Solar Irradiance Prediction using Machine Learning

Time Series Analysis for Renewable Energy Optimization

Team Members:

Aditya Choudhuri: 16010123021

Agniv Dutta: 16010123029

Amandeep Singh Rathod: 16010123036

KJ Somaiya School of Engineering

Problem Statement

Predicting solar irradiance for Ghatkopar, Mumbai enables better renewable energy management, but several challenges complicate accurate forecasting.

Seasonal Fluctuations

Monsoon patterns and seasonal cloud cover create dramatic variations in solar radiation throughout the year.

Multiple Weather Variables

Temperature, humidity, wind speed, and precipitation interact in complex ways, requiring sophisticated modeling approaches.

Long-Term Stability

Maintaining forecast accuracy over 30-day horizons while avoiding model drift and capturing temporal dependencies.

Expected Outcome: ML-based 30-day solar irradiance forecast for strategic energy planning and optimization.

Why Solar Forecasting Matters

Grid Stability

Accurate solar predictions enable proactive grid balancing and reduce reliance on fossil fuel backup generation.

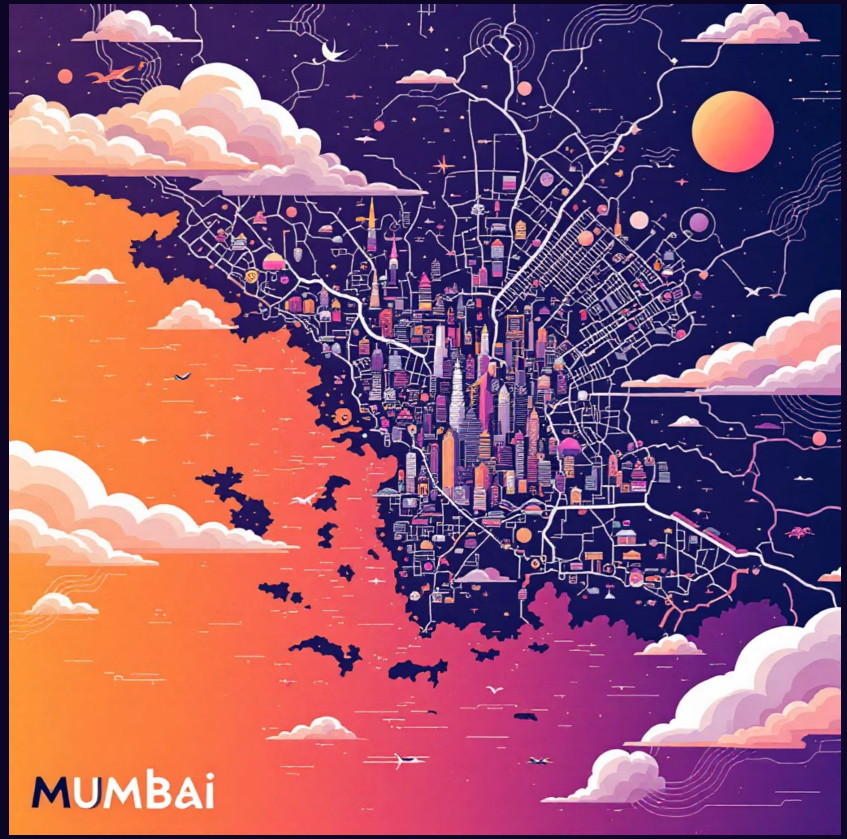
Storage Optimization

Forecasts guide battery storage charging schedules and maximize renewable energy utilization during peak generation periods.

Infrastructure Planning

Solar farm siting decisions and capacity planning benefit from long-term irradiance predictions and performance modeling.

