Solar-Cast: Solar Power Generation Prediction from Weather Forecasts using Machine Learning

Rishi Singhal
Department of Electronics & Communication Engineering
Indraprastha Institute of Information Technology, New Delhi
Delhi, India
rishi19194@iiitd.ac.in

Poonam Singhal
Department of Electrical Engineering
J.C. Bose University of Science & Technology, YMCA
Faridabad, India
poonamsinghal@jcboseust.ac.in

Shailender Gupta
Department of Electronics & Communication Engineering
J.C. Bose University of Science & Technology, YMCA
Faridabad, India
shailender@jcboseust.ac.in

2022 IEEE 10th Power India International Conference (PIICON) | 978-1-6654-5930-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/PIICON56320.2022.10045237

***Abstract*— The rapid growth of solar generation technology has become a boon in the energy sector. Smart grids have replaced the conventional Grids due to upcoming various distributed energy sources feeding the grid. The correct estimation of solar intensity according to geographical features will help in determining the capacity of smart Grids. In real-time, most smart grids are compelled to change their renewable energy production process according to the real-time availability of energy resources like wind and solar during the day. Thus, to assuage this problem, the possibility is investigated by using readily available weather data on the NSRDB website to predict solar forecasts 48 hours ahead in the future by using various Machine Learning(ML) algorithms. In this paper, a new day-night model has been designed to limit the uncertainty of solar power generation and reduce the dependability of power grids on non-renewable energy sources like fossil fuels. Further, to improve the solar forecasting prediction, multiple weather observations were taken from preceding time intervals to establish a new data set in linear regression and its subtypes (Ridge and Lasso regression).**

Keywords—*Smart Grids, Renewable, Solar, Machine Learning(ML), NSRDB*

I. Introduction

Solar energy is a non-depletable, non-polluting, and renewable energy source. With the current global energy crisis, using solar energy to generate electricity has become crucial. However, the amount of solar energy that can be collected and transformed into electricity is minimal when considering the existing techniques. The highly variable nature of solar energy production puts stress on fossil fuel-based power generation [3]. This paper aims to predict solar intensity for a given area 48 hours into the future using local time-series weather data. The goal is to provide high-confidence solar generation forecasts (via solar intensity) using readily-available weather data. This will enable better regulation of fossil fuel-based power generation. The solar intensity and weather data have been acquired from NSRDB [1] and Sunrise-Sunset-API [2]. This paper aligns with sustainable development goals and deploying machine learning techniques to solve real-world environmental problems.

In this paper, a scheme has been proposed to do the solar forecasting task in which various Exploratory Data Analysis (EDA) and data preprocessing methods on the NSRDB dataset have been performed. Then various Machine Learning (ML) models are implemented and compared based on Root Mean Square Error (RMSE) as the performance metric. The proposed scheme is depicted using the block diagram in Fig 1.

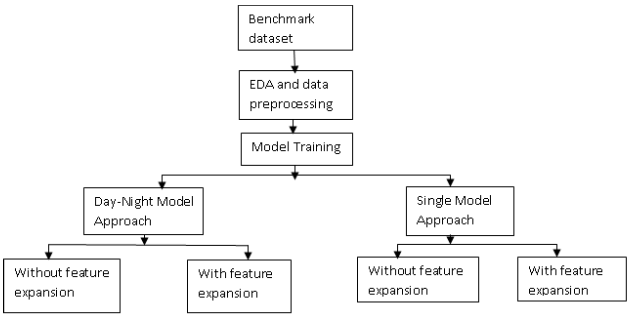


Fig 1: Proposed Scheme

The rest of the paper is organized as follows: Section II includes a comprehensive literature survey. The description of the datasets considered is given in Section III. The detailed methodology is explained in Section IV, with results represented in Section V. Section VI concludes the paper.

II. LITERATURE SURVEY

A significant amount of work has been done in the field of solar forecasting using weather data. Elsaraiti et al. [8] proposed deep learning techniques like LSTM to forecast solar power data in the short term. They compared the performance of the LSTM model with MLP using MAE and RMSE & showed that LSTM outperformed the MLP model.

Huang et al. [9] used twelve machine-learning models to predict daily and monthly solar radiation values by incorporating meteorological factors. Furthermore, they used the top five models by creating a stacked model and showed that it outperformed the single 12 models.

