Proceedings of the Fifth International Conference on Intelligent Computing and Control Systems (ICICCS 2021)
IEEE Xplore Part Number: CFP21K74-ART; ISBN: 978-0-7381-1327-2

Solar Energy Prediction using Decision Tree Regressor

Rahul
Department of Computer Science & Engineering
Delhi Technological University
Delhi, India
rahul@dtu.ac.in

Aakash Gupta
Department of Computer Science & Engineering
Delhi Technological University
Delhi, India
aakashg1999@gmail.com

Ankur Bansal
Department of Computer Science & Engineering
Delhi Technological University
Delhi, India
ankur.b1999@gmail.com

Kshitij Roy
Department of Computer Science & Engineering
Delhi Technological University
Delhi, India
roy.kshitij@gmail.com

2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS) | 978-1-6654-1272-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICICCS51141.2021.9432322

***Abstract*—The energy demand is increasing every day. This increase in demand has led us to find alternate sources of energy. The non-renewable sources have limited supply, and constant use gives rise to various types of pollution. Gradually, people started relying on renewable sources of energy due to their abundance in nature. Alternately, Solar Energy (SE) is one of the best sources of energy in India. This paper aims to predict the output production of Solar Power Plants in KWh. Also, this research paper attempts to find the relation of weather attributes with the output power generated. The Decision Tree Regressor model will be used to forecast the power generation of a Solar Power Plant. This model's predictions will help us determine how a solar power plant can be efficiently used to generate a high amount of power. The motivation behind the study was to develop a system that can effectively predict the output of these photovoltaic cells so that orders for supply and demand gaps are placed timely, ensuring the quality of service and cost savings.**

Keywords*— *Machine Learning, Solar Energy, Electrical Energy, Decision Tree, Renewable Energy, Regression, Photovoltaic cells.

I. INTRODUCTION

India has witnessed a rapid demand for non-renewable sources of energy in the past several years. Due to their limited supply and increase in pollution levels, it has been gradually attempted to find alternatives to such resources. One of the best alternatives in India is SE, considering its abundance and tropical climatic conditions. SE is generated with the help of photovoltaic cells. These energy sources highly depend on weather conditions, and they always may not be favorable for the production of electrical energy. In such cases, expensive sources of energy will be required to fulfill its shortfall. Thus, a system will be required to effectively predict the output of these photovoltaic cells so that orders for supply and demand gaps are placed timely, ensuring the quality of service and cost savings.

The efficiency of a photovoltaic cell is also a major concern. A maximum of 75% of the SE to earth can be converted to electricity by a silicon semiconductor, but that is not the case since the PV cells are able to give efficiency up to 33% only. Minimum energy is required by the photon to excite an electron, i.e., to remove it from its crystal structure, known as the band-gap energy. The storage of the generated energy is also a major issue as the current battery technology has a long way to go.

Gujarat Power Corporation Limited (GPCL) is Gujarat state's energy generation and distribution company. When faced with the challenge to cater to the ever-increasing energy demands, GPCL launched the "Gujarat Solar Park" project in December 2010. It was inaugurated in April 2012 in Patan district with Rs.4500Cr budget (Rs.3996Cr for power plant and rest for land acquisition and infrastructure). It is currently Asia's largest and Asia's first Multi-Facility Solar Park. This Solar Park accounts for 10% of India's total solar capacity. The "Gujarat Solar Park" has 224 MW of installed Solar Power capacity while 50MW more is expected to be completed soon and 100 MW of Wind Power. Developing a large cluster of SE unit in one place has resulted in savings of 40% of the cost and 3,42,400 tons of Carbon Emission Reduction[1].

Various machine learning and statistical models have been proposed to predict SE generation, each with its own shortcoming, thus leaving scope for further research since forecasting weather in itself is a challenging task. One way to revamp the accuracy and precision of the prediction is by developing efficient algorithms. The other way is to incorporate more and more relevant weather-related and PV-related parameters in our datasets. In this research, we have touched upon the scope of improvement. Our attempt includes

[1] https://gpcl.gujarat.gov.in/showpage.aspx?contentid=15

the use of a **Decision Tree Regressor** for this regression problem, and our results demonstrate that the level of accuracy obtained is satisfactory. Secondly, to accomplish our objective, we make use of several relevant weather parameters provided in a public forum by Nasa Power Project [2] like the maximum temperature at ground level, precipitation, dew/frost data, wind speed range, surface pressure, etc., in addition to the parameters in our original dataset.

