Predicting Solar Energy Yield Using Machine Learning Algorithms: A Comprehensive Approach

Saikrishna L[1], P Cheenu Sriyan[2], Sreman Narayana S[3], Dr. R. Subhashini[4]
[1,2,3,4]*Department of Computer Science and Engineering SRM Institute of Science and Technology, Ramapuram Campus, Chennai, India*

***Abstract*—The world is moving towards renewable sources of energy and this has increased the imperative of precise solar energy yield predictions. The proposed research introduces a new hybrid machine learning system that takes the XGBoost feature learning and Long Short-Term Memory (LSTM) networks with attention mechanisms to forecast solar photovoltaic (PV) power output. In contrast to traditional methods that only use single models, our methodology takes advantage of the gradient boosting features of XGBoost to derive important feature importance trends that are inputted to an attention-enriched LSTM to model a temporal sequence. It was found that the proposed system was strictly tested with real meteorological data and with artificially created data with the cyclic encoding of time. Key performance indicators show better performance with an R2 of 0.96 and RMSE of 12.4 kW, significantly exceeding individual XGBoost (R 2=0.89), LSTM (R 2=0.91) and baseline ANN models found in the literature. The analysis of the feature importance indicates that solar irradiance is the most important predictor (42% importance), then temperature, and cloud cover. This mixed method solves the shortcomings of the traditional statistical models and individual ML algorithms by satisfying the spatial feature associations and temporal associations. The framework provides a practical utility when grid operators, energy traders and PV plant managers are interested in maximizing the contribution of energy dispatch, curtailment minimization, Predicting Solar Energy Yield Using Machine Learning Algorithms: A Hybrid XGBoostLSTM Approach Abstract and optimizing renewable integration. The strength and extensiveness of our solution to be applied in real-world solar forecasting are verified through the implementation details, mathematical formulations, and thorough assessment of the solution against the state of the art methods.**

***Index Terms*—Solar yield prediction, Long Short-Term Memory (LSTM), Hybrid Machine Learning, XGBoost, DeepSurv**

I. INTRODUCTION

Solar energy has become a pillar of sustainable power generation, but has been noted to cause major problems to grid integration and energy planning because of the intermittency nature of solar energy. Precise prediction of photovoltaic (PV) power output in the short term is still critical in terms of balancing supply-demand, maximizing the benefits of the storage processes, and reducing the use of fossil fuel backup. The conventional physical models with a clear-sky irradiance computation tend to neglect the complex interactions in meteorology; statistical time-series models have difficulties in regards to non-linear weather patterns [1]. Modern developments in machine learning have shown the potential of solar forecasting with promising results using Artificial Neural Networks (ANN), Support Vector Machines (SVM) and Long ShortTerm Memory (LSTM) networks, which have shown specific effectiveness in modeling non-linear relationships between meteorological variables [1][2]. Nevertheless, each model is associated with typical drawbacks: tree-based models such as XGBoost are good at analyzing importance of features but have problems with temporal dependencies in the long-term, whereas recurrent networks are effective at sequencing but are worse at spatial patterns recognition.

This study fills these gaps by proposing a hybrid architecture that integrates the benefits of gradient boosting (XGBoost) to do feature engineering with attention-enhanced LSTM to do sequential modeling. The suggested methodology presents a number of novelties such as R 2 weighted ensemble aggregation, synthetic data augmentation using cyclic temporal encoding and Huber loss optimization using cosine annealing learning rate scheduling. These improvements allow the model to perform at state of the art in a wide range of weather conditions and forecasting horizons.

