SENSITIVITY ANALYSIS OF HYPERPARAMETER IN SOLAR ENERGY PREDICTION MODEL USING GRADIENT BOOSTING METHOD

Aska Ramadhan[1], Bertha Maya Sopha[2], Mohammad Kholid Ridwan[3]

[1]Magister Program of System Engineering, Faculty of Engineering, Gadjah Mada University
[2]Department of Mechanical and Industrial Engineering, Faculty of Engineering, Gadjah Mada University
[3]Department of Nuclear Engineering and Engineering Physics, Faculty of Engineering, Gadjah Mada University
*Correspondence : aska.adhan@mail.ugm.ac.id

Abstract

Solar energy prediction is one alternative to handling unpredicted conditions of weather and solar radiation intensity. It could be the most important factor in achieving stability in electricity generation using solar energy resources. In making predictions, the use of machine learning models has been carried out by various methods, and in this study, the method used for the algorithm model is gradient boosting. In the modeling process using gradient boosting, several hyperparameter settings are needed. Hyperparameters have an important role in producing stable predictive patterns and can avoid overfitting or underfitting conditions. In this study, the accuracy and speed of prediction of the machine learning model with the gradient boosting approach, namely XGBoost and LightGBM, were analyzed in relation to setting the hyperparameter learning rate and max depth of the model's prediction pattern. The dataset used spans 6 months at a data resolution rate of every 5 minutes and includes meteorological data at the location point of Energy Laboratory UKRIM Yogyakarta as well as the output value of PLTS power and temperature panels onsite. Setting the hyperparameter learning rate in the highest and lowest conditions generates accuracy values with a difference of 2% and about the same prediction speed. With nMAE values of 2.84% and 1.35% and nRMSE values of 6.11% and 3.68%, respectively, the higher learning rate results in lower error values for both models. The XGBoost model shown tendency for overfitting and slower prediction speeds with the highest max depth setting. The prediction speed is faster at the lowest max depth condition, but the XGBoost and LightGBM models both exhibit underfitting.

History:

Received:
Accepted:
First published online:

Keywords:

Machine learning
Hyperparameter
Solar energy prediction
Gradient boosting

1. Introduction

The implementation of machine learning applications to proceed prediction systems of solar radiation into electrical energy generated from variables, most of which come from complex meteorological data, is one of the case studies conducted to improve reliability of the PV on-grid network. In order to make data predictions for the future, machine learning might examine the connections between various large datasets and identify patterns in the data collection (Lai et al., 2020). This study develops sensitivity analysis of the hyperparameter gradient boosting method for forecasting predictions of solar energy production during the annual period using the machine learning application.

Several studies that have been conducted include the prediction of global diesel radiation in Malaysia using a regression model (Ahmed Kutty et al., 2015); forecasting of PV energy production using the LightGBM model at the Faculty of Engineering, University of Ljubljana, Slovenia (Reba et al., 2019); forecasting solar energy From PLTS in South Korea, PV output data throughout the past is used with the Ensemble Learning Model which are Random Forest, Xgboost, and LightGBM (Choi & Hur, 2020). Further research using a different algorithm, namely the Naive Bayes Classifier and K-NN, on the PV Energy Prediction of the Faculty of Industrial Technology UII in Yogyakarta (Ikhsan, 2020).

Algorithm models in this study apply the gradient boosting techniques which are XGBoost and LightGBM. Predictions on a set of datasets are currently frequently made using both models. Zhou et al. (2022), who utilized different ensemble and gradient boosting models to predict solar energy sources in China, concluded that the XGBoost and LightGBM models obtained great results. They also noted that being able to manage to overfit and the rate of repetition became important considerations. Compared to conventional models like Random Forest, it has a better level of accuracy and is more successful at preventing overfitting (Zhou et al., 2022). Therefore, it will concentrate more on the application of XGBoost and LightGBM models in this research.

Extreme-gradient boosting, the foundation of gradient boosting modeling, served as the concept for the model method known as XGBoost. The basic gradient boosting and regularization principles are applied in this algorithm model, which is intended to be used with machine learning algorithms for data that is organized into broad categories. Regression and data classification machine learning methods can be modeled using XGBoost (XGBoost developers, 2022). It is widely recognized that XGBoost can be used in machine learning applications. This is a result of its capability to adjust the algorithm's performance to various scenarios and adjustments in the number of data sets (Chen & Guestrin, 2016).