Study Area: Ghatkopar, Mumbai



Geographic & Climate Profile

Location: Ghatkopar, Mumbai (19.0860°N, 72.9081°E)

Climate Zone: Tropical monsoon with high humidity and seasonal rainfall patterns

Data Period: October 2022 – October 2025 (3 years of continuous observation)

Temporal Resolution: Daily measurements capturing seasonal and inter-annual variability

Data Collection & Characterization

This study leverages NASA's POWER API, a globally distributed dataset of meteorological and radiative parameters derived from satellite observations and reanalysis models.

Data Source	Description
NASA POWER API	Prediction of Worldwide Energy Resources—satellite-based global coverage with validation against ground stations
Sample Size	1,096 daily records spanning three years with no temporal gaps
Variables	Solar irradiance (kWh/m²/day), temperature, relative humidity, wind speed, precipitation
Frequency	Daily aggregations optimized for solar energy applications and grid-scale forecasting

Methodology: End-to-End Pipeline

Our approach integrates data engineering, feature extraction, and multi-model ensemble techniques to maximize forecast performance.

01

Data Collection

Extract 3-year daily solar and weather records from NASA POWER API with quality validation.

02

Preprocessing

Handle missing values, outliers, and ensure data consistency across all meteorological variables.

03

Feature Engineering

Construct temporal, lag-based, and rolling statistical features to capture time-series dependencies.

04

Model Training

Train ARIMA, Random Forest, XGBoost, and Prophet models on historical data with cross-validation.

05

Evaluation

Assess models using RMSE, MAE, and MAPE metrics; select best performer for deployment.

06

Forecasting

Generate 30-day ahead predictions and analyze forecast uncertainty and seasonal patterns.

Data Preprocessing: Cleaning & Normalization

Robust preprocessing ensures data quality and prepares variables for machine learning models.

- **Missing Value Imputation**

  Forward-fill followed by backward-fill preserves temporal continuity in 1,096-day records.

- **Physical Constraints**

  Remove impossible values (negative solar irradiance) and flag anomalous meteorological combinations.

- **Normalization**

  MinMax scaling transforms all features to [0, 1] range, stabilizing model training and preventing feature dominance.

- **Stationarity Testing**

  Augmented Dickey-Fuller test identifies trending patterns requiring differencing before ARIMA modeling.

Feature Engineering: Capturing Temporal Dynamics

Engineered features extract periodic patterns and dependencies inherent in solar radiation time series data.

Temporal Features

- Month of year (1–12) and day of year (1–365)
- Season classification (winter, summer, monsoon)
- Day-of-week cyclicity to capture weekly patterns

Lag & Rolling Features

- Lag-1 and lag-7 irradiance (prior-day and prior-week values)
- 7-day rolling mean smooths short-term noise
- 7-day rolling std captures volatility trends

Model Development: Multi-Algorithm Comparison

Four complementary models leverage different mathematical frameworks to capture time-series dynamics and ensemble uncertainty.

ARIMA

Classical autoregressive approach models temporal autocorrelations and moving averages in univariate irradiance data.

Random Forest

Ensemble tree method captures nonlinear relationships across weather variables without assuming distributional assumptions.

