Impact of Solar Peak Energy and Weather Classification on BiLSTM-Based Solar Irradiance Forecasting in Equatorial Regions

Lek Keng Lim, Wai Shin Ho*, Haslenda Hashim and Zarina Ab Muis

Process System Engineering Center (PROSPECT), Faculty of Chemical and Energy Engineering, Universiti Teknologi Malaysia

Abstract: Accurate solar irradiance forecasting is critical for managing solar energy systems in equatorial regions, where high solar potential is coupled with significant variability. This study investigates the influence of solar peak energy and weather classification features on a Bidirectional Long Short-Term Memory (BiLSTM) model for multi-step Global Horizontal Irradiance (GHI) prediction. The research method involved four phases: data preprocessing (including cyclical time encoding, lag features, and solar peak extraction), weather classification (pseudo-labelling refined by decision trees), BiLSTM-based forecasting with Optuna hyperparameter tuning, and model evaluation using standard error metrics. Three configurations were compared: (A) solar peak + weather classification, (B) weather classification only, and (C) core meteorological and temporal features without additional inputs. The workflow incorporated cyclical time encoding, pseudo-labelling with decision tree refinement, lag feature construction, and Optuna-based hyperparameter tuning. Model performance was assessed using MAE, RMSE, MAPE, R^2, and MASE. Scenario A achieved the lowest MAPE (28.74%), whereas Scenario C yielded the smallest MAE (103.59 W/m²) and MASE (0.786). Scenario B performed worst with a MAPE of 29.85% and MAE of 105.64 W/m², highlighting the limited standalone value of weather classification. Across all scenarios, RMSE values remained within 148–150 W/m² and R^2 around 0.68, reflecting minimal differences in variance explanation. These findings suggest that simpler models can perform as well as or even outperform more complex configurations, offering efficiency benefits for operational forecasting. The practical implication of these results is that reliable irradiance forecasts can be achieved with simpler BiLSTM configurations, reducing computational cost and supporting real-time energy management, PV system sizing, and grid stability in equatorial regions. Future work should incorporate satellite imagery and real-time cloud tracking to further enhance prediction accuracy.

Keywords: Solar irradiance forecasting, Bidirectional LSTM (BiLSTM), Feature engineering, Solar peak energy, Weather classification.

1. INTRODUCTION

The global transition toward renewable energy (RE) sources has gained significant momentum in recent years due to rising environmental concerns and the depletion of fossil fuels. Among various RE options, solar energy stands out as a major contributor, particularly in countries with high solar exposure. Many nations, including those along the equatorial belt, are increasingly adopting solar photovoltaic (PV) systems as a sustainable energy solution. In Malaysia—located near the equator—solar energy holds great potential due to year-round sunlight. However, the intermittency and variability of solar radiation, particularly Global Horizontal Irradiance (GHI), present challenges in ensuring consistent energy generation. Fluctuations in GHI caused by cloud cover, humidity, and other atmospheric conditions make accurate forecasting essential for efficient solar energy management and system design. To address these challenges, various forecasting techniques have been explored, with machine learning (ML) and deep learning (DL) models emerging as powerful tools for GHI prediction.

Pazikadin, Rifai [1] conducted a comprehensive review of solar irradiance measurement technologies and the application of Artificial Neural Networks (ANNs) for forecasting solar power generation, based on an analysis of 87 research articles published between 2014 and 2019. The study focuses on ANN and hybrid ANN systems (*e.g.*, ANN-Wavelet, ANN-GA) used for forecasting solar power output, highlighting their structure, input parameters (weather and irradiance data), and performance compared to traditional statistical models. The review finds that ANN-based methods, especially hybrid systems, outperform conventional models in forecast accuracy, with RMSE values often below 10%, and show strong adaptability to diverse conditions and data inputs. While ANN shows high accuracy, the review notes limitations including its dependency on high-quality training data, lack of use of time-sequence models like LSTM, and limited standardization in evaluation metrics across studies, making direct comparison difficult. However, one notable limitation in many of the reviewed studies is the limited incorporation of time-sequence models, such as LSTM, which are well-suited for capturing temporal dependencies in solar irradiance data. Given that solar GHI (Global Horizontal Irradiance) exhibits strong time-dependent patterns influenced by factors like time of day and weather dynamics, incorporating models that account for these temporal trends is essential for improving forecast accuracy.

Khan, Mazhar [2] compared multiple deep neural network models (LSTM, BiLSTM, GRU, CNN-LSTM)

*Address correspondence to this author at the Process System Engineering Center (PROSPECT), Faculty of Chemical and Energy Engineering, Universiti Teknologi Malaysia; E-mail: hwshin@utm.my

Received on 13-08-2025 Accepted on 17-09-2025 Published on 20-09-2025

E-ISSN: 2410-2199/25

with traditional methods (Random Forest, SVR, ARIMA) to forecast solar and wind power output using meteorological and time-based features such as irradiance, temperature, time of day, and humidity. Models were tuned using Randomized Search CV, with CNN layers extracting spatial features and LSTM layers modelling temporal patterns. SHAP (SHapley Additive exPlanations) analysis was applied for model interpretability, confirming solar angle, irradiance, and time as the most influential variables. Results showed tuned LSTM models outperforming all others, followed by CNN-LSTM, GRU, and BiLSTM, while Random Forest and SVR lagged behind. Although accuracy was high, computational demands of models like CNN-LSTM may limit their use in low-resource settings.

Umaeswari, Sonia [3] explored the integration of IoT and machine learning to optimise solar energy consumption in a duplex residential building using lithium-ion batteries and thermal cooling systems over a year-long study. Twenty-one machine learning algorithms were tested, with particular emphasis on Gaussian Probability models and optimisation techniques such as Particle Swarm Optimisation (PSO), Grey Wolf Optimisation (GWO), and Moth-Flame Optimisation (MFO) to analyse energy usage patterns and improve system efficiency. The study achieved up to 41.6 kW/month in energy savings, reduced electricity bills, extended panel lifespan, and enabled effective real-time control via IoT, with GWO emerging as the most effective optimisation method. Across these studies, incorporating meteorological and time-based features enhances forecasting performance, with LSTM consistently demonstrating the highest accuracy, albeit at the cost of greater computational requirements.

Qing and Niu [4] proposed a deep learning approach using Long Short-Term Memory (LSTM) networks for hourly day-ahead solar irradiance forecasting, utilizing only weather forecast data and time features, without relying on historical irradiance measurements .The authors design an LSTM network to model dependencies across consecutive hours in a day, treating the task as a structured output prediction problem. Inputs include 9 weather/time features, and the model is compared against linear regression, persistence, and BPNN (backpropagation neural network). The proposed LSTM model outperforms all baselines in both accuracy and generalization, showing 18.34% lower RMSE than BPNN on a two-year dataset and 42.9% lower RMSE on a large-scale 11-year dataset (compared to BPNN). While the LSTM model focuses on predictive accuracy, recent studies have extended such models to practical energy applications.

