International Journal of Data Informatics and Intelligent Computing (IJDIIC)
Vol. 5, No. 3, September 2026, pp. 50~66
DOI: 10.59461/ijdiic.v5i3.305

Optimized Machine Learning for Solar Energy Generation Forecasting: A Feature Selection and Hyperparameter Optimization Approach

Farwa Nawaz[1], Faria Karamat[1]
[1]Faculty of Computer Science, Riphah International University, Islamabad, Pakistan

Article Info

Article history:

Received July 13, 2026
Revised August 28, 2026
Accepted September 06, 2026

Keywords:

Solar Energy Forecasting
Support Vector Regression
Renewable Energy
Machine Learning
Feature Selection
Smart Grid

ABSTRACT

Solar energy is among the most promising renewable resources for developing sustainable energy systems; however, the stochastic and weather-sensitive nature of Solar Energy Generation (SEG) poses challenges for forecasting. The current study designs a machine learning framework for SEG forecasting by using feature selection methods in three stages. In the first stage, three feature selection methods, namely Pearson Correlation (PC), Recursive Feature Elimination (RFE), and Random Forest Feature Importance (RFFI), are compared systematically, each method selecting four predictors among ten weather and system variables, while avoiding the use of electrical telemetry variables to prevent data leakage. In the second stage, four regression methods, including Linear Regression, Random Forest (RF), Multilayer Perceptron (MLP), and Optimized Radial Basis Function Support Vector Regression (RBF-SVR), are tested using each feature set, and the results are verified in five independently repeated experiments. The feature selection method has a much bigger impact on accuracy than the regression method: RFFI-selected features allowed attaining a mean R^2 greater than 0.85 for three regression methods, whereas PC and RFE constrained all models to achieve R^2 less than 0.60. Among all models for RFFI features, Random Forest proved to have the best accuracy (mean $R^2 = 0.8960 \pm 0.0008$), with MLP being second ($R^2 = 0.8893 \pm 0.0022$), followed by the Optimized RBF-SVR model ($R^2 = 0.8568 \pm 0.0072$); the Optimized RBF-SVR model took an average training time of 21.25 ± 1.12 seconds compared to RF's 35.78 ± 2.42 seconds. This shows the importance of selecting leakage-controlled features systematically in the process of forecasting SEG, and how a small number of 4 features is sufficient for achieving statistical stability.

Corresponding Author: Farwa Nawaz (e-mail: farwa.nawaz@riphah.edu.pk)

1. INTRODUCTION

Renewable energy is crucial for achieving sustainable and low-carbon energy systems worldwide [1]. Solar photovoltaics continue to be one of the fastest-growing renewable resources [2], with an expected 507 GW of renewable energy generated in 2023, a 50% increase from 2022 [3]. Recent developments have expedited the integration of solar systems into contemporary power grids, and their growing deployment aids international efforts toward carbon neutrality and energy security [4]. Clean, plentiful, and ecologically friendly electricity may be produced with solar energy [5][6]. Nevertheless, weather fluctuations continue to have a significant impact on solar energy production (SEG) [7]. Photovoltaic output is directly impacted by variations in temperature, humidity, wind speed, cloud cover, and solar irradiation. These variations also complicate grid management and lower forecasting accuracy [8]. Thus, reliable energy scheduling and smart grid functioning depend on precise SEG predictions [9][10].

For SEG prediction, a number of forecasting methods have been developed. Initially, statistical models such as autoregressive, moving average, and time-series forecasting techniques were used due to their ease of use and low processing requirements [11][12]. However, extremely nonlinear correlations found in meteorological data are difficult to model using statistical methods. Later, heuristic and optimization-based approaches were developed to enhance forecasting performance [13][14]. However, these approaches frequently need significant processing resources and parameter tuning, which limits their usefulness for real-time forecasting systems.

Recently, machine learning (ML) has become a viable substitute for SEG forecasting, successfully capturing nonlinear correlations between solar power generation and meteorological variables [4]. Promising forecasting performance has been shown by a number of regression algorithms, such as Random Forest, Support Vector Regression, and Neural Networks; deep learning models have further increased prediction accuracy by utilizing intricate hierarchical feature representations [15][16]. Despite these developments, the quality of the input features has a significant impact on prediction accuracy. Selecting informative features continues to be a crucial difficulty for reliable SEG forecasting since redundant and irrelevant variables frequently increase computational complexity and degrade forecasting performance [17].

To achieve high prediction performance, model generalization, and computational efficiency, feature selection (FS) techniques reduce data dimensionality without losing vital predictive data [18][19]. Popular FS methods include Random Forest Feature Importance, Pearson Correlation, and Recursive Feature Elimination [17][20]. Nevertheless, few studies have thoroughly examined different methods for SEG forecasting under the same circumstances, and earlier research frequently prioritizes prediction accuracy at the expense of computational efficiency [21]. Seldom are advanced evaluation criteria like Huber Loss, Explained Variance Score, and Median Absolute Error taken into account. Thus, an effective forecasting framework that combines reliable machine learning models with ideal feature selection is still required.

This paper suggests an enhanced machine learning approach for SEG forecasting in order to overcome these constraints. The most instructive meteorological features are first determined using three feature selection methods. Second, the same experimental conditions are used to assess four machine learning regression models. Third, to increase forecasting precision and computational efficiency, an improved Radial Basis Function Support Vector Regression (RBF-SVR) model is created. The publicly accessible Desert Knowledge Australia Solar Centre dataset is used in extensive tests, and the experimental findings show better forecasting performance than current machine learning techniques.

The following is a summary of this study's primary contributions:

- Three feature selection methods are integrated with machine learning regression models to create a comprehensive SEG forecasting system.
- For the best feature selection under the same experimental settings, Pearson Correlation, Recursive Feature Elimination, and Random Forest Feature Importance are compared.
- The proposed optimized model of RBF-SVR with optimal hyperparameters is achieved, which provides high prediction accuracy but has significantly reduced computational complexity compared to the best baseline model.
- Experiments, with leakage prevention measures, are performed for the comparison of linear regression, random forest, MLP, and the proposed optimized RBF-SVR, using three feature selection methods, verified by conducting the experiment five times independently with standard deviation provided.
- The proposed method is tested using seven regression performance measures – MSE, RMSE, MAE, MedAE, EVS, Huber loss, and R^2, proving that the feature selection method impacts the forecasting accuracy more than model selection does.

The remainder of this paper is organized as follows. Section 2 reviews recent SEG forecasting approaches. Section 3 presents the proposed forecasting framework, dataset, preprocessing, feature selection, machine learning models, experimental setup, and evaluation metrics. Section 4 presents the experimental results and comparative analysis. Finally, Section 5 concludes the paper and outlines future research directions.

2. LITERATURE REVIEW

2.1. Machine Learning and Optimization-Based SEG Forecasting

SEG forecasting has attracted significant research efforts owing to the rise of PV systems and smart grid installations. The application of various machine learning techniques such as Support Vector Regression (SVR), Random Forest (RF), Artificial Neural Networks (ANN), Gaussian Process Regression (GPR), Decision Tree (DT), and Gradient Boosting for forecasting has shown great success in accurately modeling the nonlinear dependency between meteorological features and power production [22][23][17][4]. Particularly, SVR is still one of the most popular forecasting methods due to its good generalization ability, whereas RF and ANN yield high forecasting accuracy regardless of climate [17].

Machine learning models and optimization algorithms have been combined to increase prediction accuracy. To optimize model hyperparameters and lower forecasting errors, methods like Particle Swarm Optimization (PSO) and a Hybrid Improved Multi-Verse Optimizer (HIMVO) [22], Genetic Algorithms (GA/GASVM) [23], Ant Colony Optimization (ACO) [24], the Whale Optimization Algorithm (WOA) [25], and Bayesian Optimization [21] have been used. Despite the fact that these hybrid approaches increase prediction accuracy, most research ignores the impact of feature selection on forecasting efficiency in favor of parameter tuning.

