Photovoltaic power forecasting with Long Short Term Memory Algorithm

# Abstract

Renewable energy sources are the future in the energy industry since conventional energy sources are having negative environmental impacts like the depletion of the ozone layer and air pollution. But the big problem of renewable energy sources is their unpredictable and varying nature, and this makes their integration into the grid difficult. But with an accurate forecast of their output power, their integration into the grid can easily be done. For conventional forecasting of the output power of a Photovoltaic (PV) Plant, many input data like historical weather data, weather forecast, and historical PV data are required, but in Nigeria, we have the problem of accessing accurate data, this is the reason why it is needful to develop a model that depends on the readily available data to predict the power output of a PV plant with high accuracy. In this work, we used only the historical PV power output data which can easily be measured for any PV plant. Using feature extraction, we were able to extract many features (like the timestamp feature, aggregation of PV output of the same time for different years) from the data and used them as inputs to train a Long Short Term Memory (LSTM) algorithm. Our result using the LSTM DL algorithm shows more accurate prediction accuracy when contrasted with ANN and other reference techniques.

# Introduction

The importance of electricity cannot be over-emphasized as it gives rise to the economic and technological growth of a country. Countries’ economic growth is often measured with their Electric Power capacity. But the trend has been the dependence on dirty energy sources like hydrocarbon deposits for the production of electric power. These non-renewable energy sources which can be referred to as dirty energy sources posses a great threat to our environment as they lead to pollution and carbon dioxide emission which lead to ozone layer depletion. Diminishment of the ozonosphere raises the level of Ultraviolet (UV) rays reaching the earth’s crust, and these UV rays are hazardous to human health; it causes eye defects like Cataracts, cancer of the skin. A high level of UV radiation affects both lives in water and land, thereby changing food chains, growth, and biochemical cycles. It also affects agricultural yield through its effect on the growth of plants[1]. Due to these negative effects of non-renewable energy sources, the world is looking for other alternatives that could substitute the non-renewable sources; and renewable energy sources (RES) present themselves as alternatives to substitute these dirty energy sources. RESs are gaining ground as we see policies from the governments of the world supporting its penetration into the energy market. The European Union’s goal is to use RES to generate 100 percent of the needed energy by 2050; the Chinese and the United States president’s joint speeches on climate change is a pointer to the fact that policies that will favor the penetration of RES are on the way. With technological advancements and innovation, the coming years will see a more reduced cost of RESs which support their implementation[2].

Except for tidal and geothermal, all other forms of renewable energy sources are derived from solar energy[3]. Solar energy is derived from the solar system, it is the radiation from the sun which can be converted to other forms of energy like thermal energy. The earth’s crust is incident with a mean annual irradiance of 1367 watts per meter [4]. The earth absorbed about 71% of the total radiation it receives from the sun[5]. This amount of energy is enough to supply the global energy demand when properly harnessed.

Solar energy can be collected through two major techniques which include solar thermal collectors and PV technology. Solar thermal collector converts the radiant energy from the sun to longer wavelength radiation (heat) which can be used in the heating of fluids. These heated fluids can be hot water used in the home (as in the case of flat plate solar thermal collectors) or stem used for rotating of turbine for large scale power production (as in the case of concentrating solar power (CSP) collectors). The CSP unlike the PV technology has the drawback of only being suitable for large-scale electric power generation. Solar PV technology on the other hand uses solar cells for the conversion of the radiation from the sun directly to electricity through photovoltaic action. PV technology can be large scale or small scale, their sizes ranging from the ones mounted on the rooftop to gigantic solar PV farms. PV technology can be applied in buildings, vehicles, and solar power plants. Because of this, PV technology is becoming more popular when compared with CSP[6]. The past decade has seen tremendous growth in PV technology, it is estimated that by 2030, the world we hit the capacity of 1700 GW of PV power[2].

However, the major challenge facing the penetration of solar PV into the grid is its variability and unpredictable nature of the solar energy reaching the earth's surface. The output of the solar PV varies from time to time due to the rotation and revolution of the earth around the sun. The intensity of the solar radiation increases from the morning and reaches its peak at midday and starts to decrease till evening when there is approximately no radiation from the sun reaching that particular section of the earth's surface. Due to this variability issue of solar energy, its integration into the existing power system network poses some challenges which include: system instability, increase in cost of running the power system network, power system control problems, etc.

However, to reduce these uncertainties there is a need to forecast the future output power of the PV plant so that the power systems operator will have an idea of decisions to make before time.

# Related works

We explained earlier that we aim to use the DL technique to forecast the output power of a PV for the next hour. In this chapter, we review some key works that have been done in the area of Metrological Models, Statistical Models, ML models, DL models, and Hybrid models.

A technique was proposed by [7] for sub-kilometer, nowcasting cloud solar intensity prediction utilizing a ground-based sky imaging system. They captured images of the sky every 0.5 minutes and used clear sunlight criteria and clear sky library to calculate the sky cover. To determine cloud covers at the plane, they produced a 2D cloud map using coordinate-transformed sky cover, which was then utilized to make the predictions. Two factors influenced the forecasting accuracy the most: forecast horizon and cloud speed. The prediction error was decreased to 50% to 60% of the persistence model’s error in the 30s predictions, according to the results.

Based on several sky imaging systems, [8]created a short-term intensity of solar radiation calculation for a new three-dimensional cloud identification and tracking system. They trained a classification algorithm to detect clouds precisely even to the level of pixel and also the output of cloud mask. Then, using the images taken by the various sky imaging systems, they calculated the velocity of each cloud cover, ready to be merged into bigger views for the solar forecast. Contrasted with reference to the persistence method, the suggested model achieved at least 36 percent enhancement for all solar radiation intensity predictions between 60 seconds and 900seconds intervals.

