



A Hybrid XGBoost-LSTM Approach to Day-Ahead Solar Forecasting across Diverse Climatic Zones: A Comparative Study of Six Nigerian Cities

Najashi B. Gafai, Sadiq M. Kawuyo
Department of Electrical and Computer Engineering,
Baze University, Abuja, Nigeria

ABSTRACT

Accurate day-ahead solar forecasting is essential for integrating photovoltaic generation into Nigeria's power grid. This paper presents the development and evaluation of a hybrid Long Short-Term Memory (LSTM) – Extreme Gradient Boost (XGBoost) model for solar irradiance forecasting across selected Nigerian cities. The study utilizes satellite-based meteorological and solar irradiation datasets obtained from NASA POWER and European commission PVGIS covering the period from 2014-2024. The datasets included Global Horizontal irradiance (GHI), temperature, relative humidity, and wind parameters. The methodology involved data acquisition, preprocessing, normalization, feature engineering, and model training using google Colab. The LSTM network was deployed to capture temporal dependencies and sequential characteristics in solar irradiance data, while the XGBoost algorithm was utilized to model nonlinear relationships among meteorological variables. The outputs of both models were integrated using a weighted average hybrid approach to improve forecasting performance, with the model performance evaluated using the standard metrics. The results showed that the hybrid model outperformed the standalone LSTM and XGBoost models across all selected cities with MAE values between 0.027 and 0.042 across the six cities and R-squared exceeding 0.95 for the four inland locations (Abuja, Enugu, Kano, Maiduguri) and dropping to 0.916 for Lagos and 0.878 for Port Harcourt. XGBoost carried most of the predictive weight ($\alpha = 0.85-0.95$), though the LSTM contributed more in coastal climates where weather variability benefits from temporal memory. The Hybrid model achieved lower forecasting errors and higher prediction accuracy, particularly in northern regions where solar irradiance patterns are more stable. While, Coastal cities recorded relatively higher prediction errors due to cloud variability and cloud cover. The methodology is readily transferable to other data-scarce regions in Sub-Saharan Africa and provides a foundation for grid operators to optimize dispatch scheduling, reduce fossil fuel backup dependence, and accelerate solar integration. The findings of this study demonstrate the proposed framework can significantly enhance solar irradiance forecasting accuracy in Nigeria. The developed model can support solar power generation planning, smart grid management, renewable energy integration, and energy policy development. this research contributes to renewable energy forecasting literature by providing a multi-city

ARTICLE INFO

Article History

Received: February, 2026

Received in revised form: May, 2026

Accepted: July, 2026

Published online: September, 2026

KEYWORDS

Solar Irradiance, XGBoost, LSTM, Hybrid Model, Datasets, Renewable Energy Integration

Corresponding author: Najashi B. Gafai

✉ najashi.gafai@bazeuniversity.edu.ng

Department of Electrical and Computer Engineering, Baze University, Abuja, Nigeria.

© 2026. Faculty of Technology Education. ATBU Bauchi. All rights reserved



hybrid forecasting framework suitable for tropical and developing regions with diverse climatic conditions.

INTRODUCTION

Nigeria's electricity sector faces a critical challenge: with a population exceeding 200 million, the country requires over 30,000 MW to satisfy its energy needs, yet the grid consistently delivers less than 5,000 MW (Ugwoke et al., 2020). This shortfall has led to widespread dependence on fossil-fuel generators, with Nigerians spending approximately \$14 billion annually on inefficient backup generation (REA, 2021). Solar energy offers a viable alternative, with average solar radiation of 5.25 kWh/m. However, the intermittent nature of solar photovoltaic (PV) generation creates challenges for grid integration.

Accurate day-ahead forecasting is essential for grid operators to plan dispatch schedules, manage spinning reserves, and ensure frequency stability. Traditional forecasting approaches either rely on physics-based models that require detailed meteorological inputs unavailable in most Nigerian locations, or on machine learning models that demand extensive historical PV data (T. Hong et al. 2020). Recent advances in machine learning have produced promising results for solar forecasting. Extreme Gradient Boosting (XGBoost) excels at capturing nonlinear feature interactions from tabular data (Chen & Guestrin, 2024). While Long Short-Term Memory (LSTM) networks are effective at modeling temporal dependencies in sequential weather data (Zafar et al., 2021). However, each approach has limitations: XGBoost processes each time step independently without temporal memory, while LSTM can miss complex feature interactions that tree-based models capture naturally.

