SMU Data Science Review

Volume 6
Number 2 *Summer 2022*

Article 12

2022

Short Term Forecasting of Solar Radiation

Ashwin Thota
Southern Methodist University, athota@smu.edu

Bradley Blanchard
Southern Methodist University, bblanchard@smu.edu

Lijju Mathew
Southern Methodist University, lmathew@smu.edu

Paritosh Rai
Southern Methodist University, prai@smu.edu

Sid Swarupananda
Southern Methodist University, sswarupananda@smu.edu

Short Term Forecasting of Solar Radiation

Paritosh Rai[1], Lijju Mathew[1], Sid Swarupananda[1], Bradley Blanchard[1], Ashwin Thota[1]

[1]
Master of Science in Data Science
Southern Methodist University
Dallas, Texas USA
prai@smu.edu, lmathew@smu.edu, sswarupananda@smu.edu, bblanchard@smu.edu, athota@smu.edu

Abstract. This paper details how to predict solar radiation at a location for the next few hours using machine learning techniques like Facebook's Prophet, and Amazon's DeepAR+. Multiple techniques like AutoRegressive (ARIMA) and Exponential Smoothing (ES) have been used to forecast solar radiation, but they lack accuracy and are not scalable. Whereas Prophet, and Amazon's DeepAR+ are scalable, accurate, and easily integrated into other machine learning techniques. This will be the first time where the combination of these techniques along with Linear Regression, Random Forest, XGBoost and Decision Tree will be leveraged to forecast solar radiation for the short term.

Predicting solar energy accurately depends on multiple factors (including weather conditions) that make forecasting highly resource-intensive, and accuracy remains a challenge. Improving the accuracy of the short-term forecast of solar energy production would provide a massive value to the companies operating IoT Devices and drones to have a more efficient operation and reduced cost. The objective is to improve the accuracy of forecasting short-term solar radiation to power drones and IoT devices, leveraging the ensemble techniques by combing the outcome of Prophet and DeepAR+.

Facebook's Prophet, and Amazon's DeepAR+ used to carry out shortterm solar forecasting can be scaled by leveraging the supercomputer. Amazon's DeepAR+ runs on the AWS cloud platform, so they align well with scaling and bring in all the enhancement that comes with cloud technology.

Multiple models were used to identify the best way to forecast short-term solar radiation. Random Forest and ensemble models outperformed the Facebook Prophet and Amazon's DeepAR+, achieving a coefficient of determination R^2 of 99 % in Dallas, Texas. Ensemble Model was created to minimize the bias and variance of the outcome.

1 Introduction

Solar-powered energy generation is on the rise in the United States. According to EIA's Preliminary Monthly Electric Generator Inventory survey reports, Texas will add 10 gigawatts (GW) of utility-scale solar capacity by the end of 2022 [1] [2]. An increase in solar power generation indicates the need for more enhanced tools to manage both short-term and long-term solar power management. In this fast-changing technology and explosion in IoT, the need to power these miniature devices for a long time (10 to 15 years) remains a challenge. Solar energy is considered one of the prime sources to feed power to this evolving technology. Solar technology can play a vital role in bringing the carbon footprint down. There are multiple algorithms [3] in place to predict the long-term changing solar radiation. These IoT devices and drones need something short-term to ensure they will operate in full capacity for the coming few years. It became critical to predicting solar radiation patterns and another meteorological factor that will allow us to estimate the PV (Photovoltaic) for a short duration (say 5 to 7 hours). This research will deliver the Machine Learning techniques to take the solar radiation, and other impacting factors to predict the short-term PV power generated leveraging supervised, unsupervised, and semi-supervised deep learning techniques.

Today drones are being increasingly deployed throughout the world to streamline logistic and monitoring routines. When dispatched on autonomous missions, drones require an intelligent decision-making system for trajectory planning and tour optimization.

Let us consider a drone use case. Today we fully charge the drones 100 percent and it's predetermined flight time may be 8 hours. At the end of 8 hours, we have to return to the charging station. Imagine you are helping first responders with this drone, you have to call off the rescue mission after 8 hours of flight time. With our analytics, we can move away from these pre-charged drones and predetermined distance concept and move towards dynamic charging and a scalable solution. Our analytics will take into consideration changing weather conditions and dynamically calculate the charging capacity while it is in-flight. With this solution, we need not call off the rescue mission after 8 hours, rather we can provide the in-flight solar charge capacity and continue the rescue mission. Sunbirds design and sells solar drones that can travel for 10+ hours under ideal conditions. Their solar drones fly autonomously and can also be controlled by a remote control. Sunbirds solar drones are equipped with cameras used for mapping and aerial photographing.

Powering IoT Devices with renewable sources removed many barriers. However, the unpredictable nature of weather and solar radiation keeps the uncertainty of operational hours. Buoys are a waterborne IoT device that sends a constant data stream to monitor water conditions. These are located no dedicated power source t change the batteries to support the measurement of water condition and communicating to the server for further analysis. Powering the device by solar

energy and forecasting the solar radiation help estimate the availability of power to extend the operational time of the devices.

Solar energy generation uses photovoltaic (PV) panels that produce electricity through interaction with photons from the sun. Solar energy availability at a given location can be described by the level of solar radiation (W/m2). Solar radiation (IGH), using the pyranometer, measures the energy available to enter a solar PV system. The system configuration and power losses through the PV system are well known, and thus if the solar radiation is known, solar energy can be accurately predicted.

This paper presents a machine learning model for predicting solar radiation at a location for the next few hours. Solar energy is an efficient renewable energy source and is becoming increasingly popular as more energy companies are growing to offer this renewable source of electricity. The use of solar energy to power IoT will reduce operational costs. Due to the reduced use of fossil fuels, it becomes environment friendly, thus increasing its popularity and demand for its services. It becomes more and more critical to forecast the solar radiation to maintain the operation.

DeepAR+ is used to forecast time series using recurrent neural networks (RNN), handle a huge dataset, and support more features and scenarios. Prophet shines when applied to time-series data that have strong seasonal effects and several seasons of historical data to work.

The paper presents the application of combining Facebook's Prophet, and Amazon's DeepAR+ forecasting models for solar radiation forecasting. These two methods bring their respective strengths. Prophet gives better results for items with a long history and frequent outcomes, whereas Amazon's algorithms show superiority for items without a long history and rare occurrences. The two algorithms will process NREL data, and a new ensemble [4] outcome will be leveraged to forecast the short-term (5 to 7 hours) solar radiations.

The paper discusses the dataset and critical features. It talks about extracting information from a dataset to prepare for processing the data with Facebook's Prophet, and Amazon's DeepAR+. It explains the method used to create individual outcomes from each method, followed by creating an ensemble [5] model to deliver enhanced accuracy of short-term solar radiation.

The objective is to improve the accuracy of forecasting short-term solar radiation to power drones and IoT devices, leveraging the ensemble [4] techniques by combing the outcomes of Prophet and DeepAR+.

Main contributions of the paper are:

- Creating an ensemble model [4] [5] using Facebook's Prophet [6], and Amazon's DeepAR+ [7] [8].
- Comparing the results of multiple models against set performance indicators to determine best model to forecast short-term solar radiation.
- Leveraging other meteorological variables to predict the solar radiation.

Create an ensemble model to further enhance the prediction results. Ensemble methods aim to improve the performance of the model by combining multiple models instead of using a single model. The combined models increase the accuracy of the results significantly. Ensemble methods aim to improve model predictability by combining several models to make one very reliable model. The most famous ensemble methods are boosting, bagging, and stacking. Ensemble methods are ideal for regression and classification, where they reduce bias and variance to boost the accuracy of models

Scaling: As the data volumes are large and the IoT devices and drones will be extended across the globe, users may have run multiple forecasts at in parallel to ensure all the devices have capability and power to support the need of business and mission. Scaling of this methodology is very critical. Amazon's DeepAR+ operates on AWS platform which comes with scaling capability and all other advantage that comes with cloud platform.

The rest of this paper is organized as follows: Section 2 discusses the related work carried to predict solar radiation across the globe by multiple organizations. Section 3 describes the data set used in the study. Section 4 gives a brief background of the multiple machine learning techniques used to forecast solar radiation. In Section 5, we present the results to demonstrate the performance of our proposed methodology. Section 6 discusses future scope and conclusion.

