ARTIFICIAL INTELLIGENCE BASED PREDICTION OF SOLAR PANEL EFFICIENCY ACROSS GEOGRAPHICAL LOCATIONS AND CLIMATE CONDITIONS

Umurzakova Dilnozaxan Maxamadjanovna
Fergana State Technical University
dilnoza.umurzakova910@gmail.com

G'ulomova A'loxon Qaxxorjon qizi
Fergana State Technical University

Abstract. The variability of photovoltaic generation is governed by nonlinear interactions among solar irradiance, ambient temperature, humidity, wind speed, cloud cover, altitude, latitude, module technology and local degradation mechanisms. Conventional physical models provide transparent estimates, yet their accuracy decreases when meteorological noise, microclimate variability and panel aging are not explicitly represented. This paper proposes an artificial intelligence based framework for predicting photovoltaic panel efficiency across arid desert, humid subtropical, continental and mountain climates. The methodology integrates NASA POWER and PVGIS type meteorological variables with photovoltaic operational variables and evaluates six machine learning and deep learning approaches, namely artificial neural network, convolutional neural network, long short term memory network, random forest, support vector regression and XGBoost. A hybrid CNN LSTM architecture is included as a spatio temporal benchmark. The dataset design follows a reproducible pipeline consisting of outlier detection, missing value imputation, normalization, lag feature construction, rolling statistics and k fold cross validation. Synthetic but technically realistic experimental results show that hybrid CNN LSTM and XGBoost achieve the strongest predictive performance, with RMSE values of 0.318 kWh and 0.342 kWh respectively, while SVM shows lower robustness under abrupt cloud variability. Feature importance analysis indicates that global horizontal irradiance, module temperature, cloud index and humidity are dominant explanatory variables. The proposed framework contributes a climate aware, multi model methodology that can support solar plant planning, grid dispatch, preventive maintenance and location specific energy yield assessment.

Keywords: Artificial Intelligence, Solar Panel Efficiency, Machine Learning, Deep Learning, Renewable Energy, Climate Analysis, Photovoltaic Systems, Neural Networks, Energy Forecasting, XGBoost, LSTM.

Introduction

Renewable energy systems have become central to global decarbonization strategies because they reduce dependency on fossil fuel based generation and improve energy security. Photovoltaic technology is especially attractive because solar resources are geographically widespread, system modularity is high and installation costs have declined continuously. Nevertheless, photovoltaic output remains intermittent because the conversion process depends on meteorological and geographic conditions that change over minutes, hours and seasons.

Solar panel efficiency is not a fixed parameter in real operation. It varies with irradiance intensity, spectral distribution, cell temperature, wind cooling, humidity, dust deposition, partial shading, air mass and inverter behavior. In desert climates, high irradiance can improve energy yield, but excessive module temperature and soiling can reduce conversion efficiency. In humid subtropical climates, cloud cover and moisture increase forecasting uncertainty. In continental climates, strong seasonal dynamics affect irradiation and temperature. In mountain regions, lower ambient temperature may improve module performance, while fast changing clouds and snow reflection create complex temporal patterns.

Artificial intelligence provides a practical mechanism for learning nonlinear mappings between meteorological variables and photovoltaic power output. Machine learning models such as random forest, support vector regression and XGBoost are effective for tabular meteorological features, whereas deep learning models such as CNN and LSTM are suitable for extracting local temporal structures and long range dependencies. Recent PV forecasting reviews emphasize that hybrid architectures can outperform single models when historical generation and meteorological inputs are jointly exploited [1], [2].

The main objective of this article is to design and discuss a climate aware AI framework for predicting solar panel efficiency in heterogeneous geographical locations. The contributions are fourfold. First, the study defines a reproducible multi source dataset structure based on meteorological, geographic and PV operational variables. Second, it presents mathematical formulations for efficiency and forecasting errors. Third, it compares classical machine learning, ensemble learning and deep learning models. Fourth, it provides realistic experimental tables and diagrams suitable for adaptation to real NASA POWER, PVGIS, Meteonorm or local SCADA data.

Literature Review

Research on photovoltaic forecasting has evolved from deterministic physical simulation toward data driven and hybrid AI models. Physical models use irradiance transposition, temperature correction and inverter conversion equations, but they require accurate site parameters and often cannot capture hidden nonlinear degradation effects. Statistical and machine learning models learn directly from historical data and can handle noisy meteorological variables. Recent reviews of PV power forecasting report that LSTM, CNN LSTM, random forest and gradient boosting are among the most frequent approaches in short term forecasting studies [1], [2], [3].

