Journal of Engineering and Technology for Industrial Applications

ITEGAM-JETIA

Manaus, v.11 n.56, p.87-98. November/December., 2025.
DOI: https://doi.org/10.5935/jetia. v11i56.2698

ISSN ONLINE: 2447-0228

RESEARCH ARTICLE OPEN ACCESS

ACCURATE PHOTOVOLTAIC POWER PREDICTION USING MACHINE LEARNING AND DEEP LEARNING: A COMPARATIVE STUDY ACROSS MULTIPLE LOCATIONS

Aythem Khairi Kareem[1], Ahmed Adil Nafea[2], Manar Thayer Mansour[3], Nasrin Nadher Jamil[4]

[1]Department of Heet Education, General Directorate of Education in Anbar, Ministry of Education, Heet, 31007 Anbar, Iraq
[2]Department of Artificial Intelligence, College of Computer Science and IT, University of Anbar, Ramadi, Iraq
[3]Department of Medical Physics, College of Science, AL-Nahrain University, Jadriya, Baghdad, 10072, Iraq.
[4]Department of Vocational Education in Anbar, Ministry of Education, Anbar, Iraq.

[1]https://orcid.org/0000-0001-7855-7843, [2]https://orcid.org/0000-0003-2293-1108, [3]https://orcid.org/0009-0006-5104-0979, [4]https://orcid.org/0000-0002-8879-1788

E-mail: ayt19c1004@uoanbar.edu.iq, ahmed.a.n@uoanbar.edu.iq, manar.thaer@nahrainuniv.edu.iq, nasrinjamil@uoanbar.edu.iq.

ARTICLE INFO

Article History
Received: September 12, 2025.
Revised: October 20, 2025.
Accepted: November 1, 2025.
Published: November 30, 2025.

Keywords:
Photovoltaic Power Forecasting,
Solar Energy Prediction,
Machine Learning,
Deep Learning,
Renewable Energy.

ABSTRACT

This work evaluates machine learning (ML) and deep learning (DL) models for Photovoltaic (PV) power output prediction based on two real-world datasets. Eight models – Support Vector Regression (SVR), K-Nearest Neighbors (KNN), Linear Regression (LR), Random Forest Regression (RFR), Gradient Boosting Regression (GBR), Decision Tree (DT), one Dimensional Convolutional Neural Network (1DCNN) and Artificial Neural Network (ANN) – were evaluated using Root Mean Squared Error (RMSE), R-squared (R^2), Mean Absolute Error (MAE), and Mean Squared Error (MSE). For Dataset1 (region of Cacak) 1DCNN achieved the optimal performance, with the reduced errors and the largest R^2 values at three spatial locations, and it was closely followed by RFR and GBR. Likewise, in Dataset2 (region of Kraljevo), 1DCNN, ANN, and RFR give the best results. Conventional models, such as LR and DT, performed poorly in both data tribes. The results highlight the ability of state-of-the-art ensemble and DL models in learning nonlinear patterns in solar energy data and thus, the importance of choosing the right prediction tools to provide accurate and reliable PV power forecasts.

I. INTRODUCTION

Climate change, rising living standards, and population growth are all contributing to the energy demand, which is predicted to rise sharply in the near future. This underlines the difficulties that the current electricity infrastructures face and the necessity of continuing research and growth in the domain of Renewable Energy (RE) [1], [2], [3]. A key component of RE technologies, PV technology uses Solar Radiation (SR) to create electricity. PV technology is continuously offering a sustainable substitute for other consumables and fossil fuels. The conversion efficiency, the effects of different environmental factors, and the modification of operational settings to maximum energy generation are all part of the physical electrical output analysis via PV cells [4], [5]. However, the functioning and proficiency of PV. Solar irradiance, or the quantity of sunshine Captured by a given surface over a given time interval, has a substantial effect on the staging and efficiency of solar power systems. Predicting the quantity of this RE source is becoming more and more crucial as a result of the rising use of solar-powered electricity on the one hand and the increasing integration of this energy source into the electrical grid on the other [6], [7]. Suboptimal energy production, storage, and distribution result from traditional forecasting systems' inability to manage the complex and nonlinear interactions between these contributing variables. This leads to the following particular difficulties:

1. Weather Variability: Accurately predicting solar irradiance and energy output is challenging due to unpredictable weather variations.
2. Complex Interactions: To make precise predictions, sophisticated techniques are needed to account for the interactions between variables like temperature, panel orientation, and shading.

3. Needs for Real-Time Prediction: Real-time or near-real-time forecasts are necessary for energy firms and grid operators to maximize energy storage and preserve grid stability.
4. Absence of Granular Data: The precision of PV generation models is impacted by the scarcity of localized, high-resolution data.
5. Energy Storage Optimization: Reducing expenses and energy waste requires precise forecasts for effective energy storage system operation.
6. Traditional predicting techniques struggle to handle these influencing variables' complex and nonlinear associations, leading to suboptimal energy production, distribution and storage.

Within the field of artificial intelligence, ML applies use of data and algorithms to allow machines to simulate human thought learning, processes, and decision-making without the need for specialized programming. By using big datasets and sophisticated algorithms to evaluate both historical and current data for PV energy projection, ML presents a possible option. By modeling intricate patterns, ML can enhance: Forecasting Solar Power in the Short and Long Term and Grid Integration Efficiency. Optimization of Energy Storage: Identification of Faults and Scheduling of Maintenance. Stakeholders may improve operational effectiveness, lower energy losses, and accelerate the transition to a more sustainable and dependable solar energy future by incorporating ML models into PV systems [8], [9]. The effective operation of solar power systems depends on precise PV production forecasting, which is achieved through the use of ML techniques [8], [9]. With further investigation and progress, ML will probably become more significant in estimating PV output in the years to come [10]. The use of ML techniques has spread widely during the last few decades in a variety of businesses that address data-oriented problems. Whether or not they adhere to formal problem frameworks, these methods seek to identify relationships between input and output data.

One well-known ML technique that predicts a target value using input variables is regression, which is a supervised learning approach. It is frequently used in forecasting to comprehend the relationships between several factors. The type of link between input variables and the number of independent variables taken into account are two aspects of regression models that can differ. Finding and establishing the associations between Predictor and target variables is the main goal of this analysis[11], [12]. The proposed approach presents an intelligent approach for predicting PV output power based on multiple ML and DL techniques, including RFR, LR, DT, GBR, SVR, KNN, ANN, and 1DCNN. By showcasing the potential of cutting-edge ML algorithms in maximizing PV system performance, this research advances solar energy usage. Implications based on operational and environmental parameters, make highly accurate predictions about short-term (minutes to hours), medium-term (daily), and long-term (seasonal or annual) energy generation. By anticipating PV power generation and balancing supply and demand, you can guarantee the stability of electrical networks and facilitate seamless grid operations. This work is arranged as follows: in Section 1 presents the background and method. Section 2 reviews related work. The method and models used are presented in Section 3. Presented the results and analyzed in Section 4. Section 5 concludes the proposed methodology and suggests future work's direction.

