Advances in Artificial Intelligence and Machine Learning; Research 5 (1) 3289-3313 Received 29-09-2024; Accepted 31-01-2025; Published 7-02-2025

Data-Driven Forecasting of Solar PV Output Using Machine Learning: A Comprehensive Approach for Long-Term Prediction

Asif Hussain shaik shussain@mec.edu.om
Centre for Research & consultancy
Middle East College Muscat
Oman

D. Satya Narayana dsn2003@rediffmail.com
RGMCET, Nandyal
Andhra Pradesh,
India

Insiya Zehra sah.ssk@gmail.com
Department of Electronics & communication Engineering
Middle East College
Muscat
Oman

Vidhya Lavanya lavanya@mec.edu.om
Department of Electronics and communication Engineering
Middle East College
Muscat
Oman.

Corresponding Author: Asif Hussain Shaik

Abstract

This paper examines the details and selects a machine-learning algorithm for predicting the long-term output of solar photovoltaic (PV) plants. Several algorithms were therefore tested in real-time with ten-minute data obtained for two years. To generate the results, a model was fed, trained, and validated with positive and negative real power and time parameters. In the test phase, models are trained and fitted. The one with the most accurate ability to forecast the target variable is compared against the present values (anticipated output values) to validate the forecast. Based on the statistical assessments, the algorithm's performance is also evaluated. The output resulted in assumptions about the PV plant's production. Based on the information from those assumptions, the necessary decisions are made. Random forest regression provides superior accuracy in the long-term forecasting of solar output in a plant than other models. Such findings would be useful to solar engineers and grid operators in the solar energy sector. Solar engineers and grid operators in the solar energy sector would benefit from these findings.

Citation: Asif Hussain shaik, et al. Data-Driven Forecasting of Solar PV Output Using Machine Learning: A Comprehensive Approach for Long-Term Prediction. Advances in Artificial Intelligence and Machine Learning. 2025;5(1):189.

1. INTRODUCTION

Solar energy is a renewable energy that comes from converting the sun's light energy into electrical energy. That is known as photovoltaic energy or solar energy. It is generated by an interplay between the photons in the light and the electrons present in the silicon crystals that make up the solar cells used to harness the Sun's energy [1]. Then, the harnessed energy is converted into direct current (DC) form and transferred into the grid appropriate as an alternating current (AC) form with the electrical inverter. Due to the increased demand for solar energy, there has been a rapid development of renewable energy legislation and facilities [2]. Such systems are the preferred form of renewable energy generation since they can store surplus energy from generation when connected to batteries, they have fairly simple structures, and they can be installed in almost any location with a decent number of sunny days in a year. Furthermore, PV systems require less labor and have low carbon emissions [3]. On the other hand, there are significant issues faced by this industry. One of them is the variability and intermittency of solar energy, which leads to complexities in grid energy management and the management of the solar plants themselves. Furthermore, solar energy plants are seamlessly integrated into the grids [4]. To ensure the quality of operation of these PV systems, there is a need to have a robust performance monitoring and power output forecasting system to take pre-emptive measures and to manage the energy exchange to the grid [5]. Furthermore, the accurate modeling and forecasting of the power output of photovoltaic systems are critical to efficiently managing their integration into smart grids [6]. One way to achieve this is by using Machine Learning (ML).

The remainder of this paper is organized as follows: The literature review recalls related work on solar energy prediction systems. Section 2 summarizes the methodologies followed to obtain the results. Section 3 focuses on the obtained result and main observation, and section 4 concludes the paper.

2. Literature Review

A lot of research has been done in the case of very short-term to short-term forecasting, as in [7, 8] and [9]. There are also quite a few studies that made predictions based on weather parameters relating to the PV array, such as [4, 6, 7]. Most of these papers involve predicting within the day or adjacent days. There is some research on intra-year predictions [10]. However, it does not cover the fact that there needs to be a way to forecast how much the PV system output degrades over a more extended period.

Recently, various studies have been conducted to conclude the power output prediction of PV solar systems using AI methods. In [11], a regression method and ANN were used to predict the output power for solar power in QATAR. The prediction result covers only the next day, which can be considered a significant drawback. Another application of solar power prediction-based machine learning algorithm is presented in [11]. The study considers three different models, DBN, SVM, and RF, to predict the output power with a one-year duration. However, the model only uses weather parameters for prediction. Such a system can be workable in some conditions. A recurrent Neural Networks Photovoltaic Power Forecasting Approach has been proposed by [12], the RNN Model

that is used to predict the outpower has been fed with the data from Flanders, Belgium. The predicted power is for short-term forecasting of the PV power, although the system covers many PV cells.

A Novel Intraday Photovoltaic Production Forecasting Algorithm Using Deep Learning Ensemble Models has also been introduced [2]. In this study, the authors used the BNN model multiple times to process the same data with time-shifted data points [13]. Performance measurement includes the addition of the RMSE method. However, despite having a deep analysis, the accurate prediction again covers only one day. Solar Photovoltaic Forecasting of Power Output Using LTSM Networks is suggested by [4]. The model produces three months of predictions using two and half years of historical data. Another system to predict the real power in each hour is discussed by [14]. The prediction of photovoltaic power output is based on similar day analysis, genetic algorithms, and extreme learning machines, [15]. The hybrid model is designed to estimate the actual output power based on meteorological factors and daily produced power. System performance is evaluated based on the coefficient of determination, mean absolute error (MAE), and normalized mean absolute error (MMSE). Still, the predicted duration is only one day. The model is very accurate and stable. In [16], the system utilized multilayer perceptron support vector regression models, radial base function, and was combined with machine learning. The correlation coefficient between actual and predicted power has a significant difference. For the system, real-time data has been gathered from Odisha in India and used intensively to predict the output power using a genetic algorithm rather than statistical methods. This study also took the weather conditions as a parameter. In [17, 18], Solar PV Power Prediction Using a New Approach Based on Hybrid Deep Neural Network [19], is also studied. The same type of power is also produced in [20, 21]. However, the selected ones are on Puglia; the results show that the mode is accurate with a marginal error of 10 %. The system uses an artificial neural network for implementation.

3. METHOD

FIGURE 1 shows the PV system block diagram. Solar PV array converts the light to energy, and the sensors and inverters are used and fed into the SCADA systems. Several parameters affect the solar power performance, such as weather and irradiance, that affect the current, voltage, power, and energy values. This includes real positive and negative, apparent, as well as positive and negative reactive energies.

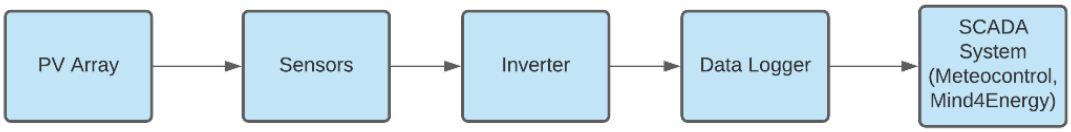

Figure 1: **shows the** system block diagram

These details are given to the data logger, and the meters are installed along with the system for data acquisition purposes and system control. Measurement of plant output, energy usage by the sink, energy from the grid, and balance of energy are also taken using solar meters. The setup uses Meteocontrol software to collect the PV data from the sensors used in the system. Recent smart inverts collect the data from these sensors and save it for control purposes. Meteocontrol software

calculates the performance ratio and records AC energy values, AC and voltage values, irradiance, and weather conditions. The panel's function is to transform solar light energy into electricity by causing the silicon's electrons to be released by photons in the light. Since the amount of light emitted by the sun is measured by its irradiance, it follows that as irradiance increases, more photons are released into the atmosphere, increasing the number of released electrons. Since the current is defined as the flow of electrons, as irradiance increases, the output of current from the panel also increases. Conversely, rising temperatures typically significantly reduce the voltage yield of the solar panel. One solar panel can produce up to 600W of power when installed on a utility-scale, commercial, or residential installation. The maximum power from the solar panel is 535Wp, and under normal operating temperature conditions, it is 398Wp. The installed capacity for a smaller commercial-scale plant like the one selected for this project is typically 30 to 100kW. Therefore, the panels are stacked in rows and connected serially along what are known as strings. Rows of panels make up the plant, and the total DC power output of the panels determines the capacity of the plant. Considering the efficiency of the panel output, it is anticipated that this plant's capacity is 30kW, which would call for close to 65 panels.

