ARTICLE | Article in Press

Machine learning estimation of prosumer net metering using optimal polynomial basis functions for dimensionality reduction

Onder Tutsoy · Kasim Zor · Ozgur Celik

Received: 10 July 2026 / Accepted: 7 September 2026
© The Author(s) 2026

Abstract

Solar energy has become an essential renewable resource for residential prosumers engaging in bidirectional electricity exchange with the grid. However, accurately estimating net consumption remains a challenge due to the prevalence of limited yet redundant local Photovoltaic (PV) data. This paper designs an input dimension reduction approach using optimal polynomial basis functions to enhance machine learning estimation. Optimal higher-order powers are determined via an L_2-norm performance function based on matrix conditioning. Additionally, Pearson correlation and the Least Absolute Shrinkage and Selection Operator (LASSO) are employed to eliminate redundancies and quantify key input–output relationships. The refined data is used to train Batch Least Squares (BLS), Neural Network (NN), and Artificial Bee Colony (ABC) algorithms. Results demonstrate that the BLS model, integrated with Pearson-based feature selection, achieves superior precision with a Mean Absolute Error (MAE) of 0.0037 and a Root Mean Square Error (RMSE) of 0.0048. Validation on unseen test data confirms the model's robustness, maintaining a low MAE of 0.0036 and RMSE of 0.0046.

Keywords Estimation · Machine learning · Photovoltaic · Rooftop · Solar energy

Introduction

The reliance on conventional fossil fuels for electricity generation not only imposes a significant economic burden due to high costs but also accelerates greenhouse gas emissions and irreversible environmental degradation[1]. Consequently, renewable energy sources such as hydropower and wind energy have gained prominence; however, their inherent intermittency and dependence on fluctuating weather conditions often challenge their reliability[2]. Among these, solar energy has emerged as a particularly sustainable, clean, and increasingly cost-effective alternative[3]. Over the past decade, solar investments have surged, especially in energy-dependent and sun-rich regions[4]. To enhance the integration of solar power, governments are incentivizing residential prosumers to deploy rooftop photovoltaic (PV) systems[5]. Nevertheless, balancing local production with demand remains a critical challenge due to the high uncertainty and inaccuracy of solar generation estimates[6]. To address these

This Article in Press is shared early to give you faster access to new research. It is citable and carries a permanent DOI. The final edited version will replace it automatically.

https://doi.org/10.1038/s41598-026-71068-2

Published online: 02 October 2026

scientific reports

issues, this paper develops advanced, robust, and computationally efficient machine learning models designed to accurately estimate the net energy exchange between prosumers and the grid.

Various machine learning techniques, including fuzzy logic-based models, have been developed to estimate prosumer net electricity consumption. An expert knowledge-based fuzzy logic model was proposed to estimate PV power generation, utilizing the K-Nearest Neighbor (K-NN) algorithm to impute missing input data[7]. While this approach yielded a 7% estimation error on a small-scale dataset, the study lacked a detailed analysis of input data properties and the specific impact of the K-NN-based imputation. Similarly, a neutrosophic fuzzy logic approach was integrated with Neural Networks (NN) for short-term solar energy estimation by incorporating external variables such as sunshine duration, humidity, and temperature[8]. High-dimensional input data can still increase the risk of overfitting and larger errors on unseen testing data[9]. Furthermore, a fuzzy logic-based reconfiguration method was introduced for PV systems under partial shading, achieving error reductions between 4.28% and 10% depending on the current and voltage input configurations[10]. While integrating expert knowledge can accelerate the learning process, these approaches often struggle with high-dimensional input spaces and the management of stochastic uncertainties.

Classification-based machine learning models have also been extensively explored in the literature. A variety of algorithms including Gradient Boosting (GBoost) and Random Forest (RF) were implemented to estimate solar radiation by utilizing maximum and minimum temperatures, relative humidity, precipitation and wind speed as inputs[11]. The study indicated that these algorithms produce comparable and promising estimates on the training dataset. Similarly, the relationship between environmental factors such as panel temperature, ambient temperature, solar flux and humidity and PV power generation was investigated using Support Vector Machines (SVM) and Gaussian Process Regression (GPR)[12]. Under these configurations, the GPR and SVM models achieved RMSE values of 7.967 and 21.72, respectively, when utilizing the complete input dataset. Furthermore, the estimation of PV generation in solar-rich regions was examined using polynomial regression, Decision Trees (DT), SVM, RF and Long Short-Term Memory (LSTM) networks[13]. By incorporating global and diffuse horizontal radiation, the study highlighted the importance of data normalization and reported that the RF model outperformed the other techniques.