Another form of meteorological method that is used for the forecast of solar power uses the NWP models, which are based on mathematical integration equations and require knowledge of the field of meteorology to describe the irradiance process and changes in the environment.

It was pointed out in [9] that the NWP is a physical deterministic model and that is its main benefit. The authors have suggested, however, that the model based on NWP is constrained by the non-linear nature of the field equations and also the insufficiency of spatial resolution of the integration network, which is too large from 100 km, to a few km compared to the size of the PV plant.

Statistical methods can be used to render direct predictions for outputs of PV without the need to estimate irradiance in the first place. As described earlier, these methods include ES, ARIMA, and ARMA.

[10] Assessed five non-exogenous input predictive models. They contrasted ARIMA with the permanent model, the k-NN, the NN, and the Genetic Algorithms Optimized NN (GA-NN) and checked the correctness of these techniques utilizing eight months of data. Although the findings revealed that GA-NN outstripped the other comparative approaches used, ARIMA as well gave good results.

[11] Suggested 3 time-series decomposition techniques to predict hourly GHI values. The first model introduced pre-processing technique (additive seasonal-trend decomposition) before the use of ES, which decreases state space and thus increases the efficiency of computation.

ML approaches use ML algorithms to predict PV performance directly. There are usually 2 ways to apply ML techniques: by constructing one prediction model or by grouping multiple forecasting techniques to create a collection of forecasting techniques.

[10] Applied the k-NN technique and proved that it exceeded the persistence technique used for contrast. In [12] they have suggested a new k-NN- approach for forecasting intra-hour GHI and DNI, and also for related uncertainty forecasting horizon. The prediction interval spans from 300 seconds to 1800 seconds and the variables were calculated on the basis of an optimal point algorithm. The findings showed that the suggested technique obtained a 10 percent to 25 percent increase when contrasted with the persistence technique. They have suggested that the optimization of sky pictures could lead to a marginal improvement of around 5%. In [13], the effect of various climate forms on forecasting perforation was studied and the suggested k-NN and neural network-based techniques for predicting global irradiation were studied. Both techniques were optimized using feature extraction techniques and the findings demonstrated that the suggested techniques greatly enhanced persistence techniques. As described earlier, some of the common ML models for prediction include: k-NN, NN, SVM.

A technique for predicting hourly Global Solar Irradiation (GSI) values was introduced in [14]. More precisely, they trained the k-NN technique to pre-process the data before training the NN to predict the GSI value 60 minutes ahead of the target PV plant. The k-NN model uses weather information from eight PV generating units close to one another and produces inputs for the neural network technique utilized in making the predictions. The findings revealed that the hybrid technique obtained 42 watts per meter of MABE and 242 watts per meter RMSE.

NN are the commonly used tools for solar energy forecasting activities [15]–[18]. NNs can be utilized to solve complicated non-linear tasks but involve a wary choice of parameters, and also involve the structure of the NN and the training algorithm [19], [20].

[21] applied a NN model to estimate tilted GSI extracted from horizontal data obtained from Algeria. The technique uses declination, horizontal global extra-terrestrial irradiance at the horizon of 300 seconds, azimuth angle as inputs, and zenith angle. It was tested utilizing 24 months' data, and it showed positive results-the lowest relative RMSE obtained was 8.82 percent.

Besides NNs, SVR [22] is another common ML technique that has been extensively utilized in PV output prediction. NN forecasting techniques and SVR forecasting models are frequently compared. [23]–[26].

[27] suggested 7 SVR techniques with separate inputs for the forecasting of daily solar irradiance. They contrasted the suggested models to five experimental sunshine-based techniques (exponential, linear, linear exponential, cubic, and quadratic) that were developed using data collected from three Chinese locations. SVM models provided an RMSE that is 10 percent lower compared to practical techniques, demonstrating the potential of SVM techniques.

Another way of obtaining an efficient model is to use ensembles of forecasting techniques that join the forecast of different ML techniques. The concept on which this is based is to make use of the strength of each of the ML models involved in the ensemble since each of the ML has both strengths and weaknesses and each has a different situation where they are more suitable. Diversity among each of the ensemble ML models can be generated by varying the forecasting models structure, training input data, and the kinds of forecasting techniques utilized in the ensemble.

[28] suggested an ensemble of neural networks to predict the output power of PV for the next day. They formed six ensembles, each of which incorporates fifteen neural networks. The last forecast was made utilizing BNN. The single neural networks belonged to 3 distinct types: cascade-forward backpropagation, feedforward, and Elman networks had varying numbers of hidden layers and utilized different variants of the backpropagation technique for training. The accuracy was measured utilizing 2 years of Australian historical PV power output data, and the findings showed that the use of the ensembles was advantageous.

There are some researches on combining meteorological, statistical, and machine learning models to set up hybrid forecasting models. This is different from ensembles; hybrid combines forecasting models of any type, while the combination of forecasting of machine learning method only is related to ensembles.

[29] introduced a hybrid architecture that combined ARIMA, SVM, NN, and adaptive neuro-fuzzy inference systems (ANFISs) utilizing GA. Historical sun’s output power, intensity sun’s radiation, and temperature data were obtained and sun’s power data for 3 PV generating units were forecasted. The findings showed that the suggested hybrid setup was highly reliable than the single-prediction model of the hybrid setup, reaching 5.64 percent, 3.43 percent, and 6.57 percent of Normalized RMSE for the three sites, respectively.