Zhang [5] develops a novel hybrid machine learning model (CEEMDAN-SE-GWO-SVR) to improve the accuracy of direct normal irradiance (DNI) forecasts and demonstrates its real-world application by simulating hydrogen production using a photoelectrochemical (PEC) device. The model combines CEEMDAN (decomposition), Sample Entropy (SE) (clustering), Grey Wolf Optimizer (GWO) (hyperparameter tuning), and Support Vector Regression (SVR) (prediction) to handle complex solar data and enhance predictive robustness. The model combines CEEMDAN (decomposition), Sample Entropy (SE) (clustering), Grey Wolf Optimizer (GWO) (hyperparameter tuning), and Support Vector Regression (SVR) (prediction) to handle complex solar data and enhance predictive robustness. The hybrid model outperformed MLP and LSTM across all seasons, achieving an R^2 of 0.97, RMSE of 43.25, and enabling a peak hydrogen production rate of 57.5 μg/s, validating both prediction accuracy and real-world energy application.

Although LSTM yields strong forecasting results, its performance can be challenged by highly variable and unpredictable weather patterns. Clustering techniques are recommended to categorize weather patterns prior to prediction, allowing models to be trained on distinct categories and better capture varying weather trends. Chen, Lin [6] proposes a hybrid forecasting model combining K-means++ clustering with a CNN-LSTM neural network to improve 1-hour-ahead global horizontal irradiance (GHI) predictions, based on multivariate time-series meteorological data. The paper addresses the limitation of static weather-type classification in prior models by introducing real-time input-side clustering, enabling adaptive model selection based on current irradiance patterns. Machine learning is applied through four cluster-specific CNN-LSTM models, each trained on data grouped by K-means++ clustering of GHI sequences; this allows better learning of both spatial patterns (via CNN) and temporal dependencies (via LSTM). Dou, Wang [7] propose a hybrid deep learning model to correct day-ahead global horizontal irradiance (GHI) forecasts from Numerical Weather Prediction (NWP) models under varying weather conditions. The framework integrates Deep Clustering (DC), Variational Mode Decomposition (VMD), and a Convolutional LSTM-based Encoder–Decoder Correction model (EDC). DC, implemented with a CLSTM encoder, clusters input data by weather condition, enabling the training of separate correction models for each cluster. VMD enhances feature representation by decomposing GHI signals to reduce redundancy, while CNN layers within the CLSTM-based encoder–decoder extract spatial features and model temporal dependencies.

The hybrid approach achieves superior performance, reducing RMSE to 75.51 W/m² at Solar Plant 1 and outperforming baseline models such as LSTM, SVR, BPNN, and newer architectures like TFT, with robust results across sunny, cloudy, and overcast scenarios. However, limitations include high computational cost from cluster-specific training, bias under high-GHI conditions (>400 W/m²), and occasional clustering errors due to noisy NWP data. The absence of satellite/cloud imagery and wind-related features also constrains performance, suggesting that incorporating such data could further improve forecast accuracy.

Dou, Wang [8] proposes a hybrid forecasting framework (MMDC-MMIF) for day-ahead global horizontal irradiance (GHI) prediction, combining deep clustering and multi-modal fusion of observed GHI, NWP GHI, and ground-based cloud images. The paper introduces a multi-modal deep clustering (MMDC) method using CNN, LSTM, and VGG/Swin Transformer to jointly cluster weather patterns and a multi-modal forecasting module (MMIF) to perform GHI predictions tailored to those clusters. The MMDC-MMIF model achieves the lowest RMSE (29.36 W/m²) and highest correlation (99.23%), outperforming all baseline models and showing strong robustness across sunny, cloudy, and overcast conditions. The model's performance still drops under overcast conditions, and it doesn't include variables like humidity, wind, or rainfall; it also assumes static model retraining, requiring future updates for evolving weather patterns. LSTM requires significant training time, and the input data must be preprocessed. Reducing data redundancy can help shorten the training duration.

Rathore, Gupta [9] proposes a hybrid deep learning model called N-FFT-AM-LSTM, combining Noise-Assisted Multivariate Empirical Mode Decomposition (NA-MEMD), Fast Fourier Transform (FFT) for dimensionality reduction, and an Attention-based LSTM network to predict day-ahead hourly Global Horizontal Irradiance (GHI) for four different Indian locations. Most existing models either don’t handle multistep GHI forecasting well or suffer from computational complexity when using raw decomposed components. This paper addresses that by combining FFT to reduce redundancy and attention mechanisms to improve long-term pattern recognition. After decomposing time-series data into Intrinsic Mode Functions (IMFs) using NA-MEMD and reducing their number with FFT, the resulting five frequency-based granularities (HIGH1, HIGH2, HIGH3, MEDIUM, LOW) are each modeled by AM-LSTM networks that prioritize important time-step features. The proposed model outperformed benchmark models like Random Forest, LSTM, and even NA-MEMD-LSTM, achieving RMSE as low as 51.89 W/m² and MAPE below 10% for over 60% of predictions. It also reduced training time by nearly 59% compared to full decomposition models. Additionally, outlier data should be identified and processed during preprocessing to improve model performance.

Pattnaik, Bisoi [10] introduces a hybrid forecasting model named IF-CEEMDAN-LSTM-REDRVFLN, which integrates Isolation Forest (IF) for outlier removal, CEEMDAN for signal decomposition, and a combination of stacked LSTM and REDRVFLN (Recurrent Ensemble Deep Random Vector Functional Link Network) to predict 30-minute and 1-hour ahead solar irradiance (DHI) data. Previous methods often lack effective handling of outliers, non-stationary signals, and temporal dependencies in solar data. Many models either neglect decomposition or apply simple neural networks that fail to generalize on complex and noisy solar datasets. The model uses a four-layer stacked LSTM to capture temporal patterns and long-term dependencies, and then replaces the dense output layer with a REDRVFLN, which uses randomized neurons with local recurrence for faster, more generalizable forecasting. It is preceded by IF to detect outliers and CEEMDAN to break signals into clean components (IMFs). While the model shows superior performance, it relies heavily on complex preprocessing (IF + CEEMDAN), does not use cloud imagery, and its performance may be impacted in extremely low irradiance or weather-unstable regions. Hyperparameter tuning for LSTM layers is also resource-intensive.

Hyperparameter tuning of LSTM layers is essential for improving accuracy, as different input data may require distinct parameter settings. Wang, Yan [11] proposes a new day-ahead GHI prediction model that uses multi-feature perspective clustering (MFPC), combining both time-domain and frequency-domain features. It clusters historical days by weather type and trains weather-specific BiLSTM sub-models, then improves prediction further using a Bayesian probabilistic reconstruction model (FCM-VAE). Previous studies on solar irradiance prediction have not extensively explored the inclusion of daily solar peak energy.

