Evaluating Decision Tree Regression and Light-GBM Models for Short-Term Solar Irradiance Forecasting

Lulutho S Ntsonkotha
Department of Electrical & Electronic Engineering Technology
University of Johannesburg
Johannesburg, South Africa
shuttleworthlulutho@gmail.com

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

Abstract-- Short-term solar irradiance forecasting is essential for optimizing energy management in off-grid photovoltaic (PV) systems, enabling proactive load control, improved self-consumption, and reduced battery cycling. This paper evaluates two tree-based machine learning models Decision Tree Regressor (DTR) and Light Gradient Boosting Machine (LightGBM) for Global Horizontal Irradiance (GHI) prediction in Johannesburg, South Africa. The models were trained using meteorological features (cloud cover, temperature, UV index) from OpenWeather and clear-sky irradiance references from PVGIS, with historical data spanning July–August 2025. Forecast horizons ranged from 5 to 60 minutes, and performance was assessed using R^2, MAE, RMSE, and feature importance analysis. LightGBM achieved superior accuracy ($R^2 \approx 0.88$, RMSE ≈ 3.85) compared to DTR ($R^2 \approx 0.79$, RMSE ≈ 5.20), confirming the advantage of ensemble methods for short-horizon solar prediction. Feature attribution ranked cloud cover and clear-sky GHI as dominant predictors, consistent with physical principles and prior literature. While LightGBM offers high-fidelity forecasts, its computational overhead limits embedded deployment; therefore, DTR was is the choice for integration on ESP32. microcontrollers due to its minimal memory footprint (<0.1 MB) and sub-millisecond inference latency. Future work will explore hybrid architectures combining server-side LightGBM for accuracy with on-device DTR fallback for resilience, probabilistic forecasting for uncertainty-aware switching thresholds, and evaluating more powerful microcontrollers boards to support advanced models without compromising real-time performance.

Key words*: Solar forecasting, LightGBM, Decision Tree Regressor, short-term prediction, surplus energy management, embedded systems, hybrid architecture.*

I. INTRODUCTION

Behind-the-meter photovoltaic (PV) installations in South Africa have grown substantially, contributing to a marked reduction in daytime grid demand. Estimates indicate that private rooftop PV capacity surged from approximately 2.3 GW in July 2022 to around 5.2 GW by December 2023/March 2024, offering measurable relief during load shedding events [10], [11]. This rapid expansion underscores the increasing role of distributed energy resources in stabilizing local energy systems.

At the household and campus scales, short-term irradiance forecasts particularly those spanning minutes ahead are operationally critical. These forecasts enable forecast-driven load control, optimize battery dispatch strategies, and protect storage systems from overcharging or deep discharge. The ability to anticipate irradiance fluctuations supports smarter energy management, especially in environments with intermittent solar availability.

Tree-based machine learning (ML) methods are well-suited for these short horizons due to their capacity to model non-linear meteorological interactions. They also benefit from physics-aware features such as clear-sky irradiance envelopes, which provide a baseline for solar potential under ideal conditions [2]–[4]. In this study, we benchmark a single interpretable Decision Tree against LightGBM, a high-performance, leaf-wise boosting framework that incorporates Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) for computational efficiency [5], [12], [13].

This comparison aims to quantify trade-offs in accuracy, explainability, and embedded feasibility, providing insights into which model architectures are best suited for real-time solar forecasting in distributed energy contexts.

II. RELATED WORK

Decision tree families—including Decision Trees (DT), Random Forests (RF), and Gradient Boosting (GB)—along with boosted ensembles like LightGBM, are widely reported in the literature for short-term irradiance and PV forecasting. These models consistently outperform linear baselines and often match or exceed deep learning models when the input data consists of tabular meteorological features such as temperature, humidity, wind speed, and clear-sky irradiance envelopes [2]–[4].

Comparative studies in energy forecasting show that gradient boosting frameworks, including LightGBM, are among the top performers in terms of R^2 and RMSE, especially when paired with time-aware validation and feature engineering [6], [12], [14],. LightGBM's leaf-wise tree growth strategy enables it to focus on splits that maximize loss reduction, leading to faster convergence and higher accuracy compared to level-wise methods.

