Advancing Prediction of Solar Irradiation using Hybrid XGBoost-LSTM Residual Learning

Ririn Andriyani[1*], Wulan Kusuma Wardani[2], Putty Yunesti[2]

[1]Department of Atmospheric and Planetary Science, Institut Teknologi Sumatera, Lampung Selatan, Indonesia
[2]Department of Energi System Engineering, Institut Teknologi Sumatera, Lampung Selatan, Indonesia
[*] Corresponding author:
Email: ririn.andriyani@sap.itera.ac.id

Abstract.
Accurate solar irradiation prediction is crucial for optimizing solar energy generation and supporting energy management systems. This study addresses the challenges of advancing the accuracy of solar irradiance forecasting by developing a Hybrid XGBoost-LSTM residual learning model, using historical solar radiation data from Lampung Selatan. The results highlight the capability of the Hybrid XGBoost-LSTM model as a powerful tool for forecasting solar irradiation, providing a more precise and reliable solution compared to standalone methods. This is demonstrated by the improvement in R^2 values from the standalone XGBoost predictions, increasing from 0.62 to 0.87 for the training data and from 0.50 to 0.77 on the testing datasets. In addition, the hybrid model demonstrates a decrease in RMSE and MAPE values when compared to standalone XGBoost model, with RMSE and MAPE dropping from 0.26 and 4.19% to 0.14 and 2.39% for the training dataset, and 0.32 and 5.28% to 0.23 and 3.70% for the testing dataset. These findings indicate that the inclusion of LSTM improves the model's ability to refine the residuals from XGBoost, capturing intricate temporal dynamics and fluctuations in solar irradiance data, making it a more reliable and effective tool for predicting solar energy potential in complex environments.
Keywords: *Solar Irradiation; Prediction; XGBoost; LSTM; Residual Learning.*

I. INTRODUCTION

The global shift towards renewable energy sources has emerged as a crucial strategy to address climate change challenges, enhance energy security, and encourage sustainable growth [1]. Solar energy, as one of the most abundant and clean energy sources [2], plays a crucial role in this transition. Solar energy systems, commonly known as photovoltaic systems, harness the intensity of radiation emitted by the sun dan thus heavily depend on atmospheric and environmental conditions, such as weather, cloud cover, humidity, and air pollution [3]. The nonlinear dynamics of atmospheric and environmental conditions can significantly influence solar radiation patterns. Solar radiation data exhibits complex nonlinear patterns and temporal dependencies influenced by various meteorological factors [4], including solar irradiance, atmospheric temperature, humidity, wind velocity, pressure, and cloud cover. These factors fluctuate over time, causing solar radiation levels to change dynamically, making accurate forecasting a challenging task [5]. Accurate prediction of solar radiation intensity is crucial for enhancing the efficiency and seamless integration of solar power systems within the energy grid, especially at the local scale. The conditions at a specific location, influenced by surrounding buildings, trees, and other nearby objects, can affect the pattern of solar radiation received at that site [6]. This poses a challenge in installing solar panels on buildings, as the intensity and distribution of sunlight can vary greatly depending on the surrounding environment.

Achieving accurate predictions of solar radiation remains a worldwide challenge, prompting researchers to utilize various modeling techniques for forecasting [7]. Conventional methods for solar radiation prediction, such as empirical models and dynamical approaches, have long been used to estimate solar irradiance. These empirical models are typically developed using parameters that correlate with solar irradiance, such as sunshine duration-first developed by Ångström in 1924 (Angstrom, 1924) and subsequently refined by numerous studies [8], [9], [10], [11], [12], [13], [14], [15], [16], air temperature [17], [18], [19], [20], cloudliness [21], [22], and combinations of these along with other related factors [23], [24]. Meanwhile, the dynamic approach in solar energy prediction is generally applied based on fundamental atmospheric physics principles and is used for large-scale forecasting. Therefore, this approach is less effective for predictions on smaller or local scales, especially over short time periods. As an alternative, soft

computing techniques are employed to model nonlinear interactions between input and output variables, making them more suitable for solar energy forecasting at local scales and shorter time frames [7].

In recent years, machine learning (ML) techniques have been widely used to overcome the limitations of conventional methods. As a field within artificial intelligence (AI), ML enables systems to identify complex nonlinear patterns between input and output data by learning directly from datasets, without requiring explicit instructions [25]. ML functions by capturing the stochastic dependency between the past and future, helping to decrease the computational complexity [7]. Technically, ML integrates database technologies with statistical methods to identify correlations, uncover hidden patterns, and extract valuable information from large datasets. Two main approaches for modeling and forecasting: the traditional technique, which depends on one independent model, and the advanced machine learning algorithms, such as boosting, bagging, and random forest [26]. One popular and effective ensemble learning method that belongs to the boosting family is Extreme Gradient Boosting (XGBoost). This method is renowned for its high efficiency, scalability, and predictive accuracy [27]. The algorithm enhances traditional gradient boosting by incorporating several advanced features that improve both model performance and computational speed [28]. Several studies have been used XGBoost for solar radiation predictions [29], [30], [31], [32] and demonstrating high accuracy for modeling complex feature interactions in structure data. Meanwhile, the LSTM (Long Short-Term Memory) method, a branch of deep learning, can learn temporal dependencies in sequential data. This method was introduced to address the limitations of Recurrent Neural Networks (RNN) in managing long-term dependency problems [33]. Since weather prediction involves complex temporal relationships in which historical conditions influence future states, LSTMs have been widely adopted owing to their capability to model long-term dependencies [34]. Numerous studies have demonstrated the effectiveness of LSTM-based models for solar radiation forecasting across different climatic conditions and prediction horizons [25], [35], [36].