II. Previous Works

Nowadays almost everything depends on the steady supply of electricity. The petroleum derivatives are a vital wellspring of dependable and reliable energy. However, they have significant expenses related to them. They discharge hazardous gases and are dependent upon the vulnerability of the global oil costs. Then again, SE is modest and a clean wellspring of energy that can be created by many countries. However, its reliance on climate conditions makes it less solid. To adapt to this lack of quality, the expectation of SE contributes towards a steady stockpile. The forecast of SE is multidisciplinary research that needs commitment from meteorology, sun-powered cell designing, electrical designing, and AI prediction.

The comparative study of the solar power generation predicting models is mentioned in [1],[2]. Weather data was taken as input to make the predictions in most of the studies conducted. Machine learning methods like Support Vector Regressor (SVR) have been used in conjunction with ensemble technique to further improve on the SVR [3]. Deep learning methods are also being used where we train Neural Networks on our training dataset to yield high accuracy results. Artificial Neural Networks were used with correlation-based feature selection, and compared with other regressive models, the model achieved 2.1436 RMSE (Root Mean Square Error) [4].

The Recurrent Neural Networks have always suffered from the vanishing and exploding gradient problem, which has been solved by modifying them and adding Gate controls; these networks are called LSTM (Long Short-Term Memory). These models were also used to predict SE Prediction [5]. Another type of Neural network, namely, Convolutional Neural Networks (CNN), where we have layers of convolutions and used mostly in computer vision, was used in this research paper [6]. Classification models like K-Nearest Neighbours (KNN) and Support Vector Machine (SVM) are used in Day Ahead Short Term (DAST) PV power prediction for weather classification [7].

Some studies compare these models, such as regression(linear, polynomial), Boosting(XGBoost, AdaBoost, Gradient Boosting), Bagging [8],[9],[10],[11]. Studies have been conducted in Gumi, South Korea, by Lee and Kim [12]. Another study based in Yeongnam, South Korea, was conducted to make day-ahead predictions [13]. They have used meteorological data from the Korea Meteorological Administration (KMA). Their Meteorological data did not include high correlation parameters like humidity, temperature, and solar elevation.

Other than these Machine Learning and Deep Learning techniques, various statistical models have also been used.
It is seen that models based on Numerical Weather Prediction (NWP) systems such as European Centre for Medium-Range Weather Forecast (ECMWF) and the Global Forecast System (GFS) are a good alternative for next day prediction of photovoltaic cells [14]. Daily and monthly solar radiation prediction have also been made with the help of statistical models such as ARIMA (Autoregressive Integrated Moving Average) [15]. Hybrid models combining deep learning and statistical methods have been used and have delivered promising results [16]. Usage of solar angles such as Zenith Angle and Azimuth Angle in conjunction with weather data improve the performance of prediction models [17].
.

III. Methodology

A. Data-set

To obtain accurate machine learning predictions, we need a good quality dataset with appropriate attributes. The dataset which we have chosen is from a 10MW solar power plant of GPCL (Gujarat Power Corporation Limited) spread across 35 acres. The dataset contains the amount of solar power generated in 5 years. The dataset contains daily values of solar power generated starting from 1st May 2015. Feature scaling was also performed on the dataset. Weather is also an important parameter for the prediction of SE output. In order to obtain higher and better accuracy, we have used weather attributes from NASA Power Project [18]. The location of the GPCL power plant was entered, which is situated in Patan District of Gujarat (23°53'57.9 "N, 71°13'21.3" E).

TABLE I. Description of Data-set Attributes

S. No.	*Attributes*	*Format*	*Description*
1	Grid Availability	Numeric	Percentage or factor of grid available for energy procurement.
2	Equipment Availability	Numeric	Percentage of time equipment is used for production
3	Solar Irradiation	Decimal	Solar radiant power incident on a surface per unit area (W-h/m2)
4	Output	Numeric	SE in KWh

The attributes for prediction were Grid Availability, Equipment Availability. Apart from these, Solar Irradiation was also used. These factors helped us to determine the output

