A Comparative Study of Machine Learning Approaches for Accurate Predictive Modeling of Solar Energy Generation

Khaled Chaaban† and Najf Alfadl††

†College of Computing, Umm Al-Qura University, 21955 Makkah, Saudi Arabia
††Riyadh Schools, Riyadh, Saudi Arabia

Summary
Solar energy prediction poses a challenging task that requires robust models and precise data to accurately forecast solar energy yield, especially in grid areas with a large share of photovoltaics. Existing methods often rely on statistical or physical models, which have limitations in capturing the complex and non-linear relationships between weather variables and solar power generation. In this paper, we address this issue by comparing and evaluating different learning models, ranging from artificial neural networks (ANNs) and random forest models to long- and short-term memory (LSTM) networks, to predict the PV energy yield based on weather forecast data. A methodology has been developed to evaluate various models using real-world datasets from a large-scale industrial solar project, incorporating historical photovoltaic data, meteorological data, and solar irradiation data. The experimental results showed that the Random Forest Algorithm (RFR) consistently outperforms other algorithms, providing a mean absolute error (MAE) of 0.06 and a root mean square error (RMSE) of 0.15 when applied to historical meteorological datasets. The accuracy of the learning model was improved by combining meteorological data with a solar irradiation dataset to obtain an MAE of 0.03 and an RMSE of 0.09. Validation analysis has shown that the proposed model is highly effective in terms of both forecast accuracy and stability. The proposed methodology has the potential to provide valuable information to PV system operators, grid managers, and energy planners, facilitating the optimization of the use of solar energy resources.

Keywords:
Machine learning, neural networks, optimization, photovoltaics.

1. Introduction

As the demand for renewable energy continues to increase, efficient solar power use plays a crucial role in sustainable energy solutions. Integrating a higher proportion of renewable sources, such as solar power, is advantageous in reducing carbon emissions and meeting future power grid requirements. However, this integration also poses new challenges related to grid management. Variability and uncertainty in photovoltaic (PV) energy generation can lead to stability and reliability issues in power system operations due to the intermittent nature of solar-generated electricity. Consequently, grid operators must incorporate these factors into their generation planning and dispatch operations [1],[2]]. This has led utility grids with numerous distributed photovoltaic systems to transition to modern, digitally enhanced technologies. These technologies facilitate the monitoring and control of distributed energy resources. The ongoing trends of electrification, decentralization, and digitalization are driving the transformation of the current paradigm of the power sector. This transformation aims to fully leverage the flexibility of the system to accommodate high levels of variable renewable energy.

Saudi Arabia has set ambitious goals of generating 50% of its energy from renewable sources by 2030[3]]. To achieve this goal, the country has actively invested in renewable energy initiatives, focusing on solar energy as a key component of its renewable energy strategy. However, the integration of PV systems into the energy grid presents challenges due to the variability of solar power generation, which is influenced by weather and environmental conditions [[4],[5]]. Accurate forecasting of solar power production is crucial to effective energy management and grid stability. Several research investigations have explored the impact of adverse weather conditions, such as wind speed/direction, temperature, relative humidity, and the frequency of dust storms, on the losses in photovoltaic cell power output. Understanding the dynamics of solar panels is crucial to ensure proper installation and achieve optimal performance [[5],[6]].

Numerous studies have explored various prediction methodologies, including the use of machine learning algorithms such as artificial neural networks and recurrent neural networks [[7],[8]]. These approaches take advantage of historical weather and environmental data to forecast solar power generation, facilitating the seamless integration of photovoltaic systems into smart grid frameworks. The main objective of this paper is to determine a robust and accurate methodology to predict PV output energy by evaluating different training methods. To do this, a comparative study of widely used machine learning methods such as artificial neural networks (ANN), support vector regression (SVR), and random forest regression (RFR) was performed to evaluate their effectiveness for PV power forecasting applications.

Manuscript received December 5, 2024
Manuscript revised December 20, 2024
https://doi.org/10.22937/IJCSNS.2024.24.12.15

The proposed methodology includes training different machine learning models in different supervised learning scenarios and benchmarking of the forecast performance. The verification was carried out using actual high-quality hourly photovoltaic operational and meteorological measurements acquired over two years from a large-scale industrial photovoltaic plant in Saudi Arabia. The settings of the test bench photovoltaic system provide a perfect opportunity to study the effect of the supervised training regimes on the accuracy of the forecasts based on commonly used metrics. Furthermore, the results obtained provide useful information for the establishment of a robust forecasting methodology that combines historical data and weather forecasts with an optimal supervised learning approach.

The primary contributions of the paper encompass the following key points:

(i) Establish a resilient PV power forecasting system through a comprehensive evaluation of various machine learning algorithms, thereby identifying the most precise model.
(ii) Elevating the precision of the prediction of the PV power output by incorporating weather-related features alongside historical data.
(iii) Enhancing the efficiency and accuracy of the performance and forecasting of the network connection.

The remainder of this paper is structured as follows. In Section 2, we provide a review of the literature and discuss related work. Section 3 delves into the case study, including its datasets and the process of feature engineering. Section 4 describes the methodology used to evaluate and compare various learning algorithms. Section 5 presents the experimental setup and presents the results of our work. Finally, in Section 6, we conclude and highlight key perspectives from this study.

