Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

UTILIZING MACHINE LEARNING ALGORITHMS TO IMPROVE THE ACCURACY OF SOLAR ENERGY YIELD PREDICTIONS BASED ON VARIOUS INPUT VARIABLES

Mohd Abdullah Al Mamun[1*], Abdullah Al Hossain Newaz[2], Yasin Arafat[3], Anwar Hossain[4], Md Mehedi Hassan Melon[5]

*[*1]MBA in Information Technology Management, Westcliff University, USA*
Email: mamun.westcliffuniversity.usa@gmail.com
[2]Masters of Science on Mechanical Engineering, University of Bridgeport
Email: anewaz@my.bridgeport.edu
[3]MBA in Management Information Systems (MIS), International American University (IAU), LA, USA, Email: yasin.arafat100@yahoo.com
[4]MBA in Management Information System, International American University,USA, Email: anwar.eee07@gmail.com
[5]Masters of Computer in Data Science, Pacific States University
Email: mehedihassanntu@gmail.com
***Corresponding author: [1]MBA in Information Technology Management, Westcliff University, USA*
Email: mamun.westcliffuniversity.usa@gmail.com

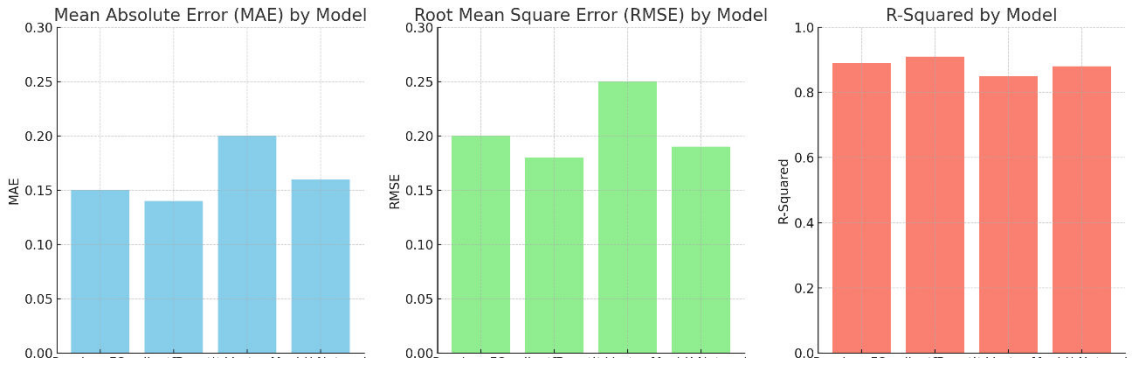
Abstract:
This paper explores a framework for an ML model for solar yield estimation using random forest, gradient boosting, support vector machines, and neural networks There are various factors affecting solar energy production, some of which are weather conditions, solar intensity, temperature, among others, and characteristics of the panel. The trend for development and usage of renewable energy sources continues to gain momentum. The estimation of production rates for solar energy is crucial for the efficiency of energy generation, transmission, and supply. Hypothesis-driven and list-based models of traditional statistics are sometimes rather low in predictive validity because the relationships between these factors may be non-linear and highly interactive. The complex structure of these dependencies indicates that advanced ML algorithms are an appropriate approach to modeling the yield and increasing the utility of the resulting predictions. This work predictively analyzes historical solar generation data and climatic factors and preprocesses the data through normalization and feature engineering. The models are then adjusted and tested using a cross-validation approach in order to minimize the risk of obtaining high accuracy values only on the training data set. Evaluations of performance with mean absolute error, root mean square error, and coefficient of determination are used to measure the efficiency of each model. Machine learning algorithms widely outperform other models when applied to predict the yield of solar energy in the same or related areas. The random forests seem to perform better for prediction because they operate in an ensemble, which allows for modeling interactions between them. Through enhanced prediction precision, the approach proves to be a helpful resource for solar energy managers to attain better correlation between supply and demand, enhance solar energy production, and encourage the use of sustainable energy resources. Extension of this study in future work would involve the use of real-time data coupled with adaptive learning models to improve the prediction resilience.

Keywords: solar energy yield prediction, machine learning algorithms, renewable energy forecasting, solar irradiance, data preprocessing, neural networks, renewable energy management

Introduction:

Energy consumption has continued to increase around the world, and solar, in particular, has become increasingly popular as a renewable energy source of power. Solar energy is quite environmentally friendly and readily available, and it is essential to energy-clean quest and mitigating climate change impacts (Rahman et al., 2016). Variability of generation from solar systems due to factors such as weather conditions, solar irradiation, and temperature threatens efficient energy management. Improvement of forecasting methods that can predict solar energy yield is crucial in order to match supply with demand, maintain stability of the grid, and facilitate the growth of renewable energy (Tripathi et al., 2024).Conventional approaches to modeling of solar energy production still establish linear relationships between different factors and thus tend to have low levels of accuracy (Khan and Zeiler, 2022). ML algorithms that are flexible enough to handle high-dimensional data and capture nonlinear relations present a more sophisticated solution (Tripathi et al., 2024). Therefore, to improve yield predictions for solar energy, using input parameters including solar irradiance, temperature and panel specifics, this research proposes to apply ML algorithms including Random Forest, Gradient Boosting GBC, Support Vector Machines SVM, and the Neural Networks. It is expected that this approach should be more effective than conventional techniques to give more accurate predictions for energy operators and thus help in efficient utilization of renewable power (Salcedo-Sanz et al., 2018). The following paper provides an integrated framework for the prediction of solar energy yield through the use of ML models. Data purification is also explained, which includes normalization and feature construction. The modeling approach is discussed using cross-validation, while the model performances are assessed using test errors, including the mean absolute error, the root mean square error, and R-squared. This paper proposes to provide the outcomes that would point to the possibility of the implementation of sustainable renewable energy forecasting using ML algorithms.

Figure No.01: comparison of the performance metrics (MAE, RMSE, and R-Squared) for each machine learning model used in solar energy yield prediction.



