Received February 23, 2021, accepted March 10, 2021, date of publication March 17, 2021, date of current version March 30, 2021.

*Digital Object Identifier 10.1109/ACCESS.2021.3066494*

# Machine Learning Based PV Power Generation Forecasting in Alice Springs

**KHIZIR MAHMUD[1], (Member, IEEE), SAMI AZAM[2], ASIF KARIM[2], SM ZOBAED[3], BHARANIDHARAN SHANMUGAM[2], AND DEEPIKA MATHUR[2]**

[1]School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, NSW 2053, Australia
[2]College of Engineering, IT and Environment, Charles Darwin University, Casuarina, NT 0810, Australia
[3]School of Computing and Informatics, University of Louisiana at Lafayette, Lafayette, LA 70504, USA

Corresponding author: Sami Azam (sami.azam@cdu.edu.au)

**ABSTRACT** The generation volatility of photovoltaics (PVs) has created several control and operation challenges for grid operators. For a secure and reliable day or hour-ahead electricity dispatch, the grid operators need the visibility of their synchronous and asynchronous generators' capacity. It helps them to manage the spinning reserve, inertia and frequency response during any contingency events. This study attempts to provide a machine learning-based PV power generation forecasting for both the short and long-term. The study has chosen Alice Springs, one of the geographically solar energy-rich areas in Australia, and considered various environmental parameters. Different machine learning algorithms, including Linear Regression, Polynomial Regression, Decision Tree Regression, Support Vector Regression, Random Forest Regression, Long Short-Term Memory, and Multilayer Perceptron Regression, are considered in the study. Various comparative performance analysis is conducted for both normal and uncertain cases and found that Random Forest Regression performed better for our dataset. The impact of data normalization on forecasting performance is also analyzed using multiple performance metrics. The study may help the grid operators to choose an appropriate PV power forecasting algorithm and plan the time-ahead generation volatility.

**INDEX TERMS** Artificial intelligence, machine learning, power systems, PV power forecasting, renewable energy, statistical regression.

## I. INTRODUCTION

### *A. OVERVIEW*

Traditionally, electrical power generation systems are dominated by fossil fuel-based generators. However, due to their negative consequences on the environment, the power industry now focuses on alternative green energy-based generation systems [1]. So, in recent years, there is a fundamental shift towards deploying various renewable energy sources, including solar, wind, tidal, and biomass energy. The mass adaptation of these renewable energy sources, especially small and large-scale photovoltaics (PVs) has some environmental and economic benefits [2]. However, the PVs' intermittent nature makes it a highly volatile generator, which poses a significant challenge for the grid operators. Due to their high variability of active power, grid operators put some restrictions on solar farms while participating in the energy market and restrict the penetration of roof-top PVs in the medium-voltage distribution networks [3]. Most of the energy markets are a day ahead, and for that reason, grid operators require to make the day-ahead unit commitment and economic dispatch decisions [4], [5]. As PV power is highly volatile and can vary between zero to a hundred per cent, the grid operators act in a conservative way to avoid any contingencies while taking power from solar farms [4]. So, it is critical for all parties in the power industry, especially the grid operators, to know the forecasted output of the PV active power to plan their spinning and regulating reserves optimally. However, solar farms' output is highly dependent on various weather parameters, including temperature, solar irradiance, precipitation, humidity, and so on [6]. While predicting the day-ahead PV power, these environmental parameters' forecasted values contain a substantial percentage of random errors. So, it is essential to consider all influencing factors while predicting solar power and offering it to the energy markets [6].

Recently, several countries, including Australia, have accelerated the integration of PVs in their low and medium voltage networks. Due to its convenient geographical

The associate editor coordinating the review of this manuscript and approving it for publication was Mohamed Elhoseny.

location, Australia receives an average of 58 million petajoules (PJ) of solar radiation per year [7]. It also has the highest uptake of solar energy globally, with more than 21% of homes with rooftop PVs, according to the data revealed on 30 June 2020 by the Department of Industry, Science, Energy and Resources, Australia [8]. Solar energy usage is expected to show a strong upward trend with an increment of 5.9 per cent per year to 24 PJ in 2029–30 [7].

As solar power adaptation is accelerating in Australia, the level of generation volatility in the power systems is also increases. So, various regulatory and research organizations including the Australian National Electricity Market (NEM), Australian Energy Market Operator (AEMO), and Commonwealth Scientific and Industrial Research Organisation (CSIRO) studied solar energy forecasting for a reliable power systems operation [9]. AEMO is responsible for managing the national electricity market, including the southern and eastern parts of Australia [10]. However, AEMO does not deal with the electrical systems in Alice Springs, where average solar radiation is significantly higher than the southern and eastern parts of Australia [7]. Much of the previous studies on solar forecasting were conducted on a broad time scale ranging from minutes to months and years. They can be generalized into three different time horizons, *e.g.,* short, medium, and long-term [11]–[13]. Various statistical and numerical approaches [9], and a combination of two, also known as a hybrid model [14], [15] were applied in the forecasting. Various input parameter, mostly weather and physical arrangements of the PVs, including module type, angle, power ratings, and efficiency, are considered in the analysis. Most of these physical arrangement data are guided by the manufacturers, where weather data are collected from the adjacent weather stations. Sometimes meteorological data are used for long-term forecasting. Some studies focus on various meteorological weather forecasts, including cloud cover, irradiance, and temperature [5], [6]. Several prior works also reported PV energy forecasting using module-specific information and weather behaviour [7], [8]. In some studies, the meteorological data-driven approach is used for PV power generation forecasting. Few studies [16], [17] considered power demand and roof-top PV forecasting and their percentage errors to investigate their impact on domestic energy management. A study [11] presented a detailed comparative analysis between machine learning and meta-heuristic methods to help researchers choose appropriate forecasting techniques based on their objectives. In recent years, various machine learning approaches including artificial neural networks (ANN), convolutional neural networks (CNN), recurrent neural network (RNN), deep learning, and long and short-term memory (LSTM) are getting attention to forecast intermittent renewable energy generation [9], [11], [12]. The forecasting model proposed by authors in [7] used a combination of astronomical, historical, and meteorological data to improve the accuracy of the prediction. Authors in [18] proposed a forecasting model using CNN and LSTM. Under Australian Solar Energy Forecasting System (ASEFS), a thorough localized study used various machine learning techniques for five minutes to six days ahead of PV power generation prediction [9]. In March 2014, they reported that one hour-ahead normalized mean absolute error (NMAE) for Black Mountain area was 7.72%. The report [9] also shows that performance varies at a different time, season and location. Although a significant number of forecasting studies are conducted, the effectiveness of the prediction is highly dependent on the datasets which varies across place, time and season. So, a high performing approach in a particular location may not be performing as expected in another area.

### B. MOTIVATION

Traditional power systems are mostly based on synchronous power generators and can provide a certain level of inertia and spinning reserve capability to the grid. The inertia and spinning reserve maintain the grid security and strength during any minor or major contingency events. However, PV power generators are asynchronous and are designed to provide synthetic inertia. Sometimes during disturbances, the phase-locked loop (PLL) of the grid following inverters loses synchronization and pushes the PV power generation out of the grid, so it cannot provide any synthetic inertia support. On the other hand, due to the intermittent nature and asynchronous generation volatility, grid operators conservatively operate solar farms through multiple constraints. For a secure and reliable day or hour-ahead electricity dispatch, the grid operators need to know the capacity forecast of solar farms and plan the appropriate spinning reserve. When the solar power generation forecasting is not accurate, the grid operators must constrain their output or otherwise compromise the system's security and stability. During any cloud or storm events, the PV power output drops quite significantly. So, in the absence of an adequate spinning reserve, the grid operators need to choose the under-frequency load shedding. Therefore, it is crucial to get accurate PV power forecasting for a reliable and secure power grid.

### C. CONTRIBUTION

The objective of this study was to investigate the PV power generation forecasting in Alice Springs. We attempted to provide a survey on short and long-term PV power prediction and comparative analysis among various existing algorithms. Considering previous studies, this research makes the following contributions to the growing area of literature.

