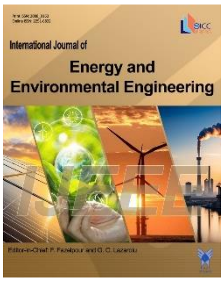

Volume 16, Issue 2, (2025) Article 6

International Journal
of
Energy and Environmental Engineering (IJEEE)

https://doi.org/10.57647/ijeee.2025.1602.06

Feature Selection and Seasonal Variability in Forecasting Global Horizontal Irradiance: Insights from Riyadh Meteorology

Debasish Sarker[1]*, Mohammad K Hossain[2], Md Arifuzzaman[3]

[1]*Department of Mechanical Engineering, International University of Business Agriculture and Technolog, Dhaka, Bangladesh*
[2]*King Fahd University of Petroleum and Minerals College of Science, Dhahran, Eastern Province, Saudi Arabia*
[3]*Civil and Environmental Engineering, King Faisal Universit, Al Ahsa, Eastern Province, Saudi Arabia*

***Corresponding author:** dsarker.me@iubat.edu

Original Research

Received:
16 December 2024

Accepted:
25 January 2025

Published in Issue:
30 June 2025

Abstract

This study investigates the seasonal variability of meteorological feature importance in forecasting Global Horizontal Irradiance (GHI) using machine learning (ML) and deep learning models. High-resolution solar and meteorological data from NREL's NSRDB (24.25°N, 45.34°E, 740 m) were seasonally partitioned into winter, spring, summer, and autumn. Feature selection was conducted using Pearson correlation (threshold 0.25), followed by dimensionality reduction through Principal Component Analysis (PCA). Six ML models XGBoost, LightGBM, Random Forest, SVR, MLP, and LSTM were trained on the processed datasets, and SHAP analysis was used to interpret feature contributions. The results revealed that clear-sky irradiance parameters (GHI, DNI, DHI) consistently dominated GHI prediction (correlation >0.95; SHAP $>10^{-1}$), while features like temperature and relative humidity varied across seasons. Wind direction, though weakly correlated, showed increased influence in winter. PCA enhanced model stability in spring and winter but slightly reduced accuracy during periods of high irradiance variability. Overall, XGBoost and Random Forest models provided the most accurate and reliable forecasts across seasons.

Keywords: Renewable energy; Solar Forecasting; Machine learning; Seasonal effect; Feature selection

Cite this article: Sarker, D., Hossain, M. K., Arifuzzaman, Md., Feature Selection and Seasonal Variability in Forecasting Global Horizontal Irradiance: Insights from Riyadh Meteorology. Int. J. Energy Environ. Eng. 2025; 16(2) : Article 6. https://doi.org/10.57647/ijeee.2025.1602.06

1. Introduction

As the global energy landscape shifts toward renewable sources, the ability to accurately forecast solar power generation has become increasingly vital [1-3]. One of the most important parameters in this context is Global Horizontal Irradiance (GHI), which measures the total solar radiation received on a horizontal surface. Accurate GHI forecasting is essential for ensuring the stability and efficiency of modern power systems, particularly those with a high penetration of solar energy [4, 5].

Without reliable forecasts, grid operators face challenges in balancing supply and demand, which can lead to inefficiencies or even instability in the power grid [5, 6]. In this regard, rooftop photovoltaic (PV) systems, as illustrated in Figure 1, have become a common feature of urban energy infrastructure.

These decentralized installations contribute significantly to overall solar power generation, increasing the variability and complexity of energy supply. Consequently, improving the accuracy of solar forecasting especially of GHI is critical to effectively integrating such systems into the grid.

To address this challenge, researchers have increasingly turned to machine learning (ML) techniques, which have shown great promise in modeling the complex, nonlinear patterns inherent in solar irradiance data [1, 7-9]. A wide variety of ML models have been applied to this task, including Long Short-Term Memory (LSTM) networks, Support Vector Regression (SVR), Extreme Gradient Boosting (XGBoost), Multilayer Perceptrons (MLP), and Random Forests [10-14].

These models are typically trained using a combination of meteorological inputs such as temperature, humidity, atmospheric pressure, precipitable water, and historical GHI values.

These features help the models learn how environmental conditions influence solar radiation over time.

While these approaches have demonstrated strong predictive performance, most of the existing literature has focused on improving model accuracy through architectural enhancements, hyperparameter tuning, or the development of new input features.

However, one area that remains relatively underexplored is the impact of seasonal variability on the importance of these input features and the overall performance of forecasting models.

Figure 1. Solar Rooftop PV Panel

Seasonal changes, such as shifts in cloud cover, humidity, and solar angle can significantly alter the atmospheric conditions that affect solar irradiance [15]. For example, during the summer months, high humidity and clear skies might make temperature and water vapor more influential predictors, whereas in winter, cloud cover and atmospheric pressure might play a larger role. Despite this, many forecasting models treat feature importance as static, assuming that the same variables are equally relevant throughout the year. Recent studies have begun to highlight the importance of considering seasonal dynamics in solar forecasting [16]. For instance, research has shown that the performance of models like SVR and LSTM can vary significantly across seasons, with different features contributing more or less to prediction accuracy depending on the time of year [10, 11]. This suggests that a one-size-fits-all approach may not be optimal and that season-specific models or adaptive feature selection strategies could lead to more accurate and reliable forecasts. While machine learning has significantly advanced the field of solar irradiance forecasting, there is a clear need for more research into how seasonal variability affects feature relevance. Understanding these dynamics could inform the development of more robust, context-aware forecasting systems that adapt to changing environmental conditions, ultimately improving the integration of solar energy into modern power grids [17].

The central motivation of this research stems from a critical gap in the current understanding of solar irradiance forecasting: the role of seasonal variability in shaping the relationship between meteorological features and GHI. Although machine learning models have been widely adopted for predicting GHI due to their ability to capture complex, nonlinear patterns, most existing studies treat the input features and model configurations as static across time. This approach overlooks the dynamic nature of atmospheric conditions, which can vary significantly from one season to another. As a result, the influence of individual meteorological variables, such as temperature, humidity, pressure, and precipitable water, on GHI may also fluctuate throughout the year. This study aims to address this oversight by conducting a comprehensive analysis of how the correlations between key meteorological features and GHI evolve across different seasons.

The study explores the hypothesis that no single model may be optimal across all seasons, and that tailoring model selection to seasonal conditions could yield significant improvements in forecasting accuracy. This perspective challenges the conventional "one-size-fits-all" approach commonly seen in prior research, where a single model is trained on year-round data and applied uniformly across all time periods [18]. To explore this hypothesis, this research employs a structured methodology that links three key components: feature relevance, seasonal variability, and model performance. First, it quantifies the correlation between meteorological features and GHI for each season. Next, it evaluates how these correlations influence the performance of different ML models when trained and tested on season-specific data. Finally, it assesses whether adapting model selection based on seasonal feature dynamics leads to measurable gains in forecasting accuracy. By establishing these connections, the study aims to lay the groundwork for the development of seasonally adaptive solar forecasting frameworks. Such frameworks could dynamically adjust both feature selection and model architecture based on the time of year, thereby enhancing the reliability and responsiveness of solar energy systems [19]. This is particularly important as power grids become more reliant on renewable sources, where variability and uncertainty are inherent challenges. Ultimately, this research contributes to the broader goal of improving the integration of solar energy into modern power systems by making forecasting tools more context-aware and performance-optimized throughout the year.