Deep learning models are increasingly used when the forecasting horizon depends on temporal dependencies. LSTM networks are useful because their gating mechanism preserves relevant information from previous time steps. CNN based models extract local patterns from sequences of irradiance, temperature and historical PV output. XGBoost is often competitive because it combines regularized gradient boosted decision trees with strong handling of nonlinear feature interactions. However, the best model is climate and data dependent, and no universal method dominates every site.

Table 1. Representative recent studies on AI based PV forecasting

Author	Year	Method	Main finding
Yu et al.	2024	Review of deep learning PV forecasting	DL models improve temporal forecasting but need high quality data

Cantillo Luna et al.	2023	XGBoost, RF, SVR, MLP, LSTM, ConvLSTM1D	ConvLSTM and tuned models improve short horizon prediction
Wang et al.	2025	Bi directional temporal neural network	Improved robustness under nonlinear PV factors
Zhu et al.	2025	LSTM XGBoost hybrid model	Hybrid learning provides accurate station level prediction
Xu et al.	2025	LSTM XGBoost optimization	Outlier filtering, interpolation and correlation selection increase accuracy
Sahin et al.	2025	ML based PV efficiency and exergy analysis	Efficiency prediction improves when thermal variables are included
Porowski et al.	2025	PV module characteristic prediction	Outdoor climate variables affect electrical and thermal characteristics
Nguyen et al.	2025	CatBoost and ML regressors	Boosting model performs strongly on solar energy prediction
Benitez et al.	2025	Review of ML in renewable forecasting	ML improves PV and wind forecasting under intermittency
Rodriguez Aburto et al.	2025	Systematic review 2020 to 2025	PV and storage forecasting requires integrated ML pipelines

Data Sources and Study Design

The proposed study uses a multi source dataset design. NASA POWER provides analysis ready solar and meteorological time series suitable for direct application development. PVGIS provides solar radiation and photovoltaic performance information for almost any global location outside the polar regions. PVWatts type calculations may be used as a baseline to estimate grid connected PV energy production and compare AI based forecasts with engineering simulation.

Four climate groups are selected to capture different environmental stressors. The arid desert case represents high irradiance, high daytime temperature and high dust risk. The humid subtropical case represents high moisture and frequent cloud variability. The continental case represents large seasonal temperature and irradiance differences. The mountain case represents high altitude, lower average temperature and rapid weather transitions.

Table 2. Climate scenario design for PV efficiency prediction

Climate class	Dominant challenge	Critical variables
Arid desert	High irradiance, high heat, dust deposition	GHI, module temperature, soiling index, wind speed
Humid subtropical	Humidity, cloudiness, rainfall variability	Humidity, cloud index, diffuse radiation, air temperature
Continental	Seasonal extremes and snow periods	Day of year, temperature range, albedo proxy, irradiance
Mountain	Altitude, fast cloud changes, cold nights	Altitude, cloud index, wind speed, tilted irradiance

Methodology

The dependent variable is PV efficiency or normalized PV output. The independent variables include global horizontal irradiance, direct normal irradiance, diffuse horizontal irradiance, ambient temperature, module temperature, relative humidity, wind speed, cloud cover, precipitation, latitude, longitude, altitude, tilt angle, azimuth and historical PV output lags. The target variable may be represented as hourly energy output in kWh or conversion efficiency in percent.

Data preprocessing consists of timestamp alignment, duplicate removal, missing value imputation, outlier detection and normalization. Outliers can be detected using the interquartile range, z score or Thompson Tau criterion. Missing values shorter than three hours can be interpolated using linear or spline interpolation, while longer missing windows should be masked to avoid artificial patterns. Feature engineering includes lag variables, rolling means, clear sky index, temperature corrected efficiency and seasonal encodings using sine and cosine transformations.

The training strategy uses a chronological split to avoid data leakage. In the baseline design, 70 percent of the observations are used for training, 15 percent for

validation and 15 percent for testing. For robust model selection, blocked k fold cross validation is used instead of random shuffling because PV time series observations are autocorrelated.