Vennila et al. [7] proposed a hybrid model incorporating machine learning and statistical approaches to predict future solar energy generation. Also, to further improve

978-1-6654-5930-3/22/$31.00 ©2022 IEEE

the accuracy, an ensemble of machine learning models was used.

Anuradha et al. [4] focused on the problem of solar power forecasting using National Weather Service (NWS) weather data. They utilized three ML algorithms: Linear Regression, SVM, and Random Forest, and compared them using RMSE as the performance measure metric. They concluded that Random Forest beat the other two models by many folds.

Pedro and Coimbra [5] used data from a solar power plant in Central California to conduct their experiments to predict the solar power output based on their own data. They used the Persistent model, ARIMA, KNNs, and ANNs for their study and were able to achieve significant improvement using ANNs in their predictions which can be further improved using a genetic algorithms-based optimization technique. They were able to achieve these results without the use of any exogenous data.

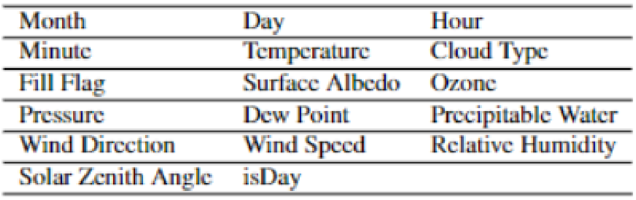
Andrade & Bessa [6] developed a forecasting framework for renewable resources (wind and solar energy) using information from Numerical Weather Predictions (NWP) grid. They used different smoothing techniques to create features from the NWB grid data. To develop this framework, a combination of gradient boosting trees and PCA. Three performance metrics (RMSE, MAE & CRPS) were used to evaluate their framework.

III. DATASET

A. General Information

The NSRDB dataset includes observed weather data (temperature, pressure, cloud cover, solar zenith angle, etc.) and solar intensity data measured in watts per square meter. The dataset includes several solar radiation measures, such as Diffused Normal Irradiance (DNI), Diffused Horizontal Irradiance (DHI), and Global Horizontal Irradiance (GHI). GHI measurement was used as the target variable since it incorporates DHI, DNI, and ambient solar radiation reflected from nearby surfaces. This makes it a good indicator for solar panel readings. The NSRDB data is measured once every 10 minutes. A single location of Las Vegas, Nevada, USA, was investigated for the year 2019 for the dataset.

The dataset contains more than 50,000 distinct observations, each with 16 features (including time values such as Month, Day, Hour, and Minute) shown in the figure below with a corresponding GHI measure.



Month	Day	Hour
Minute	Temperature	Cloud Type
Fill Flag	Surface Albedo	Ozone
Pressure	Dew Point	Precipitable Water
Wind Direction	Wind Speed	Relative Humidity
Solar Zenith Angle	isDay	

Fig 2: Features of the dataset

B. Exploratory Data Analysis

The sunrise and sunset times for each day from Sunrise-Sunset-API [2] have been obtained, and by using these times, a new boolean column 'isDay' was added. The distribution of the fill flag column using the pie chart, along with the pairwise correlation between features using the correlation matrix are plotted. Furthermore, the skew measure of each of the features is provided in the appendix section.