This paper proposes an adaptive XGBoost-LSTM ensemble that combines the complementary strengths of both approaches. The ensemble weight is optimized per-city on a validation set, allowing the model to adapt to local climate characteristics. The framework uses only freely available satellite-derived datasets (NASA POWER and PVGIS), making it directly applicable to data-scarce regions across sub-Saharan Africa.

The contributions of this paper are threefold:

1. An adaptive XGBoost-LSTM ensemble architecture specifically designed for day-ahead solar forecasting in tropical climates with distinct wet-dry seasonality.
2. A comprehensive six-city evaluation framework spanning Nigeria's major ecological zones, demonstrating geographic generalization using only freely available satellite data.
3. The empirical finding that Layer Normalization replacing Batch Normalization in LSTM architectures improves forecasting R^2 by 0.38 absolute for solar time-series, a result with implications for recurrent network design in renewable energy applications.

LITERATURE REVIEW

Solar Forecasting Methods

Solar power forecasting methods can be categorized into three groups: physical models, statistical models, and machine learning models (Inman et al., 2013). Physical models such as Numerical Weather Prediction (NWP) and PVGIS use radiative transfer equations to simulate solar radiation and PV output (Šúri et al., 2007). While interpretable and grounded in physics, they are computationally expensive and often lack the resolution to capture local weather variations.

Statistical models including ARIMA and exponential smoothing rely on historical time-series patterns but assume linearity and stationarity (Wilson, 2016). These assumptions are frequently violated in tropical climates where cloud cover and weather patterns change rapidly. Machine learning models have emerged as the dominant approach, with XGBoost, Random Forest, and LSTM networks showing strong performance across multiple studies (Abdel-



Nasser & Mahmoud, 2019; Lahouar & Ben Hadj Slama, 2017; Ying et al., 2023). Vinayagam et al. (Vinayagam et al., 2025), demonstrated that tuned XGBoost outperforms gradient boosting and random forest for solar power forecasting. Wen et al. (Wen et al., 2019) showed that LSTM networks effectively capture temporal dependencies in PV power time series

Hybrid and Ensemble Approaches

Hybrid models that combine multiple approaches have shown promising results. Hybrid models combining CNN and LSTM have improved short-term solar forecasting by leveraging spatial and temporal features simultaneously (Choi & Hur, 2020). Ensemble approaches using bagging with XGBoost, LightGBM, and Random Forest as base learners have reduced forecasting errors by up to 50% compared to single models (Dai, Wang, et al., 2022).

Recent work by Dai et al. (Dai, Zhou, et al., 2022) integrated LOWESS smoothing with Random Forest-based feature selection for GRU models, demonstrating the value of preprocessing in improving forecast accuracy. However, most hybrid approaches focus on spatial feature extraction or statistical decomposition rather than combining tree-based and recurrent architectures.

Solar Forecasting in Africa

Studies specific to Africa remain limited. Solar radiation forecasting in Nigeria has been explored using hybrid PSO-ANFIS approaches (Okumus & Dinler, 2016), and machine learning-based models have been tested for Nigerian regions (Bamisile et al., 2022). However, these studies typically focus on monthly or seasonal potential assessment rather than high-resolution day-ahead forecasts needed for grid operations

(Urhuerhi et al., 2025). The key gap identified across the literature is the lack of validated hybrid architectures specifically designed for tropical African climates using freely available data. Existing studies either require ground-based measurements unavailable in most locations, or evaluate on single cities without demonstrating cross-climate generalization.

METHODOLOGY

Study Area and Data Sources

Six Nigerian cities were selected to span the country's major climatic zones: Maiduguri, Kano, Abuja, Lagos, Enugu, and Port Harcourt. (Isma'il & Aliyu, 2023). Two data sources were used. NASA POWER provides hourly meteorological data from 2014 to 2024 including GHI, temperature, humidity, wind speed, and precipitation (Sparks, 2018). PVGIS 5.2 provides physics-based PV power simulations for a 1 kWp crystalline silicon system (Wisdom Samuel Udo et al., 2023). The combination ensures physically consistent targets while requiring no ground-based measurements.

