Journal of Engineering and Computer Sciences
Qassim University, Vol. 15, No. 2, pp. 39-57 (July 2022)

Towards a Greener Future: Machine Learning Applications in Solar Irradiance Forecasting for Renewable Energy Planning in Saudi Arabia

Talal Alharbi
Department of Electrical Engineering, College of Engineering, Qassim University, Buraydah, Qassim, Saudi Arabia
atalal@qu.edu.sa

Saeed Iqbal
Department of Computer Science, Faculty of Information Technology & Computer Science, University of Central Punjab, Lahore, Pakistan
saeediqbalkhattak@gmail.com

(Received 09/04/2024; accepted for publication 14/05/2024)

Abstract. Renewable energy planning is set to be transformed by the integration of advanced solar irradiance forecasting techniques. By harnessing the predictive power of Machine Learning (ML) algorithms, planners can gain more accurate insights into future solar irradiance levels. This study investigates the use of ML algorithms for solar irradiance forecasting, intending to enhance planning strategies for renewable energy sources (RES) in Saudi Arabia. Using datasets sourced from various regions in Saudi Arabia, several regression models are evaluated, including Gradient Boosting Regressor (GBR), Linear Regression (LR), Long Short-Term Memory (LSTM), Bidirectional Long Short-Term Memory (Bi-LSTM), and K-Nearest Neighbor (KNN). The analysis of this research reveals that ensemble techniques such as Random Forest Regression (RFR) and data-driven approaches like KNN exhibit superior performance compared to conventional regression models like LR, underscoring the significance of various ML methods in solar irradiance prediction When compared to Decision Tree Regressor (DTR) and RFR, models with high goodness of-fit metrics (R-squared, adjusted R-squared) and low error metrics (Mean Absolute Error (MAE), Root Mean Square Error (RMSE)) show better predictive power. The precision with which the proposed models forecast solar irradiance levels is further confirmed by comparison with previous studies. Planning for RES is advanced by this study's identification of the best ML methods for predicting solar irradiation. The results highlight the potential of using ML approaches to optimize solar energy system integration and accelerate the shift to sustainable energy practices.

Keywords: Deep Learning, Machine Learning, Predictive Power, Renewable Energy Planning, Solar Irradiance Forecasting, Solar Energy System.

Nomenclature

ANN	Artificial Neural Network
AR2	Adjusted R-Squared
Bi-LSTM	Bidirectional-Long Short-Term Memory
DT	Decision Tree

DTR	Decision Tree Regressor
EVT	Extreme Value Theory
EVD	Extreme Value Distribution
GBR	Gradient Boosting Regressor
GHI	Global Horizontal Irradiance
GPD	Generalized Pareto Distribution
KNN	K-Nearest Neighbor
LR	Linear Regression
LSTM	Long Short-Term Memory
ML	Machine Learning
MAPE	Mean Absolute Percentage Error
MAE	Mean Absolute Error
MSE	Mean Squared Error
MBE	Median Biased Error
MLP	Multi-Layer Perception
MICS	Multiple Indicator Cluster Survey
NRMSE	Normalized Root Mean Squared Error
SCDA	Smart Urban Demo Aspern
PV	Photovoltaic
RES	Renewable Energy Sources
RFR	Random Forest Regression
RBF	Radial Basis Function
RF	Random Forest
R2	R-Squared
RMSE	Root Mean Square Error
RFE	Recursive Feature Elimination
RFA	Recursive Feature Addition
RNN	Recurrent Neural Network
SVR	Support Vector Regression
SDG	Sustainable Development Goal
SVM	Support Vector Machine
PDF	Probability Density Function
PCA	Principal Component Analysis

WT Wind Turbine

1. Introduction

The significance of precisely projecting future energy needs is highlighted by the rising demand for power brought forth by technological breakthroughs. These projections are essential for deciding on the layout, kind, and capacity of new power plants as well as for maximizing the efficiency of already-existing ones [1]. Furthermore, these forecasts are essential for investors since they allow them to evaluate the possible effect of expected sales on stock values. In the energy industry, they also aid in management and technological planning, guaranteeing the supply of dependable and reasonably priced energy resources. Accurate forecasting also promotes the expansion of RES and supports global efforts to reduce carbon emissions. Notably, a study advocates for the use of a random effect regression model to incentivize investments in renewable energy. Relevant literature further highlights cultural perceptions of energy affordability and regional variations in power demand [2]. The aforementioned results highlight the importance of forecasting in guiding energy policy and investment decisions to meet evolving needs in a sustainable manner [3].

The impact of energy poverty on the development of young children in nations with limited access to energy resources was studied by researchers in [4]. They discovered a direct relationship between energy poverty and early development, which has an impact on things like living standards and healthcare, using data from Multiple Indicator Cluster Surveys (MICSs). The importance of Extreme Value Theory (EVT) in other domains is also highlighted in the study. Through the examination of past data and the fitting of distributions, especially the Extreme Value Distribution (EVD) and Generalized Pareto Distribution (GPD), EVT, a field of statistical analysis, assists in the prediction of energy prices. EVT helps with risk management. and well-informed financial decision-making by enabling the forecast of potentially significant changes in energy costs through the identification of patterns in data.

The authors in [5] suggested using Machine Learning (ML) technology Support Vector Machine (SVM) to anticipate load requirements for different components within a building, such as air conditioners and power, by using weather forecasts and periodic energy demand data. The SVM approach yielded reliable estimates of the total power load, with Median Biased Errors (MBEs) of 7.7% and a RMSE of 15.2%. The K-Nearest Neighbor (KNN) technique was used in another project, the Smart Urban Demo Aspern (SCDA) project, to predict data center power consumption. This required using associated data points and historical measurements (load demand curves) for KNN prediction. Nevertheless, the KNN approach's capacity to accurately forecast future values is constrained by its exclusive dependence on locating comparable occurrences across a wide feature space. As such, it requires corroboration with time-related data identification so that predictions for the next 24 hours can be made during business hours.

For short-term load forecasting, five distinct ML techniques were investigated [6]. These techniques were used after first generating 24 time series, one for each hour of the day, based on historical data. These time series initially represented each hour of the day. Multi-Layer Perception (MLP), SVM, Radial Basis Function (RBF) regressor, Reduced Error Pruning Tree (REPTree), and Poisson process were among the ML techniques used. The Moroccan electrical load data were used in the experiment. As the most accurate method, the Mean

Absolute Percentage Error (MAPE) of 0.96 was obtained from the MLP strategy. In second place, SVM performed better than the other methods even though it did not reach the MLP method's level of accuracy. A ML classification methodology was used [7] to develop and test a strategy for energy usage prediction. Using the techniques for Linear Regression (LR), Support Vector Regression (SVR), Random Forest (RF), Decision Tree (DT), and KNNs, the researchers examined historical data to create a predictive model. The addition of a one-day power usage attribute (kWh) was a new criterion in the study. The LR and SVR models had the highest accuracy rate of 85.7%, according to the data. In addition, there has been a notable advancement in ML, moving from methods of shallow learning to the training of Artificial Neural Networks (ANNs).