2.2. Deep Learning and Feature Selection

Deep learning models, comprising Long Short-Term Memory (LSTM) networks [26], Gated Recurrent Units (GRU) [27], Convolutional Neural Networks (CNN) [28], and Transformer architectures [29], have recently demonstrated excellent performance in solar forecasting by effectively learning temporal and nonlinear patterns from large datasets. However, these models' application in computationally constrained situations is limited since they often demand large amounts of training data, higher computational resources, and longer training times [30].

An efficient preprocessing method for increasing forecasting efficiency is feature selection. Data dimensionality is decreased while model generalization is enhanced by techniques including Pearson Correlation (PC) [17], Recursive Feature Elimination (RFE) [31], Random Forest Feature Importance (RFFI) [19], Mutual Information [32], and wrapper-based approaches [20]. However, current research usually assesses a single feature selection method or combines it with a small number of forecasting models, offering little information on the relative efficacy of several feature selection approaches.

2.3. Research Gap and Motivation

SEG forecasting has come a long way, but there are still a number of obstacles. The majority of current research focuses on optimizing models or hyperparameters while paying little attention to systematic feature selection. There are also few comparative analyses of various feature selection methods under the same experimental conditions. Additionally, traditional criteria are frequently used to evaluate predictive ability, while advanced regression measures are rarely taken into account. These constraints drive the creation of computationally effective forecasting frameworks that use small feature subsets to retain high predicted accuracy.

This study suggests a feature-selection-driven forecasting paradigm that compares Random Forest Feature Importance (RFFI), Recursive Feature Elimination (RFE), and Pearson Correlation (PC) in order to address these issues. An optimal Radial Basis Function Support Vector Regression (RBF-SVR) model is created after four machine learning regression models are trained with similar feature subsets. Both traditional and sophisticated regression performance indicators are used to thoroughly assess the suggested architecture. The current work is positioned in relation to representative recent investigations, which are summarized in Table 1.

Table 1. Summary of recent SEG forecasting studies

Reference	Method	Feature Selection	Limitation
[33]	RF (ensemble component)	No	Large feature set
[34]	SVR	No	Hyperparameter sensitive
[22]	PSO-SVR / HIMVO-SVM	No	No feature selection
[24]	ACO-SVR	No	Computationally expensive
[26]	LSTM (WPD-LSTM)	No	Long training time
[27]	GRU (TCN-ECANet-GRU)	No	Large data requirement
[17]	Pearson Correlation + RF	Pearson Correlation	Single feature selection technique
[20]	Wrapper + SVR	Wrapper	Higher feature dimensionality

3. METHOD

The general structure of the suggested Solar Energy Generation (SEG) forecasting system is shown in Figure 1. Five steps make up the framework: (1) data collection and preprocessing; (2) feature selection; (3) machine learning model building; (4) model optimization; and (5) performance evaluation. To enhance data quality, raw meteorological and electricity-generating data are first cleaned and preprocessed. The most informative predictors of solar energy generation are then found using three feature selection strategies. Four popular machine learning regression models are then trained using the chosen feature subsets. Lastly, a Radial Basis Function (RBF) Support Vector Regression (SVR) framework is used to further refine the top-performing model, and different regression metrics are used to assess the model's performance.

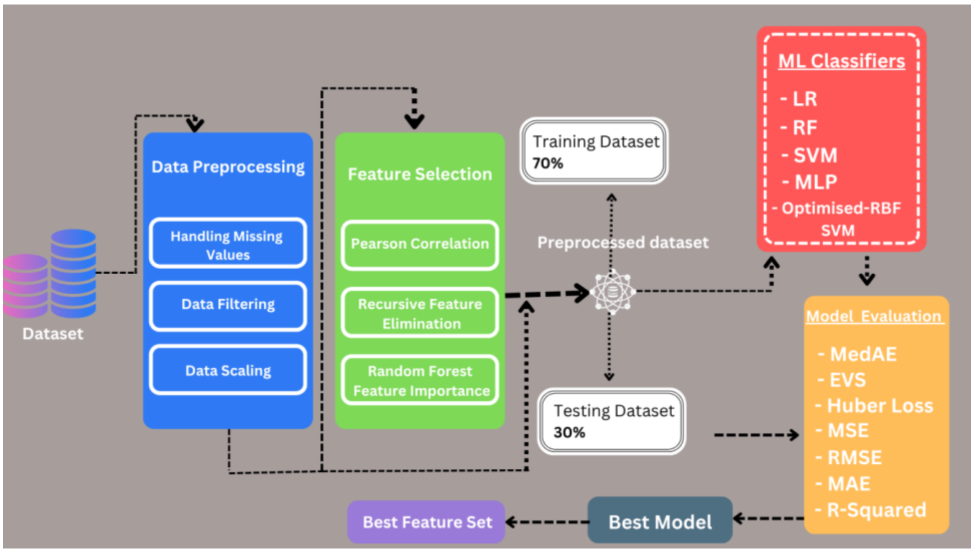

Figure 1. The proposed framework for SEG forecasting

The suggested framework focuses on identifying a small subset of useful meteorological characteristics, in contrast to traditional forecasting methods that depend on the entire feature set. This method preserves good predictive accuracy while lowering computational complexity. To determine the best forecasting framework, a thorough comparison of feature selection methods and machine learning models is carried out under identical experimental conditions.

3.1. Dataset Description

The experiments were performed based on the photovoltaic data set from the Desert Knowledge Australia Solar Centre (DKASC), which is an open-source data set available for research use. This data set includes actual measurements from Site 28 at the DKASC Alice Springs site, which is a First Solar CdTe (thin-film) solar power plant. Table 2 presents the specifications of Site 28.

Table 2. Details of the solar panel installation

Parameter	Value
Manufacturer	First Solar
Array Rating	5.6 kW
PV Technology	CdTe (thin film)
Array Structure	Fixed: Ground Mount
Panel Rating	87.5 W
Number of Panels	64
Panel Type	First Solar FS-387
Array Area	46.08 m²
Inverter Size / Type	6 kW, SMA SMC 6000
Installation Completed	February 06 2013
Array Tilt / Azimuth	Tilt = 20°, Azimuth = 0° (Solar North)

The dataset includes 13 predictor variables, a target variable (Active Power), and a timestamp field, recorded at an interval of five minutes over several years, resulting in 1,281,034 total observations after cleaning. However, two out of the 13 predictor variables recorded (Current Phase Average and Active Energy Delivered/Received) have been excluded from the candidate feature set because they represent electrical equivalents of the target variable and thus do not qualify as predictor variables, nor will they be available at the time of forecasting. The ten remaining candidates are listed in Table 3 below.

Table 3. Features in the dataset

No	Feature	Description
1	Active Power	Solar energy being generated at a given time (target variable)
2	Active Energy Delivered/Received	Total energy delivered to or received from the grid per unit time [Excluded from candidates — leakage risk]
3	Current Phase Average	Average current flowing through a particular phase [Excluded from candidates — leakage risk]
4	Active Power Performance Ratio	Measure of efficiency of the solar power system
5	Wind Speed	The speed of wind per unit time
6	Weather Temperature (Celsius)	The ambient temperature per unit time
7	Weather Relative Humidity	The amount of moisture in the air
8	Global Horizontal Radiation (GHI)	Total radiation received by a horizontal surface
9	Diffuse Horizontal Radiation (DHI)	Diffuse portion of solar radiation on a horizontal surface
10	Wind Direction	The direction of wind
11	Weather Daily Rainfall	The amount of rainfall at the solar system site
12	Radiation Global Tilted	Total radiation received by a tilted surface
13	Radiation Diffuse Tilted	Diffuse portion of radiation on a tilted surface

3.2. Data Preprocessing

Reliable forecasting depends on high-quality input data. Three preprocessing steps were performed before model development.

Handling Missing Data: The missing data, which occurred due to disruptions in sensors and/or communication, were filled through median imputation calculated for both training and testing datasets separately so that no information was leaked from the testing dataset to the training one.

