PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 21, No. 1 March 2025 | DOI: 10.33480/pilar.v21i1.6126

PREDICTING SOLAR POWER GENERATION: A MACHINE LEARNING APPROACH FOR GRID STABILITY AND EFFICIENCY

Popong Setiawati[1]; Adhitio Satyo Bayangkari Karno[2]; Widi Hastomo[3]*; Ellya Sestri[4]; Dian Kasoni[5]; Dodi Arif[6]; Fahrul Razi[7]

Informatics Engineering[1]
Esa Unggul University, Jakarta, Indonesia[1]
https://www.esaunggul.ac.id/[1]
setiawatipopong1967@gmail.com[1]

Information Systems[2, 6]
Gunadarma University, Depok, Indonesia[2, 6]
https://www.gunadarma.ac.id/[2, 6]
adh1t10.2@gmail.com[2], dodiarif8@gmail.com[6]

Information Technology[3, 4, 7]
ITB Ahmad Dahlan, Tangerang, Indonesia [3, 4, 7]
https://www.itb-ad.ac.id/[3, 4, 7]
Widie.has@gmail.com[3]*, ellyasestri24@gmail.com[4], fahrulrazi0398@gmail.com[7]

Information Systems[5]
STMIK Antar Bangsa, Tangerang City, Indonesia[5]
https://antarbangsa.ac.id/[5]
dhekalearning@gmail.com[5]
(*) Corresponding Author

Abstract— *In countries with high levels of insolation, the demand for renewable energy sources has driven the rapid emergence and growth of solar power plants. Maintaining grid stability and efficient power management in response to weather variations that affect solar radiation intensity and battery consumption limits remains a major challenge. This study aims to develop a machine learning-based prediction model to estimate the electricity generated by solar power plants using weather data. Four algorithms are utilized: Linear Regression, Random Forest Regressor, Decision Tree Regressor, and Gradient Boosting Regressor. The results show that the Random Forest algorithm produces the best model, with MAE and RMSE values of 0.1114281 and 0.3187232, respectively. This research contributes to the literature, particularly on the relatively unexplored topic of using multiple machine learning models to predict energy output from photovoltaic systems. The findings have the potential to inform more efficient energy policies and improve energy integration technologies for grid-connected solar power systems.*

Keywords: *energy forecasting, machine learning, renewable energy.*

Abstrak— *Di negara-negara dengan tingkat insolasi tinggi, permintaan akan sumber energi terbarukan telah menyebabkan kemunculan dan pertumbuhan pembangkit listrik tenaga surya yang pesat. Mempertahankan stabilitas jaringan dan efektivitas manajemen daya dalam menghadapi variasi cuaca yang mengubah intensitas radiasi matahari dan batasan konsumsi baterai merupakan tantangan utama. Tujuan dari penelitian ini adalah untuk membuat model prediksi berbasis pembelajaran mesin yang memperkirakan daya listrik yang dihasilkan dari pembangkit listrik tenaga surya menggunakan data cuaca. Penelitian ini menggunakan 4 algoritma yaitu linier regression, random forest regressor, decision tree regressor, gradient boosting regressor. Hasil penelitian menghasilkan model terbaik dari algoritma Random Forest dengan nilai MAE dan RMSE-nya masing-masing adalah 0.1114281 dan 0.3187232. Penelitian ini dapat menambah pengetahuan dalam bidang literatur, terutama berkaitan dengan topik yang belum banyak diteliti tentang penggunaan beberapa*

pembelajaran mesin untuk memprediksi keluaran energi dari sistem fotovoltaik surya. Hasil studi ini berpotensi memberikan kebijakan energi yang lebih efisien dan teknologi integrasi energi untuk sistem pembangkit listrik tenaga surya yang terintegrasi ke jaringan induk.

***Kata Kunci**: peramalan energi, pembelajaran mesin, energi terbarukan.*

INTRODUCTION

Renewable energy has become essential in reducing global carbon emissions and reliance on fossil fuels (Holechek, Geli, Sawalhah, & Valdez, 2022). Solar energy, particularly in high-insolation countries like India, is among the most promising renewable sources (Rathore & Panwar, 2022). India's investment in large-scale solar power plants has significantly boosted global clean energy capacity, with global solar capacity growing over 30% annually in recent years due to decreasing costs and improving photovoltaic (PV) module efficiency (Helveston, He, & Davidson, 2022).

Advances in PV technology, energy storage, and smart grids have helped address the challenge of solar energy's natural variability, which depends on daily solar radiation intensity changes (Tan et al., 2021). Despite these advances, weather-induced fluctuations in solar output create stability issues for power grids (Poddar, Kay, Prasad, Evans, & Bremner, 2023), demanding sophisticated prediction technologies to mitigate the impact of these fluctuations and enhance solar integration into the global grid system.

Current prediction models for solar output are often limited and lack adaptability to real-time weather changes, leading to inefficiencies in planning and grid stability (Al-Dahidi et al., 2024). Machine learning models, such as Random Forest and Gradient Boosting, offer more accurate solar output predictions based on historical weather and operational data, minimizing the risks associated with solar generation (Hanif et al., 2024). These models hold potential for improving energy planning at the grid level, ensuring a more reliable renewable energy supply (T. Ahmad, Madonski, Zhang, Huang, & Mujeeb, 2022).

Improving solar power forecasting is crucial for grid stability and efficient resource management (Bouquet, Jackson, Nick, & Kaboli, 2024). Accurate predictions enable grid operators to react proactively to energy supply changes, reducing disruptions and operational costs (Mirshekali, Santos, & Shaker, 2023). Enhanced forecasting through machine learning can thus transform renewable energy management (Aslam et al., 2021), especially for large-scale solar systems integrated within existing power networks, addressing the limitations of traditional statistical models and boosting overall system efficiency.

Previous research by (Oladapo, Olawumi, & Omigbodun, 2024) employed Long Short-Term Memory (LSTM), Random Forest, Support Vector Machines (SVM), and ARIMA to predict energy generation and demand patterns. The study by (Bashir et al., 2021) used SVM, K-Nearest Neighbor (KNN), Logistic Regression, Naive Bayes, Neural Networks, and Decision Tree classifiers. In the research conducted by (Chang, Bai, & Hsu, 2021), several regression techniques were compared for generating prediction models, including least squares and Support Vector Machines (SVM) using Multiple Short-Term Functions (MSTF).