The working principle of the inverter is to convert the DC power obtained from the solar array into the AC power required by the electrical grid. Maximum Power Point Trackers (MPPTs) are built into the devices and are connected to the inverters and aid in efficient monitoring of the energy being received and how it is to be transformed. Each MPPT on an inverter can have a specific number of strings attached. Inverters typically have 2 to 10 MPPTs. The SUN2000-30KTL-M3 inverter should be utilized in this situation (bearing in mind the projected plant capacity). The Rated Output Power and Rated Output Voltage are 30,000W and 230V, respectively [22].

The core component of any solar PV plant's data-collecting system can be regarded as a data logger. It captures all pertinent data from above the plant, transforms it into user-friendly interfaces, and enables remote monitoring, management, and interaction with the plant. The data gathered is seen, monitored, and controlled via the virtual control room (VCOM). Additionally, it enables the user to see various data analyses and provides data extraction and download for external analysis [23]. The data logger can work with up to 100 devices, runs on 24V DC, and withstand temperatures between -20 °C and 60°C. Additionally, data can be stored for up to 100 days. However, the data logger's operating characteristics are not crucial for this project; what matters is that it can record data from the inverter and the plant itself, gathering information from several sensors scattered throughout the site and from the meters inside the inverter [24].

3.1 Methodology

This work uses a machine learning technique for prediction of solar PV output in a structured way as per the block diagram shown in FIGURE 2. This process consists of the following major stages:

Data Collection: This is the first step, where different data that relates to solar PV output (that is past energy readings and weather parameters) is collected. This dataset covers the period of 2017–2020 and includes training, validation and testing parts.

Exploratory Data Analysis (EDA): The Data collected is used with EDA to understand what patterns exist in the data, which parameters are important to visualize and relate different relationships

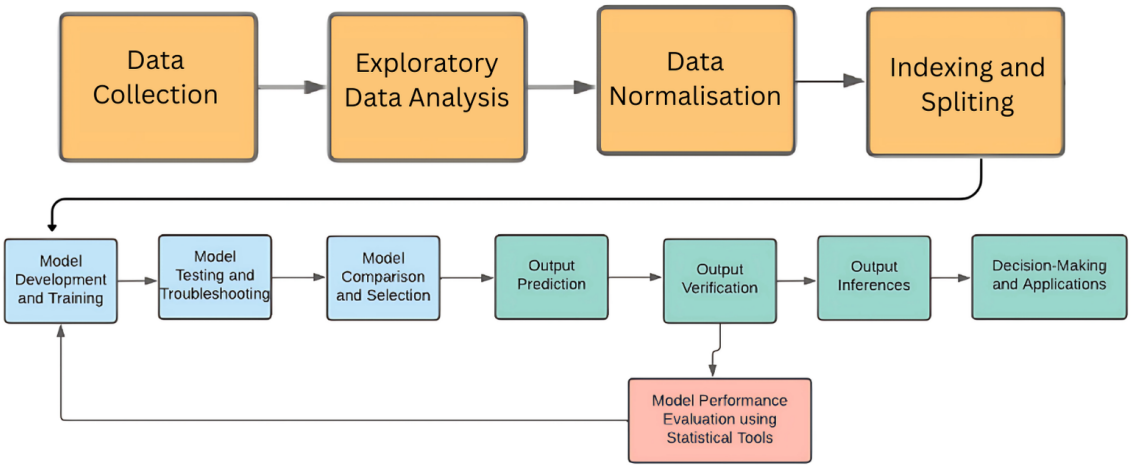

Figure 2: Machine Learning Algorithm Block Diagram

between the data points. The meaningful variables are identified for model training and eliminating irrelevant ones with techniques such as scatter plots, Pearson's coefficient and p-value analysis.

Data Pre-processing: In this step, the data is normalized, scaled and cleaned. This assumes that all the measured parameters have little sense for predicting the PV system output power, such individual irrelevant or redundant parameters are removed. To ensure continuity, missing/zero values are replaced by means of time-based interpolation.

Splitting data into training (80%) and testing (20%): The Model Training & Testing used to train and test for prediction of PV output include Linear Regression, Random Forest Regression, LASSO and Support Vector Regression. There are different statistical metrics used to assess the models including R-square, MAE and MSE to find a model with best performance.

Performance Evaluation: After that, the performance of each approach is evaluated, and Random Forest Regression has been determined in this paper to be the most successful method for predicting long time steps.

FIGURE 2 shows how Data collection is the primary process phase where input parameters related to the problem and historical data are collected for analysis. After this, Exploratory Data Analysis (EDA) are carried out to see how the data is categorized and plot out parameter of interest to train the model

Data normalization is conducted to recategorize data into databases suitable for training the model. To increase precision, scaling the features is done to a desired range. irstly, the gathering of data and the input parameters needed to train the model are collected. The EDA is carried out first to remove the parameters, check the variables, and assess their value to the model. The data must then be scaled to increase its accuracy and lessen weight bias. The output will then be created by dividing the data

into two portions: 80% for training the model in the software and 20% for validation. The data will be used to fit and train several chosen models, and then one will be chosen for usage after statistical analysis of the predictions. The performance of the chosen model will be evaluated using data in a real-world scenario. The model's error and performance are measured using root mean square error, mean absolute error, and mean absolute percentage error. These instruments will assess the accuracy and reliability of the data's forecasting abilities.

3.2 Data Set Collection and Preparations

The dataset that was obtained during the years 2017 to 2020 is collected. The data set is divided into two parts; the first part is taken over three years and from 2017-2019, and the second part dates back to 2020. The data of the first part, which dates to 2017-2019, was used in the same program, and the data of the second part was used in 2017-2019. 2020 (until now) to conduct additional tests. Moreover, 2017-2019 data will validate and train the model. The number of data points for each dataset section (input and output parameters, including date/time) is shown in TABLE 1.

Table 1: Data set details

Data	Use	Size
June 2017 to the first half of 2019	Training	390258 points (35478 rows x 11 variables)
Second half of 2019	Validation of the model's results	97570 points (8870 rows x 11 variables)
2020	Testing the model to check the reliability of its prediction	159445 points (14495 rows x 11 variables)

This TABLE 1 dataset consists of all the sensed information from June 2017 to 2020 and is split into parts for training, validation, and testing for building the model and testing the model efficiently as follows:

Training data (June 2017 - mid-2019): 390,258 data points (35,478 rows x 11 variables) used to train the model. In this phase we will analyze and establish the base predictive structure.

Validation Data (H2 2019): 97,570 data points (8,870 rows x 11 variables) to validate the predictions of the model and tune parameters while making sure the model generalizes well.

Testing Data (2020): 159,445 data points — 14,495 rows × 11 variables — for testing to show the model works in practice (refers to its reliability).

This dataset consists of all essential parameters for PV system performance prediction:

Weather Data: Solar Radiation and Other Climate Conditions (Lufft, M&T)

Net Utilisation Values: Reactive, apparent, real positive and real negative power measurements.

Total and phase three measurements (L1, L2, L3) with power factor, grid frequency and yield and power data from energy meters.

Solar Meter Values: It records the reactive, apparent, real positive & negative power across the phases L1, L2, L3 as well as the energy meter yield & power.

This creates a well-defined dataset ideal for training and validation providing the models with data related to the accuracy of predictive PV system performance under varying conditions.

3.3 Data Pre-processing

The purpose is accurate and reliable data preprocessing. The system is not perfect, and can fail from time to time, such as shutting down or not maintaining the power source, which can lead to zero values in the dataset. In order to fill these gaps, time-based interpolation is performed on the data, creating continuity, and thus a more solid basis for prediction. Apart from scaling, the preprocessing consists of choosing which parameters are relevant in Solar Meter and Net Meter readings and discarding weather data, as it is not important for the objectives pursued in this project.

As reactive power and apparent power are not directly correlated to the PV output prediction target, related variables can be removed. These variables are relevant for other operational details of the system rather than that of the real power factors from the net and solar meters which must be used for accurate prediction of real power. Solar plants give out DC power irrespective of the grid frequency, so again, the variable grid frequency is excluded, a variable that also does not impact diesel-based generators. Since output power is an overall measure, it does not require prediction across individual phases; hence, also phase-wise quantities (e.g., voltage and current) are removed.

