



KERJA PRAKTIK - IF184801

Fine-Tuning Solar Power Predictions: An In-Depth Analysis of Decomposition Techniques

CITI - Center for IOT Innovation

MA012 (NTUST Management Building), No. 43 號, Section 4,
Keelung Rd, Da'an District, Taipei City, Taiwan 106

Periode: September 12th 2023 - December 12th 2023

By:

Angela Oryza Prabowo

5025201022

Department Supervisor

Baskoro Adi Pratomo S.Kom., M.Kom.

Field Supervisor

Professor Shuo-Yan Chou

DEPARTEMEN TEKNIK INFORMATIKA

Fakultas Teknologi Elektro dan Informatika Cerdas

Institut Teknologi Sepuluh Nopember

Surabaya 2023



KERJA PRAKTIK - IF184801

Fine-Tuning Solar Power Predictions: An In-Depth Analysis of Decomposition Techniques

CITI - Center for IOT Innovation

MA012 (NTUST Management Building), No. 43 號, Section 4,

Keelung Rd, Da'an District, Taipei City, Taiwan 106

Periode: September 12th 2023 - December 12th 2023

By:

Angela Oryza Prabowo

5025201022

Department Supervisor

Baskoro Adi Pratomo S.Kom., M.Kom.

Field Supervisor

Professor Shuo-Yan Chou

DEPARTEMEN TEKNIK INFORMATIKA

Fakultas Teknologi Elektro dan Informatika Cerdas

Institut Teknologi Sepuluh Nopember

Surabaya 2023

[This page intentionally left blank]

TABLE OF CONTENTS

TABLE OF CONTENTS.....	IV
TABLE OF FIGURES.....	VIII
LIST OF TABLES.....	X
APPROVAL SHEET	XII
ACKNOWLEDGEMENTS.....	XVII
CHAPTER I INTRODUCTION.....	1
1.1. BACKGROUND AND MOTIVATION	1
1.2. OBJECTIVES	1
1.3. CONTRIBUTIONS	2
1.4. PROBLEM DEFINITION	2
1.5. LOCATION AND DURATION.....	2
1.6. METHODOLOGY	3
1.6.1. <i>Problem Formulation</i>	3
1.6.2. <i>Literature Review</i>	3
1.6.3. <i>System Analysis and Design</i>	3
1.6.4. <i>System Implementation</i>	4
1.6.5. <i>Result and Discussion</i>	4
1.6.6. <i>Conclusion</i>	4
1.7. REPORT STRUCTURE.....	4
1.7.1. <i>Chapter I Introduction</i>	4
1.7.2. <i>Chapter II Company Profile</i>	4
1.7.3. <i>Chapter III Literature Review</i>	4
1.7.4. <i>Chapter IV System Analysis and Design</i>	4
1.7.5. <i>Chapter V System Implementation</i>	5
1.7.6. <i>Chapter VI Result and Discussion</i>	5
1.7.7. <i>Chapter VII Conclusion</i>	5
CHAPTER II COMPANY PROFILE.....	7

2.1.	CENTER FOR IOT INNOVATION (CITI) PROFILE.....	7
2.2.	LOCATION.....	7
CHAPTER III LITERATURE REVIEW		9
3.1.	RECURSIVE FEATURE ELIMINATION (RFE).....	9
3.2.	DECISION TREE	10
3.3.	RANDOM FOREST	10
3.4.	EXTREME GRADIENT BOOSTING (XGBOOST).....	11
3.5.	DECOMPOSITION.....	11
3.6.	LONG-SHORT TERM MEMORY (LSTM)	12
CHAPTER IV SYSTEM ANALYSIS AND DESIGN		16
4.1.	SYSTEM ANALYSIS	16
4.1.1.	<i>General Application Definition</i>	16
4.2.	SYSTEM INFRASTRUCTURE DESIGN	16
4.2.1.	<i>System Design</i>	16
CHAPTER V SYSTEM IMPLEMENTATION		22
5.1.	EXPLORATORY DATA ANALYSIS	22
5.1.1.	<i>Similarity of Datasets</i>	22
5.1.2.	<i>Concatenating Datasets</i>	23
5.1.3.	<i>Missing Value Imputation</i>	23
5.2.	FEATURE ENGINEERING	25
5.2.1.	<i>Transformation of Wind Direction</i>	25
5.2.2.	<i>Cyclical Features of Time</i>	25
5.2.3.	<i>Features Selection</i>	27
5.3.	DECOMPOSITION.....	31
5.3.1.	<i>Split into Train, Val, and Test</i>	34
5.3.2.	<i>Scaling the Data</i>	34
5.3.3.	<i>Model Building</i>	35
5.3.4.	<i>Reconstruction of Forecasted Results</i>	36
CHAPTER VI RESULTS AND DISCUSSION.....		38
6.1.	TESTING OBJECTIVES.....	38

6.2.	RESULT OF DIFFERENT DECOMPOSITION SCENARIOS.....	38
6.2.1.	CEEMDAN DECOMPOSITION	38
6.2.2.	EWT DECOMPOSITION	40
6.2.3.	COMBINATION OF EWT AND CEEMDAN.....	42
6.2.4.	COMPARISON BETWEEN ALL DECOMPOSITION SCENARIOS.....	44
6.3.	EVALUATION	46
	CHAPTER VII CONCLUSION	51
7.1.	CONCLUSION	51
7.2.	SUGGESTION	51
	REFERENCES	53
	AUTHOR’S BIODATA.....	55

[This page intentionally left blank]

TABLE OF FIGURES

Figure 2.1 SimpleRNN Diagram.....	13
Figure 2.2 SimpleRNN with cell state.....	13
Figure 2.3 LSTM Diagram.....	14
Figure 3.1 System Design.....	17
Figure 3.2 FFT Plot for Power and Irradiance.....	26
Figure 3.3 Day Sin and Hour Plot.....	27
Figure 3.4 Decomposition Scenario 1.....	32
Figure 3.5 Decomposition Scenario 2.....	33
Figure 3.6 Decomposition Scenario 3.....	34
Figure 4.1 Loss Plot for Scenario 1.....	38
Figure 4.2 Predicted Plot for Scenario 1.....	40
Figure 4.3 Loss Plot for Scenario 2.....	40
Figure 4.4 Predicted Plot for Scenario 2.....	42
Figure 4.5 Loss Plot for Scenario 3.....	42
Figure 4.6 Predicted Plot for Scenario 3.....	44
Figure 4.7 Multi-Step Prediction Plot.....	47

[This page intentionally left blank]

LIST OF TABLES

Table 4.1. Attributes and Data Types.....	18
Table 5.1 Missing Values.....	23
Table 5.2 Imputed Values for Power.....	24
Table 5.3 Imputed Values for Irradiance.....	24
Table 5.4 Features Selection.....	29
Table 6.1 MAE Scores for Scenario 1.....	39
Table 6.2 MAE Scores for Scenario 2.....	41
Table 6.3 MAE Scores for Scenario 3.....	43
Table 6.4 Comparison of MAE Scores Between All Scenarios...	45
Table 6.5 MAE Scores for Multi-Step Prediction.....	49

[This page intentionally left blank]

[This page intentionally left blank]

**APPROVAL SHEET
INTERNSHIP REPORT**

**Fine-Tuning Solar Power Predictions: An In-Depth
Analysis of Decomposition Techniques**

By:

Angela Oryza Prabowo


5025201022

Approved by Internship Supervisor:

1. Baskoro Adi Pratomo
S.Kom., M.Kom.
NIP. 198702182014041001

(Department Supervisor)

2. Professor Shuo-Yan Chou



(Field Supervisor)

Fine-Tuning Solar Power Predictions: An In-Depth Analysis of Decomposition Techniques

Student Name : Angela Oryza Prabowo
Student Number : 5025201022
Department : Teknik Informatika FTEIC-ITS
Department Supervisor : Baskoro Adi Pratomo, S.Kom,
M.Kom.
Field Supervisor : Professor Shuo-Yan Chou

ABSTRACT

Solar power generation plays a pivotal role in Taiwan's pursuit of renewable energy, aligning with its ambitious target of generating over 27 GW by 2050 and achieving net-zero emissions by the same year. Notably, Taiwan stands as the world's second-largest producer of solar photovoltaic (PV) energy, driven by the abundance of solar radiation, particularly in southern regions where it exceeds 145 watts per square meter.

