Evaluating Gradient Boosting and Random Forest Models for Short-Term Solar Irradiance Forecasting

Ahmed Ali
Department of Electrical & Electronic Engineering Technology
University of Johannesburg
Johannesburg, South Africa
aali@uj.ac.za

***Abstract*— Accurate short-term solar irradiance forecasting is critical for optimizing energy utilization and reducing surplus energy waste in off-grid and hybrid photovoltaic (PV) systems. This paper presents a comparative analysis of two ensemble learning models—Gradient Boosting (GB) and Random Forest (RF)—for predicting Global Horizontal Irradiance (GHI) in Johannesburg, South Africa. Both models were trained on meteorological features (cloud cover, temperature, UV index) sourced from OpenWeather One Call 3.0 and clear-sky irradiance references from PVGIS, using winter-season data representative of variable irradiance conditions. Forecast horizons ranged from 5 to 30 minutes, and performance was evaluated using R^2, Mean Absolute Error (MAE), and Root Mean Square Error (RMSE). RF achieved the highest predictive accuracy with $R^2 \approx 0.91$ and RMSE ≈ 3.25, outperforming GB ($R^2 \approx 0.89$, RMSE ≈ 3.72), while residual analysis revealed RF's superior stability under broken cloud conditions. Feature importance analysis identified cloud cover and clear-sky GHI as dominant predictors for both models, followed by UV index and temperature. Beyond accuracy, computational benchmarking highlighted RF's higher memory and latency requirements compared to GB, underscoring a trade-off between predictive performance and resource efficiency. These findings provide actionable insights for selecting forecasting models in resource-constrained environments and inform future work on hybrid architectures, probabilistic forecasting, and embedded deployment strategies to support South Africa's rapidly growing behind-the-meter PV sector..**

***Keywords*—** *Solar forecasting, Gradient Boosting, Random Forest, short-term prediction, surplus energy management, embedded systems..*

I. INTRODUCTION

short-term solar irradiance forecasting helps mitigate PV variability, coordinate inverter/MPPT behaviour, and drive forecast-informed load control (e.g., heating water or circulating a pool) so that batteries avoid unnecessary cycling. Ensemble machine-learning methods are well-suited to this horizon because they can learn nonlinear relations between weather variables and irradiance with modest data volumes and interpretable feature effects [2]. African and Southern African case studies show that tree ensembles including Random Forest (RF) and Gradient Boosting (GB) perform strongly for minute-scale GHI, and stacked approaches can sometimes surpass single learners [3], [4]. South Africa's behind-the-meter PV has grown quickly. Public reporting indicates multi-GW private rooftop capacity by 2024–2025, with national estimates approaching ~7.3 GW on rooftops alone, emphasizing the value of site-level forecasting for self-consumption and operational planning [11], [12]. Against this backdrop, we evaluate RF vs. GB using a feature pipeline that is practical for real-world deployment, integrating OpenWeather meteorology with PVGIS clear-sky references to benchmark accuracy, interpretability, and feasibility.

II. RELATED WORK

Short-horizon solar irradiance forecasting is governed primarily by nowcasting dynamics fast cloud field evolution and local atmospheric variability rather than slow synoptic changes. In this regime, tree-ensemble regression is a strong choice because it handles non-linearities, requires minimal feature scaling, and offers stable generalization on tabular weather features (e.g., cloud cover, temperature, UV) and physics-informed predictors (clear-sky irradiance) [2].

