ANSWER SHEET DAC 2024

PRELIMINARY ROUND

DATA ANALYSIS COMPETITION 2024

TEAM NAME: SDM Tinggi

TEAM ID: O187

CHAPTER I: Introduction

Due to environmental sustainability concern, there's has been a massive global energy sector shift in the pas few years. Mostly with renewable energy sources especially solar energy, taking centre stage for sustainable energy solutions. Because it can produce completely pure electricity and has many environmental benefits, clean energy—which is produced by directly harvesting solar heat—has grown in popularity. Greenhouse gas emissions can be decreased since solar panels can produce electricity without emitting any harmful substances into the atmosphere. While at the same time offering more affordable energy costs. We truly believe that sustainability energy will likely depends on how we able to optimizing the solar panel in the future

Accurately forecasting solar panel generation can prevent supply-demand imbalance in the market. Accurate predictions can help improve supply efficiency and determine how the operation of the heaven energy production. How the planned operation is, how stable the grid energy is, and how well the panels are maintained. All these predictions can be done by training advanced machine learning models, using complete datasets. So with precise predictions, it is possible that we can minimize the risk of uncertain solar power generation supply due to different environmental conditions.

As leading renewable company, DAC Green Energy Company is face challenges from the unpredictable nature of energy production due to weather uncertainty, while need to keep the grid stable. To solving the problem and maintain energy sustainability targets, company need to make an accurate solar output prediction models. This report will outlines the development of robust forecasting model that utilize hybrid approach CNN and LSTM. The CNN-LSTM model is designed to capture spatial patterns and temporal trends in solar energy data, provides more reliable and precise forecast.

The goals of this report analysis are clear: to fine-tune energy production by predicting solar power output with greater accuracy, to strengthen grid stability and reliability, and to apply cutting-edge deep learning models to improve the accuracy of solar energy forecasts. By focusing on data from the fourth quarter of 2017, this study aims to offer DAC Green Energy actionable insights for enhancing their solar power production strategies, contributing to the larger objective of reducing reliance on fossil fuels and moving toward a more sustainable energy future.

CHAPTER II: Theoretical Framework

Based on the scientific article reference entitled "Deep Learning Enhanced Solar Panel Forecasting with AI-Driven IoT" [2]. Research data shows the best model framework for predicting solar panel output energy, namely the CNN-LSTM model. This method works with the best (lowest) MAPE, RMSE, MAE values, especially for short-term predictions. Based on the research conducted, this model is able to outperform traditional time series prediction methods, such as MLP, LSTM, and ALSTM for solar energy generation forecasting. CNN and LSTM Combined for a More Advanced Deep Learning Framework. In this case, the two techniques are combined to produce better forecasting results for the PV energy generation forecasting problem.

2.1 Long Short Term Memory Neural Network

LSTM cell structure effectively resolves the gradient explosion/vanishing problems. There are 4 key elements in flow chart of the LSTM model illustrated in **figure 1**. The input, forget, and output gates are used to control the update, maintenance, and deletion of information contained in cell status. The forward computation denoted as

$$\begin{gathered} f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right), \\ i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] + b_i\right), \\ O_t = \sigma\left(W_O \cdot [h_{t-1}, x_t] + b_o\right), \\ \tilde{C}_t = \tanh\left(W_C \cdot [h_{t-1}, x_t] + b_C\right), \\ C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t, \\ h_t = O_t \cdot \tanh\left(C_t\right), \end{gathered} \tag{2.1}$$

Table 2.1. Meaning of Symbols

W_f	the weight matrix of getting gate
W_i,	the weight matrix of input gate
W_o	the weight matrix of output gate
b_f	forgetting gate
b_i	the input gat
b_o	the output gate

2.2 Overall Forecasting Framework

Based on overall forecasting framework, the CNN is responsible for extracting time series features from the data, and the LSTM network responsible in capturing the high-latitude features. Both of CNN and LSTM model is a powerful approach for time series forecasting, particularly predicting the solar output energy. By integrating these two models, we can leverage the strengths of both techniques to achieve more accurate forecasting results.

As illustrated in **figure 2**, the CNN-LSTM hybrid model processes the input data through several convolutional layers to extract relevant features, which are then fed into the LSTM layers. The LSTM layers further process the data to capture long-term dependencies, making it effective for time series forecasting.

The proposed CNN-LSTM hybrid model represents a significant advancement in the field of solar energy forecasting. This approach confirms to be outperforms traditional time series prediction methods, and providing higher accuracy for both short-term and very short-term forecasting. By focusing on relevant features and tailoring the model to seasonal variations, the proposed framework offers a robust and reliable solution for PV power forecasting.

CHAPTER III: Analytical Steps

This chapter details the step-by-step process used to prepare data and build the CNN-LSTM model. Each section addresses a crucial aspect of the analytical process.

3.1. Dataset Description and Data Sources

The dataset consists of multiple files, each providing distinct but complementary information essential for solar power output analysis:

- train.csv: This file represents the data spanning from January 2014 to September 2017. It includes records of solar power output along with various features related to weather and solar irradiance.
- test.csv: This file represent test data from October 2017 to December 2017. That will be use to evaluate the model's performance by predicting the result of the data that hasn't discover during training.
- sample_submission.csv: A submission template format. It is used to ensure that the final model predictions conform to the required structure for evaluation.
- metadata.csv: Brief information about the dataset features definition, providing context and details on the variables present in the datasets.
- Solar Irradiance Data: provide solar irradiance dataset, includes: Solar_Irradiance_2014.csv, Solar_Irradiance_2015.csv, Solar_Irradiance_2016.csv, Solar_Irradiance_2017.csv.

