Performance Analysis of Machine Learning Models in Solar Energy Forecasting

Nifat Sultana and Tasnim Ahmed*

***Abstract*—Energy is essential to facilitate the social and economic growth of a society. But this energy generation using fossil fuels results in a tremendous amount of greenhouse gas emissions. As a renewable alternative, the increased competitiveness of solar PV panels has increased the number of solar energy generation stations in recent years. Since solar power generation is highly intermittent and dependent on local weather characteristics, AI can be implemented to predict solar energy output from a solar power plant. As the need to predict solar photovoltaic (PV) energy output is essential for many actors in the energy industry, Statistical Data Analysis and Machine Learning (ML) can be employed towards this end. In this study, comparative analysis of different machine learning models is performed to estimate power-plant solar energy generation from historical meteorological data. A variety of supervised machine learning techniques are implemented to predict and forecast solar energy. The implemented models include Weighted Linear Regression (WLR) with and without dimensionality reduction, Gradient Boosting Model (GBM), and Artificial Neural Networks (ANN). Findings indicate that, both the ANN and GBM models performed significantly well in short-term prediction, whereas Long-Short Term Memory (LSTM) Recurrent Neural Network (RNN) achieved reliable performance in forecasting. The trained models, therefore, may provide a way for grid-operators to Predict and balance energy generation and consumption.**

***Index Terms*—Machine learning, solar energy, prediction, forecasting**

I. Introduction

The historical accumulation of Green House Gas (GHG) in the atmosphere has resulted in concerning changes in climate such as, rising sea levels, warming oceans, rising global temperature, frequent natural calamities, and many others. These concerning changes, along with ever-growing energy demand, is gradually pushing humanity towards renewable alternatives. Solar energy prediction and forecasting can provide a way for grid-operators to predict and balance energy generation and consumption. Therefore, one of the key benefits of solar energy forecasting is to increase the efficiency of electric grid management. This is a need with significant importance, as solar energy usage continues to expand.

Utilization of Artificial Intelligence (AI), such as ML algorithms, has already been proven to be an efficient way of creating data-driven models for prediction and forecasting. Although ML techniques are nothing new, the higher availability of quality data and the improved computational capacity of modern computers have made these techniques useful for predictive analysis.

In this study, three established prediction models have been compared, and one forecasting model has been analyzed. The ML models for prediction are – Locally Weighted Linear Regression (LOWESS) as WLR model, Gradient Boosting Model (GBM), and Multilayer Perceptron (MLP) as an ANN model. LSTM-RNN has been used as the forecasting model. Each of these prediction models are unique in terms of their method of operation. WLR is regression-based method, whereas GBM uses decision-tree approach. On the other hand, ANN is a network of connected neurons where each neuron performs regression operation. Additionally, correlation-test and Principal Component Analysis (PCA) have been used to figure out the most influential meteorological parameters for predicting solar PV energy output.

Several Approaches have already been investigated for solar PV energy output prediction and forecasting. Mashud and Irena *et al.* [1] used weather-data clustering and ensemble of multiple ANN models for Solar Power forecasting. Then akuzmiakova and Colas *et al.* [2] showed the efficiency of LSTM in short-term solar energy forecasting based on weather-data. Chuluunsaikhan and Nasridinov *et al.* [3] compared ML models for predicting Power Output of Solar Panels based on Weather and Air Pollution Features. Zhang and Zou [4] utilized K-means algorithm for weather-data clustering, and the applied Support Vector Machine (SVM) for photovoltaic output prediction. Then Javier and Pastoriza *et al.* [5] showed the efficiency of ANN model in photovoltaic power prediction using numerical weather data.

In this study, the characteristics of weather-parameters will be investigated and compared with respect to solar PV output characteristics. Thus, the most influential parameters will be figured out. The potential of the four unique ML models will also be explored from different perspectives. Finally, the findings will be represented as a comparative analysis.