The authors [8] highlighted the importance of renewable energy, specifically photovoltaic (PV) energy systems, for Saudi Arabia's future. The investigators [9] performed a survey of 1498 people to investigate the factors that influence the adoption of residential PV systems in Saudi Arabia. The findings raised worries 5 about installation expenses as well as revenue production from PV systems. The researchers [10] published a study based on a survey that showed people's enthusiasm for PV systems for their homes, especially if the government subsidized capital expenditures by 40%.

Analysis of power usage has long piqued the curiosity of data scientists and ML technologists. In addition to introducing ML models for solar energy prediction, this work reviews prior research on power usage predictions. The intention is to support the energy industry in forecasting solar energy generation. Using datasets from Turaif, Qassim, and Majmaah (KSA), several regression models were tested, including elastic net regression, linear regression, random forest, KNN, Gradient Boosting Regressor (GBR), light gradient boosting regressor, extreme gradient boosting regressor, and Decision Tree Regressor (DTR). High accuracy rates are shown in the results, with some algorithms reaching 99%. This is very helpful for sectors that need to estimate production rates. It is determined that Turaif, Qassim, is a better location for solar power plants because of its consistent weather as opposed to Majmaah, which produces good results but has unpredictable weather.

This work advances solar energy forecasting by accurately predicting Global Horizontal Irradiance (GHI) to improve solar energy generation efficiency. By leveraging the predictive power of ML algorithms, planners can obtain more accurate values of the solar energy. This enhanced forecasting capability enables better anticipation of solar energy generation potential, facilitating optimized deployment of solar energy systems and infrastructure. A variety of ML data-driven models, such as Decision Tree (DT) Regressor (DTR), GBR, Light Gradient Boosting Regressor (Light GBR), Extreme Gradient Boosting Regressor (Extreme GBR), RF, K-Nearest Neighbors (KNN), and Elastic Net Regression, are used to evaluate predictive performance. Standard metrics including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), coefficient of determination R-Squared (R2), and adjusted R-squared (AR2) are used to evaluate solar electricity output prediction accuracy. The GBR, LR, Long Short-Term Memory (LSTM), and Bidirectional Long Short-Term Memory (Bi-LSTM) models outperformed Majmaah in the Turaif and Qassim regions. This discovery emphasizes the models' potential relevance for solar energy projects in Saudi Arabia, particularly when regional variations in solar irradiance patterns are taken into account.

The key contributions of this study include:

- Accurate forecasting of GHI is crucial to optimize solar energy generation effectiveness to make informed decisions regarding the sizing, placement, and operation of renewable energy assets.
- The intricate correlations inherent in solar irradiance data are captured using a varied array of data-driven models, ranging from classic regression techniques to advanced machine learning algorithms.
- A detailed review of prediction performance is conducted using different standard evaluation indicators, providing insights into each forecasting model's strengths and limits.
- Discovering enhanced efficiency of specific models, notably GBR, LR, LSTM, and Bi-LSTM, in certain geographical regions inside Saudi Arabia, which can influence decision-making processes for solar energy project development and execution.

Overall, this work advances the state-of-the-art solar energy forecasting approaches by providing significant information for stakeholders involved in the development and administration of solar energy infrastructure in Saudi Arabia and similar countries.

The study focuses on applying ML techniques for solar irradiance forecasting, including regression models (e.g., LR, GBR), deep learning architectures (e.g., LSTM, Bi-LSTM), and ensemble methods (e.g., RF). By analyzing information and forecasting solar energy generation levels, these ML algorithms enhance the planning process for RESs. The study illustrates the efficacy of sophisticated ML techniques in precisely predicting solar irradiance through a thorough evaluation and comparison of these algorithms. This is important for optimizing solar energy system integration and easing the shift to sustainable energy practices.

2. Data and Methodology

This research uses regression analysis to anticipate energy usage in the Saudi Arabian Kingdom in different regions as seen in Fig. 1. As depicted in Fig. 2, many regression approaches are used, such as linear regression, gradient boosting regressor, LSTM, Bi-LSTM, and KNN. Preprocessing the dataset, choosing a model, and assessing performance are the three steps of the methodology. The most accurate findings are obtained by linear regressions. The study tackles modeling uncertainty and attempts to support the power industry in projecting future electricity use.

Data is gathered from multiple sources and goes through preprocessing stages such as skewness, imputation, normalization, and data cleaning. Then, feature selection methods like Principal Component Analysis (PCA), Recursive Feature Elimination (RFE), and Recursive Feature Addition (RFA) are used. Performance measures like MAE, Mean Squared Error (MSE), RMSE, and MAPE are used to examine the results of the use of deep learning algorithms (LSTM, Bi-LSTM, and traditional machine learning classifiers (KNN, LR, GBR), as illustrated in Fig 2.

Fig. (1). Saudi Arabia Map

2.1. Dataset

Three datasets from various cities located in Saudi Arabia like Turaif, Qassim, and Majmaah, are examined in this study. Date, time, global horizon, clear sky, top of atmosphere, code, temperature, relative humidity, pressure, wind speed, wind direction, rainfall, and snowfall are among the 14 columns in each dataset. A total of 72, 961 entries from February 1st, 2004, to March 1st, 2006, or almost two years, make up the dataset. Global horizon is ranging from −999.00 to 275.23, temperature ranging from 274.15 to 320.01, and pressure ranging from 924.93 to 954.28 are some of the important data ranges. Snowfall ranged from 0.00 to 0.00, and rainfall from 0.00 to 1.82, as shown in Fig 3.

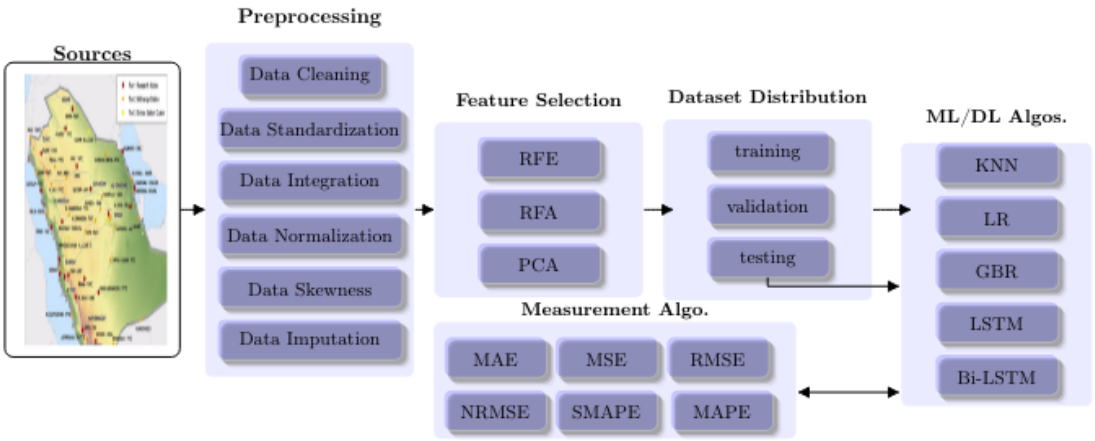


Fig. (2). Workflow for the Proposed Methodology

2.2. Proposed Methodology

This work thoroughly assesses the dataset using a variety of statistical techniques throughout the preparation phase of our study to make sure it is reliable and appropriate for analysis. Metrics are used like variance, summation, skewness, standard error, and deviation among various approaches. When evaluating the distribution, variability, and general properties of the data, each of these metrics is essential. By measuring the amount that individual data points depart from the average, the standard deviation sheds light on how the data points are distributed or dispersed around the mean. The precision of the sample mean estimation is revealed by the standard error, which calculates the variability of sample means around the population mean.