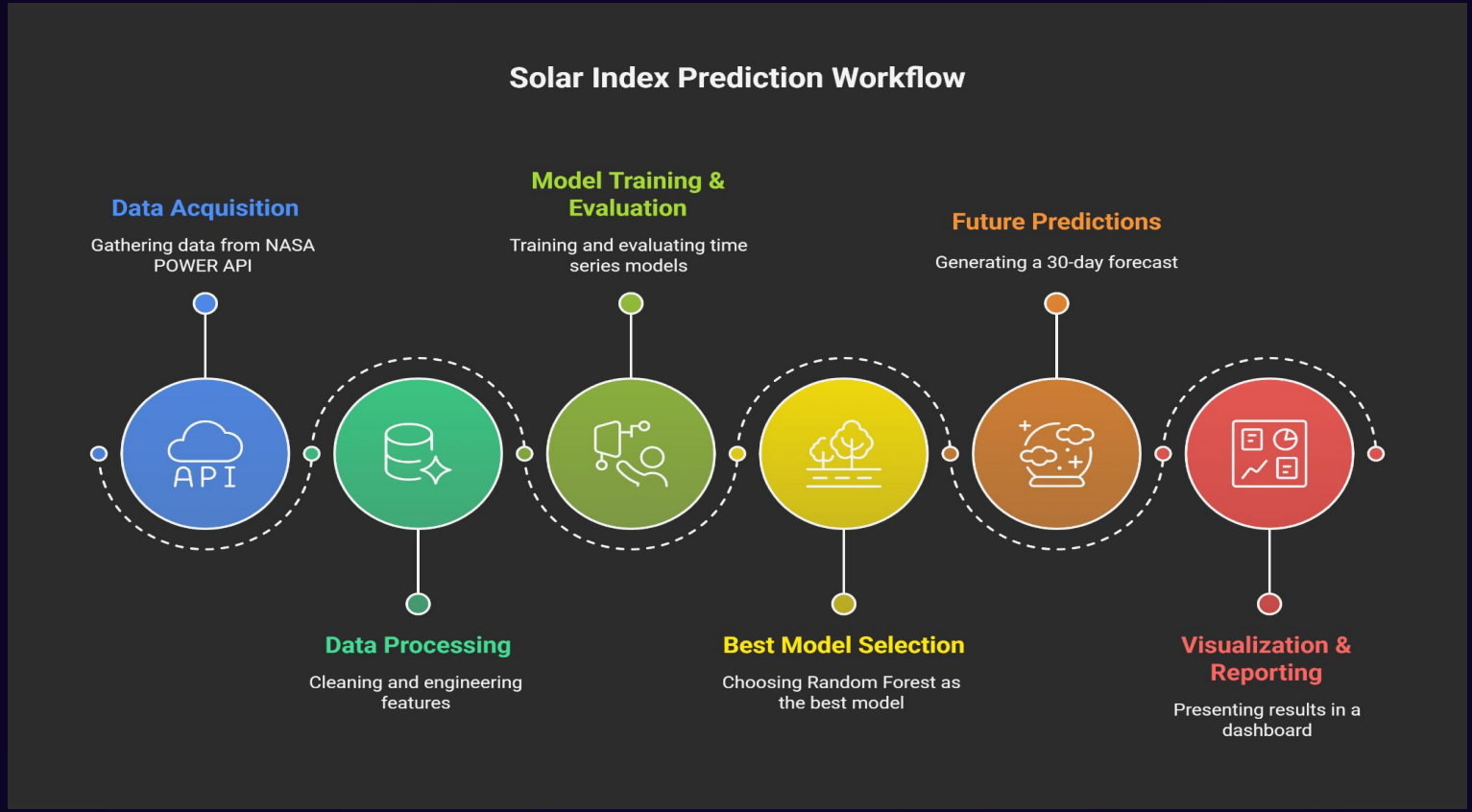
XGBoost

Gradient boosting framework iteratively refines predictions, handling feature interactions and delivering superior generalization performance.

Prophet

Facebook's seasonal decomposition model explicitly models trend, seasonality, and holiday effects for interpretable forecasts.

Block Diagram



Evaluation Metrics for Model Performance

Key Performance Indicators

We employed three critical metrics to assess and compare our machine learning models:

MAE (Mean Absolute Error): Measures average magnitude of prediction errors

RMSE (Root Mean Square Error): Penalises larger errors more heavily

R^2 Score: Indicates proportion of variance explained by the model

These metrics collectively enable robust comparison across different modelling approaches and help identify the most reliable predictor.

Model Comparison: Performance Rankings

Comprehensive evaluation across four machine learning approaches reveals clear performance distinctions:

Model	MAE	RMSE	R^2 Score	Rank
Random Forest	23.64	151.85	-0.021	1
XGBoost	23.72	151.95	-0.023	2
Prophet	24.17	152.07	-0.024	3
ARIMA	25.04	152.29	-0.027	4

Winner: Random Forest

Achieved lowest error rates across MAE and RMSE, demonstrating superior predictive accuracy for solar irradiance forecasting.