2. Methodology

For this study, the solar irradiation and meteorological data were collected from the National Renewable Energy Laboratory (NREL), specifically from the National Solar Radiation Database (NSRDB). The data files provided by NREL contain detailed metadata, including information on the source, location, geographical coordinates, time zone, elevation, and measurement units for each parameter. The specific geographic coordinates of the data collection point are latitude 24.25°N and longitude 45.34°E, at an elevation of 740 meters. The dataset encompasses multiple atmospheric and solar parameters essential for solar irradiance forecasting. The features available in the dataset include both clear-sky and observed values for Global Horizontal Irradiance (GHI), Direct Normal Irradiance (DNI), and Diffuse Horizontal Irradiance (DHI). In addition to these, several meteorological parameters were also provided: air temperature, dew point temperature, atmospheric pressure, relative humidity, precipitable water, surface albedo, solar zenith angle, wind direction, and wind speed. Each entry in the dataset is timestamped with year, month, day, hour, and minute, providing a high-resolution temporal structure suitable for time series analysis. To examine the seasonal impacts on feature relevance and model performance, the data were systematically arranged according to meteorological seasons. The seasonal divisions adopted for this study were as follows: Winter comprises December, January, and February; Spring includes March, April, and May; Summer consists of June, July, and August; and Autumn covers September, October, and November. Data from corresponding months across multiple years were aggregated under these seasonal categories to ensure robust statistical representation of seasonal variability. The research workflow was divided into two major stages. First, a feature selection process was employed to identify the most influential variables for GHI prediction, with particular emphasis on how these influences shift between seasons. Following feature selection, time series forecasting experiments were carried out using a suite of widely-used machine learning models, including LSTM, XGBoost, LightGBM, SVR, MLP, and RandomForest. These models were systematically trained and validated on seasonally partitioned datasets to assess how seasonal variations and feature relevance influence predictive performance. This methodological framework allowed for a detailed analysis of whether different seasons demand different model choices or feature sets for achieving optimal forecasting results in solar energy applications.

2.1. Model: feature selection

The methodology developed for this study followed a structured and systematic sequence of steps, aligning with the overall research objectives. The corresponding process flow is illustrated in Figure 2, and the implementation was carried out using Python-based machine learning libraries. The meteorological datasets required for this analysis were collected from the National Renewable Energy Laboratory (NREL), specifically from the National Solar Radiation Database (NSRDB). These datasets include hourly records of meteorological and solar irradiance variables. Multiple CSV and Excel files corresponding to various time periods were consolidated into a single structured dataset. During the data import, irrelevant time-related columns such as Year, Month, Day, Hour, and Minute were removed to focus on feature variables relevant for GHI prediction. To ensure data consistency, records with missing values in the target variable (Global Horizontal Irradiance, GHI) were excluded from further analysis.

Data Preprocessing

Prior to feature selection and model training, several preprocessing steps were applied to prepare the dataset for machine learning models. These steps included:

- Handling Missing Data in Features: Any missing values in the feature columns were imputed using the mean value strategy to maintain dataset integrity.

10.57647/ijeee.2025.1602.06

- Normalization: The feature values were standardized to ensure uniform scaling across variables, which is essential for models sensitive to feature magnitudes, such as Support Vector Regression (SVR) and neural networks.

These preprocessing steps were crucial to facilitate effective feature correlation analysis and to prepare the data for model training.

Feature Selection Based on Correlation Analysis

An important component of the methodology was determining the relevance of each meteorological feature to the target variable, GHI. This was achieved by computing the Pearson correlation coefficient between each feature and GHI. To filter out weakly correlated variables, a threshold of 0.25 was applied. Features exhibiting a correlation greater than this threshold were considered significant for subsequent modeling efforts. This approach helped reduce dimensionality and ensured that the forecasting models were focused on variables with meaningful predictive relationships with GHI.

Model Development and Evaluation

Multiple machine learning models were selected for forecasting GHI, each representing different algorithmic families: Tree-Based Models: XGBoost, LightGBM, Random Forest; Kernel-Based Model: Support Vector Regression (SVR); Neural Network Model: Multi-Layer Perceptron (MLP).

For models requiring scaled inputs or handling of missing values internally (e.g., SVR, MLP), pipelines were constructed integrating preprocessing steps with model fitting. Each model was trained on the prepared dataset, and predictive performance was evaluated using standard error metrics including: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Coefficient of Determination (R^2). These metrics provided insight into the accuracy and robustness of each model across different seasonal datasets.

Feature Contribution Analysis Using SHAP

To better interpret the contribution of individual features toward the model predictions, SHapley Additive exPlanations (SHAP) were employed. SHAP provides a consistent and theoretically sound framework for feature importance analysis, offering interpretable explanations for model outputs.

For ensemble models (XGBoost, LightGBM, Random Forest), SHAP explanations were generated using tree-based explainers, whereas kernel-based SHAP methods were used for SVR and MLP. The feature contributions were visualized through SHAP summary plots, illustrating the impact of each feature on the predicted GHI values. All intermediate and final outputs including feature importances, correlation plots, SHAP visualizations, and performance metrics were saved systematically for analysis and reporting.

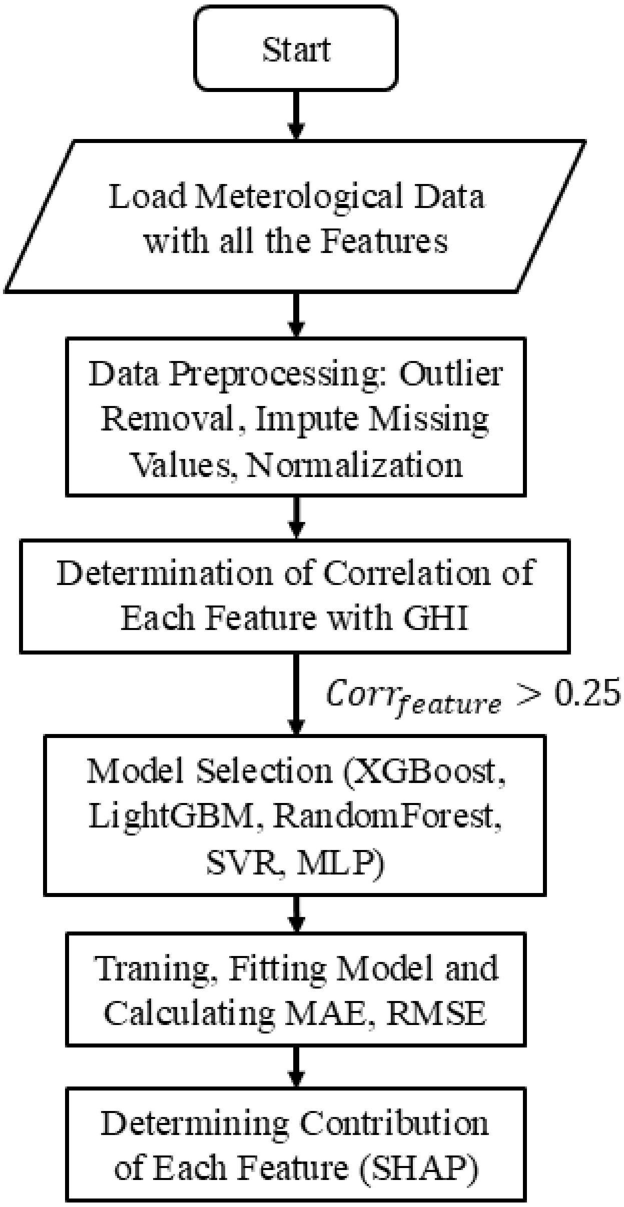

Figure 2. Feature Selection Model

2.2. Model: time series forecasting

The methodological framework employed in this study for time series forecasting is outlined in Figure 3. The process is structured into distinct stages, beginning from data acquisition to model evaluation, ensuring a systematic approach to predictive modeling.

