Machine learning based models for solar energy

Dalila Cherifi[1], Abdeldjalil Dahbi[2,3], Mohamed Lamine Sebbane[1], Bassem Baali[1], Ahmed Yassine Kadri[3], Messaouda Chaib[3]

[1]Institute of Electrical and Electronic Engineering, University of Boumerdes, Boumerdes, Algeria
[2]Unité de Recherche en Energies Renouvelables en Milieu Saharien (URERMS), Centre de Développement des Energies Renouvelables (CDER), Adrar, Algeria
[3]Laboratory of sustainable Development and computing, (L.D.D.I), University of Adrar, Adrar, Algeria

Article Info

Article history:

Received Aug 24, 2024
Revised Dec 25, 2025
Accepted Jan 22, 2026

Keywords:

Machine learning
Photovoltaics
Power forecasting
Solar generation
Weather conditions

ABSTRACT

Photovoltaic (PV) technology is one of the most promising forms of renewable energy. However, power generation from PV technologies is highly dependent on variable weather conditions, which are neither constant nor controllable, which can affect grid stability. Accurate forecasting of PV power production is essential to ensure reliable operation within the power system. The primary challenge of this study is to accurately predict photovoltaic energy production, considering that weather conditions, such as irradiance, temperature, and wind speed, are random variables. The key contribution of this article is developing a machine learning model to predict the energy production of a real PV power plant in Algeria. Using real measurements sourced from the Center of Renewable Energy Development (CDER) in Adrar, Algeria, in 2021. The data are from two PV power plants located in harsh desert climate conditions. The results presented in this study offer a comparison of several predictive methods applied to real-world data from a PV power plant situated in the Saharan Region. Our findings reveal that the artificial neural network (ANN) model yields the most accurate predictions of 94.96%, with the smallest prediction error: root mean square (RMSE) and mean absolute error (MAE) are 7.78% and 3.80%, respectively.

Corresponding Author:

Dalila Cherifi
Institute of Electrical and Electronic Engineering, University of Boumerdes
Boumerdes, Algeria
Email: da.cherifi@univ-boumerdes.dz

1. INTRODUCTION

Solar energy is one of the most promising sources for generating power for residential, commercial, and industrial applications. This is particularly true given that the cost of solar modules continues to decrease, in contrast to the rising costs of energy generation from fossil fuels and other polluting sources. Therefore, it is becoming more practical to use renewable energy resources such as solar energy, which can convert solar irradiance into electric energy through the photovoltaic effect [1], [2]. Energy generated by photovoltaic (PV) systems is directly influenced by geographical and weather conditions such as solar irradiance, temperature, and site-specific factors [3], [4]. However, the variability of PV output power poses significant challenges to the power grid's operation, including issues related to system stability, reliability, and electric power balance. To facilitate effective decision-making and ensure grid stability, solar PV power forecasting has emerged as a crucial solution to these issues. Accurate forecasting of PV power helps reduce the impact of output uncertainty on the grid, making the system more reliable and efficient while maintaining power quality.

Previous research has advanced significantly in PV power forecasting, with various approaches proposed. Antonanzas *et al.* [5] provide a comprehensive review of photovoltaic forecasting methods, covering physical, statistical, and machine-learning approaches, and underline the importance of accurate forecasts for reliable grid operation. Al Amin and Hoque [6] applied ARIMA models for short-term predictions, obtaining moderate accuracy but facing challenges with non-linear weather effects.

Machine learning (ML) techniques have shown significant contributions in overcoming these challenges, offering potential improvements in terms of accuracy and reliability compared to traditional methods. The objective of this article is to develop machine learning techniques to generate models and mathematical relationships that can forecast energy generation, as solar photovoltaic systems are subject to fluctuations and weather dependence. Despite these advances, challenges remain in generalizing models across different climates and optimizing efficiency, which this study aims to address. Our study is based on PV power generation data collected over one year at 30-minute intervals from two locations in Algeria: Kabereten (Adrar) and El Hadjira (Ouargla). Using this dataset, we developed four machine learning models: linear regression, polynomial regression, support vector regression (SVR), and artificial neural networks (ANN). We analyze and preprocess the data to optimize the performance of the model and then compare the performance of various models to identify the most effective approach for power prediction.

The article consists of three sections: i) The first section introduces PV systems and the various factors that can affect their performance, emphasizing the importance of accurate PV power forecasting in the energy industry; ii) The second section explores commonly used techniques for PV power prediction, providing an overview of the machine learning models used in this study, along with theoretical information and evaluation metrics; and iii) The third section presents the datasets used in our research, describing the pre-processing and feature engineering steps taken to ensure their suitability for analysis. This section also presents the study's results and findings, followed by a comprehensive discussion of the results.

2. RELATED WORK

Many research efforts have focused on providing more accurate forecasts for solar power generation. ML and artificial intelligence (AI) forecasting models offer the advantage of directly predicting PV power without the intermediary step of forecasting solar irradiance. This approach also provides flexibility in forecasting horizons. Most state-of-the-art forecasting models use ANN, regression models, and support vector machines (SVM). These data-driven techniques leverage historical observations to train models, enabling them to compute predictions by analyzing past values of input variables [7], [8]. Consequently, power output can be directly predicted based on the input variables used. Table 1 summarizes the current studies in the literature that are closest to the method proposed in our work for predicting solar power generation. These studies used different datasets and locations than ours, along with varying preprocessing and algorithmic techniques.

According to Table 1, previous research indicates that PV power generation primarily depends on meteorological factors such as irradiance, temperature, wind speed, and relative humidity. The current flow through solar cells increases significantly with higher irradiance, leading to a rise in power output [9]. Higher temperatures can reduce panel efficiency by decreasing power output as the voltage drops with increasing temperatures [10]. Higher wind speeds can lower air and solar cell operating temperatures, enhancing the efficiency of a solar PV system [10]. Increases in relative humidity can significantly decrease PV voltage; low relative humidity improves efficiency, while high relative humidity reduces it [11]. Due to the intrinsic nature of these factors, the output power is variable and uncertain, resulting in unstable fluctuations [12]. Previous work focused primarily on the development of solar PV power output forecasting models using traditional statistical and physical approaches, as well as machine learning techniques such as linear regression, polynomial regression, SVR, ANN, long short-term memory (LSTM), and convolutional neural networks (CNN)-LSTM. Although these studies attempted to achieve higher accuracy by applying various input parameters (e.g. temperature, solar radiation, wind speed) at multiple locations, they were likely to miss the nonlinear nature of weather-dependent PV power generation, especially in fast-changing environments. Furthermore, most studies focused on a single method or did not include a comparative study of multiple machine learning models under similar conditions, and little attention was given to localized case studies in regions such as Algeria, where environmental conditions can directly affect PV performance. In this study, we close these gaps by suggesting and contrasting four ML models, that is: linear regression, polynomial regression, SVR, and ANN, using real data for two areas in Algeria. This study provides a deeper understanding of the nature of the model under different climatic conditions and suggests better forecasting methods based on localized PV systems.