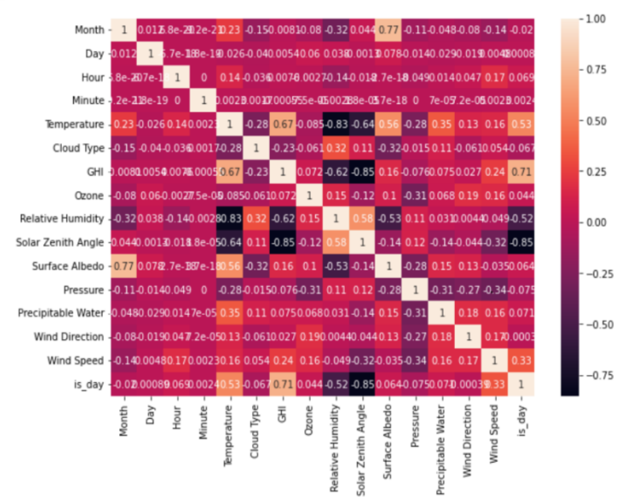


Fig 3: Correlation Matrix

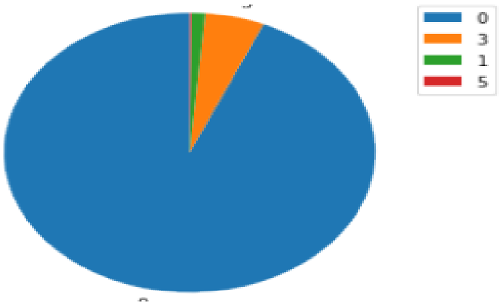


Fig 4: Fill Flag Pie Chart

C. Data Pre-Processing and Analysis

From the figures, various outliers were identified. However, it was assumed that they are the natural part of the weather observations; hence were not removed from the dataset. Also, there were no missing/NaN values in the dataset under study.

The Correlation Matrix(Fig 3) shows that features 'Dew Point' and 'Precipitable Water' are highly correlated. Since highly correlated features have almost the same effect on the dependent variable, so only one of them needs to be considered. Thus, using the baseline model, the RMSE comparison was made by keeping one of the features at a time. Thereby, the “Dew Point” column was dropped as including it in the dataset resulted in a higher RMSE value.

The Pie-Chart for the 'Fill Flag' [Fig 4] feature shows that more than 90% of observations correspond to '0'. It indicates that the value is not available. Therefore 'Fill Flag' adds negligible information to the dataset, hence it was dropped.

D. Dataset Preparation

Since the task is to predict solar intensity values 48 hours into the future. Thus, a one-to-one mapping between current weather observations and GHI values 48 hours ahead in the future has been developed.

IV. METHODOLOGY

To perform this task, two different methodologies were applied, which are as follows:

A. Model 1: Single Model Approach

In the first methodology, a single model was implemented and trained on 80% of the data and was later used for the rest 20% of the data, i.e., the testing set.

B. Model 2: Day-Night Model Approach

In the alternative approach, two different models were designed for the day data & one for the night data. In this method, the training set was divided into two sets, one corresponding to the daytime & and the other to the nighttime. This was achieved by using the *isDay* column that was discussed earlier. These proposed models: the day model and night model, were trained on the day dataset and the night dataset, respectively. And while testing, the day-time sample was predicted using the day model and the night-time sample using the night model. Finally, the predictions made by the two models were concatenated in a list and used for calculating the final RMSE value. This proposed methodology is further explained using a schematic diagram in Fig 6.

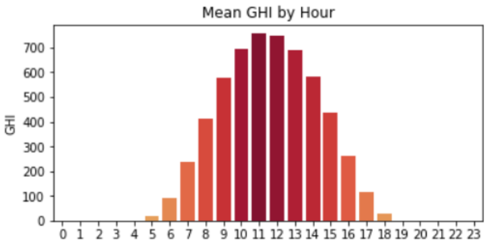


Fig 5(a)

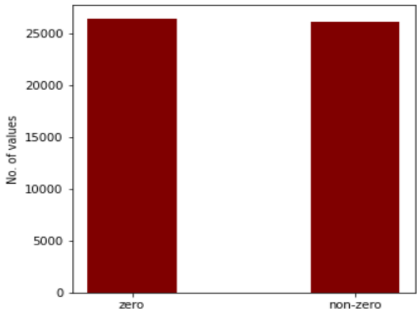


Fig 5(b)

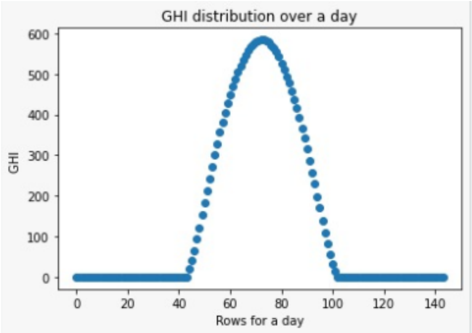


Fig 5(c)

Fig 5 (a) Mean GHI by the hour, (b) GHI Sparsity, (c) GHI distribution over a day.

