IETE TECHNICAL REVIEW
2023, VOL. 40, NO. 4, 558–573
https://doi.org/10.1080/02564602.2022.2136270

An Empirical Analysis of Machine Learning Algorithms for Solar Power Forecasting in a High Dimensional Uncertain Environment

Amit Rai[1], Ashish Shrivastava [2] and K. C. Jana[1]

[1]Electrical Engineering Department, Indian Institute of Technology (ISM), Dhanbad, India; [2]Skill Faculty of Engineering and Technology, Shri Vishwakarma Skill University, Gurugram, India

ABSTRACT

In the fabric of energy generation, solar power is the most promising clean energy solution as an alternative to non-renewable energy sources. However, solar power's dependency on environmental factors adds uncertainty to energy production. In such a scenario, solar power forecasting provides an edge to mitigate this uncertainty and improves overall system stability. Recently, machine learning (ML) models have been extensively deployed for designing and forecasting solar power. However, data pre-processing, forecast horizon, and performance evaluation of ML algorithms have to be carefully evaluated to find an accurate model. This paper provides an empirical comparison of different generation ML models for solar power forecasting, which can help understand future research on which method to adopt, depending on the ML model's strengths and weaknesses. Therefore, an effective forecasting method is designated in aspects such as performance errors, convergence time, and computational complexity. So, this work rates different ML models on error performance metrics and convergence time. Moreover, cross-fold validations and hyperparameter are also examined for the top five performing models for a comprehensive evaluation and to give more intuitive and calibrated insight into various stakeholders working in the solar power plant modeling field.

KEYWORDS

AdaBoost; Ensemble methods; LightGBM; Linear regression; Machine leaning; Regularization method; Solar power; XGBoost

ABBREVIATIONS

SPF	Solar power forecasting
AR	Auto-regressive
MA	Moving average
ML	Machine learning
ET	Extra trees regression
LIGHTGBM	Light gradient boosting machine
XGBOOST	Extreme gradient boosting
RF	Random forest
GBR	Gradient boosting regression
LR	Linear regression
BR	Bayesian ridge
LASSO	Lasso regression
MAE	Mean absolute error
MAPE	Mean absolute percentage error
RIDGE	Ridge regression
DT	Decision tree
KNN	K Neighbors regression
ADA	AdaBoost regression
EN	Elastic net
OMP	Orthogonal matching pursuit
LAR	Least angle regression
SVM	Support vector machine
RMSLE	Root mean squared logarithmic error
SP	Solar power

1. INTRODUCTION

The world's demand for electricity is increasing with growing industrialization and population and it is a crucial component for the overall development of a nation. The growth rate of energy generation jumped approximately 3% per year during 2000–2018, a period which showed economic growth worldwide [1]. In 2019, the growth was approximately 2%, showing slower economic activity worldwide due to COVID restrictions. Nowadays, energy consumption per capita is a symbol of growth for any country. Coal-based power plants have a major share of approximately 36% of overall energy generation [2]. However, coal-based power plants have anthropogenic greenhouse gas emissions as a by-product and are perishable in nature, so presently focus is shifting toward renewable energy generation systems [3]. Different climatic conventions are also stressing the use of renewable energy to generate power, like the Conference of Parties 21 (COP21), which was signed by 196

countries. These renewable energy sources (RES) include solar, wind, tidal, hydro, ocean, and geothermal, and recent advancements in these RES allow cost-effective, reliable, and sustainable means of extracting energy.

Renewable energy sources are increasing their presence in overall energy generation and increased generation of 1214529 MW from the year 2009 to 2018 [4]. Currently, in these renewable sources, hydropower has a major proportion. However, among these renewable energy sources, solar power is abundantly available and has the potential to accomplish the world's growing energy demand [5]. The solar power capacity has increased from 2.02% to 20.66% from the year 2009 to 2020, an approximate growth of 18.64% in the same duration. However, the wind percentage has increased from 10.77% and hydropower has reduced to 32.30% in overall renewable power generation worldwide in the same duration.

Solar power plants (SPP) generate electricity using semiconductor material, which directly converts the solar radiation into current, without having any moving part or carbon emission. The reduced costs of solar panels are also adding to the growth of SPP capacity across the globe [6]. In recent decades, the share of SPP production in the grid is also increasing. This increasing presence of solar power improves economic, environmental, and installation benefits in isolated locations. However, solar power generation is uncertain due to its environmental dependency which poses a threat to grid stability [7]. To accommodate this stability issue to solar power, a backup power source or battery is required [8]. But, these methods are not cost-effective alternatives, so, solar power forecasting is a pivotal step in designing and mitigating the random behavior of SPP's output with an increase in profitability [9,10].

Solar power forecasting (SPF) has gained a lot of attention from the research fraternity to address the incorporation challenges of renewable sources into the grid. It is a time series problem and can be addressed with regression analysis. Recently numerous physical, statistical, and data-driven regression models are used for accurate solar power prediction. The physical models use numerical and metrology data analysis methodology for solar power prediction [11]. These methods are simple for one variable input. However, complexity increases with the increase in independent variables. The limited data availability of SPP also limits the prediction accuracy of solar power [12]. These historical models use historical data and map it with solar power (SP) to predict the next time stamp solar power output [13]. In data-driven approaches, different inputs such as solar radiation, solar power, wind speed, humidity, sun angle, temperature, and weather parameters, are given to the algorithm which fits a regression relationship between input and output variables to predict SP [14].

The data-driven approaches are currently gaining a lot of attention in nearly all spheres of science and society. In solar power prediction, machine learning models have been widely applied with different combinations of input variables, such as solar radiation, temperature, wind speed, humidity, pressure, cloud, *etc*., for other geographical regions of the world. Nearly all the machine learning methods have been applied linear regression, support vector machine, random forest, decision tree, ensemble methods, and hybrid models to predict SP.

The statistical SPF methods are utilized for different locations of the world. The statistical techniques vary from linear to non-linear models. The statistical model extracts the pattern from input variables to predict the target variable [15]. The auto-regressive class of statistical techniques evaluates statistical components such as auto-regressive (AR), integrated (I), and moving averages (MA) from the time-series dataset. These statistical variables are then used to establish a relationship between the input and the target variable. ARIMA and SARIMA are used widely for solar power forecasting in the literature [16–19]. However, these methods perform well for short-term prediction but their performance decreases in long-term forecasting and reviewed literature also shows a skewed comparison of the same. Moreover, tuning the statistical parameters is also a challenging task. Linear regression is one step ahead of AR models as it does not require tuning the statistical parameters. Linear regression tries to find a linear relationship between the input and the target variable. These methods are simple, easy to implement, and less resource hungry as compared to advanced machine learning (ML) methods, and also extensively used in solar power forecasting [20–26]. However, solar radiation and power are non-linear time-series data, which limit the efficiency of linear regression in solar power forecasting. Next, upgradations of linear models are regularization methods or ensemble methods. The regularization methods works on L1 and L2 regularization where these algorithms take care of correlated features and errors [27–29]. Non-linear methods, such as state vector machine, random forest, ensemble learning, and gradient boost, give an edge over the aforementioned techniques in terms of tapping the non-linearity in the solar power or radiation dataset [30]. Moreover, empirical mode decomposition will further enhance the performance of ensemble models [31,32]. However, most

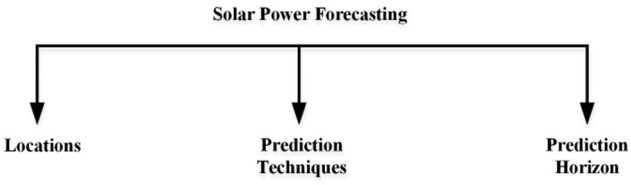

Figure 1: Division of solar power forecasting literature works

of the literary works were evaluated and their proposed methods with very fewer techniques were compared them with conventional techniques, which obliviously will underperform from advanced techniques. Stacking two or more ML techniques is an efficient way to improve the performance of any individual ML technique as it combines the virtues of two different ML techniques for solar power forecasting. However, merely combining two ML techniques without any base criterion will not give optimal solar forecasting results.

Research on solar power forecasting can be broadly categorized into three different sub-sections, as shown in Figure 1. So, different existing works on solar power prediction have focused on various aspects such as location data, prediction horizons, and different machine learning (ML) techniques. The literature series related to solar power forecasting is shown in Appendix A.

It is evident from Appendix A that the works have been done on all short, mid, and long-term prediction horizons. However, the majority of the publications have focused on short-term prediction. The reason behind this approach is driven by the policies of market operators in the energy sector. Market operators in the energy sector mostly trade for a short time or day ahead for unit commitment.