II. RELATED WORK

Solar PV panels are composed of many solar cells transforming sunlight to electrical energy. ML has been increasingly utilized to address the shortcomings of traditional forecasting methods with the widespread application of PV systems. The complicated nonlinear associations between environmental and operational variables may be learned by ML algorithms to generate more accurate and flexible prediction models of solar power generation. This subsection presents a review on last contributions that make use of ML and hybrid methods to improve the prediction of PV systems and increases their operational quality. Yi Zhou et al. suggested that a mixing model of Same Days Analysis (SDA), Genetic Algorithms (GA) and ELM (Extreme Learning Machines) could effectively forecast the hourly value of PV power[13]. SDA adopts the training samples according to meteorological similarity in terms of Pearson correlation, consequently the training time is shortened and the reference samples are focused. GA is employed thereafter to optimize these hidden layer weights of the ELM under this approach an improved prediction accuracy is achieved. Berny Carrera et al. introduced a cross-ratio-based framework to predict the solar power for the Yeongam solar plant, South Korea, 36 hours ahead [14].

The performance was evaluated in terms of 5 daily observational time slots, to ensure a consistent basis for the comparison of different advanced forecasting techniques. Wei Zhao et al. presented a hybrid approach of AML and an improved GA were employed to optimize day-ahead simulation of SPG of multiple PV plant regions [15]. The model integrates physical information of SPG dynamics and climatology in 2016–2018, outperforming a number of base-line models. Kumar Shivam et al. presented a multi-objective ML based energy optimization program for the domestic hybrid energy systems that consists of a solar array, battery storage, and electric load [16]. Their model employs a dilated CNN to predict energy generation and consumption: optimization to ensure efficient grid interaction is included as part of their framework. Sohaila Chahboun et al. analyzed six ML models using a two-year meteorological and radiation data to predict PV output [17]. Bayesian Regularized Neural Networks achieved the highest accuracy in comparison with other approach in terms of RMSE, MAE, and R^2 criteria.mRamyar Saeedi et al. introduced a novel adaptive ML system that can adapt with the help of small sensor information.

The framework estimates PV outputs at non-monitored sites by using the state of art nowca[18]sting models and estimating the total load demand through the combination of predicted PV outputs and smart meter data, which is crucial to local grid control.mKhizir Mahmud et al. used different ML models in short- and long-term predictions over Alice Springs, Australia [19]. LSTM and RF, and SVR models were compared under the normal and uncertain circumstances. RF showed the best performance in various instances, in particular. Zeming He et al. investigated some types of hybrid systems with PV cells and TEG [20]. They applied ML methods for optimal arrangement of the structure parameters and power estimation with respect to the PV/TEG area ratio for hybrid energy systems design. Hamza Mubarak et al. investigated the predictive ability of several ML methods, including KNN, SVM, ETR, and FFNN, for thin-film PV systems [21]. Extra Tree Regression had the best performance and DT Regression had the worst. Primary input characteristics were irradiance, wind speed and module temperature. Shadrack T. Asiedu et al. tested single and hybrid ML techniques to predict SE output from a 180 kWp PV system for several horizon terms in the future (from 1 day up to 1 month) [9]. ANN modelling was optimal for 1-day ahead forecasts and RF and the hybrid RF-XGBoost models achieved better prediction results for longer forecasting horizons.

Jing Zhang et al. presented a SE predicting model based upon Extreme Learning Machine with optimization through an Improved Moth-Flame Optimization (IMFO) algorithm [22]. The performance of the presented IMFO-ELM approach was also confirmed utilizing Desert Knowledge Australia Solar Center data by obtaining a MAPE below 5%. There are still some drawbacks in the previous works even though they have achieved a remarkable progress. Most of the models are either too simple to include the complex non-linear relationship in solar power generation, or are not wide applicability as validated only for limited datasets or regions. Moreover, some methods are heavily dependent on primitive overlapping functions or evaluated based on few performance indices that may not completely reflect the practical usability. By comparison, we address all these lacks in this study by testing different types of ML/DL models (e.g., 1DCNN, RF, ANN) in different spatial positions of different datasets. This all-inclusive evaluation model, with some error assessments (MAE, MSE, RMSE, R^2), validates the stability and performanceof advanced models for predicting the PV power output under different circumstances.

III. METHODOLOGY

This section describes the presented framework for implementing the PV approach based on ML and DL. Figure 1 represents the analysis's methodological workflow. It begins with the dataset description and then preprocessing operations that are applied to the dataset. Several ML and DL are employed to predict electricity production on an hourly basis. The techniques are RFR, LR, DT, GBR, SVM, ANN, and 1DCNN. Finally, the proposed approach's performance is evaluated using MAE, MSE, RMSE, and R2.

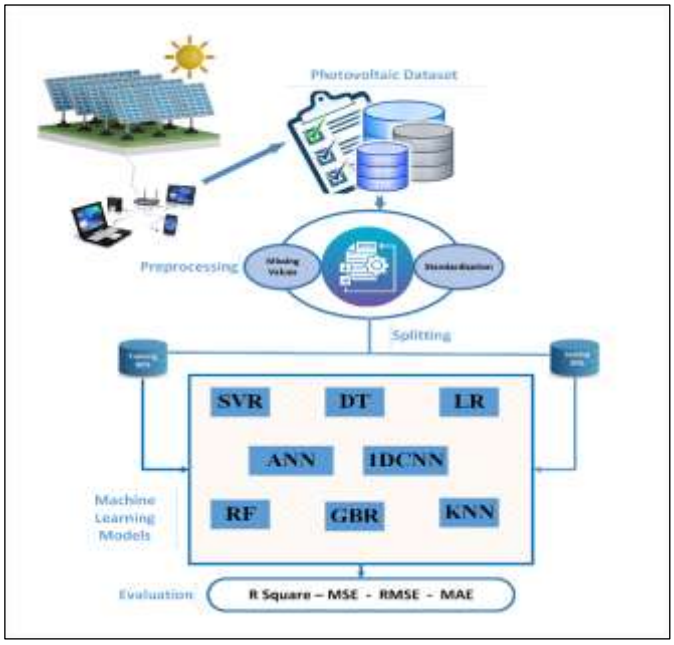

Figure 1: Block diagram of the proposed approach.
Source: Authors, (2025).

III.1 DATASET D ESCRIPTION

The presented approach utilizes an international solar energy dataset called the "Solar energy production dataset" [23]. It applies ML techniques to predict electricity production on an hourly basis. It includes two datasets called (Data_Cacak) and (Data_Kraljevo), each containing 8,274 instances with seven features in addition to the label representing the electricity production from solar panels. Table 1 indicates the feature description of the utilized dataset.

Table 1: Feature description.