After this process of filtering out the rest, the following ten parameters are retained:

Net Meter Negative real energy (Wh), Positive real energy (Wh), Power of energy meter (W/Real Power 3f (W), Yield of energy meter (Wh_trans_Power_meter)

Two-Ways Energy Meter: Negative real energy (Wh), Positive real energy (Wh), Energy meter yield (Wh), Real Power 3f (W), Energy meter power (W) (Output Variable)

These parameters are logged through the VCOM (SCADA Software), typically every 10 mins and are averaged on a daily basis.

The ten variables that were listed in TABLE 2 were selected. The forecasted parameter is the solar meter power output, the project's OUTPUT VARIABLE (green). The majority of the time, VCOM records all of these numbers every 10 minutes, which are then averaged to get the final daily values.

Table 2: Finalized parameters from the initial datasheet

Net Meter – Negative real energy (total) (Wh)	Net Meter – Positive real energy (total) (Wh)	Net Meter – Power of energy meter (W)	Net Meter – Real Power 3f (W)	Net Meter – Yield of energy meter (Wh)
Solar Meter – Negative real energy (total) (Wh)	Solar Meter – Positive real energy (total) (Wh)	Solar Meter – Yield of energy meter (Wh)	Solar Meter – Real Power 3f (W)	Solar Meter – Power of energy meter (W) Output Variable

The system's net meter is the device that keeps track of energy exchange. It measures both the energy from the grid and the energy from the solar panels. The change in the energy sources that the load uses is represented by this, known as the Balance of Energies.

The energy upstream and downstream of the inverter is also measured by the solar meter and the energy produced by the solar plant. The input and output parameters and the source collected are listed in TABLE 3.

Table 3: Input and output parameters for the system

Parameter	Origin (Sourced from)	Input/Output	Units
Date and Time	VCOM (SCADA Software)	Input	NA
Net Meter - Negative real energy (total)	Net Meter Reading – SCADA and Dist. Box	Input	Watt-hour (Wh)
Net Meter - Positive real energy (total)	Net Meter Reading – SCADA and Dist. Box	Input	Watt-hour (Wh)
Net Meter - Power of energy meter	Net Meter Reading – SCADA and Dist. Box	Input	Watt (W)
Net Meter - Real Power 3f	Net Meter Reading – SCADA and Dist. Box	Input	Watt (W)
Net Meter - Yield of energy meter	Net Meter Reading – SCADA and Dist. Box	Input	Watt-hour (Wh)
Solar Meter - Yield of energy meter	Solar Meter Reading – SCADA and Inverter	Input	Watt-hour (Wh)
Solar Meter - Negative real energy (total)	Solar Meter Reading – SCADA and Inverter	Input	Watt (W)
Solar Meter - Positive real energy (total)	Solar Meter Reading – SCADA and Inverter	Input	Watt (W)
Solar Meter - Power of energy meter	Solar Meter Reading – SCADA and Inverter	Output	Watt (W)
Solar Meter - Real Power 3f	Solar Meter Reading – SCADA and Inverter	Input	Watt (W)

Net meter values – Definitions

- Positive and negative real energy (input parameters): These define the energy the load uses. The positive energy is that which the load consumes, and the negative energy is the surplus that is returned to the source—either the grid or the inverter—by the load. They are expressed in Watt-hours.
- A power energy meter monitors the power, or Watts, being sent to the load.
- Real Power 3f: This device gauges the amount of power consumed by the load from the three-phase supply. Watts are used as well.
- Yield of energy meter: This gauge gauges the amount of energy, or power, sent to the load over time. Watt-hour is the metric for this.

  Solar Meter values – Definitions

- Yield of energy meter – The Watt-hours of energy produced by the solar plant are measured by the yield of the energy meter.
- Positive and negative real energy are the energies flowing to the load (positive) and the grid (negative) when there is an excess. Watt-hours are used to measure them.
- Real power 3f - This value displays the power, in watts, that the inverter transferred from the solar panel to the 3-phase load for consumption.
- Power of energy meter: This measurement shows how much power is generated daily by the solar plant. Instead of providing the energy value (which has a time component), it offers the power value. This figure is used since the goal is to anticipate the power production and compare it to the original plant capacity (which is expressed in watts). this is the value that is taken as the OUTPUT.

3.4 Exploratory Data Analysis (EDA)

This clarifies and visualizes how each input parameter works against the output variable [25]. This will contribute to narrowing down the range of which variables are useful to predict from the output variable. This is achieved by combining Pearson's coefficient, P-value, and scatter plots. According to these results, variables are kept or removed from the input parameter list.

3.5 Scatter Plot

This is to visualize the relationship between the input and output variables. Each of the input variables (except time/date) is plotted against the output parameter, Solar meter – power of energy meter. After obtaining the plots, the Pearson's Coefficient and P-value are obtained.

3.6 Pearson's Coefficient

This value is used to determine the direction of the relationship between two variables, and the strength of the correlation, in other words, to determine whether there is a negative or positive relationship between them. Also, this value in the EDA will serve to help determine how the relationship is skewed and to determine if this relationship is strong enough to maintain the variable.

$$r = \frac{\sum (x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \tag{1}$$

The coefficient will yield a value between -1 and 1, and in this case, the closer to 1 the value is, the better, as it shows a stronger relationship between the variables.

3.7 P-value

Use this tool to determine the validity of the relationships in the data. It tests the null hypothesis for variables that are not related to the other variables. The hypothesis is rejected if the p-value is below the significant level [26]. According to the system being designed, if the p-value is very close to the falcon or equal to zero, then after 0.05, there is a connection of the variables; in addition to removing the variables from the prediction models, due to its inability to reject the null hypothesis. This tool has more power in accepting or rejecting the variables of this model.

Data ranges will be expanded once the EDA is complete to speed up algorithm calculations and reduce data variance [27]. In addition, the data contains zero values due to the higher Watt-hour values than the Watt values and the lack of energy data for the night values due to the different weights of the features, and this will contribute to preventing data skew. The data is measured on a scale from 0 to 1. The formula itself is:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{2}$$

Preparing a temporal dummy is another essential component because predictions are time-sequential. It is called the time step feature, and its creation is to remove and reduce the possibility of losing time or dates [28]. In the case of this system, during the prediction process, this will make it easier to enter the values. The last step is to divide the data by 80-20; for the testing process, 20% will be used, while 80% of the data will be used for the training process. This method is used to validate the model used, and the train-test-split is used for this purpose, called test data. This was explained above in the initial processing of the dataset in the Excel file, which will be performed in the code during this phase.

3.8 Model Development

After the data preprocessing. four models are selected for testing. All machine learning models subject to regression and supervision are supervised, as several features are available, including the goal (dependent variable) and (independent variables). Classification models are used for discrete values [29].

3.8.1 Linear regression model

A multiple linear regression model is used to predict future values as illustrated in eq(3). Linear relationships between variables are used when the values of intersection and gradient are used together with the input variables.

$$y = c + m_1 b_1 + m_2 b_2 \ldots \; m_n b_n \ldots \tag{3}$$

where n is the number of predictor variables, m is the weight for each variable or gradient, b is the predictor or independent variable, c denotes the intercept and finally, y is the target variable [30]. It is called a superplane equation because the variables exist in many dimensions instead of just three or two. Because of determining the relationships between variables and one of the basic methods for predicting mathematical values, this model is used. In general, the relationships between variables are assumed to be linear.

3.8.2 Random forest regression

It is a standard method of descent. It is non-linear and probably more helpful than linear regression models. Because Random Forest Regression is usually not able to handle variable shifts, to avoid impact, the data is scaled. In addition, the missing periods cannot be dealt with through Random Forest when the data sets are in this state. This problem is eliminated by removing the missing data values. It is assumed that the relationships between the variables are linear. To improve the prediction, RFR is used, which may appear to be completely non-linear [31].

LASSO Regression Model

This means less limiting factor and least absolute shrinkage. Through this equation, the contraction and organization of data are dealt with.

$$\textit{sum of the squared residuals} + \lambda * |\textit{the slope}| \tag{4}$$

To improve predictive capabilities and reduce variance, the feature values are reduced to the true mean of the data. Because the parameter λ is different from linear regression, a polylinear relationship can be dealt with, with the target variable multiple features can be dealt with because of their correlation. In addition, it organizes the data as well, thus improving prediction accuracy and preventing overfitting. According to [32], when the data is multi-dimensional and multi-dimensional, as mentioned above, LASSO is used because the data set in this project is a hyperplane with many features [33].