LightGBM, the algorithm model, which is an upgrade of the XGBoost model, has ability to process training data more quickly and with less memory usage while maintaining or even improving XGBoost's accuracy and error value (Microsoft Coorporation, 2022). The most recent iteration of the gradient boosting model for machine learning algorithms, known as LightGBM, emphasizes speed in the processing of datasets (Ke et al., 2017).

2. Methodology

In this research, analysis of the machine learning model LightGBM and XGBoost as gradient boosting method

is covered to predict the energy of the Rooftop Solar Generation Plant at Laboratory Energy in the University of Kristen Immanuel (UKRIM), Yogyakarta.

Meteorological data has a strong correlation with its effect on the solar radiation produced, so this will also provide a more accurate level of prediction to determine the solar energy produced (Widodo et al., 2021). The weather data used consists of relative humidity, air temperature, wind speed, cloud density level, water precipitation level, solar radiation, and surface air pressure.

a. Variable data input

This study is modeled using meteorological time series data with an annual data range of January 2021–June 2021 in 5-minute interval data resolution obtained from the Solcast database site at the coordinates of latitude -7.775066 and longitude 110.45093. Then the sample data set is combined with the actual data of the PV system at the coordinates of the same location, namely the Energy Laboratory of UKRIM Yogyakarta. The dataset contains data rows totaling 52,128 database rows, which follow the distribution data shown in table 1.

Table 1. Distribution data of the dataset

Data Variable	Air Temp	Cloud Opacity	Dewpoint Temp	Dhi
Count	52,128	52,128	52,128	52,128
Mean	25.43	43.11	21.73	111.50
Std	2.69	34.03	0.92	159.53
Min	20.70	0.00	17.20	0.00
25%	23.20	11.90	21.20	0.00
50%	24.40	37.30	21.80	0.00
75%	27.80	75.40	22.30	195.00
Max	32.20	97.00	24.00	682.00

Table 1 (continued). Distribution data of the dataset

Data Variable	Dni	Ebh	Ghi	GtiFixedTilt
Count	52,128	52,128	52,128	52,128
Mean	137.87	103.11	214.61	216.90
Std	248.26	199.87	305.44	308.63
Min	0.00	0.00	0.00	0.00
25%	0.00	0.00	0.00	0.00
50%	0.00	0.00	0.00	0.00
75%	166.00	85.25	427.00	431.00
Max	924.00	887.00	1,042.00	1,042.00

Table 1 (continued). Distribution data of the dataset

Data Variable	Precipitable Water	Relative Humidity	Surface Pressure
Count	52,128	52,128	52,128
Mean	137.87	103.11	214.61
Std	248.26	199.87	305.44
Min	0.00	0.00	0.00
25%	0.00	0.00	0.00
50%	0.00	0.00	0.00
75%	166.00	85.25	427.00
Max	924.00	887.00	1,042.00

Table 1 (continued). Distribution data of the dataset

Data Variable	WindSpeed 10m	Pac	Temp
Count	52,128	52,128	52,128
Mean	1.77	1,468.77	34.42
Std	1.16	2,196.27	35.60
Min	0.00	0.00	0.00
25%	0.90	0.00	0.00
50%	1.50	0.00	0.00
75%	2.30	2,619.25	69.00
Max	6.80	9,259.00	99.00

b. Data pre-processing

Excel is utilized to prepare the dataset, and then it will be uploaded to a Jupyter notebook via the VS Code program for cleaning and feature selection. Cleaning missing data, selecting the proper data type for the following algorithm, and grouping data are the preliminary steps in the data processing process. This is the preparation stage for data (Martínez, 2018).

c. Train and test split data

In this research, a model will be run using the training data ratio of 70:30. This is done to examine how the training data and testing data proportionally influence the processing time and score accuracy.

d. Setting hyperparameter models

Hyperparameter settings are done by giving a range of values in advance for some of the hyperparameters used, namely learning rate, max depth, n_estimators, and feature fraction. The range of values for each hyperparameter will determine the best value of the hyperparameter that will be used by the two models. The best hyperparameter determination is carried out using the randomized searchCV method. This method is very efficient

in determining hyperparameters, from a given range of values, randomizedsearchCV will take a random pair of hyperparameters and perform modeling for each pair of hyperparameters (Li, 2020). Table 2 shows the best-obtained hyperparameters with a given range of values.