The input parameters for solar power prediction are considerably large and, are varying in different research, Appendix A. Moreover, most of the works have considered solar radiation, wind speed, and ambient temperature as prime input parameters for prediction models. Performance indices in these works are primarily root mean square error (RMSE), mean absolute error (MAE), mean square error (MSE), and mean absolute percentage error (MAPE) to evaluate the proposed models.

1.1 Research GAP and Contribution of Present Work

SPF is a key step for stable and profitable solar power plant operation in the current scenario. However, SP environmental dependency makes it a complex problem to formulate and predict. To address this issue, ML models are gaining popularity, especially for solar power and radiation prediction. The ML models are more robust and accurate than their counterparts, and hence they are the preferred choice for researchers working in the solar power prediction domain. Furthermore, it is also evident from the literature review that all the works have focused on statistical or ML techniques and their stacking, without concentrating on computational time. Moreover, standard classical methods are also crucial as, in some places, it gives more robust prediction than ML methods. So, the use of ML models without any focus on computational time will provide a skewed analysis, as it is a key parameter for the practical implementation of the model. With these views, this work further contributes to a proper insight into ML models for solar power prediction in the following ways:

(1) This work provides an exhaustive empirical analysis of sixteen conventional and contemporary ML models for SPP.
(2) In this work, conventional and contemporary ML models are evaluated for solar power forecasting for providing a multidimensional comparison. This exhaustive review with comparative analysis will guide future research in SPP design and operation.
(3) This work provides a clear insight into the computational time and K-fold cross-validation, which was grossly neglected in previous research works in SPF.
(4) This work also provides a comprehensive evaluation of ML models on five different statistical errors. The top five ML models are also evaluated based on the residual error curve and validation curves to further validate the best available regression model.
(5) Moreover, the strengths and weaknesses of different ML methods are also provided for a better understanding and adaptability of a particular method.

So, the present work can guide future research works in the optimal selection of ML models. Moreover, the test of different geographical locations can further validate the outcome of the present work.

The remaining paper is organized as follows: input data for the evaluation purpose is discussed in next section. This section also discusses the statistical learning methodology with different performance errors. Section three discusses the outcome of ML models on benchmark errors for evaluation and validation purposes. Finally, section four concludes this work.

2. MATERIAL AND METHODS

This section provides detail of data used for analysis, ML methods, and statistical errors for the comparison of ML algorithms.

2.1 Data

The prediction accuracy and performance of any prediction method largely depend on input data and the selection of appropriate ML techniques. So, the amount of dataset and its pre-processing are considered crucial factors in the success of any ML-based SPF method. This study is carried out on a publicly available dataset on PV-GIS [33]. So, input data cleaning and preparing for improving prediction is a crucial step in forecasting. The input data are acquired from PVGIS for the Ghaziabad region of India, the latitude and longitude of the location are 28.7° and 77.391°, respectively. The solar power plant output is calibrated for 10,000 kWp with a crystalline silicon panel. The database is from 1 January 2015 to 31 December 2016. The database contains seven feature vectors: solar power (watt), reflected radiation (w/m^2), direct radiation (w/m^2), diffused radiation (w/m^2), sun height (degree), temperature (degree Celsius), and wind speed at 10-meter height (m/s).

Figure 2 shows the correlation among different input variables as a heatmap. A correlation map is a statistical method to assess the correlation among feature variables. The large value of correlation shows a strong relationship and the small value reflects a low correlation or dependency among variables. PV power's highest correlation with direct and reflected radiation shows that with the variation in these radiations solar power will be affected more. However, diffused radiation and sun height also has a significantly high correlation with SP. Moreover, temperature and wind speed have a lower correlation with the target parameter, which means their variability will affect less to the solar power production.

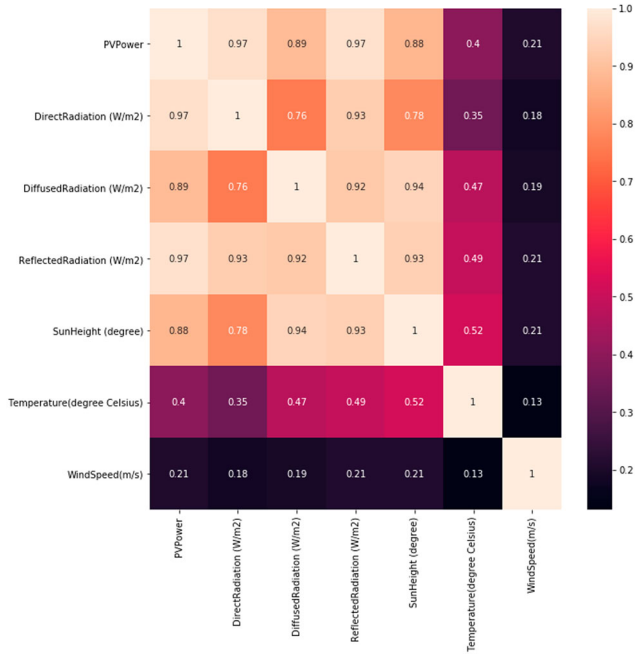

Figure 2: Correlation matrix of input parameters

2.2 Statistical Solar Power Forecasting Methods

The statistical time-series forecasting method depends on the historical datasets. It can be subdivided into classical approaches and ML approaches. Here in classical approaches ARIMA model is discussed.

2.2.1 ARIMA Method for Solar Power Forecatsing

ARIMA is a statistical way of prediction which depends on historical time-series data. Its forecasting accuracy relies on the selection of moving average (a^p) and autoregressive (b^q) parameters, in correlation dataset statistical properties. In terms of the predicted target value ($\hat{y}$), the general ARIMA representation is

$$\hat{y} = c + a^1 y_{t-1} + \ldots + a^p y_{t-p} - b^1 e_{t-1} - \ldots - b^q e_{t-q} \quad (1)$$

where e is an error with lagged values.

The pros of ARIMA lie in the requirement for datasets. However, its prediction accuracy depends on the correlation between present and past time instances, hence not preferred for long-term prediction. Moreover, the estimation of ARIMA prediction parameters is tedious work, which affects forecasting accuracy. So, due to these reasons, most research works have preferred ML methods that automatically tune the weight and biases of the model.

2.2.2 Machine Learning Methods for Solar Power Forecatsing

Machine learning is the fastest-growing computer field, and in every area of engineering and sociological decision-making, it is currently gaining popularity. In ML, computers learn from training data. With sufficiently large inputs, ML algorithms can learn and predict better than other available methods.

In analysis and prediction with ML models, a data preprocessing step is performed to clean and format the input data. Firstly, raw data are imported into the model

and they are checked for any missing values in the imported dataset. Then missing values can be replaced with the mean or median of the whole dataset. These methods of handling missing values do not affect the mean or median of the whole dataset. After managing missing values, the data are normalized or standardized. This step converts all the variables of the dataset on the same scale and reduces the variance in prediction. For the constant mean dataset, standardization will give the optimized outcome, or else normalization provides accurate results. Normalization is the most prominent step in renewable energy datasets as they have non-linear characteristics. The above steps are summarized as pre-processing steps. Now, the pre-processed data are split into training and testing parts. In the training stage, the ML models learn from training data and establish a relationship to predict the target variable. In the next step, they forecast the values on unseen data for the validation step of prediction. Approximately similar training and validation errors show the effectiveness of the ML model for prediction. The flow chart of the ML model's prediction steps is shown in Figure 3.

Regression maps input space $x^{(i)} = x_1, x_2, x_3, \ldots x_m$, to output $y^{(i)} = y_1, y_2, y_3, \ldots y_m$ through function space $f(x^{(i)})$. $F_a \ldots F_n$ are sample input features and $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ are predicted vectors by each set of input features. Equation 1 shows the predicted and input spaces.

$$\hat{y} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \\ . \\ . \\ . \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} F_a^1 & F_b^1 & . & F_n^1 \\ F_a^2 & F_b^2 & . & F_n^2 \\ . & . & . & . \\ F_a^m & F_b^m & . & F_n^m \end{bmatrix} \in \mathbb{R}^{n.m} \qquad (2)$$

Table 1 gives a comparative analysis of all the regression methods. Regularization methods come under linear regression techniques and ensemble methods include non-linear techniques. It is evident from the table that linear and regularization methods are fast but lagging in accuracy. Most real-world problems are non-linear so the ML model should adopt this non-linearity. Tree and ensemble methods have high accuracy as it sums up the decision of different trees or algorithms. So, as the model accuracy is increasing, the complexity also increases and hence the computational time. Tree-based approach for prediction is providing a better outcome than other methods as it is combining the forecast of different models.

