An estimation and novel design of future predictive model for solar cell parameters using AI/ML

A Project Report Submitted in Partial Fulfillment of the Requirements for Award of
the Degree of Bachelor of Technology in Information and Communication Technology

Submitted by
Nishan Patel, 19BIT091

Under the Supervision and Guidance of
Dr. Deepak Jarwal
Assistant Professor - EC
Pandit Deendayal Energy University

Group members
1. Prince Patel, 19BIT101
2. Shailesh Vekariya, 19BIT114
3. Vrund Chaudhari, 19BIT132
4. Himanshi Panchal, 19BIT137

Submitted to
Department of Information and Communication Technology
School of Technology
Pandit Deendayal Energy University (PDEU)
Gandhinagar, INDIA, 382007

Declaration

I hereby declare that the project work entitled **"An estimation and novel design of future predictive model for solar cell parameters using AI/ML"** is an authentic record of my own work carried out in Pandit Deendayal Energy University as requirement of B. Tech dissertation for the award of **Bachelor of Technology in Information and Communication Technology**. I have duly acknowledged all the sources from which the ideas and extracts have been taken. The project is free from any plagiarism and has not been submitted elsewhere for any degree, diploma and certificate.

Signature:..............................

Nishan Patel

(19BIT091)

Certificate of approval by HoD

Information and Communication Technology

Certificate

This is to certify that the project entitled **"An estimation and novel design of future predictive model for solar cell parameters using AI/ML"** submitted by **Nishan Patel**, Roll No. 19BIT091 to the Department of Information and Communication Technology under School of Technology, PDPU in partial fulfillment of the requirements for award of the degree of **Bachelor of Technology in Information and Communication Technology** embodies work carried out under the guidance and supervision of Dr. Deepak Jarwal, Assistant Professor, Department of Electronics and Communication.

. . . .

Dr. Ganga Prasad Pandey

(HOD, I.C.T. Department, PDEU.)

Certificate from Inhouse Mentor
Dept of ICT, PDEU

Dr. Deepak Jarwal

E-mail:
Deepak.Jarwal@sot.pdpu.ac.in

Certificate

This is to certify that the project entitled **"An estimation and novel design of future predictive model for solar cell parameters using AI/ML"** submitted to the Department of Information and Communincation Technology under School of Technology, Pandit Deendayal Energy University in partial fulfillment of the requirements for award of the degree of **Bachelor of Technology in Information and Communication Technology** is a record of work carried out by **Nishan Patel**, Roll No. 19BIT091 under my supervision and guidance in the Dr. Deepak Jarwal (EC Dept, SOT).

Signature ....

Dr. Deepak Jarwal
(Assistant Professor - EC)
Email ID: Deepak.Jarwal@sot.pdpu.ac.in

Certificate of approval by evaluators

The forgoing project entitled **“An estimation and novel design of future predictive model for solar cell parameters using AI/ML”** submitted by **Nishan Patel**, Roll No 19BIT091 to the Information and Communication Technology under School of Technology, PDPU is hereby approved as project work carried out and presented in a manner satisfactory to warrant its acceptance as a prerequisite of **Bachelor of Technology in Information and Communication Technology** degree for which it has been submitted. It has been understood that by this approval of the undersigned do not necessarily endorse or approve every statement made, opinion expressed or conclusion drawn therein but approve only for the purpose for which it is being submitted.

. . . .

Signature of Panel Members

Acknowledgement

The success and final outcome of this project required a lot of guidance and assistance from many people. It is our privilege to pledge the following few lines of dedication to those who helped us directly or indirectly in completing our project.

We would also like to thank Mr. Siddeshwar Jangid, Operations Excellence at Waaree Energies for being available and assuring the help if we need any.

Furthermore, we are thankful to and fortunate enough to the faculties of ICT. Department for all the knowledge they have imparted, which we could put to use in this project...

Abstract

With the increase in technology and its extensive use of energy in every sector like industry, agriculture, transportation, communication, etc. we need to find a solution to meet our demand and for that we need to find a solution to generate energy efficiently through resources like renewable resources for example: photovoltaic solar energy. Hence our project aims to predict the amount of photovoltaic solar energy generation while keeping in mind various parameters like humidity, temperature, ultraviolet rays and all sky shortwave downward radiation that affect the energy efficiency . We plan on using machine learning model on the dataset collection of the parameter to get the better efficient power generation in future.

Contents

Chapter 1

Introduction

1.1 Introduction

With the increase in technology and its extensive use of energy in every sector like industry, agriculture, transportation, communication, etc. we need to find a solution to meet our demand and for that we need to find a solution to generate energy efficiently through resources like renewable resources for example: photovoltaic solar energy. Hence our project aims to predict the amount of photovoltaic solar energy generation while keeping in mind various parameters like humidity, temperature, ultraviolet rays and all sky shortwave downward radiation that affect the energy efficiency . We plan on using machine learning model on the dataset collection of the parameter to get the better efficient power generation in future.

As we know non renewable resources are decreasing day by day and demand of energy consumption increases in each sector to fast economic growth. We have to fulfil demand of energy so current world is tending towards renewable energy resources like photovoltaic solar energy, wind energy, hydro energy, tidal energy, geothermal energy, biomass energy etc. but we can not this resources in efficient manner lots of energy wasted because we are not able to convert in its useful manner. So we are trying to make change in energy sector through our knowledge.

We here consider only photovoltaic solar energy as our area of interest. Nowadays institutions, commercial buildings, power plants, solar field, houses are set up solar grid to appropriate use of solar energy which converts solar energy into electricity. This buildings and home has no need to rely on external electricity sources even it gives electricity to electricity corporation ltd. Like PGVCL, DGVCL. And this corporation pays to owner of grid.

1.2 Title of Project

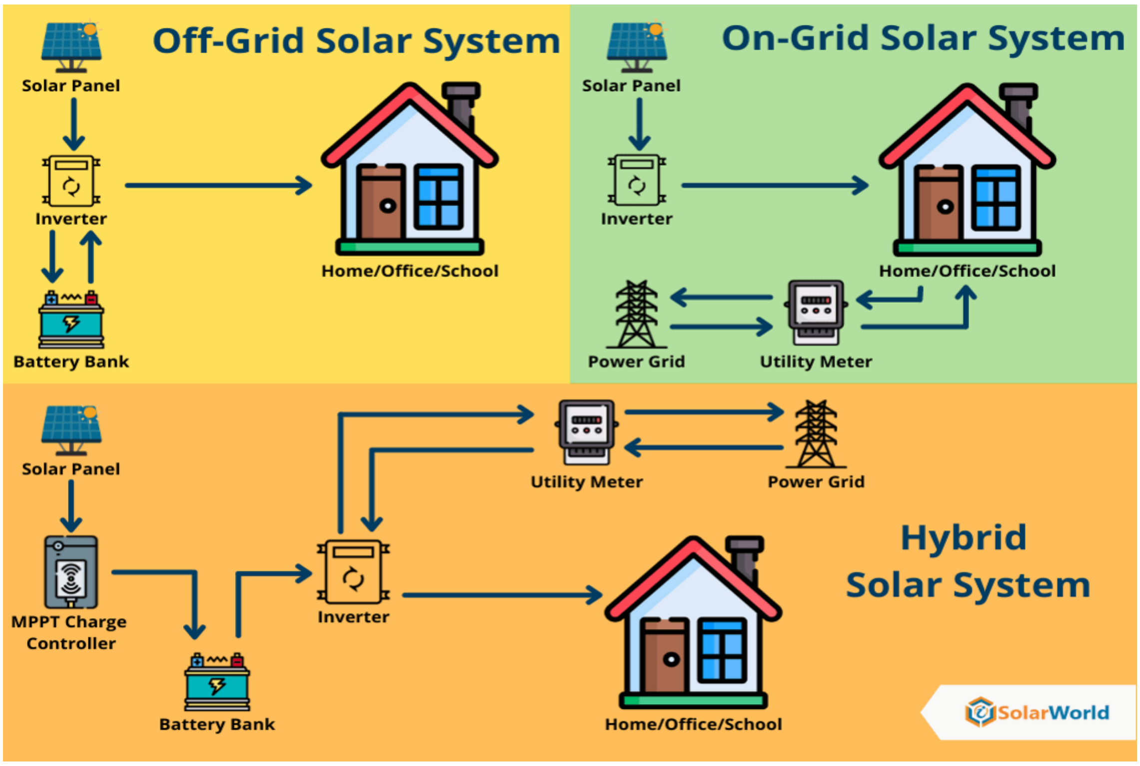
We have named our mini project **"An estimation and novel design of future predictive model for solar cell parameters using AI/ML"**. We have decided on this name as the main aim of our project is to predict solar power generation of preceding day using Machine Learning Algorithm.