[2] https://power.larc.nasa.gov/

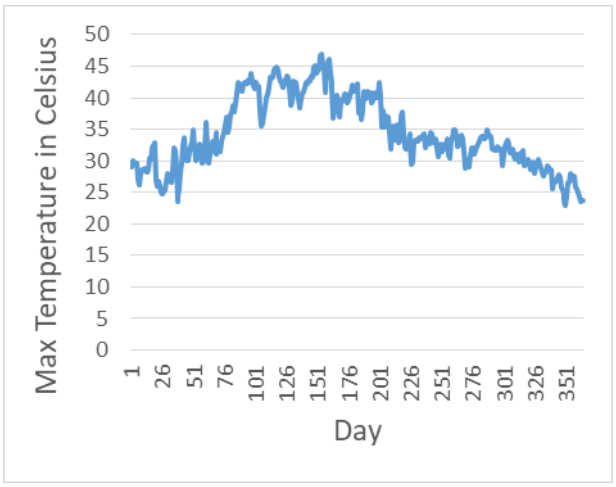

Fig. 1(a) Maximum temperature of each day

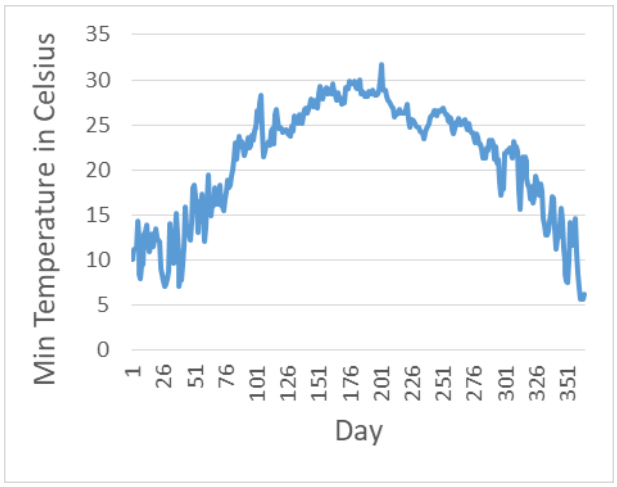

Fig. 1(b) Minimum temperature each day

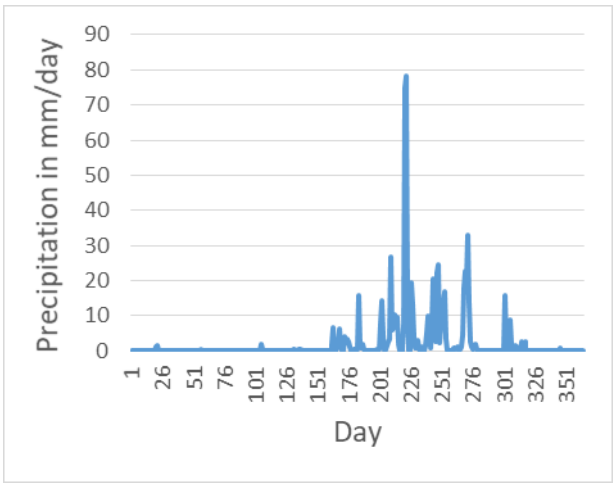

Fig. 1(c) Precipitation of each day

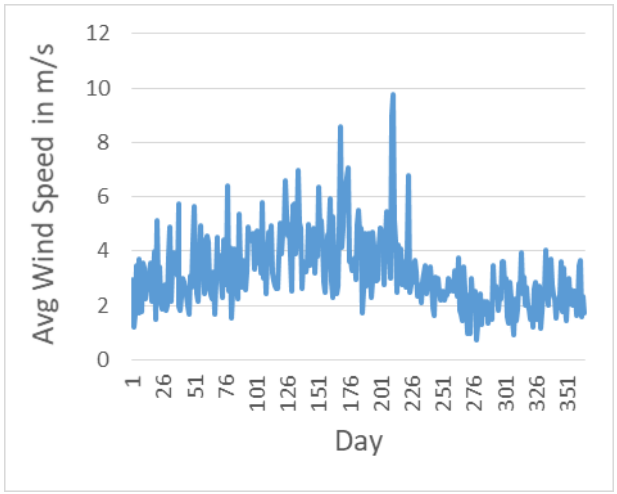

Fig. 1(d) Average Wind Speed of each day

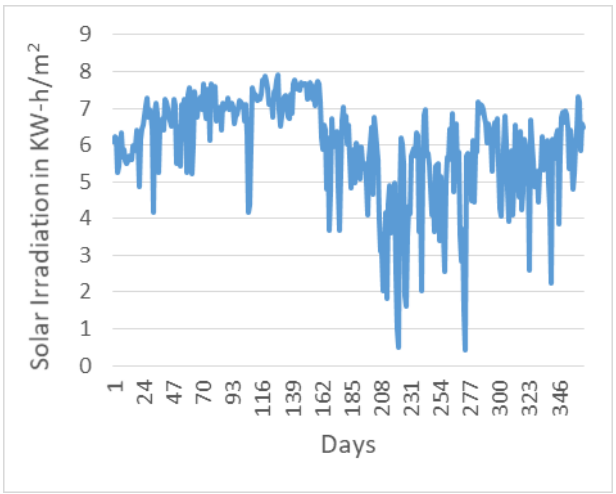

Fig. 1(e) Solar Irradiance of each day

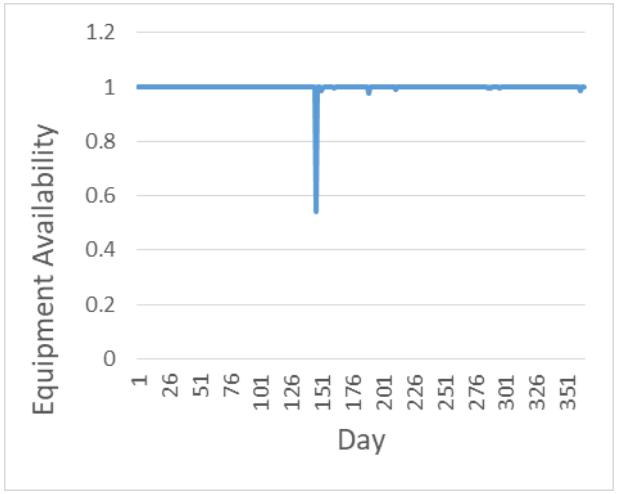

Fig. 1(f) Equipment availability of each day

Fig. 1. Parameter for the Year 2019 (District Patan, Gujarat). X-Label=Day of the Year 2019

power. The attributes of the dataset, with its data format and their description, are mentioned in Table 1. Fig 1 shows the graph of weather parameters in the year 2019.

B. Model

We have used the Decision Tree Regression Model for our study [19],[20],[21]. Decision Tree Regression is similar to Decision Tree Classifiers, but both serve different purposes, as the name suggests. In this model, we select a variable to split the nodes and evaluate the splits using splitting criteria like Mean Absolute Error (MAE), Mean Square Error (MSE), Poisson function, etc. A tree continues to split until it achieves a minimum sample split parameter which defines the minimum number of samples that are required by the tree to consider splitting a node into 2. The