3.2. Data Preprocessing

1. Data Loading and Combination

   The first stage in the preprocessing of the data was to load the several datasets that were needed for the study. Among the datasets are:

   - Train and Test Data: Include features and target values that are going to be use for training and testing process.
   - Solar Irradiance Data: Dataset files that list the solar irradiance information data, such as Direct Normal Irradiance (DNI), Diffuse Horizontal Irradiance (DHI), Global Horizontal Irradiance (GHI), etc.
   - Weather Data: A meteorological insight such as temperature, humidity, and wind speed. Which most relevant factors influencing solar energy production.

2. Datetime Conversion

   Datetime conversion is critical in time series analysis, where accurate time-based indexing is required for proper data alignment and modelling. In this step:

- Solar Irradiance Data: A column named Timestamp was created by combining the Year, Month, Day, Hour, and Minute columns. The combined datetime was then converted into a Timestamp object.
- Train and Test Data: The Timestamp column was initially stored as a string in the format '%b %d, %Y %I%p'. This string format was converted to a Timestamp to facilitate time-based operations such as merging and resampling.
- Weather Data: The date_time column was also converted into a Timestamp object. This conversion ensures that the weather data can be accurately aligned with the solar irradiance data.

3. Data Merging
   A full dataset comprising all pertinent features from sun irradiance, weather, and the initial training and testing data was produced by performing data merging. The following were involved in the merging process:
   - Merge Solar Irradiance and Weather Data: Combine The solar irradiance data with the weather data using the Timestamp column from the irradiance data. Then the date_time column from the weather data.
   - Merging Combined Data with Train and Test Data: This process resulting combined data (solar irradiance + weather) was then merged with the train and test datasets using the Timestamp column.
4. Handling Missing or Null Values
   The dataset was first examined to identify missing or null values across all columns. The presence of missing values can significantly impact the accuracy of predictive models if not handled appropriately. The extent of missing data was quantified for both the training and testing datasets.
5. Handling Duplicate Values
   Duplicate records in the dataset can introduce bias and reduce the efficiency of data analysis. Therefore, it is essential to identify and remove any duplicates before proceeding further. Our process involves:
   - Identifying Duplicates: The initial size of the dataset is recorded before checking for duplicates.
   - Evaluating the Impact: The shape of the dataset (i.e., the number of rows and columns) is compared before and after duplicate removal. A reduction in the number of rows indicates that duplicates were present and successfully eliminated.
6. Outlier Detection
   Outliers will inevitably give misleading statistical analyses and modelling efforts, leading to inaccurate insight. Detecting and handling outliers is a very critical steps in the data cleaning process:

- Visual Inspection: A boxplot is used to visually inspect the distribution of the % Baseline variable for potential outliers.
- Removing Outliers: A systematic approach is applied to remove outliers from the dataset. The interquartile range (IQR) method is used. The difference between before and after cleaning can be seen in **figure 10** and **figure 11**.
- Impact Evaluation: The number of rows in the training dataset is compared before and after outlier removal to quantify the extent of the cleaning process.

7. Train-test Split

   Prepare data for training and validation to ensure the model is evaluated on unseen data. The target variable % Baseline was separated from the features, and the features were further processed for model training. After various experiments 80:20 is the most appropriate number to get the most accurate prediction results.

8. One Hot Encoding and Data Type Conversion

   Categorical columns in the training, validation, and test datasets were transformed using one-hot encoding. The plot of categorical and the numerical columns can be seen in **figure 9**. This process creates binary columns for each category, making the data suitable for modelling. The *ColumnTransformer* was used to apply one-hot encoding to categorical features while leaving numerical features unchanged. Also, columns that representing numeric values (e.g., DHI, DNI) that were initially read as objects were converted to int or float.

9. Data Scaling and Normalization

   Data scaling was performed using the *StandardScaler* to normalize the features before inputting them into the model. This normalization is crucial as it helps in achieving faster convergence during training and generally improves model accuracy by ensuring all features contribute equally to the predictions. For instance, features like Dew Point, Solar Zenith Angle, and Wind Speed were scaled to ensure their influence on the model was balanced.

10. Modelling

    CNNs are highly efficient at capturing spatial characteristics from information. Within this framework, we can detect patterns in weather and solar irradiance data that have multiple dimensions. Using convolutional layers, the model can detect significant repetitive patterns, like daily or seasonal changes in weather that affect the efficiency of solar panels. Combined with LSTM who can monitor shifts in data over time, like the impact of past weather on current energy output in solar energy forecasting.

CHAPTER IV: Analysis of Results

We used the CNN-LSTM model combines Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) to exploit spatial and temporal patterns in data. This hybrid architecture is well suited for time series prediction, where spatial features (extracted by CNN layers) and temporal dependencies are both important for accurate predictions.

4.1. Data Exploration and Data Visualization

In this segment, we utilize a range of exploratory data analysis (EDA) methods to uncover insights in the dataset and grasp the connections between variables. Some of these methods involve looking at time series data visually, assessing the temperature distribution, studying relationships between variables, and creating scatter matrices and pairplots. Summary statistics are computed to offer a snapshot of the data's central tendencies and variability.

1. Visualizing Time Series Data

   A graph is produced for the variable % Baseline as time progresses (can be seen in **figure 4**). This visualization aids in detecting patterns, seasonal variations, and possible irregularities within the data. Understanding how solar energy output changes over time and potentially by different factors like weather conditions is essential, making the time series plot significant.