Data Acquisition and Preparation

The process begins with the acquisition of historical meteorological datasets stored in CSV format, encompassing a wide range of environmental variables pertinent to solar irradiance forecasting. These variables include, but are not limited to, global horizontal irradiance (GHI), ambient temperature, relative humidity, precipitable water, surface albedo, and pressure. The datasets were imported and transformed into numerical arrays using the Python programming language, specifically leveraging the capabilities of the

Pandas library. This preparation was essential to enable seamless integration with the machine learning and deep learning algorithms adopted in subsequent stages. Two parallel strategies were employed for feature engineering. In one pathway, the complete set of features was retained for modeling to capture potentially subtle interactions between variables. In the other, feature selection was guided by the strength of correlation with the target variable, GHI. This selection process utilized Pearson's correlation coefficient to filter out weakly correlated features, thus minimizing redundancy and reducing the potential for overfitting.

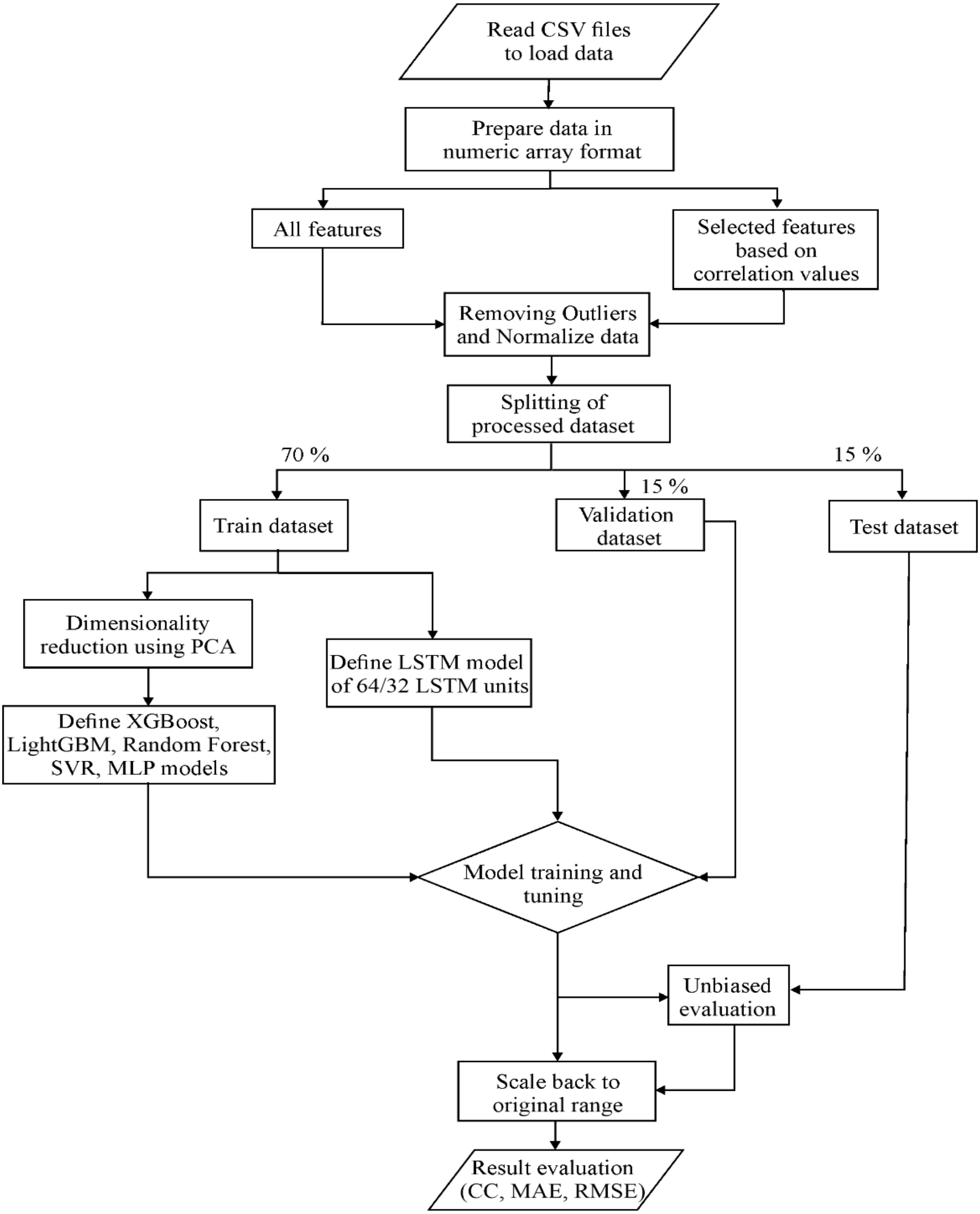

Figure 3. Time series forecasting model

Following feature selection, data cleaning was performed by removing outliers based on the interquartile range (IQR) method, thereby ensuring the dataset was free from extreme values that could bias model training. Subsequently, normalization was applied to scale the data, ensuring that each feature exhibited a mean of zero and a standard deviation of one. This preprocessing step is widely recognized for improving model convergence, particularly for algorithms sensitive to feature magnitudes. To facilitate model development and evaluation, the processed dataset was partitioned into three disjoint subsets. 70% of the data was allocated for training, 15% for validation, and 15% for testing. The training dataset was used for fitting model parameters, the validation dataset guided hyperparameter optimization and model selection, and the test dataset provided an unbiased final evaluation of model generalizability. This splitting approach is consistent with standard practices in machine learning for time series prediction.

Model Development

The modeling framework incorporated both traditional machine learning algorithms and advanced deep learning architectures. Prior to training the classical machine learning models, dimensionality reduction was conducted using Principal Component Analysis (PCA) to mitigate multicollinearity and reduce computational complexity. The machine learning models utilized included XGBoost, LightGBM, Random Forest, Support Vector Regression, and Multi-Layer Perceptron (MLP) regressors.

These models were chosen for their established capability to handle nonlinear regression tasks and their demonstrated success in forecasting applications across the environmental sciences.

Parallel to the machine learning models, a recurrent neural network architecture was developed using Long Short-Term Memory (LSTM) networks. LSTM models were configured with either 32 or 64 recurrent units per layer depending on the experimental setup. These models are particularly effective for sequential data as they are capable of capturing both short-term and long-term dependencies in temporal patterns, making them ideally suited for irradiance forecasting tasks.

Model Training and Optimization

Both classical and deep learning models underwent rigorous training using the training dataset. Hyperparameter tuning was guided by performance metrics evaluated on the validation dataset. For classical machine learning models, grid search algorithms were employed to optimize key hyperparameters. In the case of LSTM models, training was conducted using the Adam optimizer with mean squared error as the loss function to ensure smooth convergence.