Previous research by (Suanpang & Jamjuntr, 2024) employed the Light Gradient Boosting Machine (LGBM) and K-Nearest Neighbors (KNN). Research by (Abdelsattar, Ismeil, Azim Zayed, Abdelmoety, & Emad-Eldeen, 2024) utilized CatBoost, Gradient Boosting Machines (GBMs), Multilayer Perceptron (MLP) regressor, Support Vector Machine (SVM), XGBoost, and Random Forest (RF). The study by (Ibrar et al., 2022) used artificial neural network (ANN), Averaged Perceptron, Bayes Point Machine, Decision Forest, Decision Jungle, LightGBM, Locally Deep SVM, Logistic Regression (LR), SVM, and XGBoost methods.

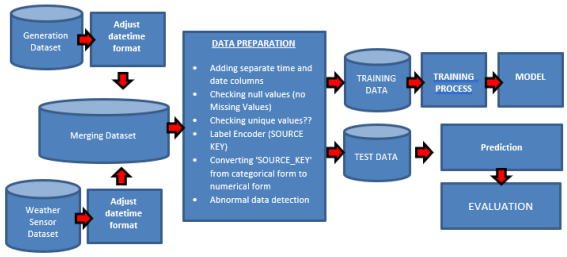
This research stands out by compare Random Forest (RF), Gradient Boosting Regressor (GBR), Decision Tree, and Linear Regression leveraging strength in handling complex data and ability to gradually enhance accuracy. These four machine learning methods were chosen because they are capable of handling continuous target data. This compare produces more stable and precise predictions, offering advantages in improving power grid efficiency and resilience.

While machine learning techniques are widely used in solar power generation prediction, few studies compare multiple techniques using the same data to find the most accurate approach. This study aims to develop a prediction model using four techniques: gradient boosting, decision tree, random forest, and linear regression.

MATERIALS AND METHODS

Starting with the combination of the original data, this study goes through numerous stages of work, including data preparation, data separation, training, model testing, and evaluation. As seen in Figure 1, each of these phases can be briefly explained as a flow chart. The next part provides an indirect description of each step.

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 21, No. 1 March 2025 | DOI: 10.33480/pilar.v21i1.6126



Source: (Research Result, 2024)

Figure 1. Research Flowchart

This section provides a detailed description of the data utilized in the study as well as several data analytics that can provide considerably more thorough insights and are highly beneficial for enhancing the electricity produced by solar power plants.

Two CSV format files Plant_Generation_Data.csv, which contains information about energy, and Plant_Weather_Sensor_Data.csv, which contains information about weather make up the dataset that was acquired from www.kaggle.com. Every 15 minutes for 34 days (2020-02-15 to 2020-06-17), data is recorded by 22 grids from both files. The attribute names for the Plant_Generation_Data.csv file, which has 67,698 rows of data overall, are DATE_TIME, PLANT_ID, DC_POWER, AC_POWER, DAILY_YIELD, and TOTAL_YIELD (Figure 2). IRRADIATION, AMBIENT_TEMPERATURE, MODULE_TEMPERATURE, SOURCE_KEY, PLANT_ID, and DATE_TIME are among the column properties included in the Plant_Weather_Sensor_Data.csv file, which has 3,259 rows of data overall (Figure 3).

	DATE_TIME	PLANT_ID	SOURCE_KEY	DC_POWER	AC_POWER	DAILY_YIELD	TOTAL_YIELD
0	2020-05-15 00:00:00	4136001	4UPUqMRk7TRMgml	0.0	0.0	9425.000000	2.429011e+06
1	2020-05-15 00:00:00	4136001	81aHJ1q11NBPMrL	0.0	0.0	0.000000	1.215279e+09
2	2020-05-15 00:00:00	4136001	9kRcWv60rDACzjR	0.0	0.0	3075.333333	2.247720e+09
...	...	...	...	...	...	...	...
67695	2020-06-17 23:45:00	4136001	vOuJvMaM2sgwLmb	0.0	0.0	4322.000000	2.427691e+06
67696	2020-06-17 23:45:00	4136001	xMblugepa2P7lBB	0.0	0.0	4218.000000	1.068964e+08
67697	2020-06-17 23:45:00	4136001	xoJJ8DcxJEcupym	0.0	0.0	4316.000000	2.093357e+08

67698 rows × 7 columns

Source: (Research Result, 2024)

Figure 2. Plant Generation Data

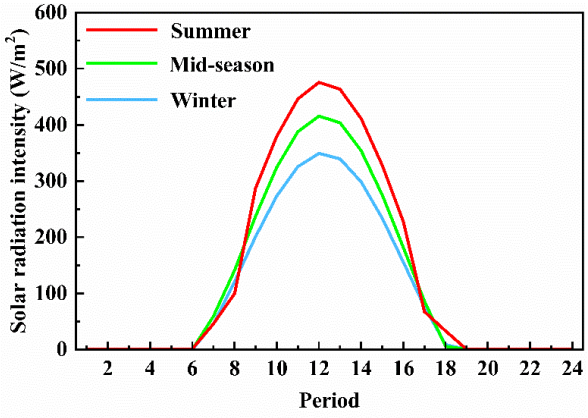
	DATE_TIME	PLANT_ID	SOURCE_KEY	AMBIENT_TEMPERATURE	MODULE_TEMPERATURE	IRRADIATION
0	2020-05-15 00:00:00	4136001	iq8k7ZNt4Mwm3w0	27.004764	25.060789	0.0
1	2020-05-15 00:15:00	4136001	iq8k7ZNt4Mwm3w0	26.880811	24.421869	0.0
2	2020-05-15 00:30:00	4136001	iq8k7ZNt4Mwm3w0	26.682055	24.427290	0.0
...	...	...	...	...	...	...
3256	2020-06-17 23:15:00	4136001	iq8k7ZNt4Mwm3w0	23.354743	22.492245	0.0
3257	2020-06-17 23:30:00	4136001	iq8k7ZNt4Mwm3w0	23.291048	22.373909	0.0
3258	2020-06-17 23:45:00	4136001	iq8k7ZNt4Mwm3w0	23.202871	22.535908	0.0

3259 rows × 6 columns

Source: (Research Result, 2024)

Figure 3. Plant Weather Sensor Data

Peak sun hours (PSH): The phrase "peak sun hours" (PSH) typically describes the amount of sunlight that occurs each day. The number of hours that 1 kW/m2 of energy would be required to generate the same amount of energy as the total for the day is known as the total PSH for the day. The terms "peak sunlight" and "peak sun hours" are interchangeable. Irradiance is the total amount of solar energy incident on a unit area over a given time period, like a day, month, or year. Insolation is another word for irradiance. The amount of sunlight that reaches a surface in a specific amount of time. Peak Sun Hour is the daily insolation measurement (kWh/m2/day). Irradiance: Solar radiation incident to a surface at a given time, in W/m2 (Figure 3).