Feature	Description	Unit of measure
Date time	The time and date	MM/DD/YYYY HH:MM
Air Temperature	The temperature of air	degrees Celsius (°C)
Cloud Opacity	The number of clouds in the sky. Cloud Opacity: Levels report the transparency of the cloud; higher units indicate higher cloud opacity (more cloud cover) resulting in less light. Higher down values indicate a clearer sky with direct sunlight.	percentage (%)
DHI	Diffuse Horizontal Irradiance (DHI): DHI is the SR that comes to a horizontal plane from the total sky. It features a combination of diffuse and direct sunlight.	kilowatts per square meter (kW/m^2)
DNI	Diffuse Horizontal Irradiance (DHI): DHI is the SR that comes to a horizontal plane from the total sky. It features a combination of diffuse and direct sunlight.	kW/m^2
EBH	Direct Normal Irradiance (DNI) is the SR received directly from the sun on a unit area perpendicular to the sun's rays.	kW/m^2
GHI	The amount of SR that would reach the surface of the Earth if the atmosphere were not present is known as the extraterrestrial horizontal irradiance (EBH).	kW/m^2

Source: Authors, (2025).

III.2 PREPROCESSING

Data analysis and ML workflows require preprocessing as a fundamental step. This step applies to transform raw data to prepare it for the following analysis and modeling tasks. This process ensures cleanliness and consistency while formatting the data correctly for its intended use. Preparing the acquired data for analysis requires multiple processing steps. This work used two preprocessing methods: Missing Values and Standardization.

III.2.1 Missing Values

A "Solar energy production dataset" contains missing values, which can pose challenges in data analysis. ML techniques cannot be trained directly with missing data. The simplest solution is to eliminate all instances with missing values, which is considered undesirable because we may lose important data. Another solution is to impute missing values with special values [24]. The proposed approach imputes missing values with a mean of the feature. This method is functional only for numeric characteristics and is usually connected with replacing missing values with the most common characteristic value for symbolic features.

III.2.2 Data Standardization

By following the range of values for each attribute, it is encountered that the data values of various features vary significantly. In ML techniques, extra dimensions have a significant impact on the forecast performance, so the data are instructed to be standardized before training. However, several methods, including Standardscaler, Z-score standardization and Min-max standardization [25], [26]. The proposed approach employs the Standardscaler method to perform this operation. The mathematical form of standardization is represented in Equation (1) [27].

$$Standardization\ x = (x - \mu)/\sigma \tag{1}$$

Where σ is standard division, μ is mean.

III.3 SPLITTING DATASET

In this step, the data is separated into two parts: the training subset and the testing subset. In the proposed approach, 70% of the data is training data, and the rest is testing data.

III.4 MACHINE LEARNING TECHNIQUES

First, The ML methodologies presented in this analysis for power predictions are based on prediction and regression approaches. Regression approaches comprehend the association between features or independent variables and an output or dependent variable. Results can then be forecasted once the association among dependent and independent variables has been assessed. Regression in statistics recreates a central role in predicting representatives in ML. It is employed as a method to forecast real-valued developments in Forecasting modeling outcomes from data. The proposed approach employs various ML techniques, Including RFR, LR, KNN, DT, SVR, ANN, and 1DCNN.

III.4.1 Linear Regression (LR)

An LR possesses the labeled data (supervised ML), which creates the association among dependent and independent variables operating straightforward scientific equations and thereby estimates the line of lowest hesitancy or the best-fit line. This line can be utilized to estimate PV-power predictions utilizing curve or graph analysis. LR is higher than most of the alternative ML techniques as it is more straightforward to implement and includes fundamental computational and mathematical theory. It implicates modeling a Linear function to the provided data points of output variable(y) and input variable(x) whose slope is (n) and error is (er). This line is enforced by decreasing the sum of squared discrepancies among absolute values comprehended in the Equations (2) with (3) and Figure 2 [28].

$$y = n_x + v + e \tag{2}$$

$$y = n_1 x + n_2 x \ldots\ldots n_k e + v + e \tag{3}$$

Where, v is the bias term composed on the y-axis for several datasets with slopes $n_1, n_2 \ldots n_k$.

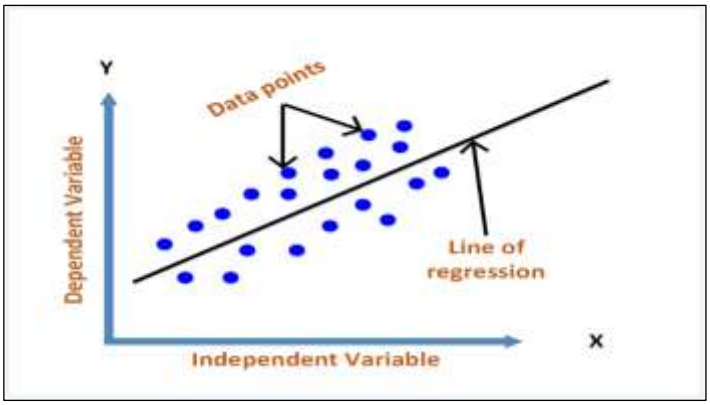

Figure 2: Illustration of best-fit line regression.
Source: Authors, (2025).

III.4.2 Decision Tree (DT)

DT technique belongs to the supervised ML. It is mainly selected to solve regression and classification issues. It contains of internal nodes describing the configurations of the branches and data, describing the finding delivered by the technique, and individually Leaf Node (LN) defining a result [29]. There are two nodes: Decision Node (DN), which operates to create a decision and has different branches, and the LN, which creates DNs and has no branches. Root Node (RN) is a initial attribute that additionally extends to different branches, creating a tree-like configuration. DT forks the tree into sub-trees on the basis of the answer to the question. The DT schematic diagram is presented in Figure 3, and the terminologies are presented in Table 2 [30].

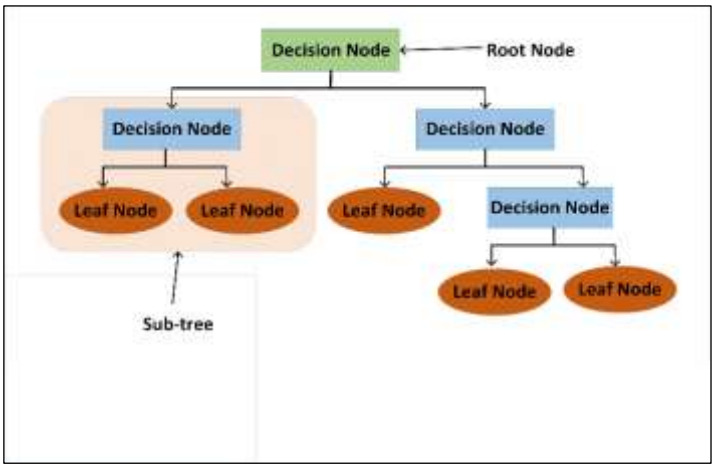

Figure 3: DT schematic diagram.
Source: Authors, (2025).

Table 2: The DT terminologies.

Terminology	Description
Root Node	The beginning portion of the DT is where the whole data begins to separate further into multiple potential homogeneous sets.
Leaf Node	It is the final result node, and no additional splitting of trees is likely.
Splitting	The process involves separating the main node additional, under the supplied conditions, into sub-nodes.
Sub Tree	It is a sub-tree produced by dividing up a hierarchy, resulting in a branch.
Pruning	It involves eliminating superfluous branches to obtain optimal outcomes. It really minimizes the size of the tree without hindering its performance. There are two kinds: error reduction and cost complexity trimming.
Child and Parent node	The parent node is the boot node, where the remaining nodes are called child nodes.