- Alice Springs-specific short-term and long-term PV power generation forecasting considering local weather patterns and behaviour.
- Investigated the weather parameter impact on PV power generation in Alice Springs.
- Provide a comparative study among various machine learning algorithm and report their performance using efficiency matrices.
- Investigate the impact of data pre-processing on prediction performance.

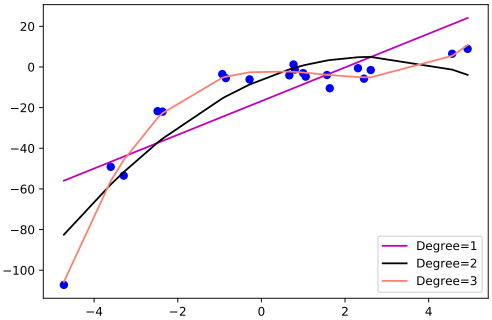


**FIGURE 1.** **Implementation of polynomial regression over training dataset.**

## II. LITERATURE REVIEW

### A. NOTABLE ALGORITHMS USED IN POWER FORECASTING

#### 1) LINEAR REGRESSION

Linear regression is a widely used algorithm to conduct predictive analysis of continuous data by determining the set of variables that influence significantly to estimate the outcome variable. Linear regression is often used to forecast the impact of changes in the dependent variable as a consequence of the variation in the independent variable(s) by any order of magnitude. The simplistic form of the regression equation with a single dependent and independent variable is defined in Equation 1, where $c,\ b,\ X$, and $Y$ denote constant, regression coefficient, independent variable, and estimated dependent or target variable respectively.

$$y = c + b \cdot x \tag{1}$$

Linear regression neither requires much time nor space overhead. This algorithm can be applied to most of the datasets [19]. There are several variants of it such as simple linear regression, multiple linear regression, logistic regression, ordinal regression, multinomial regression, and discriminant analysis.

#### 2) POLYNOMIAL REGRESSION

Polynomial regression is utilized in scenarios where the relationship between the target and the independent variable is not linear. If the data distribution is complex as depicted in Figure 1, straight lines might not capture patterns of the data. Therefore, unlike the linear equation, polynomial equations depicted in Equation 2, is more suitable to capture the data distribution. In the equation, $\theta_0$ denotes bias, $\theta_1, \theta_2, \ldots, \theta_n$ denote the weights, and $n$ denotes the degree of polynomial regression. $n$ determines the order in the equation. Increasing $n$ yields higher-order terms that turn the model into a more complex form. A complex model can usually avoid underfitting issue.

$$Y = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \ldots + \theta_n x^n \tag{2}$$

A polynomial regression is trained for $n = 1,\ 2,\ 3$ on the considered dataset. The performance of the regression is shown in Figure 1. According to the figure, linear regression ($n = 1$) fails to fit the data. On the other hand, the regression line captures the data complexity with minimal prediction error for $n = 3$. However, this affects the generalization capability of the models because of probable overfitting issue. Hence, it is recommended to consider such a degree that neither causes underfitting nor overfitting issues. The regression performs well on the dataset when $n = 2$.

#### 3) DECISION TREE REGRESSION

Decision tree regression observes features of an object and trains a model as a tree-like structure to predict upcoming data and generates meaningful continuous output. The regression is useful when the target variable is continuous. In addition, it works well compared to other regression algorithms when there are missing features in the dataset, a mixture of categorical and numerical features, and large differences in similar features [20], [21].

A decision node contains two or more branches that represent values for the attribute tested. The value of leaf nodes is the mean of the observations falling in that region. Therefore, if an unseen data point falls in that region, we predict using the mean value. Commonly, popular regression algorithms such as linear or polynomial would not be able to fit such discrete datasets, whereas decision tree regression performs well. Decision tree regression is suitable for the datasets that contain both non-linearity and non-continuity, as well.

#### 4) SUPPORT VECTOR REGRESSION

Support vector regression (SVR) is one of the widely adopted regression techniques because of its advantages over complex data distribution. The regression is based on the concept of support vector machine (SVM) that tries to find an optimal hyperplane that can classify datapoints in an N-dimensional space. SVM can be used with the datapoints of binary classes even the datapoints are non-linearly separable in two dimensional space. Hence, SVM applies a kernel to transform the datapoints into a N-dimensional space where the classes can be separated linearly. To look for an optimal hyperplane, SVM determines support vectors that are basically two marginal datapoints of different classes. Support vectors are selected considering that the hyperplane should be positioned at the least possible distance from both of them. In Figure 2, we show a high-level overview of applying SVM to the example data.

SVR differs from SVM since SVM is used to predict discrete class labels, whereas SVR is a regressor that is used to predict continuous ordered variables. In simple regression, the error rate between prediction and the actual value is minimized. SVR tries to fit the error within a certain threshold. Based on the predefined threshold, it creates a boundary space. SVR considers the datapoints that are within the boundary to provide a better fitting model. Support Vectors help determining the closest match between the data points and the function used to represent them. The following steps are required to apply SVR to the training dataset:

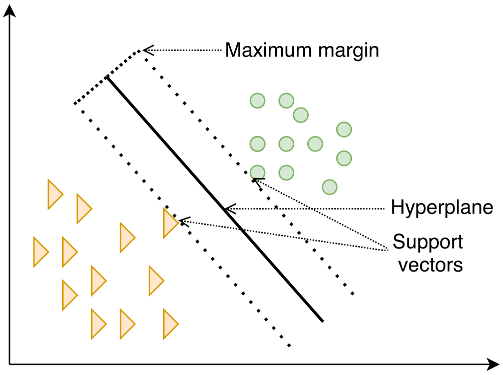


**FIGURE 2.** A high-level overview of applying support vector machine to example datapoints.

1) Kernel function selection with its parameters and any regularization if required. Inappropriate choice of kernel affects the regression model.
2) Creation of correlation matrix.
3) Training the model to get the contraction coefficients $\alpha = \alpha i$
4) Estimator creation using the coefficients.

Equation 3 is used to apply SVR on the training datapoints when they are linearly separable.

$$y = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) \cdot < x_i, x > +b \tag{3}$$

Equation 4 is used to apply SVR on the training datapoints when they are non-linearly separable, where $K(x, y)$ denotes kernel function.

$$y = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) \cdot K(x_i, x) + b \tag{4}$$

One of the notable advantages of SVR over other regression algorithms is that it can improve prediction accuracy by calculating confidence in classification. Moreover, in comparison with other algorithms, SVR is computationally less intensive.

### 5) RANDOM FOREST REGRESSION

Random forest regression is based on ensemble learning method that combines predictions from multiple machine learning algorithms to generate more accurate predictions compared to a standalone model. The following steps are required to perform random forest regression over the training dataset:

1) At first, $k$ number of datapoints are chosen from the input (training) dataset, $X$
2) A decision tree is built that is associated to these $k$ datapoints.
3) Steps 1 and 2 are repeated until generating $N$ number of decision trees during the training period.
4) For a new datapoint, each of the trees generates the prediction value of $y$ and assigns that datapoint to the average across all the predicted $y$ values.

Random forest regression performs well on diversified problems with the potentiality of handling non-linear relationships. However, the overfitting issue is observed for some datasets by applying this regression while training. In addition, it is also biased, particularly to the categorical variables that contain more levels [22].

### 6) ARTIFICIAL NEURAL NETWORK

An artificial neural network (ANN) is formed by hundreds or thousands of artificial neurons that are designed to simulate human brain cells. The network contains a vast number of connections that provide the output of one neuron as an input to another. Each connection is assigned a weight that represents its relative importance. An artificial neuron can contain various input or output connections. The neurons are classically organized into multiple layers. The layers of neurons that receive data and provide the ultimate result are the input and output layer, respectively. In between the input and output layers, there exist one or more hidden layers. Hence, such vanilla implementation of ANN is also referred as multilayer perceptron (MLP). Multiple layers and non-linear activation can classify data that are not linearly separable.

Commonly, MLP initializes a training phase where it learns to detect patterns in data, either visually or textually. The network compares its produced (predicted) outputs with the desired (actual) ones throughout the training phase. Later, the difference between the actual and predicted outputs is adjusted using backpropagation. Backpropagation means going from the output to the input units for adjusting the connection weights until the difference between the actual and predicted outputs shows the lowest possible error.