2. Histogram for Temperature Distribution

   A histogram is used to visualize the distribution of the temperature variable within the dataset. By plotting the frequency of temperature values, we can assess the overall distribution, identify the presence of outliers, and understand the central tendency and spread of the temperature data.

3. Correlation and Scatter Matrix

   To explore the relationships between multiple variables, we calculate a correlation matrix, which quantifies the degree of linear association between pairs of variables. The results of the heatmap in **figure 3** show which columns need to be used to train the model.

4. Pairplot

   A pairplot is created to visualize the pairwise relationships between selected features, including % Baseline, Temperature, GHI, and Wind Speed. This plot allows us to observe how these variables interact with each other and whether there are any apparent patterns or clusters. The results of the pairplot in **figure 15** show the relationships between selected features.

5. Summary Statistics

   Summary statistics are computed for a comprehensive set of variables, providing an overview of the central tendency (mean, median), dispersion (standard deviation, range), and distribution (min, max, quartiles) of each variable. By summarizing the data in this way, we can establish a baseline understanding of the variables before proceeding with more complex analyses.

4.2. Model Prediction and Evaluation

1. Important Variables and Their Contribution to Prediction

   Temperature, wind speed, and relative humidity play an important role in determining the level of solar energy generated. Correlation analysis shows a robust correlation between solar energy production and temperature along with relative humidity. Increased temperatures usually improve the performance of solar panels, while higher humidity can decrease their performance. Solar radiation information, comprising DNI, DHI, and GHI, exhibits the strongest relationship with solar power generation.
2. CNN-LSTM

   The model was built using a Sequential architecture that includes:
   - CNN Layers: These layers extract spatial features from the data, capturing essential patterns and structures. A dropout rate of 0.2 is applied to prevent overfitting by randomly setting a fraction of the input units to 0 during training.
   - LSTM Layer: This layer has 50 units and is used to capture temporal dependencies in the sequence data. Another dropout layer with a rate of 0.2 is added after the LSTM layer to further combat overfitting.
   - Dense Layers: A fully connected layer with 50 units and ReLU activation to introduce non-linearity and combine features learned by the previous layers.
3. Model Loss and MAE Visualization

   The following criteria were used to assess the model's performance:
   - Loss Curves: Loss curves for both training and validation datasets were plotted. For example, during training, the loss decreased from 0.12 to 0.05, indicating good convergence, while validation loss stabilized around 0.07, suggesting minimal overfitting.
   - Mean Absolute Error (MAE): Predictive accuracy was measured using the MAE. The mean absolute error (MAE) on the validation set was found to be approximately 0.03. This indicates that the model's predictions were, on average, 0.03 off from the actual values.
4. Comparison of Predictions

   Validation Data: For example, the actual vs. predicted values also can be see in **figure 14** showed the following:

- On 2014-05-26 18:00:00, the actual value was 0.0813, and the prediction was 0.0878.
- On 2016-11-23 10:00:00, the actual value was 0.4313, and the

prediction was 0.3427. This comparison indicates how well the model aligned with observed data, reflecting its overall performance.

Test Data: Predictions on the test dataset were compared against the baseline also can bee see in **figure 12** showed:

- For 2017-10-01 06:00:00, the % Baseline prediction was 0.0710.
- For 2017-10-01 07:00:00, the % Baseline prediction was 0.0992.

The % Baseline values indicate the predicted proportion of the baseline values, showing how well the model generalized to unseen data.

CHAPTER V: Conclusion and Recommendation

The goal of this project is to create a prediction model that merges CNN and LSTM to predict solar energy generation, using meteorological and solar irradiance data as inputs. Throughout the procedure, information is collected from a variety of sources such as solar energy output data, weather data, and solar irradiance data. Once the data has been processed, the CNN-LSTM model is created to merge CNN's spatial feature capturing capability with LSTM's temporal pattern analysis capability, enabling accurate prediction of solar energy output.

The process involves exploring data, preprocessing data, designing model architecture, sharing data for training and validation, and tuning hyperparameters. Throughout this procedure, a range of visualization methods is employed to analyse the data's features and the connections among variables, like temperature distribution histograms, correlation heatmaps, and pairplots. When evaluating model accuracy, the Mean Absolute Error (MAE) metric indicates that this model can offer precise predictions.

In general, our CNN-LSTM model successfully detected patterns and trends in the data analysed, offering dependable forecasts for solar energy production to be utilized by DAC Green Energy Company. It is anticipated that the adoption of this model will boost the effectiveness of generating solar energy and assist in making more informed decisions in managing energy resources, all while cutting down on operational expenses. In order to successfully implement this model on a large scale in industry, it will be crucial to provide training for teams and ensure seamless integration of technology.

REFERENCES

[1] Mobarak Abumohsen, Amani Yousef Owda, Majdi Owda, Admad Abumihsan, Hybrid Machine Learning Model Combining CNN-LSTM-RF for time series forecasting of Solar Power Generation, Ramallah: Elsevier, 2024.

[2] Hangxia Zhou, Qian Liu, Ke Yan, Yang Du, Deep Learning Enhanced Solar Energy Forecasting with AI-Driven IoT, Hangzhou: Wiley, 2021.

