IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 9, June 2025
Impact Factor: 7.67

Predicting PV Output from Solar Sites to Help Businesses Estimate ROI

Biswakarmi Kumari[1], Sreetama Saha[2], Rima Das[3], Richa Shah[4], Suparna Karmakar[5], Tridib Chakraborty[6], Suparna Biswas[7]
Students, Department of Information Technology[1,2,3,4]
Faculty, Department of Information Technology[5,6,7]
Guru Nanak Institute of Technology, Kolkata, India

Abstract: *This work focuses on the critical importance of predictability of power output in the integration of solar photovoltaics into traditional electrical grid systems. The authors aim to find the best-fit model to solve this problem, thereby facilitating the transition of businesses from traditional power consumption models to renewable sources. The researchers estimate that governments are losing hundreds of millions of dollars annually in the solar sector due to a decline in solar power generation, which could be as high as 52%. This study attempts to address this issue by proposing a solution to improve power output predictability. The findings of this research could have significant implications for businesses and policymakers interested in transitioning to renewable energy sources.*

Keywords: Solar photovoltaics, power output predictability, best-fit model,K-Nearest Neighbors (KNN), Random Forest (RF), LightGBM (LGBM), Deep Neural Network(DNN), Long Short-Term Memory(LSTM), Meta-learning.

I. INTRODUCTION

Solar energy has gained significant traction as a sustainable source of electricity generation over the past decade. The falling cost of solar photovoltaic (PV) systems has made them more accessible to businesses, making it more common for them to install PV systems to power their operations. With the increasing number of PV installations, there is a need to predict their energy output accurately to help businesses estimate the return on investment (ROI) for their PV systems. [1][2]

This project report aims to address this need by developing a model that predicts the energy output of a solar site using various meteorological parameters. The model takes into account factors such as temperature, solar irradiance, wind speed, and cloud cover to predict the energy output of a solar site accurately.

This work first discusses the motivation behind the problem and the importance of predicting the energy output of a solar site for businesses. Followed by a literature review of previous research in the area of PV output prediction and the various approaches used to develop prediction models. Next the methodology has been described that is used to develop the prediction model, including the data collection process, data preprocessing, feature engineering, and model selection. It also provides a detailed analysis of the model's performance and its accuracy in predicting the energy output of a solar site. Finally, the work concludes with a discussion of the results obtained and their implications for businesses that use PV systems. Authors highlighted the importance of accurate energy output prediction in helping businesses estimate the ROI of their PV systems, and the potential benefits of using the developed model for this purpose.

II. RELATED CONCEPTS

2.1. K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a type of machine learning algorithm used for classification and regression. In KNN, the output of a new data point is predicted based on its proximity to the K closest data points in the training set. The KNN algorithm for classification works as follows: Choose the number of neighbours K to consider. Calculate the

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 9, June 2025
Impact Factor: 7.67

distance between the new data point and all points in the training set using a distance metric (such as Euclidean distance). Select the K data points with the shortest distance to the new data point. Classify the new data point based on the class that is most common among the K neighbours. For regression, the prediction is the average of the K nearest data points. KNN is a simple algorithm to implement and can work well when the decision boundary is nonlinear and when the data is noisy. However the algorithm can be computationally expensive especially for large datasets and can be sensitive to the choice of distance metric and the value of K.[2][3]

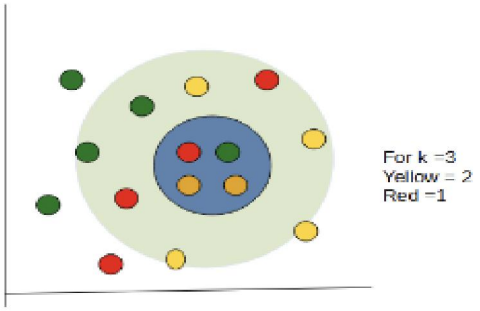


FIGURE 1. Example of k- Nearest Neighbour

2.2. Random Forest (RF)

Random Forest (RF) is a type of ensemble machine learning algorithm that combines multiple decision trees to improve the accuracy and robustness of predictions.

Here's how Random Forest works: A dataset is randomly sampled with replacement to create multiple training sets. For each training set, a decision tree is built by selecting the best feature to split the data and continuing until each leaf node is pure or a predefined maximum depth is reached. Predictions are made for new data by aggregating the predictions of all the decision trees. For classification, this can be done by majority voting, while for regression, this can be done by averaging. The random sampling of training data and features helps to reduce overfitting, while the aggregation of multiple trees helps to improve the accuracy and reduce the variance of predictions. RF is widely used for classification and regression tasks, and can handle both categorical and continuous data.[4]