Building upon previous research that attained a remarkable low Mean Absolute Error (MAE) of 0.0223 through multivariate analysis with decomposition techniques, this study aims to further refine the forecasting models by focusing on the univariate decomposition. The hypothesis posits that this approach will lead to an even lower MAE, contributing to more accurate predictions.

The research methodology involves extensive data preprocessing, which includes comparing and merging external datasets with information from the Taiwan Central Weather Bureau and Open Meteo. Feature engineering techniques are employed, incorporating transformations for time, wind direction, and wind speed. Recursive feature elimination is utilized for effective feature selection, enhancing the quality of input variables.

In conclusion, this project centers on the refinement of univariate decomposition techniques to optimize solar power generation forecasts in Taiwan. By improving the accuracy of the decomposition model, the anticipated outcome is a more precise overall prediction when the refined results are integrated into subsequent forecasting steps. This research contributes to the ongoing efforts to harness solar energy efficiently and aligns with Taiwan's commitment to a sustainable and renewable energy future.

Keywords : Decomposition, Solar Power Prediction, Time Series, Univariate

[This page intentionally left blank]

ACKNOWLEDGEMENTS

Praise and gratitude are offered to Allah SWT for His guidance and blessings, allowing the author to complete one of the obligations as a student of the Department of Informatics, ITS, namely the Internship titled: Fine-Tuning Solar Power Predictions: An In-Depth Analysis of Decomposition Techniques.

The author is aware that there are still many shortcomings in carrying out the internship and compiling this internship report. However, the author hopes that this report can enhance the readers' insights and serve as a reference source.

Through this report, the author would also like to express gratitude to those who have assisted in compiling the internship report, both directly and indirectly, including:

1. Both of the author's parents.
2. Mr. Baskoro Adi Pratomo, S.Kom, M.Kom., as the internship supervisor.
3. Professor Shuo-Yan Chou, as the field supervisor during the internship.
4. The author's friends who have always provided encouragement during the internship.

Surabaya, January 5th 2024

Angela Oryza Prabowo

[This page intentionally left blank]

CHAPTER I

INTRODUCTION

1.1. Background and Motivation

Taiwan's pursuit of solar power generation is pivotal, given its ambitious renewable energy target of over 27 GW by 2050 and the commitment to net-zero emissions. Solar energy is chosen for its abundant potential, with more than two-thirds of southern Taiwan receiving solar radiation exceeding 145 watts per square meter, establishing Taiwan as the world's second-largest solar PV producer.

Previous research achieved a remarkable Mean Absolute Error (MAE) of 0.0223 through multivariate analysis. This study aims to refine forecasting by transforming multiple sequence decomposition techniques into a univariate model, seeking an even lower MAE for enhanced precision in solar power predictions.

Reducing prediction errors is crucial for maintaining the delicate balance between energy supply and demand. As solar energy integrates into the power supply system, accurate forecasts become essential to manage fluctuations, avoiding excessive adjustments to traditional power generation units and preserving their lifespan.

This research contributes to the effective integration of solar energy, aligning with Taiwan's sustainability goals and ensuring a resilient and balanced energy future.

1.2. Objectives

The primary objective of this project is to fulfill the requirements for completing 3 SKS (academic credits)

in the internship program. The focus is on minimizing errors in decomposition results to improve the overall accuracy of solar power predictions.

1.3. Contributions

The contribution achieved through an in-depth exploration of decomposition techniques is the enhancement of the solar power prediction model's accuracy. This improvement plays a crucial role in minimizing the need for frequent adjustments to the traditional power generation system.

1.4. Problem Definition

The problems addressed in this project are outlined below:

1. How can better decomposition techniques be implemented for more accurate solar power prediction?
2. How to effectively implement the optimal architecture for a time-series model to predict univariate decomposed data?

1.5. Location and Duration

The project spanned from September 12, 2023, to December 12, 2023. The initial month was conducted online, within a Work From Home environment, while the final two months took place offline at CITI - Center for IoT Innovation, NTUST, Taiwan.

1.6. Methodology

The methodology in creating the internship report includes:

1.6.1. Problem Formulation

To understand the model requirements, an initial meeting was conducted with the supervisor, Mr. Indie. During this meeting, he emphasized the significance of accurate solar power prediction in Taiwan, a country heavily reliant on solar energy. Insights into the performance of previous models were shared, along with expectations for improvement in this project. With a clear understanding of the task at hand, the team commenced exploration, focusing specifically on the decomposition technique as a potential avenue to enhance the overall accuracy of solar power predictions.

1.6.2. Literature Review

After gaining a comprehensive understanding of the model's construction, we were instructed to review pertinent literature before commencing the project. The literature encompassed topics such as LSTM (Long Short-Term Memory), Decomposition Techniques, various EDA (Exploratory Data Analysis) techniques, including feature engineering and feature selection, among others. Additionally, guidance was provided on the specific considerations for handling time-series data.

1.6.3. System Analysis and Design

After conducting several literature reviews, it became evident that constructing a robust system would benefit from a comprehensive system architecture design. For this model, the decision was made to maintain the overall architecture from previous research, with planned modifications to the decomposition architecture for improvement.

1.6.4. System Implementation

Implementation is the tangible realization of the design phase. In this stage, we commence the process of translating the system we previously designed into an actual, functional implementation.

1.6.5. Result and Discussion

Following the implementation of various decomposition techniques and architectures, an evaluation phase is essential to determine which model yields the most accurate predictions. The assessment will be based on the Mean Absolute Error (MAE) predictions, aiding in the selection of the most accurate model.

1.6.6. Conclusion

The conducted testing has met the desired criteria and proceeded well and smoothly.

1.7. Report Structure

1.7.1. Chapter I Introduction

This chapter encompasses the background, objectives, benefits, problem formulation, location and duration of the internship, methodology, and the structure of the report.

1.7.2. Chapter II Company Profile

This chapter provides a general overview of CITI, NTUST, including its profile and location.

1.7.3. Chapter III Literature Review

This chapter contains the theoretical foundations of the technologies used in completing the internship project.

1.7.4. Chapter IV System Analysis and Design

This chapter covers the system analysis stage of the application in completing the internship project.

1.7.5. Chapter V System Implementation

This chapter describes the stages undertaken during the application implementation process.

1.7.6. Chapter VI Result and Discussion

This chapter presents the results of testing and evaluation of the developed application during the internship period.

1.7.7. Chapter VII Conclusion

This chapter contains conclusions and suggestions derived from the internship implementation process.

[This page intentionally left blank]

CHAPTER II COMPANY PROFILE

2.1. Center for IoT Innovation (CITI) Profile

The Center for IoT Innovation (CITI) at the National Taiwan University of Science and Technology (NTUST) is a leading research center specializing in the fields of IoT and AI. With a focus on technology-enabled services, CITI explores innovative solutions to address contemporary global challenges. The center is dedicated to engaging with industry and applying cutting-edge knowledge and insights to real-world issues, aiming to enhance businesses, policies, practices, and overall outcomes.

2.2. Location

MA012 (NTUST Management Building), No. 43 號, Section 4, Keelung Rd, Da'an District, Taipei City, Taiwan 106

[This page intentionally left blank]

CHAPTER III LITERATURE REVIEW

3.1. Recursive Feature Elimination (RFE)

The The effectiveness of a feature ranking criterion doesn't necessarily translate into an effective feature subset ranking criterion. Criteria such as $D_j(i)$ or $(w_i)^2$ are designed to estimate the impact of removing one feature at a time on the objective function. However, they prove suboptimal when removing multiple features simultaneously, crucial for obtaining a concise feature subset. To address this, the Recursive Feature Elimination (RFE) method is introduced:

1. Train the classifier by optimizing the weights w_i with respect to the objective function J .
2. Calculate the ranking criterion (e.g., $D_j(i)$ or $(w_i)^2$) for all features.
3. Iteratively remove the feature with the smallest ranking criterion.

