**Contents**

[**Introduction** 2](#_Toc178330336)

[**Literature Review** 2](#_Toc178330337)

[**Objectives and Hypothesis** 3](#_Toc178330338)

[**Research Method** 4](#_Toc178330339)

[**Conclusion and Results** 9](#_Toc178330340)

[**Figures** 12](#_Toc178330341)

[**References** 13](#_Toc178330342)

# **Introduction**

Reaching to the point where we all acknowledge the end of the mass fossil fuel usage is close; we lack nonetheless sufficient efforts in reaching a climate-friendly energy production-consumption cycle as a substitute. This century with its always ever-progressive technology can potentially fulfill this transition. Renewable energies such as wind, ocean wave, and solar power has been recognized as the main sources to replace fossil fuel. Solar power, among them, is the most common one (Hereher, Mohamed, and Ahmed M. El Kenawy., 2020), as the sun's rays reach almost every corner of our planet. However, the sea is only close enough to some regions, and wind is not always a consistent element.

Electrons in a Photovoltaic (PV) cell exposed to sunlight can become excited, causing them to flow through a circuit (Al-Ezzi, Athil S., and Mohamed Nainar M. Ansari, 2022). This electron flow in direct current is called solar energy, and with the use of an inverter, it can be converted to alternating current for household usage. The amount of energy generated depends mostly on the solar panel efficiency (Ishaq, H., and I. Dincer, 2021), sunlight availability (Li, G., M. Li, R. Taylor, Y. Hao, G. Besagni, and C. N. Markides., 2022), and temperature sensitivity (Zhai, Huixing, Suilin Wang, Baolin An, and Xiaoye Dai., 2020).

Sunlight fluctuation is the biggest challenge in the viability of solar energy, as the majority of researchers in the field have strived to mitigate such effects. (Ehsanul Kabira, Pawan Kumar, Sandeep Kumar, Adedeji A. Adelodun, Ki-Hyun Kim, 2017) (Shivashankar, S., Mekhilef, S., Mokhlis, H. and Karimi, M., 2016). In this research, our prediction model aims to describe weather-related regression to understand energy fluctuations, helping to maintain reliable energy distribution. By estimating the amount of electricity a solar power system can produce over a given period—and knowing the average energy consumption—we can mitigate such fluctuations and manage potential energy shortages in advance.

Enriched, detailed, and reliable historical records of weather conditions—such as daily sunlight hours, temperature, cloud cover, precipitation levels, and more—enable a more precise forecast and provide a comprehensive understanding of how each weather attribute affects energy generation. These insights can significantly improve our ability to predict and optimize solar energy output.

# **Literature Review**

Sweeney and colleagues described various forecasting approaches and introduced several advancements for the future. Among these, Statistical Modelling, in comparison to Physical Modelling, plays a crucial role in renewable energy forecasting. They attest that, although physical models such as Numerical Weather Prediction and Hydrological Models adopt a more scientific approach, the availability of large historical datasets can make Statistical Modelling a reliable alternative. The simplicity and efficiency of these models can provide researchers and businesses with valuable insights in a shorter time (Sweeney, Conor, Bessa, Ricardo J., Browell, Jethro, Pinson, Pierre, 2019). The advantage of using Machine Learning in predicting future values based on sequential and time-dependent data helps detect complex and non-linear patterns in the data, considering the inner correlation among different attributes. (El Naqa, I. and Murphy, M.J, 2015).

Based on Bochenek and colleagues' review, the most common methods of using Machine Learning for Numerical Weather Prediction are Deep Learning, Random Forest, Artificial Neural Networks, Support Vector Machines, and XGBoost, respectively (Bogdan Bochenek, Zbigniew Ustrnul, 2022). Among 500 related articles, the most common countries examined are China, the USA, Australia, India, and Germany. This research focus demonstrates the need for solar power generation prediction in populated countries aiming toward renewable energy transition.

Scher and Messori assert that although an ensemble of numerical weather simulations can provide a confidence estimate, their very expensive methodology makes them impractical for weather prediction. In contrast, they assessed machine learning methods to determine whether they are a better approach for weather prediction. Their suggested deep learning method performs efficiently and shows the potential for substituting ensemble weather forecast models with machine learning models (Sebastian Scher and Gabriele Messori , 2018).

