WIND AND SOLAR ENERGY PREDICTION using ML/DL models

Mehak Aggarwal, Sonanshi Goel, Shambhavi Rai and Princy Singhal

Indira Gandhi Delhi Technical University For Women, Kashmere Gate, Delhi, India.

Abstract

Purpose: There are numerous commercial providers of high-quality time series data that project developers rely on. However, due to their high cost, which can reach several thousand USD for a single site's hourly time series, these are unsuitable for large-scale academic studies. Another option is to use publicly available data, such as meteorological reanalyses or direct satellite measurements.
Methods: Currently we have done data preprocessing to find the null values in our table and replace them with NaN. Also exploratory data analysis has been done to understand the dataset better and plot the graphs to get a visualization of the same.
Results: After applying and considering ML/DL models, we finalized the best accuracy for SVM model for wind energy and Random forest for solar energy prediction. LSTM and time series forecasting model (i.e, ARIMA) also gave a good accuracy.
Conclusion: Weather complexity makes accurate synthesis of wind output difficult, and commercial confidentiality means that historical data is frequently limited. We present and validate a model for simulating the hourly power output of wind farms located anywhere in the world.

1 Introduction

Motivation: Due to the pressing global, economic, and political issues as well as the alarming levels of pollution in the air, water, and land, renewable and sustainable energy sources including solar, wind, and tidal energy are gaining prominence. Sustainable energy is abundant and good for the environment,

but because of its inherent variations, it appears to be particularly challenging to be integrated into power networks in terms of location and population. Solar energy is one of several sustainable sources that is becoming increasingly important in the energy sector due to its potential to cut carbon emissions and counteract growing electricity prices. The main issue with solar energy is that it cannot be used due to the constantly shifting and unpredictable weather, cloud cover, climate, and seasons. As a result, solar energy generation varies. Therefore, resource planners and businesses are looking for models that take these uncertainties into account for the daily design and management of solar energy production, which could enable them to meet consumer demand and supply regardless of weather conditions. Prediction of short-term solar energy is therefore extremely important.

About Technology: Future global energy mix could include more solar energy than any other source. The most popular solar energy conversion technology at the moment is photovoltaic (PV) [1], which has a low levelized cost of power, a high deployability, and a developed market structure. However, in terms of the collecting of PV data, the quantity of data seems to be less than its popularity, as publicly accessible databases with research-grade PV data are incredibly uncommon. There are two theories as to why. First off, the majority of PV data is retained as confidential because PV systems are primarily controlled by private entities, who don't seem to have any incentive to share the data with the scientific community. Second, data on PV power production alone has low research value because it must be combined with other types of solar data in order to be used in other energy meteorological research projects.

Globally, solar PV is expanding quickly, raising challenging challenges about how to effectively incorporate it into national electrical networks.

Photovoltaic (PV) electricity has quickly emerged as a leading renewable energy source during the past ten years, with globally installed capacity increasing from less than 1 GW in 2000 to 222 GW in 2015.Technically, it makes the functioning of electricity markets more challenging since ramping requires more flexible capacity, and financially, it does so because zero marginal cost renewables stifle meaningful pricing signals in the wholesale market, impeding rational investment decision-making. The demand for new approaches to monitoring, forecasting, and selecting the best renewable alternatives to produce power has increased along with the urgent need to forsake fossil fuels and the development of sustainable means for obtaining energy.

Solar forecasting is a technology that makes it possible to integrate variable, weather-dependent solar power generation into an electric system. 1–3 Therefore, the fact that there has been a lot of interest in the topic during the past ten years is not surprising. Even Nevertheless, there aren't many standardised datasets available for the creation and evaluation of solar forecasting techniques. In particular for those lacking the capacity to deploy and maintain their own solar and meteorological devices, the absence of such datasets restricts comparative comparison of forecasting techniques and slows the

Problems: It is not simple to synthesise time series of wind [2] and solar power as inputs for energy models to study these challenges. These data must have the right level of geographical and temporal precision, maintain cross-temporal correlations, and accurately reflect the behaviour of actual PV plants. Over the past few years, meteorological reanalyses in particular have become an important data source for studies on renewable energy modelling. This is due to a number of factors, including the fact that reanalysis data are typically available globally, cover several decades, and are typically freely available. Reanalyses have the advantage of integrating measurements and numerical models to offer data for areas or time steps where no direct observations are available. Because it depends on complex and unpredictable weather systems, it is challenging to model its time-varying power output, making it difficult to comprehend both its potential and restrictions. We discover that reanalysis data have higher stability but marginally better accuracy. Using thirty years' worth of simulated outputs, we analyse the long-term patterns, variability, and association with power demand across Europe after correcting for systematic bias by matching our simulations to the mean bias in modelling individual sites. The findings quantify how the diverse consequences of growing PV deployment across several European nations significantly alter net power consumption, affect system sufficiency, and ramping requirements. Through an interactive web platform, the simulation code and the hourly simulations for all of Europe are freely accessible. The World Resources Institute estimates that in 2017, coal, natural gas, and oil accounted for 81 percent of all energy consumed globally. Grid supply may be forecasted more precisely with more precise weather forecasts. The utilisation of our project can be used to lower the costs associated with storing extra energy and producing insufficient energy. Many academics are driven to model ideally by taking into account the inherent variations and complexity of solar energy in order to discover efficient and trustworthy solar energy forecast. However, the use of machine learning techniques for accurate prediction is growing.