### 7) LONG SHORT-TERM MEMORY (LSTM)

Unlike traditional feed-forward ANN, RNN can use their internal state to process a sequence of inputs. However, vanilla RNN leaves out important information regarding long sequences which affects the predicted output. LSTM network is a variant of recurrent neural network (RNN) capable of learning order dependency in sequence prediction related problems.

The core component of LSTM is the memory cell which can control its state over time, consisting of explicit memory (aka cell state vector) and forget, input, and output gate. Forget gate controls what information should be deleted from memory. It also decides how much of the past the network should remember. The input gate controls what portion of new information is added to the cell state from the current input. The output gate decides what to output from the memory cell.

## *B. APPLICATIONS OF COMMONLY USED ALGORITHMS*

Arce and Macabebe *et al.,* proposed a model that was used to predict the total solar energy consumption of the residence at each month by using linear regression, polynomial regression, random forest regression, and SVR algorithms [23].

**TABLE 1.** A summary of the literature studies.

| Studies | Contribution | Technique Used | Comments |
|---|---|---|---|
| [23] | Real-time solar power consumption forecasting | Linear, polynomial, random forest regression, and SVR | Although SVR outperformed other regressions, prediction accuracy of random forest was close to it. |
| [24] | Solar irradiance prediction | Linear, decision tree regression, and SVR | SVR achieved the highest prediction accuracy. The authors also identified significant attributes from data. They trained the model using these attributes only and obtained improvement in prediction accuracy. |
| [25] | Predicting output of solar photovoltaic panel | Linear regression, neural network | Comparing the accuracy between linear regression and neural network models, no significant differences was observed. |
| [26] | A density-based probabilistic model for wind and solar power prediction | A modified version of SVR | The authors discussed several fluctuation factors that affect the prediction accuracy. Their model FIG-SVQR outperformed all of the compared models. |
| [27] | Power output prediction of a grid-connected solar panel | SVR | SVR provided a subtle improvement in prediction accuracy. The authors investigated how unusual climatic changes could impact the prediction accuracy. |
| [28] | Solar irradiance prediction | LSTM | LSTM achieved the highest prediction accuracy by reducing at least 42.9% RMSE compared to others. |

Their experimental result demonstrates that linear and polynomial regressions showed significant error rates because of the non-linear trend of the training dataset. On the other hand, random forest regression and SVR only had minimal error rates. Even though random forest provided the lowest error rate, the model was unable to provide a precise continuous predictions and subsequently, it could not predict beyond the range of the training dataset compared to SVR. As a result, support vector was chosen as the appropriate model for solar energy consumption, having high accuracy and low error rate.

Javed *et al.,* proposed [24] a model for solar irradiance prediction by exploiting linear regression, decision tree regression, and SVR algorithms. To forecast the irradiance, the model leveraged essential climate information considering several significant parameters such as temperature, wind, humidity, and speed. The authors depicted that support vector with radial basis function (RBF) ensured the highest prediction accuracy. They also emphasized data preprocessing and selection of the appropriate attributes such as dew point, precipitation, sky cover coverage as they were responsible for improving the performance of the prediction model.

Shapsough *et al.,* [25] implemented linear regression-based and neural network-based models on sensor data to estimate the power output of photovoltaic systems. The models were trained and validated on actual monitoring data. In addition, to observe prediction accuracy, they also researched the original data's attributes to determine the effective ones. They suggested irradiance and temporal data are sufficient to train a model. Based on their experiments, the maximum possible prediction accuracy of power output is about 97%. However, according to the experimental results, the performance of one model did not precede the other ones.

He *et al.,* proposed [26] a model named FIG-SVQR to predict wind and solar power utilizing a modified variant of SVR namely, support vector quantile regression. They preprocessed the raw data based on fuzzy information granulation to eliminate the fluctuation, noise, and uncertainty from data. To compare their proposed model's performance, they also created a baseline version without preprocessing. They utilized the *Epanechnikov* kernel function in FIG-SVQR and a baseline to obtain the probability density curves of the prediction results. Overall, the proposed model clearly outperformed their baseline model which indicated the effectiveness of the preprocessing step. However, to rigorously check the efficiency of preprocessing, other regression algorithms should be used as well.

Nageem *et al.,* proposed [27] a forecasting model to estimate the power output from a solar panel using multi-input SVR algorithm. They investigated that any unexpected climatic change could increase the error rate in prediction accuracy. However, one of the major drawbacks of this study is that they compared the proposed model with a traditional analytic approach. Moreover, their model achieved negligible performance improvement compared to the traditional one. Due to the lack of comparison with other regression algorithms, it is not evident that the SVR-based models are the most suitable regression for their dataset.

Qing *et al.,* applied LSTM to predict households or small commercials solar irradiance [28]. Their study showed that LSTM network presented better performance than traditional networks or regressors. The study involved 10 years of historical data to predict one year of irradiance data. The implemented LSTM-based model offered 18.34% improvement in prediction and 42.9% decrement in RMSE compared to other approaches. Although the LSTM-based model provided higher accuracy than that of other algorithms, the incurred training time of LSTM is substantially higher than that of other algorithms.

In Table 1, a summary of the aforementioned literature studies is provided. Based on the studies, we conclude that LSTM is the overall winner since it can efficiently train on complex data distribution. Specifically, in studies where neural network-based models had been

**TABLE 2.** PV technology of the Array from where actual data is taken.

| Parameter | Values |
|---|---|
| Array ratings | 5.6 kW |
| PV technology | CdTe |
| Array structure | Fixed: ground mount |
| Inverter size | 6 kW |
| Array tilt | 20 |
| Azimuth | 0 (Solar north) |

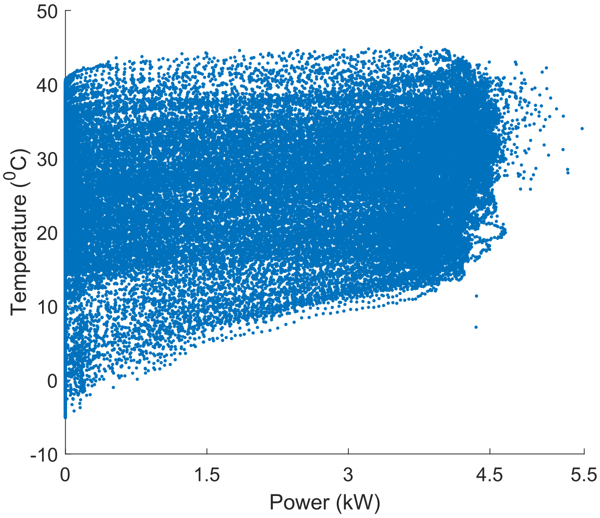


**FIGURE 3.** Relationship between temperature and PV power generation.

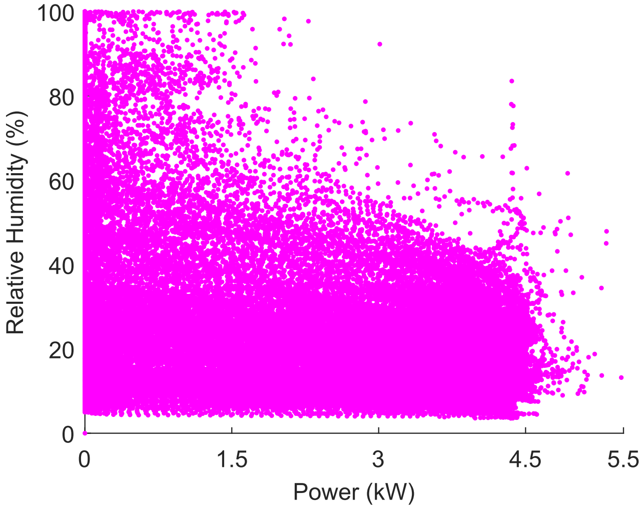


**FIGURE 4.** Relationship between relative humidity and PV power generation.

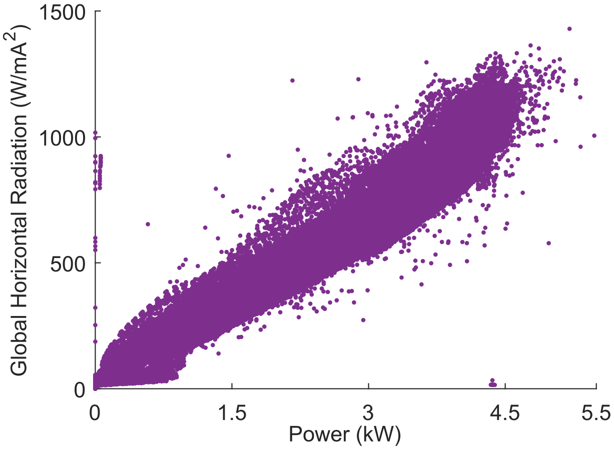


**FIGURE 5.** Relationship between global horizontal radiation and PV power generation.

used, LSTM proved itself as a good competitor of such models.