Source: (Research Result, 2024)

Figure 4. Ideal Graph of Solar Power Generation

The shape in Figure 4 is based on the sun's angle to the panel. In the morning, when the sun is low, sunlight passes through more of the atmosphere, resulting in energy loss. As the sun rises higher during the day, less atmosphere is crossed, allowing the panel to capture more energy. In winter, although the sun is lower in the sky, brighter and sunnier days can still provide good energy, despite the risk of snow reflecting sunlight.

Model

Four machine learning models were selected for this study because the target dataset is continuous; these will be briefly covered in this section.

1. Linier Regression

A linear relationship between an independent variable (X) and a dependent variable (Y) is modeled using a machine learning method. Because of its ease of use and effectiveness, the linear regression technique is frequently employed for tasks involving regression analysis and prediction. The following is the basic formula for linear regression:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n + \in \quad \text{..... (1)}$$

Where:
Y is the dependent variable,
β_0 is the intercept,

$\beta_1, \beta_2, \ldots, \beta_n$ is the regression coefficient,
$X_1, X_2, \ldots, X_n$ is the independent variable, and
ϵ is the error or residual.

This technique offers a clear interpretation of each feature's contribution and works well with datasets that have a linear relationship between variables(G. James, Witten, Hastie, Tibshirani, & Taylor, 2023). Despite being straightforward, Linear Regression is frequently used as a reference model to evaluate how well more intricate algorithms work (Huang, Ko, Shu, & Hsu, 2020). A basic framework for assessing whether a straightforward linear connection is adequate for predicting how much energy a solar power plant would generate based on meteorological factors like temperature, wind speed, and solar radiation is provided by linear regression in the context of this study.

2. Decision Tree Regressor

The decision tree is a tree-based machine learning approach for classification and regression (Yulianto et al., 2023), which splits data into smaller groups based on information gain. It uses simple "if-then" criteria to handle complex data, such as location, temperature, and event time, for tasks like crime category prediction (AlKheder & AlOmair, 2022) . Information Gain, which is based on entropy, is the fundamental formula used in Decision Trees to choose the optimal attribute (Aning & Przybyła-Kasperek, 2022) . The following is the formula:

$$Information\ Gain = E(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} E(S_i) \quad \ldots\ldots (2)$$

$$E(S) = -\sum_{i=1}^{c} p_i \log_2 p_i \quad \ldots\ldots (3)$$

The entropy prior to separation is E(S), S_i is the subset of data following its separation according to specific attributes, n is the number of subsets, and S is the initial dataset's size, The dataset's class i probability is denoted by p_i. Although it has a tendency to overfit on complex data, the Decision Tree model is frequently used due to its computational speed and intuitive interpretation, and in this study, it is used to predict the electrical power that can be generated from solar heat because of its ability to handle non-linear and interpretive data (Costa & Pedreira, 2023).

3. Random Forest Regressor

The Random Forest Regressor, an ensemble machine learning algorithm, predicts solar power plant energy output using weather variables like temperature, wind speed, and solar radiation due to its effectiveness with non-linear and dynamic data (Bakır, Orak, & Yüksel, 2024), (N. Ahmad, Ghadi, Adnan, & Ali, 2022). Random Forest was chosen for its ability to handle high-dimensional and interacting features, its robustness against noisy data, and its useful feature importance metric for identifying key factors (Ghosh & Cabrera, 2022).

4. Gradient Boosting Regressor (GBR)

Gradient Boosting Regressor (GBR) is effective for regression problems, especially with complex and non-linear data, like solar power plant energy predictions (Gareth James, Witten, Hastie, Tibshirani, & Taylor, 2023). GBR works by sequentially improving predictions, where each model corrects the errors of the previous one, a process known as boosting (Safari, Kheirandish Gharehbagh, & Nazari Heris, 2023). Gradient Boosting was chosen for its ability to iteratively enhance prediction accuracy by optimizing the errors of previous models. This algorithm excels at detecting complex patterns in data and often delivers superior performance compared to traditional methods, particularly with imbalanced datasets (Bentéjac, Csörgő, & Martínez-Muñoz, 2021).

Performance Metric

1. MAE

Mean Absolute Error (MAE) is a widely used evaluation metric for measuring the effectiveness of regression models, including forecasting energy output from solar power plants. (Karunasingha, 2022). The difference between the expected and actual outcomes of a model is measured by the mean absolute error, or MAE. The MAE is calculated mathematically using the following formula.

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \quad \ldots\ldots (4)$$

The actual value, such as the power generated by the solar panel, is denoted by y_i. The model's predicted value is $\hat{y}_i$. The total number of predictions is denoted by n. This measure is easy to use because it provides a clear indication of the inaccuracy in values predicted from the predictive data. Better model performance is indicated by a lower MAE number, which shows a smaller average difference between the expected and actual values.
In this study, the MAE metric was chosen to evaluate the prediction algorithms that predict the energy output of the solar power plants. For example, MAE calculates the difference between the expected and actual energy values when a model, like a Random Forest and Gradient Boosting, is used to analyze weather and energy output data.

2. RMSE

The effectiveness of a regression model in predicting the energy output of a solar power plant is measured by Root Mean Squared Error (RMSE)

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 21, No. 1 March 2025 | DOI: 10.33480/pilar.v21i1.6126

(Hastomo, Bayangkari Karno, Kalbuana, Meiriki, & Sutarno, 2021; Karno et al., 2023), which indicates the average prediction error in the same units as the target variable (Karunasingha, 2022). Its computation is based on the following equation:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad \text{..... (5)}$$

The actual value, such as the power generated by the solar panel, is denoted by y_i. The model's predicted value is $\hat{y}_i$. The total number of predictions is denoted by n. The square root of the error is the mean square error (RMSE) between the expected and actual output values. This suggests that outliers affect RMSE because extreme errors are penalized more severely than moderate ones.

