Proceedings of the 2024 Winter Simulation Conference
H. Lam, E. Azar, D. Batur, S. Gao, W. Xie, S. R. Hunter, and M. D. Rossetti, eds.

ADVANCED SOLAR POWER FORECASTING: A HYBRID/ENSEMBLE APPROACH UTILIZING GEOGRAPHIC AND METEOROLOGICAL DATA

Mahdi Darvishi[1] and Abdolreza Abhari[2]

[1]Fac. of Eng. and Architectural Scie., Toronto Metropolitan University, Toronto, ON, CANADA
[2]Dept. of Computer Science Eng,. Toronto Metropolitan University, Toronto, ON, CANADA

ABSTRACT

Photovoltaic (PV) systems are pivotal in the global energy transition, where accurate solar power forecasting is critical. Traditional forecasting has leaned heavily on solar irradiance data, yet such reliance carries inherent uncertainties and measurement complexities, presenting significant forecasting challenges. This paper introduces a novel hybrid/ensemble model that reduces dependence on solar irradiance data, utilizing geographic, meteorological, and temporal data to predict solar power output. Combining the strengths of XGBoost and LightGBM algorithms through a linear regression meta-model, our approach demonstrates improved prediction accuracy, evidenced by a mean absolute error (MAE) of 0.033, and an R-squared value of 0.693. This study advances solar power forecasting, enhancing PV system efficiency, and reliability, and promoting sustainable energy investments.

1 INTRODUCTION

Photovoltaic (PV) systems have emerged as a dominant technology in renewable energy due to their direct conversion of solar energy into electricity, leading to substantial growth in global installations. From 2000 to 2011, the installed capacity of PV systems worldwide increased dramatically from 1 GW to 67 GW, influenced by several pivotal factors (Das et al. 2018).

As mentioned, photovoltaic (PV) technology has rapidly evolved, offering an environmentally friendly alternative to conventional energy sources like thermal power. Unlike these traditional sources, PV systems are subject to high variability and uncertainty primarily due to fluctuating atmospheric conditions such as solar irradiance, temperature, and wind speed. This variability can occur across different timescales—from seconds to hours—posing significant challenges for accurate forecasting and balancing electricity generation with demand. Moreover, enhancements in prediction accuracy could substantially reduce operational costs (Oneto et al. 2018).

Forecasting for horizontal PV arrays is primarily driven by spatial constraints and performance characteristics under diffuse irradiation conditions. Horizontal installations, such as those on building rooftops, are often the only viable option in urban or densely built environments where space for larger, angled solar arrays is unavailable (Kaur et al. 2016) These arrays are particularly advantageous as they perform more effectively under diffuse light conditions compared to latitude-tilted panels, which are optimized for direct sunlight (Liu et al. 2019; Pedro et al. 2012). This makes horizontal arrays suitable for regions with frequent overcast conditions, thus broadening the applicability of solar energy.

2 FORECASTING METHODS OVERVIEW

Solar power forecasting can be broadly categorized into three primary methods: physical methods, time-series statistical methods, and hybrid techniques, each serving distinct forecasting needs based on the data availability and the required precision of the forecasts.

2.1 Physical Methods

These involve using mathematical models to simulate the physical states and dynamics of the atmosphere that affect solar irradiance. These methods are particularly useful for short-term forecasts where immediate local environmental changes play a significant role (Kaur et al. 2016; Liu et al. 2019; Ogliari et al. 2013). They heavily rely on accurate weather forecasting data and can be less effective when sudden meteorological changes occur.

2.2 Time-Series Statistical Methods

These are data-driven approaches that use historical data to identify patterns and predict future outcomes. They can be subdivided into Regressive Methods and Machine Learning-Based Models.

Regressive methods, establish relationships between dependent variables (solar power output) and independent predictors using statistical techniques. While they generally provide stable forecasts, their performance can deteriorate with abrupt weather changes (Antonanzas et al. 2016; Raza et al. 2016).

Machine Learning-Based Models include neural networks (NNs), support vector machines (SVM), random forests (RF), deep learning methods, and k-nearest neighbors (KNN). These models excel in capturing complex, nonlinear relationships in data and can adaptively learn from past trends to improve accuracy in solar power forecasting (Rana and Rahman 2020; Han et al. 2019; Liu et al. 2019; Vandeventer et al. 2019; Agoua et al. 2018).

2.3 Hybrid/Ensemble Models

As discussed, these models combine two or more forecasting methods along with optimization algorithms to harness the benefits of each, thereby enhancing predictive accuracy. They are particularly useful in scenarios where a single model's approach is insufficient to handle the complexity or variability of the data (Kaur et al. 2016; Ogliari et al. 2013). Hybrid models are often designed to complement the strengths and mitigate the weaknesses of their constituent models, offering a balanced approach to forecasting that can adapt to varying data qualities and environmental conditions.