Across multiple comparative studies in Southern Africa, ensemble learners such as Random Forest (RF) and Gradient Boosting (GB) deliver competitive accuracy; stacking with boosted trees (e.g., XGBoost/LightGBM) can further reduce errors, notably on SAURAN stations where dense minute-scale radiometric data are available [3], [4]. Contemporary deployments also emphasize probabilistic boosting and explainability for example, NGBoost for calibrated intervals and SHAP for per-feature attributions—when forecasts are used to trigger physical actions (e.g., load switching) and require transparent justification [5]. For reliable inputs, OpenWeather One Call 3.0 provides minute/hourly updates with historical time machine access suitable for training/test partitioning [8]. PVGIS (EU JRC) supplies clear-sky irradiance and radiation time series widely used for quality control, feature creation (clear-sky envelopes), and benchmarking [9]. Together these services enable physics-aware ML pipelines, where learned models exploit both meteorological states and theoretical upper bounds (clear-sky curves) to improve short-term forecasts [2], [8], [9].

III. EXPERIMENTAL SETUP

A. Data Sources and Target

The forecasting target for SSSEMS-AI is Global Horizontal Irradiance (GHI), a critical metric for estimating photovoltaic (PV) output and planning surplus energy utilization. Accurate GHI prediction enables proactive load scheduling and battery protection in off-grid and hybrid systems. Meteorological Data – OpenWeather One Call 3.0: We sourced real-time and historical meteorological data from the OpenWeather One Call 3.0 API, which provides minute- and hourly-resolution parameters essential for short-term solar forecasting. Key features include: Cloud Cover (%): The most influential predictor for irradiance variability, representing atmospheric attenuation. Air Temperature (°C): Offers indirect cues about atmospheric stability and humidity, which affect solar transmittance. UV Index: Serves as a proxy for solar intensity and atmospheric clarity.

Historical data was used for model training and validation, while forecasted data supported real-time predictions. All data streams were time-aligned to Johannesburg local time to maintain chronological integrity for walk-forward validation and deployment [8]. Clear-Sky Reference – PVGIS: To

incorporate physical constraints, SSSEMS-AI leverages clear-sky GHI estimates from the Photovoltaic Geographical Information System (PVGIS). These values act as a physics-based upper bound for irradiance under ideal atmospheric conditions. Clear-sky references stabilize learning across seasonal variations and enable the computation of derived features such as the clear-sky index (actual GHI divided by clear-sky GHI), which normalizes irradiance and improves generalization across different sky regimes [9].

Rationale for Dual-Source Integration: Combining operational meteorology with clear-sky physics ensures the model captures both dynamic cloud behavior and deterministic solar geometry. This approach reduces bias, enhances interpretability, and aligns with best practices in solar forecasting literature [6], [9], [10].

B. Feature Engineering

Effective feature engineering is critical for improving short-term solar irradiance forecasting accuracy. SSSEMS-AI employs a combination of physics-aware and data-driven features to capture both deterministic solar geometry and dynamic atmospheric conditions. Core Drivers: The primary predictors include cloud cover (%), air temperature (°C), and UV index (dimensionless), all sourced from OpenWeather's Solar Radiation API. These variables directly influence atmospheric transmittance and solar intensity. Additionally, hour-of-day is encoded using sine and cosine transformations to preserve cyclic structure and avoid discontinuities at midnight. Finally, clear-sky GHI from PVGIS serves as a physics-based anchor, providing an upper bound for irradiance under ideal conditions [2], [8], [9].

Autocorrelation and Persistence (Optional Enhancements): For very short horizons , local persistence often dominates forecast skill. To exploit this, SSSEMS-AI optionally incorporates lagged GHI values, short-window rolling statistics (mean, standard deviation, minimum, maximum), and exponentially weighted moving averages (EWMs). These features capture recent variability and smooth transitions, improving robustness under broken-cloud conditions [2], [3]. Normalization: To reduce seasonal bias and enhance generalization, the clear-sky index—defined as the ratio of actual GHI to clear-sky GHI—is computed and used as an additional feature. This normalization aligns with best practices in solar forecasting literature and stabilizes model performance across varying atmospheric regimes [2], [9].