## III. CASE STUDIES

### A. RELATIONSHIP BETWEEN POWER AND WEATHER PARAMETER

Various weather parameters including temperature, relative humidity, global horizontal radiation, diffuse horizontal radiation and daily precipitation are used to train the model and predict PV power output. These weather data are taken from the actual weather station of the solar installation site at Alice Springs, Australia. The historical weather data and actual PV power output of the array is provided by the Desert Knowledge Australia Centre [29]. The PV technology of the array that used as data source is listed in Table 2.

The relationship between PV power output and the temperature, relative humidity, global horizontal radiation, diffuse horizontal radiation, and daily precipitation is shown in Figures 3, 4, 5, 6, and 7. The temperature has a proportional relationship with the PV power generation as shown in Figure 3, while humidity shows inverse relationship as represented in Figure 4.

As the PV cells function based on the sunlight, the radiation from the sun is represented in a few different ways including global horizontal radiation (also known as Global Horizontal Irradiance (GHI)) and diffuse horizontal radiation. Global horizontal radiation is calculated as the total amount of shortwave radiation received by a surface horizontal to the ground, where diffuse horizontal radiation does not arrive directly from the sun, but scattered by particles in the atmosphere [30]. So, both global and horizontal diffuse radiation are important for PV power generation. Based on the analysis of the selected dataset as represented in 5 and 6, it is found that the Global Horizontal Radiation and Diffuse Horizontal Radiation has a strong proportional relationship between PV power output. On the other hand, the precipitation in Figure 7 does not show any dominant relation with the PV power output.

### B. FORECASTING OUTPUT

The forecasting is carried out using Python programming language with various library functions such as scikit-learn, keras, pandas, and numpy. The PV power prediction in comparison to the actual output, using LR, PR, SVR, RFR, and DTR are shown in Figure 8.

Figure 8 represents the long-term PV power forecasting performance. Both actual and predicted data are calculated for one year with five minutes intervals resulting in 105120

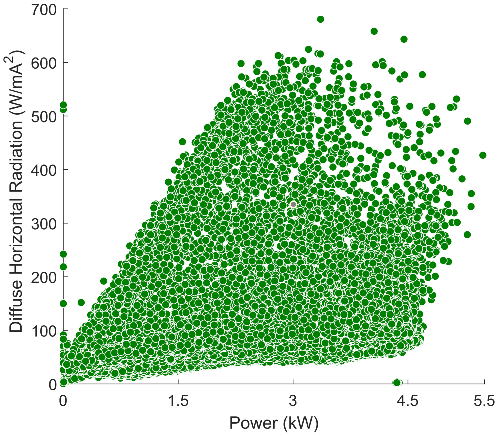


**FIGURE 6.** Relationship between diffuse horizontal radiation and PV power generation.

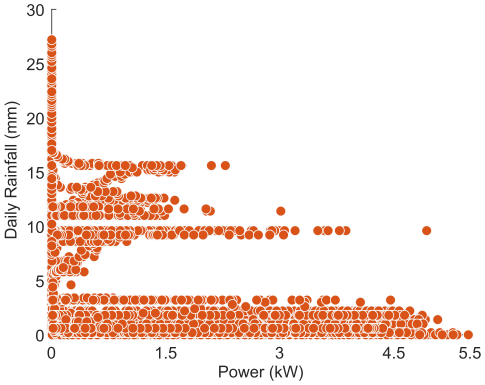


**FIGURE 7.** Relationship between daily precipitation and PV power generation.

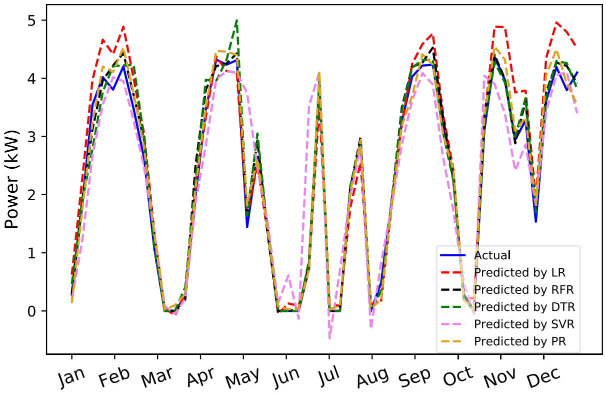


**FIGURE 8.** Long-term PV forecasting for weekly data interval and one year duration.

data points in total. Since, the number of data points are too many to show them in a single graph, we take one random data point from each week as shown in Figure 8. The line graphs of individual algorithms are shown separately in Figures 9-13. The figures show a mixed performance spectrum in the considered time horizon and data landscape. RFR, PR, and DTR show better performance than SVR and LR.

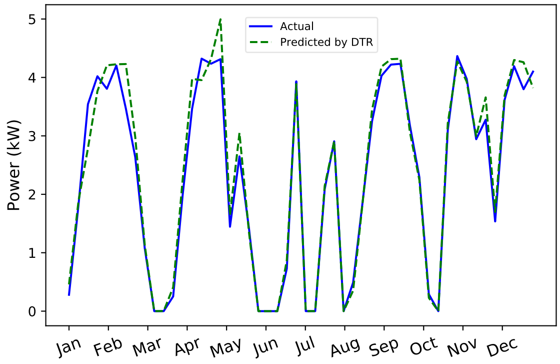


**FIGURE 9.** Long-term PV power forecasting using DTR.

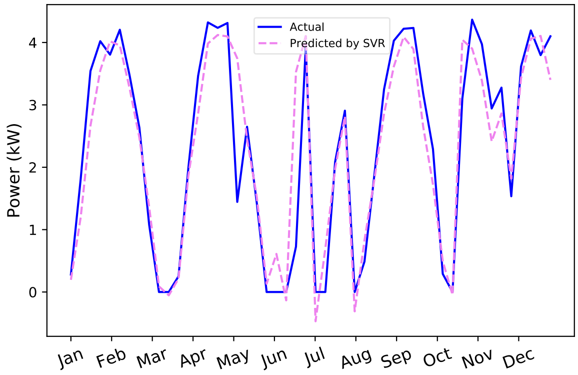


**FIGURE 10.** Long-term PV power forecasting using SVR.

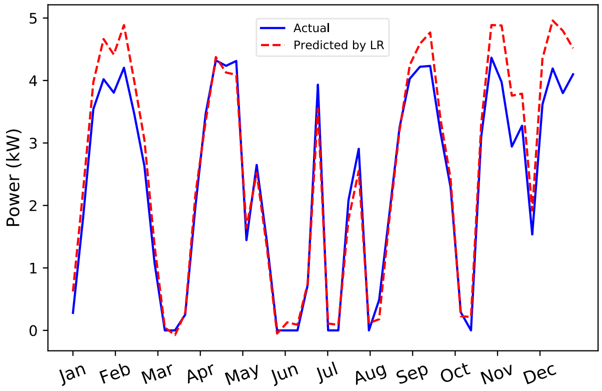


**FIGURE 11.** Long-term PV power forecasting using LR.

Although RFR, PR, and DTR perform better in predicting the PV power profile, they could provide unexpectedly higher peaks during maximum PV power generation time. However, their prediction performance during lower power generation time is noteworthy. On the other hand, LR and SVR show lower forecasting efficiencies for both the lower and upper power generation range. The performance matrix for one-year data points is listed in Table 3 and Table 4.