AbdulRaheem and colleagues demonstrated a performance evaluation of decision tree, logistic regression, and k-nearest neighbor machine learning algorithms using Python's built-in and Scikit-learn libraries. Their daily sequential data consists of daily precipitation rates, and daily maximum and minimum temperatures. As their first step, they attempted to detect data anomalies and missing values, then normalized the data using Pandas library tools. Next, they used the same data for training, accuracy checks, and error calculation for each model. They employed four common evaluation metrics in machine learning: accuracy score, precision score, recall score, and F1-score. They evaluated that the decision tree algorithm performs the best, followed by logistic regression, which performs relatively well, while the k-nearest neighbor algorithm had a relatively low accuracy rate (Muyideen AbdulRaheem, Joseph Bamidele Awotunde, Abidemi Emmanuel Adeniyi, Idowu Dauda Oladipo, Sekinat Olaide Adekola, 2022).

Seul-Gi Kim and colleagues used weather data to predict solar power generation. With a two-step modeling process using supervised machine learning methods that links unannounced weather variables with announced weather forecasts, they reached an overall R-squared value of 70.5% in their test data. Among the different models suggested by them, Adaptive Boosting and Linear Regression showed the worst performance, while ANN, K-NN, and RFR performed the best due to their higher capacity. Based on their results, the most important variables are Solar Radiation, Time Zone, Humidity, Surface Temperature, Sky Type, Vapor Pressure, Air Pressure, and Week Number, respectively (Seul-Gi Kim, Jae-Yoon Jung and Min Kyu Sim, 2019). A closer look at their research poses the idea that there might have been a strong correlation among them, which could have helped reduce the number of variables. For instance, Time Zone and Radiation or Air Pressure and Surface Temperature seem to be very related. If they had included inner correlations in their work, their results might have been different. Their final result suggests that RFR has the highest robustness and good performance in predicting solar energy generation using weather data.

# **Objectives and Hypothesis**

1. The more a solar panel is exposed to sunlight, the more energy will be generated.
2. Cloudiness has a negative impact on the amount of energy being generated.
3. Generation shortages occur more frequently where weather conditions fluctuate more.
4. Using historical data, can we provide a solar energy generation prediction regression to estimate energy generation based on weather forecasts? If yes, which weather attributes have impacts, and how significant are their impacts?
5. How can we test the accuracy of our predictions?

# **Research Method**

Understanding the factors that influence the efficiency of solar power systems is important in optimizing energy generation. Some of these important factors are the duration of exposure to sunlight, Global Horizontal Irradiance, temperature, and rain and snow precipitation. Many researchers have focused on these attributes of weather data the most (Wang, Jianxiao, Haiwang Zhong, Xiaowen Lai, Qing Xia, Yang Wang, and Chongqing Kang., 2017) (Gordo, E., Khalaf, N., Strangeowl, T., Dolino, R. and Bennett, N, 2015) (Meral, M.E. and Dincer, F., 2011).

In this article, historical weather data and energy generation information[[1]](#footnote-1) — consisting of 196,776 rows and 17 columns — have been examined. It contains hourly records of solar energy generation and various weather metrics. The key columns are:

* Energy delta [Wh]: The solar energy generated (target variable)
* GHI: Global Horizontal Irradiance (solar radiation)
* Temp: Temperature in degrees Celsius
* Clouds\_all: Cloud coverage percentage
* Wind\_speed: Wind speed in m/s
* Rain\_1h and Snow\_1h: Rain and snow precipitation in the last hour (in mm)
* SunlightTime and DayLength: Time of sunlight and total day length in seconds

Other features, such as hour and month, can help analyze temporal trends. In the first step, a search for missing values and anomaly detection was performed. For continuous quantitative attributes, such as temperature, the total average value was considered, and for binary attributes, such as IsSun (Yes or No), the total mode value was replaced (Somasundaram, R.S. and Nedunchezhian, R., 2011). Figure 1 shows the distribution of key columns.

