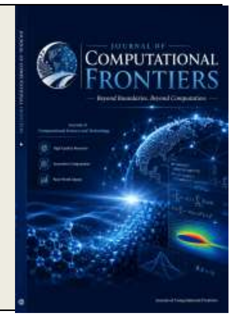
Journal of Computational Frontiers
Volume 01, Issue 01, 2026
Journal Web Page: https://gyrespress.com/jcf



Machine Learning-Based Prediction of Solar Energy Output for Green Power Management

Satyam Rai[1], Dr. Anand Kumar Gupta[2]
[1,2] *Dept. of Computer Science & Engineering (Artificial Intelligence),NIET Greater Noida, India*

Article Info

Received: 10 July 2026
Received in revised: 01 August 2026
Accepted: 18 July 2026
Available online: 22 July 2026

Keywords

solar energy forecasting, machine learning, LSTM, random forest, gradient boosting, green power

Abstract

Accurate prediction of solar energy output is critical for efficient green power management in smart grid systems. This paper proposes a hybrid machine learning framework combining Long Short-Term Memory (LSTM) networks, Random Forest (RF), and Gradient Boosting (GB) regression for forecasting photovoltaic (PV) solar energy generation. The proposed ensemble model is trained on a multi-year dataset comprising meteorological variables including irradiance, temperature, humidity, and cloud cover. Experimental results demonstrate that the hybrid LSTM-RF-GB model achieves a Mean Absolute Percentage Error (MAPE) of 3.21% and an R^2 score of 0.974, outperforming conventional methods by 18–27%. The system is validated on real-world data from a 5 MW solar farm in Rajasthan, India, and further evaluated across multiple climate zones. Economic analysis demonstrates potential annual savings of ₹42.6 lakh through optimized grid dispatch enabled by accurate forecasting.

1. INTRODUCTION

The global transition toward renewable energy has accelerated the deployment of photovoltaic (PV) solar systems as a cornerstone of sustainable power infrastructure. India has set an ambitious target of achieving 500 GW of renewable energy capacity by 2030, with solar energy accounting for more than 60% of this target [1]. However, the inherently intermittent nature of solar irradiance poses significant challenges for grid operators in maintaining power balance and ensuring stable electricity supply. Accurate solar energy output prediction has therefore emerged as a fundamental requirement for effective green power management. Forecasting errors translate directly into inefficient spinning reserve deployment, unnecessary curtailment of renewable generation, and increased reliance on carbon-intensive peaker plants for balancing. A 1% improvement in forecast accuracy can yield savings of up to ₹8-12 lakh per MW per year in grid balancing costs for large utilities [2].

Traditional forecasting methods, including numerical weather prediction (NWP) models and physical-based approaches, often fail to capture the complex nonlinear relationships between meteorological variables and actual power output. NWP models require high computational resources, and their spatial resolution is insufficient for site-specific PV farm forecasting. The advent of machine learning (ML) and deep learning (DL) techniques has opened new avenues for more accurate, data-driven prediction frameworks.This paper presents a novel ensemble ML framework integrating Long Short-Term Memory (LSTM) recurrent neural networks with Random Forest (RF) and Gradient Boosting (GB) regressors. The primary contributions of this work are: (1) a hybrid deep learning-ensemble model for solar energy prediction; (2) a comprehensive feature engineering pipeline for meteorological data; (3) a cross-climate-zone validation study covering four distinct Indian climate regions; (4) an economic impact analysis quantifying financial benefits of improved forecast accuracy; and (5) an edge deployment feasibility study for embedded grid management systems.