RFE is a form of backward feature elimination, where, for computational efficiency, it may be practical to eliminate multiple features at once, even though this could lead to a potential degradation in classification performance. In such instances, the method provides a feature subset ranking rather than a feature ranking, with nested subsets $F_1 \subset F_2 \subset \dots \subset F$. (GUYON et al., 2002). While a feature ranking is still applicable when removing features one at a time, the top-ranked features (eliminated last) may not necessarily be individually the most relevant. It is crucial to recognize that RFE doesn't impact correlation methods, as the ranking criterion is computed using information about a single feature.

3.2. Decision Tree

A Decision Tree is a versatile machine learning algorithm used for classification and regression tasks. Operating by recursively partitioning data based on the most significant attributes, it creates a tree-like structure where leaves represent predicted outcomes. Known for its transparency, Decision Trees are valuable for interpreting the decision-making process. While susceptible to overfitting, techniques like pruning can mitigate this issue. Often employed in ensemble methods like Random Forests and Gradient Boosting, Decision Trees serve as foundational concepts in machine learning. Their simplicity and interpretability make them useful for introductory education and as building blocks for more advanced models. (Friedman, 2009)

3.3. Random Forest

Random Forest is an ensemble learning algorithm that enhances predictive accuracy by constructing multiple decision trees during training. Each tree is trained on a bootstrap sample of the data, and at each node, a random subset of features is considered for splitting, mitigating overfitting and improving generalization. The algorithm outputs the mode of predictions for classification or the average prediction for regression. Known for versatility, Random Forest handles both classification and regression tasks, provides feature importance scores, and is robust against noisy data. Its ensemble nature combines the interpretability of individual decision trees with superior predictive performance, making it widely adopted across diverse domains. (BREIMAN, 2001)

3.4. eXtreme Gradient Boosting (XGBoost)

XGBoost, or Extreme Gradient Boosting, is a highly effective machine learning algorithm widely used for structured data challenges. Developed by Tianqi Chen, it operates within an ensemble learning framework, sequentially constructing decision trees to iteratively correct errors. A distinctive feature of XGBoost is its optimization of an objective function that combines a loss function measuring label prediction accuracy with regularization terms controlling model complexity. The final prediction results from a weighted sum of predictions from all trees in the ensemble. XGBoost stands out for its ability to handle missing data, incorporation of regularization techniques to prevent overfitting, and support for parallel and distributed computing. (Chen & Guestrin, 2016)

3.5. Decomposition

Signal decomposition techniques aim to break down complex signals into simpler components, providing insights into the underlying structures and patterns within the data. There are two decomposition techniques that are used in this project, Complex Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) and Empirical Wavelet Transform (EWT).

3.5.1. Complex Empirical Mode Decomposition with Adaptive Noise (CEEMDAN)

Complex Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) is a signal processing technique used for decomposing non-linear and non-stationary signals into a set of intrinsic mode functions (IMFs). Unlike traditional Empirical Mode Decomposition (EMD), CEEMDAN introduces an adaptive noise

term to enhance decomposition performance, particularly in the presence of noise. (Torres et al., 2011) The adaptive noise helps in addressing mode mixing issues, providing a more accurate representation of the signal's intrinsic components.

3.5.2. Empirical Wavelet Transform (EWT)

Empirical Wavelet Transform (EWT) is a signal processing technique that decomposes a signal into components with different frequency bands and time localization. EWT combines the concept of wavelet transforms with empirical mode decomposition, providing a flexible and adaptive approach to analyze non-stationary and non-linear signals. The method aims to capture both frequency and time information effectively, making it suitable for a wide range of applications, including signal denoising and feature extraction. (Gilles, 2013)

3.6. Long-Short Term Memory (LSTM)

In the realm of recurrent neural networks (RNNs), particularly Long Short-Term Memory (LSTM), the preservation of information over time stands out as a key mechanism, preventing the gradual fading of older signals during processing—reminiscent of the principles behind residual connections. To initiate a comprehensive exploration, let's commence with the SimpleRNN cell. Within this intricate structure, various weight matrices are denoted with the letter o such as W_o and U_o , specifically assigned for output-related computations.

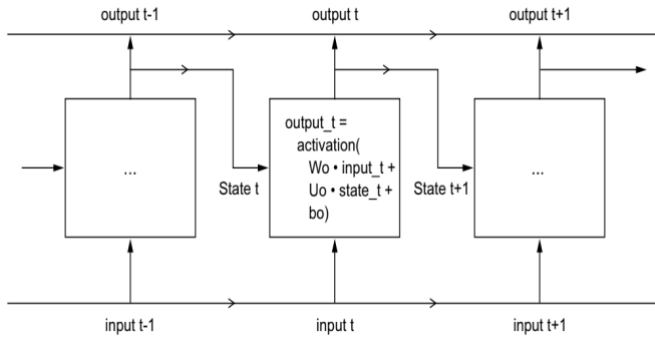


Fig 2.1 SimpleRNN Diagram

A pivotal augmentation to this framework involves introducing an additional data flow denoted as c_t , where the "C" signifies carry. This information intricately intertwines with input and recurrent connections, undergoing a dense transformation. (Chollet, 2021) This transformation is characterized by a dot product with a weight matrix, a subsequent bias addition, and the application of an activation function. Furthermore, the information from c_t plays a significant role in shaping the state transmitted to the subsequent timestep, undergoing an activation function and a multiplication operation.

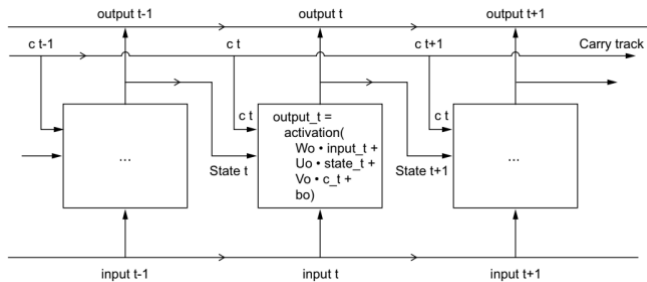


Fig 2.2 SimpleRNN with cell state

In a conceptual sense, this auxiliary carry dataflow acts as a modulating force, influencing both the subsequent output and state. Philosophically, the multiplication of c_t and f_t can be viewed as a purposeful mechanism for discarding irrelevant information within the carry dataflow, while i_t and k_t contribute insights into the present, infusing the carry track with updated information.

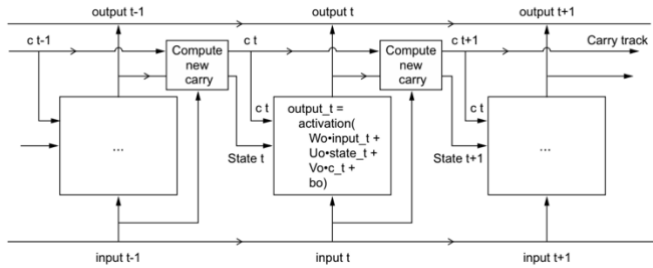


Fig 2.3 LSTM Diagram

[This page intentionally left blank]

CHAPTER IV

SYSTEM ANALYSIS AND DESIGN

4.1. System Analysis

This chapter will delineate the essential phases required for constructing a model to predict solar power generation, with a particular emphasis on the decomposition technique step.

4.1.1. General Application Definition

In general, the model for predicting solar power is built on LSTM. This model will utilize various decomposition techniques for univariate prediction, the results of which will be combined with other weather attributes to generate the final prediction of solar power generation.