Building on the strengths of both XGBoost and LSTM, this research proposes a hybrid XGBoost and LSTM through residual learning for advancing solar radiation forecasting. In this framework, XGBoost is employed to capture and predict the primary patterns and relationships present in the meteorological data. Subsequently, the LSTM model focuses on learning and correcting the residual errors left by the XGBoost predictions, especially excelling in modeling the nonlinear and sequential dependencies inherent in solar radiation series. The hybrid approach aims to enhance prediction accuracy by leveraging both algorithms' strengths, especially in addressing the challenges posed by the nonlinear and temporal nature of solar radiation [37]. This research employs a case study in Lampung Selatan to evaluate the performance of the proposed hybrid model in forecasting solar radiation.

This region is known for its significant solar energy potential and local climatic variability. This focus on Lampung Selatan aligns with broader regional analyses highlighting the province’s prominent role in renewable energy development in Sumatera. A recent study employing the Inverse Distance Weighting (IDW) interpolation technique identified Lampung Province as having the highest renewable energy potential in Sumatera, with an average solar capacity of 1416.676 kWh for a 1 kW power plant, surpassing other provinces such as the Riau Islands and North Sumatra [38]. The increasing need for renewable energy over almost four decades highlights the strategic importance of accurate solar radiation forecasting in this area. By utilizing historical meteorological variables specific to the area—such as sunshine duration, humidity, temperature, and wind speed—the research demonstrates the model’s capability to provide accurate and reliable solar radiation forecasts. Utilizing meteorological variables-related factors to forecast solar radiation is a recognized method since solar radiation reaching the ground is strongly linked to these variables.

II. RESEARCH METHODOLOGY

Study Area and Data Exploration

This study is conducted in Lampung Selatan, a regency located in the southern part of Lampung Province, Indonesia. Geographically, Lampung Selatan has a unique characteristic as it directly borders the Sunda Strait to the east and the Indian Ocean to the south. These coastal boundaries position Lampung

Selatan as a region significantly influenced by maritime factors, which affect its local climate and weather patterns. Additionally, this area is located near the equator, so the maximum possible sunshine duration typically reaches around 12 hours near the equinoxes (March and September), when day and night are approximately equal worldwide. During other periods, the duration may vary slightly due to the Earth's axial tilt, but it remains close to 12 hours daily at equatorial latitudes throughout the year.

The dataset for this research comprises two primary components: solar irradiance data, which serves as the target variable to be predicted, and meteorological data, which function as predictor variables. The solar irradiance data are obtained from NASA's Prediction of Worldwide Energy Resources (POWER) satellite database. This satellite-derived data serves as a reliable and comprehensive source for measuring solar irradiance at the spesific location, especially in areas where ground-level radiation observations are scarce or nonexistent. The meteorological data are sourced from BMKG (Indonesia's Meteorological, Climatological, and Geophysical Agency), specifically from the Radin Inten Meteorological Station located in Lampung Selatan. These data include essential atmospheric variables such as sunshine duration, temperature, humidity, precipitation, and wind speed. By integrating satellite solar irradiance data with weather information gathered locally, the research intended to create a robust predictive model for solar radiation that accurately reflects both regional atmospheric conditions and the potential for solar energy. The data used in this study consists of monthly observations spanning the period from January 1993 to June 2024. All variables, except for precipitation, represent the average of daily measurements within each month. Precipitation data, however, is recorded as the total accumulation over the entire month. A detailed description of the dataset, along with the variable names used in the modeling process, is provided in Table 1.

Table 1. Description of meteorological data set

Parameter	Dataset Name	Unit
Solar irradiance	Irradiance	kW-h/m^2
Sunshine duration	Sun_Duration	Hour
Average temperature	Temp_avg	°C
Minimum temperature	Temp_min	°C
Maximum temperatur	Temp_max	°C
Relative humidity	Humidity	%/
Precipitation	Prec	Mm
Average wind speed	Wind_avg	m/s
Maximum wind speed	Wind_max	m/s
Wind direction	wind_direct	°

Any clearly erroneous or inconsistent data in the datasets such as impossible humidity values or non-numeric text in numeric fields—are identified and addressed using interpolation methods. Interpolation is commonly applied to fill missing or problematic values by estimating them based on surrounding data, enhancing its effectiveness for time series data due to its consideration of temporal patterns (Picornell et al., 2021). Additionally, incomplete or noisy data are cleaned to prevent bias during model training.

Predictor Selection with Correlation Analy An initial step was taken to evaluate the strength and direction

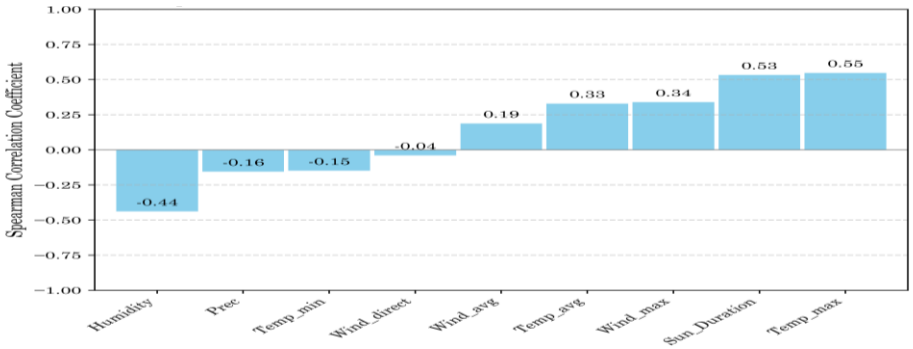

Fig. 1. The Spearman's correlation rank between solar radiation and other meteorological parameters

of the relationship between each meteorological parameter and solar irradiance by performing correlation analysis using Spearman's method. The results of this correlation analysis were subsequently used to identify which meteorological parameters have the strongest association with solar radiation. By determining these

key parameters, the study ensures that only the most relevant predictors are incorporated into the modeling process, thereby improving the accuracy and efficiency of solar radiation forecasting.

