Review on forecasting of photovoltaic power generation based on machine learning and metaheuristic techniques

Muhammad Naveed Akhter1,2, Saad Mekhilef3,4,5, Hazlie Mokhlis6, Noraisyah Mohamed Shah5

1Power Electronics and Renewable Energy Research Laboratory (PEARL), Department of Electrical Engineering, Faculty of Engineering, University of Malaya, 50603 Kuala Lumpur, Malaysia
2Department of Electrical Engineering, Rachna College of Engineering and Technology Gujranwala 52250 (A Constituent College of University of Engineering and Technology Lahore), Pakistan
3Center of Research Excellence in Renewable Energy and Power Systems, King Abdulaziz University, Jeddah 21589, Saudi Arabia
4School of Software and Electrical Engineering, Swinburne University, Victoria, Australia
5Department of Electrical Engineering, Faculty of Engineering, University of Malaya, 50603 Kuala Lumpur, Malaysia

E-mail: saad@um.edu.my

Abstract: The modernisation of the world has significantly reduced the prime sources of energy such as coal, diesel and gas. Thus, alternative energy sources based on renewable energy have been a major focus nowadays to meet the world's energy demand and at the same time to reduce global warming. Among these energy sources, solar energy is a major source of alternative energy that is used to generate electricity through photovoltaic (PV) system. However, the performance of the power generated is highly sensitive on climate and seasonal factors. The unpredictable behaviour of the climate affects the power output and causes an unfavourable impact on the stability, reliability and operation of the grid. Thus, an accurate forecasting of PV output is a crucial requirement to ensure the stability and reliability of the grid. This study provides a systematic and critical review on the methods used to forecast PV power output with main focus on the metaheuristic and machine learning methods. Advantages and disadvantages of each method are summarised, based on historical data along with forecasting horizons and input parameters. Finally, a comprehensive comparison between machine learning and metaheuristic methods is compiled to assist researchers in choosing the best forecasting technique for future research.

1 Introduction

Among the different types of energies, electrical energy has a major share in this modern day. The demand of electricity is expected to increase due to globalisation and modernisation of the world. In the past, the coal, natural gas and oil were the fossil fuels used to generate electrical energy. Although electricity demand is being fulfilled by these sources of energies, its massive usage has caused huge depletion of fossil fuels and environmental problems [1]. The estimate for the existence of oil, coal and gas reserves is ~35, 107, and 37 years, respectively [2]. The generation of electricity from fossil fuel power plants has caused major pollution in terms of CO2 emission and greenhouse gas (GHG) emission, thus leading to major climate changes around the world. Keeping in view of these facts, use of alternative sources of energy to meet the electrical demand has been explored intensively. Among these alternative resources, renewable energy sources (RESs) have gained major interest globally. Power generation from RESs is environmental friendly, with very low GHG emission, long lasting and less cost than conventional energy sources [3–6].

Among different types of RESs, photovoltaic (PV) is the most dominant, with high penetration rate in energy markets. This is due to the sun's constant solar energy of 1367 W/m² supplied over the atmosphere [3, 7]. Study conducted in [7] estimated that the total amount of power received by the earth from solar radiation at an instant is 1.8 × 10¹¹ MW. This huge potential of sustainable energy that can be generated from solar, has attracted the attention of policy makers, governments, investors, economists and environmental technologists [8, 9]. Thus PV energy has huge potential in both rural and urban electrification [10, 11]. Solar energy can be utilised in several different forms such as in power generation, heating/cooling generation, passive systems and combined power [12–16]. Other advantages of PV systems include low maintenance cost, greater lifetime, more robust and installation cost reimbursement within a specified period. Electrical power generated from PV systems has received global acceptance. This can be seen from the raising trend of installed PV capacity as shown in Fig. 1.

From the data, installed power capacity from PV has increased tremendously from 2 GW in 2007 to 64 GW in 2016. Germany, China, Japan and USA are among the main users of installed PV. Furthermore, according to International Energy Agency (IEA) [3], it is estimated that by 2030, the total installed global solar power capacity will cross 1700 GW.

The PV power production, however, is dependent upon various factors such as climatic conditions, wind pressure, humidity, solar radiation, ambient temperature, and module temperature. The natural variations in the climate may vary these parameters, altering the amount of power produced. The abrupt change in solar power output disturbs the power system reliability, stability, and planning. To avoid such circumstances, an accurate and precise solar power output forecasting is required to ensure the reliability, stability, and quality of the power system. It may reduce the power uncertainty impact on the grid.

In past, mathematical technique has been applied to forecast power generation output from PV. These methods can be categorised into Persistence model and Statistical method. Unfortunately, this technique generally produces low accuracy forecasting and also fails to work correctly with non-linear data. Due to these limitations, machine learning (such as support vector machine – SVM, artificial neural network – ANN, extreme learning machine – ELM) and metaheuristic techniques are of recent interest, with few review papers found related to these topics [3, 18–21].

Machine learning methods can deal with problems that explicit algorithm are not able to solve. The ability to develop a relationship between inputs and outputs, even when their representation is impossible, makes these models suitable for pattern recognition, classification, data mining and forecasting [18]. There are three main groups of solar irradiance forecasting...
methods, i.e. statistical/numerical methods, physical methods and hybrid or ensemble methods. Machine learning models (ANN, SVM, ELM) are grouped under statistical methods. Physical methods consist of three sub-models, (i) numerical weather prediction (NWP) model, (ii) sky imagery model, and (iii) satellite imaging or remote sensing model. Statistical methods are based on historical data and the ability to extract information from the data to forecast time series. The physical methods are based on the interaction between physical state and dynamic motion of solar radiation in atmosphere.

Statistical/numerical methods produce better solar forecasting results for time horizons between 1 and 6 h, while for medium and longer time horizons, physical methods (NWP, remote sensing etc.) become most attractive [22]. NWP models are normally used for 15 days-ahead forecasting. Three types of NWP models are global (entire earth), mesoscale (part of the earth) and regional (specific local region) [23]. Sky imagery models are used for short-term global horizontal solar irradiation (GHI) forecast (6 h) to deal with small-scale variability created by clouds variable motion. It has the advantage of having complete meteorological information for a very short term forecast of future cloud patterns in solar generation facility area [20]. Remote sensing or satellite imaging models are used to forecast the solar irradiation without any need of ground sensors. While metaheuristic methods having a combination of machine learning and physical methods provide more accurate solar forecasting for medium and long term time horizons [24].

A self-adaptive differential evolutionary extreme learning machine model outperformed all nine benchmark models with relative absolute mean error (rMAE) and relative root mean square error (RMSE) 2.6 and 2.3% (for Brisbane) and 0.8 and 0.7% (for Townsville), respectively, for long-term solar radiation forecasting by using remotely sensed MODIS satellite [24].

A support vector regression (SVR) technique with PV power measurements, NWPs and cloud motion vectors (CMVs) irradiation forecasts were developed for 15 min to 5 h ahead PV power forecasting. The SVR combined model (with all inputs) outperformed the individual models. The results found that SVR forecasts proved to be slightly better than the physical modelling approach without applying linear regression but slightly inferior after applying linear regression to physical models due to input data limitation. The proposed SVR model was able to predict with the same accuracy as statistically enhanced predictions of the PV simulation model [25].

Ogliari et al. [26] compared physical and hybrid methods for forecasting PV output power one day ahead, using real data from the existing power plant in Milan, Italy. ANN, combined with clear sky solar radiation provides the best forecasting results (NMAE 5.6%), compared to two deterministic models (NMAE three-parameter 8.5% versus NMAE five-parameter 9.0%).

Weather research and forecasting model combined with multivariate statistical learning method outperformed smart persistence, a climatological forecast and GFS for one day-ahead hourly solar irradiance forecasts in Singapore, by achieving 23% lower root mean square error (RMSE) than smart persistence [27].

In smart grids (grids connected with solar PV plants), the variable and non-controllable nature of solar irradiation production causes critical issues in power system such as voltage fluctuation, reactive power compensation, frequency response, reliability and stability. An accurate solar radiation forecasting is compulsory to compensate these concerns. Metaheuristic techniques have various combinations of machine learning and physical methods can provide better solar forecasting by reducing the forecasting errors (RMSE, mean absolute percentage error – MAPE, mean absolute error – MAE), compared to other methods [20, 23–27].

Machine learning methods have also been used widely in wind power and power system load forecasting. SVM, adaptive neuro-fuzzy inference system (ANFIS) and ANN are some of these methods used for wind power forecasting. Wind power forecasting is the prediction of intermittent wind power generated due to the unpredictable nature of wind, based on different parameters such as wind speed, air density [28–30]. FFNN, SVM and SVR were used for load forecasting. While load forecasting is performed to forecast the random behaviour of electrical loads due to insertion of distributed energy sources and certain emerging technologies in power system [31–33].