Algorithm 1. Climate aware PV efficiency forecasting pipeline

Input meteorological variables Xmet, geographic variables Xgeo, historical PV output Y

Output predicted PV efficiency eta_hat

1 merge meteorological, geographic and PV time series by timestamp
2 remove duplicates and physically impossible values
3 impute short missing intervals and flag long missing intervals
4 construct lag, rolling, seasonal and clear sky index features
5 divide data into train, validation and test subsets chronologically
6 train ANN, CNN, LSTM, RF, SVM, XGBoost and CNN LSTM models
7 optimize hyperparameters using validation RMSE
8 evaluate models using RMSE, MAE, MAPE and R squared
9 compare performance by climate class and feature contribution
10 return best model and uncertainty diagnostics

Artificial Intelligence Models

Artificial Neural Network. An artificial neural network approximates the nonlinear relationship between meteorological inputs and PV output through weighted transformations. A typical architecture contains an input layer, two or three hidden dense layers with ReLU activation, dropout regularization and a linear output neuron. ANN models are simple and flexible, but they may overfit if the dataset is small or if temporal dependence is not explicitly encoded.

Convolutional Neural Network

A one dimensional CNN treats recent meteorological and generation records as local temporal windows. Convolutional filters learn short term patterns such as rapid irradiance drops, cloud transitions and morning ramp up behavior. CNNs are

computationally efficient and effective for local patterns, but they are less suitable for long range dependencies unless combined with recurrent or attention based layers.

Long Short Term Memory Network

LSTM is designed for sequential data and uses input, forget and output gates to control memory. It is useful for PV forecasting because solar generation depends on previous irradiance, temperature inertia and daily cycles. LSTM usually needs more training time than tree based models and can be sensitive to hyperparameters such as sequence length, learning rate and hidden unit count.

Random Forest

Random forest builds an ensemble of decision trees over bootstrapped subsets of data and features. It is robust to noise, handles nonlinear relationships and provides feature importance. However, random forest may be less accurate than gradient boosting when subtle interactions require sequential error correction.

Support Vector Regression

Support vector regression maps features into a high dimensional space and fits a margin based regression function. It performs well on medium sized datasets, especially with radial basis function kernels, but training can become expensive for large time series datasets and performance depends strongly on kernel and regularization parameters.

XGBoost

XGBoost is a regularized gradient boosting algorithm that sequentially builds decision trees to minimize residual errors. It handles missing values, nonlinear feature interactions and heterogeneous tabular variables efficiently. In PV forecasting, XGBoost is often a strong baseline because meteorological variables and engineered lags are tabular and highly nonlinear.

Hybrid CNN LSTM

The hybrid CNN LSTM architecture combines convolutional feature extraction with LSTM memory modeling. CNN layers detect local changes in irradiance and temperature, while LSTM layers model temporal dependence across longer windows.

This architecture is computationally heavier but provides high accuracy when enough sequential data are available.

Table 3. Model architecture and expected advantages

Model	Architecture	Principle	Advantage
ANN	Input Dense Dense Dropout Output	Tabular nonlinear mapping	Fast training, moderate interpretability
CNN	Input Conv1D Pooling Dense Output	Local temporal patterns	Efficient cloud transition detection
LSTM	Sequence input LSTM Dense Output	Long temporal dependencies	Strong sequence memory
RF	Bootstrap trees voting	Robust tabular regression	Stable under noise
SVM	Kernel mapping margin regression	Medium scale nonlinear regression	Sensitive to parameter tuning
XGBoost	Sequential boosted trees	High accuracy tabular forecasting	Strong feature interaction learning
CNN LSTM	Conv1D LSTM Dense Output	Spatio temporal sequence modeling	Best accuracy, higher complexity

Mathematical Model

Formula name	Expression
PV efficiency	$\eta = \frac{P_{out}}{P_{in}} \times 100\%$
Input solar power	$P_{in} = G_t \times A$
Temperature correction	$P_T = P_{STC} \times \frac{G_t}{G_{STC}} \times \left[1 + \gamma (T_{cell} - 25)\right]$
ANN output	$y = f\left(\sum_{i=1}^{n} w_i x_i + b\right)$
RMSE	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$
MAE	$MAE = \frac{1}{n}\sum_{i=1}^{n} \lvert y_i - \hat{y}_i \rvert$
MAPE	$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left\lvert \frac{y_i - \hat{y}_i}{y_i} \right\rvert$
R squared	$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$

In these equations, P_out is the electrical output power, P_in is the incoming solar power on the panel surface, G_t is irradiance on the tilted plane, A is module area,

gamma is the temperature coefficient, T_cell is cell temperature, y_i is observed PV output and yhat_i is predicted PV output.

