Journal of Engineering and Applied Research (JEAR)

National University of Skill

Winter & Spring 2026, Vol. 3, No. 1, p. 208-228

Journal Homepage: https://jear.tvu.ac.ir

:10.48301/vjtve.2026.542635.1139

AI-Driven Adaptive Hybrid Forecasting of Solar PV Power via Dirichlet Process Clustering and Ensemble Regression

Ali Asghar Tehrani [1], Sania Sami [2], Tayyeb Nayghelichi[3*], Soha Sami[4], Mohammad Hajian[5]

[1] Department of Management, SR.C., Islamic Azad University, Tehran, Iran.

[2] Department of Veterinary Medicine, Ilam University.

[3, 5] Gorgan University of Agricultural Sciences and Natural Resources, Faculty of Water and Soil Engineering, Department of Biosystems Engineering, Gorgan, Iran.

[4] Mechanics of Biosystems Engineering Department, College of Aburaihan, University of Tehran, Tehran, Iran.

ARTICLE INFO

Article Type:
Original Research

Received: 2025.09.07
Revised: 2025.11.08
Accepted: 2026.05.17

Keyword:
Photovoltaics
Power Forecasting
Hybrid AI Models
Deep Learning, Renewable Energy
Machine Learning
Cluster-Based Prediction

***Corresponding Author:**
Tayyeb Nazghelichi
Email: *tnazghelichi@gmail.com*

ABSTRACT

Accurate forecasting of photovoltaic (PV) power generation is crucial for enhancing the efficiency and reliability of renewable energy in existing power grids. Despite significant progress in predicting PV generation, the nature of PV data remains heterogeneous, non-linear, and dynamic, making modeling and predictive ability challenging in their own right. This dissertation proposes a self-adaptive hybrid approach that offers a novel development avenue within the renewable energy forecast paradigm. It utilizes a Nonparametric Dirichlet Process Gaussian Mixture Model (DPGMM) to identify the distinct predictive capabilities of machine-learning algorithms from a validation perspective, thereby informing their employment in the hybrid method. DPGMM was able to identify homogeneous subgroups from a nearly 3-year PV dataset of 300 households in Sydney, Australia, enabling the development of distinct predictive models for each subgroup. From a statistical validation perspective, contrast testing was completed using a Friedman and Wilcoxon signed-rank procedure. While there exist differences between the models ($p < 0.05$), it was clear that Gradient Boosting Regression (GBR) outperformed the Support Vector Regression (SVR) and Random Forest Regression (RFR) models on all clusters. Finally, metrics of the mean absolute error (MAE), mean squared error (MSE), and root mean square error (RMSE) were employed to examine and measure improvements in accuracy compared to baseline methods. The hybrid approach yielded statistically significant improvements, which included increases of 12% for MAE, 22% for MSE, and 15% for RMSE compared to single-stage convolution implementations.

E-ISSN: 3060-6357

1)Introduction

Reliable forecasting of photovoltaic (PV) power is essential for grid reliability with growing renewable energy resource penetration. Solar generation forecasts are complicated by weather and atmosphere uncertainty, leading to potential supply–demand mismatches. Reliable forecasts can minimize supply-demand mismatches while reducing operating costs and enhancing authentic energy management, resource optimization, and successful energy transitions in a sustainable way[1].

Fractions of loads exceeding conditions have emerged in demand. Optimizing predictors can enhance forecasting accuracy, minimize forecast errors, and increase the reliability of the grid. Conventional statistical and mechanistic approaches typically do not sufficiently absorb complex, nonlinear, high-dimensional, and time-varying photovoltaic (PV) data in the presence of rapidly evolving weather conditions, due to limited flexibility and assumptions of linearity only[2; 3]. Compared to traditional forecasting methods, machine learning (ML) and deep learning (DL) methods, including supervised and unsupervised methods, are better suited for leveraging complex relationships and accounting for the temporal nature of data, which leads to greater forecasting accuracy[4]. Hybrid methods, however, produce even more accurate forecasts, fewer forecasting errors, and even less system uncertainty of errors [5].

Modern Technologies provide new opportunities through large datasets, with multiple sources of data, including comprehensive meteorological data combined with exhaustive lists of historical photovoltaic (PV) performance data to identify and analysing many environmental parameters in forecasting, revealing and adaptive towards somewhat small developmentally important forecast details and predicting (dynamically) future possibilities [3; 5].

Sophisticated preprocessing methods, feature engineering, and clustering algorithms increase model processing and efficiency, ultimately leveraging machine learning (ML) algorithms to exploit big data [4]. The integration of big data with ML provides a pathway to actionable data that yields improved energy planning and improved resilience of the grid [2]. When deploying hybrid models, which combine the algorithm components of Gradient Boosting Regressors (GBR), Random Forest Regressors (RFR), and Support Vector Regression (SVR), it is common that hybrid model performance will exceed that of single model performance due to the distributed functional characteristics of the underlying algorithms. For instance, GBRs can effectively handle large datasets; RFRs are not affected by outliers; SVRs can demonstrate feature effectiveness in high-dimensional spaces. The limitations of GBR overspecification are mitigated by RFRs, which are less sensitive to noisy data, while SVRs can potentially overcome the reduction in prediction capability related to fuzzy levels of input data [3; 4]. Optimization methods (hyperparameter optimization and clustering-based segmentation) improve our confidence and decisions around renewable energy integration [5]. In this context, unsupervised (e.g., k-means clustering, Autoencoders) learning can extract hidden patterns from solar and meteorological data and reduce prediction error [5].

Moreover, the development of solar PV system optimization and design advances, as shown by Mofidian et al. [6], underscores the value of taking advantage of simulation tools such as PVsyst in optimizing energy production and system performance. Their analysis on a 200 kW rooftop PV system near Shiraz points to the need for accurate parameter adjustment, such as panel direction and shading study, in obtaining high performance with optimum losses. This corresponds with the larger objective of implementing renewable energy systems in the grid in an effective manner. In addition, the consideration given to quantum machine learning (QML) by Taghandiki [7] points out the ability of quantum algorithms to transform data analysis and optimization problems with exponential acceleration benefits and better scale for high-complexity forecast models. Such improvements all point to the development of renewable energy forecast for the grid.

The lack of accuracy, flexibility, and new features and or patterns in complex data sets remains an issue. There is progress in the development of modeling frameworks in this area, but improvements can still be made. This study proposes a new hybrid method to progress the understanding of the models by developing a natural and implicit connection between DPGMM and machine learning. The proposed method leverages DPGMM's properties for automatically discovering data classes, thereby enhancing the estimates and prediction accuracy of a model or subset, reducing computational time and costs, and providing a scalable method for handling varying data sizes in photovoltaic forecasting. Through this sense of achievement of meeting these challenges, there is an added optimism and feeling of comfort about the extended impacts of PV power forecasting.