Our Objectives: In this paper, we explore the properties of solar energy [3] and compare linear and non-linear machine learning approaches to enhance their generalizability for greater adaptation and lower prediction error. Different machine learning-based regression models are used to forecast daily solar energy for this purpose.

Many academics are driven to model optimally by taking into account the inherent volatility and complexity of solar and wind energy in order to find effective and trustworthy solar and wind energy forecast. However, the use of machine learning techniques for accurate prediction is growing. In recent decades, machine-learning techniques have been widely applied in a variety of sectors with data-driven problems. Data mining, artificial optimization, artificial neural networks, statistics, mathematics, and other interdisciplinary fields are all included in machine-learning methodologies. Machine learning approaches look for relationships between input and output data, whether or not there are mathematical problem forms involved. Data analysis is the

process of analysing data. In order to be explicitly programmed, ML uses statistical methodologies. Regression and classification are the two main application categories for machine learning. Regression analysis is necessary for solar power predictions [4]. Lasso Regression (LR), Ridge Regression (RR), Support Vector Machine Regression (SVMR), Decision Tree (DT), and Random Forest are a few ML regression methods that can be utilised for time series forecasting (RF). Long Short Term Memory is a Deep Learning algorithm that can also be used to make these predictions.

2 Related work

There are numerous ways to classify solar energy forecasts. The persistence or smart persistence model, which forecasts future power generation over a limited time period using existing data, is the simplest way (2-3 hours). This technique can be used to establish a benchmark by which other forecasting techniques can be evaluated. A prediction is often made in two steps. A NWP is initially created for a certain area and time period. Using forecasting algorithms, the resulting NWP is then used to project electricity generation. A physical model, a statistical approach, or a machine learning approach could all be used. When ML algorithms and the Smart Persistence (SP) method are evaluated for prediction, ML models surpass the SP Model. Grid management has been challenged by the unpredictable nature of solar resources as solar diffusion rates have risen. One of the most challenging parts of incorporating renewable energy into the system is dealing with its unpredictable nature and intermittent electricity delivery. Forecasting solar power [5] is therefore becoming more and more crucial for grid stability, ideal unit commitment, and economical dispatch. We use machine learning approaches to sort among exceptional solar radiation forecasting algorithms in order to solve the issue. Different regression algorithms, such as support vector machines with different kernel functions and linear least squares, are investigated for creating prediction models. In these tests, we demonstrate that a machine learning approach can accurately forecast short-term solar power using projections of day-ahead sun radiation data. Clustering, classification, and regression techniques were used to create a hybrid or mixed forecasting methodology. Based on the weather forecast for the following day, the model with the closest weather condition is chosen to forecast the power output using cluster-wise regression. Renewable energy sources are increasingly being integrated into electric networks alongside nonrenewable energy sources due to their irregular and variable nature. Soft computing energy prediction methods are required to address these issues. We use a variety of data mining approaches, including gathering historical load data and analysing the properties of the load time series, because the use of other energy sources, such as natural gas and oil, is intertwined with the use of electricity. We compared and contrasted energy consumption trends from renewable and nonrenewable sources. Two new machine learning-based hybrid techniques are multi-layer Perceptron (MLP) and support vector regression (SVR). When

using SVM regression, solar power generation produces satisfactory results. Because it does not thoroughly examine solar power generation and meteorological data, it is limited in its ability to effectively predict other data sets by only employing various SVM kernels after some fundamental statistical data processing. Artificial intelligence (AI) techniques are used to investigate the relationship between anticipated power output and weather conditions recorded as a historical time series.