Source: Authors, (2022).

III.4.3 Random Forest Regression (RFR)

The RFR models are a group of forecasting trees, where multiple trees form a "forest" that can be utilized to split data components. A forecasting tree is a non-linear technique for formulation complicated data in which Predictor variables are partitioned into optimal separations until they reach partitions, permitting the homogeneous sub-nodes [31]. RFR is an ensemble representative. It improves predictive likelihood by integrating models. The prediction technique of RFR executes a best representative through a regression tree, as displayed in Figure 4 [32].

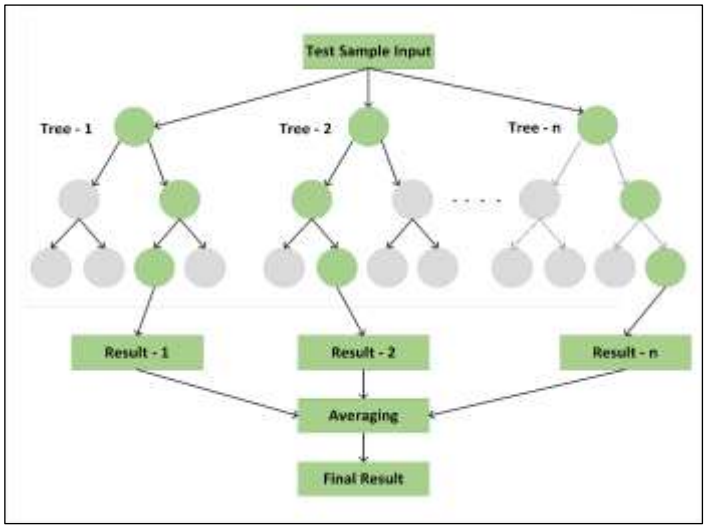

Figure 4: Schematic illustration of RFR.
Source: Authors, (2025).

The technique utilizing a regression tree is named bagging. It is applied to construct a tree-1 from the dataset, and then data are resampled and extracted from entire data. Tree-2, tree-3, and tree-n are then produced using the previous strategy. A forecast is generated using the number of average trees. The RFR technique is commonly employed in numerical forecast analysis. The performance of the RFR technique will be powerful because it emanates the prediction by pulling the representative typical significances on average utilizing the outcomes received via multiple DT.

III.4.4 Support Vector Regression (SVR)

SVR technique belongs to the supervised ML. It is Vapnik presented the idea of the SVR. The technique was initially developed for binary class and improved for forecasting numerical values, which is called SVR. The technique maps training dataset into a predefined characteristic feature space to emanate a model for quantitative or qualitative predictions via fitting a regression function. SVR allows the forecasts of numerical target values. It is built from training data x and label vector including the characteristic value y the of individually training data point. It represents a regression function of form $f(x) =< h.x > +c$ where, where $< h.x >$ is a scalar product, c is bias, and h is weight vector and attempts to map training data as near as likely to the numerical label, displayed in Figure 5. Determinate mathematical derivation from specific values is accepted by the ε-insensitive tube, where errors bigger than ε are punished. Therefore, ε represents the limitations for discrepancies among real and predicted values of training samples [33], [34].

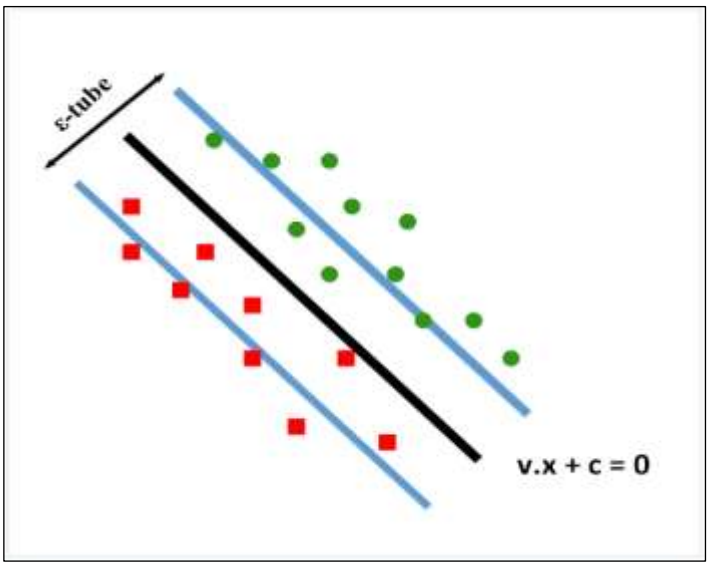

Figure 5: Schematic illustration of SVR.
Source: Authors, (2025).

The “kernel” plays an essential role in the SVR technique. Suppose LR is not achievable in an input space X. In that issue, the kernel is involved in mapping the training dataset into a transformed space, H, in which linearly separable structure possibility be feasible. Therefore, instead, the scalar product is substituted by a kernel function K(.,.). The proposed approach employs the Radial Basis Function (RBF) kernel, which can be presented in Equation (4).

$$K_{RBF}(v.u) = exp(-\gamma\|v-u\|^2) \tag{4}$$

Where, γ is hyper-parameter, and $v.u$ are two vectors.

III.4.5 Gradient Boosting Regression (GBR)

The GB is a DT-based ensemble ML technique for classification and regression problems [35]. The GB technique training methodology is similar to that of the RFR technique. However, RFR trains independently for each DT, while GBRs train DTs one at a time. By fitting individual trees in the arrangement to the residuals of the earlier tree, each subsequent tree is permitted to concentrate on the earlier tree’s missteps: fit a DT to the data $F_1(x) = y$, then the next DT was trained to the prediction errors of the earlier $G_1(x) = y - F_1(x)$, add this new tree to the algorithm $F_2(x) = f_1(x) + G_1(x)$, fit the next DT to the residuals of $F_2G_2(x) = y - F_2(x)$, Add this new tree to our algorithm $F_3(x) = F_2(x) + G_2(x)$, and continue this strategy until some mechanism informs to stop. The final representative here is a stage wise additive representative of b separate trees, as presented in Equation (5) [36]:

$$F(x) = \sum_{b=1}^{B} F^b(x) \tag{5}$$

III.4.6 K-Nearest Neighbor (KNN)

KNN is an effective, simple, and powerful nonlinear regression technique. The fundamental concept of KNN is to forecast a value to a provided input instance using a fixed number (k) of its nearest neighbors uncovered from the training instances. The k's value possesses the adaptability of the KNN technique. KNN does not need a clearly defined training phase besides the initial dataset, which describe an individual's possessions. The idea of the KNN representative can be formally described as follows.