Spearman's rank correlation coefficient is a nonparametric measure that evaluates the monotonic relationship between variables by considering their ranked values rather than their raw data. This makes it suitable for detecting consistent increasing or decreasing relationships, even if they are not linear [39]. The relationship can be characterized as either negative or positive on one side, and as either weak or strong on the other. The coefficient value ranges between -1 and +1. The value of +1 signifies a flawless positive monotonic connection, indicating that when one variable rises, the other variable also steadily rises. On the other hand, the value of -1 indicates a perfect negative monotonic relationship, meaning that an increase in one variable is consistently associated with a decrease in the other. A coefficient close to zero implies little to no monotonic relationship between the variables [39]. This flexibility allows Spearman's correlation to detect nonlinear associations, which are common in meteorological datasets.

Spearman's correlation rank between solar radiation and other meteorological parameters was used as a criterion to identify the most influential meteorological variables affecting solar irradiation, as shown in Figure 1. The graph clearly shows that sun duration (Sun_Duration) and maximum temperature (Temp_max) have the strongest positive correlations with irradiance, with correlation coefficients of approximately 0.53 and 0.55, respectively. This indicates that longer periods of sunlight and higher peak temperatures are strongly associated with higher irradiance values. This finding is understandable since irradiance indicates the strength of solar radiation that arrives at the Earth's surface. Therefore, longer periods of sunlight and elevated maximum temperatures directly lead to a greater amount of solar energy being received. This is consistent with results from a study [40], which indicates that air temperature exhibits a similar pattern, with lower temperatures in winter and higher temperatures in summer. Additionally, research [41] confirms that higher solar radiation is associated with longer sunshine durations. In contrast, humidity shows a notable negative correlation with irradiance at approximately -0.44, suggesting that higher humidity levels are linked to lower solar irradiance. This is reasonable because a higher concentration of moisture in the form of water vapor results in increased absorption and backscattering of solar radiation, which ultimately decreases the quantity of solar radiation that reaches the Earth's surface [4]. Additionally, increased moisture in the air often leads to cloud formation or fog, which obstructs sunlight from reaching the ground [42].

Other variables such as average temperature (Temp_avg) and maximum wind speed (Wind_max) exhibit moderate positive correlations with irradiance, with values around 0.33 and 0.34. The average temperature is determined based on the temperature throughout the entire day, including at nighttime hours when the sunlight is absent. Consequently, the average temperature is also affected by various factors including air quality and humidity levels. Wind speed can affect the movement of dust and clouds. The faster the wind, the more it enhances the transmission of solar radiation [4]. Conversely, it might lead the sky to become darker, reducing the quantity of solar radiation passing through the atmosphere [43]. Therefore, the relationship between these two parameters and solar radiation can be considered quite complex. Meanwhile, Precipitation (Prec), Minimum Temperature (Temp_min), and Wind Direction (Wind_direct) show the lowest correlations with solar irradiance. The increase in cloud cover and humidity caused by precipitation can lead to a decrease in the amount of solar radiation reaching the Earth's surface, but its impact on irradiation is inconsistent. Precipitation is often sporadic and only significantly blocks solar radiation during heavy or sustained rainfall. Minimum temperature occurs at night, before sunrise, reflecting cooler temperatures due to heat loss. Since solar radiation is absent during the night, minimum temperature is affected by factors like humidity and atmospheric conditions, which are not directly linked to sunlight. At the same time, while wind can influence weather patterns, the connection between wind direction and solar radiation is restricted.

Based on the Spearman Rank Correlation results that shown in Figure 1, the study selected three key parameters as predictors of solar radiation: sun duration, maximum temperature, and humidity. These three parameters show significant correlations with solar radiation, both positive and negative. Sun duration and maximum temperature have a strong positive correlation, while humidity is negatively correlated with solar

radiation. By selecting three parameters, the research developed a more accurate model for predicting solar radiation, which can be beneficial for applications like solar energy development and climate analysis.

Algorithm Implementation and Training

1. Data Preprocessing

Data preprocessing is an essential step that must be completed prior to the start of the modeling process. This step is regarded as one of the essential elements contributing to the success of ML models. Several issues in data quality can make it difficult for algorithms to extract meaningful information, which in turn impacts the model's performance. Data preprocessing transforms or encodes the data into a format that allows the machine to efficiently process and analyze it. In other words, this step ensures that the model can promptly analyze the features of the data [44]. Since missing values and outliers had already been handled in Section 2.1 and feature selection has already been conducted in Section 2.2, this step focused primarily on the following:

a. Splitting the Data: The dataset is divided into training and testing subsets. 80% of the data is allocated for training the model, while the remaining 20% is designated for testing. This guarantees that the model is assessed using other data, enabling to measure its capacity to generalize.

b. Scaling Features and Target using StandarScaler normalization methods. This step is required when the dataset includes features with varying scales. Normalization is carried out to equalize all features, ensuring that no feature dominates the learning process due to its scale (Maharana et al., 2022). The StandardScaler uses z-score normalization for transforming normal variants to standard score format. This transformation yields datasets that have a mean of 0 and a standard deviation of 1, as illustrated in the equation below:

$$y_i = z = \frac{y_i - \bar{y}_i}{\sigma} \quad (1)$$

Where $\bar{y}_i$, and σ representing the mean and standar deviation of each data point (y_i) [45].

2. XGBoost Implementation

XGBoost is an effective and widely used ML method based on tree boosting that was introduced by Chen and Guestrin [28]. Tree boosting involves combining multiple decision trees sequentially to improve the model's accuracy by correcting errors from previous trees. XGBoost, specifically, is a highly scalable and efficient system designed for large datasets. It offers an end-to-end solution, handling everything from data preprocessing to model evaluation. Decision tree-based optimization techniques in XGBoost are built upon the gradient descent method. The approach is used to optimize the loss function, and regularization parameters are used to prevent overfitting [28]. The boosting method refers to the process of training models sequentially where the outcomes from each weak learner's training affect the training of the subsequent model. The predicted value/output of XGBoost is represented in the equation below [46]:

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i),\ f_t \in F \quad (2)$$

Where:

$\hat{y}_i^{(t)}$ is the final predicted value

t is the total number of trees in the model

$f_k(x_i)$ is the prediction made by the k-th tree for the input x_i

$\hat{y}_i^{(t-1)}$ is the predicted value at the prior iteration

f_t is t-th regression trees

F denotes the collection of all regression trees.