II. Data Collection and Processing

Operation at this section consists of two sequential steps: collection, and pre-processing

A. Data Collection Process

The solar energy generation plant, selected for this analysis, is situated at the University of Illinois campus. The historical power generation data for this plant is publicly available [6]. To obtain a reliable estimation of the weather conditions around the solar power-plant, data were collected from three weather stations, closest to the plant [2], as shown in Table I. Historical weather data for these stations are publicly available at the website of National Center for Environmental

Manuscript received October 20, 2022; revised December 11, 2022; accepted Mach 7, 2023.

Nifat Sultana is with University of Dhaka, Bangladesh.
Tasnim Ahmed is with Rakuten Mobile, Inc., Japan.
*Correspondence: ahmed1302185@gmail.com (T.A.)

doi: 10.18178/ijml.2023.13.3.1140

Information [7]. Nearly two years (2016 and 2017) of historical time-series data were collected for the weather-parameters, listed in Table II. Analyzed PV output data were also for the same period.

TABLE I. WEATHER STATION DETAILS

Weather station Name	Distance from power plant (km)	City, State code	Country
Airport of Santa Clara County	5.8	Palo Alto, CA	United States
Moffett Federal Field Airport	10.4	Mountain View, CA	United States
San Carlos Airport	12	San Carlos, CA	United States

TABLE II. WEATHER PARAMETER DETAILS

Weather feature	Data Type	Unit
Sky condition	Categorical(ordinal)	-
Visibility	Continuous	Miles
Temperature	Continuous	C
Dew point	Continuous	C
Relative humidity	Continuous	%
Wind speed	Continuous	Mph
Station pressure	Continuous	Inch-Hg
Altitude(Altimeter)	Continuous	Inch-Hg

B. Data Pre-processing

The data pre-processing and wrangling approaches plays a major role in this study. These approaches are listed below:

- Feature-engineering starts the processing operation. The objective here is to transform the ordinal categorial parameters, such as sky-condition into ratio, so that it can be quantified.
- Missing time-stamps in the series are filled-up with the average of four nearest time-stamps, taken equally from both preceding and leading ones.
- Aggregation of weather parameter values from three different stations is conducted by taking their weighted average using barycenter formula, as shown in Eq. (1).

$$x_j = \frac{d_A x_{j,A} + d_B x_{j,B} + d_C x_{j,C}}{d_A + d_B + d_C} \tag{1}$$

Here d_A is the distance from weather station, A to power-plant. x_j, A is the value of feature x_j measured by weather-station A. With this formula, the closest a weather station is, the more weight it has.

- Resampling both weather data and PV output time-series to one hour granularity, so that these two series can be merged together using time as the common primary-key. The hourly resolution of the data ranges in between 6AM and 5PM, since solar energy is not produced at night, and therefore, night hours have been avoided.
- Feature scaling with min-max Normalization of all weather features and the PV output. If we consider the feature X with n observations, then normalization formula for each $X_i, i \in \mathrm{N}$ is shown in Eq. (2). This approach transforms all values into the range of 0 and 1 , and assures that all parameters have the same scale.

$$X_i' = \frac{X_i - minX}{maxX - minX} \tag{2}$$

- Finally, creating a duplicate copy of the dataset and randomizing the time-stamps for the prediction models. The sequential data will be used for the forecasting model only.

