Extreme weather conditions, which affect photovoltaic output power, can have a major impact on electricity generated by PV systems. In India, an annual PV power density of 2000kWh/m-2 may be used. Renewable energy (RE) is expected to play a rising part in the nations in coming years. The sun’s radiation is the primary source of renewable energy (RE). With the objective of predicting PV output power with the least amount of error in mind, it is vital to analyse the impact of major environmental parameters on it. The researchers looked at a variety of environmental factors in this study, including irradiance, humidity levels, meteorological conditions, wind velocity, PV global temperature and dust deposition. Countries such as India would gain immensely from this since it will increase the quantity of PV power generated in their national networks. ANN-based prediction models and multiple regression models were used to predict PV system hourly power output. There were three ANN models that predicted PV output power with RMSEs of 2.1436, 6.1555, and 5.3551, respectively, utilising all features using the correlation feature selection (CFS) or relief feature selection (ReliefF) approaches. It is possible to reduce bias to enhance accuracy by employing two distinct bias calculation methodologies, which were applied in this study. For example, the ANN model outperforms linear regression, M5P decision trees and GAUSSIAN process regression (GPR) models in terms of performance.

1. Introduction

Global warming and climate change have prompted a slew of national policies encouraging the use of renewable energy sources. Alternatives to fossil fuel-generated power, such as photovoltaic (PV) energy, among the renewable energy as a greatest advantage. You can have as much as you like, and it does not contain any dangerous contaminants [1, 2]. The Arabian Peninsula receives more than 2000kWh/m-2 of solar irradiation every year. Compared to other renewable energy sources, PV technology has a lot of potential in this location because of the significant amount of solar irradiation
(e.g., wind energy or tidal energy) [3]. Aside from its environmental benefits, solar energy is becoming increasingly popular due to other factors, like its low maintenance costs and noise and pollution-free technology [4]. However, the PV energy system’s power generation is subject to significant uncertainty due to the weather system’s chaotic and erratic character [5]. In light of these uncertainties, it will be more difficult to manage and operate the electric power and energy system, and real-time control performance may be affected [6]. Making the most of a solar power plant depends on being able to accurately predict its efficiency. We all know that solar radiation has a direct impact on the power generated by photovoltaic panels. Wind velocity and dirt deposition have also been shown to impact PV system efficacy, as have other climate variables (such as temperature) [7, 8]. Several recent studies have shown that PV panel dust deposition has a negative consequence on the forecast of PV output power [9]. For PV power forecasting, the authors believe that the right meteorological factors in a specific site can be crucial [10]. When multiple sources of energy are combined into a hybrid energy matrix, PV power prediction can be very useful. Intermittent solar electricity causes difficulties in maintaining system stability when too much of it is added [11, 12]. Predicting future solar power generation can help system controllers improve system stability. With this new knowledge in hand, utility companies may more effectively build switching controllers for hybrid energy systems [13]. The switching controller’s design parameters may be influenced by environmental factors because of the potential impact on solar power generation.

Solar power forecasting and estimation has recently been the subject of several studies [14]. PV plant power generation can be estimated using a variety of methods, including phenomenological, stochastic/statistical, and even hybrid models. System advisor models (SAM) and other deterministic software programmes like PVsyst can be used to estimate the output of a PV plant based on physical facts [15, 16]. This study used a deterministic method to predict the electrical, thermal, and optical properties of PV modules. Studies on PV power forecasting tend to focus on a single point forecast, which is known as a deterministic forecast [17]. PV power data mistakes can be overlooked by deterministic forecasting systems at times. Recent emphasis has been given to probabilistic PV power forecasting models that can mathematically explain this uncertainty. An ensemble of deterministic forecasters is one of the most common ways to generate probabilistic uncertainty [18–20]. Ensemble-based PV power forecasting models could pose a real-time challenge because of their high computing costs. The shallow learning models employed in deterministic and probabilistic PV forecasting are a downside [21]. It’s possible that shallow models will not be able to adequately extract the PV power data’s corresponding nonlinear features and static properties because of the weather system’s complexity [22]. Improved performance may be achieved with further exploration into the deterministic technique of providing high accuracy through the optimization of an artificial neural network (ANN). Because of their function approximation abilities and structured approach, ANNs are effective data-driven techniques that are frequently utilised for nonlinear system dynamic modelling and detection. [23].