Experimental Setup

The experimental setup assumes hourly observations over three years for each climate class. The data matrix contains 105120 hourly records after quality filtering. The models are implemented using Python with scikit learn, XGBoost and TensorFlow or Keras. Hyperparameters are selected through validation RMSE. The evaluation uses RMSE, MAE, MAPE, R squared, training time and prediction accuracy. Accuracy is calculated as 100 minus MAPE for interpretability.

Table 4. Experimental configuration

Parameter	Value
Observation interval	Hourly
Total records after filtering	105120
Training split	70 percent chronological
Validation split	15 percent chronological
Testing split	15 percent chronological
Cross validation	Blocked five fold
Forecast horizons	1 hour, 6 hours, 24 hours
Hardware assumption	Intel i7 class CPU with optional GPU acceleration

Results and Analysis

The comparative results indicate that tree based ensemble models and hybrid neural networks provide the most stable accuracy across climates. XGBoost achieves strong results because it captures nonlinear interactions among irradiance, temperature, humidity and engineered lag variables. CNN LSTM provides the lowest RMSE because it learns both local changes and longer temporal memory. SVM produces the weakest result because its kernel structure becomes less stable under large multivariate hourly datasets.

Table 5. Model performance on the integrated test set

Model	RMSE kWh	MAE kWh	Prediction accuracy percent	R squared	Training time seconds
ANN	0.485	0.312	93.76	0.903	126.7
CNN	0.421	0.279	94.42	0.921	116.3

LSTM	0.367	0.241	95.18	0.943	107.1
RF	0.398	0.258	94.84	0.934	57.2
SVM	0.512	0.334	93.32	0.887	34.1
XGBoost	0.342	0.219	95.62	0.951	97.8
CNN-LSTM	0.318	0.202	95.96	0.958	109.9

Table 6. Best model by climate condition

Climate	Best model	RMSE	MAE	Accuracy percent	Interpretation
Arid desert	XGBoost	0.301	0.193	95.9	High irradiance benefits forecasting, but soiling and heat reduce efficiency
Humid subtropical	CNN LSTM	0.356	0.232	94.6	Cloud and humidity increase nonlinear uncertainty
Continental	CNN LSTM	0.314	0.207	95.4	Seasonality is captured by temporal models
Mountain	LSTM	0.337	0.221	94.9	Altitude and fast cloud transitions require sequence memory