1.3 Problem statement

The day by day consumption of goods operating on electricity are increasing worldwide, and as we now conventional ways for electricity production will not be enough to meet this rising demand. That's why many countries has started using renewable resources to develop electricity. Here comes the concept of solar power prediction using the parameters from weather forecast, so that proper planning can be done in advance for distribution of power in different areas according to the consumption.

1.4 Motivation

Consumption of energy is increasing linearly every year in the world. So it is very hard to satisfy the need of major consumers only by renewable energy. So the solar power prediction will be an important part in near future to get an idea of how much to buy conventional energy to satisfy the need.



Various model of solar powered system
https://www.isolarworld.com/blog/solar-grid-tied-off-grid-system-and-hybrid-solar-systems

1.5 Literature review

In current scenario of the word energy production using renewable sources is main problem that world have face. The requirement of energy is increase day by day. In last century the main source of energy production are fossil fuels and also in present days major part of energy comes from fossil fuels. But they are not renewable and producing green house gases that polluted our environment. So now the world focus on the renewable energy sources. In renewable energy major part of energy production is depend on solar and wind energy. Because of their availability and potential. Now the government of various countries also focus on solar energy and the graph of solar energy generation is go bigger and bigger.

The main sources of solar energy are solar farms. Governments and also various companies started establishing their solar farms to generate energy. In current scenario the solar power plants are of two types solar thermal system and solar photovoltaic (PV). Among this PVs are mainly use now days due to the availability and cost efficiency. The installation of PVs are increase day by day in compare with concentrated solar power (CSP) technique. PVs output is depend upon sun position, the power generation various according to the sun's movement. In different parts of worlds the generation of power from PVs depended on local solar redaction. The local solar radiation intensity varies with latitude, season, atmospheric conditions (e.g. rain, snowfall, fog, humidity etc.), air quality and pollution index (smog) etc. the solar radiation intensity in various parts of world is given in fig.1. and the installation of PVs in compare with CSPs in last few years is given in fig.2.

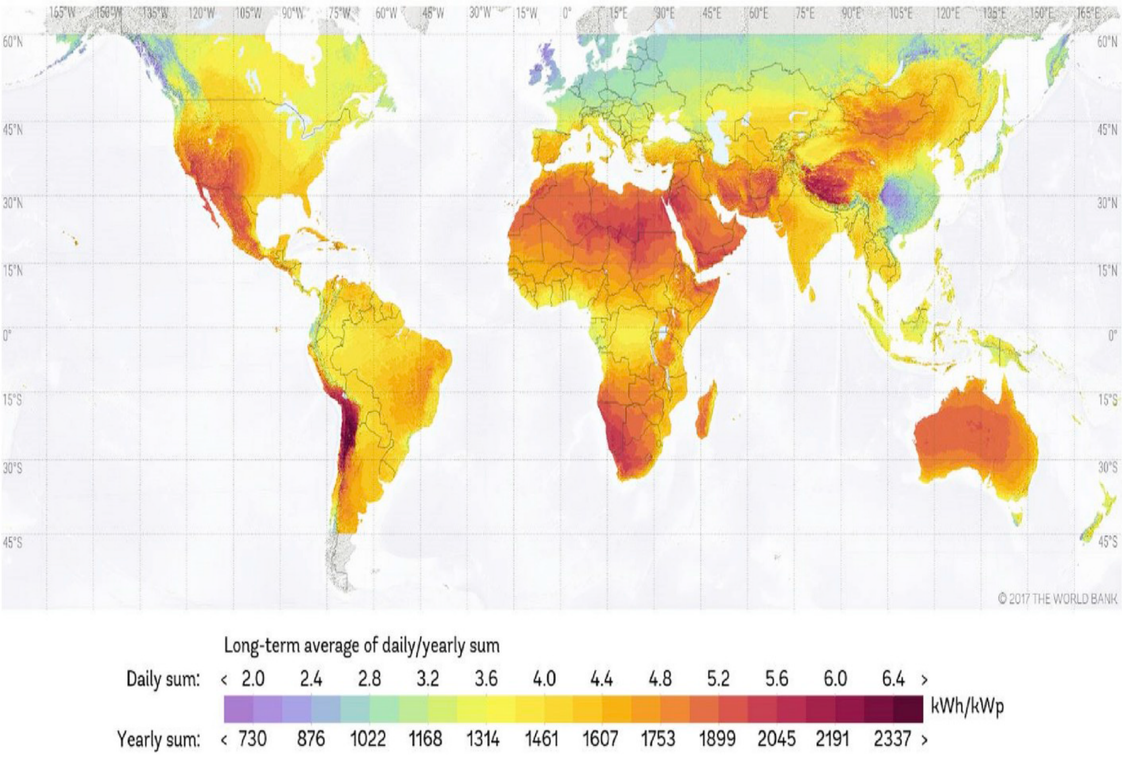


Fig. 1. The world solar resources map
http://helioscsp.com/concentrated-solar-power-had-a-global-total-installed-capacity-of-6451-mw-in-2019/

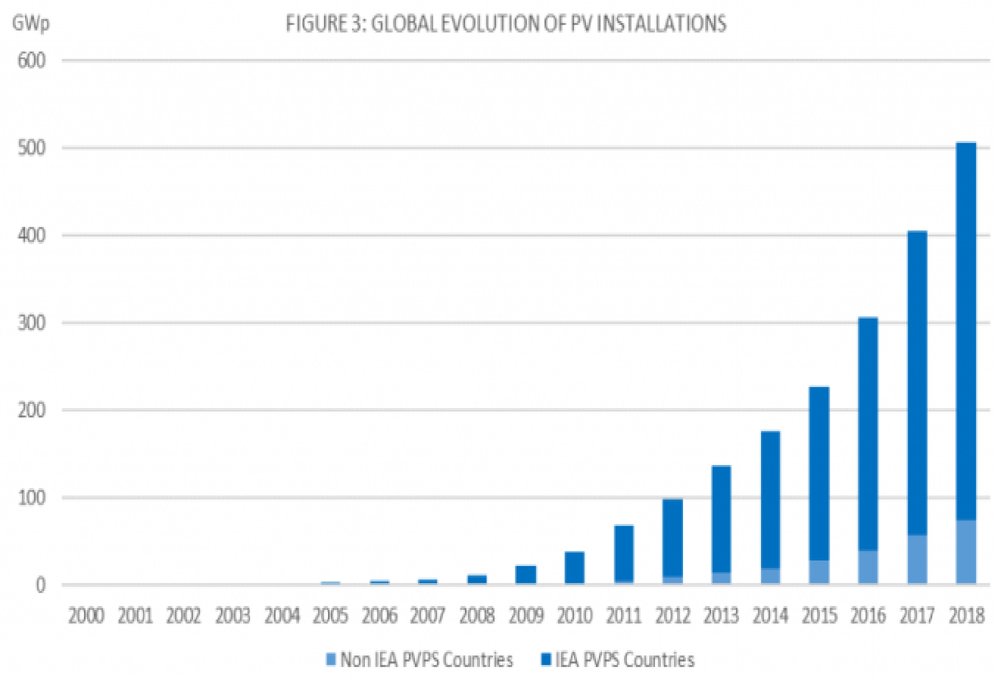


Fig. 2. Global installed solar power capacity, 2000–2018
https://solarbusinesshub.com/2019/04/21/total-installed-solar-pv-capacity-exceeded-500-gw-globally-in-2018-iea-pvps-report-finds/

As we can see that the in solar energy generation the PVs are mainly uses and the generation of power in PVs is depending on sun's movement and varies with the geological location the prediction of power generated by PVs is important. In our project we tried to focus on this problem and make a model that forecast power generation by PV's.

According to the time duration of forecasting some researchers gives three categories of forecasting technique as given below:-

1. Short term forecasting
2. Medium term forecasting
3. Long term forecasting

Short term forecasting is popular in the electricity market, where decisions be made up on economic load dispatch and power system operation. Short term forecasting also useful in control the renewable energy integrated power management systems.

The Short term forecasting is in between 30 to 360 min. there is a very short term technique that forecasting for period of some seconds to 30 min.

Medium term forecasting spans some hours. Mainly it forecasting for period of 6-12 h. but in many cases some people may take one day, one week and even one month forecasting in this category. This is used for scheduling between solar power and conventional power.

Long term forecasting is for period of more than 24 h, but some people may take it for some weeks, some months and even for the span of year. But this model have reduce accuracy due to the weather fluctuations for one or more days cant predicted accurately.

Due to the accuracy of long term model mainly medium term forecasting is used. Many researchers divided medium term in intra-hour, intra-day and day-ahead models. Among them for roof top PV generation and PV solar farms the day-ahead model is uses very significantly. Because in this day-ahead model we forecasting the generation of energy for the next day on the bases of next day weather data.