RESULTS AND DISCUSSION

Dataset preparation

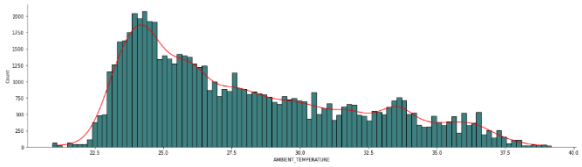
Before the data is used in the machine learning process, a number of steps are taken, specifically:

1. "%Y-%m-%d %H: %M" is the standard data format for the "DATE_TIME" feature of the "Plant_Generation_Data.csv" and "Plant_Weather_Sensor_Data.csv" files.
2. 'PLANT_ID' in the generation file and 'SOURCE_KEY' and 'PLANT_ID' in the weather file are examples of superfluous data attributes that have been removed.
3. By binding the "DATE_TIME" attribute from both Generation_Data and Sensor_Data files into a single new file, the files can be merged.
4. Dividing time information into new columns called "DATE," "TIME," "DAY," "MONTH," "WEEK," "HOURS," and "MINUTES."
5. The file has 18 columns and 67698 rows, as a result of data checking that it does not contain null data (Figure 5).
6. Changing the "SOURCE_KEY" attribute's data type from category to numeric.
7. To observe the degree of fluctuation, a plot was created displaying ambient temperature data over a 34-day period (Figure 6).