Dataset Overview & Characteristics

Key Statistics

1,096

Daily Observations

Oct 2022–Oct 2025

4.89

Mean Irradiance

kWh/m²/day

7.32

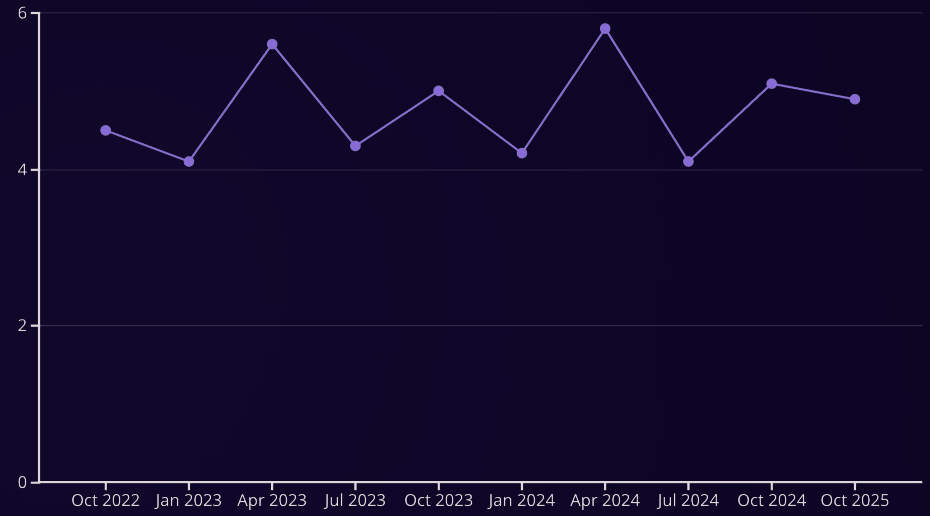
Maximum Value

Peak solar intensity

0.94

Minimum Value

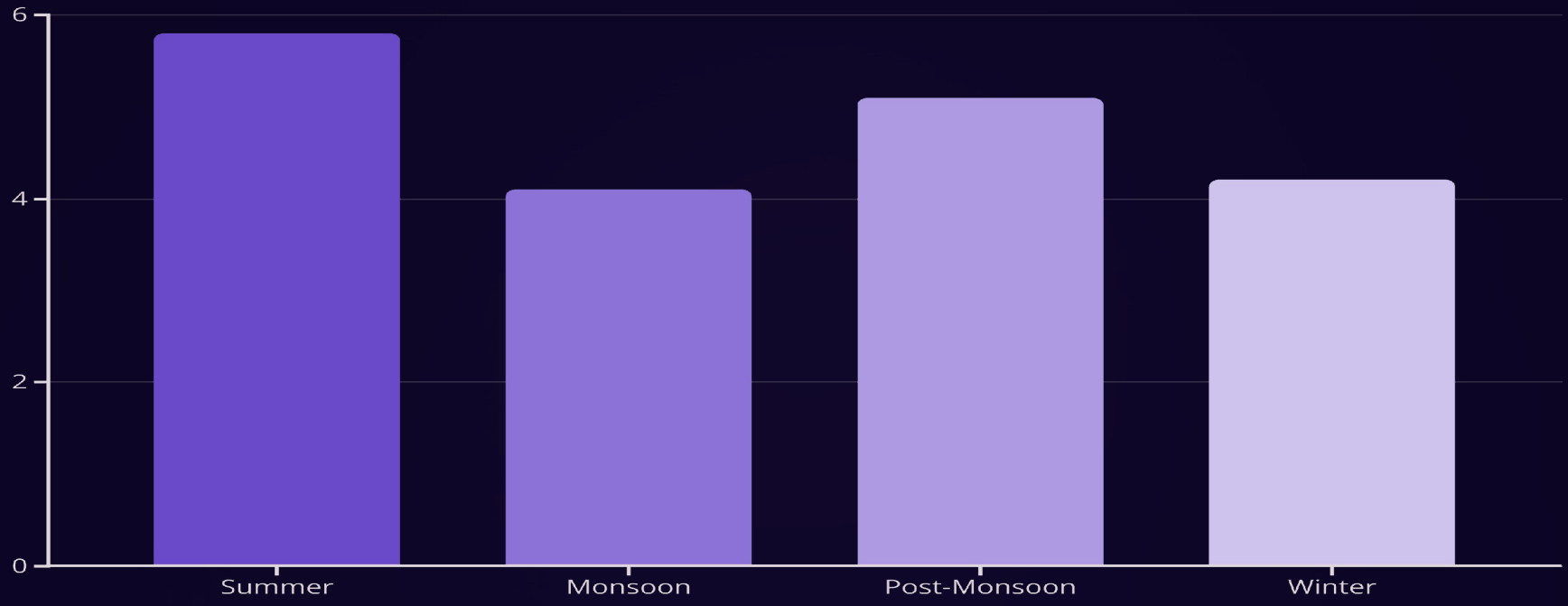
Lowest recorded



Standard Deviation: **1.45** — indicating moderate variability in daily