ATTACHMENT

Figure 1: Structure of LSTM

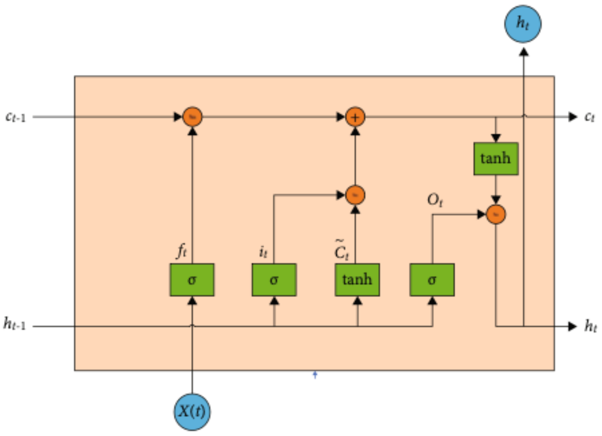


Figure 2: Proposed Model (CNN- LSTM)

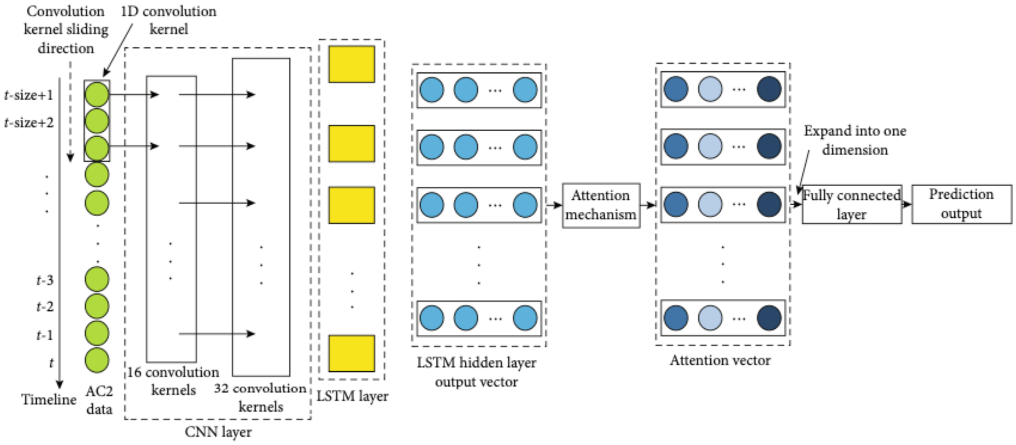


Figure 3: Feature Heatmap using Correlation Matrix

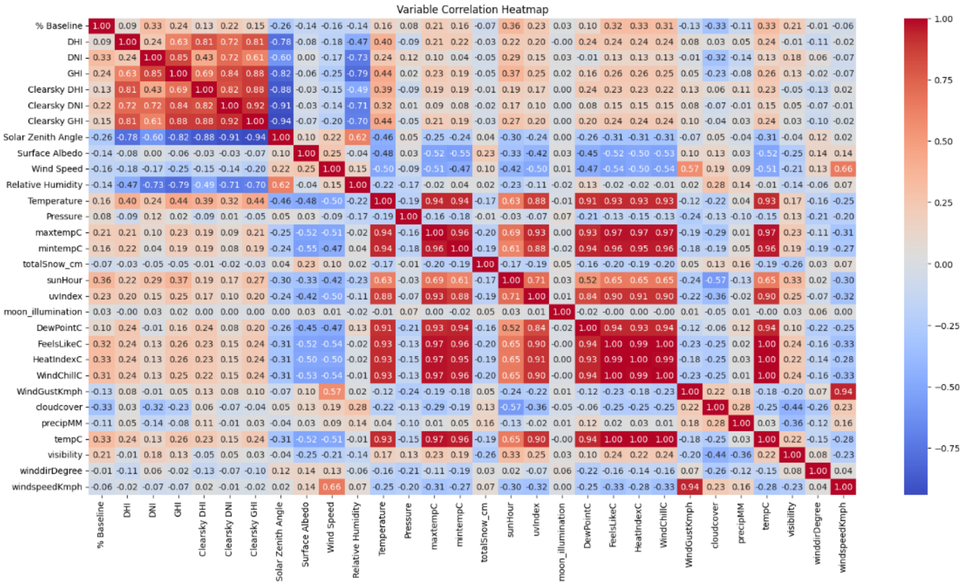


Figure 4: % Baseline Over Time (train data)

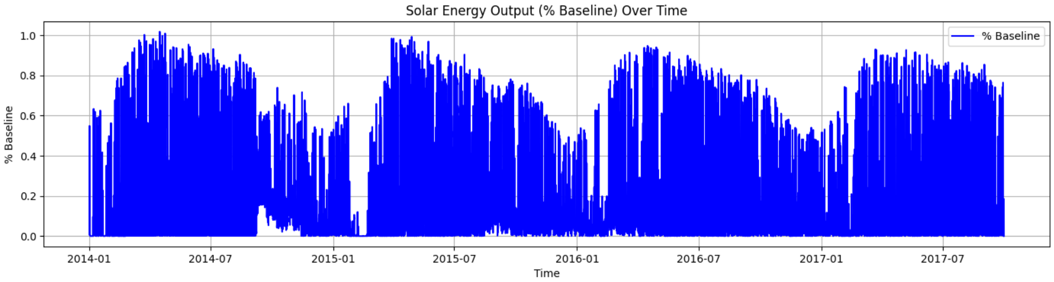


Figure 5: Null Values (train & test data)