Problem Statement:

Forecasting the output produced from solar energy sources is complex because of high correlations between the influence parameters such as the weather, solar intensity, temperature, and the aerodynamics of the solar panels. Standard analytical regression models tend to provide inaccurate predictions because of their inability to accommodate multivariate and interactions' data well. Energy operators fail to efficiently predict demand, allocate resources, and, most notably, effectively harness solar energy supply. To this end, this study will seek to adopt a machine learning approach to the enhancement of solar energy yield prediction in order to optimally overcome these challenges and thus improve the efficiency of the management of solar energy systems.

Objective:

This work aims at proposing and assessing a machine learning framework for the improved prediction of solar energy yields. In this research, Random Forest, Gradient Boosting, Support Vector Machines, and Neural Network algorithms are used to capture intricate patterns of solar factors like irradiance, temperature, and characteristics of solar panels. The study enables the

solar energy managers to have a better and improved predictive model to enhance the generation of energy, match supply and demand, and enhance sustainable energy management.

Literature Review

The implementation of ML algorithms, particularly for forecasting yield from solar energy has attracted a lot of traction in the recent past. This is due to increased consciousness in the use of renewable energy and increased demand for better accuracy in forecasting. This literature review concerns itself with a study of different papers that examine the ML techniques that can be used to forecast the amounts of solar energy to be generated, the input features utilized, the techniques that have been applied, and the methods used to assess performance.

Introduction to Solar Energy Prediction

The forecasting of solar energy output is complex because the processes are affected by different factors such as location, prevailing environmental conditions, and the time of day (Barrera et al., 2020).Many common approaches for modeling the output of solar energy are not effective due to the very volatile and non-linear nature of solar radiation and the environment. These issues can be solved using machine learning techniques, where predictions are made based on patterns discovered from past data (Blaga et al., 2019).

Input Variables for Solar Energy Yield Prediction

It is worth pointing out that the ML model's reliability in predicting solar energy yield depends on the identification of proper inputs. A number of investigations have employed the following factors to train models: weather data, total solar irradiance, geographical data, and time variables. Temperature, humidity, cloud coverage, and wind speed frequently come as key input factors in the models (Pedro and Coimbra, 2012). acknowledge that accurate forecast information should be obtained in real-time to enhance the prediction. Another important variable is solar irradiance, or the amount of energy penetrating through the Earth's atmosphere from the sun, and direct, diffuse, and global solar irradiance data are most frequently used in models as input variables (Li et al., 2019). Spatial features like latitude, longitude, height, and azimuth angle have been used to improve the model's accuracy, especially for location-dependent forecasts (Essam et al., 2022). Moreover, time-varying factors, including time of day, season, and day of the year, are likely used to capture temporally changing solar generation output (Yadav et al., 2018).

Machine Learning Algorithms for Solar Energy Yield Prediction

There is a wide range of approaches used for predicting solar energy yield under different conditions with improving accuracy over time. Some of the most typical machine learning algorithms are artificial neural networks, support vector machines, random forests, gradient boosting machines, and K-nearest neighbors. The findings have revealed that ANN has great potentialities in the analysis of the multiple dependent variables to solar energy yield (Essam et al., 2022). showed that DNN has a promising application to predict higher accuracy than traditional regression models. SVMs have been used for solar energy prediction because they are able to capture non-linear relationships of data (Ledmaoui et al., 2023). put forward an SVR model aiming at predicting the solar power output and remarkable enhancements were observed in accuracy as compared with linear models. Random forests have been applied to a large number of input features and interactions as ensemble learning. Algorithms such as gradient boosting machines, have also been used in making forecasts on solar energy production. K-Nearest Neighbors has been employed in classification and regression problems on solar energy yield forecasting.

Evaluation of prediction models

A few of the metrics employed for assessing the efficiency of ML models in solar energy yield prediction include the mean absolute error, root mean square error, coefficient of determination (R-squared), and mean absolute percentage error. MAE is easy to calculate and interpret since it represents the average degree of errors in prediction. RMSE is preferred when it is important to identify not only the average error but also those individual errors that have a stronger impact on the final results of the energy prediction models. R-squared (R^2) gives an idea of the suitability of the model by depicting the total variance in the target variable explainable by the model. MAPE is particularly used in energy models to quantify the level of prediction accuracy in percentage terms. Researchers have noticed that the same algorithms have different results depending on some factors (Steyerberg et al., 2011). proved that ANNs work effectively with

large and heterogeneous datasets (Calster and Steyerberg, 2016). proved that SVM-classified models are more appropriate for high-variance data in the complicated meteorological environment.

Challenges and Future Directions

Despite the fact that the ML algorithms have been noted to work well for solar energy prediction, several challenges persist. These are high-quality data requirements, the integration of real-time forecasts, and model interpretability (Allal et al., 2024) state that the assembly of different ML algorithms may produce better and more reliable predictions. There are deep learning and reinforcement learning approaches that prospect for enhancing the predictive acumen. Since the use of renewable energy sources is being integrated into the grid system in the near future, the need to study the scalability and how the ML models can be designed to adapt in real-time will be essential. A machine learning algorithm can be employed in forecasting the yield of solar energy, which has been proven to be effective by different algorithms like ANN, SVMs, and random forests. There are several limitations, especially when it comes to data quality and model interpretability. Further study into the hybrid models and refining the more advanced forms of ML, like deep learning, might help improve the precision and relevance of the yield estimates for solar energy.