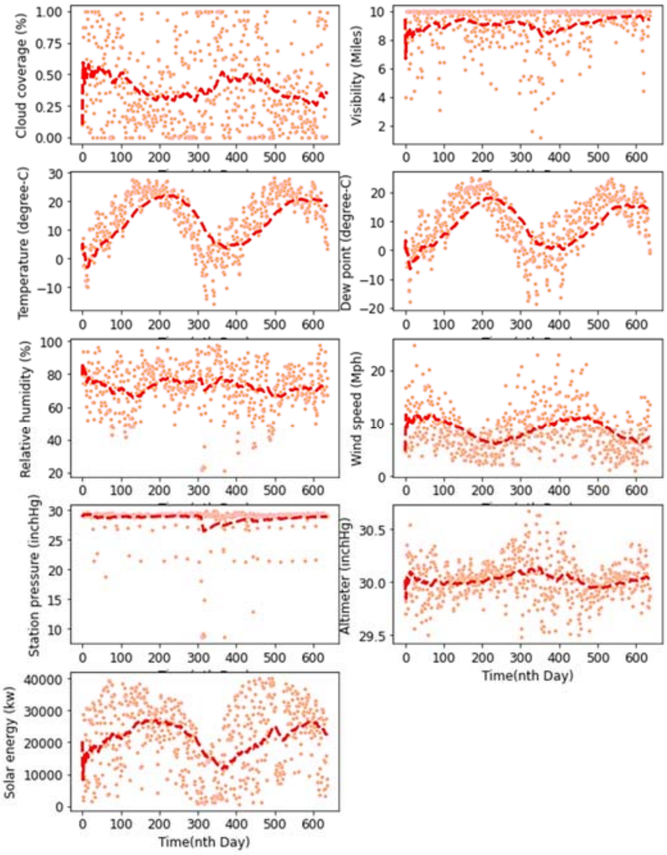

Fig. 1. Feature trend visualization.

III. STATISTICAL ANALYSIS

Statistical-analysis on Meteorological feature is crucial to understand their characteristics, and their influence on Solar PV output. Feature trend analysis was the first statistical approach in this regard. Visualization of underlying trends provided significant insight into trend-similarities among the features. Exponentially Weighted Moving Average (EWM) approach was considered for feature-trend visualization. Daily average raw values for all features for 600 consecutive days were considered for this assessment. As shown in Fig. 1, trend visualization shows periodic characteristics. The sinusoidal seasonality is significantly noticeable among the features. Therefore, there must be significant correlation among them. Especially, Temperature, visibility, and Due-point trends looks significantly similar to PV output. On the other hand, the trend of cloud coverage and wind-speed looks almost the opposite. Now, due to this similarity in seasonal characteristics, ML models can also be trained efficiently to learn from data seasonality patterns, and provide predictive solutions based on that. Then, Pearson-method based monthly correlation of weather parameters with respect to solar PV output is observed, as shown in Fig. 2. It clearly distinguishes the features that are highly correlated with solar PV output. Mathematical representation of Pearson correlation is shown in Eq. (3).

$$\rho = \frac{\sigma XY}{(\sigma X)(\sigma Y)} \tag{3}$$

Here (σXY) is the covariance between variable X and Y, whereas (σX), and (σY) are the standard deviation of X, and Y respectively. In Fig. 2, The red bars denote values where the absolute correlation is greater than 0.3, which corresponds to significant correlation.