Table **1** summarizes the features employed across the various machine learning models for solar irradiance prediction. The majority of these models consistently utilize core meteorological variables such as time, Global Horizontal Irradiance (GHI), humidity, and temperature. These features are widely recognized for their strong correlation with solar radiation and have been shown to significantly enhance prediction accuracy. Among these, GHI and temperature often

Table 1: Features of Various Machine Learning Models for Solar Irradiance Prediction

	Time	Irradiance	Humidity	Temperature	Rainfall	Cloud Opacity	Solar Peak Energy	Wind Speed
Umaeswari, Sonia [3]	✗	✔	✗	✔	✗	✗	✗	✗
Jeon, Yeon [12]	✗	✔	✔	✔	✗	✗	✗	✗
Chen, Lin [6]	✔	✔	✗	✔	✗	✔	✗	✗
Dou, Wang [8]	✔	✔	✗	✗	✗	✔	✗	✗
Qing and Niu [4]	✔	✗	✔	✔	✔	✔	✗	✔
Zhang [5]	✔	✔	✔	✔	✔	✔	✗	✔
Dou, Wang [7]	✔	✔	✔	✔	✗	✔	✗	✗
Pattnaik, Bisoi [10]	✔	✔	✗	✗	✗	✗	✗	✗
Khan, Mazhar [2]	✔	✔	✔	✔	✗	✗	✗	✔
Wang, Yan [11]	✔	✔	✔	✔	✗	✗	✗	✗

serve as primary predictors due to their direct relationship with solar energy generation patterns. Cloud opacity is included in approximately half of the reviewed models, reflecting its moderate importance in capturing short-term fluctuations caused by varying cloud cover. While cloud data can enhance temporal accuracy, its effectiveness is often limited by the availability and resolution of the input data. On the other hand, features such as rainfall and wind speed are less frequently adopted, appearing only in a small subset of studies. This may be due to their indirect influence on irradiance or the added complexity they introduce without a proportional gain in model performance.

Daily peak solar irradiation is often underutilized in solar energy prediction models. However, Alizamir, Shiri [13] employed daily solar irradiation data to estimate solar generation yield, demonstrating that accurate prediction of daily solar radiation using advanced machine learning models—particularly the wavelet long short-term memory (WLSTM) method—plays a critical role in optimizing solar energy generation and management. In the context of solar system design, peak power analysis has been effectively applied to estimate the number of solar PV modules required. When integrated with daily energy consumption data, this approach enables more accurate system sizing to meet household energy demands [14]. Furthermore, the need to improve the estimation of daily peak solar radiation has been identified as a promising direction for future research aimed at enhancing prediction accuracy [15].

Building on these insights, this study aims to investigate the impact of incorporating solar peak energy and weather classification on forecasting accuracy using a Bidirectional Long Short-Term Memory (BiLSTM) framework. Additionally, the study examines the effect of data clustering on prediction performance. To achieve this purpose, the study (1) preprocesses and structures solar and meteorological data with cyclical time encoding and lag feature construction, (2) develops weather classification through pseudo-labelling and decision tree refinement, (3) implements BiLSTM forecasting models under three different feature configurations, and (4) evaluates their predictive performance using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), coefficient of determination (R^2), and Mean Absolute Scaled Error (MASE). The novelty of this study lies in the incorporation of daily solar peak energy and refined weather classification as input features for BiLSTM-based solar irradiance forecasting, and in systematically comparing their contributions across different scenarios. Ultimately, the findings aim to provide insights for improving the efficiency and practicality of solar energy forecasting systems in equatorial regions.

2. METHODOLOGY

The overall methodology of this study is organized into four main phases, namely data processing, weather classification, BiLSTM-based forecasting, and model evaluation and reporting. Figure **1** shows the research flow chart and describes the activities in each phase. Each phase is carefully structured to prepare, transform, and utilize solar and meteorological data in order to achieve accurate predictions of Global Horizontal Irradiance (GHI).

The first phase involves comprehensive data preprocessing to improve both the quality and relevance of the input data. Only data collected between 7:00 a.m. and 7:00 p.m. is retained, corresponding to the hours when solar radiation is most

significant. This selective filtering helps to reduce noise and ensures that the model focuses on meaningful patterns related to daylight hours. To help the model interpret temporal features, time is encoded using sine and cosine transformations, which preserve its cyclical nature. In addition to time-related processing, solar data is analyzed to identify the daily peak energy, an important feature in understanding solar performance. Weather-related variables such as cloud opacity, relative humidity, and GHI are also visualized using histograms and other exploratory analysis tools. This analysis supports feature selection and model design decisions in later phases.

In the second phase, the focus shifts to weather classification, which is conducted through a two-step process. The first step involves rule-based pseudo-labelling, where daily weather characteristics are summarized using key metrics. These include the total precipitation rate and the mean values of cloud opacity and relative humidity. Based on these aggregates, initial weather types are assigned as sunny, cloudy, or rainy using predefined classification rules. In the second step, a decision tree classifier is trained to refine these labels. Selected features such as GHI, diffuse horizontal irradiance (DHI), and air temperature are used to train the model. The dataset is divided into training and testing sets, and the classifier is built with a controlled depth to avoid overfitting. The predictions from the decision tree model then replace the initial pseudo-labels, resulting in a final weather type feature that is more consistent and data-driven.

The third phase involves the construction of a GHI forecasting model based on Bidirectional Long Short-Term Memory (BiLSTM) networks. The weather type labels from the previous phase are encoded using a label encoding technique to ensure compatibility with the input format of the neural network. Temporal dependencies are emphasized by introducing lag features, which allow the model to capture the sequential relationships across time steps. Feature scaling is applied to normalize the input variables, and the dataset is organized into several categories including general features, solar peak values, GHI, and weather type. The forecasting model consists of an input layer followed by two stacked Bidirectional LSTM layers, which process temporal sequences in both forward and backward directions. This structure is followed by a dropout layer to reduce overfitting and a dense output layer that generates predictions for thirteen future GHI values. To further enhance the model's performance, hyperparameter optimization is carried out using the Optuna framework. A variety of parameters such as the number of LSTM units, dropout rate, batch size, and number of training epochs are explored. The optimization process includes early stopping and pruning techniques to avoid overfitting and improve training efficiency. Once the model is trained, its architecture and weights are saved to support future forecasting tasks.

The final phase of the methodology focuses on evaluating and reporting the model's performance. A set of standard metrics is used to quantify forecasting accuracy, including mean absolute percentage error (MAPE), mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and the coefficient of determination (R-squared). In addition to numerical evaluation, visual diagnostics such as residual histograms and scatter plots of residuals are produced. These plots provide a clearer understanding of how prediction errors are distributed and whether any systematic bias is present in the model outputs. Together, these evaluation tools ensure that the forecasting framework is both accurate and robust, and they support informed conclusions about model reliability.

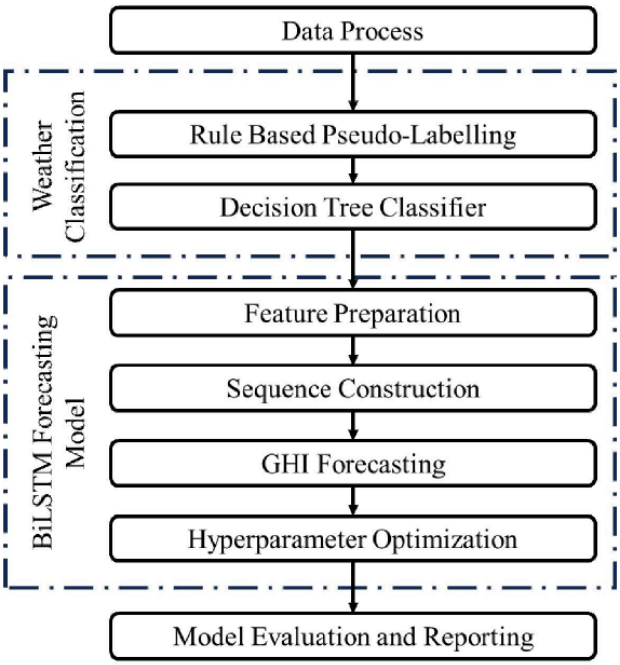

Figure 1: Research Flowchart.