2 Literature Review

There can be multiple ways solar radiation can be measured and predicted. Below are some of the articles explored for solar radiation prediction.

2.1 Short-Term Solar Power Forecasting Considering Cloud Coverage and Ambient Temperature Variation Effects

Foad H. Gandoman, Shady H.E. Abdel Aleem, Noshin Omar, Abdollah Ahmadi, Faisal Q. Alenezi. Short-term solar power forecasting considering cloud coverage and ambient temperature variation effects. Retrieved from link

This article [9] proposes a new methodology to assess the impacts of factors like cloud and ambient temperature change on the hourly output power of a PV system installed in Iran. The physical methods use diverse resources such as weathe, PV system data, and satellite and sky imagery clouds to predict the PV forecast. This paper proposes a new physical model to estimate short-term PV power based on Oktas-scale variations and temperature changes. The paper agrees that other meteorological conditions [10] impact radiation rate. However, the study focused on cloud coverage and air temperature. The study used Julian Day Number (JDN) to account for the sun's geometric position (solar elevation angle) for leap and regular years. Earth rotation on its axis was also calculated by determining hourly

solar radiation angle. Model accuracy was evaluated using standardized root-mean-square error (RMSE).

As the solar power generation strongly depends on solar radiation, which is directly influenced by cloud cover, some researchers like this case study have tried to solve the solar power forecast problem from a more physical perspective with cloud imagery. This case study applied a physical model based on short-term cloud variations as well as temperature changes to estimate hourly-averaged PV output power in small solar power plants with different climatic conditions. This was the only paper found to give some understanding of short-term forecasts. Root Mean Square (RMS) was used to determine the average solar PV power accuracy. RMS will be leveraged to measure the accuracy in this study as well. However, a more granular approach of 30 minutes will be used to predict better. The author only used temperature and cloud. However, this study will use multiple other meteorological factors [11]to get a better insight. Indeed, temperature and cloud appear to be important factors in PV generation by solar panels. However, other factors can also play a role in adjusting the GHI and PV. Higher ambient temperature reduces the conversation efficiency of solar panels from GHI to PV, and clouds contribute to reducing the GHI reaching the solar panel. The difference between the proposed models in this paper and the actual varied between 3 percent to 14 percent depending on the season, temperature, and cloud conditions.

We are using cloud and temperature as one of the key meteorological variable for estimating solar radiation.

2.2 Short-Term Solar Power Forecasting Using Linear and Non-Linear Regularization Models

S.K. Aggarwal, L.M. Saini. Solar energy prediction using linear and non-linear regularization models: A study on AMS (American Meteorological Society) 201314 Solar Energy Prediction Contest. Retrieved from link

The paper's [12] author used linear and non-linear regularization models vs. team will leverage other neural network models. Also, to measure the improvement author has used MAE (Mean Absolute Error), and in this study, RMSE will be used to measure the accuracy.

This paper shows that in global horizontal radiation forecasting the models can be used in three different ways:

- structural models which are based on other meteorological and geographical parameters
- time-series models which only consider the historically observed data of solar radiation as input features (endogenous forecasting)
- hybrid models which consider both solar radiation and other variables as exogenous variables (exogenous forecasting)

Here an ensemble of ANN (artificial neural network) and LSR (least square regression) models is presented to forecast the solar energy production and it shows that an adequate choice of the input data allows to obtain more accurate predictions. One of the main limits of ANN methods is an excessive training data approximation, aimed to increase the out-of-sample forecasting errors.

2.3 Multi-Dimensional Linear Prediction Filter Approach for Hourly Solar Radiation Forecasting

Emre Akarslan, Fatih Onur Hocaoglu, Rifat Edizkan. A novel M-D (multi-dimensional) linear prediction filter approach for hourly solar radiation forecasting. Retrieved from link

The paper [13] discusses a new approach for hourly solar radiation forecasts. The data measured hourly throughout the year are converted into 2-D image forms, and the data points are evaluated as pixels of the images. Multi-dimensional linear prediction models are designed to link the solar radiation images with different images that correlate with solar radiation data. The performance of each model is compared with each other, and the results show that the proposed approach significantly improves the prediction accuracy.

The model used by the authors gave prediction accuracy of 1 percent to 30 percent for different multi dimensional Models. RMS was used to evaluate the accuracy of the forecasting methodology.

The basic principle in this study to predict the global solar radiation for impending hours was to correlate the solar radiation of future hours with the values of past hours in applying different approaches.

The forecasting approach discussed is based on gray satellite images, where as this study will leverage actual raw data collected like GHI and other meteorological factors.

2.4 Mycielsi-Markov Model for Hourly Solar Radiation Forecasting

Fatih Onur Hocaoglu, Faith Serttas. A novel hybrid (Mycielsi-Markov) model for hourly solar radiation forecasting. Retrieved from link

The paper [14] discusses using the Mycielski-Markov model to predict solar radiation for two cities in Turkey. The approach assumes solar radiation data repeats itself. A novel ANN methodology was used to estimate the profile of data using solar radiation, air temperature, voltage, and current data for training and validation of the model. Mycielski algorithm was built to find exact matching data. Hourly solar radiation values are converted into the states. The range of the states, statistical distribution, and standard deviation of the data were taken into account to estimate future values. Based on the study's outcome, it is possible to predict new solar radiation data samples accurately by using only historical solar radiation data without any other parameters.

Mycielsi-Markov Model is expensive, both in terms of memory and compute time. Also, it needs to be trained on a set of seed sequences. They seem to work well on long term prediction, if the prediction time interval is short, then Markow models are inappropriate because the individual displacements are not random, but rather are deterministically related in time.

2.5 Solar Radiation Forecasting Using Artificial Neural Network and Random Forest Methods

Benali, G. Notton, A. Fouilloy, C. Voyant, R. Dizene. Solar radiation forecasting using artificial neural network and random forest methods: Application to normal beam, horizontal diffuse and global components. Retrieved from link The paper [15] discusses forecasting normal beam, horizontal diffuse, and global solar components at Odeillo, France, hourly using Smart Persistence (SP), Artificial Neural Networks (ANN), and Random Forest methods. Random Forest Method was found to predict more accurately looking at nRMSE. It also shows that forecasting during winter and summer is difficult due to higher solar radiation variability. The author also discussed the data cleaning methodology by removing night data and data during sunrise and sunset. They also discussed the challenge of evaluating the hourly beam and diffusion component of GHI (Global Horizontal Irradiation).

The author's discussion was great on missing data. The team planned to leverage some of the techniques discussed in the paper; however, there was no missing data found in the given database, so the team could not leverage his technique. One of the big take-ups was to predict GHI vs. PV as GHI to PV conversation depends on other factors like the material used for making a solar panel. This model motivated the team to look at Random Forest as one of the baseline models as this was giving excellent results to the authors of the paper. However, these models were used for the long-term forecast.

2.6 Predicting Solar Radiation in Tropical Environment Using Satellite Images

Ayu Wazira Azhari, Kamaruzzaman Sopian, Azami Zaharim, Mohamad Al Ghoul. School of Environmental Engineering, Universiti Malaysia. A New Approach For Predicting Solar Radiation In Tropical Environment Using Satellite Images. Retrieved from link

The paper [16] discusses how greyscale satellite images estimate solar radiation from several ground measuring stations of solar radiation. The research examines the importance of clouds and the type of clouds (thickness of clouds: very low, low, average, high cloud cover, and very high cloud cover). The study was based on low-cost solar radiation estimates and several ground measurements. Cloud images from satellites and calculated and compared with actual data. The results were extrapolated to estimate solar radiation at other places.

As stated above, the paper was based on satellite images in tropical areas. However, this paper gave the team an excellent exposure to various terminology and an essential factor impacting solar radiation (GHI). One of the critical findings encouraged the team to look at other factors beyond cloud and temperature that can impact GHI. The author did not compare the past or baseline with the outcome of his analysis with past findings. However, he shared his finding using map images from different seasons of the year. The methods used in the paper were not very relevant as it was image-based on the long-term forecast.