The model has been verified for many types of solar modules using hourly solar resource and meteorological data. Multiple linear regressions, neural inference systems with adaptive neuro-fuzziness and support vector machines, for example, do not require any prior knowledge of the system under investigation to function [24]. To “understand” how inputs and outputs are related, they conduct a thorough analysis of a dataset that includes both the input and output variables they have collected. Backpropagation technique seeks to reduce the computational complexity by modifying both weights and biases of the network. [25]. Numerous advantages of statistical learning algorithms can be cited aside from being able to use incomplete data, they can also learn from it. Second, they can generalise and make predictions after they have been trained [26]. Their characteristics allow them to be employed in a variety of situations. To forecast output power from other renewable sources than solar, many machine learning (ML) algorithms have been tested. Individual qualities in datasets are shown to be important by ML, which provides insights into the properties of data dependency. ANN methods were compared by [27] without displaying the prediction model’s features or comparing their relative performance quantitatively. Instead of relying just on PV power, machine learning has been used to estimate solar irradiance [28]. To estimate PV power, some studies used simply a single machine learning algorithm. To model and size a standalone PV plant with a small input dataset, an adaptive ANN was used. In order to represent the PV power system’s many components and their output signals, ANFIS was employed. Estimates of worldwide solar radiation were made using a linear regression model and an artificial neural network (ANN) [29]. A hybrid model can mix several models to overcome the constraints of a single technique. As an extra benefit, “ensemble” methods combine several procedures to build prediction models [30].

There has been a great deal of research into machine learning-based prediction models, but there is always potential for improvement [31]. The following new contributions have been made by the authors in this manuscript:

1. The creation of a test bed for photovoltaic and weather systems, complete with sensors that can be calibrated on-the-fly. The accuracy of the weather data is ensured by this constant calibration of the sensors
2. Two years of PV system deployment yielded a reasonably large dataset of PV and environmental factors. In order to develop and test the prediction model described in this publication, researchers used a real dataset that can be made available to the public upon request
3. ANN-based prediction models are compared to numerous other types of multiple regression models
4. Extensive study of ANNs to discover the most accurate forecasting model with the less Root Mean Squared Error (RMSE). The forecast was also tested to see if it was a biased prediction
Table 1: Features of a PV module made from polycrystalline silicon.

<table>
<thead>
<tr>
<th>Max power at STC</th>
<th>Area (m²)</th>
<th>Voc(V)</th>
<th>Isc(A)</th>
</tr>
</thead>
<tbody>
<tr>
<td>80</td>
<td>0.5914</td>
<td>22</td>
<td>5.42</td>
</tr>
</tbody>
</table>

(5) Investigating the best attributes for effectively predicting PV power

2. Experimentation

An in-house PV system was built and operated to study the link between PV performance and environmental variables. Figure 1 shows the experimental set up, which consists of two subsystems. There were two sub-systems: one on the roof, which had sensors for collecting data, and another in the lab, which stored and plotted data in real time.

To control the terrace’s solar power system, an Arduino Mega 2560 CPU, an XBee or WiFi transceiver, a DC-DC converter, and a programmable electronic resistive load were needed. MPPT was utilised to create pulse width modulation (PWM) signals that simulated fluctuating current and voltage across the load without modifying its real-life resistive value for the programmable electronic load. Using data from calibrated voltage and current sensors, power and current curves for a simulated electronic load were shown. Solar irradiation was maximised while PV output power was increased by using an MPPT controller to adjust the panel orientation. Wireless XBee/ Wifi adapters attached to a workstation were used to capture and log data in the laboratory as part of the rooftop subsystem research. The rooftop sub-system sensors sent data to these wireless adapters on-demand or at fixed intervals. The LabVIEW measurement files recorded from the received data were numerically presented on the workstation screens. An open IoT data platform called Thingspeak was used to make the data accessible to the general audience. The Arduino Mega 2560 microcontroller is used to communicate with an Xbee/EtherMega shield. The PV system’s block diagram can be found here. Table 1 shows the specifications of a polycrystalline silicon photovoltaic module.

2.1. Machine Learning-Based Prediction. Data pre-processing, training prediction models, assessing validation accuracy on the training dataset, and evaluating pre-trained models for the test dataset are the three main stages of smearing ML to any dataset in order to anticipate the unidentified output rates. In the beginning, incongruities like missing data, outliers, and spurious data values were removed from the dataset. Prior to this, the most relevant traits had already been identified and extracted. It was simpler to complete this sub-task by utilising measurements such as temperature and relative humidity, solar panel surface temperature, irradiance and dust build up. It was necessary to split the training and testing datasets
into 85 and 15 percent, respectively, using Matlab’s `scv` partition function. There were 380 training and validation cases, and 95 testing cases. Many machine learning algorithms were taught with the Regression Learner and the Neural Network Toolbox in Matlab’s Statistics and Machine Learning Toolbox.