2. Related Works

In recent years, there has been a significant increase in attention and advancement in research related to solar energy prediction. Numerous studies have focused specifically on developing accurate prediction models using machine learning techniques. An approach commonly adopted in solar power forecasting is the use of artificial neural networks (ANN) [9]. ANNs have been shown to be effective in capturing the nonlinear relationships between weather variables and solar power generation [[10],[11],[12],[13]]. In [[10]], Saberian et al. utilized artificial neural networks to forecast the output power of solar panels based on meteorological conditions, including temperature and humidity. In a similar vein, Lee et al. [[11]] explored three methods, namely, ANNs, deep neural networks (DNN), and long- and short-term memory (LSTM) models, to predict PV power output by capturing hidden relationships within meteorological data. In [[12]], the authors highlight the effectiveness of artificial neural networks in forecasting solar radiation. They emphasize the importance of incorporating time-series data and month-specific patterns, which contribute to improving the accuracy of solar energy predictions. However, the weakness in the works cited lies in the relatively high errors in the learning models, indicating room for improvement in prediction accuracy. In [[13]], the authors use artificial neural networks (ANN) and long-short-term memory (LSTM) to predict solar radiation. The objective of this paper is to help harness solar energy more efficiently, especially in regions with intermittent electrical power. The authors evaluated the experimental results by comparing their obtained results using metrics such as the coefficient of determination (R^2), mean square error (MSE), root mean square error (RMSE), mean absolute error (MAE) and mean biased error (MBE),

Another popular machine learning technique for solar power forecasting is support vector regression (SVR), which has been applied in various studies to predict solar power production based on weather data [[14],[11],[15]]. In [14], Nageem et al. introduced a multi-input support vector regression (SVR) model to forecast the performance of solar panels connected to the grid. In their model, humidity, temperature, pressure, and wind speed are incorporated as input variables. In [[11]], the authors used the HIMVO-SVM model to predict PV output power in different weather scenarios using historical data provided by a solar center in Australia. The results demonstrated the effectiveness of the SVR model in accurately predicting solar power output, particularly in capturing the nonlinear relationships between input features and output predictions. In [[15]], the authors employ various machine learning algorithms, including support vector regression, to predict the power output from building integrated photovoltaic systems. The study aims to determine the most accurate and reliable algorithm for short-term power prediction.

Ensemble methods, such as random forest regression (RFR), have also been extensively explored in solar power forecasting. RFR combines multiple decision trees to make predictions and has been shown to achieve high accuracy and robustness [[16]]. In [[17]], the authors present a methodology for forecasting solar power output using a random forest algorithm. The methodology is based on a dataset of historical solar power output and weather data, including solar radiation, temperature, and wind speed. The authors applied their methodology to a solar power production dataset in China and compared the results with those obtained using other forecasting methods. The results showed that the RFR algorithm outperforms the other

methods, achieving a mean absolute error (MAE) of 3.58 and a mean absolute percentage error (MAPE) of 18.56%.

Furthermore, from sequential and time-series data, recurrent neural networks (RNNs) can learn features and long-term dependencies [[18]]. An example of RNN is Long Short-Term Memory (LSTM) networks, which have gained popularity in solar power forecasting due to their ability to capture temporal dependencies in time series data [[19],[20]]. In [[19]], an LSTM-based model was proposed to predict solar power generation based on historical weather data. The results showed that the LSTM model achieved superior performance compared to traditional machine learning models, highlighting the importance of considering temporal dynamics in solar power forecasting. In [[20]], a hybrid model that combined LSTM with a particle swarm optimization algorithm achieved enhanced accuracy in solar power forecasting. Deep learning models, including LSTM and CNN, are often considered black-box models, which means that it can be challenging to interpret the inner workings and understand the reasons behind their predictions. Providing insights into the interpretability of the proposed model could improve its practical utility and acceptance in real-world applications.

Additionally, hybrid models that combine multiple forecasting techniques have also been investigated. These models aim to leverage the strengths of different algorithms to improve the prediction accuracy [[8],[21],[22],[23]]. In [[8]], Di Su et al. adopted a hybrid approach that combined multiple noncorrelated forecasting techniques to improve the accuracy of their predictions. Note that none of the recent studies in the literature have specifically addressed the challenges of long-term projections for such large-scale solar plants. In [[21]], Jatin et al. tested three optimizers and found that the Nadam optimiser outperformed ARIMA and SARIMA, suggesting its potential applicability even to smaller plants with reduced parameter measurements while maintaining high accuracy.

In [[22]], the authors address the complex nature of PV power forecasting by integrating machine learning and statistical post-processing techniques. This hybrid approach takes advantage of the strengths of both methods, allowing for more accurate and reliable predictions.

Table 1: Summary of Related Works

Algorithm	Variables	Metrics	References
ANN	Weather and air pollution, including temperature, humidity, wind speed, and solar radiation	MAE: 0.35 RMSE: 0.85	[4]
		RMSE: 0.055 MAE: 0.038	[8]
		MAE: 0.051 RMSE: 0.063	[11]
		MAE: 0.044 MAPE: 5.71%	[22]
		MAE: 0.17 RMSE: 0.089	[13]
		MAE: 0.042 MAPE: 4.65%	[23]
RFR	Solar radiation, temperature, and wind speed	MAE: 0.035 MAPE: 18.56%	[26]
		MAE: 0.055 RMSE: 0.066	[11]
SVR	Month, hour, horizontal diffuse irradiance, temperature, humidity.	MAPE: 36	[14]
		MAE: 0.062 RMSE: 0.073	[11]
		MAE: 0.084	[15]
HIMVO-SVM	Intensity, humidity, and atmospheric temperature of solar radiation.	MAE: 0.026 MAPE: 1.81%	[27]
LSTM	PV power output data. Meteorological parameters include temperature, humidity, solar radiation, wind speed, etc.	RMSE: 0.8619 MAE: 0.05	[19][25]
		MAE: 2.86 MAPE: 9.65%	[20]
		MAE: 0.273 MSE: 0.2877 R^2: 92.7 %	[21]
Hybrid Model	Temperature, humidity, pressure, wind speed, direction, rainfall, snowfall, and snow depth	nRMSE: 6.74	[8]
		MAE: 0.032 MAPE: 13.1%	[22]
		MAE: 0.029 MAPE: 3.13%	[23]

In [[23]], the authors of the article propose a hybrid model that combines artificial neural networks and support vector regression for the forecasting of solar power. The hybrid model is designed to take advantage of both techniques to improve the accuracy of forecasting for solar power generation. Their article lacks a detailed analysis of the model's performance compared to existing methods or benchmarks. The study evaluates machine learning models to improve day-ahead forecasting for PV power generation, with BNN proving to be highly effective with an nRMSE of 4.53% [[24]]. A study by Richard et al. compares deep learning models for accurately forecasting photovoltaic power generation in Ecuador's Galapagos Islands. In particular, the LSTM model achieved a low RMSE of 0.05, highlighting its effectiveness in short-term prediction for sustainable energy planning [[25]].