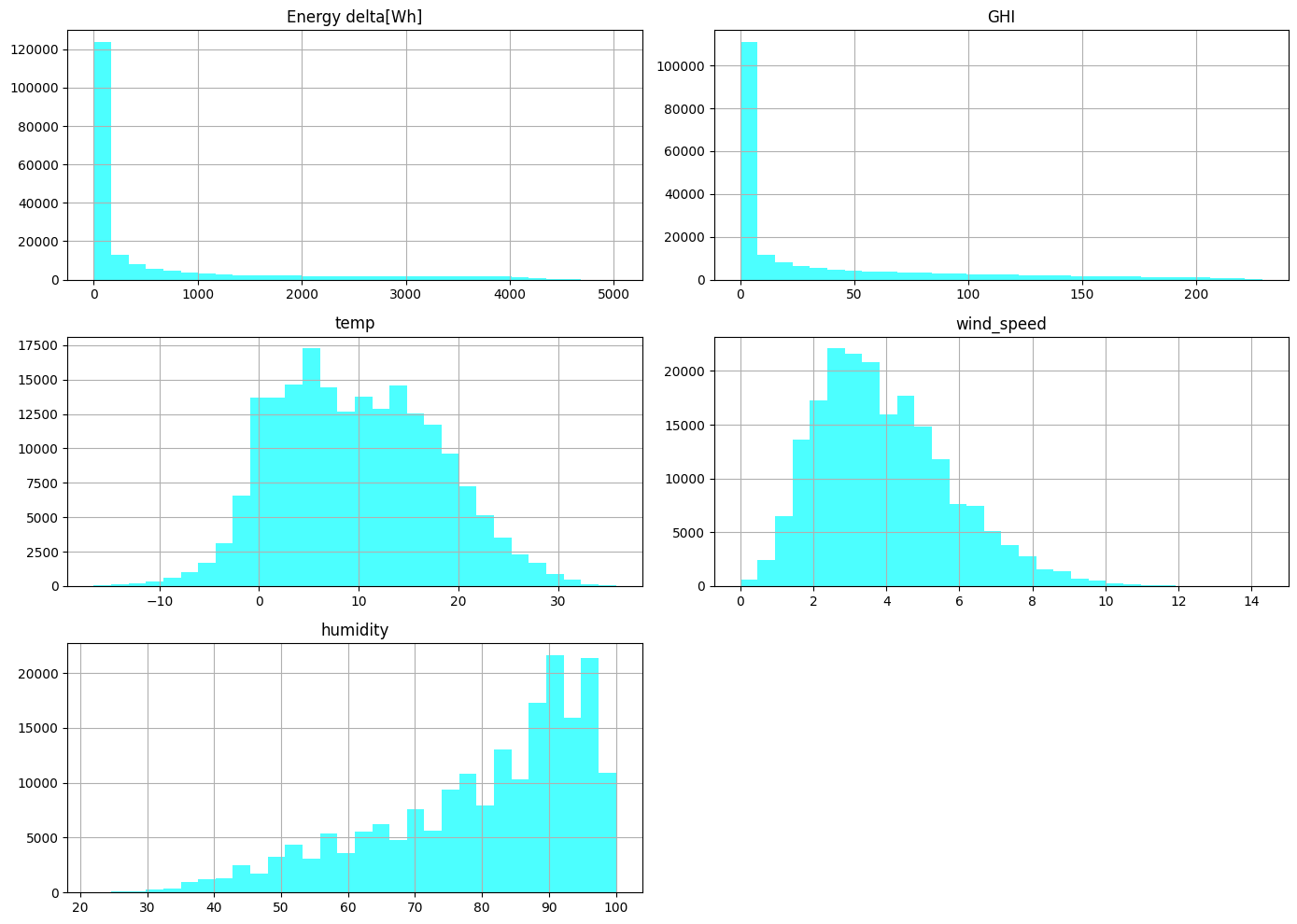


Figure 1 Distribution visualization of Key Columns

After the pre-processing, data description was examined, including the number of records, average, standard deviation, minimum, data quartiles, and maximum. This examination leads to the following insights:

* A high standard deviation for GHI (51.89, which is greater than the mean of 32.58) suggests significant fluctuations in solar irradiance, which is expected due to its dependence on time and weather conditions.
* Temperature ranges from -16.6°C to 35.8°C, indicating diverse weather conditions. The median temperature (9.4°C) is close to the mean (9.79°C), suggesting a somewhat balanced distribution. • Both rain and snow averages are close to zero; however, rain can reach up to 8.09 mm in one hour, and snow up to 2.82 mm, which could be significant for short periods.
* IsSun, with a mean of 0.524, indicates that sunlight is present roughly half of the time.
* Sunlight Time / Day Length has a mean of 0.265, suggesting that, on average, about a quarter of the day length has effective sunlight time.

Then, to comprehend relationships among the attributes, a comprehensive correlation analysis was conducted. Significant positive relationships were found between GHI and Energy delta, temperature and GHI, and temperature and Energy delta, respectively. The significant negative relationships were between humidity and GHI, humidity and Energy delta, and humidity and temperature. This analysis highlights the key factors in predicting Energy delta.

The energy generation over the years, shown in Figure 2, illustrates that throughout the year there are no significant deviations, and only a few fluctuations have occurred. At first glance, this figure supports the idea of relying on solar energy generation, as its production pattern over time seems predictable. However, further examination is still necessary.