2.1. Mathematical Equations

2.1.1. Cyclical Time Encoding

Time of day is an important feature in predicting solar irradiation, as it strongly influences the amount of incoming solar energy. However, time exhibits a cyclical nature, making it challenging to represent directly in predictive models. To address this, cyclical time encoding is employed to capture the periodicity inherent in daily solar data. In this approach, h represents the hour of the day. Equation 1 and 2 shows the cyclical time encoding in mathematical equation.

$$\text{hour}_{\sin} = \sin\left(\frac{2\pi \cdot \text{h}}{24}\right) \quad \text{Eq (1)}$$

$$\text{hour}_{\cos} = \cos\left(\frac{2\pi \cdot \text{h}}{24}\right) \quad \text{Eq (2)}$$

2.1.2. Pseudo-Labelling Rules

Several rule-based criteria are commonly used to classify weather conditions, and they have proven to be both effective and widely applicable. These rules can serve as an initial step for assigning weather labels prior to further refinement using decision tree algorithms. Equation (3) shows the condition for weather classification.

$$\text{Weather Type} = \begin{cases} \text{Rainy}, & \text{if PRR} > 7 \ \wedge RH > 0.80 \\ \text{Cloudy}, & \text{if OPA} > 39 \\ \text{Sunny}, & \text{otherwise} \end{cases} \quad \text{Eq (3)}$$

2.1.3. Feature Normalization

Min-max normalization is employed in this study to scale the input features to a uniform range prior to training the neural network. Equation 4 shows the normalization used in this study. This normalization facilitates more stable and efficient model learning. Importantly, categorical variables such as weather type, as well as key target variables like global horizontal irradiance (GHI) and solar peak values, are excluded from the scaling process to preserve their accuracy and validity of data integrity and ensure accurate interpretation during prediction.

$$\text{x}_{\text{scaled}} = \left(\frac{(\text{x} - \text{x}_{\min})}{\text{x}_{\max} - \text{x}_{\min}}\right) \cdot (\text{b} - \text{a}) + \text{a} \quad \text{Eq (4)}$$

2.1.4. Lag Features

In solar forecasting, capturing the temporal dependency between consecutive time steps is essential, as global horizontal irradiance (GHI) typically changes gradually rather than abruptly. To model this temporal continuity, lag features are introduced. These features represent previous GHI values and enable the model to learn patterns and trends over time. Incorporating lag features helps the model better understand the sequential relationships in the data. In this study, the global horizontal irradiance (GHI) values from the two preceding time steps are incorporated as lag features to enhance the model's ability to learn temporal patterns and improve forecasting performance.

$$\text{ghilag1}_{\text{t}} = \text{ghi}_{\text{t}-1} \quad \text{Eq (4)}$$

$$\text{ghilag2}_{\text{t}} = \text{ghi}_{\text{t}-2} \quad \text{Eq (5)}$$

2.1.5. Sequence Construction (Sliding Window)

Sequence construction is a crucial step in preparing time series data for forecasting models. It involves creating input-output pairs by slicing the continuous time series into fixed-length sequences as shown in Equation 6 and 7. In this study, input window size is set at 13 while the output steps are set at 91 as shown in Equation 8 and 9. Each input sequence contains historical data points, while the corresponding output sequence represents the future values the model aims to predict. This approach enables the BiLSTM model to learn temporal dependencies by observing how past patterns relate to future trends.

$$\text{X}^{(\text{i})} = [\text{X}_{\text{i}}, \text{X}_{\text{i}+1}, \ldots, \text{X}_{\text{i}+\text{T}-1}] \quad \text{Eq (6)}$$

$$\text{y}^{(\text{i})} = [\text{y}_{\text{i}+\text{T}}, \text{y}_{\text{i}+\text{T}+1}, \ldots, \text{y}_{\text{i}+\text{T}+\text{H}-1}] \quad \text{Eq (7)}$$

$$\text{T} = \text{input window size} = 13 \quad \text{Eq (8)}$$

$$\text{H} = \text{output steps} = 91 \quad \text{Eq (9)}$$

2.1.6. BiLSTM Model Structure

The architecture consists of two stacked BiLSTM layers, where the first layer returns full sequences to feed into the second, enabling deeper temporal feature extraction. Equation 10 -13 shows the BiLSTM model structure. A dropout layer follows to reduce the risk of overfitting by randomly deactivating a fraction of neurons during training. Finally, a dense output layer predicts 13 future GHI values for each input sequence. This design helps the model capture complex temporal dynamics and improve multi-step forecasting accuracy.

$$\overrightarrow{\text{h}_{\text{t}}} = \text{LSTM}_{\text{fwd}}\left(\text{x}_{\text{t}}, \overrightarrow{\text{h}_{\text{t}-1}}\right), \quad \text{Eq (10)}$$

$$\overleftarrow{\text{h}_{\text{t}}} = \text{LSTM}_{\text{bwd}}\ \left(\text{x}_{\text{t}}, \overleftarrow{\text{h}_{\text{t}+1}}\right) \quad \text{Eq (11)}$$

$$\text{h}_{\text{t}} = \left[\overrightarrow{\text{h}_{\text{t}}}; \overleftarrow{\text{h}_{\text{t}}}\right] \quad \text{Eq (12)}$$

$$\hat{\text{y}}_{\text{t}+1:\text{t}+13} = \text{Dense}(\text{Dropout}(\text{h}_{\text{T}})) \quad \text{Eq (13)}$$

2.1.7. Loss Function

The loss function used in this study is the Mean Squared Error (MSE), which calculates the average of the squared differences between actual and predicted GHI values. Is the true value and is the predicted value. Equation 14 shows the loss function. This function penalizes larger errors more heavily and guides the model during training to minimize prediction deviations.

$$\mathcal{L} = \frac{1}{\text{n}}\sum_{\text{i}=1}^{\text{n}} (\text{y}_{\text{i}} - \hat{\text{y}_{\text{i}}})^2 \quad \text{Eq (14)}$$

2.1.8. Evaluation Metrics

In this study, five evaluation metrics are used to assess model performance: MAE (Mean Absolute Error), RMSE (Root Mean Square Error), MAPE (Mean Absolute Percentage Error), R^2 (Coefficient of Determination), and MASE (Mean Absolute Scaled Error). MAE (Mean Absolute Error) measures the average absolute difference between predicted and actual values. Each metrics is given in Equation 15 -19. RMSE penalizes larger errors more heavily due to squaring, making it useful when large deviations are particularly critical. MAPE expresses errors as a percentage of actual values, providing a scale-independent metric, though it can be sensitive when actual values are near zero. R^2, or the coefficient of determination, indicates how well the model explains the variability in the data, with values closer to 1 reflecting better predictive power. Lastly, MASE scales the prediction error relative to a naive baseline, enabling consistent comparison across different datasets or forecasting models. Here are the formulas for each metric provided below.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y_1}| \quad \text{Eq (15)}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_1})^2} \quad \text{Eq (16)}$$

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y_1}}{y_1}\right| \quad \text{Eq (17)}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y_1})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$MASE = \frac{\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y_1}|}{\frac{1}{n}\sum_{i=1}^{n-s}|y_i - y_{i-s}|} \quad \text{Eq (18)}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y_1}| \quad \text{Eq (19)}$$

3. INPUT FEATURE

The forecasting model incorporates thirteen features: air temperature, dew point temperature, relative humidity, wind speed at ten meters, cloud opacity, precipitation rate, sine of the hour, cosine of the hour, global horizontal irradiance lagged by one time step, global horizontal irradiance lagged by two-time steps, solar peak energy, current global horizontal irradiance, and weather type. These features collectively serve as input variables to support accurate forecasting within the model.