depth of a decision tree must be kept in check as a deeper tree has high variance and low bias while a shorter tree will have a high bias and low variance. Thus, in our study, we have experimented with both the splitting criteria and maximum depth of the tree to obtain the maximum accuracy out of the model, as shown in Table 2.

C. Validation

To ascertain the performance of our model, we have used the Mean Absolute Percentage Error (MAPE) loss function. The mathematical formula for MAPE is given by equation 1. We preferred MAPE as our loss function because we had many outliers in our data-set [22].

We have 56 months of continuous data, and due to its chronological nature, we decided to keep one year of data-set as test data (21.4% of the data-set), as shown in Fig 2. It helped us analyze the performance of our model in all possible weather conditions of Patan District, Gujarat.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Actual\ Value_i - Prediction\ Value_i}{Actual\ Value_i}\right| \quad (1)$$

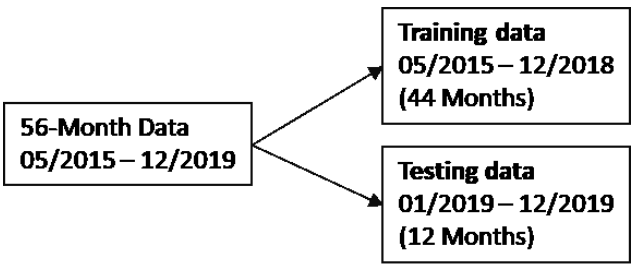

Fig. 2. Dataset split

IV. Results

A. Model Comparison

We have split our data-set into training and testing data-set. Our test data predicts solar power production (KWh) in the year 2019. Then these predictions will be compared to the actual output power in 2019. The result and their error are shown in Table 2 and Fig 3.