1.2) Research background

To place our work in context and emphasize the progress in PV power forecasting, this section presents a systematic background of the research, grouping the various methodologies into major areas to provide a concise and clear overview of the progress in the field.

1.3) Hybrid Models

Hybrid models are now commonplace for improving the accuracy and credibility of PV power forecasting. Hybrid models are built on the principles of education and cooperation, in which the strengths of multiple methods are combined to address the disadvantages of individual methods within the approach. Hybrids can combine multiple types of machine learning, signal processing, and optimization algorithms, models, and typically perform better than a single model regarding the complexities that PV power forecasting presents. For example, the WOA-VMD-SCINet model's unique combination of Whale Optimization Algorithm (WOA), Variational Mode Decomposition (VMD), and Sampled Interaction Neural Network (SCINet) is different in this domain. This hybrid assesses meteorological variables directly utilizing Pearson correlation coefficients, hence resulting in significant reductions in errors in seasonal predictions, and is increasing the overall credible forecasts made [8]. In long term forecasting, an outstanding example of fusion from

integration work is the KM-PSO-SVR model understood as the integration of K-means clustering (KM), Particle Swarm Optimization (PSO) and SVR - their fusion facilitates much higher-quality climate information, resolved using the Vine Copula methods ultimately leading to significant increases in long-term forecasting (Z. Liu et al.). Broad Learning System (BLS) hybrid models can be combined with Relevance Vector Machine (RVMs) and advanced statistical scaling models to achieve significantly lower RMSE compared to traditional methods. As the research shows [9], these BLS-RVM hybrids perform exceptionally well on regional PV cluster power generation. These examples illustrate flexibility and improved outcomes for hybrid methods for a variety of forecasting conditions and data.

1.4) Optimization Algorithms

Optimization algorithms are instrumental in developing optimized PV power forecasting models to allow them to transition to varying geographies and environmental conditions. The optimization algorithms enable adapting rapidly to changing environmental conditions and thus introduce badly needed flexibility and robustness into forecasting systems. A conventional example is the Squirrel Optimization Algorithm (SOA) adjusted Deep Belief Network (DBN) model that effectively fine-tunes the PV power forecast while accounting for several geographic variances. All optimization solutions have improved the model's applicability by enabling various environmental conditions to be adopted while improving the accuracy and reliability of predictions [10]. Furthermore, the multi-strategy improved Squirrel Search Algorithm (SSA), when combined with Support Vector Machine (SVM), provides a new way of enhancing predictive model performance. By effectively applying Principal Component Analysis (PCA) for dimensionality reduction and SSA for hyperparameter tuning, this approach achieves greater accuracy and stability when forecasting PV power, as proposed by Zhu et al [11].

1.5) Statement of the problem

The survey underscores a notable gap in the existing literature: the deficiency of self-adaptive hybrid frameworks capable of dynamically adapting to the inherent structure of photovoltaic (PV) power data. The integration of adaptive clustering techniques, such as Dirichlet Process Gaussian Mixture Models (DPGMM), with forecasting routines to facilitate automatic data segmentation and cluster-specific predictions presents substantial potential for enhancing accuracy and flexibility. This paper proposes a novel adaptive hybrid methodology employing DPGMM for automatic clustering, followed by Support Vector Regression (SVR)/Random Forest Regression (RFR) and Gradient Boosting Regression (GBR) for cluster-specific forecasting. The approach assesses performance using Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Square Error (RMSE), and employs nonparametric statistical tests, including the Friedman and Wilcoxon tests, to analyze differences. The results indicate that this innovative approach achieves significant reductions in predictor error and offers a systematic, scalable framework for photovoltaic power forecasting.

2) Methodology

2.1) Data Collection

Here, consumption and generation data of photovoltaic power from 300 Sydney homes, in kilowatt-hours, were accessed and extracted from the IPART database. Meteorological data for the Sydney region, including environmental factors such as solar radiation, temperature, humidity, and other variables, were collected in an integrated dataset. Pearson's correlation coefficient was utilized to examine the interrelationship and interdependence among these variables. Pearson's correlation coefficient provides both direction and strength information about the relationship between variables, enabling critical predictors of photovoltaic power generation to be established. The selected variables were then readied for the subsequent analysis and model building [12].

The research methodology explores the effect of clustering on the primary dataset and its subsets and employs ML and DL algorithms to forecast photovoltaic power generation.

The following supervised and unsupervised learning techniques were utilized in succession: three primary algorithms (GBR, RFR, and SVR) were implemented on the original and clustered datasets. The performance of the models was subsequently assessed based on different error measures like MAE, MSE, and RMSE. Supervised learning involved regression algorithms, while unsupervised learning involved clustering algorithms. There was a strong focus on pattern recognition and clustering in the data.

2.2) Feature Selection Method

In the analysis, Pearson's correlation was used to identify variables with substantial effects on photovoltaic power generation. These are depicted in Figure 1, which is a heat map representation. The parameters were used to generate a new dataset. Feature selection at optimal levels lowered the prediction error and computational cost and generated more precise and applicable models with lower-cost functions. This combination of properties allowed effective prediction with a reduced dataset size. Before clustering, dimension reduction was performed using PCA as shown in Figure 1.