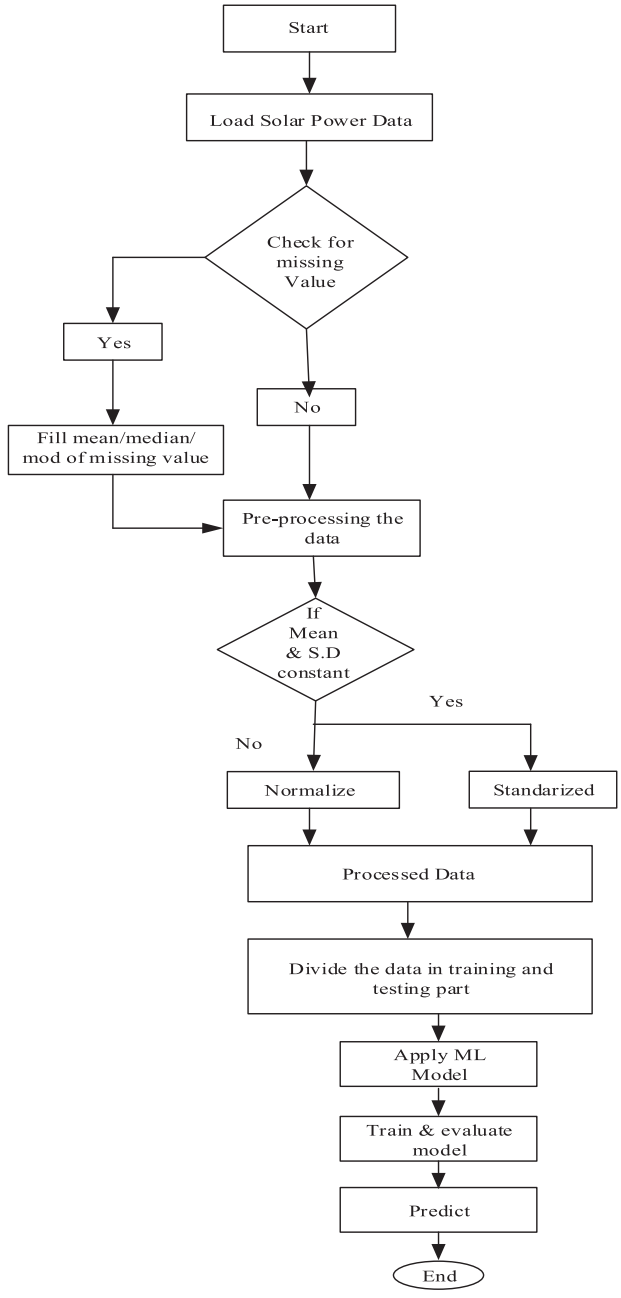

Figure 3: Flow chart of the steps involved in machine learning prediction

2.3 Statistical Errors

With a large set of available ML algorithms, selecting the most accurate model to predict solar power is the most challenging part of forecasting. It is evident from the comparison in Table 1 that targeting both bias and variance at the same time is a challenging issue for each ML technique. So, statistical errors are the most important criteria to assess the performance of any ML algorithm.

Table 1: Comparison of different ML methods

Regression techniques	Advantages/ disadvantages
Linear Regression	Advantages 1 – Implementation is easy. 2 – Training and testing are fast. Disadvantages 1 – Only applicable for linear problems. 2 – Input features should not be correlated.
k-NN	Advantages 1 – Performs well for small dimensional data. 2 – Implementation is easy. 3 – Fewer hyperparameters are required. Disadvantages 1 – The value of k is crucial. 2 – Computational requirement is high for higher dimensional data.
DT	Advantages 1 – Pre-processing of input data is not required. 2 – Gives an explainable solution. 3 – It effectively addresses the multi-collinearity problem. Disadvantages 1 – Overfits for a high number of trees. 2 – Outliers affect the solution. 3 – Complexity increases with the increase in the number of trees.
SVR	Advantages 1 – Has high accuracy. 2 – Non-linear problems can be addressed with non-linear kernel. 3 – Performs better for high-dimensional data. Disadvantages 1 – Slower training and testing. 2 – Selecting an appropriate kernel is tough. 3 – Difficult to interpret.
Regularization (Ridge, LASSO, Elastic Net)	Advantages 1 – Efficient for high-dimensional data. 2 – Computationally efficient. Disadvantages 1 – High bias error due to dimensionality reduction. 2 – Bias-variance trade-off is tough.
RF	Advantages 1 – Efficiently handles high-dimensional data. 2 – Efficiently handles missing value. 3 – New data do not affect the whole model as it has an impact on only one tree. Disadvantages 1 – Long training data. 2 – High complexity for a longer tree.
Ensemble Learning	Advantages 1 – Performance is good for high-dimensionality data. 2 – Linear and non-linear data can be handled. 3 – Less affected by noisy data. Disadvantages 1 – Has high complexity. 2 – Interpretation is tough.

So, the performance of ML models is evaluated on standard statistical indicators [34], *i.e.* mean absolute percentage error (MAPE), root mean squared error (RMSE), mean absolute error (MAE), R^2 error, and root mean squared logarithmic error (RMSLE). The mathematical formulations of discussed errors are given below:

$$\text{MAE} = \frac{\sum_{i=1}^{m} (\hat{S}^{(i)} - S^{(i)})}{m} \tag{3}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{m} (\hat{S}^{(i)} - S^{(i)})^2}{m}} \tag{4}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{m} (S^{(i)} - \hat{S}^{(i)})^2}{\sum_{i=1}^{m} (S^{(i)} - \bar{S}^{(i)})^2} \tag{5}$$

$$\text{RMSLE} = \sqrt{\frac{1}{m} \sum_{i=1}^{m} \left[\log \left(\frac{S^{(i)} + 1}{\hat{S}^{(i)} + 1} \right) \right]^2} \tag{6}$$

Here, $S_i^{(t)}$ is the actual value of solar power and $\hat{S}_i^{(t)}$ is the predicted output.

MAE shows the mean of absolute errors in prediction horizons, shown in Equation (3). R^2 is the proportion in variation from true to the predicted value is shown in Equation (5). RMSE is the squared root value of mean square error. In an unbiased estimator, RMSE just shows the square root of the variance, which is the standard deviation of predicated values, as shown in Equation (4). RMSLE estimates the log of the actual and predicted values. RMSLE doesn't penalize the huge difference and is normally taken into account when both actual and predicted values are large, as shown in Equation (6).

3. RESULTS AND DISCUSSION

Solar power is first predicted with the ARIMA model. The base parameters of the ARIMA model are evaluated with Adafuller (ADF) test and Akaike's Information Criterion (AIC), shown in Appendix B. The best model is the Best model: ARIMA (0, 1, 1) (1, 1, 0) [12] shown in Appendix C, which shows the presence of a seasonality component in the dataset, also seen in the rolling mean estimation, as shown in Figure 4.

Figure 5 shows the true solar power versus predicted solar power. The MAPE and R^2 errors of the prediction are 7.986 and 0.696, respectively.

In this part of the work, the performance evaluation of ML models has been done on the dataset for 6-hour prediction. The models are implemented on spyder notebook with python programming language. And, NumPy,

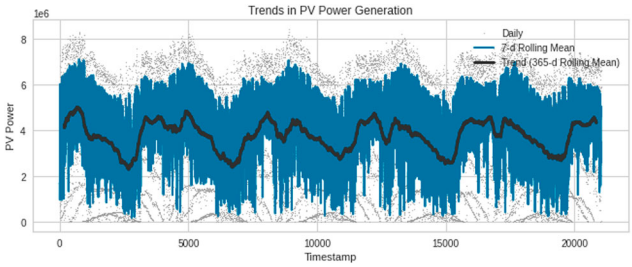

Figure 4: Seasonality trend in the dataset

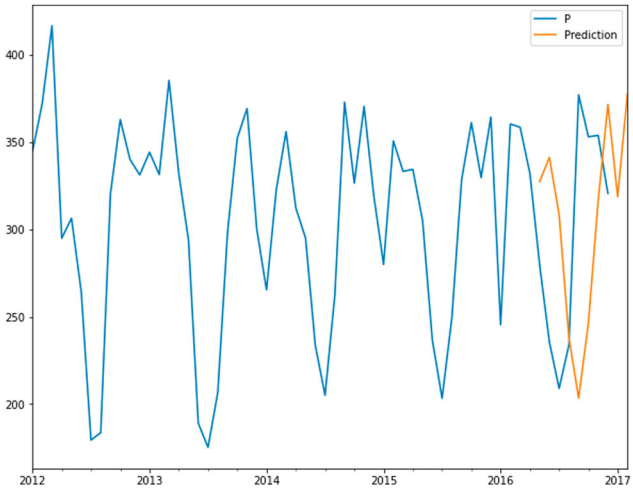

Figure 5: True vs. predicted solar power with ARIMA

pandas, and pycaret packages [35] are used to implement the solar forecasting algorithm. To evaluate the performance of ML algorithms, the models are implemented on pycaret. The models included in the analysis are extra tree (ET), Light gradient boost machine (LGBM), extreme gradient boost (XGB), random forest (RF), gradient boost regression (GBR), linear regression (LR), Bayesian ridge (BR), least absolute shrinkage and selection operator (LASSO), ridge regression (RR), decision tree (DT), k-nearest neighbor (KNN), ada boost regression (ABR), elastic net (EN), orthogonal matching pursuit (OMP), least angle regression (LAR), and state vector machine with the linear kernel (SVML). One ML model cannot outperform other models in all kinds of time series datasets. For comparing more intuitive analysis, single-step cross-validations of 5 and 10 steps are performed of the top five ML models.