The performance of a trained algorithm can be enhanced by using an additional feature selection step as in Figure 2. Sifting through high-quality, non-redundant features is an essential step in constructing more accurate learning models. By means to develop the best prediction model with only the properties we required, we experimented with a variety of feature-reduction strategies. It was necessary to create assessment parameters for pre-trained models that performed the best to ensure that a statistical evaluation could be completed. There are ways to tweak this process to improve model quality as additional data is obtained over time.

2.2. Features Selection. An automated data collection system was used to capture, process, validate and test the PV and ambient parameters as depicted in Figure 1. Consider whether all of the environmental elements are necessary, or whether they can be reduced in order to make more accurate forecasts. Feature selection approaches such as CFS and ReliefF were employed to narrow the field of potential contributors. Subset search methods are used to calculate the level of duplication between features in all subsets using the CFS technique’s correlation-based heuristic evaluation function. We then look for subsets having a low connection between them but a high correlation with the desired outcome. It is an instance-based approach called ReliefF that weighs each feature according to how well it can distinguish

### Table 2: Parameters of Neural Networks.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>No of hidden layer when all features were utilized</td>
<td>70</td>
</tr>
<tr>
<td>When CFS filtering was employed</td>
<td>270</td>
</tr>
<tr>
<td>When ReliefF filtering was employed</td>
<td>190</td>
</tr>
<tr>
<td>Training data</td>
<td>72%</td>
</tr>
<tr>
<td>Validation data</td>
<td>14%</td>
</tr>
<tr>
<td>Testing data</td>
<td>14%</td>
</tr>
<tr>
<td>Number of folding in cross validation</td>
<td>5</td>
</tr>
</tbody>
</table>

### Table 3: Information on the predictive model’s input conditions.

<table>
<thead>
<tr>
<th>Environmental parameters</th>
<th>Max</th>
<th>Min</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temperature (°C)</td>
<td>62.04</td>
<td>13.98</td>
</tr>
<tr>
<td>Relative humidity (%)</td>
<td>91.83</td>
<td>26.79</td>
</tr>
<tr>
<td>PV surface temperature (°C)</td>
<td>75.44</td>
<td>9.410</td>
</tr>
<tr>
<td>Irradiance (W/m²)</td>
<td>1142.46</td>
<td>37.64</td>
</tr>
<tr>
<td>Dust accumulation (mg/m³)</td>
<td>1.24</td>
<td>0.061</td>
</tr>
<tr>
<td>Wind speed (km/h)</td>
<td>35.43</td>
<td>0.069</td>
</tr>
<tr>
<td>Power (W)</td>
<td>109.11</td>
<td>0.041</td>
</tr>
</tbody>
</table>
between various classes. ReliefF has the ability to recognise interactions that are more than the sum of their parts. With ReliefF, progressively removing the lowest ranked features allowed us to find the best subset from the ranked features.

2.3. Prediction Models. There are a variety of ML-based predictive models, and each performs differently depending on the dataset. Solar panel power output was estimated in this study using a variety of simple regression and prediction models. A few examples include the SLR, GPR, and M5P regression trees, all of which use a simple linear regression model. It is possible to forecast the output response using regression models that employ simple linear regression. An unobserved point’s value can be predicted using GPR, which uses a Gaussian process and evaluate of point resemblance (the kernel function). In addition to providing an estimate, the forecast also provides information on the degree of uncertainty. A Gaussian distribution with only one-dimensional dimensions. Linear regression functions are used to develop a typical that predicts a target value by learning modest guidelines at the ends of each branch in the typical decision tree. “If...then...else...” statements are considered to be powerful.

PV output power was predicted using an ANN, a standard machine learning technique for classifying and regressing variables. Forward propagation and weight adjustment are both a part of the ANN. The ANN aims to mimic the human brain’s layered organisation by mimicking machine learning. The models of ANNs are artificial neurons, each of which receives a specific amount of inputs. Using these inputs, the activation function generates the activation level of neurons (the neuron output value), and training inputs and output pairs are used to offer learning information. Distinct datasets require different training functions, each with its own set of pros and downsides. The dataset used in this study necessitated a thorough examination of all of the available training functions. ANN-based PV power prediction network configurations are summarised in Table 2. All features, CFS and ReliefF, had distinct optimal no of hidden layers that provided the better mode.