PV yield is most unequivocally related with sun powered spectral irradiance; the last mentioned depending on meteorological affect components(MIF): aerosol distribution, wind speed and heading, mugginess and cloud cover. Thus, changing climate status influences PVPF exactness, and a successful PVPF model must coordinated determining with climate classification for improving vigor. Without a doubt, state-of-the-art investigate demonstrates that climate classification is a vital pre-processing step, especially for short-term PVPF (ST-PVPF).

The PVs output is directly related with the solar irradiance. Solar irradiance is the radiation power energy per area produce by the sunlight. This power in term of electromagnetic radiation by light wavelength. PV solar generation system use solar cells to generate power. PV solar cell produce electrons from the sun light, using the wavelength of sun light electrons are gain energy and from PN junction present in PV solar cell electron are separated and power is generated by solar cell. This energy gain by the electron is strongly related to the energy provided by the light wave. According to the variation in wavelength energy the production of power is varies. This solar irradiance is affected by the air sole and quality of air sole in environment, wind speed and direction and clouds present in sky. In many papers researchers proposed various ml techniques to correlate solar power generation and solar irradiance with Artificial Neural Network (ANN). They gain forecast error on past PV generation data around 12.57%, 12.60% and 10.91% respectively for input vector of PV output, solar irradiance and module temperature.

The study paper employs a variety of techniques, including the ensemble approach, fuzzy logic, SVR, HMM, ANN, and ANN-SVM. Initially, they used the SARIMA model to anticipate power needs, but it was less effective in cloudy or rainy weather. For one-day in advance electricity generation, more ML and AI were applied. they also used techniques like RNN, FBNN, FFNN, Elman Neural Network, and generalised Regression Neural Network. Following the employment of BP-ANN and GA-BP to predict PV generation, analysis of the results revealed that the root mean square error between predicted and real output power was 0.25, 0.30, and 0.426 for three separate study models. Results indicated that breaking up the data into various weather categories was more effective than using the entire data set. The final result was that we may increase the efficiency of PV power generation by applying ensemble methods (bagging and boosting).

Prediction of photovoltaic power output based on similar day analysis, genetic algorithm and extreme learning machine.Due to their robust nonlinear regression capabilities, a number of machine learning approaches have been effectively applied to predicting photovoltaic (PV) power production. In this study, a hybrid model (SDA-GA-ELM) based on extreme learning machine (ELM), genetic algorithm (GA), and customised similar day analysis (SDA) was developed to predict hourly PV power output. The performance of the proposed forecast model was evaluated using the coefficient of determination (), mean absolute error (MAE), and normalised root mean square error (RMSE) (nRMSE). The findings indicate that the SDA-GA-ELM model has more precision and stability in predicting PV power one day in advance. Researchers in the domains of predicting difficulties, classification challenges, pattern recognition, data mining, and spam filtering have recently adopted machine learning approaches in large numbers due to their exceptional performance in processing nonlinear and complicated problems. They are able to find probable linkages between input and output variables even if the representation is impossible. In the meanwhile, the capacity for regression of machine learning algorithms optimised by other optimization algorithms can be enhanced to a degree. The machine-learning-based forecasting model is capable of overcoming the aforementioned shortcomings of physical and statistical forecasting methods.The basic method to implement SDA-GA-ELM forecast method has three steps in it that are selection of the day then forecast of Pv power output by Elm is done and the parameters are optimsed by the use of GA. By this method the data having low quality are filtered out by customized SDA method and similar days are determined for the training data set.

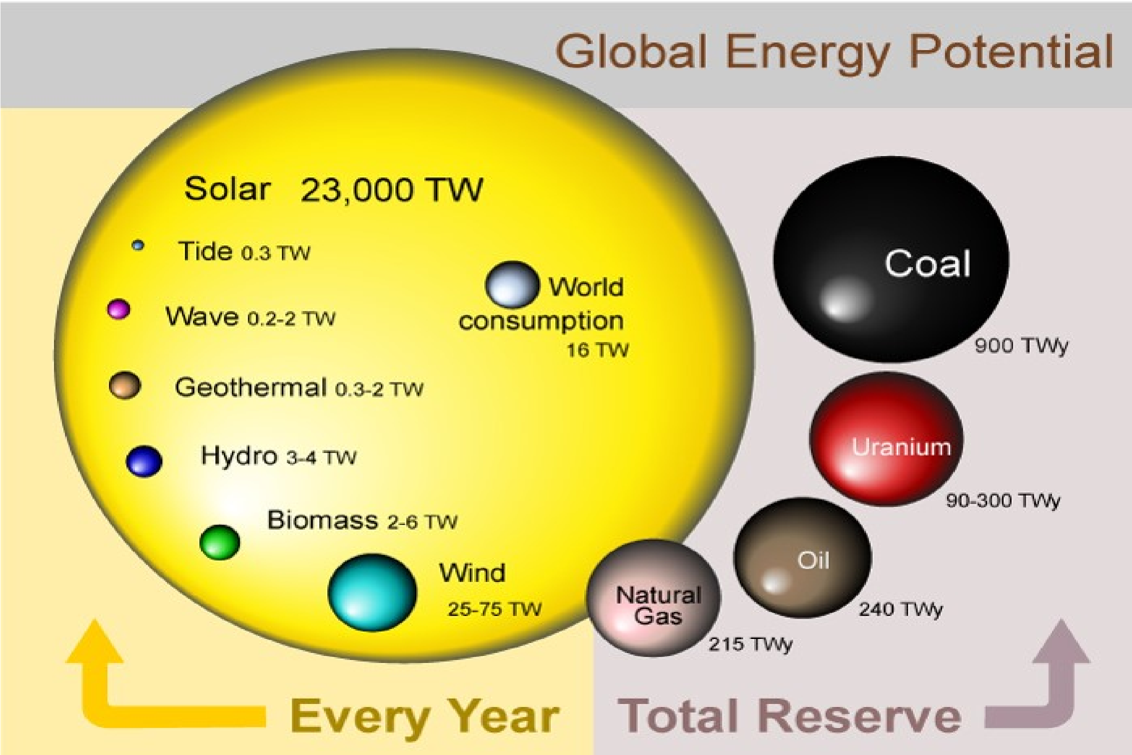
A novel competitive swarm optimized RBF neural network model for short-term solar power generation forecasting, In this paper, a novel improved radial basis function neural network model is proposed and applied in forecasting the short-term solar power generation. Numerical results demonstrate that the proposed competitive swarm optimized radial basis function neural network model could obtain higher accuracy compared to other counterparts and thus provides a useful tool for solar power forecasting. The innovative technology modifies the particles by significantly absorbing data from dual contests. The CSO method has been verified for its competitive performance in tackling large-scale issues. The canonical technique and its modifications have been utilised successfully to solve a variety of large-scale engineering challenges. The key contribution of this method can be unique CSO-optimized RBF model is adopted, in which the CSO approach is utilised for the first time in training the nonlinear parameters of the RBF model. To validate the performance, the proposed CSO-based model training approaches were fully compared with existing state-of-the-art MA methods on representative nonlinear modelling issues. To validate the model performance in solar forecasting, a real-world case study of a Dutch community's PV generation is done, and monthly data models are developed and studied. The suggested model structure may effectively calculate nonlinear and linear RBF model parameters. Multiple non-linear benchmark tests and a solar generation forecasting work in the actual world were used to validate the competitive performance of the proposed model.

Due to the variability and unpredictability of photovoltaic (PV) power outputs, not just deterministic point sources are required forecasts (DPPs), as well as related prediction Intervals (PIs), are crucial for boosting the use of PV in practise, particularly as grid connectivity grows. The results suggest that the proposed two-stage model is valid. In terms of predicting short-term PV power outputs, the article surpasses standard forecasting techniques with corresponding uncertainty.

In the past, several work were devoted to enhancing the precision of PV deterministic point forecasts (DPPs). Emanuele et al. [9] compared the PV output power day-ahead forecasts performed by physical models, based on three and five parameters electric equivalent circuit and statistical models on the basis of artificial neural network (ANN), discovering that the ANN, combined with clear sky solar radiation, generally achieves the best forecasting performance with the normalised mean absolute error(NMAE) of 5.6 percent. To meet the criteria for the safe and stable functioning of the power grid, short-term PV power predictions must not only provide the DPP value, but also a realistic estimate of the possible uncertainty of the point projections. The article brought PI estimate within the context of PV power DPP to represent the unpredictability and uncertainty of PV power in response to this demand. The research specifically presented a two-stage model consisting of an integrated WVCFM model in the first stage and an NKDE-based interval forecast approach in the second stage. Results indicate that interval forecasts are able to identify the robustness of point predictions and can augment point forecasts with more confidence. The paper's two-stage methodology is more favourable to quantifying the uncertainty of PV power outputs than conventional forecasting methods.