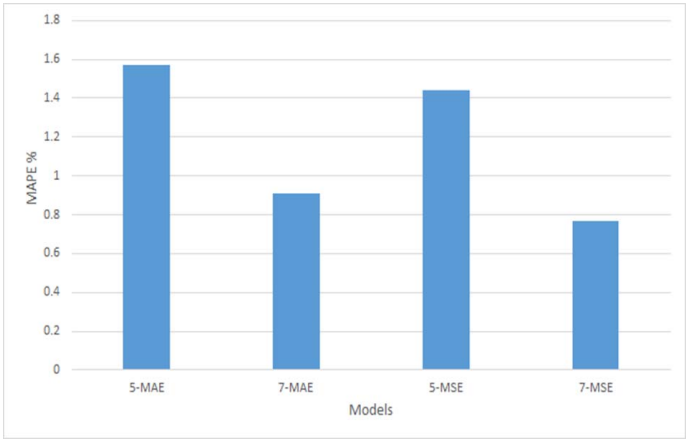

Fig. 3. Bar Graph to Compare Different Models

The Fig 3 is the bar graph of Mean Absolute Percentage Error (MAPE) on the y-axis with the different parameters of the Decision Tree Regressor model on the x-axis. We have four different models for comparison. The model labeled as 5-MAE is the Decision Tree Regressor of maximum depth five and Mean Absolute Error (MAE) with its splitting criterion. Similarly, the model labeled as 7-MSE is a Decision Tree of maximum depth seven and Mean Squared Error (MSE) as its splitting criterion. Similarly, we have labeled 7-MAE and 5-MSE in the graph. The table below shows the values of MAPE for different models.

TABLE II. Splitting Criteria of Decision Tree with its Error Percentage

S. No.	*Splitting Criteria*	*Max Depth*	*MAPE*
1	Mean Absolute Error	5	1.57
2	Mean Absolute Error	7	0.91
3	Mean Squared Error	5	1.44
4	Mean Squared Error	7	0.77

The results shown above are obtained after training our model. The model used was Decision Tree Regressor which gave us satisfactory results. We have obtained results from various splitting criteria and max depth of the Decision Tree. The splitting criteria used were MAE and MSE. The **max_depth** parameter of the Decision Tree was kept 5 and 7. After setting the above parameters, we found the difference between predicted and actual output power production in KWh. The MAPE is computed for each parameter to show deviation from the actual output. We found the least error when the **max_depth** of the tree was seven, and the splitting criteria of the Decision Tree Regressor was MSE. Fig 4 shows

the graph between actual and predicted power generated (KWh) for all four parameters in the year 2019.