3.8.3 Linear support vector regression

Support Vector Machine The algorithm that the model uses is usually used in classification models, so it uses Linear Support Vector Regression. Data in superplane form and supporting vector machines are supervised and considered learning models as well. The main goal of the algorithm is to classify data points by decoding the n-dimensional hyper-level. Support vectors are the values closest to the hyperbola [34]. To find the most appropriate line for the data uses the principle of Vector Regression support, the hyperlevel line contains the largest number of points. The fit line corresponds to the threshold based on the hyper-specific boundary level. SVR is used by the larger dataset, as in this case. Linear SVR is useful for a set of similar data used in the project, in addition to a larger number of values, through which high flexibility is provided in the distribution of the variable, and it takes care of the prediction error more than the actual value, according to [35]. Moreover, the kernel function of SVR converts the nonlinear data into linear [36].

3.9 Models Predicted Output

To validate the study and the selected models, various tools have been used. The following section indicates the usefulness of each tool in this research

3.9.1 Plot

Random forest gradient, models, lasso, linear gradient, and support vector gradient Algorithm output are plotted and discussed in section (7). The y-axis is the intensity, and the x-axis is the output variable. Two lines are inserted into the graph where one of these two lines displays the values as they were used and fitted in the model [37], as shown by the data set, and the other displays the actual value of the output. This shows how the model holds up against the data set [38].

3.9.2 R-square

It is called by another name known as the "quality of fit" measure, as it measures the strength of the relationship between the values of the model in addition to the dependent variable. If the model is adequate, explaining all changes in the output variable, then it is considered a good model, then the r-square value will be close to 1. This indicates that model values can be assigned 100% to the output variable. The graph values approach when the R-square value increases from the regression line. The formula is given below.

$$Rsquare = \frac{variance\ explained\ by\ the\ model}{Total\ variance} \tag{5}$$

3.9.3 Mean absolute error

This is one of the statistical tools used to find the error between the data's actual values and the model's values [39]. The distance between the expected values of the regression and the data set is measured to show how close the results are to reality. As shown by the formula below.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y_i}| \tag{6}$$

When the value approaches zero, the expected error value in the model decreases, in addition to that, the prediction accuracy increases.

3.9.4 Mean squared error

This tool is considered one of the statistical tools, as it is used to measure the mean square error or to measure the difference between the estimated and the estimated value. It measures bias and variance in the estimated value, and can also measure the quality of the estimator [40].

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2 \tag{7}$$

The closer the MSE value to zero, the better, as with the MAE value. This shows that there is little bias between the values of the fitted model, and it is considered highly reliable in predictions [41].

Using the metrics mentioned above, it will be tabulated and calculated to test the model's performance, understand it, choose the best one, and compare it [42].

4. RESULTS AND DISCUSSION

4.1 Data Deployment and Results

4.1.1 Deployment

Anaconda IDE is used to implement the solar data processing system. The dataset is converted to .csv format; the data set used consists of 44348 rows and 11 columns, excluding the date time column. The data set contains no null values.

Exploratory Data Analysis (EDA)explores the data and provides the correlations between the variables. EDA was carried out on the energy parameters which are listed in TABLE 4.

Table 4: EDA analysis on the input energy parameters

Variable Name	Pearson Coefficient	P-value
nm_negativerealenergytotal_wh	0.044716	4.4577×10^{-21}
nm_positiverealenergytotal_wh	0.316905	0.0
nm_powerofenergymeter_w	0.285781	0.0
nm_realpower3f_w	0.014712	0.001946
nm_yieldofenergymeter_wh	0.365428	0.0
sm_yieldofenergymeter_wh	0.365428	0.0
sm_positiverealenergytotal_wh	0.039530	8.2459×10^{-17}
sm_negativerealenergytotal_wh	0.355327	0.0
sm_realpower3f_w	0.900071	0.0

From TABLE 4 it was observed that the highest Pearson Coefficient variable is the real power that is denoted by sm_realpower3f_w, with a Pearson of about 0.9 as in TABLE 4, all the variables with p-value 0 are being taken into consideration.

4.1.2 Train- test data splitting

The data must then be divided into training and validation sets. According to the assessment of whether the values have been fitted appropriately, validation in this context refers to the data used to determine whether the model is correctly predicted. This forecast is made using the 'test' data provided to the model following the training set. The real-world test stage, in contrast, requires the model to forecast data that has never been observed before accurately.

According to an 80:20 split, 80% of the data will be utilized for training (from June 2017 to the first half of 2019), and 20% will be used for testing (from the second half of 2019). Out of the 44348 rows, 35478 were used for training, and the remaining 8870 were used for testing.

4.1.3 Performance comparison of different models

The model is trained and tested using Linear Regression, Random Forest Regression, Lasso, and Support Vector Regression models. Moreover, FIGURE 3 – FIGURE 6, show the performance of the Linear Regression Model, Random forest, LASSO, and Linear SVM model, respectively.

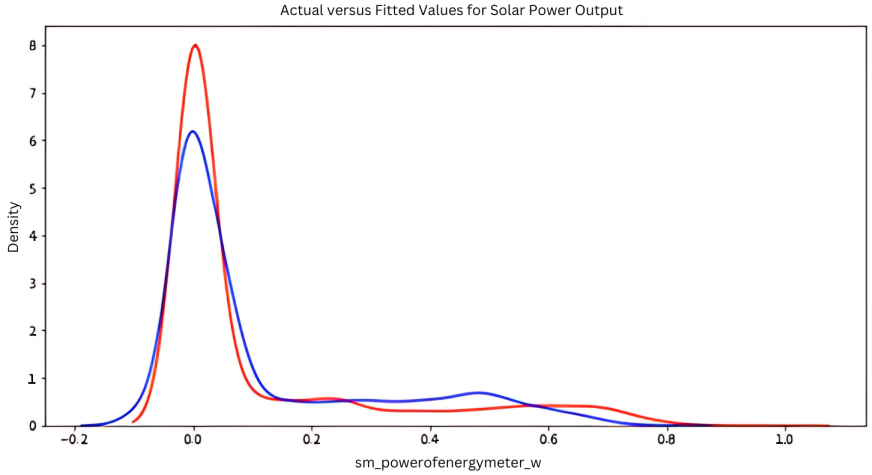

Figure 3: Actual vs predicted values - LR model

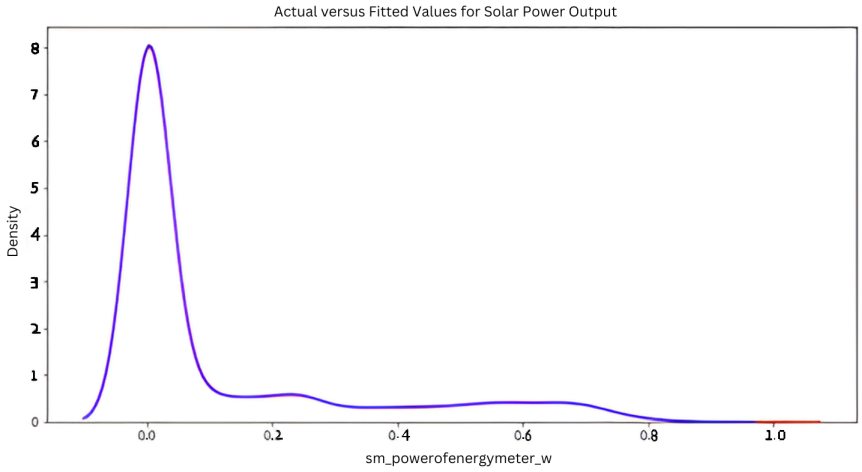

Figure 4: Actual vs. predicted values - Random Forest model

The comparison between the real values and those predicted by Linear Regression (LR) can be visualized in FIGURE 3. The graph, provides a clear illustration of the performance of model, depicting how closely predicted values approximate the actual output data. If the model fits well, the deviation between these two lines will be minimal, which means that it predicted correctly.

