**Forecasting and predictive analysis of Solar Radiation by machine learning algorithm**

ANISH SI SOUMITRA DAS

19MCA0199 19MCA0233

SITE, VIT, VELLORE SITE,VIT,VELLORE anish.si2019@vitstudent.ac.in [soumitra.das2019@vitstudent.ac.in](mailto:soumitra.das2019@vitstudent.ac.in)

***Abstract-****Solar radiation plays an important role to generate solar energy. It is an important weather feature. In this paper, we have predicted the solar radiation depends on attributes such as humidity, temperature, and wind speed, etc. on the 4 months’ meteorological data. Moreover, later, we have forecasted solar radiation on the same data. For the prediction, we have chosen many machine learning and deep learning model such as artificial neural network, multilayer perceptron, Gradient Boost, Random forest models. For the purpose of forecasting, we have used RNN model. The attributes are sent through the forecasting model and the forecasted value use to portend the value of the solar irradiance by various machine learning and deep learning method. Later, we have compared the results of both prediction and forecasting models and conclude with the best model.*

Keywords—Radiation, Temperature, RNN, Regression, ANN, prediction,forecasting

1. **INTRODUCTION:**

Now-a-days, solar radiation plays an important role in the creation of solar energy and it also plays an important role in the earth. It also depends on many natural components such as wind, temperature sun rays etc. It also has a huge role for solar energy dataset. It can also help us to predict if there a solar system can be established. But, in the reality it has been seen that many data are not available for many places due to the absence of the meteorological space stations. But it is necessary to know all the data for the help of the availability for the solar energy in that place. So It is necessary to predict those data accurately depends on the other attributes. In this scenario, forecasting is also important as depending on the correct forecasting of the data, the establishment of the solar plant, where solar energy plays an important role, can be done.

Machine learning is a part of artificial intelligence that helps to develop forecasting and prediction model. It also investigates the algorithms that can help to find future prediction and forecasting data from the help of the previous data. As we all know, that machine learning as well as deep learning and statistical models are widely used in this prediction and forecasting purpose. There are many famous statistical models to forecast a data for the future use. Those model mainly based on the auto-regressive and moving average. We had to check the trend and seasonality of the data. Here in this purpose we had to accept seasonality as the season used. For this forecasting purpose we can also use RNN (Recurrent Neural Network) for better results.

For prediction purpose, there are many machine learning models and as well as many deep learning models. Many used tree based ensemble methods such as random forest and other while some used gradient boosting methods such as ADAboost, XGboost etc. Some also used many regression methods such as Support Vector Machine, Radial basis Functions etc. Now-a-days, multilayer perceptron and ANN are also widely used as the prediction methods. They have also used some hybrid technique by combining two or three models. There are also many other network to work in this purpose.

Here in this paper, we use both forecasting and prediction techniques to predict the data of solar radiation. We use forecasting methods to know the climate feature such as humidity, temperature, wind speed, pressure, wind direction etc. and then using that forecasted data, we input that values in the predicted model to find the solar radiation in the future. In this method as all the independent attributes can be generated by the help of forecasting methods, we can use this model to predict the future where attribute and data are no present. It may help to predict where the solar plant can build and the energy it can generate in the near future.

So, in this paper, we tried do that for the help of the people in large by predict and forecast the data of the solar radiation which plays an huge role in the human life.

1. **LITERATURE SURVEY:**

In this paper[1], authors proposed a clustering approach to forecast the solar radiation. They used normal elbow method to analyze the number of clusters. This is developed to determine the input to their main approach Bayesian Neural Network (BNN). Their input is the values of the wind direction, wind speed, temperature and the predicted output is the solar radiation. They mainly used a four-stage forecasting method. They are preprocessing clustering splitting and forecasting method. Preprocessing stage is helps to decomposes the data into an appropriate resolution level by entropy-based criterion. Later they compared their results with the traditional forecasting model and find that it gives better results. So they showed that results in the tabular form to compare those results.

In this paper[2], they have proposed a hybrid method. Their dataset is on the hourly time scale. This paper is mainly focusing on the forecasting the one hour ahead solar radiation on the cloudy day. They used the combination of the Auto Regressive and the Dynamic System Model. It is mainly another type of the forecasting model. Also they have done this by two methods, one is before correction to the predicted value as the difference between the present solar radiation value and the value of the one time lag step and another is after adding the correction. They found that the after the correction the accuracy increased 30%. Later they compared their result with the other three forecasting method such as a hybrid model, neural network approach and a basic model and found that new proposed method is working good for cloudy days.