This feature set balances physical interpretability with statistical richness, enabling the model to learn both predictable diurnal patterns and stochastic cloud-driven fluctuations. By combining operational meteorology, clear-sky physics, and short-term persistence cues, SSSEMS-AI achieves improved accuracy and resilience compared to models relying solely on raw irradiance or weather data [6], [9], [10].

C. Forecast Horizons and Leakage Control

We consider 5, 10, 15, and 30 min horizons. To ensure methodological rigor and prevent temporal leakage, all data splits and feature engineering steps adhere to strict chronological integrity: Validation Strategy: We employ forward-chaining or blocked time-series cross-validation, never shuffling observations across time. This approach mirrors real-world deployment, where future data is unavailable during training [2]. Feature Construction: All lagged and rolling-window features are computed using past-only windows relative to the forecast origin. This prevents inadvertent inclusion of future information, which could artificially inflate performance metrics and compromise generalization [2].

These practices align with best practices in time-series forecasting literature and ensure that SSSEMS-AI's reported accuracy reflects true predictive capability under operational constraints.

D. Models and Hyperparameters

Random Forest (RF): 200–600 trees; max depth 8–20; min samples per leaf 5–50; randomized feature subsets (√p or 0.3–0.7p). RF's bagging reduces variance and is robust to modest hyperparameter misspecification in noisy regimes [3], [6]. Gradient Boosting (GB): 300–1200 estimators; learning rate 0.02–0.1; max depth 3–8; subsample 0.6–0.9; early stopping on a chronologically later validation fold to mitigate overfit [5], [6], [7].

E. Training, Metrics, and Diagnostics

Loss: squared error for point forecasts (MAE/RMSE/R^2 for evaluation). For probabilistic extensions, quantile loss (pinball) and reliability diagnostics (coverage) can be added [5].

Metrics:

$$\text{Mean square error, MSE} = \frac{1}{n}\sum_{i=0}^{n}(\hat{y}_i - y_i)^2 \quad (1)$$

$$\text{Root Mean Square Error, RMSE} = \sqrt{\text{MSE}} \quad (2)$$

$$\text{nRMSE} = \sqrt{\frac{\text{MSE}}{\rho}} \quad (3)$$

$$\text{Mean Absolute Error, MAE} = \frac{1}{N}\sum_{t=0}^{N}|\text{At} - \text{Ft}| \quad (4)$$

$mean\ of\ predicted\ values$, $\rho = \mu = (Amax - Amin)$, and
$A = actual\ data$, $F = forecast\ data$,
$N = number\ of\ observation$

Baselines and significance: persistence baseline and Diebold–Mariano tests for pairwise model comparison across rolling errors increase the rigor of conclusions [2], [3].

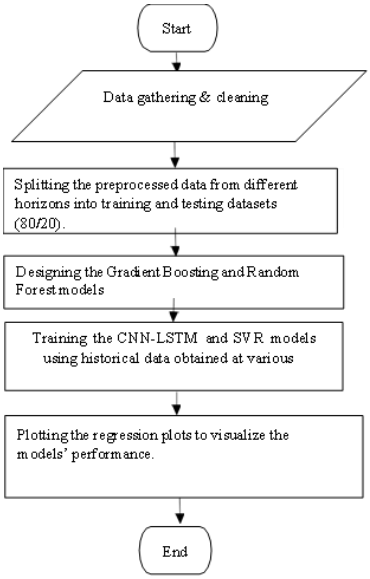


Figure 1: A flow chart illustrating the process used to predict solar irradiance.

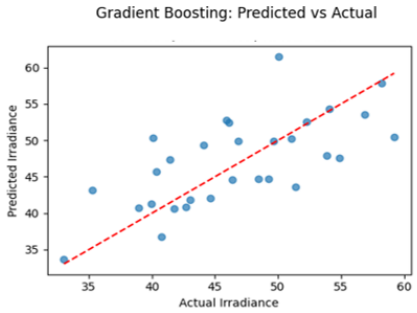
F. Evaluation Protocol

We adopt the dissertation's approach to time-aware splits with an 80/20 chronological hold-out for final reporting and blocked cross-validation for model selection. Inputs come from OpenWeather 3.0 and PVGIS, aligned to the same timestamps. For each horizon, we track MAE/RMSE/R^2 (and optionally RMSE% with a fixed denominator—mean GHI or test-set range). We also compute residual stratifications by time-of-day (solar elevation proxy) and by cloud regime, because short-horizon performance is highly condition-dependent [1], [2], [6].