A comprehensive analysis of statistical methods is crucial for understanding the properties and distributions of the dataset. This includes examining measures such as standard deviation, skewness, summation, and variance, as depicted in Fig. 4. The data's skewness, as shown in Fig. 4, indicates if the data is symmetrically distributed around the mean or skewed towards one tail. This measure of asymmetry helps determine the shape of the distribution and helps spot any possible outliers that might affect further investigation. Furthermore, Fig. 4 provides a graphic depiction of the relative value of every characteristic in the dataset, illuminating its importance. This greater comprehension of each variable's contribution to the dataset's features is made possible by these visualizations, which also serve as a basis for later modeling and analytical judgments. The dataset thoroughly was evaluated using statistical analysis and visualization tools to make sure our data pretreatment strategy is resilient. By identifying abnormalities, outliers, or discrepancies, this thorough inspection makes it possible to make well-informed decisions and produce trustworthy results for further data analysis and assessment.

A basic ML approach called linear regression fits a line to the dataset by forecasting numerical results based on numerical inputs. It is essential for its interpretability and broad application across many industries. The dependent variable in (1) is *y* below:

$$y = b_0 + b_1 x_1 + b_2 x_2 \quad (1)$$

and the independent variables are x_1, x_2, etc. The relationship between the independent and dependent variables is represented by the coefficients (b_1, b_2 etc.). The direction of the relationship is indicated by positive or negative coefficients. To capture significant correlations between variables, linear regression is useful. Nevertheless, scenario-based forecasting has difficulties since producing precise forecasts frequently necessitates knowing predictor values in the future. Despite this, linear regression is still an effective method for comprehending and forecasting results in datasets that are appropriate for its use [11].

An adaptable machine learning technique that works well for both regression and classification applications is random forest. To improve prediction accuracy, it combines several decision trees using ensemble approaches. In the random forest, every decision tree functions independently and adds to the final prediction. Random forest ensures forecast stability and reduces overfitting by combining the output from several trees. Random forests need uncorrelated trees and characteristics with predictive power to operate at their best. This approach is well-known for its ease of handling high-dimensional data and is especially useful when working with complex datasets.

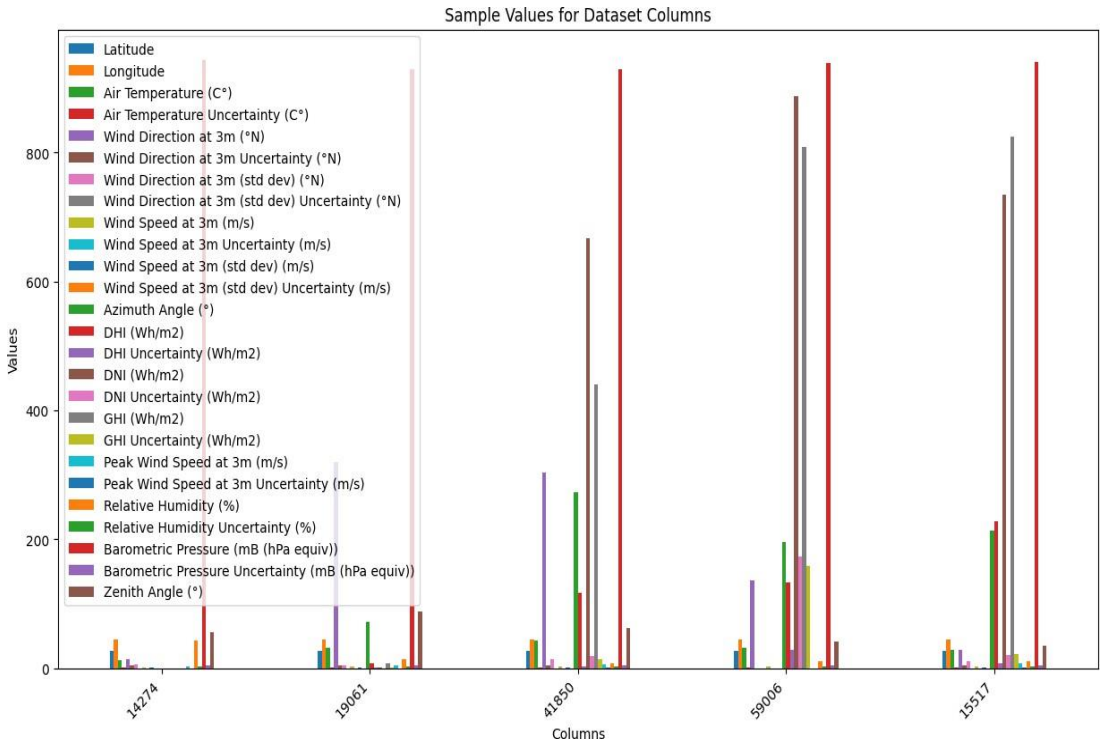


Fig. (3). Sample data

The KNN algorithm is a popular and adaptable machine learning technique that may be utilized for both regression and classification applications. The key to its efficacy is ‘‘feature comparability”, which is the idea that the algorithm uses to determine how comparable new data points are to preexisting samples in the training dataset. KNN is useful in situations that require precision but lack a predetermined solution structure because of its capacity to produce extremely accurate predictions. The success of the method depends on some variables, including the distance measure selected and the quantity of spatial data available. Even with potentially greater computing costs, KNN is still the better choice when accuracy is more important than forecast frequency. Although KNN has several drawbacks, it is flexible enough to accommodate different configurations intended to improve performance on a range of datasets [12].

$$\widehat{y(x)} = \frac{1}{k}\sum_{i=1}^{k} y_i \qquad (2)$$

Equation (2) is the KNN algorithm’s forecast for input xx. In (2), $\hat{y}(x)$ represents the projected output value for input x. The parameter k indicates the number of nearest neighbors to take into account when generating a forecast, while y_i indicates the output values of these k nearest neighbors. The equation computes the forecast by adding the output values of the nearest neighbors and then dividing by k to get the average. This averaging procedure ensures that the forecast is standardized and less influenced by outliers or the density of data points in the neighborhood.

KNN uses the average of its k nearest neighbors’ outputs in the feature space to forecast the output of a new data point as given in (2). When performing classification tasks,

it classifies a new data point according 11 to the majority class among its closest neighbors; for regression tasks, it predicts the average of the target values of its closest neighbors.