2.4. Bias Calculation and Correction in Prediction. Forecasting bias can be defined as a predisposition to either overestimate or underestimate the outcome of a given situation. In the presence of bias, forecast accurateness can be improved if bias is correctly recognised. Errors in prediction can be reduced or eliminated by making suitable upward or downward adjustments to the projection, depending on whether it’s overly conservative or conservative.

The forecast bias is calculated using two primary techniques in this study:

(i) Tracking signal-based system
(ii) NFM technique

2.5. Signal-based Tracking system. The tracking signal is another typical statistic for measuring forecast accuracy. A forecast’s “Bias” can be measured using the “Tracking Signal.” A forecast that is heavily skewed cannot be used to plan a product. The first step in determining the accuracy of a forecast is to look at the tracking signal. Each period’s tracking signal is calculated:

\[
\text{Tracking signal} = \frac{\text{Actual} - \text{Forecast}}{\text{ABS} (\text{Actual} - \text{Forecast})}
\] (1)

The overall tracking signal is then calculated by adding the figures for each time. Zero is returned for a forecast history that is completely free of bias, and +12 (under predicted) or -12 (over forecast) are the worst possible outcomes for a total of 12 observations (over-forecast). It would be regarded out of control if the predicted history returned a value more than 4.5 or lower than a negative.

2.6. The NFM Method. Bias can be measured with the Normalized Forecast Metric (NFM). NFM uses the following formula to determine bias:

\[
\text{NFM} = \frac{(\text{Forecast} - \text{Actual})}{(\text{Forecast} + \text{Actual})}
\] (2)

This statistic has a range of 1 to 1, with 0 indicating that there is no bias present. Consistently low levels suggest a proclivity to underestimate, whereas consistently high values indicate a proclivity to overestimate. A 12-period forecast is considered skewed toward over-forecasting if the additional values are greater than two. Similar to that, a forecast is deemed biased towards under-forecasting if the additional values are fewer than 2. Corrective steps are needed on a regular basis to keep a forecasting process that is biassed from going off the rails.

Bias correction and change factor techniques can be used to rectify non-stochastic data. As a result, QM eliminates any systematic bias in the predicted output and corrects for statistical downsaling errors.

2.7. Analysis. A variety of numerical examines were conducted to estimate the behaviour of machine learning procedures for the prediction of Photovoltaic output power. The following are some of the measures that can be used to evaluate a product: A number of measures of linear dependency between two variables, such as correlation coefficient (r), mean absolute error (MAE), and mean square error (MSE), were evaluated and averaged: While RMSE is similar to MAE in that it uses an average of the difference squares to obtain the square root and so gives more weight to larger errors, it differs from MAE in that it uses the square root of the MSE rather than the difference squares to find the square root. Predictor performance can be better described using these factors.
MAE, MSE, RMSE, r-value and $R^2$ value were used to compare variant ANNs and regression models. Utilizing the neural network toolbox of Matlab, the Bayesian regularisation backpropagation approach was used after a thorough investigation of the ANN training functions that gave the best prediction of the Photovoltaics. Built-in Matlab method for Bayesian regularisation backpropagation builds a network that effectively generalises by minimising the linear amalgamation of squared errors and weights. Levenberg-Marquardt optimization is used to update the weight and bias variables. There were three main situations tested to determine the optimum ANN selection strategy: using all characteristics, using the CFS technique, and using the ReliefF technique.

$$r = \frac{\text{Cor}(X, Y)}{\sigma_X \sigma_Y}$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |X_i - Y_i|$$

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (X_i - Y_i)^2$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (X_i - Y_i)^2} = \sqrt{\text{MSE}}$$

wherever $X$ is the actual data vector, $Y$ and $Y$ are the predicted and mean of the predicted data vector.

$R^2 = 1 - \frac{\text{MSE}}{\text{MSE (Baseline)}}$
3. Results and Discussion

To gather data on PV and environmental characteristics, as well as PV power output. Data required to develop a prediction model for PV power generation are summarised in Table 3.