2.7 Advanced Ensemble Model for Solar Radiation Forecasting Using Sine Cosine Algorithm and Newton's Laws

El-Kenawy, El-Sayed M ; Mirjalili, Seyedali ; Ghoneim, Sherif S. M ; Eid, Marwa Metwally ; El-Said, M ; Khan, Zeeshan Shafi ; Ibrahim, Abdelhameed Advanced Ensemble Model for Solar Radiation Forecasting Using Sine Cosine Algorithm and Newton's Laws. Retrieved from link

This paper [17] proposes optimized solar radiation forecasting ensemble model consisting of pre-processing and training ensemble phases. The training ensemble phase works on an advanced sine cosine algorithm (ASCA) using Newton's laws of gravity and motion for objects (agents). Obtained results of the proposed ensemble model are compared with those of state-of-the-art models, and significant superiority of the proposed ensemble model is confirmed using statistical analysis such as ANOVA and Wilcoxon's rank-sum tests. The proposed ensemble model shows superiority over the reference base model including LSTM, NN, and SVM. The ASCA based ensemble weights model provided better results over the average ensemble and the KNN ensemble models. Several experiments are conducted and different performance metrics are considered to conclude that the proposed ensemble weights model is the most suitable for forecasting solar radiation.

The author talked about the unique approaches leveraging sine cosine algorithm vs. standard ML-based techniques. The paper got the attention due to its unique approach; however, the team realized these are very complex, not very scalable, and very resource intensive. Therefore this method was not used in the analysis process.

2.8 DeepAR: Probabilistic forecasting with autoregressive recurrent networks

David Salinas, Valentin Flunkert, Jan Gasthaus, Tim Januschowski. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. Retrieved from link

This paper [18] is about going to forecast time-series based on the past trends of multiple factors with the help of the DeepAR algorithm, a methodology for

producing accurate probabilistic forecasts, based on training an autoregressive recurrent neural network model on a large number of related time series. The Deep Autoregressive model (DeepAR) has built-in algorithms for Amazon Sagemaker. It is a time-series forecasting using a Recurrent Neural Network (RNN) capable of producing point and probabilistic forecasts. This paper shows that forecasting approaches based on modern deep learning techniques can drastically improve forecast accuracy relative to state-of-the-art forecasting methods on a wide variety of datasets. The DeepAR focuses on experimenting with our time series to get the best possible results without worrying about the internal infrastructure. Forecasting tasks can be done quickly as there is no need to write any training code. The data can be prepared, and the necessary tunings to be performed to fine-tune the model.

The model mentioned in this paper, DeepAR, is one of the models used in our study. The work presented in the paper is applied for solving retail business use cases and the findings can be applied to our solar energy forecasting use case. This paper also talks about the advantages of using DeepAR over classical approaches.

2.9 Towards flexible groundwater-level prediction for adaptive water management: using Facebook's Prophet forecasting approach

H. Aguilera a, C. Guardiola-Albert a, N. Naranjo-Ferna´ndez and C. Kohfahl b. Department of Research in Geological Resources, Spanish Geological Survey, Madrid, Spain. Towards flexible groundwater-level prediction for adaptive water management: using Facebook's Prophet forecasting approach. Retrieved from link

The paper [19] discusses forecasting using the Facebook Prophet forecasting tool. Facebook has released the code and made it open source. Prophet is an additive model that considers nonperiodic changes and parodic components with easily interpretable parameters. The paper discusses how hydrologists and water managers used this tool to predict Ground Water Levels (GWLs) with flexibility and efficiently. Papers talk about challenges like flexibility, eases of use, and interpretability encountered with other tools were used. Prophet is robust to missing data, shifts in the trends, and significant outliers.

The model mentioned in this paper, Prophet is another model used in our study. The paper talks about approaches how prophet outperforms most methods in predicting the Ground Water Level. Prophet with optimized parameters shows slightly better results in the training period than the automatic model with default parameters, but there are no significant differences between both approaches. Training and test performance measures suggest that Prophet is robust against overfitting as the optimized models still have generalization ability

2.10 Comparison analysis of Facebook's Prophet, Amazon's DeepAR+ and CNN-QR algorithms for successful real-world sales forecasting

Emir Zunic, Kemal Korjenic, Sead Delalic, and Zlatko Subara. Comparison analysis of Facebook's Prophet, Amazon's DeepAR+ and CNN-QR algorithms for successful real-world sales forecasting. Retrieved from link

The paper [20] presents the application and comparison of the Facebook's Prophet, Amazon's DeepAR+, and CNN-QR forecasting models for sales forecasting in distribution companies. The results show that Prophet gives better results for items with a long history and frequent sales. In contrast, Amazon's algorithms show superiority for items without a long history and items that are rarely sold. The Prophet Forecasting Model is an open-source procedure for forecasting timeseries data. Facebook's Core Data Science team created the Prophet tool. The Prophet algorithm works well in the case of datasets with multiple seasons and qualitatively describes the seasonality in the data. The DeepAR+ model is based on autoregressive Recurrent Neural Networks. It is used as a supervised learning algorithm to forecast one-dimensional time series. DeepAR+ has been trained on a large number of time series. Amazon's CNN-QR (Convolutional Neural Network - Quantile Regression) forecasting algorithm is based on the application of casual convolutional neural networks to predict scalar time series data. Prophet analyzes each signal independently, while Amazon's algorithms create
a single model for all signals and find interdependencies. Therefore, Amazon's algorithms show superiority over classical methods only when they have a large number of signals over which to create a model, and in the case of articles with a short history. It has been observed that the Prophet model shows superiority in the case of items that are sold frequently, in large quantities, and have a long sales history. At the same time, Amazon's algorithms showed dominance in other cases. The paper provides a performance comparison of Facebook's Prophet algorithm, and Amazon's DeepAR+ and CNN-QR algorithms and this can be used in our study. The obtained results confirm the advantages of each algorithm. The question of the benefits of using each of the proposed approaches is raised, given the essential difference in the way Prophet works and Amazon's algorithms. As noted, Prophet analyzes each signal independently, while AWS algorithms create a single model for all signals and try to find interdependencies. Therefore, AWS algorithms show superiority over classical methods only when they have a large number of signals over which to create a model, and in the case of articles with a short history.

2.11 Thesis

Based on the data science techniques, solar radiation and meteorological data forecast the short-term solar radiation leveraging multiple machine learning techniques (linear and nonlinear) for a given location most efficiently and

effectively. The best model or ensemble of the models can be trained to enhance the forecast accuracy to generate a generic model for global or regional use.

3 Data

Working with solar radiation and meteorological data over the United States and regions of the surrounding countries acquired from the National Solar Radiation Database maintained by National Renewable Energy Laboratory. The research will use climate, weather, and other meteorological data to leverage linear and nonlinear machine learning algorithms and techniques to build a framework for assessing the performance of short-term solar radiation forecasting.

3.1 Data Sources

The data comes from the National Solar Radiation Data Base (NSRDB) [21], consisting of solar radiation and meteorological data over the United States and regions of the surrounding countries. It is a publicly open dataset that has been created and disseminated during the last 23 years. The current NSRDB provides solar radiation at a 4-km horizontal resolution for each 30-min interval from 1998 to 2016.

The data is computed by the National Renewable Energy Laboratory's (NREL's) Physical Solar Model (PSM) and products from the National Oceanic and Atmospheric Administration's (NOAA's) Geostationary Operational Environmental Satellite (GOES), the National Ice Center's (NIC's) Interactive Multisensor Snow and Ice Mapping System (IMS), and the National Aeronautics and Space Administration's (NASA's) Moderate Resolution Imaging Spectroradiometer (MODIS) and Modern-Era Retrospective analysis for Research and Applications, version 2 (MERRA-2). The NSRDB radiation data have been validated and shown to agree with surface observations with mean percentage biases within 5 percent and 10 percent for global horizontal radiation (GHI) and direct normal radiation (DNI), respectively.

The data can be freely accessed via https://nsrdb.nrel.gov or through an application programming interface (API). During the last 23 years, the NSRDB has been widely used by an ever-growing group of researchers and industry both directly and through tools such as NREL's System Advisor Model. The data is provided in high density data files (.h5) [22] [22] separated by year. The variables mentioned below are provided in 2 dimensional time-series arrays with dimensions (time x location). The temporal axis is defined by the time-index dataset, while the positional axis is defined by the meta dataset. For storage efficiency each variable has been scaled and stored as an integer. The scale-factor is provided in the psm-scale-factor attribute. The units for the variable data is also provided as an attribute psm-units.