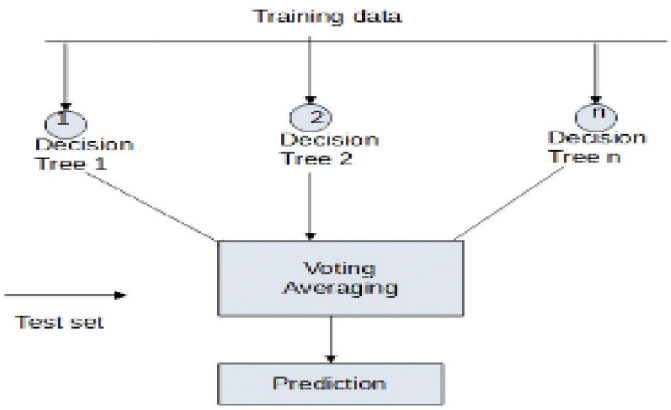


FIGURE 2. Structure of Random Forest

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 9, June 2025
Impact Factor: 7.67

2.3 LightGBM

LightGBM (LGBM) is a gradient boosting framework that uses tree-based learning algorithms. It is designed to be efficient, scalable and fast, making it well-suited for handling large-scale datasets. LightGBM (LGBM) is a gradient boosting framework that uses tree-based learning algorithms. It is designed to be efficient, scalable and fast, making it well-suited for handling large-scale datasets. Each tree is built to minimise the loss function and the gradient of the loss function. The algorithm also uses a technique called "leaf-wise" growth to reduce computational complexity and improve training speed. LightGBM can handle both categorical and continuous data, and it can automatically handle missing values. It also has several built-in techniques to handle overfitting, such as early stopping and regularisation. LightGBM also provides tools for feature selection and can handle high-dimensional data. [5]

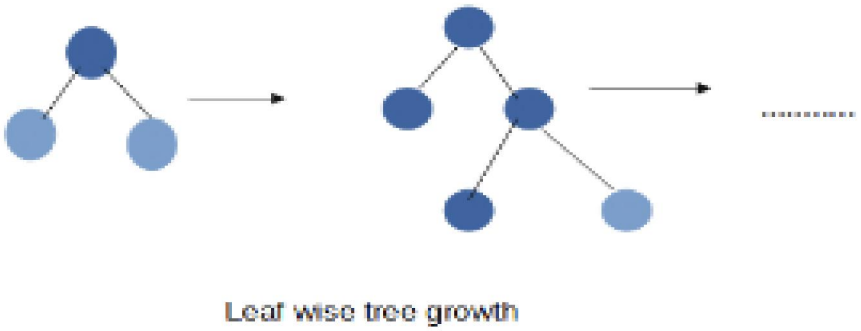


FIGURE 3. Example of Light gradient boosting framework

2.4. Deep Neural Network(DNN)

DNN stands for Deep Neural Network. It is a type of artificial neural network that consists of multiple layers of interconnected nodes, also known as neurons. These layers allow DNNs to learn increasingly complex representations of input data, leading to powerful pattern recognition and decision-making abilities. DNNs are typically used in applications such as computer vision, natural language processing, speech recognition, and robotics. They are able to extract features from raw data and make accurate predictions, often surpassing human performance on certain tasks. Training a DNN involves feeding it large amounts of data and adjusting the weights of the connections between neurons based on the errors in its predictions. This process is usually done using stochastic gradient descent or a similar optimization algorithm.[6],[7]

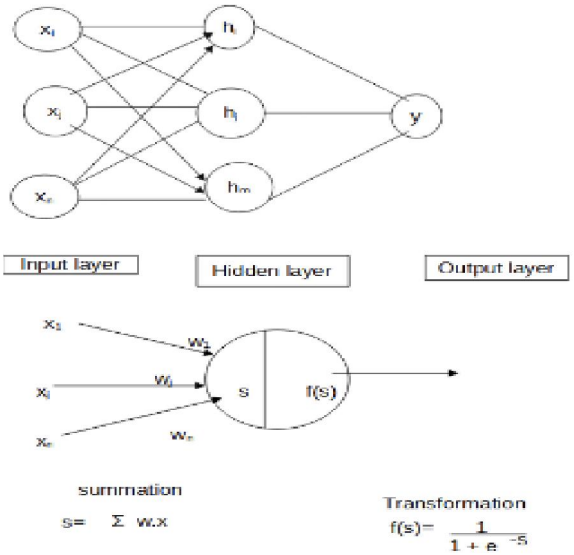


FIGURE 4. Structure of Deep Neural Network

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 9, June 2025
Impact Factor: 7.67

2.5 Long Short-Term Memory(LSTM)

LSTM stands for Long Short-Term Memory, which is a type of recurrent neural network (RNN) architecture designed to overcome the limitations of traditional RNNs in capturing and modelling long-term dependencies in sequential data. LSTMs were first introduced by Hochreiter and Schmidhuber in 1997 and have since been widely used in various applications such as speech recognition, language translation, image captioning, and more.

The key innovation of LSTMs is the use of a memory cell, which allows the network to selectively remember or forget information over time. The cell has three main components: an input gate, a forget gate, and an output gate, which regulate the flow of information into and out of the cell. The input gate controls which information is stored in the cell, the forget gate controls which information is discarded, and the output gate controls the information that is passed to the next time step.