3.1. Case Study

Scenario A is enriched with both solar peak indicators and weather classification, offering the most comprehensive contextual input. Scenario B relies solely on weather classification, whereas Scenario C operates without any contextual enhancements. Table **2** shows the description for each scenario.

Table 2: Scenario Analysis for Solar Peak and Weather Classification

	A	B	C
Solar Peak	Y	X	X
Weather Classification	Y	Y	X

3.2. Weather Information

The frequency distributions of three key weather variables—Global Horizontal Irradiance (GHI), cloud opacity, and relative humidity—were analyzed using data collected over four years. Only daytime data, from 7:00 AM to 7:00 PM, was used in this analysis. This time range was chosen because it covers the period when solar radiation is present and relevant to solar energy studies. Nighttime values were excluded, as GHI is zero during those hours and would distort the distribution.

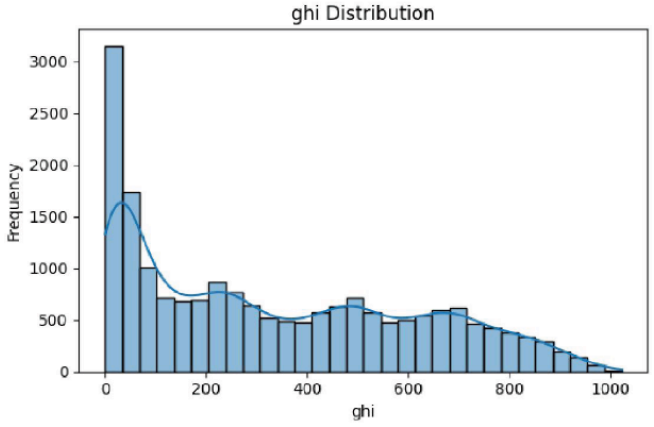

Figure 2: Global Horizontal Irradiance Distribution.

Figure **2** shows the GHI distribution. The distribution of GHI is strongly right-skewed. Most values are on the lower end, meaning that even during the day, solar irradiance is often low. High GHI values, above 800 W/m², occur less frequently. This shows that intense sunlight happens, but not very often. The shape of the distribution suggests that many daytime hours still have weak solar input, possibly due to weather conditions or seasonal changes.

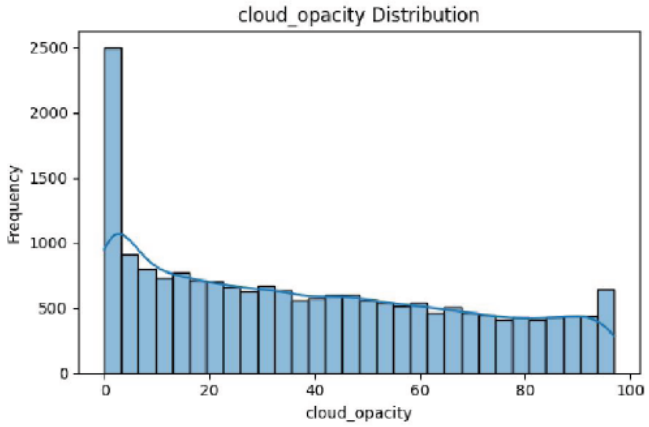

Figure 3: Cloud Opacity Distribution.

Figure **3** shows the distribution of cloud opacity. Cloud opacity also shows a skewed distribution, with many values close to zero. This means clear skies are common during the day. However, there is also a noticeable increase in values near 100, which indicates that overcast conditions are also frequent. Mid-range opacity values are less common. This suggests that skies are often either mostly clear or completely cloudy, rather than partly cloudy.

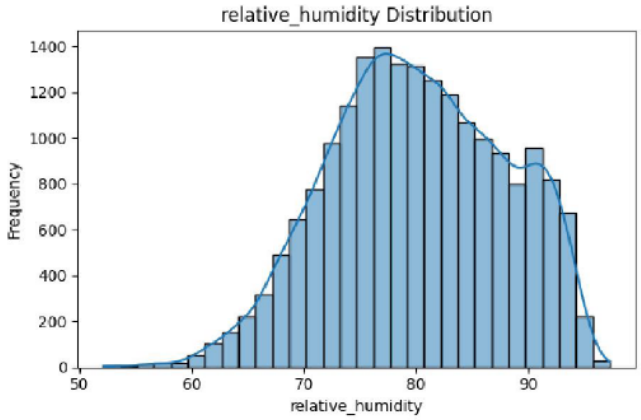

Figure 4: Distribution of Relative Humidity.

Figure **4** shows the distribution of relative humidity. Relative humidity has a different pattern. Its distribution is close to normal, with most values between 70% and 85%. This means that daytime humidity is usually high, and it does not vary as much as GHI or cloud opacity. The results suggest a humid climate, where moisture levels stay fairly consistent throughout the day.

In summary, the daytime weather patterns show high humidity, frequent clear or overcast skies, and mostly low to moderate levels of solar irradiance. These conditions are important to consider when planning or evaluating solar energy systems, as they directly affect energy generation during sunlight hours.

4. RESULTS

4.1. Results Weather Classification

Figure **5** illustrates the number of days classified under different weather types: Sunny, Rainy, and Cloudy, based on a four-year dataset. The chart shows that Sunny days are the most frequent, with approximately 760 days recorded. Rainy days follow, totalling around 540 days, while Cloudy days are the least common, with about 170 days. This distribution suggests that clear weather dominates the local climate, which is favourable for solar energy generation. However, the relatively high number of rainy days indicates significant seasonal or periodic rainfall, which may reduce solar irradiance during those times. The small number of cloudy days, when skies are overcast but without rain, further supports the idea that the region tends to experience either clear or rainy conditions with fewer intermediate states. Overall, this weather pattern aligns with the earlier histogram findings and points to a climate where solar energy is often available but still subject to seasonal limitations due to rain.

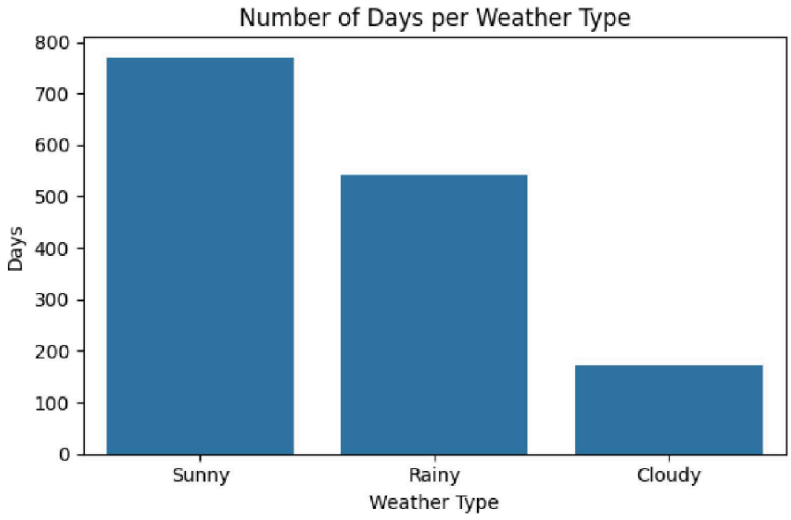

Figure 5: Number of Days per Weather Type.

