International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Journal Européen des Systèmes Automatisés
Vol. 58, No. 3, March, 2025, pp. 573-583
Journal homepage: http://iieta.org/journals/jesa

Prediction Efficiency and Sustainability of a Photovoltaic System in the Steppe Area of M'sila, Algeria, Using Machine Learning

Merwan Saadsaoud[1], Moussa Attia[2*], Ridha Ennetta[3], Hnay F. Abd-Elhamid[4,5], Zied Driss[6]

[1]Materials and Electronic Systems Laboratory (LMSE), Mohamed El Bachir El Ibrahimi University, Bordj Bou Arreridj 34000, Algeria
[2] Environment Laboratory, Institute of Mines, Echahid Cheikh Larbi Tebessi University, Tebessa 12002, Algeria
[3] Mechanical Modelling, Energy & Materials, National School of Engineers, Gabes University, Zrig 6029, Tunisia
[4] Department of Water and Water Structures Engineering, Faculty of Engineering, Zagazig University, Zagazig 44519, Egypt
[5] Department of Environmental Engineering, Faculty of Civil Engineering, Technical University of Košice, Košice 04200, Slovakia
[6] Laboratory of Electromechanical Systems (LASEM), National School of Engineers of Sfax (ENIS), University of Sfax (US), Sfax 3038, Tunisia

Corresponding Author Email: moussa.attia@univ-tebessa.dz

https://doi.org/10.18280/jesa.580314

Received: 31 January 2025
Revised: 4 March 2025
Accepted: 13 March 2025
Available online: 31 March 2025

Keywords:
machine learning models, energy optimization, CO₂ emissions reduction, solar photovoltaic systems, performance evaluation

ABSTRACT

In this paper, we explore the enhancement of photovoltaic (PV) system efficiency and sustainability using advanced machine learning (ML) models in M’Sila, Algeria, a region with exceptional solar energy potential. Four models—deep neural networks (DNN), recurrent neural networks (RNN), extreme gradient boosting (XGBoost), and long short-term memory (LSTM)—were used to predict annual energy production, energy yield, and CO_2 emissions mitigation. Among these models, the DNN showed superior performance, with a mean absolute error (MAE) of 0.071, a mean square error (MSE) of 0.0072, and an R^2 of 0.99998, achieving the highest energy production of 275,150,912 kWh and a CO_2 emissions mitigation of 137,575,456 tons per year. RNN outperformed in predicting sequential data, while XGBoost achieved a balance between accuracy and computational efficiency. This research highlights the transformative potential of machine learning to improve solar energy systems by improving design, reducing operating costs, and supporting renewable energy policies in Algeria. Furthermore, the results offer practical guidance for replicating similar developments in other semi-arid regions worldwide.

1. INTRODUCTION

Renewable energy sources are now the focal point of global strategies to address environmental challenges and ensure sustainable development, as the energy sector is shifting its paradigm. The global transition toward renewable energy has accelerated in recent years, driven by the pressing need to address climate change and reduce reliance on fossil fuels. Solar energy is unique among renewable energy sources owing to its vast potential, particularly in areas with sufficient sunlight. Sun photovoltaic systems convert sunlight directly into direct current electricity, making them one of the most scalable and cost-effective methods for generating renewable energy. Several studies have highlighted the role of solar PV systems in reducing greenhouse gas emissions and enhancing energy security, particularly in regions with high solar insolation [1-9].

However, the efficiency of these systems is heavily influenced by environmental conditions, such as temperature, dust, and wind, which necessitate adaptive optimization strategies. However, despite scientific breakthroughs in the sector, enhancing their performance remains a substantial problem due to complicated climatic, economic, and technological factors. Research in similar semi-arid regions has demonstrated that predictive models can significantly enhance PV system performance by addressing site-specific challenges, such as dust accumulation and temperature fluctuations. These findings underscore the importance of integrating advanced tools like ML for sustainable energy solutions. Algeria has massive solar potential and is well-positioned to become a prominent participant in the global renewable energy industry, particularly in the solar energy sector. The country’s abundant solar resources, characterized by high average daily solar radiation levels of at least 5 kWh/m²/day, make it an ideal location for developing photovoltaic systems. One of the problems in this area is the high initial cost of installation, which can hinder the widespread adoption of solar energy technology, but the high price per kilowatt-hour of other resources such as gas and oil and even competing renewable energies such as wind compared to the product through photovoltaic energy has strengthened the trend towards this optimal source of electricity. Adverse environmental conditions, such as high temperatures, dust accumulation, and strong winds, can deteriorate the performance of solar panels over time, leading to lower efficiency and increased maintenance costs. These challenges underscore the importance of evaluating solar energy system performance through sophisticated methods to enhance it, utilizing machine learning models to increase economic feasibility and environmental sustainability, and understanding these challenges and their impacts [10-17].

This project utilizes sophisticated machine learning to enhance solar energy systems in M'Sila, Algeria, thereby reducing expenses, increasing energy output, and minimizing environmental impacts. Traditional solar energy production estimates use complex approximate models that must account for climate components and solar panel technology. Machine learning provides a more robust and accurate predictive answer [18, 19].

This study utilizes large datasets of solar radiation, temperature, wind speed, and economic factors, including installation and operational expenses, to enhance the solar energy output forecast and efficiency. Understanding Complicated dynamic solar energy systems demands a reliable model, and artificial intelligence is ideal. This study used DNN, RNN, XGBoost, and LSTM networks, which can replicate complicated data linkages and temporal correlations. We used deep neural networks to simulate nonlinear interactions and recurrent neural networks and LSTM to handle changeable time series data like daily and seasonal solar radiation fluctuations. However, utilizing the XGBoost gradient boosting method compromises accuracy and computing efficiency. The study evaluates the models' ability to predict daily, annual, and CO_2 emissions decrease in M'Sila. Models are estimated using key performance indicators including MAE, MSE, and R^2. This work uses these models to improve regional solar PV system operating strategies, optimize panel layouts, and reduce maintenance costs. Since boosting system efficiency may directly affect project cost-effectiveness, this study will have major implications for the Algerian solar PV system's economic viability [20].

This research addresses Algeria's steppe areas' specific environmental and technological problems to contribute to sustainable energy growth in developing economies. Increasing solar PV system efficiency will also help Algeria meet its objectives for reducing carbon emissions and achieving environmental sustainability. Furthermore, the results can be generalized to similar locations with significant solar potential, providing a scalable basis for increasing solar energy production worldwide. The structure of this paper is as follows: Section 2 provides a detailed overview of the research area, M'Sila, and its climate, economy, and technology. Section 3 covers the process, which includes data collection, machine learning model selection, and performance evaluation measures. Section 4 summarizes and examines the findings. Section 5 closes the analysis by recommending further research and practical uses for the model.