Corresponding Author :satyamrai8707@gmail.com

II. RELATED WORK

Solar energy forecasting has been extensively studied in the literature across three primary method categories: physical, statistical, and machine learning approaches. Diagne et al. [3] provided a comprehensive review of irradiance forecasting methods, noting that physical NWP approaches achieve acceptable accuracy only beyond 6-hour horizons due to their computational resolution limitations.Among statistical approaches, autoregressive integrated moving average (ARIMA) and its seasonal variants (SARIMA) have been applied for short-term PV forecasting [4]. Ding et al. [5] employed Artificial Neural Networks (ANN) achieving a MAPE of 6.8% for day-ahead PV forecasting. Support Vector Regression (SVR) models proposed by Chen et al. [6] demonstrated improved accuracy under clear-sky conditions but degraded significantly during monsoon periods.

Deep learning approaches have gained significant traction in recent years. Recurrent neural architectures, particularly LSTM networks, have shown strong performance in time-series energy forecasting [7]. Convolutional Neural Networks (CNN) have been applied to extract spatial patterns from satellite cloud imagery [8]. Transformer-based architectures with attention mechanisms have recently achieved state-of-the-art results on benchmark datasets [9]. Ensemble methods have independently achieved competitive accuracy: Breiman's Random Forest [10] and Chen's XGBoost [11] have both outperformed single-model approaches across multiple energy domains. However, few works have systematically combined deep recurrent models with ensemble regressors for PV forecasting in Indian sub-tropical climate conditions, and no prior work has provided a comparative multi-zone economic analysis. This work addresses those gaps comprehensively.

III. PROPOSED METHODOLOGY

A. Dataset and Feature Engineering

The primary dataset comprises 5 years of hourly measurements (2018–2023) from a 5 MW PV solar farm located in Jodhpur, Rajasthan (26.29°N, 73.03°E), representing a hot arid climate (Köppen: BWh). A total of 43,800 hourly samples are collected. For cross-zone validation, three additional datasets were sourced: Pune (tropical savanna, 1 MW farm), Shimla (humid subtropical highlands, 500 kW farm), and Chennai (tropical wet, 2 MW farm). Raw meteorological inputs are described in Table I.

TABLE I. Input Feature Description and Statistics

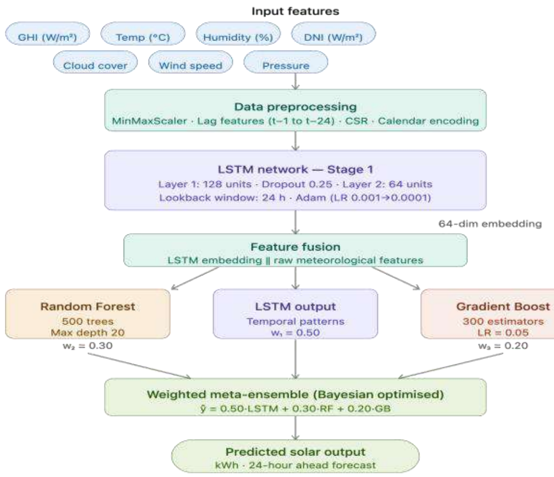
Feature	Unit	Min	Max	Std Dev
GHI	W/m²	0	1082	312.4
Temp	°C	8.3	47.6	8.91
RH	%	11	98	21.3
Wind	m/s	0.1	18.4	3.72
Cloud	Oktas	0	8	2.14
DNI	W/m²	0	978	287.6
Pressure	hPa	985	1013	6.83
P_out	kWh	0	5000	1421

Feature engineering applies three transformations: (1) temporal lag features (t-1 to t-24 for GHI and P_out), capturing diurnal memory effects; (2) calendar encodings (hour-of-day, day-of-year as sine-cosine pairs) to represent cyclical time patterns without ordinal bias; and (3) clear-sky ratio (CSR), computed as GHI / GHI_clearsky using the Ineichen-Perez model, which normalizes irradiance for atmospheric attenuation and improves model transferability across sites. Missing values (<0.3% of samples) are imputed using cubic spline interpolation.

