PERFORMANCE ANALYSIS OF LIGHTGBM, XGBOOST, RANDOM FOREST, AND GRADIENT BOOSTING IN PHOTOVOLTAIC ENERGY FORECASTING WITH HYPERPARAMETER OPTIMIZATION

[a]Muhammad Ehtsham*, [b] Marianna Rotilio, [c]Federica Cucchiella

a: Department of Civil, Construction-Architectural and Environmental Engineering, University of L'Aquila, 67100 L'Aquila, Italy.
muhammad.ehtsham@graduate.univaq.it
b: Department of Civil, Construction-Architectural and Environmental Engineering, University of L'Aquila, 67100 L'Aquila, Italy.
marianna.rotilio@univaq.it
c: Department of Industrial and Information Engineering and Economics, University of L'Aquila, 67100, L'Aquila, Italy.
federica.cucchiella@univaq.it
* Corresponding author

Abstract- Accurate photovoltaic (PV) energy forecasting is crucial for effective grid integration and for predictive usage at residential and industrial levels, especially under increasing climatic variability. This study evaluates and compares four machine learning (ML) models, LightGBM, XGBoost, Random Forest, and Gradient Boosting, for hourly PV energy forecasting using real-time data from numerical weather model (NWM), PVGIS, and historical production data from operational PV plant in Southern Italy. Three hyperparameter strategies, namely default settings, Optuna optimization, and Grid Search, were tested. Results show that LightGBM achieved the best performance with Grid Search tuning, yielding an MAE of 2.85 kWh, RMSE of 5.45 kWh, and R^2 of 0.71 over an 8-day forecasting horizon. Comparatively, XGBoost with Grid Search attained an MAE of 3.00 kWh, RMSE of 5.82 kWh, and R^2 of 0.67. The findings highlight that hyperparameter tuning significantly improved forecast accuracy and provide actionable insights for selecting ML models and optimization techniques in PV management systems. Findings are specifically of interest for practitioners, researchers, and organizations associated with PV management and operations.

Keywords- Gradient Boosting, LightGBM, Machine Learning, PV Management

1 Introduction

Extreme climatic phenomena have hugely impacted various parts of the globe in the recent past, and the consequences of environmental change are more evident than ever before. Encouraging communities and policymakers to adopt sustainable and clean practices, including green energy resources [1], [2]. Photovoltaic (PV) energy is one of the most sustainable types of energy sources, and in recent years, the reliance on it has increased notably [3]. In Figure 1, a comparative analysis of the increase in PV and wind energy in Italy as compared to hydro energy, which was previously considered as most green source of energy. However, the most challenging phenomenon related to PV energy is its uncertainty and difficulty in prediction at various parts of the globe, and specifically in diverse terrains and climatic conditions. In this regard, various researchers have adopted, evaluated, and reported the latest technologies. Many researchers nowadays are exploring and reporting the potential of machine learning in different scientific applications.[4], [5], [6]. Similarly, many researchers have exploited the potential applications of machine learning and deep learning technologies in PV-related challenges, including efficiency analysis, forecasting, predictive maintenance, faults and diagnosis, feasibility reports, etc. However, the inherent shortcoming of machine learning models is a continuous need to evaluate and assess the effectiveness of different techniques for hyperparameter optimization. Some researchers have studied the effects of hyperparameter optimization in different fields. However, literature lacks a comprehensive study that evaluates notable models like LightGBM [7], XGBoost [8], Random Forest [9], and Gradient Boosting [10] for PV forecasting and specifically provides a holistic overview of these models in terms of various hyperparameter optimization techniques with real-time data from numerical weather model (NWM), PVGIS [11], and already operational PV plant.