Each cutting-edge approach used to forecast solar power has its high points and disadvantages that must be accounted for when implementing these models in real-life scenarios.

Meteorological approaches for forecasting of sun’s power are not direct techniques that depend on weather forecasting variables such as temperature, the intensity of solar radiation, solar angle, humidity, cloud cover index, and wind speed. To forecast solar power, a useful tool is the ability to forecast meteorological variables and weather shifts, for example, cloud movements directly influence PV power output. For this purpose, meteorological models based on satellite imagery and NWP data have been commonly used. The former may be more efficient in forecasting cloud movements, while the latter incorporates global or local weather and climate information that may affect the variability of PV power output.

 However, the drawbacks of metrological models for the prediction of solar power should not be overlooked. Their effectiveness depends on the existence of reliable weather predictions that might be unavailable for the site of PV generating units. The lack of a record of parameters or the lack of precision in the prediction for these parameters could contribute to a substantial reduction in the accuracy of the forecast. This limits the realistic use of meteorological techniques for the forecasting of the sun’s power.

Statistical approaches use algorithms that can be used for direct forecasting PV output power. Compared to meteorological models, they do not depend heavily on whether there are reliable weather predictions, meteorological, or power engineering expertise. Because of this, these models are also used by data scientists and tuned to boost prediction.

However, the majority of statistical methods are more fitting for short-term predictions [30]–[32] such as intraday, because as the intervals are stretched to the next day, their accuracy declines. This often limits the realistic use of statistical models.

Machine learning techniques are also commonly used to render predictions for PV power production. Furthermore, their use does not require deep domain knowledge in meteorology. The versatile forecasting horizon is another feature of the ML models. ML models can be utilized for intra-hour, day-to-day, and month-to-month predictions [10], [33], [34].

Nevertheless, there are many fields of a former application of ML approaches to the sun’s power prediction that can be explored and enhanced. Many of the former research, using machine learning algorithms either neglect to use a DL algorithm that is very efficient for time series data or depended on weather data or does not extract time and seasonal features. These machine learning approaches need to be expanded to include time and seasonal features and to use a machine learning algorithm that is suitable for time series data. We utilized of LSTM DL algorithms to forecast PV output directly and concurrently for the next fifteen minutes.

Due to the unavailability of reliable weather data in many locations and the difficulty of accessing them especially in a country like Nigeria, we used a DL algorithm that is very suitable for time series data and we did some time and seasonal feature extraction, which we used in predicting the future output power of PV plants connected to the grid.

# Machine Learning Algorithm

In this study, we used a DL algorithm. We trained an LSTM DL Algorithm.

## Long Short Term Memory

A diagram showing the layout of an LSTM cell is shown in **Figure 1**. The LSTM is a unique type of RNN. RNN is a NN having recurrent connections between neurons, allowing it to learn from both previous and present data to find a better solution. However, due to gradient vanishing and explosion issues, it is difficult to acquire usable information when two RNN cells are far apart. Memory cells, a kind of neuron, are the answer to this problem. The LSTM can store relevant information for an arbitrary length of time due to these unique neurons[35].

There are 3 gates that direct the flow of data into and out of the neuron’s memory cell. These include the input, forget, and output gate; and all these gates have an activation function. The input will be saved in the memory cell when the input gate detects activation of 1. The stored information will be released to the next neurons if the output gate detects activation of 1. The memory cell will be emptied if the forget gate detects a high activation[36].

Another important feature of LSTM is that to forecast the output at time t, the former n examples have to be propagated through the network. Where ‘n’ is the settling time and it is defined during the set up of the network.

We utilized n =1 former examples to forecast new values in this work,

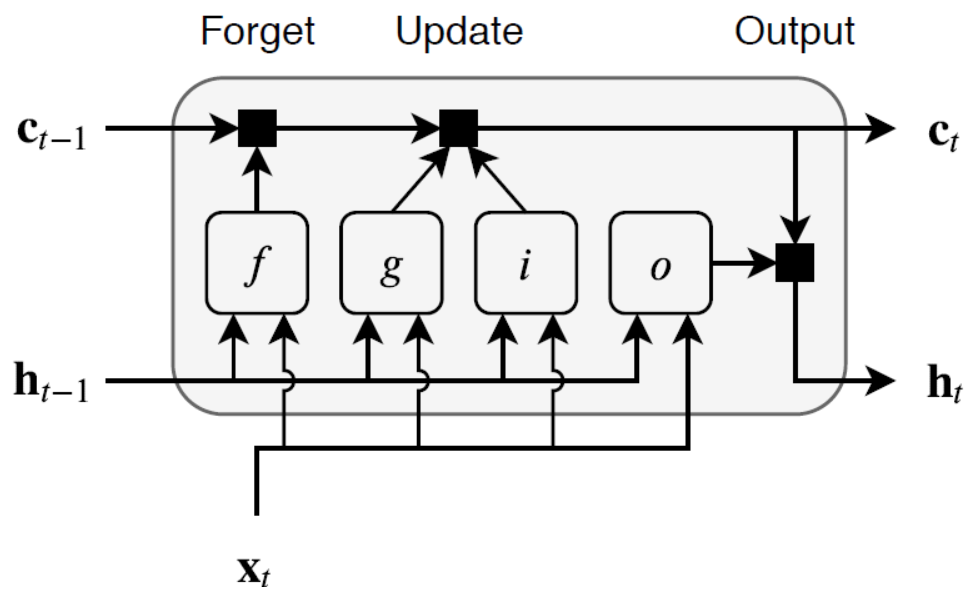


Figure 1 The structure of a long short-term memory cell[37]*.*

## Activation Functions

In this work, we used tanh activation function for the single fully connected layer,

|  |  |  |
| --- | --- | --- |
|  |  | *3.1* |

We use Sigmund as the gate activation function.

