Predicting the Clearsky GHI for solar panels

Kemal Şahin [1] **Tuncer Sivri** [2] **Burak Kurt** [3]

Abstract

This paper explores the prediction of Global Horizontal Irradiance (GHI), a critical factor influencing the efficiency of solar panels. Accurate GHI forecasting is essential for optimizing the planning and management of solar energy systems. The study investigates various machine learning techniques, including linear regression, Random Forest, 1D-CNN, and LSTM, utilizing weather data as input features. Different models are compared in terms of performance, considering their strengths and weaknesses. The proposed methodology involves ensemble voting, a technique to aggregate predictions from multiple models, aiming to enhance overall accuracy.

1. Introduction

The total amount of solar radiation that reaches the Earth's surface per unit area and time is called Global Horizontal Irradiation (GHI). It is a factor that affects the design and operation of solar panels and other renewable energy systems, as the potential power output and efficiency of these devices are determined by it. The prediction of GHI is influenced by many complex and dynamic factors, such as cloud cover, aerosols, humidity, and atmospheric conditions. Therefore, accurate and reliable GHI forecasting methods are needed to optimize the planning and management of solar energy systems. Machine learning techniques, which are methods that learn from data and make predictions based on patterns and relationships, are one way to approach GHI prediction. Nonlinear and high-dimensional data can be handled by machine learning, and it can adapt to changing environments and scenarios. Different machine learning techniques for GHI prediction, using weather data as input features, will be explored in this report. The performance of linear, nonlinear, and sequential models, such as linear regression, Random Forest, and recurrent neural networks, will be compared. The advantages and limitations of each technique will also be discussed, and possible improvements and future directions will be suggested.

2. Related Work

The problem of optimizing renewable energy sources for sustainable development has been addressed by various researchers using different methods and perspectives. In this section, some of the related works will be reviewed, and highlighted their contributions and limitations. Researches show that some machine learning methods that are not based on neural networks can perform well on this task. In (Feng et al., 2019), researchers used a random forest algorithm along with neural networks. Results are very close, even the RF algorithm outperformed NN-based models in some cases. K-nearest-neighbor algorithm also performs well on image data with proper data preprocessing (Pedro & Coimbra, 2015). However, most of the time neural networks are preferred to solve this problem. In (Chiteka & Enweremadu, 2016), researchers developed a very simple neural network with one hidden layer and seven input layers. Its performance is the best of all the models stated but its limitation is that model only works well on small feature space. For more complex inputs like images, CNN-based models are proposed in (Yang et al., 2021). Researchers developed a 3D CNN model to obtain features from image data and they used the obtained features to train their SVM, ANN, and KNN models. Results show that the CNN structure is suitable for feature-extracting tasks in this domain. Lastly, to be able to capture both spatial and temporal features, the CNN-LSTM model is proposed in (Zang et al., 2020). The proposed model first applies a convolutional neural network (CNN) to extract spatial features from a two-dimensional matrix composed of meteorological parameters associated with a target site and its neighboring sites. Then, a long short-term memory (LSTM) network is applied to extract temporal features from historical solar irradiance time series data associated with the target site. Each model explained here has strengths and weaknesses. In this study, we aim to use these advantages and drawbacks and develop a model that describes the given data best.

3. Methodology

3.1. Dataset

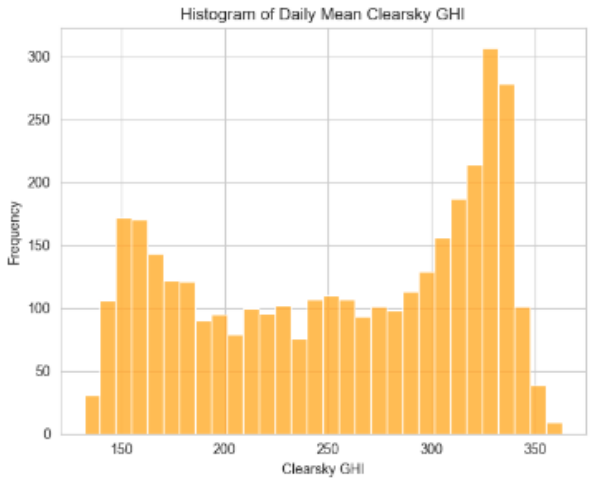
The data that will be used is from kaggle[1] and is from a company named Wipro. Data incluedes thirty minute time interval data of nearly a total of ten years with 192.816

different intervals and 18 features like temprature, pressure, wind speed etc. The main objective of the data is to highlight clearsky DHI,DNI and GHI which are all connected to each other with:

$$GHI = DHI + DNI \cdot \cos(\alpha_{\text{zenith}}) \quad (1)$$

Equation. When the data is analyzed a pattern shown in 1 can be observed for the mean distribution of GHI

Figure 1. Mean distribution of GHI



The graph shows a parabola like distribution with majority of GHI values being around 300 and 350. Finding the correct GHI value based on the features doesn't require precise work but It requires precise work to perfect it. Project's aim is to minimize the root mean square error(RMSE) so that we have the perfect model to precisely find GHI values. Thus finding the most perfect spots for solar panels which play a quite important role in sustainability. Four experiments have been done with Logistic Regressor, Random Forest Regressor, 1D-CNN and LSTM to see how far the error can be minimized. Main idea and method that is going to be followed is to create a general model which works by voting principle, It will make all of the experiment models work together and make them vote for the best result. This way it is planned to achieve the least error rate for predictions.