AI techniques, rather than formal statistical analysis, employ algorithms that may implicitly characterise the extremely complex, nonlinear relationship between input data (NWP predictions) and output power. The ANN is a biologically based brain model. They are used in a variety of applications that employ AI techniques such as supervised, unsupervised, and reinforcement learning. In order to learn from data, the ANN is trained to approximate and estimate the function or connection in the supervised learning strategy. Their models have been improved to forecast the power generation from PV plants. Due to the high unpredictability in important components, particularly the diffuse component from the sky hemisphere, solar irradiance is far less predictable than temperature, even with the cloud graph from synchronous meteorological satellites. Some people thought of using weather forecasts from meteorological websites. Others have attempted to streamline the solar prediction model using nonlinear modelling techniques like artificial neural networks (ANN) [6]. RBF and multilayer perception are two types of networks that are commonly used to forecast global solar radiation, solar radiation on titled surfaces, daily solar radiation, and short-term solar radiation (MLP) [7]. In a three-layer feed forward model, the neural network training method is back propagation. In order to reduce forecast error, the input layer provides an error correction factor based on the expected output for the previous five minutes. An LSTM network [8] will learn a function that takes a series of prior solar irradiance values as input and outputs a solar irradiance value. Deep neural networks, such as the Deep Belief Network, will learn a function that takes a sequence of past sun irradiance values as input and outputs a solar irradiance value (DBN). An LSTM network may learn from a set of observations that are transformed into a variety of events. For the purpose of prediction, the sequence is divided using LSTM.

3 Dataset Description

3.1 Source of the dataset

Data was collected from https://open-power-system-data.org/ which is a free open source platform with data on power systems for 37 European countries. But we chose to focus on a specific country, Germany, due to having the highest proportion of renewable energy than any other country(about 46 percent of its energy come from solar,wind,biomass) and hence it is a good indicator of where the rest of the world is headed. Research paper :

hyperref Link

Utc_timestamp	cet_cest_timestamp	DE_load_actual_ent	DE_load_forecast_e	DE_solar_capacity	DE_solar_generation	DE_solar_profile	DE_wind_capacity	DE_wind_generation	DE_wind_profile	DE_wind_offshore_c	DE_wind_offshore_g	DE_wind_offshore_p	DE_wind_onshore_c	DE_wind_onshore_g	DE_wind_onshore_p
2014-12-31T23:00:0	2015-01-01T00:00:0	1080		37246			27913			667			27246		
2015-01-01T00:00:0	2015-01-01T01:00:0	41151	39723	37246			27913	6852	0.3171	667	517	0.7744	27246	8336	0.3059
2015-01-01T01:00:0	2015-01-01T02:00:0	40135	38813	37246			27913	6054	0.3244	667	514	0.771	27246	8549	0.3134
2015-01-01T02:00:0	2015-01-01T03:00:0	39106	38490	37246			27913	6070	0.3249	667	518	0.7761	27246	8552	0.3139
2015-01-01T03:00:0	2015-01-01T04:00:0	39765	38644	37246			27913	6163	0.3283	667	520	0.7793	27246	8643	0.3172
2015-01-01T04:00:0	2015-01-01T05:00:0	39941	38773	37246			27913	6231	0.3307	667	520	0.779	27246	8712	0.3197
2015-01-01T05:00:0	2015-01-01T06:00:0	39045	37247	37246			27913	6689	0.3471	667	521	0.7815	27246	9167	0.3365
2015-01-01T06:00:0	2015-01-01T07:00:0	40206	40371	37246			27913	10331	0.3761	667	520	0.7801	27246	9811	0.3601
2015-01-01T07:00:0	2015-01-01T08:00:0	41133	42522	37246	71	0.0019	27913	10266	0.3657	667	525	0.7874	27246	9683	0.3554
2015-01-01T08:00:0	2015-01-01T09:00:0	42963	45020	37246	773	0.0207	27913	10029	0.3593	667	527	0.7907	27246	9502	0.3487
2015-01-01T09:00:0	2015-01-01T10:00:0	45088	47101	37246	2117	0.0568	27913	10550	0.378	667	525	0.7872	27246	10025	0.3679
2015-01-01T10:00:0	2015-01-01T11:00:0	47013	49003	37246	3364	0.0903	27913	11390	0.408	667	528	0.7908	27246	10862	0.3987
2015-01-01T11:00:0	2015-01-01T12:00:0	48158	49910	37246	4196	0.1127	27913	12103	0.4335	667	528	0.7912	27246	11575	0.4248
2015-01-01T12:00:0	2015-01-01T13:00:0	47164	47528	37246	3500	0.094	27913	12500	0.448	667	528	0.7912	27246	11977	0.4396
2015-01-01T13:00:0	2015-01-01T14:00:0	46702	47097	37246	2279	0.0612	27913	11921	0.4271	667	525	0.7864	27246	11396	0.4183
2015-01-01T14:00:0	2015-01-01T15:00:0	47403	47383	37246	746	0.02	27913	11851	0.4246	667	524	0.7858	27246	11327	0.4157
2015-01-01T15:00:0	2015-01-01T16:00:0	49440	48936	37246	50	0.0013	27913	13582	0.4866	667	524	0.786	27246	13058	0.4793
2015-01-01T16:00:0	2015-01-01T17:00:0	53410	52288	37246	0	0	27913	15436	0.553	667	524	0.7863	27246	14911	0.5473
2015-01-01T17:00:0	2015-01-01T18:00:0	53672	53367	37246	0	0	27913	16868	0.6043	667	524	0.7857	27246	16344	0.5999
2015-01-01T18:00:0	2015-01-01T19:00:0	53012	52589	37246	0	0	27913	16936	0.6451	667	524	0.7851	27246	17465	0.6417
2015-01-01T19:00:0	2015-01-01T20:00:0	50313	50020	37246	0	0	27913	19181	0.6872	667	522	0.783	27246	18659	0.6846
2015-01-01T20:00:0	2015-01-01T21:00:0	48736	48086	37246	0	0	27913	20353	0.7292	667	521	0.7814	27246	19831	0.7278
2015-01-01T21:00:0	2015-01-01T22:00:0	48624	47803	37246	0	0	27913	21235	0.7608	667	519	0.7787	27246	20716	0.7604
2015-01-01T22:00:0	2015-01-01T23:00:0	45868	44876	37246	0	0	27913	22083	0.7912	667	494	0.7407	27245	21589	0.7924
2015-01-01T23:00:0	2015-01-02T00:00:0	42424	40078	37250	0	0	27926	22472	0.8047	667	305	0.4572	27259	22167	0.8132
2015-01-02T00:00:0	2015-01-02T01:00:0	40577	38711	37250	0	0	27926	22434	0.8391	667	276	0.414	27259	23158	0.8495
2015-01-02T01:00:0	2015-01-02T02:00:0	39707	37820	37250	0	0	27926	24248	0.8683	667	282	0.4228	27259	23966	0.8792
2015-01-02T02:00:0	2015-01-02T03:00:0	39004	37328	37250	0	0	27926	25034	0.8964	667	325	0.4869	27259	24709	0.9064
2015-01-02T03:00:0	2015-01-02T04:00:0	40991	38925	37250	0	0	27926	25721	0.921	667	335	0.502	27259	25386	0.9313
2015-01-02T04:00:0	2015-01-02T05:00:0	42382	40503	37250	0	0	27926	26172	0.9372	667	368	0.5519	27259	25784	0.9459