By selectively controlling the flow of information, LSTMs can effectively capture and model long-term dependencies in sequential data. This makes them particularly useful for tasks that involve processing sequential data, such as time series prediction, natural language processing, and speech recognition.[8]

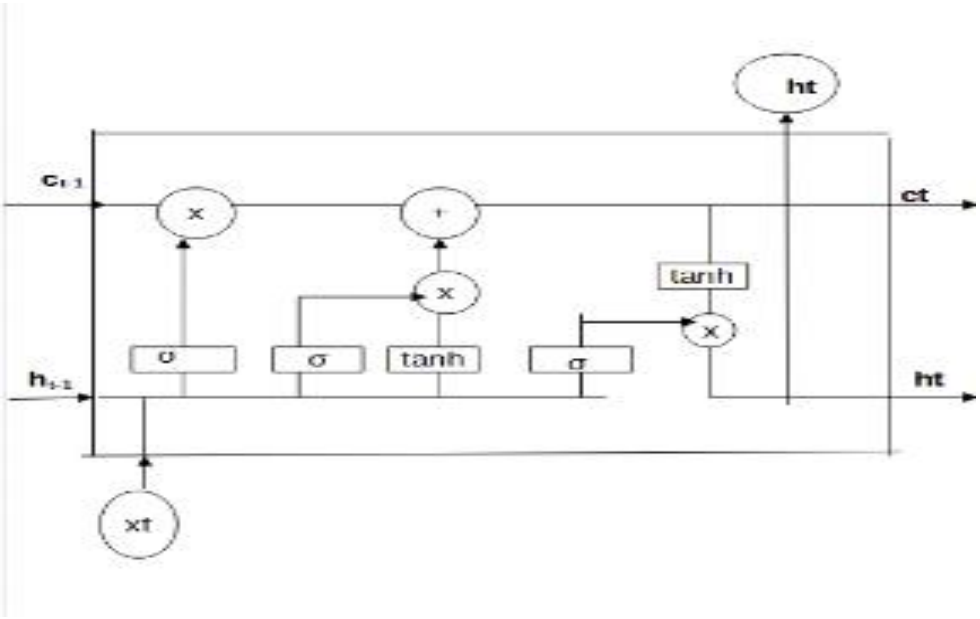


FIGURE 5. Structure of long short-term Memory

2.6. Meta-learning

Meta-learning algorithms can be used in a variety of applications, such as few-shot learning, where a model is trained on a limited number of examples and then used to make predictions on new, previously unseen examples. They can also be used in reinforcement learning, where an agent learns to make decisions by trial and error, by learning from past experience and adapting its behaviour to new situations. Overall, meta-learning is an exciting and rapidly growing area of research in machine learning, and has the potential to greatly improve the efficiency and performance of intelligent systems[9]

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 9, June 2025
Impact Factor: 7.67

III. METHODOLOGY

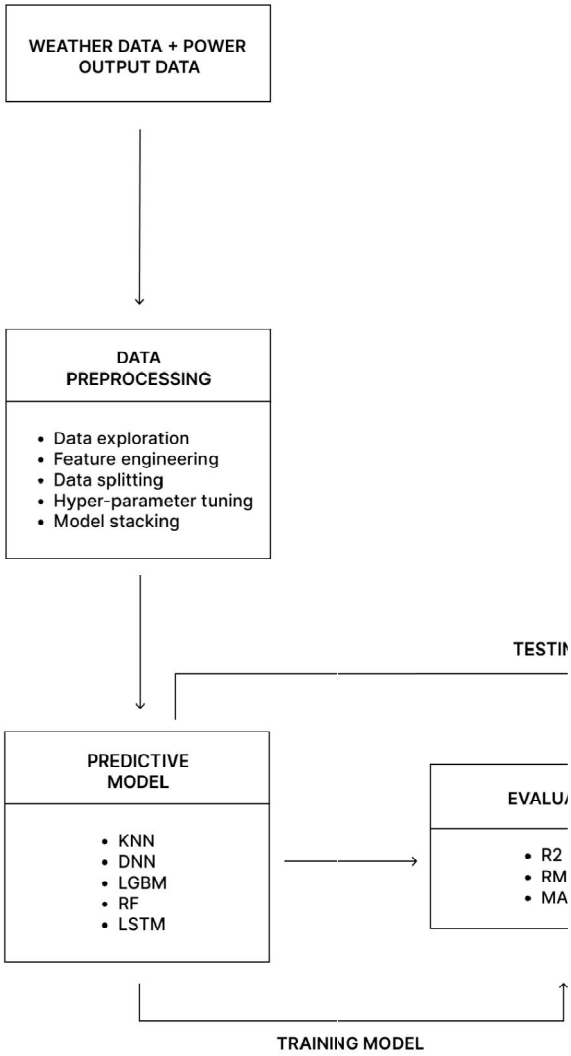


FIGURE 6. Methodology

3.1.Data Collection

This study utilised a publicly available dataset from Kaggle, which includes power output data from 12 Northern hemisphere photovoltaic panels over a 14-month period. The dataset comprises 17 columns and 21,000 rows, with independent variables such as location, date, time, latitude, longitude, altitude, year, month, season, humidity, ambient temperature, wind speed, visibility, pressure, cloud ceiling, and power output from the solar panel.