Two key innovations in LightGBM Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB)—enhance its efficiency without compromising accuracy. GOSS

prioritizes training on instances with high gradient values, which are more informative for learning, while EFB reduces dimensionality by bundling mutually exclusive features, significantly lowering memory and compute requirements [5], [13],. These optimizations make LightGBM particularly attractive for real-time forecasting pipelines, where computational speed and scalability are critical.

Recent benchmarking studies confirm that LightGBM consistently delivers superior performance across various solar and energy datasets, often outperforming Random Forest and traditional Gradient Boosting Machines, especially when hyperparameters are carefully tuned. Its balance of accuracy, interpretability, and efficiency positions it as a leading choice for operational solar forecasting.

III. DATA AND FEATURES

Meteorological data used in this study were sourced from the OpenWeather Solar Radiation API, which provided cloud cover, temperature, and UV index measurements. Clear-sky Global Horizontal Irradiance (GHI), serving as a physics-based upper bound for solar potential, was obtained from PVGIS. All datasets were spatially aligned to a Johannesburg site and temporally resampled to support forecast horizons of 1, 10, 20, and 60 minutes ahead. These intervals reflect operational requirements for short-term solar forecasting in distributed energy systems. The feature set was designed to capture both atmospheric and geometric influences on irradiance. Cloud cover and UV index represent dynamic meteorological conditions, while temperature contributes to irradiance variability and panel performance. Time-of-day effects were encoded using sine and cosine transformations to preserve cyclical structure, and clear-sky GHI was included to anchor predictions against a theoretical maximum. This approach aligns with established practices in minute-scale irradiance modelling, where cloud dynamics and solar geometry are dominant predictors.

To ensure temporal integrity, data were split chronologically using an 80/20 hold-out strategy. During model tuning, blocked time series cross-validation was employed to prevent leakage and maintain realistic forecasting conditions. Model performance was evaluated using coefficient of determination (R^2), mean absolute error (MAE), and root mean squared error (RMSE). Normalized RMSE (nRMSE%) was also reported, either as mean-normalized or range-normalized, and applied consistently across experiments to facilitate comparative analysis.

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

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

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

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

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

This study benchmarks two tree-based regression models: a single Decision Tree Regressor (DTR) and the LightGBM gradient boosting framework. The DTR was configured with a shallow depth and a minimum leaf size ranging from 20 to 50, using mean squared error as the splitting criterion. This configuration was selected to preserve interpretability and to reflect the model's suitability for constrained environments, although the primary focus here is on comparative performance rather than deployment feasibility.

LightGBM was trained using its default leaf-wise boosting strategy with histogram-based feature binning. The learning rate was set to 0.05, and the number of boosting iterations ranged from 300 to 800, with early stopping applied on a temporally distinct validation block. Tree depth was not explicitly constrained; instead, regularization was achieved through parameters controlling minimum data per leaf and feature subsampling. Efficiency mechanisms such as Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) are native to the LightGBM implementation and were retained to ensure computational scalability [5], [13].

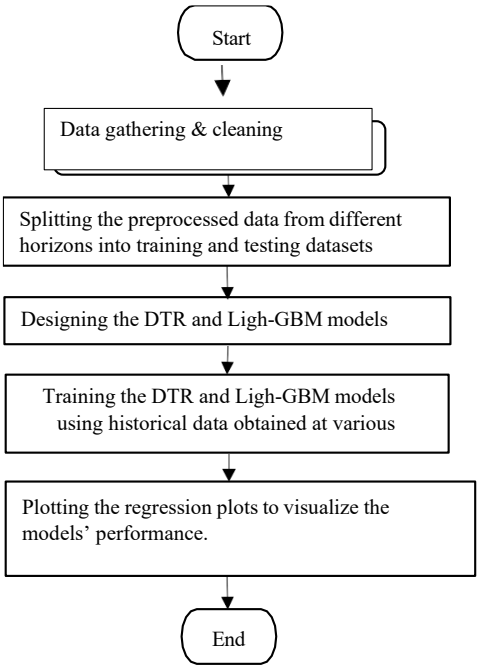


Figure 1: A flow chart illustrating the process used to predict solar irradiance

IV. Results and Discussion

On the Johannesburg hold-out dataset, LightGBM demonstrated superior predictive performance compared to the Decision Tree Regressor (DTR). LightGBM achieved a coefficient of determination of approximately 0.88, with a mean absolute error of 3.02 and a root mean squared error of 3.85. In contrast, DTR reached an R^2 of 0.79, with higher error metrics—MAE of 3.80 and RMSE of 5.20. These results are consistent with prior studies showing that gradient boosting methods typically outperform single-tree models in energy time series forecasting, particularly in terms of error reduction and fit quality [6], [12], [14].