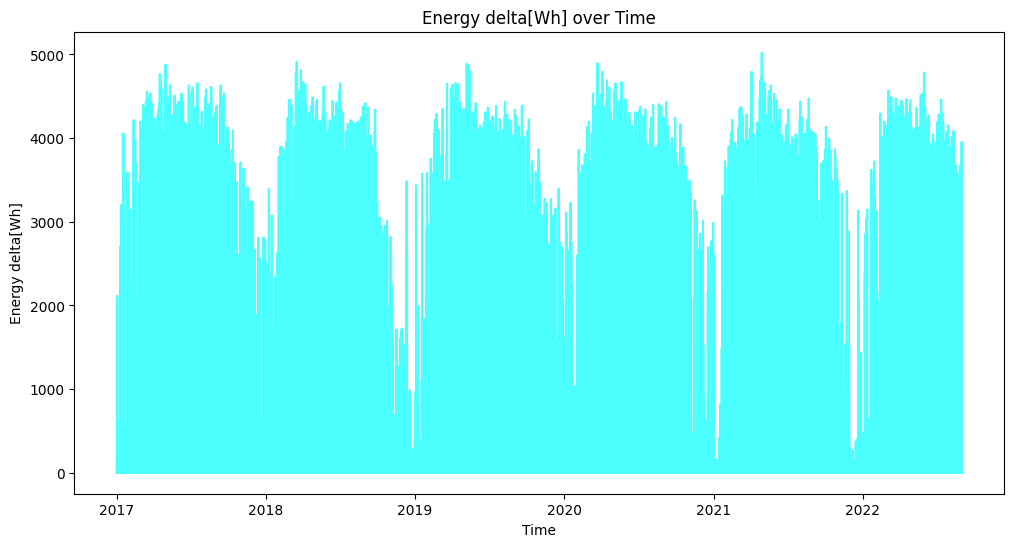


Figure 2 Energy generation level over time

The impact of solar radiation (GHI) on energy generation throughout different times of the day has been demonstrated by grouping the data by hour and calculating the mean GHI and energy generation for each hour, as shown in Figure 3. This figure illustrates that the majority of energy is produced between 9:00 AM and around 11:30 AM. Closer examination shows that the maximum energy produced occurs before the peak solar radiation is reached. This aligns with other research describing the impact of the temperature sensitivity characteristic of photovoltaic materials (Zhai, Huixing, Suilin Wang, Baolin An, and Xiaoye Dai., 2020).

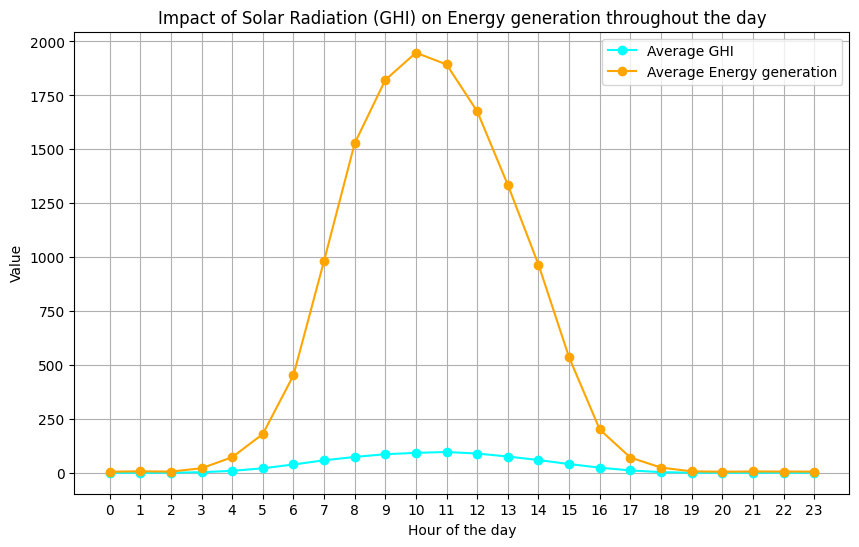


Figure 3 Impact of solar Radiation (GHI) on Energy Generation throughout the day

The question, 'Which months have the highest and lowest solar energy generation, and how do weather patterns influence these trends?' can be addressed by grouping the data by month and summing the energy generation for each month, as shown in Figure 4. This analysis reveals that from April to August, for five consecutive months, energy generation remains relatively stable, with a minimum of at least 14,000,000 Wh. However, the significantly lower energy generation from October to February raises sustainability concerns. To address this issue for household usage, energy providers must compensate for the shortage by sourcing energy from other methods. This analysis suggests that while solar energy should be considered a primary energy source, sectors with year-round energy demand could benefit from encouraging consumption during peak generation months and reducing or halting usage during lower generation periods to mitigate the imbalance.