Fig. 1 Fig1:Dataset used for analysis

3.2 Size of the dataset

Number of rows: 50,000 Number of columns: 16

3.3 Attribute types

1. Utc-timestamp: Start time for the duration for which energy output has been recorded (in timestamp)
2. cet-cest-timestamp: End time for the duration (1 hour plus start time) for which energy output has been recorded
3. Solar-generation-actual: Solar energy that is actually produced in megawatts
4. Load Actual entso-e transparency: It depicts the actual load in megawatts.
5. Load forecast entso-e transparency: It depicts the predicted load in megawatts.
6. Solar capacity: It is the solar energy potential measured in megawatts.
7. Solar profile: It's value ranges from 0 to 1. It is heavily influenced by the tilt and orientation of your panel, your geographic location, the weather on that particular day, and the season.
8. Wind-generation-actual: Wind energy that is actually produced in megawatts.
9. Wind capacity: It is the wind energy potential measured in megawatts.
10. Wind profile: It's value ranges from 0 to 1. It depends on climatic conditions and wind speeds.
11. Wind offshore capacity: It is the offshore wind energy potential measured in megawatts.
12. Wind offshore generation actual: Offshore Wind energy that is actually produced in megawatts.
13. Wind offshore profile: It's value ranges from 0 to 1. It is the wind profile measured from offshore.
14. Wind onshore capacity: It is the onshore wind energy potential measured in megawatts.

15. Wind onshore generation actual: Onshore Wind energy that is actually produced in megawatts.
16. Wind onshore profile: It's value ranges from 0 to 1. It is the wind profile measured from onshore.

4 Dataset Analysis

4.1 Removing missing values in ground truth values (if any)

1. For solar generation actual, wherever there is a null value, it has been filled with the value for a day before.
2. For the first day of the month, since there is no previous day value, it has been filled with 0 assuming there is no solar generation before 6 am.
3. Replacing the leftover Nan with mean of the data of solar generation actual
4. Repeating the same process for Wind Energy

For wind generation actual, wherever there is a null value, it has been filled with the mean value of the entire wind generation actual column.

4.2 Exploratory Data Analysis

In order to run time series machine learning models,data frame is divided into two. On the two distinct data frames,independent Exploratory Data Analysis was performed operations and examined distributions and means. And also examined the time series' seasonal decomposition, paying attention to the trend and seasonal component. The production of solar and wind energy seems to have increased over time, and this is undoubtedly attributable to Germany's continued addition of new solar and wind farms. This analysis came handy in time series forecasting.