With respect to machine learning algorithms integrated with input dimension reduction, a Bayesian bidirectional Long Short-Term Memory (BiLSTM) network was applied for estimation alongside a Variational Auto-Encoder (VAE) for dimensionality reduction[14]. This approach reduced model parameters by 97.3% and computational time by 37.9%. Similarly, NN were combined with metaheuristic algorithms to estimate solar energy under seasonal and cloud cover uncertainties by utilizing Principal Component Analysis (PCA) for dimensionality reduction[15]. The resulting training RMSE values were reported as 12.06, 6.72 and 9.49 for the winter, summer and monsoon seasons, respectively. Furthermore, a Supervised Local Maximum Variance Preserving (SLMVP) approach was proposed for solar data reduction and compared against PCA, Linear Optimal Low-Rank (LOL) and Locality Preserving Projection (LPP) methods[16]. Experimental results across two distinct locations verified that SLMVP outperformed the other techniques, with a 3.12% MAE improvement at Lisbon. Experimental results across two distinct locations verified that SLMVP outperformed other techniques with a 3.12% MAE. In a different approach, a Statistical Machine Learning (SML) model was employed to conduct probabilistic solar power estimation using Singular Value Decomposition (SVD) for dimensionality reduction under uncertainty[17]. However, this study focused on a multi-scenario-based assessment for energy consumption planning rather than providing a direct and accurate estimation of solar energy production.

A review of deep learning methods for solar photovoltaic power output forecasting compared LSTM, Convolutional Neural Network (CNN) and transformer networks and noted that the reported models still rely on meteorological inputs[18]. Next-day solar irradiance and PV generation at the Bhadla Solar Park were also forecasted by comparing Gaussian and Weibull clustering of Markov scenarios, where the Weibull distribution yielded smaller MAE and RMSE[19]. In a microgrid setting, a learning-enhanced stochastic model predictive controller was developed with trustworthy forecasting for multi-energy operation[20], while a physics-informed machine

learning framework was utilized for climate-aware digital twins of decentralised energy systems, with Extreme Gradient Boosting (XGBoost) used for photovoltaic and wind power forecasting[21]. Search-based metaheuristics such as particle swarm optimization have been employed to tune unknown parameters, and the same literature has noted that the Artificial Bee Colony (ABC) search is prone to local optima and a slower convergence[22]. Neural networks have also been utilized as Nonlinear Autoregressive with Exogenous inputs (NARX) models for input–output mapping[23]. However, these studies focused on plant-level PV forecasting, microgrid energy management or actuator modelling rather than providing a direct estimation of household net-metering from redundant local PV measurements.

Even though these dimensionality reduction approaches have been successfully integrated with machine learning algorithms to achieve high estimation accuracies, the combined role of a matrix-conditioned polynomial basis, Pearson and LASSO reduction, and three learners with different operating characters (batch gradient BLS, iterative gradient NN, and search-based ABC) has not yet been examined for household net-metering from local PV measurements. To address these gaps, the major contributions of this paper are as follows:

1. **Construction of a polynomial basis function framework:** A polynomial basis function is constructed to expand the unknown parameter space, where an L_2-norm performance function optimizes the polynomial powers. These optimal powers are determined based on maximizing the information content of the input data.
2. **Redundancy elimination via Pearson correlation and LASSO:** Statistical Pearson correlation and L_1-norm parameter fitting-based LASSO approaches are utilized to quantify the relationship between the raw input dataset and the target output. The results verify that a 5-dimensional Pearson-selected input subset, taken from the original 14-dimensional dataset, yields the lowest BLS MAE of 0.0037 on training (Table 3) and 0.0036 on testing (Table 4). A 3-dimensional subset remains usable (BLS test MAE 0.0088 in Table 4) but is not the best set.
3. **Comparative analysis of diverse machine learning architectures:** The study applies batch and gradient-driven BLS, iterative and gradient-driven NN, and iterative search-based ABC algorithms to analyze the performance of machine learning models with distinct operational characteristics.
4. **Robustness validation through unseen data:** The developed models are evaluated not only on training performance but also on testing performance using unseen data to ensure robustness for real-time applications.