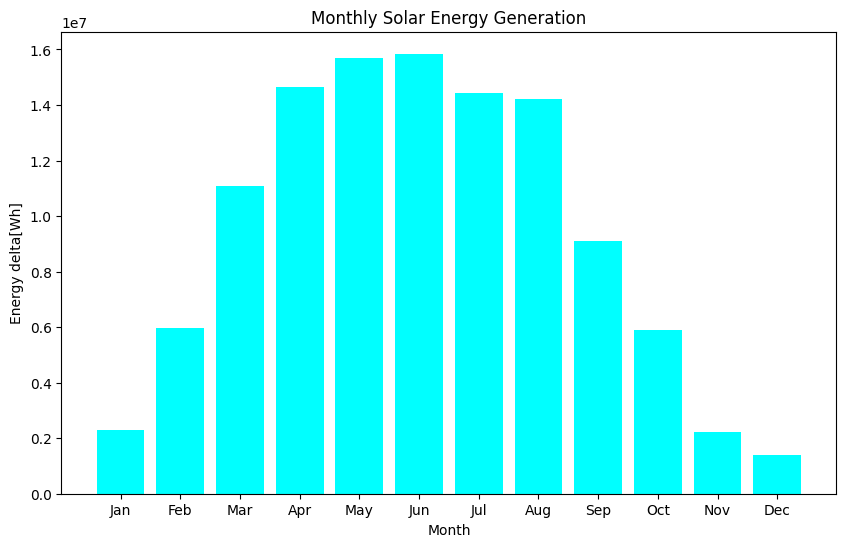


Figure 4 Monthly Solar Energy Generation

A seasonal analysis shown in Figure 5 reveals that Fall and Winter have lower energy generation due to reduced sunlight exposure on the solar panels. Conversely, energy generation in Spring is higher than in Summer, providing further evidence of the effect of temperature sensitivity on solar panel efficiency.

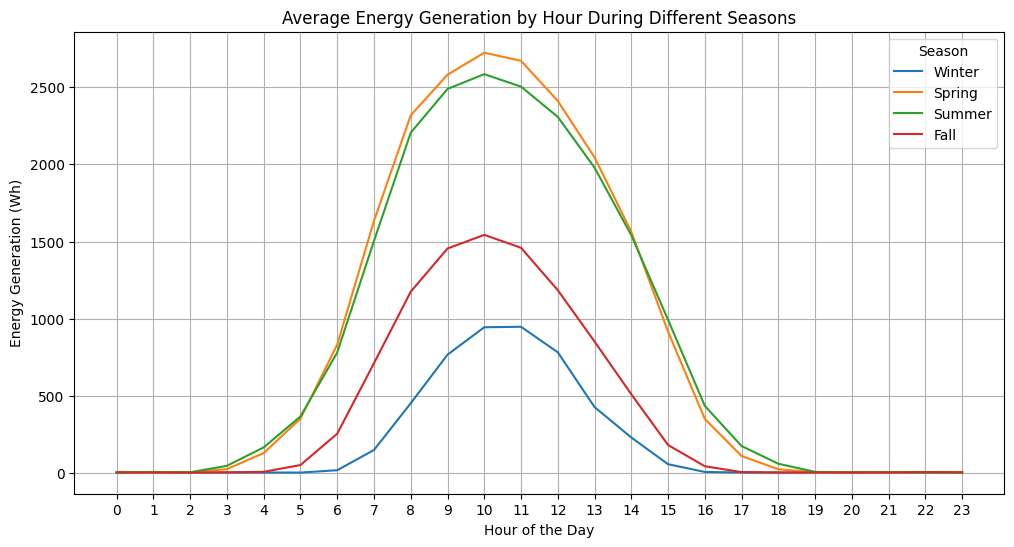


Figure 5 Hourly-average Seasonal Energy generation

For the purpose of sustainability analysis, the generation level during extreme weather conditions has been explored. Extreme weather has been defined as wind speeds over 10 m/s, rainfall greater than 5 mm in the last hour, and cloud coverage. Analyzing the Figure 6 shows that the hourly short-term impact of extreme weather conditions can significantly reduce energy generation, particularly during peak generation hours (9:00 AM to 11:30 AM), posing another challenge to the reliability of solar energy. On the other hand, Figure 7 illustrates the seasonal overall impact of such extreme weather conditions. While the energy generation reduction occurs similarly throughout the year, these extreme weather conditions affect energy generation more significantly during April and August, the boundary months of the higher generation season.