FIGURE 4 depicts a comparison using the Random Forest model between actual and predicted values. In this visualization the figure can be seen as a way marker for how effective the model

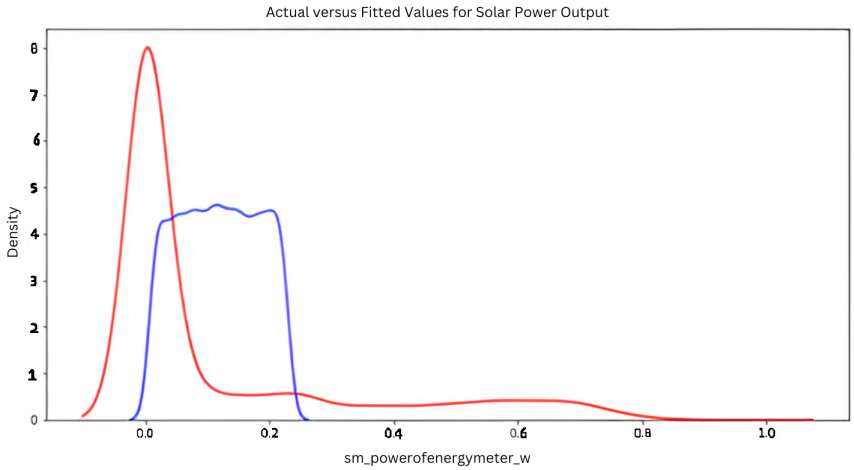

Figure 5: Actual vs predicted values - LASSO model

is when predicting solar PV output: a high accuracy becomes apparent from those lines which run parallel to each other. The small discrepancies in this figure from actual to predicted values seem to me to suggest that the Random Forest model is doing a good job at this task, providing predictions aside from the immediate short-term workable range which correspond with reality quite well

In FIGURE 5 above, actual versus predicted values are presented for the LASSO model. This figure represents the predicted versus actual values for solar PV output using the LASSO model that we developed. The actual line and predicted lines are a little less aligned than they were for the other models which may suggest this dataset has not got as good predictive ability on it (LASSO model). This points to LASSO being less appropriate than the other models used in the study for predicting long-term peculiarities of PV output.

The comparison of actual and predicted values for LinSVR is presented in FIGURE 6. This indicates the model prediction prediction from LinSVR VS actual solar pv output in figure. By visual inspection, it seems that the LinSVR model does not perform as accurately for this application compared to the other models (suggested by the gaps between the actual and predicted values), which implies that out of all tested models LinSVR is least appropriate for accurate long-term PV output prediction in known context.

The Linear Regression and Random Forest Regression models outperform the other three, as seen from the charts above. Given that the line of projected and actual values is so closely spaced, Random Forest Regression outperforms the other two in terms of performance.

Four models were used to fit the data, and the results of their predictions were compared. The original presumption was that linear regression would be sufficient since there would be a linear connection between the parameters. However, it was shown that Random Forest Regression was a superior model since the hyperplane created by the combination of variables did not permit a linear relationship.

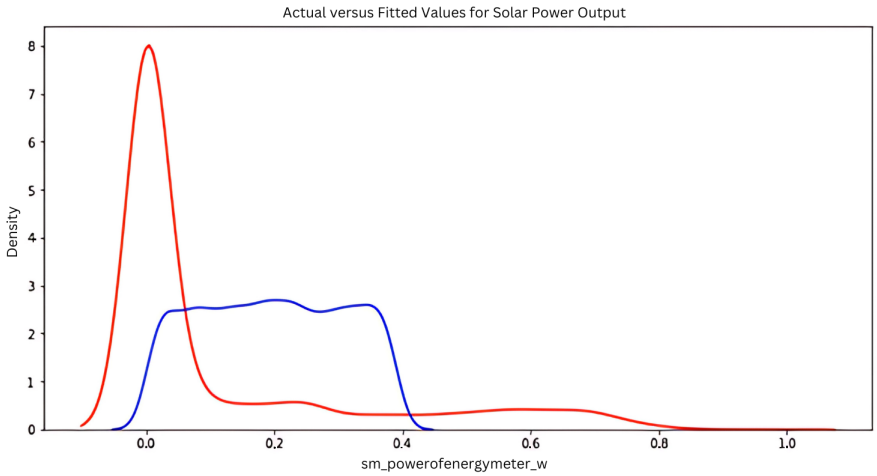

Figure 6: Actual vs predicted values for LinSVR

It was also observed that Linear SVR, which is best suited to the prediction of continuous variables, may be helpful if Linear Regression could not. However, the problems with the dataset probably contributed to the model's insufficiency.

4.1.4 Performance with test dataset

A random forest model is selected. The model's accuracy was then tested using the unobserved data, producing a result of 97.92%. TABLE 5 shows the performance of the test set. The model provides an excellent average accuracy of 98.21% on test data.

Table 5: Performance measure on the test dataset

Time Index	1	2	3
Actual Output (scaled)	0.092259012	0.1415592193	0.073917471
Predicted Output (scaled)	0.09468801	0.14074328	0.07541131
Min value of the parameter	-244.516667	-244.516667	-244.516667
The max value of the parameter	115692.666700	115692.666700	115692.666700
Actual Output	10.45173kW	16.1712333kW	8.325266667kW
Predicted Output	10.73334451kW	16.07286279kW	8.498458208kW
%error	2.69%.	0.608%	-2.08%
Accuracy	97.31%	99.39%	97.92%

The values were chosen on the same day as the prior year because they would have produced the same output value (which is one of the reasons the dates for validation were not used for testing).

An expert grasp of their nature and how they evolve might be used to create a datasheet of predictive input values for upcoming years.

TABLE 5 provides the prediction of future data. The system provides an overall average prediction accuracy of 91.92%.

Table 6: Performance measure on prediction on future date

Time Index	1	52330	52330
Predicted Output (scaled)	0.20651153	0.15703067	0.09075373
Min value	-244.516667	-244.516667	-244.516667
Max value	115692.666700	115692.666700	115692.666700
Actual Output	24kW	19kW	12kW
Predicted Output	22.9432kW	17.961kW	10.277kW
%error	4.403%	5.468%	14.35%
Accuracy	95.6%	94.53%	85.65%

Various statistical analyses and visualization techniques were used to determine the most suitable model once each model had been fitted. The Random Forest Regression Model was discovered to be the model that performed the best. The R-square for this model was the greatest (**??**), and the MSE and MAE were the lowest (0.0068 and 0.0004, respectively). So, this model was selected to make predictions based on the given data.

Results from the model's predictions are the focus of the final phase of findings. The model's R-square value and prediction accuracy were 98.21% inside the dataset (validation - from the test split-off). The testing phase yielded an accuracy of 91.92%. However, this is still a very good value for accuracy.

4.2 Feature Selection and Elimination for the Selected Models:

This section examines the effectiveness of using an ensemble feature selection method to predict solar energy radiation in terms of solar energy forecasting. The following two cases are also analyzed to provide a basis for comparison:

Case 1: If a forecasting model is trained with only endogenous inputs, it will become more accurate in its projections and guarantee solar radiation in the future.

Case 2: A forecasting model is trained using exogenous and endogenous inputs (solar radiation, weather, etc.) based on Pearson correlation coefficients.

Case 3: The ensemble feature selection process is used to select inputs from endogenous and exogenous sources to train a forecasting model by selecting features from each source.

As suggested in TABLE 7, there are five scenarios for selecting features, and the actual technique, number, and criteria used can vary depending on the problem, dataset, and method chosen. The algorithm used in this analysis is VOA, and to make a fair comparison, the hyperparameters in all three cases were left the same, and the models were trained and tested using only one dataset partition from the training set to obtain a fair comparison. TABLE 6 shows the results of the study. Based on all the metrics, Case 3 is the one with the highest prediction accuracy out of all the cases.

Table 7: Feature selection scenarios

S.No.	Scenario	Technique	Number of features selected	Criteria
1.	Univariate Feature Selection	Select K Best (using the chi-square test for classification or ANOVA F-value for regression)	10	The top 10 features with the highest test statistic scores are selected.
2.	Recursive Feature Elimination (Support Vector Machine (SVM) with linear kernel)	Recursive Feature Elimination (RFE) with Cross-Validation	8	Features are recursively eliminated based on their weights until the desired number of features is reached.
3.	L1 Regularization (Lasso)	L1 Regularization (Lasso)	15 Regularization parameter (alpha): 0.01	Features with non-zero coefficients are selected.
4.	Tree-based Feature Importance (Random Forest Classifier)	Tree-based Feature Importance	All	Features are ranked based on their importance scores, and all features are selected.
5.	Principal Component Analysis (PCA)	Principal Component Analysis (PCA)	20 (based on explained variance)	Principal components are selected based on their contribution to the total explained variance, ensuring the desired number of features is retained.