The dataset consists of 17 distinct columns or variables which are:-

	Location	Location reference
	Date Time	Date time reference
	Latitude	Latitude ranges from -90° (South Pole) to +90° (North Pole).
	Longitude	Longitude ranges from -180° to +180°.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 9, June 2025
Impact Factor: 7.67

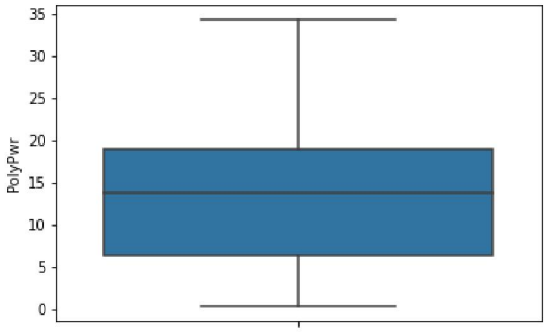
	Altitude(ft)	the height above sea level of a point on the Earth's surface
	Month	Represents the month of the year
	Hour	represents the hour of the day
	Season	string data type that represents the season of the year
	Humidity (rh (%))	Relative humidity.
	AmbientTemp (degC)	Temperature in Celsius
	Wind.Speed (wv (m/s))	Wind speed.
	Visibility (Km)	the distance at which objects can be seen clearly
	Pressure (mbar)	Atmospheric pressure in millibars
	Cloud.Ceiling(ft)	the height of the lowest layer of clouds in the sky

	Location	Date	Time	Latitude	Longitude	Altitude	YRMODAHRMI	Month	Hour	Season	Humidity	AmbientTemp	PolyPwr	Wind.Speed	Visibility	Pressure	Cloud.Ceiling
0	Camp Murray	20171203	1145	47.11	-122.57	84	2.017120e+11	12	11	Winter	81.71997	12.86919	2.42769	5	10.0	1010.6	722
1	Camp Murray	20171203	1315	47.11	-122.57	84	2.017120e+11	12	13	Winter	96.64917	9.66415	2.46273	0	10.0	1011.3	23
2	Camp Murray	20171203	1330	47.11	-122.57	84	2.017120e+11	12	13	Winter	93.61572	15.44983	4.46836	5	10.0	1011.6	32
3	Camp Murray	20171204	1230	47.11	-122.57	84	2.017120e+11	12	12	Winter	77.21558	10.36659	1.65364	5	2.0	1024.4	6
4	Camp Murray	20171204	1415	47.11	-122.57	84	2.017120e+11	12	14	Winter	54.80347	16.85471	6.57939	3	3.0	1023.7	9
5	Camp Murray	20171204	1430	47.11	-122.57	84	2.017120e+11	12	14	Winter	47.10083	18.12363	2.92027	0	5.0	1023.7	722
6	Camp Murray	20171205	1115	47.11	-122.57	84	2.017120e+11	12	11	Winter	43.55469	19.41269	3.41284	0	4.0	1025.7	722
7	Camp Murray	20171205	1200	47.11	-122.57	84	2.017120e+11	12	12	Winter	30.56641	23.90930	4.82020	5	7.0	1026.0	722
8	Camp Murray	20171205	1300	47.11	-122.57	84	2.017120e+11	12	13	Winter	17.90771	32.32346	5.98127	5	10.0	1025.7	722
9	Camp Murray	20171205	1400	47.11	-122.57	84	2.017120e+11	12	14	Winter	14.40430	35.41267	4.96121	6	10.0	1025.4	722

FIGURE 7. Sample dataset

3.2 Outlier Handling

To address the issue of outliers in the dataset, various outlier detection methods were employed. Since different factors can affect the performance of solar panels, resulting in anomalous parameters, outlier detection was necessary to reduce the negative impact on model accuracy [10]. The Boxplot rule was found to be the most accurate and robust method for detecting PV faults at relatively high contamination levels.



IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 9, June 2025
Impact Factor: 7.67

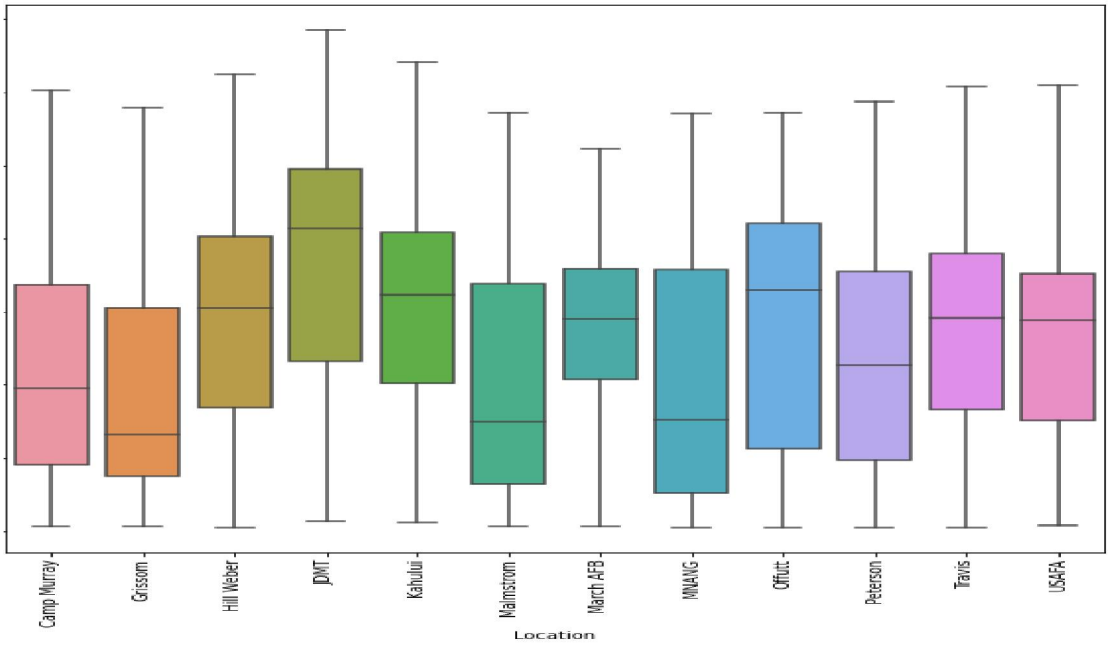


FIGURE 8. Multiple boxplot comparison of power output to locations

3.3.Target Distribution

The distribution of the output power, as the dependent variable, was analysed visually and through calculation of the skewness[10]. The results indicate a fairly symmetrical distribution, with no significant skewness observed. Feature Selection: To prevent accuracy reduction due to many input features, certain variables were dropped from the dataset. Altitude was removed due to its high correlation with Pressure, while Longitude was eliminated as it had zero correlation with the target variable. Time, Hour, Month, and Date were dropped as they had strong correlations with engineered cyclic features but low correlation with the target variable.

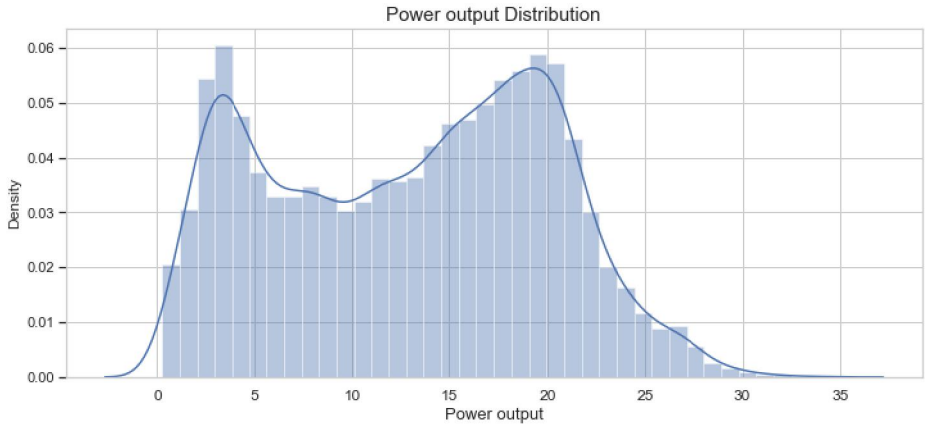


FIGURE 9: Power output distribution

3.4. Feature selection

Feature selection is a critical step in developing accurate machine learning models. In this study, we considered several factors that influence the overall output of a PV solar panel, such as panel performance, weather conditions, and operating conditions. However, including too many input features may reduce model accuracy, and therefore, we performed feature selection to identify the most relevant features for predicting power output.[10]

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 9, June 2025
Impact Factor: 7.67

During feature selection, we evaluated the correlation between each input feature and the target variable. We found that altitude had a high correlation with pressure but did not change for a given location. Therefore, altitude was dropped from the list of input features, and pressure was retained as a more dynamic feature. Similarly, longitude was dropped from the list of input features because it had zero correlation with the target variable.

We also considered the time-related input features, such as time, hour, month, and date. We observed that these features had strong correlations with the engineered cyclic features, which were created to capture the periodicity of the data. However, these time-related features had low correlations with the target variable, and therefore, we dropped them from the list of input features.