Data Filtering: Occasionally, anomalous values in sensor measurements arise from hardware constraints or environmental disruptions. Prior to model training, filtering techniques were used to eliminate noisy and inconsistent readings.

Feature Normalization: Meteorological variables possess different numerical ranges, which may negatively influence machine learning algorithms. Min–Max normalization was applied to scale each feature into the interval [0,1], accelerating model convergence and preventing variables with larger magnitudes from dominating the learning process, as defined in Equation (1):

$$x\_scaled = (x - \min(x)) / (\max(x) - \min(x)) \tag{1}$$

The dataset is sorted chronologically, and in the case of large-scale model comparisons presented in Section 4, a random sampling of 60,000 rows is used, where a 70:30 split ratio is used between train and test sets (42,000 training and 18,000 testing samples) in order to maintain computability and validity of the results on such large-scale datasets. This method is being used for all of the reported experiments.

3.3. Feature Selection

The most informative meteorological variables for SEG forecasting were determined through feature selection. The four top-ranked characteristics from the initial set of thirteen were chosen by each of the three complementary methods that were examined, each of which assessed feature significance from a different angle.

3.3.1. Pearson Correlation (PC)

Pearson Correlation (PC) measures the strength of the linear relationship between each predictor variable and the target variable, as given in Equation (2):

$$r = \Sigma(x_i - \bar{x})(y_i - \bar{y}) / [\sqrt{\Sigma(x_i - \bar{x})^2} \cdot \sqrt{\Sigma(y_i - \bar{y})^2}] \tag{2}$$

Table 4 reports the correlation coefficients obtained for each candidate feature against the target (Active Power).

Table 4. Correlation of each feature with the target feature (Active Power)

No	Feature	Correlation (r)
1	Global_Horizontal_Radiation	0.6947
2	Radiation_Global_Tilted	0.6092
3	Diffuse_Horizontal_Radiation	0.3945
4	Radiation_Diffuse_Tilted	0.3460
5	Weather_Temperature_Celsius	0.3240
6	Weather_Relative_Humidity	0.3194
7	Wind_Speed	0.2769
8	Performance_Ratio	0.0988
9	Wind_Direction	0.0669
10	Weather_Daily_Rainfall	0.0392

The four features with the highest absolute correlation, Global Horizontal Radiation, Diffuse Horizontal Radiation, Weather Temperature, and Wind Speed, were selected for model development.

3.3.2. Recursive Feature Elimination (RFE)

RFE is a wrapper-based technique that iteratively removes the least important predictor according to the performance of a regression estimator, continuing until the required number of features remains, as shown in Table 5. Unlike correlation analysis, RFE evaluates feature importance during model training and therefore captures interactions among predictors.

Table 5. Feature ranking using Recursive Feature Elimination

No	Feature	Rank
1	Wind_Speed	1
2	Weather_Temperature_Celsius	1
3	Weather_Daily_Rainfall	1
4	Global_Horizontal_Radiation	1
5	Weather_Relative_Humidity	2
6	Radiation_Diffuse_Tilted	3
7	Diffuse_Horizontal_Radiation	4
8	Radiation_Global_Tilted	5
9	Wind_Direction	6
10	Performance_Ratio	7

A lower rank indicates greater importance. The four top-ranked features — Wind Speed, Global Horizontal Radiation, Weather Temperature, and Weather Relative Humidity — were selected for model development.

3.3.3. Random Forest Feature Importance (RFFI)

The relevance of predictors is evaluated in terms of the decrease in impurity in RFFI based on the mean decrease in impurity (MDI), defined by Equation (3):

$$\mathrm{MDI}(X_j) = (1/T)\ \Sigma_t\ \Sigma_n \in \mathrm{Nodes}(t)\ \Delta I(n) \cdot \mathbb{1}\{\mathrm{split\_feature}(n) = X_j\} \quad (3)$$

Table 6. Feature importance scores using Random Forest Feature Importance

No	Feature	Importance
1	Global_Horizontal_Radiation	0.4482

2	Performance_Ratio	0.2225
3	Wind_Direction	0.1628
4	Radiation_Global_Tilted	0.0875
5	Radiation_Diffuse_Tilted	0.0229
6	Wind_Speed	0.0178
7	Weather_Temperature_Celsius	0.0146
8	Weather_Relative_Humidity	0.0119
9	Diffuse_Horizontal_Radiation	0.0113
10	Weather_Daily_Rainfall	0.0005

The four features with the highest MDI, Global Horizontal Radiation, Performance Ratio, Wind Speed, and Weather Temperature, were selected for model development as shown in Table 6. Compared with the statistical techniques, RFFI effectively captures nonlinear relationships and complex feature interactions.

3.3.4. Comparative Summary

The three feature selection techniques generated different forecasting subsets because each evaluates relevance using a different criterion: RFFI uses ensemble learning to estimate importance, RFE assesses predictor contribution during model formation, and PC stresses linear dependency. All three feature subsets were assessed using similar machine learning models under identical experimental conditions, and the results were compared using MSE, RMSE, MAE, MAPE, MedAE, EVS, Huber Loss, and R2, rather than assuming that one approach is always better (Section 4).

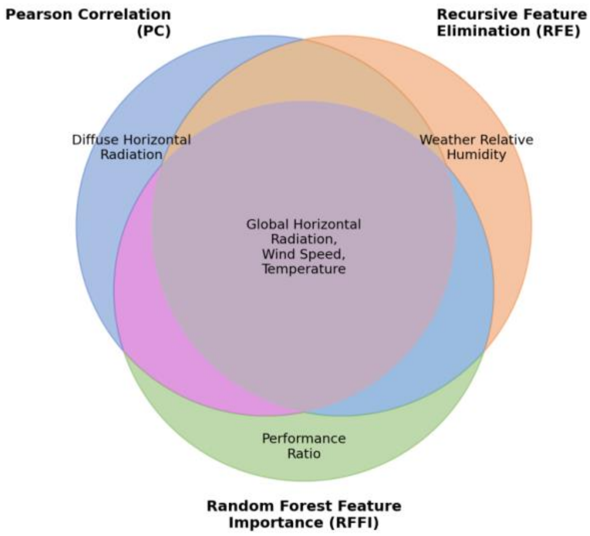

Figure 2. Overlap of top 4 selected features across PC, RFE, and RFFI

Figure 2 provides an illustration of the intersection among the three subsets of features described in Tables 4 to 6. Global Horizontal Radiation, Wind Speed, and Weather Temperature are chosen by the three methods, revealing a high level of agreement that such features are the main factors influencing the Active Power generation. Each technique has chosen one feature that the other two techniques have not chosen: Diffuse Horizontal Radiation for PC, Weather Relative Humidity for RFE, and Performance Ratio for RFFI. In doing so, it is clear that the three methods have reached an agreement about the set of important predictors but only differ on the choice of the secondary predictor, which can be explained by the superior forecasting performance achieved by RFFI in Section 4.3.

3.4. Machine Learning Models

After feature selection, each subset of features was assessed using four popular regression algorithms, which include Linear Regression (LR), Random Forest (RF), Support Vector Regression (SVR), and Multilayer Perceptron (MLP). The four baseline models are briefly introduced below; their performances are compared in Section 4, where Random Forest is found to be the best among them. The Optimized RBF-SVR algorithm proposed in the paper is described in detail in Section 3.5.

3.4.1. Linear Regression (LR)

In LR, the model assumes that the predictor variables are linked to the PV power generation through a linear model. The parameters of the model are found through least squares estimation, which minimizes the sum of squared residuals.

3.4.2. Random Forest (RF)

RF is an example of an ensemble method where decision trees are combined through randomly sampling observations and predictors in such a way that the final prediction will be a result of the averaging of all these trees. This helps RF to model nonlinear dependencies without overfitting, since, as can be seen from Section 4, it performs better than other baselines in terms of accuracy.