The complete XGBoost prediction equation combines both the loss function and the regularization term to optimize the model during training, which is known as the objective function $(Obj(\theta))$. This study used reg:squarederror as the objective function which is shown in the following equation:

$$Obj(\theta) = \sum_{i=1}^{N}[(\hat{y}_i - y_i)^2] + \left(\gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2\right) \quad (3)$$

Where:

N is number of training data samples

γ is a regularization parameter controlling the complexity of the model
T is the number of leaves (or nodes) in the tree
w_j is the weight associated with leaf j
λ is the regularization parameter for the tree's weights.

The XGBoost model was configured with the following parameters: the objective function was set to "reg:squarederror" for regression tasks, the learning rate was 0.01 to control the step size in minimizing the loss, the maximum depth of trees was set to 3, and 300 trees were built (N_estimators). After training, the XGBoost model generated predicted solar irradiance values for both the training and test datasets. The forecasts were subsequently compared to the real values to determine the residuals, indicating the variations between the expected and actual irradiance. The residuals represented the portion of the variance in solar irradiance predictions that could not be captured by XGBoost and were passed to the LSTM model for further refinement. Before being input into the LSTM model, the residuals were scaled using StandardScaler to standardize the data.

3. LSTM Implementation

LSTM (Long Short-Term Memory) networks were developed to overcome the challenges faced by traditional Recurrent Neural Network (RNN) in managing long-term dependencies in sequential data [33]. While RNNs have a hidden state updated at each time step and can process sequential information, they struggle to retain long-term memory due to the vanishing gradient problem. LSTMs improve upon this by introducing additional gates and memory units that control the flow of information, allowing them to effectively remember and process longer sequences. This architecture allows LSTMs to arrest more complex temporal patterns and long-term dependencies, making them appropriate for time series analysis and other tasks involving sequential data [47].

An LSTM unit consists of four main components, each playing a critical role in managing the flow of information over time. The first three components, the forget gate, input gate, and output gate, are responsible for selecting which information is relevant at each time step. The Forget Gate determines which information should be removed from the memory cell, utilizing a sigmoid function to generate values ranging from 0 to 1. The input Gate identifies the new information that should be incorporated into the memory cell, also outputting a value between 0 and 1. The output Gate regulates the information to be output from the memory cell, allowing the model to utilize relevant information for subsequent predictions. The fourth component, known as the memory cell, retained information to be added to the cell state [48].

Figure 2 shows the working of an LSTM, that begins with the input data (x_t) and the hidden state (h_{t-1}) from the previous step, which are carried by the cell state and processed by all the gates. The forget gate (f_t) first decides which information should be retained or discarded. This is processed using the sigmoid function (σ), producing values between 0 and 1. The input gate (i_t) then receives the information to be combined and processed into the cell state, also using the sigmoid function (σ), which generates values between 0 and 1 to control how much new information will be added to the memory. The candidate cell $(\tilde{C}_t)$ generates new information that can be inserted into memory using the $tanh$ function, producing values within the range of -1 to 1. The cell state is updated by combining the information retained by the forget gate, new information from the input gate, and the candidate memory. Next, the output gate (o_t) selects which part of the cell state will be output at a given time. The sigmoid function (σ) is utilized again to generate values ranging from 0 and 1, which regulates the amount of cell state information that will be forwarded as the output. The result of the output gate is subsequently utilized as the hidden state (h_t), which will be passed to the next phase in the LSTM or serve as the final output of the model.

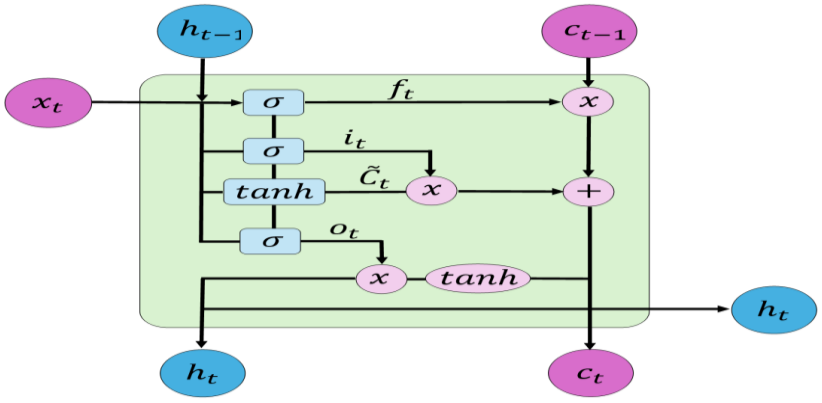

Fig. 2. LSTM architecture

The objective of this section is to combine the predictive power of XGBoost with the sequential learning ability of LSTM networks to create a hybrid model that improves the accuracy of solar radiation prediction. This hybrid approach leverages the strengths of both models: XGBoost is effective at capturing non-linear relationships and handling structured data, while LSTM networks are capable of modeling time dependencies in residuals that can improve predictions over time. LSTM model was configured with a lookback length of 12, meaning the model used the past 12 time steps to predict future residuals. The batch size was set to 32, determining how many samples are processed before updating the model's weights. The LSTM model is composed of two layers: Both LSTM layers with units are set based on hyperparameter tuning, followed by a Dropout layer to prevent overfitting. The model was assembled utilizing the Adam optimizer with a learning rate set through a search for optimal hyperparameters, alongside the mean squared error as the loss function. Early stopping was utilized to monitor validation loss, halting the training if no improvement occurred after 10 epochs. To optimize hyperparameters, Keras Tuner was utilized to identify the most suitable architecture for the LSTM model by conducting a random search on parameters including the quantity of LSTM units, dropout rate, and learning rate.