4.2. System Infrastructure Design

4.2.1. System Design

We commence the system design process with an initial phase dedicated to data exploration, termed Exploratory Data Analysis (EDA). In this stage, our focus lies in assessing the similarity between datasets and external data sources to augment our understanding. Upon establishing dataset similarities, we consolidate them to form a more comprehensive dataset. Subsequently, we meticulously cleanse the dataset, addressing missing values, detecting duplicates, and employing visualization techniques to enhance data comprehension.

Transitioning from the exploration stage, we undertake data pre-processing to bolster the model's robustness. This involves pivotal steps such as feature selection and feature engineering. In the realm of

feature engineering, we transform the wind attribute and convert time data into a cyclical format. Simultaneously, in feature selection, Recursive Feature Elimination (RFE) aids in systematically narrowing down features until only the essential ones remain.

With confidence in the cleanliness and readiness of the data for training, we progress to the decomposition stage. Various decomposition techniques are applied to attributes, enhancing the model's ability to discern signals effectively. The resultant prediction is then assimilated as a new attribute.

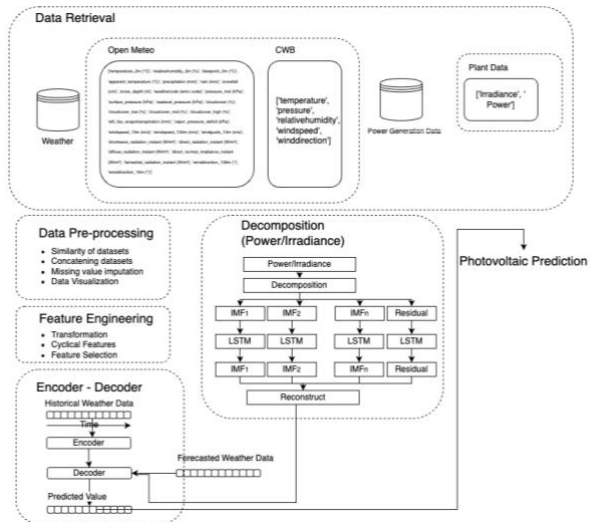


Fig 3.1 System Design

The concluding phase of the system involves constructing the Encoder-Decoder model, opting for the Seq2Seq model. In this architecture, historical weather data is input into the encoder, while the predicted outcome from the decomposition model and forecasted weather data are supplied to the decoder. The amalgamation of these components yields the final

prediction for solar photovoltaic output. Figure 3.1 picturing comprehensive system design encapsulates the entire process.

In this project, our primary emphasis lies on the decomposition technique. Therefore, the final stage, involving the construction of the encoder-decoder model, is not a requisite for our objectives. The project is tailored to encompass Data Retrieval, Data Pre-Processing, Feature Engineering, with a specific focus on Decomposition.

Table 4.1. Attributes and Data Types

No	Variabel	Description	Data Type
1.	timestamp	Date and time of data acquisition	Datetime64
2.	year	Year of data acquisition	int64
3.	month	Month of data acquisition	int64
4.	day	Day of data acquisition	int64
5.	hour	Hour of data acquisition	int64
6.	power	The solar power generated at a specific timestamp.	float64
7.	irradiance	The irradiance level captured from the sun at a specific timestamp.	float64
8.	temperature_2m (°C)	Air temperature at 2 meters above ground	float64
9.	relativehumidity_2m (%)	Relative humidity at 2 meters above ground	float64
10.	dewpoint_2m (°C)	Dew point temperature at 2 meters above ground	float64

11.	apparent_temperature (°C)	Apparent temperature is the perceived feels-like temperature combining wind chill factor, relative humidity and solar radiation	float64
12.	precipitation (mm)	Total precipitation (rain, showers, snow) sum of the preceding hour	float64
13.	rain (mm)	Rain from large scale weather systems of the preceding hour in millimeter	float64
14.	snowfall (cm)	Snowfall amount of the preceding hour in centimeters. For the water equivalent in millimeter, divide by 7. E.g. 7 cm snow = 10 mm precipitation water equivalent	int64
15.	snow_depth (m)	Snow depth on the ground	int64
16.	weathercode (wmo code)	Weather condition as a numeric code. Follow WMO weather interpretation codes	int64
17.	pressure_msl (hPa)	Atmospheric air pressure reduced to mean sea level (msl) or pressure at surface. Typically pressure on mean sea level is used in meteorology.	float64
18.	surface_pressure (hPa)	Surface pressure gets lower with increasing elevation.	float64
19.	cloudcover (%)	Total cloud cover as an area fraction	int64
20.	cloudcover_low (%)	Low level clouds and fog up to 3 km altitude	int64
21.	cloudcover_mid (%)	Mid level clouds from 3 to 8 km altitude	int64
22.	cloudcover_high (%)	High level clouds from 8 km altitude	int64
23.	et0_fao_evapotranspiration (mm)	ET ₀ Reference Evapotranspiration of a well watered grass field.	float64

24	vapor_pressure_deficit (kPa)	Vapour Pressure Deficit (VPD) in kilopascal (kPa).	float64
25	windspeed_10m (m/s)	Wind speed at 10 meters above ground.	float64
26	windspeed_100m (m/s)	Wind speed at 100 meters above ground.	float64
27	winddirection_10m (°)	Wind direction at 10 meters above ground.	int64
28	winddirection_100m (°)	Wind direction at 100 meters above ground.	int64
29	windgusts_10m (m/s)	Gusts at 10 meters above ground as a maximum of the preceding hour	float64
30	is_day ()	1 if the current time step has daylight, 0 at night.	int64
31	shortwave_radiation_instant (W/m ²)	Shortwave solar radiation as average of the preceding hour.	float64
32	direct_radiation_instant (W/m ²)	Direct solar radiation as average of the preceding hour on the horizontal plane and the normal plane (perpendicular to the sun)	float64
33	diffuse_radiation_instant (W/m ²)	Diffuse solar radiation as average of the preceding hour	float64
34	direct_normal_irradiance_instant (W/m ²)	Direct solar radiation as average of the preceding hour on the horizontal plane and the normal plane (perpendicular to the sun)	float64
35	terrestrial_radiation_instant (W/m ²)	Amount of solar radiation received at the Earth's surface at a specific moment in time	float64

[This page intentionally left blank]

CHAPTER V

SYSTEM IMPLEMENTATION

This chapter discusses the implementation of the system we created. The implementation will be divided into several parts, namely Exploratory Data Analysis (EDA), Feature Engineering, and Decomposition.

5.1. Exploratory Data Analysis

5.1.1. Similarity of Datasets

To assess the similarity of datasets, we employ two distinct methods: Kullback-Leibler divergence and Cosine Similarity. Our analysis focuses on comparing the similarity of specific columns present in both datasets, CWB and Open Meteo. These shared columns encompass pressure, temperature, Relative Humidity, Wind Speed, and Wind Direction.

Kullback-Leibler (KL) divergence serves as a measure to quantify the difference between two probability distributions. It serves as an initial step to determine whether the data distributions are congruent or divergent. A KL score close to zero indicates a high degree of similarity between the probability distributions of the two datasets. The average KL divergence across corresponding attributes from the CWB dataset and Open Meteo dataset is calculated to be **0.10325426163906415**. This result suggests a noteworthy similarity between the datasets, given the proximity of the KL score to zero.

Concurrently, Cosine Similarity is employed to gauge the similarity between two vectors, providing a precise calculation of the distance between each pair of corresponding data points. A score of 1 denotes perfect similarity, 0 signifies no similarity, and -1 indicates

perfect dissimilarity. The average cosine similarity score for corresponding attributes in both datasets is determined to be **0.94245929**. This outcome underscores a high degree of similarity between the datasets, as the cosine similarity score approaches 1.

5.1.2. Concatenating Datasets

Given the observed high similarity between both datasets, we can confidently conclude that it is safe to concatenate them. To achieve this, we employ the `pd.concat` function, a functionality offered by the pandas library. This process allows for a seamless integration of the CWB and Open Meteo datasets, leveraging their shared attributes and ensuring a unified dataset for further analysis.

5.1.3. Missing Value Imputation

In the dataset, two attributes, **Power** and **Irradiance**, contain missing values, as outlined in the Table 5.1.