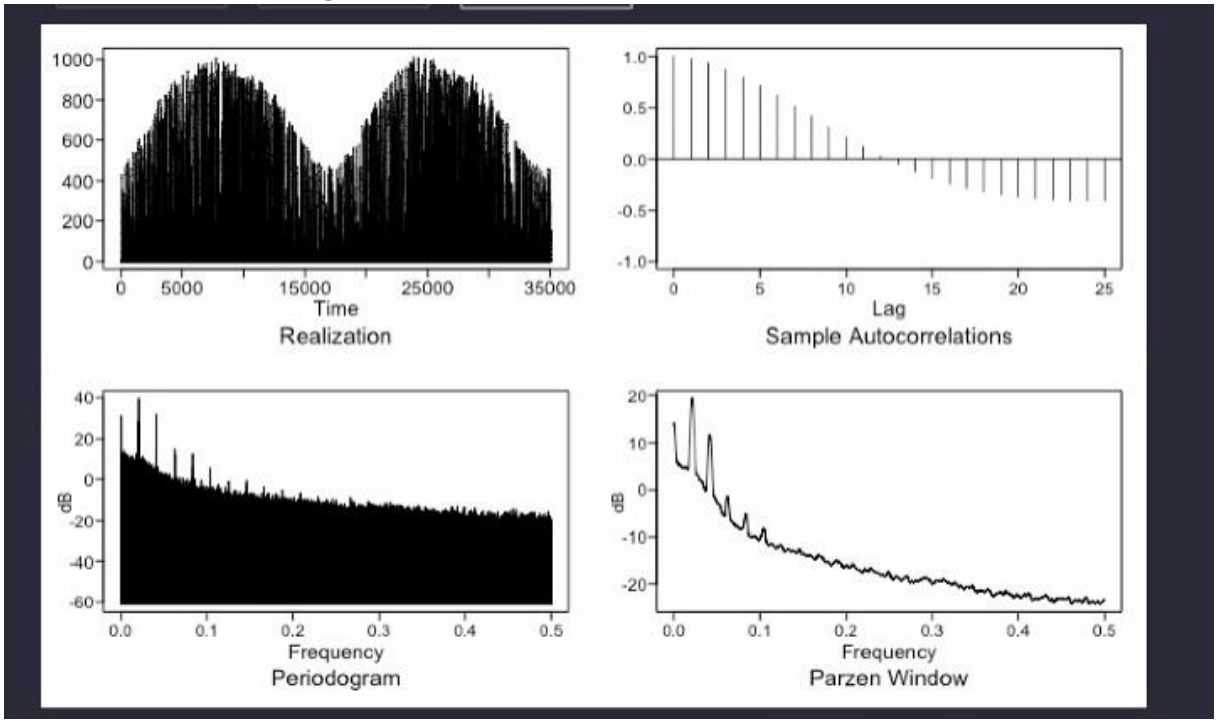
Below are some of the key attributes from the data that were explored based on the correlation of the variables as seen in the heatmap.

- **Wind Speed** - Horizontal motion of air near the surface of the earth. Measured in meters per second. Source: MERRA
- **Precipitable Water** - The amount of water in a vertical column of atmosphere. Measured in millimeters. Source: MERRA. The unit of measure is typically the depth to which the water would fill the vertical column if it were condensed to a liquid. For example, 6 centimeters of precipitable water (in the absence of clouds) indicates a very moist atmosphere. Precipitable water is often used as a synonym for water vapor.
- **Relative Humidity** - The amount of water vapor in the air expressed as the ratio between the measured amount and the maximum possible amount (the saturation point at which water condenses as dew). Measured in percentage. Calculated from specific humidity
- **Direct Normal radiation** - The amount of solar radiation from the direction of the sun. Measured in Watts per square meter. Modeled solar radiation obtained from the direction of the sun
- **Diffuse Horizontal radiation** - The radiation component that strikes a point from the sky, excluding circumsolar radiation. In the absence of atmosphere, there should be almost no diffuse sky radiation. High values are produced by an unclear atmosphere or reflections from clouds. Measured in Watts per square meter. Modeled solar radiation on a horizontal surface received from the sky excluding the solar disk
- **Global Horizontal radiation** - Measured in Watts per square meter. Modeled solar radiation on a horizontal surface received from the sky. Also called Global Horizontal radiation; total solar radiation; the sum of Direct Normal radiation (DNI), Diffuse Horizontal radiation (DHI), and ground-reflected radiation; however, because ground-reflected radiation is usually insignificant compared to direct and diffuse, for all practical purposes global radiation is said to be the sum of direct and diffuse radiation only:

$$GHI = DHI + DNI * cos(Z) \quad (1)$$

  where Z is the solar zenith angle.
- **Clearsky Diffuse Horizontal radiation** - Measured in Watts per square meter. Modeled solar radiation on a horizontal surface received from the sky excluding the solar disk. This is assuming clear sky condition
- **Clearsky Direct Normal radiation** - Measured in Watts per square meter. Modeled solar radiation obtained from the direction of the sun. This is assuming clear sky condition
- **Clearsky Global Horizontal radiation** - Measured in Watts per square meter. Modeled solar radiation on a horizontal surface received from the sky. This is assuming clear sky condition

Fig.1. Solar Radiation Trends



In the graph above we observe the following:

- Realization shows seasonality
- ACF is a diminishing sinusoidal
- Spectral density shows few frequency peaks

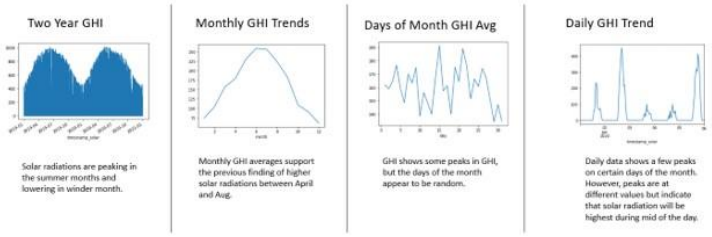
3.2 Exploratory Data Analysis

Exploratory Data Analysis (EDA) was carried out to understand the data and its distribution better. EDA helps detect obvious errors, identify outliers in datasets, understand relationships, unearth important factors, find patterns within data, and provide new insights.

Data Preparation: The data set used for the analysis is in a high-density data file, also known as "dot h5" (.h5) files. The team worked through the extraction process by converting the data to CSV. Data collected was in intervals of 30 minutes. Data for two years, 2019 and 2020, was secured, and both the CSV files were combined in one single data frame to carry out further analysis.