Results for the three various methods of ANN-based prediction modelling training and validation are displayed in Figures 3 and 4. Different approaches performed well at different epochs of training, as illustrated in Figure 3, which shows the validation performance. The best epochs were 707, 91, and 598, respectively, among the many models established by the ANN using variant sets of structures. A model for estimating PV power could be developed from this data in the future. Bins are represented by the graph’s vertical bars in Figure 4, which displays an error histogram with 20 bins. (25.84 to 16.88), (49.08 to 26.43) and (43.83-20.03) are the relative ranges of total error for each of the three neural networks. Samples from the associated dataset are represented by the number of vertical bars in each bin. As you can see, over eighty percent of the errors fall inside a 10-watt range. If an algorithm is able to forecast the output with 80% of the error within 10% of the target value, it is considered to be a Excellent predictive model.

Figure 5 indicates the link among the initial power output and the anticipated power output by the best ANN epochs. The ANN’s best linearized predictive model can be used to predict the actual target, and the line shows the better linear relationship between the original output and that model. Correlation coefficient, r, shows the successful linearized model built by ANN employing three distinct strategies, as shown in Figure 5 (dotted line).

After testing a number of properties selection methods, it was discovered that Artificial Neural Network had the least Root mean square error of all of the techniques tested, at 2.1436, while the Relief feature selection method came in second with an RMSE of 5.5351 during validation. The same results were found in the testing dataset. It was found that the best RMSE of 5.4784 was achieved using the all features strategy. On the other hand, it is clear from Figure 6 that the expected production was biased in several cases due to a continuous over- or under-forecasting. It is shown in Figure 6 that all feature-based forecasts have a bias. This shows that tracking
signal-based bias computation is more accurate than NFM in identifying bias.

To see if the ANN-derived models can reliably anticipate the power of a PV system, we utilised assessment data (which was not included in our training data). In order to improve prediction accuracy, the training dataset needs to have more data. The bias correction has been cited in the works by numerous groups, although a different paper demonstrated a

Figure 5: In training and authentication, the relative among the original output power and the Predicted output power: (a) With all features; (b) Correlation Feature Selection; (c) Relief Feature selection.

Figure 6: All features-based Photovoltaic power prediction: Bias calculation for a biased forecast (a) Employing tracking signal method; (b) Employing NFM method.
different performance. The overall computing complexity and cost will dramatically increase as a result of bias correction being incorporated into the prediction method. A CNN can also be employed to remove PV power forecast bias (CNN). A CNN or deep learning technique may be more computationally intensive for real-time prediction than an ANN. The photovoltaic error histogram using test data for various features of predicted and actual is shown in Figure 7.

4. Conclusions

In order to monitor, analyse, and assess PV performance across a wide range of climatic conditions. We were talented to predict how much PV electricity will be created in the upcoming using the system’s PV and environmental data.

(i) Multiple regressions and ANN-based prediction models were built using data from the PV system, in summary. As a means of determining which attributes are most important, we used two feature selection methods (CFS and ReliefF).

(ii) The ANN predicted output power more accurately than the three top regression models with an RMSE of 2.1436. There were reported RMSEs of 6.1555 and 5.5351 when PV power was estimated utilizing feature selection method and the ANN. The output power of PV systems can be accurately predicted using trained ANN models, which are less complex and use less processing effort.

(iii) As a result of this effort, Indian researchers can benefit from the improved set of rules and its presentation in forecasting for this section.

(iv) As a result of this, we believe that the region’s solar industry will gain tremendously. More PV and environmental data is being collected and will be evaluated with the Convolution Neural Network-based technique in the forthcoming in order to train a more accurate predictive model.

(v) Reducing dust build-up and increasing efficiency are two possible outcomes of integrating the chilled

---

**Figure 7:** PV Error Histogram using Test Data: Difference between Predicted and Actual: (a) With all features; (b) Correlation Feature Selection (c) Relief Feature selection method.
PV with the existing PV system. There is still extent for more study in this area, given the five years of system data collected and additional aspects like the cooling and cleaning impacts of the PV system.

**Data Availability**

The data used to support the findings of this study are included within the article. Further data or information is available from the corresponding author upon request.

**Conflicts of Interest**

The authors declare that there is no conflict of interest regarding the publication of this paper.

**Funding**

This research work is not funded from any organisation.

**Acknowledgments**

The authors thank Mizan Tepi University, Ethiopia, for the research and preparation of the manuscript. The authors thank ICFAI University, Odisha University Technology and Research, and Chaitanya Bharathi Institute of Technology, for providing assistance to complete this work. The authors would like to acknowledge the Researchers Supporting Project number (RSP-2021/373), King Saud University, Riyadh, Saudi Arabia.

**References**