Table 1. An overview of methods employed in PV power prediction

Authors	Location	Data parameters	Method	Accuracy	Error	
					MAE	RMSE
Verma *et al.* [13]	India	Temperature	Linear regression	74.4%	6%	/
		Cloud cover	Logarithmic regression	47.4%	15%	/
		Wind speed	Polynomial regression	75.1%	6.1%	/
		Humidity	ANN	92%	3%	/
		Rainfall				
Kuriakose *et al.* [14]	India	Solar radiance	ANN	80.97%	6.53%	/
		Temperature	Linear regression	83.21%	6.66%	/
		Wind speed	SVR	83.88%	6.74%	/
		Relative humidity				
Abuella and Chowdhury [15]	USA	Temperature	ANN	97.09%	/	5.54%
		Cloud cover	MLR	96.98%	/	5.71%
		Pressure				
		Humidity				
		Wind component				
		Solar radiation				
		Thermal radiation				
		Net solar radiation				
		Liquid water				
		Ice water				
Aslam *et al.* [16]	Germany	Day	LSTM	86.8%	3.57%	7.07%
		Temperature	LSTM-attention	86.44%	3.67%	7.2%
		Wind	CNN-LSTM	85.25%	3.78%	7.38%
		Sky cover	Ensemble method	87.4%	3.69%	6.85%
		Humidity				
		Precipitation				
Uddin *et al.* [17]	Indonesia	Radiation	K-NN	64.9%	/	/
		Air temperature				
		Wind speed				
		Sunshine (minutes)				
		Air humidity				
		Air pressure				

3. METHODOLOGY

This section focuses on the machine learning models used for PV power forecasting [18]–[20]. We examine both linear and non-linear models, evaluating their performance and complexity. The models are organized in a hierarchy, from the simplest to the most complex, to identify the most suitable approach for accurate and reliable PV power forecasting. Specifically, we employed four models:

a) Linear regression: Assumes a linear relationship between a variable of input weather parameters and dependent variables [21].

b) Polynomial regression: Allows for modeling non-linear relationships between variables as nth-degree polynomials [22]. We have tested different polynomial degrees from n = 0 to n = 10 in order to find the optimal degree that fits the data to avoid overfitting while effectively capturing the underlying patterns in the data.

c) SVR: Outputs an optimal hyperplane with at most ε deviation to perform regression tasks, fitting the error within a threshold [23]. SVR excels at modeling intricate, non-linear relationships using kernel functions that transform the input space into higher-dimensional feature spaces. In this work, we studied three distinct SVR kernel functions: a linear kernel, a polynomial kernel, and a radial basis function (RBF) kernel. We found that the linear and polynomial kernels performed poorly compared to the RBF kernel. As a result, we focused on testing SVR using the RBF kernel on our two datasets.

d) ANN: Mimics brain neurons and excels at learning patterns from training data to predict output variables. It consists of layers of interconnected nodes: an input layer, one or more hidden layers for processing, and an output layer [24]. The nodes are interconnected, with the input layer containing a number of nodes equal to the dataset's features and only one output node.to introduce non-linearity into the model, we use an activation function allowing it to learn complex patterns. In our experiment, we used the linear activation function to test the performance of the model and we approved the rectified linear unit (ReLU) activation function to capture the non-linearity in the results obtained.

4. EXPERIMENTS AND RESULTS

In this section, we delve into the datasets used for predicting the power output of two solar power plants, analyzing the relationships between various environmental factors and power output. We also detail the different experiments conducted, including parameter selection and evaluation of predictive algorithms, to identify the most suitable approach for achieving accurate and reliable PV power forecasting.

4.1. Data description and analysis

The methodology begins with the collection of solar energy data from the Renewable Energies Research Unit in Saharan Environment (URERMS), Center of Renewable Energy Development (CDER), covering two PV power plants in Algeria. The raw data went through a cleansing process, where negative and missing values were processed to maintain integrity. Following this, exploratory data analysis was used to identify correlations and patterns, which led the feature selection process. Key features such as solar irradiance, temperature, wind speed, and relative humidity were selected based on their relevance to PV performance. Multiple regression and machine learning models including linear regression, polynomial regression, SVR, and ANN were trained using a 70/15/15 data split for training, validation, and testing. Model performance was evaluated using metrics like mean absolute error (MAE), root mean square error (RMSE), and the coefficient of determination (R²). Figure 1 illustrates the overall workflow adopted in this study for solar energy prediction using machine learning models. This structure ensured that each model was evaluated on consistent and reliable data, providing an accurate comparison of predicted accuracy.

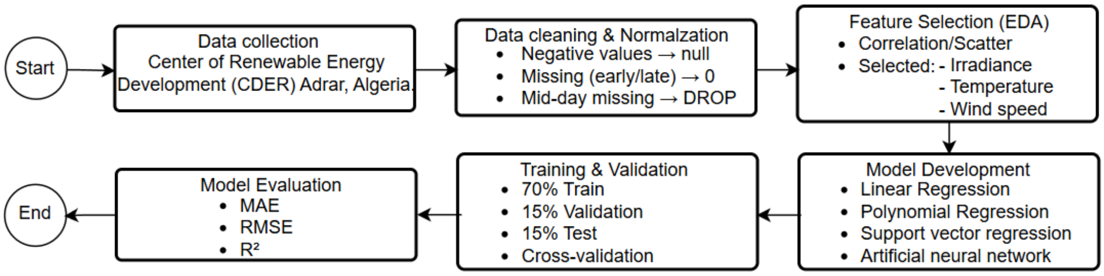


Figure 1. Workflow for proposed solar energy prediction using machine learning models

4.1.1. Data collection

The data used in the models and ANN were collected from a meteorological weather station installed at the PV power plant site. This data was carefully processed to ensure its suitability for the proposed application in the Algerian energy market. Given that the datasets are based on real measurements from an actual PV power plant operating in desert climate conditions, the results obtained are highly relevant and can serve as a valuable reference for similar applications in other PV power plants within Saharan regions. The meteorological data was sourced from Renewable Energies Research Unit in Saharan Environment (URERMS), CDER Adrar, Adrar, Algeria for the year 2020. The dataset includes half-hourly measurements gathered by multiple sensors connected to PV systems at two stations located in Ouargla, Algeria, and Adrar, Algeria. The first dataset is sourced from the Kaberetene photovoltaic power plant in Adrar, which spans 6 hectares with a capacity of 3 MWp. Located near Ksar Kabertene, about 60 km from the wilaya of Adrar, Algeria (31° 50' N, 0° 78' E), the facility comprises three sub-fields, each with a 1 MWp capacity. It uses 93 matrices, each containing 44 panels organized into 2 strings of 22 panels connected in series.