Table 2. Hyperparameter result using randomizedsearchCV

Hyperparameter	LightGBM model	XGBoost model	Range of value setting
Max depth	11	10	1-12
Learning rate	0.046	0.026	0.014-0.048
N_estimator	440	500	20-580
Feature fraction	0.7	-	0.1-0.9

The fraction hyperparameter feature is not used in the use of the XGBoost model in the package, so the default value is 1.0. Furthermore, hyperparameters are set in the model to make predictions using training data and data testing.

e. Model prediction and analysis

The next step is to set the learning rate, max depth, n estimators, and feature fraction parameters in machine learning to the range of values that will produce the best results for each model after the dataset is ready to be used. For each LightGBM and XGBoost model, data evaluation was performed using a methodology based on accuracy level, predicting time speed, MAE, and RMSE values.

To evaluate the model, this research accommodates the method using:

1. MAE, or mean absolute error, is another method for assessing forecasting approaches. Each error value or leftover amount is squared. Next, multiply the total by the number of observations. This method accommodates high predicting residuals (Barrera et al., 2020).

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}(i) - y(i)|$$

2. RMSE, or root mean squared error, which is based on the amount of data, is a measurement of the variance between the anticipated value and the actual value. When used on data with minimal outliers, this measurement will be more sensitive to big error values (Voyant et al., 2017).

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\big(\hat{y}(i) - y(i)\big)^2}$$

3. Predicting time speed is measured while model start to fit training data and predict the testing data. The recorded time has been done by Jupyter notebook in the time unit.

4. As for the level of accuracy of the two models using the method of determination coefficient R2, which measures the level of correlation between the independent variable and the dependent variable in the dataset (Kim et al., 2019).

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\big(y(i) - \hat{y}(i)\big)^2}{\sum_{i=1}^{N}(y(i) - \bar{y})^2}$$

Where, $y(i)$ is the actual value for i-th, $\hat{y}(i)$ is the predicted value for i-th, $\bar{y}$ is mean number for the actual value $y(i)$, and *N* is amount of total sample data in dataset (Kim et al., 2019). The process stages in the research using the XGBoost and LightGBM models for the machine learning algorithm are described in the flowchart in Figure 1.

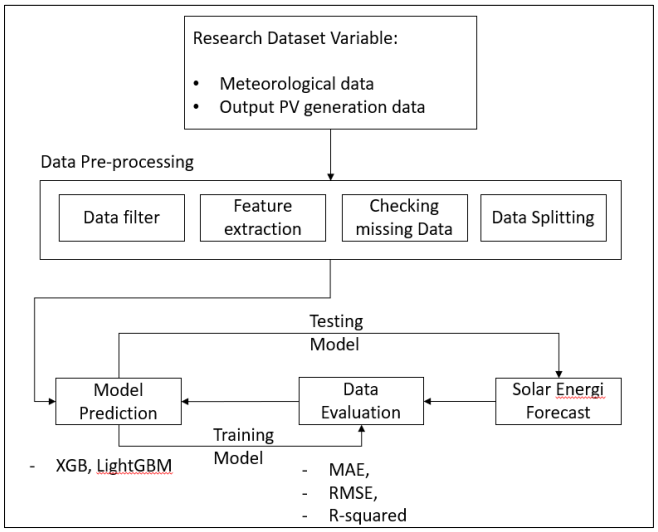

Figure 1. Flow diagram of solar prediction model using machine learning gradient boosting

3. Results & Discussion

The solar energy prediction modeling process has been carried out using a Jupyter Notebook with Python language, the dataset used has the following data analysis:

a. Solar radiation potential

Solar radiation is only available periodically, which makes it difficult to employ as an energy source. The energy that can be produced is obviously less than that time span due to typical circumstances, in which solar radiation is only available for a period of 8–10 hours (IESR, 2021).

The distribution pattern of solar radiation for the UKRIM area can be seen in Figure 2 below.