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 67698 entries, 0 to 67697
Data columns (total 18 columns):
 #   Column               Non-Null Count  Dtype         #   Column               Non-Null Count  Dtype
---  ------               --------------  -----         ---  ------               --------------  -----
 0   DATE_TIME            67698 non-null  object        10  TIME                 67698 non-null  object
 1   SOURCE_KEY           67698 non-null  object        11  DAY                  67698 non-null  int32
 2   DC_POWER             67698 non-null  float64       12  MONTH                67698 non-null  int32
 3   AC_POWER             67698 non-null  float64       13  WEEK                 67698 non-null  UInt32
 4   DAILY_YIELD          67698 non-null  float64       14  HOURS                67698 non-null  object
 5   TOTAL_YIELD          67698 non-null  float64       15  MINUTES              67698 non-null  int32
 6   AMBIENT_TEMPERATURE  67698 non-null  float64       16  TOTAL MINUTES PASS   67698 non-null  int32
 7   MODULE_TEMPERATURE   67698 non-null  float64       17  DATE_STRING          67698 non-null  object
 8   IRRADIATION          67698 non-null  float64     dtypes: UInt32(1), float64(7), int32(4), object(6)
 9   DATE                 67698 non-null  object      memory usage: 8.1+ MB
```

Source: (Research Result, 2024)

Figure 5. Null Data Checking

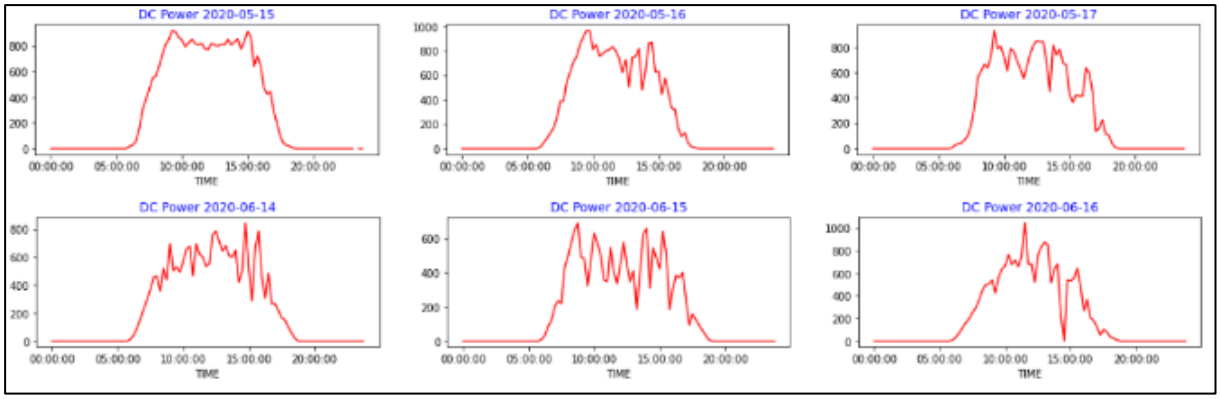


Source: (Research Result, 2024)

Figure 6: Count Data Visualization Using AMBIENT_TEMPERATURE

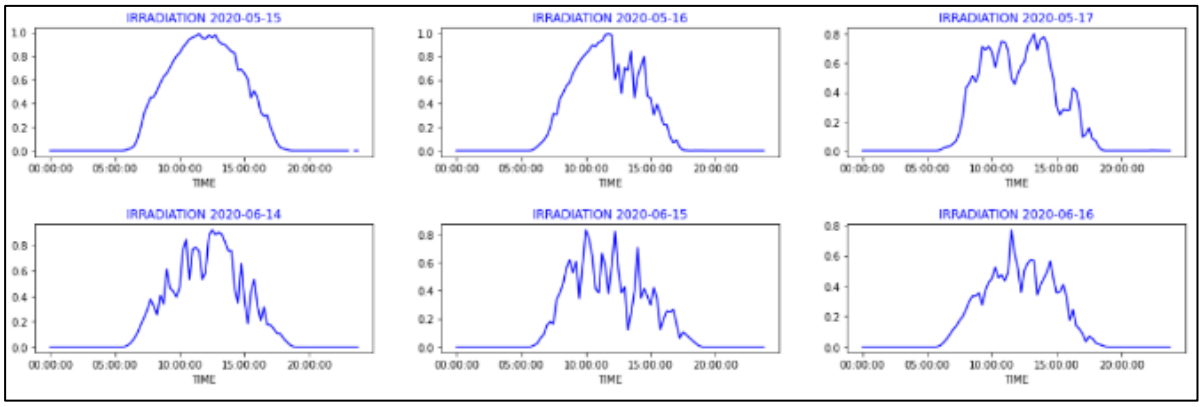
Solar Power Plant Disturbance and Abnormality Detection

DC_POWER generation is abnormal, as shown by the daily DC_POWER generation g raph (Figure 7), which displays daily power generation variations. Figure 7 is simply a part of the full image that should be displayed in order to conserve pages. There is less fluctuation in DC_POWER generation on the days listed below: May 15, May 18, May 22, May 23, May 24, May 25, May 26, and May 26, 2020. The following days saw a significant variation in DC_POWER generation: 2020-05-19, 2020-05-28, 2020-05-29, 2020-06-02, 2020-06-03, 2020-06-04, 2020-06-13, 2020-06-14, and 2020-06-17.



Source: (Research Result, 2024)

Figure 7. Daily DC Power Generation Graph



Source: (Research Result, 2024)

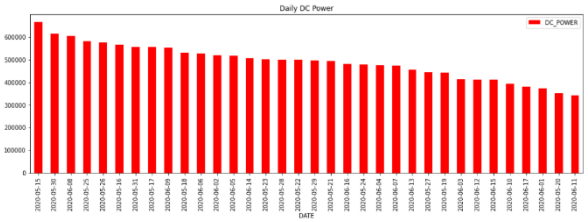
Figure 8: Graph of Daily IRRADIATION

DC power generation experienced significant fluctuations and decreases on 2020-06-03, 2020-06-11, 2020-06-12, and 2020-06-15. The incredibly large fluctuations and decreases in DC_POWER generation could be due to a system failure, weather variations, or cloud cover.

The DC_POWER generation per day bar chart (Figure 9) shows the average power generation per day;

1. The highest average DC_POWER production occurred on May 15, 2020.
2. The peak in average DC power generation was recorded on 2020-06-11.

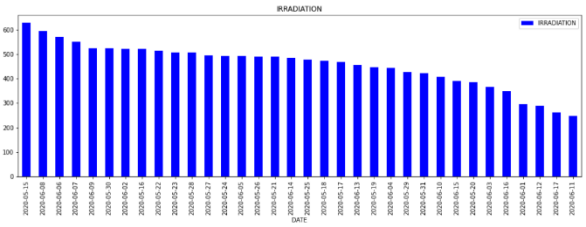
This large variation in DC_POWER generation is caused by weather-related changes or system malfunctions. However, this bar chart (Figure 9) allows us to identify the day that generated the highest and lowest DC_POWER.



Source: (Research Result, 2024)

Figure 9. Bar Chart of Daily DC_POWER Generation

The daily DC_POWER generation and the IRRADIATION graph pattern appear to be quite similar. IRRADIATION plays a major role in DC_POWER, or output power, in solar power plants. Alternatively, it can be said to be directly proportional (Figure 8). Figure 8 is simply a part of the full image that should be displayed in order to conserve pages. As with DC_POWER generation, 2020-05-15 and 2020-06-11 saw the highest and lowest average IRRADIATION generation, respectively (Figure 10).



Source: (Research Result, 2024)

Figure 10. Bar Chart of Daily IRRADIATION

An Analysis of the Top and Poorest Solar Power Facilities

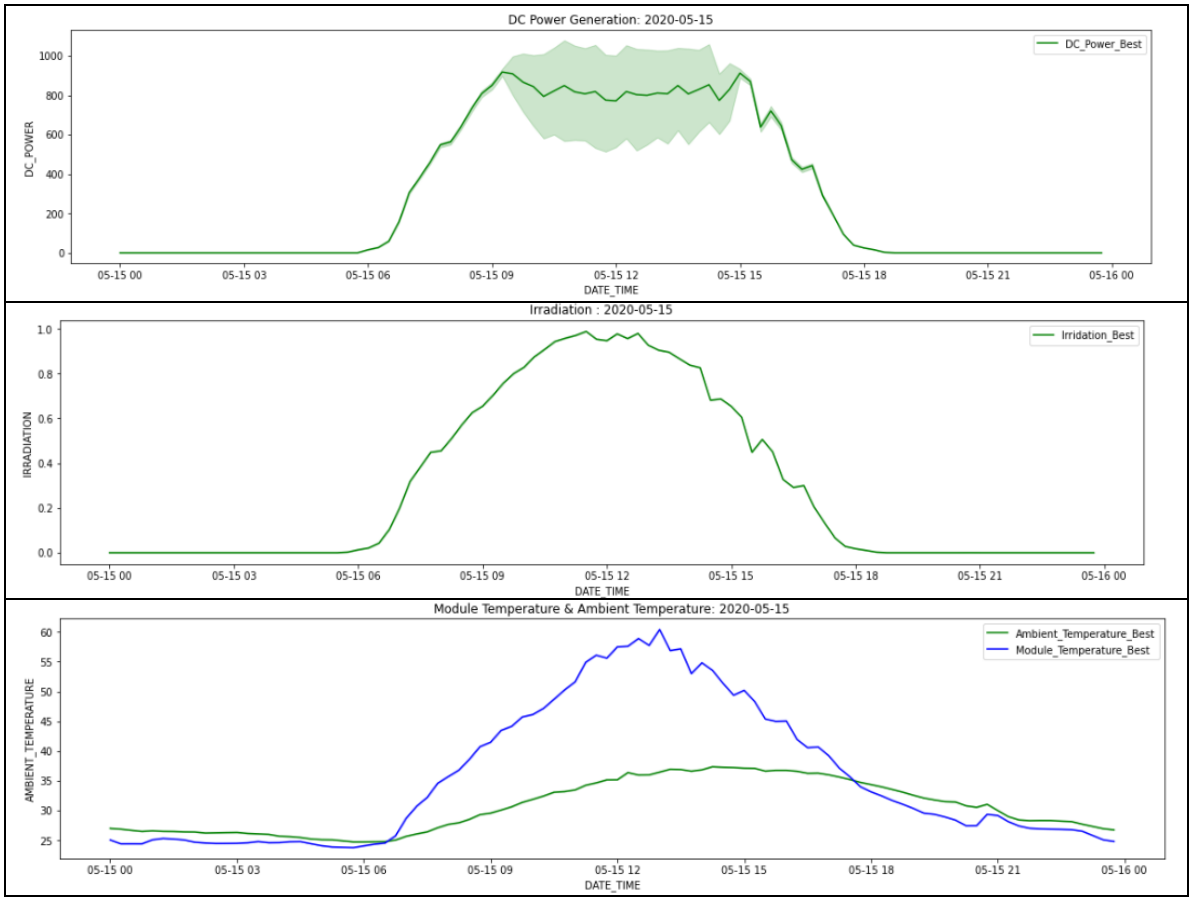
The following are the main environmental factors that affect solar power production. Cloud cover, especially thicker layers in winter, reduces sunlight and lowers solar panel output. In addition to the sun's position, factors like panel temperature also affect performance. The production of solar energy relies directly on the intensity of solar radiation. Both the DC_POWER and IRRADIATION graphs are similar to the previously mentioned ideal graphs. There are no clouds in the sky, and the weather looks to be great due to the small variations in IRRADIATION, solar panel temperature, and ambient temperature (Figure 11).