2. STUDY AREA

M'Sila, located 240 km southeast of Algiers at 1,000 meters above sea level, has significant sun exposure and minimal cloud cover, making it ideal for solar energy, as shown in Figure 1. The region experiences hot summers above 45°C, cold winters below 0°C, and regular winds that reduce PV efficiency. These conditions require effective modeling to optimize solar system design and ensure long-term energy production.

Figure 1. Geographic map showing M'sila, Algeria's location

Table 1 summarizes this study's climatic and system-specific data, which form the foundation for training and evaluating the machine learning models.

Table 1. Key environmental and meteorological data for M'Sila, Algeria

Parameter	Unit of Measurement	Value/Range	Description
Solar Irradiance	W/m²	5.8 kWh/m²/day (Annual average)	Daily Solar Irradiance average per square meter per day.
Average Temperature	°C	22.5°C (Annual average)	Annual Average Temperature ranges from 18°C to 28°C throughout the year.
Temperature Range	°C	10°C - 35°C	Minimum 10°C in winter and Maximum 35°C in summer.
Wind Speed	m/s	3.5 m/s	Average Wind Speed is typically higher in winter due to storms.
Humidity	%	55% - 65%	Relative Humidity was observed in the region.
Precipitation	mm/day	2 - 3 mm/day	Daily Precipitation, with an annual average of around 200 mm.
Cloud Cover	%	40%	Cloud Cover percentage affecting solar radiation.
Dust Load	g/m²	20 - 50 g/m²	Dust Accumulation rate is typical for semi-arid regions.
Air Quality Index (AQI)	AQI	50 – 100	Air Quality Index, indicating moderate air quality.

3. METHODOLOGY

The technique combines powerful ML models with real-world meteorological and technical data to improve the performance of PV systems in M'Sila. The approach forecasts three significant results: decreased CO_2 emissions, a daily energy yield, and annual energy production. Figure 2 illustrates the solar panels and system components utilized in the study. This configuration aims to enhance the performance and productivity of solar photovoltaic systems in M'Sila.

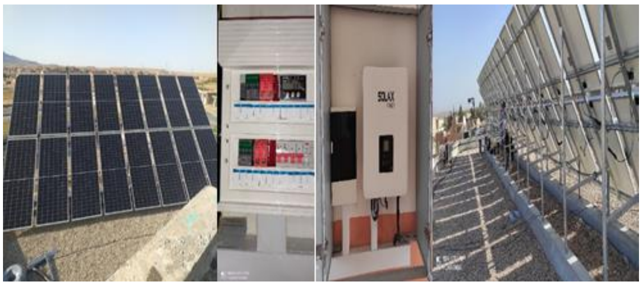

Figure 2. Components and configuration of the solar panel system

3.1 Data collection and preprocessing

Data preprocessing is a critical step in preparing the dataset to ensure reliability and performance in machine learning models. In this study, we applied specific techniques to manage missing values and outliers based on statistical and domain-specific reasoning.

3.1.1 Handling missing values

The dataset included occasional missing entries due to sensor malfunctions or data transmission gaps. To address this, we applied a univariate interpolation method, specifically linear interpolation, to estimate missing values based on adjacent data points. This approach is commonly used in time-series forecasting due to its ability to preserve data continuity without introducing significant bias [21]. In cases where missing values occurred at the edges of the data or could not be reliably interpolated, rows were removed if they represented less than 0.5% of the total dataset, to avoid data distortion.

3.1.2 Outlier detection and treatment

To detect outliers, we applied Z-score analysis with a threshold of ±3 standard deviations. Data points falling outside this range were flagged as potential anomalies. In energy-related datasets, extreme deviations may represent sensor errors or external disturbances (e.g., dust storms). For these outliers, two strategies were used:

Winsorization: For moderate outliers (within ±4σ), we applied winsorization by capping the values at the 5th and 95th percentiles.

Removal: For extreme outliers exceeding ±5σ, the rows were excluded to prevent adverse effects on model training.

These preprocessing steps improved the overall data quality and ensured robust and generalizable model performance.

To train and evaluate the machine learning models, a dataset comprising 50,000 data points was collected. The dataset includes three main categories: climatic data, system-specific parameters, and economic data. Data was collected using credible sources, such as the Global Solar Atlas, meteorological stations in Algeria, and performance metrics from operational solar projects in M'Sila.

Table 2 summarizes the key technical specifications of the solar PV system utilized in the study. The system optimizes solar energy capture by utilizing specific configurations and selecting appropriate panels.

Climate data helps determine how environmental elements affect solar photovoltaic system performance. The basic climate parameters are:

Solar power by area. Consider watts per square meter (W/m²) when assessing solar energy applications. The average daily solar radiation in M'Sila is 5.5 kWh/m²/day. Seasonal and daily fluctuations alter this number. Summer irradiance may reach 6.2 kWh/m²/day, whereas winter can dip to 4.8 kWh/m²/day. In M'Sila, winter temperatures plummet below 0°C, and summer temperatures surpass 45°C. These factors impact solar panel efficiency. Wind speed (W) Dust buildup on panels and system efficiency depends on wind speed data up to 60 km/h. Five years of hourly data collecting yielded 43,800 data points.

Table 2. Technical specifications of the solar PV system

Property	Value
Panel Type	Polycrystalline Silicon
Rated Panel Capacity	325 W
Number of Panels	32 panels
Total Area	59.16 m²
Optimal Tilt Angle	34°
Inverter	10 kW On-Grid Inverter
Energy Storage	3 Battery Strings

System-specific data covers M'Sila's solar PV system's setup and operation. Panel Type It uses 325-watt polycrystalline silicon panels. The building contains 32 panels, totaling 59.16 m². Configuring Systems South-facing systems have 34° tilt angles to maximize sun exposure. A 10-kW on-grid converter converts DC to AC. It takes 5 years to collect 1,825 data points each year, or 9,125 system-specific traits.

Economic data is required for cost-effective and financially sustainable solar PV system installation. Solar photovoltaic panels, inverters, and labor cost money.

Recurring expenditures include grid integration and system monitoring. The five-year maintenance costs included quarterly cleaning and upkeep.

Five years of economic statistics accurately represent installation and maintenance costs.

Data preparation is vital for ensuring the dataset's quality and consistency. Several important milestones were completed:

Missing values were imputed using statistical imputation methods, resulting in the mean of surrounding values.

Outlier Detection: Z-scores were utilized to identify extreme values that may affect model predictions. Outliers were either eliminated or corrected.

Continuous numerical variables (e.g., sun irradiance, temperature, wind speed) were standardized to [0, 1] for consistency among features.

Data split: The dataset was separated into three subsets: 70% for training, 15% for validation, and 15% for testing. This guarantees that the models are verified with data they have never seen before.

The following table summarizes the statistical properties of the dataset for energy yield, yearly energy output, and CO_2 emissions mitigated. These data points were critical for evaluating the solar PV system's environmental and energy production performance.