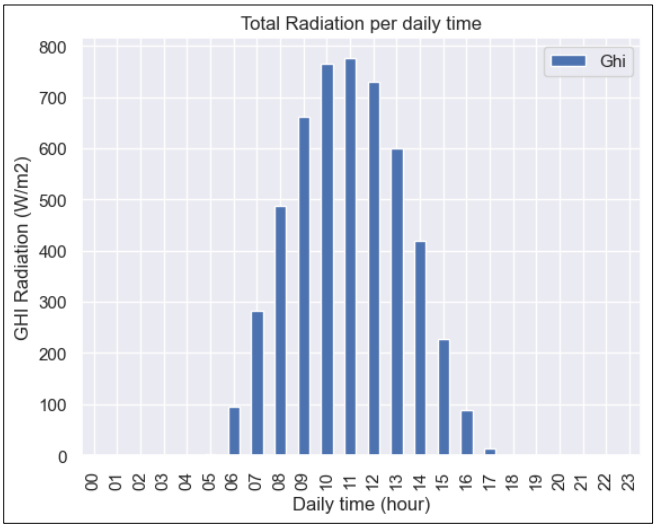

Figure 2. Pattern of solar radiation per daily time (GHI)

The level of solar radiation starts around 6 to 16 hours with peak conditions in the range of 10–12 hours during the day. This condition is in accordance with the energy pattern produced by rooftop solar power plants in the UKRIM laboratory. The amount of energy that is available in solar power plants is significantly impacted by changes in the weather and solar radiation intensity. Therefore, it requires a machine learning prediction system with a high level of stability.

b. Model evaluation on best hyperparameter setting

The modeling outcomes are as follows, based on the simulation of data testing utilizing the best parameters found using the randomizedsearchCV approach.

Table 3. Result of model evaluation using best hyperparameter

Model	LightGBM	XGBoost
Prediction time (sec)	0.86	19.90
Accuracy score, R2 (%)	94.60	97.19
MAE (Watt)	237	149
nMAE (%)	2.86	1.75
RMSE (Watt)	510	368
nRMSE (%)	6.15	4.31

The two algorithm models differ by about 2% in accuracy. The XGBoost model has a greater accuracy value than LightGBM, but this does not necessarily imply that XGBoost is superior to LightGBM. LightGBM is still taking into account the accuracy value of its performance, which is similarly above 90% with a lot faster speed. The XGBoost model also has a little lower error value than LightGBM in terms of the final error value.

c. Effect of learning rate parameter on gradient boosting models

A phenomenon known as overfitting, when the resulting pattern does not generalize to the actual value, starts to occur at a high learning rate setting, which is 0.048 in both models.