Estimating Solar Power Plant Inverter Efficiency

Inverter efficiency is the ratio of AC output power (pac) to DC input power (pdc), accounting for heat loss during conversion. Standby power, also known as idle power consumption, is used to maintain the inverter's power mode (Figure 12).

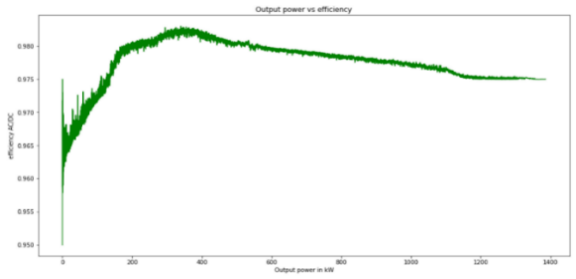
The efficiency of low-quality modified sine wave inverters ranges from 75% to 85%, while high-quality pure sine wave inverters have an efficiency of 90% to 95%. Efficiency improves with higher load power capacity, reaching its peak before exceeding the inverter's output capacity. Below 15% load, efficiency is typically low. Proper matching of the inverter's capacity with the load enhances efficiency, resulting in higher AC output for the same DC input power (Figure 13).

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 21, No. 1 March 2025 | DOI: 10.33480/pilar.v21i1.6126



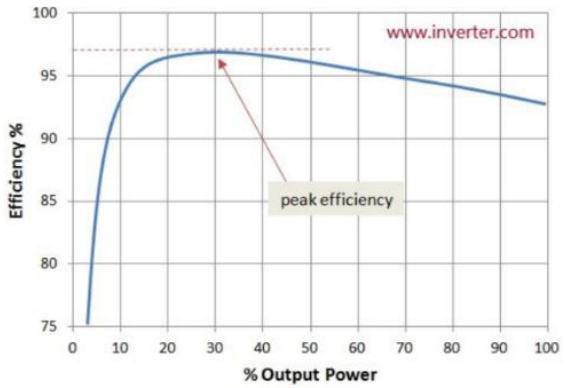
Source: (Research Result, 2024)

Figure 11: DC_POWER, IRRADIATION, and Temperature Graph Comparison



Source: (Research Result, 2024)

Figure 12. This Study's AC/DC Efficiency Graph

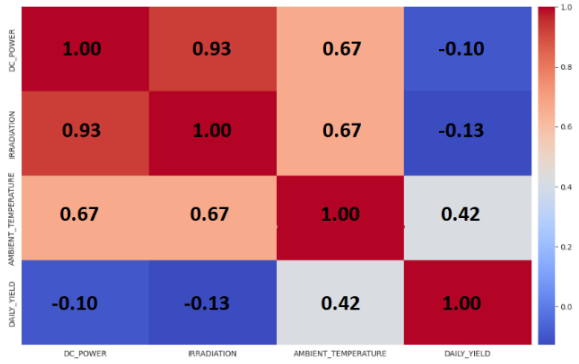


Source: (Research Result, 2024)

Figure 13. Graph of AC/DC Efficiency

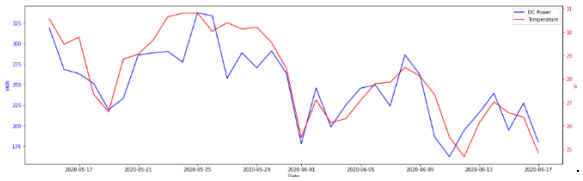
Correlation

The correlation map (Figure 14) shows a strong relationship between DC_POWER and IRRADIATION of 0.93. To further confirm this connection, we multiply IRRADIATION by 1000 to get the same scale. When the DC_POWER and IRRADIATION graphs in Figure 15 are combined, the patterns in the fluctuations of the two graphs generated by 22 power grids are identical. The predictive potential of any model will be fairly high due to the strong correlation between DC output power and IRRADIATION.



Source: (Research Result, 2024)

Figure 14. Map of Correlation



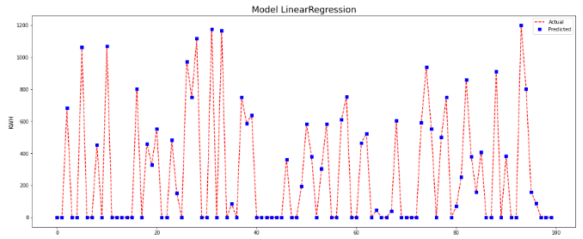
Source: (Research Result, 2024)

Figure 15. Merging the DC_POWER and IRRADIATION

Forecasting Solar Power

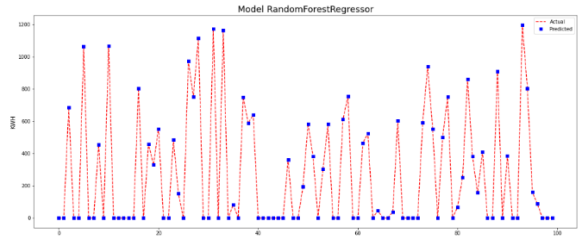
Four models created during the training process were tested using the test dataset. Actual facts and predicted results are shown in graphs (Figures 16, 17, 18, and 19). To save pages, the four graphs only show 100 of the 13.540 test data points. To assess these four models, the prediction error is estimated using MAE and RMSE (Table 1). With the least MAE and RMSE values of 0.1114281 and 0.3187232, respectively, Random Forest is the best model.

The Random Forest model in this study improves power grid efficiency by enhancing energy consumption forecasts, supporting better distribution, optimizing renewable energy use, reducing costs, and improving reliability. It can be integrated into smart grids for faster decision-making, early warnings during demand surges, and better green energy integration for a sustainable system.



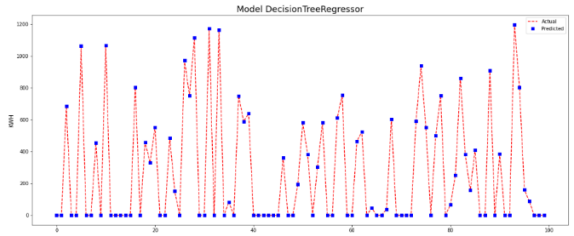
Source: (Research Result, 2024)

Figure 16. Linier Regression Prediction Results Plot



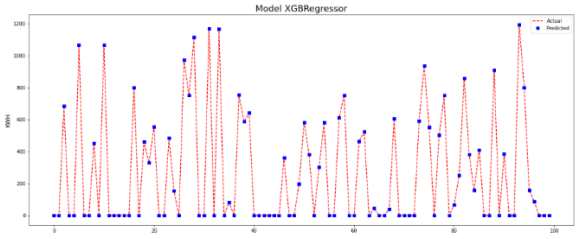
Source: (Research Result, 2024)

Figure 17. Random Forest Prediction Results Plot



Source: (Research Result, 2024)

Figure 18. Decision Tree prediction results Plot



Source: (Research Result, 2024)

Figure 19. Gradient Boosting prediction results Plot

Table 1. The Outcomes f Four Machine Learning Algorithms' Error Measurements

	MAE (kW)	RMSE (kW)
Linier Regression	0.5966207	0.8637416
Random Forest	0.1114281	0.3187232
Decision Tree	0.1380701	0.4269832
Gradient Boosting	0.6495836	1.4015379

Source: (Research Result, 2024)

Random Forest demonstrates the best performance (MAE 0.1114 kW, RMSE 0.3187 kW), significantly more accurate than Linear Regression and Gradient Boosting, reflecting a substantial reduction in error. The small gap between MAE and RMSE in Random Forest indicates a stable and consistent model, minimizing the risk of overfitting or underfitting. Random Forest is more effective than Gradient Boosting (RMSE 1.4015 kW), highlighting a more optimal selection and refinement of the ensemble model. Hyperparameter optimization and improved data preprocessing enhance the model's performance. Adaptation to the specific characteristics of solar data also supports more accurate predictions.

CONCLUSION