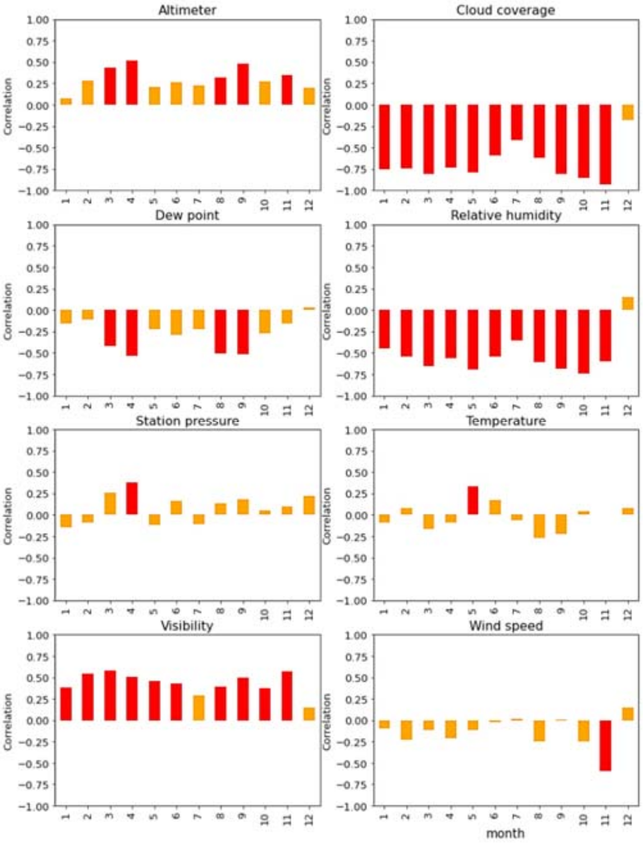

Fig. 2. Monthly correlation of weather parameters with PV output.

It also shows that the parameters such as Cloud-coverage, Visibility, Relative Humidity, Altitude(Altimeter), and Dew point have quite strong correlation with PV energy generation. Therefore, these features can be considered the most influential. Internal correlation among all features is visualized in Fig. 3.

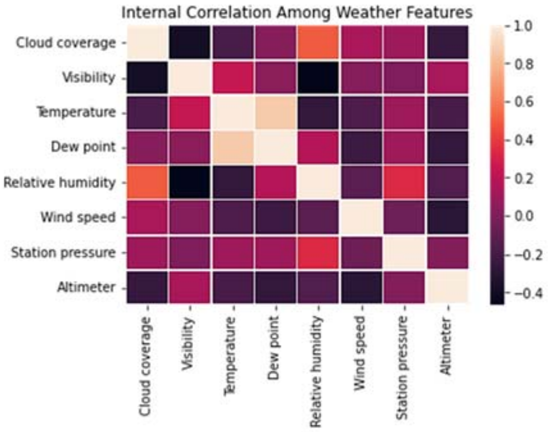

Fig. 3. Correlation among the weather features in hourly time-series.

Fig. 3 shows that, some features are highly correlated to each other, such as Temperature and Due point. Also cloud-coverage and relative-humidity are correlated significantly. That is why PCA is applied to the features to reduce dimensionality and eliminate internal correlation among features. Commutative explained variance of PCs is visualized in Fig. 4. There it is observed that, the first six (6) PCs can combinedly explain more than 90% of the overall variance. Now, training an ML model with all PCs may have the tendency to overfit to some extent, as the test-data may not always be 100% correlated with train-data. Therefore, by using 6 PCs, we can introduce a small amount of uncertainty to the model, so that it can better adapt the diversities in test data. The relative influence of weather features on the most influential PC is visualized in Fig. 5. It is observed that, two features such as, Temperature and Wind-speed, have been downgraded the most. The remaining six features have significant influence on the PC.

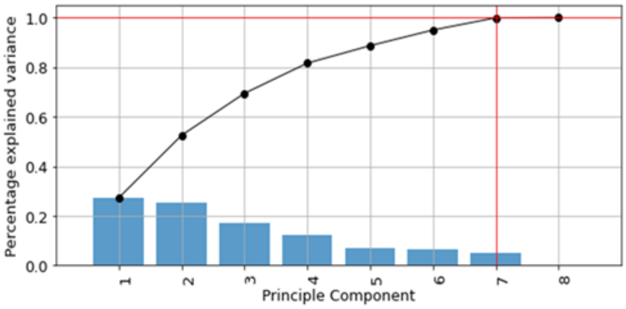

Fig. 4. PCA Cumulative explained variance.

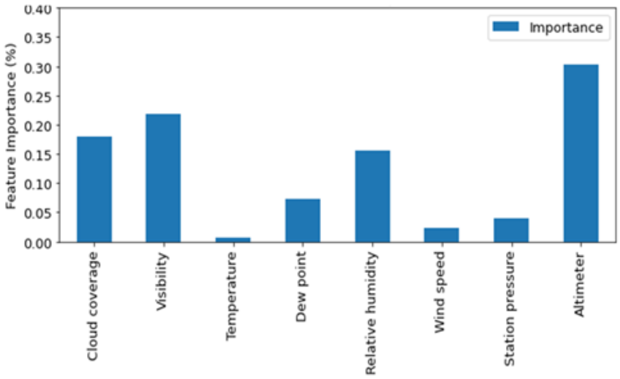

Fig. 5. PCA based feature importance.