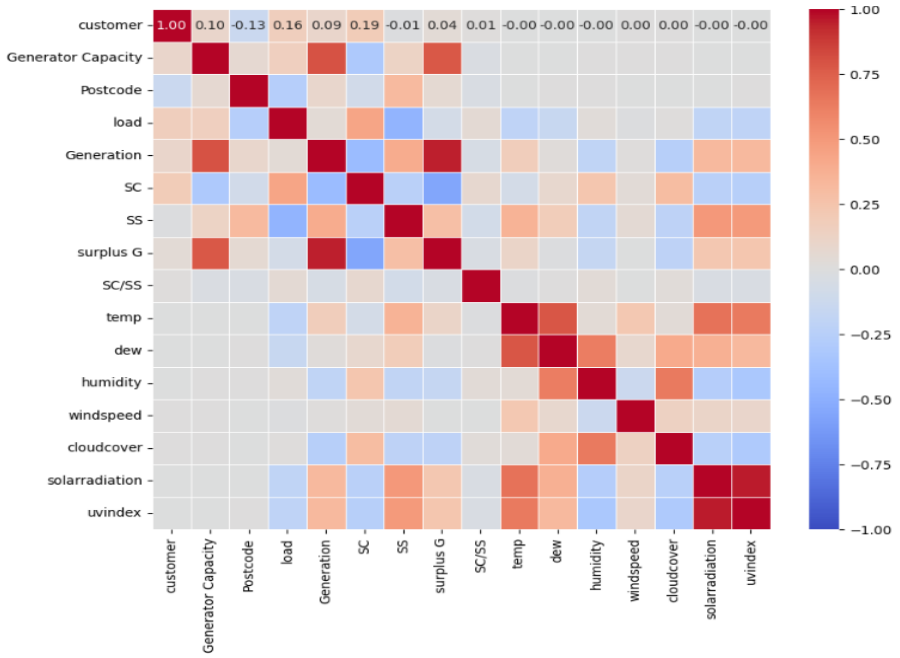

Figure 1. Data Correlation

The most significant parameters selected for photovoltaic power forecasting were temperature, photovoltaic generation capacity, ultraviolet radiation index, solar radiation, and cloud cover. The parameters are in units given in Table 1. Ultraviolet radiation was used as a dimensionless parameter.

Table 1. displays the units of parameters used

Parameter	Unit of Measurement
Air Temperature (T)	Celsius (°C)
Precipitation (Precip)	Millimeters (mm)
Cloud Cover	Percentage (%)
Solar Radiation	Watts per square meter (W/m^2)
Solar Energy	Megajoules per square meter (MJ/m^2)
Actual Generation	Kilowatts per hour (KW/h)

2.3) Data Clustering

Clustering is an unsupervised method for discovering data structures by grouping similar instances. We compared KNN, DBSCAN, hierarchical clustering, and the Dirichlet Process Gaussian Mixture Model (DPGMM); the latter most effectively captured the underlying distribution of photovoltaic

power data. Unlike fixed-component GMMs, DPGMM uses a Dirichlet process to estimate the number of clusters adaptively, making it suitable for dynamic and diverse datasets. Assigning clusters divided the data into interpretable subclusters, each modeled with specialized supervised predictors, creating a hybrid, cluster-specific forecasting model that exploits local patterns. Quantitatively, the DPGMM achieved the highest average Silhouette score, indicating better intra-cluster cohesion and inter-cluster separation.

DL methods, including Autoencoders, were incorporated into the clustering algorithm to enhance accuracy and performance. Autoencoders remove noise and redundancy, enabling the DPGMM algorithm to better identify similarities and underlying patterns in the data [13]. The original dataset was split into 10 clusters of varying sizes. Clusters having more than 1,000 data points were selected for algorithm deployment. Figure 2 illustrates the overview of significant parameters as they are plotted on two axes, with PCA used to reduce the dimension. Groups with similar characteristics were considered for identifying unique patterns. Similarity facilitated an accurate examination of data and made explicit comparisons between clusters.

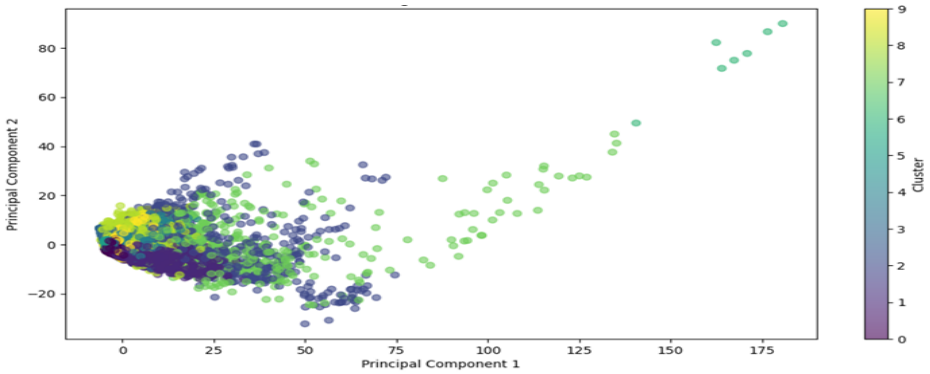

Figure 2. Displaying data based on clustering methodologies

2.4) Supervised Learning Models

As mentioned, three supervised learning models of GBR, RFR, and SVR were employed, each described in detail in the following. GBR is an ensemble learning process that generates a strong regressor through a stagewise optimization process combining multiple weak learners, typically decision trees. It works by minimizing a pre-defined loss function, and Model m is fit to the residuals of the previous model, which allows for an increase in predictive accuracy [14]. Therefore, this model is constructed as an additive function [15]:

$$F_M(x) = \sum_{m=1}^{M} \gamma_m h_m(x) \quad (1)$$

Where $F_M(x)$ is the final predictive model, h_m represents the weak learner (regression tree) at iteration m, γ_m is the learning rate (step size), and M is the total number of boosting iterations.

The GBR model was implemented using the scikit-learn library in Python, with hyperparameters such as the number of trees M, learning rate γ, and tree depth optimized via cross-validation.

RFR was utilized as a nonparametric ensemble learning method to model photovoltaic power generation. RFR is an extension of decision tree regression that mitigates overfitting by aggregating multiple decision trees, each trained on a randomly sampled subset of the data with feature randomness. This approach enhances model generalization and reduces variance while maintaining interpretability [16].

The RFR model consists of an ensemble of T independent regression trees, each denoted as $h_t(x)$. The final prediction is obtained by averaging the outputs of all trees [17]:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_t(x) \tag{2}$$

Where $\hat{y}$ is the predicted response, $h_t(x)$ represents the individual regression tree, and T is the total number of trees in the ensemble. Each tree $h_t(x)$ is constructed using Bootstrap Aggregation (Bagging), where training subsets D_t are drawn randomly with replacement from the original dataset D, ensuring diversity among trees. Each tree is trained by recursively partitioning the feature space to minimize an impurity measure. In regression tasks, MSE is used as the splitting criterion:

$$\mathrm{MSE} = \frac{1}{\mathrm{N}}\sum_{i=1}^{N}(y_i - \bar{y})^2 \tag{3}$$

Where N is the number of samples in the node, y_i represents the target variable, and $\bar{y}$ is the mean value of the target variable in the node. At each split, the algorithm selects the feature and threshold that yield the maximum reduction in impurity:

$$\Delta\mathrm{MSE} = \mathrm{MSE}_{\mathrm{parent}} - (w_{\mathrm{left}} .\ \mathrm{MSE}_{\mathrm{left}} + w_{\mathrm{right}} .\ \mathrm{MSE}_{\mathrm{right}}) \tag{4}$$

Where w_{left} and w_{right} denote the proportions of samples assigned to the left and right child nodes. To decorrelate the trees, RFR selects a random subset of features m at each split. The feature importance score for each feature j is computed as:

$$I_j = \sum_{t=1}^{T}\sum_{s\in\mathrm{splits}} \Delta\mathrm{MSE}_s .\ 1(\text{split uses } x_j) \tag{5}$$

Where 1 is an indicator function that counts how often feature x_j is used in tree splits, weighted by impurity reduction. The RFR model was implemented using scikit-learn in Python. Hyperparameters, including the number of trees T, the maximum depth of trees, and the number of features considered at each split, were optimized via cross-validation.