The second dataset comes from the El Hadjira PV power plant in Ouargla, which covers 60 hectares with a capacity of 30 MW. Situated near El Hadjira, about 99 km from the wilaya of Ouargla, Algeria (32.6016° N, 5.8339° E), the plant consists of 30 subfields, each equipped with polycrystalline silicon modules. Each subfield generates 1 MWp, housing 4004 modules organized into 91 strings of 44 modules each. Each module is rated at 250 W with an efficiency of 15%. The photovoltaic field array data from both plants contain time-series data collected by several sensors linked to the PV systems, measured in 2020 at 30-minute intervals from 6:00 AM to 8:00 PM. The Kaberetene dataset contains 10,364 entries, while the El Hadjira dataset contains 9,570 entries. Both datasets include 7 columns or features: total power (kW), TSA, R Globale (W/m²), temperature

(°C), wind speed (m/s), humidity (%), and pressure (HPA). The dataset is highly relevant to the Algerian energy market, as it includes data from the southern region of Algeria which is known for its strong solar irradiance, the dataset represents a variety of weather conditions in desert environments, where solar energy can vary significantly. This data is crucial for evaluating solar energy potential, improving forecasting models, and optimizing renewable energy integration into Algeria's grid, helping to reduce reliance on fossil fuels. The dataset's half-hourly resolution allows for a detailed analysis of energy generation which is crucial for improving the integration of solar power into the national grid.

4.1.2. Data exploration and preprocessing

We began with thorough data exploration and pre-processing to ensure data quality and suitability for regression modeling. This exploratory data analysis (EDA) involved understanding distributions and relationships using statistical analyses and visualizations to define the most appropriate forecasting models for our dataset. A scatter plot matrix and a correlation matrix were created to identify the most relevant input variables for modeling power output. Irradiance showed a strong correlation with power output: when irradiance is high, power is likely to be high as well. However, at low irradiance levels, there was more significant variation in power values. This observation aligns with the PV cell working principle, suggesting that a linear regression model would be appropriate for predicting power based on irradiance. Temperature and wind speed demonstrated moderate correlations with power, indicating complex (non-linear) relationships. Relative humidity had a high negative correlation with temperature, as increased humidity can lead to precipitation and subsequently lower ambient temperatures. Pressure showed nearly zero correlation with power output and was therefore excluded from further analysis.

Based on this analysis, irradiance, temperature, wind speed, and relative humidity were chosen as the input variables due to their direct or indirect effects on PV cell performance and power output. The dataset contained negative values for power and solar irradiance, which were measured during the night when there is no solar irradiance, and power is drawn from the battery or grid. These values were set to null to sanitize the data. Missing values were found in solar radiation and power data during the early and late hours of the day, likely due to sensor offsets and inverter failures. These were set to zero. For missing values during mid-day periods, likely due to sensor or inverter breakdowns, those data points were excluded from processing to ensure accurate analysis. Finally, the dataset was split into 70% training, 15% validation, and 15% testing sets, with techniques like cross-validation employed to ensure robust model evaluation.

4.2. Model evaluation: performance metrics

Performance metrics are statistical measures used to evaluate the effectiveness of a model. They offer a means of evaluating a model's efficacy by contrasting its forecasts with actual outcomes. The effectiveness of the method is determined by the error between the actual output power values and the predicted values, with the most accurate method being the one that produces the smallest error. We analyzed and compared machine learning-based forecasting methods for PV power generation. The evaluation criteria we defined include error rates, specifically MAE, RMSE, and R² score. These metrics offer a comprehensive assessment of the methods' effectiveness and applicability [25]. The advantage of utilizing MAE loss function lies in providing the average size of the error in the target variable's units, making it simple to analyze and comprehend. The RMSE is calculated as the square root of the average of the squared differences between the actual and predicted values. R² score indicates goodness of fit, therefore measures how well unseen samples are likely to be predicted by the model, through the proportion of explained variance.

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y_i}|}{n} \tag{1}$$

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{|y_i - \hat{y_i}|^2}{n}} \tag{2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} |y_i - \hat{y_i}|^2}{\sum_{i=1}^{n} |y_i - \bar{y_i}|^2} \tag{3}$$

Where y_i represents the actual value,$\hat{y_i}$ denotes the predicted value, $\bar{y}_i$ is the mean of the actual values, and n is the total data points number. Values closer to 1.00 indicate a better model, noting that it can be negative (because the model can be arbitrarily worse).

4.3. Experiments

The objective of our study is to predict the power output of a solar power plant based on various weather factors. Regression analysis is well-suited for this task because it quantitatively captures the relationships between the input variables (irradiance, temperature, wind speed) and the output variable (power). Given the linear relationship between power and irradiance where power output increases proportionally with irradiance we suggest that a linear regression model would be appropriate for predicting power. Although temperature and wind speed have a nonlinear relationship with power, irradiance is considered the dominant feature. Note that the models developed are applied to two datasets.