4. Metric Evaluation

Following the training of the models, their effectiveness was assessed on the test set through various established metrics to evaluate the precision of the predictions. These metrics consist of R^2 (Coefficient of Determination), MAE (Mean Absolute Error), and RMSE (Root Mean Squared Error), with each offering distinct perspectives on the model's ability to make predictions. The R^2 value measures the proportion of the variance in the target variable that is explained by the model [49]. An R^2 value close to 1 suggests that the model accounts for a significant portion of the variance, indicating that the model's predictions are in close agreement with the real values. RMSE gives an estimate of the standard deviation of the residuals (errors in prediction), indicating how much the errors vary, offering a sense of how spread out the errors are. Unlike RMSE, MAE treats all errors equally, without penalizing larger errors more than smaller ones. This makes MAE a useful metric when it is important to assess how far off the predictions are on average [50]. Mathematically, R^2, RMSE, and MAE are calculated as:

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(x_i - y_i)^2}{\sum_{i=1}^{m}(\bar{y} - y_i)^2} \quad (4)$$

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(x_i - y_i)^2} \quad (5)$$

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|x_i - y_i| \quad (6)$$

Where:

x_i is the actual values

$\bar{y}$ as mean of the actual values

y_i is the predicted values

m as number of data points

III. RESULT AND DISCUSSION

The prediction performance on the training and testing datasets, based on evaluation metric results from the XGBoost, Hybrid XGBoost-LSTM, Random Forest (RF), and Support Vector Regression (SVR) methods, is shown in Table 2. The performance results of the standalone XGBoost model are presented as the primary comparison, and the two other methods, namely RF and SVR, are used as additional comparisons to assess the effectiveness of the hybrid XGBoost-LSTM model in predicting solar irradiation.

Tabel 2. Evaluation of metrics values

Methods	Training Dataset			Testing Dataset		
	R^2	RMSE	MAPE	R^2	RMSE	MAPE
XGBoost	0.62	0.26	4.19%	0.50	0.32	5.28%
Hybrid XGBoost-LSTM	0.87	0.14	2.38%	0.77	0.23	3.70%
RF	0.75	0.21	3.41%	0.51	0.31	5.12%
SVR	0.52	0.30	4.64%	0.37	0.33	5.31%

As an additional comparison, RF and SVR yield lower results on both the training and testing datasets. While RF performs relatively well in capturing data relationships, showing an R^2 of 0.75 on the training data, it experiences a significant drop on the testing data, with R^2 falling to only 0.51. Higher MAPE and RMSE values on the testing data also suggest that RF struggles with generalization compared to XGBoost or the Hybrid XGBoost-LSTM model. SVR, on the other hand, shows even poorer performance, with an R^2 of just 0.37 on the testing data and a higher MAPE of 5.31%, reflecting the model's difficulty in capturing the data's complexity.

The standalone XGBoost model demonstrates more reliable performance compared to the RF and SVR methods. XGBoost achieves an R^2 of 0.62 on the training data, indicating that the model explains around 62% of the variance in the data, and 0.50 on the testing data, capturing half of the data variation. While XGBoost's R^2 value is lower than that of RF on both training and testing data, the difference between the training and testing values suggests that XGBoost is more stable and reliable in generalizing predictions. This is also reflected in its lower RMSE and MAPE values.

On the other hand, the Hybrid XGBoost-LSTM residual learning approach significantly improves prediction accuracy. The addition of the LSTM layer to process residual data for further prediction proved to provide more reliable results compared to the three standalone methods. The results from this hybrid model show significant improvement, with the R^2 on the training data increasing drastically to 0.87, and much lower RMSE (0.14) and excellent MAPE (2.38%). Even on the testing data, the model still maintains a very good R^2 of 0.77, with low RMSE (0.23) and MAPE of 3.70%. The substantial increase in R^2 and the decrease in RMSE and MAPE indicate that this hybrid approach effectively addresses the limitations of XGBoost when working independently. Therefore, while XGBoost performs well on its own, the Hybrid XGBoost-LSTM model shows a significant improvement in solar irradiation prediction accuracy, both on the training and testing data. The use of LSTM to capture temporal patterns that XGBoost alone struggles to handle makes this approach more effective in addressing the challenges in solar irradiation prediction.

Figure 3 illustrates a comparison between the actual solar irradiance values (blue line) and the predicted values (green dashed line for XGBoost and red dashed line for the Hybrid XGBoost-LSTM model). Both models perform well in following the overall trend of the data, but there are notable differences in the level of accuracy and their ability to respond to rapid changes in the data. The XGBoost model (green dashed line) tends to provide more stable and smoother predictions. While it captures the general patterns of the solar irradiance, it shows larger deviations during sudden spikes or drops in the data. This model appears slower to react to significant changes, which suggests that XGBoost might have limitations in detecting and responding to rapid fluctuations. This could be due to its inability to fully capture complex, non-linear relationships within the data, as it relies on tree-based methods that perform well with simpler, more predictable trends.