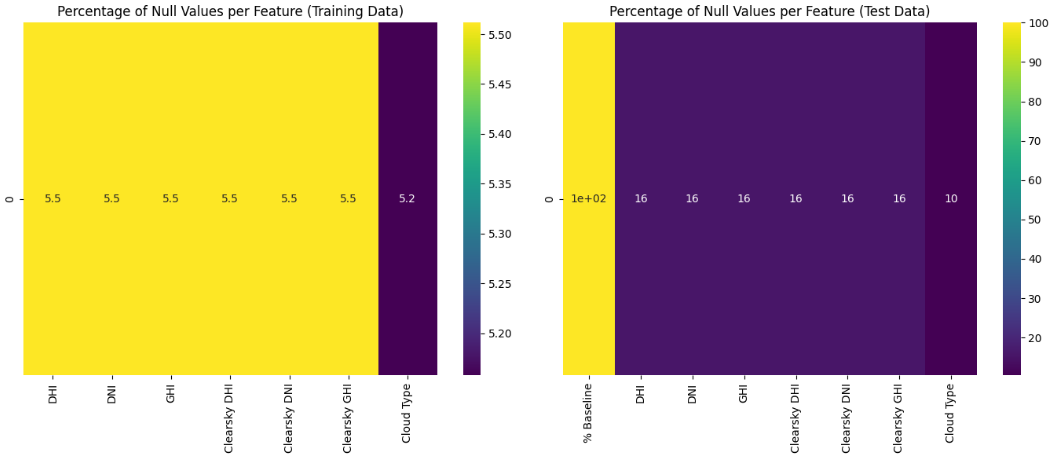


Figure 6: Missing Values (before filled)

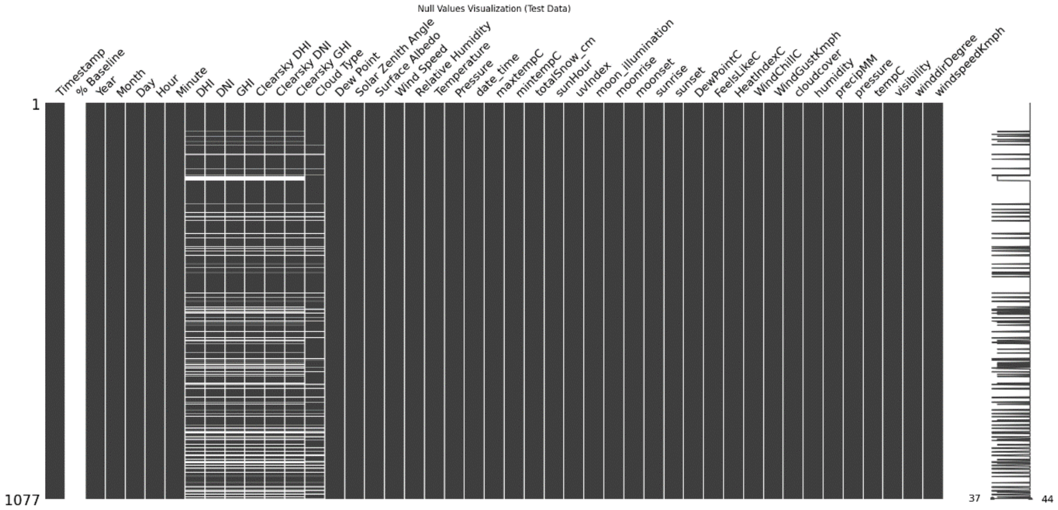


Figure 7: Missing Values (after filled)