SVR was utilized as a supervised learning algorithm to predict PV power output. SVR is based on SVM in that it learns a regression model by mapping the input features into a high-dimensional space and optimizing a hyperplane that maximizes the margin while minimizing the prediction error. Unlike previous regression approaches, SVR uses an ε-insensitive loss function, which provides additional flexibility in approximating nonlinear relationships while also keeping track of model complexity [18].

Given a training dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$, where $x_i \in \mathbb{R}^d$ represents the input features and $y_i \in \mathbb{R}$ denotes the target values, the objective of SVR is to find a function $f(x)$ that approximates y with minimal deviation. The standard SVR formulation is expressed as [19]:

$$f(x) = w^T \phi(x) + b \tag{6}$$

Where $w \in \mathbb{R}^d$ is the weight vector, $\phi(x)$ is a nonlinear mapping function that projects the input features into a higher-dimensional space, and b is the bias term. The model is trained by solving the following convex optimization problem:

$$\min_{w,b,\xi_i,\xi_i^*} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) \tag{7}$$

Subject to:

$$\begin{aligned} y_i - w^T\phi(x_i) - b &\le \epsilon + \xi_i \\ w^T\phi(x_i) + b - y_i &\le \epsilon + \xi_i^* \\ \xi_i, \xi_i^* &\ge 0\,, i = 1, \dots, N \end{aligned} \tag{8}$$

Where ϵ is the margin of tolerance, within which errors are ignored, ξ_i and ξ_i^* are slack variables that allow violations beyond ϵ, and C is a regularization parameter controlling the trade-off between margin width and error tolerance. Since real-world data often exhibit nonlinear patterns, SVR employs the kernel trick to implicitly transform the input space without explicitly computing $\phi(x)$. This enables solving the problem in the dual form, where predictions are calculated as:

$$f(x) = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) K(x_i, x) + b \tag{9}$$

Where α_i and α_i^* are Lagrange multipliers and $K(x_i, x) = \phi(x_i)^T \phi(x)$ is the kernel function, which defines the similarity between samples. The SVR model was implemented using scikit-learn with hyperparameter tuning performed via grid search cross-validation. By leveraging the structural risk minimization principle, SVR effectively balances model complexity and generalization performance, making it a powerful tool for nonlinear regression tasks.

2.5) Hybrid Model Development

In this study, a hybrid model was developed by integrating DPGMM with one of the machine learning regression-based algorithms: GBR, RFR, or SVR. The primary objective in this integration was to design an adaptive and independent prediction model that could change its structure based on the underlying data distribution. The DPGMM was employed as an unsupervised probabilistic clustering algorithm to identify latent patterns and group similar points, without prior knowledge of the number of clusters. Once the data was segmented, a unique supervised regression model (GBR, RFR, or SVR) was trained in each identified cluster. This enabled the model to learn cluster-specific patterns, thereby enhancing both local accuracy and overall stability.

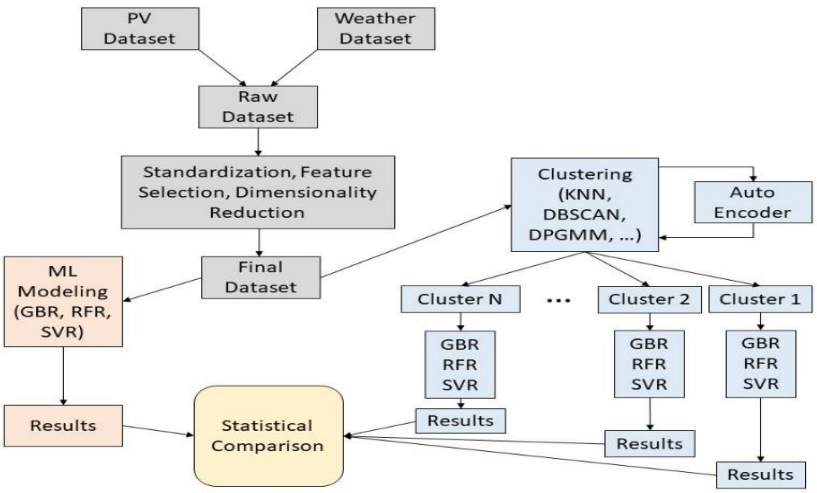

Figure 3. Process flowchart of hybrid model development: The diagram outlines the integration of the DPGMM clustering technique with supervised regression algorithms (GBR, RFR, SVR) to form adaptive cluster-specific predictive models.

Among the hybrid models experimented with, GBR+DPGMM was the top-performing model, as indicated by its high statistical performance across various error measures (MAE, MSE, RMSE), and confirmed by Friedman and Wilcoxon signed-rank tests. The statistical tests confirmed that GBR+DPGMM performed significantly better than the other hybrid models in prediction accuracy at a confidence level of more than 95%. A process flowchart illustrating the hybrid model development process is given in Figure 3. It is a straightforward illustration of the sequential methodology employed in integrating clustering and regression methods for adaptive prediction, based on the inherent data structure.

3) Results and discussion

As mentioned earlier, five clusters (Clusters 1, 2, 5, 6, and 7), each containing more than 1000 data points, were selected from the clustering process. ML algorithms were applied to each cluster in the original dataset. After evaluating and analyzing nine unsupervised and three supervised algorithms, and integrating the results, an effective hybrid algorithm was identified.

The predictive performance of the proposed hybrid model was evaluated using three error metrics (MAE, MSE, and RMSE) across three different learning algorithms (SVR, GBR, and RFR), applied both to the original dataset and to clustered subsets (using DPGMM). Experiments were conducted on two datasets containing 1000 and 500 data samples, respectively. The corresponding error values are reported in Table 2 and Table 3.