In [3], direct forecasting techniques for PV power generation have been reviewed. While, machine learning methods for PV output prediction have been reviewed in [18] and a brief review of ANN is performed for short-term load forecasting in [19]. Statistical, physical and ensemble methods have been studied as solar PV generation forecasting methods in [20]. Antonanzas et al. [21] highlighted some issues related to forecast planning. From these reviews, it is clear that there are no comprehensive analysis and comparative discussion on machine learning together with metaheuristic techniques for PV power output forecasting. To fill this gap, this paper presents a comprehensive literature review on machine learning and metaheuristic (hybrid) methods. A brief comparative discussion is also provided at the end. This review can be used as a guideline and potential research opportunity for future researchers, working in the areas of smart grids and smart buildings.

Section 2 elaborates on the types of PV power output forecasting methods based on time horizons. Section 3 provides a discussion on mathematical techniques. Machine learning techniques ANN, SVM and ELM are critically analysed in Section 4 with support of comparison tables. Section 5 provides a critical analysis of metaheuristic techniques together with a comparative table of hybrid techniques. A comprehensive comparative discussion is performed for all these techniques in Section 6, together with the advantages and disadvantages of each technique. Finally, Section 7 concludes with the major findings of this paper.

2 Forecasting time

The time duration for which the forecasting of PV power output is performed is known as the forecasting horizon. In the forecasting horizon, the time duration is the main factor that determines its classification. Proper selection of time horizon is compulsory before the design of the model, to maintain the accuracy of PV forecasting at an acceptable level, as the forecasting accuracy is highly sensitive to the forecast horizon. Prediction intervals (PIs) and confidence intervals (CIs) are valuable tools to minimise the effect of weather uncertainties, hence improve the efficiency of forecasting models. CIs represent statistical intervals, calculated from existing data. While PIs develop upper and lower bounds of future realisations for a random variable with corresponding coverage probability. PIs point out prediction values and provide information to make decision makers ready for best and worst cases in future [22, 34]. In recent literature, several methods are used to generate PIs [35] and optimise [36] further in order to enhance their efficiency against uncertainties due to seasonal or geographical variations [37].

The accuracy of forecasting is affected with the change of forecasting horizon, even with similar parameters in the same
model. Rawat et al. [38] analysed the forecasting accuracy of the multivariable neural network (NN) for three types of forecasting horizons (1, 26, 51 h ahead). The forecast error (MSE) for the proposed forecasting model is in the range of 1.068–7.8909 during training session and 0.99253–8.1365 during testing session for the above said forecasting horizons. The forecasting accuracy decreases with the increase of forecast horizon even for the same forecasting technique. Thus, selection of proper time horizon is compulsory before designing of forecasting model, in order to maintain the accuracy of PV forecasting at an acceptable level.

In a multi-time-scale data driven spatio–temporal PV power forecast, average RMSE found was 106.9, 154.8, 163.2, 187.6 for forecasting horizon of 5, 15 min, 1 and 2 h ahead forecasts, respectively [39].

Very short-term forecasting (1 s to <1 h) is useful for real-time electricity dispatch, optimal reserves, and power smoothing, whereas short-term forecasting (1–24 h) is useful to increase the security of the grid. Medium-term forecasting (1 week to 1 month) maintains the power system planning and maintenance schedule by predicting the available electric power in near future. Long-term forecasting (1 month to 1 year) helps in electricity generation planning, transmission and distribution authorities in addition to energy bidding and security operations. Fig. 2 describes the recently mentioned four types of PV forecasting based on time horizon. Table 1 describes the relation between forecasting horizon, forecasting model and related activities.

### Table 1 Relation between forecasting horizon, forecasting model and related activities

<table>
<thead>
<tr>
<th>Forecasting horizon</th>
<th>Intra-hour</th>
<th>Intra-day</th>
<th>Day ahead</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time step</td>
<td>15 min to 2 h</td>
<td>1–6 h</td>
<td>1 day–3 days</td>
</tr>
<tr>
<td>Related to</td>
<td>ramp events</td>
<td>load following forecast</td>
<td>unit commitment, transmission, scheduling, day ahead markets</td>
</tr>
<tr>
<td>Forecasting models</td>
<td>Total sky imager and/or time series</td>
<td>—</td>
<td>satellite imagery and/or NWP</td>
</tr>
</tbody>
</table>

### Fig. 2 Classification of PV power forecasting based on time

![Types of PV Forecasting](image1)

### Fig. 3 Types of PV power forecasting based on historical data

![Types of forecasting techniques based on historical data](image2)

### 3 Mathematical forecasting techniques

Mathematical forecasting techniques can be broadly divided into two techniques: (i) persistence method, and (ii) statistical techniques. Fig. 3 describes both techniques of solar power forecasting based on historical data.
Fig. 4 General schematic diagram for ARMA based prediction [20]

3.1 Persistence model

Persistence model is used as a standard model to test the forecasting accuracy of any proposed model and acts as a benchmark for other models [3]. Only the historical data is required in this model. From previous studies [40–43], other proposed model showed better performance when compared to the persistence model. It is also known as naïve predictor. The persistence model is the mostly used model in forecasting of solar power output for a time span of 1 h [44]. The solar power forecasting output is assumed to be similar to the power value measured on last or coming day [23]. The forecasted output for the next 24 h can be expressed as [45]

\[ P_f(t+h) = P_{pd}(t) \]  

(1)

where \( P_f(t+h) \) is the forecasted power and \( P_{pd}(t) \) is the power output of the day prior to the forecasted day, at the same time. This method is not suitable for forecasting \( > 1 \) h and can only be used for comparative analysis with other advanced techniques [43].

3.2 Statistical approaches

3.2.1 Time series models: Autoregressive moving average (ARMA) model: This model is a combination of two models known as the AR and MA models. It has an important role in forecasting of PV power output from time series data and is represented by the following equation [46]:

\[ X(t) = \sum_{i=1}^{p} \alpha_i X(t-i) + \sum_{j=1}^{q} \beta_j \epsilon(t-j) \]  

(2)

where \( X(t) \) is the forecasted solar output, \( p, q = \) order, \( \alpha_i, \beta_j = \) coefficients, \( \epsilon(t) \) is white noise that produces random uncorrelated variables with zero mean and constant variance [47].

This model is usually applied to auto-correlated time series data. ARMA is a promising tool to understand and predict the future values of specific time series. The stationary time series requirement of the ARMA model is the major limitation [23]. The key fact for its importance that it has the ability to extract the statistical properties and its adoption of Box–Jenkins method. ARMA was used in conjunction with TDNN for 10 min-ahead prediction of hourly solar radiation, resulting in a better forecasting accuracy. Augmented dickey-fuller method was used to find the stationary behaviour of the residual for judgment of different de-trending models. ARMA is used to predict the linear component of trend time series only, with a limitation that it cannot deal non-linear components [48].

ARMA models are very flexible, so it can represent different types of time series by using a different order. The deterministic component is removed to ensure the stationarity of solar irradiance series in case of solar energy forecasting. This is performed by dividing the measured value of solar irradiance at ground \( G \) to the corresponding quantity at the top of the atmosphere \( G_{ext} \) [49]

\[ K_t = \frac{G}{G_{ext}} \]  

(3)

\( K_t \) is the clearness index, which separates the stochastic component of solar irradiance time series. ARMA models forecast the clearness index. A general schematic diagram is shown for ARMA based prediction in Fig. 4.

A combination of ARMA and GARCH model was analysed in short-term solar forecasting to assess the PIs associated with the point forecasts. Furthermore ARMA-GARCH needs lower computational requirement than an ensemble method based on NWP [50]. Applications of ARMA model in solar forecasting energy field have also been discussed in [51].

ARIMA is an extended version of the ARMA model with an added integrated element. ARIMA models can process non-stationary time series data and was used as a reference estimator in forecasting of the global irradiance field [52]. The ARIMA model \((a, b, c)\) of the time series \((X_1, X_2, X_3)\) can be expressed as

\[ \phi(B) \Delta^h X_t = \theta(B) \epsilon_t \]  

(4)

where

\[ \phi(B) = 1 - \phi_1 B - \phi_2 B^2 \cdots \phi_p B^p \]  

(5)

\[ \theta(B) = 1 - \theta_1 B - \theta_2 B^2 \cdots \theta_q B^q \]  

(6)

\( B \) is the backward shift operator, \( BX_t = X_{t-1} \Delta \) is the backward difference, \( \phi_\theta \) and \( \theta_\epsilon \) are polynomials of order \( a \) and \( b \), respectively. ARIMA \((a, b, c)\) model is the combination of autogressive part \( AR(a) \) (see (5)) and integral part \( I(b) = \Delta^b \) and MA \((c)\) is moving average part (see (6)). For stable operation (bounded outputs for bounded inputs), both polynomials are designed in such a way that their zeros lie outside the unit circle. White noise process \((\epsilon_t, \epsilon_{t-1}, \ldots)\) is used for some random values (deviated from the time series average) drawn from a fixed distribution with zero mean and variance \( \sigma_\epsilon \). \( \alpha_h \) is the independent step time variation of the white noise process. After differenting at appropriate step times to remove any trends, AR component is stochastically coupled with MA component [20].