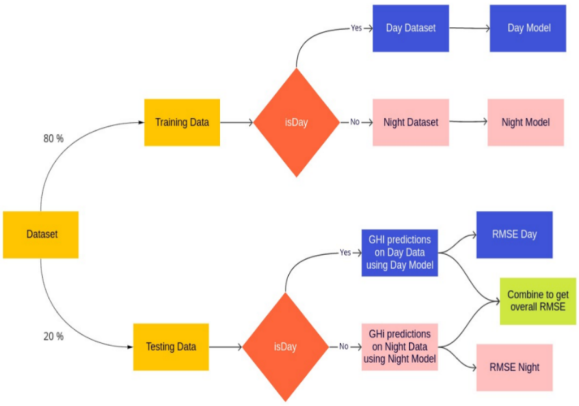

Dataset
80 %
20 %
Training Data
Testing Data
isDay
Yes
No
Day Dataset
Night Dataset
Day Model
Night Model
GHI predictions on Day Data using Day Model
GHI predictions on Night Data using Night Model
RMSE Day
RMSE Night
Combine to get overall RMSE

Fig 6: Schematic diagram of the Day-Night model

As from Fig 5, it is observed that the GHI column is highly sparsed with around 50% values equal to zero. Also, most of the zero values correspond to the time before sunrise and after sunset. Therefore, this sparseness was also considered while designing the second methodology.

Finally, To perform this regression task, the following models were used: Linear Regression & its subtypes (Lassor & Ridge), Polynomial Regression, Regression Trees, SVM Regression, and Artificial Neural Networks (ANNs), PCA. Apart from this, a new feature expansion technique has been implemented in the case of Linear Regression (with and without PCA) to decrease the RMSE value.

C. Regression Models and Their Details

a. Linear Regression & Its Subtypes

It is a supervised learning algorithm that uses a linear function to predict a target variable based on independent variables, also known as features. It minimizes the mean squared error to come up with the optimal solution, which includes finding the optimal parameter values for weights and bias. Apart from this, two more Linear Regression variants were used, i.e., Ridge Regression and Lasso Regression.

- Ridge Regression: The cost function also includes an added penalty corresponding to the squared value of coefficients.
- Lasso Regression: The cost function also includes an added penalty corresponding to the absolute value of coefficients.

These techniques help in reducing overfitting and help in reducing the model complexity.

b. Polynomial Regression

It is a supervised learning algorithm that uses a polynomial function to predict a target variable based on independent variables, also known as features. The degree of the polynomial is a hyperparameter in this technique. It minimizes the mean squared error to come up with the optimal solution.

c. Regression Trees

It is a supervised learning, tree-based model used for regression tasks to predict continuous valued outputs. At each node, the idea is to minimize the mean squared error at the point and split the data into two parts.

d. Support Vector Machine (SVM) Regression

It is a supervised learning algorithm used for regression tasks to predict real, discrete values. It involves finding the optimal hyperplane with maximum data points within the decision boundary. The kernel function is a hyperparameter in this model and is used to transform the non-linearly separable data to linearly-separable data in higher dimensions.

e. Artificial Neural Network (ANN)

Nowadays, neural networks are used a lot because they can find patterns that might not be easy to detect for other ML classifiers. Thus, an ANN model was also incorporated by finetuning its various parameters, i.e., the number of layers, layer sizes & activation functions.

D. Feature Expansion Technique

As part of the proposed feature expansion method, the feature set was expanded for predicting the GHI values by adding weather observations from the adjacent past time points. In the starting, each data sample had 15 features, and after the nth expansion, the number of features increased to 15*(n+1), where n = {0,1,2,3}

E. Performance Metrics

To evaluate the various models, Root Mean Square Error (RMSE) as the performing metric.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (1)$$

Here,

n = *number of samples*
y_i = *ground truth target variable value*
$\hat{y}_i$ = *predicted target variable value*

Generally, low RMSE values are required for regression tasks, which is also applicable to the GHI value prediction task as the measurements need to be as close as the original values.

V. RESULTS & ANALYSIS

For all the models, the train set size was kept as 80%, and the test size was 20% of the original dataset.

A. Linear Regression: Baseline Model

Linear Regression(LR) model and its subtypes, i.e., Lasso and Ridge Regression were trained, for both the mentioned methodologies: the single model approach and the day-night model approach. The performance of these models on the test set is summarized in Table I.

Table I: Linear Regression: Comparison of Single model and Day-Night Model (Baseline)

Model	Train RMSE		Test RMSE	
	Single model	Day-night model	Single model	Day-night model
Linear Regression	151.54	81.135	152.01	82.182
Lasso Regression	151.54	81.135	152.01	82.182
Ridge Regression	151.54	81.137	152.01	82.178

Through the above table, it can be seen that there is low variance in the baseline LR models (both in the single as well as in the day-night model approach). However, the day-night model performed better than the single model because the sparsity in the dataset was not considered in the case of the single-model approach. Furthermore, since the models are not overfitting, applying Lasso and Ridge did not help to improve the performance.