Let $D = \{x_j, y_j\}_{j=1}^{M}$ be a training data with M instances, where $x_j = \{x_1^j, x_2^j, \dots \dots x_n^j\} \in R^n$ is an input instance J from n-dimensional feature space, and its output value is $y_j \in Y$, where $Y = \{y_1, y_2, \dots \dots \dots . y_m\}$represents the output values set. For a provided new data instance X, the objective is to comprehend the predictor function f(x) from the training data such that $\hat{y} \approx f(x)$, where $\hat{y}$ is the assessed value for the output y of X. The KNN begins by estimating the distance (T) among test instance X and each instance x_j in D. In this case, the distance of Euclidean is the most generally assumed distance metric, and its formulation for the distance among $X = \{x_1, x_2, \dots \dots \dots . x_n\}$ and X_j is presented by Equation (6) [37].

$$t(X, X_j) = \sqrt{\sum_{j=1}^{n} (x_j - x_j^i)^2} \tag{6}$$

Next, the set of K, $M_X^k = \{(X_j, y_j)\}_{j=1}^{k}$, is uncovered from the reordered training instances in T according to the boosting distances of Euclidean. Ultimately, the forecasting value y for is calculated by carrying the arithmetic of the forecasting values mean $\{y_1, y_2, \dots \dots \dots . y_m\}$ of the nearest neighbors as Equation (7):

$$\hat{y} = \frac{\sum_{j=1}^{k} y_j}{j} \tag{7}$$

This is using the assumption that training instances in the M_X^k have matching forecasting values to f(x), and all nearest neighbors in the M_X^k have equivalent significance in the prediction.

III.4.7 Artificial Neural Network (ANN)

An ANN is a data preparation technique built upon biological neuron approaches, such as the brain. Its objective is to employ internal estimations to calculate output values from input inputs. One commonly operated mathematical modeling mechanism is ANN. It is widely for forecasting and predicting values in complicated issues. Typically, it functions like human brain activity. Here, Hidden Layers (HLs) with neurons are current between the Input Layer (IL) and Output Layer (OL). The weights or coefficients with which each neuron is connected to the individual of its neighbors select the proportionate influence of different neuron inputs on other neurons. The number of HLs and hidden node selection are typical issues in ANN configuration. Development trial runs and hits are utilized to select appropriate values.

The number of HLs available is not known beforehand. Using a larger number of HLs outcomes in noise, which is generated by over parameterization. It generates inferior generalizations of the untrained dataset and requires n high amount of training time. Among multifarious network kinds of ANNs, in this analysis, we use Multi-Layer Perceptron (MLP) to forecast the power values of PV parameters. In this analysis, we use MLP to predict the power values of PV parameters throughout several network kinds of ANNs. In this kind, the data transmission is unidirectional. Here, the outputs of the HL are sequentially served as the inputs of the output neurons, where they proceed via additional changes. This MLP has various applications in different fields. For MLP, each $j$node of L hidden layer reserves by encountering the effects of the weight $v_{j,i}$ of x_j input (i.e., the node's output i in $L-1$ layer). The node's output is calculated as a function of the sum, as a formula in Equation (8) [38].

$$net_i^{(L)} = \sum_{j=1}^{m} v_{ji}^{(L-1)} x_j^{(L-1)} + c_i^{(L)} \; and \; x_i^{(L)} = f\left(net_i^{(L)}\right) = \frac{1}{1 + e^{-net_i^L}} \tag{8}$$

Where $net_i^{(L)}$ denotes the activation of the i^{th} node in layer L, n denotes the number of nodes in $L-1$ layer and $c_i^{(L)}$ is the bias or offset; $v_{ji}^{(L-1)}$ denotes the connection weight between node i Layer L and node j in layer $L-1$. f is the purelin activation function employed in this analysis. Figure 6 show the architecture of proposed ANN model.

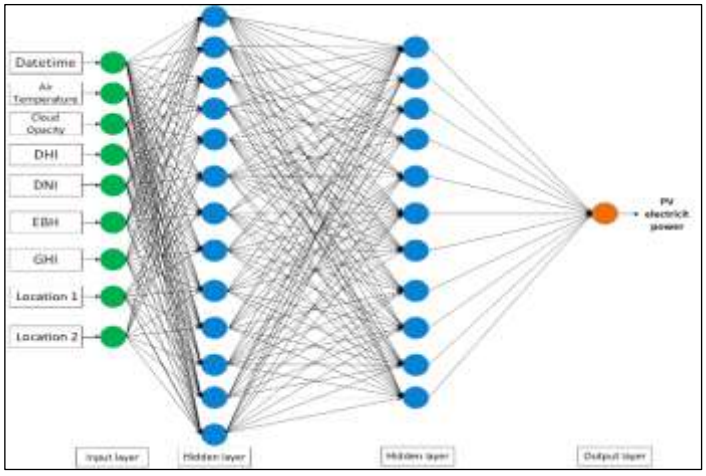

Figure 6: The architecture of the proposed ANN model.
Source: Authors, (2022).

III.4.8 One Dimension Convolutional Neural Network (1DCNN)

A CNN is a DL network that can understand high-level features from input features. A CNN is conceptually similar to an MLP, where each neuron maintains an activation function that cracks the weighted inputs into the outputs. 1DCNN is a modified understanding of CNN developed for processing one-dimensional data. Moreover, 1DCNNs use 1D convolutional filters to extract features from the dataset. 1DCNNs also have rarer parameters than 2DCNNs, which makes them more computationally efficient. A proposed 1DCNN is generally comprised of one IL, one Convolution Layer (CL), one Pooling Layer (PL), one Fully Connected Layer (FCL), and several dense layers [39]. Figure 7 and Table 3 shows the architecture of the proposed 1DCNN model.

Table 3: The architecture of the proposed 1DCNN model.

Layer	Description
IL	IL carries in the input data, which could be a series of numerical data (nine attributes).
CL	CL is a core part of a 1DCNN, which involves a several filters to the input data to generate significant characteristics. we set filters=64, padding='same', strides=1, kernel size=3, and the activation function= Relu.
PL	PL minimize the feature map dimension, reducing the data into a unified representation. It assists in decreasing the computational complication of the network and constructs it more vital to small contrasts in the input data. we set strides=1 and pool size=2.
FCL and dense layers	These layers construct the final layer in a 1DCNN. It carries the efficient model created by the earlier layers and involves a set of fully connected weights to create the final output.

Source: Authors, (2025).

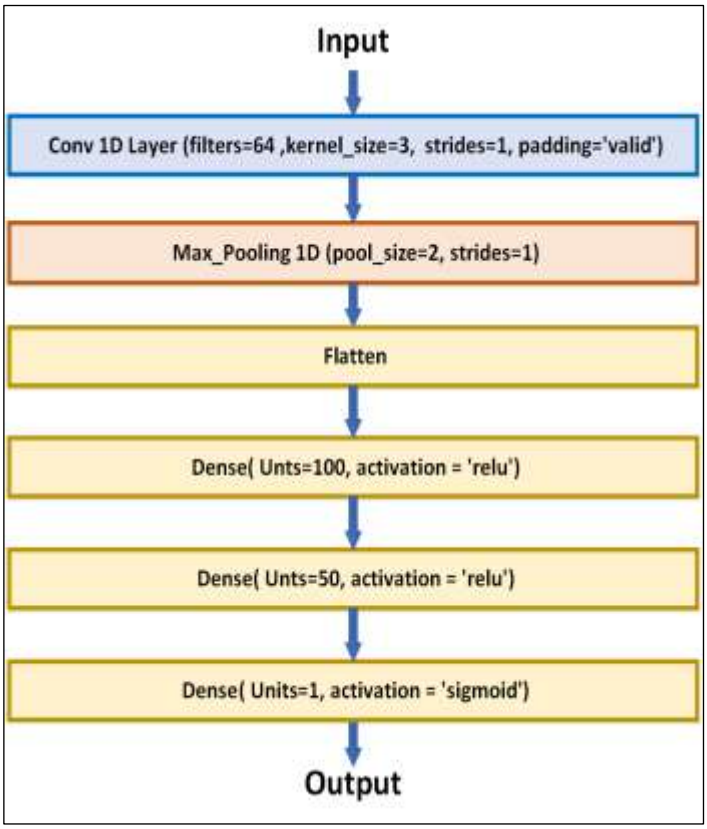

Figure 7: The architecture of the presented 1DCNN model.
Source: Authors, (2025).