Hybrid or ensemble models represent a significant advancement in the domain of solar power forecasting due to their ability to amalgamate the strengths of multiple forecasting techniques, thereby enhancing overall accuracy. These models leverage diverse methodologies, including statistical, machine learning, and physical approaches, to capture a comprehensive range of dynamics and relationships within solar power data. For example, integrating Support Vector Regression (SVR) models optimized with Genetic Algorithms (GA) has been shown to yield higher accuracy by effectively tuning model parameters and adapting to non-linear patterns in solar power generation data (Kaur et al. 2016; Liu et al. 2019). Furthermore, hybrid models can mitigate the individual limitations of single-method forecasts by combining their predictive capabilities, thus providing a more robust and reliable forecast even under variable and complex environmental conditions. The strategic combination of different models helps reduce the bias and variance associated with standalone models, leading to improved predictive performance, especially in the inherently unpredictable field of solar power due to weather variabilities (Acar and Rais-Rohani 2009). The advancement of hybrid/ensemble models in PV system forecasting represents a significant leap forward, particularly for managing the complexities of horizontal PV arrays in varied environmental conditions. These models enhance forecast accuracy by integrating diverse methodologies, proving essential in adapting to the dynamic nature of solar power generation. This project contributes a novel aspect by quantifying the efficacy of machine learning algorithms to predict horizontal photovoltaic power output without direct irradiation data, employing a hybrid/ensemble approach. By utilizing geographic, meteorological, and temporal data, our study advances the predictability and operational efficiency of solar power systems, setting new benchmarks for the industry.

3 HYBRID/ENSEMBLE MODELS IN PV SYSTEMS FORECASTING

Hybrid/Ensemble modeling has emerged as a pivotal approach in forecasting PV system power output, employing various sophisticated techniques to manage the complex dynamics involved. Artificial Neural

Networks (ANN) represent a foundational meta-modeling approach, utilized extensively for their robust predictive capabilities across multiple scenarios (Omar et al. 2016) These models are particularly effective in handling nonlinear data and adapting to new input patterns through learning algorithms. For instance Omar et al. (2016) enhanced the day-ahead power prediction accuracy of a solar facility using an MLP feed-forward ANN, trained on the historical power output data of the plant.

Further advancements in meta-modeling include the integration of genetic algorithms and particle swarm optimization with neural networks to fine-tune prediction accuracies (Li et al. 2016). These techniques adjust the network parameters optimally, navigating the complex landscape of possible solutions more efficiently than traditional methods

The Analog Ensemble (AnEn) method, another significant development, offers short-term probabilistic forecasts based on historical data sets from Numerical Weather Prediction models. This method, combined with ANNs, has demonstrated superior forecasting performance in large-scale computations (Alessandrini et al. 2015).

Other notable methods include the Multivariate Adaptive Regression Spline (MARS) model, which builds relationships between variables without prior assumptions, and Support Vector Regression (SVR), which has been shown to correlate well with seasonal variations in power output. The integration of discrete wavelet transformation with the SVM framework illustrates another innovative approach, where data are decomposed to enhance the prediction quality (Deo et al. 2016).

Models often segment the forecasting process into distinct stages, employing specific techniques tailored to each phase. For example, Yang et al. (2014) proposed a three-step model using classification, training, and forecasting stages, applying different methodologies like Self-Organizing Map (SOM), Learning Vector Quantization (LVQ) networks, SVR, and fuzzy inference methods across these stages.

The development and refinement of hybrid models represent a significant trend in meta-modeling, combining multiple approaches to optimize prediction accuracy. This strategy is rooted in earlier works by Perrone and Cooper (1993), which advocated for committees of neural networks. Subsequent studies have expanded on this concept, using surrogates and optimized weight coefficients to integrate various models into a cohesive forecasting framework (Zerpa et al. 2005; Acar and Rais-Rohani 2009).

4 PROBLEM STATEMENT

Solar irradiance is a pivotal factor in forecasting solar power output; however, its utilization faces significant challenges. Firstly, the measurement of solar irradiance can be intricate and time-consuming, especially at specific sites. This results in data that may contain uncertainties ranging from 8% to 25%, leading to potential inaccuracies in power output forecasts (Qing and Niu 2018; Yaniktepe and Genc 2015; Letendre et al. 2014). These uncertainties stem from variabilities in weather conditions and the inherent limitations of the models used to predict irradiance. Moreover, reliance on irradiance data can lead to forecast errors that complicate the evaluation of photovoltaic (PV) performance under varied environmental conditions (Pasion et al. 2020).

This study showcases a hybrid/ensemble model that adeptly predicts solar power output. This methodology addresses the challenges associated with the dependency on irradiance data by testing the hypothesis that machine learning algorithms like XGBoost and LightGBM can accurately predict power output while avoiding the uncertainties of irradiance measurements. This method combines the strengths of both physical and statistical forecasting techniques to address the variability in solar power generation.

5 DATASET DESCRIPTION

In this research, we utilized a comprehensive dataset used in another article (Pasion et al. 2020), which was collected from 12 strategically chosen locations across Department of Defense (DoD) installations within 25 regions. This selection was informed by a meticulous application of climate classification matrices and Pareto analysis to ensure a representative sample of climate regions for our study (UCAR Center for Science Education 2020).