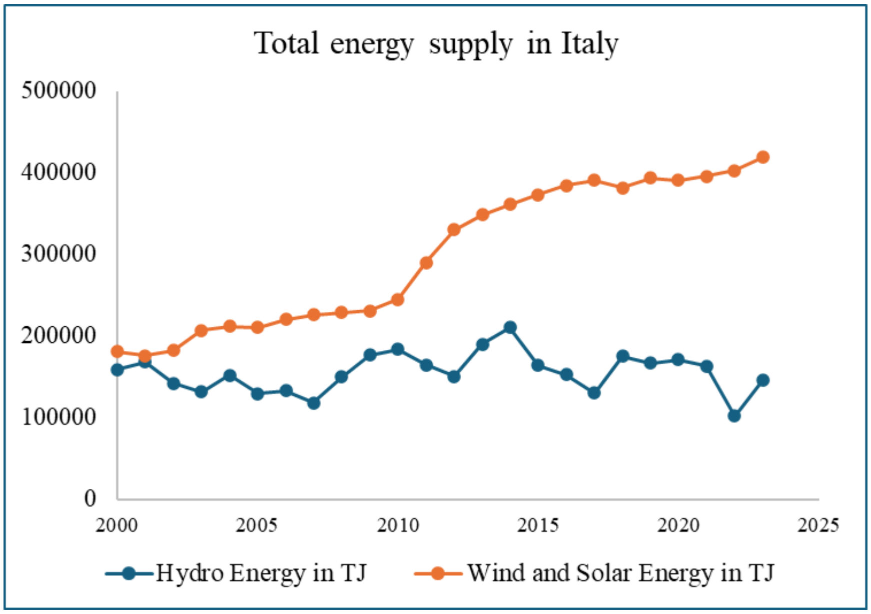

Figure 1: Share of solar, wind, and hydropower energy in Italy. Source: IEA [12].

This study not only provides a framework for utilizing the PVGIS and NWM for hourly PV energy forecasting but also provides a comparative analysis of machine learning models: LightGBM, XGBoost, Random Forest, and Gradient Boosting. In addition, the study tests and evaluates three hyperparameter approaches, namely, Default hyperparameters, Optuna tuning technique [13], and the Grid Search tuning technique [14]. Results of this study can be utilized by the organizations and researchers associated with the management and operation of PV facilities and can provide a clearer path in terms of choosing and deploying of specific hyperparameter techniques.

2 Research Methodology

2.1 PV, NWM, and PVGIS Data

The dataset was constructed by integrating meteorological and PV production data to enable robust machine learning model development for PV energy forecasting. Initially, the PV energy production data of a PV plant, located in southern Italy, were collected. the dataset was accessed from a MySQL database, using specifically tailored queries to convert the 5-minute data recorded by data loggers to hourly cumulative values. Meteorological variables were collected from the OpenWeather [15] NWM, providing high-resolution hourly weather forecasts and reanalysis data. The weather-related features include atmospheric pressure, relative humidity, wind speed, cloud cover, maximum and minimum temperature, as well as solar irradiance components, including direct beam irradiance Gb(i), diffuse irradiance Gd(i), and ground-reflected irradiance Gr(i), along with sunshine duration (hours of sun per day), air temperature measured at 2 meters above ground level (T2m), and wind speed measured at 10 meters (WS10m).

To complement the weather data, hourly data for the same location were retrieved from the PV Geographical Information System (PVGIS), which provides reliable PV performance estimates based on solar radiation and system characteristics. The dataset thus captures both the environmental predictors and the corresponding PV output, facilitating supervised learning. In addition, temporal features were engineered to reflect daily and diurnal patterns, including the full timestamp, date-only fields (to capture day-to-day variability), and time-only fields (to account for intra-day dynamics). This enriched dataset allows the models to learn complex interactions between weather conditions, time-of-day, and PV output, providing a comprehensive basis for evaluating forecasting performance under realistic operational scenarios.

2.2 *Data validation and outlier removal*

Before using the dataset for model training and evaluation, a thorough data validation and cleaning process was carried out to ensure data quality and consistency. The PV production data extracted from the MySQL database were checked for completeness and alignment with the meteorological and PVGIS datasets. Any missing or duplicated records were identified and removed. To address potential measurement errors and anomalous values in the PV and meteorological variables, an outlier detection procedure was applied. For continuous variables such as PV output, irradiance components, and temperature, statistical thresholds were defined based on domain knowledge and descriptive statistics (e.g., interquartile range and physical plausibility limits). Observations falling outside these acceptable ranges were flagged as outliers.

Specifically, PV output values were constrained to be non-negative and not to exceed the rated capacity of the PV plant. Irradiance components were limited to their respective theoretical maxima under clear-sky conditions. Meteorological variables such as temperature, humidity, and wind speed were also cross-checked against climatological norms for the location and time of year. Outliers were treated by either removing the affected records entirely or imputing them with more plausible values based on temporal neighbors or climatological averages, depending on the nature and extent of the anomaly. This validation and cleaning step ensured that the dataset fed into the forecasting models was reliable and representative of real-world operational conditions.