Fig. 2 A sample screenshot of Average Daily Solar Production with target variable solar energy

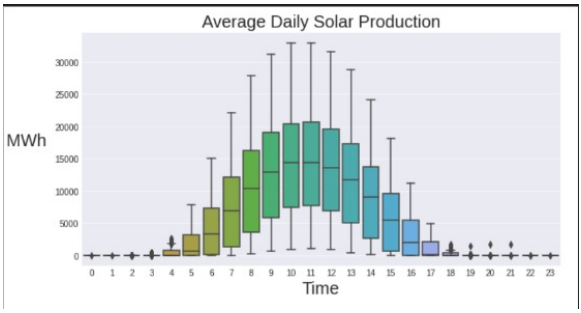


Fig. 3 A sample screenshot of Average Daily Wind Production with target variable wind energy

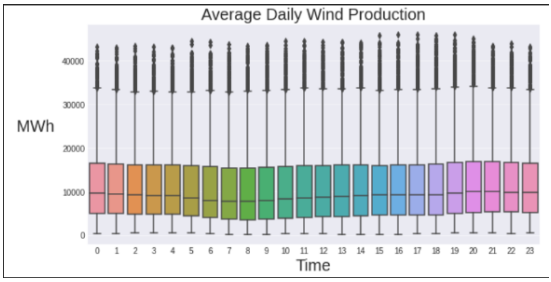


Fig. 4 A sample screenshot of Solar Generation in MW with target variable solar energy

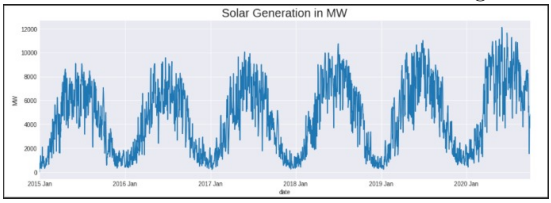
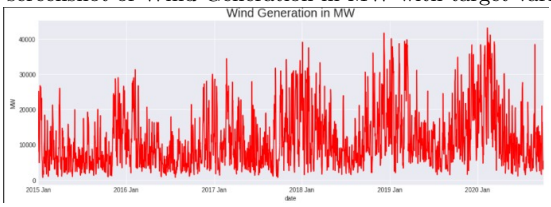
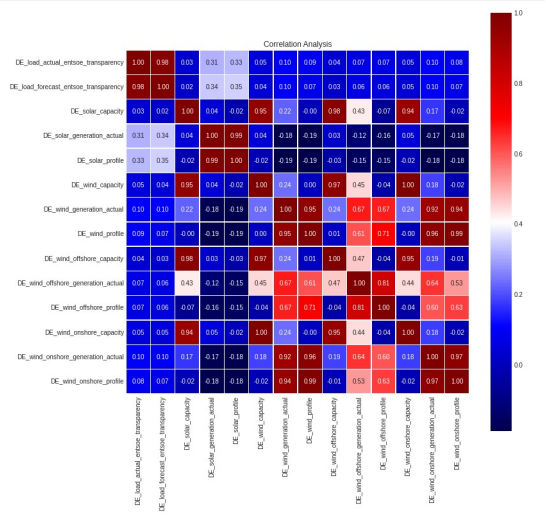


Fig. 5 A sample screenshot of Wind Generation in MW with target variable wind energy



Correlation analysis show that some features are more correlated to target variables. Solar profile is most correlated with solar generation output. In case of wind energy, wind profile, wind onshore profile, wind onshore generation followed by wind offshore profile and wind offshore generation.

Fig. 6 histogram depicting the correlation between variables



Also analysis was done to get an idea of the most frequently occurring wind and solar output value through Histogram.

Fig. 7 A sample screenshot of frequency vs solar energy values

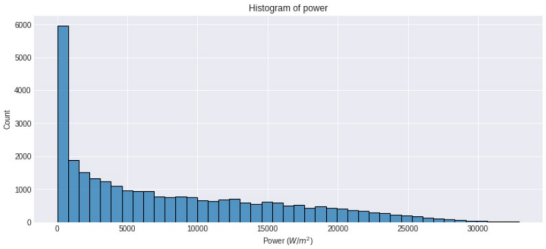
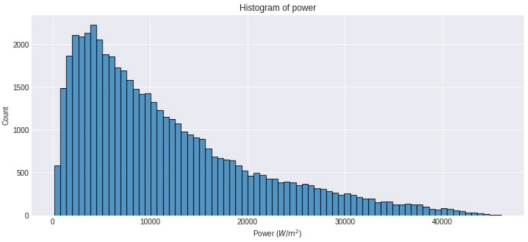


Fig. 8 A sample screenshot of frequency vs wind energy values



5 Experimental Design

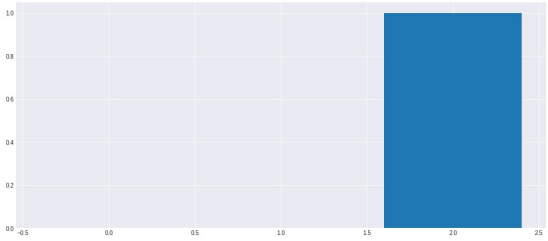
Train-Test Split: A machine learning model's performance can be measured using the train-test split approach. It is used to evaluate how well a certain model performs on new data as opposed to the data used to train the model. The train-test split entails dividing a data set into two parts: the train data set and the test data set. The train data-set and the test data-set are used to fit the machine learning model and train the algorithm, respectively, and are divided into two subsets. The train data-set is used to train the algorithm and fit the machine learning model, while the test data-set is used to produce predictions using the train data as input. The split percentage depends on factors like cost of training and testing the model, size of the data set, etc.