The dataset comprises several key variables, each chosen for its relevance to solar power generation the following is the list of variables:

- Cloud Ceiling: Clouds scatter and block sunlight, directly affecting solar irradiance. The cloud ceiling measures this effect by recording the altitude at which significant cloud coverage occurs, indicating how much sunlight reaches the PV panels (Lave and Kleissl 2011).
- Latitude: influences the angle at which sunlight hits the Earth, affecting the intensity and duration of solar exposure on the panels, which is essential for determining potential energy output (Faine et al. 1991).
- Month and Hour: Essential for capturing seasonal and daily variations in sunlight, impacting solar panel efficiency and power output predictions (Faine et al. 1991).
- Humidity: Impacts panel efficiency through effects on sunlight refraction, diffraction, and reflection, and influences dust accumulation on panels (Mekhilef et al. 2012).
- Temperature: Directly impacts PV panel efficiency; generally, higher temperatures decrease performance (Kayri et al. 2017).
- Wind Speed: Can improve panel efficiency by cooling, or contribute to dust accumulation, affecting irradiance absorption and efficiency (Lu and Zhao, 2018).
- Visibility: Measures atmospheric clarity, influencing the amount of sunlight reaching the panels.
- Pressure: Helps forecast weather-related impacts on PV performance by indicating impending weather changes (UCAR Center for Science Education 2020).
- Altitude: At higher altitudes, the thinner atmosphere allows more sunlight to reach the panels, potentially increasing solar power output.

6 EQUIPMENT & MATERIALS

The study involved deploying photovoltaic (PV) systems across various United States Air Force (USAF) installations globally to monitor and analyze the performance of these renewable energy sources.

The primary equipment included:

Renogy 50-watt, 12-volt, polycrystalline PV panels: Selected for their cost-effectiveness and durability, suitable for different environmental conditions.

Raspberry Pi 3, model B, version 1.2 computers: Used to record essential data such as panel power output and environmental conditions at 15-minute intervals.

These devices are housed in Waterproof Pelican cases: These cases protect the Raspberry Pi computers from environmental elements, ensuring data integrity.

Connectivity was maintained through CAT cables and power cables, and data storage was handled using SD cards (Pasion et al. 2020).

7 METHODOLOGY

7.1 Data Preprocessing

In advancing our solar power forecasting model, a rigorous data preprocessing strategy was essential to prepare the dataset for effective machine learning analysis. This section delves into the nuanced preprocessing techniques we employed.

Our preprocessing began with converting date strings into Date-Time objects, which facilitated the extraction of useful temporal features such as the day of the week and month of the year. This initial step was crucial for the subsequent creation of interaction terms, enabling our model to capture interactions between different temporal units that could significantly affect solar power generation.

Environmental factors like temperature and wind speed are critical for predicting solar power output. To address this, we employed rolling averages and exponentially weighted moving averages for these variables to smooth out fluctuations and highlight longer-term trends. This treatment mitigates the impact of transient anomalies on the prediction model, focusing the learning process on underlying trends that are more predictive of solar output.

To capture non-linear relationships and interactions between features, we applied polynomial transformations to our data, generating new features that represent interaction terms up to the second degree.

This approach allows our models to consider complex interdependencies between variables, which is often crucial for accurate energy forecasting.

7.1.1 Cyclic Feature Transformation

Solar power generation is profoundly influenced by seasonal and diurnal patterns due to the Earth's orbit around the sun and its axial tilt. To adequately model these temporal dynamics, we transformed time-related attributes into cyclic features. This transformation was accomplished by applying trigonometric functions—sine and cosine—to time variables such as month, day, and hour (1), (2):

$$x_{\sin} = \sin\left(\frac{2\pi \times x}{T}\right) \tag{1}$$

$$x_{cos} = \cos\left(\frac{2\pi \times x}{T}\right) \tag{2}$$

Where: x represents the time unit(month, day, or hour) and T is the total number of units in one cycle (12 for months, 30 for days, and 24 for hours).

These cyclic transformations help preserve the continuity of time, where the end of one cycle seamlessly connects back to the beginning (e.g., 23:00 hours is close to 00:00 hours), thereby facilitating the model's ability to learn and predict seasonal trends and daily patterns effectively.

7.1.2 Solar Position Calculations

The availability and intensity of sunlight, which directly impact solar power output, vary primarily with the geographical position and the time of the year. To quantitatively assess this variation, we computed solar position metrics, including solar elevation and azimuth angles, using established solar geometry algorithms. These metrics are essential for determining how sunlight interacts with solar panels at different times and locations.

The solar elevation angle (α) is calculated using the formula (3) and The solar azimuth angle (γ) from the north, turning positive towards the east is calculated as (4).

$$\alpha = \arcsin(\sin(\delta) \times \sin(\phi) + \cos(\delta) \times \cos(\phi) \times \cos(H)) \tag{3}$$

$$\gamma = \arctan\left(\frac{\sin(H)}{\cos(H) \times \sin(\phi) - \tan(\delta) \times \cos(\phi)}\right) \tag{4}$$

Where: δ is the solar declination, ϕ is the latitude of the location.

H is the hour angle, which is 15° per hour away from solar noon (positive in the afternoon and negative in the morning).

These calculations provide a precise measure of the sun's position relative to the specific location of the photovoltaic (PV) installations. By incorporating these geographical dynamics into the model, we enhance its capability to predict solar power output by aligning the predicted values more closely with the actual solar path and its interaction with the Earth's surface. This approach not only leverages the cyclic nature of solar movement but also accounts for geographical variations that can significantly influence solar power generation.