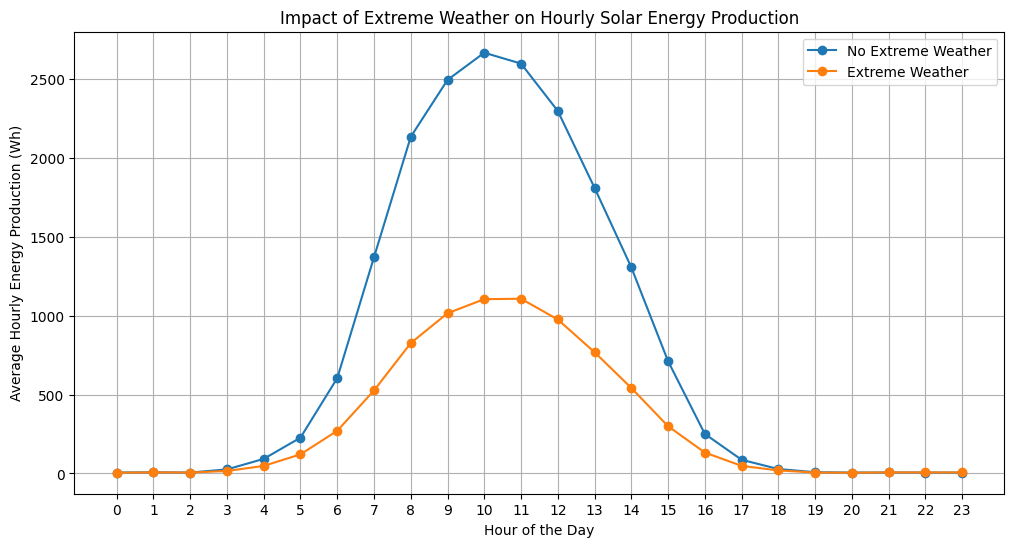


Figure 6 Hourly short-term impact of extreme weather conditions on energy generation

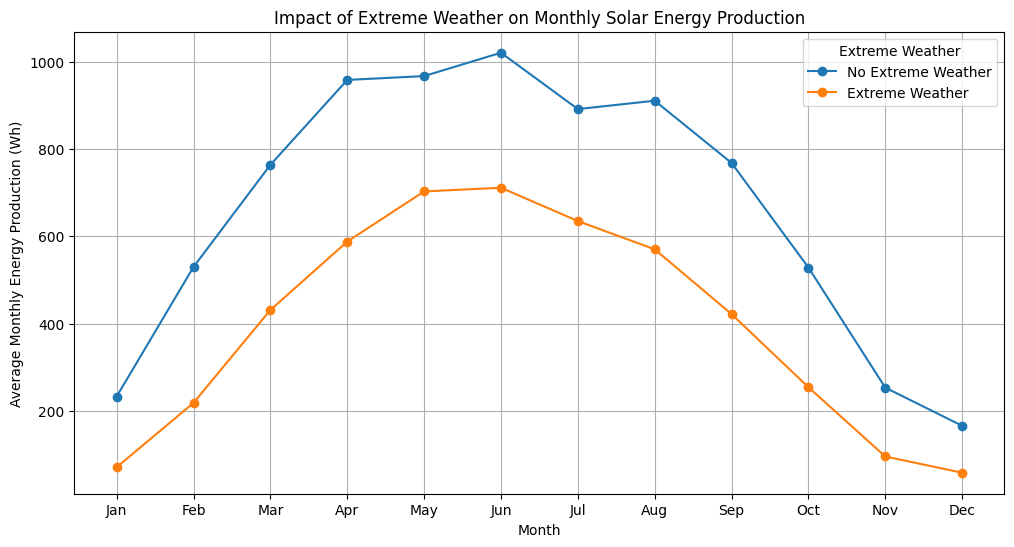


Figure 7 Seasonal impact of extreme weather conditions on energy generation

# **Conclusion and Results**

To predict energy generation and evaluate this prediction, multiple Machine Learning models, including linear regression, KNN, SVM, decision tree, and random forest, have been used. To evaluate the performance of these models, key indicators include Root Mean Squared Error, R² score, and Mean Squared Error (Sebastian Scher and Gabriele Messori , 2018) (Muyideen AbdulRaheem, Joseph Bamidele Awotunde, Abidemi Emmanuel Adeniyi, Idowu Dauda Oladipo, Sekinat Olaide Adekola, 2022) (Seul-Gi Kim, Jae-Yoon Jung and Min Kyu Sim, 2019).

The Linear Regression model, as shown in Figure 8, provides a medium fit to the dataset, with an RMSE of approximately 398.35. This indicates that the average deviation between the predicted and actual values is around 398.35 units. This significant amount of MSE suggests that there are considerable prediction errors, possibly due to the model's inability to handle nonlinear relationships in the data.