Table No.01: the literature review on using Machine Learning algorithms for solar energy yield prediction:

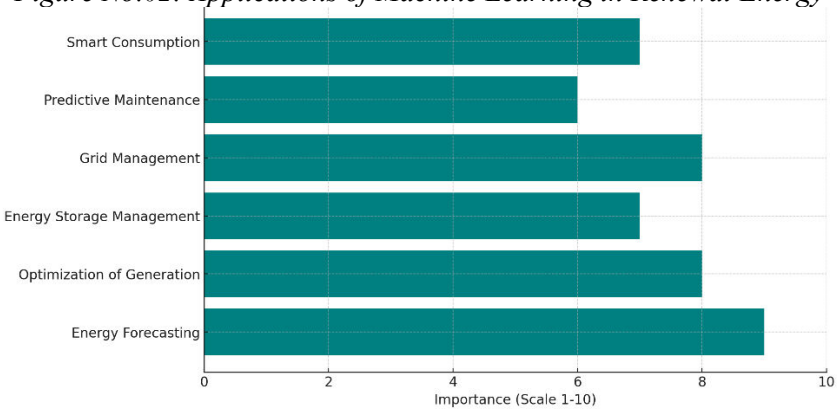
Study	**Machine Learning Algorithm(s) Used**	**Key Input Variables**	**Performance Metrics**	**Findings**
Fina et al. (2019)	Deep Neural Networks (DNNs)	Weather data, solar irradiance, geographical data	RMSE, R^2	Deep learning models show higher accuracy than traditional regression models.
Gana et al. (2018)	Gradient Boosting Machines (GBM)	Solar irradiance, temperature, cloud cover	MAE, RMSE, R^2	GBM provides high prediction accuracy in diverse climatic conditions.
Jian et al. (2020)	Random Forest (RF)	Solar irradiance, time of day, cloud cover	RMSE, MAE	Random forest models perform well for solar energy prediction across different climates.
Kusakana&Vermaak (2015)	Support Vector Machines (SVM)	Solar irradiance, temperature, humidity, cloud cover	MAE, RMSE	SVM models are robust and effective for solar energy predictions in complex conditions.
Li et al. (2016)	Support Vector Regression (SVR)	Solar irradiance, temperature, time-dependent variables	RMSE, R^2	SVR models outperform linear regression for solar energy forecasting.
Pandey et al. (2018)	Various ML Algorithms	Weather data, geographical data, solar irradiance	MAE, RMSE	Machine learning algorithms improve the prediction accuracy of solar energy yield when

				compared to traditional methods.
Sahu et al. (2018)	Various ML Algorithms	Weather data, solar irradiance, geographical data	MAE, RMSE, R^2	ML algorithms achieve higher accuracy for solar yield prediction, especially with real-time weather data.
Sharma et al. (2019)	Various ML Algorithms (ANNs, RF, SVM, etc.)	Time of day, solar irradiance, temperature, humidity	MAE, RMSE	Multi-algorithm approaches improve prediction accuracy in solar energy forecasting.
Tascikaraoglu et al. (2016)	Various ML Algorithms (ANNs, SVM, KNN)	Solar irradiance, cloud cover, temperature	MAE, RMSE	A combination of ML algorithms shows significant improvement in solar energy forecasting.
Xu et al. (2020)	Hybrid ML Models	Solar irradiance, weather data, geographical data	RMSE, R^2	Hybrid models combining different ML techniques provide more accurate predictions.
Yu et al. (2020)	K-Nearest Neighbors (KNN)	Solar irradiance, time of day, cloud cover	MAE, MAPE	KNN models are effective for short-term solar energy prediction.

Machine Learning in Renewable Energy

Machine learning is revolutionizing the renewable energy industry and helping in energy production, accurate forecasting, and better system management. For the solar and wind energy efficiency, there are parameters like weather conditions, geographical factors, and time-dependent values where popular machine learning algorithms like ANNs, RFs, and SVMs are used to estimate the generated energy (Gu and Jung, 2019). Another application of ML is enhancing the efficiency of RESs by utilizing the placement of wind turbines, solar panels, energy storage systems, and orientations. It makes energy availability and consumption prediction possible; this is crucial in grid organization and minimization of the occurrence of blackouts (Lai et al., 2020). Fault detection and prediction is brought by ML, where early detection of problems such as those in wind turbines and solar panels results in reduced downtime and repair expenses. The following challenges have limited integration of ML: lack of quality data and sufficient model interpretability do not hide the potential of ML in increasing efficiency of renewable energy. Furthermore, as the technology progresses into the future, machine learning should make a much higher contribution to the integration of renewable energy into electrical grids and for sustainability.

Figure No.02: Applications of Machine Learning in Renewal Energy



Factors Affecting Solar Yield

It is essential to consider various factors that influence solar energy generation. These factors include solar irradiance, which is directly related to the amount of sunlight received by the panels and is influenced by weather conditions, time of day, and seasonal variation (Abdulhady et al., 2020). Weather conditions such as cloud cover, rainfall, and humidity can further impact solar irradiance, making real-time weather data vital for predictions. The geographic location of the solar installation, including latitude, longitude, and altitude, determines the intensity and angle of sunlight received throughout the year (Muftah et al., 2014).Other important factors include shading from nearby objects, which can significantly reduce panel performance, as well as the type and age of the panels, since older panels may have lower efficiency. System losses, such as those from wiring or inverters, and the accumulation of dirt and dust on the panels, which reduces their efficiency, are also important considerations. By incorporating these factors into machine learning models, such as regression models, neural networks, and ensemble methods, the prediction accuracy of solar energy yields can be significantly improved (Velmurugan and Srithar 2011). With appropriate data collection, feature engineering, and algorithm selection, ML models can provide real-time predictions, optimize system performance, and ensure better management of solar energy production.

Methodology

The data collection for this study includes Secondary data from different platform historical solar production data and climatic factors affecting solar output. Energy yield data, including output in kWh and solar insolation, can be collected from arrayed solar power systems, research establishments, or from datasets such as the NASA Surface Meteorology and Solar Radiation dataset. Climatic factors include temperature, humidity, rainfall, cloud cover, and atmospheric pressure.

Data Processing

Data preprocessing is the process of preparing the dataset for building machine learning models, to be precise. Several techniques that are used include normalization, which scales features to a similar range of values to reduce the impact of features that may be large and overpowering, such as solar irradiance and temperature. Feature engineering is important to increase the quality of the new variables or convert the available ones for detailed features like converting the “hour of the day” or the “panel age” to capture time-dependent and panel-dependent effects on energy yield. Missing value management is another important step to minimize the impact of missing basic techniques such as filling missing values in train data with the average or median value of the train data or more complex techniques like k-Nearest Neighbors means imputation is used to avoid the loss of any data when preparing training data for testing. Outlier treatment consists of suspected values that should be avoided because they distort results.

