Solar Photovoltaic Power Forecasting via Advanced Machine Learning Methodologies for Enhanced Electrical Grid Stability

Gorle.Sai Kethan[1], Chappa.Jaswanth Naidu[2], Jagu.Sravan Kumar[3], Jalla Rajya Lakshmi[4], Galla.Venkataswamy[5]
[1,2,3,4]B.Tech (Final Year Student), [5]Assistant Professor, Department of Computer Science and Engineering (Data Science), Raghu Engineering College (Autonomous), Dakamarri, Visakhapatnam

Abstract

This paper presents a machine learning-based system for forecasting solar photovoltaic (PV) power generation to support enhanced electrical grid stability. Solar energy, while abundant and environmentally sustainable, is inherently variable due to its dependence on meteorological conditions. Accurate forecasting is therefore critical for grid management. The proposed system leverages weather-related parameters — including temperature, humidity, cloud cover, solar radiation, wind speed, and zenith angle — to predict solar energy output in kilowatts. Multiple machine learning models are trained and evaluated, including traditional regression methods (Linear Regression, Ridge Regression) and advanced ensemble algorithms such as XGBoost, LightGBM, CatBoost, K-Nearest Neighbors (KNN), and a stacking ensemble. Feature selection using f_regression identifies the ten most predictive meteorological parameters. LightGBM achieved the highest performance with an R^2 of 0.8345, RMSE of 383.93 kW, and MAE of 241.87 kW. The trained model is deployed via a Flask-based web application backed by a MySQL database, enabling real-time predictions from user-supplied weather inputs. Results confirm that advanced gradient boosting methods substantially outperform traditional regression approaches for solar power forecasting.

Keywords: *solar power forecasting, machine learning, LightGBM, XGBoost, CatBoost, meteorological parameters, energy prediction, Flask web application, renewable energy, predictive analytics.*

1. Introduction

The global transition toward renewable energy sources is accelerating, driven by the urgent need to reduce carbon emissions and address climate change. Solar photovoltaic (PV) energy stands at the forefront of this shift, offering abundant, clean power generation at rapidly declining costs. However, solar energy generation is inherently intermittent — output fluctuates with sunlight intensity, temperature, humidity, cloud cover, and wind speed. This variability introduces significant challenges for grid operators who must balance supply and demand in real time.

Accurate solar power forecasting is therefore a foundational requirement for modern smart grids. Without reliable forecasts, utilities must maintain costly reserves of backup power, often from fossil-fuel sources, to compensate for unexpected drops in solar output. Forecasting errors also increase curtailment of renewable energy and reduce the efficiency of energy distribution.

Traditional statistical models such as linear regression have been applied to this forecasting problem, but they frequently fail to capture the complex, nonlinear relationships between atmospheric variables and PV output. As a result, their prediction accuracy is limited, particularly under dynamic or extreme weather conditions. Machine learning methods have emerged as powerful alternatives, capable of modelling intricate data patterns across large meteorological datasets.

This paper proposes a comprehensive solar PV forecasting system employing multiple machine learning algorithms — KNN Regression, XGBoost, LightGBM, CatBoost, and a stacking ensemble — evaluated against standard regression baselines. Feature selection via f_regression ensures that only the most informative meteorological variables are used, reducing dimensionality and improving generalisation. The full pipeline is deployed as a Flask web application with a MySQL backend, enabling practical real-time use by energy planners and utilities.

The primary objectives of this work are to: (1) preprocess and select features from a meteorological-solar generation dataset; (2) train and rigorously compare multiple ML models; (3) identify the best-performing algorithm using MAE, RMSE, and R^2 metrics; and (4) deploy the resulting model in a production-ready web interface.

2. Review of Literature

A substantial body of research has investigated data-driven approaches to solar energy forecasting. Sharma, Singh, and Kumar compared Decision Tree, Random Forest, and Support Vector Machine (SVM) models for solar power prediction using temperature, humidity, and cloud cover as inputs. Their study found that ensemble methods, particularly Random Forest and Gradient Boosting, yielded the most accurate forecasts when dealing with complex, nonlinear weather patterns.