Table 2 shows the performance measures of 16 ML regression models. Algorithms are compared on six different performance parameters: MAE, RMSE, R^2, RMSE, MAPE, and computational time in seconds. The table is sorted according to the MAE from a lower to a higher value. It is clear from Table 3 that ETR is outperforming other ML models in terms of MAE, RMSE, R^2, and

Table 2: Performance comparison of ML models

Model	MAE	RMSE	R^2	RMSLE	MAPE
ET	18,676	41,385	0.9997	0.0651	0.0235
LGBM	23,413	46,766	0.9996	4.454	0.0309
XGB	22,900	46,500	0.9996	3.5818	0.0256
RF	25,542	55,359	0.9995	0.0255	0.0197
GBR	43,524	79,746	0.999	5.8542	0.052
LR	60,700	84,000	0.9989	7.3134	0.1186
BR	60,088	83,576	0.9989	7.2949	0.1124
LASSO	60,700	84,000	0.9989	7.3134	0.1186
RR	60,800	84,000	0.9989	7.3207	0.1187
DT	40,600	88,600	0.9987	0.0394	0.0305
KNN	54,700	107,000	0.9981	0.631	0.1289
ABR	234,000	255,000	0.9894	8.9636	0.5745
EN	364,000	506,000	0.9584	8.7903	0.9472
OMP	322,000	557,000	0.9495	7.3893	0.4293
LAR	475,000	692,000	0.9221	8.5216	1.1188
SVML	1,790,000	3,060,000	0.5177	6.2468	0.9971

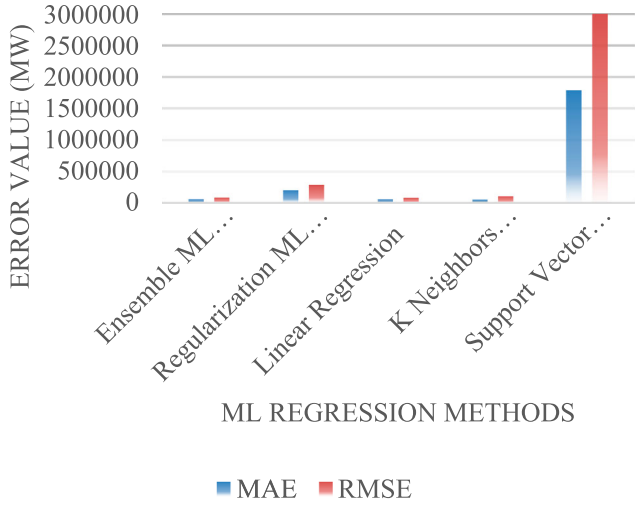

Figure 6: Error performance of different ML regression methods

MAPE with values of $1.89*10^4$ W, $4.12*10^4$ W, 0.9997, and 0.0136, respectively. The value of RMSLE is 0.1007 which is the log of actual and predicted values, so it doesn't penalize the huge difference and is normally taken when both actual and predicted values are large. Moreover, it is clearly visible that ensemble ML models are performing better than other ML models in terms of MAE, RMSE, RMLE, and MAPE. The SVML methods regression performance is last in statistical error values.

Figure 6 shows the performance of different classes of ML regression techniques in terms of MAE and RMSE. It is concluded from the figure that the average value of MAE and RMSE of ensemble techniques (ETR, LightGBM, XGboost, RF, AdaBoost, and GBR) has 61237.33 and 87019.41 W, respectively, which is better than other ML regression classes.

Ensemble regression methods are performing better in error performance, although their computational time is

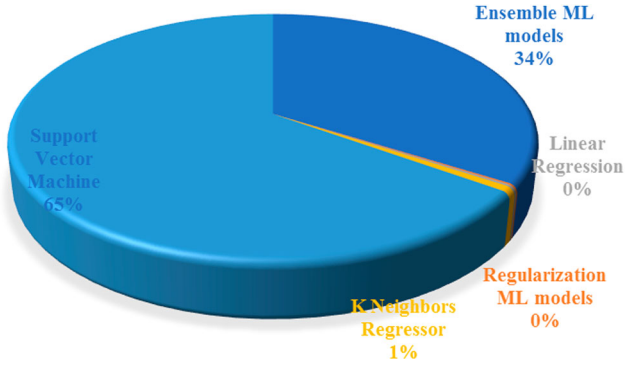

Figure 7: Computational time of different ML methods

Table 3: 5-Fold performance of the top five models

	ET		GBR		XGB
SM	Mean	SD	Mean	SD	Mean
MAE	19,833	618	44,208	1186	23,891
RMSE	42,807	1930	79,439	2448	49,053
R^2	0.9997	0	0.999	0.0001	0.9996
RMSLE	0.0823	0.0653	6.0776	0.0709	3.8788
MAPE	0.0142	0.0003	0.0538	0.0039	0.0262
	LightGBM		**RF**		**XGB**
SM	**Mean**	**SD**	**Mean**	**SD**	**SD**
MAE	23,832	913	26,724	772	960
RMSE	47,444	3003	57,572	2600	1757
R^2	0.9996	0	0.9995	0	0
RMSLE	4.4058	0.0962	0.056	0.0542	0.1185
MAPE	0.0317	0.0031	0.021	0.0008	0.0018

higher than other regression methods due to multiple tree-based decision approach for prediction, as shown in Figure 6. However, in computational time context regularization, linear regression methods are outperforming other ML methods with average performance errors depicted in Figure 7.

K-fold cross-validation is dividing the dataset into K parts for the training of the ML models. It uses each part once in the training and validation of algorithms. This cross-validation shuffles the dataset and splits it into K-groups. Each performance is retained in the individual group till the training is over. It reduces the bias of the predicting method. For a more comprehensive review, 5- and 10-fold cross-validations are compared in this work. Tables 3 and 4 show the performance of the top five ML models for 5- and 10-fold, respectively. All these top five performing models are ensemble methods and have acceptable performance metrics. However, ETR has the highest forecasting accuracy among all ensemble methods in both 5- and 10-fold cross-validation with MAE, RMSE, and MAPE of 19833.939W, 42807.125W, and 0.0142, respectively. It also has the lowest standard deviation (SD) of these errors, which reflects its robustness

Table 4: 10-Fold performance of the top five models

	ET		GBR		XGB
SM	Mean	SD	Mean	SD	Mean
MAE	18,925	1013	43,437	1710	22,860
RMSE	41,162	2508	78,744	3249	46,545
R^2	0.9997	0	0.999	0.0001	0.9996
RMSLE	0.1007	0.0768	5.9577	0.1539	3.5818
MAPE	0.0136	0.0008	0.0533	0.0066	0.0256
	LightGBM		**RF**		**XGB**
SM	**Mean**	**SD**	**Mean**	**SD**	**SD**
MAE	23,184	1427	25,205	1117	1164
RMSE	45,962	3784	54,425	3189	2677
R^2	0.9997	0.0001	0.9995	0.0001	0
RMSLE	4.4607	0.0647	0.041	0.0314	0.404
MAPE	0.033	0.0112	0.0196	0.0013	0.0037

and precision in prediction. Moreover, ML models performance ar ealso evaluated on manual division of dataset in three equal parts, for the purpose of training and validation. The outcome also supports the K-fold approche result, shown in Appendix D.

In the regression method of prediction, there are multiple curves of performance to judge the virtue of prediction. In this work, residual error plots and validation score curves are used to estimate the performance of the top five performing ML models.

Residuals of regression show the vertical distance of the predicted value from the regression fit line. It clearly reflects the error between the observed and forecasted values. A good residual value of any ML model should be normally distributed and independent. Moreover, the density of points should lie close to the origin, and should be symmetric or normally distributed across the origin.

The residual variation of the R^2 value for training data of ensemble methods varies from 1 to 0.979; for the test, it varies from 0.974 to 0.977. The R^2 values show the goodness of these top five ensemble regression techniques. However, ETR regression has the highest training R^2 value of 1 and the lowest testing R^2 value of 0.974. The low test R^2 value is, however, only 0.3% lower than the highest value, which can be considered approximately comparable to other ensemble methods. Although ETR is outperforming other ensemble methods, RF is also giving promising outcomes with R^2 values of training and testing being 0.997 and 0.975, respectively, and also normally distributes errors across the origin. The other three models, *i.e.* XGBoost, LightGBM, and Gradient boost have close R^2 values but errors are not normally distributed. It reflects that the remaining three models are not as good as ETR and RFR in capturing the prediction variable information.

