T. Sri Sudha, U. Sreenivasulu, Noti Pandu, Ranga Reddy, K. Jamal, Dayadi Lakshmaiah, Pasupuleti Jagadeesh, and Assefa Belay. Forecasting solar energy production using machine learning. *International Journal of Photoenergy*, 2022.
- [10] Sandeep Sunori, Shilpa Jain, Pradeep Juneja, and Amit Mittal. Ann and random forest based prediction of power output of solar panel. *2024 5th International Conference on Smart Electronics and Communication (ICOSEC)*, pages 882–887, 2024.
- [11] Venkata Lakshmi Yerrabolu, Idamakanti Kasireddy, K M Jasmine, T B Murali Krishna Vamsi, N Joshua, V Shyam Kumar, and Dsnm Rao. Performance comparison of random forest regressor and support vector regression for solar energy prediction. *IOP Conference Series: Earth and Environmental Science*, 2024.
- [12] Bryan Lim, Sercan Ö. Arik, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *ArXiv*, abs/1912.09363, 2019.
- [13] Abderraouf Bouakkaz, Adel Lahsasna, Antonio José Gil Mena, Salim Haddad, Mario Luigi Ferrari, and Rafael Jiménez Castañeda. Evaluating and analyzing the performance of pv power output forecasting using different models of machine-learning techniques considering prediction accuracy. *International Journal of Renewable Energy Development*, 2024.
- [14] Aleksandr Gevorgian, Giovanni Pernigotto, and Andrea Gasparella. Addressing data scarcity in solar energy prediction with machine learning and augmentation techniques. *Energies*, 17(14), 2024.
- [15] PVOutput. Pvoutput.org api, 2024. Data retrieved via API access, July 2024 - December 2024.

- [16] H. Hersbach, B. Bell, P. Berrisford, G. Biavati, A. Horányi, J. Muñoz Sabater, J. Nicolas, C. Peubey, R. Radu, I. Rozum, D. Schepers, A. Simmons, C. Soci, D. Dee, and J.-N. Thépaut. Era5 hourly data on single levels from 1940 to present, 2023. Accessed on DD-MMM-YYYY.
- [17] L Breiman. Random forests. *Machine Learning*, 45:5–32, 10 2001.
- [18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [19] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM.
- [20] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *Neural Information Processing Systems*, 2017.
- [21] Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. Catboost: gradient boosting with categorical features support. *ArXiv*, abs/1810.11363, 2018.
- [22] Bryan Lim, Sercan Ö. Arik, Nicolas Loeff, and Tomas Pfister. Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4):1748–1764, 2021.
- [23] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.
- [24] Janek Thomas. Pytorch forecasting: Time series forecasting with pytorch, 2020.