4.2. Results of GHI Forecasting

Table **3** shows the optimal hyperparameters for scenarios A, B, and C, determined using Optuna tuning. In this context, the "class" refers to the classification of initial data based on weather type. Scenario A includes both weather-based classification and solar peak information, resulting in the highest number of training epochs (149) and the largest dropout rate (0.484). This suggests that the combined input introduced greater complexity, requiring stronger regularisation and extended training to prevent over fitting. Scenario B includes only the weather-based classification and used slightly fewer epochs (140) and the lowest dropout rate (0.329), indicating relatively simpler data patterns that the model could learn more efficiently. Scenario C, which excludes weather classification, required only 67 epochs and a moderate dropout rate (0.385), suggesting faster convergence due to reduced input variability. Both B and C employed a smaller batch size (32), while A used a larger batch size (64), possibly to stabilise updates over more complex input. The number of units remained constant at 256 across all scenarios, indicating this parameter was less sensitive to the changes in input structure. These differences highlight how the inclusion of weather classification and solar peak data influences model complexity and training dynamics.

Table 3: Optimal Hyperparameters for Scenarios A, B, and C

	A	B	C
Epochs	149	140	67
Dropout	0.484369	0.329301	0.385006
Batch Size	64	32	32
Units	256	256	256

Figure **6** presents the cosine similarity matrix of the weight vectors associated with each input feature for Scenario A. This visualisation captures the degree to which different input features are treated similarly by the model in terms of their learned representations. High similarity values (closer to 1, shown in lighter colours) indicate that the model assigns comparable weight patterns to the respective features, suggesting that they may convey related or overlapping information from the model's perspective. Conversely, lower or near-zero similarity values (darker regions) imply that the model processes those features in a more distinct manner.

From Figure **6**, we observe that features such as GHI, and exhibit strong mutual similarity, which aligns with their temporal and physical correlation as successive measurements of global horizontal irradiance. Similarly, temperature-related features like air temperature and dew point temperature also show a moderate degree of similarity, likely reflecting their

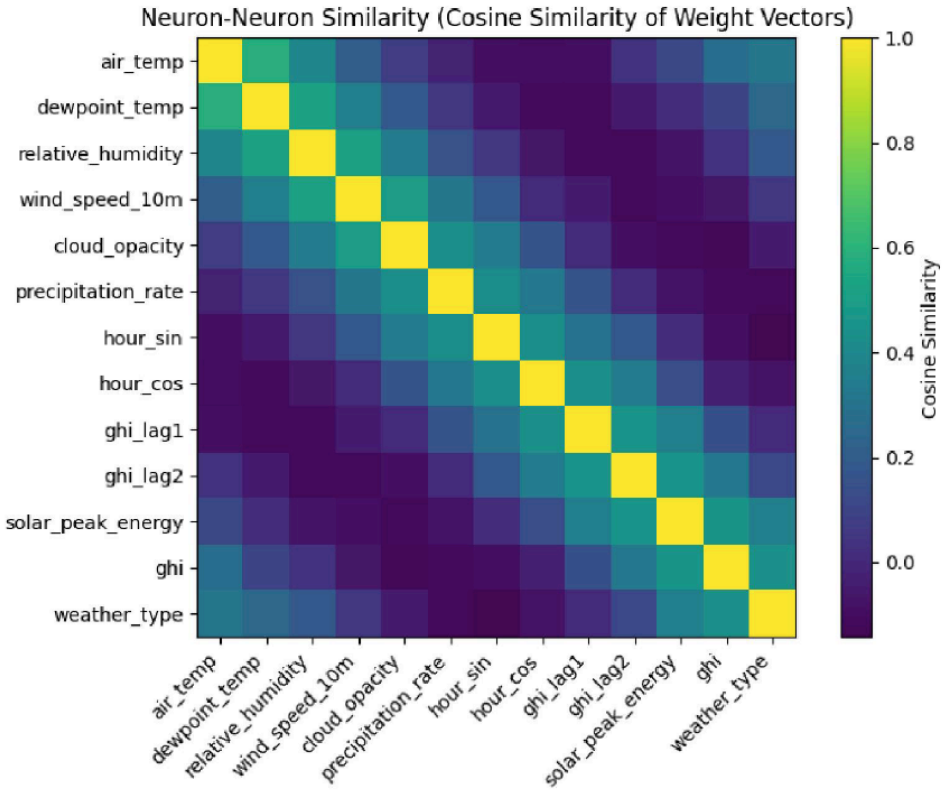

Figure 6: Cosine Similarity matrix of the weight vectors.

meteorological interdependence. In contrast, categorical or cyclic time features such as, and weather class appear less similar to most continuous weather variables, indicating that they contribute uniquely to the model's internal representation.

Figure **7** presents the residual error distribution for Scenario A as a histogram, showing an approximately symmetric and bell-shaped pattern centered near zero. This indicates that the model does not have a significant bias toward over- or under-prediction. Most residuals are tightly clustered around the center, suggesting consistent accuracy in the majority of predictions. However, longer tails on both sides reveal the presence of a few large errors, including some extreme positive and negative residuals, which may be caused by sudden weather changes or rare, complex atmospheric conditions that are difficult to capture. The sharp peak and gradual tapering of the histogram reflect the model's overall precision and stability, with many small residuals indicating good performance. Despite this, the occasional large errors highlight room for improvement.

Figure **8** shows a similarly symmetric distribution but with a wider spread compared to Figure **7**, indicating more variability in residuals. Figure **9** also exhibits symmetry but with heavier tails, suggesting a higher frequency of larger errors. The peak frequencies are 11,700 for Scenario A, 11,000 for Scenario B, and 11,500 for Scenario C. Scenario A's narrowest spread and highest peak reflect better model precision, with residuals more tightly clustered near zero, suggesting the best overall performance among the three scenarios. Scenario B's broader peak and greater number of residuals farther from zero indicate a higher error variance and slightly weaker predictive performance compared to Scenario A. Scenario C performs moderately, with a peak frequency higher than Scenario B but lower than Scenario A, indicating better performance than Scenario B but still not reaching the precision of Scenario A.

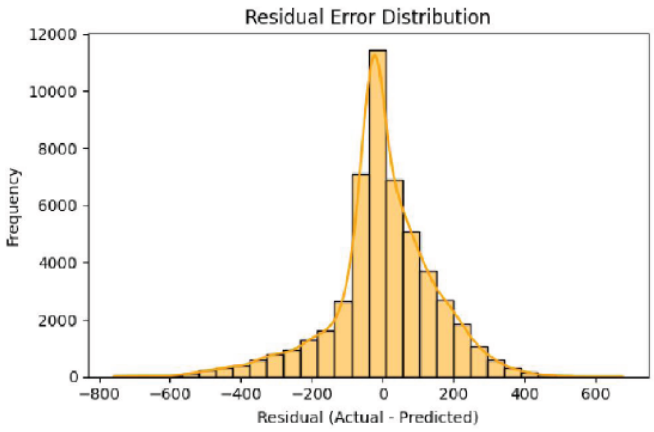

Figure 7: Residual Error Distribution Of Scenario A.