2.3 *Machine learning models and optimization techniques*

For the PV energy forecasting task, four machine learning models were employed: Random Forest, Gradient Boosting, Extreme Gradient Boosting, and Light Gradient Boosting Machine. These tree-based ensemble methods are well-suited for capturing complex, nonlinear relationships in high-dimensional datasets such as the one constructed for this study.

Each model was initially trained using default hyperparameters, as implemented in their respective Python libraries (scikit-learn, xgboost, and lightgbm), to establish baseline performance. To improve forecasting accuracy and assess the sensitivity of the models to hyperparameter settings, two optimization techniques were applied:

- Grid Search: A systematic, exhaustive search over a predefined set of hyperparameter values, combined with 5-fold cross-validation to select the best configuration based on validation performance.
- Optuna: An advanced, efficient hyperparameter optimization framework based on a Bayesian approach, which automatically explores the hyperparameter space and prunes unpromising trials to accelerate convergence to an optimal set of parameters.

2.4 *Efficiency analysis*

These optimization techniques were applied individually to each machine learning model. The objective function in both cases was to minimize forecasting error, evaluating using metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Correlation Coefficient (R^2). The inclusion of both traditional Grid Search and modern Optuna optimization in addition to default settings provided a comprehensive analysis of model performance and robustness under different tuning strategies. By comparing the default, Grid Search optimized, and Optuna-optimized versions of each model, this study highlights the benefits of hyperparameter tuning in improving PV energy forecasting accuracy.

3 Results

3.1 Application of machine learning models for hourly forecasts

All the selected machine learning models were deployed for the forecasting of PV energy production of the selected PV plant for the next 8 days. The models were trained with the hourly meteorological records as training features and hourly PV production as the target variable. In the initial phase, the models were deployed for real-time data, and forecasts were generated using the default hyperparameters of the models. In the successive stages, the hyperparameter tuning was performed with Optuna and Grid Search techniques, and forecasts were generated for the same forecasting horizon. The generated forecasts were compared with the actual recorded data by data loggers. Figure 2 provides an overview of the difference between the actual and recorded forecasts generated by models using default hyperparameters. Figures 3 and 4 provide an overview of the forecasting efficiency of the models using the Optuna and Grid Search methods, respectively.

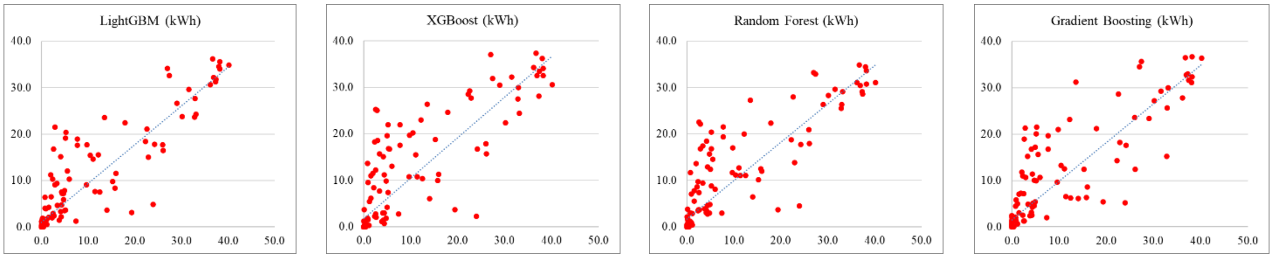

Figure 2: Comparison of forecasted vs recorded values using default hyperparameters