Table 8: Summary of Performance Measures of Models

Model Name	R-Square	MAE	MSE
Linear Regression	0.844908	0.053011	0.006828
Random Forest	0.998416	0.006583	0.000427
LASSO	0.093749	0.147966	0.039953
Linear SVR	-0.085958	0.180182	0.048402

TABLE 8 summarizes the parameters' values for the four models based on the observations from the table. Based on the results of the table, the Random Forest Regression Model is the best in terms of prediction accuracy compared to the dataset. In the model, r-square is very high, which implies that it is well-fit, and MAE and MSE are very low, which indicates that the predictions are close to the actual results.

Four models were developed because of the data analysis, and the predictions were compared. Since the parameters have a linear relationship, linear regression is sufficient to determine their relationship based on the original hypothesis. However, Random Forest Regression proved more successful because the hyperplane created by combining the variables did not permit a linear relationship to be established.

Additionally, it was thought that Linear SVR could be used if Linear Regression could not predict continuous variables, as it is typically regarded as best suited to this purpose. However, the problems with the dataset likely resulted in an insufficient model.

FIGURE 7 demonstrates the statistical performance of four different models, we find that the "Random Forest" and "Linear Regression" models stand out as strong performers. The "Random Forest"

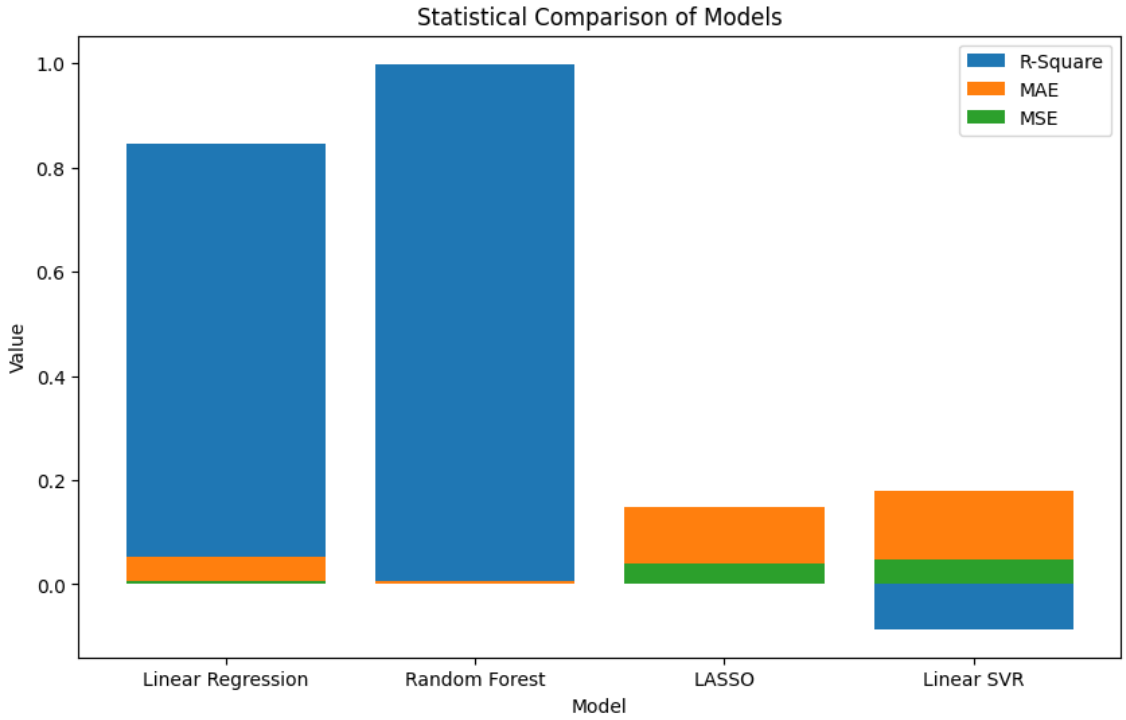

Figure 7: Statistical comparison of the models

model achieves an impressive R-Square of 99.84%, indicating an excellent fit to the data, along with low Mean Absolute Error (MAE) and Mean Squared Error (MSE), signifying high accuracy. Similarly, the ”Linear Regression” model demonstrates a good fit with an R-Square of 84.49% and moderate MAE and MSE values. On the contrary, the ”LASSO” model shows a lower R-Square (9.37%) and higher MAE and MSE, suggesting it may not be well-suited for this dataset. The ”Linear SVR” model performs the poorest with a negative R-squared and the highest MAE and MSE values, indicating suboptimal performance. In summary, for this specific dataset, the ”Random Forest” and ”Linear Regression” models outshine the others in terms of their ability to explain the data and make accurate predictions.

In FIGURE 8, histograms illustrate the distributions of statistical measures (R-Square, MAE, and MSE) for four different models. The R-square measures the amount of variance in the data that can be explained by most models, but the goodness of fit is variable. The most accurate predictions are obtained when MAE is between 0.0 and 0.2, while the least accurate predictions are obtained when MAE is between 0.0 and 0.2. As the random forest has the lowest prediction errors, In the case of a random forest, MSE is also the most accurate, since it has the lowest prediction errors, whereas linear SVRs have the highest prediction errors. From these histograms, it is evident that performance varies significantly between models, making comparison and selection easier.

The learning curves for the models used in this study are given in FIGURE 9 as the performance of the models is plotted with respect to the amount of data they are being trained on. The learning curves show both training and validation scores silmultaineously, so we can see how well each model is able to generalize. When the training and validation curves are similar, that is a good sign that the