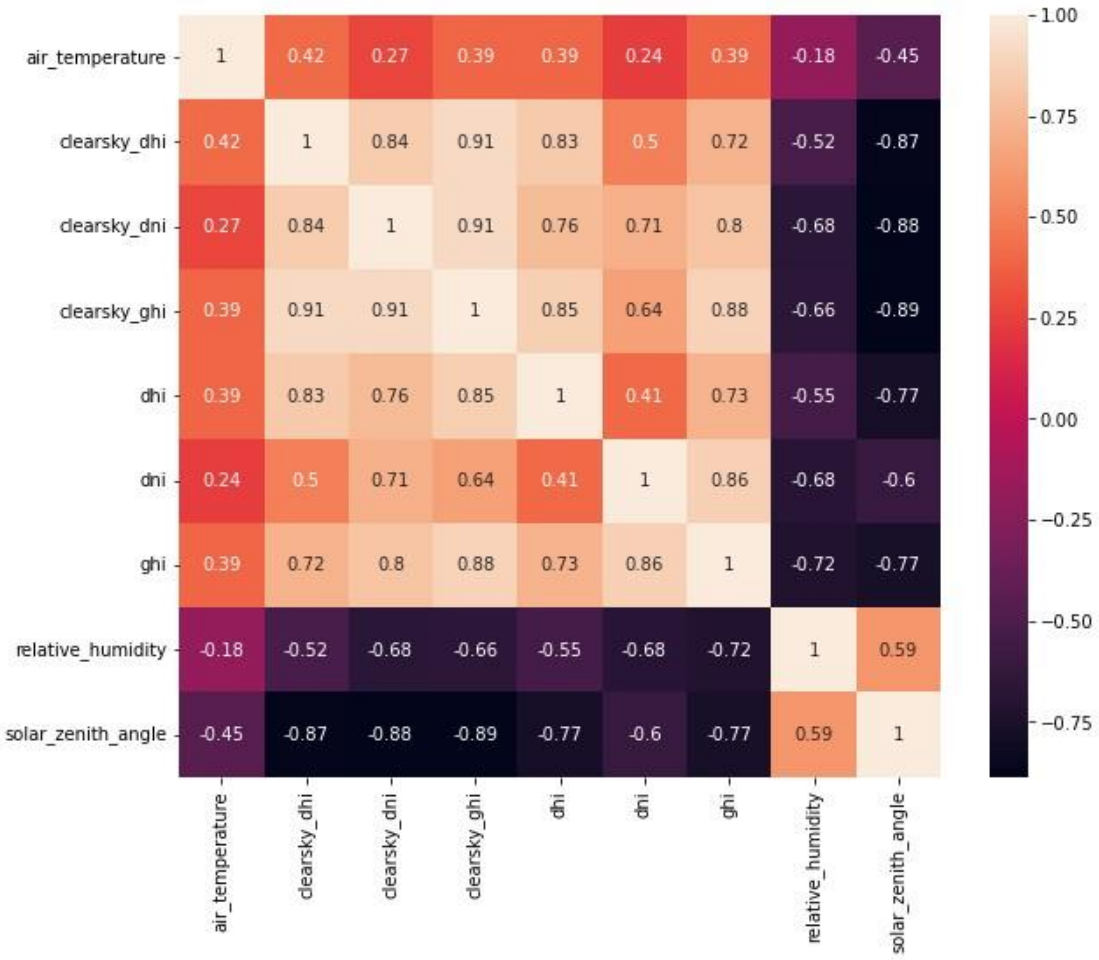
Data Size: Data-frame used for analysis contains 35088 rows and 26 features. The number of rows perfectly aligns with the expected sample size over two years. There are 365 days in 2019 and 366 days in 2020 (leap year), with 24 hours in a day, and each hour has two samples (sample collected every 30 minutes). 26 variables were related to meteorological data like temperature, wind direction and speed, dew point and few more. It has data related to solar data such as GHI, DHI, DNI, solar zenith angle, and others explained above in the paper. GHI is the measure of solar radiation, so it will remain the prime focus of this paper. GHI and solar radiation will be used interchangeably throughout the paper.

Fig.2. GHI plot



This is a time series data plotted for hourly, daily, monthly trends to identify trends.

Fig.3. A Heatmap showing the correlation of the variables in the model



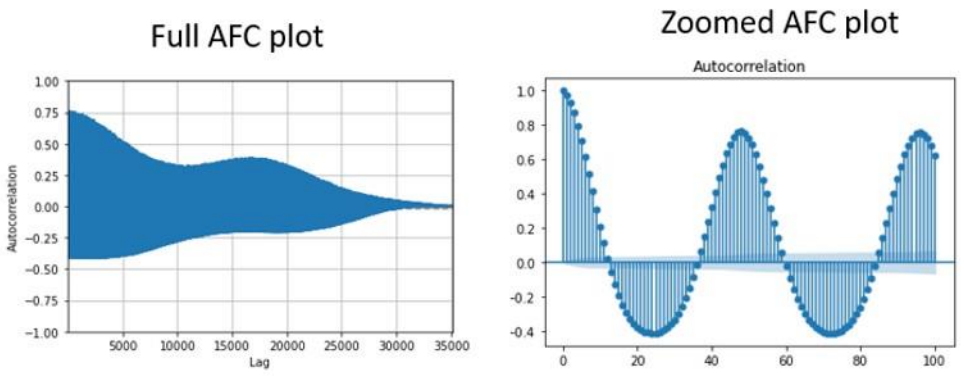
Correlation [23] 1 shows a high positive correlation, and -1 shows a high negative correction. Zero value shows no correlation.

Data was reviewed for correlation across all 26 samples. Some variables have a higher correlation (greater than 0.3) with Global Horizontal radiation. Following is the correlation chart of highly correlated variables with Global Horizontal radiation. Air temperature is positively correlated, and relative humidity is negatively correlated to Global Horizontal radiation. On solar data, Diffuse Horizontal radiation , Direct Normal radiation, Clearsky Direct Normal radiation and Clearsky Diffuse Horizontal radiation show a positive correlation, and the solar zenith angle shows a negative correlation.

Autocorrelation (ACF): The autocorrelation function [24] gives a good insight into your time-series. As time-Series data is ordered by time. There are periodical events that impact current data heavier than the rest of the data. AFC is a vital tool to give insight into these kinds of information. Following is the AFC plot of solar data.

ACF [24] graph is slowly dimensioning sinusoidal. Certain time values are positively correlated, and some show negative values. Negative values show that these values are below average. As expected, GHI will be high during the daytime and low during the night. AFC gives a good perspective on the randomness and stationarity of the time series data. AFC captures trends and seasonality. Each bar represents the size and direction of the correlation. For random data, ACF bars will be close to zero (i.e., white noise). The AFC graph above is a slowly diminishing sinusoidal chart. The diminishing AFC behavior indicates a trend (as correlation is solely tapering), and its sinusoidal nature demonstrates seasonality.

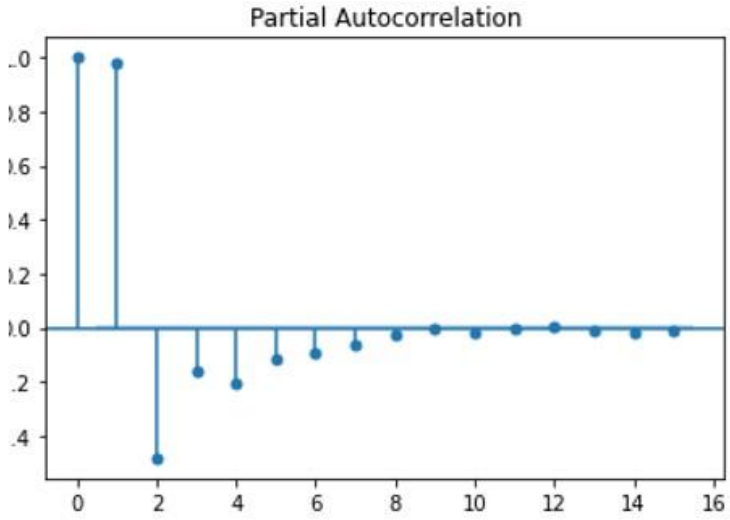
Fig.4. ACF plot

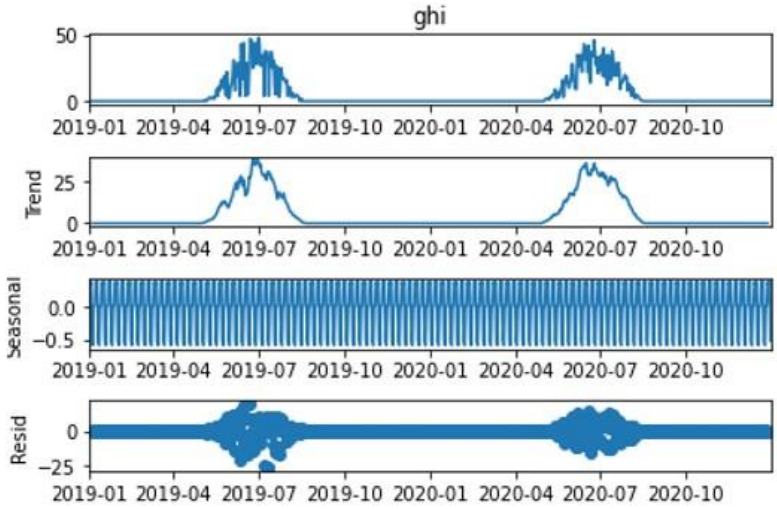


Partial Autocorrelation Function (PACF): the partial correlation [25] for each lag is the unique correlation between those two observations. The PACF

graph of solar data below lags 1, 2, and 3 are statically significant. The subsequent lags are nearly insignificant.

Fig.5. PACF plot





To further support the finding of ACF [24] , the PACF [25] data was decomposed [26], and subdivided into trend, seasonality and residue (noise) for

GHI trends. Findings from decomposing the data are aligned with AFC and PACF. Data shows trends, seasonality, and residue.