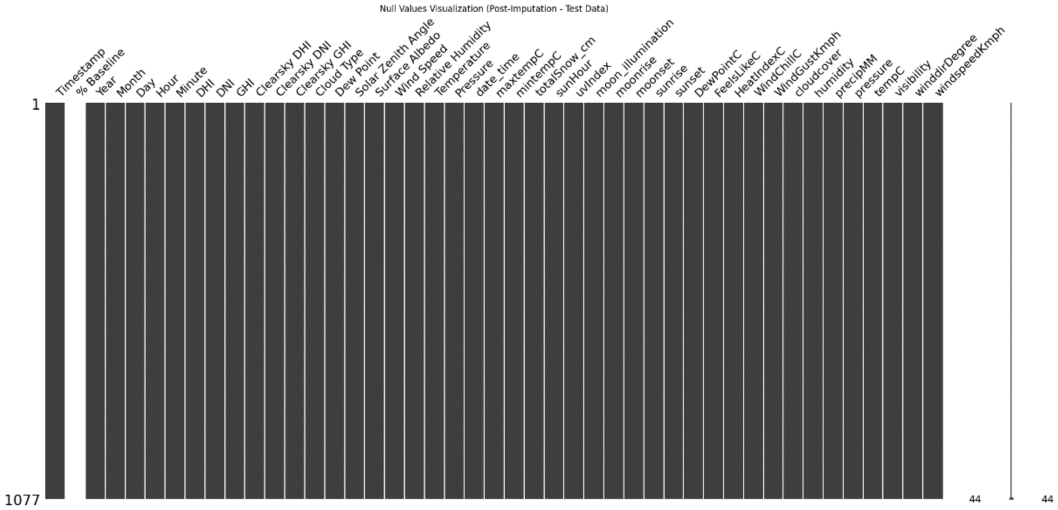


Figure 8: Number of Samples in Training and Testing Datasets

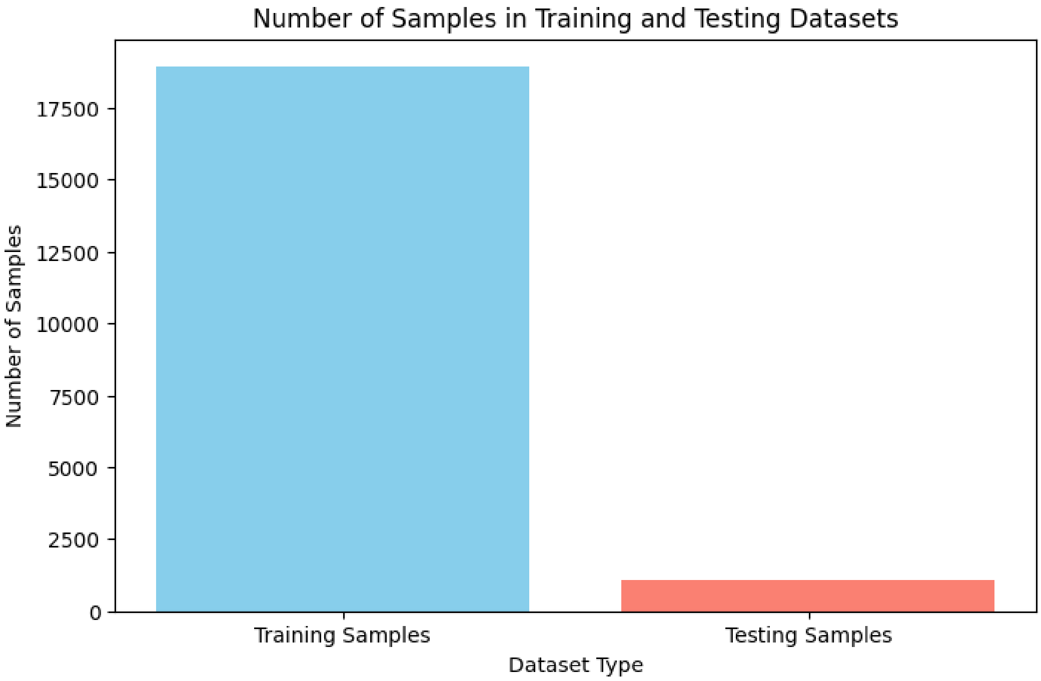


Figure 9: Column Types in Dataset

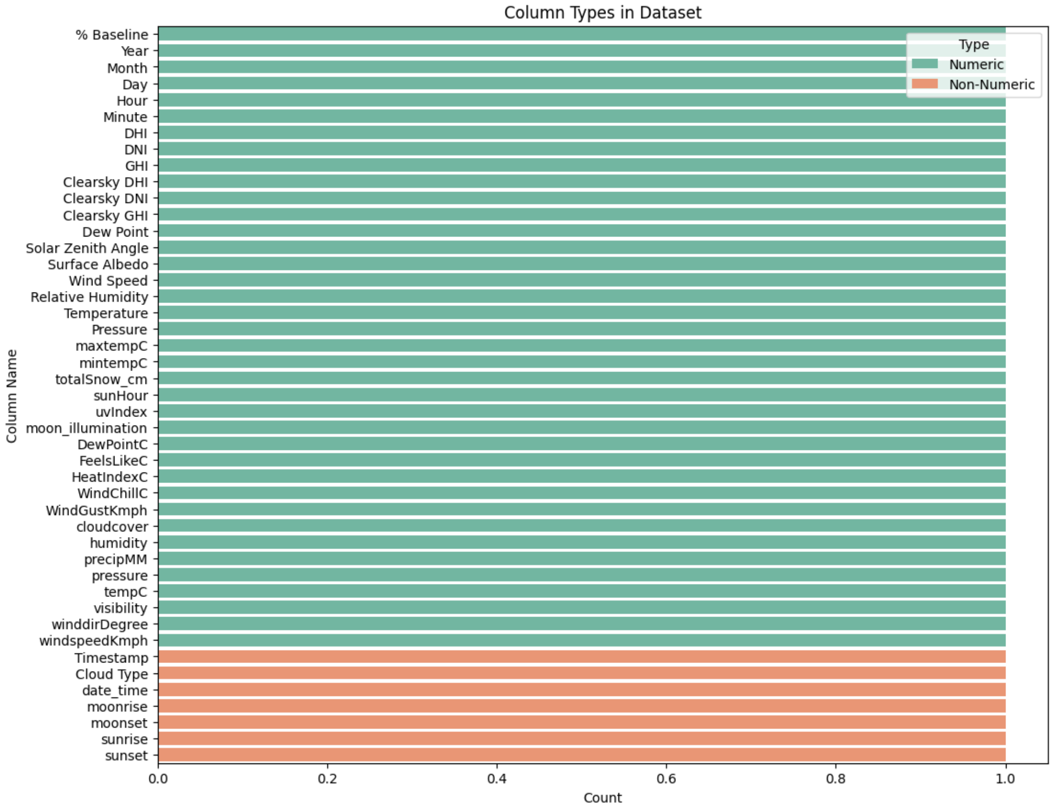


Figure 10: Box Plot Outlier (before cleaned)