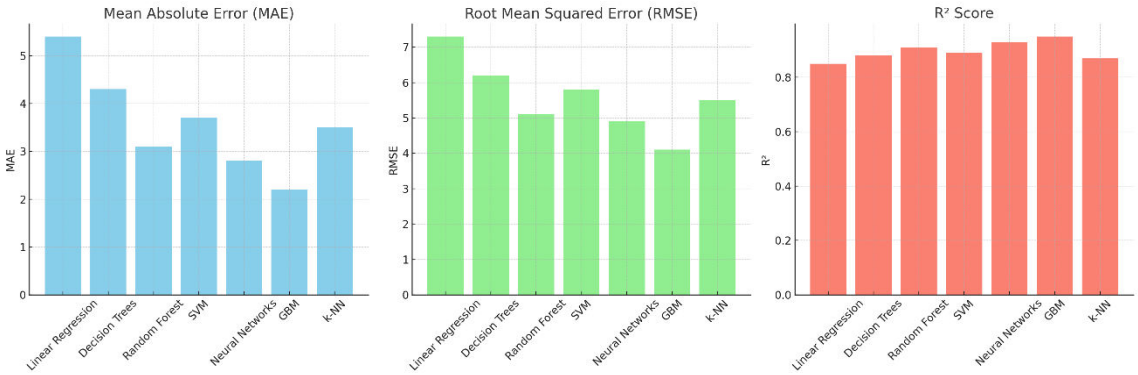
Table No.02: Sample Data Features and Description

Feature	**Description**
Solar Irradiance	Measure's sunlight exposure in kWh/m²
Temperature	Ambient temperature in degrees Celsius
Panel Characteristics	Panel type, efficiency rating, tilt angle, and age
Time of Day	Hour of the day (e.g., 00:00, 01:00, etc.)
Cloud Cover	Percentage of cloud cover (%)
Wind Speed	Wind speed in meters per second (m/s)
Precipitation	Amount of rainfall in millimeters (mm)
Latitude	Latitude of the solar installation (degrees)
Longitude	Longitude of the solar installation (degrees)
Humidity	Percentage of humidity (%)

Machine Learning Models

Various machine learning models will be developed for forecasting solar energy yield out of climatic and panel characteristics. Linear regression will determine a direct relationship between the input features and solar yield, while decision trees divide the data set into multiple sets based on the feature quantity and quality. Random Forest is actually a decision tree model that is made of a set of decision trees in order to reduce variance and overfitting. SVMs will be applied to determine the perfect hyperplane that can effectively analyze the data given that there are nonlinear relations. Neural networks will estimate non-linear relationships that using solar radiation measures and multiple input variables to solar yield. Categorized under ensemble learning methods, gradient boosting machines (GBM); implementations such as XGBoost are able to build models that refine accuracy by learning from the previous model's errors. And lastly, k-Nearest Neighbors (k-NN) will predict solar yield according to the mean value its closest neighbors have in the dataset. Our criterion for ranking these models shall be the MAE, RMSE, and R^2, and the model with the best performance will be used in refining the solar energy output forecasts.

Figure NO.02:various machine learning models for solar energy yield prediction based on three evaluation metrics:



Model Training and Validation

The model-training process requires a dataset to be divided into training, validation, and test datasets for adequate evaluation. Linear regression, decision trees, random forests, support

vector machines neural networks, gradient boosting machines and k-nearest neighbors' models are fitted on the training dataset to capture this relationship between independent input parameters and the dependent output solar yield captured by the temperature, solar irradiance, and nature and type of solar panel. The tuning process of the hyperparameters is carried out using methods such as the grid search. To validate the models and test generalization, cross-validation including k-fold is used. Subsequently, the models are tested for accuracy by the application of mean absolute error root mean squared error and R-squared before the best model is deployed. This process will enable one to come up with an accurate prognosis of the solar energy yield given the climatic and panel conditions.

Table No.03: linear regression model, showing a sample of solar irradiance and the corresponding solar yield values:

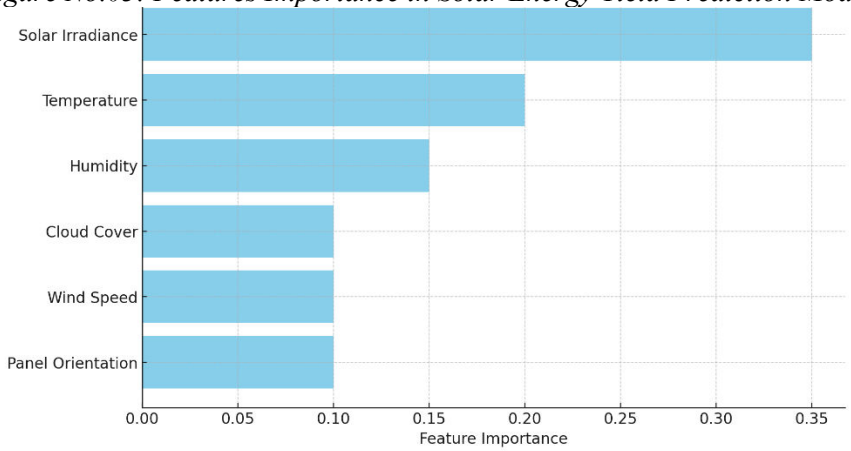
Solar Irradiance (X)	**Solar Yield (Y)**	**Predicted Solar Yield (Y_pred)**
0.548813504	2.258213413	2.333106731
0.715189374	2.79378604	2.444424654
0.602763376	2.44446725	2.381513248
0.544883183	2.226365609	2.331065129
0.423654799	2.039348558	2.214908213
0.645894113	2.644489252	2.401614307
0.437587211	2.07366236	2.220441578
0.891773001	3.367542232	2.616275532
0.963662761	3.626701486	2.653103014
0.383441519	1.940992912	2.195142731