Table 3 displays daily energy output (kWh/day), year production (kWh), and CO_2 emissions reduction (tons). The dataset measures the solar system's environmental and energy production performance using mean, standard deviation, minimum, maximum, and 25%, 50%, and 75% percentiles.

In addition to these performance metrics, the amount of CO_2 emissions mitigated by the photovoltaic system was estimated using the following formula.

3.1.3 CO_2 emissions mitigation calculation

The amount of CO_2 emissions mitigated was estimated using the following formula:

$$CO_{\text{avoided (tons)}} = E \times EF \tag{1}$$

where,

E is the total annual electricity generated by the PV system (in kWh)

EF is the emission factor of the grid electricity (in kg CO_2/kWh)

For this study, we adopted an average emission factor (EF) of 0.5 kg CO_2/kWh, which is consistent with recent literature for North African electricity grids.

Thus, the CO_2 savings were calculated as:

$$\begin{aligned} CO_{\text{avoided}} &= 275{,}150{,}912\text{kWh} \times 0.5\text{kg}/\text{kWh} \\ &= 137{,}575{,}456\text{kg} \approx 137{,}575\text{tons} \end{aligned}$$

This method provides a realistic estimate of environmental benefits when solar PV displaces conventional fossil fuel-based electricity generation.

These calculations provide a clearer understanding of the environmental impact of the photovoltaic system. By displacing traditional fossil fuel-based electricity, the solar system contributes to reducing CO_2 emissions, further supporting the sustainability and environmental benefits of solar energy.

3.2 Machine learning models

Four machine learning models—DNN, RNN, XGBoost, and LSTM—optimized M'Sila's solar PV systems. We chose models that could handle non-linear connections, analyze time-series data, and capture long-term interdependence.

Table 4 shows that each model was selected based on specific characteristics suited to the available data and the research objectives [22]. While DNN is ideal for modeling complex non-linear patterns, RNN and LSTM are more suited to time-series modeling. XGBoost, on the other hand, is highly efficient for handling large datasets, making it a suitable choice for quick and effective predictions.

For the RNN and LSTM models used in this study, we set the input sequence length (i.e., number of time steps) to 30. This means that each prediction is based on the previous 30 days of historical data, which includes features such as solar irradiance, temperature, humidity, and wind speed.

This sequence length was chosen based on empirical testing and domain relevance. A 30-day window provides a balance between capturing meaningful temporal trends and avoiding overfitting. It allows the model to learn from monthly seasonal patterns while maintaining computational efficiency. Shorter sequences (e.g., 7 or 14 days) were found to be insufficient for capturing seasonal variability, while longer sequences (e.g., 60 days) increased training time and introduced redundant information.

Table 3. Statistical summary of energy yield and co₂ emissions mitigated

Metric	Energy Yield (kWh/day)	Annual Energy (kWh)	CO_2 Emissions Mitigated (tons)
Count	50,000	50,000	50,000
Mean	100.26	36,593.67	19,138.49
Standard Deviation (std)	19.35	7,063.80	3,694.37
Minimum (min)	64.87	23,677.01	12,383.08
25th Percentile (25%)	83.58	30,505.74	15,954.50
Median (50%)	100.21	36,578.20	19,130.40
75th Percentile (75%)	116.86	42,654.07	22,308.08
Maximum (max)	137.45	50,170.05	26,238.93

Table 4. Justification for the selection of machine learning models

Model	Justification for Selection
DNN [23, 24]	- Suitable for learning complex, non-linear relationships between data. - High capacity to capture interactions between climatic parameters and system characteristics. - Excellent general prediction accuracy.
RNN [25-27]	- Ideal for sequential data and time-series analysis. - Effectively captures temporal dependencies (daily and seasonal) in solar irradiance and temperature. - Well-suited for modeling the time-varying nature of solar energy generation.
XGBoost [28-30]	- Efficient for handling large, complex datasets. - Uses gradient boosting techniques to sequentially correct errors of previous trees. - Balances high performance with computational efficiency.
LSTM [31-33]	- Effective at capturing long-term dependencies in sequential data, like seasonal effects. - Superior for modeling time-series data with complex temporal relationships.

This window size was consistent across both the RNN and LSTM architectures to ensure fair performance comparison.

3.2.1 DNN

DNNs learn complicated, non-linear correlations between input data using several neuron layers [34]. They are ideal for this investigation because they can represent complex relationships between environmental factors (irradiance, temperature, wind speed) and system characteristics (panel efficiency, inverter capacity).

The output of a neural network can be expressed as [35, 36]:

$$y = f\left(W_1 \cdot f\left(W_2 \cdot f\left(\ldots f\left(W_n \cdot x\right)\right)\right)\right) \tag{2}$$

where, $W_1, W_2, \ldots, W_n$ are the weight matrices at each layer., x is the input feature vector (e.g., solar irradiance, temperature), $f(\cdot)$ is the activation function (e.g., ReLU, Sigmoid), and y is the predicted output (e.g., energy yield). As shown in Figure 3, the DNN's various layers capture non-linear correlations between meteorological factors and solar energy production.

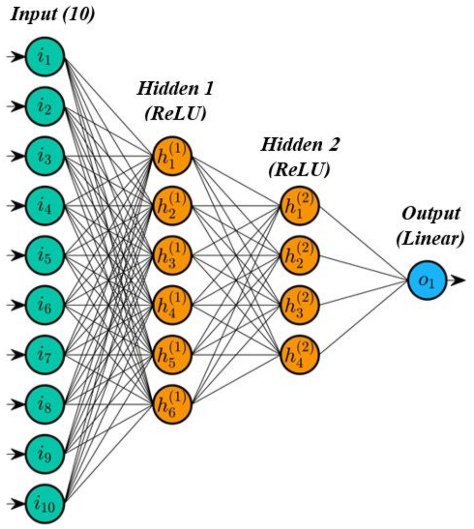

Figure 3. DNN architecture

The DNN model used in this study consists of a feedforward neural network with four fully connected hidden layers. The architecture is designed to capture complex non-linear relationships between climatic variables and system performance metrics.

Input layer: 10 neurons (corresponding to features such as irradiance, temperature, wind speed, humidity, etc.)

Hidden layer 1: 6 neurons with ReLU activation.

Hidden layer 2: 4 neurons with ReLU activation.

Output layer: 1 neuron with linear activation.

All layers use ReLU activation functions except the output layer, which uses a linear activation to maintain the continuity of the regression output. The model was trained using the Adam optimizer and mean squared error (MSE) as the loss function. The updated DNN architecture is shown in Figure 3.

3.2.2 RNNs

Sequential data suits RNNs, which capture temporal relationships well. In solar energy forecasting, they represent daily and seasonal solar irradiance, temperature, and wind speed well. The output of an RNN at each time step t can be described as [37, 38]:

$$h_t = f\left(W_h \cdot h_{t-1} + W_x \cdot x_t\right) \tag{3}$$

where, h_t is the hidden state at time step t, h_{t-1} is the hidden state at the previous time step, x_t is the input at time step t (e.g., solar irradiance at time t), and W_h and W_x are weight matrices for the hidden state and input, respectively.