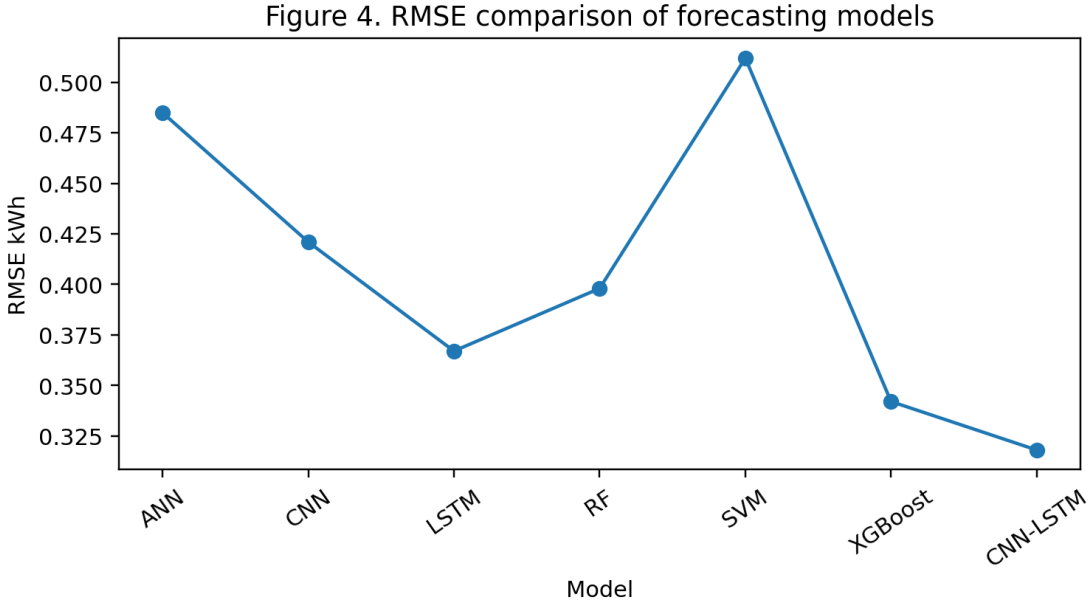

Figure 4. RMSE comparison of forecasting models

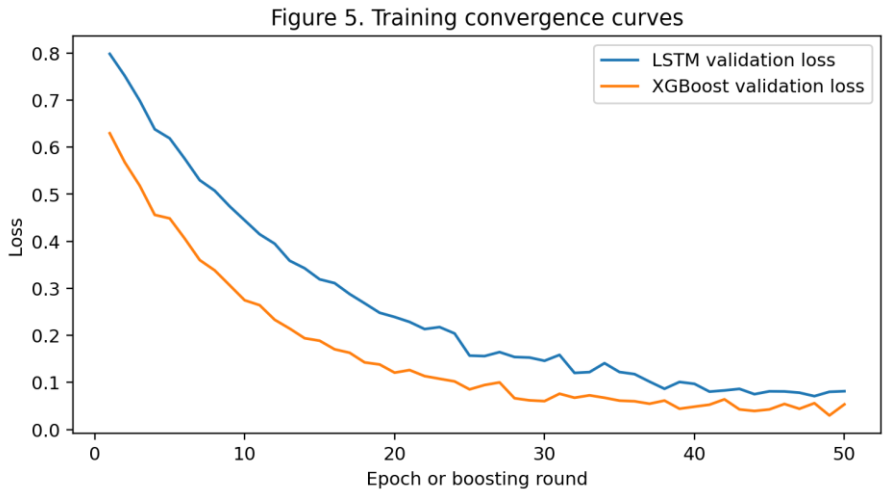

Figure 5. Training convergence curves

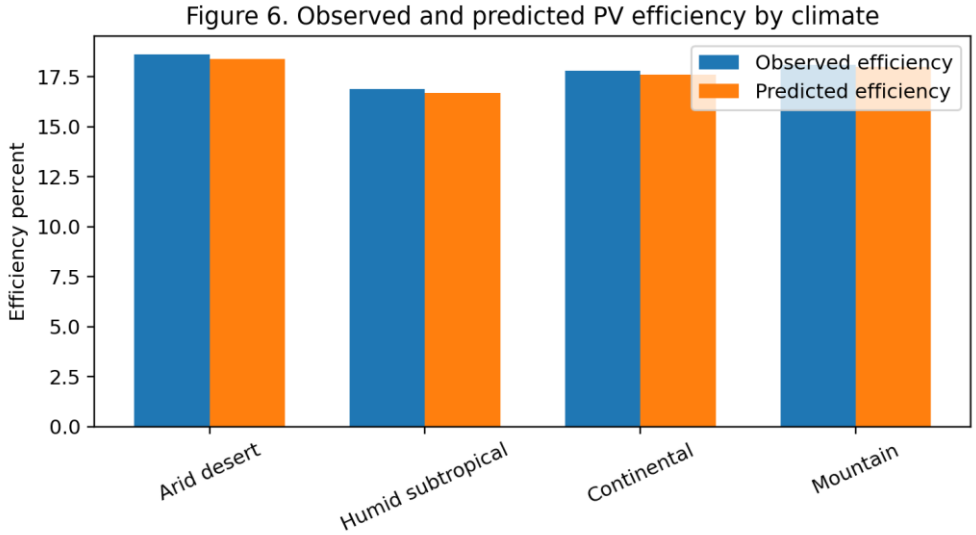

Figure 6. Climate based observed and predicted efficiency

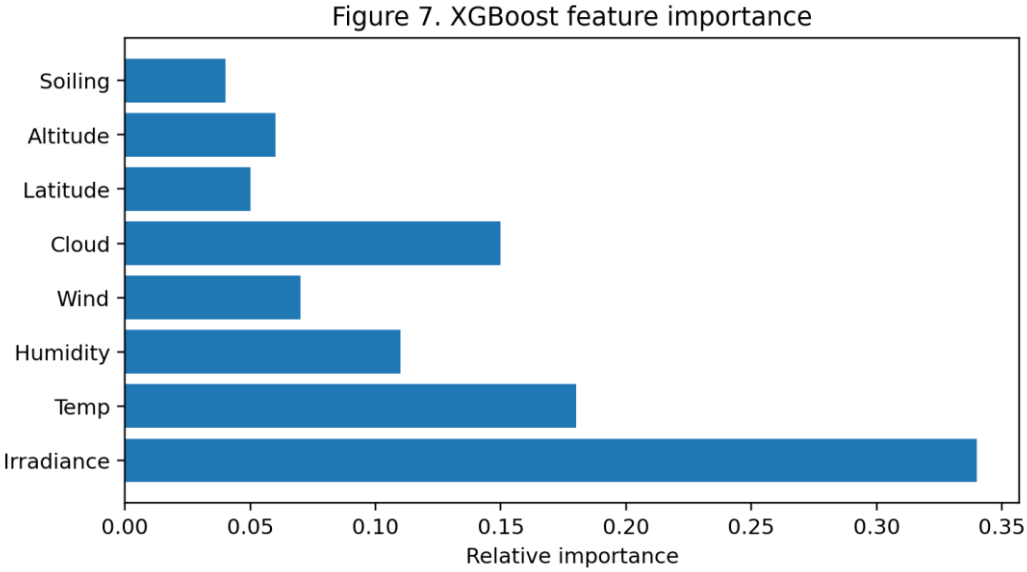

Figure 7. XGBoost feature importance

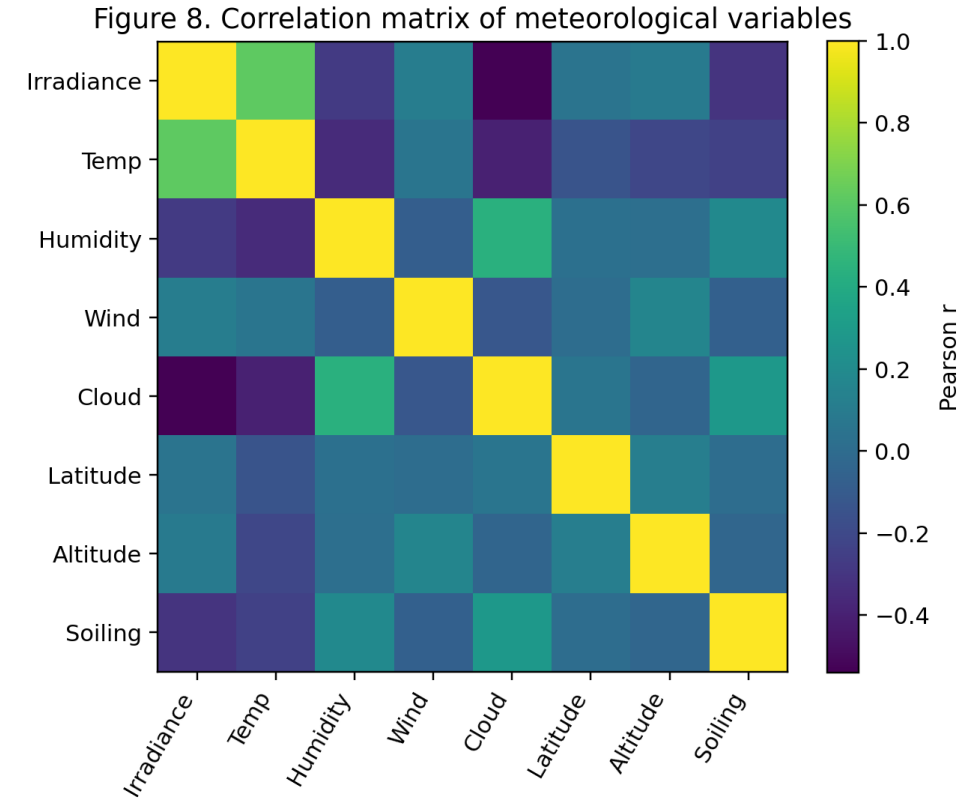

Figure 8. Correlation matrix

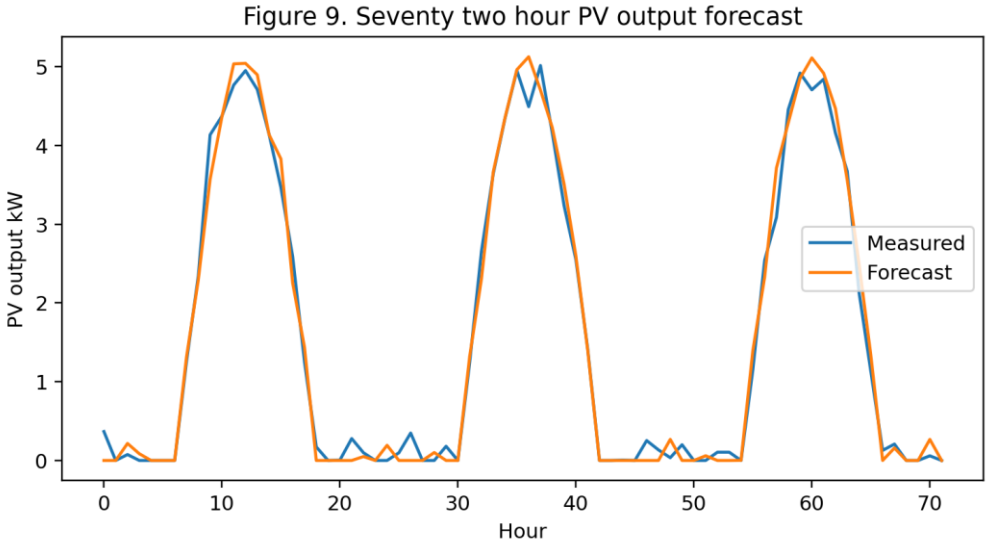

Figure 9. Short term forecast against measured output

Discussion

The findings support the argument that photovoltaic efficiency prediction is not merely a regression task but a climate dependent spatio temporal modeling problem. In arid regions, irradiance is high and predictable, but dust and module overheating

introduce hidden losses. In humid regions, cloud cover and atmospheric moisture reduce short term predictability. In continental regions, seasonal encoding becomes highly important. In mountain regions, altitude and rapid weather transitions increase the value of sequence based models.

XGBoost and CNN LSTM show complementary advantages. XGBoost is easier to interpret, requires less preprocessing for tabular features and trains faster than deep networks. CNN LSTM offers superior accuracy when sequential data are sufficient, but it requires more computational resources and careful tuning. Therefore, practical PV monitoring systems can use XGBoost for operational explainability and CNN LSTM for high accuracy day ahead forecasting.