Upon completing model training and validation, the predictions were transformed back to their original scales to ensure interpretability in terms of physical irradiance values. Model performance was assessed using three widely accepted statistical metrics: the correlation coefficient (CC), mean absolute error (MAE), and root mean squared error (RMSE). These metrics provided comprehensive insight into both the accuracy and reliability of the forecasting models. The use of multiple metrics ensured that model evaluation was not biased by the limitations of a single criterion.

The methodological framework presented here represents a comprehensive and hybridized approach to time series forecasting of solar irradiance. By combining feature selection techniques, dimensionality reduction, diverse machine learning algorithms, and powerful deep learning models, the proposed methodology achieves a balance between predictive accuracy and computational efficiency. This multi-layered approach offers flexibility for adaptation to other environmental forecasting challenges beyond solar energy applications.

3. Results and discussion

3.1. Feature selection

Figure 4 shows the correlation of various meteorological parameters with Global Horizontal Irradiance (GHI) across four seasons: (a) Autumn, (b) Spring, (c) Summer, and (d) Winter. Pearson's correlation coefficient was used, with correlation values ranging from -1 to +1 on the vertical axis. The horizontal axis lists the parameters: Clearsky GHI, Clearsky DNI, DNI, Clearsky DHI, DHI, Temperature, Wind Speed, Wind Direction, Dew Point, Precipitable Water, Surface Albedo, Pressure, Relative Humidity, Solar Zenith Angle, and cloud-related variables.

In all seasons, Clearsky irradiance parameters (GHI, DNI, DHI) exhibit strong positive correlations ($p < 0.001$) with GHI, confirming their dominant influence on surface irradiance. Temperature also shows positive correlations, particularly in summer, due to enhanced solar insolation and atmospheric clarity. By contrast, Relative Humidity consistently displays moderate to strong negative correlations ($p < 0.001$), likely due to the attenuation of solar radiation by increased moisture content in the atmosphere. Similarly, Solar Zenith Angle shows negative correlations, reflecting the reduced effective irradiance when the sun is positioned lower in the sky. While most other parameters exhibit weak correlations, Wind Direction shows increasing significance from autumn to winter ($p < 0.05$), likely related to seasonal shifts in prevailing winds that influence cloud formation and atmospheric clarity.

10.57647/ijeee.2025.1602.06

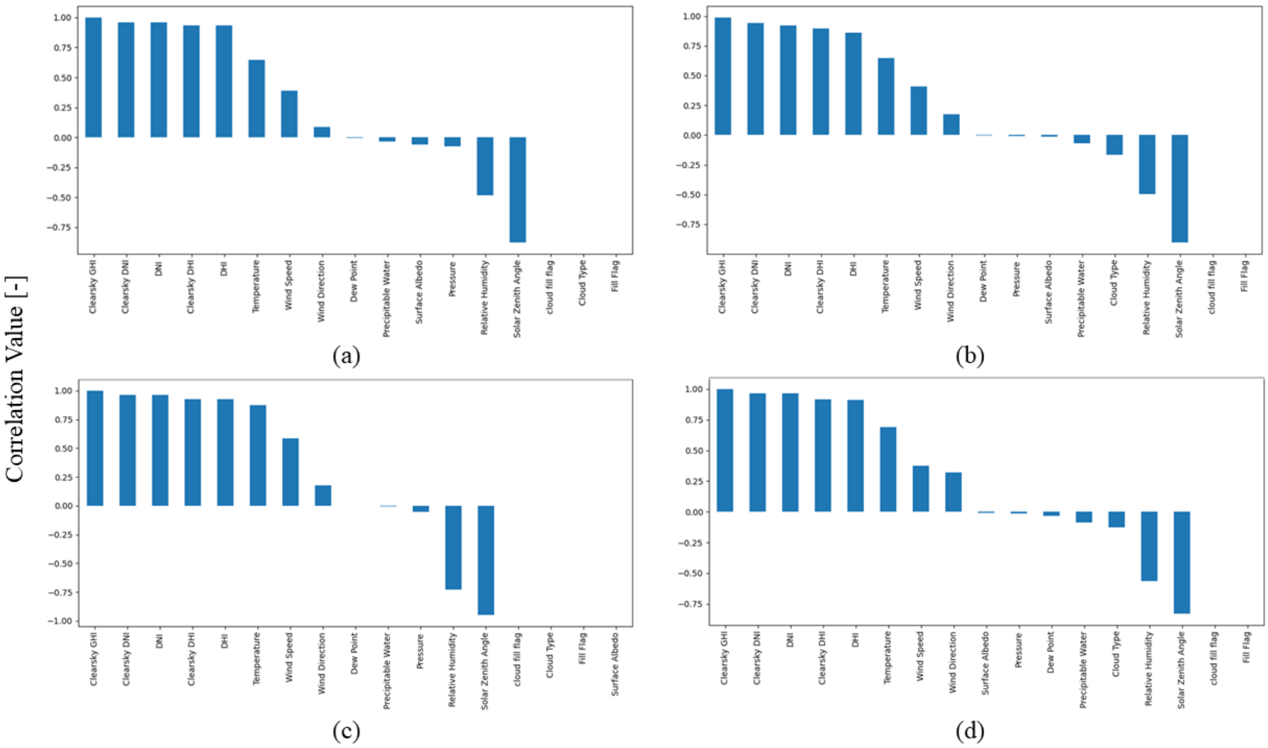

Figure 4. Correlation of meteorological parameters with GHI across seasons: (a) Autumn, (b) Spring, (c) Summer, and (d) Winter

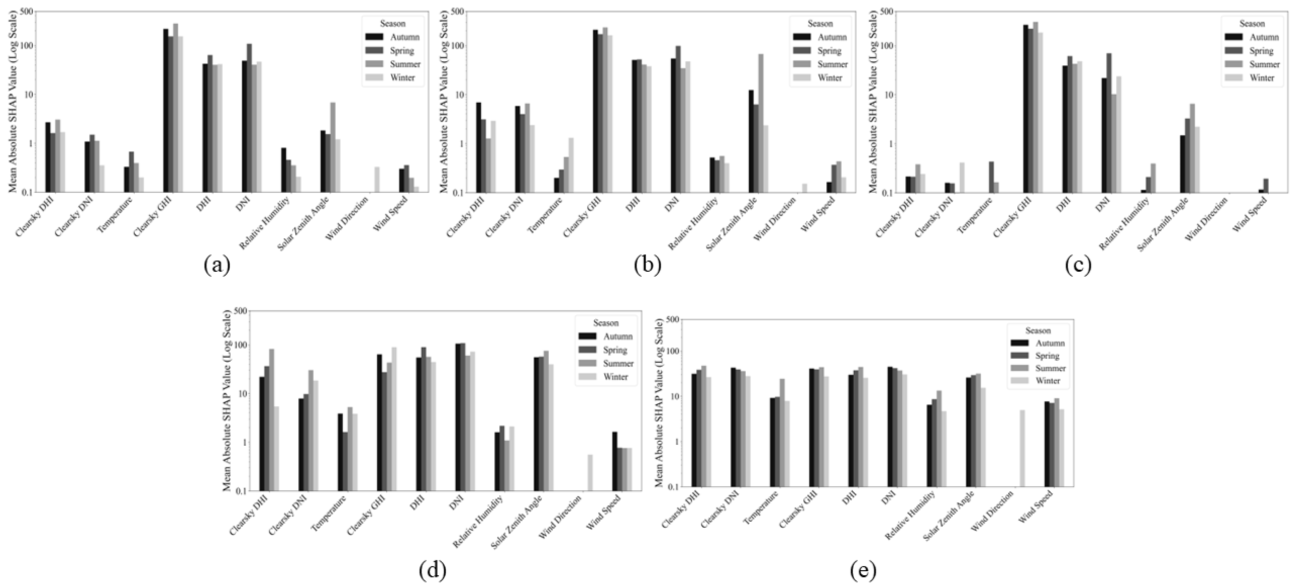

Figure 5. SHAP Analysis of Feature Importance for (a) XGBoost, (b) LightGBM, (c) Random Forest, (d) MLP, and (e) SVR