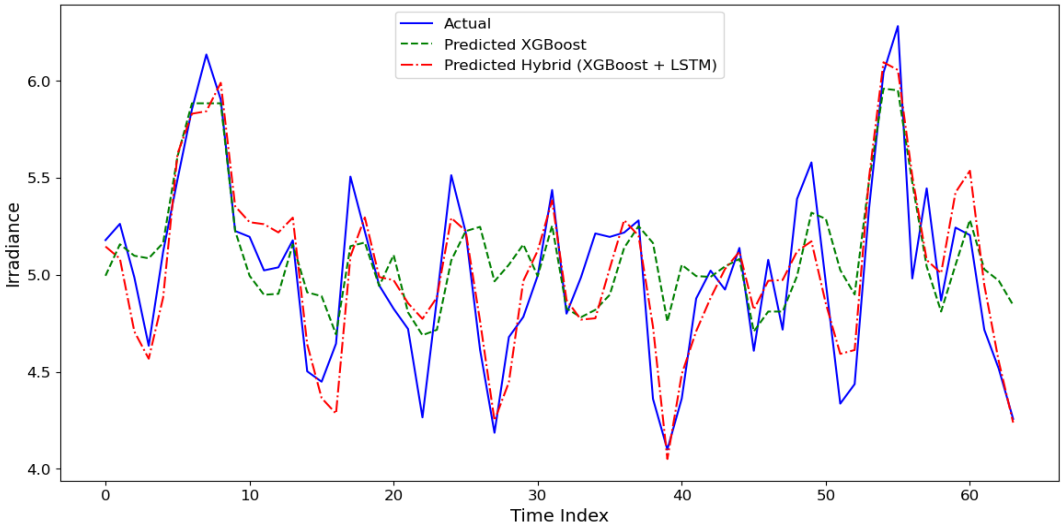

Fig. 3. Predicted versus actual solar radiation by the standalone XGBoost model (green dashed line) and the Hybrid XGBoost-LSTM Residual Learning Model (red dashed line)

On the other hand, the Hybrid XGBoost-LSTM model (red dashed line) demonstrates better performance, especially in capturing rapid and complex fluctuations in the data. The predictions from the hybrid model align more closely with the actual solar irradiance values, particularly during moments of sharp changes. This shows that the LSTM component of the hybrid model is more adaptive and sensitive to non-linear patterns in the data. The LSTM model's ability to learn from sequential data allows it to detect and react to dynamic changes more effectively than XGBoost alone. Overall, the Hybrid XGBoost-LSTM model outperforms XGBoost in predicting solar irradiance, particularly in scenarios involving rapid or complex changes in the data. However, the choice of the model should depend on the specific needs of the task. XGBoost may be a more efficient option when the goal is to prioritize prediction stability and faster computations, while the Hybrid XGBoost-LSTM model is more suitable for cases that require deeper modeling of non-linear and sequential patterns in the data.

IV. CONCLUSION

In this study, the Hybrid XGBoost-LSTM model was employed to forecast solar irradiance using the most influential historical meteorological data. The most influential features, including maximum temperature, humidity, and sun duration, were selected through Spearman's correlation analysis. The results demonstrated that this hybrid approach effectively captures both linear and non-linear patterns in the solar irradiance data, outperforming other methods. The hybrid model achieved the highest R^2 values on both the training and testing data compared to the three standalone methods (XGBoost, RF, and SVR), with 0.87 on the training data and 0.77 on the testing data. This indicates its ability to capture both general trends and complex fluctuations in solar irradiance. Similarly, the RMSE and MAPE values were notably lower, with 0.14 and 2.38% on the training set, and 0.23 and 3.70% on the testing set, respectively, highlighting the hybrid model's higher prediction accuracy compared to other models.

The combination of XGBoost and LSTM proved to enhance prediction accuracy compared to using XGBoost alone, with XGBoost handling the general trends in the data while LSTM captures the temporal dynamics of the residuals, i.e., the errors in the initial predictions. This improvement is clearly seen in the increase in R^2 values on both the training and testing data in the hybrid model compared to the standalone XGBoost model. Specifically, the R^2 increased from 0.62 for XGBoost on the training set to 0.87 in the hybrid model, and from 0.50 on the testing set for XGBoost to 0.77 in the hybrid model, showing a significant improvement in the model's ability to generalize and predict solar irradiance more accurately.

This study also highlights that the hybrid approach is more reliable than standalone methods when dealing with complex datasets, such as solar irradiance data, which involves intricate time-dependent

relationships. When faced with intricate, high-dimensional, or non-linear data, using a single model might not be sufficient to fully capture the information of the data. This is especially true for solar irradiance, where variations can be influenced by multiple factors like weather conditions, time of day, and seasonal changes. With the presence of LSTM for learning from the residuals, the prediction values that were initially provided by XGBoost can be further refined by capturing the temporal patterns and fluctuations that XGBoost may not fully account for.

Despite the superior performance of the Hybrid XGBoost-LSTM model compared to other standalone methods, prediction accuracy still has potential for further enhancement. Given that this study has already used an advanced and reliable hybrid method, prediction accuracy could potentially be enhanced by incorporating a wider range of predictors, such as factors related to environmental activities, or by using a longer time span, allowing the model to learn more about the data trends.

REFERENCES

[1] A. Q. Al-Shetwi, I. Z. Abidin, K. A. Mahafzah, and M. A. Hannan, "Feasibility of future transition to 100% renewable energy: Recent progress, policies, challenges, and perspectives***," J. Clean. Prod***., vol. 478, p. 143942, Nov. 2024, doi: 10.1016/j.jclepro.2024.143942.

[2] N. T.R., P. V, S. A. Sridharan, and N. Ramrao, "Harnessing Renewable Energy at Kalasalingam Academy of Research and Education – A Role Model Case," ***Journal of Sustainability Perspectives***, vol. 1, no. 1, pp. 7–13, Jun. 2021, doi: 10.14710/jsp.2021.11201.

[3] E. H. Sepúlveda-Oviedo, "Impact of environmental factors on photovoltaic system performance degradation," Energy Strategy Reviews, vol. 59, p. 101682, May 2025, doi: 10.1016/j.esr.2025.101682.

[4] G. Zhang et al., "Solar radiation estimation in different climates with meteorological variables using Bayesian model averaging and new soft computing models," Energy Reports, vol. 7, pp. 8973–8996, Nov. 2021, doi: 10.1016/j.egyr.2021.10.117.

[5] N. Krishnan, K. R. Kumar, and C. S. Inda, "How solar radiation forecasting impacts the utilization of solar energy: A critical review," J. Clean. Prod., vol. 388, p. 135860, Feb. 2023, doi: 10.1016/j.jclepro.2023.135860.