4 Methods

The objective is to compare various machine learning techniques to find which features are most likely to predict solar radiation accurately. Below are some of the methods being considered based on preliminary analysis of data. [27]

- Linear Regression
- Random Forest
- XG Boost
- Decision Tree
- Facebook Prophet
- Amazon's Deep Autoregressive model (DeepAR+)

Linear Regression model

Linear Regression [28] model used is a model with single independent variable x that has a relationship with a response variable y that is a straight line. The simple linear regression model is given by

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\epsilon}_i \quad (2)$$

where $\hat{\beta}_0 is the intercept$

$\hat{\beta}_1 is the slope\ \hat{\epsilon}_i is the random error$

The errors are assumed to have mean zero and unknown variance σ^2

Random Forest

Random forest [29] is a collection method that performs regression or classification by establishing multiple unrelated decision trees. Random forest mainly uses the idea of Bagging, and adopts random bootstrap method in sample selection. The weighting aspect adopts the method of uniform sampling, and all the prediction function weights are equal, which supports the parallel calculation of each prediction function. Through randomized forests, the random forest is not easy to overfit, extremely strong anti-noise ability and excessively fast calculation speed. Suppose the initial sample size is N. The sample feature dimension is M, and the number of decision trees in the artificially designated random forest is k. The specific modeling steps are as follows:

1. Constructing k decision trees from the original samples by means of bootstrap
2. Select m features in the M dimension as training for different decision trees,and m > M
3. The decision trees are not pruned and grow as much as
4. Random forest results are averaged from the results of each decision tree

XG Boost

XGBoost is a classification and regression algorithm based on Gradient Boosting Decision Tree (GBDT) [30]. XGBoost first builds a certain number of weak learners, most of which are classification regression trees, to train the vulnerable learners. It also performs weighted summation after training to obtain the final regression model. In the constructing process, a new learner is always added based on the residual error obtained from the last weak learner iteration. The new learner is built on the gradient to ensure the overall model error is reduced.

A model with more vital regression prediction capabilities is finally made.

Decision Tree

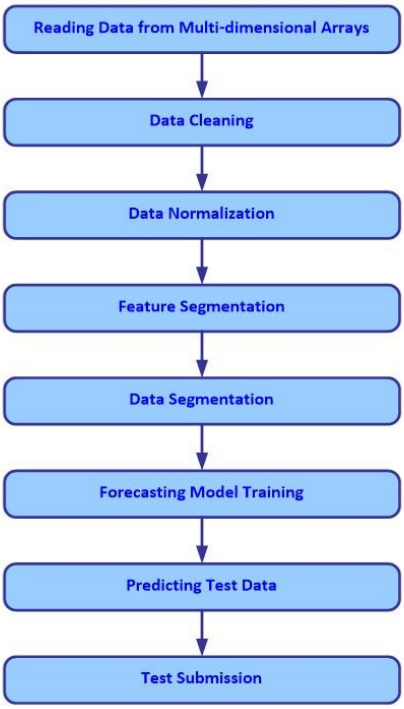
Decision tree is tree-like structure where at every node we make a decision and continue doing it till we reach a conclusion. A decision tree uses a training set of different predictors and target. The core algorithm for building decision trees is called ID3. The ID3 algorithm can be used to construct a decision tree for regression by replacing Information Gain with Standard Deviation Reduction.

Facebook Prophet

Facebook Prophet [6] is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data. Prophet is robust to missing data and shifts in the trend, and typically handles outliers well. Prophet is open source software released by Facebook's Core Data Science team and they have designed this package so that an analyst could do time series forecasting, in addition to data scientists.

Amazon's DeepAR+

The Amazon SageMaker DeepAR [7] [8] forecasting algorithm is a supervised learning algorithm for forecasting scalar time series using recurrent neural networks (RNN). During training, DeepAR+ uses a training dataset and an optional testing dataset. It uses the testing dataset to evaluate the trained model. In general, the training and testing datasets don't have to contain the same set of time series. You can use a model trained on a given training set to generate forecasts for the future of the time series in the training set, and for other time series **Fig.6. Methods of Analysis**



Reading Data from Multi-dimensional Arrays:

Data collected for solar radiation is very complex. Meteorological conditions play a vital role in Solar forecasting. Processing hourly data and multiple variables that may depend on each other or correlate with each other and time, is a complex task. It is critical to carry out detailed EDA to understand these variables and how (or if) they are related to each other.

Data Cleaning: [31]

Sometimes, sensor outages can lead to missing/incorrect data. To prevent the impact of missing data on predicating results, the gap should be filled appropriately. As solar radiation data exhibits yearly periodicity, forward and backward data filling will be applied to handle the missing values. Wherever the data corresponding to D – (365*24*4) was available, the missing data were replaced by backward filling. Wherever it was not available, it was filled by forward filling, i.e., data corresponds to D + (365*24*4) as solar data used for the short-term forecast is collected every 15 minutes.

Data Normalization:

As variables used in the data are in different units. Normalization [32] is required when dealing with attributes on a different scale; otherwise, it may dilute the effectiveness of a critical, equally important attribute (on a lower scale) because

of other attributes having values on a larger scale. When multiple attributes are there, but attributes have values on different scales, this may lead to poor data models while performing data mining operations. So, they are normalized to bring all the attributes on the same scale. It is vital to get the data on the same scale so their impact can be quantified and understood better. Data normalization consists of numeric remodeling columns to a standard scale. Data normalization is generally considered the development of clean data. Diving deeper, however, the meaning or goal of data normalization is twofold. Data normalization is the organization of data to appear similar across all records and fields. It increases the cohesion of entry types, leading to cleansing, lead generation, segmentation, and higher quality data.

Feature Segmentation:

As observed in the correlation table above, some features are highly correlated with the target variable (ghi). Data will be studied in two phases one will need all the features, and the other review data only with highly correlated variables. **Data Segmentation:**

Solar radiation and meteorological data exhibit a seasonal pattern. Data in different seasons may belong to a different distribution. Instead of using a single segment, the data may be broken into different segments to build a robust model with relevant data. The team will be leveraging the Prophet algorithm to capture seasonal trends. Prophet shines when applied to time-series data that have substantial seasonal effects and several seasons of historical data to work.

Forecasting Model Training:

Data will be split in a 70/30 ratio for training and test data set before normalizing. The team will ensure that no data leakage occurs between training and test data. The model will be built and train data.

The following techniques will be applied to forecast:

- Linear Regression
- Random Forest
- XG Boost
- Decision Tree
- Facebook's Prophet
- Amazon's DeepAR+

Initially, Solar radiation will be forecasted leveraging these methods independently. Efforts will be made to build an ensemble model with equal-weighted or adjusted weighted to improve the performance. A key objective will be to improve the performance of the solar model.

Predicting the Test Data:

[33] A model built on train data will be leveraged to predict the outcome. The predicted outcome will measure for accuracy.

Test Submission:

The outcome of test models will be measured independently and weighted and non-weighted ensemble model outcomes. Training and test times will be observed to keep an eye on processing time.
The key Performance metrics used for evaluating the models are coefficient of determination (R2), mean absolute error (MAE), and root-mean-squared-error (RMSE).

R^2 is the proportion of the variation in the dependent variable that is predictable from the independent variable. It is a statistic used in the context of statistical models whose primary purpose is either the prediction of future outcomes or the testing of hypotheses based on other related information. It measures how well-observed outcomes are replicated by the model based on the proportion of total variation of outcomes explained by the model. [34]

The R^2 value is obtained as follows: [35]

$$R^2 = \frac{\Sigma_i(y_i - \hat{y_i})}{\Sigma_i(y_i - \bar{y_i})} \tag{3}$$

where: y_i is the actual value $\hat{y}_i$ *meanoftheactualvalues* $\hat{y}_i$ *predictedvalues*

MAE has the same units as the predicted value and thus represents the expected absolute error, which is calculated by [36]

$$MAE = \frac{\sum_{i=1}^{D} |x_i - y_i|}{} \tag{4}$$

where: N is the total number of samples

The RMSE value squares the difference between actual and predicted values, emphasising larger errors. This is appropriate for solar prediction as larger errors lead to disproportionally higher costs. [37] RMSE can be calcuated as follows: [38]

$$RMSE = \sqrt{\Sigma_{i=1}^{N} \frac{\left(y_i - \hat{y_i}\right)^2}{N}} \tag{5}$$

5 Results

With this research, the expectation is to predict the short-term solar radiation. This research can influence determining the right combination of solar panel

material and rectifiers to ensure the desired application of IoT devices, drones, and other solar-powered appliances can be built at an optimal cost.