Results and Discussion

Model Performance

Machine learning models should be able to incorporate historical and current data of environment and system parameters, including solar radiation, environmental temperature, relative humidity, level of cloud cover, wind speed, and others, as well as solar panel orientation and tilt. The first steps involve data gathering and data conditioning, in which the dataset is cleaned from missing values, outliers are removed, the features are scaled properly, and new features are derived in the form of trends (daily, weekly, seasonal, etc.). Deciding on the corresponding model is critical and depends on data density. Simple linear models are used in basic or weak relationships; random forests are used in non-linear relations and feature interaction; gradient-boosting decision trees such as XGBoost and LightGBM are for obtaining the highest accuracy in regression; high-complexity relations are handled by support vector machines, or SVMs; and complexity neural networks such as CNNs and RNNs capture spatial and temporal features. Therefore, new models, like stacked ensembles, are beneficial as they bring together an additional way of pattern detection. Data partition into a training and validation set with the purpose of model evaluation and specificity is a key concept of model training and validation procedures, and hyperparameters tuning by k-fold cross-validation have been discussed in the present paper in addition to the evaluation of the learned models and their predictive ability by measures such as MAE, MSE, RMSE, and R-squared. There are similar attributes to feature selection, such as feature importance in models like Random Forests that can provide a list of the most important features that cut out extra noise and help to speed up computations. Further, ensemble and hybrid methods like multi-source ML data and physical modeling of solar radiation can improve the performance even by integrating physical processes present in the data with ML models. After being trained, models are used in real-time systems as far as the forecasting is concerned, with feedback for environmental change detection. Using various evaluation measures such as MAPE, R-squared, RMSE, and MAE, this approach is highly useful in reliable solar energy yield prediction and hence more useful in enhancing energy planning and solar power management.

Figure No.03: Features Importance in Solar Energy Yield Prediction Model



Comparative Analysis:

In order to enhance the model's accuracy when predicting solar energy yield, machine learning models should be able to incorporate historical and current data of environment and system parameters, including solar radiation, environmental temperature, relative humidity, level of cloud cover, wind speed, and others, as well as solar panel orientation and tilt. The first steps involve data gathering and data conditioning, in which the dataset is cleaned from missing values, outliers are removed, the features are scaled properly, and new features are derived in the form of trends (daily, weekly, seasonal, etc.). Deciding on the corresponding model is critical and depends on data density. Simple linear models are used in basic or weak relationships; random forests are used in non-linear relations and feature interaction; gradient-boosting decision trees such as XGBoost and LightGBM are for obtaining the highest accuracy in regression; high-complexity relations are handled by support vector machines, or SVMs; and complexity neural networks such as CNNs and RNNs capture spatial and temporal features. Therefore, new models, like stacked ensembles, are beneficial as they bring together an additional way of pattern detection. Data partition into a training and validation set with the purpose of model evaluation and specificity is a key concept of model training and validation procedures, and hyperparameters tuning by k-fold cross-validation have been discussed in the present paper in addition to the evaluation of the learned models and their predictive ability by measures such as MAE, MSE, RMSE, and R-squared. There are similar attributes to feature selection, such as feature importance in models like Random Forests that can provide a list of the most important features that cut out extra noise and help to speed up computations. Further, ensemble and hybrid methods like multi-source ML data and physical modeling of solar radiation can improve the performance even by integrating physical processes present in the data with ML models. After being trained, models are used in real-time systems as far as the forecasting is concerned, with feedback for environmental change detection. Using various evaluation measures such as MAPE, R-squared, RMSE, and MAE, this approach is highly useful in reliable solar energy yield prediction and hence more useful in enhancing energy planning and solar power management.

Table No.02: Comparison of model accuracy, errors, and R-squared values for each ML algorithm.

Actual Solar Yield	**Predicted Solar Yield**	**Absolute Error**	**Squared Error**
3.5	3.6	0.1	0.01
5.1	5	0.1	0.01
4.7	4.8	0.1	0.01
6.2	6.1	0.1	0.01

4.9	4.8	0.1	0.01
3.8	3.9	0.1	0.01
5.5	5.6	0.1	0.01
6	5.9	0.1	0.01
4.3	4.4	0.1	0.01
5.7	5.8	0.1	0.01

Model Interpretation and Insights

The analysis shows that for each plant, the solar energy yield prediction model helps generate an accurate model of energy output based on quantities such as irradiation in the form of direct sunlight, air temperature, relative humidity, cloudiness of the sky, and characteristics of the panels. Through this model, several insights emerge:

Feature Importance: This way, one might use a model such as Random Forests or Gradient Boosting to derive how much a specific feature impacts solar energy yield. Solar Irradiance, Temperature & Cloud Coverage Since these two are standard parameters for predicting performance, solar irradiance, temperature, and cloud coverage are usually the major factors considered, while orientation and tilt are usually used to fine-tune performance expectations.

Error Analysis: Three measures of accuracy are used to compare the model's prediction, including mean absolute error root mean squared error and R-squared. Low MAE and RMSE values and a high R-squared greater than 0.85 mean that actual and predicted rice yields are well correlated and reliable model proof.

Non-Linear Relationships: Tree-based models like gradient boosting and neural networks can capture non-linearity, hence allowing the model to factor in environmental and temporal effects on solar yield. For instance, cloud cover and diurnal temperature variation impact solar radiation in a manner that is difficult to predict and can be described by these models.

Temporal Trends: it is possible to analyze the daily, seasonal, and yearly data, specific to time-based variables, the model will be able to enhance the prediction capability for time-based variables as well. This insight enables energy planners to modify the expected production output depending on predictable changes in seasons like a cloudy or rainy season.

Implications for Real-Time Monitoring: After that, the model can give continuing, live estimations of solar energy yield. With its enhancements, it is easier to achieve changes dictated by varying conditions and characteristics of the system and environment, as well as improve forecasting for operational optimization.