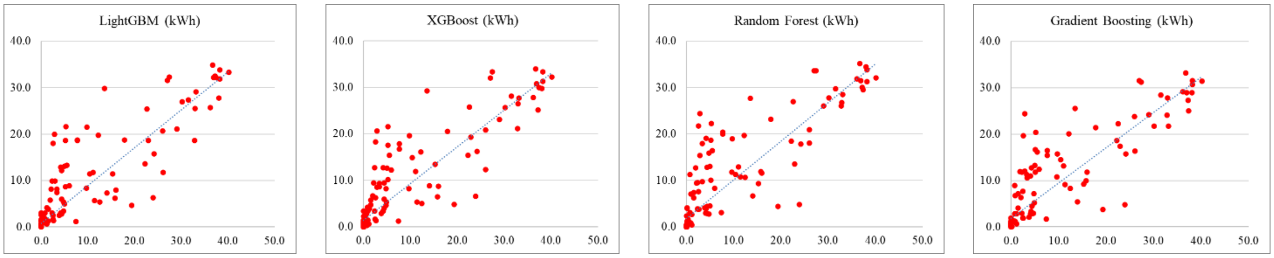

Figure 3: Comparison of forecasted vs recorded values using Optuna hyperparameters

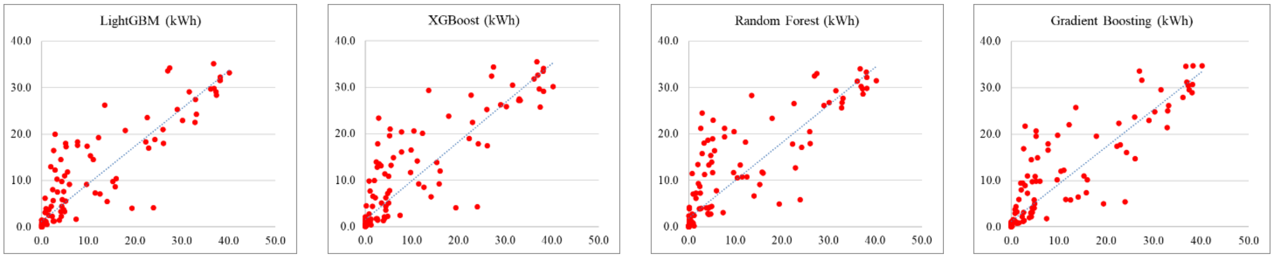

Figure 4: Comparison of forecasted vs recorded values using Grid Search hyperparameters

3.2 Performance of various models and techniques

The performance of models needs to be assessed based on universally accepted and well-known performance metrics. In this regard, the authors decided to calculate MAE, R2, and RMSE of all the models and forecasts generated. Table 1 provides an overview of the estimated performance metrics, creating an opportunity for the readers to assess the accuracy of various models and techniques using the same dataset and forecasting horizon.

Table 1 Evaluation Metrics using all the techniques

Model	MAE (kWh)	R^2	RMSE (kWh)
Default Hyperparameters			
LightGBM	2.81	0.70	5.54
XGBoost	3.68	0.57	6.70
Random Forest	3.01	0.66	5.85
Gradient Boosting	3.41	0.64	6.06
Optuna Hyperparameter Optimization			
Model	**MAE (kWh)**	**R^2**	**RMSE (kWh)**
LightGBM	3.25	0.70	5.63
XGBoost	3.20	0.70	5.62
Random Forest	3.09	0.65	6.00
Gradient Boosting	3.19	0.69	5.66
Grid Search Hyperparameter Tuning			
Model	**MAE (kWh)**	**R^2**	**RMSE (kWh)**
LightGBM	2.85	0.71	5.45
XGBoost	3.00	0.67	5.82
Random Forest	3.15	0.65	6.06
Gradient Boosting	2.92	0.72	5.48

The deployed optimization techniques, Optuna and Grid Search, automatically tuned key hyperparameters for each model to improve their predictive performance. The optimal parameter configurations identified through these methods are summarized in Table 2. Optuna generally favored lower learning rates and higher numbers of estimators, with shallower tree depths for controlling overfitting. Grid Search, on the other hand, selected more conservative settings, balancing model complexity and generalization. These tuned hyperparameters highlight the differences in search strategies and their impact on model configuration

- n_estimators: Specifies the number of boosting rounds or trees in the ensemble. A higher value can improve learning but may increase computation and risk overfitting if not combined with regularization.
- learning_rate: Controls how much each tree contributes to the overall prediction. Lower learning rates (e.g., 0.0138 for Gradient Boosting with Optuna) slow down learning, often requiring more estimators but improving generalization.
- max_depth: Limits the maximum depth of each tree, helping to prevent overfitting by reducing model complexity. Optuna tended to choose shallower depths (e.g., 3) compared to Grid Search.
- num_leaves (LightGBM): Defines the maximum number of leaves per tree. A higher number allows more complex splits, while lower values improve regularization.
- subsample (XGBoost): Specifies the fraction of the training data to use when growing each tree, acting as a regularization mechanism to prevent overfitting by introducing randomness.
- min_samples_split (Random Forest): Sets the minimum number of samples required to split an internal node, which controls the granularity of splits and impacts overfitting.