IV. Machine Learning Models

In this section, ML models for PV output predicting and forecasting are demonstrated along with their hyperparameters.

A. *Splitting Train and Test Data*

About 90% if the processed data was used for training ML models, which is equivalent of around twenty months. The remaining 10% data was used for testing models. Both train and test data were randomly sequenced for prediction models.

B. *Hyperparameter Optimization*

Grid-search algorithm was used for optimizing hyperparameters. The hyperparameter ranges for LOWESS, GBM and MLP are shown in Table III, IV, and V respectively. Here, LOWESS is a regression method, where the data is split into small partitions, so that each partition fits to a liner line.

The key parameter to tune LOWESS model, is Sigma and span. Sigma indicates how widely the data will be smoothened. And the span indicates what percentage of the data is to be used.

TABLE III. LOWESS HYPERPARAMETERS

Hyperparameter	Search grid range	Optimal
Sigma	0.01 to 0.2	0.1
Span	0.1 to 0.9	0.8

On the other hand, GBM is a Decision-tree based model, where the number of estimators controls the amount of total sequential trees, and maximum depth ensures the number of branches each tree may have. Minimum sample split controls how many data is needed as minimum to split it into two branches. Smaller learning rate ensures smooth training performances but longer computation time.

TABLE IV. GBM HYPERPARAMETERS

Hyperparameter	Search grid range	Optimal Value
Number of estimators	100 to 1000	750
Maximum Depth	3 to 12	9
Minimum sample split	3 to 12	9
Learning rate	0.005 to 0.05	0.01

The ANN model is having three hidden layers. Number of Nodes per layer changed in between 25 to 105 with a regular interval of 20. The input layer of the network is having 8 neurons, for 8 meteorological inputs. For forecasting operation, a standard vanilla LSTM model [8] was trained with 50 training epochs and batch-size of 26. Only these two parameters were optimized with grid-search algorithm.

TABLE V. ANN HYPERPARAMETERS

Hyperparameter	Search grid range	Optimized value
Nodes per layer	25 to 105	100
Weight decay	0.01 to 1	0.1
Weight Initializer		Xavier-He
Hidden layers	1 to 3	3
Data batch-size	5-25	16
Number of epochs	50 - 130	100
optimizer		Adam

V. RESULT AND DISCUSSION

In this section, the predictive and forecasting performances of machine learning models have been compared in terms of Root Mean Squared Error (RMSE) and R-squared (R2) score parameter. The RMSE equation is given by:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2}{n}} \qquad (4)$$

TABLE VI. PERFORMANCE COMPARISON OF ML MODELS

ML Model Name	Number of PC used	Explained Variance (%)	RMS Error	R2 score
WLR	7	100	956.9	0.6
WLR	6	95	954.74	0.6
WLR	5	85	976.12	0.6
WLR	Non-PCA		955.52	0.6
ANN	Non-PCA		944.49	0.6
GBM	Non-PCA		940.74	0.6
LSTM (with non-Temporal Data)	Non-PCA		1048.38	0.5
LSTM (with Temporal Data)	Non-PCA		1005.6	0.5
LSTM (with Sequential Data)	Non-PCA		868.48	0.7

Here Y_i corresponds to the true value and $\hat{Y}_i$ is the forecast. On the other hand, R2 score is the proportion of variance (%) in the dependent variable that can be explained by the independent variable. Table VI shows that, using six PCs has reduced RMSE error the most among the WLR models, trained with PCs.

PC based WLR model also performed better than the non-PC WLR model. Therefore, it is proven, that PCA can improve the ML model efficiency as well. The overall performance of GBM and ANN were the best among all prediction models. The performance of LSTM on sequential data is significantly better, suggesting that, the LSTM based forecasting produces better results than the prediction models.