3. Case Study and Datasets

3.1 Plant specification

Data for this study were obtained from a commercial solar photovoltaic plant in Sakaka operated by SAKAKA Solar Energy Company (Figure Fig. 1). Sakaka is a city in northwest Saudi Arabia that is the capital of the Al Jawf province. The climate in Sakaka is a tropical and subtropical desert climate. The average temperature for the year is 22 ° C. The hottest month is July with an average temperature of 32 ° C. January is the most humid, with an average relative humidity of 57%.

Fig. 1 IPP photovoltaic plant, Sakaka, Saudi Arabia.

The plant has a capacity of 300 MW, the first utility-scale renewable energy project in Saudi Arabia under the country's National Renewable Energy Program. With a budget of 302 million USD, this project is the first in a series of projects under the Saudi Arabian National Renewable Energy Program, which aims to achieve 50% renewable energy generation in Saudi Arabia by 2030 [3]]. The Sakaka PV IPP plant started commercial operation at the end of 2019. We used two years of data from July 2020 to June 2022.

3.2 Datasets

Solar irradiation data play a crucial role in the analysis and optimization of solar energy systems. In this study, comprehensive solar irradiation data were collected for the entire plant and stored in the data file 'Sakaka_WMS.csv'. The dataset consists of three variables: global horizontal pyranometer irradiance, solar irradiance from the POA (Plane of Array) solar irradiance, and module temperature.

To ensure accurate measurements, ten pyranometers were strategically placed to cover the entire surface area of the solar plant. These pyranometers measure the total amount of solar radiation that reaches the Earth's surface horizontally. Furthermore, five POA modules (POA1 to POA5) were used to measure the solar radiation that falls directly on the solar panels. These measurements consider various factors, such as the angle of the sun, atmospheric effects on sunlight, and shading caused by surrounding objects. POA solar irradiance is a critical factor in determining the efficiency and power output of a solar panel system.

The solar irradiation data collected span a two-year period, from July 2020 to June 2022. Table Table 2 provides a summary of the characteristics of the data and their descriptions that were recorded during this period. The dataset serves as a valuable resource for analyzing the performance and optimization of the energy output of the solar plant.

The second data file 'Sakaka_Energy_Data.csv' contains the cumulative solar power data per hour exported by two modules (PT1 and PT2), for the period from July 2020 to June 2022 as depicted in Table Table 1.

Table 2: Solar Irradiance Data

Name	Description	Unit	Type
Datetime	Timestamp	Time	Datetime
WMS_i/POA1 (i=1...5)	Irradiation module 2	Watt/ m^2	float64
WMS_i/POA2 (i=1...5)	Irradiation module 2	Watt/ m^2	float64
WMS_i/Module 1 Temp (i=1...5)	Module 1 Temperatures	Watt/ m^2	float64
WMS_i/Module 2 Temp (i=1...5)	Module 2 Temperatures	Watt/ m^2	float64
Pyranometer_i (i=1...10)	Solar radiation	Watt/ m^2	float64

Table 3: Cumulative solar power output per hour for the PT1 and PT2 modules

Name	Description	Unit	Type
Datetime	Hourly timestamp	Time	Datetime
PT1	Cumulative photovoltaic energy for module 1	KW (kilowatt)	float64
PT2	Cumulative photovoltaic energy for module 2	KW (kilowatt)	float64

The third data file 'Sakaka_Weather_Energy_Data.csv' combines the power of the photovoltaic system and the type of weather for the period from November 2020 to July 2022 as described in Table Table 4.

The file does not include the night data as the PV system does not produce any power at night. The times when the system starts and stops are given to show its operation.

Table 4: PV system power and weather type

Name	Description	Unit	Type
Datetime	Hourly timestamp	Time	Datetime
Startup time	System startup	Time	Datetime
Shutdown time	System shutdown	Time	Datetime
Energy	PV system power per hour	KW (kilowatt)	float64
Weather	Weather type: Sunny, cloudy, etc.	Text	Categorical

3.3 Features Engineering

To reduce the feature space and select the most relevant variables, we used Pearson's correlation to identify the dependencies and independencies between variables. For the first dataset of solar irradiation, we identified a very high correlation of 99% among the five values of module temperature, pyranometer, and POA modules, and this is for both modules. Therefore, we computed an average value for each of these three physical parameters and added it to the dataset, while dropping the original columns to simplify the dataset. We also set the date and time columns as the index of the dataset. The resulting dataset has the following description (Table Table 5):

Table 5: Solar data frame after feature selection

Name	Description	Unit	Type
Datetime	Hourly timestamp	Time	Datetime
POAAvg	Mean value of plane of array (POA) modules	Watt/ m^2	float64
TempAvg	Mean Value of Temperature Modules	° C	float64
IrradiationAvg	Mean Value of Pyranometers	Watt/ m^2	float64

In general, very little maintenance was required on the instruments during the campaign. The pyranometers partially malfunctioned for two periods between 10 June 2020 and 13 June 2020 and between 14 July and 18 July 2021, and the corresponding data are omitted.