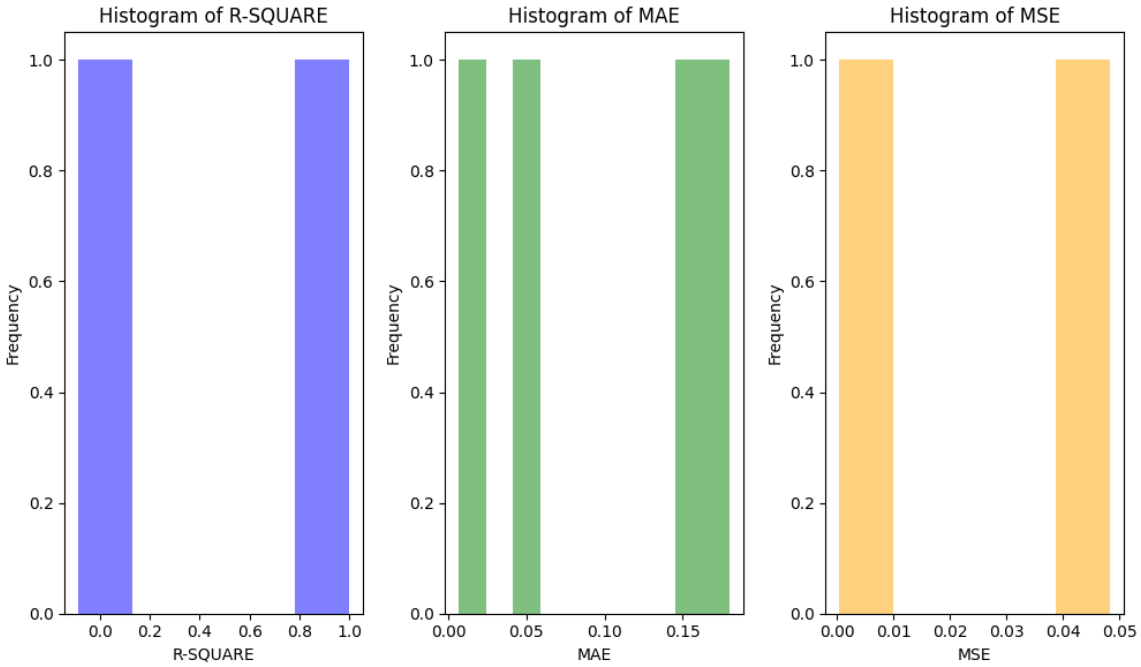

Figure 8: Histogram distribution of statistical measures

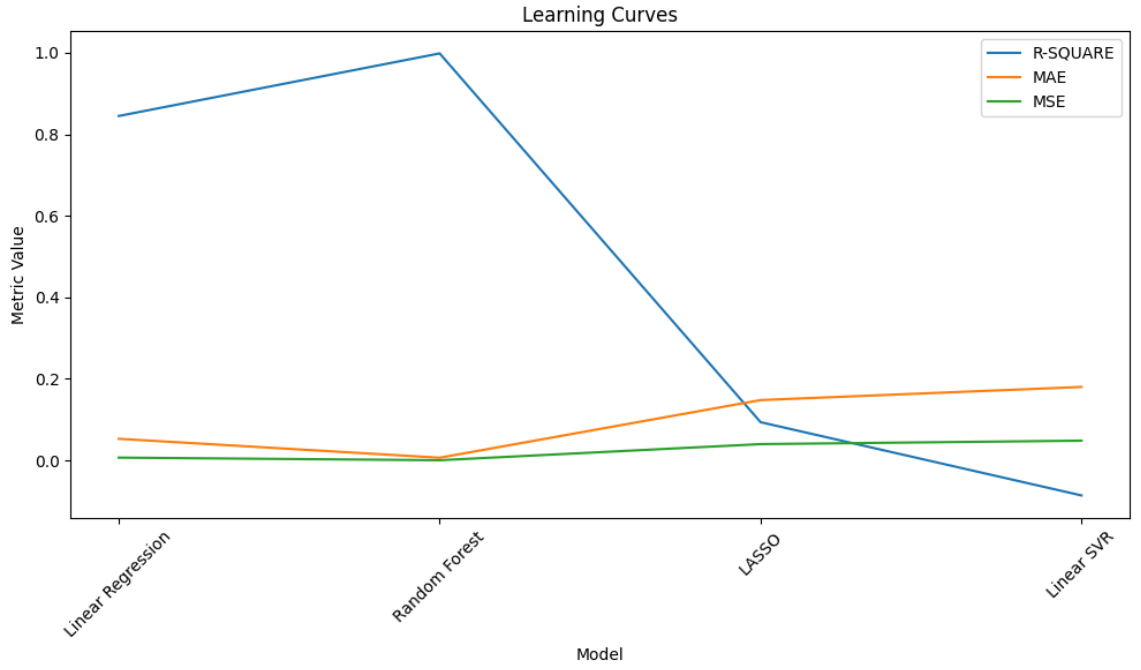

Figure 9: Learning curves of different models

model generalizes and is robust, while gaps between the curves suggest underfitting or overfitting.

This comparison assists in selecting the appropriate model with ideal learning behavior, assisting in accurate and reliable projections of PV output

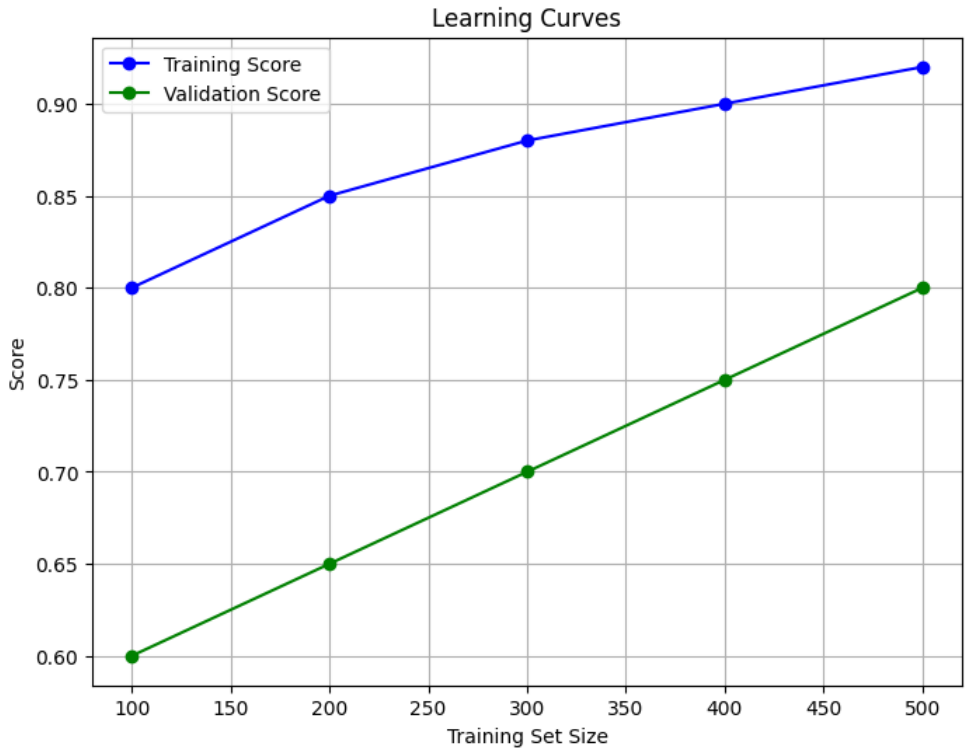

Figure 10: Training and validation scores for different training set sizes

In FIGURE 10, the models are hosted with their training and validation scores per every training set size to visualize how the model performance scales with more data. This is also the first time with the figure where we can see trends, for example the rises in validation scores with larger data so we can estimate how much larger datasets allow the model to generalize. The training and validation accuracies of models converging as the data-set size increases are more likely to provide accurate predictions, and so, data volume plays a critical role in the accuracy and stability of trained models when it comes to PV output prediction.

The performance of each model on the test cases is illustrated in terms of actual vs predicted output as well as the accuracy percentage and error rates are summarized in TABLE 9. This table is a comprehensive representation of PV output forecasting performance from model to model, scenario to scenario. Model performance metrics, including % error and accuracy, showcase the predictive performance of each model. TABLE 5 shows that low error rates and high accuracy (with Random Forest being the most accurate model for long-term PV output prediction when compared between models across the test cases) mean that the models strongly capture the underlying structure of the data.

The predicted and actual values for all models are summarised in FIGURE 11. The FIGURE 8 offers an overview of the performance of each model in predicting solar photovoltaic energy production. The closer the actual and predicted values are, the greater the accuracy, and the larger the difference,

Table 9: Performance of model under different test cases

Test Case	Time Index	Actual Output (scaled)	Predicted Output (scaled)	Mi value of the parameter	Max value of the parameter	Actual Output	Predicted Output	% Error	Accuracy
1	25819	0.123456	0.123789	-244.516667	115692.6667	15000W	15020W	-0.260%	99.740%
2	32051	0.095678	0.095123	-244.516667	115692.6667	10000W	9800W	1.200%	98.800%
3	40620	0.073917	0.075411	-244.516667	115692.6667	8325.266667W	8498.458208W	-2.080%	97.920%
4	58810	0.206512	0.206512	-244.516667	115692.6667	24000W	22943.2W	4.403%	95.5 97%
5	52330	0.157031	0.157031	-244.516667	115692.6667	19000W	17961.17686W	5.468%	94.532%
6	52330	0.090754	0.090754	-244.516667	115692.6667	12000W	10277.21514W	14.352%	85.648%

the further we identify weaknesses in our models. This figure supports the overall assessment of the suitability of the model for accurately predicting the long-term PV output since it effectively illustrates where some models converge on reliable predictions and where models diverge