7.1.3 Feature Selection

In our study, we applied Recursive Feature Elimination (RFE) to fine-tune the feature set for our solar power forecasting models, specifically LightGBM and XGBoost. This technique progressively eliminates the least significant features based on their impact on model performance, thereby optimizing the feature space.

Our initial RFE evaluation covered the entire set of available features. This comprehensive assessment indicated a noticeable improvement in model performance within the feature count range of 95 to 125. The performance not only improved dramatically in this range but also stabilized, suggesting a concentration of critical predictive features. Motivated by this observation, we conducted a focused analysis within this range to pinpoint the most effective number of features. This detailed exploration identified optimal counts of 109 features for LightGBM and 103 features for XGBoost, where each model reached its lowest negative mean squared error (Neg MSE). The results for LightGBM are shown in Figure 1, as an example of the analysis process.

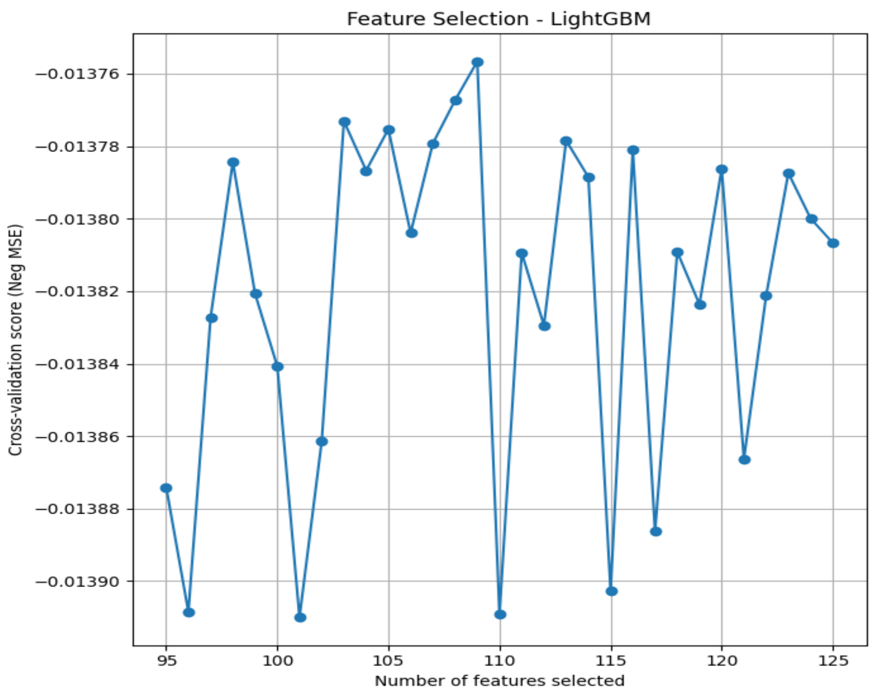

Figure 1: Analysis of MSE across varying numbers of RFE features.

7.2 Model Training

In the development of the hybrid/ensemble model presented in this study, XGBoost (Extreme Gradient Boosting) and LightGBM (Light Gradient Boosting Machine) were selected for their exceptional ability to handle complex, non-linear relationships inherent in environmental data.

7.2.1 Selection Rationale of Base Models

XGBoost (Extreme Gradient Boosting) is a decision-tree-based ensemble machine learning algorithm that operates within a gradient boosting framework. The core of XGBoost's functionality is its objective function (5), which it seeks to minimize:

$$L(\theta) = \sum_{i=1}^{n} \left[L\left(y_i , \hat{y}_i^{(t)}\right) + \sum_{k=1}^{t} \Omega(f_k) \right] \quad (5)$$

In this function: $L(y_i, \hat{y}_i^{(t)})$ represents the loss function, quantifying the prediction error between the forecasted $\hat{y}_i^{(t)}$ and the actual y_i values.

$\Omega(f_k)$ denotes the regularization term for each tree f_k, which helps prevent overfitting by penalizing the complexity of the model.

The regularization component is further broken down as $\Omega(f) = \gamma T + 1/2\,\lambda\|w\|^2$, where T is the number of leaves, w is the leaf weights, γ controls the contribution of the number of leaves, and λ moderates the leaf weights.

XGBoost incorporates both L1 (lasso regression) and L2 (ridge regression) regularization techniques through its regularization terms. These methods help enhance the performance and generalizability of the model by constraining the complexity of the learned structures. Moreover, XGBoost is uniquely equipped to handle missing data, automatically attributing default directions on nodes of trees when encountering missing values. This capability is particularly beneficial in fields like meteorology where data can be incomplete but critical for accurate forecasting.

The integration of these features makes XGBoost exceptionally effective for datasets where complete data collection poses a challenge, such as in solar forecasting applications. By robustly addressing both missing data and the risk of model overfitting, XGBoost provides a reliable, high-performance tool for predictive tasks involving complex environmental data.

LightGBM (Light Gradient Boosting Machine) advances the gradient boosting framework by employing a histogram-based technique that buckets continuous feature values into discrete bins. This method effectively reduces memory usage and accelerates computations, which are vital when dealing with extensive datasets. The discretization process can be described mathematically by the formula (6):

$$Feature_k = \left\lceil \frac{value - \min(value)}{width\ of\ bin} \right\rceil \quad (6)$$

In this formula: $value$ represents the continuous feature value, min ($value$) is the minimum value of the feature across the dataset, and $width\ of\ bin$ defines the interval size into which the values are grouped.