Gradient boosting is a potent ML method that may effectively capture nonlinear relationships seen in datasets. With no need for preprocessing, it can effectively handle datasets with a high cardinality of features, missing values, and outliers. Gradient boosting is an ensemble technique that combines several weak models to improve performance. It makes predictions more accurate by methodically lowering prediction errors and modifying forecasts according to how they affect the overall error through repetitive iterations. The gradient of the prediction error for every sample, which directs the model efficiently minimizing prediction errors, is where the technique gets its name [13].

$$F(x) = \sum_{m=1}^{M} y_m \, h_m(x) \tag{3}$$

The GBR algorithm's ensemble estimate for input *x* is represented in (3) which indicates the GBR model's ultimate estimate for the input x. It is the aggregate of individual predictions from several weak learners, each with a weight assigned throughout the training process. *M* represents the total number of weak learners (decision trees) in the ensemble. Each poor learner is identified by the index mm, which ranges from *1* to *M*. The y_m is the weight or coefficient provided to the estimate of the *m-th* weak learner. It calculates the impact of the *m-th* weak learner on the final prediction. $h_m(x)$ represents the prediction generated by the *m-th* weak learner (decision tree) given input *x*. Each weak learner usually makes binary decisions depending on the characteristics of the incoming data.

GBR successively constructs a collection of weak decision trees. Every tree undergoes training to forecast the residuals, or mistakes, of the antecedent trees. The total of all the trees' predictions, weighted by a learning rate that establishes each tree's contribution, is the final forecast, as given in (3).

Time-series forecasting tasks like energy consumption prediction using the K.A.CARE dataset are well suited for the effective Recurrent Neural Network (RNN) architectures of LSTM and Bi-LSTM. These architectures are good for sequential data analysis. With the use of memory cells, input gates, forget gates, and output gates, LSTM networks are particularly good at identifying long-term dependencies in sequential data. This architecture provides greater performance in modeling complicated temporal correlations found in energy consumption data by enabling LSTM to learn and retain patterns across extended sequences [14]. The LSTM can be explained by the following (4) to (9):

$$f_t = \sigma(W_f . [h_{t-1}, x_t] + b_f) \tag{4}$$
$$i_t = \sigma(W_i . [h_{t-1}, x_t] + b_i) \tag{5}$$
$$O_t = \sigma(W_o . [h_{t-1}, x_t] + b_o) \tag{6}$$
$$\widetilde{C_t} = tanh(W_c . [h_{t-1}, x_t] + b_C) \tag{7}$$
$$C_t = f_t \odot C_{t-1} + i_t \odot \widetilde{C_t} \tag{8}$$
$$h_t = O_t \odot tanh(C_t) \tag{9}$$

f_t indicates the forget gate activation vector at time step *t*. It specifies how much of the cell state from the previous interval of time should be forgotten. i_t refers to the input gate activation vector at time step *t*. It regulates how much of the novel input will be integrated into the cell state. O_t is the output gate activation vector at time step *t*. It estimates the cell's output at time step *t* depending on its current condition. $\widetilde{C_t}$ cell state activation vector at time step *t*. It calculates new candidate values for the cell state. C_t indicates the cell state vector at time step *t*. It holds information from past time steps that has been selectively updated or forgotten depending on the gates. h_t is the hidden state vector at time step *t*. It is the output of the LSTM cell at time step *t*, computed using the current cell state and output gate activation.

An LSTM extension called Bi-LSTM processes input sequences in both forward and backward orientations at the same time, improving model performance. Bi-LSTM can now capture dependencies from both past and future time steps thanks to its bidirectional processing, giving researchers a more thorough understanding of the temporal patterns in the K.A.CARE dataset. The LSTM and Bi-LSTM architectures' advantages can be used by researchers to create reliable and accurate models for forecasting energy usage. With the ability to incorporate seasonal patterns, historical trends, and other temporal dependencies, these models can successfully handle the dynamic nature of energy consumption data and produce accurate projections that are crucial for energy planning and management.

3. Results and Discussions

To assess model performance and convergence, it is essential to compare different ML algorithms such as (LR, KNN, GBR) and deep learning techniques such as (LSTM, Bi-LSTM), taking into account factors like training accuracy, loss, validation accuracy, and validation loss.

Figure 5 presents an assessment of the performance of different regression models in estimating solar irradiance over three major regions: Turaif, Qassim, and Majmaah. This research Starts with the Turaif and Qassim datasets, which indicate improved predictive capabilities due to the higher performance of LSTM and Bi-LSTM models over other models. In particular, the LR model has a notable MAE of 5.69, which indicates that it can reasonably predict real values with little variation. Moreover, the RMSE of 15.50 is rather good and highlights the accuracy with which it captures variability in the dependent variable.

The results of our investigation showed that the performance of different regression models varied significantly depending on the geographical location. More specifically, in both the Turaif and Qassim regions, the LR model performed the worst when it came to forecasting sun irradiance levels. The notably greater MAE and RMSE numbers in comparison to other models demonstrate this as shown in Tables 1 to 3. Because of the limitations of the LR model, it is crucial to use more sophisticated machine learning methods that can identify the nonlinear correlations present in solar irradiance data.

The predictive effectiveness of several models on the Turaif and Qassim datasets is shown in Fig. 5 to give a visual comparison of model performance.