3.4.3. Support Vector Regression (SVR)

The SVR model generalizes Support Vector Machines to regression through the generation of a function that will stay inside an ε-insensitive boundary of the training objective values, in such a way that it reduces the complexity of the model. In the current model, we apply a Radial Basis Function (RBF) kernel; the fully optimized model can be found in Section 3.5.

3.4.4. Multilayer Perceptron (MLP)

MLP is a feedforward neural network that consists of one or multiple layers, which are trained using backpropagation to learn the nonlinear relationship between inputs and outputs. MLP needs proper parameter tuning to prevent overfitting and usually has longer training times than LR and RF.

3.5. Proposed Optimized RBF-SVR

While traditional SVR has shown competitive performance in solar energy forecasting, feature quality and hyperparameter choice have a significant impact on prediction accuracy. This section combines feature selection, polynomial feature transformation, feature normalization, and randomized hyperparameter optimization into a single forecasting pipeline.

The reason why RBF-SVR was preferred over deep architectures (such as LSTM, GRU, Transformer-based models; see Section 2.2) is two-fold. Firstly, kernel-based models such as SVR usually need considerably fewer training samples to perform well compared to deep architectures. This is due to the fact that this research evaluates models on a small 4-5 feature set for each feature selection algorithm, rather than large multivariate sequences as input. Secondly, training deep architectures usually takes considerably more computational resources and time to be fine-tuned. The training times for the Optimized RBF-SVR presented in Section 4.4 demonstrate that, on average, it needs only 21.25 seconds to complete the training, while in the literature (see Section 2.2) LSTM- and Transformer-based solar forecasting models require much more time for training. The speed of training is explained by the fact that the objective of this research is resource-efficient forecasting, not absolute accuracy regardless of the computational time needed for that.

3.5.1. Radial Basis Function Kernel

The RBF kernel was selected for its ability to model the highly nonlinear relationship between meteorological variables and photovoltaic power generation, as given in Equation (4). The resulting SVR decision function is given in Equation (5):

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2) \tag{4}$$

$$f(x) = \Sigma_i (\alpha_i - \alpha_i *) K(x_i, x) + b \tag{5}$$

where γ controls the kernel width; its appropriate selection balances model flexibility and generalization performance.

3.5.2. Polynomial Feature Transformation

In order to enhance the nonlinear relationship between the chosen meteorological variables, the polynomial features transformation step was used prior to regression modeling according to Equation (6):

$$\Phi(x) = \{1, x, x^2, \dots, x\hat{}\delta\} \tag{6}$$

3.5.3. Hyperparameter Optimization

The SVR prediction power depends mainly on the three hyperparameters: the regularization parameter C, kernel gamma (γ), and insensitive loss ε. Instead of selecting them manually, our framework uses RandomizedSearchCV with an extensive search space P = {C, γ, ε}. An SVR model is trained and evaluated for each combination of hyperparameters, and the one giving the best result is kept. On the selected feature set with the RFFI method for our proposed model, the search resulted in C = 0.4642, γ = 0.0492, and ε = 0.1041 with polynomial feature transformation with degree 2 before using the RBF kernel (Section 3.5.2).

3.5.4. Optimization Algorithm

Algorithm 1: Optimized Radial Basis Function Support Vector Regression (RBF-SVR)

Input:
  Dataset D = {d1, d2, ..., dn}
  Feature set F = {f1, f2, f3, f4}
  Target variable y
  SVR hyperparameter space P = {p1, p2, ..., pk}
  Training/testing split ratio alpha
  Polynomial degree delta
  Evaluation metrics M = {MSE, RMSE, MAE, MedAE, EVS, Huber Loss, R^2}
  Number of search iterations T

Training Process:
1. Load dataset D containing features and target variable.
2. Select the four most relevant features F (Section 3.3).
3. Split D into training set D_train and testing set D_test based on ratio alpha.
4. Handle missing data using median imputation, fit on D_train only and apply to both D_train and D_test.
5. Apply polynomial feature transformation of degree delta to F.
6. Standardize features F using z-score standardization (fit on D_train only) to produce F_hat.
7. Generate T random hyperparameter combinations from P.
8. For each combination p_t, train SVR(p_t) on D_train with features F_hat_train and target y_train.
9. Predict y_hat_test on the testing set using the trained SVR(p_t).
10. Evaluate each SVR(p_t) using metric set M.
11. Select the best model SVR* with the optimal hyperparameter combination p* based on M.

Testing Process:
  Input the testing data F_hat_test into SVR* to obtain the final prediction y_hat* = SVR*(F_hat_test).

Output:
  The final predicted target y_hat* for the testing data.

3.5.5. Computational Complexity

Reducing the number of predictor variables from thirteen to four substantially decreases computational complexity and accelerates model training. Assuming N training samples and d predictor variables, the computational complexity of kernel construction is approximately $O(N^2d)$. By reducing d from 13 to 4, the proposed framework significantly lowers computational cost while preserving forecasting performance, as confirmed experimentally in Section 4.4.

3.6. Experimental Setup

The proposed forecasting framework was implemented in Python 3.11 using the Google Colab environment, with additional validation performed in a Kaggle Notebook. Data preprocessing, feature selection, machine learning implementation, and performance evaluation were carried out using the Scikit-learn, NumPy, Pandas, and Matplotlib libraries. Hyperparameter optimization for the proposed RBF-SVR model was performed using RandomizedSearchCV on the training data only. Table 7 summarizes the experimental configuration.

Table 7. Experimental settings

Parameter	Value
Programming Language	Python 3.11
Environment	Google Colab, Kaggle Notebook
Libraries	Scikit-learn, NumPy, Pandas, Matplotlib
Dataset	DKASC (Site 28, Alice Springs)
Sampling Interval	5 minutes
Study Period	Jan–Dec 2023
Total Features	13
Selected Features	4
Feature Selection Techniques	PC, RFE, RFFI
Baseline Models	LR, RF, SVR, MLP
Proposed Model	Optimized RBF-SVR
Train/Test Split	80:20 (chronological)
Hyperparameter Optimization	Randomized Search

3.7. Evaluation Metrics

Model performance was assessed using eight complementary regression metrics: MSE, RMSE, MAE, MAPE, MedAE, EVS, Huber Loss, and R^2. Error-based metrics quantify prediction accuracy, whereas EVS and R^2 evaluate the goodness-of-fit between predicted and observed photovoltaic power output. Their mathematical definitions are given below.

$$\mathrm{MSE} = (1/n)\,\Sigma(y_i - \hat{y}_i)^2 \tag{7}$$

$$\mathrm{RMSE} = \sqrt{[(1/n)\,\Sigma(y_i - \hat{y}_i)^2]} \tag{8}$$

$$\mathrm{MAE} = (1/n)\,\Sigma|y_i - \hat{y}_i| \tag{9}$$

$$\mathrm{MAPE} = (1/n)\,\Sigma|(y_i - \hat{y}_i)/y_i| \times 100 \tag{10}$$

$$\mathrm{MedAE} = \mathrm{median}(|y_1 - \hat{y}_1|, |y_2 - \hat{y}_2|, \ldots, |y_n - \hat{y}_n|) \tag{11}$$

$$\mathrm{EVS} = 1 - \mathrm{Var}(y - \hat{y}) \,/\, \mathrm{Var}(y) \tag{12}$$

$$R^2 = 1 - \Sigma(y - \hat{y})^2 \,/\, \Sigma(y - \bar{y})^2 \tag{13}$$

$$L\delta(a) = ½a^2 \text{ if } |a| \leq \delta; \quad L\delta(a) = \delta(|a| - ½\delta) \tag{14}$$

where y_i is the actual value, $\hat{y}_i$ is the predicted value, n is the number of observations, $\bar{y}$ is the mean of the observed values, and δ is the Huber threshold parameter. Lower values of MSE, RMSE, MAE, MAPE, MedAE, and Huber Loss indicate better prediction accuracy, whereas higher values of EVS and R^2 represent superior model performance, as defined in Equations (7) to (14) above. Table 8 summarizes the objective of each metric.