This quantization simplifies the model's calculations during tree building by reducing the complexity of decisions at each node split. The ability to efficiently process these discretized values enhances LightGBM's performance, particularly in large-scale applications.

Moreover, LightGBM seamlessly integrates support for categorical features, further enhancing its utility in diverse applications. This direct support avoids the overhead typically associated with non-numeric data processing in machine learning tasks.

LightGBM's design is optimized for high computational and memory efficiency without compromising the accuracy of the model. This makes it exceptionally suitable for real-time applications such as solar power forecasting, where quick data processing is crucial. The speed with which LightGBM handles large volumes of data makes it an invaluable tool for dynamic environments that require rapid and accurate predictive capabilities.

7.2.2 Rationale for Employing Linear Regression as Meta-Model

The deployment of linear regression as the meta-model within our ensemble framework is a strategic decision aimed at enhancing the model's interpretability and operational efficacy. This choice is critical for ensuring the accuracy, simplicity, and interpretability of solar power forecasting, which are pivotal for real-world applications.

Linear regression excels in aggregating outputs from multiple models, such as LightGBM and XGBoost, which capture different aspects of the dataset's complexity. This integration uses a statistically robust weighted average approach, optimizing prediction accuracy by combining diverse model insights. This method not only mitigates the error variance across predictions but also provides a transparent rationale for each prediction's contribution to the final output.

Operational advantages of linear regression include its computational efficiency, involving fewer parameters and less computational overhead than more complex algorithms. This simplicity enables quicker recalculations and updates, essential in real-time forecasting scenarios where prompt decision-making is crucial. Additionally, the reduced computational demand facilitates scalability and efficient deployment across various technological platforms.

Lastly, the transparency of linear regression greatly enhances the interpretability of the model. Stakeholders can easily understand the derivation of predictions, fostering trust and providing actionable insights. This clarity also simplifies the process of model validation and auditing, crucial in settings where understanding the basis of algorithmic decisions is necessary.

8 RESULTS

8.1 Model Evaluation Metrics

The meta-model achieved promising results with a Mean Absolute Error (MAE) of 0.033 and a Root Mean Squared Error (RMSE) of 0.116, indicating a tight grouping of predicted values around the actual data points. The model's R-squared value stood at 0.693, reflecting that approximately 69.3% of the variance in the dependent variable is predictable from the independent variables.

8.1.1 Cross-Validation Consistency

To ascertain the stability and reliability of the model across different subsets of data, we employed a 10-fold cross-validation approach. The average R-squared value across all folds was 0.692, with individual folds ranging from a minimum of 0.665 to a maximum of 0.717, as depicted in Figure 2. This analysis demonstrates the model's consistent performance, even with variations in the data subsets, underscoring its robustness.

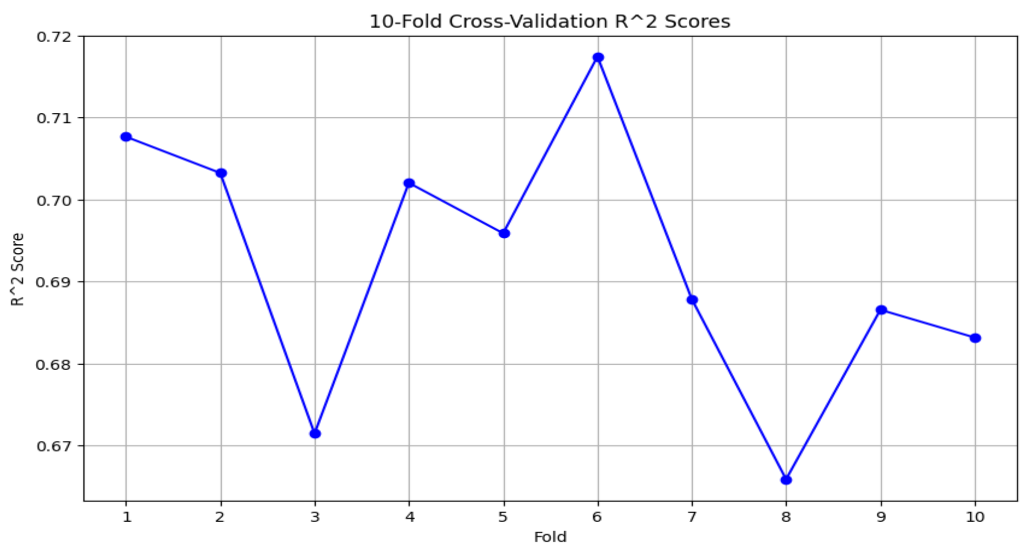

Figure 2: 10-Fold cross-validation R^2 scores.

8.2 Predictive Performance Visualization

Further insight into the model’s predictive accuracy is garnered through a plot comparing actual and predicted values of solar power output Figure 3. The visualization illustrates a close alignment between the predicted and actual values across 4,000 samples, with the model effectively capturing the trends and fluctuations in solar power output throughout the dataset.