In order to reduce the effects of climate change and global warming, the usage of renewable energies is steadily growing. Among the possible renewable energy, photovoltaic (PV) has undergone an exponential increase in popularity for electricity production. Nonetheless, the fluctuation of PV power output has a variety of detrimental effects on the electric grid system, including as a reduction in the system's stability the operation's dependability and planning, in addition to its economic benefits. Therefore, precise forecasting of PV power output is crucial for grid stability and promotion of large-scale PV systems.The integration of electricity. Numerous studies have been undertaken to anticipate PV power generation from various angles. This paper examines the performance evaluation of a number of PV power forecasting models based on several categories. Additionally, the possible advantages of model optimization are explored. This study reviewed short-term direct PV power generation forecasting methods based on historical data. Section 2 of this work discusses the necessity of input-output data correlation and preprocessing for a PV power forecasting model. Section 3 classifies PV forecasting methods by prediction horizon, historical data, and technique. Section 4 discusses the strengths and limitations of statistical and AI models based on historical data. Section 5 discusses model optimization. Section 6 introduces predicting performance matrices. Section 7 analyses current model performance and techniques-based developments. Section 8 finishes the paper. Weather circumstances caused a large error range. Each forecast sub-model must be accurate to reduce mistakes. In this situation, evaluate the model's computing cost and complexity. The forecasting models' performance matrices have also been introduced. RMSE was utilised more than others, according to findings. Model optimization's benefits were also mentioned. Optimizing the algorithm improved the model's predicting accuracy. GA is a promising optimization approach in this procedure.

1.6 Why Solar Energy?



Global Energy Potential
https://commons.wikimedia.org/wiki/File:Global_Energy_Potential_2014_08_09.svg

We can see in above image generation of energy from different sources.

According to toatl reserves of coal, it will last for approx. 56 years. for uranium, it will last for maximum 20 years.

For oil and natural gas it will last for approx. 15 years according to the data.

But Solar Energy generation per year is of 23,000 TW, which will never exhaust.

Chapter 2

Experimental Methods and Results

In this chapter methods to be used and a glimse of data is include at present. In future we will include algorithms used and results produced.

2.1 Objective

The main objective of the project is to forecast the solar power generation of the next day by using the estimated weather data of the next day. The PV-system installed at various location can generate power according to the weather and sunshine at that location. We can see that in the day at every point of time the intensity of sunshine is changing as well as the weather is also changing with time. So if we focus on day then at every time interval the weather variables have different values and changing dynamically.

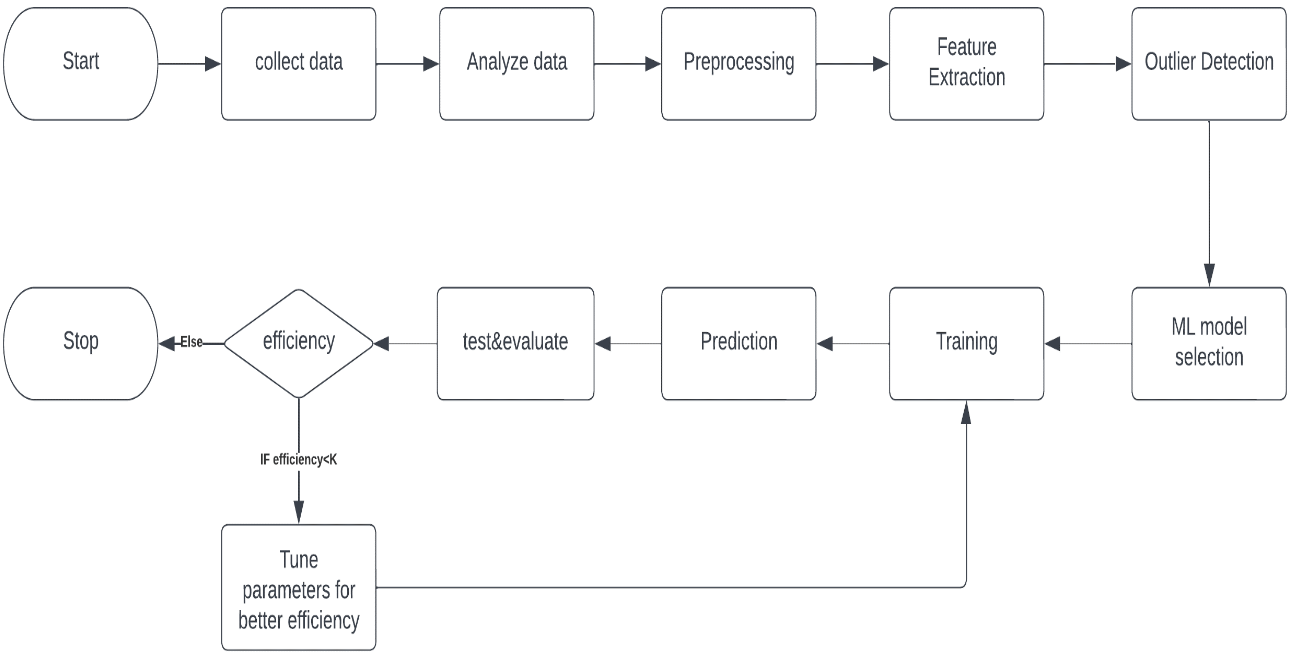
Now for the PV solar power generating system, the power generation is dependent on the weather parameters like temperature, solar irradiance, humidity etc. at that location. So we can say that the power generation rate in PV- system is dynamically change with the weather condition s. So, we can't easily estimate the next day power generation for the system for that we have to use certain forecasting techniques which have complex calculation. To make these type of calculation easer we use machine learning algorithm and produce results.

To comment about the feature power generation of the PV system we use various machine learning algorithm and develop model that generate information on the bases of the estimated temperature variables of the next day. For that we use past data of every day and map the power generation with the weather condition and train our model to generate value with minimum error between actual power generation and estimated value from the model.

Prediction of feature power generation in PV system is very much useful in terms of large scale power production. These prediction can help us to check whether the PV system will fulfill our energy consumption or we have to arrange some other resource to complete our need or the generation is more than the consumption so we can use power for other works too. So, on the basis of use fullness and real time application we choose this objective for our project.

2.2 Plan of execution

Flow chart:



2.3 Algorithms applied

Methodology:
Linear regression

Library used:
Pandas, Numpy, Matplotlib, Seaborn, Scikitlearn

Tools used:
Jupyter Notebook:

Data collection:
So gathering data is the first step in solving any data science problem. We gathered the data for day, month, year, hour, minute, DHI, DNI, clear sky DHI, clear sky DNI, cloud type, UV Index, dew point, humidity, wind speed, temperature, and pressure as we have created a solar power forecasting model taking into account various attributes that affect the efficiency or power production of the solar panel. We gathered information on the aforementioned characteristics from the years 2014 to 2018 and integrated it. The dataset has about 43000 entries overall when all the years data are combined.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Year	Month	Day	Hour	Minute	DHI	DNI	GHI	Clearsky D	Clearsky D	Clearsky G	Cloud Typ	Dew Point	Solar Zenit	Fill Flag	Surface All	Wind Spee	Relative H	Temperatu	Pressure
2	2014	1	1	0	0	0	0	0	0	0	0	1	-6	124.02	0	0.12	3.5	78.43	-3	1010
3	2014	1	1	1	0	0	0	0	0	0	0	1	-5	135.09	0	0.12	3.8	81.74	-3	1010
4	2014	1	1	2	0	0	0	0	0	0	0	1	-5	145.77	0	0.12	4.2	81.25	-3	1010
5	2014	1	1	3	0	0	0	0	0	0	0	1	-6	155.07	0	0.12	4.5	74.14	-3	1010
6	2014	1	1	4	0	0	0	0	0	0	0	1	-8	160.55	0	0.12	4.6	72.8	-4	1010
7	2014	1	1	5	0	0	0	0	0	0	0	1	-8	158.86	0	0.12	4.6	68.48	-4	1010
8	2014	1	1	6	0	0	0	0	0	0	0	1	-9	151.26	0	0.12	4.5	71.57	-5	1010
9	2014	1	1	7	0	0	0	0	0	0	0	1	-9	141.17	0	0.12	4.4	69.95	-5	1020
10	2014	1	1	8	0	0	0	0	0	0	0	4	-9	130.24	0	0.12	4.3	74.24	-6	1020
11	2014	1	1	9	0	0	0	0	0	0	0	4	-10	119.14	0	0.12	4.3	73.07	-6	1020
12	2014	1	1	10	0	0	0	0	0	0	0	4	-10	108.2	0	0.12	4.1	72.43	-6	1020

Data Analysis:
The analysis of the data is currently the most important step in the methodology. How does the particular characteristic impact the solar panels' ability to produce energy? which features have a greater and lesser impact on electricity production.

We aggregated the attributes such as day, month, year, hour, and minute into one column and gave it the name date_time after analysing the dataset.