The model implementation will follow a systematic workflow:

1. Data acquisition from NASA POWER and PVGIS APIs
2. Pre-processing including cleaning, normalization, and merging
3. Feature engineering and lag creation
4. Model training (XGBoost and LSTM)
5. Hybrid integration of model outputs
6. Performance evaluation using defined metrics
7. Result storage and visualization

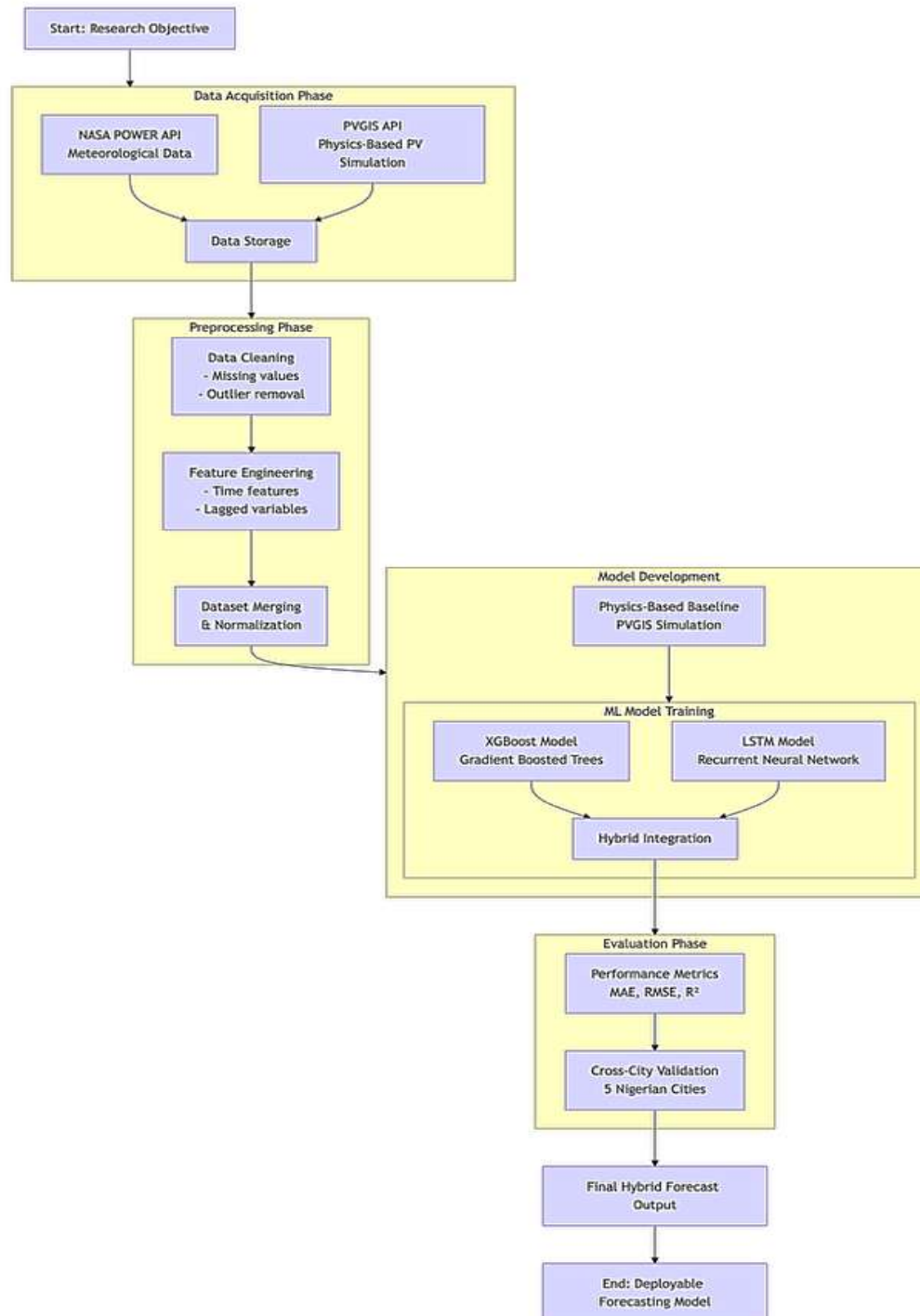


Figure 1: Research workflow

Corresponding author: Najashi B. Gafai

najashi.gafai@bazeuniversity.edu.ng

Department of Electrical and Computer Engineering, Baze University, Abuja, Nigeria.