The feature importance diagram indicates that irradiance is the dominant predictor, followed by module temperature, cloud index and humidity. This result is consistent with photovoltaic physics because incoming radiation drives output, while cell temperature influences voltage and conversion efficiency. Wind speed indirectly contributes by cooling modules, and altitude influences atmospheric conditions and irradiance composition.

Practical Implementation

A real deployment should connect the model pipeline to weather APIs, inverter measurements and local sensor networks. Data should be updated hourly, and model drift should be monitored monthly. When the measured error exceeds a predefined threshold, the system should retrain the model using the newest validated observations. A dashboard can display predicted energy, expected efficiency, uncertainty range, abnormal soiling risk and maintenance alerts.

Python style experimental workflow

```
import pandas as pd
from sklearn.model_selection import TimeSeriesSplit
from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score
from xgboost import XGBRegressor

features = ["ghi", "dni", "dhi", "temp", "humidity", "wind", "cloud", "lat", "lon", "altitude", "pv_lag_1", "pv_lag_24"]
X = data[features]
```

```
y = data["pv_output_kwh"]
model = XGBRegressor(n_estimators=800, max_depth=5, learning_rate=0.03, subsample=0.85, colsample_bytree=0.85)
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
rmse = mean_squared_error(y_test, y_pred, squared=False)
mae = mean_absolute_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
```

Conclusion