4.3.1. Experiment 1: Power modeling using linear regression

Linear regression was employed to model and predict the amount of solar power generated based on various weather-related features. It was utilized to model and predict the amount of solar power generated based on various weather-related features. The process began with data scaling, where the input datasets: X_{train}, X_{val}, and X_{test}, along with their corresponding target vectors, Y_{train}, Y_{val}, and Y_{test}, were preprocessed to ensure consistency in feature magnitudes. A linear regression() was initialized and trained on the scaled training data. Following training, the model was used to generate predictions for the training, validation, and testing sets. To evaluate the models performance and its ability to generalize to new data, several metrics were calculated, including MAE, RMSE, and the coefficient of determination (R^2 score). By modeling power output as a linear function of irradiance and temperature, we gained initial understandings into the data and the relationships between variables. The linear function with two inputs was learned as (4).

$$Power = 136.84 + 2518.19 \cdot irradiance - 336.52 \cdot temperature \tag{4}$$

We added wind speed as a third input variable to our model in order to handle outliers and improve its predictive accuracy. This additional variable was expected to improve the model's performance and reduce differences between the actual and predicted values by capturing complex interactions affecting power output. The linear function with three inputs was learned as (5).

$$Power = 122.66 + 2511.72 \cdot irradiance - 335.43 \cdot temperature + 49.45 \cdot windspeed \tag{5}$$

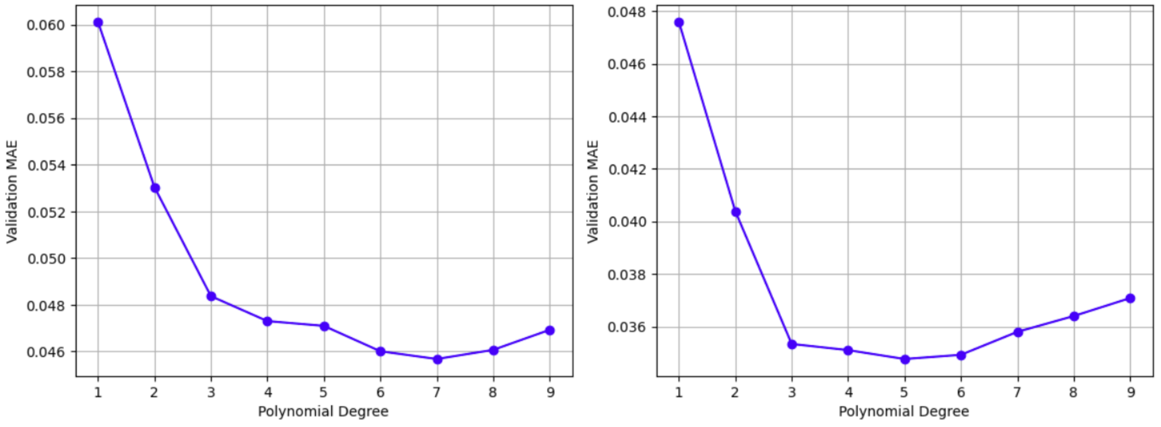
The performance metrics for the linear regression models with two and three inputs are summarized in Table 2. Although the three-input model demonstrates greater accuracy and less error compared to the two-input model. the linear regression model failed to capture the non-linear relationship of the power with both temperature and wind speed. Therefore to enhance the accuracy of our model, given the complexity observed in the relationships between the input variables and power output, we have selected polynomial regression.

Table 2. Performance metrics for linear regression models with two and three inputs

Metrics	Kaberetene			El Hadjira		
	Training set	Validation set	Testing set	Training set	Validation set	Testing set
RMSE (2 inputs)	8.98%	8.40%	8.61%	8.63%	7.35%	8.67%
RMSE (3 inputs)	8.98%	8.40%	8.59%	8.62%	7.33%	8.66%
MAE (2 inputs)	5.42%	6.02%	5.04%	5.41%	4.77%	5.76%
MAE (3 inputs)	5.42%	6.01%	5.03%	5.41%	4.75%	5.76%
R² (2 inputs)	91.95%	93.65%	92.29%	92.81%	93.99%	93.27%
R² (3 inputs)	91.97%	93.66%	92.31%	92.82%	94.02%	93.30%

4.3.2. Experiment 2: Power modeling using logostic (polynomial) regression

Polynomial regression's ability to capture non-linear relationships and interactions effectively, this approach allows for modeling non-linear patterns as nth-degree polynomials by incorporating higher-order and interaction terms, providing a more accurate fit for the data. Input features were transformed into polynomial features using polynomial regression(), where n is the degree of the polynomial. The validation MAE was computed for each degree, and the best-performing degree was selected as the optimal model. The model was retrained on the training data using the best degree, and performance was evaluated on the test set using the following metrics: MSE, MAE, and R2 score for training, validation, and test sets. The optimal polynomial degree is 7 for the Kaberetene dataset and 9 for the El Hadjira dataset as shown in Figure 2, where we perform different polynomial degrees to identify the optimal degree. It is observed that those polynomial degrees provided the best balance between model complexity and prediction accuracy. The performance of the selected polynomial regression models was evaluated using several metrics. The results are summarized in Table 3. Based on the result obtained, it is noticed that polynomial regression shows better accuracy than the linear regression model for the two data sets and its flexibility to deal with the complexity of the temperature and wind speed. To improve the accuracy we have used new models such as support vector regression and see this model can capture the complex relationship better than polynomial regression.



(a) (b)

Figure 2. Validation MAE vs polynomial degree for (a) Keberatene dataset and (b) El Hadjira dataset

Table 3. Performance metrics for polynomial regression models with two and three inputs

Metrics	Kaberetene			El Hadjira		
	Training set	Validation set	Testing set	Training set	Validation set	Testing set
RMSE (2 inputs)	8.13%	7.42%	7.94%	7.67%	6.49%	7.61%
RMSE (3 inputs)	8.06%	7.43%	8.35%	7.69%	6.49%	7.62%
MAE (2 inputs)	3.83%	4.59%	3.59%	3.76%	3.44%	4.11%
MAE (3 inputs)	3.81%	4.56%	3.77%	3.81%	3.47%	4.21%
R^2 (2 inputs)	93.41%	95.04%	93.43%	94.31%	95.31%	94.81%
R^2 (3 inputs)	93.52%	95.03%	92.75%	94.28%	95.32%	94.81%

4.3.3. Experiment 3: Power modeling using SVR

SVR is an approach that handles non-linearity and complex relationships similar to polynomial regression. To reduce training time, a subset of 1000 samples was randomly selected from the scaled training set using resample() with a fixed random seed. A randomized search was performed over the following parameter grid: Kernel = 'linear', 'rbf', Regularization parameter C = 1, 10, 100, Kernel coefficient γ = 'scale', 0.01, 0.1, Epsilon ε = 0.1, 0.5. The search tested 10 random combinations using 3-fold cross-validation, optimized for negative mean squared error. In this model, the best estimator was selected based on the lowest average validation error across folds. The resulting optimal parameters were the RBF as the kernel function we established the optimal parameters to be C = 10, γ = 0.01, and ε = 0.1. Using these parameters, the following table provides an overview of the predictive performance of the SVR model across different datasets. From Table 4, we have observed a low validation accuracy compared to training accuracy, which means that there is overfitting, basically the model has learned the training data very well and failed to capture the underlying patterns effectively, leading to poor predictive performance. To address overfitting and low performance of support vector regression (SVR), we have applied the artificial neural networks (ANN) model to get better performance and accuracy.