Table 5: Summary of hyperparameter tuning of the top-performing ML models

Model	H^t	MAE	RMSE	R^2	MAPE
ET	T	31,319	64,694	0.9993	0.0235
	D	**18,676**	**41,385**	**0.9997**	**0.0133**
GBR	T	**21,556**	**42,530**	**0.9997**	**0.0311**
	D	43,524	79,746	0.999	0.052
Light GBM	T	**23,007**	**45,531**	**0.9997**	**0.0276**
	D	23,413	46,766	0.9996	0.0309
RF	T	33,136	69,092	0.9992	0.0303
	D	**25,542**	**55,359**	**0.9995**	**0.0197**
BR	T	179,6220	179,8170	0.4754	4.2282
	D	**60,088**	**83,576**	**0.9989**	**0.1124**

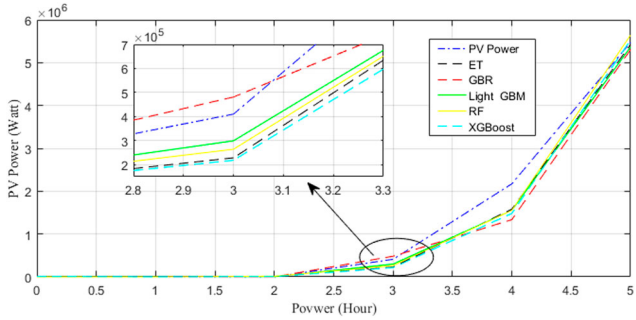

Figure 8: PV power prediction of 6 h.

The ML model's performance depends on the selection of model parameters and hyperparameters. The choice of hyperparameter values directs the training process on the dataset. So, it affects the convergence of the ML model toward global value. Every ML model has a different hyperparameter value whose selection and value can be diverse according to the characteristics of the datasets. The models' hyperparameters have been evaluated with the PyCaret package for this study. The random search CV approach has been adapted to search default combinations of hyperparameters provided by the sci-kit-learn library. The motivation for selecting the default sci-kit-learn search space is that it allows for it to have a considerable value of hyperparameter latent space.

Table 5 depicts the result of hyperparameter tuning of most of the ensemble methods. Only ensemble methods are selected for hyperparameter tuning due to their superiority over other ML models. The hyperparameter tuning improves the results of GBR and LightGBM by 50.47% and 1.75% in terms of the loss function. $\mathbf{H^t}$ shows hyperparameter tuning, where T stands for tuned model output and D stands for default parameter.

The training and validation scores of extra tree regression and random forest regression are approximately 0.98 with very less difference between both the scores. This shows the models are efficiently capturing the input data trend. XGBoost regression has the highest training score but the difference between training and validation scores is large which shows the overfitting of the model. And, the same training and validation patterns are in LightGBM and gradient boosting regression, which shows slight overfitting, as the difference between both the scores is not large.

These models are tested on short-time forecasting for a 6-hour period. All the top five models are providing good predictions of solar power. However, GBR is providing better output in the first part of the prediction as the output increases from zero, and in the exponential region, ET is forecasting better than other models, as shown in Figure 8.

4. CONCLUSION

The selection of an accurate machine learning model for solar power forecasting is a challenging issue, and numerous approaches are investigated in different works involving different ML models. So, this work provides an in-depth review and empirical analysis of solar power forecasting techniques, which includes ARIMA-based classical time-series methods to 16 different ML models. ARIMA-based solar power prediction models require comparatively fewer data points. However, their inability in long-term predictions and parameter estimation for accurate prediction limits their capability.

Although ML models are already used for solar power prediction, several advanced ML techniques are still not investigated and compared in one work, which is also analyzed in this work for a more comprehensive review. The performance evaluation of the ML models shows that the ensemble models have better forecasting accuracy than other ML models. The multiple training across nodes and combining the prediction of individual nodes gives an edge to ensemble methods over other ML models. Ensemble ML models also have the lowest mean and standard deviation in prediction errors for 5- and 10-fold cross-validation, which shows lower variation in prediction. Among ensemble methods, extra tree regression has outperformed other models. However, the computational time is a crucial constraint in the real-time implementation of ML models. Regularization ML models have the lowest computational time among all ML models. Ensemble models give accurate results in forecasting but have high computational time.

The practical approach of this work provides a systematic review of future studies and opportunities for stakeholders to select an accurate ML model for solar power prediction.

DISCLOSURE STATEMENT

No potential conflict of interest was reported by the author(s).

ORCID

Ashish Shrivastava http://orcid.org/0000-0002-2465-0708

REFERENCES

1. IEA, International Renewable Energy Agency, United Nations Statistics Division, The World Bank, and World Health Organization, "The energy progress report," *IEA, IRENA, UNSD, WB, WHO (2019), Track. SDG 7 Energy Progress Report. 2019, Washingt. DC*, 2019.

2. S. F. Singer, "World energy outlook," Symposium Papers – Energy Model., 567–75, 1982.

3. Z. Şen, "Solar energy in progress and future research trends," *Prog. Energy Combust. Sci.*, Vol. 30, no. 4, pp. 367–416, 2004. DOI: 10.1016/j.pecs.2004.02.004.

4. BP, "Energy outlook 2022 edition 2022 explores the key uncertainties surrounding the energy transition," 2022, p. 109.

5. IRENA, *Renewable Capacity Statistics 2019*. Abu Dhabi: International Renewable Energy Agency (IRENA), 2019. Available: https://www.irena.org/publications/2019/Mar/Renewable-Capacity-Statistics-2019.

6. IRENA, *International Renewable Energy Agency. Renewable Power Generation Costs in 2017*. 2018. Available: https://www.irena.org/publications/2018/Jan/Renewable-power-generation-costs-in-2017.

7. S. B. Nam, and J. Hur, "A hybrid spatio-temporal forecasting of solar generating resources for grid integration," *Energy*, Vol. 177, pp. 503–10, 2019. DOI: 10.1016/j.energy.2019.04.127.

8. V. Bagalini, B. Y. Zhao, R. Z. Wang, and U. Desideri, "Solar PV-battery-electric grid-based energy system for residential applications: system configuration and viability," *Research*, Vol. 2019, pp. 1–17, 2019. DOI: 10.34133/2019/3838603.

9. N. Dong, J. F. Chang, A. G. Wu, and Z. K. Gao, "A novel convolutional neural network framework based solar irradiance prediction method," *Int. J. Electr. Power Energy Syst.*, Vol. 114, no. July 2019, pp. 105411, 2020. DOI: 10.1016/j.ijepes.2019.105411.

10. A. E. Gürel, Ü Ağbulut, and Y. Biçen, "Assessment of machine learning, time series, response surface methodology and empirical models in prediction of global solar radiation," *J. Clean. Prod.*, Vol. 277, pp. 122353, 2020. DOI: 10.1016/j.jclepro.2020.122353.

11. K. Bakker, K. Whan, W. Knap, and M. Schmeits, "Comparison of statistical post-processing methods for probabilistic NWP forecasts of solar radiation," *Sol. Energy*, Vol. 191, no. August, pp. 138–50, 2019. DOI: 10.1016/j.solener.2019.08.044.

12. Y. Hao, and C. Tian, "A novel two-stage forecasting model based on error factor and ensemble method for multi-step wind power forecasting," *Appl. Energy*, Vol. 238, no. July 2018, pp. 368–83, 2019. DOI: 10.1016/j.apenergy.2019.01.063.

13. U. K. Das, *et al.*, "Forecasting of photovoltaic power generation and model optimization: A review," *Renew. Sustain. Energy Rev.*, Vol. 81, no. June 2017, pp. 912–28, 2018. DOI: 10.1016/j.rser.2017.08.017.

14. A. Ahmed, and M. Khalid, "A review on the selected applications of forecasting models in renewable power systems," *Renew. Sustain. Energy Rev.*, Vol. 100, no. September 2018, pp. 9–21, 2019. DOI: 10.1016/j.rser.2018.09.046.

15. S. Ferrari, M. Lazzaroni, V. Piuri, L. Cristaldi, and M. Faifer, "Statistical models approach for solar radiation prediction," in *Conf. Rec. – IEEE Instrumentation Measurement Technology Conference*, 2013, pp. 1734–9. DOI: 10.1109/I2MTC.2013.6555712.

16. G. Reikard, "Predicting solar radiation at high resolutions: A comparison of time series forecasts," *Sol. Energy*, Vol. 83, no. 3, pp. 342–9, 2009. DOI: 10.1016/j.solener.2008.08.007.