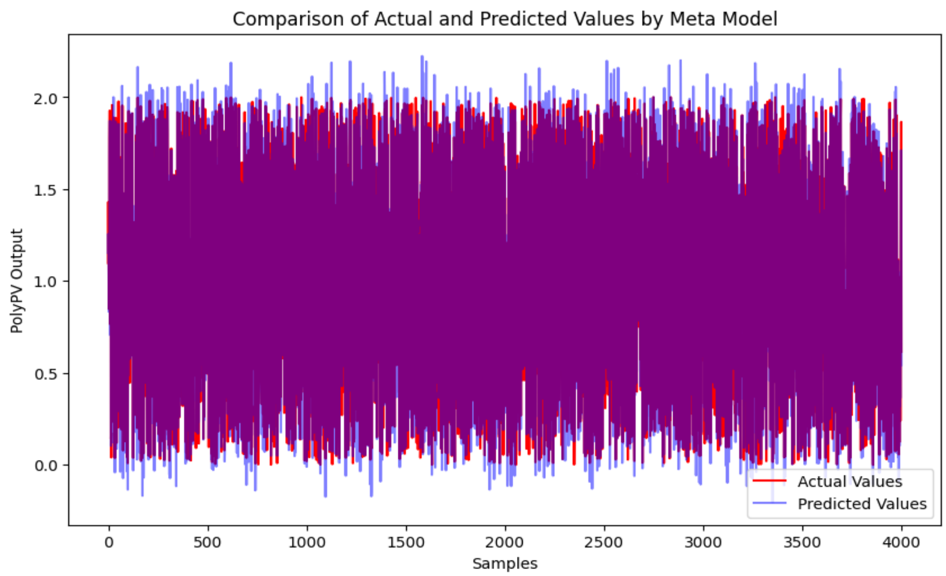

Figure 3: Comparison of actual and predicted values by meta model.

9 DISCUSSION

Comparative analysis in solar power forecasting is challenging for our study due to the common reliance on solar irradiance data in previous research. This reliance complicates direct comparisons, as our methodology does not utilize solar irradiance data. Fortunately, the study by Pasion et al. (2020), which also forecasts solar power without relying on solar irradiance data, provides a suitable basis for comparison. This section will discuss the relative effectiveness of our methodology in relation to their approach.

In the evaluation of predictive models for solar power forecasting, understanding the trade-offs between different performance metrics and their implications on real-world applications is essential. This analysis compares two models based on multiple criteria: R-squared, Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and cross-validation R-squared, See Table 1, providing a comprehensive overview of each model's strengths and weaknesses.

Table 1: Comparison metrics.

Metrics	Present Work	(Pasion et al. 2020)
Model	Meta-Model	DRF
R^2	0.693	0.939
MAE	0.033	1.176
RMSE	0.116	1.754
Cross-Validation R^2	0.692	0.673

9.1 Pasion et al. (2020) Results

Pasion et al. report a high R-squared value of 0.93, indicating an excellent capacity to capture the variance in the dependent variable from the predictors within their training dataset. This high R-squared is preferable in scenarios where the primary goal is to explain the influence of independent variables on the dependent variable, relevant in studies aimed at understanding the driving factors behind observed phenomena. However, this model also exhibits high MAE and RMSE values of 1.176 and 1.754, respectively, indicating substantial average prediction errors. Critically, the cross-validation R-squared for this model drops to 0.673, a significant decrease from its initial R-squared. This discrepancy suggests that while the model performs well on the training data, it fails to generalize effectively to unseen data, a likely indication of overfitting. Overfitting occurs when a model fits the training data too closely, capturing noise as if it were true underlying patterns. This typically happens with overly complex models, but it can also occur due to insufficient training data or high variability within the data. Overfitting diminishes the model's predictive reliability in practical scenarios, highlighting the importance of balancing model complexity with data quality and quantity.

9.2 Present Work Results

Contrastingly, our meta-model demonstrates a slightly lower but highly consistent R-squared of 0.693, closely mirrored by the cross-validation R-squared of 0.692. This consistency indicates a robust model performance across different data subsets, avoiding the pitfalls of overfitting observed in the DRF model. Additionally, the significantly lower MAE (0.033) and RMSE (0.116) reinforce the model's accuracy in predicting values that are close to the actual data points. The lower MSE and RMSE values make this model more suitable for predictive tasks where accuracy is paramount, particularly advantageous in operational settings or predictive analytics where the cost of prediction errors is high. The use of an ensemble of LightGBM and XGBoost, meta-modeled with simple linear regression, contributes to this performance by effectively balancing the bias-variance trade-off. The linear regression component of the meta-model, by averaging the outputs of the base models, not only reduces variance but also curtails the overfitting tendency, thus enhancing the model's generalizability. The alignment of the R-squared values and the low error metrics suggest a balanced model that avoids overfitting and maintains its predictive accuracy across both the training dataset and unseen data.

10 CONCLUSION

The comparative analysis provided in this study demonstrates the superiority of our meta-model in terms of generalization and predictive accuracy, specifically within the context of solar power forecasting. While the model by Pasion et al. showcases high explanatory power within the training dataset, it suffers from poor generalizability and higher prediction errors due to potential overfitting. In contrast, our methodology, which combines LightGBM and XGBoost with a meta-model based on simple linear regression, exhibits stable performance metrics across both training and unseen data, thereby establishing itself as a more reliable and accurate tool for forecasting.