Table 2. Error metrics (MAE, MSE, RMSE) of SVR, GBR, and RFR models using DPGMM clustering (1000-sample dataset)

Cluster	Metric	SVR+dpgmm	GBR+dpgmm	RFR+dpgmm
All Data	MAE	1.561	0.99	1.475
	MSE	8.854	1.8	7.775
	RMSE	2.976	1.342	2.788
Cluster 1	MAE	0.794	0.702	0.752
	MSE	1.289	0.756	1.072
	RMSE	1.135	0.869	1.035
Cluster 2	MAE	1.114	0.595	1.498
	MSE	2.751	0.629	4.876
	RMSE	1.659	0.793	2.208
Cluster 3	MAE	0.937	0.46	1.117
	MSE	1.761	0.355	2.087
	RMSE	1.327	0.596	1.445

Cluster 4	MAE	0.731	0.643	0.871
	MSE	1.257	0.656	1.533
	RMSE	1.121	0.81	1.238
Cluster 5	MAE	0.997	0.796	1.003
	MSE	2.022	1.167	1.929
	RMSE	1.422	1.08	1.389

Table 3. Error metrics (MAE, MSE, RMSE) of SVR, GBR, and RFR models using DPGMM clustering (500-sample dataset)

Cluster	Metric	SVR+dpgmm	GBR+dpgmm	RFR+dpgmm
All Data	MAE	1.652	1.212	1.136
	MSE	8.913	4.957	2.218
	RMSE	2.986	2.226	1.489
Cluster 1	MAE	0.959	0.718	0.834
	MSE	1.803	0.743	0.942
	RMSE	1.343	0.862	0.971
Cluster 2	MAE	1.321	0.608	1.105
	MSE	4.751	0.682	1.115
	RMSE	2.18	0.826	1.165
Cluster 3	MAE	1.175	0.582	0.887
	MSE	2.362	0.538	1.127
	RMSE	1.537	0.734	1.062
Cluster 4	MAE	0.982	0.762	0.966
	MSE	2.021	0.965	1.049
	RMSE	1.421	0.982	1.431
Cluster 5	MAE	1.131	0.776	1.003
	MSE	2.84	0.972	1.689
	RMSE	1.685	0.986	0.74

To statistically verify differences in performance between the models, the Friedman test was performed. Results for both the 1000-sample and 500-sample datasets indicated statistically significant differences between the models in all three error measures (p-value = 0.0316 for MSE and RMSE; p = 0.0176 for MAE in the 500-sample dataset). This suggests that model selection has a significant impact on predictive accuracy.

Further pairwise comparisons were conducted using the Wilcoxon signed-rank test. Outcomes confirmed statistically significant differences among any two models in all measurements and dataset sizes ($p < 0.05$ in each case). The interesting part here is that GBR performed better than SVR and RFR in nearly every cluster and the aggregations in general, having fewer average errors. For instance, in the 1000-sample dataset, the lowest MAE was achieved by GBR at 0.4595, and the lowest RMSE was also achieved at 0.5961, both of which belong to cluster 3, thus demonstrating the model's good local adjustability.

The robustness of GBR on complete and clustered datasets proves its suitability for PV power forecasting tasks, especially when combined with unsupervised clustering techniques such as DPGMM. The results validate the effectiveness of hybrid approaches that combine data decomposition and ensemble learning to enhance forecasting performance. These findings highlight the importance of model selection and data stratification in renewable energy forecasting, particularly for devices with nonlinear and stochastic behavior, such as residential photovoltaic generation.

This hybrid algorithm enabled more precise analysis of the results, as well as greater efficiency and prediction accuracy, when applied to the original dataset. The effect of the number of data points per cluster was tested using the hybrid algorithm. Random subsets of 500, 600, 700, 800, 900, and 1000 data points were sampled from the clusters.

In the five selected clusters, all error measures under consideration increased when 500 random data points were tested within each cluster, as shown in Figure 4. Column 1 in the figure represents the output from using the original dataset, whereas the other columns show the outputs from applying ML algorithms to the five selected clusters. For the MAE error measure, decreases varied from a minimum of 12% in Cluster 6 to a maximum of 50% in Cluster 4. Similarly, for the RMSE error measure, decreases varied from 40% in Cluster 6 to 76% in Cluster 4. For the MSE measure, error decreases of at least 22% were found in Cluster 6 and up to 60% in Cluster 4.

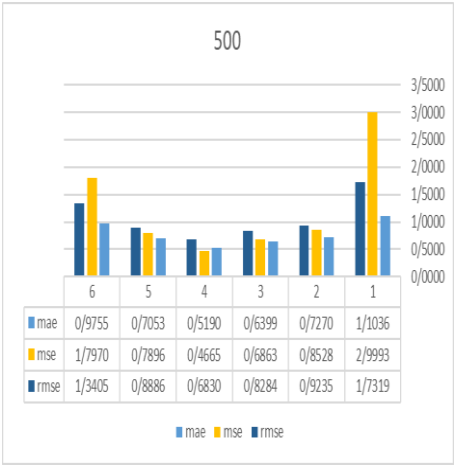

Figure 4. A Comparative Analysis of Clustered and General Data Across 500 Random Data Sets

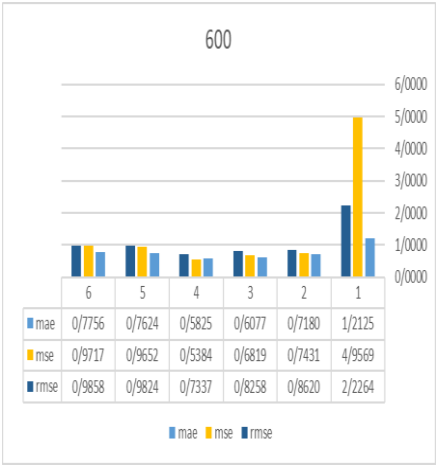

Figure 5. A Comparative Analysis of Clustered and General Data Across 600 Random Data

This was then compared for a random set of 600 points, and the results, as shown in Figure 5, consistently demonstrated improved performance using the hybrid algorithm. All clusters displayed similar error reduction patterns. These were also compwere randomly selected from sets of 700, 800, 900, and ultimately 1000 points, as illustrated in Figure 6, Figure **7**, **Figure 8**, and Figure **9** respectively.