17. R. Huang, T. Huang, R. Gadh, and N. Li, "Solar generation prediction using the ARMA model in a laboratory-level micro-grid," in *2012 IEEE 3rd International Conference on Smart Grid Communication SmartGridComm 2012*, 2012, pp. 528–33. DOI: 10.1109/SmartGridComm.2012.6486039.

18. M. David, F. Ramahatana, P. J. Trombe, and P. Lauret, "Probabilistic forecasting of the solar irradiance with recursive ARMA and GARCH models," *Sol. Energy*, Vol. 133, pp. 55–72, 2016. DOI: 10.1016/j.solener.2016.03.064.

19. M. Bouzerdoum, A. Mellit, and A. Massi Pavan, "A hybrid model (SARIMA-SVM) for short-term power forecasting of a small-scale grid-connected photovoltaic plant," *Sol. Energy*, Vol. 98, no. PC, pp. 226–35, 2013. DOI: 10.1016/j.solener.2013.10.002.

20. D. Yang, P. Jirutitijaroen, and W. M. Walsh, "Hourly solar irradiance time series forecasting using cloud cover index," *Sol. Energy*, Vol. 86, no. 12, pp. 3531–43, 2012. DOI: 10.1016/j.solener.2012.07.029.

21. M. Abuella, and B. Chowdhury, "Solar power probabilistic forecasting by using multiple linear regression

analysis," in *Conference proceedings – IEEE SOUTHEASTCON*, vol. 2015-June, no. June, 2015, pp. 5–9, DOI: 10.1109/SECON.2015.7132869.

22. H. Sheng, J. Xiao, Y. Cheng, Q. Ni, and S. Wang, "Short-term solar power forecasting based on weighted Gaussian process regression," *IEEE Trans. Ind. Electron.*, Vol. 65, no. 1, pp. 300–8, 2018. DOI: 10.1109/TIE.2017.2714127.

23. C. Persson, P. Bacher, T. Shiga, and H. Madsen, "Multi-site solar power forecasting using gradient boosted regression trees," *Sol. Energy*, Vol. 150, pp. 423–36, 2017. DOI: 10.1016/j.solener.2017.04.066.

24. M. Zamo, O. Mestre, P. Arbogast, and O. Pannekoucke, "A benchmark of statistical regression methods for short-term forecasting of photovoltaic electricity production. Part II: Probabilistic forecast of daily production," *Sol. Energy*, Vol. 105, pp. 804–16, 2014. DOI: 10.1016/j.solener.2014.03.026.

25. G. Wang, Y. Su, and L. Shu, "One-day-ahead daily power forecasting of photovoltaic systems based on partial functional linear regression models," *Renew. Energy*, Vol. 96, pp. 469–78, 2016. DOI: 10.1016/j.renene.2016.04.089.

26. R. Amaro e Silva, and M. C. Brito, "Impact of network layout and time resolution on spatio-temporal solar forecasting," *Sol. Energy*, Vol. 163, no. November 2017, pp. 329–37, 2018. DOI: 10.1016/j.solener.2018.01.095.

27. N. Tang, S. Mao, Y. Wang, and R. M. Nelms, "Solar power generation forecasting With a LASSO-based approach," *IEEE Internet Things J.*, Vol. 5, no. 2, pp. 1090–9, Apr. 2018. DOI: 10.1109/JIOT.2018.2812155.

28. T. C. Carneiro, P. A. C. Rocha, P. C. M. Carvalho, and L. M. Fernández-Ramírez, "Ridge regression ensemble of machine learning models applied to solar and wind forecasting in Brazil and Spain," *Appl. Energy*, Vol. 314, pp. 118936, May 2022. DOI: 10.1016/j.apenergy.2022.118936.

29. D. Nikodinoska, M. Käso, and F. Müsgens, "Solar and wind power generation forecasts using elastic net in time-varying forecast combinations," *Appl. Energy*, Vol. 306, pp. 117983, Jan. 2022. DOI: 10.1016/j.apenergy.2021.117983.

30. M. H. D. M. Ribeiro, and L. dos Santos Coelho, "Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series," *Appl. Soft Comput. J.*, Vol. 86, pp. 105837, 2020. DOI: 10.1016/j.asoc.2019.105837.

31. P. Ray, R. K. Lenka, and M. Biswal, "Frequency mode identification using modified masking signal-based empirical mode decomposition," *IET Gener. Transm. Distrib.*, Vol. 13, no. 8, pp. 1266–76, Apr. 2019. DOI: 10.1049/iet-gtd.2018.5527.

32. M. Babu, and P. Ray, "A review on energy forecasting algorithms crucial for energy industry development and policy design," *Energy Sources Part A Recover. Util. Environ. Eff.*, 1–24, Nov. 2021. DOI: 10.1080/15567036.2021.2006370.

33. T. a Huld, M. Súri, E. D. Dunlop, M. Albuisson, and L. Wald, "Integration of HELIOCLIM-1 database into PV-GIS to estimate solar electricity potential in Africa," in *20th European Photovoltaic Solar Energy Conference Exhibition*, 2005, pp. 2989. Available: c:/pdflib/00018470.pdf

34. M. Despotovic, V. Nedic, D. Despotovic, and S. Cvetanovic, "Review and statistical analysis of different global solar radiation sunshine models," *Renew. Sustain. Energy Rev.*, Vol. 52, pp. 1869–80, 2015. DOI: 10.1016/j.rser.2015.08.035.

35. M. Ali, "PyCaret." 2020. Available: https://pycaret.org/.

36. J. Huang, M. Korolkiewicz, M. Agrawal, and J. Boland, "Forecasting solar radiation on an hourly time scale using a coupled autoregressive and dynamical system (CARDS) model," *Sol. Energy*, Vol. 87, no. 1, pp. 136–49, 2013. DOI: 10.1016/j.solener.2012.10.012.

37. C. W. Chow, B. Urquhart, M. Lave, A. Dominguez, J. Kleissl, J. Shields, and B. Washom, "Intra-hour forecasting with a total sky imager at the UC San Diego solar energy testbed," *Sol. Energy*, Vol. 85, no. 11, pp. 2881–93, 2011. DOI: 10.1016/j.solener.2011.08.025.

38. J. Zeng, and W. Qiao, "Short-term solar power prediction using a support vector machine," *Renew. Energy*, Vol. 52, pp. 118–27, 2013. DOI: 10.1016/j.renene.2012.10.009.

39. C. Yang, and L. Xie, "A novel ARX-based multi-scale spatio-temporal solar power forecast model," in *2012 North America Power Symposium NAPS 2012*, 2012, DOI: 10.1109/NAPS.2012.6336383.

40. M. De Felice, M. Petitta, and P. M. Ruti, "Short-term predictability of photovoltaic production over Italy," *Renew. Energy*, Vol. 80, pp. 197–204, 2015. DOI: 10.1016/j.renene.2015.02.010.

41. M. G. De Giorgi, M. Malvoni, and P. M. Congedo, "Comparison of strategies for multi-step ahead photovoltaic power forecasting models based on hybrid group method of data handling networks and least square support vector machine," *Energy*, Vol. 107, pp. 360–73, 2016. DOI: 10.1016/j.energy.2016.04.020.

42. G. I. Nagy, G. Barta, S. Kazi, G. Borbély, and G. Simon, "GEFCom2014: probabilistic solar and wind power forecasting using a generalized additive tree ensemble approach," *Int. J. Forecast.*, Vol. 32, no. 3, pp. 1087–93, 2016. DOI: 10.1016/j.ijforecast.2015.11.013.

43. R. Meenal, and A. Immanuel Selvakumar, *Assessment of Solar Energy Potential of Smart Cities of Tamil Nadu Using Machine Learning with Big Data*, vol. 750. Singapore: Springer, 2019. DOI: 10.1007/978-981-13-1882-5_3.

44. M. A. M. Ramli, S. Twaha, and Y. A. Al-Turki, "Investigating the performance of support vector machine and artificial neural networks in predicting solar radiation

on a tilted surface: Saudi Arabia case study," *Energy Convers. Manag.*, Vol. 105, pp. 442–52, 2015. DOI: 10.1016/j.enconman.2015.07.083.

45. B. Wolff, J. Kühnert, E. Lorenz, O. Kramer, and D. Heinemann, "Comparing support vector regression for PV power forecasting to a physical modeling approach using measurement, numerical weather prediction, and cloud motion data," *Sol. Energy*, Vol. 135, pp. 197–208, 2016. DOI: 10.1016/j.solener.2016.05.051.

46. J. Alonso-Montesinos, F. J. Batlles, and C. Portillo, "Solar irradiance forecasting at one-minute intervals for different sky conditions using sky camera images," *Energy Convers. Manag.*, Vol. 105, pp. 1166–77, 2015. DOI: 10.1016/j.enconman.2015.09.001.