B. *Weather Based Observation*

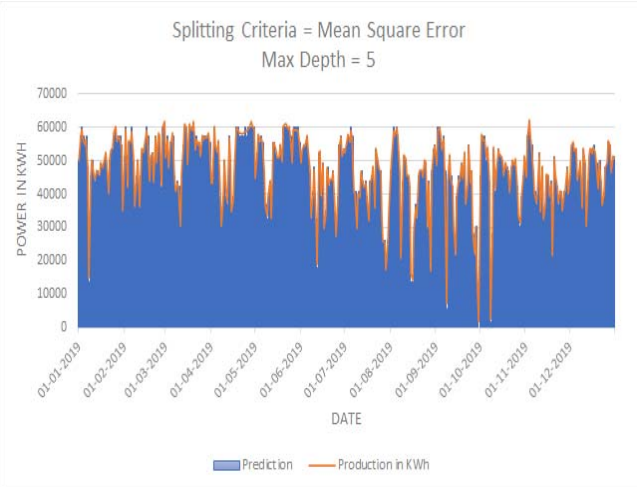

Fig. 4(a) MSE with max depth = 5

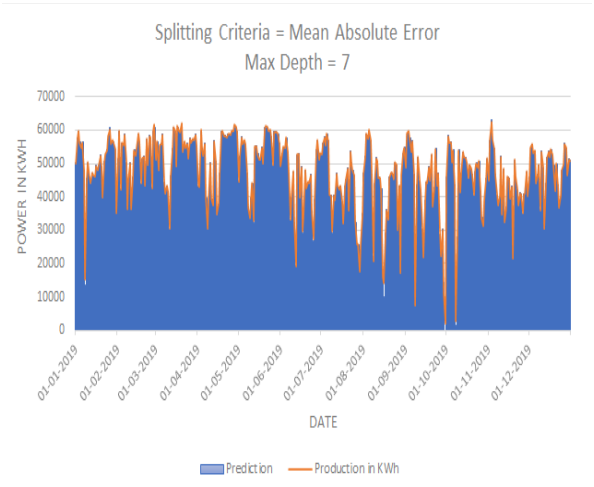

Fig. 4(b) MAE with max depth = 7

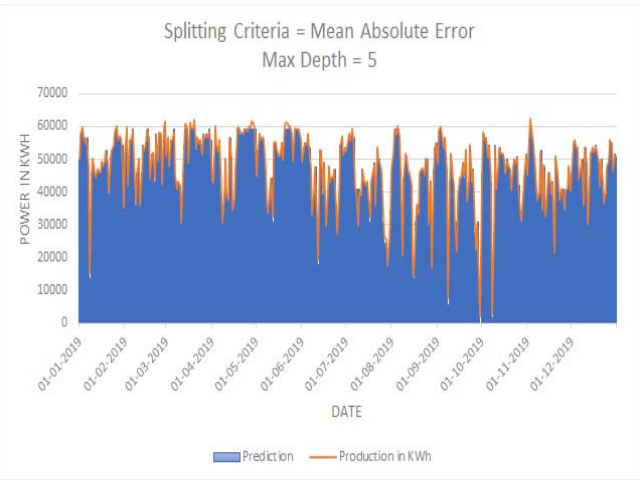

Fig. 4(c) MAE with max depth = 5

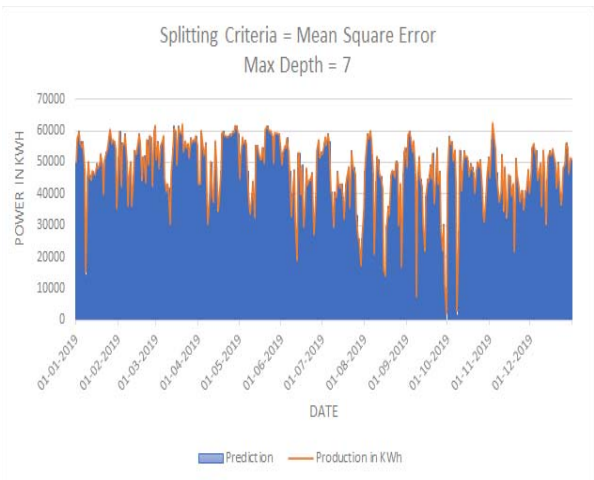

Fig. 4(d) MSE with max depth = 7

Fig. 4. Graph between Actual and Predicted Power Generation for 2019(test data)

The output power production of a SE plant highly depends on the weather conditions of the area. In Fig 5, we have plotted Mean Absolute Percentage Error (MAPE) against all the months of 2019(our test data) using the parameters, maximum depth as seven, and the splitting criteria of the Decision Tree Regressor as Mean Squared Error (MSE)**.** We can observe that the MAPE for the months of September and October is above the MAPE of 2019, the reason being higher precipitation in these two months. During these two months, the Patan district received high rainfall, which increased the average precipitation. Also, the average wind speed was very low for this period. Due to this, solar irradiance was also impacted. It resulted in low power generation compared to other months of the year. High solar irradiation and low average precipitation were observed in the months of February and March 2019. These conditions are highly favorable for a solar power plant, and the error percentage was comparatively lesser for these months as the weather was largely stable.

V. CONCLUSION

SE has been one of the fastest growing renewable energy in recent times, and with the cost coming down to just Rs.36/W for 100KW to 500KW systems[3], we see no indications of a slowdown in the near future. The main challenges that one encounters while planning for a solar-based power generation system are firstly Storage of Energy as energy generation only takes place when solar radiations are present and secondly, the weather-dependent nature of this

[3] https://mercomindia.com/mnre-benchmark-costs-rooftop-solar/

arrangement. The first challenge will eventually be solved when the battery technology will evolve, especially with the Booming Electric Vehicle industry and the second challenge is what we aim to solve using this study. An accurate prediction ahead in time can help us save carbon emissions as one can reduce dependency on other fossil fuel-based sources of energy and also save up economically. This study on 56 months of data collected from a 10MW installation in Patan District, Gujarat, India, has yielded satisfactory results. This study demonstrates that we can develop a stable grid even with renewable sources of energy which currently is not possible.