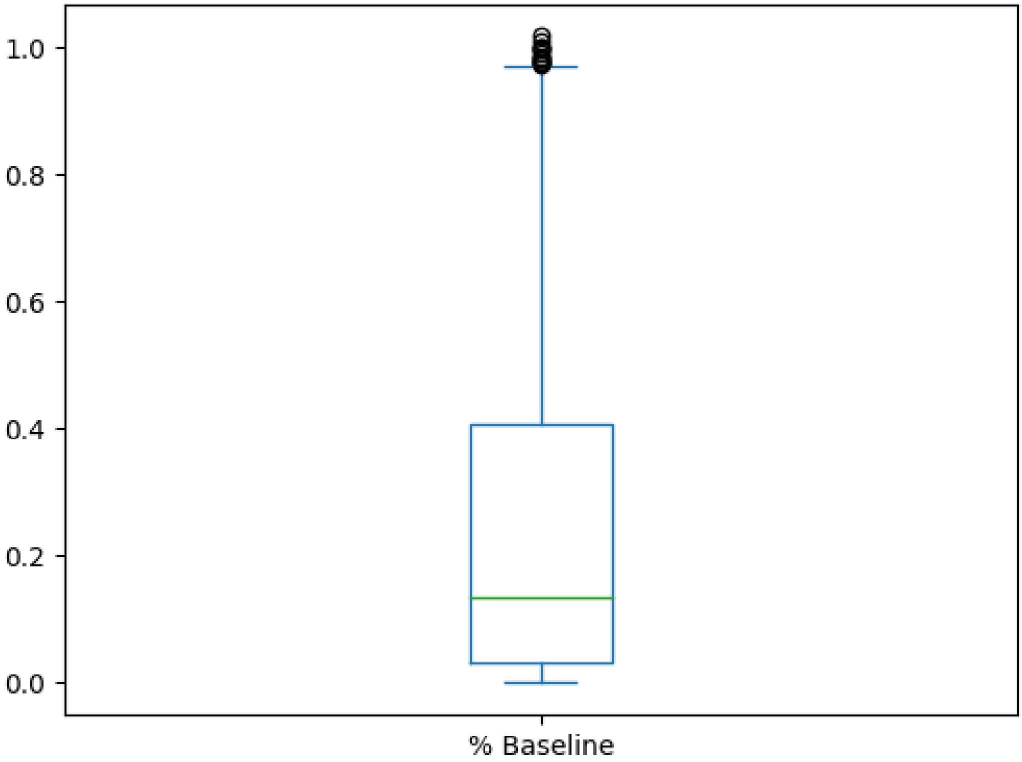


Figure 11 : Box Plot Outlier (after cleaned)