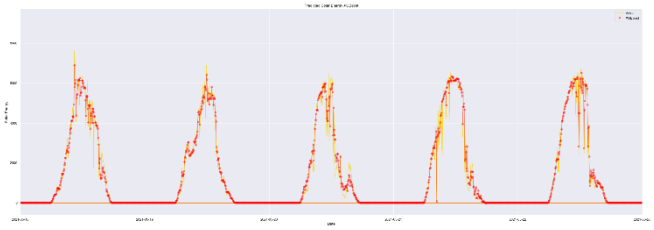

Figure 3. Solar energy prediction using XGBoost model in higher learning rate

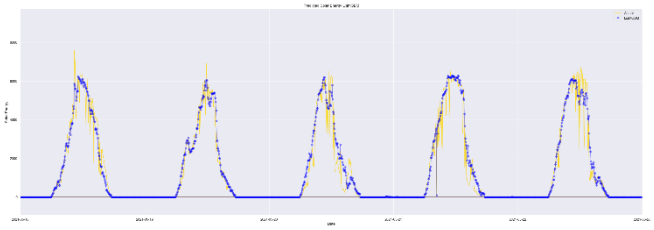

Figure 4. Solar energy prediction using LightGBM model in higher learning rate

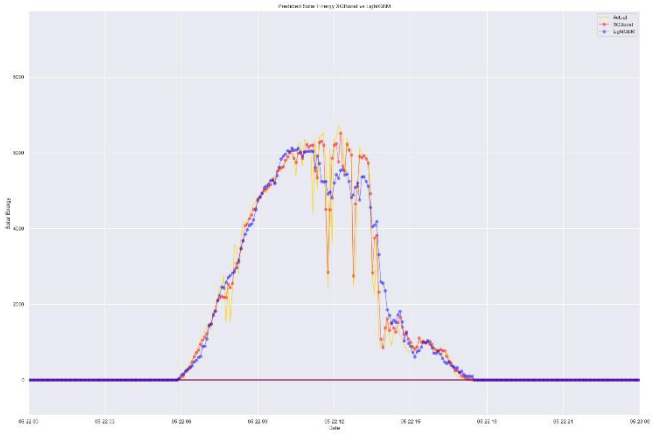

Figure 5. Comparation between XGBoost and LightGBM using higher learning rate

From Figure 5, it can be seen that the XGBoost model tends to follow actual values more closely than the LightGBM model. It is indicated that XGBoost has the potential to get overfitting predictions. Both models produce accuracy and error values under similar settings when the parameters for the learning rate are set at higher and lower conditions, as indicated in the table as follows.

Table 4. Hyperparameter with higher learning rate

	Result on learning rate 0.048	
Model	LightGBM	XGBoost
Prediction time (sec)	1.16	18.0
Accuracy score, R2 (%)	94.65	97.73
MAE (Watt)	236	121
nMAE (%)	2.84	1.35
RMSE (Watt)	507	330
nRMSE (%)	6.11	3.68

In Table 4, both models have a very high level of accuracy, above 90%, with the learning rate set at a position of 0.048. The nMAE and nRMSE error values of the XGBoost model are lower, with values of 1.35% and 3.68%, respectively, while the LightGBM model has an error value of 2.84% and 6.11%. Nevertheless, the comparison of the two models in their prediction patterns gives a different opinion. The LightGBM model still produces a better generalization pattern than the XGBoost model.

Furthermore, the two gradient boosting models were adjusted to the learning rate at a low condition, which is 0.014. The results of these settings provide a predictive pattern that is not too much different from a high learning rate condition.

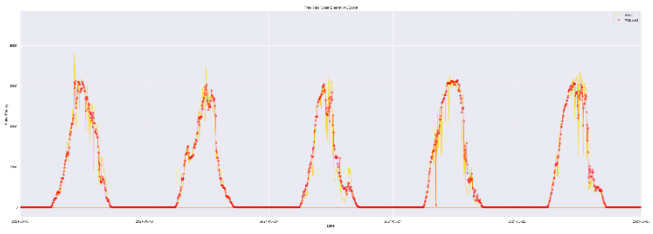

Figure 6. Solar energy prediction using XGBoost model in lower learning rate

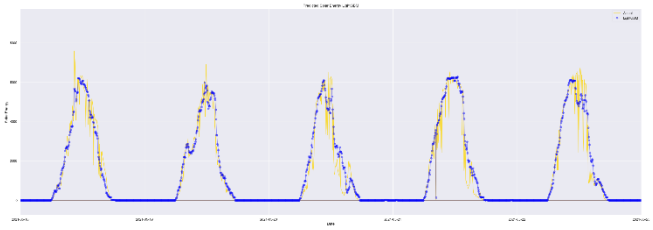

Figure 7. Solar energy prediction using LightGBM model in lower learning rate

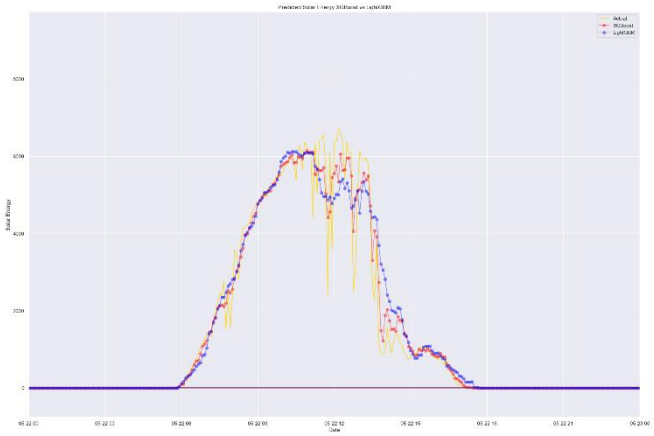

Figure 8. Comparation between XGBoost and LightGBM using lower learning rate

According to the given graph, using a low learning rate of 0.014 results in a different pattern between LightGBM and XGBoost when compared to the actual pattern. While the XGBoost model still suffers from an overfitting situation that forces the model to produce a pattern that is closer to the actual value, LightGBM can nevertheless identify patterns that are generalized.