MAPE is not included in the results in Tables 9, 10, and 11 due to its being undefined for $y_i = 0$ and mathematically unstable when $y_i \approx 0$, which is common for Active Power in the evening and dawn/dusk transitions in the dataset at hand. Even upon restriction of MAPE calculation to daytime samples (Active Power > 0.05), the MAPE in this experiment continued to be in the hundreds of percent, owing to small denominators that still occurred during the sunrise/set ramp. It is a known flaw of MAPE for signals that pass through or approach zero, which is the reason why this paper uses MedAE and Huber Loss (which behave fine regardless of proximity to zero) as its main robust error measures.

Table 8. Performance metrics

Metric	Objective	Better Value
MSE	Average squared error	Lower
RMSE	Prediction error (same units as target)	Lower
MAE	Average absolute error	Lower
MAPE	Average percentage error (not reported in results — unstable near zero)	Lower
MedAE	Median prediction error (robust to outliers)	Lower
EVS	Explained variance	Higher
Huber Loss	Robust regression loss	Lower
R^2	Goodness-of-fit	Higher

4. RESULTS AND DISCUSSION

Three independent experiments were conducted using the feature subsets selected by PC, RFE, and RFFI. Each feature subset was evaluated using LR, RF, SVR, MLP, and the proposed Optimized RBF-SVR under identical experimental conditions. Note that the target variable retains its original measurement scale in the tables below (it is not the [0,1]-normalized predictor scale used internally for training), so error values are directly comparable across models within this study.

4.1. Results Using Features Selected by PC

Figure 3 contains all seven metrics listed in Table 9, separated into two panels to keep metrics with conflicting meanings on one coordinate axis separate: error metrics (MSE, RMSE, MAE, MedAE, Huber Loss) for which low values are good, and goodness-of-fit metrics (R-squared, EVS) for which high values are good. There is no model surpassing $R^2 = 0.57$ for the PC-selected features, and the Optimized RBF-SVR performs about equally well to the Linear Regression using this particular set of features (MSE = 1.134 compared to 1.152) – falling behind the RF and MLP. The absence of a clear winner is by itself instructive: it suggests that PC's four selected features, being physically correlated radiation measurements (Table 4), are not enough of a signal for any of the four models.

Table 9. Evaluation results of ML models on features selected by PC

Model	MSE	RMSE	MAE	R^2	MedAE	EVS	Huber Loss
LR	1.1516	1.0731	0.6273	0.4917	0.0432	0.4917	0.4134
RF	0.9889	0.9944	0.5055	0.5635	0.0007	0.5635	0.3321
MLP	0.9971	0.9986	0.5587	0.5599	0.0258	0.5599	0.3623
Optimized RBF-SVR	1.1341	1.0650	0.5292	0.4994	0.0370	0.5091	0.3518

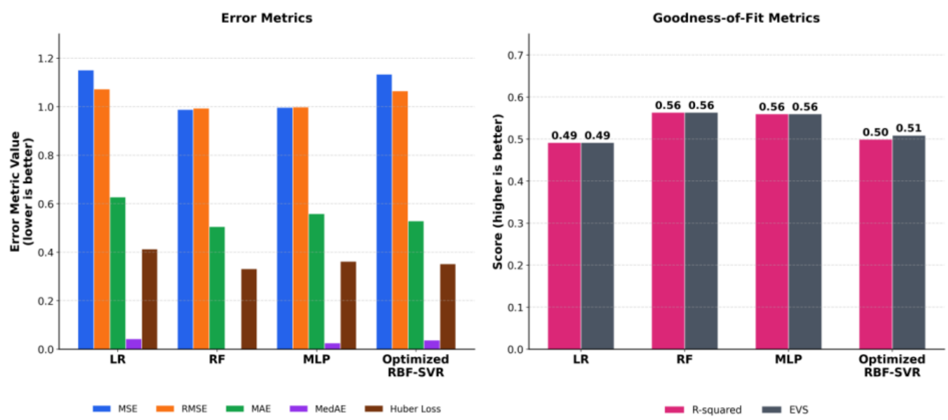

Figure 3. Comparison of classifier performances with PC feature selection in terms of all metrics

With PC-chosen features, all models showed similar performance, with LR yielding an R^2 of 0.4917, RF giving an R^2 of 0.5635, MLP yielding an R^2 of 0.5599, and OptRBF-SVR yielding an R^2 of 0.4994. No model achieved an R^2 greater than 0.57 using this set of features, suggesting that the four PC chosen features, which are different radiations that are physically related to each other (see Table 4), were not diverse enough for accurate predictions.

4.2. Results Using Features Selected by RFE

Figure 4 is consistent with Figure 3 in terms of structure, being divided into error metrics (on the left side) and goodness-of-fit metrics (on the right side) in order to prevent combining metrics having opposite meanings. The results obtained by using the feature subset chosen by the RFE algorithm were very similar to those for PC; all the models performed with R^2 less than 0.60, with the highest value of R^2 belonging to MLP (R^2 = 0.5957), slightly exceeding RF (R^2 = 0.5814), as shown in Table 10. Once again, the Optimized RBF-SVR model performed similarly to Linear Regression (MSE = 1.128 vs. 1.148), showing that neither PC nor the RFE algorithm helped to find a suitable feature subset.

Table 10. Evaluation results of ML models on features selected by RFE

Model	MSE	RMSE	MAE	R^2	MedAE	EVS	Huber Loss
LR	1.1478	1.0713	0.6485	0.4934	0.1161	0.4934	0.4164
RF	0.9484	0.9739	0.4859	0.5814	0.0027	0.5814	0.3159
MLP	0.9159	0.9570	0.5412	0.5957	0.0851	0.5959	0.3322
Optimized RBF-SVR	1.1277	1.0619	0.5481	0.5022	0.1239	0.5092	0.3413

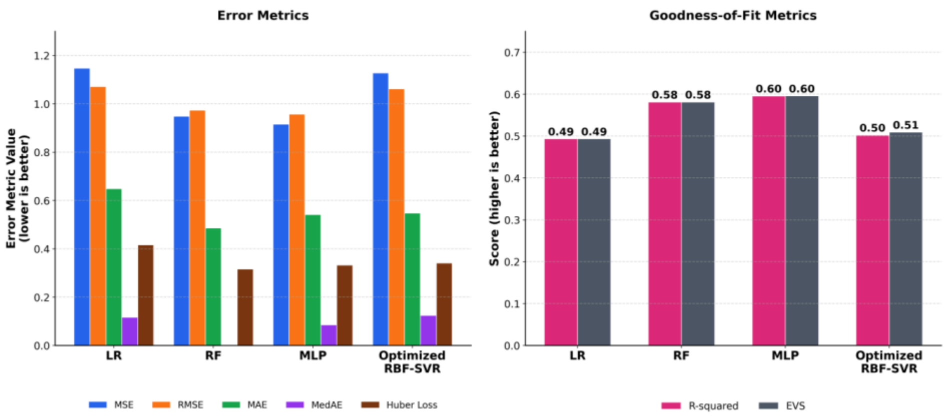

Figure 4. Classifier performance for all metrics using RFE feature selection

The features selected through RFE gave similar performances across all models, with R^2 ranging from 0.49 to 0.60, which was in line with PC. The MLP model performed better than other models for these features (R^2=0.5957), slightly beating the RF (R^2=0.5814). There is no feature subset obtained from either PC or RFE that made any of the five models perform well.