Table 4. Performance metrics for SVR models with two and three inputs

Metrics		Kaberetene			El Hadjira	
	Training set	Validation set	Testing set	Training set	Validation set	Testing set
RMSE (2 inputs)	9.02%	34.91%	8.44%	8.60%	33.28%	8.69%
RMSE (3 inputs)	9.03%	34.91%	8.4%	8.62%	33.28%	8.73%
MAE (2 inputs)	5.89%	30.93%	5.30%	5.76%	28.86%	6.22%
MAE (3 inputs)	5.95%	30.93%	5.34%	5.83%	28.86%	6.31%
R² (2 inputs)	91.89%	36.97%	92.58%	92.86%	34.02%	93.24%
R² (3 inputs)	91.87%	39.97%	92.64%	92.82%	37.21%	93.18%

4.3.4. Experiment 4: Power modeling using ANNs

ANNs allow the modeling of intricate variables through multiple layers of neurons through neural network architecture, We aim to achieve improved predictive performance and generalization. It contains an input layer with 2 or 3 input variables, two hidden layers with 64 Neurons and 32 Neurons respectively, and an output layer. The ReLU activation function is used to capture the complex relationships in the data. The model was trained using the full scaled training dataset with the following hyperparameters: Epochs: 100, Batch size: 32, Validation set: A separate validation split was used during training to monitor generalization. The model was evaluated on the training, validation, and test datasets using the following metrics: MSE, MAE, R2, as shown in Table 5 summarize the performance metrics of our ANN model. These metrics provide a detailed overview of the model's accuracy and generalization capabilities across the training, testing, and validation sets. The performance metrics tables indicate that the ANN model successfully captures complex relationships. Its architecture allows learn from intricate patterns, enhancing predictive accuracy and reducing error across different datasets compared to other models.

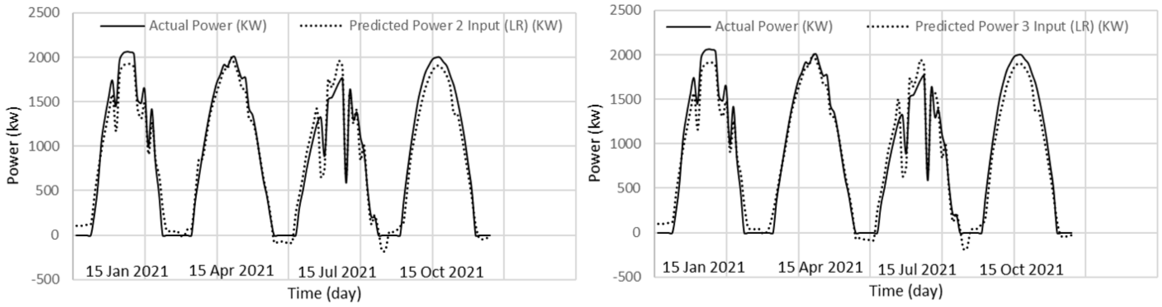
Table 5. Performance metrics for ANNs models with two and three inputs

Metrics		Kaberetene			El Hadjira	
	Training set	Validation set	Testing set	Training set	Validation set	Testing set
RMSE (2 inputs)	8.41%	8.27%	7.78%	7.89%	3.12%	8.14%
RMSE (3 inputs)	8.16%	7.09%	8.22%	7.63%	6.49%	7.50%
MAE (2 inputs)	4.51%	5.55%	3.80%	4.25%	6.12%	4.80%
MAE (3 inputs)	3.58%	4.18%	3.69%	3.57%	3.38%	3.91%
R² (2 inputs)	92.94%	93.85%	93.69%	93.98%	95.83%	94.08%
R² (3 inputs)	93.36%	95.47%	92.97%	94.37%	95.31%	94.96%

4.3.5. Experiment 5: Power testing using different modules

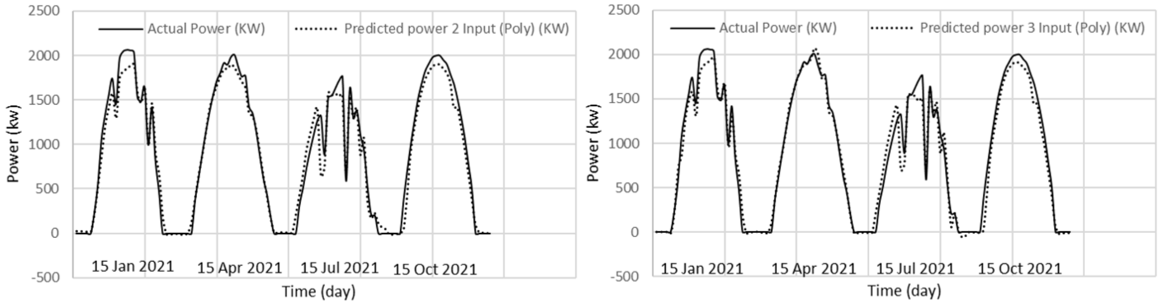
In this study, we explore the performance of various regression models, including linear regression, polynomial regression, SVR, and ANN, in predicting the power output of a solar power plant. Each model was evaluated according to standard performance metrics (RMSE, MAE, and R2 score) between training, testing, and validation sets. To further assess the generalizability and strength of our models, we used them to predict power output with a new dataset that was not included in the original dataset of the Kaberetene data set. This dataset contains data collected over four days in 2021: January 15, April 15, July 15, and October 15 with each day representing a different season of the year.

Figures 3–6 presents a comparison between the actual power generated and the predicted power using regression models developed with two input variables (irradiance and temperature) and three input variables (irradiance, temperature, and wind speed). In the graph, the blue line represents the actual power generated, while the orange line represents the predicted power. Figure 6 highlights the output efficiency over the four days compared with the real power, where it is observed that the ANN provide the most accurate prediction power and the lowest error compared to the models presented in Figures 3–5. These results support the hypothesis that ANN is the best suitable approach for the Algerian data.