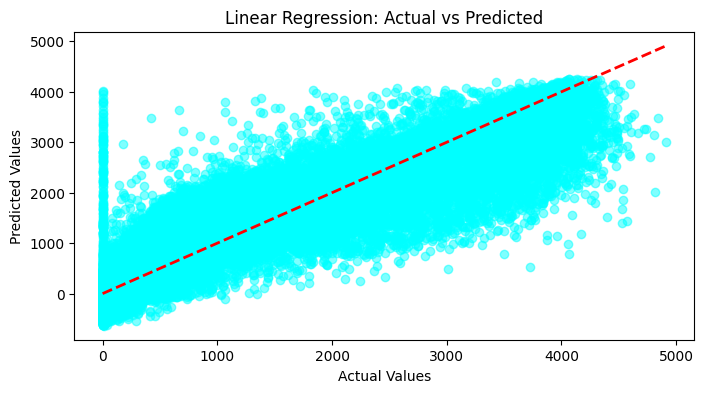


Figure 8 Linear Regression Model

The amount of MSE for the K-nearest neighbors model, shown in Figure 9, is lower compared to the linear regression model, indicating that the KNN model makes more accurate predictions on average. The R² score reveals that KNN explains 91.3% of the variance in the data, which is higher than linear regression’s 85.6%. Additionally, the RMSE of 308.55 shows that the average prediction error is smaller than that of linear regression (398.35), indicating more precise predictions. After tuning the hyperparameter for the number of neighbors in this model, the best results were achieved when 3 neighbors were considered for the prediction.

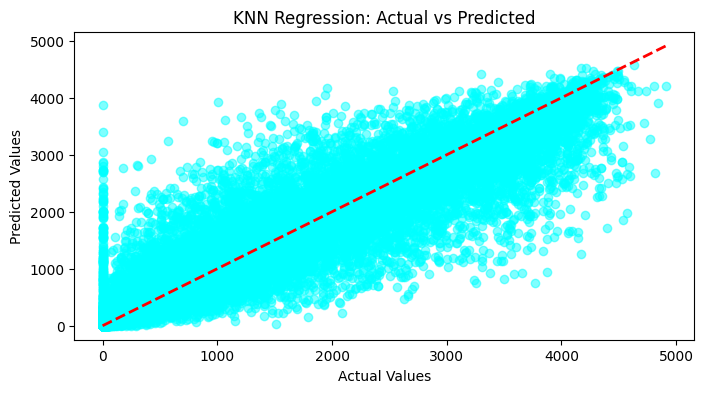


Figure 9 K-Nearest Neighbors Model

After examining the SVM model, the high MSE and RMSE values, along with the very low R² score, indicate that the SVM model predictions are not accurate and do not generalize well to this dataset. The Decision Tree model has shown a good level of accuracy and fit to the data. However, the RMSE indicates that there are still notable deviations between the predicted and actual values. Table 1 presents a sorted overview for the key indicators of the evaluation for each Machine Learning model.

Table 1 Key Evaluation Metrics for each Machin Learning model

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Mean Squared Error** | **R^2 score** | **Root Mean Squared Error** |
| Random Forest with 200 estimators in 20 depths | 71319.206 | 0.935 | 267.057 |
| Random Forest with 100 estimators in 20 depths | 71610.949 | 0.935 | 267.602 |
| Random Forest | 71625.136 | 0.935 | 267.629 |
| Random Forest with 200 estimators in 15 depths | 72206.448 | 0.934 | 268.713 |
| Random Forest with 50 estimators in 20 depths | 72353.201 | 0.934 | 268.986 |
| Random Forest with 100 estimators in 15 depths | 72544.377 | 0.934 | 269.341 |
| Random Forest with 50 estimators in 15 depths | 73026.714 | 0.934 | 270.235 |
| Random Forest with 200 estimators in 10 depths | 80149.079 | 0.927 | 283.106 |
| Random Forest with 100 estimators in 10 depths | 80282.35 | 0.927 | 283.341 |
| Random Forest with 50 estimators in 10 depths | 80332.302 | 0.927 | 283.43 |
| Decision Tree with 10 depths | 91149.726 | 0.917 | 301.91 |
| KNN with 3 neighbors | 93761.94 | 0.915 | 306.206 |
| KNN with 5 neighbors | 95201.399 | 0.913 | 308.547 |
| KNN with 7 neighbors | 98129.581 | 0.911 | 313.256 |
| KNN with 10 neighbors | 99259.579 | 0.91 | 315.055 |
| Decision Tree with 15 depths | 115573.563 | 0.895 | 339.961 |
| Decision Tree with 5 depths | 119647.231 | 0.891 | 345.901 |
| Decision Tree with 20 depths | 136067.938 | 0.876 | 368.874 |
| Decision Tree | 145957.799 | 0.867 | 382.044 |
| Linear Regression | 158683.759 | 0.856 | 398.351 |
| SVM Optimized | 327718.769 | 0.699 | 572.467 |
| SVM | 927143.511 | 0.149 | 962.883 |

# **Figures**

[Figure 1 Distribution visualization of Key Columns 3](#_Toc178267496)

[Figure 2 Energy generation level over time 4](#_Toc178267497)