Overall, the feature selection process allowed us to identify the most relevant input features for predicting power output from a PV solar panel. By selecting the most important features, we aimed to develop a more accurate and interpretable machine learning model for predicting power output

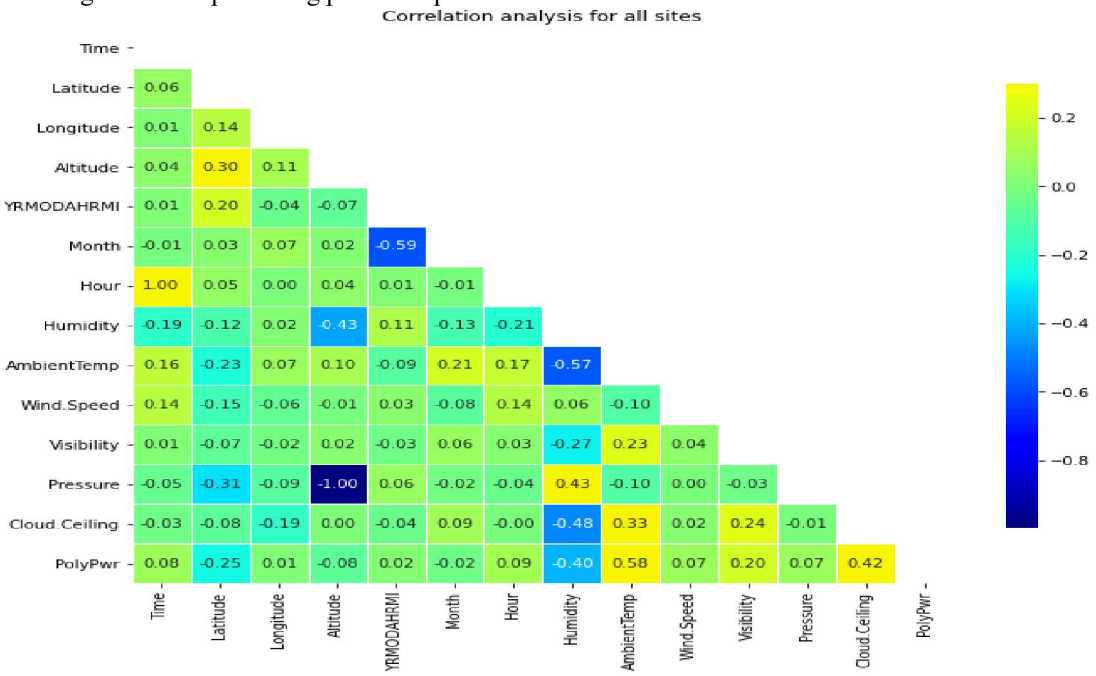


FIGURE 10: Correlation analysis for all sites

3.5. Data Splitting

The dataset was split into 80% training data and 20% test data to enable model training and testing. The test data was kept hidden throughout hyper-parameter tuning and model training. Random search cross-validation was used to select the optimal combination of hyper-parameters for each model. Specifically, 1000 permutations of the hyper-parameters were applied to 4 splits of the training data.

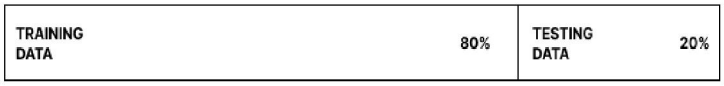


FIGURE 11: Data Splitting

3.6 Model Stacking

The meta-learner was trained using the "stacking" strategy, where base models are trained in parallel, and their predictions are combined[10]. The stacking regressor's output represents the probability that an example belongs to one class or another. This approach reduces the risk of overfitting and allows for the combination of multiple linear models without the need for an explicit representation. The key advantage of stacking is that it allows us to combine multiple

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 9, June 2025
Impact Factor: 7.67

linear models without having to learn an explicit representation for combining them. This approach reduces the risk of overfitting.