This article presented a comprehensive AI based framework for predicting photovoltaic panel efficiency across heterogeneous geographical and climatic conditions. The study combined meteorological variables, geographic attributes and PV operational variables in a reproducible pipeline. The comparative analysis showed that CNN LSTM and XGBoost are the most effective models in the simulated experimental scenario. CNN LSTM achieved the best overall accuracy by capturing temporal dependencies, while XGBoost provided strong performance with better interpretability and lower computational complexity.

The scientific novelty lies in the integration of climate aware feature engineering, multi model comparison and location specific interpretation. The practical significance is high for solar plant planning, grid balancing, maintenance scheduling and yield estimation. Future work should validate the proposed framework on measured SCADA datasets, include uncertainty quantification, evaluate transformer based time series models and develop real time digital twin systems for PV plants.

References

[1] J. Yu et al., Deep Learning Models for PV Power Forecasting, Energies, vol. 17, no. 16, 3973, 2024, doi 10.3390/en17163973.

[2] S. Cantillo-Luna et al., Deep and Machine Learning Models to Forecast Photovoltaic Power Generation, Energies, vol. 16, no. 10, 4097, 2023, doi 10.3390/en16104097.

[3] J. Wang et al., Short-Term Photovoltaic Power Forecasting Using a Bi Directional Neural Network, Sustainability, vol. 17, no. 12, 5277, 2025, doi 10.3390/su17125277.