In the rest of the paper, Section II introduces the training data and prediction model basis, Section III provides the raw input data dimension reduction approaches, Section IV reviews the machine learning algorithms, Section V analyses the learning results and finally Section VI discusses the major findings of the paper and states the future works.

Training data and prediction basis construction

This section first presents the characteristics of the input–output training data, followed by the implementation of data pre-processing techniques. Subsequently, a polynomial basis function framework is constructed, detailing the optimization of its unknown polynomial powers to enhance the feature space.

Input–output training data

Table 1 presents the 14-dimensional raw input data obtained from Pecan Street through real-time measurements[24]. The records have 8619 samples in total. The data are split into training samples and testing samples[24].

As observed in Table 1, the input features consist entirely of internal PV measurements. Consequently, the developed estimation models eliminate the need for external meteorological data, such as humidity or wind speed, thereby reducing sensor dependency. The corresponding target output for the machine learning models is

Table 1 Training and testing input data.

Data	Definition	Data	Definition
Solar	Generated power	MoY	Month of year
Leg1V	Phase A voltage	DoM	Day of month
Leg2V	Phase B voltage	DoW	Day of week
Angle1	Phase A angle	HoD	Hour of day
Angle2	Phase B angle	QoH	Quarter of hour
Current1	Phase A current	Lag15m	Lagged 15 min
Current2	Phase B current	Lag1h	Lagged 1 h

the "grid" variable, which represents the net power either drawn from or injected into the utility grid. The subsequent section details the data pre-processing techniques employed in this study.

Data pre-processing approaches

Initially since the raw input data $\mathbf{x}^i$ comes from different sources with different magnitudes, min–max normalized $\mathbf{x}^n$ data is formed as

$$\mathbf{x}^n = \frac{\mathbf{x}^i - \min(\mathbf{x}^i)}{\max(\mathbf{x}^i) - \min(\mathbf{x}^i)} \tag{1}$$

In addition, as the input data has continuous nature, the re-sampled data $\mathbf{x}$ is obtained as

$$\mathbf{x}_k = \mathbf{x}(t)\,\delta(t - kt_s) \tag{2}$$

where t is the data time index, t_s is the sampling time, k is the sampled data index and δ is the impulse signal. Next sub-section introduces the parametrized prediction model.

Prediction model construction

The prediction model produces an estimated output $\hat{y}_k$ formed as

$$\hat{y}_k = \mathbf{w}_k^T \mathbf{b}_k \tag{3}$$

where $\mathbf{b}_k$ is the polynomial basis function vector of the raw input data and $\mathbf{w}^i$ is the corresponding unknown parameter vector. The polynomial basis function vector is structured as

$$\mathbf{b}_k = \begin{bmatrix} \mathbf{1} & |\mathbf{x}_k|^{p^1} & |\mathbf{x}_k|^{p^2} \end{bmatrix} \tag{4}$$

where $\mathbf{1}$ is the bias vector, $|\mathbf{x}_k|$ is absolute value of the input data vector, p^1 and p^2 are the unknown higher-order polynomial powers which are optimally determined with Algorithm 1, provided next.

Polynomial basis function power optimization

To optimally determine the unknown higher-order powers of the polynomial basis function, consider the L_2-norm performance function J built as

$$J_k = \|y_k - \hat{y}_k\|_2^2 \tag{5}$$

where y_k is the real output. Substituting Eq. (3) in Eq. (5) and expanding it gives

$$J_k = y_k^T y_k - \mathbf{w}_k \mathbf{b}_k^T y_k - y_k \mathbf{w}_k^T \mathbf{b}_k + \mathbf{w}_k \mathbf{b}_k^T \mathbf{w}_k^T \mathbf{b}_k \tag{6}$$

Taking the derivative of Eq. (6) with respect to the unknown parameter vector $\mathbf{w}$ results in

$$\frac{\partial J_k}{\partial \mathbf{w}_k} = -2\mathbf{b}_k^T y_k + 2\mathbf{b}_k^T \mathbf{b}_k \mathbf{w}_k \tag{7}$$