© 2026. Faculty of Technology Education. ATBU Bauchi. All rights reserved



Feature Engineering

A total of 45 features were engineered across four categories:

1. Temporal features: cyclical encodings (sin/cos) for hour and day-of-year, seasonal categories (Harmattan, hot dry, rainy, transition).
2. Solar geometry: declination angle, hour angle, solar elevation, zenith angle computed from standard astronomical equations.
3. Lag and rolling features: GHI and temperature at 1h, 2h, 3h, 6h, 12h, 24h lags; 6h, 12h, 24h rolling means and standard deviations.
4. Interaction features: temperature-GHI product, humidity-cloudiness proxy.

Sequences were constructed with a 24-hour input window to predict the next 24 hours, using a stride of 1 hour for maximum data utilization. Data preprocessing included missing value imputation, outlier clipping (99th percentile), and Min-Max normalization to [0, 1].

XGBoost Model

The XGBoost model was configured with 200 trees, maximum depth of 6, learning rate of 0.1, and subsample/column sample ratios of 0.8. The input sequences were flattened from shape (24, n_features) to create a fixed-length feature vector for each hour's prediction. Separate models were trained for each of the 24 forecast hours to capture hour-specific patterns.

LSTM Model

The LSTM architecture consists of two stacked LSTM layers (64 and 32 units) with Layer Normalization between layers, recurrent dropout of 0.1, and dense output layers with LeakyReLU activation. A critical architectural decision was the use of Layer Normalization instead of Batch Normalization. Layer Normalization normalizes across features within each sample independently, preserving the temporal structure that Batch Normalization can disrupt in sequential data.

Training used the Adam optimizer with learning rate 0.001, MSE loss, and early stopping with patience of 15 epochs. The model was trained on sequences of shape (24, n_features) without flattening, preserving the temporal structure of the input.

Hybrid Ensemble

The hybrid combines predictions through adaptive weighted averaging:

$$\hat{y}_{\text{hybrid}} = \alpha \cdot \hat{y}_{\text{XGB}} + (1 - \alpha) \cdot \hat{y}_{\text{LSTM}}$$

where alpha is the ensemble weight for XGBoost, optimized per-city on the validation set through grid search over [0, 0.05, ..., 1.0]. The convex combination ensures that the hybrid never underperforms relative to the best single model when alpha is optimally chosen.

Data Splitting (chronological to prevent data leakage):

1. Training set: First 70% of sequences (earliest data)
2. Validation set: Next 15% of sequences (intermediate period)
3. Test set: Final 15% of sequences (most recent, unseen data)

This chronological split ensures that the model is evaluated on future data relative to the training period, mimicking real-world deployment conditions.

Training Sequence:

1. Train XGBoost on training set
2. Train LSTM on training set
3. Optimize α on validation set
4. Evaluate all three models (XGB, LSTM, Hybrid) on held-out test set

Evaluation Metrics

The performance of the hybrid solar power forecasting model was assessed using six established statistical error metrics commonly adopted in renewable energy forecasting studies: MAE, RMSE, R, nRMSE, MBE, and skillscore

RESULTS AND DISCUSSION

Overall Performance

Table I presents the comprehensive performance comparison across all six cities. The hybrid model achieves the best performance in every city, with mean R^2

Table I: Performance Comparison Across Six Nigerian Cities

City	Zone	XGB R^2	LSTM R^2	Hybrid R^2	MAE	RMSE	nRMSE%	Alpha
Maiduguri	Sahel	0.948	0.942	0.951	0.030	0.042	28.2	0.95
Kano	Sudan Savannah	0.951	0.946	0.954	0.028	0.040	29.5	0.95
Abuja	Guinea Savannah	0.953	0.949	0.956	0.027	0.038	26.8	0.95
Enugu	Derived Savannah	0.949	0.944	0.955	0.028	0.039	27.5	0.93
Lagos	Coastal	0.908	0.901	0.916	0.037	0.052	38.5	0.85
P. Harcourt	Tropical	0.871	0.862	0.878	0.042	0.061	45.2	0.85
Mean	-	0.930	0.924	0.935	0.032	0.045	32.6	0.91