Four models were created by training 54,158 training data using four machine learning algorithms. 13,540 test datasets were then used to evaluate these four models. MAE and RMSE were used to measure the model testing outcomes; the model with the lowest error value is the best model. Random Forest emerged as the strongest model from this investigation, with RMSE and MAE values of 0.3187232 and 0.1114281, respectively. A

machine learning system called Random Forest may be used to forecast power output with greater accuracy, which is crucial for enhancing the electrical grid's stability and efficiency.

This study focuses on testing the Random Forest model with limited data and comparing only four algorithms. Future research should use more diverse data, explore algorithms like Gradient Boosting or Deep Learning, and develop a real-time prediction system to improve electricity grid management.

REFERENCE

Abdelsattar, M., Ismeil, M. A., Azim Zayed, M. M. A., Abdelmoety, A., & Emad-Eldeen, A. (2024). Assessing Machine Learning Approaches for Photovoltaic Energy Prediction in Sustainable Energy Systems. *IEEE Access*, *12*(August), 107599–107615. https://doi.org/10.1109/ACCESS.2024.3437191

Ahmad, N., Ghadi, Y., Adnan, M., & Ali, M. (2022). Load Forecasting Techniques for Power System: Research Challenges and Survey. *IEEE Access*, *10*, 71054–71090. https://doi.org/10.1109/ACCESS.2022.3187839

Ahmad, T., Madonski, R., Zhang, D., Huang, C., & Mujeeb, A. (2022). Data-driven probabilistic machine learning in sustainable smart energy/smart energy systems: Key developments, challenges, and future research opportunities in the context of smart grid paradigm. *Renewable and Sustainable Energy Reviews*, *160*, 112128. https://doi.org/10.1016/j.rser.2022.112128

Al-Dahidi, S., Madhiarasan, M., Al-Ghussain, L., Abubaker, A. M., Ahmad, A. D., Alrbai, M., … Zio, E. (2024). Forecasting Solar Photovoltaic Power Production: A Comprehensive Review and Innovative Data-Driven Modeling Framework. *Energies*, Vol. 17. https://doi.org/10.3390/en17164145

AlKheder, S., & AlOmair, A. (2022). Urban traffic prediction using metrological data with fuzzy logic, long short-term memory (LSTM), and decision trees (DTs). *Natural Hazards*, *111*(2), 1685–1719. https://doi.org/10.1007/s11069-021-05112-x

Aning, S., & Przybyła-Kasperek, M. (2022). Comparative Study of Twoing and Entropy Criterion for Decision Tree Classification of Dispersed Data. *Procedia Computer Science*, *207*, 2434–2443. https://doi.org/10.1016/j.procs.2022.09.301

Aslam, S., Herodotou, H., Mohsin, S. M., Javaid, N., Ashraf, N., & Aslam, S. (2021). A survey on deep learning methods for power load and renewable energy forecasting in smart microgrids. *Renewable and Sustainable Energy Reviews*, *144*, 110992. https://doi.org/10.1016/j.rser.2021.110992

Bakır, R., Orak, C., & Yüksel, A. (2024). Optimizing hydrogen evolution prediction: A unified approach using random forests, lightGBM, and Bagging Regressor ensemble model. *International Journal of Hydrogen Energy*, *67*, 101–110. https://doi.org/10.1016/j.ijhydene.2024.04.173

Bashir, A. K., Khan, S., Prabadevi, B., Deepa, N., Alnumay, W. S., Gadekallu, T. R., & Maddikunta, P. K. R. (2021). Comparative Analysis of Machine Learning Algorithms for prediction of Smart Grid Stability. *International Transactions on Electrical Energy Systems*, *31*(9). https://doi.org/10.1002/2050-7038.12706

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, *54*(3), 1937–1967. https://doi.org/10.1007/s10462-020-09896-5

Bouquet, P., Jackson, I., Nick, M., & Kaboli, A. (2024). AI-based forecasting for optimised solar energy management and smart grid efficiency. *International Journal of Production Research*, *62*(13), 4623–4644. https://doi.org/10.1080/00207543.2023.2269565

Chang, R., Bai, L., & Hsu, C.-H. (2021). Solar power generation prediction based on deep Learning. *Sustainable Energy Technologies and Assessments*, *47*, 101354. https://doi.org/10.1016/j.seta.2021.101354

Costa, V. G., & Pedreira, C. E. (2023). Recent advances in decision trees: an updated survey. *Artificial Intelligence Review*, *56*(5), 4765–4800. https://doi.org/10.1007/s10462-022-10275-5

Ghosh, D., & Cabrera, J. (2022). Enriched Random Forest for High Dimensional Genomic Data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *19*(5), 2817–2828. https://doi.org/10.1109/TCBB.2021.3089417