3.2. Root Mean Square Error (RMSE)

Root Mean Square Error (RMSE) is a performance metric commonly used in predictive modeling to assess the accuracy of a model's predictions. It measures the square root of the average squared differences between observed (y_i) and predicted ($\hat{y}_i$) values across a dataset of N observations. The formula is:

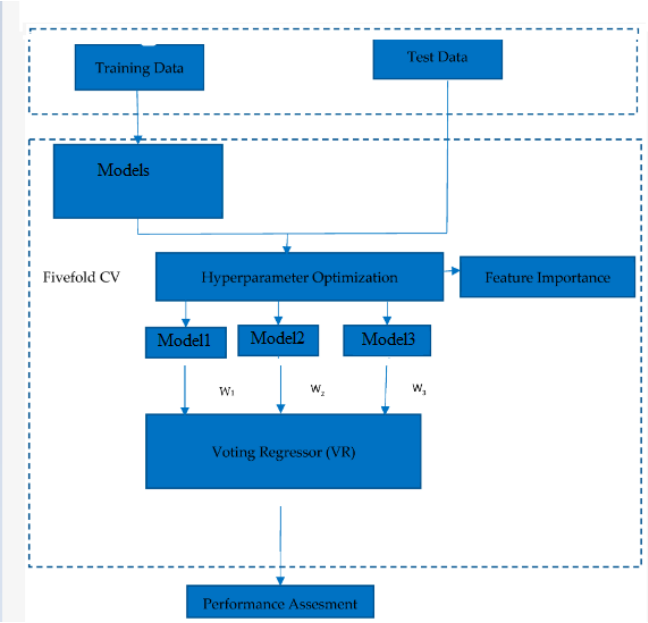
$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \quad (2)$$

Smaller RMSE values indicate better predictive accuracy. RMSE is preferred for its sensitivity to the magnitude of errors and its ability to penalize larger errors more heavily, offering a comprehensive evaluation of a model's performance. In a scientific paper, reporting RMSE provides readers with a clear understanding of the predictive capabilities of the model and its room for improvement.

3.3. Ensemble Voting

This is the main method that is planned to be implemented. Existing methods will be extended by the use of this to achieve the lowest error rate. The Flowchart of the plan looks like this:

Figure 2. Chart of Ensemble Voting



Ensemble Voting is a widely utilized technique in machine learning, it involves aggregating predictions from multiple models to improve overall performance. Different base models often employing different algorithms or model architectures contribute predictions and a voting mechanism combines their outputs. Common approaches include majority voting for classification tasks and averaging for regression(which is the case for the project). Ensemble Voting leverages the collective intelligence of models and improving generalization, therefore achieving robust and more accurate predictions by the sacrifice of computational time

4. Experiment Results

Data has been splitted into train and test sets of 80/100 and 20/100. Total of 4 experiment had been made. The main objective and approach here is to get as diverse and as

accurate as possible for the Ensemble Voting that is going to be made later. Different models of different algorithms had been tested to find which would work great and which wouldn't. Further improvements are going to be made

4.1. Linear Regression

Linear Regressor is applied and an error rate of 0.048 is obtained at lowest, also the error rates are checked for the most corrolated and first 2 most corrolated columns to see how much computational time can be neglected to gain the result.. Also polynomial regression is used on first most corrolated colmumns to fine tune it. The reason for this low error rate is likely due to the effective capture of underlying patterns and relationships in the data by employing linear regression on the most correlated columns.

4.2. Random Forest Regression

Random Forest is another machine learning method that operates by constructing multiple decision trees. The final decision is made based on the majority of the trees and is chosen by the random forest. The advantages of using a random forest algorithm are; it reduces the overfitting, gives higher accuracy than 1 decision tree since it reduces the variance by ensembling different decision trees, and runs efficiently on large datasets. In experiments, random forest resulted in 0.151 RMSE normally and 0.126 RMSE with grid search with
'nestimators': [50, 100],
'maxdepth': [10, 20],
'minsamplessplit': [2, 5],
'minsamplesleaf': [1, 2]

No further tests are made on this because of computational time. The reason it has a higher error rate is that as the complexity increases, there is a risk of overfitting in RF, especially if the model is not properly tuned. Also the while a grid search was performed on hyperparameters the chosen hyperparameter values might not be optimal for the specific dataset.