[6] J. Tian and R. Ooka, "Prediction of building-scale solar energy potential in urban environment based on parametric modelling and machine learning algorithms," Sustain. Cities Soc., vol. 119, p. 106057, Feb. 2025, doi: 10.1016/j.scs.2024.106057.

[7] V. Demir, "Evaluation of Solar Radiation Prediction Models Using AI: A Performance Comparison in the High-Potential Region of Konya, Türkiye," Atmosphere (Basel)., vol. 16, no. 4, Apr. 2025, doi: 10.3390/atmos16040398.

[8] L. Achour, M. Bouharkat, O. Assas, and O. Behar, "Hybrid model for estimating monthly global solar radiation for the Southern of Algeria: (Case study: Tamanrasset, Algeria)," Energy, vol. 135, pp. 526–539, Sep. 2017, doi: 10.1016/j.energy.2017.06.155.

[9] J. Almorox and C. Hontoria, "Global solar radiation estimation using sunshine duration in Spain," Energy Convers. Manag., vol. 45, no. 9–10, pp. 1529–1535, Jun. 2004, doi: 10.1016/j.enconman.2003.08.022.

[10] D. B. Ampratwum and A. S. S. Dorvlo, "Estimation of solar radiation from the number of sunshine hours," Appl. Energy, vol. 63, no. 3, pp. 161–167, Jul. 1999, doi: 10.1016/S0306-2619(99)00025-2.

[11] V. Bahel, H. Bakhsh, and R. Srinivasan, "A correlation for estimation of global solar radiation," Energy, vol. 12, no. 2, pp. 131–135, Feb. 1987, doi: 10.1016/0360-5442(87)90117-4.

[12] K. Bakirci, "Correlations for estimation of daily global solar radiation with hours of bright sunshine in Turkey," Energy, vol. 34, no. 4, pp. 485–501, Apr. 2009, doi: 10.1016/***j.energy***.2009.02.005.

[13] A. A. El-Sebaii, A. A. Al-Ghamdi, F. S. Al-Hazmi, and A. S. Faidah, "Estimation of global solar radiation on horizontal surfaces in Jeddah, Saudi Arabia," Energy Policy, vol. 37, no. 9, pp. 3645–3649, Sep. 2009, doi: 10.1016/***j.enpol***.2009.04.038.

[14] M. E. Fernández, J. O. Gentili, and A. M. Campo, "Sunshine Duration Analysis as a First Step to Estimate Solar Resource for Photovoltaic Electricity Production in Middle Latitudes," Environmental Processes, vol. 5, no. 2, pp. 313–328, Jun. 2018, doi: 10.1007/s40710-018-0298-3.

[15] K. Lan, L. Wang, Y. Zhou, Z. Zhang, S. Fang, and P. Cao, "The applicability of sunshine-based global solar radiation models modified with meteorological factors for different climate zones of China," Front. Energy Res., vol. 10, Jan. 2023, doi: 10.3389/fenrg.2022.1010745.

[16] F. J. Newland, "A study of solar radiation models for the coastal region of South China," Solar Energy, vol. 43, no. 4, pp. 227–235, 1989, doi: 10.1016/0038-092X(89)90022-4.

[17] M. A. Ali, A. Elsayed, I. Elkabani, M. E. Youssef, and G. E. Hassan, "Modeling global solar radiation using ambient temperature," Cleaner Energy Systems, vol. 7, p. 100101, Apr. 2024, doi: 10.1016/j.cles.2023.100101.

[18] J. Almorox, M. Bocco, and E. Willington, "Estimation of daily global solar radiation from measured temperatures at Cañada de Luque, Córdoba, Argentina," Renew. Energy, vol. 60, pp. 382–387, Dec. 2013, doi: 10.1016/j.renene.2013.05.033.

[19] J. Fan, B. Chen, L. Wu, F. Zhang, X. Lu, and Y. Xiang, "Evaluation and development of temperature-based empirical models for estimating daily global solar radiation in humid regions," Energy, vol. 144, pp. 903–914, Feb. 2018, doi: 10.1016/j.energy.2017.12.091.

[20] N. Ghazouani et al., "Performance Evaluation of Temperature-Based Global Solar Radiation Models—Case Study: Arar City, KSA," Sustainability, vol. 14, no. 1, p. 35, Dec. 2021, doi: 10.3390/su14010035.

[21] L. Gu et al., "Cloud modulation of surface solar irradiance at a pasture site in southern Brazil," Agric. For. Meteorol., vol. 106, no. 2, pp. 117–129, Jan. 2001, doi: 10.1016/S0168-1923(00)00209-4.

[22] A. Radovan, V. Šunde, D. Kučak, and Ž. Ban, "Solar Irradiance Forecast Based on Cloud Movement Prediction," Energies (Basel)., vol. 14, no. 13, p. 3775, Jun. 2021, doi: 10.3390/en14133775.

[23] S. Mujabar and R. Chintaginjala Venkateswara, "Empirical models for estimating the global solar radiation of Jubail Industrial City, the Kingdom of Saudi Arabia," SN Appl. Sci., vol. 3, no. 1, Jan. 2021, doi: 10.1007/s42452-020-04043-9.

[24] A. Vernet and A. Fabregat, "Evaluation of Empirical Daily Solar Radiation Models for the Northeast Coast of the Iberian Peninsula," Energies (Basel)., vol. 16, no. 6, p. 2560, Mar. 2023, doi: 10.3390/en16062560.

[25] X. Qing and Y. Niu, "Hourly day-ahead solar irradiance prediction using weather forecasts by LSTM," Energy, vol. 148, pp. 461–468, Apr. 2018, doi: 10.1016/j.energy.2018.01.177.

[26] A. Aldrees, H. H. Awan, M. F. Javed, and A. M. Mohamed, "Prediction of water quality indexes with ensemble learners: Bagging and boosting," Process Safety and Environmental Protection, vol. 168, pp. 344–361, Dec. 2022, doi: 10.1016/j.psep.2022.10.005.