A medium-term PV power forecasting considering a week time horizon and all 5-minute time interval data points is shown in Figure 14. For a better graphical representation, all data points at nighttime (which are zero) are omitted. The figure shows that the SVR has higher efficiency but

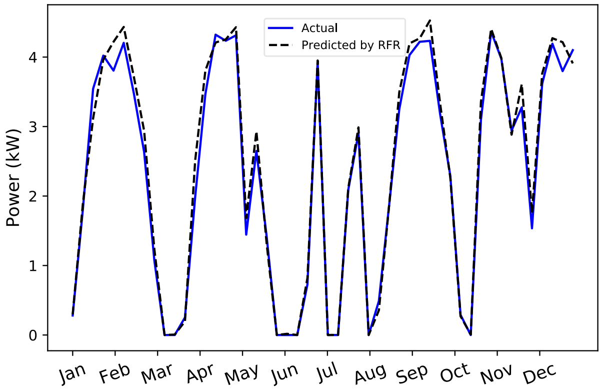


**FIGURE 12.** Long-term PV power forecasting using RFR.

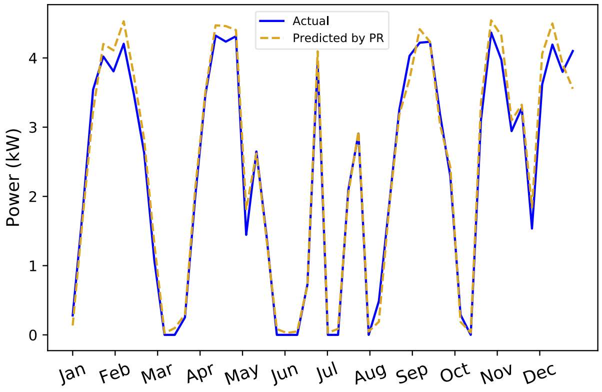


**FIGURE 13.** Long-term PV power forecasting using PR.

**TABLE 3.** Comparison of performance matrices between various algorithms using normalized data.

| Algo. | Normalized Data | | | | |
|---|---|---|---|---|---|
| | MAE | MSE | MedAE | EVS | $R^2$ Score |
| LT | 0.0282 | 0.0018 | 0.0163 | 0.9695 | 0.9657 |
| PR | 0.0159 | 0.0006 | 0.0087 | 0.9886 | 0.9880 |
| SVR | 0.0157 | 0.0056 | 0.0507 | 0.8994 | 0.8944 |
| DTR | 0.0137 | 0.0011 | 5.6668 | 0.9804 | 0.9799 |
| RFR | 0.0098 | 0.0004 | 2.5818 | 0.9924 | 0.9919 |
| MLP | 0.1492 | 0.0389 | 0.1532 | -1.3907 | -1.4027 |
| LSTM | 0.0447 | 0.0048 | 0.5618 | 0.5599 | 0.5618 |

**TABLE 4.** Comparison of performance matrices between various algorithms using un-normalized data.

| Algo. | Data Without Normalization | | | | |
|---|---|---|---|---|---|
| | MAE | MSE | MedAE | EVS | $R^2$ Score |
| LT | 0.2148 | 0.0938 | 0.1375 | 0.9683 | 0.9638 |
| PR | 0.8051 | 1.0725 | 0.7104 | 0.5857 | 0.5853 |
| SVR | 1.2196 | 3.8276 | 0.1002 | -2.9016 | -0.4799 |
| DTR | 0.1172 | .0647 | 3.9668 | 0.9768 | 0.9750 |
| RFR | .0959 | .0415 | 4.9972 | 0.9858 | 0.9840 |

DTR and LR perform poorly. The forecasting performance representation using line graphs is further narrowed down to a short-term scale, *i.e.,* a 24-hour time horizon, and shown in Figure 15. The figure showed all 5-minutes interval data points for 24 hours. The performance is discussed using various standard matrices in Section III(C).

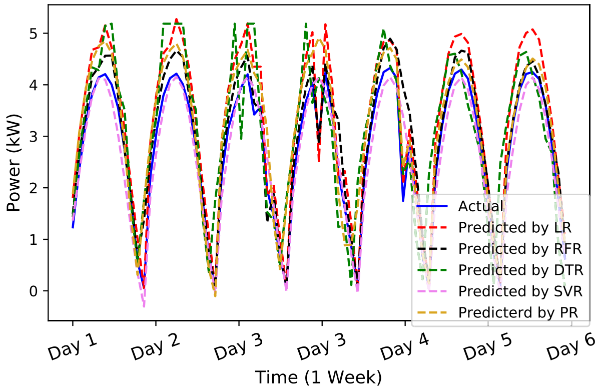


**FIGURE 14.** Medium-term PV power generation forecasting.

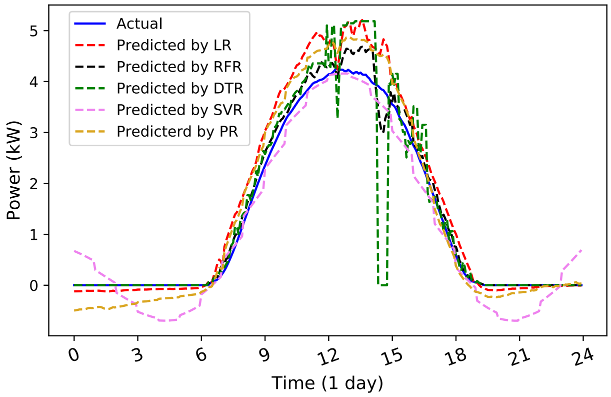


**FIGURE 15.** Short-term PV power generation forecasting.

### C. FORECASTING PERFORMANCE DURING UNCERTAINTY

PV power forecasting is necessary for grid operators to understand the capacity of the intermittent asynchronous generators and manage their daily operations through a reliable and economic unit commitment (UC). However, it is crucial to know the generation volatility of PVs during the rapid changes of environmental parameters. This section investigates the forecasting performance during PV power generation fluctuation conditions. Two types of fluctuations, medium and fast fluctuation, are selected to investigate the forecasting performance during uncertainty. From the study, it is found that during medium-level of uncertainty LR, PR, and RFR perform better than DTR, whereas SVR performs poorly and even becomes negative during zero output cases, as shown in Figure 16. During the higher-level of uncertainty, SVR and DTR struggle to follow the fluctuating tracks, while LR and PR perform better to track the profile, as shown in Figure 17.

### D. COMPARISON OF PERFORMANCE MATRICES OF FORECASTING ALGORITHMS

The performance of the considered forecasting algorithms is analyzed using various performance matrices including $R^2$ score, mean absolute error (MAE), mean squared error

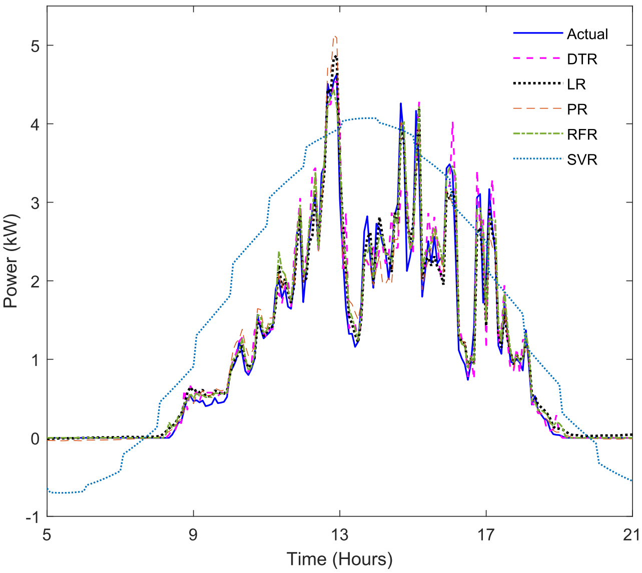


**FIGURE 16.** **Forecasting performance during medium fluctuation case.**

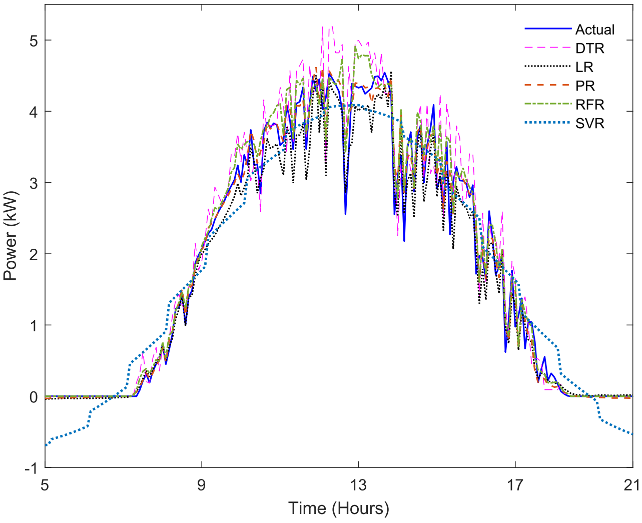


**FIGURE 17.** **Forecasting performance during fast fluctuation case.**

(MSE), median absolute error (MedAE) and explain variance score (EVS).