4.3. 1D-CNN Regression

In the experimentation with 1D-CNN (One-Dimensional Convolutional Neural Network), two convulation layers and one pooling layer had been used with a window size of 24. the following results were obtained over 10 epochs:

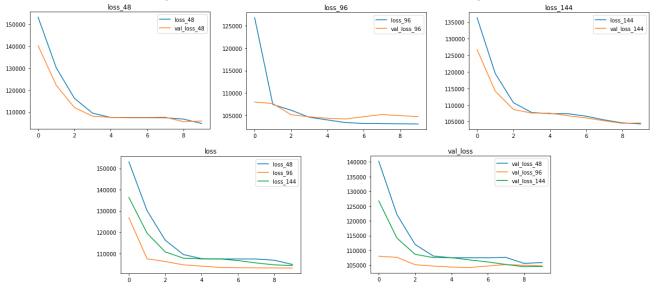
- Epoch 1, Loss: 0.0005654108244925737, Validation Loss: 0.015202730106917762
- Epoch 2, Loss: 0.0006166600505821407, Validation Loss: 0.01511872393708127
- Epoch 3, Loss: 0.0004682322032749653, Validation Loss: 0.014384918490570909
- Epoch 4, Loss: 0.000824643939267844, Validation Loss: 0.01292010450900632
- Epoch 5, Loss: 0.0004908135160803795, Validation Loss: 0.01380245900771407
- Epoch 6, Loss: 0.00023989556939341128, Validation Loss: 0.013743289268261892
- Epoch 7, Loss: 0.00035523934639059007, Validation Loss: 0.01318761679816896
- Epoch 8, Loss: 0.00046317523811012506, Validation Loss: 0.013201200424174888
- Epoch 9, Loss: 0.00023900083033367991, Validation Loss: 0.01317400226777206
- Epoch 10, Loss: 0.00026675552362576127, Validation Loss: 0.013964214367803774

The loss values for both training and validation sets are provided for each epoch. These values indicate how well the 1D-CNN model is performing on the given data. It seems that the model is learning over the epochs, as the training loss is decreasing, and the validation loss is also relatively low, suggesting good generalization to unseen data. Which at the end gives $\sqrt{0.0139} = 0.117$ RMSE to work. 1D-CNNs are well-suited for tasks involving sequential data, such as time series or signal data. The model's ability to learn hierarchical features and capture patterns within the specified window size (24 in this case) allows it to effectively recognize and exploit the sequential nature of the input data.

4.4. Long Short-Term Memory(LSTM)

Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) architecture that is well-suited for sequence prediction problems. In the context of regression, LSTM can be used to predict numerical values based on sequential input data. In your experimentation with LSTM, you likely used a network configuration suitable for regression tasks. Three different time intervals had been used to make the model. Periods of 1 days, 2 days and 3 days. Among three models these results had been obtained:

Figure 3. Chart of Ensemble Voting



The least loss gotten is from the model with 2 day periods:

- 545/545 [==============================] - 6s 10ms/step - loss: 0.008466

It can be said that from the least loss of best sequence model that it gives $\sqrt{0.008466} = 0.091$ RMSE to work with

The reason for this low error rate is the fact that LSTMs are specifically designed to address the vanishing gradient problem in traditional RNNs, allowing them to capture long-range dependencies in sequential data. The memory cells and gating mechanisms in LSTMs enable the model to retain information over extended periods, facilitating better learning of sequential patterns.

Table 1. RMSE results of every model tested

METHOD	RMSE	VIABLE?
LINEAR	0.048	+
RANDOM FOREST	0.126	?
1D-CNN	0.117	+
LSTM	0.091	+

References

Chiteka, K. and Enweremadu, C. Prediction of global horizontal solar irradiance in zimbabwe using artificial neural networks. *Journal of Cleaner Production*, 135:701–711, 2016. ISSN 0959-6526. doi: https://doi.org/10.1016/j.jclepro.2016.06.128. URL https://www.sciencedirect.com/science/article/pii/S0959652616308058.

Feng, Y., Cui, N., Chen, Y., Gong, D., and Hu, X. Development of data-driven models for prediction of daily global horizontal irradiance in northwest china. *Journal of Cleaner Production*, 223:136–146, 2019. ISSN 0959-6526. doi: https://doi.org/10.1016/j.jclepro.2019.03.091. URL https://www.sciencedirect.com/science/article/pii/S0959652619307826.

Pedro, H. T. and Coimbra, C. F. Nearest-neighbor methodology for prediction of intra-hour global horizontal and direct normal irradiances. *Renewable Energy*, 80:770–782, 2015. ISSN 0960-1481. doi: https://doi.org/10.1016/j.renene.2015.02.061. URL https://www.sciencedirect.com/science/article/pii/S0960148115001792.

Yang, H., Wang, L., Huang, C., and Luo, X. 3d-cnn-based sky image feature extraction for short-term global horizontal irradiance forecasting. *Water*, 13(13), 2021. ISSN 2073-4441. doi: 10.3390/w13131773. URL https://www.mdpi.com/2073-4441/13/13/1773.

Zang, H., Liu, L., Sun, L., Cheng, L., Wei, Z., and Sun, G. Short-term global horizontal irradiance forecasting based on a hybrid cnn-lstm model with spatiotemporal correlations. *Renewable Energy*, 160:26–41, 2020. ISSN 0960-1481. doi: https://doi.org/10.1016/j.renene.2020.05.150. URL https://www.sciencedirect.com/science/article/pii/S0960148120308557.