In this paper[3], they have used gradient booster regression, Support Vector Regression(SVR) and random forest regression to forecast solar radiation values. They have also used a new hybrid approach by combining those three to downscale. This new method improves the accuracy of 3 hour accumulated radiation forecasting which are provided by the Numerical Weather Prediction for their dataset which are taken for the Spain. They also showed the disaggregation of the 3-hour forecast into hourly values using interpolation method which is mainly based on the theoretical and experimental radiation model.

Later they showed their results in the tabular format and it shows that the disaggregated forecasts are the better than the basic NWP. They also tried to show the way of future work with the help of this method.

In this paper[4], authors proposed a benchmarking of various supervised machine learning model and two forecasting statistical model to predict the Global Horizontal Solar Irradiance. Those models are Gaussian regressors, SVMs and multilayer perceptron. They also used AR model and two naïve bayes models on the persistence of that irradiance and the clear sky index. Here they did on the dataset on the taken from the three French islands called Crosica, Guadeloup and Reunion. Their main focus is on an hour ahead solar forecasting methods. They also found that their work improves the accuracy than the method of linear AR model and scaled persistence models. Later they have compared all the results by graphically and tabular format.

In this paper[5], the authors reviewed many papers and stated about those machine learning models which are used vastly to predict the solar radiation. Mainly this forecasting can be done by the two different approaches such as physical model and machine learning models. So their main focus in this paper to give an overview of the machine leaning models. They have shown here almost every regression and forecasting techniques such as support vector machine, ensemble learning, neural network etc. It has been seen that the mostly this prediction is done by the ANN now-a-days. They have shown it graphically and at the end they have shown the results of each paper they have reviewed in the tabular format and also shown the dataset types which they have used.

In this paper[6], author used multivariate logistic regression method to predict the solar irradiance. They have also showed that beside clearness index, it is also necessary to use predictors such as air temperature, cloud visibility and cover, solar altitude etc. to predict the diffuse components. They also found that the Mean Absolute Error (MAE) for this logistic regression with the use of above mentioned predictors was less than 21.5 W/m2 and 30 W/m2 respectively for the two dataset of Hong Kong and USA. At the end of the paper, they have concluded by the graphical analysis of their model and comparison of the different predictors.

In this paper[7], the author used two forecasting method for the solar radiation. They are ARMA and Nonlinear Autoregressive Neural Network Model. Their main target in this paper is to forecast the solar radiation multiple hour ago by using an hourly data of solar radiation. Their approach is of the 915 hour earlier. They here showed three models. Firstly, in ARMA model they have predicted the future values. For this they converted the non-stationary time series into stationary one. They have also used NAR models for nonlinear data. Later they calculated NRMSE vales for both the methods. The first one’s value is 0.3241 and the second one’s value is 0.2634. Later they graphically showed the comparison of their model with other models.

This paper[8] is mainly a case study of the forecasting models for the data taken in China. This paper is mainly focusing on the two machine learning algorithms such as Support Vector Machine (SVM) and extreme gradient boosting methods. It is the data of daily solar radiation. From the comparison they have shown that XGboost has better accuracy as well as it is very stable with average increase 6.3% RMSE while SVM has the stability of10.5% RMSE. But according to them, XGboost works better in the training phase while SVM works better in the testing phase. XGboost has also greater computation speed. So from their comparison, they recommended XGboost than SVM. Later, in the paper they showed these all data and results in the tabular form to show the comparison.

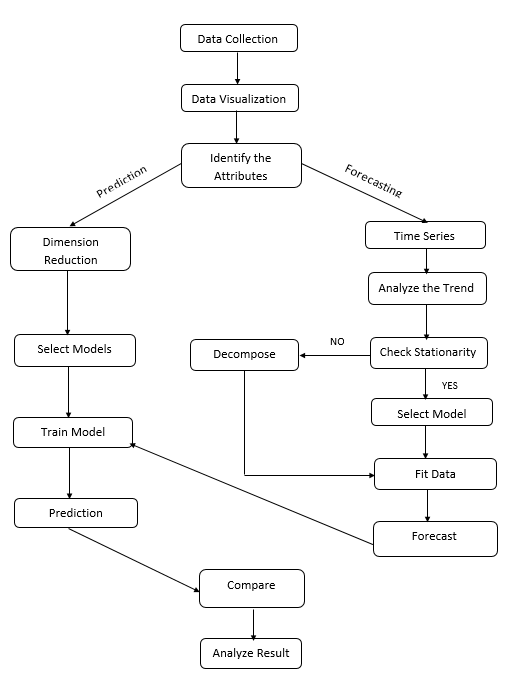
In this study[9], they have taken hourly meteorological dataset and proposed a novel hybrid model by deploying random forests algorithm and firefly algorithm for predicting hourly global solar radiation. They have used firefly algorithm to compute the random forests technique. Later, they have compared the results of the proposed model with various deep learning techniques such as conventional ANN network, optimize ANN and find that proposed model shows better performance than a fore mentioned model in terms of prediction accuracy.