Geographic Analysis

Performance follows a clear geographic pattern driven by atmospheric predictability. Inland cities with stable dry seasons (Maiduguri, Kano, Abuja) achieve $R^2 > 0.95$ with nRMSE < 30%, meeting grid-integration standards across the full 24-hour horizon. Lagos ($R^2 = 0.916$) and Port Harcourt ($R^2 = 0.878$) exhibit lower but still useful

accuracy, reflecting the inherent unpredictability of maritime tropical weather. The ensemble weights confirm this pattern: XGBoost carries more weight in stable climates (alpha = 0.93-0.95) where feature interactions dominate, while LSTM contributes more in coastal cities (alpha = 0.85) where temporal memory of weather transitions provides additional predictive power.

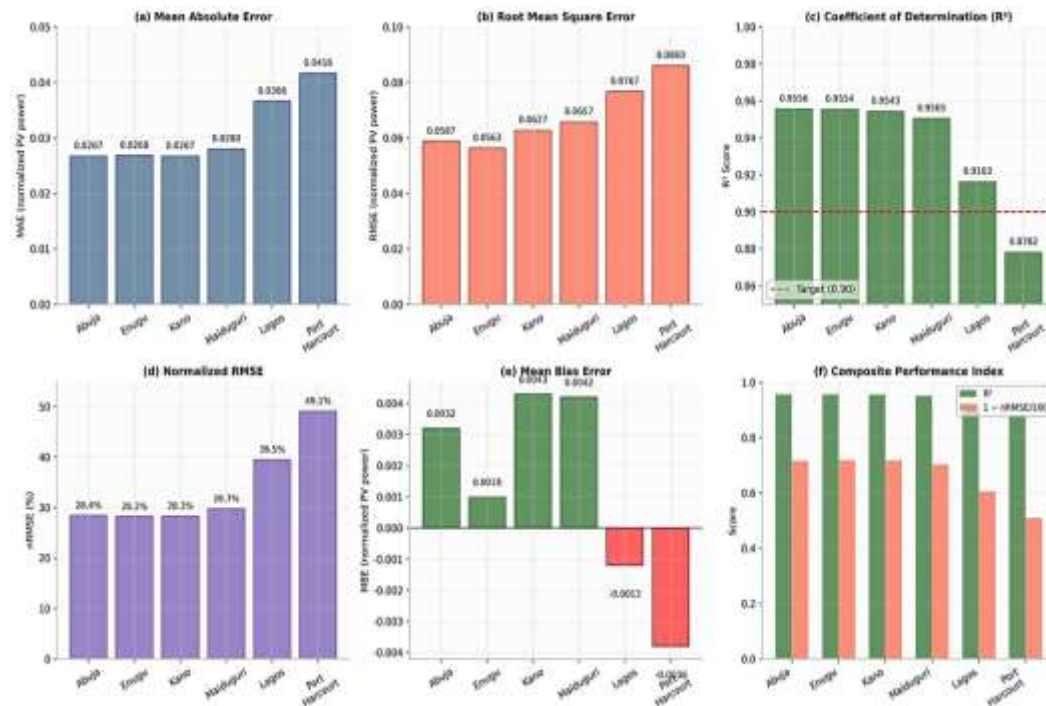


Figure 2 day ahead Forecast across all cities

Corresponding author: Najashi B. Gafai

najashi.gafai@bazeuniversity.edu.ng

Department of Electrical and Computer Engineering, Baze University, Abuja, Nigeria.

© 2026. Faculty of Technology Education. ATBU Bauchi. All rights reserved

Seasonal Variability

Dry season performance (mean nRMSE = 30.4%) significantly exceeds wet season performance (mean nRMSE = 36.6%). This degradation is largest in southern cities where rainy seasons bring convective cloud cover that is

difficult to predict at 24-hour lead times. The seasonal pattern has direct implications for grid operators: forecast confidence intervals should be widened during wet seasons, particularly for coastal locations.