This correlation analysis highlights the need to prioritize highly correlated irradiance-based features in solar forecasting models. Furthermore, season-specific features, such as Wind Direction during winter, may provide additional predictive value, enhancing model robustness across varying atmospheric conditions. The use of statistical significance testing supports the inclusion of these predictors in data-driven approaches to solar energy forecasting. Figure 5 presents the SHAP (SHapley Additive exPlanations) analysis of feature importance for five machine learning models: (a) XGBoost, (b) LightGBM, (c) Random Forest, (d) MLP, and (e) SVR. The y-axis displays the mean absolute SHAP value on a logarithmic scale, representing the average contribution of each feature to the model's predictions. The features evaluated are Clearsky DHI, Clearsky DNI, Temperature, Clearsky GHI, DHI, DNI, Relative Humidity, Solar Zenith Angle, Wind Direction, and Wind Speed. The bars are color-coded by season, with black for Autumn, dark gray for Spring, gray for Summer, and light gray for Winter. In all models, the Clearsky irradiance parameters (DHI, DNI, and GHI)

10.57647/ijeee.2025.1602.06

consistently exhibit the highest SHAP values across seasons, indicating their dominant influence on GHI predictions. Among these, Clearsky GHI and DNI generally contribute most strongly, confirming their predictive significance irrespective of the algorithm used.

Temperature also emerges as an important feature, particularly for the tree-based models (XGBoost, LightGBM, Random Forest), with notable contributions across all seasons. Relative Humidity and Solar Zenith Angle show moderate contributions, aligning with their known inverse correlation with solar radiation, as previously observed in the correlation analysis. Seasonal variability in feature importance is clearly visible, particularly for features such as Relative Humidity and Solar Zenith Angle, whose SHAP values tend to fluctuate more across seasons compared to irradiance parameters. Additionally, Wind Direction and Wind Speed generally exhibit lower SHAP values, suggesting limited predictive influence in the models examined, although small seasonal variations are still noticeable.

Comparison of Correlation and SHAP Results

The correlation analysis (Figure 4) and SHAP feature importance (Figure 5) show strong agreement regarding the dominant predictors of GHI. Clearsky GHI, Clearsky DNI, and DNI exhibit the highest correlation values across all seasons, consistently exceeding 0.95, confirming their primary influence on GHI. Correspondingly, SHAP analysis reveals that these features also yield the largest contributions to model predictions, with mean absolute SHAP values exceeding 10^{-1} on the logarithmic scale for all five machine learning models. Among these, Clearsky GHI and Clearsky DNI consistently rank as the most influential across models and seasons.

Temperature shows positive correlations with GHI (0.40–0.65) depending on the season, with SHAP confirming its predictive relevance, particularly in the tree-based models (XGBoost, LightGBM, Random Forest). Negative correlations for Relative Humidity and Solar Zenith Angle (typically between -0.5 and -0.8) are reflected in their moderate SHAP values, with noticeable seasonal fluctuations for both parameters. Interestingly, while Wind Direction and Wind Speed exhibit weak correlations (|< 0.2), SHAP identifies small but discernible contributions, particularly for Wind Direction in winter, where the SHAP value increases by approximately one order of magnitude compared to other seasons in some models. This suggests that machine learning models can capture weak but nonlinear interactions involving these variables, which are not evident in correlation analysis alone. Overall, while correlation highlights features with strong linear relationships, SHAP provides a model-aware perspective that accounts for nonlinear effects and feature interactions. The strong alignment for the primary irradiance features validates the physical relevance of the models, while the SHAP-specific findings for secondary features (e.g., Wind Direction) highlight opportunities for seasonal refinement in predictive modeling.

3.2. Evaluation of feature selection on model performance

To comprehensively assess the influence of feature selection on the performance of machine learning models in forecasting Global Horizontal Irradiance (GHI), a series of comparative experiments were conducted across different feature selection strategies. The models under investigation included XGBoost, LightGBM, SVR, MLP, RandomForest, and LSTM, evaluated over four seasons Autumn, Spring, Summer, and Winter using training, validation, and testing datasets. Figures 6 and 7 summarize the prediction performance using two commonly used error metrics: Root Mean Square Error (RMSE) and Normalized Mean Absolute Error (nMAE).

The seasonal variation of RMSE across different feature selection methods is presented in Figure 6. Three distinct feature selection strategies were applied: (a) All Features: All available meteorological and derived variables were used as predictors. (b) Limited Features: Important features were selected based on combined analysis using SHAP (SHapley Additive exPlanations) values and correlation analysis. (c) PCA Features: The original feature space was transformed using Principal Component Analysis (PCA), retaining the principal components that explained the majority of variance.

Each subplot displays the seasonal RMSE values for all six models, providing insight into both seasonal prediction difficulty and the effect of feature selection. It was observed that: All features (a): Generally produced reasonable performance, yet possibly suffered from multicollinearity and inclusion of irrelevant features.

Limited features (b): Showed notable improvements in predictive accuracy for most seasons, emphasizing the importance of informative feature selection. PCA features (c): Delivered mixed results. While dimensionality reduction improved performance in certain seasons (e.g., Autumn), it sometimes degraded predictive skill due to the potential loss of physical interpretability of the features. Seasonally, RMSE tended to be higher in Summer, potentially due to increased cloud cover variability, and lower in Autumn and Winter, where solar irradiance is generally more stable.