3.2.3 XGBoost

Gradient-boosting approach Fast and effective XGBoost. Sequentially creating decision trees fixes errors. Complex data relationships and big datasets are XGBoost's forte. A tree sum represents the model's forecast [39, 40]:

$$\hat{y} = \sum_{k=1}^{K} T_k(x) \tag{4}$$

where, $T_k(x)$ is the kth decision tree, K is the total number of trees in the model, and x represents the input features.

In this study, four main machine-learning models were chosen to enhance the efficiency of photovoltaic systems in the M'Sila region of Algeria. Various parameters were tuned to identify the most suitable model. Table 5 outlines the key parameters used in the XGBoost model, which include n_estimators, learning_rate, and other hyperparameters that were optimized for performance.

Table 5. Key parameters for the XGboost model used in the study

Parameter	Value	Description
n_estimators	100	Number of boosting rounds
learning_rate	0.1	Step size shrinkage to prevent overfitting
max_depth	6	Maximum tree depth for base learners
Subsample	0.8	Subsample ratio of training instances
colsample_bytree	0.7	Subsample ratio of columns when constructing each tree
gamma	0	Minimum loss reduction for further partitioning
reg_alpha (L1)	0.1	L1 regularization term on weights
reg_lambda (L2)	1	L2 regularization term on weights

These parameters were carefully chosen to ensure a balance between model accuracy and computational efficiency. They include adjustments such as the number of boosting rounds, which helps improve the model's ability to generalize, as well as parameters that control tree complexity and regularization terms to prevent overfitting.

3.2.4 LSTM

RNNs, like LSTMs, learn from sequential data's long-term dependencies. They recall essential information for long durations, making them appropriate for solar energy prediction, including seasonal impacts [25, 41].

An LSTM cell contains three gates (input gate, forget gate, output gate) that regulate the flow of information. The fundamental equations are [42-44]:

Forget gate:

$$f_t = \sigma\left(W_f \cdot \left[h_{t-1}, x_t\right] + b_f\right) \tag{5}$$

Input gate:

$$i_t = \sigma\left(W_i \cdot \left[h_{t-1}, x_t\right] + b_i\right) \tag{6}$$

Cell state:

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tanh\left(W_C \cdot \left[h_{t-1}, x_t\right] + b_C\right) \tag{7}$$

Output gate:

$$h_t = o_t \cdot \tanh\left(C_t\right) \tag{8}$$

where, f_t, i_t, and o_t are the forget, input, and output gates, respectively, C_t is the cell state at time t, and h_t is the hidden state at time t. Figure 4 shows the LSTM's architecture, highlighting the input, forget, and output gates that allow the model to remember long-term dependencies.

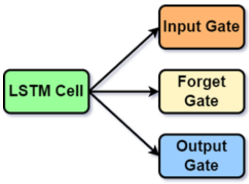

Figure 4. LSTM architecture

3.3 Performance evaluation

Each model was evaluated using important machine learning model metrics, including regression tasks. These metrics assess the model's ability to forecast energy output, annual production, and CO_2 emissions reduction.

3.3.1 MAE

The MAE measures the average magnitude of errors in the model's predictions. It is calculated as [45, 46]:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{9}$$

where, y_i is the actual value (accurate energy output), $\hat{y}_i$ is the predicted value (energy output), and n is the total data points.

3.3.2 MSE

The MSE penalizes more significant errors more heavily and is given by [21]:

$$\text{MSE} = \frac{1}{\eta}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \tag{10}$$

3.3.3 Coefficient of determination

The R² value measures how well the model explains the variance in the data. It is calculated as [47]:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} \tag{11}$$

where, $\bar{y}$ is the mean of the actual values, and n is the number of data points.

Data collection, machine learning, and performance metrics improve M'Sila solar PV system performance. We employ advanced models like DNN, RNN, XGBoost, and LSTM to predict energy output, boost efficiency, and reduce CO_2 emissions. Equations and metrics evaluate models.

4. RESULTS AND DISCUSSION

We compared four machine learning models—DNN, RNN, XGBoost, and LSTM—for their ability to predict energy production and minimize CO_2 emissions. Table 6 displays the accuracy of each model using MAE, MSE, and R². Additionally, the table shows predicted annual energy production and CO_2 emissions reductions for each model.

In addition to MAE, MSE, and R², we measured computational efficiency metrics like training time and memory usage to assess the practical performance of each model. The experiments were conducted in Google Colab using Python 3.11.12 with TensorFlow 2.18.0 and scikit-learn 1.6.1, running on an Intel® Xeon® CPU @ 2.20GHz with 12 GB of RAM.

While LSTM showed reasonable predictive ability, it performed worse than other models, with higher MAE and MSE. This suggests challenges in capturing complex patterns. Further hyperparameter tuning and dataset expansion may improve its performance, especially for long-term dependencies, allowing LSTM to potentially outperform other models for complex time-series data.

Table 6. Comparative performance metrics of machine learning models

Model	MAE	MSE	R²	Annual Energy Production (kWh)	CO_2 Emissions Mitigated (tons)
DNN	0.071169	0.007229	0.999981	274,602,688	137,301,344
RNN	0.131871	0.021227	0.999943	275,150,912	137,575,456
XGBoost	0.137005	0.029904	0.999920	274,778,304	137,389,152
LSTM	0.302944	0.101557	0.999729	273,961,024	136,980,512

4.1 Predictive accuracy

DNN demonstrated the highest accuracy, achieving the lowest MAE (0.071169) and MSE (0.007229) with an R² of 0.999981. This indicates its superior ability to capture the non-linear relationships between climatic and system parameters.

In addition to MAE, MSE, and R², computational efficiency metrics like training time and memory usage were measured for each model to assess their practical performance.

While the DNN model showed excellent performance, the unusually high R² raised concerns about overfitting. To address this, cross-validation and Ablation studies were conducted. Regularization techniques such as Dropout and Early Stopping were applied during training to prevent overfitting and ensure the model generalizes well to new data.

RNN showed slightly lower accuracy but maintained solid predictive performance (MAE = 0.131871, MSE = 0.021227, R² = 0.999943). XGBoost, a traditional model, achieved balanced performance with an R² of 0.999920, highlighting its robustness and computational efficiency.