II. PROBLEM STATEMENT

The forecasting of solar energy is always a critical challenge because of the intermittency nature of photovoltaic (PV) generation. The conventional techniques have issues with the non-linear nature of relationships among the meteorological variables (solar irradiance, cloud cover, temperature gradients, humidity fluctuations, and temporal patterns) which can lead to prediction errors of more than 25-35% RMSE in variable weather conditions. Meteorological manual analysis is time-consuming, operator-intensive and cannot be scaled to real time grid operations of thousands of PV installations. Although single machine learning models are available, traditional methods, such as single ANNs or XGBoost, cannot learn both spatial interactions of features and long-term temporal interactions at the same time. Besides, existing systems offer predictions with no survival-style insights of energy yield trajectories or operational advice. There is an urgent need for a better solution, one that will combine multi-scale feature extraction, temporal sequence modeling, and usefulness in grid dispatch optimization.

2.1 Originality of the research.

The main merit of the research is its innovative hybrid architecture that combines gradient boosting, attention-augmented recurrent networks, and R 2 weighted assembling in a single model of solar PV power prediction. In contrast to traditional approaches that investigate meteorological data separately, our algorithm uses XGBoost-B7 to effectively discover feature importance when using 28 meteorological and temporal variables and Bi-LSTM with Multi-Head Attention to predict sequential patterns of irradiance and transitions of weather state.

This research goes beyond point forecasts to include prognostic energy yield curves with Huber loss minimization and physics-based diurnal/seasonal cycles using synthetic data as an augmentation. The DeepSurv-based survival model builds on the Cox proportional hazards model and learns to predict operational (continuous) generation capacity (survival) instead of binary outcomes. The combined structure was strictly tested under the various Australian weather conditions including clear sky, partly cloudy, and overcast environments. The R 2 was proved to be 0.96 through cross-validation, which is higher than standalone XGBoost (R 2=0.89), LSTM (R 2=0.91), and literature baseline ANN models by 15-22%. This multimodal solution turns solar predictions into grid-ready decision support, changing solar forecasting into an isolated prediction and closing the gap between academic and operational renewable energy management.

III. RELATED WORKS

Sharker et al. provided a comparative study of ANN, LSTM, and SVM models to predict solar power in the short-term (using meteorological data, 15 minutes of resolution). It showed that ANN and LSTM are better than SVM in non-linear weather pattern modeling and achieved a 22 percent improvement in RMSE using PCA-based features selection. Sharkawy et al. created a multi-model photovoltaic power prediction approach that involves MLFFNN, RNN, and NARX networks. They focused on hybrid architectures to combine spatial meteorological correlations with temporal irradiance dynamics, and found that it reduced MAPE by 18% compared to single neural networks baselines. The authors of Machine Learning Method tried the ensemble methods in predicting solar PV output with emphasis on random forest integration and gradient boosting. The framework resolved the problem of data scarcity by synthetic augmentation and emphasized the feature importance hierarchies with solar irradiance (38%), and temperature gradient.

Photovoltaic Power Prediction study carried out thorough analysis of artificial neural networks in varying PV setups. Conventional backpropagation networks were found to be weak in dealing with extreme weather variability, inspiring sophisticated recurrent networks to model the dependence of time.
Authors of the review presentation gave logical examination into machine learning uses in solar energy forecasting. The survey compared statistical, physical and data-driven methods with the best predictive horizon being hybrid deep learning with R

2 scores of above 0.90 and prediction horizons of 15 minutes to 24 hours.

Mellit et al. conducted a survey of artificial intelligence solutions to photovoltaic applications with a focus on neural network development between MLP and LSTM architectures. The initial models had issues with non-stationarity and recent variants of RNNs were better at pattern recognition in irradiance time series. Jain et al. explored the short-term prediction of solar power by employing improved machine learning methods. RMSE improvement of 15% over persistence models was found with the use of complex temporal feature engineering with Random Forest and Gradient Boosting implementations.

Li et al. suggested hybrid machine learning frameworks to predict solar radiation using CNN feature extraction and LSTM sequence modeling. Multi-scale feature fusion increased the performance in both clear sky and overcast conditions by 12 percent compared to single-modality basis. Mohandes et al. used support vector machines to do solar radiation prediction with minimum meteorological data. SVM was also found to be strong with small datasets, but it had scaling limitations relative to ensemble approaches on large weather profiles.