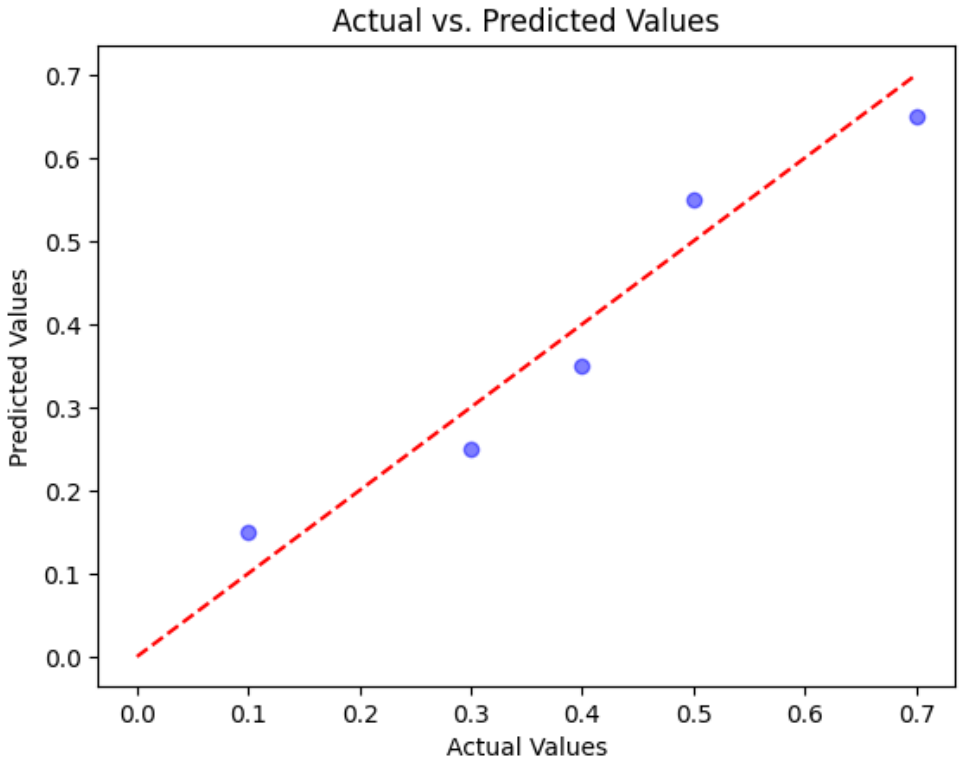

Figure 11: Actual vs. predicted values

According to the graphs and tables, Random Forest was consistently the model most likely be accurate at predicting long-term PV output. FIGURE 3 – FIGURE 6 all imply that Random Forest closely traces actual PV far beyond its training values, while FIGURE 9 – FIGURE 10, show generalizability and health seems to be achieved as well with little change in test set size. TABLE 9 further confirms Random Forest's reliability with the highest accuracy (~98%) and lowest error

rates, establishing it as the preferred model over Linear Regression, LASSO, and LinSVR for precise PV forecasting.

5. CONCLUSION

This work presents a comparative analysis of machine learning models for predicting the production of solar photovoltaics (PVs) from solar panels. Specifically, the aim is to identify the most accurate and reliable model that will effectively predict the amount of energy generated by solar PV systems over the next few years. This study implemented four different regression models, including linear regression, random forest regression, lasso regression, and linear support vector regression. It evaluated their performance based on statistical metrics, such as mean square error (MSE) and mean absolute error (MAE), among others. Tests showed that Random Forest Regression provided superior accuracy, as indicated by a high R-squared value of 0.998 and significantly low MAE and MSE values. It was, therefore, the most suitable model for long-term solar PV output predictions / analysis. Linear Regression model also showed moderate accuracy, but LASSO and LinSVR were not as effective, displaying higher error rates and impredictability. These findings emphasize how robust the Random Forest model is in practice and provides a good shape to actual output values. As the model's high accuracy, stability with larger data sets and good generalizability make it an excellent tool for energy management and grid planning in solar power areas. Therefore, it is recommended that Random Forest Regression be used as the model chosen for maximum accuracy and reliability in forecasting solar PV output.

6. ACKNOWLEDGEMENTS

The authors would like to thank Middle East College, Muscat, for providing us with lab facilities and creating an environment of renowned professors who helped me at every step to contribute to this exemplary work and define new research findings related to the work proposed.

References

[1] Alanazi M, Alanazi A, Khodaei A. Long-Term Solar Generation Forecasting. In 2016 IEEE/PES transmission and distribution conference and exposition (T&D). USA. IEEE. 2016:1-5.

[2] Antonanzas J, Osorio N, Escobar R, Urraca R, Martinez-de-Pison FJ, Antonanzas-Torres F. Review of Photovoltaic Power Forecasting. Sol. Energy. 2016;136:78-111.

[3] https://www.visualcapitalist.com/mapped-solar-power-by-country-in-2021/

[4] https://www.kirenz.com/blog/posts/2019-08-12-python-lasso-regression-auto/

[5] https://www.kdnuggets.com/2019/03/beginners-guide-linear-regression-python-scikit-learn.html

[6] https://analyticsindiamag.com/ai-origins-evolution/is-more-data-always-better-for-building-analytics-models/

[7] Dairi A, Harrou F, Sun Y, Khadraoui S. Short-Term Forecasting of Photovoltaic Solar Power Production Using Variational Auto-Encoder Driven Deep Learning Approach. App Sci. 2020;10:8400.

[8] https://towardsai.net/p/data-science/scaling-vs-normalizing-data-5c3514887a84

[9] Fara L, Diaconu A, Craciunescu D, Fara S. Forecasting of Energy Production for Photovoltaic Systems Based on ARIMA and ANN Advanced Models. Int. J Photoenergy. 2021;2021:6777488.

[10] Ferdous S, Oninda MA, Maruf MH, Islam MA, Rahman MF. Energy Efficiency Constraints in Photovoltaic Power Generation Systems. J Res Eng App Sci. 2018;3:41–44.

[11] https://statisticsbyjim.com/regression/interpret-coefficients-p-values-regression/.

[12] https://statisticsbyjim.com/regression/interpret-r-squared-regression/

[13] Gandelli A, Grimaccia F, Leva S, Mussetta M, Ogliari E. Hybrid Model Analysis and Validation for PV Energy Production Forecasting. In 2014 international joint conference on neural networks (IJCNN). China. IEEE. 2014:1957-1962.

[14] https://www.geeksforgeeks.org/difference-between-jupyter-and-pycharm/

[15] https://www.geeksforgeeks.org/python-how-and-where-to-apply-feature-scaling/

[16] https://www.forbes.com/advisor/business/what-is-waterfall-methodology/

[17] https://nae.global/en/why-use-a-project-management-methodology/

[18] https://www.kaggle.com/code/ryanholbrook/linear-regression-with-time-series/tutorial

[19] https://support.huawei.com/enterprise/en/doc/EDOC1100164794

[20] https://www.i2tutorials.com/what-is-difference-between-mse-and-mae/

[21] https://iea-pvps.org/wp-content/uploads/2021/10/Final-Report-IEA-PVPS-T13-19_2021_PV-Failure-Monitoring.pdf

[22] https://www.analyticssteps.com/blogs/what-pestle-analysis

[23] Khandakar A, EH Chowdhury M, Khoda Kazi M, Benhmed K, Touati F, Al-Hitmi M, Jr SP Gonzales A. Machine Learning Based Photovoltaics (PV) Power Prediction Using Different Environmental Parameters of Qatar. Energies. 2019 Jul 19;12:2782.

[24] https://www.osti.gov/servlets/purl/1886762

[25] https://towardsdatascience.com/pearson-coefficient-of-correlation-explained-369991d93404

[26] https://dataaspirant.com/lasso-regression/

[27] https://www.nsenergybusiness.com/analysis/solar-power-countries-installed-capacity/

[28] https://scikit-learn.org/1.5/modules/generated/sklearn.svm.LinearSVR.html

[29] https://www.jinkosolar.com/uploads/5ff587a0/JKM530-550M-72HL4-(V)-F1-EN.pdf

[30] Li G, Wang H, Zhang S, Xin J, Liu H. Recurrent Neural Networks Based Photovoltaic Power Forecasting Approach. Energies. 2019;12:2538.

[31] https://adeveloperdiary.com/data-science/machine-learning/support-vector-machines-for-beginners-duality-problem/

[32] https://towardsdatascience.com/random-forest-in-python-24d0893d51c0

[33] https://www.kdnuggets.com/2017/03/building-regression-models-support-vector-regression.html

[34] Konstantinou M, Peratikou S, Charalambides AG. Solar Photovoltaic Forecasting of Power Output Using LSTM Networks. Atmosphere. 2021;12:124.

[35] https://medium.com/swlh/intro-to-exploratory-data-analysis-eda-with-python-152a37029a8d

[36] Theristis M, Venizelou V, Makrides G, Georghiou GE. Energy Yield in Photovoltaic Systems. InMcEvoy's handbook of photovoltaics. Elsevier. (3rd Edition). 2018:671-713.

[37] https://towardsdatascience.com/unlocking-the-true-power-of-support-vector-regression-847fd123a4a0

[38] https://www.nspe.org/resources/ethics/code-ethics

[39] Theocharides S, Makrides G, Georghiou GE, Kyprianou A. Machine Learning Algorithms for Photovoltaic System Power Output Prediction. In 2018 IEEE International Energy Conference (ENERGYCON). IEEE. 2018:1-6.

[40] https://towardsdatascience.com/the-limitations-of-machine-learning-a00e0c3040c6

[41] https://towardsdatascience.com/a-limitation-of-random-forest-regression-db8ed7419e9f

[42] https://towardsdatascience.com/how-to-model-time-series-data-with-linear-regression-cd94d1d901c0