The $R^2$ score is represented as the measure of the ratio of variability that the model can capture vs the natural variability in the target variable. Ideally, a value closer to 1 is better. The mean absolute error (MAE), mean squared error (MSE), and median absolute error (MedAE) are calculated with comparison to the forecasted and actual values. Smaller values mean better prediction performance. Variance score closer to 1 represents the discrepancies between a model and actual data. From the performance analysis, it is found that the training data normalization for prediction performs slightly better than the training data without any normalization.

In terms of MAE and MSE, as shown in Table 3 and Figure 18, LSTM and MLP show poor performance, while PR and DTR had satisfactory outcomes. RFR, according to these two measures of error, are the optimum performer, with an improvement of close to 34% from the averages of PR and DTR. Lower values indicate better performance.

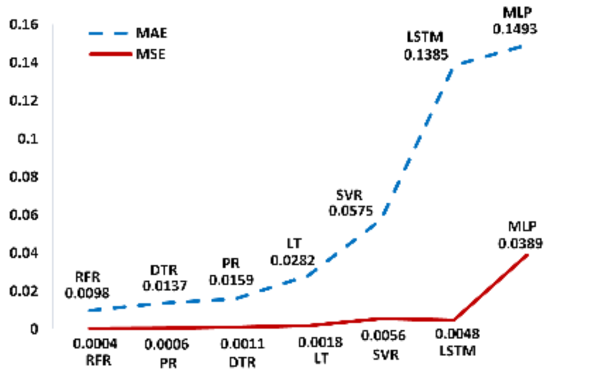


**FIGURE 18.** **Mean absolute and mean squared error representation with normalized data.**

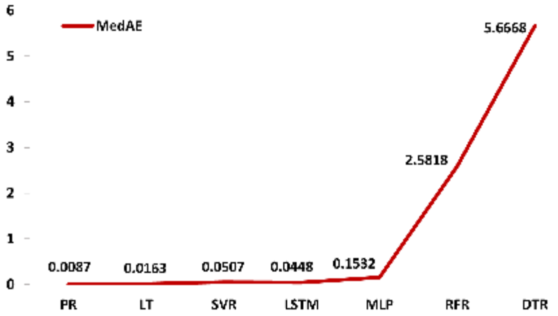


**FIGURE 19.** **Median absolute error (MedAE) representation with normalized data.**

Figure 19 shows quite the opposite trend when it comes to MedAE, where DTR and RFR have not demonstrated any strong outcomes at all whereas PR has been the top performer showing the least MedAE. LT and SVR achieve reasonably satisfactory outcomes as these are quite close to PR. Lower values indicate better performance.

EVS and $R^2$ (scores are in Table 3) show the same trends, where RFR demonstrates the best results while MLP is the least desirable, as shown in Figure 20. Values close to 1 indicate better performance.

In terms of MAE and MSE (Figure 21) for un-normalized data, PR and SVR demonstrate unsatisfactory performance while LT and DTR are acceptable outcome. RFR, according to these two measures of error, is the optimum performer, with an improvement of over 42% from the averages of LT and DTR. Lower values indicate better performance.

Figure 22 visualizes the opposite trend when it comes to MedAE, where DTR and RFR have not demonstrated any strong outcomes whereas SVR is the top performer showing the least MedAE. LT and PR achieve an average outcome. Lower values indicate better performance.

According to Figure 23, EVS and $R^2$ show trends that are similar to MSE. According to the figure, RFR demonstrates the best results but SVR performs worst. Values close to 1 indicate better performance.

Figure 24 represents a comparison with the average results for each metric obtained with or without normalization.

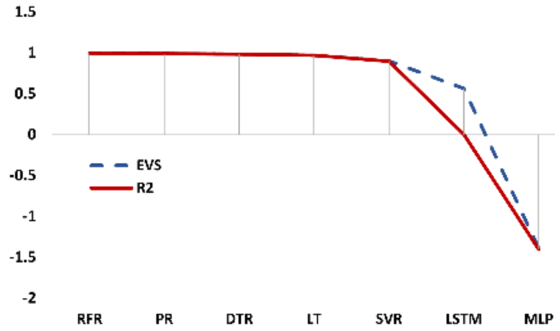


**FIGURE 20.** Explain variance score and $R^2$ representation for normalized data.

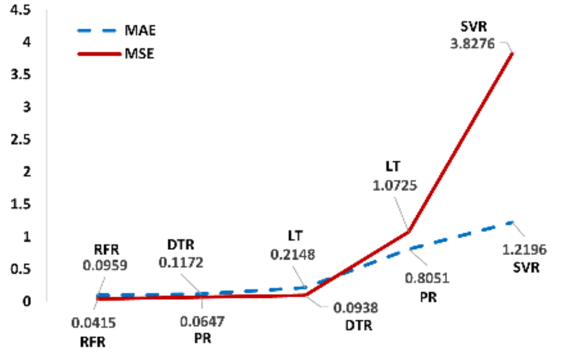


**FIGURE 21.** Mean absolute error (MAE) and mean squared error (MSE) representation for un-normalized data.

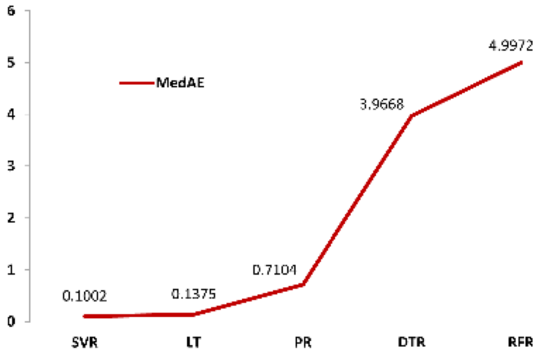


**FIGURE 22.** Median absolute error (MedAE) for un-normalized data.

Clearly in all respects, normalization achieves good performance.

Various performance matrices listed in Table 3 is further categorized based on their efficiencies, as shown in Figure 25. For our dataset, it is found that RFR delivers the highest efficiencies, while MLP has the least performance. Besides, PR has also demonstrated commendable performance as it was close to RFR in multiple observations. DTR has performed reasonably, however, with compromised reliability. On the other hand, SVR, LT and LSTM performed poorly under most of the measurements. The figure portrays the stronger performances of multiple general machine learning algorithms over the Deep Learning counterparts in predicting our target variable. The current research is conducted using 5-minutes granular datasets. However, a higher granular dataset may provide a better prediction performance.