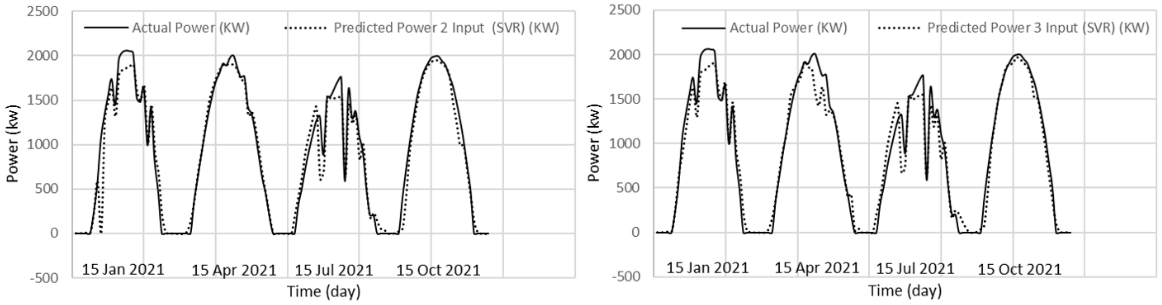
(a) (b)

Figure 3. Actual vs predicted values linear regression: (a) 2 input and (b) 3 input



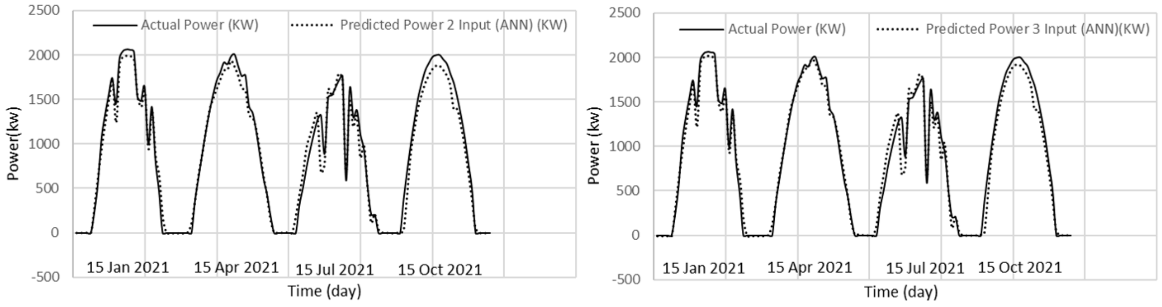
(a) (b)

Figure 4. Actual vs predicted values polynomial regression: (a) 2 input and (b) 3 input



(a) (b)

Figure 5. Actual vs predicted values SVR: (a) 2 input and (b) 3 input



(a) (b)

Figure 6. Actual vs predicted values ANNs: (a) 2 input and (b) 3 input

4.4. Discussion

This study presents the development and performance of several regression models for predicting solar power based on weather conditions (irradiance, temperature, and wind speed). Several performance criteria were used in the solar power prediction method literature as prediction accuracy and error. In this context, the prediction model's performance was evaluated in terms of prediction accuracy (R^2) and prediction error (MAE, RMSE). The linear regression models provided a limited performance due to their inability to capture non-linear relationships in the data. It is observed that the increase in the number of input features enhances their performance as shown in Table 2. However, this improvement was limited by an inability to capture non-linear relationships in the data. Polynomial regression enhanced the model's ability to account for non-linear relationships (Table 3). However, the increased complexity of higher-degree polynomials introduced overfitting, reducing their generalizability.

SVR provided a more flexible approach, handling non-linear relationships better than linear models. Despite this, the SVR model achieved low validation accuracy compared to the training accuracy. More specifically, the R^2 values on the validation sets were much lower than on the training sets, meaning that the model is overfitting the training data. This suggests that the SVR model overfitted the training data, but it failed to generalize well to new unseen data and thus performed extremely poorly on the validation and test datasets in terms of predictive power. The possible reason for failure is overfitting: the large difference between training and validation accuracy is an extremely strong sign of overfitting. The model could have memorized the training data, learning noise, and irrelevant patterns rather than the underlying patterns. This could be due to the extremely high complexity of the SVR model, which could have been too flexible for the specific details of the training data.

Another reason is data complexity: the model might be too simple to capture more complex relationships between the variables in the data. Even employing the radial basis function (RBF) kernel, typically potent for handling non-linear relationships, the model might still be too lacking in sophistication to be able to utilize this type of non-linear relationship in this dataset. ANN might suit the dataset better. In contrast, ANN excelled in capturing more complex non-linear relationships effectively. The testing plots (Table 5) and Figure 6 show that the ANN model provides a competitive accuracy when compared to the other models since it has a lower error and the highest accuracy between the actual power and the predicted power. The results showed that during different weather conditions from the season, the ANN model closely approximated the real power values with minimal error making it a reliable tool for solar power forecasting. This study was able to develop a machine learning-based model to estimate the solar power generated based on natural data, such as solar radiation, temperature, and wind speed. Machine learning was developed by implementing the ANN algorithm and resulted in estimation accuracies of 93.69% and 94.96% in the two datasets respectively. The accuracy result is comparable to other similar studies.

The studies in [13] and [14] utilized different datasets from various locations and applied multiple machine learning algorithms to develop solar power forecasting models. These models achieved accuracies ranging from 64.9% to 97.097%. Our results demonstrate that the proposed model outperforms the existing approaches reported in the literature [13], [14]. Since the proposed model is efficient in forecasting, this model will contribute to photovoltaic systems to optimize energy generation. Additionally, it can be applied in different multi-horizon forecasting applications such as a grid or microgrid demand to reduce the use of gas energy and maintain the balance different multi-horizon forecasting applications such as a grid or microgrid demand, enabling reduced reliance on gas energy and improved power plant balance. This model will also enhance the integration and reliability of photovoltaic systems in the Algerian energy market. This is particularly significant as the Algerian government has launched a program to implement 15,000 MWc of PV capacity by 2035. Accurate solar power generation forecasting using ANN models is essential for optimizing energy production, distribution, and storage. In Algeria, where solar energy has significant potential but grid stability is a challenge, precise forecasting can improve grid management, reduce energy waste, and prevent power outages, especially during peak demand periods.

Future work could improve accuracy by adding more input features, such as humidity, to assess their impact on model performance, especially for the Algerian dataset. By integrating an ANN-based model that learns from historical data and environmental factors, Algeria can increase its reliance on renewable energy while maintaining grid stability. This reduces the need for expensive fossil fuel backup plants, lowering operational costs and offering environmental benefits.