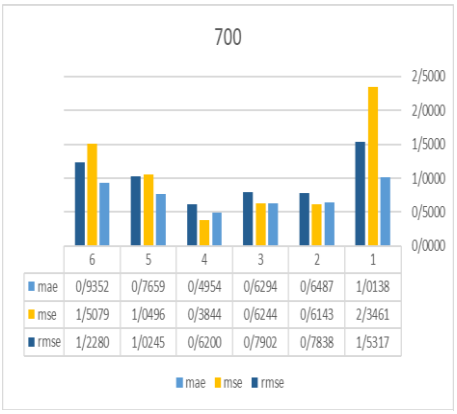

Figure 6. A Comparative Analysis of Clustered and General Data Across 700 Random Data Sets

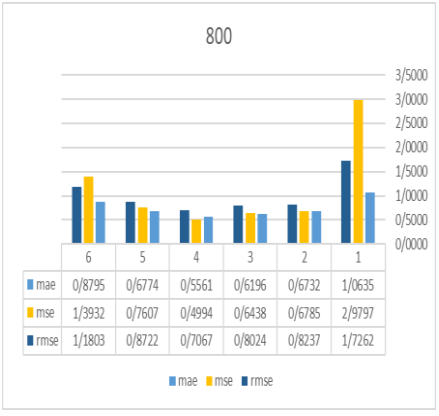

Figure 7. A Comparative Analysis of Clustered and General Data Across 800 Random Data Sets

For a random sample of 800 data points, reductions in all three analyzed error metrics ranged from 45% to 13% for all clusters, as shown in Figure 7. This reflects the high contribution of the model employed in error

minimization. When the data points were 900 (**Figure 8**), reductions in the MAE measure were between approximately 13% and 56%, while for the MSE measure, reductions were between 35% and 80%.

Lastly, with the largest dataset of 1000 data points, the results in Figure 9 show reductions in the MAE metric to the range between 13% to 56%, with 46% from Cluster 6, and 93% from Cluster 4, respectively. The RMSE metric shows values reduced from 15% to 71%, revealing the power of the hybrid model.

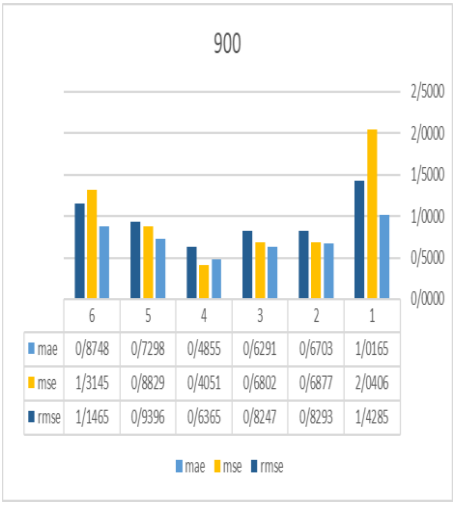

Figure 8. A Comparative Analysis of Clustered and General Data Across 900 Random Data Sets

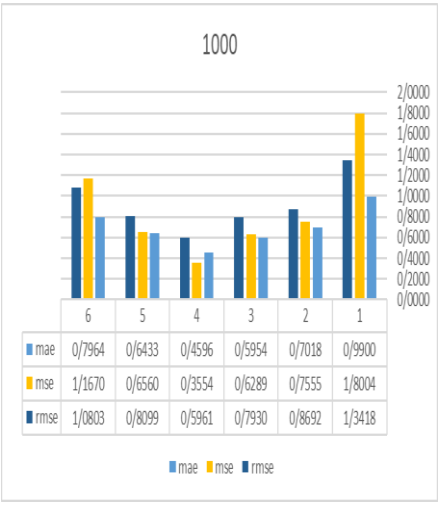

Figure 9. A Comparative Analysis of Clustered and General Data Across 1000 Random Data Sets

The hybrid model produced substantial reductions in prediction errors across all experiments, as evidenced by marked decreases in MAE, RMSE, and MSE, confirming its strong capability to simulate photovoltaic power generation and yielding results that are both satisfactory and practically meaningful. Partitioning the data into homogeneous clusters played a pivotal role in reducing errors because it allowed the model to focus on localized patterns within each group, rather than being confounded by the variability of the entire dataset. This pre-processing reduced noise and outliers, providing supervised learners with more consistent and informative training data and thereby enhancing predictive performance.

The positive effect of clustering on gradient boosting regression is particularly notable. As a gradient-boosting method, GBR is highly sensitive to input data quality. When trained on cluster-specific data, it benefits from reduced complexity and fewer irrelevant fluctuations, which facilitates the detection of nonlinear relationships and accelerates convergence through more predictable residuals. Consequently, the iterative residual-correction mechanism of GBR is more effective within homogeneous clusters, and the

combination of clustering and GBR demonstrates a clear complementary relationship that amplifies the strengths of each approach.

Empirical comparisons across clusters revealed substantial heterogeneity in performance gains. For instance, Cluster 4 exhibited the most significant improvements, with reductions in MAE of up to 93% and RMSE reductions of up to 76%, likely reflecting greater homogeneity in environmental variables such as irradiance and temperature, which allowed GBR to identify patterns more readily. Conversely, more variable clusters, exemplified by Cluster 6, showed smaller gains. These findings underscore the central importance of data properties in determining model efficacy and suggest that the hybrid strategy, by organizing data and leveraging the predictive power of GBR, offers a robust and practical solution for improving photovoltaic power forecasts. At the same time, its generalizability remains contingent on the underlying characteristics of the data.

The ability of the hybrid model to reduce prediction errors significantly suggests its potential for learning valuable information from diverse subsets of data. Performance differences among clusters emphasize the importance of data structure and quality for achieving maximum predictive model accuracy. This highlights the hybrid approach as a strong and scalable method of improving predictions of photovoltaic power generation.

The behavior of the hybrid model was further probed by measuring the variation of prediction errors (MAE, MSE, and RMSE) with a growth in the number of data points from 500 to 1000. The results, shown in figures 11, 12, 13, 14, 15 and 16, represent the trend of error measures across the five selected clusters and the primary dataset.

In Cluster 1 (figures 11), the error measures have a clear pattern with a sharp minimum point at 700 data points. In this, the lowest points of the errors of MAE, MSE, and RMSE are experienced, indicating the optimal data quantity vs. model performance balance.