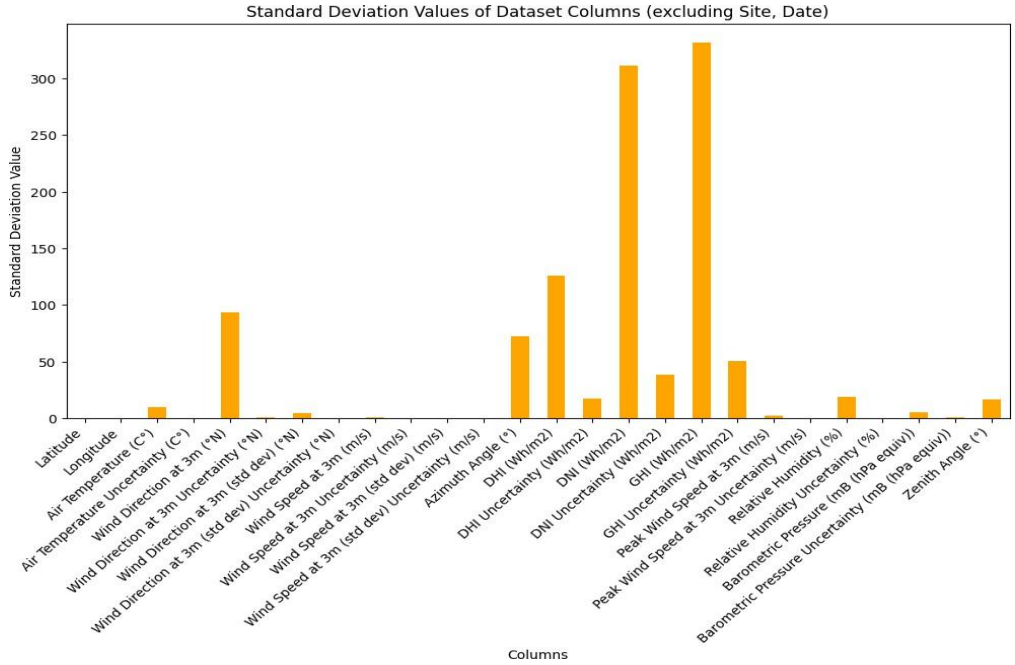


Fig. (4). Exploring Statistical Methods of the Data: (a) Standard Deviation

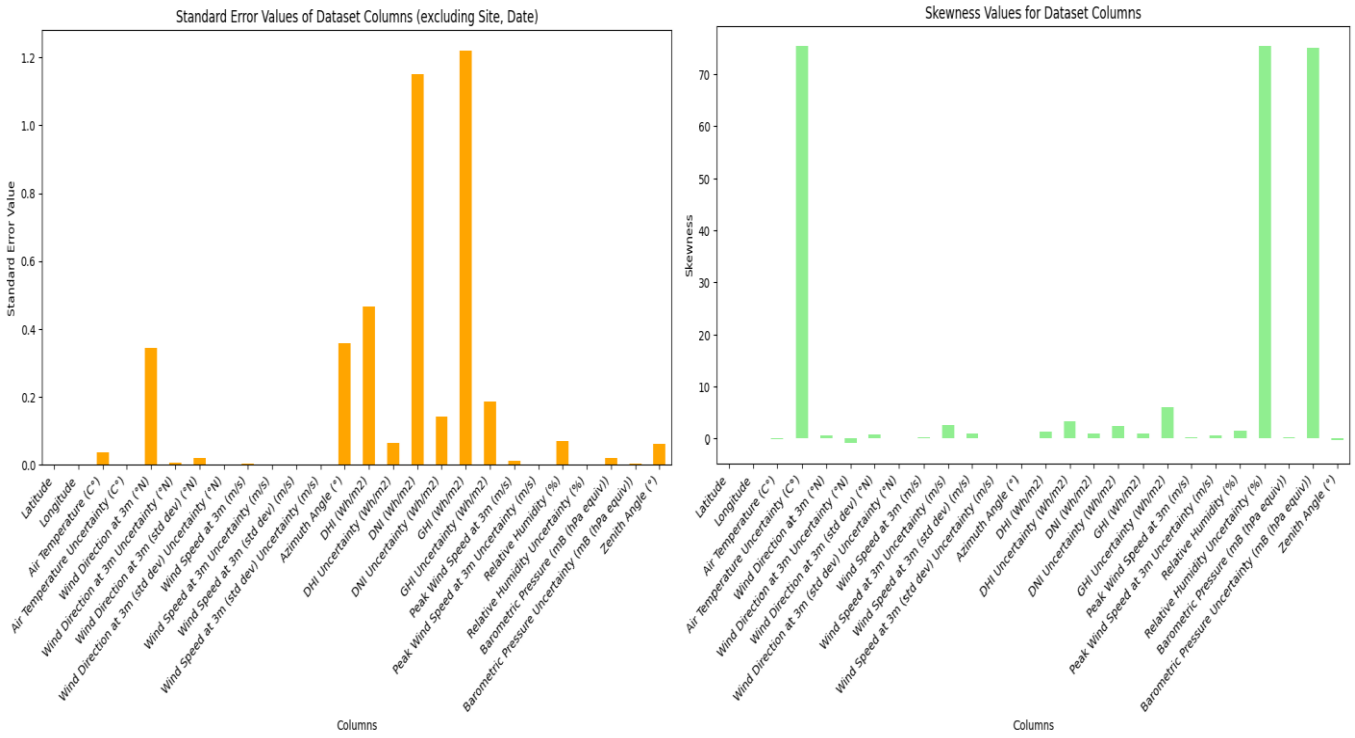
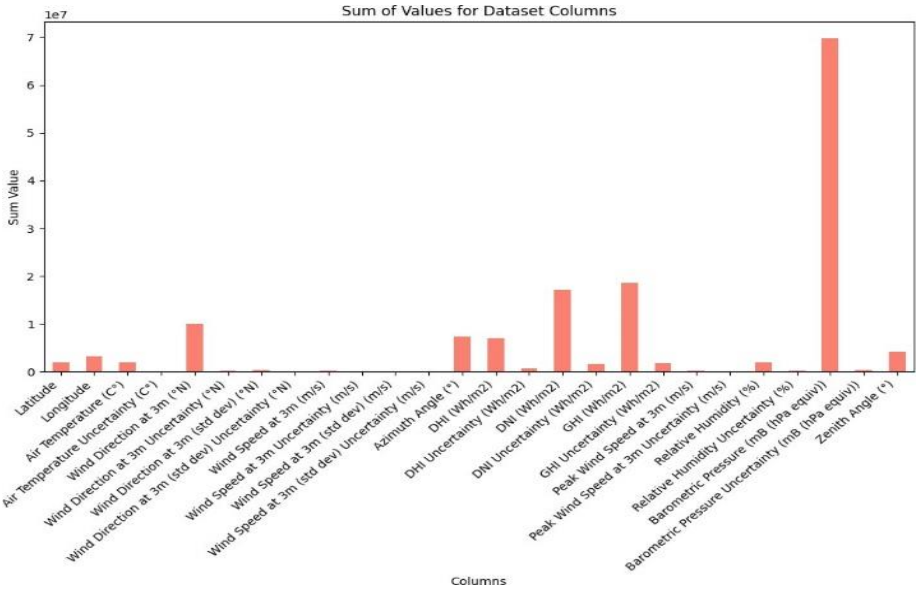


Fig. (4). Exploring Statistical Methods of the Data:

(b) Standard Error (c) Skewness



(d) Summation

Fig. (4). Exploring Statistical Methods of the Data: (d) Variance

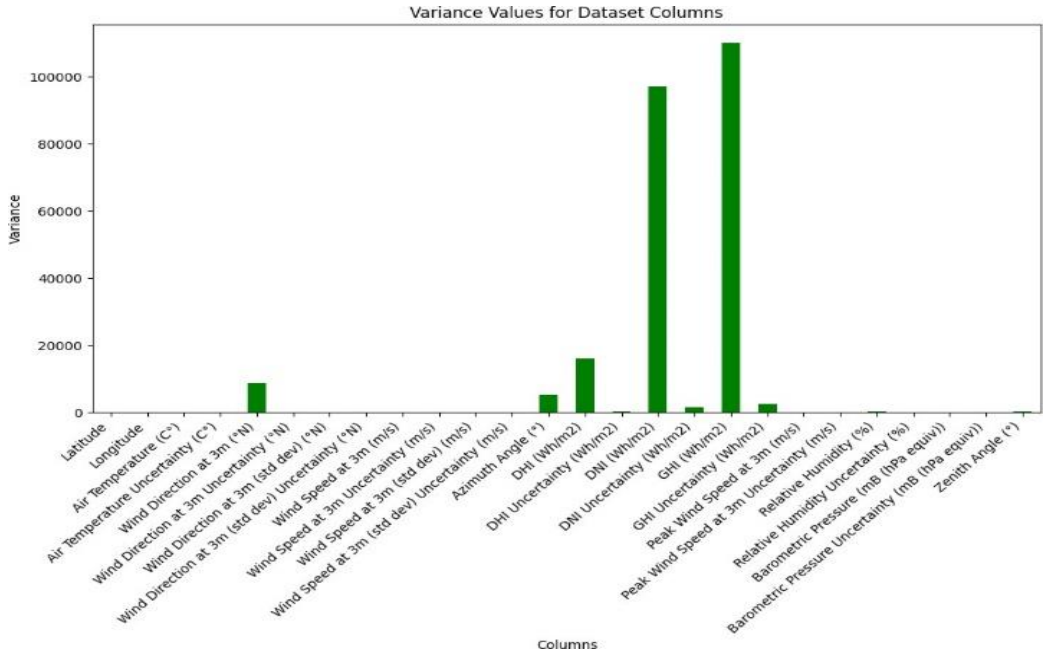


Fig. (4). Exploring Statistical Methods of the Data: (e) Variance

This graphical depiction helps stakeholders identify the best method for solar irradiance forecast in each region by providing insightful information about the relative performance of models. The differences in model performance draw attention to the necessity of customized modeling approaches that take into consideration the particular qualities of each geographic area.

The results showed that ensemble learning methods - specifically, the GBR model - performed better in the Majmaah region. In terms of solar irradiance levels predictions, the GBR model demonstrated impressive resilience, with a comparatively low MAE of *31.95* and RMSE of *80.61* as shown in Table 2. In a similar vein, the KNN model performed admirably,

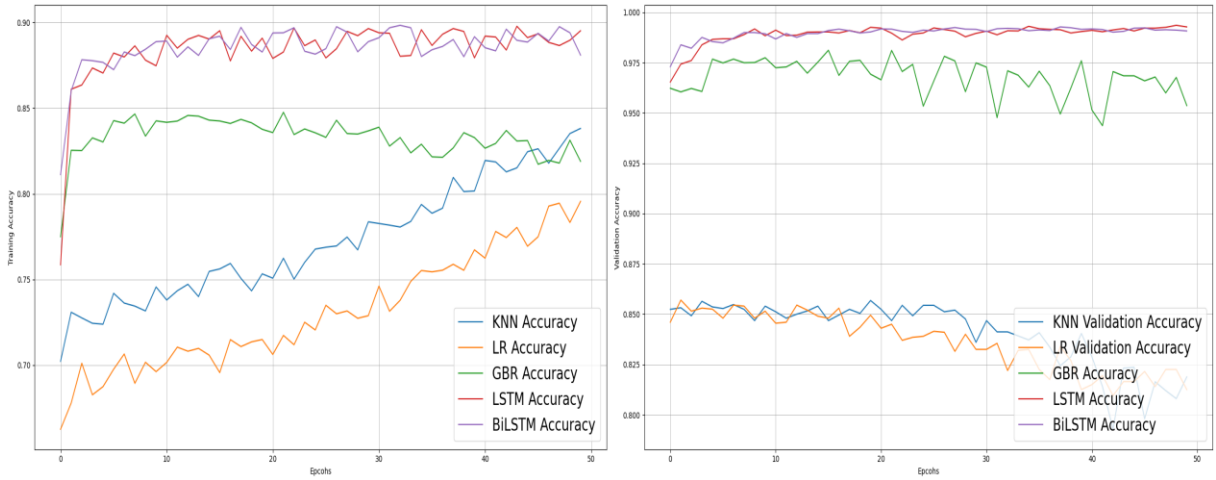
lagging the GBR model in terms of prediction accuracy. On the other hand, Majmaah demonstrated yet another failure of the LR model, confirming its restricted effectiveness in representing the intricacies of solar irradiance dynamics in this area.

These findings highlight how crucial it is to choose modeling strategies that are suited to the unique features of each geographic area. While simpler regression models like LR may not be sufficient for precise forecasting, ensemble learning techniques like GBR are excellent at capturing the nonlinear correlations present in sun irradiance data. Stakeholders may maximize planning and deployment strategies for renewable energy and improve the accuracy of solar irradiance projections by utilizing sophisticated machine learning algorithms that are customized to the distinct characteristics of each region.

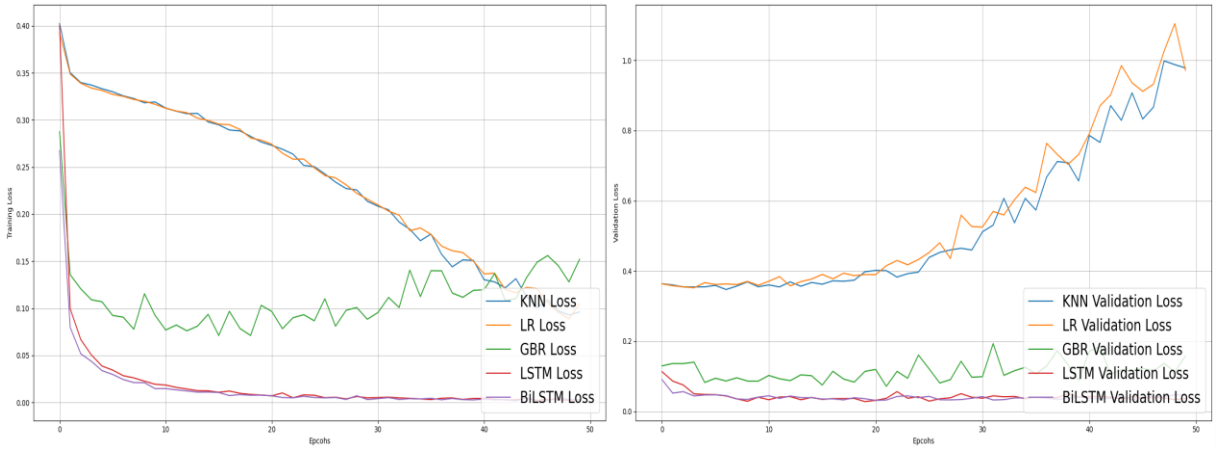
By performing a thorough comparison with previous research findings, the study goes beyond simply evaluating the performance of the model. All datasets, including Turaif, Qassim, and Majmaah, showed that the LSTM, Bi-LSTM, and GBR models consistently performed better than alternative approaches, as presented in Tables 1 to 3. These models demonstrated better prediction ability as evidenced by their lower MAE and RMSE values as well as their higher R2 and AR2 values. There was a huge difference in the performance of the model when results compared with those of earlier studies. Although the accuracy metrics of the recommended models were consistently remarkable, several traditional methods that were previously used for solar irradiance prediction did not perform up to par. These models demonstrated lower R2 values and greater error metrics, highlighting their insufficiency in capturing the complex interactions present in solar irradiance data.