Feature Importance: Along with correlation analysis(Fig6), feature importance was used to select the most appropriate features of our dataset for robust model training. Useless data must be removed for low bias. It is a technique that calculates the score of each input feature. The score indicates the importance of that feature. The method followed for the analysis was:

1. Training dataset on random forest algorithm .
2.Conducting feature importance, once the model is created
3. Plotting the importance on a graph.

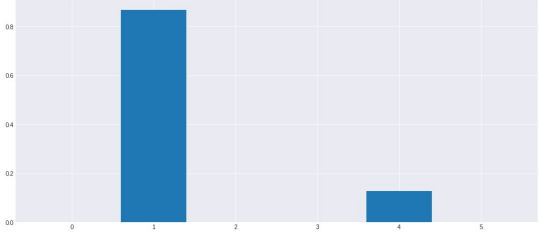
This analysis was used in different models later.

Fig. 9 Feature score for solar energy



This clearly depicts the importance as-Feature 1:load actual entsoe transparency has Score: 0.00002. Feature 2:load forecast entsoe transparency has score-0.00002 and Feature 3 or solar profile has score 0.99996. Hence most important feature is solar profile.

Fig. 10 Feature score for wind energy



This clearly depicts the importance as-Feature1:wind capacity has score 0.00136.Feature 2: wind profile has score 0.86836. Feature 3:wind offshore generation actual has score 0.00167.Feature 4:wind offshore profile has score 0.00019.Feature 5:wind onshore generation actual has score 0.12836.Feature 6:wind onshore profile has score: 0.00006. Hence, most important feature are wind profile and actual wind onshore generation.

5.1 Exploring Machine Learning logics

The sentences present in the article are given as input to the ML/DL model to detect the specific line in which the URL appears. We used five ML models to calculate accuracy and errors .

- **Lasso Regression**
  The term Lasso regression is an acronym for Least Absolute Shrinkage and Selection Operator. It includes a penalty term in the cost function. The absolute sum of the coefficients is represented by this term. As the value of the coefficients increases from 0 to 1, this term penalises the model, causing it to reduce the value of the coefficients in order to reduce loss.

  For predicting solar energy, the model was trained on the feature 'Solar Profile' having 'Solar Generation Actual' as the target column. For predicting wind energy, the model was trained on the feature 'Wind Profile' having 'Wind Generation Actual' as the target column. The hyperparameter tuning in this algorithm was done on 'alpha.' The constant alpha multiplies the regularisation term. When the learning rate is "optimal", it is also used to calculate the learning rate.

- **Ridge Regression**
  Ridge regression includes a penalty term equal to the square of the coefficient. The L2 term is equal to the square of the coefficient magnitude. We also include a lambda coefficient to control the penalty term. In this case if lambda is zero then the equation is the basic OLS else if lambda is greater than 0 then it will add a constraint to the coefficient.

  For predicting solar energy, the model was trained on the feature 'Solar Profile' having 'Solar Generation Actual' as the target column. For predicting wind energy, the model was trained on the feature 'Wind Profile' having 'Wind Generation Actual' as the target column. The hyperparameter tuning in this algorithm was done on 'alpha.' The constant alpha multiplies the regularisation term. When the learning rate is "optimal", it is also used to calculate the learning rate.

- **Decision Tree**
  The most powerful and widely used tool for classification and prediction is the Decision Tree. A Decision tree is a tree structure that looks like a

flowchart, with each internal node representing a test on an attribute, each branch representing a test outcome, and each leaf node (terminal node) holding a class label.
We have taken random state as the only parameter currently, helps in splitting the node searching for the best feature. Our classifier used 0 and 1 as random states and more accuracy was given by random state=1 for both solar and wind energy prediction.

- **SVM**
  Support Vector Machines refer to supervised learning algorithms used for regression analysis [38]. Ideally suited to our purpose, it is a robust classification mechanism that can perform a linear classification based on choosing a boundary that separates data points correctly with maximum gap, allowing for least error as depicted in Fig., and also allowing for non linear classification by incorporating mapping of data points to a higher dimension. Some notable standard applications of SVM are text, image and biological classification. Features included were solar profile for solar energy.For wind energy, all the 5 features were considered.
  Parameters in SVR are C, gamma, Kernel, epsilon and verbose. During hyperparameter tuning, Kernel was fixed 'rbf','sigmoid','poly' and 'linear' one by one. Best R2 score was predicted in linear kernel while training the model on wind energy dataset.Gamma was kept to default. C was kept between 1 to 100 to avoid overfitting and epsilon as 0.5.