Finally, R2 scores for all prediction models are almost similar (0.6), representing moderately well fit to test data. One superior characteristic of ANN model can be explored from Table VII. It shows that, the ANN has predicted the highest number of samples in low error range (within 1000 kw). So, ANN predicted the highest number of samples more accurately.

TABLE VII. NUMBER OF SAMPLES PER PREDICTION ERROR RANGE

	Number of samples per Prediction Error Range		
Model Name	Within 500 kw	Within 1000 kw	Within 1500kw
WLR	320 (44.08%)	514 (70.8%)	640 (88.15%)
GBM	311 (42.7%)	509 (70.1%)	648 (89.3%)
ANN	348 (48%)	532 (73.3%)	639 (88.2%)

Fig. 6 shows the residual plot for ANN, where it is observed that, majority of the predicted samples are almost uniformly scattered within 1000kw range from the ground truth. Only a handful of outliers have gone far beyond this range. Fig. 7 shows that, the predicted PV output closely matches to the actual values, for majority of the randomized test samples.

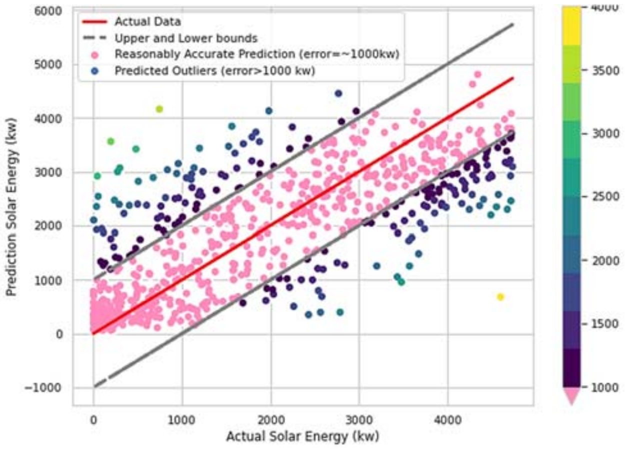

Fig. 6. Residual plot for ANN model.

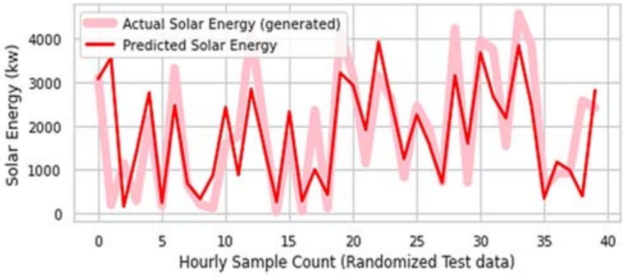

Fig. 7. Predicted and actual solar PV output comparison for ANN.

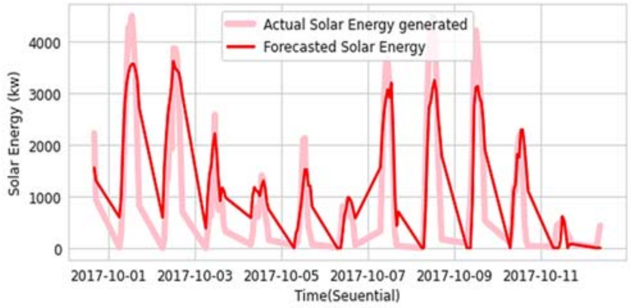

Fig. 8. Predicted and actual solar PV output comparison for LSTM.

For LSTM, Fig. 8 shows that the forecasted PV output time-series closely matches with the actual values. Although there is slight difference in amplitude occasionally, but the trend is followed quite accurately. Fig. 9 confirms that, showing the quantity of outliers, that is significantly low (around 8%).