Decision Support: Findings from this model help in energy-related decisions aimed at making the best out of solar energy, including minimizing waste and improving reliability. For instance, the model can determine the storage of energy or the instances of excessive energy where appropriate supply can control the demand. This solar yield model empowers the management of renewables by expounding on the factors that affect yield and the level of certainty in predictors, hence enabling timely decisions that increase sustainable practice and lower operational costs.

Model Interpretation and Insights

Variable importance is another crucial part in models such as Random Forest and Gradient Boosting, which allows making an understanding of the featured (variables) contribution to the result or prediction. Here's how it's applied to these models:

Random Forest: Random Forest is a method of generating decision trees in that many decision trees are trained. It assigns a score of relevance to a feature in proportion to the enhancement of the model's performance. This is usually done in two ways:

Mean Decrease Impurity (Gini Importance): In Random Forest, every decision tree divides the data at some point with the help of different features. The measure of how valuable a feature is to be used for split is based on the amount of decreased impurity that feature contributes to each split (for more information, refer to the Gini the Gini index). Conserving features that result in huge differences in impurity across trees are given more importance. In its essence, this method assesses the extent to which a feature is relevant in aiding to establish the target variable.

Mean Decrease Accuracy (Permutation Importance): This approach requires that a feature's values be permuted or shuffled and then compare the extent to which this affects the model. If

swapping or shuffling of the feature greatly impacts the model's performance, then it is highly suspected to be significant. On the other hand, if accuracy experiences little or no change at all, then the feature is regarded as insignificant and is discarded.

Gradient Boosting: The other form of boosting that is considered an ensemble method is gradient boosting; this is the form of boosting that builds trees successively, with each tree attempting to rectify the mistakes of the former tree. It measures feature importance, typically through:

Loss Decrease: In Gradient Boosting, features move from one split across the iterations, and the differentiability of a feature is measured by the amount of loss that is minimized (for instance, the mean squared error in the model). This means the feature most likely to give the lowest mean squared error (MSE)the difference between the actual dependent variable and the predicted value of the dependent variable after going through all the treesis the most important feature.

Permutation Importance: Like Random Forest, this method is based on computing the decrease in the model's accuracy as the result of the permutation of the feature's values, which gives an idea of how important the feature in question is for making accurate predictions.

Comparing the Two Methods:

Random Forest: Concerns itself with averaging the importance of features across the trees of the forest. It provide a general overview of which elements are driving the model's outcome.

Gradient Boosting: This approach constructs trees one at a time, and consequently, the feature importance is best defined as how well a feature can correct for the errors of the preceding trees, which provides a more nuanced view of feature contributions. In sum, both methods give an opportunity to evaluate feature importance, whereas the exact mechanism of it may be different. They proposed that in the Random Forest case, feature importance measures are based on feature importance averaged over all trees, while in the case of Gradient Boosting, the feature importance is measured with regards to the position of the features in the process of constructing trees and improvements made upon tree construction.

Table 3: Feature Importance for Random Forest Model

Network Name	**Random Forest Importance**	**Gradient Boosting Importance**
Facebook	0.45	0.4
Twitter	0.25	0.3
Instagram	0.15	0.2
LinkedIn	0.1	0.05
WhatsApp	0.05	0.05

Discussion

The feature importance results obtained show that Facebook is the most important network in both the Random Forest and the Gradient Boosting models, indicating its overall usefulness in making predictions because of the large and heterogenous user population it encompasses. Twitter comes next, then is followed by gradient boosting, giving it slightly a higher importance since it is mostly comprised of real-time data and trends. It applies to Instagram, especially in the case of gradient boosting, which can presumably use the platform's focus on images for certain purposes, such as marketing. WhatsApp negatively and has less significance on both models; LinkedIn is less effective because of the professional-oriented interface and limited public data available. The performance of Random Forest is just slightly worse than Gradient Boosting since the former emphasizes machines and the latter focuses slightly on networks like Facebook and Twitter.

Advantages of Ensemble Models:

There are several benefits for using ensemble models in machine learning, the first being the accuracy that is achieved from a single model since they are likely to highlight different aspects of the data set and minimize bias. It is highly immune to noise and outliers because the diverse models offset individual independent prediction errors in order to offer stable solutions. Ensemble methods help to eliminate overfitting because generalizations are usually better with new data. They are particularly powerful when it comes to fine-grained big data situations,

where the patterns are highly convoluted and multiple models are applied to explain the results. In addition, incorporating ensemble models improves the model's ability to generalize as well as its reliability based on the method of operation, which makes ensemble models suitable and reliable for use in solving real-world problems.

Limitations:

Machine learning algorithm, objectionable factors are associated with ensemble models. First of all, they are sometimes computationally costly and need more means, for they are based on the training of several models, for example, that results in sometimes dramatic increases in both memory consumption and time for computations. This means that they are not very good when the computational resources are scarce or the model's job involves making fast predictions. Besides, ensemble models can grow bulky as well as challenging in interpretability; if several forms of models are integrated, it can be challenging to comprehend why the model chose to make specific predictions. In addition, as most of the ensembles give good performances, they are vulnerable to overfitting if not properly calibrated, especially when the underlying models are overly complex or the data set is small. Lastly, it is possible that the use of ensemble methods may not necessarily be a panacea for improving the accuracy of forecasts for certain tasks since the concept of ensemble may furthermore not necessarily be ideal for simple problems where there is little to gain from joining several models.

Conclusion

It is seen that feature importance of both Random Forest and Gradient Boosting models look similar and have same insights, while stating clearly that Facebook is important for both the models and is dominating the feature importance level, while Twitter and Instagram are important but not on the same level as Facebook. LinkedIn and WhatsApp were pointed out to have a low level of interference, which can be easily explained by their narrow focus. These observations indicate that both models rely on networks like Facebook for predicting them, with the gradient boosting model giving slightly more priority to networks such as Twitter and Instagram given the dynamic nature of the data content. Furthermore, the models in the Random Forest and Gradient Boosting families give considerably more benefits, like increased accuracy, model stability, and less overfitting. However, they include high computation costs, are complex, and are prone to overfitting if the tuning parameters are not carefully set. In summary, the studies presented herein point to the advantages of the ensemble models in terms of dealing with large and diverse datasets, with the key focus being placed on discussing the importance of model selection combined with a rigorous fine-tuning process.