In this experiment[10], they design and implement Support Vector Machine (SVM) model for the management of energy generation emerged on experimental work. To do so, they have taken two input data, solar radiation and ambient temperature and photovoltaic current is taken as output data. They have also build models using various machine learning technique on the same dataset. Moreover, the result of the proposed model is compared to others model and find that proposed model shows better performance than others do.

1. **Methodology:**
2. ***Data Collection and description:***

This data is taken from the Kaggle dataset. This data is given by the Nasa observatory in a hackathon. This is the dataset taken for the Hawaii Island of Pacific time zone. This is a time series data having the features temperature, pressure, wind direction in degrees, wind speed, humidity, radiation, etc. The dataset has 32686 rows and 11 columns. Data is from September,2016 to January,2017.

1. ***Work-flow:***



1. ***Data preprocessing:***
2. **Data visualization:**

This is a time series data. So we have to visualize this data through that time format such as by the month, week, day etc. and we should also visualize the data according to the dependency of the different attributes through a hit map among the non time-series features. It will help us to find the correlation among the different features of the dataset. The column datatype is either integer/float or object type. But for time-series purpose we identified the date-time column and try to convert that into dateTime format.

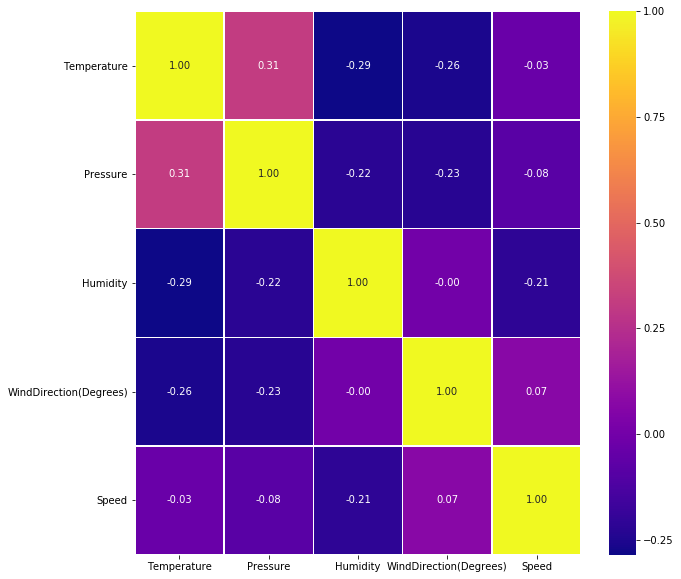


FIG: 1 Hit map plot

From this plot we can show that all the features are quiet uncorrelated, so we have to use all the features.

1. **Time series data indexing:**

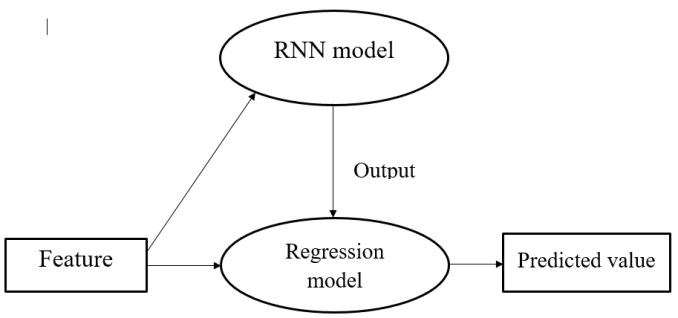
In this dataset one column is named as UnixTime. So we have to change that time to the time series that means year-day-month format along with the hour, minutes and seconds. For this purpose, we set that UnixTime column as the index and convert that data into the string. Later from that string data we changed that column into the Date-Time format.