Table 5.1 Missing Values

Attributes	Timestamp			
	2020-01-01 06:00:00	2020-01-01 07:00:00	2021-03-24 16:00:00	2021-09-23 15:00:00
Power	NaN	NaN	0.0	NaN
Irradiance	NaN	NaN	NaN	NaN

Given the limited number of missing values (3 in Power and 4 in Irradiance), employing complex machine learning models for imputation is deemed unnecessary. Instead, we opt for a more straightforward approach using interpolation, a

method that smoothly connects data points within a sequence to estimate missing values.

Three types of interpolation methods are employed for filling in the missing values:

- Linear Interpolation: Estimates missing values by assuming a straight line.
- Polynomial Interpolation: Fits a single polynomial function.
- Spline Interpolation: Divides the data range into smaller segments and fits a separate polynomial function for each segment.

Following the application of these interpolation methods, the detailed values that have been filled in are summarized in the Table 5.2 and Table 5.3

Table 5.2 Imputed Values for Power

Attributes (Power)	Timestamp		
	2020-01-01 06:00:00	2020-01-01 07:00:00	2021-09-23 15:00:00
Linear	13.733333	27.466666	5.088
Polynomial	4.247753	16.763111	6.260022
Spline	4.834627	17.133257	6.72753

Table 5.3 Imputed Values for Irradiance

Attributes (Irradiance)	Timestamp			
	2020-01-01 06:00:00	2020-01-01 07:00:00	2021-03-24 16:00:00	2021-09-23 15:00:00
Linear	31.708009	63.416018	180.17327	400.476829
Polynomial	12.878807	43.844149	229.42008	403.799687

Spline	16.166958	44.309713	229.82424	404.048076
--------	-----------	-----------	-----------	------------

The selection of spline interpolation was based on the observation that the previous pattern did not conform to a straight line; instead, it exhibited a more polynomial shape. Spline interpolation was preferred due to its ability to capture the smoother curve evident in the data, providing a more accurate representation of the missing values.

5.2. Feature Engineering

5.2.1. Transformation of Wind Direction

The collected data incorporates wind direction, represented in degrees. However, analyzing angles in their raw degree form poses challenges, as 0° and 360° denote the same wind direction. This lack of distinction makes the interpretation of the original meaning problematic. To address this issue and enhance interpretability, we have undertaken a transformation of the wind direction unit, converting it from degrees to radians. This transformation preserves the directional information while overcoming the ambiguity associated with the cyclic nature of wind direction data.

5.2.2. Cyclical Features of Time

In time series data, the date information typically represents only the time component. However, hidden patterns or valuable information may be embedded within the dataset. Consequently, it becomes imperative to transform the original temporal information into more interpretable features. Recognizing that time exhibits cyclic behavior, a normalization approach using the cosine function is employed. This method

may result in the same cosine value for different time points, necessitating the introduction of an additional cyclic feature to distinguish between such points.

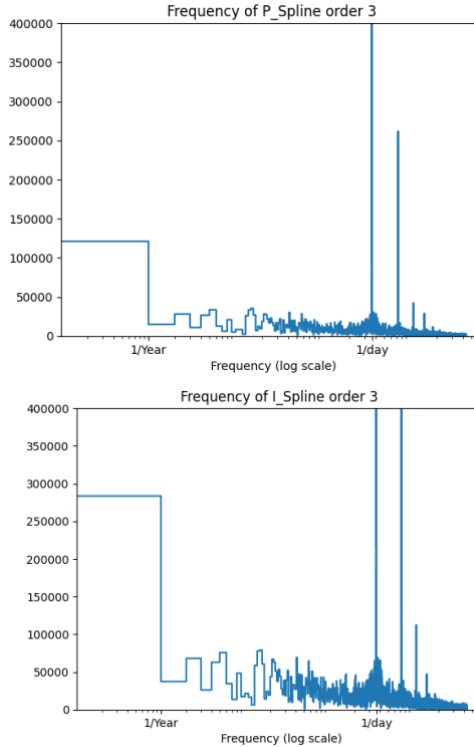


Fig 3.2 FFT Plot for Power and Irradiance

Before proceeding with the transformation of specific time data, it is crucial to identify the most important period or frequency within the dataset. Fast-Fourier Transform (FFT) is utilized for this purpose, breaking down the time-domain signal into its

constituent frequency components. The results of FFT for some attributes are illustrated in the Figure 3.2.

Analysis of the plot suggests that the high-frequency component occurs at intervals of either one year or one day. Given that the training data spans only one year, the focus is directed toward the one-day interval. Subsequently, the plot results for a one-day (24-hour) span of time data are presented in Figure 3.3

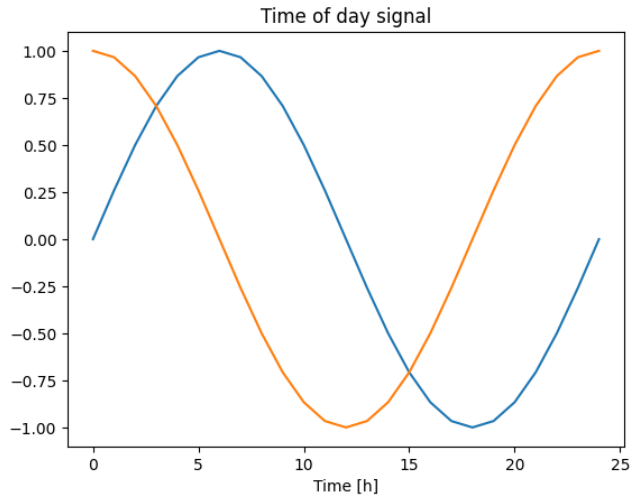


Fig 3.3 Day Sin and Hour Plot

5.2.3. Features Selection

In the original weather dataset, the multitude of variables introduces challenges related to redundancy and the potential for overfitting during the machine learning process. To enhance the efficiency of the machine learning model, a feature selection technique is employed to identify and choose the most pertinent

features from the dataset. It is crucial to distinguish feature selection from dimension reduction, as the former does not modify the fundamental content of the data beneath the selected features. Recursive Feature Elimination (RFE) is chosen as the feature selection method for this research.

Three core machine learning methods are employed within the RFE framework:

1. **Decision Tree:** A fast and basic algorithm chosen for its expediency.
2. **Random Forest:** Selected for its improved accuracy compared to a single Decision Tree.
3. **XGBoost:** Chosen for its superior accuracy in modeling complex relationships.

The number of attributes chosen will be systematically tested, ranging from 2 attributes to the total number of attributes, which is 31.

Three distinct scenarios are explored in this project concerning feature selection:

1. **Scenario 1:** The attribute "Power" is chosen as the target, with the remaining attributes used to aid in predicting the target. This scenario is particularly relevant as Power is the primary target for prediction.
2. **Scenario 2:** The attribute "Irradiance" is selected as the target, with the remaining attributes aiding in predicting this target. This is motivated by the similar signal pattern between Irradiance and Power, the main class target.
3. **Scenario 3:** Similar to Scenario 2, Irradiance is chosen as the target, but without attributes directly related to Irradiance. Attributes such as **'shortwave_radiation_instant,'**

'direct_radiation_instant,'
 'diffuse_radiation_instant,'
 'direct_normal_irradiance_instant,' and
 'terrestrial_radiation_instant' are excluded.
 These attributes are theoretically closely related
 to Irradiance, but belong to a different category.
 This exclusion aims to avoid potential bias in
 the results.

For each scenario, the result with the lowest MAE will
 be presented in the Table 5.4.