Reikard [53] develops an ARIMA model for forecasting of solar irradiance by applying regression in logs to the inputs of this model. Best results were obtained by using the ARIMA in logs at low resolutions, by producing RMSE (W/m²) of 0.1321, 0.1897, 0.1865 for 5, 15 and 30 min horizons, respectively, where the data was dominated by diurnal cycle. An ARIMA model is analysed on high-resolution GHI (W/m²) data to test its capabilities in Abu Dhabi, by using the performance parameters such as coefficient of determination \( R^2 \) and RMSE. The best fit model showed \( R^2 \) of 88.63% and RMSE of 72.88 W/m² [54].

Another version of the ARMA model is autoregressive moving average with exogenous inputs (ARMAX). It does not rely on solar irradiance as ARIMA, but takes into account the climatic information, unlike ARIMA. Li et al. [55] proposed a time series (ARMAX) model with exogenous inputs (temperature, air pressure, humidity, insolation duration, precipitation amount, wind speed and direction) for one-day-ahead prediction of the power output for a grid-connected PV system, showing better prediction performance in terms of performance parameters (RMSE (125.84 W/m²), MAPE (82.69%), MAD (98.61%)), when compared with ARIMA, RBFBNN and other techniques.

An Autoregressive with Exogenous Input based Statistical model was proposed to enhance the prediction accuracy of solar PV power production. The proposed models utilise both local and geographically correlated information of solar PV production from other sites. From the simulation results, the suggested time-scales for the ST forecast is at 1 h and 2 h ahead by using the real solar
data. For 1 h ahead, MAE of the ST model is 50.79, 41.8, and 5.15% lower than the PSS, BPNN, and AR models, respectively. For 2 h ahead, MAE of the ST model is 60.2, 47.27, and 8.09% lower than the PSS, BPNN, and AR model, respectively [39].

CARD S model: A coupled autoregressive and dynamical system (CARD S) model was developed by Jing, for hourly and intra-hourly forecasting of solar irradiance. Lucheroni [56] presents a model for the power market, which exploits the simultaneous presence of Hopf critical point and noise in two-dimensional non-autonomous stochastic differential equation system for log-price and derivative of log-price. The equation for the dynamic system part, based on the Lucheroni model is given as

\[ R = \frac{1}{3} \sum_{i=1}^{3} R_i \]

(7)

\[ eZ = k(c + R) - \lambda(3R^2 + R^3) - eZ - \gamma R - b + \zeta \]

(8)

where \( k, e, \lambda, \gamma \) are the adjustable parameters and \( \zeta \) is the noise term. In the above equation, \( R \) is the first derivative and \( \zeta \) is the second derivative of \( R \). For de-seasoned solar radiation time series \( R_t \), the following version of the model is used:

\[ R_{t+1} = R_t + \Delta R_t + \alpha_t \]

(9)

\[ z_{t+1} = z_t + \left[ k(z_t + R_t) - \lambda(3R_t^2 + R_t^3) - e z_t - \gamma R_t - b \right] \frac{\Delta t}{\varepsilon} \]

(10)

where \( \alpha_t \) and \( \alpha_t \) are noise terms. The ordinary least square method is used to estimate the other parameters. Fourier series was used to perform de-seasoning. Autoregressive process cannot model the residual series individually, formed by subtracting the Fourier series component from the original series. The reason is less efficiency of AR process to reach the highest values in the series at mean reversion. Hence a resonating model was introduced by Lucheroni [57] plus a proxy for curvature, which provide a much superior fitting to this residual series. After comparing the results of the CARDS model with Kostylev and Pavlovsk [58] from their literature survey, the best performing model for mostly clear days had 17% rRMSE and 32% for mostly cloudy, whereas it was 16.5% for all days for the CARDS model.

Huang et al. [56] developed a combination of AR and dynamical system model for 1 h-ahead forecasting of global solar radiation (GSR) in Mildura Town located in Australia. The forecasting ability of mixture of AR and Lucheroni model, were identified. The proposed model was analysed using MeAPE, MBE, KSI, and NRMSE. The analysis showed that the combined model outperforms the other models, hence selected as the model that was upgraded by including added components, which led to the development of CARDS models. The error analysis shows that the CARDS model has successfully decreased forecasting error MeAPE in the combined model by 33.4%.

3.2.2 Regression method: This method is used to develop a connection between the dependent variable and the explanatory variable. Explanatory variables are required to determine the dependent variable. For example, the solar irradiation forecast is the dependent variable, and climatic variable is the explanatory variable.

Keshetgar et al. [59] investigated the accuracy of different regression methods such as RSM, MARS, M5 model tree and Kriging method for modelling of solar radiation in Adana and Antakya, Turkey. The best MARS model was better than the best Kriging model for solar radiation in the Adana station. In Antakya, Kriging was best compared to all other three methods in terms of RMSE, d, NSE statistics. Model accuracies for estimating solar radiation were improved by including the periodic component in the input variables. The conclusion was that Kriging method was found superior for the modelling of solar radiation than other heuristic techniques such as RSM, MARS and M5 model tree.

Two models, simple and multiple linear regression models were used for PV power generation forecasting. The regression model using two inputs proved to be better than with only one input. Therefore, the requirement for a large number of explanatory variables and a mathematical model is the limitation for this method [60].

Wang et al. [61] developed a regularised partial functional linear regression model (PFLRM) for one day-ahead forecasting of PV power output. The results showed improved performance in forecasting by generating a low RMSE value of 63.1742, MAD (35.0702) and MAPE (0.0886). MAPE of regularised PFLRM is 11.34%, when compared with 20.92% in multi-linear regression (MLR) model and 63.88% in RBFNN.

A multi-linear adaptive regression spline was developed to predict the day-ahead power output of a PV plant in Germany, by using NWP and past historical data as inputs. The forecasted power output has a high correlation with the measured value of 0.706 and relatively low errors (RMSE of 177.8 kW, MAE of 125.9 kW) despite the low number of features and training samples. More accurate results can be obtained by increasing the temporal resolution in the near future [62].

Trapero et al. [63] developed a dynamic harmonic regression model in the state space framework for short-term forecasting (1–24 h) of solar irradiation, illustrated by hourly aggregated time series of GHI and DNI. The DHR dynamic harmonic regression achieved the lowest RMSE of 30 and 47% for GHI and DNI respectively, in comparison with other models.

LASSO (least absolute shrinkage and selection operator) was developed for a sub-5 min solar irradiation forecasting by collecting irradiance time series data for every one second, from a monitoring network in Oahu Hawaii. The proposed model showed better results as compared to univariate models, especially with few training data and many predictors [64].

Guo et al. [65] proposed an ensemble model having a combination of MLR, NN, boosting and random forecast (RF) for day-ahead and week-ahead prediction of PV output power. The results showed improvement in forecasting accuracy with largest RMSE of 7.98, 10.27% and smallest RMSE of 1.00, 4.42% for the day-ahead and weak-ahead forecast, respectively. The forecasting accuracy decreases with the increase of horizon. The problem of ramping up and down was managed by enhancing the sampling rate of historical measurement data.

3.2.3 Regression trees (RTs): RTs are used particularly in machine learning, automatic learning and data mining. In these trees, the target variables are represented by the leaves and branch lines represent the input layer combinations. RT learning methods are dependent on decision trees as a model for prediction. Decisions trees are not for decision purpose, they only represent data. The tree performance is validated by extrapolating its results to the test dataset. There are certain types of classic RT methods such as boosted regression trees and bagged regression trees. Boosted and bagged RTs are used for improvement of classical RTs [22].

Boosting: In boosted RT, there are successive buildings of trees. It converts a combination of weak trees into powerful committee [18]. An extra coefficient is used to represent the weight of trees used for prediction improvement. The prediction is performed by the weighted linear combination of the trees [22]. A quantile gradient boosting method, combined with NWP utilising principal component analysis (PCA), was used for day-ahead hourly forecasting of solar irradiance. The proposed model developed 43% and 39% reduction in MAE and RMSE as compared to the NWP model [66]. The gradient boosted RTs proved to be an attractive method and showed comparable results in terms of RMSE with other conventional forecasting techniques on all forecasting horizons (1–6 h) for multi-site prediction of solar power generation [67]. Huang and Perry [68] proved that the combination of gradient boosting and k-nearest neighbours regression was accurate for probabilistic forecasting of URG solar power for Global Energy Forecasting Competition 2014.

Bagging: In bagged RT, there is no dependency on earlier trees and the bootstrap sampling method is utilised to construct each node. The decision from the majority of nodes will be used as prediction. Voyant et al. [22] used RT methods (normal, pruned, boosted and bagged) for GHI prediction (1–6 h) and estimated
**4.1 Artificial neural networks**

The limitation of statistical techniques in dealing with non-linearity data due to weather variation had led to the application of ANNs in forecasting PV power output. ANNs have three main parts, the input layer, hidden layer and output layers. The input layer takes the input information. The hidden layer analyses the input information, and consists of several layers. The output layer gives the output after acquiring analysed information from the hidden layer [3]. Fig. 5 shows the schematic of ANN architecture and Fig. 6 gives its equivalent model.