|  |  |  |
| --- | --- | --- |
|  |  | *3.2* |

## Data collection

We collected data from the following:

1. **Elia Group Solar Photovoltaic (PV) Power Generation Data**

We collected historical solar Photovoltaic (PV) generation data from Elia Group, a transmission system operator based in Belgium. The data is based on Belgium's historical solar PV data. The data is based on quarter-hour intervals. The monitored capacity increases as more PV generating units are added to the network. At the time of writing this work, the recent monitored capacity is 4787.56 MW. This data contains only the historical power output of the monitored PV plants. The dataset can be downloaded from [38].

1. **Reference Dataset**

There are 21 photovoltaic facilities in the dataset scattered throughout Germany. Their generating units capacity ranges from 0.1 to 8.5 megawatts. PV installations span from residential solar panels to full-fledged solar farms. This dataset contains not only the historical output data but also has weather data. The data can be downloaded from [39]

## Feature extraction

One of the objectives of this work is to show that with the extraction of the right feature, we can obtain a good forecast with high accuracy using the historical power output data of the PV plant. We extracted important features using the timestamp by observing which component of the timestamp affects the power output. We plotted the following:

1. 15 minutes interval power output measurement for some time from January 2013 to December 2020, the graph is shown in **Figure 2**.

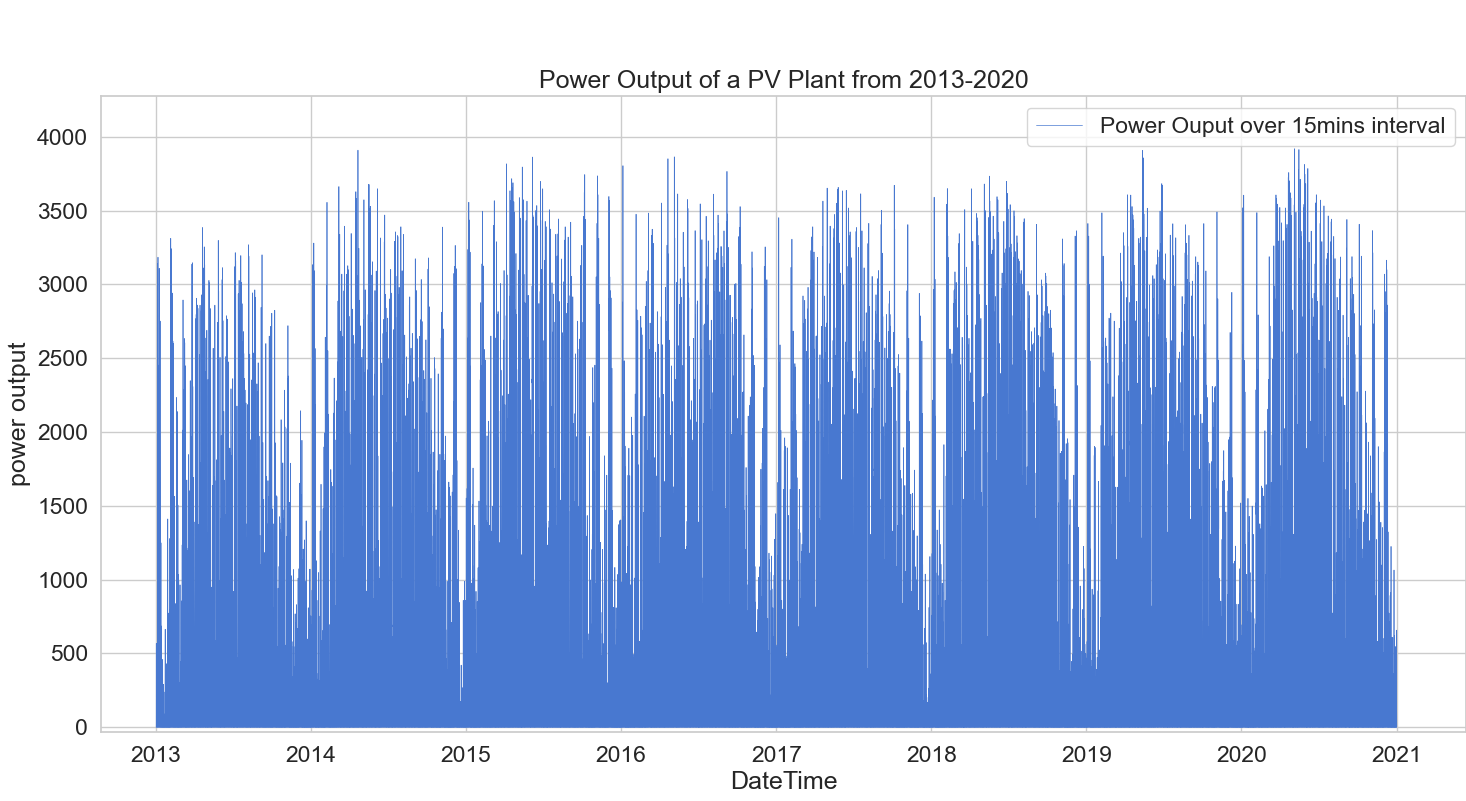
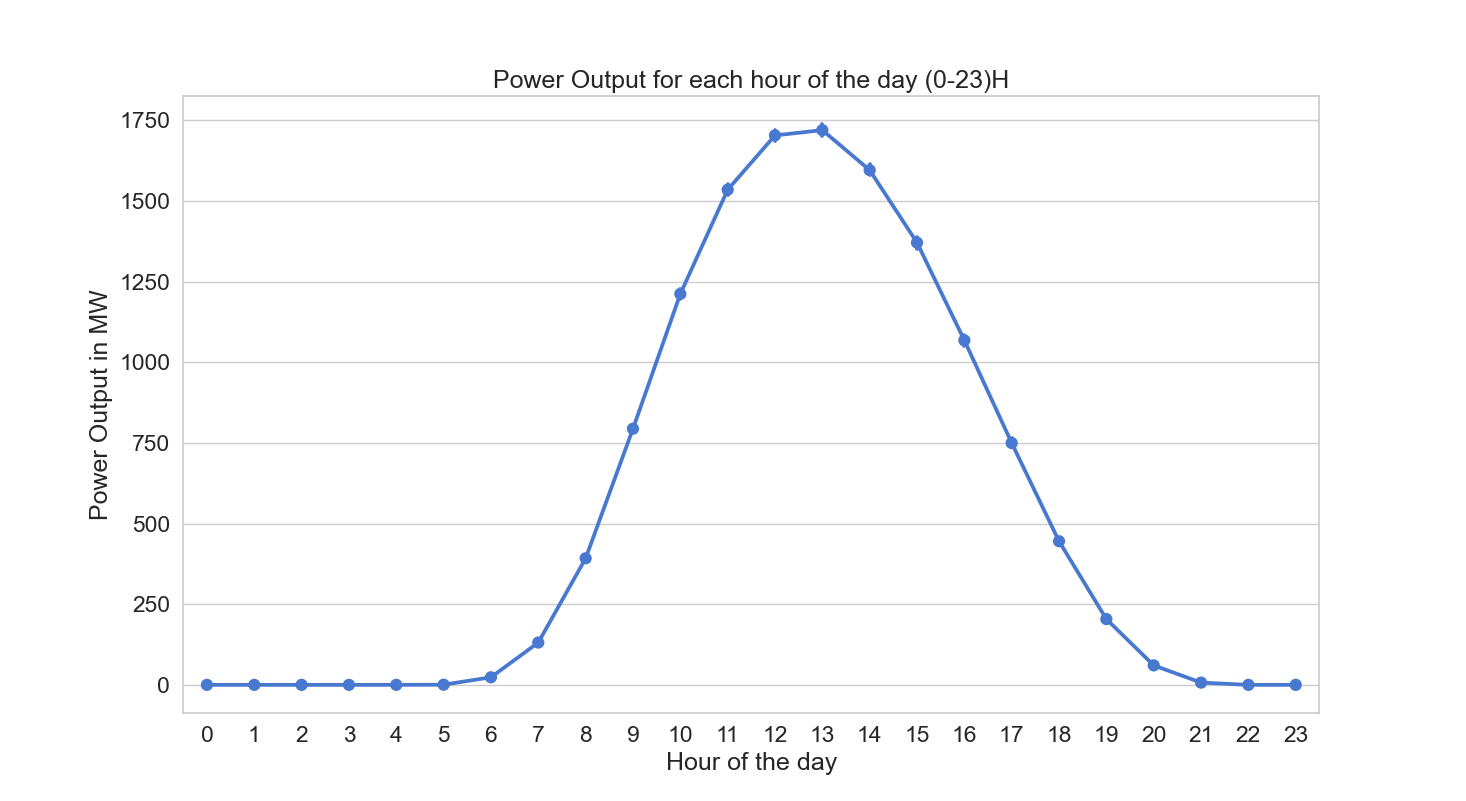