47. J. Antonanzas, N. Osorio, R. Escobar, R. Urraca, F. J. Martinez-de-pison, and F. Antonanzas-torres, "Review of photovoltaic power forecasting," *Sol. Energy*, Vol. 136, pp. 78–111, 2016. DOI: 10.1016/j.solener.2016.06.069.

48. A. T. Eseye, J. Zhang, and D. Zheng, "Short-term photovoltaic solar power forecasting using a hybrid wavelet-PSO-SVM model based on SCADA and meteorological information," *Renew. Energy*, Vol. 118, pp. 357–67, 2018. DOI: 10.1016/j.renene.2017.11.011.

49. Y. Feng, D. Gong, Q. Zhang, S. Jiang, L. Zhao, and N. Cui, "Evaluation of temperature-based machine learning and empirical models for predicting daily global solar radiation," *Energy Convers. Manag.*, Vol. 198, no. July, pp. 111780, 2019. DOI: 10.1016/j.enconman.2019.111780.

50. A. Javed, B. K. Kasi, and F. A. Khan, "Predicting solar irradiance using machine learning techniques," in *2019 15th International Wireless Communication Mobile Computing Conference IWCMC 2019*. 2019, pp. 1458–62. DOI: 10.1109/IWCMC.2019.8766480.

51. H. Zang, *et al.*, "Hybrid method for short-term photovoltaic power forecasting based on deep convolutional neural network," *IET Gener. Transm. Distrib.*, Vol. 12, no. 20, pp. 4557–67, 2018. DOI: 10.1049/iet-gtd.2018.5847.

52. W. Yin, Y. Han, H. Zhou, M. Ma, L. Li, and H. Zhu, "A novel non-iterative correction method for short-term photovoltaic power forecasting," *Renew. Energy*, Vol. 159, pp. 23–32, 2020. DOI: 10.1016/j.renene.2020.05.134.

53. R. Zhu, W. Guo, and X. Gong, "Short-term photovoltaic power output prediction based on k-fold cross-validation and an ensemble model," *Energies*, Vol. 12, no. 7, pp. 1220, 2019. DOI: 10.3390/en12071220.

54. K. Doubleday, S. Jascourt, W. Kleiber, and B.-M. Hodge, "Probabilistic solar power forecasting using Bayesian model averaging," *IEEE Trans. Sustain. Energy*, Vol. 99, no. 99, pp. 1, 2020. DOI: 10.1109/tste.2020.2993524.

55. U. Munawar, and Z. Wang, "A framework of using machine learning approaches for short-term solar power forecasting," *J. Electr. Eng. Technol.*, Vol. 15, no. 2, pp. 561–9, 2020. DOI: 10.1007/s42835-020-00346-4.

56. W. VanDeventer, *et al.*, "Short-term PV power forecasting using hybrid GASVM technique," *Renew. Energy*, Vol. 140, pp. 367–79, 2019. DOI: 10.1016/j.renene.2019.02.087.

57. M. Rana, and A. Rahman, "Multiple steps ahead solar photovoltaic power forecasting based on univariate machine learning models and data re-sampling," *Sustain. Energy, Grids Networks*, Vol. 21, pp. 100286, 2020. DOI: 10.1016/j.segan.2019.100286.

58. D. Niu, K. Wang, L. Sun, J. Wu, and X. Xu, "Short-term photovoltaic power generation forecasting based on random forest feature selection and CEEMD: A case study," *Appl. Soft Comput. J.*, Vol. 93, pp. 106389, 2020. DOI: 10.1016/j.asoc.2020.106389.

59. Z. F. Liu, L. L. Li, M. L. Tseng, and M. K. Lim, "Prediction short-term photovoltaic power using improved chicken swarm optimizer – extreme learning machine model," *J. Clean. Prod.*, Vol. 248, pp. 119272, 2020. DOI: 10.1016/j.jclepro.2019.119272.

APPENDIX A

Sr. No.	Location	Input variable	Target variable	Time horizon	Forecast method	Statistical errors
[21]	U.S.A	SR, W, T, H, P, Pr	SP	24 h	Multiple linear regression	RMSE
[22]	Singapore	SR, W, T, H, Pr	SP	5 min	Weighted Gaussian regression	MSE, RMSE, MPIW
[23]	Japan	W, T, H, Pr	SP	1–6 h	Gradient Boost regression	RMSE
[24]	France	SP	SP	2 days	Statistical regression	Histogram
[20]	U.S.A	SR, Cloud Cover	SR	1 h	Autoregressive integral moving average	RMSE, MBE
[36]	Australia	–	SR	1 h	Dynamic auto-regression	MBE, nRMSE, MeAPE, KSI
[37]	U.S.A	SR, Sky Image	SR	30 sec to 5 min	Sky imager	Cap error
[18]	Locations of U.S.A	SR, Climate type	SR	60 min	ARMA generalized auto-conditional heteroscedasticity	RMSE, MAE, MBE
[25]	Macau	SR, W, T, P, H, Pr	SP	1 d	Partial function linear model	RMSE, MAPE, MAD
[26]	U.S.A	SR, SP	SR	1 month	Autoregressive model with exogenous input	Weighted average distance
[19]	Italy	SR, T	SP	1 h	SARIMA-SVM	nRMSE, nMBE, nMPE, R^2
[17]	U.S.A	GHI, DNI, DHI	SP	24 h	ARMA	MSE, MAE
[15]	Italy	SR, Seasonality	SR	1 Week	Statistical model	Absolute error
[16]	U.S.A	SR, H, Cloud Cover	SR	5 min to 30 min	ARIMA	Statistical error
[38]	China	SP, Cloud condition	SP	5 d	SVM with weather classification	RMSE, MRE
[39]	U.S.A	GHI	SR	5–15 min, 1–24 h	Autoregressive ARX with exogenous input	RMSE, MAE
[40]	Italy	GHI, T	SP	1–10 days	SVR	RMSE, MdAPE
[41]	Italy	SP	SP	1, 3, 6, 12, 24 h	LS-SVR	nMAE, nMBE
[42]	Budapest	SR, W, H, P, Cloud	SP	24 h	QRF and RF ensemble	Weighted pinball loss
[43]	Tamil Nadu India	SR, T	SR	24 h	RF	MAE, R^2, RMSE
[44]	Saudi Arabia	GHI, DNI	SR	–	SVM	RMSE, MRE, Computational speed
[45]	Germany	SP, SR	SP	15 min and 5 h	SVR	RMSE, Bias
[46]	Spain	Sky image	SR	1, 15 min	Sky camera imager	RMSE
[47]	Spain	GHI, W, T, Sun position	SP	1 d	Single and stacked model: SVR, DNN, XGB, RF	RMSE, MAE, MBE
[48]	China	Metrological input	SP	24 h	Wavelet-PSO-SVM	nMAE, MAPE, SSE, SDE
[49]	China	GHI, T, Climatic variables	SR	–	ANN, WNN, RF, and empirical models	RMSE, MAE, MBE, R^2
[50]	Pakistan	SR, W, T, P, H	SR	–	LR, Regression Tree, SVM	R^2
[51]	China	SP, metrological input	SP	70 h	Variational mode decomposition-CNN	MAE, RMSE
[52]	China	NWP	SP	12 h	Correlation models: BP, ELM, SVM	MAE, RMSE
[53]	China	Solar Panel data, SP, W, T	SP	24 h	Ensemble model	RMSE, MAE, MAPE
[54]	U.S.A	Metrological input	SP	24 h	Bayesian model averaging	Skill score
[55]	U.S.A	SR, W, T, P, sunrise time	SP	24 h	XgBoost	RMSE
[56]	Australia	SP, SR, T, Solar panel parameters	SP	5 d	GA-SVM	RMSE, MAPE, Accuracy
[57]	Australia	SP	SP	3 h	Re-sampled ML model	MAE, MRE, Skill score
[58]	China	SR, T, Cloud cover	SP	15 min	RF+ Ensembled decomposition	MAPE, MAE, R^2
[59]	China	SP, Weather data	SP	600 min	Chicken swarm optimization + RF	RMSE, MAPE, R^2

APPENDIX B. ADAFULLER TEST RESULT

ADF statistic:	−11.30061
n_lags:	1.31E-20
p-value:	1.31E-20
Critical values:	
1%,	−3.4306615
5%,	−2.8616777
10%,	−2.5668433