Preprocessing:

In [37]:
```
# view summary statistics in numerical variables
round(merged_df.describe(),2)
```

Out[37]:

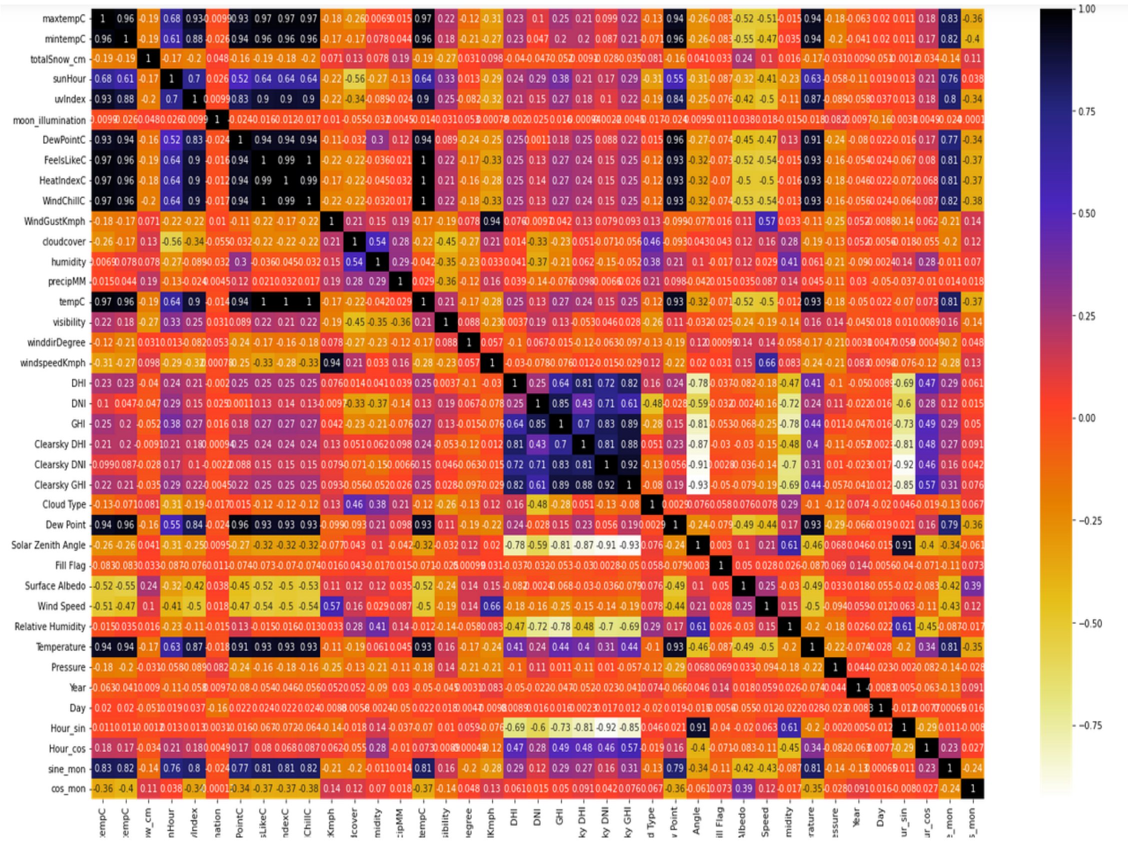
	% Baseline	maxtempC	mintempC	totalSnow_cm	sunHour	uvIndex	moon_illumination	DewPointC	FeelsLikeC	HeatIndexC	...	Surface Albedo	Wind Speed	Relativ Humidi
count	20571.00	20571.00	20571.00	20571.00	20571.00	20571.00	20571.00	20571.00	20571.00	20571.00	...	20571.00	20571.00	20571.0
mean	0.24	14.76	8.36	0.12	10.41	3.77	50.25	7.22	10.83	12.89	...	0.21	2.59	81.4
std	0.25	9.53	8.56	0.85	3.40	1.74	28.91	9.28	11.74	9.74	...	0.23	1.33	16.9
min	0.00	-12.00	-20.00	0.00	3.40	1.00	0.00	-22.00	-32.00	-20.00	...	0.11	0.10	30.3
25%	0.03	7.00	2.00	0.00	8.70	2.00	25.00	1.00	2.00	5.00	...	0.12	1.60	68.3
50%	0.13	15.00	9.00	0.00	11.20	4.00	50.00	9.00	12.00	13.00	...	0.13	2.40	85.2
75%	0.39	23.00	16.00	0.00	13.50	5.00	75.00	15.00	20.00	21.00	...	0.14	3.30	97.7
max	1.02	32.00	23.00	22.20	14.50	7.00	100.00	25.00	36.00	36.00	...	0.87	11.00	100.0

Basic noise reduction step In this stage, the entries containing NaN values are either deleted or replaced with the closest value, such as the mean, median, and mode. We had about 400 entries in our dataset with NaN values, but since there were about 43000 entries overall—as I said above—I decided to eliminate them entirely.

After that, I updated the attribute datetime's type from "date" to "datetime," as only that datatype's values can be used to train our machine learning system. I also eliminated the duplicate values that were produced as a result of combining rows such as day, month, and year.

Outlier Removal:
An outlier is a data point that is considerably different from the dataset's average value for all of the data points for specific feature. In our example, we eliminated the entries where the panel's power output exceeded three times the average power output over the course of four months.

Feature Extraction:



We used the cyclic feature of the hour and month at first since, without doing so, it would display the difference between 23:00 and 1:00 as 22:00 when, in fact, it is only 2:00. I transformed the hour and month features to cyclic features in the form of sine and cosine at the beginning.

```
In [198]: # Create Cyclic date features
merged_df['Hour_sin'] = np.sin(2 * np.pi * merged_df['Hour']/max_value)
merged_df['Hour_cos'] = np.cos(2 * np.pi * merged_df['Hour']/max_value)
merged_df['sine_mon']= np.sin((merged_df.Month - 1)*np.pi/12)
merged_df['cos_mon']= np.cos((merged_df.Month - 1)*np.pi/12)
```

After completing all of these steps, I plotted correlation metrics to determine the correlation coefficient between various attributes and preserve the feature using pearson correlation.To preserve feature 0.9 was taken as the threshold value for co-efficients.

```
In [54]: corr_features = correlation(X_train, 0.9)
         len(set(corr_features))

Out[54]: 12
```

```
In [55]: corr_features

Out[55]: {'Clearsky GHI',
          'Dew Point',
          'DewPointC',
          'FeelsLikeC',
          'HeatIndexC',
          'Hour_sin',
          'Temperature',
          'WindChillC',
          'mintempC',
          'tempC',
          'uvIndex',
          'windspeedKmph'}
```

Model Selection:

```
In [57]: from sklearn.linear_model import LinearRegression
```

```
In [58]: model = LinearRegression()
```

I chose a linear regression model for my problem because I want to estimate the electricity produced by solar panels by taking into account a variety of factors, including month, day, hour, temperature, humidity, windspeed, DNI, DHI, clearsky DNI, UV index, and pressure.

Model Training:

```
In [51]: from sklearn.model_selection import train_test_split
         X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=101)
```

After that, imported the train_test_split method was imported from the scikit learn.I split the data into train and test data using train_test_split method.I declared the independent and target variable.I imported the linear Regression model from scikitlearn and trained it using fit method.I assigned test_size to be 0.3.After completing the training,predicted the target variable for given values and found out the accuracy of the model.

Following that, the train test split method from Scikit Learn was imported. Using the train test split technique, I divided the data into train and test data.I used 0.3 for test_size.The independent and target variables were declared. I used the fit technique to train the linear regression model that I imported from scikit-learn. After finishing the training, predicted the target variable for the supplied values and evaluated the model's correctness.

```
In [57]: from sklearn.linear_model import LinearRegression

In [58]: model = LinearRegression()

In [59]: model.fit(X_train,y_train)

Out[59]: LinearRegression()
```

Tuning of parameters for better accuracy:

Now that the model has been trained and its accuracy has been determined, I have adjusted the parameters of the linear regression model and determined the accuracy for those parameters. I tweaked the parameters until I achieved the necessary level of precision.

If I still didn't get accuracy even after adjusting the parameters, I'll have to start over with the preprocessing and evaluate the data to identify the features that matter most for the model training and have the most impact on the output.

Methodology:
Gradient Boosting

Library used:
Numpy ,Pandas, Matplotlib, Seaborn,Lightgbm, Scikitlearn

Tools used:
Jupyter Notebook

Data collection:
So gathering data is the first step in solving any data science problem. We gathered the data for day, month, year, hour, minute, DHI, DNI, clear sky DHI, clear sky DNI, cloud type, UV Index, dew point, humidity, wind speed, temperature, and pressure as we have created a solar power forecasting model taking into account various attributes that affect the efficiency or power production of the solar panel.