Figure 2 15 minutes interval Power Output Measurement for a period from 2013 to 2020

1. An hour interval power output measurement for some time from January 2013 to December 2020, and magnitude of the output power for each hour in a day (summed across the years) are shown in **Figure 3**. It can be seen that a regular trend can be spotted from the graph in **Figure 3**. This feature shows that magnitude of the power output varies directly with the magnitude of the solar irradiation. During the day when solar intensity is high, the output is also high.

From the Figure it can be seen that the magnitude of the output is within the same range for all the days except on Saturday, this implies that this feature has little or no significance to forecast and was ignored to save training time and reduce the complexity of our model.

Figure 3 Magnitude of the output power for each hour in a day (summed across the years)

1. The sum of the measured power output for each month across 2013 to 2020 is shown in **Figure 4**. The Figure shows seasonality, that output power is higher during the summer than any other season, this is a very key component of our forecasting, our model will forecast based on the season.

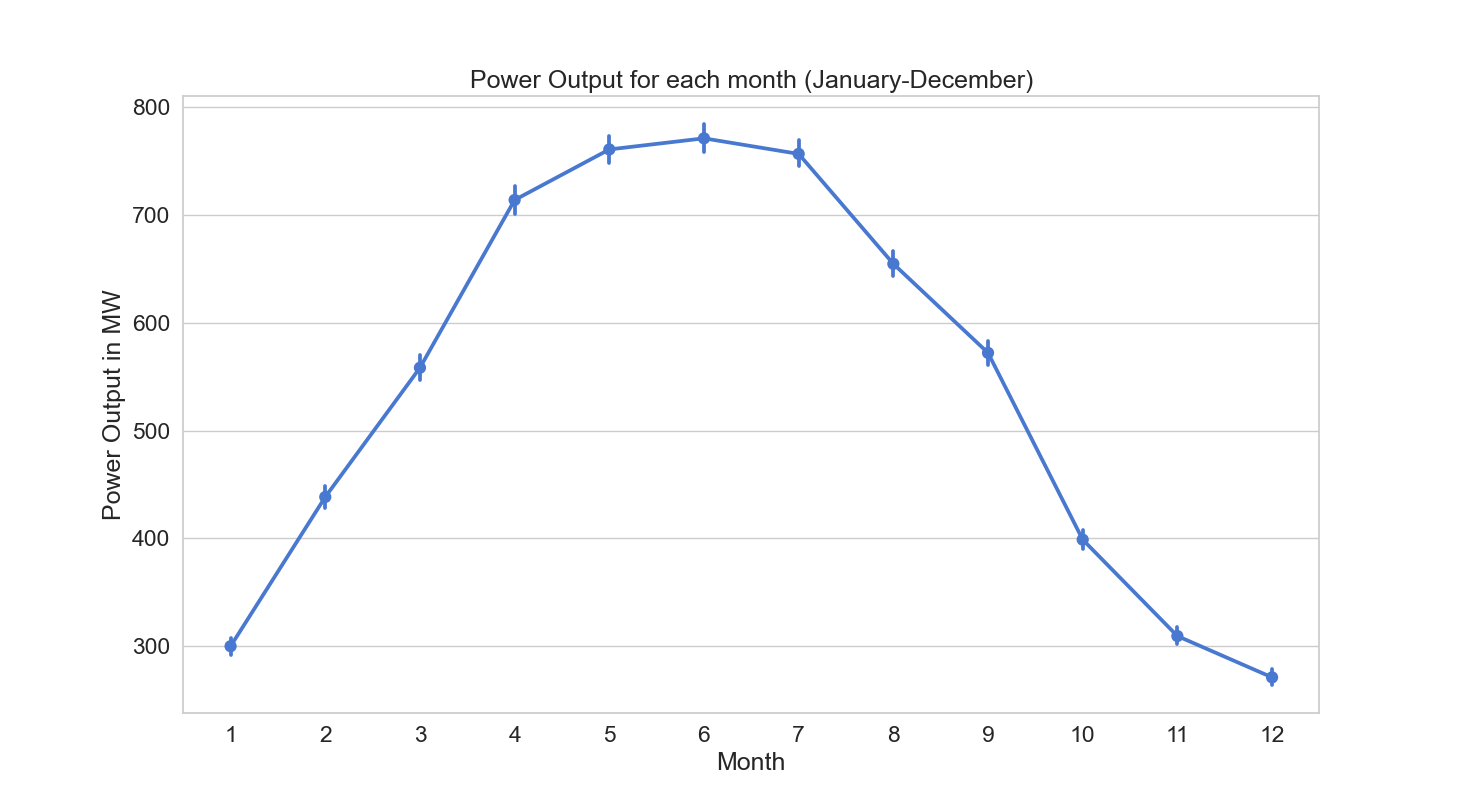


Figure 4 the sum of the measured power output for each month across 2013 to 2020

For German solar farm dataset, it contains 49 features and there is no need to extract extra features.

# Experimental Evaluation

This section will experimentally study to show that extracting the right feature will yield a good forecast which is comparable to the dataset that has multiple data sources (like weather data and weather forecast).

## Reference Model

The reference technique is the LSTM technique trained on the data in [39], in contrast to the model trained on the data in [38], the reference technique is trained on multiple data sources. For the forecast, it uses meteorological variables such as the temperature and the diffuse and direct normal irradiation. Our model uses only the historical PV power output data to train an LSTM algorithm.

## Data Description

Our data contain the historical PV data of the sum of the PV plant integrated into the grid in Belgium. It has a date-time column which has the year, month, day of month, hour, minute, and seconds. The data is based on 15 minutes interval time frame. It also contains the capacity of the plant per each measurement (because the capacity of the PV plant(s) increases with time). For a better prediction, we divided the measured value with the capacity at the point of measurement and multiplied it by a base value (the base value is the maximum capacity of the PV plant over time), i.e.:

|  |  |  |
| --- | --- | --- |
|  |  | *1* |

Our base capacity is the current capacity of the monitored PV plants which is 4787.56 MW. This allows us to compare of the prediction accuracy without accounting for the capacity of the PV plant.

We normalized the output power to be in the range of 0 – 1 using the maximum corrected power output of the PV plant, i.e.:

|  |  |  |
| --- | --- | --- |
|  |  | 2 |

## Error Measures

We used four different error indices to compare the accuracy of our model with the reference model. These include:

1. The root mean squared error: shown in equation 4.3 as

|  |  |  |
| --- | --- | --- |
|  |  | *4.3* |

1. The BIAS: shown in equation *4.4* as

|  |  |  |
| --- | --- | --- |
|  |  | *4.4* |

1. The mean absolute error (MAE): shown in equation 4.5 as:

|  |  |  |
| --- | --- | --- |
|  |  | 4.5 |

1. The average absolute deviation: shown in equation *4.6* as:

|  |  |  |
| --- | --- | --- |
|  |  | *4.6* |

Where the predicted power is output and is the measured power output for N samples.

## Experimental Setup

To evaluate the effectiveness of this technique we performed many experimental simulations of 15 minutes ahead forecast using the two datasets. All the data cleaning, preprocessing, and feature extraction was done with pandas[40] API for python 3.9v. All training was done with Matlab[41]. We implemented the LSTM using 3 layers: one input layer, one hidden layer with 200 neurons, and one output layer.

The training parameters of our model are stated in **Table 1**, we used 90 percent of the data for training and 10 percent for testing, our model is trained using the training dataset. After the training, the errors are calculated using the testing dataset, the network that has the least error is selected.