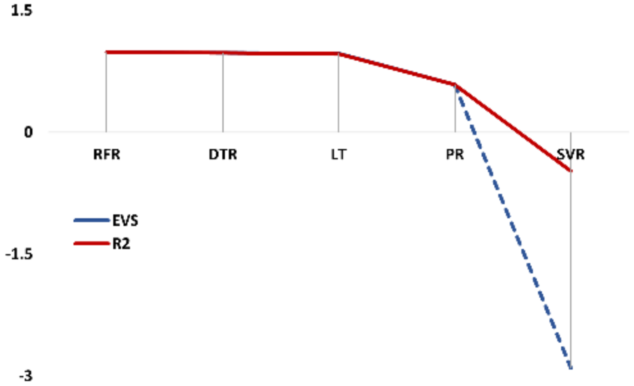


**FIGURE 23.** EVS and $R^2$ representation for un-normalized data.

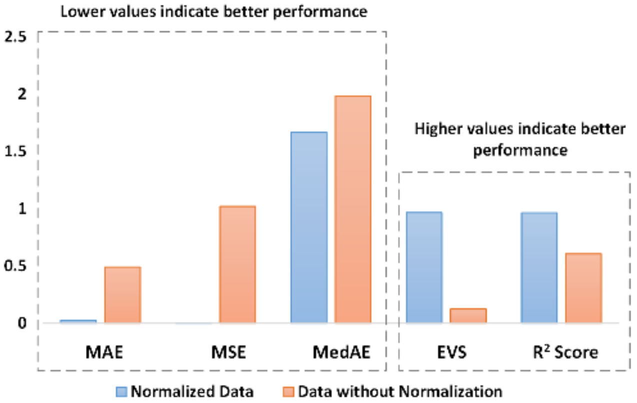
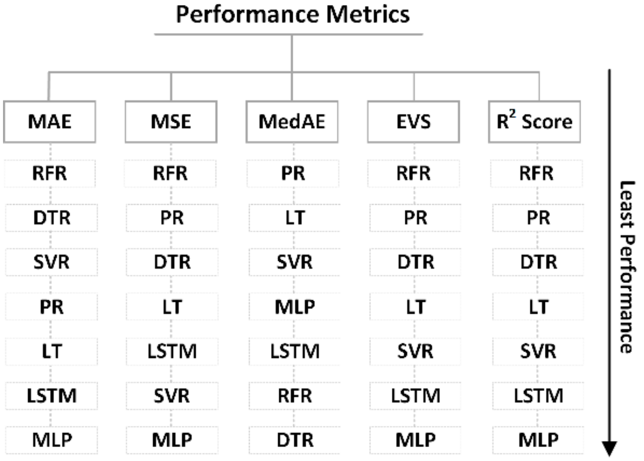


**FIGURE 24.** Comparison of results between Normalized Data and Data without Normalization.



Performance Metrics

| MAE | MSE | MedAE | EVS | $R^2$ Score |
|---|---|---|---|---|
| RFR | RFR | PR | RFR | RFR |
| DTR | PR | LT | PR | PR |
| SVR | DTR | SVR | DTR | DTR |
| PR | LT | MLP | LT | LT |
| LT | LSTM | LSTM | SVR | SVR |
| LSTM | SVR | RFR | LSTM | LSTM |
| MLP | MLP | DTR | MLP | MLP |

**FIGURE 25.** Algorithms based on their performance.

### *E. IMPACTS OF THE PV POWER FORECAST IN POWER SYSTEMS*

For a reliable and secure operation of power systems, it is essential to have full information about the networks and generators. The grid operators access all substation busbars, transformers, protection devices' status, network voltage, frequency, and spinning reserve conditions using Supervisory control and data acquisition (SCADA), as shown in Figure 26. Likewise, the grid operator also uses synchronous generators

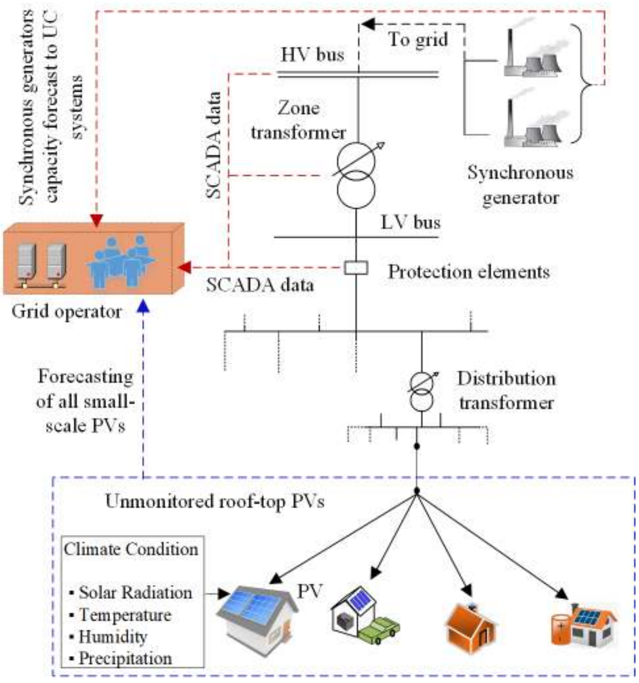


**FIGURE 26.** Significance of the small-scale PV power forecasting in power systems.

capacity forecast for a secure and economic unit commitment (UC). However, small-scale power generators such as roof-top photovoltaics are usually intermittent and are usually stationed behind the meter. So, they largely remain unmonitored and uncontrolled for grid operators. If there are few roof-top PVs, their impacts are negligible for a bigger grid. However, their higher penetration can impact grid stability and security if grid operators cannot forecast their behaviour. As Alice Springs is a geographically solar energy enriched area with a smaller grid size, it is crucial to precisely forecast the small-scale behind the meter and unmonitored PVs and get their penetration status in aggregated form. This research will help the grid operators understand the short, medium to long-term aggregated PV power generation and plan the grid security and reliability options.

## IV. DISCUSSION AND CONCLUSION

The purpose of the current study was to provide a machine learning-based short-term and long-term PV power generation forecasting for Alice Springs, Australia. A comparative analysis using various machine learning approaches, including Linear Regression, Polynomial Regression, Decision Tree Regression, Support Vector Regression and Random Forest regression, Long Short Term Memory and Multilayer Perceptron Regression, is provided. Detailed analysis of various performance matrices such as $R^2$ score, mean absolute error, mean squared error, median absolute error have been calculated. An analysis of the potential impacts of weather parameters on PV power prediction reveals that the relative humidity, temperature, diffuse horizontal radiation, and global horizontal radiation substantially impact PV power output, where daily precipitation appears to be a less significant dominating factor of PV power prediction. For the short-term PV power generation forecast, the cloud covering condition is a crucial parameter. However, we could not verify the algorithm's performance under cloud-cover or fast cloud moving conditions due to the data unavailability. In the actual implementation, incorporating this cloud status data in the training set may change the forecasting performance reported in this research. A comparative study with and without pre-processing shows that machine learning with data pre-processing performs better than the approach that directly feeds raw data to the engine. While pre-processing the historical PV outputs, non-zero values during night time are filtered, and it has improved the forecasting performance substantially. The short-term PV forecast requires forecasted weather parameters. This study could not test its performance using predicted weather values due to the data unavailability, and it might degrade the forecasting performance during real implementation. Besides, the time horizon based on which performance matrics are considered consists of both day and night time. However, PV has no active power output during night time. So, the exclusion of night-time output from the calculated time horizon may degrade performance matrics slightly. The insights gained from this study may be of assistance to plan long term renewable and non-renewable energy generation mix. Further work needs to be done to understand the large- scale PV power generation forecasting for panels with maximum power point tracking.

## REFERENCES

[1] M. J. E. Alam, K. M. Muttaqi, and D. Sutanto, "Effective utilization of available PEV battery capacity for mitigation of solar PV impact and grid support with integrated V2G functionality," *IEEE Trans. Smart Grid*, vol. 7, no. 3, pp. 1562–1571, May 2016.

[2] S. Sobri, S. Koohi-Kamali, and N. A. Rahim, "Solar photovoltaic generation forecasting methods: A review," *Energy Convers. Manage.*, vol. 156, pp. 459–497, Jan. 2018.

[3] Y. Karimi, H. Oraee, M. S. Golsorkhi, and J. M. Guerrero, "Decentralized method for load sharing and power management in a PV/battery hybrid source islanded microgrid," *IEEE Trans. Power Electron.*, vol. 32, no. 5, pp. 3525–3535, May 2017.

[4] G. Mohy-ud-din, K. M. Muttaqi, and D. Sutanto, "Transactive energy-based planning framework for VPPs in a co-optimised day-ahead and real-time energy market with ancillary services," *IET Gener., Transmiss. Distrib.*, vol. 13, no. 11, pp. 2024–2035, Jun. 2019.