VI. Future Works

In the future, it will be fascinating to expand this research further. Studies can be done on various solar power plants to predict their power generation. Similar studies can be performed for different parts of India where there are variations in climatic conditions because the demand for cleaner sources of energy is increasing every day [23],[24]. Several other Artificial Intelligence techniques can be applied to predict the production of Photovoltaic cells like Reinforcement Learning, Deep Learning, etc. We can also use these prediction methods to forecast the power generated through other renewable sources of energy since other sources of energy could also be available in abundance due to differences in geographical conditions.

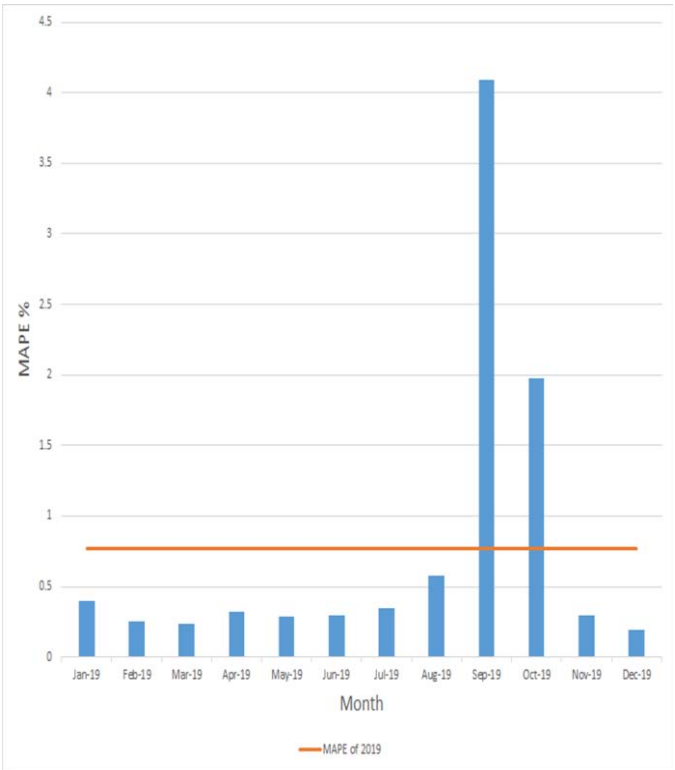

Fig. 5. MAPE for each month of 2019 using model with Model 4 in Table II

References

[1] J. Antonanzas, N. Osorio, R. Escobar, R. Urraca, F. J. Martinez-de-Pison, and F. Antonanzas-Torres, "Review of photovoltaic power forecasting," *Solar Energy*, vol. 136. pp. 78–111, 2016, doi: 10.1016/j.solener.2016.06.069.

[2] C. Voyant *et al.*, "Machine learning methods for solar radiation forecasting: A review," *Renewable Energy*, vol. 105. pp. 569–582, 2017, doi: 10.1016/j.renene.2016.12.095.

[3] A. Torres-Barrán, Á. Alonso, and J. R. Dorronsoro, "Regression tree ensembles for wind energy and solar radiation prediction," *Neurocomputing*, vol. 326–327, pp. 151–160, 2019, doi: 10.1016/j.neucom.2017.05.104.

[4] A. Khandakar *et al.*, "Machine Learning Based Photovoltaics (PV) Power Prediction Using Different Environmental Parameters of Qatar," *Energies*, vol. 12, no. 14, p. 2782, 2019, doi: 10.3390/en12142782.

[5] M. Aslam, J.-M. Lee, H.-S. Kim, S.-J. Lee, and S. Hong, "Deep Learning Models for Long-Term Solar Radiation Forecasting Considering Microgrid Installation: A Comparative Study," *Energies*, vol. 13, no. 1, p. 147, 2019, doi: 10.3390/en13010147.

[6] Y. Sun, V. Venugopal, and A. R. Brandt, "Short-term solar power forecast with deep learning: Exploring optimal input and output configuration," *Sol. Energy*, vol. 188, pp. 730–741, 2019, doi: 10.1016/j.solener.2019.06.041.