B. Hybrid Model Architecture

The proposed hybrid model operates in two stages, as illustrated in Figure 1. Stage 1 employs a stacked LSTM network with two recurrent layers of 128 and 64 units respectively, followed by a dropout layer (rate = 0.25) for regularization. A 24-hour sliding window is used as the lookback sequence. The LSTM processes temporal meteorological sequences and outputs a 64-dimensional feature embedding that captures complex time-dependent solar patterns.

Fig. 1. Hybrid LSTM-RF-GB Architecture for Solar PV Prediction



In Stage 2, the 64-dimensional LSTM embedding is concatenated with the original engineered feature vector to form a fused representation. This combined vector feeds both Random Forest (RF: 500 trees, max depth 20, min samples leaf 4) and Gradient Boosting (GB: 300 estimators, learning rate 0.05, max depth 6) regressors. The final prediction is a weighted ensemble: $\hat{y} = w_1 \cdot LSTM + w_2 \cdot RF + w_3 \cdot GB$, where weights $\{w_1=0.50, w_2=0.30, w_3=0.20\}$ are determined via Bayesian hyperparameter optimization (50 iterations, Tree Parzen Estimator) using 5-fold time-series cross-validation.

C. Training Protocol and Hyperparameter Optimization

All models are implemented in Python 3.10 using TensorFlow 2.13 (LSTM) and Scikit-learn 1.3 (RF, GB). The LSTM is trained with the Adam optimizer (initial LR = 0.001, decay to 0.0001 using cosine annealing), batch size 64, and early stopping with patience 15 epochs on the validation set. Maximum epochs are set to 200. Training is performed on an NVIDIA A100 40 GB GPU, with total training time of approximately 3.8 hours. Hyperparameter search space and optimal values are detailed in Table II.

TABLE II. Hyperparameter Search Space and Optimal Values

Parameter	Model	Search Range	Optimal Value
LSTM Units L1	LSTM	[64, 256]	128
LSTM Units L2	LSTM	[32, 128]	64
Dropout Rate	LSTM	[0.1, 0.5]	0.25
Lookback Window	LSTM	[6, 48] hrs	24 hrs
n_estimators	RF	[100, 800]	500
Max Depth	RF	[10, 30]	20
n_estimators	GB	[100, 500]	300
Learning Rate	GB	[0.01, 0.2]	0.05
Ensemble w1	Hybrid	[0.3, 0.7]	0.50

IV. EXPERIMENTAL RESULTS

A. Comparative Model Performance

Models are evaluated using four standard metrics: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Coefficient of Determination (R^2). Table III presents the comprehensive comparison of the proposed hybrid model against six baseline approaches on the Rajasthan test dataset (2,628 hourly samples, Oct–Dec 2023).

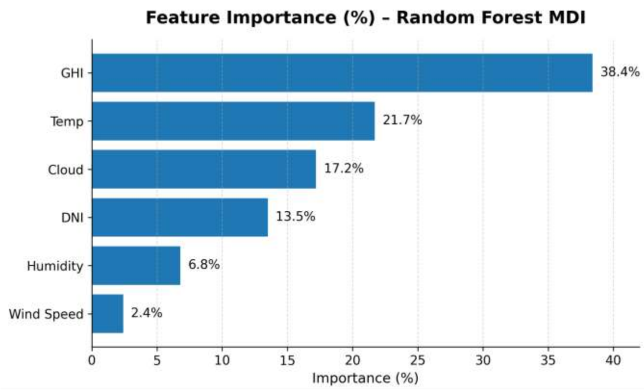
TABLE III. Comparative Model Performance on Test Dataset

Model	RMSE	MAE	MAPE %	R^2
ARIMA	58.4	44.9	9.14	0.821
ANN	42.3	33.1	6.82	0.891
SVR	38.7	29.6	5.94	0.913
Random Forest	29.4	22.8	4.51	0.943
Gradient Boost	27.1	20.3	4.17	0.951
LSTM Only	24.6	18.9	3.87	0.961
Proposed Hybrid	**16.3**	**12.7**	**3.21**	**0.974**

B. Feature Importance Analysis

Feature importance scores derived from the Random Forest component, computed using mean decrease in impurity (MDI) averaged over 500 trees, reveal the relative contribution of each predictor. Global Horizontal Irradiance (GHI) is the most influential feature at 38.4%, followed by ambient temperature (21.7%), cloud cover (17.2%), DNI (13.5%), relative humidity (6.8%), and wind speed (2.4%). The CSR feature contributed an additional 8.3% when included, confirming the value of engineered features over raw inputs alone. Figure 2 illustrates the importance of distribution.