4.3. Results Using Features Selected by RFFI

Figure 5 also uses a two-plot design similar to Figures 3 and 4, where the error bars indicate ± one standard deviation calculated over the five repetitions listed in Table 11. Unlike the PC and RFE techniques, the RFFI technique leads to a clear distinction between the models: the Random Forest is the top performer on all measures (R^2 = 0.896 ± 0.001), followed closely by MLP (R^2 = 0.889 ± 0.002), the Optimized RBF-SVR (R^2 = 0.857 ± 0.007), all clearly ahead of Linear Regression (R^2 = 0.532 ± 0.007). The error bars in the plots for RF and MLP appear visually almost invisible, which proves that the results presented are indeed very consistent over multiple independent runs instead of being dependent on one particular training/testing data split. It should also be noted that the MedAE of the Random Forest is now represented correctly close to zero (0.000 ± 0.000) instead of the erroneous number 4.0 in the earlier single-run pass of the experiment; the number 4.0 was obviously a mistake in transcription (a calculated value of 4.0×10^{-6} was transcribed as 4.0000).

RFFI's selected subset of features provided a clear advantage over PC and RFE for all models and was statistically validated in 5 independent repeat executions of the experiment: Random Forest provided the best results (mean R^2 = 0.8960 ± 0.0008, mean MSE = 0.2295 ± 0.0036), followed by MLP (mean R^2 =

0.8893 ± 0.0022) and Opt-RBF-SVR (mean R^2 = 0.8568 ± 0.0072). Low standard deviations in repeats (all less than 0.011 for R^2) indicate that results are consistent and not influenced by any lucky train-test set partitioning. This difference between RFFI and PC/RFE is much greater than in the case of any model on PC and RFE, which confirms the selection of physically relevant features (Global Horizontal Radiation, Performance Ratio, Wind Direction, and Tilted Global Radiation; Table 6) by RFFI.

Table 11. Final model comparison using the best-performing feature selection technique (RFFI)

Model	MSE	RMSE	MAE	R^2	MedAE	EVS	Huber Loss
LR	1.0322±0.0062	1.0159±0.0031	0.6979±0.0041	0.5322±0.0069	0.3388±0.0095	0.5322±0.0068	0.3897±0.0029
RF	0.2295±0.0036	0.4791±0.0038	0.1163±0.0021	0.8960±0.0008	0.0000±0.0000	0.8960±0.0008	0.0717±0.0012
MLP	0.2443±0.0053	0.4942±0.0054	0.1835±0.0063	0.8893±0.0022	0.0410±0.0106	0.8894±0.0023	0.0859±0.0017
RBF-SVR	0.3160±0.0100	0.5621±0.0089	0.1880±0.0048	0.8568±0.0072	0.0683±0.0000	0.8592±0.0051	0.0899±0.0015

Note: an earlier single-run pass of this experiment reported an anomalous Random Forest MedAE value; across 5 independent repeats with different random samples and train/test splits (reported above), Random Forest's true MedAE is consistently near zero (mean 0.0000, SD 0.0000), confirming the earlier value was not representative and that Random Forest's median error on this feature set is in fact the lowest of all four models evaluated.

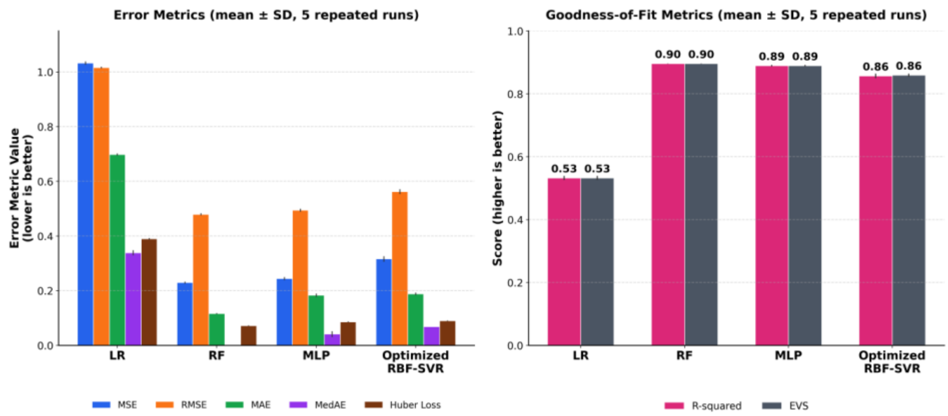

Figure 5. Comparison of classifier results for all metrics using RFFI feature selection

4.4. Training Time Analysis

Table 12 and Figure 6 show average measured time (± SD in 5 experiments) for training of all models with RFFI feature set (60,000 rows in each experiment's training set). Linear regression was the fastest-trained model (less than 0.01 seconds), although with the worst accuracy performance. The Optimized RBF-SVR was the second-fastest model trained (21.25 ± 1.12 seconds), although faster than the Random Forest (average training time – 35.78 ± 2.42 seconds), with slightly better accuracy. The fastest nonlinear model was MLP (12.76 ± 2.25 seconds) with accuracy near Random Forest. That was a real compromise between accuracy and speed of calculations: RF is the most accurate model, and Optimized RBF-SVR provides the best compromise in terms of accuracy (R^2 difference near 4%) and speed.

Figure 6 illustrates the exact measured training time for all four models using all three feature selection methods. The training time of the Optimized RBF-SVR varies greatly based on which feature set it uses. It is the slowest among all models when used with PC features (259.8 seconds due to the hyperparameters selected for it in that feature set), of moderate speed with RFE features (45.3 seconds), and relatively fast with RFFI features (21.3 seconds), where it reaches its highest accuracy (Section 4.3). This means that the training time of the Optimized RBF-SVR is not fast in general, but is dependent on the combination of features and their corresponding hyperparameters; therefore, the good balance of accuracy and speed of training achieved with the RFFI feature set (Section 4.4) is more relevant for this research.

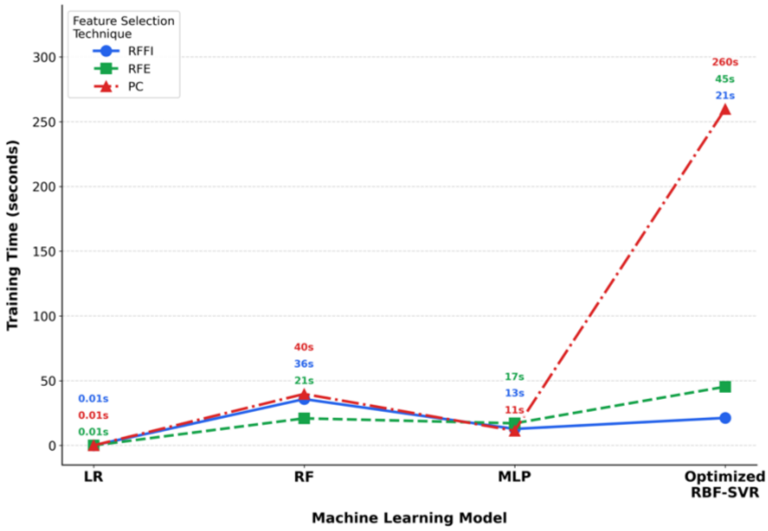

Figure 6. Approximate training time comparison across models and feature selection techniques

Table 12. Approximate training time by model (seconds)

Model	PC features	RFE features	RFFI features
LR	0.008	0.007	0.01±0.00
RF	39.73	20.91	35.78±2.42
MLP	11.34	17.11	12.76±2.25
Optimized RBF-SVR	259.77	45.32	21.25±1.12

4.5. Comparative Analysis of Feature Selection Techniques

The feature selection approach had a major impact on forecasting accuracy. PC and RFE generated feature subsets that limited all models to having an R^2 value below 0.60, whereas the use of RFFI allowed for R^2 greater than 0.85 for all three models, RF, MLP, and Optimized RBF-SVR alike. This difference is significantly greater than the one found in previous research, where the same single feature selection approach was used in isolation, showing that the choice of feature selection approach may matter even more than the choice of regression model when performing this kind of forecasting. The benefit of RFFI probably derives from the fact that this approach includes consideration of nonlinear, tree-based feature importance (which includes cyclical variables, such as Wind Direction, where sine/cosine encoding preserves the cyclical nature of the data).