During the training process for PV power prediction, 50 epochs were assigned, the batch size was selected to 32, and the optimization function was "ADAM".

III.5 EVALUATION METRICS

In this analysis, several statistical criteria were used to evaluate the ML and DL performance of techniques, including MSE, determination coefficient (R2 score), MAE, and RMSE. Table 4 shows mathematical equations for performance metrics [40].

Table 4: Mathematical equations for performance metrics.

Evaluation metrics and their equations	Parameters' definitions
$MAE = \sum_{j=1}^{m} \frac{y_{j,exp} - y_{j,pred}}{m}$	

$R^2 = \frac{\sum_{j=1}^{m}(y_{j,exp} - y_{j,pred})^2}{\sum_{j=1}^{m}\left(y_{j,exp} - \left(\frac{1}{m}\sum_{j=1}^{m} y_{j,exp}\right)\right)^2}$	m= describes the entire number of data values. $y_{j,exp}$=describes the actual value. $y_{j,pred}$=describes predicted value
$RMSE = \sqrt{\frac{1}{m}\sum_{j=1}^{m}(y_{j,exp} - y_{j,pred})^2}$	
$MSE = \frac{1}{m}\sum_{j=1}^{m}(y_{j,exp} - y_{j,pred})^2$	

Source: Authors, (2025).

IV. RESULTS AND DISCUSSION

The MSE and R² of the models were used for their evaluation. These scores help us understand the validity of our predictions. The optimal mean squared error (MSE) across all positions was obtained in Dataset1 and was 0.0857, 0.0441, 0.0703, which came from 1DCNN model, that is, it performed most favorably as shown in Table 5 ANN was the best performing followed by RFR and GBR. In contrast, we found that LR and DT had the worst MSE, meaning that they were the least accurate. ANN gave the lowest MSE (0.1384) in Dataset2, followed by 1DCNN with an MSE of 0.1485. RFR and GBR showed good performance with lower MSE, while DT reported highest MSE (0.3102) indicating high prediction errors as shown in Table 6.

Regarding the R² scores, which quantifying the variance of the data that is accounted for by the model, the best model was 1DCNN (0.91, 0.95, 0.93) and ANN (0.90, 0.95, 0.92) for Dataset1. RFR and GBR received also good results, while LR and DT obtained low scores. For Dataset2, both ANN (0.87) 1DCNN (0.86) were the best performing models, followed by RFR (0.85) and GBR (0.84). The weakest was the decision tree with R² such as 0.71, which was the worst performance. In the general case, 1DCNN and ANN models were best for both datasets, although RFR and GBR also emerged as good alternatives. The underlying patterns in the data were difficult to be modeled by DT and LR.

Table 5: Results of proposed models for Dataset1.

Model	Pos1				Pos2				Pos3			
	MAE	MSE	RMSE	R2	MAE	MSE	RMSE	R2	MAE	MSE	RMSE	R2
RFR	0.1462	0.1045	0.3233	0.89	0.0984	0.0497	0.2230	0.95	0.1201	0.0748	0.2736	0.92
LR	0.2811	0.2169	0.4657	0.78	0.1669	0.0742	0.2724	0.92	0.2241	0.1211	0.3481	0.88
DT	0.1919	0.2041	0.4518	0.79	0.1470	0.1175	0.3429	0.88	0.1630	0.1437	0.3791	0.86
GBR	0.1785	0.1227	0.3503	0.87	0.1088	0.0491	0.2217	0.95	0.1322	0.0745	0.2729	0.92
SVR	0.1807	0.1397	0.3738	0.86	0.1170	0.0496	0.2227	0.95	0.1361	0.0754	0.2746	0.92
KNN	0.1685	0.1380	0.3715	0.86	0.1155	0.0573	0.2394	0.94	0.1419	0.0873	0.2954	0.91
ANN	0.1666	0.0934	0.3057	0.90	0.1156	0.0497	0.2231	0.95	0.1379	0.0747	0.2733	0.92
1DCNN	0.1388	0.0857	0.2928	0.91	0.1030	0.0441	0.2100	0.95	0.1244	0.0703	0.2652	0.93

Source: Authors, (2025).

Table 6: Results of proposed models for Dataset2.

Model	MAE	MSE	RMSE	R2
RFR	0.2068	0.1619	0.4024	0.85
LR	0.3127	0.2363	0.4861	0.78
DT	0.2683	0.3102	0.5569	0.71
GBR	0.2248	0.1677	0.4096	0.84
SVR	0.2273	0.1719	0.4146	0.84
KNN	0.2260	0.1859	0.4312	0.82
ANN	0.2088	0.1384	0.3728	0.87
1DCNN	0.2104	0.1485	0.3853	0.86

Source: Authors, (2025).

Table 7 shows a comparison with earlier studies e.g., Zhou et al. [13] and Zhao et al. [15], the developed models in this research—especially 1DCNN, ANN, RF and GB—performed competitively or outperformed previously reported R², which varies from 0.91 to 0.95. This indicates that our framework does not only equal but, in many cases, even surpasses the performance of the state-of-the-art models such as SDA-GA-ELM, PSO-LSTM, and Indicator-AML methods. Summary This discussion suggests that DL models such as 1DCNN and ensemble methods such as GBR and RFR are robust and efficient for the PV power prediction in diverse conditions. Their better results in the two datasets indicate their application prospect on practical solar forecasting systems. However, simple models like LR and DT are interpretable and fast but often not precise and flexible enough to achieve high-accuracy energy forecasting in complex environments. The proposed models thus offer a scalable and accurate means for contemporary PV power prediction work.

Table 7: Comparison with earlier studies.

Reference	Year	Technique	R^2
Yi Zhou et al. [13]	2020	SDA-GA-ELM	0.9225
		SDA-ELM	0.8940
		SDA-BPNN	0.7909
		SDA-SVM	0.9166
		ELM	0.8465
		BPNN	0.8172
		SVM	0.8971
		Persistence	0.7922
Wei Zhao et al. [15]	2021	Indicator-AML	0.948
		DT	0.853
		LR K-Neighbors	0.856
		Regression	0.699
		RF	0.932
		Ada-Boost	0.821
		GB	0.939
		Bagging	0.931
		PSO-LSTM	0.922
		PSO-BI-LSTM	0.923
		PSO-SVR	0.933
Proposed approach	2025	RFR	0.95
		LR	0.92
		DT	0.88
		GBR	0.95
		SVR	0.95
		KNN	0.94
		ANN	0.95
		1DCNN	0.95

Source: Authors, (2025).