Regarding the second dataset that comprises the power of the photovoltaic system (PV), it should be noted that there exists a flawless correlation between PT1 and PT2, as shown in Figure Fig. 2. Consequently, only the PT1 column is retained for the subsequent phases.

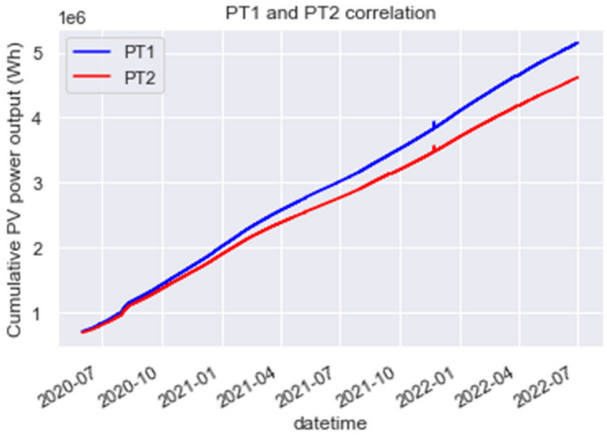


Fig. 2 Correlation between PT1 and PT2.

The hourly output energy of the photovoltaic system can be deduced from the cumulative value of PT1. Furthermore, due to an error in the system timestamping, the datetime has been shifted by 12 hours. All negative values in the dataset have been eliminated, and outliers have been filtered using the z-score method.

To improve the accuracy in predicting solar panel power output, the date, time, day of the year, and hour columns are generated from DateTime and are appended to the dataset. After collecting, cleaning, and merging the data, the attributes needed to forecast the PV power output are determined. To determine the importance of each feature in predicting PV power output, a principal component analysis (PCA) is performed. The analysis revealed that the three most significant characteristics are irradiation, temperature, and type of weather. Furthermore, the time of day, such as midday or dawn, and other factors have an impact on the output of photovoltaic power. All selected characteristics are dependent dominant variables. Numerical features include things such as temperature, day, month, and hour of the day. The type of weather is a categorical variable with the following values: cloudy, hazy, sunny, and mostly sunny. Figure Fig. 3 describes the correlation between the different features used and selected for the learning algorithms.

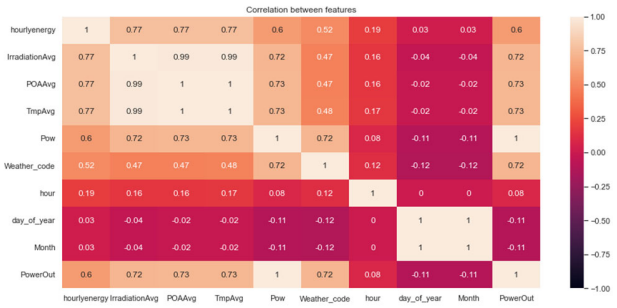


Fig. 3 Correlation between features.

3.3 Performance Metrics

This section defines the evaluation metrics to compare the learning models developed in this study.

Mean Absolute Error (MAE): MAE measures the average absolute difference between the predicted values, $\hat{y}_i$, and the actual values, y_i, without considering the direction of the error. It provides an indication of the average magnitude of the errors. The formula for calculating MAE is as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \quad (1)$$

where n is the number of samples. Lower values indicate better performance.

RMSE (Root Mean Square Error): RMSE calculates the square root of the average of the squared differences between the predicted values $\hat{y}_i$, and the actual values y_i. It considers both the magnitude and direction of the errors. The formula to calculate RMSE is as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \quad (2)$$

R^2 score (Coefficient of Determination): The R^2 score represents the proportion of variance in the dependent variable that can be explained by the independent variable(s). It ranges from 0 to 1, where a value closer to 1 indicates a better fit. The formula to calculate the R^2 score is as follows:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2} \quad (3)$$

Where $\bar{y}$ is the mean of the actual values. An R^2 score of N/A suggests that the model did not learn from the data.

4. Proposed Methodology

This section describes the proposed methodology used to evaluate and compare the different learning algorithms, as described in Figure Fig. 4.

The methodology involves several steps as follows:

(i) Data collection: Data are collected from various sources, such as Sakaka WMS (Weather Monitoring System), Sakaka Energy Data, and Sakaka Weather Data. These datasets provide the information necessary to train and test machine learning models.

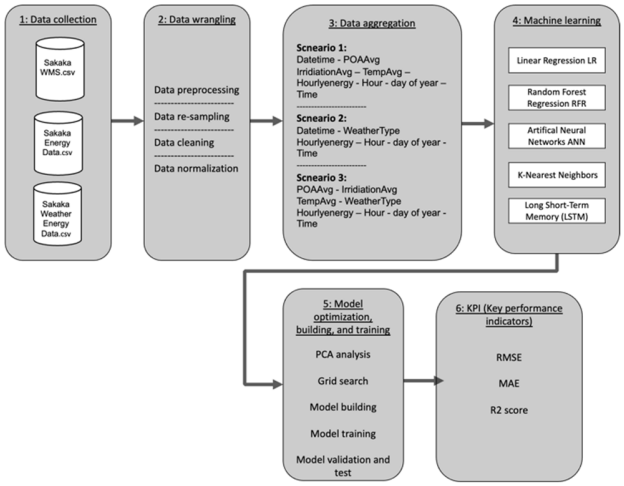


Fig. 4 Overview of the methodology.

(ii) Data Wrangling: The collected data undergoes several pre-processing steps, including cleaning, normalization, resampling, and aggregation. These steps ensure that the data are in a suitable format and ready for further analysis.