10.57647/ijeee.2025.1602.06

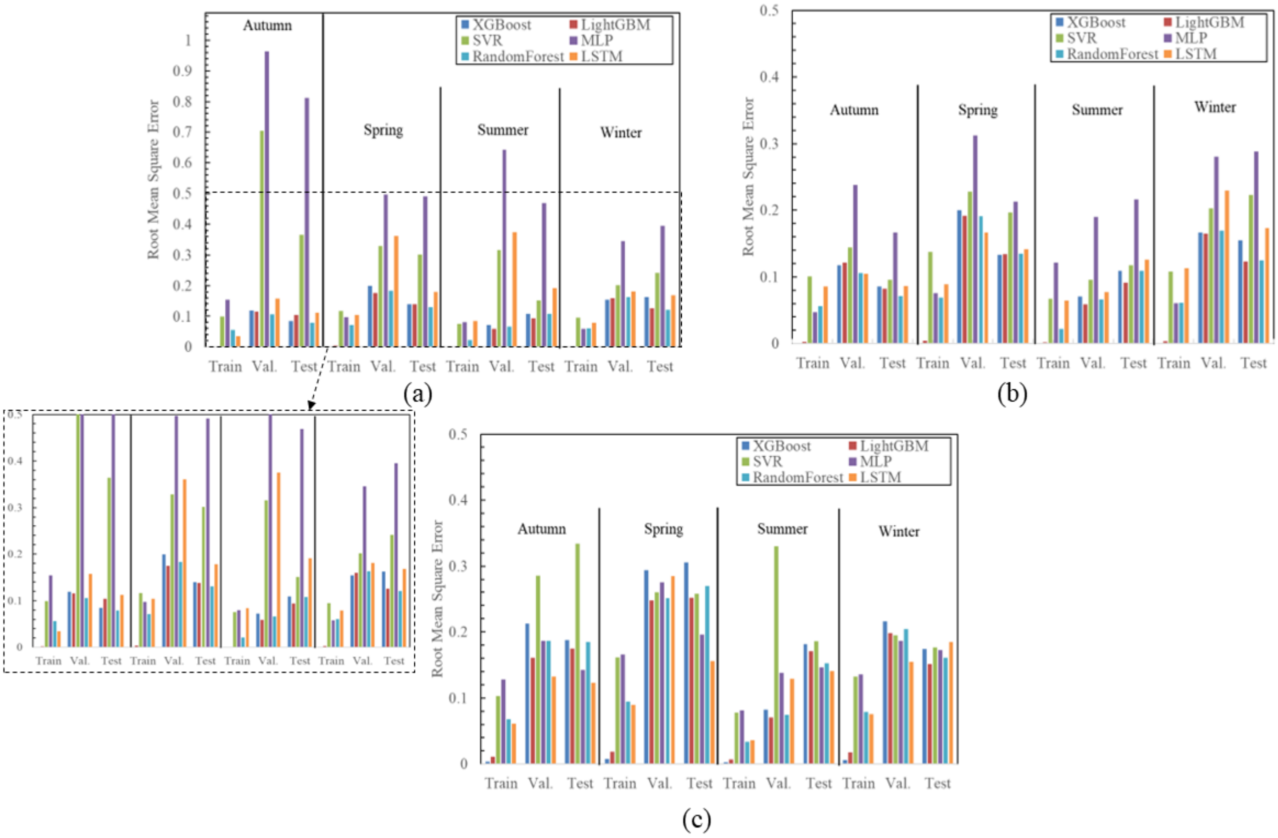

Figure 6. Seasonal variation of RMSE for different machine learning models applied to GHI forecasting under three feature selection approaches: (a) All features, (b) Limited features selected via correlation analysis, and (c) PCA features

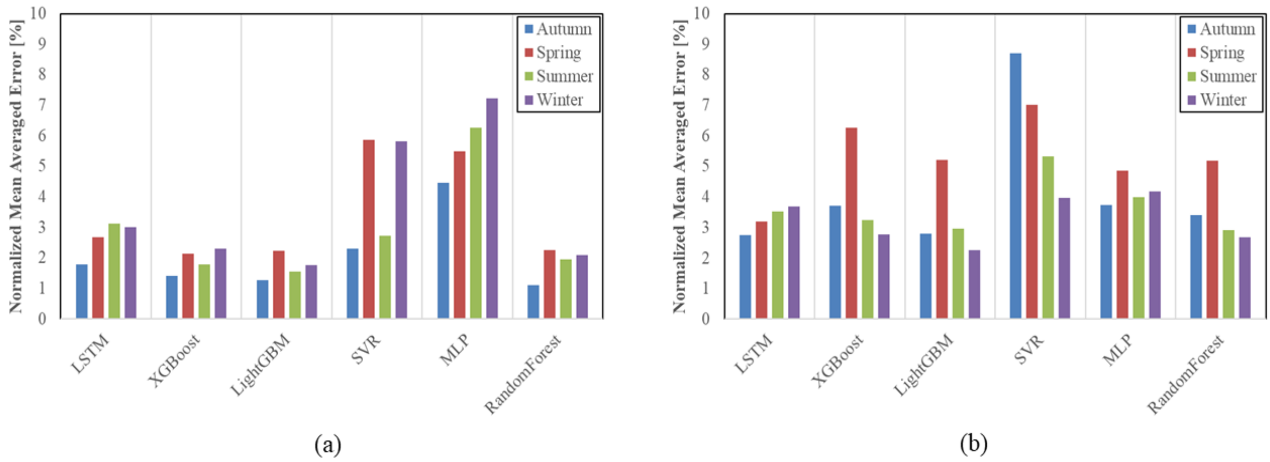

Figure 7. Seasonal variation of nMAE for different machine learning models applied to GHI forecasting under three feature selection approaches: (a) Limited features selected via correlation analysis, and (b) PCA features

Complementing the RMSE analysis, Figure 7 depicts the normalized Mean Absolute Error (nMAE) under the same conditions.

By normalizing MAE against the range of actual GHI values, direct comparisons of prediction errors across seasons with differing irradiance magnitudes were facilitated. Consistent with the RMSE results, the application of feature selection (a) led to reductions in normalized errors for most models, affirming that excluding irrelevant or redundant features can significantly enhance model generalization. The PCA-based approach again displayed variable results depending on the season and dataset used. The comparative analysis across RMSE and nMAE suggests that: Gradient boosting models (XGBoost, LightGBM) generally outperformed other models in terms of both RMSE and nMAE. LSTM models performed competitively in datasets exhibiting strong temporal dependencies, highlighting their suitability for time series problems. Support Vector Regression (SVR) and RandomForest exhibited reasonable but slightly inferior performance relative to boosting algorithms.

3.3. Time series of seasonal GHI predictions: limited vs. PCA features

Figure 8 presents the time series comparison between observed (Actual GHI) and predicted GHI values for six machine learning models LSTM, XGBoost, LightGBM, SVR, MLP, and Random Forest across the four seasons: (a) Autumn, (b) Spring, (c) Summer, and (d) Winter. The x-axis represents the number of time steps, while the y-axis indicates irradiance in W/m². These predictions were generated using a limited set of features selected through prior correlation and SHAP analyses.

Across all seasons, the models effectively capture the overall diurnal pattern of GHI, characterized by a symmetric rise and fall of irradiance during daylight hours. However, seasonal differences in predictive performance are noticeable. In Autumn (Figure 8a) and Spring (Figure 8b), most models track the actual GHI closely, with XGBoost, LightGBM, and Random Forest showing excellent agreement, particularly around peak irradiance values. LSTM tends to slightly underestimate GHI near midday, while MLP occasionally overestimates.

In Summer (Figure 8c), deviations between predicted and actual GHI become more pronounced, particularly during periods of peak irradiance, where LSTM and MLP tend to overestimate, likely due to higher atmospheric variability associated with convective activity. In contrast, tree-based models maintain relatively stable predictions, although slight underestimations are observed near the peak. Winter (Figure 8d) exhibits improved overall agreement across all models, with lower irradiance magnitudes contributing to reduced variance in prediction errors. Notably, XGBoost and Random Forest consistently provide close fits to the observed GHI across all seasons. These results demonstrate that while all models effectively learn the general irradiance pattern, XGBoost and Random Forest outperform others in terms of closely matching actual GHI values, especially during transitional seasons (autumn and spring). LSTM and MLP exhibit slightly higher deviations during high irradiance periods, indicating sensitivity to intra-day variability or possible limitations in capturing atmospheric transients with the available input features. The consistent predictive strength of tree-based models aligns with their superior feature importance rankings in the SHAP analysis and supports their suitability for robust, seasonally adaptive GHI forecasting.