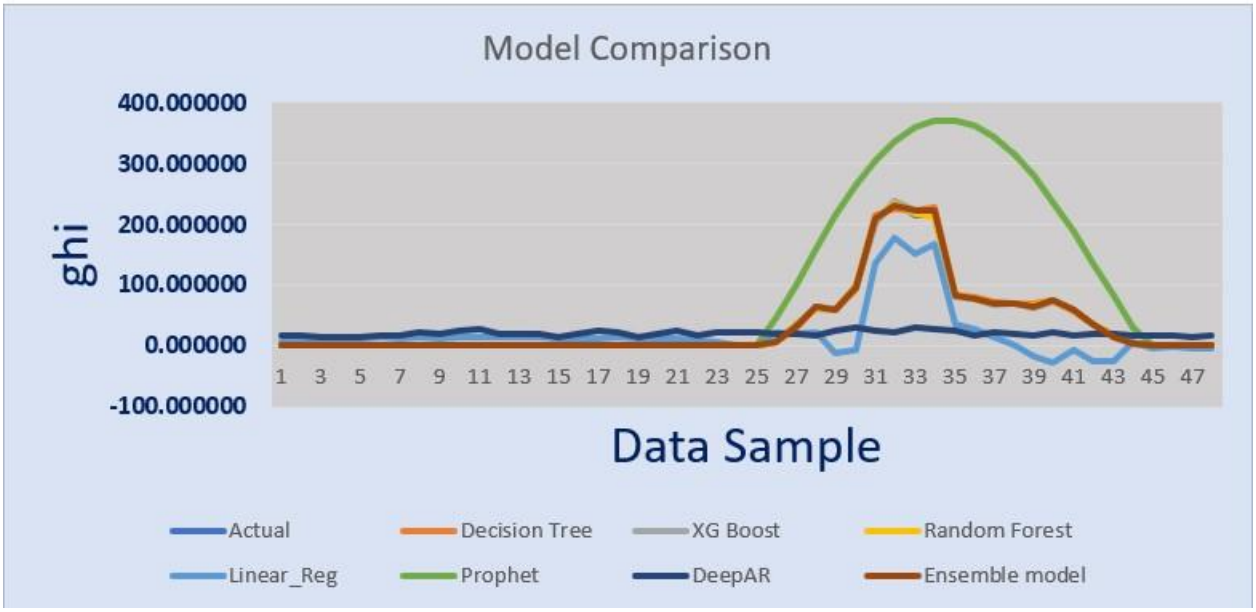
With this research and multi-layer data, the team is continuing to investigate to answer the following questions to enhance the accuracy of prediction:

- Impact and limitation of technology to convert GHI (solar radiation) to PV
- Extraction of multi-layered data from H5 files
- What will be the right combination of linear and nonlinear models to extract the outcome in the most effective and efficient way?
- Based on papers, the factor impacting solar radiation is cloud and temperature. The research team will continue to evaluate and quantify the improvement in predicting solar radiation by adding other meteorological factors.
- How the findings from long term forecast can be leveraged to enhance short term predictions
- What other multi-dimensional approaches to improve the prediction
- How Ensemble Model or hybrid model can impact Solar Radiation Forecasting

There are different methods for evaluating the performance of a forecasting model. For our study, we will use R^2, MAE and RMSE as metrics to evaluate our model.

To evaluate the prediction accuracy, the data were trained on solar radiation data for all values from 00:00 on 1 January 2018 up to 23:30 on 31 December 2019 and was used to predict 00:00 to 23:30 on 1 January 2020 (48 samples). Cross validation was performed, showing that the models generalise well. The models were run and the acutal results were compared with prediction results for each of the models.

Fig.7. Model Comparison



Visual inspection of the forecasting figures above shows Decision Tree, XG Boost, and Random Forest closely followed the actual. Linear regression followed the pattern but still ran a little short.

The Facebook Prophet captured seasonality in a distinct way. The Prophet model overshoots the actual values. However, it formed a very smooth parabolic rise and fall. Facebook Prophet has a different algorithm structure, the feature is the time series of solar radiation up to the values that are predicted. Facebook Prophet generated large negative values for short-term predictions. For all negative predictions (which only occurred in winter), the target value was zero. This shows that Facebook Prophet only forecasted negative values during the night hours. In summer, all night hour predictions were positive. As there could not be negative solar radiation and most negative predictions occurred at night, all negative values were eliminated and set to zero. For most predictions with negative values, the target value was zero. For the non-zero target values, the radiation was very low. Therefore, here too, all negative values were set to zero. Surprisingly the Amazon Deep AR appeared to follow the average path and did not account for any seasonality trends. Deep AR behavior following average was more evident as MAE was zero for the Deep AR model.

The team built an Ensemble model averaging Decision Tree and XG Boost. That also stayed very close to actual and gave MAE as zero. However, the team went beyond a visual inspection and used quantitative metrics to evaluate the performance of each model.

Fig.8. Metrics used to evaluate the models

	Decision Tree	XGBoost	Random Forest	Linear_Reg	Prophet	DeepAR	Ensemble model
RMSE	2.933030	1.693062	1.046156	40.058722	105.682482	62.368580	1.791438
MAE	1.146674	0.774938	0.292375	28.118928	58.780489	0.000000	0.000000
R-Square	0.998153	0.999318	0.999734	0.669356	0.747656	0.237471	0.999328

Based on the above metrics Random Forest came out to be best model predicting near term values and very closely followed by Ensemble model.

In the final section of our results, would like to discuss about some operational challenges we had with Facebook Prophet and Amazon's DeepAR+
The Prophet forecasting model is an open source procedure for forecasting timeseries data. It is based on an additive model where the first component is the trend of data behavior. Nonlinear data behavior is described by seasonality on a daily, weekly, or annual basis. The algorithm also observes the effects of holidays, which increases its accuracy. So, in our case it picks the seasonality correctly but because of it additive nature it tries to smooth the predicted values.
Amazon Deep AR+ is not a open source model, one of the possible reason why this model did not perform could be the way it is trained and how it works. Unlike standard approaches, such as ARIMA (autoregressive integrated moving average) or ETS (exponential smoothing), one model was used to fit multiple time series internally in Deep AR+. Since solar radiation time series has multiple seasonality(daily, quaterly and yearly), it was not able to pick the seasonality properly and it gave a flat prediction. [20]

6 Conclusion

The conclusion [39] will highlight the successful implementation of computationally efficient techniques (leveraging linear and nonlinear models) to do shortterm forecast of solar radiation. This paper presented multiple ML algorithms like Facebook Prophet, Amazon's DeepAR+, Linear Regression, RandomForest, XGBoost and Decision Tree to enhance the prediction of solar radiation. Facebook Prophet and Amazon's DeepAR+ is a novel algorithm that has rarely been used in the field of solar radiation forecast. Nevertheless, its design characteristics seemed inherently promising for solar prediction. Firstly, a basic model was constructed for both algorithms with 30 minute interval solar radiation as input. The data were taken from the National Solar Radiation Database (NSRDB). Adding other meteorological factors led to the largest improvement in R2, MAE, and RMSE for the ML models. After visual examination and evaluating

quantitative metrics it is recommended to use Random Forest along with ensemble model to eliminate bias and variance.

7 Acknowledgements

The researchers would like to thank Prof. Bradley Blanchard, Mr. Ashwin Thota, and Mr. Grant Buster from NREL for their insight, guidance, and support.

Bibliography

[1] David Feldman, Kevin Wu, and Robert Margolis. H1 2021 solar industry update. Technical report, National Renewable Energy Lab.(NREL), Golden, CO (United States), 2021.

[2] Zhu Zhongming, Lu Linong, Yao Xiaona, Zhang Wangqiang, Liu Wei, et al. Texas likely to add record utility-scale solar capacity in the next two years. 2021.