In developing our methodology, we intentionally diverged from the more complex approaches typically found in the literature. By integrating LightGBM and XGBoost, we optimize processing speed and reduce computational overhead, which are crucial for operational efficiency in real-time applications. Unlike other studies that rely on more intricate and computationally intensive methods, our straightforward use of linear regression for the meta-model not only enhances update speeds but also maintains high interpretability. This aspect is particularly vital for gaining stakeholder trust and ensuring regulatory compliance.

Our strategic emphasis on simplicity and efficiency, without compromising accuracy, positions the presented methodology as particularly effective and adaptable to the dynamic, data-intensive demands of modern renewable energy management. This approach not only challenges traditional complex modeling strategies but also underscores the importance of developing robust, efficient, and transparent models in predictive analytics, enhancing both the integration and management of solar power systems in the energy sector.

REFERENCES

Das, U. K., K. S. Tey, M. Seyedmahmoudian, S. Mekhilef, M. Y. I. Idris, and W. Van Deventer. 2018. "Forecasting of Photovoltaic Power Generation and Model Optimization: A Review." *Renewable and Sustainable Energy Reviews* 81:912–928.

Messner, J. W., and P. Pinson. 2019. "Online Adaptive Lasso Estimation in Vector Autoregressive Models for High Dimensional Wind Power Forecasting." *International Journal of Forecasting* 35:1485–1498.

Koster, D., F. Minette, C. Braun, and O. O. Nagy. 2019. "Short-term and Regionalized Photovoltaic Power Forecasting, Enhanced by Reference Systems, on the Example of Luxembourg." *Renewable Energy* 132:455–470.

Antonanzas, J., N. Osorio, R. Escobar, R. Urraca, F. J. Martinez-de-Pison, and F. Antonanzas-Torres. 2016. "Review of Photovoltaic Power Forecasting." *Solar Energy* 136:78–111.

Raza, M. Q., M. Nadarajah, and C. Ekanayake. 2016. "On Recent Advances in PV Output Power Forecast." *Solar Energy* 136:125–144.

Rana, M., and A. Rahman. 2020. "Multiple Steps Ahead Solar Photovoltaic Power Forecasting Based on Univariate Machine Learning Models and Data Re-sampling." *Sustainable Energy, Grids Networks 21:100286.*

Han, Y., N. Wang, M. Ma, H. Zhou, S. Dai, and H. Zhu. 2019. "A PV Power Interval Forecasting Based on Seasonal Model and Nonparametric Estimation Algorithm." *Solar Energy* 184:515–526.

Liu, L., M. Zhan, and Y. Bai. 2019. "A Recursive Ensemble Model for Forecasting the Power Output of Photovoltaic Systems." *Solar Energy* 189:291–298.

Vandeventer, W., E. Jamei, G. Sidarth, M. Seyedmahmoudian, T. Kok, B. Horan *et al*. 2019. "Short-term PV Power Forecasting Using Hybrid GASVM Technique." *Renewable Energy* 140:367–379.

Agoua, X. G., R. Girard, and G. Kariniotakis. 2018. "Short-term Spatio-temporal Forecasting of Photovoltaic Power Production." *IEEE Transactions Sustainable Energy* 9:538–546.

Almeida, M. P., O. Perpiñán, and L. Narvarte. 2015. "PV Power Forecast Using a Nonparametric PV Model." *Solar Energy* 115:354–368.

Alkandari, M., and I. Ahmad. 2020. "Solar Power Generation Forecasting Using Ensemble Approach Based on Deep Learning and Statistical Methods." *Applied Computing and Informatics* 16(4):243-259.

Kaur, A., L. Nonnenmacher, H. T. C. Pedro, and C. F. M. Coimbra. 2016. "Benefits of Solar Forecasting for Energy Imbalance Markets." *Renewable Energy* 86:819–830.

Liu, L., M. Zhan, and Y. Bai. 2019. "A Recursive Ensemble Model for Forecasting the Power Output of Photovoltaic Systems." *Solar Energy* 189:291–298.

Ogliari, E., F. Grimaccia, S. Leva, and M. Mussetta. 2013. "Hybrid Predictive Models for Accurate Forecasting in PV Systems." *Energies* 6:1918–1929.

Rafati, A., M. Joorabian, E. Mashhour, and H. R. Shaker. 2020. "High Dimensional Very Short-term Solar Power Forecasting Based on a Data-driven Heuristic Method." *Energy* 202:117-129.

Pedro, H. T. C. 2012. "Assessment of Forecasting Techniques for Solar Power Output with No Exogenous Inputs." *Solar Energy* 86(7):1767-1779.

Oneto, L., F. Laureri, M. Robba, F. Delfino, and D. Anguita. 2018. "Data-Driven Photovoltaic Power Production Nowcasting and Forecasting for Polygeneration Microgrids." *IEEE Systems Journal* 12(3):2842–2853.

Pasion, C., T. Wagner, C. Koschnick, S. Schuldt, J. Williams, and K. Hallinan. "Machine Learning Modeling of Horizontal Photovoltaics Using Weather and Location Data." *Energies* 13(10):2570.