- **Random Forest**
  Random forests is a supervised learning approach. It constructs many decision trees and integrates them to provide more accurate and consistent forecasts. It is adaptable enough to be used for regression. We utilized this strategy to create the model because it prevents overfitting in decision trees and also reduces variance, which increases accuracy. It is based on the bagging algorithm and employs the Ensemble Learning technique. Parameters in Random Forest are n estimators and random state.Better R2 score was acheived for random state=1 .

5.2 Deep Learning Model

In our experiment, we also used the Long Short Term Memory (LSTM) [9] network to analyze our dataset. It is a recurrent neural network that can learn sequence prediction order dependence.

- **LSTM**

  LSTM
  It has a chain structure with four neural networks and numerous memory blocks called cells; it is a repeating neural network (RNN) version that is fairly excellent at forecasting long arrangements of information such as

words and market values over time. It differs from a typical feedforward network in that it incorporates a feedback loop. It also includes an unusual element known as a memory cell that stores previous data for a longer period of time in order to make a good forecast.

The Dense Layer
It may be a broadly utilized Keras layer for making a profoundly associated layer within the neural network where each of the neurons of the dense layers gets input from all neurons of the previous layer. At its core, it performs dot items of all the input values together with the weights for getting the yield. A visible layer with one input, one hidden layers of LSTM and dense layer as an output layer that predicts a single value comprise the network. With a batch size of 1, the network is trained for epochs of 15.

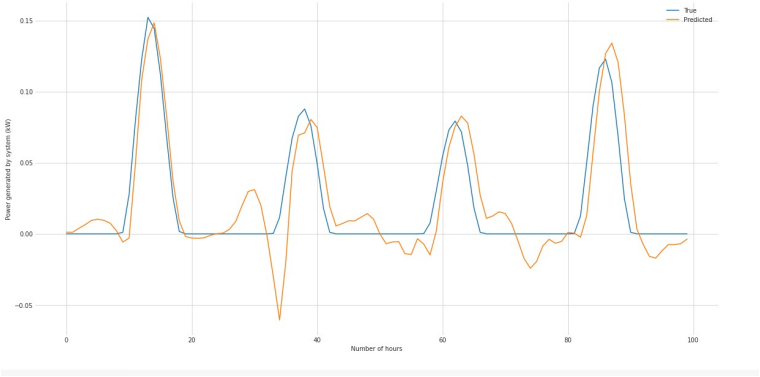


Fig. 11 Power generated by system(KW) Vs Numbers of Hours

5.3 Time Series Forecasting Model

It is an acronym for 'AutoRegressive Integrated Moving Average' [10]. It is a forecasting algorithm that is based on the idea that information in past values of a time series can be used to predict future values.
ARIMA is a type of regression analysis that indicates the strength of a dependent variable in relation to other variables that change. The model's ultimate goal is to predict future time series movement by examining differences between values in the series rather than actual values. Following steps have been followed for the same:

1. Stationery check: Series with difference is stationary because of the p-value less than 0.05
2. Finding the order of differencing for the ARIMA model: to analyse the differencing in two halves of the dataset.
3. Finding order of the ar term: For the ARIMA model we need to first find the the order (p,d,q).

4. Train test split: the dataset is divided into parts i.e, training data and testing data(on which model is tested)
5. Repeat the same steps for wind energy.

6 Results

A comparison table is made to compare the different ML algorithms used on the problem statement. Accuracy, mean square error, root mean square error, mean absolute error are calculated for each algorithm and different hyperparameters and the best one is shown. Two tables each for solar and wind are formed. Further the results are compared with our base research paper and shows that our model has higher accuracy in comparison. Moreover, LSTM model was also applied which gave the R2 Score of 0.62. Time series forecasting(ARIMA) has also been incorporated to but it did not give much good results.

Table 1 ML model training accuracy variation for solar energy

Evaluation Metrics	**R2 Score**	**MSE**	**RMSE**	**Mean Absolute Error**
Lasso Regression	0.99	3449.94	58.73	32.91
Ridge Regression	0.69	15006320.42	3873.79	1818.73
Decision Tree	0.57	14537712.012	3812.83	1800.10
Random Forest	0.998	4639.75	68.11	33.07
SVM	(-0.3157)	46489990.48	6818.35	6304.49

Table 2 ML model training accuracy variation for wind energy

Evaluation Metrics	**R2 Score**	**MSE**	**RMSE**	**Mean Absolute Error**
Lasso Regression	0.99	299772.88	547.51	347.25
Ridge Regression	0.45	43912122.65	6626.62	3458.21
Decision Tree	0.61	24471996.43	4946.91	2912.71
Random Forest	0.993	103377453.54	10167.47	7906.52
SVM	0.9936	0.316	0.526	0.316