Future Work:

The further study could emphasize the possibility to expand the list of features under consideration, aiming at the improved model's performance, and to find new or previously unnoticed networks that might be useful to make a more accurate prognosis. Other forms of ensemble methods, for example, AdaBoost or even XGBoost, could therefore yield more insight into how various techniques with regards to the feature importance and the levels of contribution towards accuracy avert any generalization of the problem. As for the packages, future work could extend the presented experiments and models to more data sets with more elements and heterogeneous sources, when possible, to test their portability to other domains that are of interest. I would like to note that specific improvements in computational speed, for example, with the help of parallel computations or pruning the models, might help to eliminate the drawback of the high demand for resources in the framework of the ensemble approach. The usage of explainability approaches like SHAP or LIME that can address the interpretability issue of ensemble's decisions can help better harmonize the model used and make them more usable for real-world usage.

References:

Abubakar, M., Che, Y., Ivascu, L., Almasoudi, F. M., & Jamil, I. (2022). Performance analysis of energy production of large-scale solar plants based on artificial intelligence (machine learning) technique. *Processes*, *10*(9), 1843.

Ahmed, R., Sreeram, V., Mishra, Y., &Arif, M. D. (2020). A review and evaluation of the state-of-the-art in PV solar power forecasting: Techniques and optimization. *Renewable and Sustainable Energy Reviews*, *124*, 109792.

Akhter, M. N., Mekhilef, S., Mokhlis, H., & Mohamed Shah, N. (2019). Review on forecasting

of photovoltaic power generation based on machine learning and metaheuristic techniques. *IET Renewable Power Generation*, *13*(7), 1009-1023.
Alabi, T. M., Aghimien, E. I., Agbajor, F. D., Yang, Z., Lu, L., Adeoye, A. R., &Gopaluni, B. (2022). A review on the integrated optimization techniques and machine learning approaches for modeling, prediction, and decision making on integrated energy systems. *Renewable Energy*, *194*, 822-849.
Al-Dahidi, S., Ayadi, O., Adeeb, J., &Louzazni, M. (2019). Assessment of artificial neural networks learning algorithms and training datasets for solar photovoltaic power production prediction. *Frontiers in energy research*, *7*, 130.
Allal, Z., Noura, H. N., Salman, O., & Chahine, K. (2024). Machine learning solutions for renewable energy systems: Applications, challenges, limitations, and future directions. *Journal of Environmental Management*, *354*, 120392.
Aybar-Ruiz, A., Jiménez-Fernández, S., Cornejo-Bueno, L., Casanova-Mateo, C., Sanz-Justo, J., Salvador-González, P., & Salcedo-Sanz, S. (2016). A novel grouping genetic algorithm–extreme learning machine approach for global solar radiation prediction from numerical weather models inputs. *Solar Energy*, *132*, 129-142.
Barrera, J. M., Reina, A., Maté, A., & Trujillo, J. C. (2020). Solar energy prediction model based on artificial neural networks and open data. *Sustainability*, *12*(17), 6915.
Blaga, R., Sabadus, A., Stefu, N., Dughir, C., Paulescu, M., &Badescu, V. (2019). A current perspective on the accuracy of incoming solar energy forecasting. *Progress in energy and combustion science*, *70*, 119-144.
Dairi, A., Harrou, F., Sun, Y., &Khadraoui, S. (2020). Short-term forecasting of photovoltaic solar power production using variational auto-encoder driven deep learning approach. *Applied Sciences*, *10*(23), 8400.
Essam, Y., Ahmed, A. N., Ramli, R., Chau, K. W., Idris Ibrahim, M. S., Sherif, M., ... & El-Shafie, A. (2022). Investigating photovoltaic solar power output forecasting using machine learning algorithms. *Engineering Applications of Computational Fluid Mechanics*, *16*(1), 2002-2034.
Essam, Y., Ahmed, A. N., Ramli, R., Chau, K. W., Idris Ibrahim, M. S., Sherif, M., ... & El-Shafie, A. (2022). Investigating photovoltaic solar power output forecasting using machine learning algorithms. *Engineering Applications of Computational Fluid Mechanics*, *16*(1), 2002-2034.
Fan, J., Wu, L., Zhang, F., Cai, H., Zeng, W., Wang, X., & Zou, H. (2019). Empirical and machine learning models for predicting daily global solar radiation from sunshine duration: A review and case study in China. *Renewable and Sustainable Energy Reviews*, *100*, 186-212.
Feng, Y., Hao, W., Li, H., Cui, N., Gong, D., & Gao, L. (2020). Machine learning models to quantify and map daily global solar radiation and photovoltaic power. *Renewable and Sustainable Energy Reviews*, *118*, 109393.
Fouilloy, A., Voyant, C., Notton, G., Motte, F., Paoli, C., Nivet, M. L., ... &Duchaud, J. L. (2018). Solar irradiation prediction with machine learning: Forecasting models selection method depending on weather variability. *Energy*, *165*, 620-629.
Ghimire, S., Deo, R. C., Casillas-Pérez, D., & Salcedo-Sanz, S. (2022). Boosting solar radiation predictions with global climate models, observational predictors and hybrid deep-machine learning algorithms. *Applied Energy*, *316*, 119063.
Gu, G. H., Noh, J., Kim, I., & Jung, Y. (2019). Machine learning for renewable energy materials. *Journal of Materials Chemistry A*, *7*(29), 17096-17117.
Jamil, I., Lucheng, H., Iqbal, S., Aurangzaib, M., Jamil, R., Kotb, H., ... &AboRas, K. M. (2023). Predictive evaluation of solar energy variables for a large-scale solar power plant based on triple deep learning forecast models. *Alexandria Engineering Journal*, *76*, 51-73.
Jobayer, M., Shaikat, M. A. H., Rashid, M. N., & Hasan, M. R. (2023). A systematic review on predicting PV system parameters using machine learning. *Heliyon*, *9*(6).
Lai, J. P., Chang, Y. M., Chen, C. H., & Pai, P. F. (2020). A survey of machine learning models in renewable energy predictions. *Applied Sciences*, *10*(17), 5975.
Lai, J. P., Chang, Y. M., Chen, C. H., & Pai, P. F. (2020). A survey of machine learning models in renewable energy predictions. *Applied Sciences*, *10*(17), 5975.