Figure 9 illustrates the time series comparison between observed (Actual GHI) and predicted GHI for six machine learning models LSTM, XGBoost, LightGBM, SVR, MLP, and Random Forest using PCA-transformed features across the four seasons: (a) Autumn, (b) Spring, (c) Summer, and (d) Winter. The x-axis denotes the number of time steps, while the y-axis represents irradiance in W/m². Principal Component Analysis (PCA) was applied to reduce the dimensionality of the feature set prior to model training, aiming to improve generalization and reduce redundancy.

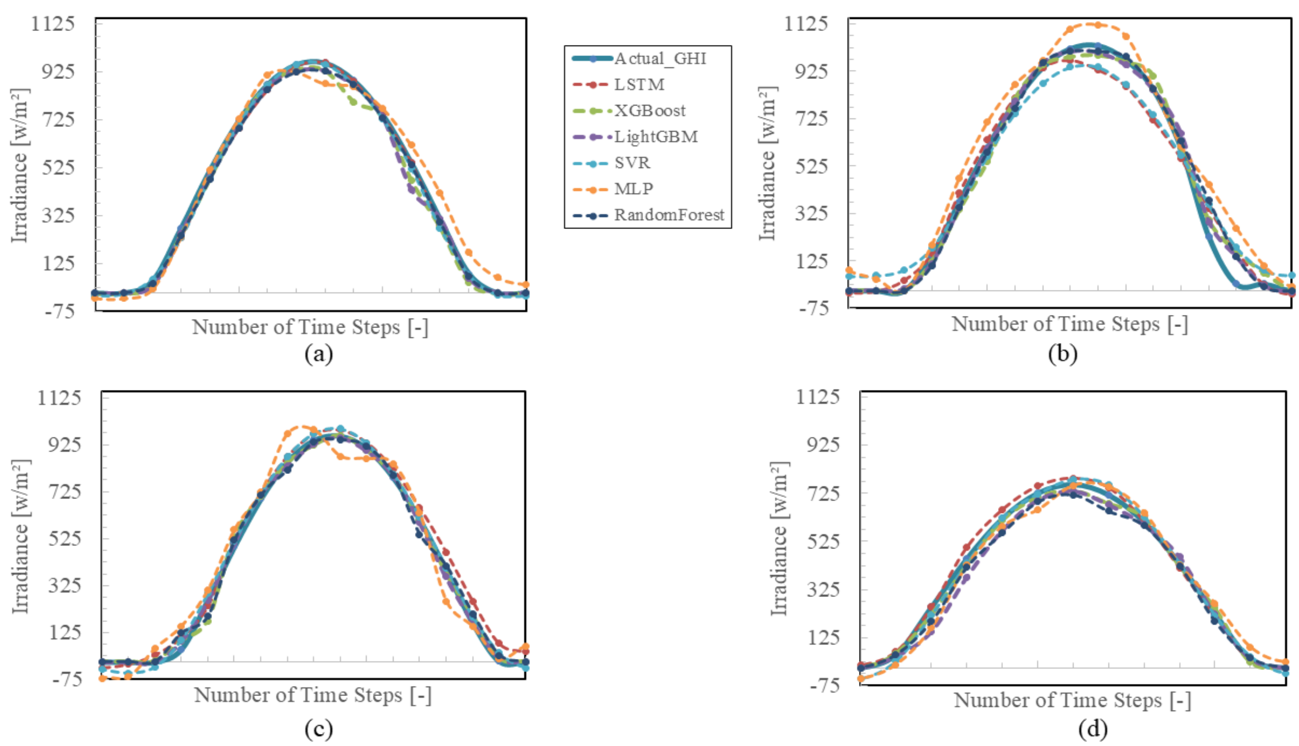

Figure 8. Time series plots of observed versus predicted GHI for machine learning models across limited features for (a) autumn, (b) spring, (c) summer, and (d) winter season

10.57647/ijeee.2025.1602.06

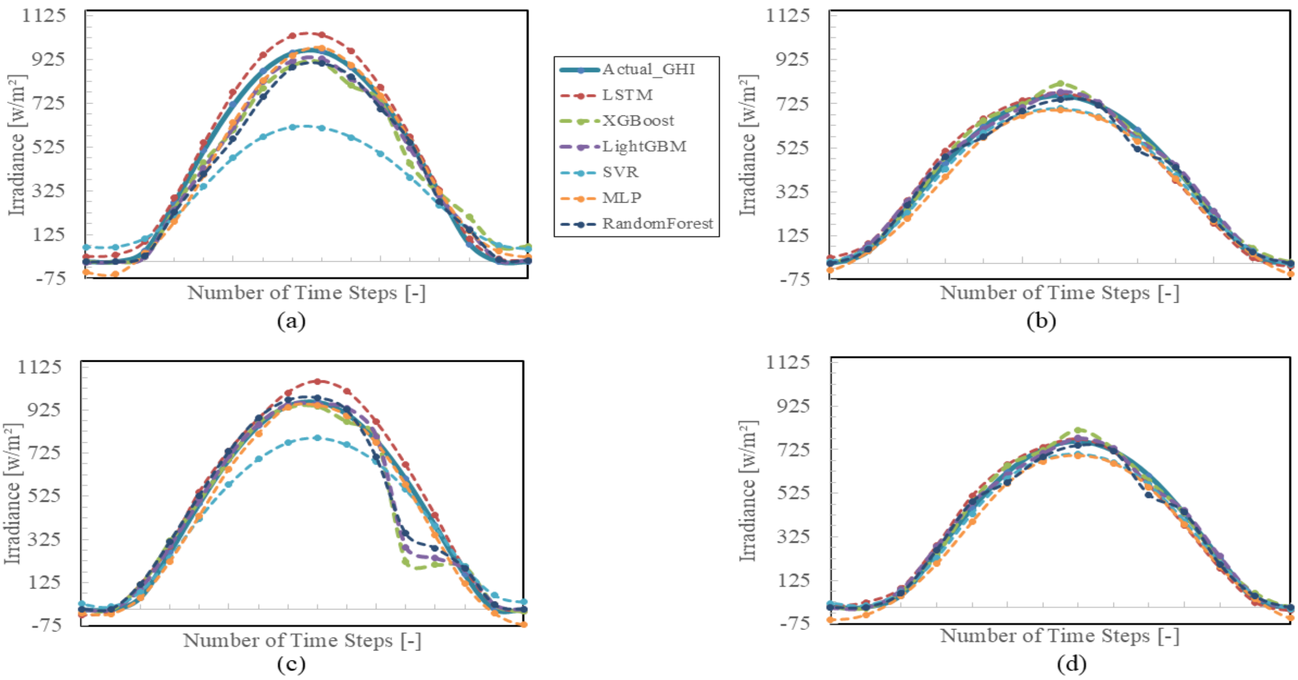

Figure 9. Time series plots of observed versus predicted GHI for machine learning models across PCA features for (a) autumn, (b) spring, (c) summer, and (d) winter season

Compared to the results with limited features (Figure 8), the PCA-based models exhibit both improvements and degradations in performance depending on the season and algorithm. In Autumn (Figure 9a) and Summer (Figure 9c), discrepancies between predicted and observed GHI become more pronounced for certain models, particularly SVR and LightGBM, which show visible deviations from the actual curve, especially during peak irradiance. Overestimation by LSTM is again observed during midday periods, while SVR tends to consistently underestimate, most prominently in summer. Conversely, in Spring (Figure 9b) and Winter (Figure 9d), the use of PCA leads to better alignment between predicted and actual values, with XGBoost and Random Forest continuing to provide the closest fits across most of the diurnal cycle. The clustering of predicted curves around the actual GHI in winter reflects improved stability under lower irradiance conditions, suggesting PCA's potential for simplifying the feature space when dealing with less variable atmospheric conditions. Overall, the results indicate that while PCA helps improve model compactness and stability in some seasons (notably spring and winter), it may introduce performance degradation during periods of high irradiance variability, particularly affecting models like SVR and LightGBM. XGBoost and Random Forest maintain robust predictive performance across all seasons, confirming their adaptability both with limited feature sets and PCA-based dimensionality reduction. These findings highlight that while PCA can be beneficial for reducing feature complexity, its impact on prediction accuracy is model- and season-dependent, necessitating careful evaluation when applied in operational forecasting frameworks.