Radial basis function (RBF), sigmoid and hyperbolic tangent sigmoid functions are commonly used for forecasting PV power output. The ANN model can be expressed as

\[ U_N = b + \sum_{j=1}^{N} (W_j \times I_j) \]  

where \( U_N \), \( W_j \), \( I_j \), \( b \), \( N \) are the final network output, connection weight, input number, bias weight and number of inputs, respectively. Several different architectures of ANNs has been reviewed in the literature to solve complex systems, such as multilayer feedforward NN (MLFFNN), multi-layer perceptron NN (MLPNN), RBF NN (RBFFNN), recurrent NN (RNN), general RNN (GRNN) and ANFISs [3, 73–76].

**4.1.1 Multilayer feedforward NN:** MLFFNN is a supervised feedforward ANN which consists of more than one hidden layer. Its architecture is less complex because the information do not travel via feedback path. The information takes the straight path only from input to output layer. The selection of hidden layers is dependent on the complexity of the problem [3]. Three multilayer FFNN techniques with BP were developed for designing a model forecaster to forecast the daily average solar radiation in five different cities in Kuwait. ANN1 (gradient descent method) and ANN2 (LM algorithm) with MAPE of 86.3 and 85.6, were proven to be more feasible for prediction purposes as compared to ANN3 having MAPE of 94.75 [77].

Mellit et al. [78] developed a comprehensive comparison of three ANN techniques (AFFNN, RBNN, DRNN) for the short-term hourly prediction of power production in a large-scale grid connected PV plant (LS-GCPV). They concluded that the performance of AFFNN (sunny, overcast) model is best, having a lowest MAPE (1.92%) and highest \( R \) (0.9986) as compared to RBFFNN and DRNN.

An MLFFNN based on the backpropagation (BP) algorithm produced less MAPE (5.9%) for the day before forecasting as compared to MAPE (7.6%) for the day after forecasting of global horizontal solar irradiation [79]. ANFIS is used to forecast non-linear values, in which previous sample information is used to forecast the sample in advance [80]. It has a wide area of application among fuzzy systems because of its robust results, less expensive and transparent. Its performance can be enhanced by proper tuning of the membership function [23].

**4.1.2 Multi-layer perceptron (MLP) NN:** MLP is a branch of FFNN. Among the three layers of nodes (input, output, hidden), each node except the input node is a neuron that uses a non-linear activation function. It differentiates from linear perceptron due to its multilayer structure and non-linear activation function.

Mellit et al. [81] used MLPNN technique for 24 h-ahead forecasting of solar irradiance and showed a good correlation coefficient of 0.98–0.99 for sunny days and 0.94–0.96 for cloudy days. A methodology was developed for the daily prediction of GSR on a horizontal surface by using MLPNN and an ad-hoc time series preprocessing. The proposed approach has reduced forecasting error to 6% as compared to Markov chains or Bayesian inference. Validation of the proposed prediction methodology was done using six other prediction methods. The cumulated DC PV energy for a 6-month period showed great similarity between simulated and measured data (\( R^2 > 0.99 \) and nRMSE <2% on 1.175 kWc mono-Si PV power grid [82].

**4.2 Machine learning techniques**

These techniques consist of three main types, ANN, SVM and ELM as shown in Fig. 3. A comprehensive discussion is given below for all these three techniques.

**Random forests:** Random forests put an extra layer of randomness to bagging. The growth of each regression tree is different, although there are equally divided samples from the given data. Random forests enhance the robustness of the model and diminish the over training risks [69]. The main points of random forests are (i) random feature selection, (ii) bootstrap sampling, (iii) out of bag error estimation and (iv) full depth decision tree growing [70]. Random forest approach and bagged regression tree were proved best among eleven models for high weather variability in solar irradiation forecasting for a time horizon of (1–6 h) [69]. Benali et al. [71] found that the RF method has the highest efficiency with the data used in the study during the summer season. RF produced less MAPE (5.9%) for the day before forecasting as compared to other models such as AFFNN, RBFFNN, DRNN and AFFNN. RF can be used as a benchmark to assess the performance of other models.

**ARMA:** ARMA requires stationary time series [23] and is able to handle non-linear systems to a certain extent [52]. ARMAX is also an extended version of ARMA using meteorological variables as an exogenous input [55]. Regression trees have been used widely for weather variability in solar irradiation forecasting for a time horizon of (1–6 h) [69]. Benali et al. [71] found that the RF method is the most efficient in comparison with smart persistence and ANN2 (LM algorithm) with MAPE of 18.98, 6.38, 2.86%, respectively [72].

**ANN model**

- **Good prediction bands for Ajaccio (France) with mean interval length (MIL) of 113 Wh/m², 70% PI coverage probability and lower (0.9) gamma index value.**

- **Random forests:** Random forests put an extra layer of randomness to bagging. The growth of each regression tree is different, although there are equally divided samples from the given data. Random forests enhance the robustness of the model and diminish the over training risks [69]. The main points of random forests are (i) random feature selection, (ii) bootstrap sampling, (iii) out of bag error estimation and (iv) full depth decision tree growing [70]. Random forest approach and bagged regression tree were proved best among eleven models for high weather variability in solar irradiation forecasting for a time horizon of (1–6 h) [69]. Benali et al. [71] found that the RF method is the most efficient in comparison with smart persistence and ANN method, with an nRMSE from 19.65 for h + 1 to −27.78% for h + 6 for GHI; an nRMSE from 34.11 for h + 1 to −49.08% for h + 6 for BNI; an nRMSE from 35.08 for h + 1 to −49.14% for h + 6 for the hourly prediction of solar irradiation. Random forests were used along with the firefly algorithm for prediction of hourly GSR and showed minimum RMSE, MAPE, MBE of 18.98, 6.38, 2.86%, respectively [72].

**3.3 Summary of mathematical techniques**

Mathematical techniques have been used for the solar irradiation forecasting of linear systems and data patterns. Persistence model is used as a benchmark to assess the performance of other models. ARMA requires stationary time series [23] and is able to handle non-linear systems to a certain extent [52]. ARMAX is also an extended version of ARMA using meteorological variables as an exogenous input [55]. Regression trees have been used widely for boosting and bagging purpose in the latest articles. Boosting and bagging phenomena are used to enhance the performance of classical regression trees [22, 67].

The forecasting accuracy of these techniques however decreases with the increase in forecasting horizon. Therefore, these techniques are suitable for short-term forecasting horizons [53, 65].
4.1.3 Radial basis function NN: RBFNN uses a RBF as an activation function. RBFNN is a bilayer NN. There are two stages of learning process based on synaptic weight [83]. RBFNN has achieved a good performance accuracy, with more computational speed for learning, less computing power and time [84]. Structural simplicity and universal approximation property are the main reasons for choosing this technique [76]. It is strictly limited to exactly one hidden layer called feature selection. RBFNN was used for a 24-h-ahead forecasting of power generation in an experimental system using input parameters such as daily air temperature, mean daily wind speed, pressure, mean daily relative humidity, mean daily solar irradiance and mean daily power output of the system. RBFNN was found suitable for sunny and cloudy days with a very high correlation coefficient (sunny (99.39%), cloudy (99.48%)) and very low MAPE (sunny (8.29%), cloudy (8.89%)) [75]. In [76], RBFNN outperformed the result of FFNN in some months for 24 h-ahead forecasting for PV system power output based on insolation prediction.

4.1.4 Recurrent NN: It is a type of NN in which directed graph is constructed based on connection between nodes. It is utilised in certain tasks such as un-segmented data and speech recognition due to its ability to use internal memory. Different complex structures and computational relationships are learnt well by RNN. It is mostly deliberated for time series forecasting [3]. In [76], RNN outperformed the result of FFNN in some months for 24-h-ahead forecasting of PV power output and is validated by simulation. The RNN method proposed by Yona et al. [85] showed minimum MAE (0.1567 kW) as compared to FFNN for 24 h-ahead forecasting using weather data, fuzzy theory and NN.

Rana et al. [86] presented a cooperative neuro-evolutionary RNN technique for half an hour ahead forecasting of PV power output. The advantage of using this approach is the ability to predict the PV power output directly as compared to other techniques where solar irradiance is first predicted and then converted to power output. The proposed technique has 0.40–35.18% accuracy enhancement over the other three baseline models. Multivariate models also showed better forecasting results (MRE: 7.33%) than univariate models (MRE: 7.45%).

R-DNN was proved best in electricity load forecasting in comparison with FF-DNN. Both techniques showed less MAPE, RMSE for day-ahead as compared to week-ahead forecasting horizon [87].

4.2 Support vector machine

SVM is a supervised learning technique in the field of machine learning theory and structural risk minimisation. It is used to enhance its generalisation capability of by reducing the empirical risk and CI of the learning machine [88]. SVM has a basic principle of applying non-linear data mapping in some spaces and linear mapping in future space. Another method, which is developed to deal with different regression problems, is known as support vector regression. It is based on statistical learning theory and structural risk minimisation method [89].