Table 5.4 Features Selection

Scenario Number	Method	MAE	Number of Features	List of Features
1	XGBoost	4.341154	12	['temperature_2m (°C)', 'apparent_temperature (°C)', 'weathercode (wmo code)', 'cloudcover_mid (%)', 'vapor_pressure_deficit (kPa)', 'windspeed_10m (m/s)', 'windgusts_10m (m/s)', 'diffuse_radiation_instant (W/m²)', 'terrestrial_radiation_instant (W/m²)', 'winddirection_100m (rad)', 'I_Spline order 3', 'Day sin']

2	Random Forest	83.95916	27	['temperature_2m (°C)', 'relativehumidity_2m (%)', 'dewpoint_2m (°C)', 'apparent_temperature (°C)', 'precipitation (mm)', 'rain (mm)', 'weathercode (wmo code)', 'pressure_msl (hPa)', 'surface_pressure (hPa)', 'cloudcover (%)', 'cloudcover_low (%)', 'cloudcover_mid (%)', 'cloudcover_high (%)', 'et0_fao_evapotranspiration (mm)', 'vapor_pressure_deficit (kPa)', 'windspeed_10m (m/s)', 'windspeed_100m (m/s)', 'windgusts_10m (m/s)', 'shortwave_radiation_instant (W/m²)', 'direct_radiation_instant (W/m²)', 'diffuse_radiation_instant (W/m²)', 'direct_normal_irradiance_instant (W/m²)', 'terrestrial_radiation_instant (W/m²)', 'winddirection_100m (rad)', 'winddirection_10m
---	---------------	----------	----	---

				(rad)', 'Day sin', 'Day cos']
3	Random Forest	87.62310	22	['temperature_2m (°C)', 'relativehumidity_2m (%)', 'dewpoint_2m (°C)', 'apparent_temperature (°C)', 'precipitation (mm)', 'rain (mm)', 'weathercode (wmo code)', 'pressure_msl (hPa)', 'surface_pressure (hPa)', 'cloudcover (%)', 'cloudcover_low (%)', 'cloudcover_mid (%)', 'cloudcover_high (%)', 'et0_fao_evapotranspiration (mm)', 'vapor_pressure_deficit (kPa)', 'windspeed_10m (m/s)', 'windspeed_100m (m/s)', 'windgusts_10m (m/s)', 'winddirection_100m (rad)', 'winddirection_10m (rad)', 'Day sin', 'Day cos']

5.3. Decomposition

Following the selection of Irradiance as the attribute for decomposition and forecasting, it becomes

imperative to leverage decomposition techniques for a nuanced understanding of its signal patterns. Irradiance serves as the target attribute due to its established similarity to the main class target, Power, and the documented strong correlation between the two variables. The decomposition stage encompasses three distinctive scenarios, each employing a specific technique:

1. CEEMDAN Decomposition:

The Irradiance attribute undergoes decomposition using the Complex Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) technique. This method maximizes the number of result components, providing a detailed breakdown of the signal patterns.

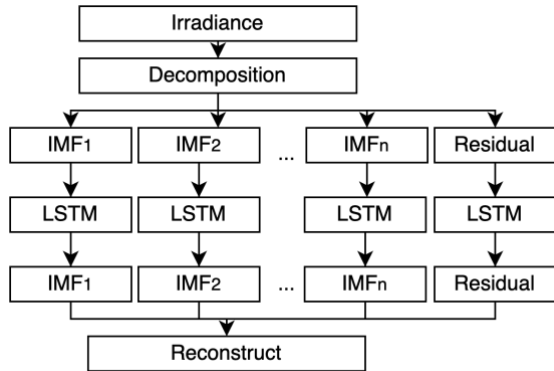


Fig 3.4 Decomposition Scenario 1

2. EWT Decomposition:

The Empirical Wavelet Transform (EWT) technique is applied to decompose the Irradiance attribute. In this scenario, the number of result components from the decomposition is deliberately set to three, allowing for a focused analysis of key frequency components.

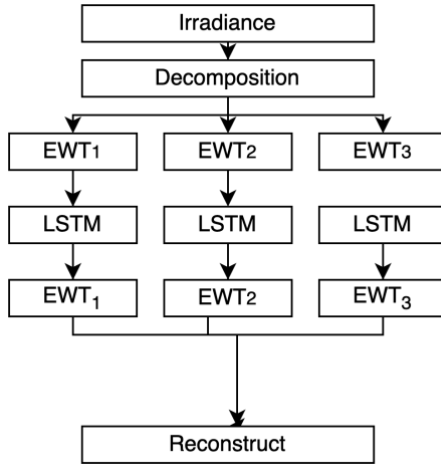


Fig 3.5 Decomposition Scenario 2

3. Combination of EWT and CEEMDAN:

A combination approach involves an initial decomposition using CEEMDAN, followed by a secondary decomposition of the first and last components using EWT. This method addresses challenges encountered in forecasting the initial and final components, presenting a nuanced solution.

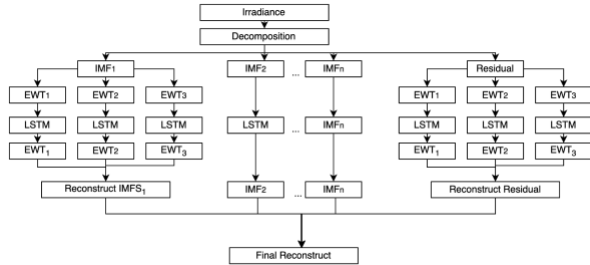


Fig 3.6 Decomposition Scenario 3

The subsequent steps in the analysis involve systematic handling of the decomposed components. This includes data splitting into training, validation, and testing sets, scaling the splitted data, model building tailored to the characteristics of each component, and the meticulous reconstruction of forecasted results. Detailed explanations of each step will follow.

5.3.1. Split into Train, Val, and Test

The Irradiance data for the year 2020 is divided into monthly intervals to mimic the "Cross Validation" method, promoting model robustness. The dataset is split into 75% for training and validation and 25% for testing. Within the training and validation set, an additional split of 75% for training and 25% for validation is applied. This method ensures diverse data training to prevent overfitting. The plot illustrates the distribution of train (blue), validation (orange), and test (green) data.

5.3.2. Scaling the Data

To enhance the model's ability to learn, the data undergoes scaling using the "Standard Scaler." This

scaler transforms the data to a standardized range (0 to 1 or -1 to 1) without altering their intrinsic values. The use of the "Standard Scaler" maintains the original data's standard deviation and variance. The plot presents the result of the scaled data.

5.3.3. Model Building

The LSTM (Long Short-Term Memory) model is employed for forecasting the decomposed data components. The configuration for LSTM is tailored for all decomposed components, with an exception made for the last imfs component due to unique characteristics. The adjusted configuration is presented below.

- LSTM Configuration (except for the last imfs component):
 - Number of LSTM layer: 1
 - Number of neurons in LSTM layer: 300
 - Regularizer of LSTM: L2 Regularizer
 - Number of Dense layer: 2
 - Number of neurons in each Dense Layer respectively: 100, 50
 - Loss Function: Huber
 - Optimizer: Adam
 - Learning Rate: 0.00004
 - Metrics: MAE and MSE
- LSTM Configuration (last imfs component):
 - Number of LSTM layer: 1
 - Number of neurons in LSTM layer: 2400
 - Regularizer of LSTM: L2 Regularizer
 - Number of Dense layer: 2
 - Number of neurons in each Dense Layer respectively: 2400, 1600
 - Loss Function: Huber
 - Optimizer: Adam

- Learning Rate: 0.000005
- Metrics: MAE and MSE

Plots illustrate the forecasted values of the imfs-0 component for the month of January compared to the original data.

5.3.4. Reconstruction of Forecasted Results

Once each component has been forecasted, a reconstruction process aggregates all forecasted component values. The Mean Absolute Error (MAE) is then calculated by comparing the reconstructed values to the original data.

[This page intentionally left blank]

CHAPTER VI

Results and Discussion

This chapter explains the testing phase of several architectures and decomposition models that have been created. Testing is conducted to determine which decomposition technique produces more accurate predictions.

6.1. Testing Objectives

Testing is carried out on various decomposition techniques and architectures to determine which model produces more accurate predictions.