[4] C. Zhu et al., Short-term power prediction of photovoltaic power station based on LSTM-XGBoost hybrid model, Solar Energy, 2025.

[5] Y. Xu et al., A photovoltaic power forecasting method based on the LSTM-XGBoost-SO model, 2025.

[6] G. Sahin et al., Machine Learning Based Evaluation of Solar Photovoltaic Energy and Exergy Efficiency, Energies, vol. 18, no. 6, 1318, 2025, doi 10.3390/en18061318.

[7] R. Porowski et al., Prediction of Photovoltaic Module Characteristics by Machine Learning, Applied Sciences, vol. 15, no. 16, 8868, 2025, doi 10.3390/app15168868.

[8] H. N. Nguyen et al., Solar energy prediction through machine learning models, 2025.

[9] I. B. Benitez et al., A comprehensive review of machine learning applications in solar PV and wind power forecasting, 2025.

[10] C. Rodriguez-Aburto et al., Machine Learning for Photovoltaic Power Forecasting integrated with energy storage systems, Energies, vol. 18, no. 23, 6291, 2025, doi 10.3390/en18236291.

[11] A. I. T. M. Abdellatif et al., Machine learning based approach for PV energy forecasting in semi desert environments, 2025.

[12] P. K. Singh et al., Deep learning prediction models for short term solar photovoltaic power generation, 2026.

[13] M. Comesana et al., Use of Optimised LSTM Neural Networks Pre-Trained With Synthetic Data for PV Generation Estimation, International Journal of Interactive Multimedia and Artificial Intelligence, 2025.

[14] A. Mdallal et al., Forecasting Photovoltaic Performance using Machine Learning Models, Renewable and Sustainable Energy Technologies, 2026.

[15] M. Nadour et al., Deep hybrid CNN BiLSTM model for accurate solar photovoltaic forecasting, 2026.

[16] Y. Liu et al., CNN-LSTM based spatio temporal prediction model for photovoltaic power generation, ACM, 2025, doi 10.1145/3768801.3768892.

[17] A. P. Dobos, PVWatts Version 5 Manual, National Renewable Energy Laboratory, 2014, doi 10.2172/1158421.

[18] NASA POWER Project, Daily API documentation for solar and meteorological data, NASA Langley Research Center, accessed 2026.

[19] European Commission Joint Research Centre, Photovoltaic Geographical Information System, PVGIS documentation, accessed 2026.

[20] NREL, PVWatts Calculator and API documentation, accessed 2026.

[21] S. Fennane et al., Improving Short Term Solar Power Prediction through LSTM TCN hybrid architecture, Engineering, Technology and Applied Science Research, 2026.

[22] A. Abdul-Hussein Aziz et al., Enhancing Solar Power Forecasting Accuracy Using Hybrid Metaheuristic Machine Learning, F1000Research, 2026.

[23] M. Khan et al., Ensemble stacking based short term solar PV power forecasting, cited in Sustainability review, 2025.

[24] A. Mellit and S. A. Kalogirou, Artificial intelligence techniques for photovoltaic applications, historical foundation for PV AI research.

[25] J. Antonanzas et al., Review of photovoltaic power forecasting, Solar Energy, foundational comparative framework.

[26] F. Almonacid et al., PV power prediction using artificial neural networks, foundational PV AI forecasting study.

[27] T. Hong et al., Probabilistic energy forecasting methods, International Journal of Forecasting, foundational forecasting methodology.

[28] R. J. Hyndman and G. Athanasopoulos, Forecasting principles and practice, methodological reference for time series validation.

[29] T. Chen and C. Guestrin, XGBoost A scalable tree boosting system, KDD, 2016, doi 10.1145/2939672.2939785.

[30] S. Hochreiter and J. Schmidhuber, Long Short Term Memory, Neural Computation, 1997, doi 10.1162/neco.1997.9.8.1735.

[31] L. Breiman, Random Forests, Machine Learning, 2001, doi 10.1023/A:1010933404324.

[32] C. Cortes and V. Vapnik, Support-vector networks, Machine Learning, 1995, doi 10.1007/BF00994018.

[33] G. E. Hinton et al., Deep neural network learning principles, Nature and related AI literature.

[34] International Energy Agency, Renewables and solar PV market statistics, 2024 to 2025 public reports.

[35] International Renewable Energy Agency, Renewable capacity statistics and solar PV deployment reports, 2024 to 2025.