B. Linear Regression with Feature Expansion

As an extension to the baseline linear regression model, the proposed feature expansion technique was included for predicting solar intensity. The number of feature expansions that were tried as part of this paper is mentioned in the table given below.

Table II: Expansion v/s Features

Expansions	1	2	5	30	60
Features	30	45	90	465	915

The performance report of the models is depicted in the following figure.

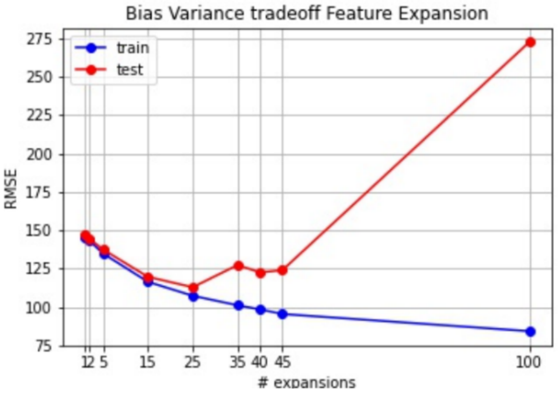
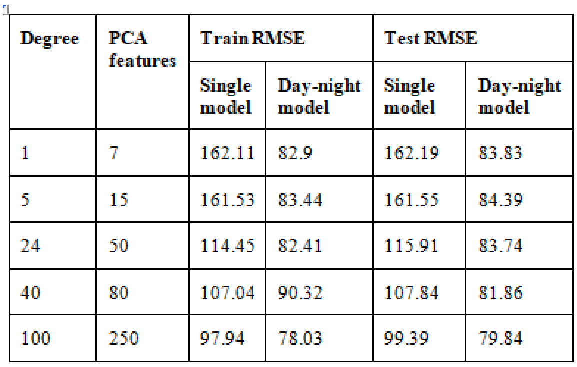


Fig 7: RMSE v/s Feature Expansion without PCA

Table III: Linear Regression with feature expansion:

Single & Day-Night Models with PCA



Degree	PCA features	Train RMSE		Test RMSE	
		Single model	Day-night model	Single model	Day-night model
1	7	162.11	82.9	162.19	83.83
5	15	161.53	83.44	161.55	84.39
24	50	114.45	82.41	115.91	83.74
40	80	107.04	90.32	107.84	81.86
100	250	97.94	78.03	99.39	79.84

From the above RMSE plots, it is inferred that on increasing the number of expansions, both train and test RMSE decreased till 25 feature expansions. But on further expansions, there is an increase in the variance due to the high dimensionality of the data. Apart from this, the Day-Night Model approach performs better than the Single-Model approach.

Therefore, to reduce the model's dimensions and complexity, the Principal Component Analysis (PCA) technique was used to get the features with high explained variance. Thus, after applying the feature expansion with PCA, the complexity of the model decreased. Therefore with the increase in the number of expansions, both the train and test RMSE decreased, as shown in the table above.

C. Polynomial Regression

Before jumping to more complex models, polynomial regression was used to achieve better RMSE values with a good bias-variance tradeoff. Through this method, the feature set is further expanded by including the square of each feature, as well as the pairwise interaction between each pair of features. Since the expansion in the number of features was exponential, polynomials for degrees 2, 3, and 4 were only used. The results for the same are described below.

Table IV: Polynomial Regression: Comparison of Single Model and Day Night Model

Degree	Train RMSE		Test RMSE	
	Single model	**Day-Night model**	**Single model**	**Day-Night model**
2	78.95	75.66	79.96	76.81
3	69.49	63.36	70.53	65.58
4	51.98	45.11	55.99	55.63

It can be observed in both the Day-Night and Single-Model approaches that with an increase in the degree of the polynomial regression, the train and test RMSE gets reduced. Although at degree 4, see the variance started to increase because of over fitting. The degree could not be increased beyond four because of limited computational resources.