solar irradiance values across the three-year observation period.

Seasonal Trends in Solar Irradiance

Clear seasonal patterns emerge across the annual cycle, with distinct variations driven by atmospheric and meteorological conditions:

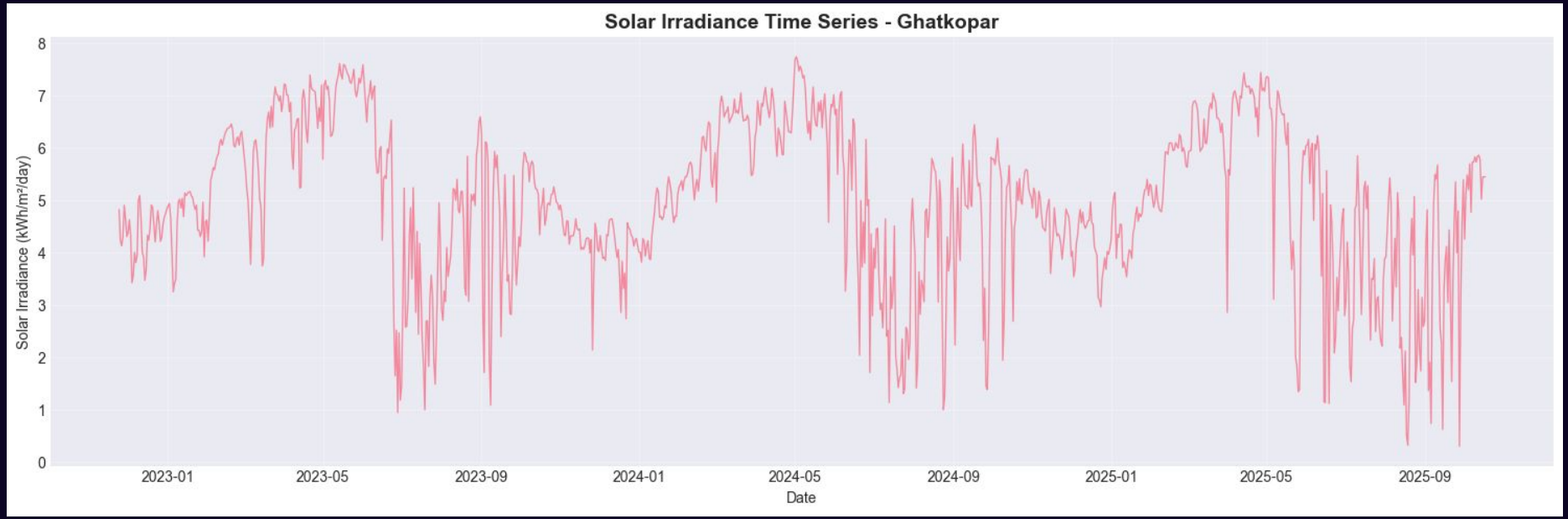


Summer Peak
Highest irradiance (~5.8 kWh/m²/day) due to clear skies and maximum sun angle