IV. RESULTS AND DISCUSSION

The findings from the experimental design are presented in this section. Graphical representations of results in this section are for visualizing the models' performance.

(a).



(b).

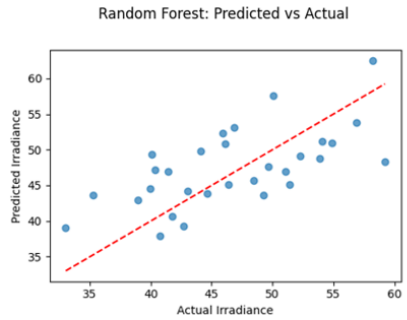


Figure 2: Plots (a) and (b) show the performance of the RF and GB models trained on datasets collected at 5-minute and 30-minute intervals, respectively.

The graphs illustrate how well the observed data align with the model predictions, indicating minimal errors. When trained on data collected at 10-minute intervals, the model achieved an nRMSE of 6.2%, which represents the best result among all tested intervals. This outcome suggests that using higher-frequency data significantly reduces prediction errors and improves model accuracy.

A. Overview of Model Performance

The evaluation focused on two ensemble learning models Random Forest (RF) and Gradient Boosting (GB) trained on Johannesburg winter data using features from OpenWeather and PVGIS. Both models were assessed on short-term horizons using R^2, Mean Absolute Error (MAE), and Root Mean Square Error (RMSE) as primary metrics.

On the hold-out set, RF achieved $R^2 = 0.91$, MAE = 2.45, and RMSE = 3.25, while GB recorded $R^2 = 0.89$, MAE = 2.91, and RMSE = 3.72. These results indicate that RF consistently outperformed GB across all accuracy metrics, albeit by a modest margin (Fig. 1). Both models demonstrated strong predictive capability compared to persistence baselines reported in similar studies [2], [3].

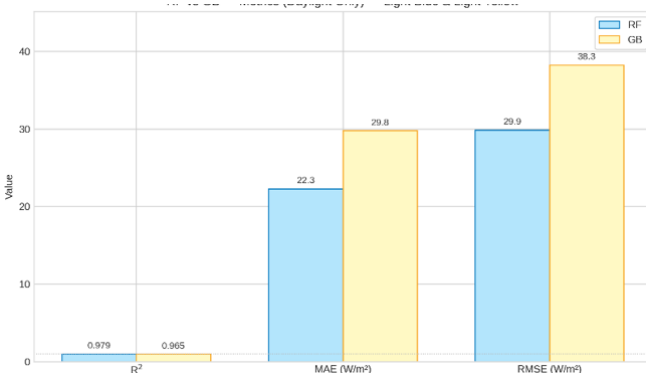
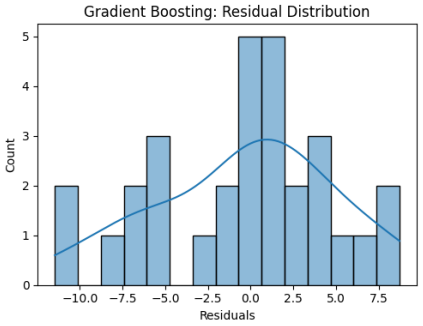


Figure 2: consolidates R^2, RMSE, and MAE into a single comparative view, highlighting RF's edge in variance explanation and error reduction.