[5] Z. Csereklyei, S. Qu, and T. Ancev, "The effect of wind and solar power generation on wholesale electricity prices in Australia," *Energy Policy*, vol. 131, pp. 358–369, Aug. 2019.

[6] R. Ahmed, V. Sreeram, Y. Mishra, and M. D. Arif, "A review and evaluation of the state-of-the-art in PV solar power forecasting: Techniques and optimization," *Renew. Sustain. Energy Rev.*, vol. 124, May 2020, Art. no. 109792.

[7] Australian Government and BREE, *Australian Energy Resource Assessment*, 2nd ed. Australia: Government of Australia, 2014.

[8] Energy.gov.au is a Department of Industry, Science, Energy and Resources website. (2020). *Solar PV and Batteries*. Accessed: Aug. 22, 2020. [Online]. Available: https://www.energy.gov.au/households/solar-pv-and-batteries

[9] C. Arena, "Australian solar energy forecasting system final report: Project results and lessons learnt," Commonwealth Sci. Ind. Res. Org., Canberra, ACT, Australia, Tech. Rep., May 2016.

[10] *Fact Sheet, the National Electricity Market*, AEMO, Melbourne, VIC, Australia, Jul. 2020.

[11] M. N. Akhter, S. Mekhilef, H. Mokhlis, and N. Mohamed Shah, "Review on forecasting of photovoltaic power generation based on machine learning and Metaheuristic techniques," *IET Renew. Power Gener.*, vol. 13, no. 7, pp. 1009–1023, May 2019.

[12] W. VanDeventer, E. Jamei, G. S. Thirunavukkarasu, M. Seyedmahmoudian, T. K. Soon, B. Horan, S. Mekhilef, and A. Stojcevski, "Short-term PV power forecasting using hybrid GASVM technique," *Renew. Energy*, vol. 140, pp. 367–379, Sep. 2019.

[13] U. K. Das, K. S. Tey, M. Seyedmahmoudian, S. Mekhilef, M. Y. I. Idris, W. Van Deventer, B. Horan, and A. Stojcevski, "Forecasting of photovoltaic power generation and model optimization: A review," *Renew. Sustain. Energy Rev.*, vol. 81, pp. 912–928, Jan. 2018.

[14] A. Y. Alanis, L. J. Ricalde, C. Simetti, and F. Odone, "Neural model with particle swarm optimization Kalman learning for forecasting in smart grids," *Math. Problems Eng.*, vol. 2013, pp. 1–9, Jan. 2013.

[15] D. Zhang, X. Han, and C. Deng, "Review on the research and practice of deep learning and reinforcement learning in smart grids," *CSEE J. Power Energy Syst.*, vol. 4, no. 3, pp. 362–370, Sep. 2018.

[16] K. Mahmud, J. Ravishankar, M. J. Hossain, and Z. Y. Dong, "The impact of prediction errors in the domestic peak power demand management," *IEEE Trans. Ind. Informat.*, vol. 16, no. 7, pp. 4567–4579, Jul. 2020.

[17] K. Rahbar, J. Xu, and R. Zhang, "Real-time energy storage management for renewable integration in microgrid: An off-line optimization approach," *IEEE Trans. Smart Grid*, vol. 6, no. 1, pp. 124–134, Jan. 2015.

[18] B. Ray, R. Shah, M. R. Islam, and S. Islam, "A new data driven long-term solar yield analysis model of photovoltaic power plants," *IEEE Access*, vol. 8, pp. 136223–136233, 2020.

[19] M. Fernández-Delgado, M. S. Sirsat, E. Cernadas, S. Alawadi, S. Barro, and M. Febrero-Bande, "An extensive experimental survey of regression methods," *Neural Netw.*, vol. 111, pp. 11–34, Mar. 2019.

[20] *Linear Regression vs Decision Trees*. Accessed: Sep. 2020. [Online]. Available: https://mlcorner.com/linear-regression-vs-decision-trees/

[21] A. Karim, S. Azam, B. Shanmugam, K. Kannoorpatti, and M. Alazab, "A comprehensive survey for intelligent spam email detection," *IEEE Access*, vol. 7, pp. 168261–168295, 2019.

[22] *Random Forest Regression*. Accessed: Sep. 2020. [Online]. Available: https://towardsdatascience.com/random-forest-and-its-implementation-71824ced454f

[23] J. M. M. Arce and E. Q. B. Macabebe, "Real-time power consumption monitoring and forecasting using regression techniques and machine learning algorithms," in *Proc. IEEE Int. Conf. Internet Things Intell. Syst. (IoTaIS)*, Nov. 2019, pp. 135–140.

[24] A. Javed, B. K. Kasi, and F. A. Khan, "Predicting solar irradiance using machine learning techniques," in *Proc. 15th Int. Wireless Commun. Mobile Comput. Conf. (IWCMC)*, Jun. 2019, pp. 1458–1462.

[25] S. Shapsough, R. Dhaouadi, and I. Zualkernan, "Using linear regression and back propagation neural networks to predict performance of soiled PV modules," *Procedia Comput. Sci.*, vol. 155, pp. 463–470, 2019.

[26] Y. He, Y. Yan, and Q. Xu, "Wind and solar power probability density prediction via fuzzy information granulation and support vector quantile regression," *Int. J. Electr. Power Energy Syst.*, vol. 113, pp. 515–527, Dec. 2019.

[27] R. Nageem and R. Jayabarathi, "Predicting the power output of a grid-connected solar panel using multi-input support vector regression," *Procedia Comput. Sci.*, vol. 115, pp. 723–730, 2017.

[28] X. Qing and Y. Niu, "Hourly day-ahead solar irradiance prediction using weather forecasts by LSTM," *Energy*, vol. 148, pp. 461–468, Apr. 2018.

[29] DKA Solar Centre. (2020). *Solar Data Download, Location: Alice Springs*. Accessed: Aug. 1, 2020. [Online]. Available: http://dkasolarcentre.com.au/

[30] L. Benali, G. Notton, A. Fouilloy, C. Voyant, and R. Dizene, "Solar radiation forecasting using artificial neural network and random forest methods: Application to normal beam, horizontal diffuse and global components," *Renew. Energy*, vol. 132, pp. 871–884, Mar. 2019.

**KHIZIR MAHMUD** (Member, IEEE) received the Ph.D. degree in electrical engineering from the University of New South Wales (UNSW), Sydney, NSW, Australia. He is a member of the Energy Systems Research Group, UNSW. He is serving as a Section Editor for the *Smart Science* journal (Taylor & Francis). His research interests include smart grid, renewable energy integration to the grid, internet of energy, and virtual power plant.

**SAMI AZAM** is currently a Leading Researcher and a Senior Lecturer with the College of Engineering and IT, Charles Darwin University, Australia. He is actively involved in the research fields relating to computer vision, signal processing, artificial intelligence, and biomedical engineering. He has number of publications in peer-reviewed journals and international conference proceedings.

**ASIF KARIM** is currently a Ph.D. Researcher with Charles Darwin University, Australia. His research interests include machine intelligence and cryptographic communication. He is currently working towards the development of a robust and advanced email filtering system primarily using machine learning algorithms. He has considerable industry experience in IT, primarily in the field of software engineering.

**SM ZOBAED** received the M.S. degree in computer science (CS) from the University of Louisiana at Lafayette (UL), USA, in 2019, where he is currently pursuing the Ph.D. degree. Before joining UL, he worked as a System Engineer with Huawei Technologies for a period of two years. His research interests include cloud computing, confidential data processing, and natural language processing.

**BHARANIDHARAN SHANMUGAM** is currently a Research Intensive Lecturer with the College of Engineering and IT, Charles Darwin University, Australia. He has a large number of publications in several different journals and conference proceedings. His research interest includes cybersecurity.

**DEEPIKA MATHUR** is currently a Research Fellow with the Northern Institute, Charles Darwin University, Australia and is based at the Alice Springs campus. Her area of research is examining ways regional towns can be made more sustainable and healthy through the built environment. In particular she has been conducting research on minimizing construction waste generation and ways of recycling and reusing this waste in regional towns, such as Alice Springs.

• • •