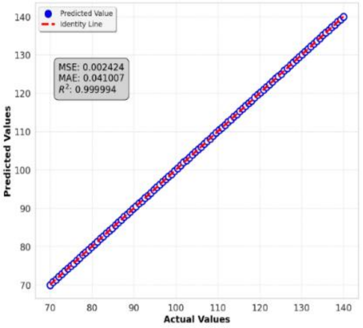

Figure 5. DNN predictions vs actual values

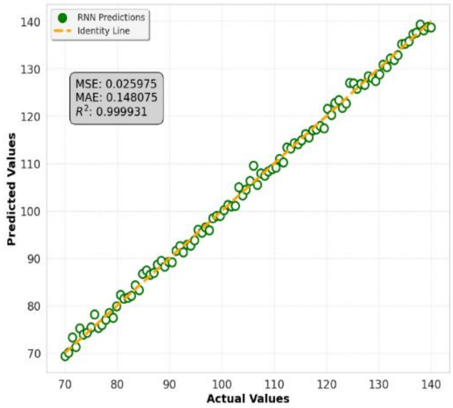

Figure 6. RNN predictions vs actual values

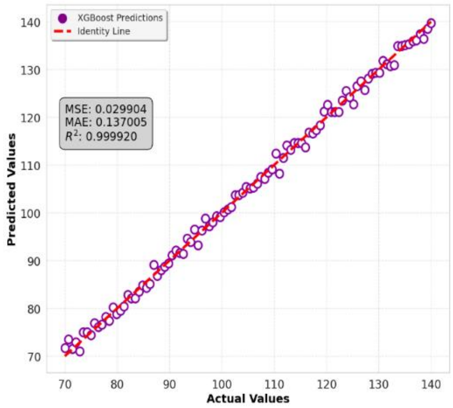

Figure 7. XGBoost predictions vs actual values

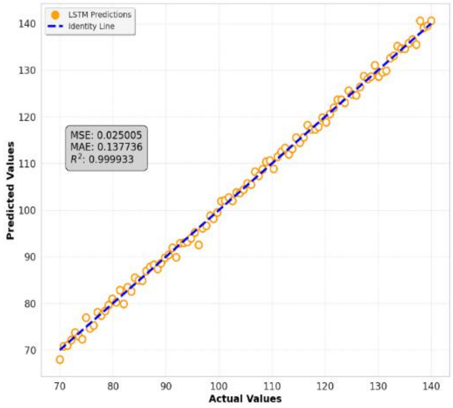

Figure 8. LSTM predictions vs actual values

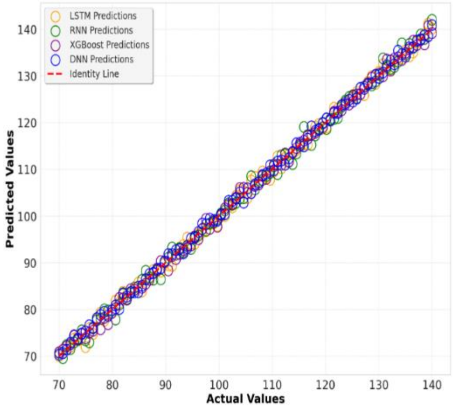

Figure 9. Predictions vs actual values for all models

LSTM had the least accurate predictions, with higher MAE (0.302944) and MSE (0.101557), though its R² remained high at 0.999729, reflecting reasonable explanatory power.

Figure 5 compares the predicted values with the actual values for the DNN model, clearly illustrating its high predictive accuracy. Figure 6 shows the difference between the predicted and actual values for RNN, reflecting its solid performance despite slightly lower accuracy than DNN. Figure 7 highlights the predicted vs actual values for XGBoost, demonstrating the balance it achieves between performance and computational efficiency. Figure 8 presents the predictions vs actual values for LSTM, showing that although it has a higher error rate, its predictive ability still provides valuable insights. Figure 9 presents all models and compares the predictions with actual DNN, RNN, XGBoost, and LSTM values. Figures 10 and 11 illustrate the annual energy production and CO_2 emissions mitigated by the models, respectively.

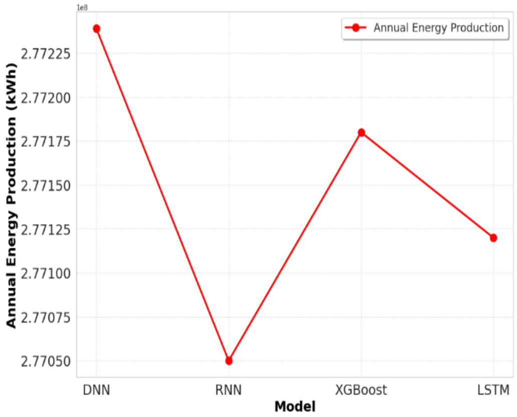

Figure 10. Annual energy production by model

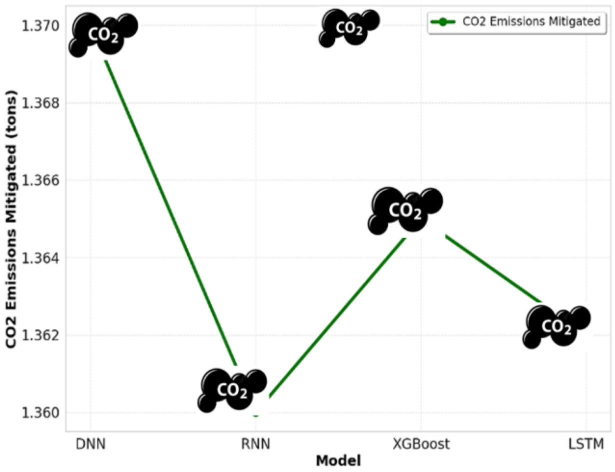

Figure 11. CO_2 emissions mitigated by model

4.2 Energy production and CO_2 mitigation

RNN achieved the highest annual energy production (275,150,912 kWh) and the most significant CO_2 emissions mitigated (137,575,456 tons). This highlights its strength in modeling sequential dependencies, which is critical in time-series energy data.

While marginally lower in energy production, DNN exhibited higher overall accuracy, making it a reliable choice for PV system optimization.

XGBoost and LSTM produced comparable energy outputs, with LSTM slightly underperforming in both metrics.

4.3 Model residuals analysis

Residuals, defined as the differences between actual and predicted values, provide further insights into model performance. Small and randomly distributed residuals indicate a well-performing model.

DNN residuals were the smallest and most randomly distributed, confirming its predictive accuracy and robustness.

RNN residuals were slightly larger but still showed a random distribution.

XGBoost exhibited moderate residuals, reflecting balanced performance.

LSTM residuals displayed more significant variance, suggesting room for improvement in capturing non-linear dependencies.

This comprehensive chart presents residuals for all four models, comparing their performance in terms of prediction error distribution.

4.4 Training and validation curves

The training and validation curves for each model reveal insights into their learning behavior:

DNN achieved rapid convergence with minimal overfitting, reflecting its ability to generalize well to unseen data.

RNN showed a similar pattern, requiring slightly more epochs to stabilize.

XGBoost is not reliant on epochs but maintains consistent performance throughout.

LSTM displayed a slower convergence rate, likely due to its complexity and hyperparameter sensitivity.

Figure 12 compares the DNN's training and validation loss/accuracy, showing rapid convergence and minimal overfitting.