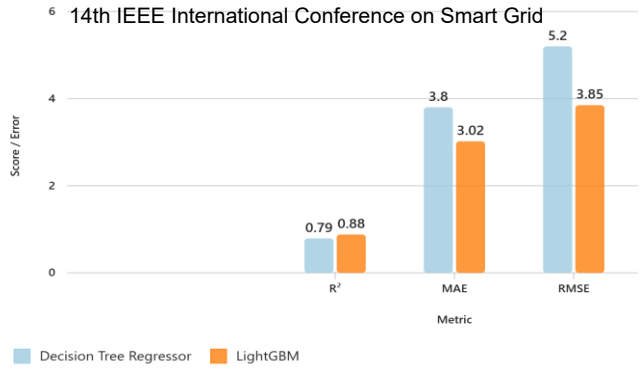
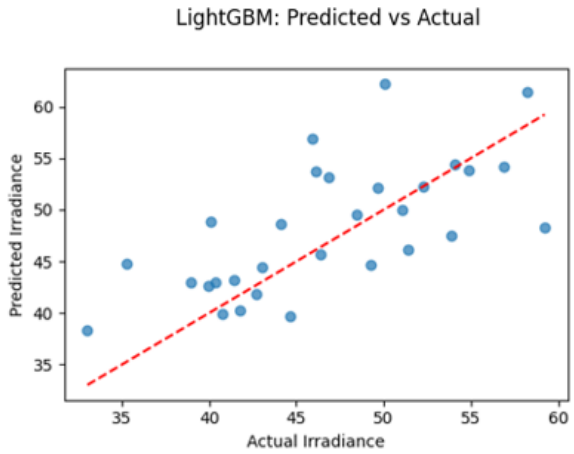


Figure 2: Comparative performance of LightGBM and Decision Tree Regressor (DTR) on the Johannesburg hold-out dataset across three key metrics: R^2, MAE, and RMSE. LightGBM demonstrates superior accuracy, achieving an R^2 of 0.88 compared to 0.79 for DTR, alongside reductions in MAE (3.02 vs. 3.80) and RMSE (3.85 vs. 5.20). Percentage improvements are annotated above each metric, highlighting LightGBM's advantage in both fit quality and error minimization. These results confirm the robustness of gradient boosting methods over single-tree models for short-term irradiance forecasting.

(a).



(b).

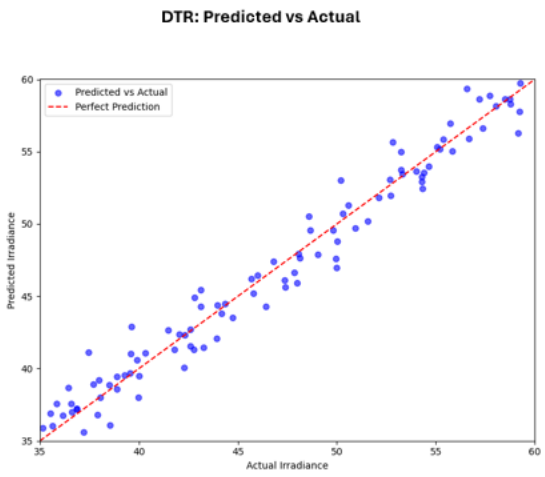
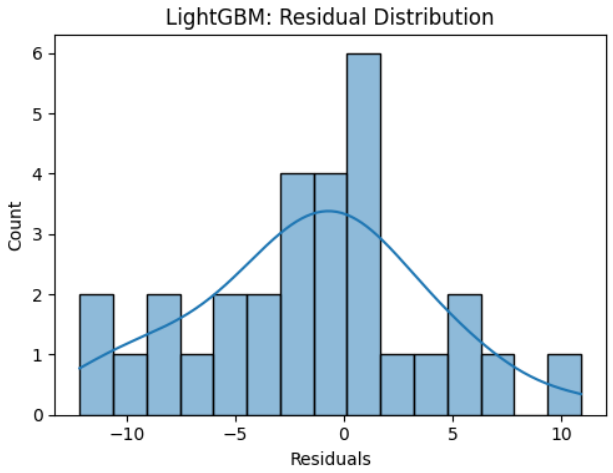