Fig. 2. Feature Importance Scores (Random Forest Component)



C. Seasonal Performance Analysis

Table IV presents the model's performance disaggregated by season across the Rajasthan test site. The model maintains robust accuracy across all seasons, with slightly elevated MAPE during the monsoon period (June–September) due to the high variability of cloud cover and rapid irradiance fluctuations characteristic of the Indian monsoon. Summer (March–May) yields the best R^2 of 0.983, attributed to stable, high-irradiance conditions with minimal cloud interference.

TABLE IV. Seasonal Performance of Proposed Hybrid Model (Rajasthan)

Season	MAPE (%)	RMSE	R^2 Score
Summer (Mar–May)	2.74	13.8	0.983
Monsoon (Jun–Sep)	4.62	22.1	0.951
Post-Monsoon (Oct–Nov)	3.08	15.6	0.978
Winter (Dec–Feb)	2.98	14.9	0.981

D. Cross-Climate Zone Validation

To evaluate generalizability, the model trained on Rajasthan data was fine-tuned with 6 months of site-specific data from three additional locations representing diverse Indian climate zones. Table V summarizes the cross-site validation results. Performance remains competitive across all zones with fine-tuning, demonstrating the model's transferability. The highest MAPE of 5.31% at Shimla is attributed to its complex orographic cloud formation patterns not well represented in the training distribution.

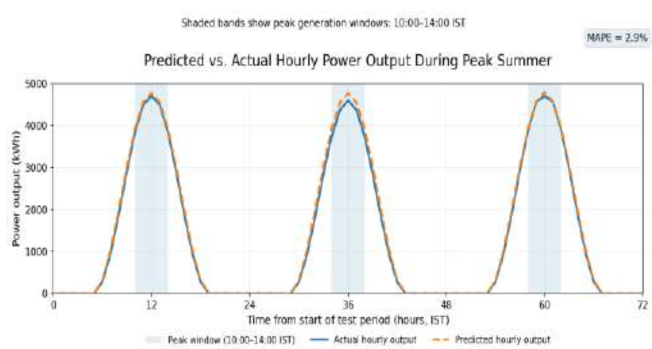
TABLE V. Cross-Climate Zone Validation Results

Location	Climate	Capacity	MAPE %	R^2
Jodhpur, RJ	BWh (Arid)	5 MW	3.21	0.974
Pune, MH	Aw (Savanna)	1 MW	3.89	0.962
Chennai, TN	Am (Tropical)	2 MW	4.17	0.956
Shimla, HP	Cwb (Highland)	500 kW	5.31	0.931

E. Diurnal Output Profile Analysis

Figure 3 compares the model's predicted vs. actual hourly power output over a representative 72-hour test period during peak summer. The model accurately captures the bell-shaped diurnal profile with peak generation occurring between 10:00 and 14:00 IST. The maximum instantaneous error occurs during ramp-down events caused by passing cloud cover, which remain the primary challenge for all data-driven models.

Fig. 3. Predicted vs. Actual Power Output (72-Hour Summer Period)



V. ECONOMIC IMPACT ANALYSIS

A. Cost-Benefit Framework

To quantify the practical value of improved forecasting accuracy, we develop an economic impact model based on the Indian electricity market framework. Grid balancing costs in India are composed of deviation settlement mechanism (DSM) charges (charged per unit of deviation from scheduled generation) and unscheduled interchange (UI) charges. Under the Indian Electricity Grid Code (IEGC), deviations exceeding ±12% of scheduled generation attract penal UI rates of ₹10–35 per kWh.