In Tables 4 and 5, the prediction speeds of the two models for both high and low learning rate settings produce values that are close to the same. It can be seen that the change in the learning rate does not really affect the speed of the prediction of the machine learning model. However, both models require prediction speed with values of 1.19 seconds and 19.9 seconds while using low learning rate conditions.

Meanwhile, the R2 accuracy value in both XGBoost and LightGBM models experienced a slight decrease at low learning rate settings, with values of 93.02% and 96.6%, respectively. This is because the use of a low learning rate causes machine learning models to reduce their complexity for each iteration in a decision tree. Nevertheless, it can reduce the tendency for model overfitting to occur. The nMAE and nRMSE values for both models also experienced less significant up. For the XGBoost model, the results obtained were 2.08% and 4.90%, while for the LightGBM model they were 3.65% and 7.73%, respectively.

Table 5. Hyperparameter with lower learning rate

	Result on learning rate 0.014	
Model	LightGBM	XGBoost
Prediction time (sec)	1.19	19.9
Accuracy score, R2 (%)	93.02	96.6
MAE (Watt)	274	171
nMAE (%)	3.65	2.08
RMSE (Watt)	580	404
nRMSE (%)	7.73	4.90

d. Effect of max depth parameter on gradient boosting models

The analysis is performed in this part using high and low maximum depth settings. Max depth is one of the settings that control how many decision trees the model creates. Therefore, overfitting circumstances can be managed by max depth settings, and the max depth option can affect how quickly the model makes predictions on each decision tree employed (XGBoost developers, 2022). The following figures show the outcomes of various maximum depth settings.

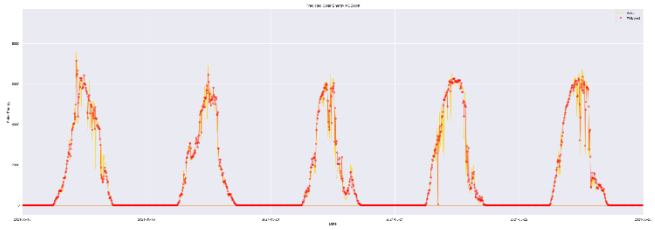

Figure 9. Solar energy prediction using XGBoost model in higher max depth

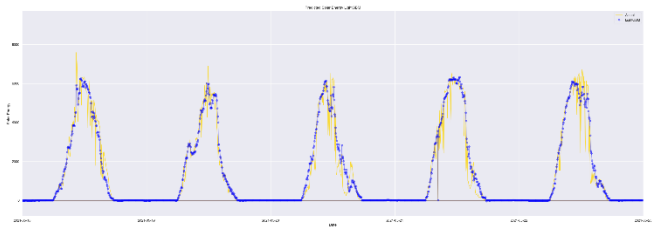

Figure 10. Solar energy prediction using LightGBM model in higher max depth

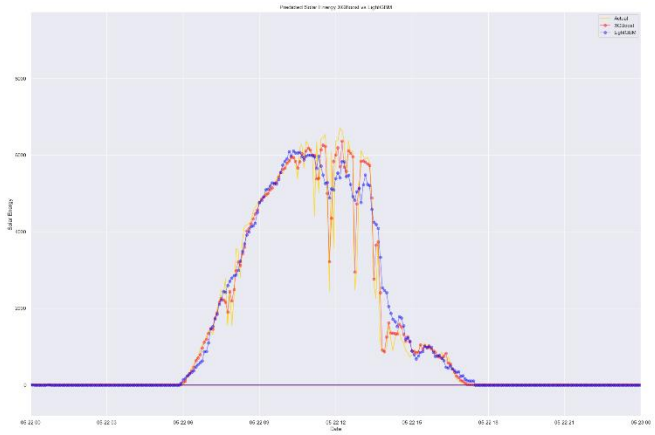

Figure 11. Comparation between XGBoost and LightGBM using higher max depth

At a high max depth setting, a predictive pattern is produced which is increasingly susceptible to overfitting. By increasing the max depth, the number of leaves in decision trees is increased. It would affect to both models to produce a predictive value that is getting closer to each point of the actual value. In figures 9 and 11, it can be seen that the XGBoost model has a more severe level of overfitting than the LightGBM model.

Furthermore, both models were analyzed using a lower max depth setting, to see the level of stability between the two models against a lower level of tree creation.