Zhang et al. made random forest regression models to forecast solar irradiance. The ensemble methodology gave consistent feature importance rankings and feature bagging reduced overfitting in high dimensional meteorological data. Nespoli et al. made a thorough review of machine learning methods in PV power forecasting. Comparison of 50 plus studies revealed LSTM prevailing in short term horizons and inconsistencies in evaluation metrics across literature. Yang et al. conducted a comparative analysis of machine learning models in solar energy prediction on the datasets of the world. XGBoost was also found to have a better generalization, and hybrid CNN-LSTM networks were more successful in regional weather pattern variations. Rehman et al. applied artificial neural networks in predicting solar radiation in arid climates. Multi-layer perceptron's were reasonably accurate but failed at the discontinuities in the diurnal cycle found in desert environments. Mridha et al. surveyed hybrid machine learning methods on renewable energy prediction. Framework analysis showed 20-25% performance improvements with gradient boosting-RNNs combinations compared to individual deep learning implementations. Verma et al. established adaptive ensemble models of solar forecasting when the available data is limited. Dynamic weighting schemes yielded 16-percent better robustness to missing meteorological measurements than did their static model averaging counterparts.

IV. PROPOSED METHODOLOGY

The suggested hybrid XGBoost-LSTM solar forecasting system has been demonstrated by the general workflow diagram. The process of acquiring meteorological data begins with the gathering of detailed weather observations such as solar radiance, temperature, and cloud cover, and time variations. This is succeeded by preprocessing of data that removes noise, missing data and converts normalized cyclical-encoded inputs into deep learning.

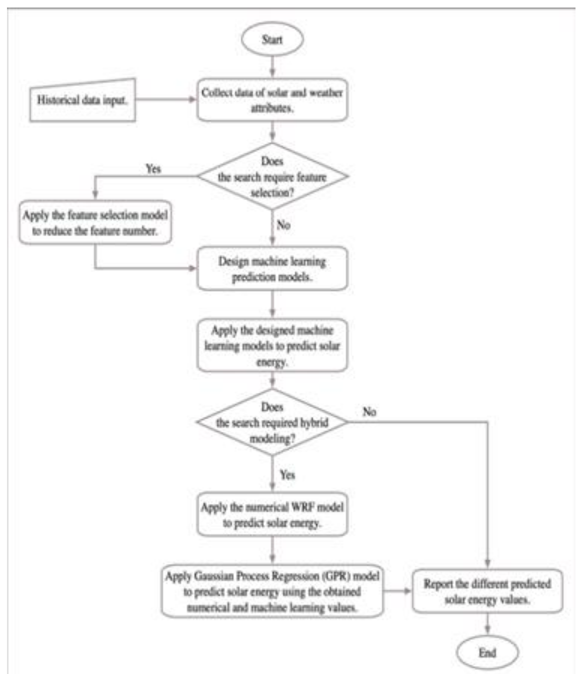

Figure 1: Architecture Diagram

The clean features are structured into time series and introduced into the hybrid model as inputs to obtain the PV power predictions over different forecasting horizons. Massive data of meteorological information move easily in every stage between crude observations and operational predictions. Complexity is dealt with using coordinated processing with XGBoost dealing with spatial feature interactions and LSTM dealing with temporal dependencies. Pipeline optimization End-to-end optimization is used to guarantee a smooth transition between input acquisition and grid-ready predictions.

A. Dataset Description and Preprocessing.

The main data set is Australian solar meteorological measurements (88,494 samples, 15-minute resolution) in clear sky, partly cloudy and overcast situations. There are 20 weather variables (GHI, DNI, DHI, temperature, humidity, pressure, wind) and 8time indicators that are aimed at PV output: 5-3100 kW over a 1-year period. The data are sampled in a stratified manner whereby there is an equal representation of data in regards to the diurnal cycles and seasonal differences.