(iii) Scenario definition: Three different scenarios are defined according to the variables used for the prediction, as detailed in Table Table 6. These scenarios involve different combinations of features such as POAAvg (Plane of Array Average), IrradiationAvg (Irradiation Average), TempAvg (Temperature Average), WeatherType, HourlyEnergy, Hour, Day of Year, and Time. These scenarios provide flexibility in exploring different combinations of features to obtain a more accurate prediction.

(iv) Model Optimization, Building, and Training: The study employs several machine learning algorithms for solar prediction, including linear regression, random forest regression (RFR), artificial neural networks (ANN), closest neighbors, KNN, and long-short-term memory (LSTM). Before building the model, PCA (Principal Component Analysis) analysis and grid search are performed to optimize the model parameters. This helps to improve the model's performance and accuracy.

(v) Validation and testing of the model: The trained models are validated and tested using key performance indicators (KPIs), such as the root mean square error (RMSE), mean absolute error (MAE), and score (R^2). These metrics evaluate the accuracy and performance of the models for predicting solar energy. The models

are evaluated against a separate test dataset to ensure a valid comparison.

Table 6: Different scenarios for merging data sources

Datasets	Scenarios		
	1: Solar data	2: Meteo	3: Hybrid
Energy.csv	✓	✓	✓
WMS.csv	✓	—	✓
Weather.csv	—	—	✓
#Samples	18,263	8,679	8,679
Features	DateTime	DateTime	DateTime
	POAAvg	WeatherType	POAAvg
	IrradiationAvg	Hourlyenergy	IrradiationAvg
	TempAvg	Hour	TempAvg
	Hourlyenergy	day_of_year	WeatherType
	Hour	Time	Hourlyenergy
	day_of_year		Hour
	Time		day_of_year
			Time

5. Experimental Results

In this section, the experimental results are presented and analyzed to evaluate the different learning models. Models considered for evaluation are linear regression, random forest regression (RFR), K-nearest neighbors (KNN), artificial neural networks (ANN), and long-short-term memory (LSTM). The evaluation metrics used in the evaluation were the mean absolute error (MAE), the mean square error (RMSE), and the R^2 score.

In Table Table 7, the optimized hyperparameters are presented for each learning algorithm used in the experiments. The selection of these hyperparameters involved performing a grid search with Principal Component Analysis (PCA) to identify the best parameters for each learning algorithm.

For neural network models, the Nadam optimizer is used. It is a computationally efficient optimization solver designed for neural network algorithms. It is specifically well-suited for handling large-scale problems characterized by a substantial amount of data or parameters. Additionally, Nadam's memory requirements are minimal, making it an ideal choice for optimizing neural networks with limited computational resources. It combines the concepts of Root Mean Square Propagation (RMSprop) and Stochastic Gradient Descent with Momentum (SGDM). By merging the strengths of these two optimization methods, Nadam aims to provide an efficient and effective approach to training neural networks.

Table 7: Hyperparameters of algorithms

Algorithm	Parameters
Random Forest Regression (RFR)	Number of trees: 150
	Min sample split: 2
	Min sample leaf: 1
	Random state: 100
K Nearest Neighbors (KNN)	Number of neighbors (k): 5
Artificial Neural Network (ANN)	Number of hidden layers: 2
	Epochs: 50
	Activation function: 'relu'
	Optimizer: Nadam
	Learning rate = 0.001
Long-Short-Term Memory	Number of hidden layers: 2
	Batch: 32
	Epochs: 50
	Activation function: 'relu'
	Optimizer: Nadam
	Learning rate: 0.001

Table Table 8 shows the different results obtained for the three scenarios using the different datasets described in Section 3.

Table 8: Experimental Results of Different Machine Learning Models

Model	Scenario	MAE	RMSE	R^2 score
Linear Regression	Scenario 1	0.46	0.61	37.32%
	Scenario 2	0.86	0.97	N/A
	Scenario 3	0.37	0.26	63.08%
RFR	Scenario 1	0.06	0.15	97.73%
	Scenario 2	0.05	0.11	98.55%
	Scenario 3	0.03	0.09	**99.05%**
KNN	Scenario 1	0.08	0.20	95.52%
	Scenario 2	0.07	0.15	97.38%
	Scenario 3	0.05	0.15	97.49%
ANN	Scenario 1	0.08	0.19	95.90%
	Scenario 2	0.10	0.16	97.03%
	Scenario 3	0.05	0.11	98.71%
LSTM	Scenario 1	0.16	0.32	88.08%
	Scenario 2	0.15	0.28	90.86%
	Scenario 3	0.09	0.19	95.72%

The linear regression model produced relatively high error values (MAE: 0.46, RMSE: 0.61) and a low R^2 score (37.32%) for Scenario 1 and slightly better for Scenario 3

with an MAE of 0.37, an RMSE of 0.26 and an R^2 score of 63.08% . This is expected and is due to the nonlinear nature of the problem. The RFR model outperforms all other learning models, and this is for all scenarios. An MAE of 0.03, an RMSE of 0.09, and an R^2 score of 99.05% were obtained for scenario 3. The model also performed well in Scenarios 1 and 2.

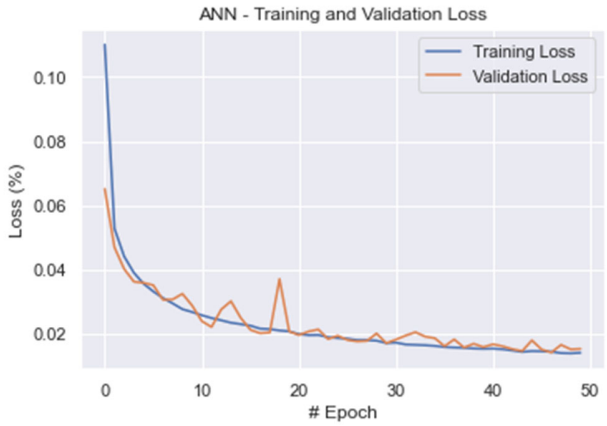


Fig. 5 Convergence of the ANN model.