Fig. 12 Training Accuracy of based Research Paper

	Solar dataset A		
Models	**MAE**	**RMSE**	**R2_score**
Linear regression-I	**4880032.24**	**6798604.39**	**0.2638796**
Linear regression-II	5486982.49	7252964.41	0.1621314
Ridge regression-I	2398962.75	3376774.70	0.8183589
Ridge regression-II	**2283598.75**	**3220330.28**	**0.8347891**
Ridge regression-III	2232447.59	3146332.09	0.8422805
Lasso regression-I	2920022.00	4022538.93	0.7422443
Lasso regression-II	2919462.23	4021716.91	0.7423496
Lasso regression-III	**2918064.42**	**4019665.79**	**0.7426123**
Decision tree-I	3501301.25	4600911.51	0.6624358
Decision tree-II	**2772250.97**	**3946272.34**	**0.7517212**
Random forest-I	**2275017.03**	**3245764.16**	**0.8320232**
Random forest-II	2339102.77	3317371.13	0.8217654
Random forest-III	2316722.90	3284232.78	0.8252453
Artificial neural network-I	**2651832.01**	**3775846.00**	**0.7726286**
Artificial neural network-II	6441782.03	7522451.15	0.0825467

It can be seen that our model has less variability and also performs better than already existing model used in the base research paper

7 Limitations and Conclusion

1. The algorithms which have been used have been limited to certain number of columns like 1,2.. so on depending upon the correlation of the attributes. In future more columns can be involved while testing.
2. More precise dataset can we loaded with more accurate and less missing values.
3. More deep learning models can be used and results can be compared in future to get precise results.
4. The missing/null values in the dataset can be more efficiently replaced to increase the accuracy.
5. Expanding our analysis by applying state-of-art time-forecasting models.
6. A significant obstacle to the simulations we ran was the absence or inaccuracy of site-level metadata. This could be mitigated by include improved information from additional data sources for measured data in future validation, or by inferring metadata.

Conclusion: The pattern of wind farm power generation has become crucial for both the study and management of power systems. Accurate wind output synthesis is difficult due to meteorological complexity, and little historical data is frequently due to commercial exclusivity. Through this paper we describe a model that can simulate the power output from wind farms anywhere in the world on an hourly basis, and we test it using data from Germany.

Our paper presents a comparative study by applying machine learning algorithms like Random Forest, Ridge and Lasso Regressions, SVM regressor etc along with LSTM Deep learning model to predict the power generation as accurately as possible. It is widely acknowledged that various models function

more effectively in particular regions of the world than others. This could only be determined by comparing results from different models and/or by collecting more measured data from around the world to use as a benchmark for simulated results.

References

[1] Pfenninger, S., Staffell, I.: Long-term patterns of european pv output using 30 years of validated hourly reanalysis and satellite data. Energy **114**, 1251–1265 (2016)

[2] Pasari, S., Shah, A., Sirpurkar, U.: Wind energy prediction using artificial neural networks. In: Enhancing Future Skills and Entrepreneurship, pp. 101–107. Springer, ??? (2020)

[3] Yao, T., Wang, J., Wu, H., Zhang, P., Li, S., Wang, Y., Chi, X., Shi, M.: A photovoltaic power output dataset: Multi-source photovoltaic power output dataset with python toolkit. Solar Energy **230**, 122–130 (2021)

[4] Pedro, H.T., Larson, D.P., Coimbra, C.F.: A comprehensive dataset for the accelerated development and benchmarking of solar forecasting methods. Journal of Renewable and Sustainable Energy **11**(3), 036102 (2019)

[5] Wimalaratne, S., Haputhanthri, D., Kahawala, S., Gamage, G., Alahakoon, D., Jennings, A.: Unisolar: An open dataset of photovoltaic solar energy generation in a large multi-campus university setting. In: 2022 15th International Conference on Human System Interaction (HSI), pp. 1–5 (2022). IEEE

[6] Abuella, M., Chowdhury, B.: Solar power forecasting using artificial neural networks. In: 2015 North American Power Symposium (NAPS), pp. 1–5 (2015). IEEE

[7] Shahid, F., Zameer, A., Afzal, M., Hassan, M.: Short term solar energy prediction by machine learning algorithms. arXiv preprint arXiv:2012.00688 (2020)

[8] Gensler, A., Henze, J., Sick, B., Raabe, N.: Deep learning for solar power forecasting—an approach using autoencoder and lstm neural networks. In: 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 002858–002865 (2016). IEEE

[9] Kuzmiakova, A., Colas, G., McKeehan, A.: Short-term memory solar energy forecasting at university of illinois. University of Illinois: Champaign, IL, USA (2017)

[10] Shadab, A., Ahmad, S., Said, S.: Spatial forecasting of solar radiation using arima model. Remote Sensing Applications: Society and Environment **20**, 100427 (2020)