Table (1). Performance Metrics of Regression Models on Turaif DataSolar energy market in Saudi Arabia

Model	**MAE**	**MSE**	**RMSE**	**NRMSE**	**AR2**
LR	23.99	42.11	89.01	0.43	0.43
GBR	5.69	40.14	55.50	0.98	0.98
KNN	23.93	22.56	89.01	0.43	0.43
RFR	44.51	51.31	41.46	0.99	0.99
LSTM	74.19	94.94	91.74	0.99	0.99



(a) Training Accuracy (b) Validation Accuracy



(c) Training Loss (d) Validation Loss

Fig. (5). Comparison of Models Performance and Convergence

The significance of advanced machine learning methods like LSTM and Bi-LSTM in transforming solar irradiance forecasting and enhancing solar energy system deployment is highlighted by the obtained results. Through the utilization of these sophisticated techniques, interested parties may allocate resources and choose sites with knowledge, which will help solar energy technology become widely adopted in the Middle East and beyond. The obtained results also demonstrate how crucial it is for legislators and business professionals to adopt cutting-edge machine learning techniques for forecasting solar irradiation. Stakeholders may reduce uncertainty related to solar energy production and increase its efficiency by utilizing the predictive power of LSTM and Bi-LSTM models. This will ultimately accelerate the shift towards sustainable energy solutions.

The results of this work indicate that the Middle East, and especially Saudi Arabia, have enormous potential for solar energy harvesting since their yearly solar radiation rates are more than 2100 kWh/m^2. To help with solar power plant siting and forecasting, the proposed study concentrated on three important Saudi Arabian regions: Turaif, Qassim, and Majmaah. It was done by studying the relationships between 15 several weather factors and sun intensity. The ML techniques that are suggested for forecasting GHI showed different levels of performance in different areas. The R-squared (R^2) values in Turaif, Qassim, are highest (98%, 99%, and

99%) and the RMSE values are lowest (15.5%, 9.74%, and 11.46%, respectively) for LSTM, GBR, and Bi-LSTM. Algorithms such as KNN, and LR, on the other hand, performed less well in Turaif, Qassim, suggesting that they are not very useful in GHI predictions for this area.

Table (2). Performance Metrics of Regression Models on Majmaah Data

Model	MAE	MSE	RMSE	NRMSE	AR2
LR	20.55	42.29	80.01	0.51	0.51
GBR	8.69	36.14	59.50	0.88	0.88
KNN	19.93	28.56	85.01	0.49	0.49
RFR	43.51	47.31	48.46	0.92	0.92
LSTM	64.19	84.94	81.74	0.96	0.96
Bi-LSTM	71.95	92.91	90.2	0.84	0.84

Significant differences were found in the performance of different machine learning methods for solar irradiance prediction in the Majmaah region. In particular, the models that performed the best were KNN, LSTM, and Bi-LSTM, with R2 values of 93%, 94%, and 94%, respectively. These strong R2 values underscore the algorithms' effectiveness in GHI forecasting by demonstrating their resilience in capturing the unpredictability of solar irradiance levels. By comparison, the results of Majmaah showed that conventional approaches like LR, KNN, and GBR performed worse. The disparity in model performance highlights the shortcomings of traditional methods in precisely forecasting solar irradiance levels in this particular region. The significantly lower R2 values corresponding to LR, KNN, and GBR demonstrate their insufficiency in encapsulating the complex correlations between climatic factors and sun intensity in Majmaah.

Table (3). Performance Metrics of Regression Models on Qassim Data

Model	MAE	MSE	RMSE	NRMSE	AR2
LR	21.27	47.39	83.27	0.52	0.52
GBR	9.84	39.68	61.52	0.89	0.89
KNN	18.73	29.84	86.71	0.50	0.50
RFR	42.16	48.21	47.63	0.91	0.91
LSTM	65.82	82.37	80.02	0.95	0.95
Bi-LSTM	70.49	90.17	88.32	0.83	0.83

Due to their innate capacity to adjust to nonlinear patterns and relationships in the data, the KNN, LSTM, and Bi-LSTM models in Majmaah have demonstrated remarkable performance. These models, particularly LSTM and Bi-LSTM are effective at capturing the dynamic variations in solar irradiance by leveraging the temporal dependencies that are stored in the sequential data, thus enhancing the predictive accuracy. Similarly, Majmaah's experience with the KNN algorithm highlights how important it is to use nearest neighbor relationships and localized data to accurately estimate solar irradiance levels. The design and implementation of solar energy in the Majmaah region are going to be significantly impacted by these findings. Stakeholders can make more informed decisions by adopting sophisticated

ML approaches like KNN, LSTM, and Bi-LSTM, which provide more accurate and consistent estimates of solar irradiance. Moreover, the success of these top-performing algorithms highlights the importance of employing data driven strategies to navigate the difficulties of solar energy forecasting in different geographical contexts.

The proposed algorithms performed better for both Turaif, Qassim, and Majmaah when compared to state of-the-art methods, such as Bi-LSTM, LSTM, GBR, and KNN as depicted in Tables 1 to 3. Due to their stable weather, Turaif and Qassim, are suitable locations for the construction of solar power plants, as shown by reduced RMSE and higher R2 values, which highlight the combined influence of weather characteristics on solar intensity. On the other hand, Majmaah also offers similar favorable conditions most of the time; however, its occasionally varying weather suggests that solar energy projects there require more detailed consideration. It is concluded that to optimize solar energy production and deployment techniques in places such as Saudi Arabia, machine learning algorithms must be utilized to accurately estimate GHI.

4. Concluesion

The ability to accurately predict solar irradiance levels enables planners to make informed decisions about the sizing, placement, and operation of renewable energy assets, such as solar PV systems and concentrating solar power plants. In summary, the proposed work clarifies the effectiveness of ML algorithms in accurately forecasting solar irradiance levels, which is essential for efficiently planning RESs. Through comprehensive testing in various regions, including Saudi Arabian cities like Turaif, Qassim, and Majmaah, regression models were evaluated such as LSTM, Bi-LSTM, and KNN. The results emphasize the critical role of precise solar irradiance forecasts in maximizing the integration and deployment of solar energy systems.

High goodness-of-fit metrics (R-squared, adjusted R-squared) and low error metrics (MAE, RMSE) demonstrate that models such as LR and GBR perform better. Conversely, traditional regression models such as LR show poor predictive power, highlighting the necessity for complex ML methods specifically designed for solar irradiance prediction. Additionally, the superiority of the proposed models in terms of accuracy and reliability is reaffirmed by the comparison with previous research in the literature.

This study advances solar irradiance forecasting and RES planning by utilizing data-driven approaches like KNN and ensemble methods like LSTM. To improve prediction accuracy, future research should concentrate on improving already-existing models, investigating cutting-edge ML strategies, and incorporating new data sources. Addressing these challenges can accelerate the transition to a future powered by RES and drive technological progress in the field. Overall, the synergy between solar irradiance forecasting and renewable energy planning is crucial for the transition to a sustainable and resilient energy landscape.

ACKNOWLODGEMENT

The authors express would like to thank the Smart Grids and Smart Cities (SGSC) Lab, Qassim University, Qassim, Saudi Arabia, for their consistent support and valuable contributions to this research.

Also, the author would like to extend their appreciation to the team at King Abdullah City for Atomic and Renewable Energy (K.A.CARE) for their invaluable assistance in providing access to the required data and for their support.