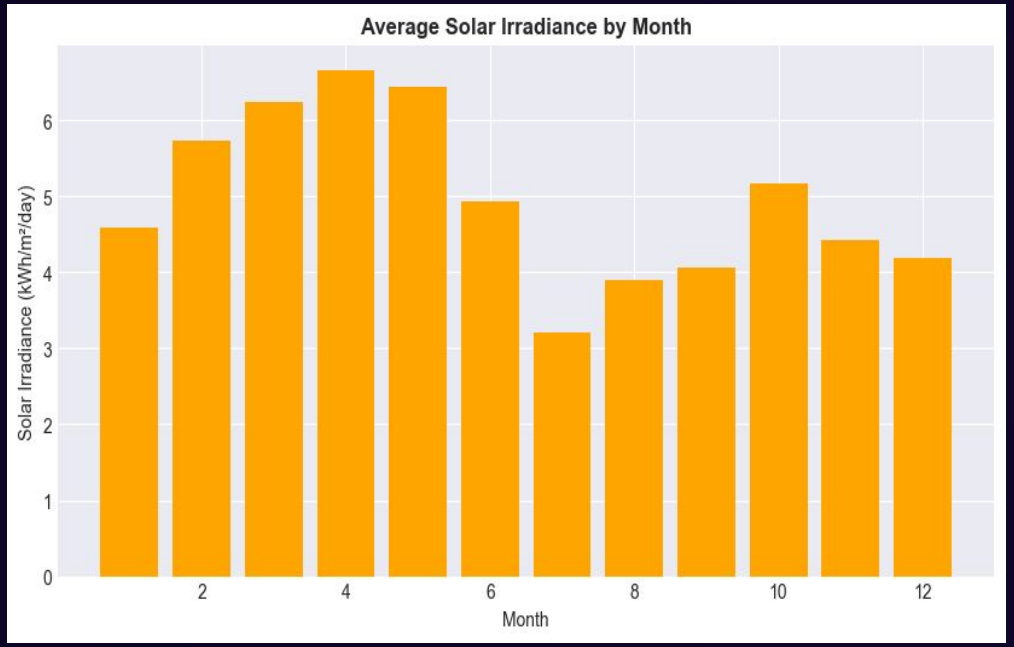
Monsoon Dip
Lowest values (~4.1 kWh/m²/day) caused by heavy cloud cover and rainfall