Gupta and Desai conducted a systematic comparison of regression models — Multiple Linear Regression, Ridge Regression, and Lasso Regression — for solar energy forecasting. Their findings indicated that regularisation techniques improve generalisation and that Lasso Regression performed best in high-dimensional feature spaces, highlighting the importance of feature control in regression-based forecasting.

Zhang and Chen provided a comprehensive review of machine learning applications in solar generation forecasting, covering neural networks, SVMs, and traditional regression models. The authors identified data sparsity and the inherently nonlinear nature of PV output as the key challenges, and underscored the need for models that can adapt to variable atmospheric conditions.

Patel and Kumar specifically investigated XGBoost and LightGBM for solar forecasting, demonstrating that gradient-boosted tree models handle nonlinear feature interactions more effectively than conventional regression techniques. Their results showed measurable accuracy gains over linear baselines when applied to meteorological datasets of similar scope to the one used in this work.

Mehta and Rathi applied CatBoost and K-Nearest Neighbours (KNN) to solar power prediction. CatBoost's native handling of categorical variables and its ordered boosting mechanism were shown to reduce overfitting, particularly on smaller datasets, while KNN provided a useful non-parametric baseline. Singh and Gupta proposed a hybrid model combining weather data with temporal features, demonstrating improved long-term forecast accuracy through ensemble integration.

Kumar and Rani benchmarked a wider set of ML models — including Bagging, AdaBoost, and Gradient Boosting — and found that stacking ensembles consistently outperformed individual models by combining their complementary strengths. Kumar, Yadav, and Sharma explored neural network architectures for solar output prediction, reporting strong performance when sufficient training data was available, though at higher computational cost than gradient-boosted alternatives.

Taken together, the literature confirms that advanced ensemble and boosting methods are the current state of the art for meteorological-driven solar forecasting, and that feature selection is critical to maintaining model efficiency and avoiding overfitting. The present work builds on these findings by systematically comparing the leading approaches within a unified pipeline and deploying the winning model in a production web environment.

3. Methodology

3.1 Dataset and Preprocessing: The dataset used in this work is the Solar Energy Power Generation Dataset sourced from Kaggle, containing meteorological variables including temperature at 2 m above ground, relative humidity, total precipitation, total cloud cover, medium cloud cover, low cloud cover, shortwave radiation, wind speed at 80 m, angle of incidence, and zenith angle, alongside the corresponding generated solar power in kW. Data preprocessing involved three main steps. Missing values were addressed through imputation or row removal as appropriate. The dataset was then split into training (80%) and test (20%) sets. Finally, outliers were mitigated using Winsorisation, and feature values were standardised to ensure consistent scaling across all models.

3.2 Feature Selection: Feature selection was performed using the SelectKBest method with the f_regression scoring function. This statistical approach computes the F-statistic for the linear relationship between each feature and the target variable (generated power), selecting the top k features with the highest variance explained relative to error variance. A correlation matrix was additionally computed to detect and remove multicollinear variables, ensuring that the final ten-feature input set was both informative and non-redundant.

3.3 Machine Learning Models: Five machine learning models were trained and evaluated in this study:

3.3.1 K-Nearest Neighbours (KNN) Regression: A non-parametric instance-based learner that predicts the target as the average of the k nearest training samples. KNN captures local patterns without assumptions about the underlying data distribution, though its performance depends strongly on the choice of distance metric and k value.

3.3.2 XGBoost (Extreme Gradient Boosting): An ensemble of decision trees trained sequentially to minimise a regularised loss function. Each tree corrects the residual errors of its predecessors. Regularisation terms penalise model complexity, controlling overfitting, and the algorithm is highly parallelisable for efficient training on large datasets.

3.3.3 LightGBM (Light Gradient Boosting Machine): A gradient-boosted framework optimised for speed and memory efficiency through Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB). The histogram-based split-finding algorithm significantly reduces training time, and the method natively handles categorical features.

3.3.4 CatBoost (Categorical Boosting): A gradient-boosting algorithm designed to handle categorical variables efficiently via ordered target encoding. Its ordered boosting process reduces prediction shift and overfitting, making it particularly robust on datasets with complex categorical interactions.

3.3.5 Stacking Ensemble: A meta-learning approach in which the predictions of XGBoost, LightGBM, Random Forest, and CatBoost base models are used as inputs to a Linear Regression meta-model. Stacking combines the complementary strengths of diverse base learners and can achieve better generalisation than any individual model.