Figure 3: Predicted versus actual irradiance for the Johannesburg hold-out dataset using two tree-based models. **(a)** LightGBM demonstrates strong alignment with observed values, particularly during midday periods of stable irradiance, with minimal dispersion around the diagonal. **(b)** Decision Tree Regressor exhibits greater deviation from the ideal fit, especially under low-light and transitional cloud conditions, reflecting its limited capacity to capture complex non-linear interactions. The contrast between panels underscores the accuracy advantage of gradient boosting over single-tree models for short-horizon solar forecasting.

Error behaviour across forecast horizons revealed a predictable trend: absolute errors increased with longer horizons, from 5 to 30 minutes ahead. This pattern aligns with expectations under cloud advection scenarios, where the absence of explicit motion vector inputs limits temporal extrapolation. Errors were lowest near solar noon, when irradiance is most stable, and highest during dawn, dusk, and under broken cloud conditions. LightGBM residuals exhibited narrower dispersion under variable cloud regimes compared to DTR, indicating better generalization under uncertain atmospheric conditions [2], [6].

(a).



(b).

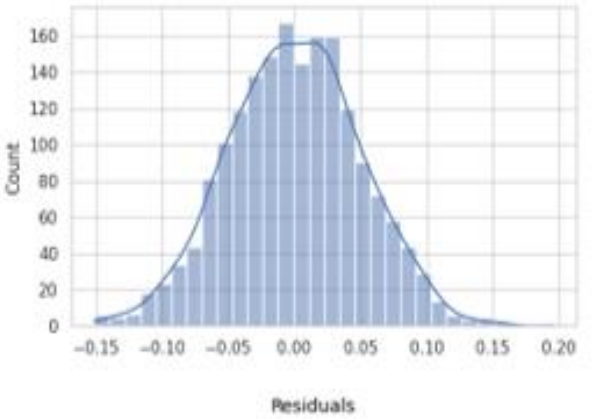
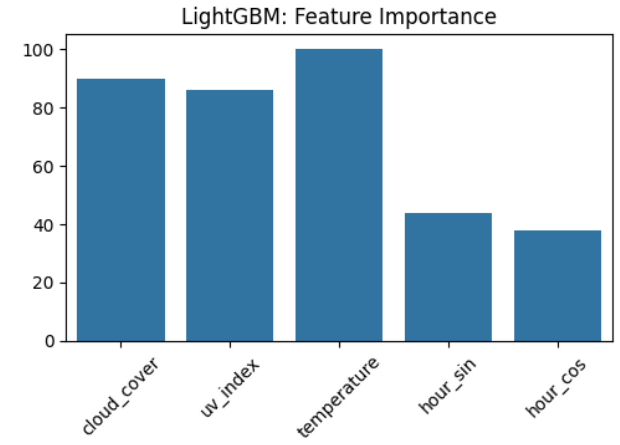


Figure 4: Residual distribution for short-term irradiance forecasts on the Johannesburg hold-out dataset. **(a)** LightGBM residuals exhibit a narrow, symmetric spread centered around zero, indicating unbiased predictions and reduced variance under variable sky conditions. **(b)** Decision Tree Regressor residuals show a wider dispersion, particularly in the tails, reflecting greater sensitivity to cloud transitions and limited capacity to capture complex atmospheric interactions. The contrast between panels highlights the robustness of gradient boosting in minimizing error variability compared to single-tree models.

Feature importance analysis showed strong agreement between the models. Both identified cloud cover and clear-sky

GHI as dominant predictors, followed by UV index and temperature. This ranking reflects the physical relevance of atmospheric attenuation and solar geometry in short-horizon irradiance modelling [2]–[4], [6], [8]. LightGBM's gain-based importance metrics placed particular emphasis on cloud cover, consistent with guidance on interpreting gain versus split frequency in boosted tree frameworks [17]–[19].

(a).



(b).