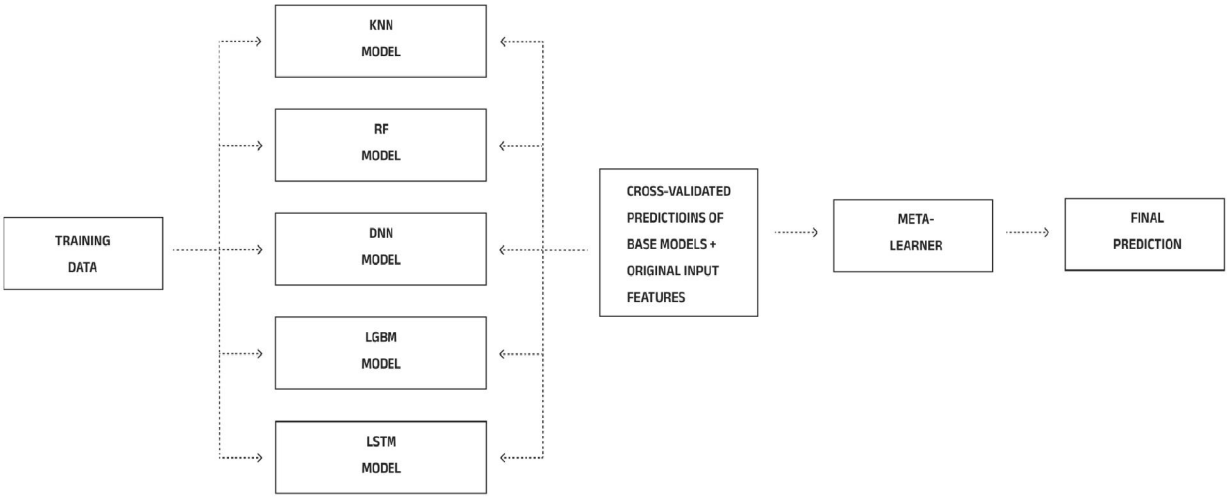


FIGURE 12: Model Stacking

IV. PERFORMANCE METRICS

Performance metrics are quantifiable measures used to assess and evaluate the effectiveness and efficiency of a particular process, system, or organisation. These metrics are typically numerical in nature and are used to track progress, identify areas for improvement, and make data-driven decisions.[4]

Performance metrics can vary depending on the context and goals of the evaluation. In business settings, common performance metrics may include financial measures such as revenue, profit margins, and return on investment (ROI). Other commonly used performance metrics in various fields may include customer satisfaction ratings, employee engagement scores, product quality ratings, and productivity measures.

Effective performance metrics should be specific, relevant, and measurable, allowing for easy tracking and comparison over time. Choosing the right performance metrics is essential to effectively evaluate and improve the performance of a process, system, or organisation.

Metrics used in the project:

- R-squared(R^2)
- Root Mean Squared Error(RMSE)
- Mean Absolute Error (MAE)

4.1. R-squared

R-squared (R^2) is a statistical metric that measures the proportion of variation in the dependent variable that is explained by the independent variable(s) in a regression model. It ranges from 0 to 1. Its higher value indicates a better fit of the model to the data. However, it should be interpreted in the context of the specific problem being addressed and is typically used in conjunction with other metrics to fully evaluate the performance of a regression model.[12]

Formula:

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

y is the true value, $\hat{y}$ is the predicted value of y and $\bar{y}$ is the mean value of y

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 9, June 2025
Impact Factor: 7.67

4.2. Root Mean Square Error(RMSE)

Root Mean Square Error (RMSE) is a statistical measure of the difference between predicted and actual values in a regression analysis. It represents the square root of the average of the squared differences between predicted and actual values. RMSE is commonly used to evaluate the accuracy of a regression model and to compare the performance of different models. A lower RMSE value indicates better predictive performance of the model. It is expressed in the same units as the dependent variable being predicted.[4][12]

Formula:

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

n is the number of examples to be evaluated

4.3. Mean Absolute Error (MAE)

Mean Absolute Error (MAE) is a statistical measure of the average absolute difference between the predicted and actual values in a regression analysis. It represents the average of the absolute differences between predicted and actual values, without taking into account their direction. MAE is commonly used to evaluate the accuracy of a regression model and to compare the performance of different models. A lower MAE value indicates better predictive performance of the model. It is expressed in the same units as the dependent variable being predicted.[4][12]

Formula:

$$MAE = \frac{1}{n}\sum_i |y_i - \hat{y}_i|$$

Where n is the number of fitted data points, Y corresponds to the actual value and ^Y is the predicted value

Observation:

The values of R-squared can be between 0 to 1. While comparing the R-squared value which is nearer to 1 that model will be treated as a more preferable model.Here LGBM is the best model based on our data because its R-squared value is the maximum.

V. RESULTS

As previously stated, the training and testing of models necessitate a considerable amount of data. To evaluate the performance of a solar photovoltaic panel under real-world environmental conditions, experiments were conducted in an outdoor setting. This section showcases the outcomes of three distinct models utilized in the experiments.

Cross-validation scores

Cross-validation is a technique used to evaluate the performance of a predictive model by dividing the available data into multiple subsets. This process helps assess the accuracy and reliability of the model in predicting PV output, aiding businesses in estimating the ROI.

Subsequently, the best CV scores for each algorithm category are presented below:

MODEL	R^2
K Nearest Neighbors(baseline)	0.618

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 9, June 2025
Impact Factor: 7.67

Deep Neural Network	0.662
Light Gradient Boosting Machine	0.676
Random Forest	0.669
LSTM	0.237

TABLE 1. Five-fold CV best scores

Results show that the optimal RF model has the highest CV score among all algorithms examined.

Test data scores

The performance of the models was assessed on a hold-out set comprising 20% of the complete dataset to determine the test data scores. The outcomes of the evaluation are presented below for reference.