[27] S. Hakkal and A. A. Lahcen, "XGBoost To Enhance Learner Performance Prediction," Computers and Education: Artificial Intelligence, vol. 7, p. 100254, Dec. 2024, doi: 10.1016/j.caeai.2024.100254.

[28] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[29] Y. Kim and Y. Byun, "Predicting Solar Power Generation from Direction and Tilt Using Machine Learning XGBoost Regression," J. Phys. Conf. Ser., vol. 2261, no. 1, p. 012003, Jun. 2022, doi: 10.1088/1742-6596/2261/1/012003.

[30] X. Li et al., "Probabilistic solar irradiance forecasting based on XGBoost," Energy Reports, vol. 8, pp. 1087–1095, Aug. 2022, doi: 10.1016/j.egyr.2022.02.251.

[31] Q.-T. Phan, Y.-K. Wu, and Q.-D. Phan, "Short-term Solar Power Forecasting Using XGBoost with Numerical Weather Prediction," in 2021 IEEE International Future Energy Electronics Conference (IFEEC), IEEE, Nov. 2021, pp. 1–6. doi: 10.1109/IFEEC53238.2021.9661874.

[32] C. Zhang, Y. Zhang, J. Pu, Z. Liu, Z. Wang, and L. Wang, "An hourly solar radiation prediction model using eXtreme gradient boosting algorithm with the effect of fog-haze," Energy and Built Environment, vol. 6, no. 1, pp. 18–26, Feb. 2025, doi: 10.1016/j.enbenv.2023.08.001.

[33] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Comput., vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[34] J. Xu, Z. Wang, X. Li, Z. Li, and Z. Li, "Prediction of Daily Climate Using Long Short-Term Memory (LSTM) Model," International Journal of Innovative Science and Research Technology (IJISRT), pp. 83–90, Jul. 2024, doi: 10.38124/ijisrt/ijisrt24jul073.

[35] S. Tajjour, S. S. Chandel, M. A. Alotaibi, H. Malik, F. P. Garcia Marquez, and A. Afthanorhan, "Short-Term Solar Irradiance Forecasting Using Deep Learning Techniques: A Comprehensive Case Study," IEEE Access, vol. 11, pp. 119851–119861, 2023, doi: 10.1109/ACCESS.2023.3325292.

[36] W. K. Wardani and R. Andriyani, "Evaluating the Performance of LSTM in Photovoltaic Energy Optimization using Sunshine Duration Prediction," ***Jurnal Fokus Elektroda***, vol. 10, no. 3, 2025.

[37] Y. Mariappan, K. Ramasamy, and D. Velusamy, "An optimized deep learning based hybrid model for prediction of daily average global solar irradiance using CNN SLSTM architecture," Sci. Rep., vol. 15, no. 1, p. 10761, Mar. 2025, doi: 10.1038/s41598-025-95118-3.

[38] A. Vatresia, E. Lestari, F. Utama, and N. Daratha, "Spatial Interpolation Long-Term Patterns Capacity of Renewable Energy in Sumatera," Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control, Jun. 2024, doi: 10.22219/kinetik.v9i3.1929.

[39] J.-S. Kang, D.-H. Shin, J.-W. Baek, and K. Chung, "Activity Recommendation Model Using Rank Correlation for Chronic Stress Management," Applied Sciences, vol. 9, no. 20, p. 4284, Oct. 2019, doi: 10.3390/app9204284.

[40] A. K. Shrestha, A. Thapa, and H. Gautam, "Solar Radiation, Air Temperature, Relative Humidity, and Dew Point Study: Damak, Jhapa, Nepal," International Journal of Photoenergy, vol. 2019, pp. 1–7, Dec. 2019, doi: 10.1155/2019/8369231.

[41] Q. Cao, Y. Liu, K. Lyu, Y. Yu, D. H. W. Li, and L. Yang, "Solar radiation zoning and daily global radiation models for regions with only surface meteorological measurements in China," Energy Convers. Manag., vol. 225, p. 113447, Dec. 2020, doi: 10.1016/j.enconman.2020.113447.

[42] F. Besharat, A. A. Dehghan, and A. R. Faghih, "Empirical models for estimating global solar radiation: A review and case study," Renewable and Sustainable Energy Reviews, vol. 21, pp. 798–821, May 2013, doi: 10.1016/j.rser.2012.12.043.

[43] P. E. Bett and H. E. Thornton, "The climatological relationships between wind and solar energy supply in Britain," Renew. Energy, vol. 87, pp. 96–110, Mar. 2016, doi: 10.1016/j.renene.2015.10.006.

[44] K. Maharana, S. Mondal, and B. Nemade, "A review: Data pre-processing and data augmentation techniques," Global Transitions Proceedings, vol. 3, no. 1, pp. 91–99, Jun. 2022, doi: 10.1016/j.gltp.2022.04.020.

[45] I. B. Mohamad and D. Usman, "Standardization and Its Effects on K-Means Clustering Algorithm," Research Journal of Applied Sciences, Engineering and Technology, vol. 6, no. 17, pp. 3299–3303, 2013.

[46] D. N. Gono, H. Napitupulu, and Firdaniza, "Silver Price Forecasting Using Extreme Gradient Boosting (XGBoost) Method," Mathematics, vol. 11, no. 18, p. 3813, Sep. 2023, doi: 10.3390/math11183813.

[47] M. Pacella and G. Papadia, "Evaluation of deep learning with long short-term memory networks for time series forecasting in supply chain management," Procedia CIRP, vol. 99, pp. 604–609, 2021, doi: 10.1016/j.procir.2021.03.081.

[48] Ottavio Calzone, "An Intuitive Explanation of LSTM," Medium.

[49] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," PeerJ Comput Sci, vol. 7, pp. 1–24, 2021, doi: 10.7717/PEERJ-CS.623.

[50] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature," Geosci Model Dev, vol. 7, no. 3, pp. 1247–1250, Jun. 2014, doi: 10.5194/gmd-7-1247-2014.

.