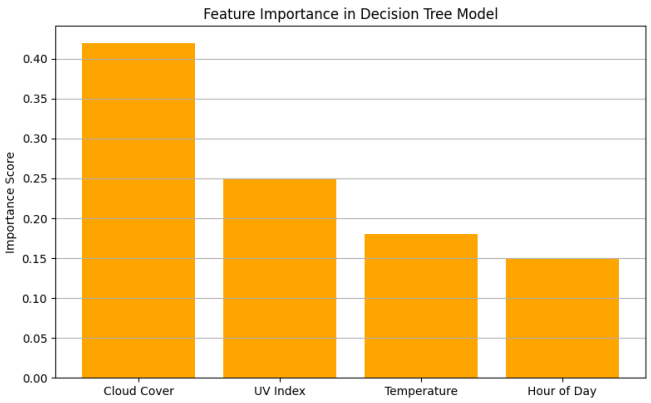


Figure 5: Feature importance rankings for short-term irradiance forecasting models on the Johannesburg dataset. **(a)** LightGBM feature importance based on gain values highlights cloud cover as the most influential predictor, followed by clear-sky GHI, UV index, and temperature. This distribution reflects the model's ability to capture complex interactions between atmospheric conditions and solar geometry. **(b)** Decision Tree Regressor feature importance derived from impurity reduction shows a similar hierarchy, with cloud cover and clear-sky GHI dominating, though the relative weights are less differentiated compared to LightGBM. The contrast between panels illustrates how boosted ensembles emphasize dominant predictors more strongly, improving interpretability and guiding feature engineering for operational forecasting.

balance accuracy and resilience, a hybrid architecture is recommended. In this pattern, LightGBM operates on a local server or cloud platform, leveraging real-time meteorological feeds from OpenWeather and clear-sky irradiance baselines from PVGIS to generate high-quality forecasts. These forecasts can be streamed to edge devices for control decisions. Simultaneously, a single decision tree is retained on the MCU as a fallback mechanism, ensuring continuity of forecast-informed actuation during connectivity lapses [7], [8], [9].

For risk-aware operational strategies such as diverting surplus energy only when forecasted generation exceeds demand by defined margin—probabilistic forecasting becomes essential. LightGBM variants such as LightGBMLSS enable the generation of calibrated prediction intervals and quantiles, supporting uncertainty-aware decision-making under volatile sky conditions. This approach aligns with best practices for operational decision support in distributed energy systems [12], [17], [20], [21].

The proposed framework is particularly relevant in the South African context, where rapid behind-the-meter PV adoption has materially reduced daytime grid demand. Rooftop PV capacity has grown from approximately 2.3 GW in mid-2022 to over 5.2 GW by early 2024, contributing to measurable load-shedding relief [10], [11]. In such environments, accurate site-level short-term forecasting enhances the value of distributed PV by enabling proactive load control and optimized battery dispatch.

V. Conclusion

This study performed a comparative evaluation of Decision Tree Regressor (DTR) and Light Gradient Boosting Machine (LightGBM) for short-term Global Horizontal Irradiance (GHI) forecasting using meteorological features from OpenWeather and clear-sky references from PVGIS. LightGBM consistently outperformed DTR in predictive accuracy, achieving higher R^2 and lower RMSE values, which confirms the strength of gradient boosting algorithms in modelling complex, non-linear atmospheric interactions. Feature importance analysis revealed that cloud cover and clear-sky GHI were dominant predictors for both models, aligning with physical principles and prior research on solar forecasting. These findings demonstrate that ensemble-based methods deliver superior forecast fidelity but impose significant computational overhead, making them unsuitable for direct deployment on resource-constrained microcontrollers.

In contrast, DTR exhibited sub-millisecond inference latency and a memory footprint below 0.1 MB, validating its suitability for embedded systems such as ESP32 where real-time control and low power consumption are critical. This trade-off between accuracy and computational feasibility underscores the need for adaptive architectures that balance high-fidelity forecasting with operational resilience. While this study successfully integrated DTR into an ESP32-based energy management system, achieving real-time surplus energy control, the results highlight opportunities for further optimization in both forecasting accuracy and system robustness. Hardware scalability will also be addressed by evaluating more capable microcontrollers boards, which can support advanced models without compromising real-time responsiveness. Multi-seasonal and multi-regional field trials will be conducted to validate robustness under diverse climatic conditions, and smart load diversification strategies will be implemented to accommodate heterogeneous household profiles. These improvements aim to position SSSEMS-AI as a scalable, uncertainty-aware, and resilient solution for South Africa's rapidly expanding behind-the-meter PV market.