REFERENCES

[1] A. Azadeh and Z. S. Faiz, "A meta-heuristic framework for forecasting household electricity consumption," *Appl. Soft Comput. J.*, vol. 11, no. 1, pp. 614–620, 2011, doi: 10.1016/j.asoc.2009.12.021.

[2] T. Hong and S. Fan, "Probabilistic electric load forecasting: A tutorial review," *Int. J. Forecast.*, vol. 32, no. 3, pp. 914–938, 2016, doi: 10.1016/j.ijforecast.2015.11.011.

[3] M. Santhakumari and N. Sagar, "A review of the environmental factors degrading the performance of silicon wafer-based photovoltaic modules: Failure detection methods and essential mitigation techniques," *Renewable and Sustainable Energy Reviews*, vol. 110. Elsevier Ltd, pp. 83–100, Aug. 01, 2019. doi: 10.1016/j.rser.2019.04.024.

[4] S. C. Karmaker, K. K. Sen, B. Singha, S. Hosan, A. J. Chapman, and B. B. Saha, "The mediating effect of energy poverty on child development: Empirical evidence from energy poor countries," *Energy*, vol. 243, 2022, doi: 10.1016/j.energy.2021.123093.

[5] Y. Fu, Z. Li, H. Zhang, and P. Xu, "Using Support Vector Machine to Predict Next Day Electricity Load of Public Buildings with Sub-metering Devices," *Procedia Eng.*, vol. 121, pp. 1016–1022, 2015, doi: 10.1016/j.proeng.2015.09.097.

[6] F. M. Butt, L. Hussain, A. Mahmood, and K. J. Lone, "Artificial Intelligence based accurately load forecasting system to forecast short and medium-term load demands," *Math. Biosci. Eng.*, vol. 18, no. 1, pp. 400–425, 2021, doi: 10.3934/MBE.2021022.

[7] M. K. M. Shapi, N. A. Ramli, and L. J. Awalin, "Energy consumption prediction by using machine learning for smart building: Case study in Malaysia," *Dev. Built Environ.*, vol. 5, no. July 2020, p. 100037, 2021, doi: 10.1016/j.dibe.2020.100037.

[8] M. Zubair, "PV energy penetration in Saudi Arabia: current status, residential, and commercial users, local investment, use in modern agriculture," *Int. J. Sustain. Eng.*, vol. 17, no. 1, pp. 1–13, 2024, doi: 10.1080/19397038.2023.2297262.

[9] N. Samargandi, M. Monirul Islam, and K. Sohag, "Towards realizing vision 2030: Input demand for renewable energy production in Saudi Arabia," *Gondwana Res.*, no. June, 2023, doi: 10.1016/j.gr.2023.05.019.

[10] M. S. Islam, M. M. Islam, A. U. Rehman, M. F. Alam, and M. Taique, "Mineral production amidst the economy of uncertainty: Response of metallic and non-metallic minerals to geopolitical turmoil in Saudi Arabia," *Resour. Policy*, vol. 90, no. February, p. 104824, 2024, doi: 10.1016/j.resourpol.2024.104824.

[11] M. D. Alanazi *et al.*, "Enhancing Short-Term Electrical Load Forecasting for Sustainable Energy Management in Low-Carbon Buildings," *Sustainability*, vol. 15, no. 24, p. 16885, 2023, doi: 10.3390/su152416885.

[12] G. Hong, G. S. Choi, J. Y. Eum, H. S. Lee, and D. D. Kim, "The Hourly Energy Consumption Prediction by KNN for Buildings in Community Buildings," *Buildings*, vol. 12, no. 10, 2022, doi: 10.3390/buildings12101636.

[13] L. F. M. Sepulveda *et al.*, "Forecasting of individual electricity consumption using Optimized Gradient Boosting Regression with Modified Particle Swarm Optimization," *Eng. Appl. Artif. Intell.*, vol. 105, no. August, p. 104440, 2021, doi: 10.1016/j.engappai.2021.104440.

[14] J. Huang, M. Algahtani, and S. Kaewunruen, “Energy Forecasting in a Public Building: A Benchmarking Analysis on Long Short-Term Memory (LSTM), Support Vector Regression (SVR), and Extreme Gradient Boosting (XGBoost) Networks,” *Appl. Sci.*, vol. 12, no. 19, 2022, doi: 10.3390/app12199788.

نحو مستقبل أخضر: تطبيقات التعلم الآلي في التنبؤ بإشعاع الشمس لتخطيط الطاقة المتجددة في المملكة العربية السعودية

طـــلال الحربي

قسم الهندسة الكهربائية – كلية الهندسة – جامعة القصيم
المملكة العربية السعودية

atalal@qu.edu.sa

سعيد إقبال

قسم علوم الحاسب الآلي – كلية تقنية المعلومات و علوم الحاسب الآلي – جامعة وسط البنجاب
باكستان

saeediqbalkhattak@gmail.com

ملخص البحث. إن الأمور التي تؤثر إيجاباً على تخطيط مصادر الطاقة المتجددة هي دمج تقنيات التنبؤ بالإشعاع الشمسي في المستقبل. لذلك تبحث هذه الدراسة استخدام ودمج تقنيات خوارزميات التعلم الآلي من خلال التنبؤ بالإشعاع الشمسي من خلال تسخير القوة التنبؤية لهذه الخوارزميات بهدف تعزيز استراتيجيات التخطيط لمصادر الطاقة المتجدد في المملكة العربية السعودية والحصول على رؤية أكثر دقة لمستويات الإشعاع الشمسي في المستقبل. تم استخدام مجموعة من البيانات من مناطق مختلفة في المملكة العربية السعودية ومن خلالها تم تقييم العديد من نماذج الانحدار مثل نموذج الانحدار الخطي ونموذج الانحدار المعزز للتدرج والذاكرة طويلة وقصيرة المدى والذاكرة طويلة المدى ثنائية القطب وكذلك خوارزمية أقرب جار.

ويكشف تحليل هذا البحث أن تقنيات التجميع مثل الانحدار العشوائي للغابات وكذلك الخوارزميات التي تعتمد على البيانات مثل خوارزمية أقرب جار تظهر أداءً متفوقاً في توقع اشعاع الشمس مقارنةً بنماذج الانحدار الخطي التقليدية هذه النتائج تؤكد أهمية وتفوق أساليب التعلم الآلي المختلفة في التنبؤ بالإشعاع الشمسي مقارنة مثلاً بخوارزمية شجرة القرار. كذلك تظهر النماذج ذات المقاييس العالية لجودة الملاءمة والمقاييس المنخفضة للخطأ قدرة تنبؤية أفضل. ويتم تأكيد ذلك بالنتائج على الدقة التي توقعت بها النماذج المقترحة لمستويات إشعاع الشمس مقارنةً مع الدراسات السابقة. وتسلط النتائج الضوء على الإمكانات المتاحة لاستخدام نهج التعلم الآلي من اجل تحسين تكامل نظم الطاقة الشمسية وتسريع عملية التحول نحو ممارسات الاستدامة في مجال الطاقة