The kernel functions are key features of SVM, which maps data into higher dimensional space. The non-linear kernel function is defined as

\[ k = \exp \left( -\frac{1}{\sigma^2} \| X - X' \|^2 \right) \]  

(12)

where \( X \) and \( X' \) are the vectors in input space and the vector of features computed from training or test samples, respectively. The targets in higher dimensional space show resemblance with the targets of similar or lower dimensional input space [90].

Jang et al. [91] developed an intra-day (15–300 min) power forecasting model based on SVM. The findings showed that the proposed SVM based model produced better forecasting accuracy, with the lowest RMSE (10.8661), MRE (9.9677%) and the highest \( R^2 \) (0.9104) than NAR and ANN. However, the prediction during the presence of medium clouds was found to be difficult. RMSE and MAE ranges are 5.7367–24.7855 and 5.2825–26.6654% for time horizon 15 and 300 min, respectively. This means that forecasting accuracy decreases with the increase of time horizon.

In [92], three strategies were evaluated for one day-ahead forecast of PV power generation. Strategy 3 (PCA) was best with 10.24 kWh RMSE, 2.7% lower than the one achieved with strategies 1 and 2 had the worst performance with RMSE of 11.16 kWh. It was deduced that the use of PCA technique along with feasible forecasting strategies, increases the forecasting accuracy by having a low RMSE. In [25], SVR with PV power measurements, NWPs and CMVs irradiation forecasts were analysed for 15 min to 5 h ahead forecast horizon based on RMSE for PV power forecasting. SVR based on PV measurements has better accuracy for 1 h-ahead forecast. NWP based predictions were better for horizon >3 h. The combined model is better than all models for all forecast horizons.

4.3 Extreme learning machine

EML has a property of simple training and anti-jamming [93]. The training algorithms of the NN, slows down the learning speed of FFNN. To increase the computational speed, an ELM was developed by Huang et al. [94]. The selection procedure of ELM for hidden nodes is random and determines the output weights of SLFNs. A brief discussion of ELM structure is given as follows. For \( N \) arbitrary distinct samples \((x_i, t_i)\), \( x_i = [x_{i1}, x_{i2}, \ldots, x_{in}]^T \in \mathbb{R}^n \) and \( t_i = [t_{i1}, t_{i2}, \ldots, t_{in}]^T \in \mathbb{R}^m \) are the input and ideal output, respectively. For a standard SLFN with \( L \) hidden neurons and activation functions, \( g(x) \) can be modelled mathematically as shown in Fig. 7

\[ o_i = \sum_{j=1}^{L} \beta_j g(w_j \cdot X_i + b_j) \]  

(13)

for \( i = 1, \ldots, N \)

where \( w_i = [w_{i1}, w_{i2}, \ldots, w_{iL}]^T \) is a weight vector connecting the \( j \)th hidden node and input nodes and \( \beta_j = [\beta_{j1}, \beta_{j2}, \ldots, \beta_{jo}]^T \) is a weight vector connecting the \( j \)th hidden node and output nodes, \( w_j \cdot X_i \) is the inner product of \( w_j \) and \( X_i \), while \( b_j \) is the threshold of the \( j \)th hidden node. For the standard SLFN, we can approximate those targets of similar or lower dimensional input space [90].

\[ \sum_{i=1}^{N} \left\| \sum_{j=1}^{L} \beta_j g(w_j \cdot X_i + b_j) - t_i \right\| = \sum_{i=1}^{N} \left\| o_i - t_i \right\| = 0 \]  

(14)

The compact form of (14) can be written as follows:

\[ H\beta = 1 \]  

(15)
where \( \beta = [\beta^T \cdots \beta^T]_{x \times M} \) and \( T = [T^T \cdots T^T]_{y \times M} \) and the hidden layer output matrix \( H \) [96] is defined as follows:

\[
H = \begin{bmatrix}
\begin{align*}
g(w_1 \cdot X_1 + b_1) & \cdots & g(w_L \cdot X_1 + b_L) \\
\vdots & \ddots & \vdots \\
g(w_1 \cdot X_N + b_1) & \cdots & g(w_L \cdot X_N + b_L)
\end{align*}
\end{bmatrix}_{N \times L}
\]

After the input weights \( \beta \) and biases are assigned randomly, the output vector \( \beta \) can be calculated by the following equation:

\[
\beta = H^{-1}T
\]

5 Hybrid methods

Hybrid methods are a mix of two or more methods together with some optimisation theorem included. This mixing increases the overall forecasting accuracy of the hybrid by incorporating the benefits of individual techniques. From a detailed review, it is clear that in most of the cases a single technique or method is not enough to fulfil the demand for better PV forecasting accuracy and reliability for the system. Several methods were combined by recent research to improve the forecasting accuracy of PV system and showed better results in comparison with the use of a single method [97]. Fig. 4 lists the hybrid forecasting methods available.

The advantages of the hybrid system include the positive points of both methods, excluding their limitations. As a result, the PV forecasting accuracy is better improved as compared to a single method. However, the combination of two or more methods increases the computational complexity of the hybrid systems.

<table>
<thead>
<tr>
<th>Table 2 Summary of ANN methods</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ref</strong></td>
</tr>
<tr>
<td>[38] multiple cases (1, 26, 51 h) ahead</td>
</tr>
<tr>
<td>[99] daily GR and hourly DNI</td>
</tr>
<tr>
<td>[100] 15 min ahead</td>
</tr>
<tr>
<td>[101] hourly</td>
</tr>
<tr>
<td>[79] day-ahead and day before</td>
</tr>
<tr>
<td>[102] hourly</td>
</tr>
<tr>
<td>[87] days and weeks ahead</td>
</tr>
<tr>
<td>[103] 24-ahead</td>
</tr>
</tbody>
</table>
Certain issues that need to be observed include cost, space, structural maintenance, robustness, reliability for proper running of systems and improvement to the PV forecasting accuracy. Another reason is that the performance of the hybrid model is totally dependent on the performance of the individual models. In a mix of two or more methods, if one method has very poor performance, then it will also lower the overall efficiency of the complete hybrid system [3]. Therefore a trade-off exists between the limitations and benefits of a hybrid in improving its forecasting accuracy. A one-time capital cost for the installation of these hybrid systems is required, in addition to a significantly lower maintenance cost, but in the long run, the benefits of PV forecasting accuracy improvement outweighs any limitation.

Hybrid systems also have the capability to improve the performance of complex systems by addressing individual issues and using a combination of best suited techniques. A number of combinations have been used as hybrid methods by different researchers. These combinations are (i) genetic algorithms (GAs) and ANNs, (ii) fuzzy and ANNs, (iii) ANFIS, (iv) ANNs and physical model, (v) ANNs and ARMA, (vi) wavelets and ANNs, (vii) ANN and optimisation algorithms, (viii) WT and SVM, (ix) SVM and optimisation algorithms, and (x) seasonal auto-regressive integrated moving (SARIMA) and SVM [3, 4, 19, 110]. From a comprehensive literature review, these combinations are described as follows.

### 5.1 GAs and ANNs

GA is developed as a computer based technique for search and optimisation purposes, based on natural genetics and natural selection. By using a stochastic and non-deterministic operator, this GA will generate optimal solution at each iteration of the population. This is the reason for its higher efficiency than other optimisation algorithms in searching discontinuous and non-linear spaces. The genetic operations are crossover, reproduction and mutation. Among ANN, most of its applications are in FFNN along with BP algorithm. The main advantage of FFNN is that it does not require a user specified problem solving algorithm [73]. GA together with ANN alone showed better results than other forecasting models in comparison with five techniques for 1 and 2 hour ahead forecasting of the average power output of a 1 MWp PV power plant in California without using any exogenous input such as solar irradiance telemetry, thus the solar panels were the only sensor used to generate the input data. GA/ANN showed an improvement in forecasting by having the lowest RMSE of 32.2 and 35.1% for 1 and 2 hour forecasting horizon than ARIMA, KNN and ANN, but with more expensive gains [111]. Forecasting accuracy has decreased with increase of horizon.

A hybrid approach using a GA for 1 h-ahead PV power output forecasting was reported in [112]. The hybrid model showed higher forecasting accuracy by producing the lowest forecasting error of 5.64, 3.43 and 6.57% in comparison to other methods. However, there is still a need of additional and accurate data to monitor the prediction process for larger data variations of PV power output.

A GA based NN approach was proposed for a distributed PV forecasting method. The weights and thresholds for the BP NN were optimised using GA, which enhances the forecasting accuracy by reducing error [113].

A genetic approach combined with multi-model framework (GAMMF) for the prediction of solar radiation time series was proposed in [114]. From the findings the GAMMF model has shown greater forecasting accuracy in terms of having low SMAPE of 19.60 as compared to ARMA, TDNN and hybrid models. SMAPE value was higher than ARMA only for the range of 35,000–40,000 Whm².