We gathered information on the aforementioned characteristics from the years 2014 to 2018 and integrated it. The dataset has about 43000 entries overall when all the year's data are combined.

Data Analysis:
The analysis of the data is currently the most important step in the methodology. How does the particular characteristic impact the solar panel's ability to produce energy? which features have a greater and lesser impact on electricity production.

We aggregated the attributes such as day, month, year, hour, and minute into one column and gave it the name date_time after analyzing the dataset.

Preprocessing:
Basic noise reduction step In this stage, the entries containing NaN values are either deleted or replaced with the closest value, such as the mean, median, and mode. We had about 400 entries in our dataset with NaN values, but since there were about 43000 entries overall—as I said above—I decided to eliminate them entirely.

After that, I updated the attribute datetime's type from "date" to "datetime," as only that datatype's values can be used to train our machine learning system.

I also eliminated the duplicate values that were produced as a result of combining rows such as day, month, and year.

Outlier Removal:
An outlier is a data point that is considerably different from the dataset's average value for all of the data points for a specific feature. In our example, we eliminated the entries where the panel's power output exceeded three times the average power output over the course of four months.

Feature Extraction:
We used the cyclic feature of the hour and month at first since, without doing so, it would display the difference between 23:00 and 1:00 as 22:00 when, in fact, it is

only 2:00. I transformed the hour and month features to cyclic features in the form of sine and cosine at the beginning.

After completing all of these steps, I plotted correlation metrics to determine the correlation coefficient between various attributes and preserve the feature using Pearson correlation. To preserve the feature 0.9 was taken as the threshold value for coefficients.

```
In [206]: def correlation(dataset, threshold):
              col_corr = set()
              corr_matrix = dataset.corr()
              for i in range(len(corr_matrix.columns)):
                  for j in range(i):
                      if (corr_matrix.iloc[i,j]) > threshold:
                          colname = corr_matrix.columns[i]
                          col_corr.add(colname)
              return col_corr
```

```
In [207]: corr_features = correlation(X_train, 0.9)
          len(set(corr_features))

Out[207]: 12
```

```
In [208]: corr_features

Out[208]: {'Clearsky GHI',
           'Dew Point',
           'DewPointC',
           'FeelsLikeC',
           'HeatIndexC',
           'Hour_sin',
           'Temperature',
           'WindChillC',
           'mintempC',
           'tempC',
           'uvIndex',
           'windspeedKmph'}
```

Model Selection:

As you may recall, I initially used a linear regression model for my machine learning model, but I was unable to achieve the needed accuracy, so I switched to a Light Gradient Boosting Machine. It is also referred to as lgbm.

```
In [ ]: from lightgbm import LGBMRegressor
        lgbm_base = LGBMRegressor()
```

LightGBM:
LightGBM is a gradient boosting framework based on decision trees to increases the efficiency of the model and reduces memory usage. It has various parameter such as boosting type,num_leaves,max_depth,learning rate,max_depth,class_weight,random_state and many more.

The main difference between Light GBM and other gradient boosting frameworks is that Light GBM stretches vertically. That is, it grows leaf by leaf. Other algorithms scale horizontally incrementally. Light GBM selects the leaf with the lowest error and the highest efficiency. This method is much more helpful in reducing the error rate. That is, it grows leaf by leaf, and the other leaves gradually expand.

Model Training:
Then I defined the independent and target variable for the problem and imported train_test_split method from the scikitlearn library. I assigned test_size to be 0.3

```
In [204]: from sklearn.model_selection import train_test_split
          X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=101)
```

I used the fit method to train the model on training data. determined the model's accuracy and predicted the power for a specific attribute.

```
In [214]: %%time
          # Fit the random search LGBM model
          lgbm_random.fit(X_train, y_train)

          Fitting 4 folds for each of 1000 candidates, totalling 4000 fits
          CPU times: total: 3min 34s
          Wall time: 1h 50min 49s

Out[214]: RandomizedSearchCV(cv=4, estimator=LGBMRegressor(), n_iter=1000, n_jobs=-1,
                             param_distributions={'feature_fraction': [0.1, 0.2, 0.3, 0.4,
                                                                       0.5, 0.6, 0.7, 0.8,
                                                                       0.9],
                                                  'learning_rate': [0.004, 0.006, 0.008,
                                                                    0.01, 0.012, 0.014,
                                                                    0.016, 0.018, 0.02,
                                                                    0.022, 0.024, 0.026,
                                                                    0.028, 0.03, 0.032,
                                                                    0.034, 0.036, 0.038,
                                                                    0.04, 0.042, 0.044,
                                                                    0.046, 0.048],
                                                  'max_depth': [3, 4, 5, 6, 7, 8, 9, 10,
                                                                11, 12, -1],
                                                  'n_estimators': [100, 150, 200, 250,
                                                                   300, 350, 400, 450,
                                                                   500, 550, 600, 650,
                                                                   700, 750, 800, 850,
                                                                   900, 950, 1000, 1050,
                                                                   1100, 1150, 1200, 1250,
                                                                   1300, 1350, 1400,
                                                                   1450],
                                                  'num_leaves': [100, 150, 200, 250, 300,
                                                                 350, 400, 450, 500, 550,
                                                                 600, 650, 700, 750, 800
```

Tuning parameters for better accuracy:
Tuning the parameters allows for more accuracy. I employ randomsearchcv to adjust the algorithmic parameters. I obtained the parameter values necessary to achieve the highest level of effectiveness and trained the system only using those values.

```
In [215]: # Get optimal hyper-params
          lgbm_random.best_params_

Out[215]: {'objective': 'rmse',
           'num_leaves': 350,
           'n_estimators': 950,
           'max_depth': -1,
           'learning_rate': 0.046,
           'feature_fraction': 0.6}
```

If I still didn't get accuracy even after adjusting the parameters, I'll have to start over with the preprocessing and evaluate the data to identify the features that matter most for the model training and have the most impact on the output.

Methodology:
Neural Network

Library used:
Numpy ,Pandas, Matplotlib, Seaborn,Lightgbm, Scikitlearn, Tensorflow

Tools used:
Jupyter Notebook

Data collection:
So gathering data is the first step in solving any data science problem. We gathered the data for day, month, year, hour, minute, DHI, DNI, clear sky DHI, clear sky DNI, cloud type, UV Index, dew point, humidity, wind speed, temperature, and pressure as we have created a solar power forecasting model taking into account various attributes that affect the efficiency or power production of the solar panel. We gathered information on the aforementioned characteristics from the years 2014 to 2018 and integrated it. The dataset has about 43000 entries overall when all the years data are combined.

Data Analysis:
The analysis of the data is currently the most important step in the methodology. How does the particular characteristic impact the solar panels' ability to produce energy? which features have a greater and lesser impact on electricity production. We aggregated the attributes such as day, month, year, hour, and minute into one column and gave it the name date_time after analysing the dataset.

Preprocessing:
Basic noise reduction step In this stage, the entries containing NaN values are either deleted or replaced with the closest value, such as the mean, median, and mode. We had about 400 entries in our dataset with NaN values, but since there were about 43000 entries overall—as I said above—I decided to eliminate them entirely.

After that, I updated the attribute datetime's type from "date" to "datetime," as only that datatype's values can be used to train our machine learning system.

I also eliminated the duplicate values that were produced as a result of combining rows such as day, month, and year.

Outlier Removal:
An outlier is a data point that is considerably different from the dataset's average value for all of the data points for a specific feature. In our example, we eliminated

the entries where the panel's power output exceeded three times the average power output over the course of four months.

Feature Extraction:
We used the cyclic feature of the hour and month at first since, without doing so, it would display the difference between 23:00 and 1:00 as 22:00 when, in fact, it is only 2:00. I transformed the hour and month features to cyclic features in the form of sine and cosine at the beginning.

After completing all of these steps, I plotted correlation metrics to determine the correlation coefficient between various attributes and preserve the feature using Pearson correlation.To preserve the feature 0.9 was taken as the threshold value for coefficients.

Model Selection:
As you may recall, while choosing a machine learning model for model selection, I initially went with a linear regression model and light gradient boosting, but now I'm using a neural network, which is a combination of Lightgraidentboosting, Random Forest, and K-NN.

LightGBM:
With minimal memory usage and improved model performance, LightGBM is a gradient boosting framework based on decision trees.

In addition to many other parameters, it has boosting type, num_leaves, max_depth, learning rate, class_weight, random_state, and many others.

The vertical extension of Light GBM is the main feature that sets it apart from other gradient-boosting frameworks. Leaf after leaf, it grows in this way. Horizontal scaling is used by other algorithms. By using Light GBM, the leaf with the lowest error and most efficiency is selected. Lowering the mistake rate with this approach is far more successful. To put it another way, it grows one leaf at a time, with the other leaves gradually enlarging.

Random Forest:
Random forest is a supervised learning model. It is a combination of multiple trees. It can be used for both classification and regression problems.It has parameters like n_estimators,max_depth,criterion,min_samples_split,max_features,random_state and many more.