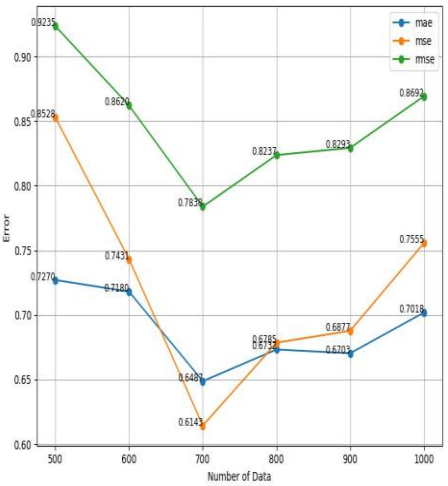

Figure 10. Errors for GBR with DBGMM1

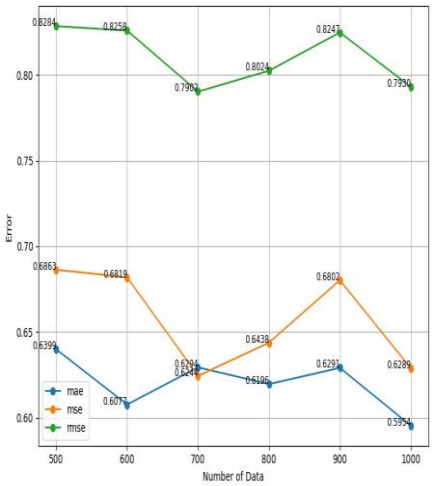

Figure 11. Error for GBR with DBGMM2

In Cluster 2 (figures 12), all three error measures decrease uniformly as the number of data points increases. The reduction is linear, showing the benefits of higher data samples in making predictions more accurate without drastic swings.

In Cluster 3 (figures 13), while there is a general trend of reducing error metrics as data increases, fluctuations are found at specific data sizes. Such a trend indicates varying data quality or distribution, but the model has an overall trend towards improvement in accuracy.

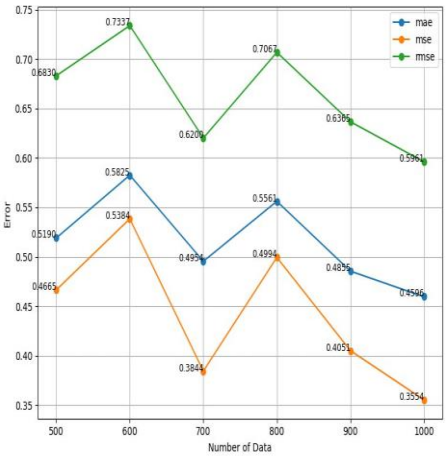

Figure 12. Errors for GBR with DBGMM3

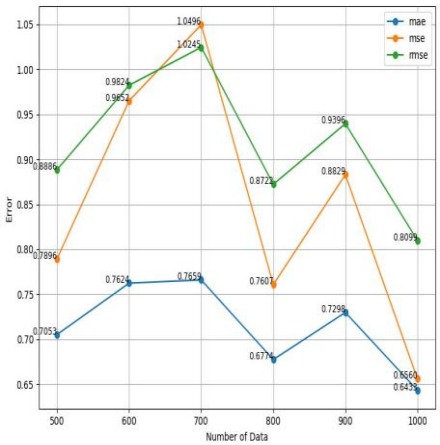

Figure 13. Errors for GBR with DBGMM4

Cluster 4 exhibits the same behavior as Cluster 3 (figures 14), but with errors declining and a slightly increased fluctuation. Despite the fluctuation, overall, the trend is positive and reflects a clear improvement in precision with larger sets.

Cluster 5 demonstrates a steady decrease in error metrics with increasing data points (figures 15). The relative minimum is observed at 600 data points, and the downward trend continues at a modest rate, indicating consistent improvement in prediction performance.

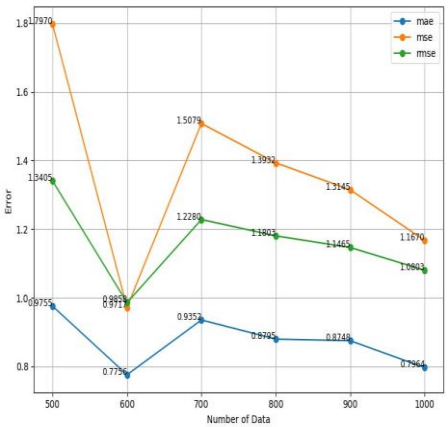

Figure 14. Errors for GBR with DBGMM5

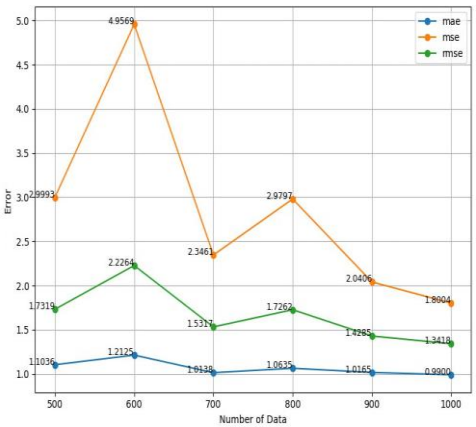

Figure 15. Errors for GBR with DBGMM6

As data size increases, the clustering with 1,000 data points demonstrates a uniform and even decrease across all three indicators of error (figures 16). In contrast to the rest of the clusters, the decrease is more consistent and lacks significant fluctuations, which suggests a more uniform and expansive data distribution.

Error Reductions: Among all the measures that were taken into consideration, the maximum reduction was observed in the MSE. The actual value of this reduction is provided in Table 4**Error! Not a valid bookmark self-reference.**.

Table 4. Initial MSE and Final MSE

Cluster	Initial MSE	Final MSE
Cluster 1	0.85	0.61
Cluster 2	0.68	0.62
Cluster 3	0.54	0.38
Cluster 4	1.05	0.65
Cluster 5	1.8	0.97
Cluster 6	4.97	1.8

The hybrid model revealed that, overall, predictive performance is enhanced by increasing the size of the training data, and when the quantity of random samples was doubled to 1000 from 500, MAE, MSE, and RMSE were greatly improved in every cluster and the overall dataset. The overall dataset revealed a more consistent reduction in error, which indicates that a large and varied dataset enables more consistent prediction. Cluster behavior was not homogeneous, however: two clusters exhibited a monotonically decreasing error with increasing sample sizes, reflecting that additional data reduced variance and noise in learning, whereas two other clusters exhibited oscillations that reflect that the usefulness of additional data depends on factors such as homogeneity and the presence of outliers. One cluster reached a minimum error at a point of approximately 700 data points, which likely represents a point of diminishing returns, while other clusters were still benefiting from further data increases. Such trends embody a trade-off between predictive accuracy and computational cost since larger datasets take longer to process and require more resources and may not lead to proportionate increases in accuracy. Therefore, finding an optimal data size for each cluster based on its intrinsic properties might be an effective means to achieve high accuracy with restricted computational expenses. From an application perspective, the reduction of errors in photovoltaic power prediction improves grid operations by enabling more accurate supply–demand balancing, reducing the requirement for nonrenewable backup capacity, facilitating energy storage

design, and supporting operational planning, which is critical to enhancing the sustainability and resilience of renewable energy infrastructure. The results may, however, be biased by the specific nature of data collected, including environmental conditions and geographical location, that limits generalizability. Future research needs to apply the algorithm to more diverse, geographically scattered data sets and across longer time frames in order to further establish the model's robustness and flexibility across varying conditions.