Table 2 Optimized hyperparameters of the machine learning models using Optuna and Grid Search optimization techniques.

Optuna Hyperparameter Tuning	
	Selected Parameters

LightGBM	'n_estimators': 52, 'learning_rate': 0.114, 'max_depth': 3, 'num_leaves': 23
XGBoost	'n_estimators': 185, 'learning_rate': 0.025, 'max_depth': 3, 'subsample': 0.806
Random Forest	'n_estimators': 173, 'max_depth': 15, 'min_samples_split': 5
Gradient Boosting	'n_estimators': 214, 'learning_rate': 0.0138, 'max_depth': 5
Grid Search Hyperparameter Tuning	
	Selected Parameters
LightGBM	'n_estimators': 100, 'learning_rate': 0.05, 'max_depth': 6, 'num_leaves': 31
XGBoost	'n_estimators': 100, 'learning_rate': 0.05, 'max_depth': 6, 'subsample': 1.0
Random Forest	'n_estimators': 200, 'max_depth': 8, 'min_samples_split': 5
Gradient Boosting	'n_estimators': 100, 'learning_rate': 0.05, 'max_depth': 4

4 Practical Implementation

The findings of this study evaluate and report the effectiveness of machine learning models, particularly tree-based ensemble methods, for forecasting PV energy production using meteorological and temporal features. Among the evaluated models, LightGBM and Gradient Boosting consistently achieved the best performance across all optimization strategies, with MAE as low as 2.76 kWh and R^2 reaching up to 0.72, a similar trend was reported by [16], [17]. This level of accuracy highlights the feasibility of integrating such models into real-world PV system management and energy planning applications. Specifically, grid operators and energy managers can utilize these models for short-term PV generation forecasts to optimize grid balancing, reduce reliance on fossil fuel-based backup generation, and improve scheduling of energy storage systems. Furthermore, the ability to tune models using optimization techniques such as Grid Search and Optuna enhances forecasting robustness, ensuring that models can be adapted to specific sites and operational requirements. The findings also underscore the importance of using high-resolution weather forecasts and properly engineered time-based features to capture the variability inherent in solar energy production, making these approaches highly applicable for smart grid and sustainable energy initiatives, in different parts of the globe.

5 Conclusion

The following conclusions can be drawn from the conducted study:

- The study highlights the potential of machine learning, particularly tree-based ensemble models, in effectively forecasting PV energy production using remotely sensed meteorological and historical PV production data.

- By comparing default settings with two hyperparameter optimization techniques, the study demonstrates that tuning significantly improves model performance, with LightGBM and Gradient Boosting showing the most promising results.

- The findings offer a practical framework for PV operators and researchers to select fine-tune models for more reliable hourly energy forecasts, supporting better planning and integration of renewable energy into the grid and for energy usage optimization at residential and industrial levels.

Future work can explore incorporating additional data sources, like more advanced NWM, satellite imagery, IoT sensors, and applying advanced deep learning models to capture complex patterns. Researchers can also explore the latest hybrid models and incorporate long-term climatic trends. Developing adaptive algorithms to handle climate anomalies and creating practical tools for PV operators.

Acknowledgment

This research is carried out as part of the XXXVIII Cycle of the PhD programme in Civil, Building-Architecture and Environmental Engineering at the University of L'Aquila. The research topic is "Monitoring System for the Production of Renewable Energy and Data Acquisition System" co-financed by the company Solis SpA. The authors express their gratitude to Solis SpA for their invaluable cooperation in this study. This encompasses providing technical and production data for BIPVs and BAPVs, supplying the necessary computer processing system and required software, as well as the contributions of the monitoring team and the company's technical head in validating the research findings. The authors would also like to express their gratitude to the developers of OpenWeather for providing the meteorological data. We would like to express our sincere appreciation to PVGIS (Photovoltaic Geographical Information System) and European Commission's Joint Research Centre for providing data resources. The research supervisor is Prof. Marianna Rotilio, co-supervisors are Prof. Federica Cucchiella and Prof. Gianni Di Giovanni, and the industrial tutor is Massimo Innaurato. The Ph.D. student in charge of the research is Eng. Muhammad Ehtsham

References

[1] F. Cucchiella, M. Rotilio, L. Capannolo, and P. De Berardinis, "Technical, economic and environmental assessment towards the sustainable goals of photovoltaic systems," *Renewable and Sustainable Energy Reviews*, vol. 188, p. 113879, 2023.