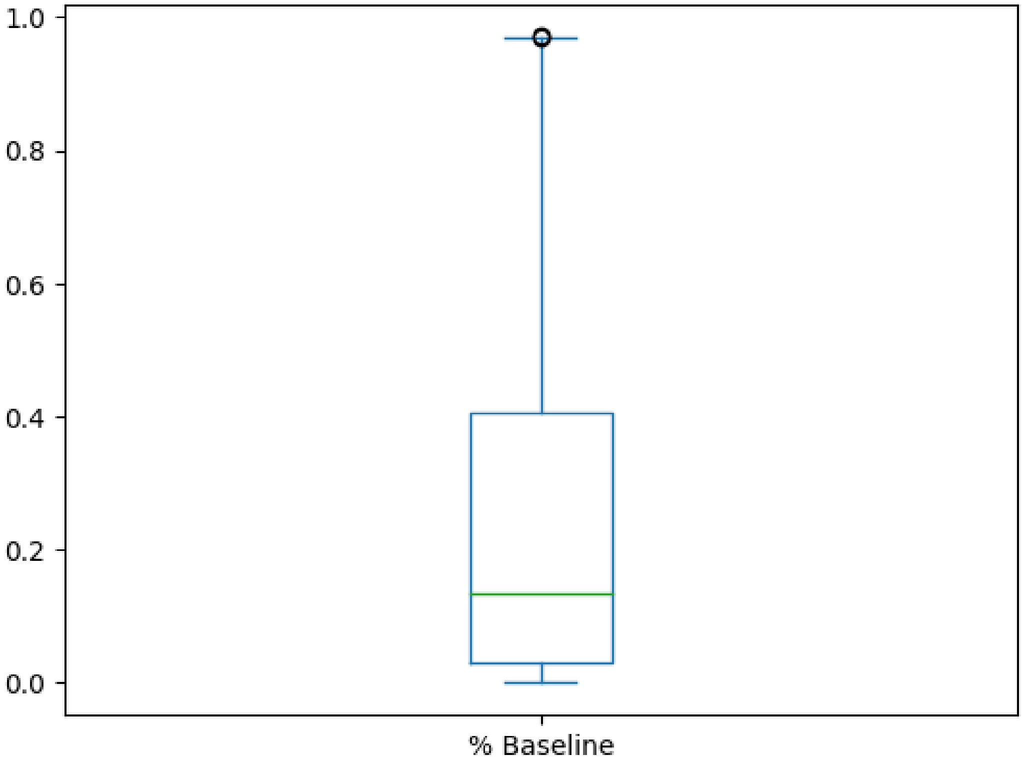


Figure 12 : Actual Predicted Baseline

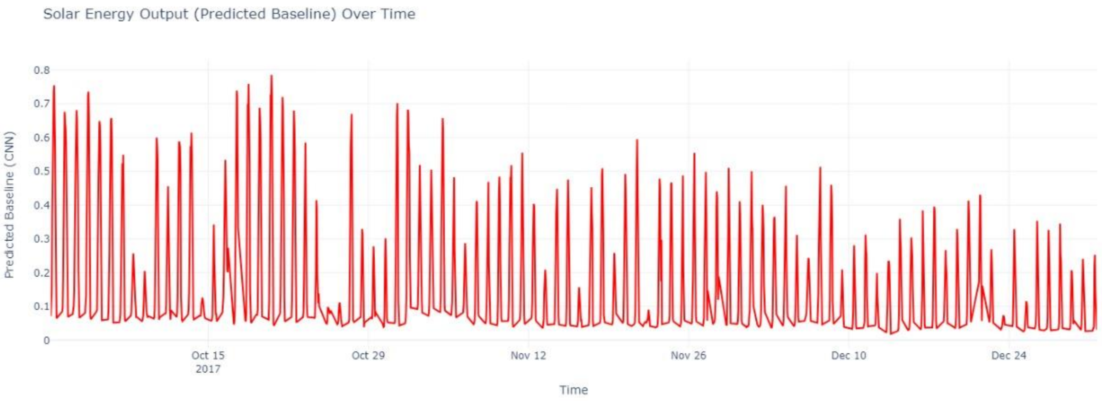


Figure 13 : Model Loss and Model MAE

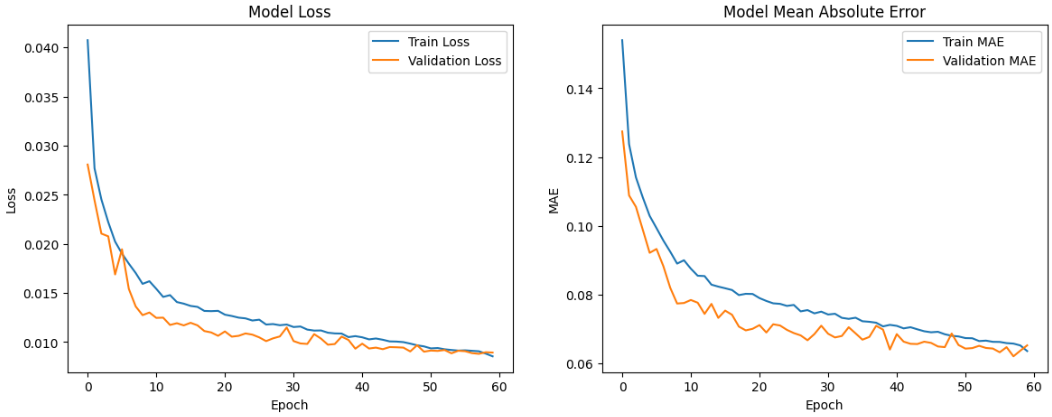


Figure 14 : Predictions based on Train & Validation Data

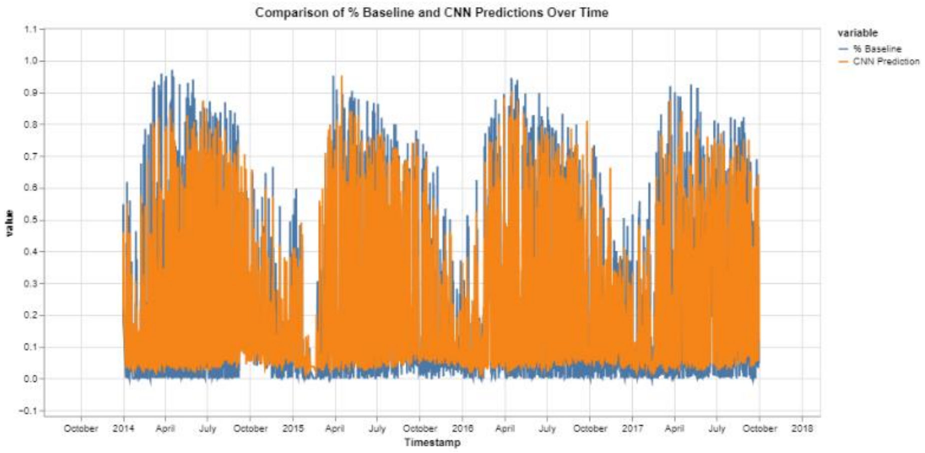
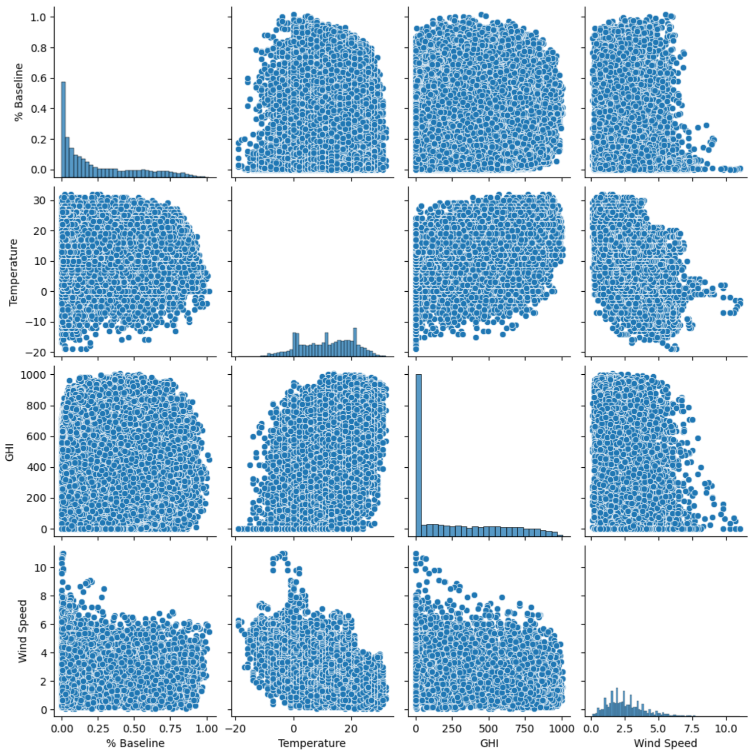


Figure 15 : Pairplot

Pairplot of Selected Features


% Baseline
Temperature
GHI
Wind Speed
% Baseline
Temperature
GHI
Wind Speed