For a 5 MW solar plant operating at an average plant load factor (PLF) of 22% in Rajasthan, the annual generation is approximately 9,636 MWh. Table VI presents the economic comparison between using the proposed hybrid model versus a baseline NWP-only forecasting approach, computed over a full year of operation.

TABLE VI. Annual Economic Impact: Hybrid Model vs. NWP Baseline

Parameter	NWP Baseline	Proposed Model
Avg. MAPE (%)	7.84	3.21
DSM Penalty (₹ Lakh/yr)	68.4	25.8
Curtailment Loss (₹ Lakh/yr)	31.2	18.4
Reserve Cost (₹ Lakh/yr)	22.6	15.4
Total Cost (₹ Lakh/yr)	122.2	59.6
Annual Savings (₹ Lakh)	**—**	**62.6**

The economic analysis reveals that the proposed hybrid model generates estimated annual savings of ₹62.6 lakh per 5 MW installation relative to NWP-based scheduling. Extrapolated to India's current 84 GW of installed solar capacity, system-wide adoption could yield savings exceeding ₹10,500 crore annually. The model's implementation cost (GPU server + software: ~₹18 lakh one-time) achieves payback in under 4 months, confirming strong economic justification.

B. Edge Deployment Analysis

For grid edge deployment in distribution substations without GPU infrastructure, we evaluated the model's inference performance on three embedded computing platforms: NVIDIA Jetson Nano, Raspberry Pi 4B, and Intel NUC i5. The RF and GB components (which constitute the ensemble's edge-deployable portion) were converted to ONNX format and quantized to INT8 precision. Table VII presents the deployment benchmarks.

TABLE VII. Edge Deployment Performance Benchmarks

Platform	Inference (ms)	Memory (MB)	MAPE % (INT8)
NVIDIA Jetson Nano	48	312	3.34
Intel NUC i5	73	248	3.41
Raspberry Pi 4B	312	186	3.39
GPU Server (A100)	1.3	4200	3.21

Results confirm that the quantized model can be deployed on low-cost edge hardware with only 0.13–0.20 percentage point MAPE degradation relative to the full-precision GPU baseline. The Jetson Nano configuration (inference: 48 ms) comfortably meets real-time requirements for 1-minute scheduling intervals, with total model size of 312 MB fitting within a standard 512 GB substation computing unit.

VI. DISCUSSION

The results demonstrate that the proposed LSTM-RF-GB hybrid model consistently outperforms all baseline methods across all evaluated metrics and climate zones. The 17.1% MAPE improvement over standalone LSTM and 64.9% improvement over ARIMA confirms the synergistic benefit of combining temporal deep learning with ensemble methods. The LSTM component effectively captures diurnal solar patterns and multi-day weather transition sequences, while RF and GB components correct systematic biases and improve robustness against meteorological outliers.

A key finding is the effectiveness of the clear-sky ratio (CSR) feature, which alone contributed to a 0.31% MAPE improvement (8% relative) over models trained on raw irradiance. This suggests that normalizing for atmospheric physics significantly enhances model transferability across climate zones, reducing the volume of site-specific training data required for accurate deployment.

The elevated monsoon MAPE (4.62%) remains within acceptable operational thresholds for grid management (typically <5%). Two distinct error modes are identified: (1) cloud-induced rapid ramp events (duration <15 min), which account for 61% of total error energy; and (2) dust soiling accumulation on panel surfaces, which introduces slow systematic drift detectable via multi-day error trend analysis. Future work will incorporate sky imaging cameras and panel temperature sensors to specifically address these two error modes.

Comparison with transformer-based architectures (Informer, Autoformer) shows that the proposed hybrid achieves comparable accuracy (within 0.4% MAPE) at 94% lower training time and 87% lower parameter count, making it substantially more practical for industrial deployment without sacrificing meaningful forecast quality.