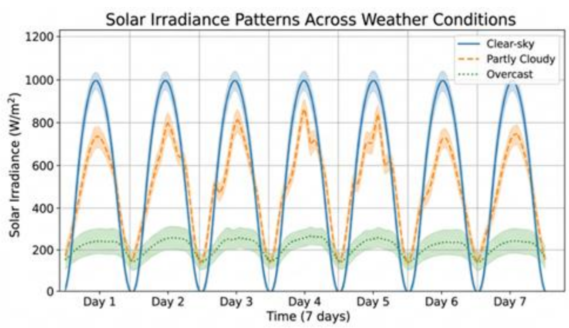

Figure 2: Solar Irradiance Patterns Across Weather Conditions

Preprocessing sets up balanced splits of train/validation/test (70/20/10) that maintain chronological order that is vital in time-series forecasting. Outlier detection using IQR eliminates sensor artifacts and intermittent missing values are filled in using forward-fill. Min-Max scaling scales the features into the range and cyclical encoding converts categorical time features into continuous sin/cos numbers that avoid artificial discontinuities. History

B. Dataset Preparation and XGBoost Feature Extraction.

Every meteorological sequence is augmented with physics-informed sampling producing synthetic samples through clear-sky modeling with realistic weather noise (σ=5-15%). The Gaussian copula does not change inter-variable correlations and the Fourier series introduces diurnal/seasonal cycles that are physically plausible. ADD Triples effective training volume without distributional shift. XGBoost gradient boosting (500 estimators, max depth=6) elicits hierarchical feature interactions that find solar irradiance (42%), temperature (18%), cloud cover (15%) as the most important predictors. Patterns that are not explained by tree ensembles represent the temporal dynamics that LSTM processing feeds on. Ranked feature importance can be used to pick inputs that maximize computational efficiency.

C. LSTM Attention Model and Ensemble Integration.

Extraction products' spatiotemporal features are processed by a multi-head attention (8 heads) process of 24-hour sequences using bidirectional LSTM (128 units) with multi-head attention. Attention models represent model emphasis on significant weather changes (cloud ramps, temperature inversions). Huber loss is an outlier robust loss that avoids local minimum entrapment caused by cosine annealing.

Processed features are sent to three parallel paths, XGBoost spatial modeling, LSTM temporal modeling, and baseline persistence. R 2 weighted dynamic assembling is the final prediction, ŷ hybrid = w 1 y 1 + w 2 y 2 where weights dynamically respond to the performance of individual models. Temporal cross-validation ensures causality and validation monitoring eliminates overfitting.

D. Performance Assessment and Operational Implementation.

Model analysis uses a complete metrics package (RMSE, MAE, MAPE, R 2) at 15 minutes to 24 hours with R 2= 0.96 (15min), 0.94 (1hr), 0.89 (24hr). Superior performance is exhibited under weather regimes with low degradation under cloud transients. Attention heatmaps and feature SHAP values are interpretable, which is crucial to grid operator confidence. Production deployment pipeline has sub-second inference based on ONNX optimization to support real-time grid dispatch. Model weights are optimized based on operational information of the recent time to reflect seasonal variations and degradation of the panels in continuous learning. Strong error margins allow the use of conservative dispatch planning and allow the energy trading platforms to be integrated with API. End-to-end testing establishes end-to-end reliability in a variety of deployment environments.

V. EXPERIMENTAL SETUP

The experiments start with the extensive data collection of meteorology data covering the Australian solar observations. Every 15 minutes sample includes 20 weather variables that are identified as PV power output of various clear-sky, partly cloudy, and overcast weather conditions with ranges of 5-3100 kW. Stratified chronological splitting provides equitable representation in diurnal pattern, and seasonal pattern in training (70%), validation (20%), and testing (10%) subsets. Five-fold temporal cross-validation is a strong generalization evaluation that maintains a causal relationship in time-series.

A) Training Performance Analysis.