6.2. Result of Different Decomposition Scenarios

6.2.1. CEEMDAN Decomposition

To ensure the robustness and efficacy of the model in learning our data, we delve into the learning process for each decomposed component. For clarity and comprehensibility, our analysis focuses specifically on the first component (imfs-0) during the first month (January). The Mean Absolute Error (MAE) for component "imfs-0" in January is computed as **25.2688407897949**. Figure 4.1 illustrates the learning curve of the model, represented by the loss curve, providing insights into the model's performance during the training process.

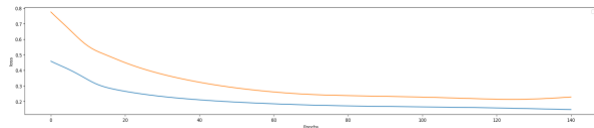


Fig 4.1 Loss Plot for Scenario 1

The comprehensive MAE values are provided in Table 6.1 for the reconstructed forecasted values of each decomposed component across all months from January to December in 2020.

Table 6.1 MAE Scores for Scenario 1

Month	Reconstructed	Reconstructed Clean
1	33.14865112	28.72327232
2	55.48101807	36.06703568
3	61.75348663	47.80811691
4	43.56692886	29.65018082
5	51.69427109	45.71088028
6	39.01631546	31.31210327
7	65.79121399	58.19694519
8	53.09525681	39.75491333
9	37.68969345	33.81318665
10	34.40693283	28.39392662
11	29.80125618	21.57585907
12	29.91107941	24.82316589

The "Reconstructed" values represent the sum of each decomposed component without additional processing steps. In contrast, the "Reconstructed_clean" values result from

summing up each decomposed component and subsequently nullifying any negative values, as Irradiance cannot be negative, with a minimum value defaulting to 0. To enhance clarity, Figure 4.2 visually depicts the relationship between the "Reconstructed" values for January 2020 and the corresponding actual data, along with the comparison of "Reconstructed_clean" values with the actual data.

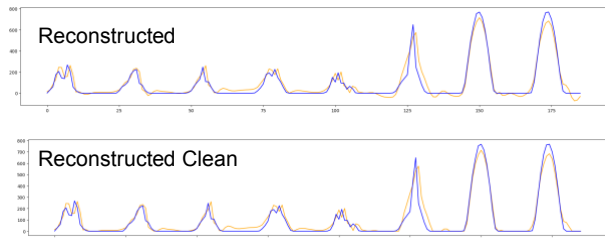


Fig 4.2 Predicted Plot for Scenario 1

6.2.2. EWT Decomposition

The Mean Absolute Error (MAE) for component "ewt-0" in January is computed as **9.6764030456543**. Figure 4.3 illustrates the learning curve of the model, represented by the loss curve, providing insights into the model's performance during the training process.

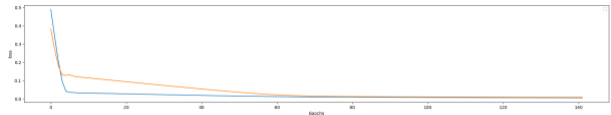


Fig 4.3 Loss Plot for Scenario 2

The comprehensive MAE values are provided in Table 6.2 for the reconstructed forecasted values of each decomposed component across all months from January to December in 2020.

Table 6.2 MAE Scores for Scenario 2

Month	Reconstructed	Reconstructed Clean
1	33.96727371	24.777174
2	44.69127274	38.68215942
3	43.04001236	37.54597855
4	36.20227432	31.50611687
5	55.57238007	49.32036972
6	31.6230526	27.00141335
7	67.2972641	59.67206955
8	56.0909996	53.75978088
9	47.98468399	38.05131912
10	37.05989838	29.90947342
11	27.06363297	22.0721035
12	27.66106415	25.4824543

To enhance clarity, Figure 4.4 visually depicts the relationship between the "Reconstructed" values for January 2020 and the corresponding actual data, along with the comparison of "Reconstructed_clean" values with the actual data.

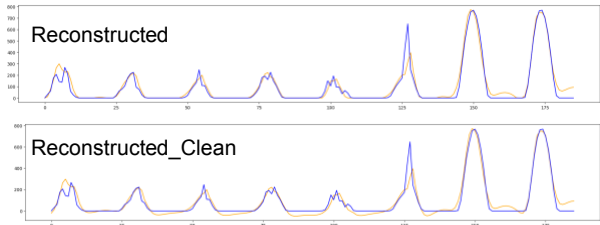


Fig 4.4 Predicted Plot for Scenario 2

6.2.3. Combination of EWT and CEEMDAN

The Mean Absolute Error (MAE) for component "imfs-0" in January is computed as **18.8876750789723**. Figure 4.5 illustrates the learning curve of the model, represented by the loss curve, providing insights into the model's performance during the training process.

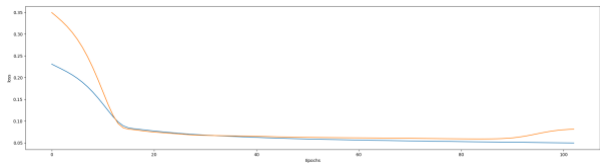


Fig 4.5 Loss Plot for Scenario 3

The comprehensive MAE values are provided in Table 6.3 for the reconstructed forecasted values of each decomposed component across all months from January to December in 2020.

Table 6.3 MAE Scores for Scenario 3

Month	Reconstructed	Reconstructed Clean
1	24.60662261	21.34564159
2	43.78833257	29.28154414
3	58.95116753	46.42009368
4	28.95876533	21.75088235
5	36.90345347	30.85139233
6	41.48620413	35.31233607
7	47.43954366	39.81791305
8	38.52378984	30.30232567
9	19.83969885	15.87964439
10	32.43910852	25.66001539
11	25.76553588	21.57624035
12	22.76674438	17.41699576

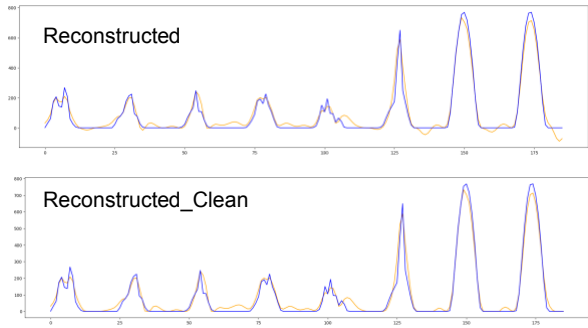


Fig 4.6 Predicted Plot for Scenario 3

To enhance clarity, Figure 4.6 visually depicts the relationship between the "Reconstructed" values for January 2020 and the corresponding actual data, along with the comparison of "Reconstructed_clean" values with the actual data.

6.2.4. Comparison between All Decomposition Scenarios

For a comprehensive understanding and analysis of the results, Table 6.4 will present a comparative overview of the methods, including their individual performance metrics, with a focus on the average for enhanced clarity.

Table 6.4 Comparison of MAE Scores Between All Scenarios

Month	Scenario 1 - Reconstructed (CEEMDAN)	Scenario 2 - Reconstructed (EWT)	Scenario 3 - Reconstructed (CEEMDAN + EWT)	Scenario 1 - Reconstructed Clean (CEEMDAN)	Scenario 2 - Reconstructed Clean (EWT)	Scenario 3 - Reconstructed Clean (CEEMDAN + EWT)
1	33,14865112	33,96727371	24.60662261	28,72327232	24,777174	21.34564159
2	55,48101807	44,69127274	43.78833257	36,06703568	38,68215942	29.28154414
3	61,75348663	43,04001236	58.95116753	47,80811691	37,54597855	46.42009368
4	43,56692886	36,20227432	28.95876533	29,65018082	31,50611687	21.75088235
5	51,69427109	55,57238007	36.90345347	45,71088028	49,32036972	30.85139233
6	39,01631546	31,6230526	41.48620413	31,31210327	27,00141335	35.31233607
7	65,79121399	67,2972641	47.43954366	58,19694519	59,67206955	39.81791305
8	53,09525681	56,0909996	38.52378984	39,75491333	53,75978088	30.30232567
9	37,68969345	47,98468399	19.83969885	33,81318665	38,05131912	15.87964439
10	34,40693283	37,05989838	32.43910852	28,39392662	29,90947342	25.66001539
11	29,80125618	27,06363297	25.76553588	21,57585907	22,0721035	21.57624035
12	29,91107941	27,66106415	22.76674438	24,82316589	25,4824543	17.41699576
Mean	44,6130087	42,3544841	35,1224139	35,4857988	36,4817011	27,96791873

6.3. Evaluation

Based on the findings in the Table 6.4, it can be inferred that the third scenario exhibits the lowest Mean Absolute Error (MAE) scores compared to the other scenarios. This conclusion is drawn from the observation that the average "Reconstructed" score for 12 months in the third scenario is **35.1224139**, whereas the first scenario reaches **44.6130087**, and the second scenario reaches **42.3544841**. Another noteworthy observation is evident after the MAE scores have been cleaned, wherein negative values are nullified (Reconstructed Clean). In this case, the average for 12 months in the third scenario is **27.9679187**, while the first scenario reaches **35.4857988**, and the second scenario reaches **36.4817011**.