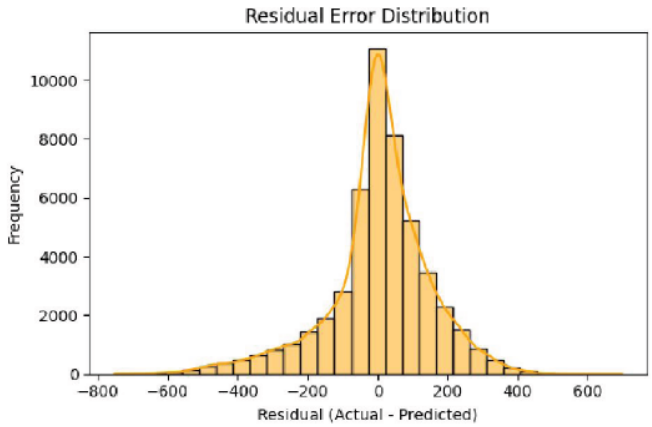

Figure 8: Residual Error Distribution of Scenario B.

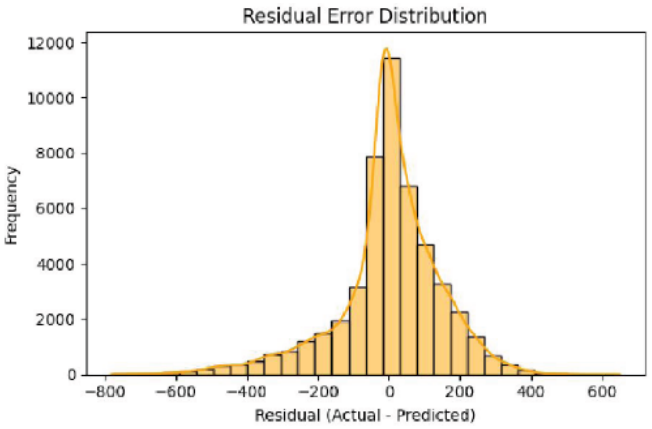

Figure 9: Residual Error Distribution of Scenario C.

Table **4** shows the forecasting performance of solar irradiation across each scnearios. The forecasting performance of solar irradiation was assessed using three model configurations with varying levels of input complexity: Scenario A incorporated both weather classification and solar peak data; Scenario B included only weather classification; and Scenario C relied solely on core features without any additional inputs. As shown in Table **1**, Scenario A achieved the lowest Mean Absolute Percentage Error (MAPE) at 28.74%, indicating a marginal improvement in relative forecasting accuracy. However, this improvement was slight compared to Scenario C, which recorded a MAPE of 29.34%. Notably, Scenario C also achieved the lowest values for Mean Absolute Error (MAE) and Mean Absolute Scaled Error (MASE), at 103.59 and 0.786 respectively, suggesting that the simplest model, without any auxiliary input variables, can be more effective in minimizing absolute and scaled forecast errors. Scenario B, which relied solely on weather classification, consistently demonstrated the poorest performance across all evaluated metrics, with a MAPE of 29.85%, MAE of 105.64, and MASE of 0.802. These findings imply that weather classification, when used in isolation, may not contribute meaningful predictive information and could potentially introduce noise or irrelevant variability into the model.

Table 4: Forecasting Performance of Each Scenarios

	A	B	C
MAPE (%)	28.74123	29.85167	29.34299
MAE (W/m^2)	105.2149	105.6391	103.5868
RMSE (W/m^2)	148.1502	149.7987	148.4406
R2	0.681467	0.674339	0.680217
MASE	0.79847	0.801689	0.786115

Further analysis of Root Mean Square Error (RMSE) and the coefficient of determination (R^2) reinforces the observation that the inclusion of additional features provides minimal practical benefit. All three scenarios yielded nearly identical RMSE values, hovering around 148–150 W/m², and R^2 values approximately 0.68, indicating similar levels of predictive accuracy and variance explanation. Although Scenario A showed the highest R^2 (0.681), the difference was not substantial when compared to Scenario C (0.680), while Scenario B again lagged slightly (0.674). These results suggest that while the inclusion of both weather classification and solar peak data in Scenario A may offer a slight edge in certain relative accuracy measures, the overall impact on forecasting performance is minimal. In contrast, the strong performance of Scenario C across multiple error metrics highlights the potential sufficiency of a simpler model structure without added complexity. Therefore, from both a practical and computational perspective, it may be more effective to adopt a parsimonious modeling approach. The limited gains from additional features underscore the need for careful feature selection and suggest that further improvements in forecasting accuracy may depend more on advanced modeling techniques or better-quality data, rather than simply increasing input dimensionality.

Figure **10** presents a comparison between actual values and predictions generated by three BiLSTM model configurations, each using different combinations of input features. Scenario A integrates both solar peak information and weather classification and shows the closest alignment with the actual data. The model accurately tracks the progression of values throughout the time series, especially during the peak period. This suggests that combining these features significantly enhances predictive performance. Scenario B uses only weather classification and shows moderate accuracy. It captures the overall trend but underestimates the peak and responds less effectively to rapid changes. In contrast, Scenario C excludes both contextual features and produces the lowest accuracy. Its output is overly smoothed and fails to reflect the dynamic patterns in the actual data, particularly around peak hours. These results highlight the importance of including both weather-related and solar peak features to improve the temporal accuracy and responsiveness of deep learning models in solar or energy-related time-series forecasting tasks.

4.3. Discussion

The results show that adding more features does not always make the forecasting model better. Although Scenario A had slightly lower percentage error, the simple model in Scenario C gave the best overall accuracy. This means that using basic meteorological and time features, together with good preprocessing, can be enough for reliable forecasts.

Scenario B, which used only weather classification, gave the weakest results. This suggests that weather classification alone does not provide much useful information for prediction in equatorial regions, where weather changes quickly. However, when combined

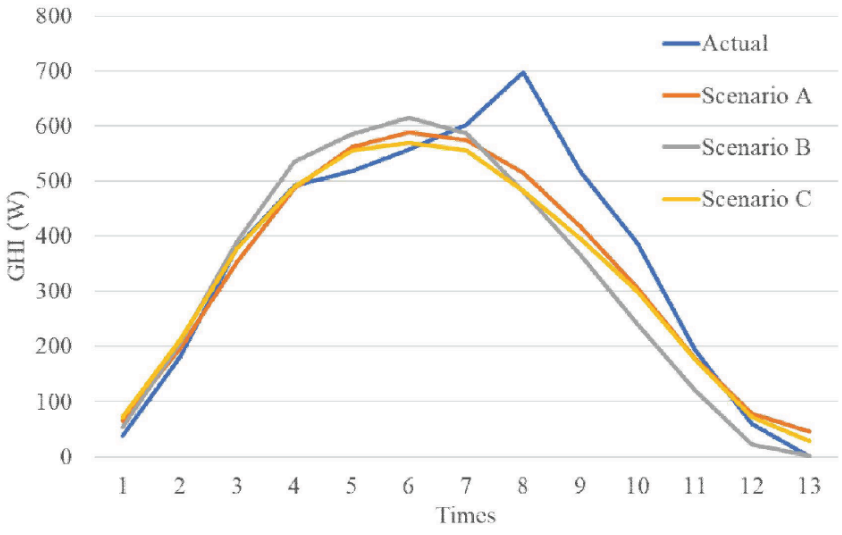

Figure 10: Comparison Between Actual Values and Predictions

with solar peak energy, it gave small improvements, showing that it may still play a supporting role.