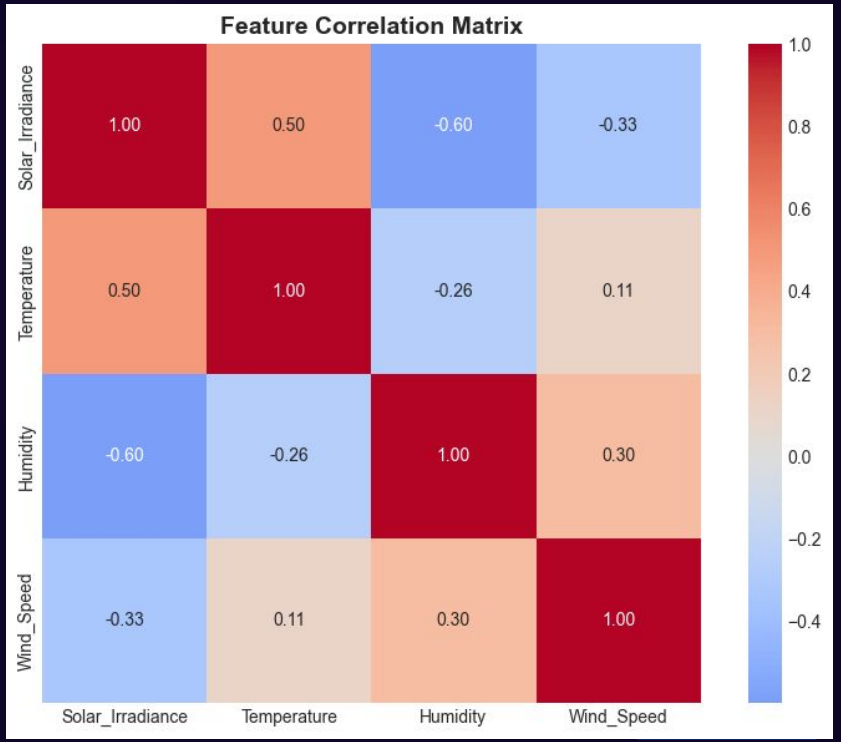
Winter Moderate
Lower angles and shorter days result in reduced irradiance (~4.2 kWh/m²/day)

Visualizations



Visualizations

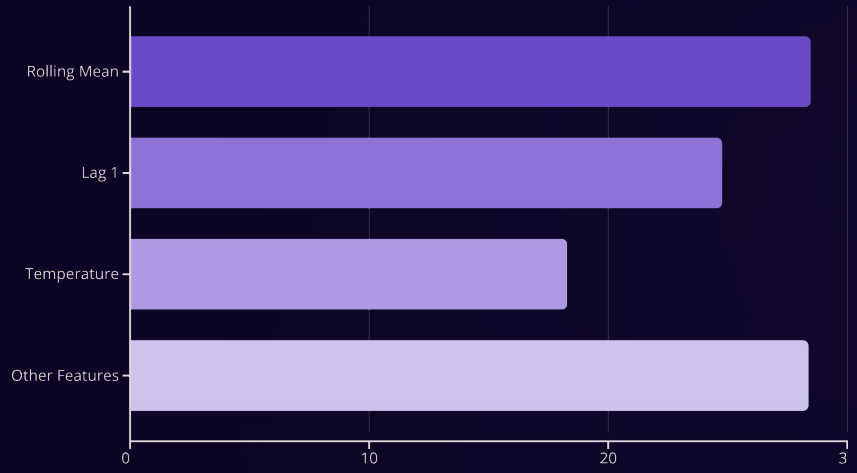




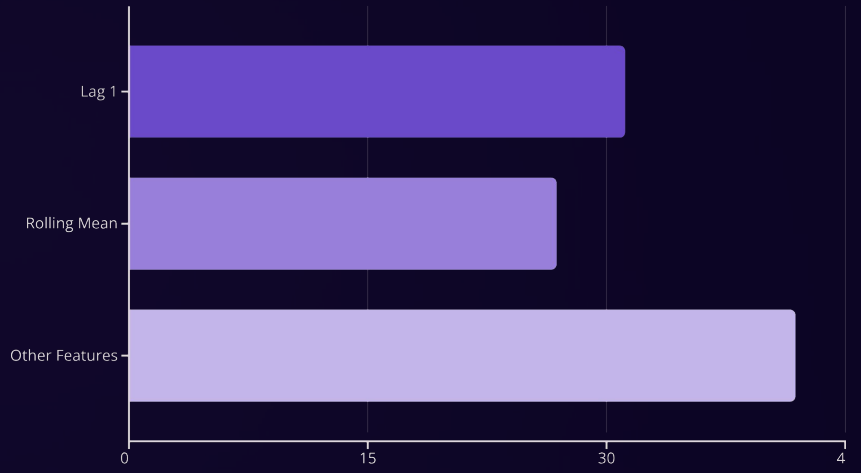
Feature Importance Analysis

Understanding which features drive predictive power helps refine future model architectures:

Random Forest Top Predictors



XGBoost Top Predictors



Key Insight: Both models prioritise temporal features (lag values and rolling averages), highlighting the strong autocorrelation in solar irradiance time series data.