Table 1 Training Parameters of our Model

|  |  |
| --- | --- |
| Parameter | Value |
| Learning Rate Drop Period | 125 |
| Initial Learning Rate | 0.005 |
| Gradient Threshold | 1 |
| Number of Epochs | 450 |
| Learning Rate Drop Factor | 0.2 |

We trained 50 LSTM models and selected the model with the best performance. The training progress dashboard which has the graph RMSE and the loss function as the training progressed is shown in **Figure 5**.

We trained our reference model with the same training parameters as our model, and these parameters were selected based on the ones with the best performance.

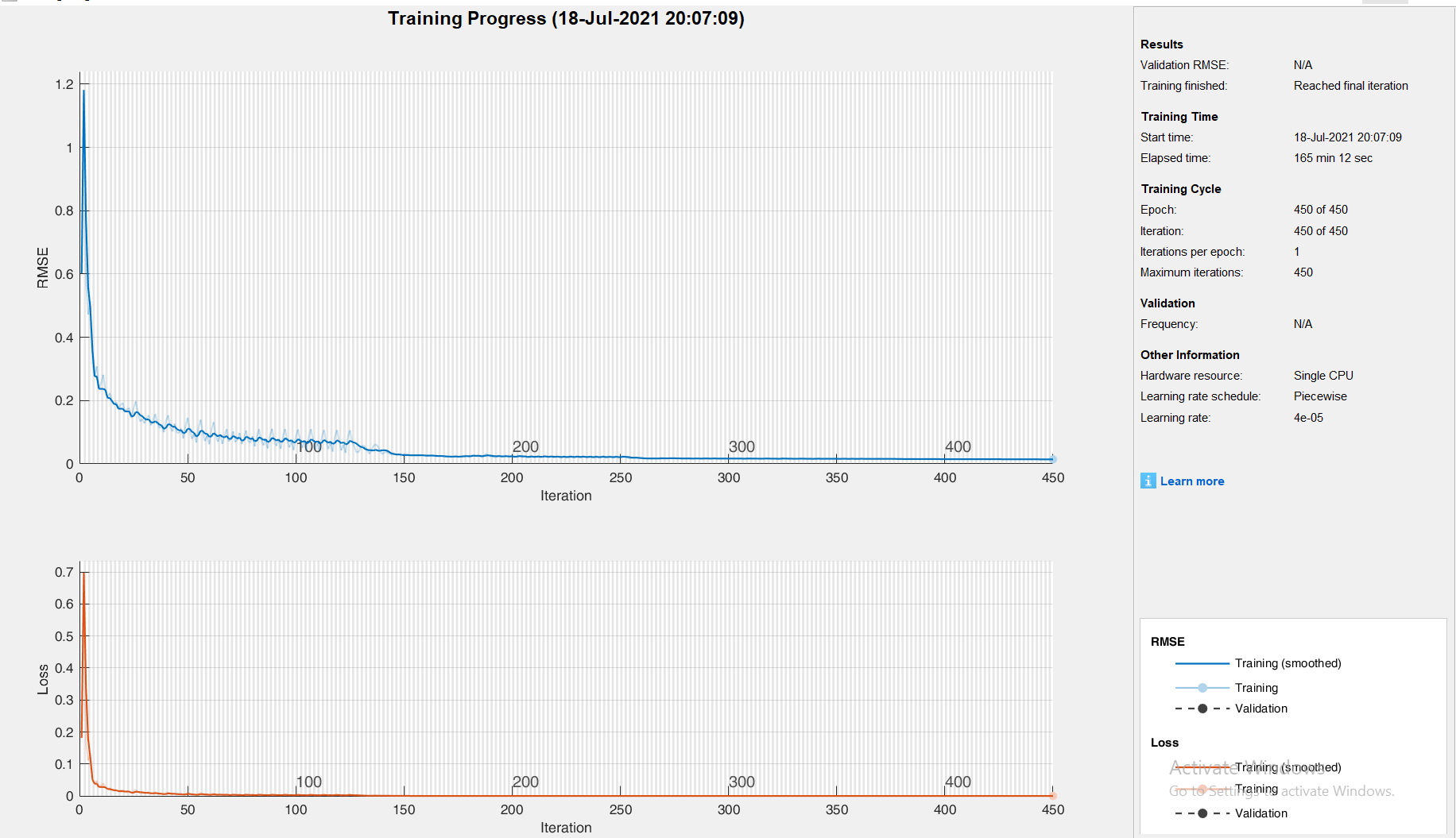


Figure 5 Training progress graph of our model

## Experimental Results

**Table 2** shows the error scores of our model during the experimental simulation; the RMSE, MAE, AbsDev, and BIAS are shown.

Table 2 Error scores of our model and the reference model

|  |  |  |
| --- | --- | --- |
| Error Type | Our Model | The Reference Model |
| RMSE | 0.0075 | 0.0445 |
| MAE | 0.0038 | 0.0368 |
| BIAS | 5.6771e-05 | 0.2786 |
| AbsDev | 9.8781e-07 | -0.0073 |

The graph showing the forecasted values compared with the measured values is shown in **Figure 6**. The measured values are plotted with a blue line while the predicted values are plotted with the red line and it can be seen that the measured and predicted values are almost the same. The difference between the predicted and measured values for each data point is also shown in **Figure 6**.