### 5.2 Fuzzy and ANNs

Neuro-fuzzy computing is a combination of NN recognising patterns and fuzzy inference systems (FISs). Fuzzy system contains human knowledge and implements decision making and differentiation. It is very beneficial to mix several methods to give a synergistic way, instead of focusing on one method exclusively [115]. FIS does not need information of the main physical process as a pre-condition for its operation. 24 h-ahead forecasting of PV power output was conducted based on insulation. Yang et al. [116] proposed one-day ahead hourly forecasting using a hybrid method (SOM + LVQ + SVM + FIS) for the prediction of PV power output based on one-year weather information collected from Taiwan Weather Central Bureau (TWCB). The results showed better prediction accuracy (3.295% MRE and 350.2 RMSE (W)) for the proposed hybrid scheme as compared to ANN (5.412% MRE and 529.2 RMSE (W)) and SVR (4.017% MRE and 402.5 RMSE (W)) for weather types other than sunny days (similar performance), with worst results on typhoon days (August 3 and 4).

### 5.3 Adaptive neuro-FIS

ANFIS is a combination of FIS and BP NN learning algorithm. It splits the previous knowledge into certain subsets in order to reduce the search space and BP NN is used to adjust the fuzzy parameters. It is similar to the fuzzy network with distributed parameters. Its structure consists of five layers named as fuzzy layer, product layer, normalised layer, de-fuzzy layer and total firing strength. The fourth layer normalise the firing strengths. Finally the fifth layer combines all the inputs from de-fuzzy layer and converts fuzzy data into the final output. Olatomiwa et al. [115] proposed an ANFIS model for estimation or prediction of solar radiation in Iseyin, Nigeria. The parameters considered are maximum mean temperature, minimum mean temperature and sunshine duration. Findings showed that the proposed model

---

**Table 3** Summary of SVM and ELM techniques

<table>
<thead>
<tr>
<th>Ref</th>
<th>Methods</th>
<th>Forecasting horizon</th>
<th>Advantages</th>
<th>Disadvantages</th>
<th>Input parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>[107]</td>
<td>SVM model</td>
<td>not given</td>
<td>SVM with two inputs has 26% accuracy, SVM model with three inputs has 18% accuracy</td>
<td>large data required</td>
<td>air temperature, sunshine ratio, relative humidity and atmospheric water vapour pressure</td>
</tr>
<tr>
<td>[88]</td>
<td>ANFIS, ANN, SVM</td>
<td>daily</td>
<td>SVM better than ANN having lowest RMSE (2.523 MJ/m²/d), MAE (1.76 MJ/m²/d), R² (0.719)</td>
<td>low value of R² with abundance of rainfall or cloudiness</td>
<td>daily air temperature, extra-terrestrial solar radiation, rainfall, precipitation</td>
</tr>
<tr>
<td>[108]</td>
<td>K-means clustering algorithm, SVM regression scheme</td>
<td>1 h-ahead</td>
<td>low RMSE (99.95 W/m²), MRE (5.45%), and high R² (0.8413) Power capacity (743 KW), energy capacity (9404 KWh), decreases cost</td>
<td>a slight degradation in prediction of accuracy of ANN and SVM due to the use of forecasted meteorological data as the inputs</td>
<td>measured meteorological data including the cloud cover as input</td>
</tr>
<tr>
<td>[109]</td>
<td>NNs, SVM</td>
<td>hour ahead</td>
<td>have rMAE (26.15%) as compared with persistence model</td>
<td>2% rRMSE for unstable sky conditions</td>
<td>measured GHI values</td>
</tr>
</tbody>
</table>

---


© The Institution of Engineering and Technology 2019
(ANFIS) has a coefficient of determination $R^2 = 0.8544$ and RMSE = 1.0854 in the training phase, while $R^2 = 0.6567$ and RMSE = 1.7585 in the testing phase. As a result, ANFIS model proved to be the best technique as compared to other empirical models based on $R^2$. The hybrid model developed the ability to learn for fuzzy systems.

The proposed method, which combines the optimised multivariable regression model for sunny weather and bi-level model consisting of an optimised regression model and ANFIS for cloudy weather, outperformed the other methods such as ANN, ANFIS, LSE-regression models and LS-SVM, by achieving an average MAPE of 8.56% and MAE 10.22 in 24 h-ahead forecasting of solar irradiation. FPA outperformed other algorithms in tuning the model parameters and convergence time [117].

5.4 ANNs and physical model

Physical hybridised NN is a combination of ANN and a physical model, known as the clear sky solar radiation model (CSRM). CSRM is a theoretical model used to find the solar radiation with respect to geographical coordinates of specific site. The purpose of CSRM is to limit the maximum available daily solar radiation by computing the time duration between sunrise and sunset of each day. A hybrid model (PHANN) was proposed and compared with the conventional ANN, for a day-ahead forecasting of PV plant power output. Hybrid model PHANN was found to be more accurate by producing NRMSE of 13.4% as compared to 17.4% for ANN. The ANN is statistical method which requires an appropriate size of historical dataset and suitable choice of network parameters [97].

Ogliari et al. [118] developed a model for a day-ahead hourly power prediction, named as physical hybrid neural network (PHANN), which is a combination of CSRM and statistical ANN. The improvement in the results was noticeable, but there exists a chance that even a properly trained network can produce wrong results due to the variation in weather forecast. This issue was addressed in their research by: (i) the use of larger NN (having a larger number of neurons), (ii) the inclusion of tolerance threshold in error assessment, and (iii) the addition of correction factor, to reduce the final error.

5.5 ARMA and ANNs

ARMA is a linear model and deals only with stationary, whereas ANN is able to handle non-linear models. Therefore, a mix of these two techniques combines the specialties of the two techniques. This combination is used to deal with the non-stationary time series with some preprocessing. Without pre-processing it can be less beneficial for many non-stationary problems. The solution to this problem is to convert a non-stationary to stationary one (weak or strong if possible) and then model the remainder as stationary process. Employing the meteorological forecasts of the ALADIN NWP model, a hybrid technique ARMA/ANN was used in the hourly forecasting of global radiation in five sites in the Mediterranean. The proposed model was found to: (i) improved the forecasting accuracy by reducing the NRMSE by 11.3% in comparison to other persistence model, (ii) used CIs to increase the reliability of forecasting, and (iii) make corrections for forecasting the future values of insolation (average nRMSE gain up-to 1.7%), due to its ability to determine stationarity. However the drawback of the proposed method is the increase in cost and complexity [41].

Both ARMA and TDNN were used in a 10-min-ahead prediction phase for a specific day because ARMA is more sensitive, and is used to forecast linear part of series, while TDNN is less stable and is used to predict non-linear component of series. The hybrid technique generated better prediction results by overcoming the drawbacks of individual schemes, especially the stationary time series requirement for ARMA [48].

Azimi et al. [40] proposed a new hybrid technique having a combination of K-means algorithm (for clustering purpose) with a time series analysis and MLPNN for forecasting of solar radiation, where the RMSE has been varied from 58.5 to 93 W/m$^2$ for forecasting horizon of 1–3 days ahead.

5.6 Wavelets and ANNs

Wavelets are used to decompose the sample data sequence of solar radiations into various time–frequency domain components. RNN is used to forecast all domains and a relatively better forecast is achieved by incorporating an algebraic sum. A wavelet RNN was proposed for 2 days-ahead forecasting of solar radiation. The obtained simulation showed a very low mean square error in simulations as compared to hybrid NNs in previous studies [119].

A hybrid model (DA-GRBFN-EPSO-WD) was developed for forecasting of PV power output in Japan. The proposed method significantly reduces errors in comparison with the conventional ANN (MLP, RBFN, GRBFN) and has a minor improvement over the hybrid model (DA-RBFN-PSO-WD) [120].

WT-DCNN hybrid model with quantile regression was used for forecasting of PV power output. The average values of interval sharpness (IS) for proposed approach were −5.04, −6.73, −11.65, and −21.84 for 30, 60, 90, and 120 min forecasting horizons. It was concluded that sharpness of predicted quantiles goes worse with the increase of time horizon [121].

5.7 ANNs and optimisation algorithms

An hour-ahead forecasting of solar power output was done for three simulated PV sites in the state of Florida, using a novel technique (BP-SFLA-ANN). The results showed that WMAPE for the first site was 8.8% in comparison to previous models (WMAPE = 9.57%) and classical ANN (WMAPE = 10.78%). In order to resolve the issue of high computational burden by SFLA, BP is used initially to find the ANN parameters. These solutions are used as the initial values of optimisation processes. It was concluded that the proposed model is able to produce more accurate predictions and faster convergence when BP was employed in the initial step to decide on the starting population of SFLA [122].

WT-GRNN-PSO hybrid model was proposed for solar forecast with bootstrap CIs to quantify the uncertainties. Analysis was performed for four seasons for three different horizons. NMAE was found 6.72% (1 h forecast) and NRMSE was found to be 16.89% (3 h forecast) and 3.33% (6 h forecast). From the observation table, it was clear that both errors (NRMSE, NMAE) increased with the increase of horizon, hence decreasing the forecasting accuracy as a result [123].