Figure 13 illustrates RNN's training vs validation loss/accuracy, with a slightly slower convergence than DNN.

Figure 14 presents LSTM's slower convergence, reflecting its sensitivity to hyperparameters.

DNN was the most accurate model, with the lowest errors and R². Modeling complicated, non-linear connections makes it excellent for PV system optimization. Both residual analysis and prediction accuracy demonstrate that it performs best among all models.

In energy production, RNN excelled in sequential data processing, leading to the most considerable yearly energy output and CO_2 mitigation. This makes it ideal for energy system time-series forecasts, mainly when modeling sequential relationships.

XGBoost may be a useful option for settings with limited resources because to its balance between accuracy and processing efficiency. Although not the most accurate, its solid performance and efficiency make it excellent for real-time applications.

Improvement Potential for LSTM: LSTM underperformed other models, although hyperparameter adjustment and more datasets may improve performance. With additional tuning, LSTM might outperform other models in complicated time-series data, making it useful in energy prediction models.

Research indicates that machine learning models like DNN, RNN, XGBoost, and LSTM may effectively predict energy production and reduce CO_2 emissions. DNN was the most accurate model, but RNN predicted energy production best. XGBoost balanced efficiency and precision, but LSTM needed more tweaking. Machine learning in renewable energy forecasting systems is better understood with these findings, which might improve energy sustainability.

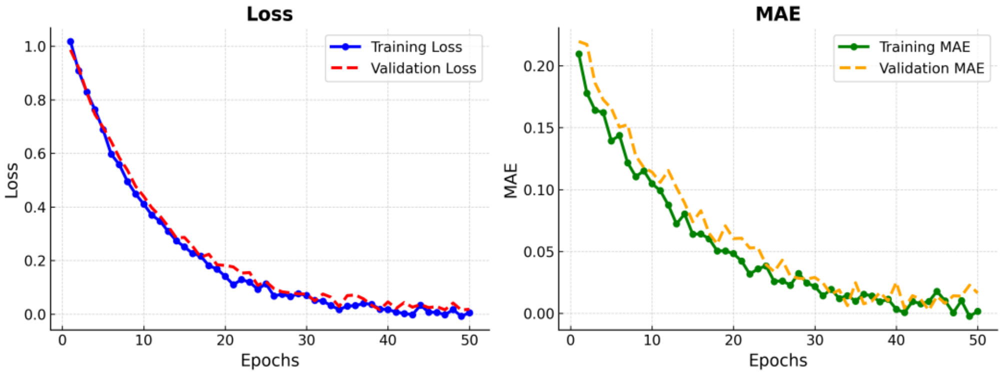

Figure 12. DNN training vs. validation (Loss, MAE)

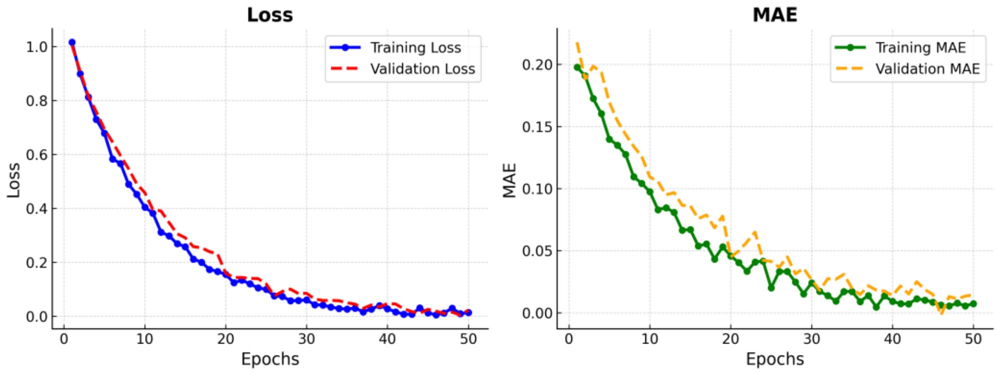

Figure 13. RNN training vs. validation (Loss, MAE)

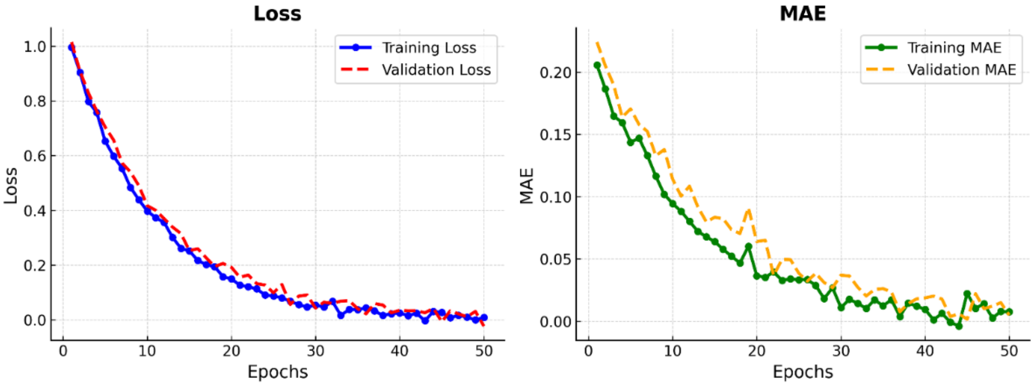

Figure 14. LSTM training vs. validation (Loss, MAE)

5. CONCLUSION

This study illustrates the considerable potential of ML techniques for enhancing the performance of PV systems, specifically in the M'Sila region of Algeria, known for its high solar radiation levels. This research employs advanced machine learning models, including DNN, RNN, XGBoost, and LSTM, to comprehensively analyze energy production, CO_2 emissions reduction, and overall system efficiency.

The DNN model outperformed other models in predicting accuracy, with the lowest MAE, MSE, and greatest R^2 value. This shows the ability to capture complex, non-linear environmental-system connections. The RNN model excels in predicting yearly energy output and CO_2 emission reductions, demonstrating its capacity to handle sequential data. XGBoost is ideal for real-time applications due to its precision and processing efficiency. While LSTM performed somewhat worse, its ability to capture long-term relationships in time-series data suggests that more tuning might improve energy predictions.

These findings highlight the importance of machine learning in improving photovoltaic system efficiency, sustainability, and cost-effectiveness, especially in solar-rich places like M'Sila. This research uses machine learning models and real-world data to maximize energy production, save costs, and decrease environmental impact. The results improve Algeria's renewable energy policy and provide a scalable model for other solar-rich locations.

This study represents a notable progression in implementing machine learning within renewable energy systems. Enhancing prediction accuracy and system performance through these techniques is essential for developing more efficient and sustainable solar energy solutions globally. This initiative aligns with Algeria's renewable energy objectives and serves as a framework for international endeavors to enhance energy security and combat climate change.