Future Forecast: 30-Day Projection

Forecast Period: October 19 – November 17, 2025

4.65

Average Irradiance

kWh/m²/day

4.30

Minimum Expected

Lower bound

4.95

Maximum Expected

Upper bound

Trend Analysis: The forecast indicates a gradual decline in solar irradiance as we transition deeper into autumn. This aligns with expected seasonal patterns where day length decreases and sun angle lowers.

Model Limitations & Challenges

Negative R^2 Scores

All models exhibited R^2 values slightly below zero, indicating predictions marginally underperform a simple mean-based baseline. This suggests the underlying data contains significant unpredictable noise.

Limited Feature Set

Our models lacked access to critical meteorological variables such as cloud cover data, atmospheric pressure, humidity levels, and satellite-derived solar measurements that could substantially improve accuracy.

Dataset Noise

The NASA POWER dataset, while comprehensive, contains inherent measurement uncertainties and interpolation artefacts that introduce variance unrelated to true solar patterns.

Future Work & Enhancements

Several promising directions could significantly advance our solar prediction capabilities:

Deep Learning Models

Implement LSTM (Long Short-Term Memory) and CNN (Convolutional Neural Networks) architectures specifically designed for complex time series pattern recognition.

Multi-Location Expansion

Extend the forecasting system across diverse geographical regions to capture spatial variations and enable comparative regional analysis.

Real-Time Forecasting

Develop continuous prediction pipelines that integrate live meteorological data feeds for up-to-the-minute solar irradiance forecasts.

Streaming API Development

Build scalable REST/GraphQL APIs enabling third-party applications to access predictions seamlessly for grid management and energy planning.

References

[1] NASA POWER Project, “Prediction of Worldwide Energy Resources.” [Online]. Available: https://power.larc.nasa.gov/

[2] C. Voyant *et al.*, “Machine learning methods for solar radiation forecasting: A review,” *Renew. Energy*, vol. 105, pp. 569–582, 2017.

[3] N. Sharma *et al.*, “Predicting solar generation from weather forecasts using machine learning,” *Proc. IEEE SmartGridComm*, 2011.

[4] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” *Proc. 22nd ACM KDD*, 2016, pp. 785–794.

[5] S. J. Taylor and B. Letham, “Forecasting at scale,” *Amer. Statistician*, vol. 72, no. 1, pp. 37–45, 2018.

[6] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[7] G. Box *et al.*, *Time Series Analysis: Forecasting and Control*, 5th ed., Wiley, 2015.

[8] F. Pedregosa *et al.*, “Scikit-learn: Machine learning in Python,” *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[9] W. McKinney, “Data structures for statistical computing in Python,” *Proc. 9th Python Sci. Conf.*, 2010.

[10] J. D. Hunter, “Matplotlib: A 2D graphics environment,” *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 90–95, 2007.

[11] M. L. Waskom, “Seaborn: Statistical data visualization,” *J. Open Source Softw.*, vol. 6, no. 60, p. 3021, 2021.

[12] NREL, “Solar Resource Assessment and Forecasting,” *Tech. Rep. NREL/TP-5D00-78583*, 2021.

Conclusion

Key Takeaways from Our Research

Random Forest emerged as the top performer among four ML models, achieving the lowest prediction errors (MAE: 23.64, RMSE: 151.85)

Strong seasonal patterns were identified, with summer showing highest irradiance and monsoon the lowest

Temporal features dominated importance rankings across both Random Forest and XGBoost models

Current limitations point to opportunities for incorporating richer meteorological data and deep learning approaches

This foundation establishes a robust baseline for advancing solar energy forecasting through enhanced features, expanded geographic coverage, and sophisticated neural architectures.