5.8 Wavelet transform and SVM

The wavelet transform is used to divide the time series signal into separate components. The continuous wavelet transform (CWT) is the integral (sum) of all the signals over the entire time period. CWT is defined mathematically as a following equation:

$$W(a, b, \psi) = \frac{1}{a^{1/2}} \int_{-\infty}^{\infty} f(t) \psi^*(\frac{t-b}{a}) \, dt$$

where $a$ is the scale index parameter, $b$ is the time shifting parameter known as translation, $\psi(t)$ is a mother wavelet function, $\psi^*(t)$ is the complex conjugate of $\psi$ and $t$ is the time. While discrete wavelet transform (DWT) is described as follows [9]:

$$a = a_0^n, b = n d b_0, a_0 > 1, b_0 \in R$$

where $n$ and $m$ are the integer numbers.

The separated components act as an input for the SVM model. An SVM-WT method was modelled for forecasting of diffused solar radiation using cloudiness index as an input variable, which was correlated with the clearness index. SVM-WT (MABE of 0.5757 MJ/m$^2$, RMSE of 0.9640 MJ/m$^2$, $R$ of 0.9631) showed better prediction accuracy as compared to SVM-RBF (MABE of 1.0877 MJ/m$^2$, RMSE of 1.2583 MJ/m$^2$, $R$ of 0.8599), ANN (MABE of 1.1267, RMSE of 1.3184, $R$ of 0.8392) and third degree empirical model (MABE of 1.2171, RMSE of 1.4548, $R$ of 0.8156). The limitation is that the minimum predicted diffused radiation deviation was high [9].
Mohammadi et al. [124] developed a hybrid (SVM-WT) technique for daily and monthly prediction of horizontal GSR in an Iranian city. From the findings, the hybrid model gave promising forecasting results for daily [MAPE (6.9996%), MABE (0.8405 MJ/m²), RMSE (1.4245%), rMSE (7.9467%), R² (0.9086)] and monthly mean estimation [MAPE (3.2601%), MABE (0.5104 MJ/m²), RMSE (0.6618 MJ/m²), rMSE (3.6935%), R² (0.9742)] as compared to GA, ANN, and ARMA.

5.9 SVM and optimisation algorithms

SVM uses non-linear mapping to correlate data. In direct computation method, the kernel function, denoted by $K$ enables it to operate in a high-dimensional space without computing the data coordinates in that space. There are four types of kernel functions for SVM, namely, linear, sigmoid, polynomial and RBFs. RBF is better known for its stability, computational efficiency and ease of adaptation to optimisation. Different optimisation theorems are used to enhance the efficiency of SVM such as firefly algorithm (FFA), ant colony optimisation (ACO), ensemble empirical mode decomposition (EEMD), feedforward NN (FFNN), evolutionary seasonal decomposition least square (ESDLS), and RBF [25, 125–129].

FFA is based on social dashing behaviour of fireflies in nature. In FFA, there are two main parameters; variation of light intensity and formulation of attractiveness. For optimal design, the objective function is proportional to the emitted light intensity of the firefly. Olatomiwa et al. [125] developed an SVM with the FFA for forecasting of monthly mean daily GSR on a horizontal surface. Findings showed that the proposed model (SVM-FFA) has the best forecasting results such as $[R^2 (0.8024), r (0.8956), \text{RMSE (0.6988)}, \text{MAPE (6.1768)}$ in the training phase and $[R^2 (0.53), r (0.7280), \text{RMSE (1.8661)}, \text{MAPE (11.592)}$ in the testing phase] as compared to both ANN and GP. This model is only feasible for sites with similar climatic conditions.

An SVM model based on ACO for forecasting of short-term power load is established in this research. The proposed method (ACO-SVM) achieved greater forecasting accuracy (1.50% error) in comparison with SVM (2.01% error) and BP NN (2.18% error). This hybrid approach has the ability to overcome the disadvantage of very large data and low processing speed [126].

A new approach having a combination of clustering and classification algorithms was developed to enhance the next-day prediction of hourly GSR. The combination SVM-C and SVM-R was best, having rMAE of 16.7% and RMSE of 23.5%, with independent variables for previous day values of meteorological parameters and rMAE of 15.2% and RMSE of 22.9%. It also has independent meteorological variables for the same day forecast except daily clearness index that relates to the previous day. The estimated value of the forecasting skill was 49.3% greater than previous values but the computing time was large [127].

Mao et al. [128] proposed a new hybrid technique (EEMD-SVM) for a day-ahead hourly output forecasting for large PV plant. The results showed that the proposed method reduced the MAPE (0.0813, 0.118), RMSE (3.95, 5.5) for both normal and abnormal days, respectively, when compared with the traditional SVM [MAPE (0.112, 0.162), RMSE (5.1, 7.2)] and BPNN method [MAPE (0.102, 0.151), RMSE (5.23, 7.9)] respectively. However, randomness has a negative impact on its forecasting accuracy.

A group of two techniques named as LS-SVR and FFNN, trained with LM, was used for the intra-hour (15 min) ahead prediction of solar power output for the PV field located in EMSI School Morocco. The results showed that both methods have a better prediction accuracy but LS-SVR was the better, with MSE MAE, RMSE, $R^2$ and RMSE of 0.0043, 0.047, 0.0653, 0.96, 15.23%, respectively [130].

Lin and Pai [129] analysed an ESDLS SVR (ESDLS-SVR) for the monthly forecast of power output. ESDLS-SVR (DS) model with linear kernel function proved to be superior in its forecasting performance (0.1618 RMSE, 7.8434% MAPE) in comparison with ESDLS-SVR (DT), ARIMA, SARIMA, GRNN and LS-SVR models.

Two models named as ‘SVM-RBF’ and ‘SVM-POLY’ were considered for prediction of GSR. SVM-RBF outperformed ANN and SVM-POLY by producing low RMSE (3.2), high coefficient of determination (0.900), low computation time and has approximately similar performance to ANFIS [25].

5.10 SARIMA average and SVM

ARIMA represents an important example of Box and Jenkins approach to the time series modelling. SARIMA model adds a seasonal component and is used for time series analysis and forecasting. A hybrid method consisting of ARIMA-SVM, was developed for hourly forecasting of a grid-connected PV plant. SARIMA estimated the linear part with the SVR estimated the non-linear part of the produced power. The proposed model produced low NRMSE (9.5678%), NMBE (−0.3552%), MPE (5.1951%) and higher $R$ (0.9905) than SARIMA and SVM, with slightly reduced NRMSE than individual SARIMA and SVR [131].

5.11 Summary of hybrid methods

Hybrid models are a combination of two or more techniques used in conjunction with each other to minimise the forecasting error. These methods have produced the best forecasting results as compared to individual statistical and machine learning techniques for all types of time horizons.

However, accuracy is affected with the increase of forecasting horizons [39, 40, 132, 133]. It is found that the combination of machine learning techniques with some physical techniques (NWP, clear sky, satellite images) produces better forecasting results for long-term time horizon also [24, 25, 133]. An utmost care is required when choosing the individual techniques to form the composite hybrid model. Poor performance of any single technique will affect the total performance of the composite hybrid model, thus limiting the overall forecasting accuracy as a result. As the number of techniques incorporated in the hybrid system increases, it becomes more complex and costly [98]. But enhancement of forecasting accuracy is the prime factor that has a balance between accuracy, computational complexity and cost is required.

Selection of input parameters is also critical. An accurate historical data with higher number of input parameters such as temperature, pressure, humidity, will result in better forecasting results, but complexity increased with the increase of input parameters, so an optimum number of input parameters must be selected based on the strong correlation with PV power output for proper forecasting results [3]. From a comprehensive review, it is clear that by keeping all these constraints within limits, the hybrid models are prime and efficient models as compared to any other individual techniques. A brief summary of several hybrid models is listed with forecasting horizons, advantages and disadvantages in Table 4.

6 Comprehensive comparative discussion for machine learning and meta-heuristic methods

Machine learning techniques (ANN, SVM, ELM) can deal with non-linear systems. MLFFNN is found to be relatively less complex [3, 143]. Several techniques such as Bayesian technique or pruning can be utilised to control the ANN's complexity [18]. GRNN requires high computational power due to its larger size [23]. RBFNN however, has more computational speed for learning, less computing power and time, hence can achieve a good performance accuracy [85]. BP NN is one of the important supervised learning algorithm used for non-linear mapping [144]. Limitations include slow convergence rate, less resistant to oscillations and the problem of easy falling into the local minima [145]. However, it is suitable for solving complex regression problems due to its non-linear mapping function [23].

ANN has several disadvantages such as local minima, huge dependence upon prior knowledge, over-fitting problem and large training data requirement. SVM can overcome the limitations of ANN but is highly sensitive to the parameters selected such as tube radius $(\alpha)$, kernel function parameter and penalty factor $(C)$ [132, 146]. In ELM, random selection of input biases and hidden nodes.
is typical. Thus hybrid models are highly investigated by researchers, with efforts to overcome the limitation of both ANN and SVM by combining the benefits of two or more techniques.