D. Support Vector Machine (SVM) Regression

Since, Support Vector Machines (SVM) can handle the nonlinearity in the data. Thus the SVM model is also being used in this paper. The performance of the SVM depends on selecting an appropriate kernel function and parameters. The SVM model uses a kernel function that transforms the data from the input space to a high-dimensional feature space. Thus, for this paper, four distinct SVM kernel functions: Linear Kernel, Polynomial Kernel, Sigmoid Kernel, and the Radial Basis Function (RBF) were considered. After testing both the models (single model approach and day-night model approach) using the test set, the following results were achieved through which it was inferred that the RBF kernel gave the best possible RMSE value on the test set with a good bias-variance tradeoff in the Day-Night model approach.

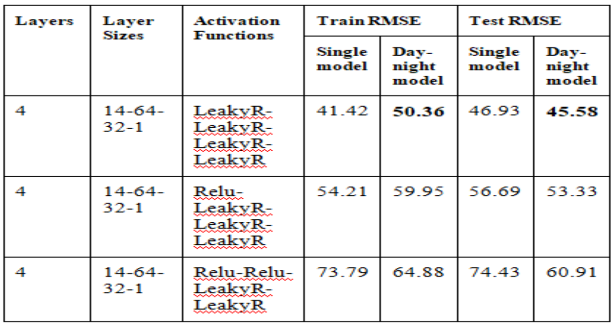
Table V: SVM Regression: Comparison of Single Model & Day-Night Model

Kernel	Train RMSE		Test RMSE	
	Single model	**Day-Night model**	**Single model**	**Day-Night model**
rbf	118.73	90.52	119.5	90.64
sigmoid	221.71	113.21	221.05	113.81
linear	153.62	85.63	153.97	87.71
polynomial	171.53	111.65	169.9	110.97

E. Artificial Neural Network (ANN)

As the above-mentioned models are not too complex, thus ANN architecture has also been incorporated as part of this study. To implement the ANN models, the Keras library has been used. Furthermore, finetuning of the model parameters is being done and these model combinations were tested and compared based on the test RMSE value, the bias-variance tradeoff, and the model complexity. The results of the same are mentioned in Table VI along with the loss plots of the best case. Here, LeakyR refers to the LeakyReLu activation function.

Table VI: ANN Results: Single model & Day-Night Model Comparison



Layers	Layer Sizes	Activation Functions	Train RMSE		Test RMSE	
			Single model	Day-night model	Single model	Day-night model
4	14-64-32-1	LeakyR-LeakyR-LeakyR-LeakyR	41.42	**50.36**	46.93	**45.58**
4	14-64-32-1	Relu-LeakyR-LeakyR-LeakyR	54.21	59.95	56.69	53.33
4	14-64-32-1	Relu-Relu-LeakyR-LeakyR	73.79	64.88	74.43	60.91