1. **Null value visualization:**

After that we need to find that Null value in each column. As this dataset is taken from the genuine source of NASA observatory, it has no NULL data.

1. ***Model overview:***

This whole model for prediction has two main parts. One is for forecasting and other is for prediction purpose. For this forecasting purpose, we would use RNN a deep learning model for each individual selected features and pour that into the prediction model. Here from that Heat map we can see all the features are temperature, humidity, pressure, wind direction and wind speed. Now we give input those value into that prediction model.



1. ***Feature Selection:***

From this dataset, we choose temperature, humidity, wind direction, speed and pressure as the features, by which we can predict the solar radiation of a day. So, these are the features of the dataset and we take them as input of our model.

1. ***Train and test split:***

We then divided our whole dataset by training and testing data. We use 80% of our dataset as training purpose while other in testing purpose we will use 20% of our data. The train set is comprised of 26148 rows while test set is comprised of 6538 rows.

1. ***Model Selection:***
2. **Forecasting purpose:**

For forecasting purpose, we use RNN model to forecast the values of the features for the future prediction. Here for the greater accuracy we need to use two RNN models. For the features temperature, humidity and pressure we use LSTM architecture having the unit 4 and the input size of 1. It has different number of layers for different types of data. Here we use RNN because it gives maximum accuracy and lowest RMSE error than the other statistical model, which are mainly forecast from the trend of the data.

For the forecasting of the data of column wind speed and wind direction we use another LSTM architecture with 256 units as input. It is also an RNN model. It has also the input size as 1 unit.The main reason behind using RNN is it gives the minimum error which helps it to evaluate the model as the best one.

1. **Prediction model:**

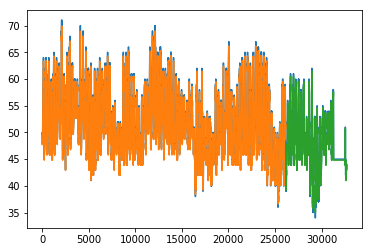
Here for the prediction purpose we will use gradient boosting, random forest model, multiple linear regression model and Artificial Neural Network. We will use these model as it gives better accuracy and minimum error among those previous values and the predicted values.

1. ***Build data frame and prediction:***

Firstly, we need to build the dataframe for each of the features. Each dataframe has one feature variable and another column of time series. So for those, each data frame, we will generate the forecasted data of the time having in the test dataset. This data frames are inputted in the RNN model built for each of those respective features and generate the output forecasted data.

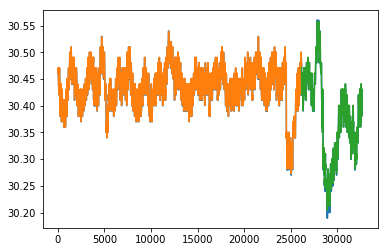
After that, we use prediction model and in that models, we will use those forecasted values to predict the values of the target variable (Solar radiation).

1. **RESULTS:**
2. ***Temperature Forecasting:***

Here, we forecast temperature with the help of the RNN model with LSTM architecture. The output, which comes as the forecasted value has the RMSE of 0.53 in the test set while for the training set it gives the RMSE of 0.63. And we save the forecasted values in a list for future use.

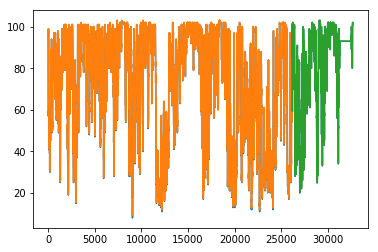
**Fig 2**: Forecasted and training value of temperature

1. ***Pressure Forecasting:***

Now, we have to forecast the values of another feature variable named pressure with the help of the same RNN model. In this case, the RMSE of the test dataset is 0.01 and the RMSE value of the train dataset is 0.00. This means this model almost perfectly forecast the pressure value for the future. Here also we store those values in a list.

**Fig 3:** Forecasted and training value of pressure

1. ***Humidity Forecasting:***

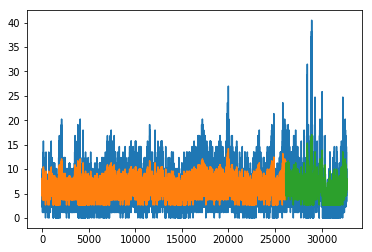
After this we have to check the forecasted value of humidity feature. This is also done by the same RNN architecture. In this case, the RMSE value of test dataset is 2.45 and the RMSE value of train dataset is 3.00. Here also we store the forecasted values in the form of a list.