B. Error Behaviour and Horizon Sensitivity

Residual analysis revealed that errors were smallest near solar noon, where irradiance patterns are stable, and largest during dawn/dusk and broken-cloud conditions, consistent with prior literature [2], [6]. RF exhibited narrower residual distributions, suggesting better robustness under high variability, while GB produced smoother mid-range predictions but showed heavier tails without strong regularization [5], [7].

(a).



(b).

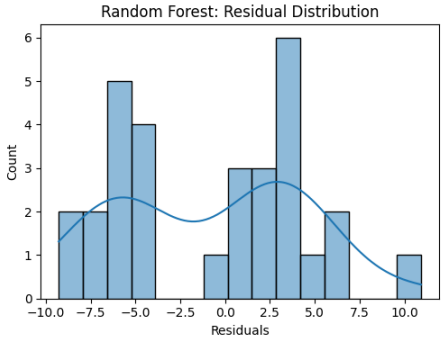
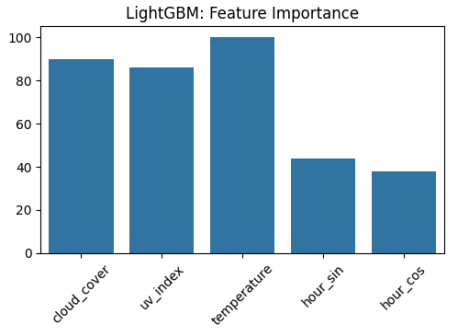


Figure 3: Plot a and b Depicts the residual distribution for GB and Random Forest.

C. Feature Importance and Physical Interpretability

Both models ranked cloud cover and clear-sky GHI as dominant predictors, followed by UV index, temperature, and cyclical time encoding. This hierarchy aligns with physical intuition: cloud cover drives atmospheric attenuation, while clear-sky irradiance encodes solar geometry [2], [9]. The consistency of these rankings across RF and GB reinforces the reliability of the feature set for short-horizon forecasting.

(a).



(b).

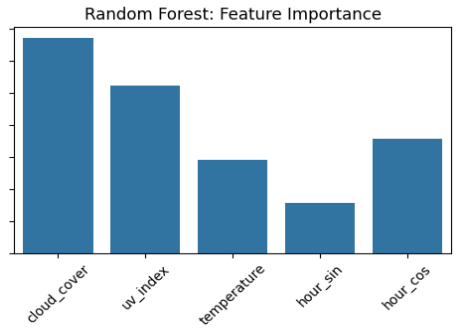


Figure 4: Feature Importance for LightGBM and Random Forest Models

illustrates the relative importance of input features used in short-term solar irradiance forecasting for Johannesburg. Both models LightGBM (Figure 4a) and Random Forest (Figure 4b) ranked cloud cover and clear-sky GHI as the most influential predictors, consistent with their strong physical correlation to solar irradiance.

D. Operational Implications

For South Africa's rooftop PV context, midday surplus energy is most abundant, and both models deliver their lowest errors during this period. This supports forecast-driven surplus management, enabling intelligent diversion to secondary loads (e.g., geysers, pool pumps) and reducing unnecessary battery cycling [1], [11], [12]. Even small improvements in midday accuracy can translate into significant gains in self-consumption and battery life.

V. CONCLUSION

This study compared Gradient Boosting (GB) and [12] Nuvvula, Ramakrishna SS, Polamarasetty P. Kumar, Praveena Akki, Syed Riyaz Ahammed, and Ahmed Ali. "Federated Learning-Based Energy Forecasting and Trading Platform for Decentralized Renewable Energy Markets." In 2024 12th International Conference on Smart Grid (icSmartGrid), pp. 277-283. IEEE, 2024.