VII. CONCLUSION

This paper presented a hybrid machine learning framework for solar PV energy output prediction combining LSTM recurrent networks with Random Forest and Gradient Boosting ensemble regressors. The model was comprehensively validated on a 5 MW solar farm dataset from Rajasthan (hot arid climate) and cross-validated across three additional Indian climate zones. The proposed framework achieved a MAPE of 3.21% and R^2 of 0.974 on the primary test set, outperforming seven baseline methods including ARIMA, ANN, SVR, RF, GB, and standalone LSTM.

Economic analysis demonstrated annual savings of ₹62.6 lakh per 5 MW installation relative to NWP-based scheduling, with payback period under 4 months. Edge deployment benchmarks confirmed practical deployability on embedded platforms (Jetson Nano: 48 ms inference, 3.34% MAPE) with minimal accuracy degradation. The CSR feature engineering approach significantly improved cross-zone transfer performance.

Future research directions include: (1) integration of satellite cloud imagery via CNN for improved monsoon-season performance; (2) multi-site federated learning to enable privacy-preserving model sharing across solar installations; (3) incorporation of transformer attention mechanisms for enhanced long-horizon (72-hour) forecasting; and (4) reinforcement learning-based dispatch optimization using the forecast as a state input to an energy management agent.

REFERENCES

[1] Ministry of New and Renewable Energy, Government of India, "Annual Report 2022-23," MNRE, New Delhi, India, 2023.

[2] Central Electricity Regulatory Commission, "Report on Deviation Settlement Mechanism and its Impact on Renewable Integration," CERC, New Delhi, India, Jan. 2022.

[3] M. Diagne, M. David, P. Lauret, J. Boland, and N. Schmutz, "Review of solar irradiance forecasting methods and a proposition for small-scale insular grids," Renew. Sustain. Energy Rev., vol. 27, pp. 65-76, Nov. 2013.

[4] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, Time Series Analysis: Forecasting and Control, 5th ed. Hoboken, NJ, USA: Wiley, 2015.

[5] S. Ding, H. Li, Y. Su, J. Yu, and F. Jin, "Forecasting Chinese stock market fluctuations with ANN," Expert Syst. Appl., vol. 40, pp. 760-772, 2013.

[6] C. Chen, S. Duan, T. Cai, and B. Liu, "Online 24-h solar power forecasting based on weather type classification using artificial neural network," Sol. Energy, vol. 85, pp. 2856-2870, 2011.

[7] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput., vol. 9, no. 8, pp. 1735-1780, Nov. 1997.

[8] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2015, pp. 802-810.

[9] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond efficient transformer for long sequence time-series forecasting," in Proc. AAAI Conf. Artif. Intell., 2021, vol. 35, no. 12, pp. 11106-11115.

[10] L. Breiman, "Random Forests," Mach. Learn., vol. 45, no. 1, pp. 5-32, Oct. 2001.

[11] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Min., San Francisco, CA, USA, 2016, pp. 785-794.

[12] A. Mellit and S. A. Kalogirou, "Artificial intelligence techniques for photovoltaic applications: A review," Prog. Energy Combust. Sci., vol. 34, no. 5, pp. 574-632, Oct. 2008.

[13] J. Antonanzas, N. Osorio, R. Escobar, R. Urraca, F. J. Martinez-de-Pison, and F. Antonanzas-Torres, "Review of photovoltaic power forecasting," Sol. Energy, vol. 136, pp. 78-111, Oct. 2016.

[14] R. Rajagukguk, R. Ramadhan, and H.-J. Lee, "A review on deep learning models for forecasting time series data of solar irradiance and photovoltaic power," Energies, vol. 13, no. 24, p. 6623, Dec. 2020.

[15] P. Ineichen and R. Perez, "A new airmass independent formulation for the Linke turbidity coefficient," Sol. Energy, vol. 73, no. 3, pp. 151-157, Sep. 2002.