APPENDIX C. PERFORMING STEPWISE SEARCH TO MINIMIZE AIC

Performing stepwise search to minimize AIC		
ARIMA(0,1,0)(1,1,1)[12]	AIC = 491.217,	Time = 0.25 sec
ARIMA(0,1,0)(0,1,0)[12]	AIC = 494.629,	Time = 0.02 sec
ARIMA(1,1,0)(1,1,0)[12]	AIC = 476.326,	Time = 0.12 sec
ARIMA(0,1,1)(0,1,1)[12]	AIC = 465.324,	Time = 0.21 sec
ARIMA(0,1,1)(0,1,0)[12]	AIC = 468.646,	Time = 0.04 sec
ARIMA(0,1,1)(1,1,1)[12]	AIC = 467.157,	Time = 0.30 sec
ARIMA(0,1,1)(0,1,2)[12]	AIC = 467.160,	Time = 0.56 sec
ARIMA(0,1,1)(1,1,0)[12]	AIC = 465.232,	Time = 0.17 sec
ARIMA(0,1,1)(2,1,0)[12]	AIC = 467.158,	Time = 0.65 sec
ARIMA(0,1,1)(2,1,1)[12]	AIC = 469.157,	Time = 1.18 sec
ARIMA(0,1,0)(1,1,0)[12]	AIC = 489.240,	Time = 0.08 sec
ARIMA(1,1,1)(1,1,0)[12]	AIC = 467.231,	Time = 0.35 sec
ARIMA(0,1,2)(1,1,0)[12]	AIC = 467.231,	Time = 0.37 sec
ARIMA(1,1,2)(1,1,0)[12]	AIC = 469.171,	Time = 0.50 sec
ARIMA(0,1,1)(1,1,0)[12] intercept	AIC = inf, Tim	e = 0.42 sec
Best model: ARIMA(0,1,1)(1,1,0)[12]		

APPENDIX D. PART 1 DIVISION DATA PERFORMANCE

Model	MAE	MSE	RMSE	R^2	RMSLE	MAPE
ET	3.98E+04	3.84E+09	6.18E+04	0.9993	0.0225	0.0133
LIGHTGBM	4.56E+04	4.31E+09	6.55E+04	0.9993	0.0846	0.033
RF	5.29E+04	6.63E+09	8.13E+04	0.9989	0.0339	0.0193
LASSO	7.88E+04	1.03E+10	1.01E+05	0.9982	0.2989	0.1291
LR	7.88E+04	1.03E+10	1.01E+05	0.9982	0.2988	0.1291
BR	7.88E+04	1.03E+10	1.01E+05	0.9982	0.2983	0.1291
RIDGE	7.89E+04	1.03E+10	1.01E+05	0.9982	0.2949	0.1303
GBR	8.24E+04	1.29E+10	1.13E+05	0.9978	0.1667	0.0815
DT	8.32E+04	1.65E+10	1.28E+05	0.9972	0.0561	0.03
KNN	1.14E+05	2.49E+10	1.58E+05	0.9957	0.2166	0.1221
ADA	2.39E+05	7.97E+10	2.82E+05	0.9863	0.4901	0.5974
PAR	3.61E+05	2.92E+11	5.40E+05	0.9499	0.403	0.245
EN	4.59E+05	3.14E+11	5.61E+05	0.9461	0.6484	0.9708
OMP	5.35E+05	4.89E+11	6.99E+05	0.916	0.7904	1.6405
LAR	7.58E+05	8.88E+11	9.42E+05	0.8475	0.7489	0.9444
SVML	8.62E+05	9.24E+11	9.89E+04	0.7432	0.8431	1.3244

APPENDIX E. PART 2 DIVISION DATA PERFORMANCE

Model	MAE	MSE	RMSE	R^2	RMSLE	MAPE
ET	3.88E+04	3.65E+09	6.02E+04	0.9994	0.0248	0.0135
LIGHTGBM	4.53E+04	4.11E+09	6.40E+04	0.9993	0.0939	0.0384
RF	5.29E+04	6.29E+09	7.92E+04	0.9989	0.0379	0.0205
GBR	7.43E+04	1.01E+10	1.00E+05	0.9983	0.1654	0.0863
LASSO	8.10E+04	1.05E+10	1.02E+05	0.9982	0.3112	0.1924
LR	8.10E+04	1.05E+10	1.02E+05	0.9982	0.3112	0.1924
BR	8.10E+04	1.05E+10	1.02E+05	0.9982	0.3112	0.1924
RIDGE	8.10E+04	1.05E+10	1.02E+05	0.9982	0.311	0.1935
DT	8.64E+04	1.77E+10	1.33E+05	0.997	0.0591	0.0318
KNN	1.23E+05	2.74E+10	1.65E+05	0.9953	0.2337	0.1478
ADA	2.28E+05	7.38E+10	2.71E+05	0.9874	0.5129	0.7544
EN	4.88E+05	3.32E+11	5.76E+05	0.9432	0.6663	1.1396
PAR	4.40E+05	3.64E+11	6.03E+05	0.9377	0.4315	0.327
OMP	4.81E+05	4.14E+11	6.43E+05	0.9293	0.7709	1.8228
LAR	6.95E+05	7.64E+11	8.74E+05	0.8694	0.6987	1.015
SVML	7.23E+05	8.41E+11	9.44E+05	0.7321	0.7821	1.328

APPENDIX F. PART 3 DIVISION DATA PERFORMANCE

Model	MAE	MSE	RMSE	R^2	RMSLE	MAPE
ET	4.04E+04	3.75E+09	6.11E+04	0.9993	0.022	0.0135
LIGHTGBM	4.50E+04	4.01E+09	6.33E+04	0.9993	0.1042	0.0394
RF	5.24E+04	5.98E+09	7.72E+04	0.9989	0.0331	0.019
LASSO	7.60E+04	9.05E+09	9.51E+04	0.9983	0.2945	0.1582
BR	7.60E+04	9.05E+09	9.51E+04	0.9983	0.2945	0.1582
LR	7.60E+04	9.05E+09	9.51E+04	0.9983	0.2945	0.1582
RIDGE	7.61E+04	9.05E+09	9.51E+04	0.9983	0.2976	0.1589
GBR	7.82E+04	1.13E+10	1.06E+05	0.9979	0.1933	0.1049
DT	8.12E+04	1.48E+10	1.21E+05	0.9973	0.049	0.0287
KNN	1.17E+05	2.54E+10	1.59E+05	0.9953	0.2502	0.166
ADA	2.24E+05	7.03E+10	2.65E+05	0.987	0.523	0.8548
PAR	3.47E+05	2.50E+11	4.99E+05	0.9539	0.3858	0.2529
EN	4.53E+05	2.96E+11	5.43E+05	0.9455	0.664	1.2482
OMP	5.04E+05	4.39E+11	6.62E+05	0.919	0.7774	2.0388
LAR	7.31E+05	8.24E+11	9.07E+05	0.8474	0.7325	1.1339
SVML	8.43E+05	9.32E+11	9.07E+05	0.7861	0.8123	1.291

APPENDIX G. NMAE AND NRMSE OF THE ML MODELS

Models	nMAE	nRMSE
ET	0.0081	0.0176171
LIGHTGBM	0.0099227	0.0196713
XGBOOST	0.0097841	0.0199209
RF	0.0107877	0.0232938
GBR	0.018591	0.0337018
LR	0.0259694	0.0359544
BR	0.0259697	0.0359544
LASSO	0.0259694	0.0359544
RIDGE	0.026004	0.0359581
DT	0.017367	0.037932
KNN	0.0233988	0.0459221
ADA	0.1000686	0.1092561
EN	0.155992	0.2166242
OMP	0.1378107	0.2385287
LAR	0.2032593	0.2961162
SVM	0.7649063	1.3079442

AUTHORS

Amit Rai received his BTech degree in electronics engineering from UP Technical University, Uttar Pradesh, India in 2006 and his MTech degree in electronics engineering from HBTI Kanpur, India in 2009. He is currently pursuing PhD at the Department of Electrical Engineering, Indian Institute of Technology (ISM), Dhanbad, India.His areas of interest are renewable energy, deep learning, and machine learning.

Email: rai.amit21@gmail.com

Ashish Shrivastava received his BE degree in electrical engineering from Government Engineering College Rewa, India, in 1999 and his MTech degree in hydroelectric engineering from Maulana Azad National Institute of Technology (NIT), Bhopal, India, in 2001. He obtained his PhD in electrical engineering from the Indian Institute of Technology Delhi (IITD), New Delhi, India, in 2013. His areas of interest are power electronics, power quality, solar PV, electric vehicles, electronic ballast, DC/DC converters, SMPS PFC LED drivers, *etc*.

Corresponding author. Email: rewa.ashish@gmail.com

Kartick Chandra Jana received his BE and MTech degrees in electrical engineering from Regional Engineering College (presently NIT Durgapur), Durgapur, India, in 2000 and 2003, respectively, and the PhD degree in engineering from Jadavpur University, Kolkata, India, in 2011. He is currently working as associate professor, in the Electrical Engineering Department at the Indian Institute of Technology (ISM), Dhanbad, India. His areas of interest are power electronics, multilevel converters, motor drives, and renewable energy extraction.

Email: kartick_jana@yahoo.com