Omar, M., A. Dolara, G. Magistrati, M. Mussetta, E. Ogliari, and F. Viola. 2016. "Day-Ahead Forecasting for Photovoltaic Power Using Artificial Neural Networks Ensembles." *2016 IEEE International Conference on Renewable Energy Research and Applications (ICRERA),* Birmingham, UK, 1152-1157.

Li, J., R. Wang, T. Zhang, X. Zhang, and T. Liao. 2016. "Predicating Photovoltaic Power Generation Using an Improved Hybrid Heuristic Method." *In 2016 Sixth International Conference on Information Science and Technology (ICIST),* Dalian, China, 383-387.

Alessandrini, S., L. Delle Monache, S. Sperati, and G. Cervone. 2015. "An Analog Ensemble for Short-Term Probabilistic Solar Power Forecast." *Applied Energy 157:95-110*.

Fonseca Jr., J. G. da S., T. Oozeki, T. Takashima, G. Koshimizu, Y. Uchida, and K. Ogimoto. 2012. "Use of Support Vector Regression and Numerically Predicted Cloudiness to Forecast Power Output of a Photovoltaic Power Plant in Kitakyushu, Japan." *Progress in Photovoltaics: Research and Applications* 20(6):575-584.

Deo, R. C., X. Wen, and F. Qi. 2016. "A Wavelet-Coupled Support Vector Machine Model for Forecasting Global Incident Solar Radiation Using Limited Meteorological Dataset." *Applied Energy* 184:709-726.

Yang, H.-T., C.-M. Huang, Y.-C. Huang, and Y.-S. Pai. 2014. "A Weather-Based Hybrid Method for 1-Day Ahead Hourly Forecasting of PV Power Output." *IEEE Transactions on Sustainable Energy* 5(3):917-926.

Zerpa, L. E., N. V. Queipo, S. Pintos, and J.-L. Salager. 2005."An Optimization Methodology of Alkaline–Surfactant–Polymer Flooding Processes Using Field Scale Numerical Simulation and Multiple Surrogates." *Journal of Petroleum Science and Engineering* 47(3-4): 197-208.

Acar, E., and M. Rais-Rohani. 2009. "Ensemble of Metamodels with Optimized Weight Factors." *Structural and Multidisciplinary Optimization* 37:279–294.

Qing, X. and Y. Niu. 2018. "Hourly Day-Ahead Solar Irradiance Prediction Using Weather Forecasts by LSTM." *Energy* 148:461–468.

Yaniktepe, B. and Y. A. Genc. 2015. "Establishing New Model for Predicting the Global Solar Radiation on Horizontal Surface." *International Journal of Hydrogen Energy* 40:15278–15283.

Letendre, S., M. Makhyoun, and M. Taylor. 2014. "Predicting Solar Power Production: Irradiance Forecasting Models, Applications and Future Prospects." *Solar Electric Power Association*, Washington, DC, USA. Available online: (accessed on 16 April 2020).

Lave, M. and J. Kleissl. 2011. "Optimum Fixed Orientations and Benefits of Tracking for Capturing Solar Radiation in the Continental United States." *Renewable Energy* 36:1145–1152.

Faine, P., S. R. Kurtz, C. Riordan, and J. M. Olson. 1991. " The Influence of Spectral Solar Irradiance Variations on the Performance of Selected Single-Junction and Multijunction Solar Cells." *Solar Cells* 31:259–278.

Mekhilef, S., R. Saidur, and M. Kamalisarvestani. 2012. "Effect of Dust, Humidity and Air Velocity on Efficiency of Photovoltaic Cells." *Renewable and Sustainable Energy Reviews* 16:2920–2925.

Kayri, M., I. Kayri, and M. T. Gencoglu. 2017. "The Performance Comparison of Multiple Linear Regression, Random Forest and Artificial Neural Network by Using Photovoltaic and Atmospheric Data." *In Proceedings of the 2017 14th International Conference on Engineering of Modern Electric Systems (EMES 2017)*, Oradea, Romania, 1–2 June 2017, 1-4. IEEE: Piscataway, NJ, USA.

Lu, H. and W. Zhao. 2018. "Effects of Particle Sizes and Tilt Angles on Dust Deposition Characteristics of a Ground-Mounted Solar Photovoltaic System." *Applied Energy* 220:514–526.

UCAR Center for Science Education. 2020. "The Highs and Lows of Air Pressure." Available online: https://scied.ucar.edu/shortcontent/highs-and-lows-air-pressure (accessed on 16 April 2020).

AUTHOR BIOGRAPHIES

MAHDI DARVISHI is a Master’s student at the Faculty of Engineering and Architectural Science of Toronto Metropolitan University, Toronto, ON, CANADA. His research interests include Environmental modeling, visualization, and optimizations, Renewable resources, and related processes his email address is mahdi.darvishi@torontomu.ca.

ABDOLREZA ABHARI is a Professor and Undergraduate Program Director in the Department of Computer Science of Toronto Metropolitan University, Toronto, ON, CANADA. His research interests include Data Science and Simulation, Data analytics for simulation, Machine learning methods to generate synthetic data, Simulating massive data processing systems, Machine learning and data mining in modeling and simulation. His email address is aabhari@torontomu.ca and his website is https://cs.torontomu.ca/~aabhari/.