[2] F. Cucchiella *et al.*, "Renovation Wave: a bioeconomy panel produced with waste.," *J Clean Prod*, p. 142868, 2024.

[3] M. Ehtsham, M. Rotilio, and F. Cucchiella, "Deep learning augmented medium-term photovoltaic energy forecasting: A coupled approach using PVGIS and numerical weather model data," *Energy Reports*, vol. 13, pp. 4299–4317, Jun. 2025, doi: 10.1016/J.EGYR.2025.03.058.

[4] G. Di Giovanni, M. Rotilio, L. Giusti, and M. Ehtsham, "Exploiting building information modeling and machine learning for optimizing rooftop photovoltaic systems," *Energy Build*, p. 114250, 2024.

[5] Y. Li *et al.*, "A machine-learning approach for regional photovoltaic power forecasting," in *2016 IEEE Power and Energy Society General Meeting (PESGM)*, IEEE, 2016, pp. 1–5.

[6] T. S. Qaid, H. Mazaar, M. Y. H. Al-Shamri, M. S. Alqahtani, A. A. Raweh, and W. Alakwaa, "Hybrid deep-learning and machine-learning models for predicting COVID-19," *Comput Intell Neurosci*, vol. 2021, no. 1, p. 9996737, 2021.

[7] G. Ke *et al.*, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," *Adv Neural Inf Process Syst*, vol. 30, 2017, Accessed: Jul. 14, 2025. [Online]. Available: https://github.com/Microsoft/LightGBM.

[8] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 13-17-August-2016, pp. 785–794, Aug. 2016, doi: 10.1145/2939672.2939785/SUPPL_FILE/KDD2016_CHEN_BOOSTING_SYSTEM_01-ACM.MP4.

[9] L. Breiman, "Random forests," *Mach Learn*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324/METRICS.

[10] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Front Neurorobot*, vol. 7, no. DEC, p. 63623, Dec. 2013, doi: 10.3389/FNBOT.2013.00021/BIBTEX.

[11] "Photovoltaic Geographical Information System (PVGIS) Available at: https://joint-research-centre.ec.europa.eu/photovoltaic-geographical-information-system-pvgis_en, Last accessed on: Feb. 15 2025."

[12] "International Energy Agency (IEA). Finland. Available at: https://www.iea.org/countries/finland. Accessed: 03 March 2025."

[13] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2623–2631, Jul. 2019, doi: 10.1145/3292500.3330701.

[14] H. Alibrahim and S. A. Ludwig, "Hyperparameter Optimization: Comparing Genetic Algorithm against Grid Search and Bayesian Optimization," *2021 IEEE Congress on Evolutionary Computation, CEC 2021 - Proceedings*, pp. 1551–1559, 2021, doi: 10.1109/CEC45853.2021.9504761.

[15] "OpenWeatherMap. Available at: https://openweathermap.org/ Last accessed on January 16, 2025.".

[16] S. Rezashoar, E. Kashi, and S. Saeidi, "A hybrid algorithm based on machine learning (LightGBM-Optuna) for road accident severity classification (case study: United States from 2016 to 2020)," *Innovative Infrastructure Solutions*, vol. 9, no. 8, pp. 1–22, Aug. 2024, doi: 10.1007/S41062-024-01626-Y/METRICS.

[17] K. Ng and P. Lei, "A Lightweight Method using LightGBM Model with Optuna in MOOCs Dropout Prediction," *ACM International Conference Proceeding Series*, pp. 53–59, Jul. 2022, doi: 10.1145/3551708.3551732;TAXONOMY:TAXONOMY:CONFERENCE-COLLECTIONS;WGROUP:STRING:ACM.