REFERENCES

[1] Sebastian, P.J., Gamboa, S.A., Campos, J. (2018). Design and development of a real-time characterization system for energy conversion devices. Journal of New Materials for Electrochemical Systems, 21(1): 7-13. https://doi.org/10.14447/jnmes.v21i1.515

[2] Almajeed, L.A., Fadhil, L., Rasheed, A.N., Gaeid, K.S. (2024). Enhancing photovoltaic panel performance through artificial neural network and maximum power point tracking. Journal Européen des Systèmes Automatisés, 57(3): 877-886. https://doi.org/10.18280/jesa.570327

[3] Agbakwuru, V., Obidi, P.O., Salihu, O.S., MaryJane, O.C. (2024). The role of renewable energy in achieving sustainable development goals. International Journal of Engineering Research Updates, 7(2): 13-27 https://doi.org/10.53430/ijeru.2024.7.2.0046

[4] Attia, M., Bechouat, M., Sedraoui, M., Aoulmi, Z. (2022). An optimal linear quadratic regulator in closed loop with boost converter for current photovoltaic application. European Journal of Electrical Engineering/Revue Internationale de Génie Electrique, 24(2): 97-103. https://doi.org/10.18280/ejee.240204

[5] Kumar, N.B., Veeranjaneyulu, J., Chandra, B.M., Venkatesh, P.M. (2024). An optimized superconducting magnetic energy storage for grid connected systems. Journal of New Materials for Electrochemical Systems, 27(1): 52-59. https://doi.org/10.14447/jnmes.v27i1.a08

[6] Nosheen, M., Akbar, A., Sohail, M., Iqbal, J., Hedvicakova, M., Ahmad, S., Gul, A. (2024). From fossil to future: The transformative role of renewable energy in shaping economic landscapes. International Journal of Energy Economics and Policy, 14(4): 606-615. https://doi.org/10.32479/ijeep.16006

[7] Samadi, S., Fischer, A., Lechtenböhmer, S. (2023). The renewables pull effect: How regional differences in renewable energy costs could influence where industrial production is located in the future. Energy Research & Social Science, 104: 103257. https://doi.org/10.1016/j.erss.2023.103257

[8] Ouada, M., Meridjet, M.S., Saoud, M.S., Tlbi, N. (2013). Increase efficiency of photovoltaic pumping system based BLDC motor using fuzzy logic MPPT control. WEAS Transactions on Power Systems, 8.

[9] Saadsaoud, M., Ahmed, A., Er, Z., Rouabah, Z. (2017). Experimental study of degradation modes and their effects on reliability of photovoltaic modules after 12 years of field operation in the steppe region. Acta Physica Polonica A, 132(3): 930-935. https://doi.org/10.12693/APhysPolA.132.930

[10] Borah, P., Micheli, L., Sarmah, N. (2023). nalysis of soiling loss in Photovoltaic modules: A review of the impact of atmospheric parameters, soil properties, and mitigation approaches. Sustainability, 15(24): 16669. https://doi.org/10.3390/su152416669

[11] Garg, A., Sarojwal, A., Sharma, D.D. (2024). Probing the impact of dust on solar photovoltaic performance using cutting edge techniques for performance optimization. In 2024 IEEE Region 10 Symposium (TENSYMP), New Delhi, India, pp. 1-7. https://doi.org/10.1109/TENSYMP61132.2024.10752232

[12] Barrie, I., Agupugo, C.P., Iguare, H.O., Folarin, A. (2024). Leveraging machine learning to optimize renewable energy integration in developing economies. Global Journal of Engineering and Technology Advances, 20(3): 80-93. https://doi.org/10.30574/gjeta.2024.20.3.0170

[13] Moussa, A., Aoulmi, Z. (2025). Improving electric vehicle maintenance by advanced prediction of failure modes using machine learning classifications. Eksploatacja i Niezawodność – Maintenance and Reliability, 27(3). https://doi.org/10.17531/ein/201372

[14] Benasla, M., Boukhatem, I., Allaoui, T., Berkani, A., Korba, P., Sevilla, F.R.S., Belfedel, M. (2024). Algeria's potential to supply Europe with dispatchable solar electricity via HVDC links: Assessment and proposal of scenarios. Energy Reports, 11: 39-54. https://doi.org/10.1016/j.egyr.2023.11.039

[15] Babalola, T.V., Nafada, A.I., Dala, H.A. (2024). An Assessment of the Impact of Accumulated dust on efficiency and performance output of solar photovoltaic panels. Nigerian Journal of Physics, 33(2): 87-94. https://doi.org/10.62292/njp.v33i2.2024.232

[16] Pronichev, A.V., Shishkov, E.M. (2023). Assessing and predicting degradation of solar panels using machine learning approach. In 2023 6th International Scientific and Technical Conference on Relay Protection and Automation (RPA), Moscow, Russian Federation, pp. 1-16. https://doi.org/10.1109/RPA59835.2023.10319847

[17] Priyadarshini, R., Manoharan, P.S., Usha, N., Kalimuthu, M., Suriyakumari, D., Deepamangai, P. (2024). Integrating AI and energy systems: LSTM model for photovoltaic performance prediction. In 2024 International Conference on IoT Based Control Networks and Intelligent Systems (ICICNIS), Bengaluru, India, pp. 752-756. https://doi.org/10.1109/ICICNIS64247.2024.10823319

[18] Deka, K., Rabha, M., Hazarika, H., Gourisaria, M.K., Bilgaiyan, S., Patra, S.S. (2024). Harnessing machine learning for improved solar radiation prediction. In 2024 Second International Conference on Networks, Multimedia and Information Technology (NMITCON), Bengaluru, India, pp. 1-6. https://doi.org/10.1109/NMITCON62075.2024.10699121

[19] Djeldjeli, Y., Taouaf, L., Alqahtani, S., Mokaddem, A., Alshammari, B.M., Menni, Y., Kolsi, L. (2024). Enhancing solar power forecasting with machine learning using principal component analysis and diverse statistical indicators. Case Studies in Thermal Engineering, 61: 104924. https://doi.org/10.1016/j.csite.2024.104924

[20] Prajapati, D.K., Joshi, R. (2024). A comparison of machine learning methods for forecasting solar energy production. In 2024 International Conference on Advances in Computing Research on Science Engineering and Technology (ACROSET), Indore, India, pp. 1-6. https://doi.org/10.1109/ACROSET62108.2024.10743318

[21] Amjad, M., Asim, M., Azhar, M., Farooq, M., et al. (2021). Improving the accuracy of solar radiation estimation from reanalysis datasets using surface measurements. Sustainable Energy Technologies and Assessments, 47: 101485. https://doi.org/10.1016/j.seta.2021.101485

[22] Mariprasath, T., Cheepati, K.R., Rivera, M. (2024). Practical Guide to Machine Learning, NLP, and Generative AI: Libraries, Algorithms, and Applications. CRC Press.