K nearest regressor:
K nearest neighbours, a fundamental method, records all potential cases and predicts the numerical target based on a similarity metric, like distance.

KNN regression is easily implemented by calculating the numerical target as the average of the K nearest neighbours. The inverse distance-weighted average of the K nearest neighbours is a different approach. Both KNN classification and KNN regression employ the same distance functions.

Train NN Model using optimal hyper-parameters from above

```
In [63]: dl_best_params = {'kr__n_units3':320, 'kr__n_units2':240, 'kr__n_units1':280, 'kr__lr':0.001,
                           'kr__ker_init':tf.keras.initializers.GlorotUniform(seed=32), 'kr__epochs':70, 'kr__dr1':0.1, 'kr__dr0':0.1,
                           'kr__dl_loss':custom_loss,
                           'kr__batch_size':600, 'kr__act3':'tanh', 'kr__act2':'relu', 'kr__act1':'relu', 'kr__verbose':0}
```

```
In [64]: dl_model = pipeline.set_params(**dl_best_params)
```

Build LGBM Model

```
In [65]: lgbm_best_params = {'objective':'rmse', 'num_leaves':1400, 'n_estimators':900, 'max_depth':11, 'learning_rate':0.008,
                             'feature_fraction':0.6, 'random_state':42}
```

```
In [66]: lgbm_model = LGBMRegressor(**lgbm_best_params)
```

Build RandomForest Model

```
In [67]: rf_best_params = {'max_features':'sqrt', 'min_samples_split':5, 'min_samples_leaf':2, 'n_estimators':1800, 'max_depth':20, 'boots
```

```
In [68]: rf_model = RandomForestRegressor(**rf_best_params)
```

Model Training:

```
In [74]: # Define the stacking ensemble
         stacked_model = StackingRegressor(estimators=base0, final_estimator=base1, cv=4, passthrough=True)

In [75]: %%time
         # Fit the model on the training data
         stacked_model.fit(X_train, y_train)

         [LightGBM] [Warning] feature_fraction is set=0.6, colsample_bytree=1.0 will be ignored. Current value: feature_fraction=0.6
         [LightGBM] [Warning] feature_fraction is set=0.6, colsample_bytree=1.0 will be ignored. Current value: feature_fraction=0.6
         [LightGBM] [Warning] feature_fraction is set=0.6, colsample_bytree=1.0 will be ignored. Current value: feature_fraction=0.6
         [LightGBM] [Warning] feature_fraction is set=0.6, colsample_bytree=1.0 will be ignored. Current value: feature_fraction=0.6
         CPU times: total: 28min 23s
         Wall time: 12min 22s

Out[75]: StackingRegressor(cv=4,
                           estimators=[('dl',
                                        Pipeline(steps=[('standardize',
                                                         StandardScaler()),
                                                        ('kr',
                                                         <keras.wrappers.scikit_learn.KerasRegressor object at 0x0000015F403D9B20>)])),
                                       ('lgbm',
                                        LGBMRegressor(feature_fraction=0.6,
                                                      learning_rate=0.008, max_depth=11,
                                                      n_estimators=900, num_leaves=1400,
                                                      objective='rmse',
                                                      random_state=42)),
                                       ('rf',
                                        RandomForestRegressor(max_depth=20,
                                                              max_features='sqrt',
                                                              min_samples_leaf=2,
                                                              min_samples_split=5,
                                                              n_estimators=1800,
                                                              random_state=42)),
                                       ('knn',
                                        Pipeline(steps=[('knn_standardize',
                                                         StandardScaler()),
```

Then I defined the independent and target variable for the problem and imported train_test_split method from the scikitlearn library. I assigned test_size to be 0.3. I used the fit method to train the model on training data. determined the model's accuracy and predicted the power for a specific attribute.

Tuning parameters for better accuracy:
Tuning the parameters allows for more accuracy. I employ randomsearchcv to adjust the algorithmic parameters. I obtained the parameter values necessary to achieve the highest level of effectiveness and trained the system only using those values.

If I still didn't get accuracy even after adjusting the parameters, I'll have to start over with the preprocessing and evaluate the data to identify the features that matter most for the model training and have the most impact on the output.

2.4 Detail Analysis & Result

The forecasting of PV system power generation in totally based on the location where the PV system have installed and the on time weather condition of that location. The power generation rate is changing dynamically with the various weather parameters so the forecasting of the total power generation of day is very difficult. In our model we use machine learning algorithm for estimation of the power generation.

If we focus on the physics of the PV system then we can able to know that the voltage difference is generated on the photodiode. The photodiode is a PN-junction diode that that generate electric current by consuming light energy. The electric voltage different generated on the terminal of the PN-junction diode depended on the energy of light wave apply on it. Various other environmental parameters are also affecting the voltage generation like surrounding temperature, humidity and other.

Let's take a larger view and understand these parameters for the PV solar panel. The power generation of the panel is based on the intensity of the sun light appear on the surface of the panel. Now this intensity is continuously changing in day time. It is also affected by the presence of the cloud at that particular location. The generation of the power is also affected by the temperature at that location. So we can say that the power generation from the PV solar panel is the function of various weather parameters at that time of the particular location. We have to predict the total generated power in day with the help of the estimated values of the weather parameters of the t day.

From the previous research papers and the past data of power generation and the weather data we can able to know that the power generation of PV system is strongly depended on the solar irradiance of that location. The solar irradiance is varying from location to location and time. In fig.1 we can see that in the clear sky day that is for normal days the power generated by the PV solar panel is highly correlated with the solar irradiance and strongly follows the solar irradiance graph. For the cloudy day or days in rainy season the graph is not strongly matched but it highly follows. Therefor the solar irradiance is the main input parameter that we used to forecasting the feature power generation for the PV system.

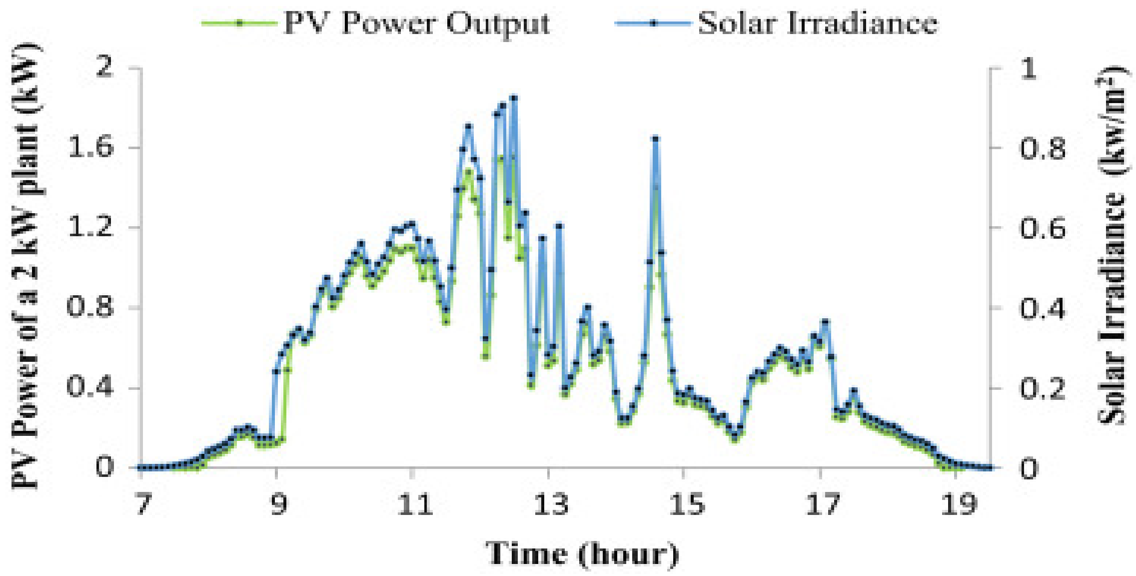
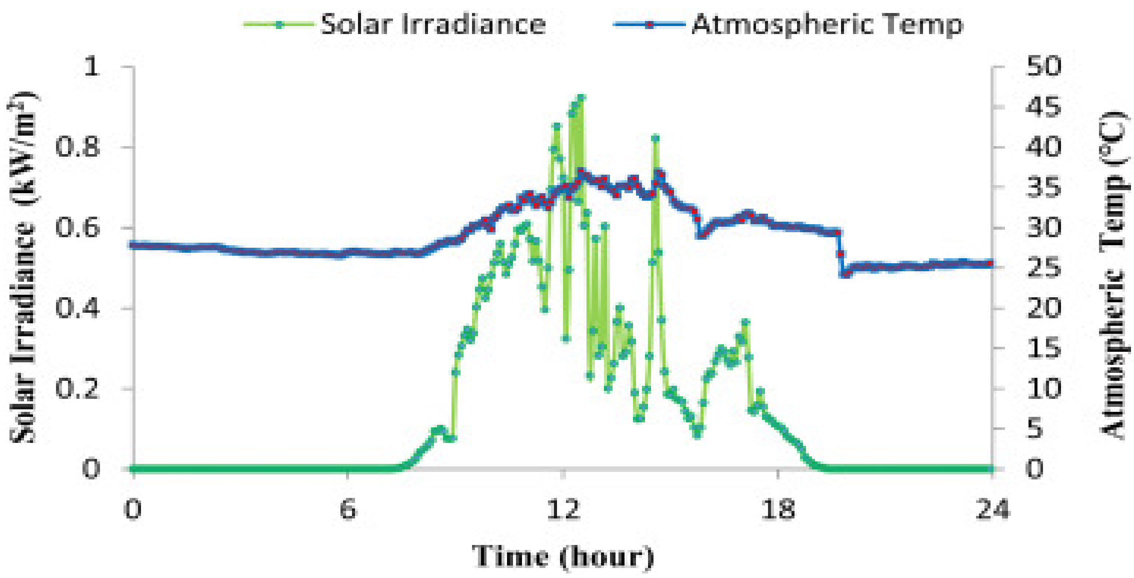