V. CONCLUSIONS

This research assessed and compared performance of various ML and DL models i.e. RFR, LR, DT, GBR, SVR, KNN, ANN and 1DCNN for PV power output prediction based two real-life datasets of multiple locations. Our findings indicate that for all three types of advanced models, namely 1DCNN, ANN, RFR, and GBR, the predictive performance was significantly better than that for traditional models such as LR and DT. These sophisticated models exhibited the lower prediction errors (MSE, RMSE, MAE) and higher R^2 values to demonstrate better predictive accuracy and robustness. The 1DCNN and ANN models in particular, had robust generalizability between spatial and regional datasets, which was well-suited to the wide spectrum of operationalization. In the future, we give priority to DL models such as 1DCNN and ANN on PV forecasting. Ensemble methods, most notably RFR and GBR, are also good contenders. Hybrid methods, real-time data integration and explainable AI, cross regional validation are also suggested to make it more accurate, reliable, and transparent in solar energy forecasting.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Aythem Khairi Kareem, Ahmed Adil Nafea, Manar Thayer Mansour, Nasrin Nadher Jamil
Methodology: Aythem Khairi Kareem, Ahmed Adil Nafea, Manar Thayer Mansour, Nasrin Nadher Jamil.
Investigation: Aythem Khairi Kareem, Ahmed Adil Nafea, Manar Thayer Mansour, Nasrin Nadher Jamil.
Discussion of results: Aythem Khairi Kareem, Ahmed Adil Nafea, Manar Thayer Mansour, Nasrin Nadher Jamil.
Writing – Original Draft: Aythem Khairi Kareem, Ahmed Adil Nafea, Manar Thayer Mansour, Nasrin Nadher Jamil.
Writing – Review and Editing: Aythem Khairi Kareem, Ahmed Adil Nafea, Manar Thayer Mansour, Nasrin Nadher Jamil.
Resources: Aythem Khairi Kareem, Ahmed Adil Nafea, Manar Thayer Mansour, Nasrin Nadher Jamil.
Supervision: Aythem Khairi Kareem, Ahmed Adil Nafea, Manar Thayer Mansour, Nasrin Nadher Jamil.
Approval of the final text: Aythem Khairi Kareem, Ahmed Adil Nafea, Manar Thayer Mansour, Nasrin Nadher Jamil.

VII. REFERENCES

[1] B. Shboul et al., 'Energy and economic analysis of building integrated photovoltaic thermal system: Seasonal dynamic modeling assisted with machine learning-aided method and multi-objective genetic optimization', Alexandria Engineering Journal, vol. 94, pp. 131–148, May 2024, doi: 10.1016/j.aej.2024.03.049.

[2] M. F. Tahir, M. Z. Yousaf, A. Tzes, M. S. El Moursi, and T. H. M. El-Fouly, 'Enhanced solar photovoltaic power prediction using diverse machine learning algorithms with hyperparameter optimization', Renewable and Sustainable Energy Reviews, vol. 200, Aug. 2024, doi: 10.1016/j.rser.2024.114581.

[3] M. Nur-E-Alam et al., 'Machine learning-enhanced all-photovoltaic blended systems for energy-efficient sustainable buildings', Sustainable Energy Technologies and Assessments, vol. 62, Feb. 2024, doi: 10.1016/j.seta.2024.103636.

[4] S. S. Shijer, A. H. Jassim, L. A. Al-Haddad, and T. T. Abbas, 'Evaluating electrical power yield of photovoltaic solar cells with k-Nearest neighbors: A machine learning statistical analysis approach', e-Prime - Advances in
Electrical Engineering, Electronics and Energy, vol. 9, Sep. 2024, doi: 10.1016/j.prime.2024.100674.

[5] A. M. Shaban, M. S. Ibrahim, A. K. Kareem, N. N. Jamil, S. A. Aliesawi, and A. A. Nafea, 'Advanced fault detection in photovoltaic systems using artificial neural network', in IET Conference Proceedings CP906, IET, 2024, pp. 374–379.

[6] W. Tercha, S. A. Tadjer, F. Chekired, and L. Canale, 'Machine Learning-Based Forecasting of Temperature and Solar Irradiance for Photovoltaic Systems', Energies (Basel), vol. 17, no. 5, Mar. 2024, doi: 10.3390/en17051124.

[7] N. Merabti, D. T. Cotfas, and P. A. Cotfas, 'Enhancing Photovoltaic Efficiency: Optimal Tilt Angle Estimation Using Advanced Machine Learning Techniques', in 2024 International Conference on Applied and Theoretical Electricity, ICATE 2024 - Proceedings, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/ICATE62934.2024.10749082.

[8] E. E. Ogar, S. Chaitusaney, and W. Benjapolakul, 'Performance Assessment of Supervised Machine Learning-Based Approaches for Photovoltaic Diagnosis', in 2024 21st International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, ECTI-CON 2024, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/ECTI-CON60892.2024.10594847.

[9] S. T. Asiedu, F. K. A. Nyarko, S. Boahen, F. B. Effah, and B. A. Asaaga, 'Machine learning forecasting of solar PV production using single and hybrid models over different time horizons', Heliyon, vol. 10, no. 7, Apr. 2024, doi: 10.1016/j.heliyon.2024.e28898.

[10] S. Qaadan and A. Alshare, 'Forecasting Solar Photovoltaic Power Output in the German Jordanian University in Amman Using Artificial Intelligence and Machine Learning Algorithms', in 2022 10th International Conference on Control, Mechatronics and Automation, ICCMA 2022, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 243–248. doi: 10.1109/ICCMA56665.2022.10011464.

[11] M. K. Shakya, V. N. Pande, R. S. Kulkarni, and S. Kakade, 'Estimation of Solar PV Power Plant Output Using Machine Learning Algorithms', in 2023 International Conference on Digital Applications, Transformation and Economy, ICDATE 2023, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ICDATE58146.2023.10248843.

[12] X. Li, 'Research on New Energy Power Generation Power Prediction Method Based on Machine Learning', in 2022 4th International Conference on Communications, Information System and Computer Engineering, CISCE 2022, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 331–334. doi: 10.1109/CISCE55963.2022.9851053.

[13] Y. Zhou, N. Zhou, L. Gong, and M. Jiang, 'Prediction of photovoltaic power output based on similar day analysis, genetic algorithm and extreme learning machine', Energy, vol. 204, p. 117894, 2020.

[14] B. Carrera and K. Kim, 'Comparison analysis of machine learning techniques for photovoltaic prediction using weather sensor data', Sensors, vol. 20, no. 11, p. 3129, 2020.

[15] W. Zhao et al., 'A point prediction method based automatic machine learning for day-ahead power output of multi-region photovoltaic plants', Energy, vol. 223, p. 120026, 2021.