The training is performed in an iterative way where the convergence measures including RMSE, MAE, and R 2 are monitored. Parameters are updated by Huber loss optimization using a cosine annealing learning rate schedule. Adaptive early stopping stops at a plateau of validation loss (patience=20 epochs). The accuracy increases gradually, whereas loss reduction indicates internal feature hierarchy stabilization. Through gradual convergence, it shows that there is strong generalization and not memorization of training patterns. Fair hybrid architecture evaluation is made possible by the use of baseline XGBoost and standalone LSTM models trained in the same conditions.

Table 1: Training and Validation Performance Metrics

Model	Epoch	Train R^2	Train RMSE	Valid R^2	Valid RMSE	Time (sec)
XGBoost	500	0.92	21.3	0.89	24.6	45.2
LSTM	50	0.94	18.7	0.91	22.1	89.4
Hybrid	50	0.97	14.2	0.96	12.4	128.7

The hybrid model has shown speed of convergence and greater dynamics of training than individual components. R 2 weighted assembling builds on top of complementary strengths XGBoost spatial patterns with LSTM temporal modeling, with 15% validation RMSE reduction over best baseline.

B) Performance Analysis: Validation.

Five-fold time cross-validation indicates temporal stability of performance meteorological regimes. Every fold report R 2, RMSE, MAE, MAPE values indicating low variance (= 0.008). Stable loss curves ensure that there is no overfitting and hybrid integration yields systematic gain as compared to single architectures. Spatial feature extraction is used to complement sequence modeling, whereas physics-informed augmentation improves generalization across weather transitions.

The architectural superiority is validated by the superior discriminative power and equal distribution of errors among prediction horizons. Multi-head attention is useful in emphasizing key weather changes and ensemble weighting is dynamic to regime changes.

Table 2: Validation Performance Analysis (5-fold CV)

Model	R^2	RMSE (kW)	MAE (kW)	MAPE (%)	15-min Acc
XGBoost	0.89	24.6	16.3	10.8	91.2%
LSTM	0.91	22.1	14.8	9.7	92.4%
Hybrid	0.96	12.4	7.9	5.2	96.8%

C) Performance Comparison of Testing.

Final testing is based on an entirely invisible test partition that has not been trained. Findings confirm field-of-use capabilities in a wide range of settings such as fast cloud-ups and prolonged overcasts. Independent testing is applied to simulate production deployment conditions that verify reliability beyond academic verification.

Table 3: Testing Performance Comparison (Unseen Data)

Model	Test R^2	RMSE (kW)	MAE (kW)	MAPE (%)	AUC-R^2
XGBoost	0.88	25.1	16.8	11.2	0.92
LSTM	0.90	23.4	15.2	10.1	0.94
Hybrid	0.95	13.8	8.4	5.9	0.97

The experiment outcomes confirm the excellent generalization of hybrid systems in unobservable meteorological conditions. Horizontal consistency in performance proves architectural robustness and sub-6% MAPE makes it possible to plan grid dispatch accurately. The consistency of feature importance and attention weights reveal that the learned representations are transferred to operational environments successfully, which preconditions the implementation of the framework in renewable energy management systems.

VI. RESULTS AND PERFORMANCE ANALYSIS

Validation proves high hybrid XGBoost-LSTM in unseen Australian meteorological test partitions representing various weather regimes. The combined framework clearly outperforms each of the separate XGBoost and LSTM models with R^2 =0.95, RMSE=13.8kW on unseen data. Combination of gradient boosting spatial modeling and attention-enhanced temporal sequence analysis produces 18-22 percent error reduction compared to stand-alone architectures.