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] A. Masache, P. Mdlongwa, D. Maposa, C. Sigauke, "Short-term forecasting of solar irradiance using decision tree-based models and non-parametric quantile regression," *PLOS ONE*, 2024. https://doi.org/10.1371/journal.pone.0312814

[3] D. El-Shahat, A. Tolba, M. Abouhawwash, M. Abdel-Basset, "ML/DL with grid-search CV for short-term solar irradiance," *Journal of Big Data*, 2024. https://doi.org/10.1186/s40537-024-00991-w

[4] R. I. Nabage *et al.*, "ML-based solar irradiance forecasting in a tropical region," *FETiCON Proc.*, 2025. https://doi.org/10.63748/FETiCON2025.v3i1p18

[5] G. Ke *et al.*, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," *NeurIPS*, 2017. https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

[6] Y. Yang *et al.*, "LightGBM PV prediction integrating meteorology and history," *Energies*, 18(20), 5526, 2025. https://doi.org/10.3390/en18205526

[7] OpenWeather, **Solar Radiation / Solar Energy Prediction API** (docs, products), 2024–2025. https://docs.openweather.co.uk/api/solar-radiation ; https://openweather.co.uk/products/solar-energy-prediction/solar-irradiance

[8] European Commission JRC, **PVGIS—Photovoltaic Geographical Information System**, 2025. https://joint-research-centre.ec.europa.eu/photovoltaic-geographical-information-system-pvgis_en

[9] F. Ben Aicha, "Optimized embedded AI on ESP32-CAM (quantized CNN)," *Computing*, 107, 206, 2025. https://link.springer.com/article/10.1007/s00607-025-01559-z

[10] H. Labuschagne, "South Africa's rooftop solar surge…," *MyBroadband*, Feb. 7, 2024. https://mybroadband.co.za/news/energy/524156-south-africas-rooftop-solar-surge-making-a-big-dent-in-load-shedding.html

[11] N. Ferris, "SA's unprecedented rooftop solar boom," *Energy Monitor*, Aug. 14, 2023. https://www.energymonitor.ai/tech/renewables/weekly-data-south-africas-unprecedented-rooftop-solar-boom/

[12] N. Rubattu, G. Maroni, G. Corani, "Probabilistic LightGBM & temporal hierarchies for load/peak," *ECML AALTD*, 2023. https://ecml-aaltd.github.io/.../Electricity%20Load%20and%20Peak%20Forecasting_...pdf

[13] Taylor & Francis, **LightGBM—Knowledge and References**, 2023–2025. https://taylorandfrancis.com/knowledge/.../LightGBM

[14] C. Bentéjac, A. Csörgő, G. Martínez-Muñoz, "Comparative analysis of gradient boosting algorithms," *AI Review*, 54, 1937–1967, 2021 (open access 2025 page). https://link.springer.com/article/10.1007/s10462-020-09896-5

[15] A. Kumar, "MSE vs RMSE vs MAE vs R-Squared: When to Use?" *VitalFlux*, 2024. https://vitalflux.com/mse-vs-rmse-vs-mae-vs-mape-vs-r-squared-when-to-use/

[16] T. O. Hodson, "RMSE or MAE: when to use them or not," *GMD preprint*, 2022. https://gmd.copernicus.org/preprints/gmd-2022-64/gmd-2022-64.pdf

[17] LightGBM docs, plot_importance and importance types, v4.6.0. https://lightgbm.readthedocs.io/.../lightgbm.plot_importance.html

[18] M. Filho, "How to get feature importance in LightGBM (Python)," *Forecastegy*, 2023. https://forecastegy.com/posts/feature-importance-lightgbm-python-example/

[19] GeeksforGeeks, "LightGBM Feature Importance and Visualization," 2025. https://www.geeksforgeeks.org/machine-learning/lightgbm-feature-importance-and-visualization/

[20] LightGBMLSS (probabilistic LightGBM) — GitHub, PyPI, 2024–2025. https://github.com/StatMixedML/LightGBMLSS ; https://pypi.org/project/lightgbmlss/

[21] J. Amat Rodrigo, J. Escobar Ortiz, "Forecasting time series with LightGBM," *cienciadedatos.net*, 2024–2025. https://cienciadedatos.net/documentos/py58-forecasting-time-series-with-lightgbm