[16] K. Shivam, J.-C. Tzou, and S.-C. Wu, 'A multi-objective predictive energy management strategy for residential grid-connected PV-battery hybrid systems based on machine learning technique', Energy Convers Manag, vol. 237, p. 114103, 2021.

[17] S. Chahboun and M. Maaroufi, 'Performance comparison of k-nearest neighbor, random forest, and multiple linear regression to predict photovoltaic panels' power output', in Advances on Smart and Soft Computing: Proceedings of ICACIn 2021, Springer, 2021, pp. 301–311.

[18] R. Saeedi, S. K. Sadanandan, A. K. Srivastava, K. L. Davies, and A. H. Gebremedhin, 'An adaptive machine learning framework for behind-the-meter load/PV disaggregation', IEEE Trans Industr Inform, vol. 17, no. 10, pp. 7060–7069, 2021.

[19] K. Mahmud, S. Azam, A. Karim, S. Zobaed, B. Shanmugam, and D. Mathur, 'Machine learning based PV power generation forecasting in alice springs', IEEE access, vol. 9, pp. 46117–46128, 2021.

[20] Z. He, M. Yang, L. Wang, E. Bao, and H. Zhang, 'Concentrated photovoltaic thermoelectric hybrid system: An experimental and machine learning study', Engineered Science, vol. 15, 2021, doi: 10.30919/es8d440.

[21] H. Mubarak, A. Abdellatif, S. Ahmad, A. Hammoudeh, S. Mekhilef, and H. Mokhlis, 'Prediction of Solar Photovoltaic Energy Output Based on Thin-Film Technology Utilizing Various Machine Learning Techniques', in 2022 IEEE Global Conference on Computing, Power and Communication Technologies, GlobConPT 2022, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/GlobConPT57482.2022.9938166.

[22] J. Zhang and J. Gao, 'Photovoltaic power prediction based on improved moth-flame optimization algorithm and extreme learning machine', in 2022 7th International Conference on Intelligent Computing and Signal Processing, ICSP 2022, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 785–788. doi: 10.1109/ICSP54964.2022.9778360.

[23] 'Solar energy production dataset | AI-on-Demand'. Accessed: Jul. 13, 2025. [Online]. Available: https://www.ai4europe.eu/research/ai-catalog/solar-energy-production-dataset

[24] J. Kaiser, 'Dealing with Missing Values in Data'.

[25] A. Saboor, M. Usman, S. Ali, A. Samad, M. F. Abrar, and N. Ullah, 'A Method for Improving Prediction of Human Heart Disease Using Machine Learning Algorithms', Mobile Information Systems, vol. 2022, 2022, doi: 10.1155/2022/1410169.

[26] A. K. Kareem, A. M. Shaban, A. A. Nafea, M. Aljanabi, S. A. S. Aliesawi, and M. Mal-Ani, 'Detecting Routing Protocol Low Power and Lossy Network Attacks Using Machine Learning Techniques', in 2024 21st International Multi-
Conference on Systems, Signals & Devices (SSD), IEEE, 2024, pp. 57–62.

[27] H. Chen, N. Wang, X. Du, K. Mei, Y. Zhou, and G. Cai, 'Classification Prediction of Breast Cancer Based on Machine Learning', Comput Intell Neurosci, vol. 2023, no. 1, Jan. 2023, doi: 10.1155/2023/6530719.

[28] S. Antad et al., 'Stock Price Prediction Website Using Linear Regression - A Machine Learning Algorithm', ITM Web of Conferences, vol. 56, p. 05016, 2023, doi: 10.1051/itmconf/20235605016.

[29] A. A. Nafea, M. S. Ibrahim, A. A. Mukhlif, M. M. AL-Ani, and N. Omar, 'An ensemble model for detection of adverse drug reactions', ARO-The Scientific Journal of Koya University, vol. 12, no. 1, pp. 41–47, 2024.

[30] S. N. H. Bukhari, J. Webber, and A. Mehbodniya, 'Decision tree based ensemble machine learning model for the prediction of Zika virus T-cell epitopes as potential vaccine candidates', Sci Rep, vol. 12, no. 1, Dec. 2022, doi: 10.1038/s41598-022-11731-6.

[31] A. M. Austin et al., 'Using a cohort study of diabetes and peripheral artery disease to compare logistic regression and machine learning via random forest modeling', BMC Med Res Methodol, vol. 22, no. 1, Dec. 2022, doi: 10.1186/s12874-022-01774-8.

[32] S. Kwak et al., 'Machine learning prediction of the mechanical properties of γ-TiAl alloys produced using random forest regression model', Journal of Materials Research and Technology, vol. 18, pp. 520–530, May 2022, doi: 10.1016/j.jmrt.2022.02.108.

[33] R. Rodríguez-Pérez and J. Bajorath, 'Evolution of Support Vector Machine and Regression Modeling in Chemoinformatics and Drug Discovery', J Comput Aided Mol Des, vol. 36, no. 5, pp. 355–362, May 2022, doi: 10.1007/s10822-022-00442-9.

[34] A. K. Kareem and K. M. Ali Alheeti, 'Multimodal Approach for Fall Detection Based on Support Vector Machine', in AIP Conference Proceedings, 2022. doi: 10.1063/5.0115534.

[35] A. M. Shaban, M. AL-Mahdawi, M. S. Ibrahim, A. K. Kareem, S. A. Aliesawi, and A. A. Nafea, 'Predicting air quality index using an ensemble model based on Internet of Things observations', in IET Conference Proceedings CP906, IET, 2024, pp. 522–526.

[36] R. Nyirandayisabye, H. Li, Q. Dong, T. Hakuzweyezu, and F. Nkinahamira, 'Automatic pavement damage predictions using various machine learning algorithms: Evaluation and comparison', Results in Engineering, vol. 16, Dec. 2022, doi: 10.1016/j.rineng.2022.100657.

[37] M. Mailagaha Kumbure and P. Luukka, 'A generalized fuzzy k-nearest neighbor regression model based on Minkowski distance', Granular Computing, vol. 7, no. 3, pp. 657–671, Jul. 2022, doi: 10.1007/s41066-021-00288-w.

[38] D. K. Jana, P. Bhunia, S. Das Adhikary, and B. Bej, 'Optimization of Effluents Using Artificial Neural Network and Support Vector Regression in Detergent Industrial Wastewater Treatment', Cleaner Chemical Engineering, vol. 3, p. 100039, Sep. 2022, doi: 10.1016/j.clce.2022.100039.

[39] T. Guo, F. Xu, J. Ma, and F. Ge, 'Component Prediction of Antai Pills Based on One-Dimensional Convolutional Neural Network and Near-Infrared Spectroscopy', Journal of Spectroscopy, vol. 2022, 2022, doi: 10.1155/2022/6875022.

[40] A. A. Mahamat et al., 'Decision Tree Regression vs. Gradient Boosting Regressor Models for the Prediction of Hygroscopic Properties of Borassus Fruit Fiber', Applied Sciences (Switzerland), vol. 14, no. 17, Sep. 2024, doi: 10.3390/app14177540.