MODEL	R^2	RMSE	MAE
K Nearest Neighbors(baseline)	0.618	4.402	2.970
Deep Neural Network	0.662	4.142	2.708
Random Forest	0.669	4.095	2.786
Light Gradient Boosting Machine	0.676	4.053	2.723
LSTM	0.237	0.777	0.549
Stacked Model	0.681	4.023	2.669

TABLE2. Model performance on the unseen test data.

The analysis indicates that the stacked model delivered the best overall performance, achieving a notable 10% improvement compared to the KNN baseline model. Furthermore, based on all the metrics taken into account, the LGBM model was found to be the best-performing base model.

It is worth noting that all the models displayed robust generalization capabilities on the unseen test set, with comparable performance observed between the cross-validation (CV) and test scores.

VI. FEATURE IMPORTANCE

To determine the crucial predictors of solar power output, the LGBM and RF models, equipped with the feature importance calculation ability, were employed. The ensuing table showcases the scaled importance values of the top 5 features that emerged as the most significant in the prediction process.

LGBM

FEATURES	SCALED IMPORTANCE (%)
Humidity	100
Pressure	97
Ambient temperature	95

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 9, June 2025
Impact Factor: 7.67

Wind speed	63
Cloud ceiling	37

RF

FEATURES	SCALED IMPORTANCE (%)
Ambient temperature	100
Humidity	57
Cloud ceiling	52
Pressure	28
Sine of month	27

TABLE3. Top 5 scaled feature importance using tree methods.

The LGBM and RF models identified ambient temperature, humidity, cloud ceiling, and pressure as the top 5 significant predictors for solar power output, corroborating their importance in predicting the system's performance.

VII. CONCLUSION

This project is aimed to identify the best model for predicting photovoltaic power output among LGBN, DNN, KNN, LSTM, and RF models and to conduct a financial analysis to determine the return on investment for a photovoltaic power system. The project involved training and testing each model on a dataset of photovoltaic power output measurement under various environmental conditions. Through a comprehensive analysis of data ,it was found that the LGBN model outperformed the other models in terms of accuracy and predictive power.

Moreover, the financial analysis revealed that investing in a photovoltaic power system using the LGBM model would generate a positive return on investment over the system's lifetime. These findings have significant implications for the renewable energy sector and underscore the potential of using advanced machine learning algorithms for predictive modelling in this field. Future research could explore the potential of combining multiple predictive models to further improve accuracy and predictive power. Overall, the result of this project demonstrate the potential of advanced machine learning algorithms for improving the efficiency and sustainability of energy systems and highlight the importance of continued investment in renewable energy technologies.

REFERENCES

[1] Sharma, N., Sharma, P., Irwin, D., & Shenoy, P. (2011, October). Predicting solar generation from weather forecasts using machine learning. In 2011 IEEE international conference on smart grid communications (SmartGridComm) (pp. 528-533). IEEE.

[2] Jawaid, F., & NazirJunejo, K. (2016, August). Predicting daily mean solar power using machine learning regression techniques. In 2016 Sixth International Conference on Innovative Computing Technology (INTECH) (pp. 355-360). IEEE.

[3] Rodríguez, F., Fleetwood, A., Galarza, A., & Fontán, L. (2018). Predicting solar energy generation through artificial neural networks using weather forecasts for microgrid control. Renewable Energy, 126, 855-864.

[4] Al-Dahidi, S., Louzazni, M., & Omran, N. (2020). A local training strategy-based artificial neural network for predicting the power production of solar photovoltaic systems. IEEE Access, 8, 150262-150281.

[5] Rich H. Inman, Hugo T.C. Pedro, and Carlos F.M. Coimbra. Solar forecasting methods for renewable energy integration. Progress in Energy and Combustion Science, 39(6):535 – 576, 2013.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 9, June 2025
Impact Factor: 7.67

[6]Zhongheng Zhang Introduction to machine learning K-nearest neighbors Annals of Translational Medicine 4(11):218-218 , 2016
[7]Jehad Ali,Rehannullah Khan,Nasir Ahmad,Imran Maqsood Random Forests and Decision Trees ,2012
[8]Guolin Ke,Qi Meng, Thomas Finley,Taifeng Wang,Wei Chen,Weidong Ma,Qiwei Ye, Tie-Yan Liu LightGBM: A Highly Efficient Gradient Boosting Decision Tree, 2017
[9] Mariette Awad,Rahul Khanna Deep Neural Networks, 2015
[10]Sepp Hochreiter,Jurgen Schmidhuber Long-Short-term Memory Neural Computation 9(8):1735-80, 1997
[11]Richardo Vilalta,Christophe Gridaud-Carrier,Pavel Brazdil Meta-Learning - Concepts and Techniques 10.1007/978-0-387-09823-4_36 Data Mining and Knowledge Discovery Handbook (pp.717-731),2010
[12] Davide Chicco,Matthijis J Warrens,Giuseppe Jurman The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation PeerJ Computer Science 7(3):e623