5. CONCLUSION

The ANN predictive model for PV power plants holds significant importance in Algeria's energy market, where hybrid energy systems combining fossil fuels and PV power plants are utilized. Since PV power generation is influenced by weather conditions, fluctuations in its output can impact the balance of energy supply and demand. Accurately predicting PV energy production is therefore crucial to compensating for any energy shortfalls with fuel-based generators, ensuring that load demands are consistently met.

Additionally, this approach enables optimal utilization of renewable energy resources, benefiting both the environment and the economy. In this article, we presented a prediction of hourly and daily solar power output using four data-mining algorithms: linear regression, polynomial regression, SVR, and ANN. These models were developed based on one year of daily data collected in 2020 and tested over four days in 2021. We detailed the parameter selection process for each model and evaluated their performance using metrics such as MSE, RMSE, MAE, and R² score. Two prediction scenarios were considered: one using two input parameters (irradiance, temperature) and another using three input parameters (irradiance, temperature, wind speed). The computational results demonstrated that the ANN model achieved the most accurate predictions of solar power with three input parameters on the testing dataset, showing reasonable generalization with few outliers. However, there is potential for further improvement in forecasting accuracy.

Our experimental findings suggest that incorporating humidity as an input variable to study its direct and indirect effects on prediction could enhance model performance. Additionally, exploring advanced methods such as recurrent neural networks (RNN) and LSTM for improved accuracy in solar power prediction will be a focus of our future work. In Algeria energy market, which uses both fossil and PV power plants, weather conditions affect PV output and, consequently, load demand. Accurate PV forecasting ensures that any energy gaps are filled by backup generators, ensuring a reliable energy supply and supporting environmental sustainability.

ACKNOWLEDGMENTS

We would like to extend our sincere thanks to the members of the cooperation agreement between the Research Unit for Renewable Energies in the Sahara medium (URERMS) in Adrar, and the SER (SKTM) team at SONELGAZ, whose contribution played a very important role in the completion of this research, through technical support and providing the necessary facilities to achieve this research.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Dalila Cherifi	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
Abdeldjalil Dahbi	✓	✓		✓		✓	✓	✓		✓	✓	✓		
Mohamed Lamine Sebbane		✓	✓	✓	✓	✓		✓	✓	✓	✓			
Bassem Baali		✓	✓	✓	✓	✓		✓	✓	✓	✓			
Ahmed Yassine Kadri						✓	✓	✓		✓				
Messaouda Chaib						✓	✓	✓		✓				

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal Analysis

I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting

Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject Administration
Fu : **Fu**nding Acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

Data availability is not applicable to this paper as no new data were created and analyzed in this study.

REFERENCES

[1] L. El Chaar, L. A. Lamont, and N. El Zein, "Review of photovoltaic technologies," *Renewable and Sustainable Energy Reviews*, vol. 15, no. 5, pp. 2165–2175, 2011, doi: 10.1016/j.rser.2011.01.004.

[2] G. K. Singh, "Solar power generation by PV (photovoltaic) technology: a review," *Energy*, vol. 53, pp. 1–13, 2013.

[3] M. Chaib, D. Abdeldjalil, A. Benatillah, N. Hachemi, E. Sakher, and B. Ben Abdelkarim, "Modeling, simulation and analysis of the input climat parameter effect on the photovoltaic panel," *2nd International Conference on Energy Transition and Security, ICETS 2023*, pp. 1–5, 2023, doi: 10.1109/ICETS60996.2023.10410824.

[4] M. Chaib *et al.*, "Long-term performance analysis of a large-scale photovoltaic plant in extreme desert conditions," *Renewable Energy*, vol. 236, 2024, doi: 10.1016/j.renene.2024.121426.

[5] J. Antonanzas, N. Osorio, R. Escobar, R. Urraca, F. J. Martinez-de-Pison, and F. Antonanzas-Torres, "Review of photovoltaic power forecasting," *Solar Energy*, vol. 136, pp. 78–111, Oct. 2016, doi: 10.1016/j.solener.2016.06.069.

[6] M. A. Al Amin and M. A. Hoque, "Comparison of ARIMA and SVM for short-term load forecasting," *IEMECON 2019 - 9th Annual Information Technology, Electromechanical Engineering and Microelectronics Conference*, pp. 205–210, 2019, doi: 10.1109/IEMECONX.2019.8877077.

[7] C. Wan, J. Zhao, Y. Song, Z. Xu, J. Lin, and Z. Hu, "Photovoltaic and solar power forecasting for smart grid energy management," *CSEE Journal of Power and Energy Systems*, vol. 1, no. 4, pp. 38–46, 2016, doi: 10.17775/cseejpes.2015.00046.

[8] Y. J. Zhong and Y. K. Wu, "Short-term solar power forecasts considering various weather variables," *Proceedings - 2020 International Symposium on Computer, Consumer and Control, IS3C 2020*, pp. 432–435, 2020, doi: 10.1109/IS3C50286.2020.00117.

[9] I. Sarbu and C. Sebarchievici, *Solar heating and cooling systems: fundamentals, experiments and applications*. Academic Press, 2017.

[10] Chandra Subhash, Agrawal Sanjay, and Chauhan D.S., "Effect of ambient temperature and wind speed on performance ratio of polycrystalline solar photovoltaic module: an experimental analysis," *International Energy Journal*, vol. 18, pp. 171–180, 2018.

[11] A. O. Njok and J. C. Ogbulezie, "The effect of relative humidity and temperature on polycrystalline solar panels installed close to a river," *Physical Science International Journal*, vol. 20, no. 4, pp. 1–11, 2019, doi: 10.9734/psij/2018/44760.

[12] A. Dahbi *et al.*, "Performance evaluation of a real polycrystalline photovoltaic field under desert conditions," in *International Conference on Artificial Intelligence in Renewable Energetic Systems*, 2023, pp. 490–501. doi: 10.1007/978-3-031-60629-8_47.

[13] T. Verma, A. P. S. Tiwana, C. C. Reddy, V. Arora, and P. Devanand, "Data analysis to generate models based on neural network and regression for solar power generation forecasting," in *Proceedings - International Conference on Intelligent Systems, Modelling and Simulation, ISMS*, 2016, pp. 97–100. doi: 10.1109/ISMS.2016.65.