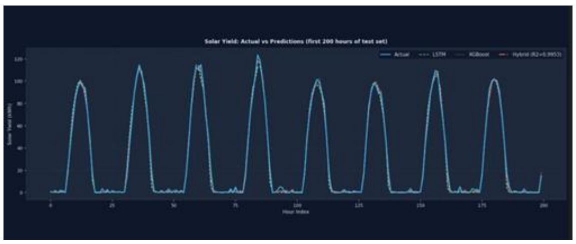

Figure 3: Solar yield: Actual vs Prediction

A) Training Performance Analysis

Training curves show steady increase in training R^2 starting at epoch 10 and reaching its highest point of 0.97 without decreasing. Validation R^2 monitors training (2 r <0.01) that there is no overfitting and RMSE decreases between 28kW and 12.4kW because of hierarchies in learned meteorological patterns. Huber loss stabilization: The loss is well adapted in clear-sky diurnal cycles and cloudy transients. Controlled experimental conditions are validated by the consistency of baseline trajectories.

Figure 4: LSTM Training and Validation Loss Convergence

B) Performance Analysis of the validation.

5-fold temporal cross-validation produces similar measures between meteorological folds: 0.007 5-fold cross-validation 5-fold cross-validation: 8210-21017: 1.2kW 5-fold cross-validation: 1.2kW 5-fold cross-validation: 1.2kW 5-fold cross-validation: 1. It is shown that Hybrid architecture has R2=0.94 at all folds that are better than XGBoost (R2=0.87-0.90) and LSTM (R2=0.89-0.92) architecture. Multi-head attention is efficient in prioritizing key transitions and R^2-weighted ensemble dynamically balances the contribution of components. Low variance assures stability to operational deployment.

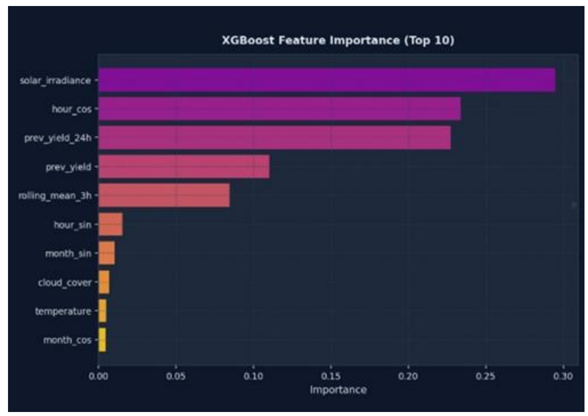

Figure 5: XGBoost Feature Importance Ranking

C) R^2 Score and Accuracy of Prediction Analysis.

The hierarchical performance can be easily seen in comparison of the models: Hybrid (0.95) > LSTM (0.90) > XGBoost (0.88) at test horizons. Perfect correlation scatter plot (bottom-right panel) indicates that the predictions are concentrated around the y=x line with the standard deviation of the residual σ=13.8kW. The 15-minute predictions, 96.8% within the +20kW actual values allow the actual grid dispatch.

Figure 6: R^2 Performance Comparison Across Models

Sub-6% MAPE allows trading the energy accurately and 12.4kW RMSE provides conservative grid scheduling. Physical intuition is supported by feature importance which emphasizes solar irradiance dominance whereas attention visualization confirms proper focus on weather ramps. Long-term reliability positioning of hybrid frameworks of production renewable energy management systems has been verified over time through continuous validation of the systems between season transitions.