4) Conclusion

This research demonstrated the effectiveness of a hybrid algorithm in estimating photovoltaic power generation with high accuracy and low errors (MAE, RMSE, and MSE). The model achieved optimal gains in performance by leveraging both supervised and unsupervised approaches, particularly in clustered data sets. These findings showcase the model's capability to improve the management of renewable energy systems more effectively. Despite superior potential results, the research is limited by the scarcity of data and environmental circumstances. The future lies in using the model with larger datasets, other geographical locations, and longer intervals of time to test its validity. The hybrid technique generally offers a computationally practical and worthy approach for application in real scenarios in renewable energy prediction.

5) Acknowledgments

We would like to give special thanks to all the participants in this study.

References

[1] El Hendouzi, A., Bourouhou, A., & Ansari, O. (2020). The Importance of Distance between Photovoltaic Power Stations for Clear Accuracy of Short-Term Photovoltaic Power Forecasting. *Journal of Electrical and Computer Engineering*, *2020*. https://doi.org/10.1155/2020/9586707

[2] Park, T., Song, K., Jeong, J., & Kim, H. (2023). Convolutional Autoencoder-Based Anomaly Detection for Photovoltaic Power Forecasting of Virtual Power Plants. *16*(14).

[3] Phan, Q. T., Wu, Y. K., Phan, Q. D., & Lo, H. Y. (2023). A Novel Forecasting Model for Solar Power Generation by a Deep Learning Framework With Data Preprocessing and Postprocessing. *IEEE Transactions on Industry Applications*, *59*(1), 220–231. https://doi.org/10.1109/TIA.2022.3212999

[4] Huang, H., Zhu, Q. L., Zhu, X., & Zhang, J. (2023). An Adaptive, Data-Driven Stacking Ensemble Learning Framework for the Short-Term Forecasting of Renewable Energy Generation. *Energies*. https://doi.org/10.3390/en16041963

[5] Elsaraiti, M., & Merabet, A. (2022). Solar Power Forecasting Using Deep Learning Techniques. *IEEE Access*, *10*, 31692–31698. https://doi.org/10.1109/ACCESS.2022.3160484

[6] Mofidian, R., Hassankhani, I., Jahanshahi, M., Hosseini, S. S., & Miansari, M. (2024). Cost Effective Design of a 200 kW On-grid Rooftop Photovoltaic System Using

PVsyst Software in Shiraz. *Journal of Engineering and Applied Research*, *1*(1), 13–24. https://doi.org/10.48301/jear.2024.194109

[7] Taghandiki, K. (2024). Quantum Machine Learning Unveiled: A Comprehensive Review. *Journal of Engineering and Applied Research*, *1*(2), 29–48. https://doi.org/10.48301/jear.2024.446673.1021

[8] Wang, S., Yan, S., Li, H., Zhang, T., Jiang, W., Yang, B., Li, Q., Li, M., Zhang, N., & Wang, J. (2024). Short-term prediction of photovoltaic power based on quadratic decomposition and residual correction. *Electric Power Systems Research*, *236*, 110968–110968. https://doi.org/10.1016/j.epsr.2024.110968

[9] Qiu, L., Ma, W., Feng, X., Dai, J., Dong, Y., Duan, J., & Chen, B. (2024). A hybrid PV cluster power prediction model using BLS with GMCC and error correction via RVM considering an improved statistical upscaling technique. *Applied Energy*, *359*, 122719–122719. https://doi.org/10.1016/j.apenergy.2024.122719

[10] Hu, W., Zhang, X., Zhu, L., & Li, Z. (2021). Short-Term Photovoltaic Power Prediction Based on Similar Days and Improved SOA-DBN Model. *IEEE Access*, *9*, 1958–1971. https://doi.org/10.1109/ACCESS.2020.3046754

[11] Zhu, R., Li, T., & Tang, B. (2024). Research on short-term photovoltaic power generation forecasting model based on multi-strategy improved squirrel search algorithm and support vector machine. *Scientific Reports*, *14*(1), 14348–14348. https://doi.org/10.1038/s41598-024-65159-1

[12] Akoglu, H. (2018). User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, *18*(3), 91–93. https://doi.org/10.1016/j.tjem.2018.08.001 https://doi.org/10.1016/j.tjem.2018.08.001

[13] Tao, Z., Li, J., Fu, H., Kong, Y., & Fu, Y. (2021). From ensemble clustering to subspace clustering: Cluster structure encoding. *IEEE Transactions on Neural Networks and Learning Systems*, *34*(5), 2670–2681. https://doi.org/10.1109/TNNLS.2021.3107354

[14] Persson, C., Bacher, P., Shiga, T., & Madsen, H. (2017). Multi-site solar power forecasting using gradient boosted regression trees. *Solar Energy*, *150*, 423–436. https://doi.org/10.1016/j.solener.2017.04.066

[15] Ridgeway, C. L., & Nakagawa, S. (2017). Is Deference the Price of Being Seen as Reasonable? How Status Hierarchies Incentivize Acceptance of Low Status. *Social Psychology Quarterly*, *80*(2), 132–152. https://doi.org/10.1177/0190272517695213

[16] Liu, D., & Sun, K. (2019). Random forest solar power forecast based on classification optimization. *Energy*, *187*, 115940–115940. https://doi.org/10.1016/j.energy.2019.115940

[17] Borup, D., Christensen, B. J., Mühlbach, N. S., & Nielsen, M. S. (2023). Targeting predictors in random forest regression. *International Journal of Forecasting*, *39*(2), 841–868. https://doi.org/10.1016/j.ijforecast.2022.02.010

[18] Ahmed, S. I., Bhuiyan, K. A., Rahman, I., Salehfar, H., & Selvaraj, D. F. (2024). Reliability of regression based hybrid machine learning models for the prediction of solar photovoltaics power generation. *Energy Reports*, *12*, 5009–5023. https://doi.org/10.1016/J.EGYR.2024.10.060

[19] Awad, M., & Khanna, R. (2015). Support Vector Regression BT - Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers. In M. Awad & R. Khanna (Eds.), (pp. 67–80). Apress. https://doi.org/10.1007/978-1-4302-5990-9_4