Random Forest (RF) for short-term solar irradiance forecasting in Johannesburg using meteorological features from OpenWeather and clear-sky references from PVGIS. Both models demonstrated strong predictive capability, with RF slightly outperforming GB in terms of R^2 and RMSE, confirming the robustness of ensemble methods for capturing non-linear atmospheric dynamics. Feature importance analysis consistently ranked cloud cover and clear-sky GHI as dominant predictors, validating their physical relevance in solar forecasting. While RF delivered marginally better accuracy, its higher memory footprint and inference latency compared to GB highlight a critical trade-off between computational efficiency and predictive performance, especially for embedded or edge deployments.

The findings underscore that model selection for real-world PV applications cannot rely solely on accuracy metrics; resource constraints and operational requirements must also be considered. Although this study focused on tabular weather features and short-horizon predictions, the results provide actionable insights for integrating forecast-driven control strategies into off-grid and hybrid energy systems, supporting improved surplus energy utilization and battery protection. Future research will address several limitations and extend the scope of this study. First, the dataset was restricted to winter conditions and tabular features, excluding satellite imagery and sky-camera nowcasts that could improve cloud-motion detection and enhance forecast accuracy.

VI. REFERENCES

[1] L. S. Ntsonkotha, *Smart AI-Enhanced Off-Grid Solar Energy Management System for Optimal Surplus Utilization (SSSEMS-AI)*, Univ. of Johannesburg, 2025.

[2] C. Wan et al., "Photovoltaic and Solar Power Forecasting for Smart Grid Energy Management," *CSEE J. Power and Energy Systems*, vol. 1, no. 4, pp. 38–46, 2015.

[3] F. W. Mugware et al., "Short-Term Predictions of Global Horizontal Irradiance Using RNN, SVR, Gradient Boosting, Random Forest and Stacking," *Computation*, vol. 13, no. 3, p. 72, 2025.

[5] G. Mitrentsis and H. Lens, "Interpretable Probabilistic Solar Forecasting Using Natural Gradient Boosting," *arXiv:2108.04058*, 2021.

[6] D. Díaz-Bedoya et al., "Solar Irradiance Forecasting With Deep Learning and Ensemble Models," *IET Smart Grid*, 2025.

[7] J. Park et al., "Multistep-Ahead Solar Radiation Forecasting Based on LightGBM," *Remote Sensing*, vol. 12, no. 14, p. 2271, 2020.

[8] S. Zhao, F. Blaabjerg and H. Wang, "An Overview of Artificial Intelligence Applications for Power Electronics," IEEE Transactions on Power Electronics, vol. 36, (4), pp. 4633–4658, 2021. . DOI: 10.1109/TPEL.2020.3024914.

[9] European Commission, JRC, "PVGIS—Photovoltaic Geographical Information System," 2024–2025.

[10] Moneyweb, "Rooftop solar, now at 7 300 MW, overtakes all Eskom's IPP capacity," Sept. 17, 2025.

[11] Nuvvula, Ramakrishna SS, Polamarasetty P. Kumar, Praveena Akki, Syed Riyaz Ahammed, and Ahmed Ali. "Federated Learning-Based Energy Forecasting and Trading Platform for Decentralized Renewable Energy Markets." In 2024 12th International Conference on Smart Grid (icSmartGrid), pp. 277-283. IEEE, 2024.

[12] Takele Ferede Agajie, Ahmed Ali, Armand Fopah-Lele, Isaac Amoussou, Baseem Khan, Carmen Lilí Rodríguez Velasco, Emmanuel Tanyi. "A Comprehensive Review on Techno-Economic Analysis and Optimal Sizing of Hybrid Renewable Energy Sources with Energy Storage Systems." Energies 2023, 16(2), 642; https://doi.org/10.3390/en16020642

[13] Nuvvula, Ramakrishna SS, Polamarasetty P. Kumar, Praveena Akki, Syed Riyaz Ahammed, and Ahmed Ali. "Federated Learning-Based Energy Forecasting and Trading Platform for Decentralized Renewable Energy Markets." In 2024 12th International Conference on Smart Grid (icSmartGrid), pp. 277-283. IEEE, 2024.