[Figure 3 Impact of solar Radiation (GHI) on Energy Generation throughout the day 5](#_Toc178267498)

[Figure 4 Monthly Solar Energy Generation 6](#_Toc178267499)

[Figure 5 Hourly-average Seasonal Energy generation 6](#_Toc178267500)

[Figure 6 Hourly short-term impact of extreme weather conditions on energy generation 7](#_Toc178267501)

[Figure 7 Seasonal impact of extreme weather conditions on energy generation 7](#_Toc178267502)

[Figure 8 Linear Regression Model 8](#_Toc178267503)

[Figure 9 K-Nearest Neighbors Model 8](#_Toc178267504)

# **References**

Al-Ezzi, Athil S., and Mohamed Nainar M. Ansari, 2022. Photovoltaic solar cells: a review. *Applied System Innovation,* 67(5.4).

Bogdan Bochenek, Zbigniew Ustrnul, 2022. Machine Learning in Weather Prediction and Climate Analyses—Applications and Perspectives. *Atmosphere ,* 180(13(2)).

Ehsanul Kabira, Pawan Kumar, Sandeep Kumar, Adedeji A. Adelodun, Ki-Hyun Kim, 2017. Solar energy: Potential and future prospects. *Renewable and Sustainable Energy Reviews,* Volume 80, pp. 894-900.

El Naqa, I. and Murphy, M.J, 2015. What is machine learning?. *Springer International Publishing,* pp. 3-11.

Gordo, E., Khalaf, N., Strangeowl, T., Dolino, R. and Bennett, N, 2015. FACTORS AFFECTING SOLAR POWER PRODUCTION EFFICIENCY.

Hereher, Mohamed, and Ahmed M. El Kenawy., 2020. Exploring the potential of solar, tidal, and wind energy resources in Oman using an integrated climatic-socioeconomic approach.. *Renewable Energy,* Volume 161 , pp. 662-675.

Ishaq, H., and I. Dincer, 2021. Comparative assessment of renewable energy-based hydrogen production methods.. *Renewable and Sustainable Energy Reviews,* 135 (110192).

Li, G., M. Li, R. Taylor, Y. Hao, G. Besagni, and C. N. Markides., 2022. Solar energy utilisation: Current status and roll-out potential.. *Applied Thermal Engineering,* 209(118285).

Meral, M.E. and Dincer, F., 2011. A review of the factors affecting operation and efficiency of photovoltaic based electricity generation systems. *Renewable and Sustainable Energy Reviews,* 15(5), pp. 2176-2184.

Muyideen AbdulRaheem, Joseph Bamidele Awotunde, Abidemi Emmanuel Adeniyi, Idowu Dauda Oladipo, Sekinat Olaide Adekola, 2022. Weather prediction performance evaluation on selected machine learning algorithms. *IAES International Journal of Artificial Intelligence,* 11(4), pp. 1535-1544.

Sebastian Scher and Gabriele Messori , 2018. Predicting Weather Forecast Uncertainty with Machine Learning. *Quarterly Journal of the Royal Meteorological Society,* 144(717), pp. 2830-2841.

Seul-Gi Kim, Jae-Yoon Jung and Min Kyu Sim, 2019. A Two-Step Approach to Solar Power Generation Prediction Based on Weather Data Using Machine Learning. *sustainability,* pp. 11, 1501.

Shivashankar, S., Mekhilef, S., Mokhlis, H. and Karimi, M., 2016. Mitigating methods of power fluctuation of photovoltaic (PV) sources–A review. *Renewable and Sustainable Energy Reviews,* Volume 59, pp. pp.1170-1184.

Somasundaram, R.S. and Nedunchezhian, R., 2011. Evaluation of three simple imputation methods for enhancing preprocessing of data with missing values.. *International Journal of Computer Applications,* 21(10), pp. 14-19.

Sweeney, Conor, Bessa, Ricardo J., Browell, Jethro, Pinson, Pierre, 2019. The future of forecasting for renewable energy. *WIREs Energy Environ,* Issue 365.

Wang, Jianxiao, Haiwang Zhong, Xiaowen Lai, Qing Xia, Yang Wang, and Chongqing Kang., 2017. Exploring key weather factors from analytical modeling toward improved solar power forecasting. *IEEE Transactions on Smart Grid,* 10(2), pp. 1417-1427.

Zhai, Huixing, Suilin Wang, Baolin An, and Xiaoye Dai., 2020. Performance and parameter sensitivity comparison of CSP power cycles under wide solar energy temperature ranges and multiple working conditions. *Energy Conversion and Management ,* 218(112996).

1. https://www.kaggle.com/datasets/sheemazain/renewable-power-generation-weather-condition-2024/code [↑](#footnote-ref-1)