In addition to the above, Multiple Linear Regression, Ridge Regression, Support Vector Regression, Random Forest, Bagging Regressor, AdaBoost Regressor, and Gradient Boosting Regressor were also trained as comparison baselines.

3.4 Model Evaluation Metrics: Three standard regression metrics were used for evaluation. Mean Absolute Error (MAE) measures the average magnitude of prediction errors. Root Mean Squared Error (RMSE) gives greater weight to large deviations, making it sensitive to outlier predictions. The R^2 (coefficient of determination) quantifies the proportion of variance in the target variable explained by the model, with values closer to 1.0 indicating superior fit.

3.5 System Deployment: The best-performing model was serialised using joblib and integrated into a Flask web application. The backend exposes REST API endpoints that accept weather parameter inputs, perform the same preprocessing and scaling pipeline applied during training, and return a solar power prediction. The frontend, built with HTML, CSS, and JavaScript, provides a user-friendly form for data entry and result display. A MySQL database stores user accounts and prediction history, enabling longitudinal tracking of forecasts.

4. Results

Table 1 summarises the performance of the five primary models on the held-out test set, ranked by R^2 score.

Model	R^2	RMSE (kW)	MAE (kW)
LightGBM	0.8345	383.93	241.87
XGBoost	0.8272	392.28	249.43
CatBoost	0.8197	400.72	263.92
Stacking	0.7957	426.56	278.91
KNN	0.7137	504.97	345.89

Table 1: Model Performance Comparison

LightGBM achieved the highest R^2 score of 0.8345, explaining approximately 83.5% of the variance in solar power generation. Its RMSE of 383.93 kW and MAE of 241.87 kW were the lowest among all evaluated models, confirming it as the most accurate predictor in this study. XGBoost ranked second with $R^2 = 0.8272$, RMSE = 392.28 kW, and MAE = 249.43 kW — a small but consistent gap behind LightGBM, likely attributable to LightGBM's more efficient handling of the histogram-based splitting strategy on this dataset.

CatBoost placed third ($R^2 = 0.8197$), followed by the stacking ensemble ($R^2 = 0.7957$). Notably, the stacking ensemble underperformed the individual boosting models, suggesting that the meta-model's linear regression layer introduced a bottleneck when combining predictions from diverse base learners on this particular dataset. KNN performed least well ($R^2 = 0.7137$, RMSE = 504.97 kW), consistent with its known sensitivity to feature scale and its inability to generalise beyond locally dense training regions.

Feature selection identified the ten meteorological parameters with the strongest linear association with generated power: temperature at 2 m, relative humidity, total precipitation, total cloud cover, medium cloud cover, low cloud cover, shortwave radiation, wind speed at 80 m, angle of incidence, and zenith angle. These features collectively captured the dominant physical drivers of PV output variation.

On the deployed web application, sample predictions from the LightGBM model demonstrated consistent results: clear-sky, low-humidity conditions produced high power estimates (e.g., 788.98 MW for the sample inputs shown in the system output screens), while high cloud cover and humidity inputs yielded substantially lower predictions, aligned with the expected physical behaviour of solar panels.

5. Discussion

The results confirm that gradient-boosted tree algorithms — LightGBM in particular — are well-suited to solar PV forecasting tasks that rely on tabular meteorological data. LightGBM's superior performance over XGBoost and CatBoost can be attributed to its histogram-based learning strategy and GOSS sampling, which together reduce overfitting and training time while maintaining high predictive accuracy. The narrow gap between LightGBM and XGBoost ($\Delta R^2 \approx 0.007$) suggests that both are robust choices for this problem, and practitioner preference may be guided by infrastructure or latency requirements.

The underperformance of the stacking ensemble relative to its individual base models is somewhat counter-intuitive but aligns with findings in the literature for datasets of moderate size. When base model predictions are highly correlated — as they are for tree-based boosting methods applied to the same feature set — the meta-model gains limited additional signal and may instead overfit to training-set idiosyncrasies. Exploring more diverse base learners (e.g., combining gradient boosting with neural networks) may improve stacking performance in future work.

KNN's comparatively poor performance illustrates a key limitation of instance-based methods in high-dimensional regression: without an appropriate distance metric tuned to the relative importance of each meteorological variable, the nearest neighbours in feature space may not correspond to physically similar conditions. The ten-dimensional feature space used here likely amplifies this effect.