[23] Tesch, T., Kollet, S., Garcke, J. (2023). Causal deep learning models for studying the earth system. Geoscientific Model Development, 16(8): 2149-2166. https://doi.org/10.5194/gmd-16-2149-2023

[24] Soldatenko, S., Angudovich, Y. (2024). Using machine learning for climate modelling: Application of neural networks to a slow-fast chaotic dynamical system as a case study. Climate, 12(11): 189. https://doi.org/10.3390/cli12110189

[25] Jayasankar, K.C., Anandhakumar, G., Kalaimurugan, A. (2024). Prediction of solar radiation using deep LSTM-based machine learning algorithm. Journal of Environmental Nanotechnology, 13(3): 1-8. https://doi.org/10.13074/jent.2024.09.242585

[26] La Fata, A., Amin, M.A., Invernizzi, M., Procopio, R. (2024). Structurally tuned LSTM networks to nowcast photovoltaic power production. In 2024 IEEE International Conference on Environment and Electrical Engineering and 2024 IEEE Industrial and Commercial Power Systems Europe (EEEIC / I&CPS Europe), Rome, Italy, pp. 1-6. https://doi.org/10.1109/EEEIC/ICPSEurope61470.2024.10751363

[27] Nketiah, E.A., Chenlong, L., Yingchuan, J., Aram, S.A. (2023). Recurrent neural network modeling of multivariate time series and its application in temperature forecasting. Plos One, 18(5): e0285713. https://doi.org/10.1371/journal.pone.0285713

[28] Song, C.E., Li, Y., Ramnani, A., Agrawal, P., et al. (2024). 52.5 TOPS/W 1.7 GHz reconfigurable XGboost inference accelerator based on modular-unit-tree with dynamic data and compute gating. In 2024 IEEE Custom Integrated Circuits Conference (CICC), Denver, CO, USA, pp. 1-2. https://doi.org/10.1109/CICC60959.2024.10529017

[29] Chimphlee, W., Chimphlee, S. (2024). Hyperparameters optimization XGBoost for network intrusion detection using CSE-CICIDS 2018 dataset. IAES International Journal of Artificial Intelligence, 13(1): 817-826. https://doi.org/10.11591/ijai.v13.i1.pp817-826

[30] Niazkar, M., Menapace, A., Brentan, B., Piraei, R., Jimenez, D., Dhawan, P., Righetti, M. (2024). Applications of XGBoost in water resources engineering: A systematic literature review (Dec 2018–May (2023). Environmental Modelling & Software, 174: 105971. https://doi.org/10.1016/j.envsoft.2024.105971

[31] Raghuvanshi, K.P. (2024). A systematic literature review on the role of LSTM networks in capturing temporal dependencies in data mining algorithms. International Journal for Research in Applied Science and Engineering Technology, 12(10): 1219-1224. https://doi.org/10.22214/ijraset.2024.64761

[32] Waqas, M., Humphries, U.W., Hlaing, P.T., Ahmad, S. (2024). Seasonal WaveNet-LSTM: A deep learning framework for precipitation forecasting with integrated large scale climate drivers. Water, 16(22): 3194. https://doi.org/10.3390/w16223194

[33] Kong, Y., Wang, Z., Nie, Y., Zhou, T., et al. (2025). Unlocking the power of LSTM for long term time series forecasting. Proceedings of the AAAI Conference on Artificial Intelligence, 39(11): 11968-11976. https://doi.org/10.1609/aaai.v39i11.33303

[34] Mohanta, S.K., Mohapatra, A.G., Mohanty, A., Nayak, S. (2024). Deep learning is a State-of-the-Art Approach to Artificial Intelligence. In Deep Learning Concepts in Operations Research, pp. 27-43.

[35] Abdulla, H., Sleptchenko, A., Nayfeh, A. (2024). Photovoltaic systems operation and maintenance: A review and future directions. Renewable and Sustainable Energy Reviews, 195: 114342. https://doi.org/10.1016/j.rser.2024.114342

[36] Aftabi, N., Moradi, N., Mahroo, F. (2025). Feed-forward neural networks as a mixed-integer program. Engineering with Computers. https://doi.org/10.1007/s00366-025-02114-2

[37] Jin, H., Zhou, Y., Hussain, Y. (2023). Enhancing code completion with implicit feedback. In 2023 IEEE 23rd International Conference on Software Quality, Reliability, and Security (QRS), Chiang Mai, Thailand, pp. 218-227. https://doi.org/10.1109/QRS60937.2023.00030

[38] Ahlawat, S. (2022). Recurrent neural networks. In Reinforcement Learning for Finance: Solve Problems in Finance with CNN and RNN Using the TensorFlow Library, pp. 177-232. https://doi.org/10.1007/978-1-4842-8835-1_4

[39] Kiriakidou, N., Livieris, I.E., Diou, C. (2024). C-XGBoost: A tree boosting model for causal effect estimation. In Artificial Intelligence Applications and Innovations: 20th IFIP WG 12.5 International Conference, AIAI 2024, Corfu, Greece, pp. 58-70. https://doi.org/10.1007/978-3-031-63219-8_5

[40] Li, H., Cao, Y., Li, S., Zhao, J., Sun, Y. (2020). XGBoost model and its application to personal credit evaluation. IEEE Intelligent Systems, 35(3): 52-61. https://doi.org/10.1109/MIS.2020.2972533

[41] Mangaonkar, K.R. (2024). Integrating system dynamics and multivariate forecasting in a simulation-based dynamic inventory optimization model. Master's thesis, State University of New York at Binghamton.

[42] Okut, H. (2021). Deep learning for subtyping and prediction of diseases: Long-short term memory. In Deep Learning Applications. IntechOpen.

[43] Gao, H., Oates, T. (2018). On finer control of information flow in LSTMs. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 527-540. https://doi.org/10.1007/978-3-030-10925-7_32

[44] Levy, O., Lee, K., FitzGerald, N., Zettlemoyer, L. (2018). Long short-term memory as a dynamically computed element-wise weighted sum. arXiv preprint arXiv:1805.03716. https://doi.org/10.48550/arXiv.1805.03716

[45] Correndo, A.A., Rosso, L.H.M., Hernandez, C.H., Bastos, L.M., Nieto, L., Holzworth, D., Ciampitti, I.A. (2022). Metrica: An R package to evaluate prediction performance of regression and classification point-forecast models. Journal of Open Source Software, 7(79): 4655. https://doi.org/10.21105/joss.04655

[46] Yang, D. (2022). Correlogram, predictability error growth, and bounds of mean square error of solar irradiance forecasts. Renewable and Sustainable Energy Reviews, 167: 112736. https://doi.org/10.1016/j.rser.2022.112736

[47] Settu, P., Ramaiah, M. (2024). Estimation of Sentinel-1 derived soil moisture using modified Dubois model. Environment, Development and Sustainability, 26(11): 29677-29693. https://doi.org/10.1007/s10668-024-05460-1