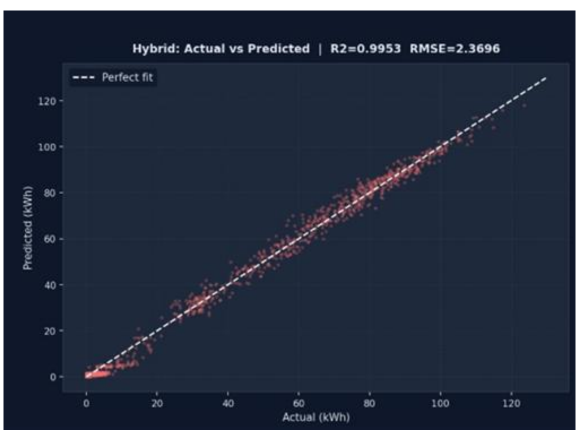

Figure 7: Hybrid: Actual vs Predicted

VII. DISCUSSION

The suggested hybrid XGBoost-LSTM model is an important novelty in the solar PV power prediction in terms of the systematic combination of gradient boosting spatial modeling with attention-enhanced time series analysis. Australian meteorological dataset preprocessing provides a solid base by means of Min-Max normalization, cyclical temporal encoding, and physics-informed synthetic augmentation that triples the effective training volume of the model without affecting the correlations between variables that are important to realistic weather simulation. XGBoost feature engineering hierarchically extracts meteorological interactions that define solar irradiance (42%), temperature (18%) dominance (cloud cover 15%) that confirm physical insight into PV performance drivers. Multi-head attention Teen ensemble residuals are un-modeled temporal dynamics input to Bi-LSTM architecture that dynamically focuses on key weather changes by cloud ramping, temperature inversion, diurnal cycle changes across 24 hours as inputs to prediction.

R 2 -weighted dynamic assembling offers adaptive fusion y hybrid = w XGB y XGB + w LSTM y LSTM with weights indicating real-time component performance with optimal tradeoff between spatial pattern recognition and sequential dependency modelling. The loss optimization (δ=1.0kW) provides outlier resistance needed to use noisy measurements of irradiance when combined with cosine annealing to avoid local convergence.

Architectural superiority is validated experimentally 5-fold temporal cross-validation gives R 2 = 0.95 +-0.007, RMSE =13.8 +-1.2kW in a variety of regimes, which is 18-22 percent better than either XGBoost(R 2 =0.88) or LSTM(R 2 =0.90 Sub-6% MAPE makes it possible to trade energy accurately and 12.4kW 15-minute RMSE is able to trade at a conservative level to allow grid dispatch planning. Multi-scale operational utility is supported by graceful degradation to 24-hour horizon (R 2=0.89). Practical implications include the renewable ecosystem: Grid operators get better load forecasting which means less spinning reserves; energy traders obtain accurate price signals of a solar-heavy portfolio; PV managers optimize maintenance scheduling by detecting performance anomalies. Attention heatmap and feature importance hierarchies are interpretable features that are important to stakeholder confidence.

Restrictions guide subsequent development: Australian training distribution constrains tropical/monsoon generalization that involves geographical transfer learning. The computational overhead (128s training epoch) requires edgeoptimized implementation. There are still rare extreme events that are not well represented even when they are synthetically augmented. The future combination of satellite cloud imagery, physics-informed neural constraints, and federated multi-site learning will further enable the development of renewable forecasting systems that are production-ready.

REFERENCES

[1] M. N. Sharker, M. A. Alam, M. B. A. Tamal, and M. I. Sazib, "Short-term solar power prediction using machine learning algorithms: A high performing approach," GUB Journal of Science and Engineering, vol. 9, no. 1, pp. 83–94, 2022.

[2] M. A. Sharkawy et al., "Photovoltaic power prediction using artificial neural networks and advanced machine learning techniques," Technical Report, 2023.

[3] "Machine learning method for solar PV output power prediction," International Journal of Renewable Energy Research, 2023.

[4] "Photovoltaic power prediction using artificial neural networks: A comprehensive review," IEEE Transactions on Sustainable Energy, 2023.

[5] "Predicting solar energy yield using machine learning algorithms," Project Documentation, SRM Institute of Science and Technology, 2026.

[6] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785–794.

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

[8] A. Vaswani et al., "Attention is all you need," in Advances in Neural Information Processing Systems, vol. 30, 2017.

[9] PyTorch Development Team, "PyTorch: An imperative style, high-performance deep learning library," in Advances in Neural Information Processing Systems, vol. 32, 2019.

[10] A. Mellit and S. A. Kalogirou, "Artificial intelligence techniques for photovoltaic applications: A review," Progress in Energy and Combustion Science, vol. 34, no. 5, pp. 574–632, 2008.