The superior performance of the third scenario can be attributed to its approach of decomposing components that prove challenging to predict. This strategic decomposition enhances the model's understanding of intricate signals, consequently leading to more accurate predictions

While the average Mean Absolute Error (MAE) score strongly indicates the superior forecasting capability of the third scenario, a more detailed examination of individual monthly results is crucial. The monthly MAE results reveal that, in 10 out of the 12 months, the third scenario demonstrates more accurate predictions. This is evident in both "Reconstructed" and "Reconstructed_Clean" scores, where the third scenario consistently exhibits lower scores compared to the other two scenarios.

However, it's noteworthy that in March and June, the second scenario outperforms the others, achieving the lowest scores. The second scenario employs only the Empirical Wavelet Transform (EWT) technique. The specific reasons why EWT (Scenario 2) outperforms the

other scenarios in March and June are yet to be determined. This anomaly highlights an area for further investigation in future research, aiming to uncover the conditions or characteristics under which EWT decomposition proves more effective than CEEMDAN.

Given that the third scenario consistently demonstrates the most accurate prediction results for the majority of our data (10 out of 12 months), we leverage this scenario for further predictions. Previous predictions were made for each decomposition scenario, with a focus on a one-hour forecast after the input data. Table 6.5 presents the results for the next 3-hour prediction.

Upon examination of Table 6.5 becomes evident that the prediction accuracy diminishes as the model forecasts multi-step further into the future beyond the input data. This is substantiated by the observation that the model encounters increasing difficulty in predicting 3 hours ahead compared to 2 hours or 1 hour after the input data. The Mean Absolute Error (MAE) consistently rises as the time interval increases, indicating a decline in predictive performance with an extended forecast horizon.

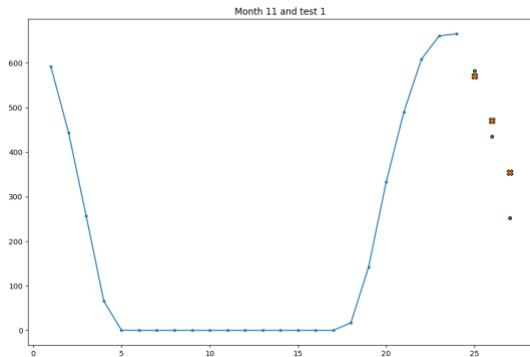


Fig 4.7 Multi-Step Prediction Plot

To provide a visual representation of the diminishing prediction accuracy with extended forecast horizons, Figure 4.7 illustrates the plot for data points at $t+24$, $t+25$, and $t+26$ based on the input data from t to $t+23$ in month November.

Table 6.5 MAE Scores for Multi-Step Prediction

Month	Reconstructed	Reconstructed Clean
1	[41.87842036544779, 57.24819747211171, 79.60328083582337]	[36.15591490873973, 50.26286508408373, 67.88735718599666]
2	[43.57846315843028, 61.45683529370894, 78.64361984568002]	[31.064097406242436, 44.57518325335551, 57.507566670971165]
3	[84.29885966678745, 82.96855168866031, 148.993261423679]	[62.30203982800183, 65.97937337554521, 112.80962838414918]
4	[47.73099644333601, 60.91945478875778, 71.90141584784529]	[35.476656578310376, 48.753432910251504, 60.21449813773968]
5	[49.95187751138506, 79.0888467026741, 103.46112062669403]	[44.31978150799664, 71.83574408005475, 94.3548517464366]
6	[40.75797383036858, 57.969726275419696, 93.3074683267508]	[35.766719877702606, 49.839266775323935, 77.76083758589733]
7	[62.10751302450047, 89.22295434370159, 121.56589248582658]	[53.46963775977511, 75.01552837402814, 100.73265669314566]
8	[47.96898645815802, 67.06859926490823, 80.620853385582]	[36.38188360906697, 51.362521200816104, 60.726247806778524]

9	[33.70004205242214, 48.132446104514194, 54.049625823870045]	[24.80022772355784, 35.52049235783853, 39.37926769434087]
10	[32.73233478946002, 39.80376825282982, 46.262254582411416]	[25.786977377728466, 32.25886576868561, 37.597324779148686]
11	[33.77999706922027, 44.77855868566453, 56.112338053103514]	[26.57083638538509, 35.83619133171333, 44.907667585923264]
12	[29.656028700312802, 43.89570067043935, 54.79801975408355]	[21.800606990924464, 34.85957333105303, 44.3720940854381]

The observed increase in Mean Absolute Error (MAE) with extended prediction horizons is understandable given the model's data input constraints. When predicting one hour ahead, the model has access to sufficient preceding data. However, for two hours ahead, the model lacks the data from the intervening hour, and for three hours ahead, the model misses data from both the first and second hours after the input data. This limitation in available historical data explains the rising MAE, as the model faces increasing challenges in accurately predicting further into the future without complete information from the preceding hours.

[This page intentionally left blank]

CHAPTER VII

Conclusion

7.1. Conclusion

The study has demonstrated the effectiveness of advanced decomposition techniques, particularly in the third scenario, to achieve more accurate solar power predictions. The strategic use of CEEMDAN and EWT contributed to consistently lower Mean Absolute Error (MAE) scores, highlighting the importance of thoughtful decomposition in addressing challenging prediction components.

7.2. Suggestion

Suggestions for the design of the solar power forecasting model architecture are as follows:

- a. **Explore Diverse Decomposition Techniques:** Investigate alternative methods for enhanced solar power prediction.
- b. **Refine Time-Series Models:** Optimize architectures to address temporal dependencies and data limitations.
- c. **Incorporate ML Advances:** Integrate newer machine learning techniques for improved accuracy and robustness.

[This page intentionally left blank]

REFERENCES

- BREIMAN, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32. 10.1023/A:1010933404324
- Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, 785-794. 0.1145/2939672.2939785
- Chollet, F. (2021). *Deep Learning with Python, Second Edition*. Manning.
- Friedman, J. (2009). *The Elements of Statistical Learning*. Springer.
- Gilles, J. (2013). Empirical wavelet transform. *IEEE Transactions on Signal Processing*, 61(16), 3999-4010. 10.1109/TSP.2013.2265222
- GUYON, I., WESTON, J., & BARNHILL, S. (2002). Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning*, 46(1-3), 389-422.
- Torres, M. E., Colominas, M. A., Schlotthauer, G., & Flandrin, P. (2011). A complete ensemble empirical mode decomposition with adaptive noise. *Signal Processing*, 91(12), 2781-2799.

[This page intentionally left blank]

AUTHOR'S BIODATA

Name : Angela Oryza Prabowo
Place, Date of Birth : Samarinda, July 20th 2002
Gender : Female
Phone Number : +6282298295978
Email : angelaoryza@gmail.com

ACADEMIC

University : Departemen Teknik Informatika –
FTEIC , ITS
Batch : 2020
Semester : 7 (Seven)