[7] F. Wang, Z. Zhen, B. Wang, and Z. Mi, "Comparative Study on KNN and SVM Based Weather Classification Models for Day Ahead Short Term Solar PV Power Forecasting," *Appl. Sci.*, vol. 8, no. 1, p. 28, 2017, doi: 10.3390/app8010028.

[8] M. W. Ahmad, J. Reynolds, and Y. Rezgui, "Predictive modelling for solar thermal energy systems: A comparison of support vector regression, random forest, extra trees and regression trees," *J. Clean. Prod.*, vol. 203, pp. 810–821, 2018, doi: 10.1016/j.jclepro.2018.08.207.

[9] C. Persson, P. Bacher, T. Shiga, and H. Madsen, "Multi-site solar power forecasting using gradient boosted regression trees," *Sol. Energy*, vol. 150, pp. 423–436, 2017, doi: 10.1016/j.solener.2017.04.066.

[10] J. Fan *et al.*, "Comparison of Support Vector Machine and Extreme Gradient Boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: A case study in China," *Energy Convers. Manag.*, vol. 164, pp. 102–111, 2018, doi: 10.1016/j.enconman.2018.02.087.

[11] A. Ahmed Mohammed and Z. Aung, "Ensemble Learning Approach for Probabilistic Forecasting of Solar Power Generation," *Energies*, vol. 9, no. 12, p. 1017, 2016, doi: 10.3390/en9121017.

[12] D. Lee and K. Kim, "Recurrent Neural Network-Based Hourly Prediction of Photovoltaic Power Output Using Meteorological Information," *Energies*, vol. 12, no. 2, p. 215, 2019, doi: 10.3390/en12020215.

[13] S.-G. Kim, J.-Y. Jung, and M. Sim, "A Two-Step Approach to Solar Power Generation Prediction Based on Weather Data Using Machine Learning," *Sustainability*, vol. 11, no. 5, p. 1501, 2019, doi: 10.3390/su11051501.

[14] R. Aler, R. Martín, J. M. Valls, and I. M. Galván, "A study of machine learning techniques for daily solar energy forecasting using numerical weather models," *Stud. Comput. Intell.*, vol. 570, pp. 269–278, 2015, doi: 10.1007/978-3-319-10422-5_29.

[15] M. Alsharif, M. Younes, and J. Kim, "Time Series ARIMA Model for Prediction of Daily and Monthly Average Global Solar Radiation: The Case Study of Seoul, South Korea," *Symmetry (Basel).*, vol. 11, no. 2, p. 240, 2019, doi: 10.3390/sym11020240.

[16] M. AlKandari and I. Ahmad, "Solar power generation forecasting using ensemble approach based on deep learning and statistical methods," *Applied Computing and Informatics*. 2019, doi: 10.1016/j.aci.2019.11.002.

[17] F. Jawaid and K. Nazirjunejo, "Predicting daily mean solar power using machine learning regression techniques," in *2016 6th International Conference on Innovative Computing Technology, INTECH 2016*, 2017, pp. 355–360, doi: 10.1109/INTECH.2016.7845051.

[18] A. Sparks, "nasapower: A NASA POWER Global Meteorology,

Surface Solar Energy and Climatology Data Client for R," *J. Open Source Softw.*, vol. 3, no. 30, p. 1035, 2018, doi: 10.21105/joss.01035.

[19] F. Mola, "Classification and Regression Trees Software and New Developments," 1998, pp. 311–318.

[20] T. Hastie, R. Tibshirani, and J. Friedman, "Linear Methods for Regression," 2009, pp. 43–99.

[21] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[22] B. Jiang and C. Ding, "Revisiting L2,1-Norm Robustness with Vector Outlier Regularization," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 31, no. 12, pp. 5624–5629, Dec. 2020, doi: 10.1109/TNNLS.2020.2964297.

[23] D. S. V., "Data Mining based Prediction of Demand in Indian Market for Refurbished Electronics," *J. Soft Comput. Paradig.*, vol. 2, no. 2, pp. 101–110, 2020, doi: 10.36548/jscp.2020.2.007.

[24] A. Chandy, "SMART RESOURCE USAGE PREDICTION USING CLOUD COMPUTING FOR MASSIVE DATA PROCESSING SYSTEMS," *J. Inf. Technol. Digit. World*, vol. 01, no. 02, pp. 108–118, 2019, doi: 10.36548/jitdw.2019.2.006.