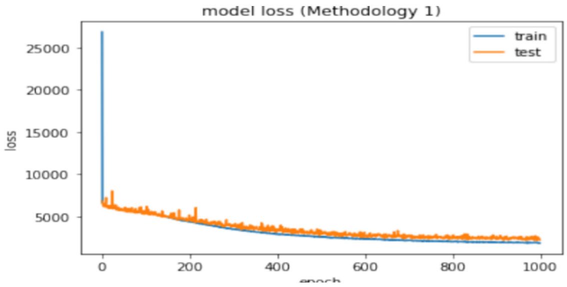


Fig 8 : ANN Loss Plot: Single Model

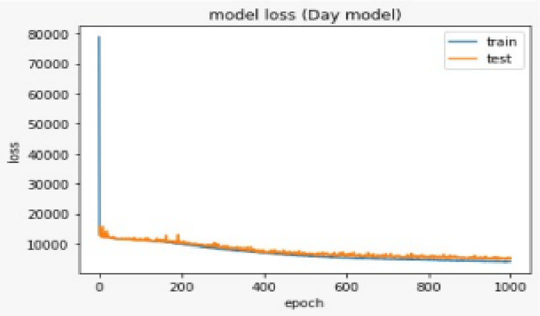


Fig 9: ANN Loss Plot: Day Model

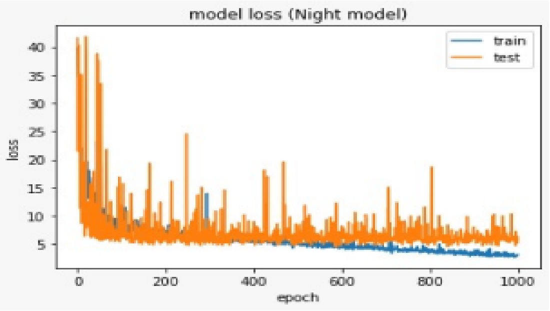


Fig 10: ANN Loss Plot: Night Model

After observing Table VI, it is clear that the best results, in this case, were achieved when four layers of size {14,64,32,1} with activation functions as [LeakyReLU, LeakyReLU, LeakyReLU, LeakyReLU] in the respective layers were used. This was applicable to both the single model approach and the day-night model approach. Thus, after comparing all the model's results, it is concluded that the ANN model performed the best because of its lowest test RMSE value of 45.58 and an optimal bias-variance tradeoff.

VI. CONCLUSION & FUTURE WORK

Since the number of people utilizing renewable energy sources is increasing daily in response to the quick depletion of non-renewable resources like fossil fuels. Thus, it becomes necessary to estimate when would be the best time to keep the renewable energy plants open to have maximum energy production. Thus, this paper focuses on predicting the solar power intensity that a region would have 48 hours into the future. The proposed scheme used various EDA methods, data-preprocessing techniques, and other visualization methods, followed by feature expansion. Finally, machine learning models like linear regression, regression trees, polynomial regression, SVM regression, and ANN were trained on the resultant feature set. Apart from this, a day-night model approach was also proposed in addition to the single model approach. These models were evaluated on the test set using the RMSE value as the performance metric. The proposed neural network having four layers, with layer sizes 14-64-32-1 and leaky ReLU activation function in all layers, gave the best results for both the single-model and the day-night model approach. However, the day-night model approach performs better than the single-model approach by achieving the least test RMSE value of 45.58 and an optimal bias-variance tradeoff.

As part of future work, more complex deep learning models and different regularization techniques can be used to achieve better RMSE values. The dataset can be further expanded by acquiring the weather data of various other geographical locations and satellite image data.

APPENDIX

Skew index measure of the dataset features:

Month -0.01, Day 0.07, Hour 0, Minute 0, Temperature 0.55, Cloud Type 1.3, Dew Point -1.34, Fill Flag 3.96, Ozone 0.78, Relative Humidity -0.091, Solar Zenith Angle -0.0001, Surface Albedo -0.71, Pressure 0.046, Precipitable Water 0.62, Wind Direction -0.73, Wind Speed 0.99, isDay -0.038

REFERENCES

[1] Google (2022) *NSRDB* Available at: https://nsrdb.nrel.gov/ (Accessed 21 June 2022)

[2] Google (2022) *Sunrise-Sunset-API* Available at: https://sunrise-sunset.org/api (Accessed 22 June 2022)

[3] Inman, R.H., Pedro, H.T. and Coimbra, C.F., 2013. Solar forecasting methods for renewable energy integration. *Progress in energy and combustion science*, *39*(6), pp.535-576.

[4] Anuradha, K., Erlapally, D., Karuna, G., Srilakshmi, V. and Adilakshmi, K., 2021. Analysis Of Solar Power Generation Forecasting Using Machine Learning Techniques. In *E3S Web of Conferences* (Vol. 309, p. 01163). EDP Sciences.

[5] Pedro, H.T. and Coimbra, C.F., 2012. Assessment of forecasting techniques for solar power production with no exogenous inputs. *Solar Energy*, *86*(7), pp.2017-2028.

[6] Andrade, J.R. and Bessa, R.J., 2017. Improving renewable energy forecasting with a grid of numerical weather predictions. *IEEE Transactions on Sustainable Energy*, *8*(4), pp.1571-1580.

[7] Vennila, C., Titus, A., Sudha, T., Sreenivasulu, U., Reddy, N., Jamal, K., Lakshmaiah, D., Jagadeesh, P. and Belay, A., 2022. Forecasting Solar Energy Production Using Machine Learning. *International Journal of Photoenergy*, *2022*.

[8] Elsaraiti, M. and Merabet, A., 2022. Solar power forecasting using deep learning techniques. *IEEE Access*, *10*, pp.31692-31698.

[9] Huang, L., Kang, J., Wan, M., Fang, L., Zhang, C. and Zeng, Z., 2021. Solar radiation prediction using different machine learning algorithms and implications for extreme climate events. *Frontiers in Earth Science*, *9*, p.596860.