Hanif, M. F., Siddique, M. U., Si, J., Naveed, M. S., Liu, X., & Mi, J. (2024). Enhancing Solar Forecasting Accuracy with Sequential Deep Artificial Neural Network and Hybrid Random Forest and Gradient Boosting Models across Varied Terrains. *Advanced Theory and Simulations*, *7*(7), 2301289.

https://doi.org/10.1002/adts.202301289

Hastomo, W., Bayangkari Karno, A. S., Kalbuana, N., Meiriki, A., & Sutarno. (2021). Characteristic Parameters of Epoch Deep Learning to Predict Covid-19 Data in Indonesia. *Journal of Physics: Conference Series*, *1933*(1). https://doi.org/10.1088/1742-6596/1933/1/012050

Helveston, J. P., He, G., & Davidson, M. R. (2022). Quantifying the cost savings of global solar photovoltaic supply chains. *Nature*, *612*(7938), 83–87. https://doi.org/10.1038/s41586-022-05316-6

Holechek, J. L., Geli, H. M. E., Sawalhah, M. N., & Valdez, R. (2022). A Global Assessment: Can Renewable Energy Replace Fossil Fuels by 2050? *Sustainability*, 14(8), 4792. https://doi.org/10.3390/su14084792

Huang, J.-C., Ko, K.-M., Shu, M.-H., & Hsu, B.-M. (2020). Application and comparison of several machine learning algorithms and their integration models in regression problems. *Neural Computing and Applications*, *32*(10), 5461–5469. https://doi.org/10.1007/s00521-019-04644-5

Ibrar, M., Hassan, M. A., Shaukat, K., Alam, T. M., Khurshid, K. S., Hameed, I. A., … Luo, S. (2022). A Machine Learning-Based Model for Stability Prediction of Decentralized Power Grid Linked with Renewable Energy Resources. *Wireless Communications and Mobile Computing*, *2022*. https://doi.org/10.1155/2022/2697303

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An introduction to statistical learning: With applications in python* (1st ed.). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-38747-0

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). Tree-Based Methods. *Springer Nature: An Introduction to Statistical Learning*, 331–366. https://doi.org/10.1007/978-3-031-38747-0_8

Karno, A. S. B., Hastomo, W., Surawan, T., Lamandasa, S. R., Usuli, S., Kapuy, H. R., & Digdoyo, A. (2023). Classification of cervical spine fractures using 8 variants EfficientNet with transfer learning. *International Journal of Electrical and Computer Engineering*, *13*(6), 7065–7077. https://doi.org/10.11591/ijece.v13i6.pp7065-7077

Karunasingha, D. S. K. (2022). Root mean square error or mean absolute error? Use their ratio as well. *Information Sciences*, *585*, 609–629. https://doi.org/10.1016/j.ins.2021.11.036

Mirshekali, H., Santos, A. Q., & Shaker, H. R. (2023). A Survey of Time-Series Prediction for Digitally Enabled Maintenance of Electrical Grids. *Energies*, Vol. 16. https://doi.org/10.3390/en16176332

Oladapo, B. I., Olawumi, M. A., & Omigbodun, F. T. (2024). Machine Learning for Optimising Renewable Energy and Grid Efficiency. *Atmosphere*, Vol. 15. https://doi.org/10.3390/atmos15101250

Poddar, S., Kay, M., Prasad, A., Evans, J. P., & Bremner, S. (2023). Changes in solar resource intermittency and reliability under Australia's future warmer climate. *Solar Energy*, *266*, 112039. https://doi.org/10.1016/j.solener.2023.112039

Rathore, N., & Panwar, N. L. (2022). Outline of solar energy in India: advancements, policies, barriers, socio-economic aspects and impacts of COVID on solar industries. *International Journal of Ambient Energy*, *43*(1), 7630–7642. https://doi.org/10.1080/01430750.2022.2075925

Safari, A., Kheirandish Gharehbagh, H., & Nazari Heris, M. (2023). DeepVELOX: INVELOX Wind Turbine Intelligent Power Forecasting Using Hybrid GWO–GBR Algorithm. *Energies*, 16(19), 6889. https://doi.org/10.3390/en16196889

Suanpang, P., & Jamjuntr, P. (2024). Machine Learning Models for Solar Power Generation Forecasting in Microgrid Application Implications for Smart Cities. *Sustainability*, 16(14), 6087. https://doi.org/10.3390/su16146087

Tan, K. M., Babu, T. S., Ramachandaramurthy, V. K., Kasinathan, P., Solanki, S. G., & Raveendran, S. K. (2021). Empowering smart grid: A comprehensive review of energy storage technology and application with renewable energy integration. *Journal of Energy Storage*, 39, 102591. https://doi.org/10.1016/j.est.2021.102591

Yulianto, R., Rusli, M. S., Satyo, A., Karno, B., Hastomo, W., & Kardian, A. R. (2023). Innovative UNET-Based Steel Defect Detection Using 5 Pretrained Models. *Joint Journal of Novel Carbon Resource Sciences & Green Asia Strategy*. *10*(4), 2365–2378.