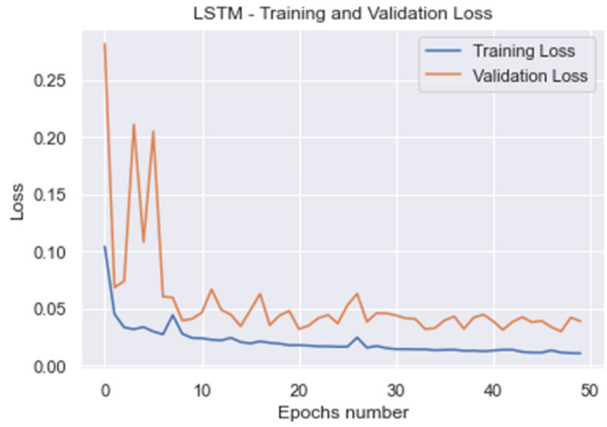


Fig. 6 LSTM model convergence.

The KNN model demonstrates good performance in general with low error values obtained for Scneario 3 (MAE = 0.05, RMSE = 0.15, and R2 score = 97. 38%).

The ANN model provided low error values (MAE: 0.08, RMSE: 0.19) and a good R^2 score (95.90%) in scenario 1 and performed slightly worse in scenario 2, with higher errors (MAE: 0.10, RMSE: 0.16) and a slightly lower R^2 score (97.03%). However, in Scenario 3, the model performed better than in the other scenarios, with lower errors (MAE: 0.05, RMSE: 0.11) and a high R^2 score (98.71%).

The LSTM model had moderate error values (MAE: 0.16, RMSE: 0.32) and a lower R^2 score (88.08%) compared to other models, with the best performance in scenario 3 (MAE: 0.09, RMSE: 0.19, R^2 score (95.72%).

In general, the Random Forest Regressor (RFR) provides the best performance for all three scenarios, consistently producing the lowest error values (MAE and RMSE) and the highest R^2 scores. The KNN and ANN models also showed good performance, while the linear regression and LSTM models had higher error values and lower R^2 scores, making them less suitable for this solar energy prediction task.

The ANN and LSTM models were trained for several epochs and quickly converged without overfitting the training data (Figures Fig. 5 and Fig. 6). The performance of the ANN model was shown to be better than that of the LSTM model by overfitting the training data and allowing for an extensive learning process.

Figure 오류! 참조 원본을 찾을 수 없습니다. shows the scatterplots of the observed energy power versus the predicted power obtained from all models for Scenario 3. The comparison reveals that the RFR model has the highest variability with an R^2 score of 99%. The ANN, LSTM, and KNN models had slightly lower variability, with R^2 scores of 0.98%, 96%, and 97.5% respectively. Furthermore, we show how KNN is too sensitive to outliers. The RFR model can handle outliers to some extent because of their ensemble nature, and thus it is less sensitive to individual data points. The ANN and LSTM models are more robust and can handle outliers better than other learning models.

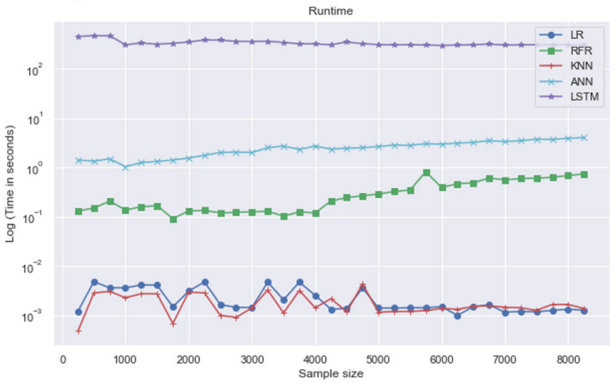
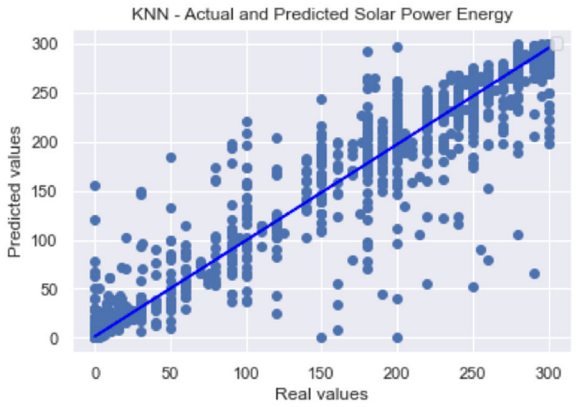
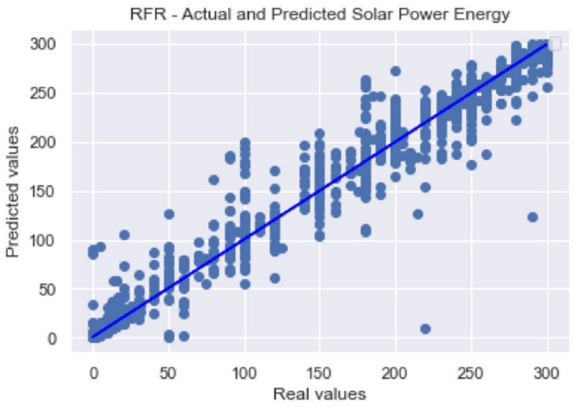
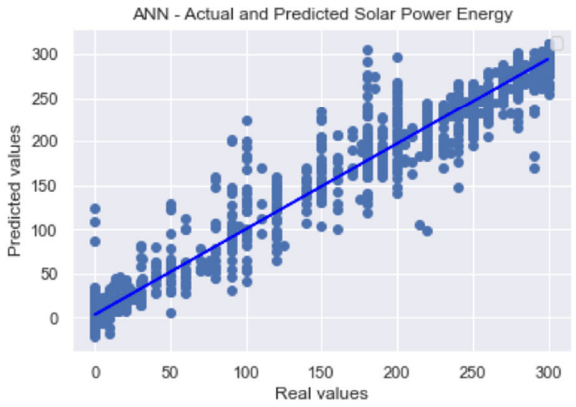