4. Conclusion

As global energy systems increasingly integrate renewable sources, accurate forecasting of solar power generation has become critical for ensuring grid stability. Global Horizontal Irradiance (GHI), representing total solar radiation on a horizontal surface, is essential for solar energy forecasting. While machine learning (ML) models such as LSTM, XGBoost, LightGBM, SVR, MLP, and Random Forest have shown promise, the impact of seasonal variability on feature relevance and model performance remains underexplored. This study addresses that gap by analyzing how meteorological feature importance changes across seasons and whether season-specific models can improve forecasting accuracy. High-resolution solar and meteorological data from NREL's (24.25°N, 45.34°E, 740 m elevation) were used, covering observed and clear-sky irradiance with key meteorological features. Seasonal partitioning was applied, followed by feature selection using Pearson correlation (threshold 0.25) and dimensionality reduction via PCA. ML and LSTM models were developed, with feature contributions interpreted through SHAP analysis. The results confirm that clear-sky irradiance parameters consistently dominate GHI predictions (correlation >0.95; SHAP $>10^{-1}$).

Temperature, relative humidity, and solar zenith angle showed significant seasonal variation, while wind direction had greater influence in winter. PCA improved stability in spring and winter but reduced accuracy during high irradiance variability. XGBoost and Random Forest models provided the most accurate, stable results across seasons. These findings underscore the need for adaptive, season-specific forecasting frameworks to enhance the integration of solar energy. Future work should explore automated seasonally adaptive systems for operational deployment.

Funding

This work has not received any financial support.

Authors Contribution

All authors have contributed equally to prepare the paper.

Availability of data and materials

Data and code will be available upon request.

Conflict of interests

There is no conflict of interest.

References

1. Barhmi, Khadija, Chris Heynen, Sara Golroodbari, and Wilfried van Sark. "A review of solar forecasting techniques and the role of artificial intelligence." In Solar, vol. 4, no. 1, pp. 99-135. MDPI, 2024.

2. Tercha, Wassila, Sid Ahmed Tadjer, Fathia Chekired, and Laurent Canale. "Machine learning-based forecasting of temperature and solar irradiance for photovoltaic systems." Energies 17, no. 5 (2024): 1124.

3. Dos Reis, Josivan R., Jonathan M. Tabora, Matheus C. Lima, Flávia P. Monteiro, Suzane CA Monteiro, Ubiratan H. Bezerra, and Maria EL Tostes. "Medium and Long Term Energy Forecasting Methods: A Literature Review." IEEE Access (2025).

4. Boubii, Chaymae, Ismail El Kafazi, Rachid Bannari, Brahim El Bhiri, Badre Bossoufi, Hossam Kotb, Kareem M. AboRas, Ahmed Emara, and Badr Nasiri. "Synergizing wind and solar power: An advanced control system for grid stability." Sustainability 16, no. 2 (2024): 815.

5. Shair, Jan, Haozhi Li, Jiabing Hu, and Xiaorong Xie. "Power system stability issues, classifications and research prospects in the context of high-penetration of renewables and power electronics." Renewable and Sustainable Energy Reviews 145 (2021): 111111.

6. Alajrash, Ban H., Mohamed Salem, Mahmood Swadi, Tomonobu Senjyu, Mohamad Kamarol, and Saad Motahhir. "A comprehensive review of FACTS devices in modern power systems: Addressing power quality, optimal placement, and stability with renewable energy penetration." Energy Reports 11 (2024): 5350-5371.

7. Gaboitaolelwe, Jwaone, Adamu Murtala Zungeru, Abid Yahya, Caspar K. Lebekwe, Dasari Naga Vinod, and Ayodeji Olalekan Salau. "Machine learning based solar photovoltaic power forecasting: A review and comparison." IEEe Access 11 (2023): 40820-40845.

8. AlKandari, Mariam, and Imtiaz Ahmad. "Solar power generation forecasting using ensemble approach based on deep learning and statistical methods." Applied Computing and Informatics 20, no. 3/4 (2024): 231-250.

9. Singh, Upma, Shekhar Singh, Saket Gupta, Majed A. Alotaibi, and Hasmat Malik. "Forecasting rooftop photovoltaic solar power using machine learning techniques." Energy Reports 13 (2025): 3616-3630.

10. Jung, Yoonhwa, Jaehoon Jung, Byungil Kim, and SangUk Han. "Long short-term memory recurrent neural network for modeling temporal patterns in long-term power forecasting for solar PV facilities: Case study of South Korea." Journal of Cleaner Production 250 (2020): 119476.

11. Das, Utpal Kumar, Kok Soon Tey, Mohd Yamani Idna Bin Idris, Saad Mekhilef, Mehdi Seyedmahmoudian, Alex Stojcevski, and Ben Horan. "Optimized support vector regression-based model for solar power generation forecasting on the basis of online weather reports." IEEE Access 10 (2022): 15594-15604.

12. Obiora, Chibuzor N., Ahmed Ali, and Ali N. Hasan. "Implementing extreme gradient boosting (XGBoost) algorithm in predicting solar irradiance." In 2021 IEEE PES/IAS PowerAfrica, pp. 1-5. IEEE, 2021.

13. Muhammad Ehsan, R., Sishaj P. Simon, and P. R. Venkateswaran. "Day-ahead forecasting of solar photovoltaic output power using multilayer perceptron." Neural Computing and Applications 28 (2017): 3981-3992.

14. Liu, Da, and Kun Sun. "Random forest solar power forecast based on classification optimization." Energy 187 (2019): 115940.

15. Sharadga, Hussein, Shima Hajimirza, and Robert S. Balog. "Time series forecasting of solar power generation for large-scale photovoltaic plants." Renewable Energy 150 (2020): 797-807.

16. Bett, Philip E., Hazel E. Thornton, Alberto Troccoli, Matteo De Felice, Emma Suckling, Laurent Dubus, Yves-Marie Saint-Drenan, and David J. Brayshaw. "A simplified seasonal forecasting strategy, applied to wind and solar power in Europe." Climate services 27 (2022): 100318.

17. Chattopadhyay, Sameep, Pulkit Paliwal, Sai Shankar Narasimhan, Shubhankar Agarwal, and Sandeep P. Chinchali. "Context Matters: Leveraging Contextual Features for Time Series Forecasting." arXiv preprint arXiv:2410.12672 (2024).

18. Wogrin, Sonja. "Time series aggregation for optimization: One-size-fits-all?." IEEE Transactions on Smart Grid 14, no. 3 (2023): 2489-2492.

19. Castangia, Marco, Alessandro Aliberti, Lorenzo Bottaccioli, Enrico Macii, and Edoardo Patti. "A compound of feature selection techniques to improve solar radiation forecasting." Expert Systems with Applications 178 (2021): 114979.