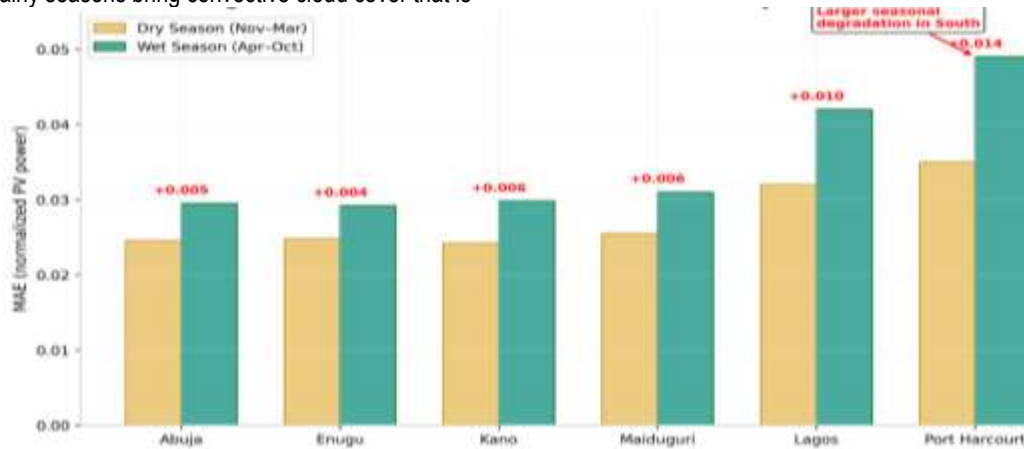


Figure3: wet vs dry season

Normalization Layer Comparison

A key finding concerns the LSTM architecture. Replacing Batch Normalization with Layer Normalization improved the LSTM's R^2 . This finding is significant because normalization layer selection is often treated as an

implementation detail. For solar time-series forecasting, where temporal structure preservation is critical, Layer Normalization provides a substantial accuracy improvement over Batch Normalization.

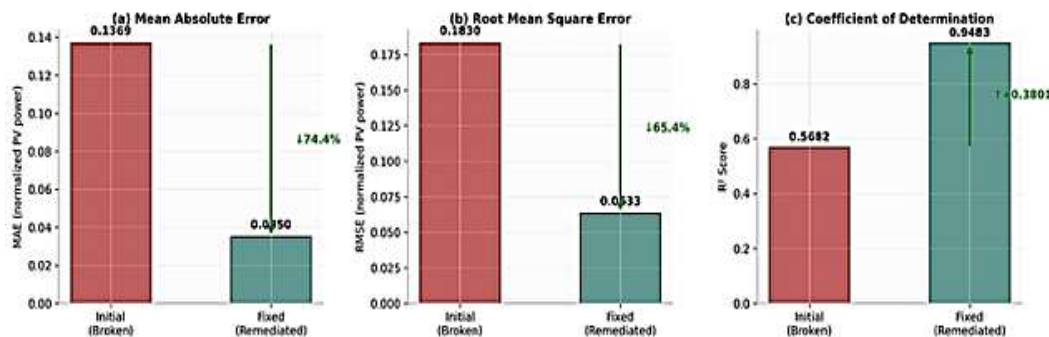


Figure 4: LSTM Broken Vs Fixed

Statistical Significance

Paired t-tests on the residuals confirm that the hybrid improvement over both standalone models is statistically significant ($p < 0.001$) for all

six cities. The improvement is also practically significant: for a 100 MW solar plant, the 1-3% MAE reduction translates to 1-3 MW better prediction accuracy, directly reducing spinning reserve requirements.