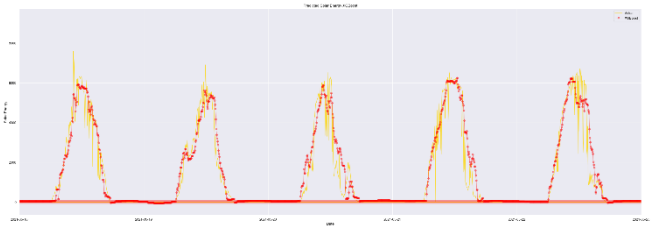

Figure 12. Solar energy prediction using XGBoost model in lower max depth

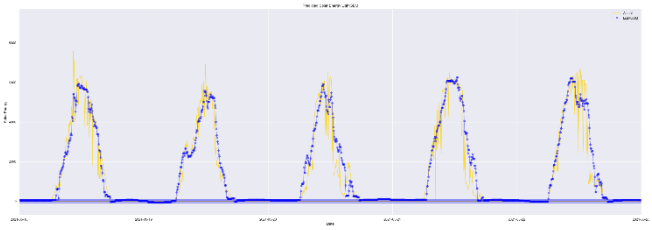

Figure 13. Solar energy prediction using LightGBM model in lower max depth

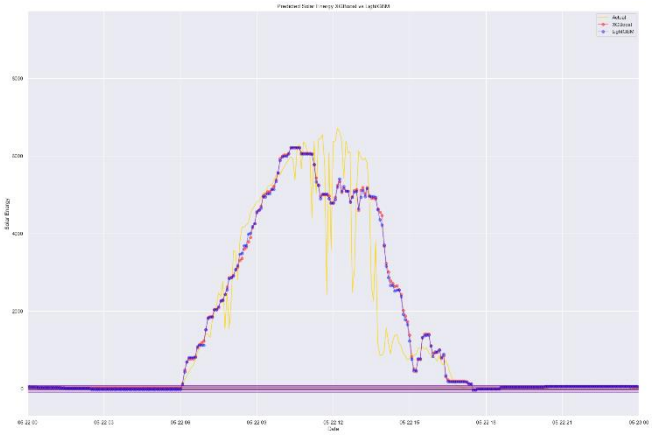

Figure 14. Comparation between XGBoost and LightGBM using lower max depth

A drastically different pattern is observed at a low max depth of 1, which results in a pattern that is underfitting. This condition exists in both models. The newly created prediction pattern begins to resemble a straight line. Then, by setting various parameters on max depth for the highest and lowest values, both models were put to the test. Based on the above conditions, it shows that the max depth has a significant influence in terms of model complexity. This is especially true for the XGBoost model, which uses the concept of a level-wise decision tree. The results are shown in the Table 6.

Table 6. Hyperparameter with higher max depth

	Result on max depth 12	
Model	LightGBM	XGBoost
Prediction time (sec)	1.04	23.7
Accuracy score, R2 (%)	94.62	97.75
MAE (Watt)	237	113
nMAE (%)	2.92	1.29
RMSE (Watt)	509	329
nRMSE (%)	6.28	3.77

The results of the high max depth setting are shown in Table 6, for the speed and accuracy values are still relatively the same when setting a high learning rate. While in Table 7 with a lower max depth setting, the results are much different. At the level of prediction speed, both models produce much faster predictions.

Significant changes occurred in the XGBoost model with a prediction speed of 2.92 seconds, and the speed is increased by about 10 times. Meanwhile, the accuracy value on XGBoost has decreased quite drastically to a level below 90%, with an increase in the error value of more than 2 times. Different conditions can be seen in the LightGBM model which is classified as experiencing a more stable change, for the accuracy value is still above 90% with the error value slightly increasing below 2 times.

Table 7. Hyperparameter with lower max depth

	Result on max depth 1	
Model	LightGBM	XGBoost
Prediction time (sec)	0.58	2.92
Accuracy score, R2 (%)	90.07	89.91
MAE (Watt)	366	368
nMAE (%)	4.82	5.02
RMSE (Watt)	691	697
nRMSE (%)	9.10	9.50

As observed from the data above, both models' accuracy and error values decreased when the max depth parameter was reduced. The prediction speed in both models has slowed down while the accuracy and error values have increased at high max depth. In terms of accuracy stability, the LightGBM model outperforms the XGBoost model at the testing stage of the learning rate and max depth parameters.