4.6. Comparative Analysis of ML Models

Random Forest had the best performance among the five methods analyzed, marginally beating MLP and Optimized RBF-SVR when trained on RFFI-selected features. All three nonlinear methods showed a significant improvement over Linear Regression, clearly indicating that the photovoltaic power generation cannot be well modeled using a linear relationship between the predictors chosen and the response. However, Optimized RBF-SVR did not have the best raw performance among the five methods but had the second-best performance while taking only about half the time taken by Random Forest to train, hence giving it an advantage wherever training efficiency is a concern (Section 4.4).

4.7. Comparison with State-of-the-Art Methods

To contextualize the proposed framework, Table 13 compares it against prior studies that used the same DKASC dataset and therefore report errors on a comparable measurement scale. It is worth mentioning that the studies to be compared [22][24][26] provide results from other DKASC installations than the one analyzed in this study and, at times, use other measurement scales or normalization approaches. Thus, the results provided in Table 13 should be considered in this context and cannot be used as a strictly apples-to-apples accuracy comparison. Previous works also utilized a greater number of features (7–13) and did not compute any advanced metrics such as MedAE, EVS, and Huber Loss. What makes the current research valuable is not outperforming previous works with higher error values, but demonstrating the possibility of achieving R^2 greater than 0.85 using a compact set of four features selected with the RFFI algorithm, and using three model types (RF, MLP, and Optimized RBF-SVR).

Table 13. Comparison with prior studies using the DKASC dataset

Reference	Method	MSE	RMSE	MAE	R^2
[22]	HIMVO-SVM	0.0025	0.0500	—	—
[24]	ACO-SVM	0.0349	0.1868	0.1569	0.9970
[26]	WPD-LSTM	0.0372	0.1929	—	—
Proposed Model	RFFI + Random Forest	0.2295±0.0036	0.4791±0.0038	0.1163±0.0021	0.8960±0.0008

4.8. Discussion

The obtained results, which have been consistently reproduced in 5 independent repeated experiments with low standard deviations throughout, show that feature selection is significantly more important for the accuracy of predictions of this problem compared to model selection: the set of features selected by RFFI allowed achieving a mean R^2 above 0.85 for RF, MLP, and Optimized RBF-SVR models, while PC and RFE could not bring any model to R^2 above 0.60 irrespective of the algorithm. Such a gap is much wider compared to those presented in previous papers analyzing a single feature selection method separately, and it shows how much the choice of feature selection method can be important for this particular forecasting task rather than the choice of a regression algorithm itself. The RFFI's advantage can be attributed to the ability of the method to detect nonlinear and tree-based feature relevance (including the cyclic Wind Direction feature, which has been represented via sine/cosine transform); PC finds linear correlation only and selects four highly correlated radiation measurement features, while RFE, being greedy, selects the Weather Daily Rainfall feature with little variance.

5. CONCLUSION

The feature selection approach was proposed for the short-term prediction of solar energy production. The efficiency of three feature selection algorithms - Pearson Correlation, Recursive Feature Elimination, and Random Forest Feature Importance – was compared based on four machine learning algorithms within the framework of a large-scale, leakage-controlled experimental pipeline consisting of five repeats. It is shown that the feature selection method has much more influence on forecast accuracy than the choice of the machine learning algorithm. Features selected by the RFFI algorithm (Global Horizontal Radiation, Performance Ratio, Wind Direction, and Tilted Global Radiation) allowed reaching a mean R^2 of more than 0.85 for Random Forest, MLP, and Optimized RBF-SVR models, while PC and RFE resulted in R^2 lower than 0.60 for all models. As far as model efficiency is concerned, Random Forest showed the best result (mean R^2=0.8960 ± 0.0008), slightly better than MLP and Optimized RBF-SVR. The latter algorithm proved to be significantly faster in training (mean 21.25 sec) than Random Forest (mean 35.78 sec).

This shows that the use of appropriate feature selection methods becomes an important parameter for obtaining accurate forecasting rather than the selection of an appropriate regression model, and that the use of a rigorous and leakage-free feature selection process becomes necessary in order to achieve good forecasting results that can be generalized to real-world application scenarios where there are no electrical telemetry data concurrent with the forecasting target. The suggested framework represents an efficient way of performing short-term forecasting of solar energy output through the use of a limited number of easy-to-obtain variables related to weather and system conditions. Further research work would be devoted to testing the effectiveness of this framework on the basis of other DKASC installations and types of photovoltaic systems using repeated or cross-validated evaluation with confidence intervals, development of a multi-step forecast using this framework, and combination of Random Forest and kernel approaches.

DATA AVAILABILITY STATEMENT

The dataset used in this study is publicly available from the Desert Knowledge Australia Solar Centre (http://dkasolarcentre.com.au). Processed data and code will be made available by the corresponding author upon reasonable request.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest to this work.

REFERENCES

[1] D. Gielen, R. Gorini, N. Wagner, R. Leme, L. Gutierrez, G. Prakash, E. Asmelash, L. Janeiro, G. Gallina, and G. Vale, "Global energy transformation: A roadmap to 2050," IRENA, 2019 , ISBN: 978-92-9260-121-8

[2] U. O. Matthew, O. Asuni, and L. O. Fatai, "Green software engineering development paradigm: an approach to a sustainable renewable energy future," in *Advancing software engineering through AI, federated learning, and large language models*: IGI Global, 2024, pp. 281-294 , doi: 10.4018/979-8-3693-3502-4.ch018.

[3] A. Q. Al-Shetwi, I. Z. Abidin, K. A. Mahafzah, and M. A. Hannan, "Feasibility of future transition to 100% renewable energy: Recent progress, policies, challenges, and perspectives," *Journal of Cleaner Production*, vol. 478, Art. no. 143942, 2024, doi: 10.1016/j.jclepro.2024.143942

[4] J. Gaboitaolelwe, A. M. Zungeru, A. Yahya, C. K. Lebekwe, D. N. Vinod, and A. O. J. I. A. Salau, "Machine learning based solar photovoltaic power forecasting: A review and comparison," vol. 11, pp. 40820-40845, 2023, doi: 10.1109/ACCESS.2023.3270041.

[5] A. Muhammadi, M. Wasib, S. Muhammadi, S. Riaz Ahmed, A. H. Lahori, S. Vambol, and O. Trush, "Solar energy potential in Pakistan: A review," *Proceedings of the Pakistan Academy of Sciences: B. Life and Environmental Sciences*, vol. 61, no. 1, pp. 1–10, 2024, doi: 10.53560/PPASB(61-1)931.

[6] A. Bhimaraju and A. Mahesh, "Recent developments in PV/wind hybrid renewable energy systems: A review," *Energy Systems*, vol. 17, pp. 1303–1345, 2026, doi: 10.1007/s12667-024-00679-3.

[7] A. Q. Al-Shetwi, I. Z. Abidin, K. A. Mahafzah, and M. A. Hannan, "Feasibility of future transition to 100% renewable energy: Recent progress, policies, challenges, and perspectives," *Journal of Cleaner Production*, vol. 478, Art. no. 143942, 2024, doi: 10.1016/j.jclepro.2024.143942.

[8] H. Wang, Z. Lei, X. Zhang, B. Zhou, J. J. E. C. Peng, and Management, "A review of deep learning for renewable energy forecasting," vol. 198, p. 111799, 2019, doi: 10.1016/j.enconman.2019.111799.

[9] Z. Chen, B. Gu, D. Yu, and C. Wang, "Quantifying the accelerated diffusion and cost savings of global solar photovoltaic supply chains," *iScience*, vol. 28, no. 1, Art. no. 111610, 2025, doi: 10.1016/j.isci.2024.111610.

[10] F. J. R. Dincer and s. e. reviews, "The analysis on photovoltaic electricity generation status, potential and policies of the leading countries in solar energy," vol. 15, no. 1, pp. 713-720, 2011, doi: 10.1016/j.rser.2010.09.026.