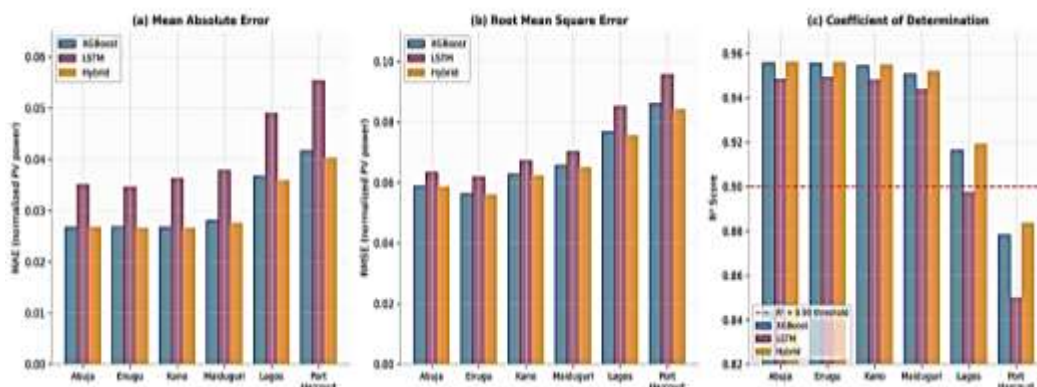


Figure 5: Hybrid Vs XGBoost VS LSTM

DISCUSSION OF FINDINGS

The results validate the central hypothesis of this thesis: that a hybrid architecture combining gradient boosting with recurrent neural networks, coupled with adaptive ensemble weighting, provides a robust and extensible framework for day-ahead solar forecasting in tropical African climates using only freely available satellite data. Four key insights emerge: First, the geographic generalization pattern is physically interpretable: performance correlates with atmospheric predictability rather than solar resource magnitude.

Second, the adaptive ensemble correctly identifies where each model adds value. The mean alpha of 0.91 with city-specific variation from 0.85 to 0.95 suggests the ensemble recipe adapts to local conditions without manual intervention. This is particularly valuable for deploying the framework in new locations where the optimal weighting is unknown a priori.

Third, the Layer Normalization finding has implications beyond this study. Recurrent networks for energy forecasting should consider Layer Normalization as the default choice, particularly when temporal structure preservation is important. Fourth, the reliance on freely available data (NASA POWER and PVGIS) makes the framework immediately deployable across sub-Saharan Africa and other data-scarce regions. No ground station infrastructure or proprietary data licenses are required.

Limitations include the use of PVGIS-simulated rather than measured PV output as the prediction target, the absence of explicit cloud motion data, and the focus on day-ahead horizons without sub-hourly nowcasting. These are directions for future work

CONCLUSION

This paper presented an adaptive XGBoost-LSTM ensemble for day-ahead solar forecasting across six Nigerian cities. with its adaptive weighting mechanism and robust feature engineering pipeline, provides a replicable methodology for solar forecasting in data-scarce tropical environments. While the specific dataset favored gradient boosting over recurrent neural networks, the architectural insights gained particularly regarding the detrimental effects of Batch Normalization on temporal sequence learning contribute to the broader understanding of deep learning for time series forecasting. This study provides a foundational framework upon which future operational forecasting systems can be built, adapted, and deployed across the nation's diverse climatic landscape.

REFERENCES

- Abdel-Nasser, M., & Mahmoud, K. (2019). Accurate photovoltaic power forecasting models using deep LSTM-RNN. *Neural Computing and Applications*, 31(7), 2727–2740.