These outcomes agree with earlier studies that highlight how data quality and preparation are often more important than adding extra features. In this work, methods like lag features and cyclical time encoding helped the BiLSTM model capture time patterns more effectively.

From a practical view, the results are useful because they show that accurate forecasts can be achieved with simpler models that use less computing power. This makes them easier to apply in real-time systems, such as PV sizing, power scheduling, and energy storage planning. For future improvements, adding inputs like satellite images, cloud tracking, or more advanced clustering could help capture the fast-changing conditions typical in equatorial regions.

5. CONCLUSION

This study explored the effectiveness of incorporating solar peak irradiation and weather This study explored the effectiveness of incorporating solar peak irradiation and weather classification into machine learning models for forecasting global horizontal irradiance (GHI), particularly in equatorial regions like Malaysia, where solar variability poses significant challenges. While the inclusion of solar peak data and weather classification (Scenario A) showed slight improvements in relative forecasting accuracy, the simplest model configuration (Scenario C), relying solely on core meteorological and temporal features, consistently outperformed others in terms of absolute and scaled error metrics. Specifically, Scenario A achieved the lowest MAPE (28.74%), whereas Scenario C recorded the smallest MAE (103.59 W/m²) and MASE (0.786). In contrast, Scenario B performed worst with a MAPE of 29.85% and MAE of 105.64 W/m². Across all scenarios, RMSE values remained within 148–150 W/m² and R^2 around 0.68, indicating minimal differences in variance explanation.

These findings highlight that increased model complexity does not necessarily lead to significant gains in performance and that simpler models may offer more robust, efficient, and computationally practical solutions for solar irradiance forecasting. Moreover, the limited impact of weather classification on prediction accuracy suggests the need for more meaningful feature engineering or the integration of alternative data sources such as satellite imagery or real-time cloud tracking to enhance model input. Importantly, the results also emphasize that data preprocessing plays a critical role in model performance, as properly cleaned, de-noised, and structured input data can significantly improve accuracy and reduce training complexity. Furthermore, shifting the focus toward models with shorter output windows may improve responsiveness and accuracy in real-world energy management systems, making them more suitable for dynamic, high-resolution applications such as smart grid optimization and real-time load balancing. The practical implication of these findings is that reliable solar irradiance forecasts can be achieved using simpler BiLSTM configurations, which reduces computational cost and makes forecasting more accessible for real-time energy management systems. In practice, this enables more efficient PV system sizing, improved scheduling of solar power generation, enhanced grid stability, and better planning for energy storage integration in equatorial regions.

CONFLICTS OF INTEREST

The author declared no conflicts of interest.

REFERENCES

[1] Pazikadin AR, Rifai D, Ali K, Malik MZ, Abdalla AN, Faraj MA. Solar irradiance measurement instrumentation and power solar generation forecasting based on Artificial Neural Networks (ANN): A review of five years research trend. Sci Total Environ. 2020; 715: 136848. https://doi.org/10.1016/j.scitotenv.2020.136848

[2] Khan S, Mazhar T, Khan MA, Shahzad T, Ahmad W, Bibi A, *et al*. Comparative analysis of deep neural network architectures for renewable energy forecasting: enhancing accuracy with meteorological and time-based features. Discover Sustainability. 2024; 5(1). https://doi.org/10.1007/s43621-024-00783-5

[3] Umaeswari P, Sonia R, Saravanan TR, Poyyamozhi N. IoT-Enabled energy conservation in residential Buildings: Machine learning models for analyzing annual solar power consumption. Solar Energy. 2024; 281. https://doi.org/10.1016/j.solener.2024.112890

[4] Qing X, Niu Y. Hourly day-ahead solar irradiance prediction using weather forecasts by LSTM. Energy. 2018; 148: 461-8. https://doi.org/10.1016/j.energy.2018.01.177

[5] Zhang Y. Enhancing solar irradiance prediction for sustainable energy solutions employing a hybrid machine learning model; improving hydrogen production through Photoelectrochemical device. Applied Energy. 2025; 382. https://doi.org/10.1016/j.apenergy.2025.125280

[6] Chen Y, Lin C, Liu J, Yu D. One-hour-ahead solar irradiance forecast based on real-time K-means++ clustering on the input side and CNN-LSTM. Journal of Atmospheric and Solar-Terrestrial Physics. 2025; 266. https://doi.org/10.1016/j.jastp.2024.106405

[7] Dou W, Wang K, Shan S, Li C, Wang Y, Zhang K, *et al*. Day-ahead Numerical Weather Prediction solar irradiance correction using a clustering method based on weather conditions. Applied Energy. 2024; 365. https://doi.org/10.1016/j.apenergy.2024.123239

[8] Dou W, Wang K, Shan S, Chen M, Zhang K, Wei H, *et al*. A multi-modal deep clustering method for day-ahead solar irradiance forecasting using ground-based cloud imagery and time series data. Energy. 2025; 321. https://doi.org/10.1016/j.energy.2025.135285

[9] Rathore A, Gupta P, Sharma R, Singh R. Day ahead solar forecast using long short term memory network augmented with Fast Fourier transform-assisted decomposition technique. Renewable Energy. 2025; 247. https://doi.org/10.1016/j.renene.2025.123021

[10] Pattnaik SR, Bisoi R, Dash PK. Solar Irradiance Forecasting using Hybrid Long-Short-Term-Memory based Recurrent Ensemble Deep Random Vector Functional Link Network. Computers and Electrical Engineering. 2025; 123. https://doi.org/10.1016/j.compeleceng.2025.110174

[11] Wang Y, Yan G, Xiao S, Ren M, Cheng L, Zhu Z. Day-ahead solar irradiance prediction based on multi-feature perspective clustering. Energy. 2025; 320. https://doi.org/10.1016/j.energy.2025.135216

[12] Jeon HS, Yeon SH, Park JK, Kim MH, Yoon Y, Kim CH, *et al*. ANN based solar thermal energy forecasting model and its heating energy saving effect through thermal storage. Applied Thermal Engineering. 2025; 267. https://doi.org/10.1016/j.applthermaleng.2025.125740

[13] Alizamir M, Shiri J, Fard AF, Kim S, Gorgij AD, Heddam S, *et al*. Improving the accuracy of daily solar radiation prediction by climatic data using an efficient hybrid deep learning model: Long short-term memory (LSTM) network coupled with wavelet transform. Engineering Applications of Artificial Intelligence. 2023; 123. https://doi.org/10.1016/j.engappai.2023.106199

[14] Abolarin SM, Shitta MB, Aghogho M, Emmanuel, Nwosu BP, Aninyem MC, *et al*. An impact of solar PV specifications on module peak power and number of modules: A case study of a five-bedroom residential duplex. IOP Conf Series: Earth and Environmental Science. 2022; 983. https://doi.org/10.1088/1755-1315/983/1/012056

[15] Kurniawan A, Shintaku E. Estimation of Hourly Solar Radiations on Horizontal Surface from Daily Average Solar Radiations Using Artificial Neural Network. Internation Journal on Advanced Science Engineering Infomation Technology. 2022; 12. https://doi.org/10.18517/ijaseit.12.6.12940

https://doi.org/10.31875/2410-2199.2025.12.08