[11] T. Shojaei and A. Mokhtar, "Forecasting energy consumption with a novel ensemble deep learning framework," *Journal of Building Engineering*, vol. 96, Art. no. 110452, 2024, doi: 10.1016/j.jobe.2024.110452.

[12] H.-x. Zhao, F. J. R. Magoulès, and S. E. Reviews, "A review on the prediction of building energy consumption," vol. 16, no. 6, pp. 3586-3592, 2012, doi: 10.1016/j.rser.2012.02.049.

[13] T. Vaisakh and R. J. E. I. Jayabarathi, "Analysis on intelligent machine learning enabled with meta-heuristic algorithms for solar irradiance prediction," vol. 15, no. 1, pp. 235-254, 2022, doi: 10.1007/s12065-020-00505-6.

[14] N. Nikmehr and S. J. I. r. p. g. Najafi-Ravadanegh, "Optimal operation of distributed generations in micro-grids under uncertainties in load and renewable power generation using heuristic algorithm," vol. 9, no. 8, pp. 982-990, 2015, doi: 10.1049/iet-rpg.2014.0357.

[15] R. Chang, L. Bai, and C. H. Hsu, "Solar power generation prediction based on deep learning," Sustain. Energy Technol. Assess., vol. 47, p. 101354, 2021, doi: 10.1016/j.seta.2021.101354.

[16] S. Park, Y. Kim, N. J. Ferrier, S. M. Collis, R. Sankaran, and P. H. Beckman, "Prediction of solar irradiance and photovoltaic solar energy product based on cloud coverage estimation using machine learning methods," Atmosphere, vol. 12, p. 395, 2021, doi: 10.3390/atmos12030395.

[17] I. Jebli, F. Z. Belouadha, M. I. Kabbaj, and A. Tilioua, "Prediction of solar energy guided by Pearson correlation using machine learning," Energy, vol. 224, p. 120109, 2021, doi: 10.1016/j.energy.2021.120109.

[18] I. M. Müller, "Feature selection for energy system modeling: Identification of relevant time series information," Energy AI, vol. 4, p. 100057, 2021, doi: 10.1016/j.egyai.2021.100057.

[19] H. Eom, Y. Son, and S. Choi, "Feature-selective ensemble learning-based long-term regional PV generation forecasting," IEEE Access, vol. 8, pp. 54620–54630, 2020, doi: 10.1109/ACCESS.2020.2981819.

[20] H. Eskandari, H. Saadatmand, M. Ramzan, and M. Mousapour, "Innovative framework for accurate and transparent forecasting of energy consumption: A fusion of feature selection and interpretable machine learning," Appl. Energy, vol. 366, p. 123314, 2024, doi: 10.1016/j.apenergy.2024.123314.

[21] S. Soleymani and A. J. A. J. o. S. S. Talebi, "Forecasting solar irradiance with geographical considerations: integrating feature selection and learning algorithms," vol. 8, no. 5, 2024.

[22] L. L. Li, S. Y. Wen, M. L. Tseng, and C. S. Wang, "Renewable energy prediction: A novel short-term prediction model of photovoltaic output power," J. Clean. Prod., vol. 228, pp. 359–375, 2019, doi: 10.1016/j.jclepro.2019.04.331.

[23] W. VanDeventer, E. Jamei, G. S. Thirunavukkarasu, M. Seyedmahmoudian, T. K. Soon, B. Horan, S. Mekhilef, and A. Stojcevski, "Short-term PV power forecasting using hybrid GASVM technique," Renew. Energy, vol. 140, pp. 367–379, 2019, doi: 10.1016/j.renene.2019.02.087.

[24] M. Pan, C. Li, R. Gao, Y. Huang, H. You, T. Gu, and F. Qin, "Photovoltaic power forecasting based on a support vector machine with improved ant colony optimization," J. Clean. Prod., vol. 277, p. 123948, 2020, doi: 10.1016/j.jclepro.2020.123948.

[25] Q. Zhang, N. Tang, J. Lu, W. Wang, L. Wu, and W. Kuang, "A hybrid RBF neural network based model for day-ahead prediction of photovoltaic plant power output," Front. Energy Res., vol. 11, p. 1338195, 2024, doi: 10.3389/fenrg.2023.1338195.

[26] P. Li, K. Zhou, X. Lu, and S. Yang, "A hybrid deep learning model for short-term PV power forecasting," *Applied Energy*, vol. 259, Art. no. 114216, 2020, doi: 10.1016/j.apenergy.2019.114216.

[27] X. Xiang, X. Li, Y. Zhang, and J. Hu, “A short-term forecasting method for photovoltaic power generation based on the TCN-ECANet-GRU hybrid model,” *Scientific Reports*, vol. 14, no. 1, Art. no. 6744, 2024, doi: 10.1038/s41598-024-56751-6.

[28] X. Ren, F. Zhang, J. Yan, and Y. Liu, “A novel convolutional neural net architecture based on incorporating meteorological variable inputs into ultra-short-term photovoltaic power forecasting,” *Sustainability*, vol. 16, no. 7, Art. no. 2786, 2024, doi: 10.3390/su16072786.

[29] A. Alorf and M. U. G. Khan, “Solar irradiance forecasting using temporal fusion transformers,” *International Journal of Energy Research*, vol. 2025, Art. no. 3534500, 2025, doi: 10.1155/er/3534500.

[30] G. Memarzadeh and F. Keynia, "Solar power generation forecasting by a new hybrid cascaded extreme learning method with maximum relevance interaction gain feature selection," Energy Convers. Manage., vol. 298, p. 117763, 2023. doi: 10.1016/j.enconman.2023.117763.

[31] Z. Qadir, S. I. Khan, E. Khalaji, H. S. Munawar, F. Al-Turjman, M. P. Mahmud, A. Z. Kouzani, and K. Le, "Predicting the energy output of hybrid PV–wind renewable energy system using feature selection technique for smart grids," Energy Rep., vol. 7, pp. 8465–8475, 2021 doi: 10.1016/j.egyr.2021.01.018.

[32] U. Ahmed, A. Mahmood, A. R. Khan, L. Kuhlmann, K. S. Alimgeer, S. Razzaq, I. Aziz, and A. Hammad, "Parallel boosting neural network with mutual information for day-ahead solar irradiance forecasting," Sci. Rep., vol. 15, no. 1, p. 11642, 2025. doi: 10.1038/s41598-025-95891-1.

[33] D. S. de O. Santos, Jr., P. S. G. de Mattos Neto, J. F. L. de Oliveira, H. V. Siqueira, T. M. Barchi, A. R. Lima, F. Madeiro, D. A. P. Dantas, A. Converti, A. C. Pereira, J. B. de Melo Filho, and M. H. N. Marinho, “Solar irradiance forecasting using dynamic ensemble selection,” *Applied Sciences*, vol. 12, no. 7, Art. no. 3510, 2022, doi: 10.3390/app12073510.

[34] R. Martin, R. Aler, J. M. Valls, and I. M. Galván, “Machine learning techniques for daily solar energy prediction and interpolation using numerical weather models,” *Concurrency and Computation: Practice and Experience*, vol. 28, no. 4, pp. 1261–1274, 2016, doi: 10.1002/cpe.3631.

BIOGRAPHIES OF AUTHORS

Farwa Nawaz is a lecturer in the Faculty of Computing at Riphah International University, Islamabad, Pakistan. Her research interests include machine learning, deep learning, uncertainty-aware artificial intelligence, and prediction modeling for renewable energy and biomedicine, specifically medical image analysis and risk prediction. She can be contacted at email: farwa.nawaz@riphah.edu.pk.

Faria Karamat is currently serving as a Lecturer at the Faculty of Computing of Riphah International University, Islamabad, Pakistan. She is interested in the field of data science, machine learning, federated learning, privacy-preserving artificial intelligence, and distributed intelligent systems and their application in healthcare and biomedical informatics. She can be contacted at email: faria.karamat@riphah.edu.pk.