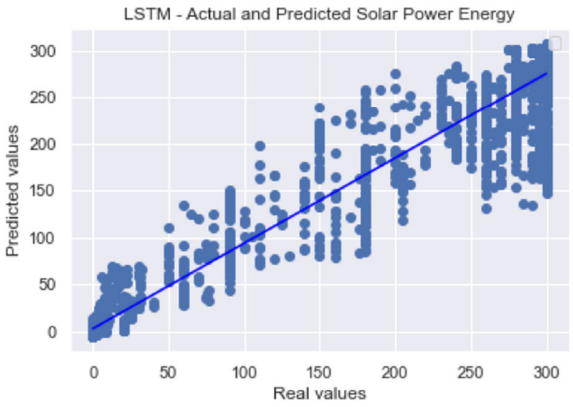
Fig. 7 Runtine experimental results.

Figure 오류! 참조 원본을 찾을 수 없습니다. shows the performance at run-time of the five algorithms. It is clear that LR and KNN have the lowest run times. However, the linear models fit the data poorly given the nonlinear nature of the problem. RFR has slightly higher runtime compared to LR and KNN but remains relatively









fast. The ability of RFR to handle complex relationships and provide accurate predictions with relatively fast computations makes it a promising choice for solar energy prediction tasks. On the other hand, ANN and LSTM have significantly higher run-times compared to the other algorithms. Both ANN and LSTM involve more complex computations due to their deep learning architectures, which require more time to process and train the models. The run-time of ANN ranges from seconds to a few minutes, while the run-time of LSTM exceeds minutes in some cases. Taking into account the runtime analysis, RFR is the most efficient algorithm in terms of both computational speed and prediction accuracy.

6. Conclusions

Forecasting the energy yield of photovoltaic systems (PV) is an effective and economical way to manage the grid and is a key factor in ensuring the dependability of the system when photovoltaic systems are widely used. In this study, a comparative analysis of machine learning models is presented to forecast solar energy generation. The analysis

The study was carried out using three years of meteorological and solar data obtained from a large-scale industrial project. The proposed methodology evaluates the performance of several machine learning models, namely linear regression, random forest regression, K nearest neighbours, artificial neural networks, and long- and short-term memory. Common performance metrics are used for the evaluation as follows: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Coefficient of Determination R^2. Different scenarios have been defined to measure the effect of training features, training data sequence, and data quality on prediction accuracy.

The experimental results revealed that the Random Forest Regression model outperformed other learning models for all scenarios. It produced an MAE error of 0.06 when using solar irradiance data and of 0.03 when using the hybrid regime that combines solar data with meteorological data. Although artificial neural networks exhibit strong performance, they require more comprehensive hyperparameter adjustments and a greater demand for computational resources.

Furthermore, the combination of meteorological data with solar irradiance data further improved the prediction accuracy of all models and scenarios by at least 0.03.

Acknowledgment

The authors express their gratitude to SAKAKA Solar Energy Company for furnishing the essential data for the experiments. The generous support they have received has been indispensable for the completion of this study.

References

[1] Mahoor, M.; Majzoobi, A.; Khodaei, A. Distribution asset management through coordinated microgrid scheduling. *IET Smart Grid* **2018**, *1*, 159–168, [https://ietresearch.onlinelibrary.wiley.com/doi/pdfstg.2018.0076].https://doi.org/https://doi.org/10.1049/ietstg.2018.0076.

[2] Sadeghian, H.; Wang, Z. A novel impact-assessment framework for distributed PV installations in low-voltage secondary networks. *Renewable Energy* **2020**, *147*, 2179–2194. https://doi.org/10.1016/j.renene.2019.09.117.

[3] Sekhar, V. Saudi Arabia Vision 2030: Solar Energy Can Complement, Not Rival, Oil And Gas. *Arab News* **2020**. https://www.arabnews.com/node/1708961/amp.

[4] Chuluunsaikhan, T.; Nasridinov, A.; Choi, W.S.; Choi, D.B.; Choi, S.H.; Kim, Y.M. Predicting the Power Output of Solar Panels based on Weather and Air Pollution Features using Machine Learning. *Journal of Korea Multimedia Society* **2021**, *24*, 222–232.

[5] Almonacid-Olleros, G.; Almonacid, G.; Fernandez-Carrasco, J.I.; Espinilla-Estevez, M.; Medina- Quero, J. A new architecture based on IoT and machine learning paradigms in photovoltaic systems to nowcast output energy. *Sensors* **2020**, *20*, 4224.

[6] Sayyah, A.; Horenstein, M.N.; Mazumder, M.K. Energy yield loss caused by dust deposition on photovoltaic panels. *Solar Energy* **2014**, *107*, 576–604. https://doi.org/https://doi.org/10.1016/j.solener.2014.05.030.

[7] Lee, D.; Kim, K. Recurrent neural network-based hourly prediction of photovoltaic power output using meteorological information. *Energies* **2019**, *12*, 215.

[8] Su, D.; Batzelis, E.; Pal, B. Machine learning algorithms in forecasting of photovoltaic power generation. *2019 International Conference on Smart Energy Systems and Technologies (SEST)* **2019**, pp. 1–6.

[9] Da Silva, I.N.; Spatti, D.H.; Flauzino, R.A.; Liboni, L.H.B.; Alves, S.F.d.R. *Artificial neural networks*; Vol. 39, Springer International Publishing: Cham, **2017**.