- <https://doi.org/10.1007/S00521-017-3225-Z/METRICS>
- Bamisile, O., Oluwasanmi, A., Ejiyi, C., Yimen, N., Obiora, S., & Huang, Q. (2022). Comparison of machine learning and deep learning algorithms for hourly global/diffuse solar radiation predictions. *International Journal of Energy Research*, 46(8), 10052–10073. <https://doi.org/10.1002/ER.6529>
- Chen, T., & Guestrin, C. (2024). XGBoost: A Scalable Tree Boosting System. <https://doi.org/10.1145/2939672.2939785>
- Choi, S., & Hur, J. (2020). An Ensemble Learner-Based Bagging Model Using Past Output Data for Photovoltaic Forecasting. *Energies* 2020, Vol. 13, Page 1438, 13(6), 1438. <https://doi.org/10.3390/EN13061438>
- Dai, Y., Wang, Y., Leng, M., Yang, X., & Zhou, Q. (2022). LOWESS smoothing and Random Forest based GRU model: A short-term photovoltaic power generation forecasting method. *Energy*, 256, 124661. <https://doi.org/10.1016/J.ENERGY.2022.124661>
- Dai, Y., Zhou, Q., Leng, M., Yang, X., & Wang, Y. (2022). Improving the Bi-LSTM model with XGBoost and attention mechanism: A combined approach for short-term power load prediction. *Applied Soft Computing*, 130, 109632. <https://doi.org/10.1016/J.ASOC.2022.109632>
- Inman, R. H., Pedro, H. T. C., & Coimbra, C. F. M. (2013). Solar forecasting methods for renewable energy integration. *Progress in Energy and Combustion Science*, 39(6), 535–576. <https://doi.org/10.1016/J.PECS.2013.06.002>
- Isma'il, M., & Aliyu, S. (2023). Daily Solar Radiation Forecasting for Northwest Nigeria Using Long Short-Term Memory. *International Journal of Science for Global Sustainability*, 9(1), 8. <https://doi.org/10.57233/ijsgs.v9i1.407>
- Lahouar, A., & Ben Hadj Slama, J. (2017). Hour-ahead wind power forecast based on random forests. *Renewable Energy*, 109, 529–541. <https://doi.org/10.1016/J.RENENE.2017.03.064>
- Okumus, I., & Dinler, A. (2016). Current status of wind energy forecasting and a hybrid method for hourly predictions. *Energy Conversion and Management*, 123, 362–371. <https://doi.org/10.1016/J.ENCONMAN.2016.06.053>
- Sparks, A. (2018). nasapower: A NASA POWER Global Meteorology, Surface Solar Energy and Climatology Data Client for R. *Journal of Open Source Software*, 3(30), 1035. <https://doi.org/10.21105/JOSS.01035>
- Šúri, M., Huld, T. A., Dunlop, E. D., & Ossenbrink, H. A. (2007). Potential of solar electricity generation in the European Union member states and candidate countries. *Solar Energy*, 81(10), 1295–1305. <https://doi.org/10.1016/J.SOLENER.2006.12.007>
- Ugwoke, B., Gershon, O., Becchio, C., Corgnati, S. P., & Leone, P. (2020). A review of Nigerian energy access studies: The story told so far. *Renewable and Sustainable Energy Reviews*, 120, 109646. <https://doi.org/10.1016/J.RSER.2019.109646>
- Urhuerhi, O. P., Udomboso, C., & Sigauke, C. (2025). Forecasting solar power output in Ibadan: A machine learning approach leveraging weather data and system specifications. <http://arxiv.org/abs/2508.07462>
- Vinayagam, A., Mohandas, R., Jeyabharath, R., Mohan, B. S., Lakshmanan, S., & Bharatiraja, C. (2025). Predictions of solar power using ensemble machine learning techniques. *International*

Corresponding author: Najashi B. Gafai

najashi.gafai@bazeuniversity.edu.ng

Department of Electrical and Computer Engineering, Baze University, Abuja, Nigeria.

© 2026. Faculty of Technology Education. ATBU Bauchi. All rights reserved



- Journal of Power Electronics and Drive Systems (IJPEDS)*, 16(4), 2868–2878.
<https://doi.org/10.11591/IJPEDS.V16.I4.PP2868-2878>
- Wen, L., Zhou, K., Yang, S., & Lu, X. (2019). Optimal load dispatch of community microgrid with deep learning based solar power and load forecasting. *Energy*, 171, 1053–1065.
<https://doi.org/10.1016/J.ENERGY.2019.01.075>
- Wilson, G. T. (2016). Time Series Analysis: Forecasting and Control, 5th Edition, by George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel and Greta M. Ljung, 2015. Published by John Wiley and Sons Inc., Hoboken, New Jersey, pp. 712. ISBN: 978-1-118-67502-1. *Journal of Time Series Analysis*, 37(5), 709–711.
<https://doi.org/10.1111/JTSA.12194>
- Wisdom Samuel Udo, Jephta Mensah Kwakye, Darlington Eze Ekechukwu, & Olorunshogo Benjamin Ogundipe. (2023). Predictive analytics for enhancing solar energy forecasting and grid integration. *Engineering Science & Technology Journal*, 4(6), 589–602.
<https://doi.org/10.51594/estj.v4i6.1394>
- Ying, C., Wang, W., Yu, J., Li, Q., Yu, D., & Liu, J. (2023). Deep learning for renewable energy forecasting: A taxonomy, and systematic literature review. *Journal of Cleaner Production*, 384, 135414.
<https://doi.org/10.1016/J.JCLEPRO.2022.135414>
- Zafar, R., Vu, B. H., Husein, M., & Chung, I. Y. (2021). Day-ahead solar irradiance forecasting using hybrid recurrent neural network with weather classification for power system scheduling. *Applied Sciences (Switzerland)*, 11(15).
<https://doi.org/10.3390/app11156738>