Compared with prior work, the LightGBM R^2 of 0.8345 is broadly consistent with state-of-the-art results for weather-driven solar forecasting on similarly sized datasets. The end-to-end Flask deployment is a practical contribution beyond what most academic studies provide, enabling real-world use by grid operators and energy planners without specialist machine learning expertise.

Current limitations include dependence on the static Kaggle dataset, which does not capture geographically specific or real-time weather dynamics. Prediction accuracy may degrade under extreme weather events not well represented in training data. The model also does not account for temporal dependencies (e.g., diurnal cycles or weather persistence), which could be addressed in future iterations using LSTM or transformer-based architectures.

6. Conclusion

This paper presented a machine learning pipeline for solar photovoltaic power forecasting aimed at supporting electrical grid stability. Among the algorithms evaluated — KNN, XGBoost, LightGBM, CatBoost, and a stacking ensemble — LightGBM achieved the best overall performance with an R^2 of 0.8345, RMSE of 383.93 kW, and MAE of 241.87 kW. Feature selection using f_regression identified the ten most predictive meteorological parameters, improving both model efficiency and generalisation.

The full system was deployed as a Flask web application with a MySQL backend, enabling real-time predictions from user-supplied weather data. The modular architecture supports future enhancements

including integration of real-time weather APIs, deep learning-based temporal models (RNN/LSTM), cloud deployment for scalability, and IoT sensor integration for direct measurement feeds.

These findings confirm that advanced gradient-boosting methods substantially outperform traditional regression approaches for meteorological-driven solar forecasting and that such systems can be deployed in production environments accessible to non-specialist end users. This work contributes a validated, reproducible baseline for solar PV forecasting research and a practical tool for renewable energy management.

7. References

Sharma, A., Singh, R., & Kumar, M. (2022). Solar power prediction using machine learning techniques. *Renewable Energy Journal, 45*(7), 1023–1035. https://doi.org/10.1016/j.renene.2022.03.045

Gupta, P., & Desai, S. (2021). A comparative study of regression models for solar energy forecasting. *Energy Reports, 7*, 457–468. https://doi.org/10.1016/j.egyr.2021.01.025

Zhang, L., & Chen, Y. (2020). Machine learning for solar energy generation forecasting: A review. *Renewable and Sustainable Energy Reviews, 117*, 109540. https://doi.org/10.1016/j.rser.2019.109540

Patel, S., & Kumar, R. (2021). Improving solar energy forecasting with XGBoost and LightGBM. *Energy AI, 3*(1), 100–110. https://doi.org/10.1016/j.egyai.2021.100045

Mehta, S., & Rathi, P. (2021). Solar power generation forecasting using CatBoost and K-nearest neighbors. *International Journal of Machine Learning and Data Science, 9*(4), 212–223. https://doi.org/10.1007/s13042-021-01345-6

Singh, J., & Gupta, R. (2021). Hybrid model for predicting solar power generation using weather data. *Renewable Energy, 38*(8), 1354–1367. https://doi.org/10.1016/j.renene.2021.04.032

Kumar, V., & Rani, S. (2020). Comparative performance of machine learning models in solar power prediction. *Renewable Energy, 125*, 356–364. https://doi.org/10.1016/j.renene.2020.02.041

Kumar, P., & Sharma, T. (2021). A neural network-based approach to predict solar energy output. *Journal of Renewable and Sustainable Energy, 13*(4), 457–466. https://doi.org/10.1063/5.0054321

Choudhury, S., & Jain, S. (2022). Real-time solar energy prediction using random forests. *International Journal of Renewable Energy, 28*(5), 88–94. https://doi.org/10.1016/j.ijre.2022.05.012

Bhatia, A., & Yadav, N. (2020). Forecasting solar power generation using ensemble methods. *Energy and Power Engineering, 8*(7), 174–183. https://doi.org/10.4236/epe.2020.87015

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). https://doi.org/10.1145/2939672.2939785

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems* (pp. 3146–3154). https://doi.org/10.48550/arXiv.1711.08893

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research, 9*, 2579–2605. http://jmlr.org/papers/v9/vandermaaten08a.html