[10] Saberian, A.; Hizam, H.; Radzi, M.; Kadir, M.; Mirzaei, M. Modeling and prediction of photovoltaic power output using artificial neural networks. *International Journal of Photoenergy* **2014**.

[11] Lee, D.; Kim, K. Recurrent neural network-based hourly prediction of photovoltaic power output using meteorological information. *Energies* **2019**, *12*, 215.

[12] Ozoegwu, C.G. Artificial neural network forecast of monthly mean daily global solar radiation of selected locations based on time series and month number. *Journal of Cleaner Production* **2019**, *216*, 1–13. https://doi.org/https://doi.org/10.1016/j.jclepro.2019.01.096.

[13] Ozdemir, T.; Taher, F.; Ayinde, B.O.; Zurada, J.M.; Tuzun Ozmen, O. Comparison of Feedforward Perceptron Network with LSTM for Solar Cell Radiation Prediction. *Applied Sciences* **2022**, *12*. https://doi.org/10.3390/app12094463.

[14] Nageem, R.; Jayabarathi, R. Predicting the power output of a grid-connected solar panel using multi-input support vector regression. *Procedia computer science* **2017**, *115*, 723–730.

[15] Kabilan, R.; Chandran, V.; Yogapriya, J.; Karthick, A.; Gandhi, P.; Vinayagam, M.; Rahim, R.; Subramanian, D.M. Short-Term Power Prediction of Building Integrated Photovoltaic (BIPV) System Based on Machine Learning Algorithms. *International Journal of Photoenergy* **2021**, *2021*. https://doi.org/10.1155/2021/5582418.

[16] Czajkowski, M.; Kretowski, M. The role of decision tree representation in regression problems– An evolutionary perspective. *Applied Soft Computing* **2016**, *48*, 458–475.

[17] Villegas-Mier, C.G.; Rodriguez-Resendiz, J.; Alvarez-Alvarado, J.M.; Jimenez-Hernandez, H.; Odry, A. Optimized Random Forest for Solar Radiation Prediction Using Sunshine Hours. *Micromachines* **2022**, *13*. https://doi.org/10.3390/mi13091406.

[18] Salehinejad, H.; Sankar, S.; Barfett, J.; Colak, E.; Valaee, S. Recent advances in recurrent neural networks. *arXiv preprint arXiv:1801.01078* **2017**.

[19] Kumari, P.; Toshniwal, D. Long short term memory–convolutional neural network based deep hybrid approach for solar irradiance forecasting. *Applied Energy* **2021**, *295*, 117061. https://doi.org/10.1016/j.apenergy.2021.117061.

[20] Lee, W.; Kim, K.; Park, J.; Kim, J.; Kim, Y. Forecasting Solar Power Using Long-Short Term Memory and Convolutional Neural Networks. *IEEE Access* **2018**, *PP*, 1–1. https://doi.org/10.1 109/ACCESS.2018.2883330.

[21] Sharma, J.; Soni, S.; Paliwal, P.; Saboor, S.; Chaurasiya, P.K.; Sharifpur, M.; Khalilpoor, N.; Afzal, A. A novel long term solar photovoltaic power forecasting approach using LSTM with Nadam optimizer: A case study of India. *Energy Science & Engineering* **2022**, *10*, 2909–2929.

[22] Theocharides, S.; Makrides, G.; Livera, A.; Theristis, M.; Kaimakis, P.; Georghiou, G.E. Day- ahead photovoltaic power production forecasting methodology based on machine learning and statistical post-processing. *Applied Energy* **2020**, *268*, 115023. https: doi.org/10.1016/j.apenergy.2020.115023.

[23] Liu, S.; Huang, Z.; Chen, G.; Li, X.; Li, Y.; Chen, X. Hybrid model combining artificial neural networks and support vector regression for solar power forecasting. *Applied Energy* **2019**, *242*, 766–775.

[24] Theocharides, S.; Theristis, M.; Makrides, G.; Kynigos, M.; Spanias, C.; Georghiou, G.E. Comparative Analysis of Machine Learning Models for Day-Ahead Photovoltaic Power Production Forecasting. *Energies* **2021**, *14*, 1081. https://doi.org/10.3390/en14041081.

[25] Guanoluisa, R.; Arcos-Aviles, D.; Flores-Calero, M.; Martinez, W.; Guinjoan, F. Photovoltaic power forecast using deep learning techniques with hyperparameters based on Bayesian optimization: A Case Study in the Galapagos Islands. *Sustainability* **2023**, *15*, 12151. https://doi.org/10.3390/su151612151.

[26] Zhang, L.; Zhao, Y.; Wang, J.; Chen, Y. Solar Power Forecasting Using Random Forest Algorithm. *Energies* **2018**, *11*, 610.
[27] Li, L.L.; Wen, S.Y.; Tseng, M.L.; Wang, C.S. Renewable energy prediction: A novel short-term prediction model of photovoltaic output power. *Journal of Cleaner Production* **2019**, *228*, 359–375. margin of safety," 550/9-74-004, 1974.

Khaled Chaaban is an associate professor at the College of Computer & Information Systems in Umm Al-Qura University in Makkah Saudi Arabia. He received his PhD degree from UTC University (Compiegne, France) in 2006. His main research interests include embedded real-time systems optimization, design space exploration, and model-based design for system engineering. He has published over 30 peer-reviewed papers in international journals and conferences.

Najd Alfadl Najd is currently a student at Riyadh Schools. She has been appointed as a student representative on the SDG4 Youth & Student Network for the Arab States. Notably, Najd has demonstrated academic excellence by achieving top positions in various national research competitions, including the Scientific Research Olympiad, KACST RSI, and Samsung Solve for Tomorrow. Her outstanding research proposals have earned her multiple grants, highlighting her dedication and success in the field of research.