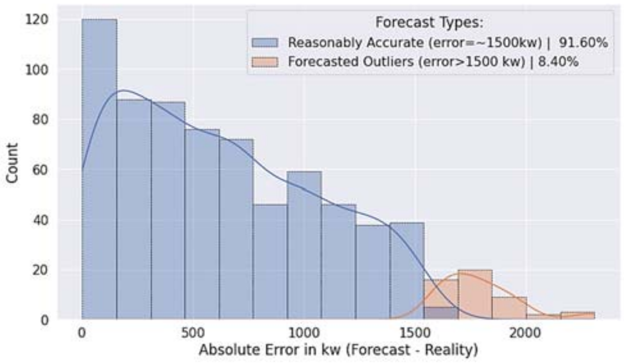

Fig. 9. Forecast error histogram for LSTM.

VI. Conclusion

The study showed, that it is possible to predict or forecast Solar PV output from weather data efficiently, using ML models. If the models are trained properly with decent amount of historical data, reliable performance can be expected. Among the selected weather parameters, Cloud Coverage, Relative Humidity, Visibility, Dew point, and Altitude are found to be the most influential for solar energy prediction. PCA has been proven as an essential tool for dimensionality reduction of the weather features, and thus, improving the predictive efficiency of the WLR model. ANN model performance is found to be the most reliable within 1000kw error range, and therefore, it can be the bast choice for scenarios where data samples for all seasonality are adequately available. Otherwise, GBM could be one of the best alternatives to consider, due to its superior adaptability to generalize the unusual data samples. Therefore, an ensemble of different ML models could be evaluated as a continuation of this research. Also, alongside the tabular-formatted weather data points, the satellite imagery can be incorporated as features, to further increase the knowledge base of the ML models, and thus improving the prediction and forecasting efficiency.

Conflict of Interest

The research was carried out without any conflict of interest among the authors. Therefore, the authors declare no conflict of interest.

Author Contributions

This research work is by Nifat Sultana, and Tasnim Ahmed, and they are the corresponding authors of this paper. Nifat sultana was focused in data collection, sampling, pre-processing, ML model training, evaluation of the model performances, and documentation. Whereas, Tasnim Ahmed contributed in literature review, weather feature analysis, ML model system architecture design, and development of that architecture. All authors had approved the final version of the paper that is submitted.

References

[1] R. Md, K. Irena, and A. Vassilios, "Solar power forecasting using weather type clustering and ensembles of neural networks," *IJCNN*, 2016.

[2] A. Kuzmiakova, G. Colas, and A. M. McKeehan, "Short-term Memory Solar Energy Forecasting at University of Illinois," *Environmental Science, Computer Science*, 2017.

[3] T. Chuluunsaikhan, A. Nasridinov, W. S. Choi, D. B. Choi, S. H. Choi, and Y. M. Kim, "Predicting the power output of solar panels based on weather and air pollution featuresusing machine learning," *Journal of Korea Multimedia Society*, vol. 24, no. 2, pp. 222-232, February 2021.

[4] K. K. Zhang and G. B. Zou,"Photovoltaic output prediction method based on weather forecast and machine learning," *J. Phys.: Conf. Ser. 2320 012032*, 2022.

[5] G. J. Martínez, A. Pastoriza, F. F. Garrido *et al.*, "Photovoltaic power prediction using artificial neural networks and numerical weather data," *Sustainability*, vol. 12, p. 10295, 2020.

[6] University of Illinois Solar Power Plant Dashboard. [Online]. Available: http://s35695.mini.alsoenergy.com/Dashboard/2a5669735065572f4a42454b772b714d3d

[7] Local Climatological Database (LCD) for weather data collection here. [Online]. Available: https://www.ncdc.noaa.gov/cdo-web/datatools/lcd

[8] V. Houdt, G. Mosquera, and C. N. Gonzalo, "A Review on the Long Short-Term Memory Model. Artificial Intelligence Review," *Artificial Intelligence Review*, vol. 53, pp. 5929–5955, 2020.