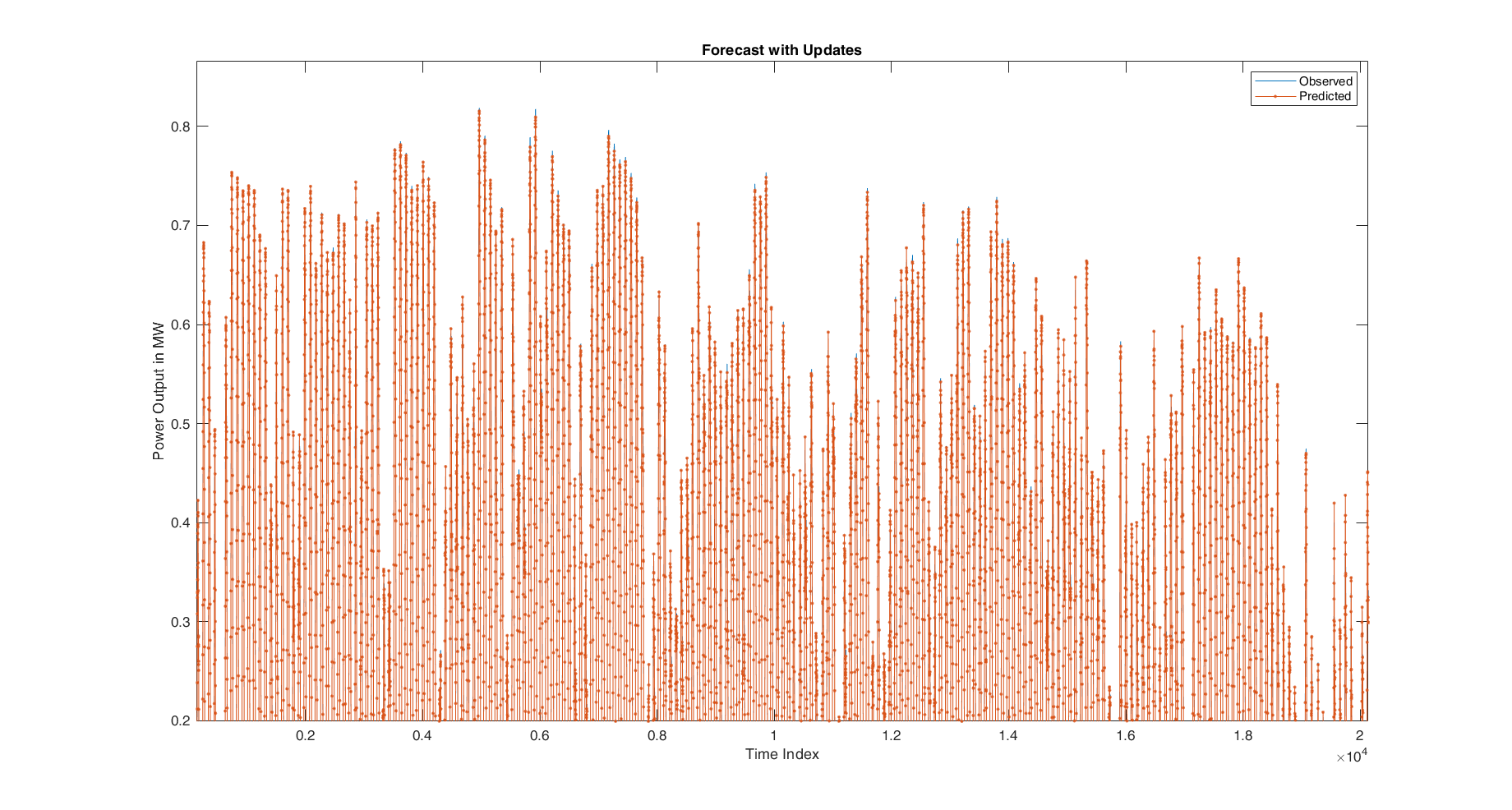


Figure 6 Comparison of the plot of the measured values to predicted values

# Discussion

From *Table 4.2* it can be seen that our model outperformed the reference model, the RMSE our model is approximately 6 times less than the reference model. When the MAE of the reference model is compared to our model, it can be seen that the MAE of our model is approximately 9.7 times less than the reference model. This shows the superiority of our model, and this is a result of selecting the right model with the right feature extraction, although the size of our data contributed to the superiority of our result. This shows that for a particular PV plant, as time goes, the forecasting model becomes better (i.e. is improves with time), which is an advantage. The more the training example, the better the model’s performance.

The RMSE and MAE of our model are 0.0075 and 0.0038, this implies that the RMSE is approximately 1.97 times more than MAE, and as we stated that is if MAE is not far less than the RMSE, the prediction has little deviations to the observed values. 1.97 is not a big number; therefore it implies that our forecasted value only has a little deviation from the observed values, thus showing how accurate our model is.

We were able to obtain a superior result because we chose the right algorithm for the time horizon we are studying, we extracted and chose the right and relevant features, and our right choice of the model parameters.

# Conclusion and Recommendation

We employed in thesis LSTM deep learning algorithm for prediction the output power of PV generating units connected to the grid for 15 minutes ahead, using only the historical PV output data

Conclusively, Integration of PV plant to the grid is possible with the right power output forecast, and this forecast can be obtained with high accuracy even for places with limited weather data if the right algorithm, features, and model parameters are used. And in this thesis, we were able to develop such a model.

In this thesis, we consider short-term solar PV forecasting, and our model did not perform well for long-term forecasting. We used a single deep learning model for our forecasting, but if our model is ensembled with another appropriate DL algorithm, the ensemble will be able to improve the result for long-term solar PV output power forecast.

Our model can be extended to the forecasting of the output power of wind power generating units, and other RES that the power output data are time-series data.