[3] Ahmet Teke, H Ba¸sak Yıldırım, and "Ozg"ur C¸elik. Evaluation and performance comparison of different models for the estimation of solar radiation. *Renewable and sustainable energy reviews*, 50:1097–1107, 2015.

[4] Wendy S Parker. Ensemble modeling, uncertainty and robust predictions. *Wiley Interdisciplinary Reviews: Climate Change*, 4(3):213–223, 2013.

[5] D Laucelli, O Giustolisi, V Babovic, and M Keijzer. Ensemble modeling approach for rainfall/groundwater balancing. *Journal of Hydroinformatics*, 9(2):95–106, 2007.

[6] K Krishna Rani Samal, Korra Sathya Babu, Santosh Kumar Das, and Abhirup Acharaya. Time series based air pollution forecasting using sarima and prophet model. In *proceedings of the 2019 international conference on information technology and computer communications*, pages 80–85, 2019.

[7] Syama Sundar Rangapuram, Matthias W Seeger, Jan Gasthaus, Lorenzo Stella, Yuyang Wang, and Tim Januschowski. Deep state space models for time series forecasting. *Advances in neural information processing systems*, 31, 2018.

[8] Yuyang Wang, Alex Smola, Danielle Maddix, Jan Gasthaus, Dean Foster, and Tim Januschowski. Deep factors for forecasting. In *International conference on machine learning*, pages 6607–6617. PMLR, 2019.

[9] Foad H Gandoman, Shady HE Abdel Aleem, Noshin Omar, Abdollah Ahmadi, and Faisal Q Alenezi. Short-term solar power forecasting considering cloud coverage and ambient temperature variation effects. *Renewable Energy*, 123:793–805, 2018.

[10] Yixiao Yu, Xueshan Han, Ming Yang, and Jiajun Yang. Probabilistic prediction of regional wind power based on spatiotemporal quantile regression. *IEEE Transactions on Industry Applications*, 56(6):6117–6127, 2020.

[11] Taku Yamamoto. Predictions of multivariate autoregressive-moving average models. *Biometrika*, 68(2):485–492, 1981.

[12] SK Aggarwal and LM Saini. Solar energy prediction using linear and nonlinear regularization models: A study on ams (american meteorological society) 2013–14 solar energy prediction contest. *Energy*, 78:247–256, 2014.

[13] Emre Akarslan, Fatih Onur Hocao˘glu, and Rifat Edizkan. A novel md (multi-dimensional) linear prediction filter approach for hourly solar radiation forecasting. *Energy*, 73:978–986, 2014.

[14] Fatih Onur Hocaoglu and Fatih Serttas. A novel hybrid (mycielski-markov) model for hourly solar radiation forecasting. *Renewable Energy*, 108:635–643, 2017.

[15] Lamara Benali, Gilles Notton, A Fouilloy, Cyril Voyant, and Rabah Dizene. Solar radiation forecasting using artificial neural network and random forest methods: Application to normal beam, horizontal diffuse and global components. *Renewable energy*, 132:871–884, 2019.

[16] Ayu Wazira Azhari, Kamaruzzaman Sopian, Azami Zaharim, and Mohamad Al Ghoul. A new approach for predicting solar radiation in tropical environment using satellite images-case study of malaysia. *WSEAS Transactions on Environment and Development*, 4(4):373–378, 2008.

[17] El-Sayed M El-Kenawy, Seyedali Mirjalili, Sherif SM Ghoneim, Marwa Metwally Eid, Mohammed El-Said, Zeeshan Shafi Khan, and Abdelhameed Ibrahim. Advanced ensemble model for solar radiation forecasting using sine cosine algorithm and newton's laws. *IEEE Access*, 9:115750–115765, 2021.

[18] David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3):1181–1191, 2020.

[19] H Aguilera, C Guardiola-Albert, N Naranjo-Fern´andez, and C Kohfahl. Towards flexible groundwater-level prediction for adaptive water management: using facebook's prophet forecasting approach. *Hydrological sciences journal*, 64(12):1504–1518, 2019.

[20] Emir Zunic, Kemal Korjenic, Sead Delalic, and Zlatko Subara. Comparison analysis of facebook's prophet, amazon's deepar+ and cnn-qr algorithms for successful real-world sales forecasting. *arXiv preprint arXiv:2105.00694*, 2021.

[21] Manajit Sengupta, Yu Xie, Anthony Lopez, Aron Habte, Galen Maclaurin, and James Shelby. The national solar radiation data base (nsrdb). *Renewable and sustainable energy reviews*, 89:51–60, 2018.

[22] Sandeep Koranne. Hierarchical data format 5: Hdf5. In *Handbook of open source tools*, pages 191–200. Springer, 2011.

[23] Leland Wilkinson and Michael Friendly. The history of the cluster heat map. *The American Statistician*, 63(2):179–184, 2009.

[24] Alexander D"urre, Roland Fried, and Tobias Liboschik. Robust estimation of (partial) autocorrelation. *Wiley Interdisciplinary Reviews: Computational Statistics*, 7(3):205–222, 2015.

[25] Serge D'egerine. Sample partial autocorrelation function of a multivariate time series. *Journal of multivariate analysis*, 50(2):294–313, 1994.

[26] Qingsong Wen, Jingkun Gao, Xiaomin Song, Liang Sun, Huan Xu, and Shenghuo Zhu. Robuststl: A robust seasonal-trend decomposition algorithm for long time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5409–5416, 2019.

[27] Karl G J¨oreskog. A general method for analysis of covariance structures. *Biometrika*, 57(2):239–251, 1970.

[28] S Ibrahim, I Daut, YM Irwan, M Irwanto, N Gomesh, and Z Farhana. Linear regression model in estimating solar radiation in perlis. *Energy Procedia*, 18:1402–1412, 2012.

[29] Da Liu and Kun Sun. Random forest solar power forecast based on classification optimization. *Energy*, 187:115940, 2019.

[30] Xianglong Li, Longfei Ma, Ping Chen, Hui Xu, Qijing Xing, Jiahui Yan, Siyue Lu, Haohao Fan, Lei Yang, and Yongqiang Cheng. Probabilistic solar irradiance forecasting based on xgboost. *Energy Reports*, 8:1087–1095, 2022.

[31] Joseph M Hellerstein. Quantitative data cleaning for large databases. *United Nations Economic Commission for Europe (UNECE)*, 25:1–42, 2008.

[32] Miodrag Lovri'c, Marina Milanovi'c, and Milan Stamenkovi'c. Algoritmic methods for segmentation of time series: An overview. *Journal of Contemporary Economic and Business Issues*, 1(1):31–53, 2014.

[33] Zaiqiao Meng, Richard McCreadie, Craig Macdonald, and Iadh Ounis. Exploring data splitting strategies for the evaluation of recommendation models. In *Fourteenth ACM conference on recommender systems*, pages 681–686, 2020.

[34] Osnat Israeli. A shapley-based decomposition of the r-square of a linear regression. *The Journal of Economic Inequality*, 5(2):199–212, 2007.

[35] Lanre Olatomiwa, Saad Mekhilef, Shahaboddin Shamshirband, Kasra Mohammadi, Dalibor Petkovi´c, and Ch Sudheer. A support vector machine–firefly algorithm-based model for global solar radiation prediction. *Solar Energy*, 115:632–644, 2015.

[36] Victor H Quej, Javier Almorox, Javier A Arnaldo, and Laurel Saito. Anfis, svm and ann soft-computing techniques to estimate daily global solar radiation in a warm sub-humid environment. *Journal of Atmospheric and Solar-Terrestrial Physics*, 155:62–70, 2017.

[37] Bj¨orn Wolff, Jan Ku¨hnert, Elke Lorenz, Oliver Kramer, and Detlev Heinemann. Comparing support vector regression for pv power forecasting

to a physical modeling approach using measurement, numerical weather prediction, and cloud motion data. *Solar Energy*, 135:197–208, 2016.

[38] Huan Long, Zijun Zhang, and Yan Su. Analysis of daily solar power prediction with data-driven approaches. *Applied Energy*, 126:29–37, 2014.

[39] Iram Naim, Tripti Mahara, and Ashraf Rahman Idrisi. Effective short-term forecasting for daily time series with complex seasonal patterns. *Procedia computer science*, 132:1832–1841, 2018.