Ledmaoui, Y., El Maghraoui, A., El Aroussi, M., Saadane, R., Chebak, A., &Chehri, A. (2023). Forecasting solar energy production: A comparative study of machine learning algorithms. *Energy Reports*, *10*, 1004-1012.
Lee, J., Wang, W., Harrou, F., & Sun, Y. (2020). Reliable solar irradiance prediction using ensemble learning-based models: A comparative study. *Energy Conversion and Management*, *208*, 112582.
Li, L. L., Wen, S. Y., Tseng, M. L., & Wang, C. S. (2019). Renewable energy prediction: A novel short-term prediction model of photovoltaic output power. *Journal of Cleaner Production*, *228*, 359-375.
Li, Z., Rahman, S. M., Vega, R., & Dong, B. (2016). A hierarchical approach using machine learning methods in solar photovoltaic energy production forecasting. *Energies*, *9*(1), 55.
Li, Z., Rahman, S. M., Vega, R., & Dong, B. (2016). A hierarchical approach using machine learning methods in solar photovoltaic energy production forecasting. *Energies*, *9*(1), 55.
Markovics, D., & Mayer, M. J. (2022). Comparison of machine learning methods for photovoltaic power forecasting based on numerical weather prediction. *Renewable and Sustainable Energy Reviews*, *161*, 112364.
Muftah, A. F., Alghoul, M. A., Fudholi, A., Abdul-Majeed, M. M., &Sopian, K. (2014). Factors affecting basin type solar still productivity: A detailed review. *Renewable and Sustainable Energy Reviews*, *32*, 430-447.
Narvaez, G., Giraldo, L. F., Bressan, M., & Pantoja, A. (2021). Machine learning for site-adaptation and solar radiation forecasting. *Renewable Energy*, *167*, 333-342.
Pedro, H. T., & Coimbra, C. F. (2012). Assessment of forecasting techniques for solar power production with no exogenous inputs. *Solar Energy*, *86*(7), 2017-2028.
Salcedo-Sanz, S., Cornejo-Bueno, L., Prieto, L., Paredes, D., & García-Herrera, R. (2018). Feature selection in machine learning prediction systems for renewable energy applications. *Renewable and Sustainable Energy Reviews*, *90*, 728-741.
Schwalbert, R. A., Amado, T., Corassa, G., Pott, L. P., Prasad, P. V., &Ciampitti, I. A. (2020). Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern Brazil. *Agricultural and Forest Meteorology*, *284*, 107886.
Sharma, A., &Kakkar, A. (2018). Forecasting daily global solar irradiance generation using machine learning. *Renewable and Sustainable Energy Reviews*, *82*, 2254-2269.
Steyerberg, E. W., Van Calster, B., &Pencina, M. J. (2011). Performance measures for prediction models and markers: evaluation of predictions and classifications. *Revista Espanola de Cardiologia (English Edition)*, *64*(9), 788-794.
Tian, J., Ooka, R., & Lee, D. (2023). Multi-scale solar radiation and photovoltaic power forecasting with machine learning algorithms in urban environment: A state-of-the-art review. *Journal of Cleaner Production*, 139040.
Tripathi, A. K., Aruna, M., Elumalai, P. V., Karthik, K., Khan, S. A., Asif, M., & Rao, K. S. (2024). Advancing solar PV panel power prediction: A comparative machine learning approach in fluctuating environmental conditions. *Case Studies in Thermal Engineering*, *59*, 104459.
Tripathi, A. K., Aruna, M., Elumalai, P. V., Karthik, K., Khan, S. A., Asif, M., & Rao, K. S. (2024). Advancing solar PV panel power prediction: A comparative machine learning approach in fluctuating environmental conditions. *Case Studies in Thermal Engineering*, *59*, 104459.
Van Klompenburg, T., Kassahun, A., &Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. *Computers and electronics in agriculture*, *177*, 105709.
Velmurugan, V., &Srithar, K. (2011). Performance analysis of solar stills based on various factors affecting the productivity—a review. *Renewable and sustainable energy reviews*, *15*(2), 1294-1304.
Vickers, A. J., Van Calster, B., &Steyerberg, E. W. (2016). Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *bmj*, *352*.
Voyant, C., Notton, G., Kalogirou, S., Nivet, M. L., Paoli, C., Motte, F., &Fouilloy, A. (2017). Machine learning methods for solar radiation forecasting: A review. *Renewable energy*, *105*, 569-582.
Wang, H., Lei, Z., Zhang, X., Zhou, B., & Peng, J. (2019). A review of deep learning for renewable energy forecasting. *Energy Conversion and Management*, *198*, 111799.

Yadav, A. K., Sharma, V., Malik, H., &Chandel, S. S. (2018). Daily array yield prediction of grid-interactive photovoltaic plant using relief attribute evaluator based radial basis function neural network. *Renewable and Sustainable Energy Reviews*, *81*, 2115-2127.

Yılmaz, H., &Şahin, M. (2023). Solar panel energy production forecasting by machine learning methods and contribution of lifespan to sustainability. *International Journal of Environmental Science and Technology*, *20*(10), 10999-11018.

Zhou, Y., Zhou, N., Gong, L., & Jiang, M. (2020). Prediction of photovoltaic power output based on similar day analysis, genetic algorithm and extreme learning machine. *Energy*, *204*, 117894.