4. Conclusion

1. From different learning rates and hyperparameter settings, the LightGBM model can generate stable data prediction patterns and can overcome the tendency for overfitting conditions.
2. Meanwhile, in the XGBoost model to test the sensitivity of learning rate changes, the model is increasingly experiencing overfitting conditions when using a high learning rate.
3. Changes to the max depth hyperparameter have a significant effect on the XGBoost and LightGBM models. When the max depth is too low, the resulting pattern is close to underfitting. However, the LightGBM model still has a better level of stability than the XGBoost model, which is seen at an accuracy rate that is still above 90%.
4. The usage of max depth and a higher learning rate will affect the prediction speed level, with slower results. While the value of the accuracy level and the resulting error value will increase, there can be a risk of overfitting the model.

References

Ahmed Kutty, H., Masral, M. H., & Rajendran, P. (2015). Regression model to predict global solar irradiance in Malaysia. *International Journal of Photoenergy*, *2015*(January). https://doi.org/10.1155/2015/347023

Barrera, J. M., Reina, A., Maté, A., & Trujillo, J. C. (2020). Solar energy prediction model based on artificial neural networks and open data. *Sustainability (Switzerland)*, *12*(17). https://doi.org/10.3390/SU12176915

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-Augu*, 785–794. https://doi.org/10.1145/2939672.2939785

Choi, S., & Hur, J. (2020). An ensemble learner-based bagging model using past output data for photovoltaic forecasting. *Energies*, *13*(6). https://doi.org/10.3390/en13061438

IESR. (2021). Indonesia's Largest of Solar Farm: Key Opportunities and Challenges. *Indonesia's Largest of Solar Farm*.

Ikhsan, M. (2020). *Peramalan Energi Photovoltaic Dengan Menggunakan Algoritma Naive Bayes Classifier dan K-Nearest Neighbors*. Universitas Islam Indonesia.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural*

Information Processing Systems, *2017-Decem*(Nips), 3147–3155.

Kim, S. G., Jung, J. Y., & Sim, M. K. (2019). A two-step approach to solar power generation prediction based on weather data using machine learning. *Sustainability (Switzerland)*, *11*(5). https://doi.org/10.3390/SU11051501

Lai, J. P., Chang, Y. M., Chen, C. H., & Pai, P. F. (2020). A survey of machine learning models in renewable energy predictions. *Applied Sciences (Switzerland)*, *10*(17). https://doi.org/10.3390/app10175975

Li, B. (2020). *Random Search Plus: A more effective random search for Random Search Plus: A more effective random search for machine learning hyperparameters optimization machine learning hyperparameters optimization*. https://trace.tennessee.edu/utk_gradthes/5849

Martínez, C. F. (2018). Bachelor Degree Thesis - Machine Learning for Solar Energy Prediction. *University of Gavle*, *May*, 1–80.

Microsoft Coorporation. (2022). lightgbm-readthedocs-io-en-latest. *LightGBM Documentation*. https://lightgbm.readthedocs.io/en/v3.3.2/

Reba, K., Bevc, J., Vásquez, A.-, & Jankovec, M. (2019). Photovoltaic Energy Production Forecasting using LightGBM. *55th International Conference on Microelectronics, Devices and Materials with the Workshop on Laser Systems and Photonics*, 36–39.

Voyant, C., Notton, G., Kalogirou, S., Nivet, M. L., Paoli, C., Motte, F., & Fouilloy, A. (2017). Machine learning methods for solar radiation forecasting: A review. *Renewable Energy*, *105*, 569–582. https://doi.org/10.1016/j.renene.2016.12.095

Widodo, D. A., Iksan, N., Udayanti, E. D., & Djuniadi. (2021). Renewable energy power generation forecasting using deep learning method. *IOP Conference Series: Earth and Environmental Science*, *700*(1). https://doi.org/10.1088/1755-1315/700/1/012026

XGBoost developers. (2022). xgboost-readthedocs-io-en-latest. *XGBoost Documentation*. https://xgboost.readthedocs.io/en/stable/

Zhou, Z., Lin, A., He, L., & Wang, L. (2022). Evaluation of Various Tree-Based Ensemble Models for Estimating Solar Energy Resource Potential in Different Climatic Zones of China. *Energies*, *15*(9). https://doi.org/10.3390/en15093463