[14] A. M. Kuriakose, D. P. Kariyalil, M. Augusthy, S. Sarath, J. Jacob, and N. R. Antony, "Comparison of artificial neural network, linear regression and support vector machine for prediction of solar PV power," in *2020 IEEE Pune Section International Conference, PuneCon 2020*, 2020, pp. 53–58. doi: 10.1109/PuneCon50868.2020.9362442.

[15] M. Abuella and B. Chowdhury, "Solar power forecasting using artificial neural networks," *2015 North American Power Symposium (NAPS)*, pp. 1–5, 2015, doi: 10.1109/NAPS.2015.7335176.

[16] M. Aslam, S. J. Lee, S. H. Khang, and S. Hong, "Two-stage attention over LSTM with Bayesian optimization for day-ahead solar power forecasting," *IEEE Access*, vol. 9, pp. 107387–107398, 2021, doi: 10.1109/ACCESS.2021.3100105.

[17] N. Uddin, E. Purwanto, and H. Nugraha, "Machine learning based modeling for estimating solar power generation," *E3S Web of Conferences*, vol. 475, 2024, doi: 10.1051/e3sconf/202447503009.

[18] C. Voyant *et al.*, "Machine learning methods for solar radiation forecasting: A review," *Renewable Energy*, vol. 105, pp. 569–582, 2017, doi: 10.1016/j.renene.2016.12.095.

[19] A. K. Yadav, H. Malik, and S. S. Chandel, "Selection of most relevant input parameters using WEKA for artificial neural network based solar radiation prediction models," *Renewable and Sustainable Energy Reviews*, vol. 31, pp. 509–519, 2014, doi: 10.1016/j.rser.2013.12.008.

[20] I. Kasireddy, V. M. Reddy, P. Naveen, and G. H. Vardhan, "Exploring machine learning models for solar energy output forecasting," in *International Conference on Cognitive Computing and Cyber Physical Systems*, 2023, pp. 210–217. doi: 10.1007/978-3-031-48888-7_18.

[21] M. S. Acharya, A. Armaan, and A. S. Antony, "A comparison of regression models for prediction of graduate admissions," in *2019 International Conference on Computational Intelligence in Data Science (ICCIDS)*, Feb. 2019, pp. 1–5. doi: 10.1109/ICCIDS.2019.8862140.

[22] Y. Chen, P. He, W. Chen, and F. Zhao, "A polynomial regression method based on Trans-dimensional Markov Chain Monte Carlo," in *2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, Oct. 2018, pp. 1781–1786. doi: 10.1109/IAEAC.2018.8577769.

[23] M. Awad and R. Khanna, "Support vector regression," in *Efficient Learning Machines*, Berkeley, CA: Apress, 2015, pp. 67–80. doi: 10.1007/978-1-4302-5990-9_4.

[24] J. Brownlee, *Deep learning with Python: develop deep learning models on Theano and TensorFlow*. 2016. [Online]. Available: http://web.stanford.edu/class/cs224n/readings/cs224n-2019-notes06-NMT_seq2seq_attention.pdf.

[25] M. Steurer, R. J. Hill, and N. Pfeifer, "Metrics for evaluating the performance of machine learning based automated valuation models," *Journal of Property Research*, vol. 38, no. 2, pp. 99–129, 2021, doi: 10.1080/09599916.2020.1858937.

BIOGRAPHIES OF AUTHORS

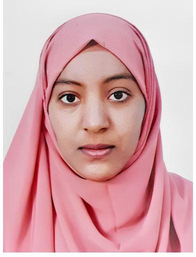
Prof. Dalila Cherifi is a professor and doctor in Electronics, with a specialization in image and signal processing at the Institute of Electrical Engineering and Electronics, University of Boumerdes. She got her Habilitation to Direct Research (HDR) from the same university and obtained her Ph.D. from Telecom-Paris (Paris-Tech) in France in 2005. As an expert in artificial intelligence, machine learning, data mining, and information retrieval, she applies her knowledge to various fields, including biomedical and biometrics applications with a particular emphasis on anomalies detection and classification. She has authored numerous research papers and articles in prestigious international journals and books, delivered many conferences and workshops, and received multiple awards and recognitions for her contributions to research and innovation. She can be contacted at email: da.cherifi@univ-boumerdes.dz.

Prof. Abdeldjalil Dahbi presently serving a principal researcher at the Research Unit in Renewable Energies in the Saharan Medium (URER-MS), Adrar, Algeria. He holds degrees in Electromechanical Engineering, Technical English, Electrical Controls, Energetic Physics, and a Ph.D. in Electric Controls. He became a professor in June 2023. His expertise includes renewable energy systems, wind energy, electric smart control, photovoltaic systems, remote control, meteorology, and IoT. He has published several papers, patents, and books. He also reviews for international conferences and journals, including IEEE and Elsevier. He can be contacted at email: dahbi_j@yahoo.fr.

Mohamed Lamine Sebbane received a B.S. degree in Electrical Engineering from the Institute of Electrical Engineering and Electronics, University of Boumerdes, Algeria, in 2022 and an M.S. degree in Power Engineering from the same university in 2024. His research interests include the application of AI in power systems, renewable energy, and time-series analysis/forecasting. He can be contacted at email: m.sebbane@univ-boumerdes.dz.

Bassem Baali obtained his B.S. degree in Electrical Engineering from the Institute of Electrical Engineering and Electronics, University of Boumerdes, Algeria, in 2022, followed by an M.S. degree in Power Engineering from the same institution in 2024. His research focuses on applying artificial intelligence to power systems. He can be contacted at email: baalibassem@gmail.com.

Ahmed Yassine Kadri is an assistant professor in electrical equipments and regulation systems at Kasdi Merbah University, Ouargla, Algeria. He holds degrees in electrical engineering, Technical English, Law, and a magister in electrical industry. He is a Ph.D. student in electrical engineering. His expertise includes renewable energy systems, wind energy, electric networks, photovoltaic systems, electrical services protection. He has published some papers in renewable energies. He can be contacted at email: aykadr@gmail.com.



Messaouda Chaib is a Ph.D. Student in material's sciences department at the University of Adrar, Algeria. She received her B.Math. in 2016. Then, Licence degree in physics of material from University of Adrar, in 2019. After that, she obtained the Master degree in Energetic physics and renewable energy from the same university, in 2021. Her research interests includes the field of engineering physics, renewables energy, environmental engineering, artificial intelligence, and intelligent control. She can be contacted at email: chaib.messaouda@univ-adrar.edu.dz.