**Fig 4:** Forecasted and training value of humidity

1. ***Wind direction Forecasting:***

After this we have to check the forecasted value of wind direction feature. This is also done by the second RNN architecture. In this case, the RMSE value of test dataset is 62.52 and the RMSE value of train dataset is 68.28. Here also we store the forecasted values in the form of a list.

1. ***Wind speed Forecasting:***

After this we have to check the forecasted value of wind direction feature. This is also done by the same RNN architecture. In this case, the RMSE value of test dataset is 3.49 and the RMSE value of train dataset is 2.72. Here also we store the forecasted values in the form of a list.

**Fig 5:** Forecasted and training value of speed

Now we have to build the data frame with the help of the stored value and then fit the train dataset into the prediction models we choose.

1. ***Prediction:***

After fitting the data we predict the values with respect to that stored values and find the predicted values from the different model. It will help us to evaluate our models with RMSE error of the model and their accuracy score.

For gradient boosting model we use n\_estimators as 10000, maximum depth and minimum sample split as 4 and the learning rate is 0.01. by this model we get the accuracy of and RMSE as 161.1733

For second model random forest we use n\_estimators as 1000 with the random state 12. This model gives the accuracy of and RMSE of 167.29

Third model multiple linear regression gives the RMSE of 172.5978.

The neural network model gives us the maximum accuracy of 83.225% and also gives the RMSE value of 148.24. This model has eight(8) dense layer and each layer has one less node than its previous layer. So according to all the basic model we have used ANN is the best model to predicting the solar radiation with the help of the RNN model.

1. **CONCLUSION:**

Here we have used the forecasting model to predict the future data from the current data of one hour before. After that, we have used the different prediction model to see the better model in the terms of error and accuracy score. Here we can see that after forecasting the data with RNN model, artificial neural network gives the better accuracy. As, this model has a RNN component, we can easily predict the feature variable for the future dates. And its prediction component can easily predict the solar radiation of those dates. This neural model will give better accuracy but due to the inconsistency of the data its accuracy shows a little less. But overall this model can give correct prediction regarding solar radiation forecasting for near future.

# **REFRENCES**

1. *Ghayekhloo, M., Ghofrani, M., Menhaj, M. B., & Azimi, R. (2015). A novel clustering approach for short-term solar radiation forecasting. Solar Energy, 122, 1371-1383.*
2. *Huang, Jing, et al. "Forecasting solar radiation on an hourly time scale using a Coupled AutoRegressive and Dynamical System (CARDS) model." Solar Energy 87 (2013): 136-149.*
3. *Gala, Y., Fernández, Á., Díaz, J., & Dorronsoro, J. R. (2016). Hybrid machine learning forecasting of solar radiation values. Neurocomputing, 176, 48-59.*
4. *Lauret, Philippe, et al. "A benchmarking of machine learning techniques for solar radiation forecasting in an insular context." Solar Energy 112 (2015): 446-457.*
5. *Voyant C, Notton G, Kalogirou S, Nivet ML, Paoli C, Motte F, Fouilloy A. Machine learning methods for solar radiation forecasting: A review. Renewable Energy. 2017 May 1;105:569-82.*
6. *Lou, Siwei, et al. "Prediction of diffuse solar irradiance using machine learning and multivariable regression." Applied energy 181 (2016): 367-374.*
7. *Benmouiza, Khalil, and Ali Cheknane. "Small-scale solar radiation forecasting using ARMA and nonlinear autoregressive neural network models." Theoretical and Applied Climatology 124.3-4 (2016): 945-958.*
8. *Fan, Junliang, et al. "Comparison of Support Vector Machine and Extreme Gradient Boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: A case study in China." Energy Conversion and Management 164 (2018): 102-111.*
9. *Ibrahim, Ibrahim Anwar, and Tamer Khatib. "A novel hybrid model for hourly global solar radiation prediction using random forests technique and firefly algorithm." Energy Conversion and Management 138 (2017): 413-425.*
10. *Kazem, Hussein A., Jabar H. Yousif, and Miqdam T. Chaichan. "Modeling of daily solar energy system prediction using support vector machine for Oman." International Journal of Applied Engineering Research 11.20 (2016): 10166-10172.*