Fig. 2 provides the graph of the atmospheric temperature and the PV power output. From that we can get idea about how the atmospheric temperature is related to the PV power output. We can see that the correlation accrues in the daylight period. In the region of absence of day light we can find the absence of the power generation. In this region power generation not follows the atmospheric temperature. In the presence of daylight we can find the correlation between power and temperature. Though it is not strongly follows as the solar irradiance but not so week also. So we have to consider the atmospheric temperature as one variable in the process of forecasting the generated power.



These how we can see various weather parameters correlated with the power output and take them in as one of the input parameter as in our model. In our machine learning model we use various machine learning algorithms to forecasting the day ahead power generation and take various weather parameters in consideration by taking them as an input parameter. The priority or the weightage of any parameter in output is based on how strongly it correlate the power output. In our study we use three machine learning algorithm for the output generation. These three algorithms are "Linear Regression, Bagging and Neural Network". We derived output on the basis of these three algorithm and minimize the error to achieve 92percent accuracy in forecasting the power generation of the next day.

2.5 Conclusion

In this project we tried to predict the solar power generated for a specific time. And to achieve our aim we used three machine learning models namely linear regression, gradient boosting and neural network. By implementing different model we came across the different accuracy we achieved for different model using the same data set. For say by using linear regression model we got accuracy of 73% which was the lowest amongst all there models. Then by using gradient boosting model we achieved 91.8% accuracy and by the use of neural network model which is combination of random forest method, gradient boosting and linear regression together we got accuracy of 91.7%.

2.6 Future Work

To get more accuracy we can use a better and more promising material as an alternative of silicone material that is perovskite. As the scientist have been working on perovskite material over a decade there has been multiple research paper written on it but the physical use of this material is yet under development so it is rather difficult to get hands on the dataset for the solar power generation by using the perovskite as an alternative.
The main reason to use perovskite is that it has more stability, efficiency and better testing terms than silicone. The research has also shown that by the combination of different layer of perovskite has much lesser carbon footprint than that of a silicone layer cell or silicone over perovskite cell. And even another perk of using perovskite is that it's panel has very less manufacturing cost and can be easily recycled. Basically if we plan on using perovskite in future we need to take care of few parameters which can we difficult to work with perovskite such as air, moisture

and heat. So air and moisture can be handled with the help of encapsulation but the task gets difficult when we need to maintain the heat of the panels when in directly placed in sun for long hours. So due to this factor a best working perovskite cell has only functioning period of approx. 4000 hours at max from continuous UV rays or heat. But we can also not neglect that by the use of perovskite the efficiency as increases from 4 percent to almost 25 percent and that is why it has been a great area for research.
As perovskite-perovskite tandem cell needs very less energy for manufacturing that is approx. 78% percent less than that of perovskite-silicon tandems and can help us save 90 percent energy that is merely used to purify silicon in production of silicon panels. So it can be a very great subject to advance in future. Also we are aiming to publish a research paper on this.

2.7 Project Milestones

Milestone	Date
Understanding of problem by doing literature survey.	08/08/2022 to 14/08/2022
Explored working principles of Photo voltic cells.	15/08/2022 to 21/08/2022
Solar cell, Device Physics, Affecting factor.s	22/08/2022 to 28/08/2022
Concerned about attributes to be used in training our model.	29/08/2022 to 04/09/2022
Data collection and ask help from Waari Energies Ltd.	05/09/2022 to 11/09/2022
First defence	30/09/2022
Data cleaning, Feature extraction and build the model.	03/10/2022 to 09/10/2022
Training & Testing the data in the model and analyse the result.	10/10/2022 to 16/10/2022
Prediction and efficiency analysis, Tune parameters for better efficiency.	17/10/2022 to 20/10/2022
Remaining Parts, extra features, completion of the project.	01/11/2022 to 08/11/2022
Aiming to get data on perovskite material for best accuracy & publish a research paper on this.	Future

Bibliography

[1] Al-Dahidi, Sameer Muhsen, H. Sari, Ma'en Alrbai, Mohammad Louzazni, Mohamed Omran, Nahed. (2022). An adaptive approach-based ensemble for 1 day-ahead production prediction of solar PV systems. Advances in Mechanical Engineering. 14. 1-19. 10.1177/16878132221089436.

[2] Markovics, Dávid Mayer, Martin. (2022). Comparison of machine learning methods for photovoltaic power forecasting based on numerical weather prediction. Renewable and Sustainable Energy Reviews. 161. 112364. 10.1016/j.rser.2022.112364.

[3] Kim, Jae-Gon Kim, Dong-Hyuk Yoo, Woo-Sik Lee, Joung-Yun Kim, Yong Bae. (2017). Daily prediction of solar power generation based on weather forecast information in Korea. IET Renewable Power Generation. 11. 10.1049/iet-rpg.2016.0698.

[4] Dolara, Alberto Leva, S. Manzolini, Giampaolo. (2015). Comparison of different physical models for PV power output prediction. Solar Energy. 119. 83-99. 10.1016/j.solener.2015.06.017.

[5] Ahmed, Razin Sreeram, Victor Mishra, Yateendra Arif, Muammer. (2020). A review and evaluation of the state-of-the-art in PV solar power forecasting: Techniques and optimization. Renewable and Sustainable Energy Reviews. 124. 10.1016/j.rser.2020.109792.

[6] Das, Utpal Tey, Kok Soon Idris, Mohd Mekhilef, Saad Seyedmahmoudian, Mehdi Horan, B. Stojcevski, Alex. (2017). Forecasting of Photovoltaic Power Generation and Model Optimization. Renewable and Sustainable Energy Reviews.

[7] Fjelkestam, Cornelia Cai, Zuansi. (2022). Novel machine learning approach for solar photovoltaic energy output forecast using extra-terrestrial solar irradiance. Applied Energy. 306. 118152. 10.1016/j.apenergy.2021.118152.

[8] Houchati, Mahdi Beitelmal, Monem Khraisheh, Marwan. (2021). Predictive Modeling for Rooftop Solar Energy Throughput: A Machine Learning-Based Optimization for Building Energy Demand Scheduling. Journal of Energy Resources Technology. 144. 1-15. 10.1115/1.4050844.

[9] Mukilan, Mukilan Krishnan, K.Thaiyalnayaki Dwivedi, Yagya Isaac, J. Poonia, Amarjeet Sharma, Arvind Al-Ammar, Essam Wabaidur, Saikh Subramanian, B. Kassa, Adane. (2022). Prediction of Rooftop Photovoltaic Solar Potential Using Machine Learning. International Journal of Photoenergy. 2022. 1-8. 10.1155/2022/1541938.

Group 24 Mini Project Report

ORIGINALITY REPORT

15%	8%	13%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	R. Ahmed, V. Sreeram, Y. Mishra, M.D. Arif. "A review and evaluation of the state-of-the-art in PV solar power forecasting: Techniques and optimization", Renewable and Sustainable Energy Reviews, 2020 Publication	2%
2	Luyao Liu, Yi Zhao, Dongliang Chang, Jiyang Xie, Zhanyu Ma, Qie Sun, Hongyi Yin, Ronald Wennersten. "Prediction of short-term PV power output and uncertainty analysis", Applied Energy, 2018 Publication	2%
3	Zhile Yang, Monjur Mourshed, Kailong Liu, Xinzhi Xu, Shengzhong Feng. "A novel competitive swarm optimized RBF neural network model for short-term solar power generation forecasting", Neurocomputing, 2020 Publication	1%
4	Utpal Kumar Das, Kok Soon Tey, Mehdi Seyedmahmoudian, Saad Mekhilef et al.	1%