From this comprehensive literature review, it is clear that hybrid models produced higher accurate forecasting results as demonstrated by analysing the RMSE, MAPE, R, \( R^2 \) and MABE as performance parameters. RMSE \([41, 65, 84, 121, 134–136]\) and MAPE \([117, 121, 129, 136, 137]\) values are reduced to <15% for composite metaheuristic models in comparison with other different algorithms is to find the best solution for the hybrid method of interest. Hybrid models with GA was found to produce less error (5.64%) in comparison with individual models \([ARIMA (9.52\%)\), LS-SVM (6.42\%), ANN (7.78\%), ANFIS(7.93\%)] for 1 h-ahead forecasting \([112]\). DCGSO-LASSO has produced better forecasting results \([lowest MAPE (13.247\%), \text{RMSE} (28.058 \text{W/m}^2), \text{RMSE/avg} (6.345\%)] as compared to SVM, GRESH, and LASSO \([136]\). In recent articles \([121, 139]\), greater forecasting accuracy is being achieved by the use of certain combinations of machine learning techniques instead of statistical techniques.

Based upon comparison with respect to forecasting horizon, it is clear that hybrid methods having combinations of individual machine learning methods (ANN, SVM, ELM etc.) better accuracy for all forecasting horizons. While a combination of machine learning method with some physical method (NWP, clear sky, satellite imaging) provide more fruitful forecasting results than the individual methods for long-term forecasting horizons, because physical methods are preferable for better forecasting for long-term horizons \([25, 69]\).

### Table 4 Summary of hybrid models

<table>
<thead>
<tr>
<th>Ref</th>
<th>Forecasting horizon</th>
<th>Methods</th>
<th>Advantages</th>
<th>Disadvantage</th>
<th>Input parameters used</th>
</tr>
</thead>
<tbody>
<tr>
<td>[40]</td>
<td>1 h, 2 h ... 48 h ahead</td>
<td>k-means algorithm + MLPNN</td>
<td>(i) Less processing time. (ii) Reduced RMSE (58.5 W/m²), ( n\text{RMSE} (28.91%)</td>
<td>forecasting accuracy decreases with the increase of horizon</td>
<td>historical solar data</td>
</tr>
<tr>
<td>[73]</td>
<td>not given</td>
<td>FFNN-GA</td>
<td>exponential model shows better results than conventional regression models</td>
<td>increased complexity</td>
<td>latitudes, longitudes, altitudes and LLP</td>
</tr>
<tr>
<td>[65]</td>
<td>day ahead and weak ahead</td>
<td>MLR + NN + RF</td>
<td>smallest RMSE of 1 and 4.42% for day-ahead and weak-ahead, respectively</td>
<td>ramping affects accuracy</td>
<td>historical dataset</td>
</tr>
<tr>
<td>[134]</td>
<td>not given</td>
<td>WD + ANN</td>
<td>RMSE of 7.193, less convergence time (2977 epochs)</td>
<td>not given</td>
<td>solar irradiance, temperature, humidity and wind speed</td>
</tr>
<tr>
<td>[135]</td>
<td>each individual hour</td>
<td>GTSOP + NG + CHL</td>
<td>lower RMSE (53.929), ( n\text{RMSE} (0.845)</td>
<td>accuracy decreases with the horizon increase</td>
<td>temperature, wind speed, wind direction as inputs. Solar data</td>
</tr>
<tr>
<td>[136]</td>
<td>5 days-ahead on daily basis</td>
<td>DCGSO-LASSO</td>
<td>lowest MAPE (13.247%), ( \text{RMSE} (28.058 \text{W/m}^2), \text{RMSE/avg} (6.345% )</td>
<td>week agreement with the actual data</td>
<td>temperature, pressure, relative humidity, solar zenith angle, precipitation, wind speed and direction, GHR PV output power data and GSR, solar irradiance, temperature, humidity, wind speed</td>
</tr>
<tr>
<td>[137]</td>
<td>not given</td>
<td>FFNN + PSO + WT</td>
<td>minimum MAPE (9.17%), 20.93% reduction in forecast error as compared to BPNN and persistence models</td>
<td>variable PV power output due to seasonal behaviour</td>
<td>temperature, ambient temperature, plane of array irradiance, hourly values of PV power</td>
</tr>
<tr>
<td>[138]</td>
<td>day ahead</td>
<td>PCA + LS-SVM</td>
<td>70% reduce the computation time of 70% without dimensionality reduction data</td>
<td>computational complexity is higher than simple method</td>
<td>module temperature, ambient temperature, plane of array irradiance, hourly values of PV power</td>
</tr>
<tr>
<td>[133]</td>
<td>time horizons &lt;7 h</td>
<td>AAKR + clear sky model</td>
<td>90% CI ranging from –5.31 to 5.00%</td>
<td>dependent on microclimatic conditions</td>
<td>solar radiation, produced electrical power average (temperature, atmospheric pressure, wind speed and direction, rainfall, sunshine duration, solar irradiance, humidity)</td>
</tr>
<tr>
<td>[139]</td>
<td>24 h ahead</td>
<td>GRNN + FFNN + MLR</td>
<td>A stepwise regression FFNN model better than all others RMSE, MAE, MBE, correlation coefficient of 2.74, 2.09, 0.01, 0.932, respectively</td>
<td>SS-GRNN has a little higher RMSE than conventional GRNN</td>
<td>solar radiation, produced electrical power average (temperature, atmospheric pressure, wind speed and direction, rainfall, sunshine duration, solar irradiance, humidity)</td>
</tr>
<tr>
<td>[140]</td>
<td>one day ahead</td>
<td>MARS</td>
<td>RMSE (119), MAD (89.8), MAPE (92.6%) in testing stage</td>
<td>In training stage, MARS is at third number after SVR and KN</td>
<td>daily temperature, dew temperature, wind speed, precipitation amount, insolation duration, humidity and air pressure</td>
</tr>
<tr>
<td>[39]</td>
<td>5, 15 min, 1 and 2 h ahead forecast</td>
<td>(ST-ARX) forecast model</td>
<td>reduced RMSE (154.5) and MAE (111.4) than PSS method</td>
<td>spatial temporal method not better than PSS for 15 min forecast</td>
<td>historical data of both local and nearby solar sites</td>
</tr>
<tr>
<td>[141]</td>
<td>1 h ahead</td>
<td>(EMD, EEMD and WD) + hybrid model</td>
<td>RMSE decreases 16.91% for EMD- hybrid model, 14.06% for EEMD-hybrid model and 7.86% for WD- hybrid model</td>
<td>(i) These results are for clear day only. (ii) increased complexity</td>
<td>GHI</td>
</tr>
<tr>
<td>[142]</td>
<td>hourly</td>
<td>linear + empirical model</td>
<td>lowest RMSE (34.86%)</td>
<td>accuracy still needs improvement</td>
<td>solar radiation data from different regions</td>
</tr>
</tbody>
</table>
7 Conclusions

Accurate and precise forecasting of solar power output is a mandatory requirement to ensure the stability and reliability of the grid system. It also addresses the variable seasonal and environmental issues to fulfill global energy demand. From this review, the hybrid techniques were found to produce better forecasting accuracy for solar power output, in comparison to individual machine learning (ANN, SVM, ELM) and mathematical techniques.

Firstly, four types of forecasting models were defined, based on their forecasting horizon. The findings show that the forecasting accuracy of the models is reduced with the increase of the forecasting horizon. Secondly, several mathematical techniques were presented. Findings show that the persistence model is used as a benchmark model and ARIMA has a stationary input data requirement, which is partially reduced by ARIMA. ARMAX is another model with no solar irradiance as exogenous inputs but takes into account climatic conditions, unlike ARIMA.

Then machine learning methods (ANN, SVM, ELM) were explored in detail. It is concluded that ANN methods can handle non-linear systems due to its effective training, but problems of over-fitting, local minima, random initial data, long data requirement for training and increased complexity due to the multi-layered structures, are the limitations. SVM was found to be the best layered structures, are the limitations. SVM was found to be the best technique for sizing photovoltaic systems: a review', 'A review on artificial intelligence based load demand forecasting techniques for smart grid and buildings', 'Short-term forecast of generation of electric energy in photovoltaic systems', 'Solar photovoltaic generation forecasting methods: a review', 'Comparing support vector machines and quantum-beat model for global solar irradiation forecasting: a review', 'Review of photovoltaic power generation of electric energy in photovoltaic systems', 'Estimating solar irradiance forecasting in the tropics using numerical weather prediction and statistical learning', 'An aggregate machine learning approach for output power prediction of wind turbines', 'Investigation of potential hybrid renewable energy applications and solar radiation models', 'Performance analysis of hybrid PV/diesel/battery system using HOMER: a case study Sahabi, Malaysia', 'Energy Concordia, 2017, 144, pp. 322–339

8 Acknowledgments

Authors would like to acknowledge the financial support received from the University of Engineering and Technology Lahore through Faculty Development Program and from the University of Malaya, Malaysia, through FRGS Grant No. FRGS/1/2018/TK07/UM/01/3, Frontier Research Grant No. FG007-17AFR and Postgraduate Research Grant (PPP) No. PG192-2015B.

9 References


[121] Li, Y., He, Y., Su, Y.: et al. 'Forecasting the daily power output of a grid-connected photovoltaic system based on multivariate adaptive regression splines', Appl. Energy, 2016, 180, pp. 392–401


The Institution of Engineering and Technology 2019