Imposing the stationary condition on Eq. (7) through equating it to zero yields

$$\mathbf{w}_k = \left(\mathbf{b}_k^T \mathbf{b}_k\right)^{-1} \mathbf{b}_k^T y_k \tag{8}$$

The transformed basis $\mathbf{b}_k^t$ is

$$\mathbf{b}_k^t = \left(\mathbf{b}_k^T \mathbf{b}_k\right)^{-1} \mathbf{b}_k^T \tag{9}$$

Condition of the transformed basis in Eq. (9) quantifies the richness of the information for the learning problem. Its condition c is defined as

$$c_k = \left\|\mathbf{b}_k^t\right\|_2 . \left\|\mathbf{b}_k^{t^{-1}}\right\|_2 \tag{10}$$

It is important to note that since both the transformed basis in Eq. (9) and the matrix condition in Eq. (10) necessitate the matrix inversion, row and column reduction based matrix singularity must be hindered. Therefore, a random noise σ_k with small magnitude is added in one of the polynomial terms in Eq. (4) as

$$\mathbf{b}_k = \begin{bmatrix} \mathbf{1} & |\mathbf{x}_k|^{p^1} & |\mathbf{x}_k + \sigma_k|^{p^2} \end{bmatrix} \tag{11}$$

Algorithm 1 presents the conditioning based unknown power optimization for the polynomial basis function.

Input:

Test data $\mathbf{x} \in \mathbf{R}^{N \times M}$, output $\mathbf{y} \in \mathbf{R}^{N \times 1}$, initial condition c^i, p^1 and p^2 candidates.

Output:

Optimal p_o^1 and optimal p_o^2

Main Loop:

for i^th power $p^1(i)$

for j^th power $p^2(j)$

1.Re-shape the basis $\mathbf{b}_k^{ij}$ in Equation (4) as,

$$\mathbf{b}_k^{ij} = \mathbf{b}_k = \left[\mathbf{1} \quad \left|\mathbf{x}_k\right|^{p^1(i)} \quad \left|\mathbf{x}_k\right|^{p^2(j)}\right] \tag{12}$$

2.Form the transformed basis $\mathbf{b}_k^{i,t}$ in Equation (9) as,

$$\mathbf{b}_k^{ij,t} = \left(\mathbf{b}_k^{ij^T} \mathbf{b}_k^{ij}\right)^{-1} \mathbf{b}_k^{ij^T} \tag{13}$$

3. Determine the condition c_k^{ij} as

$$c_k^{ij} = \left\|\mathbf{b}_k^{ij,t}\right\|_2 . \left\|\mathbf{b}_k^{ij,t^{-1}}\right\|_2 \tag{14}$$

4.Compare the conditions for the optimal powers as

if $c_k^{ij} < c^i$

$c^i = c_k^{ij}$, $p_o^1 = p^1(i)$ and $p_o^2 = p^2(j)$

end if

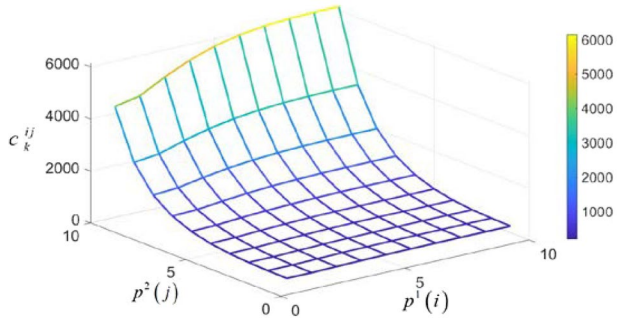
end for i^th power $p^1(i)$

end for j^th power $p^2(j)$

Algorithm 1 Polynomial basis function power optimization.

Figure 1 shows the polynomial power optimization surface obtained with Algorithm 1.

Fig. 1 Polynomial power optimization surface.



As can be seen from Fig. 1, the polynomial power surface is smooth and its unique minimum occurs at $p^1 = 1$ and $p^2 = 2$. Henceforth, the optimal polynomial basis is

$$\mathbf{b}_k = \begin{bmatrix} \mathbf{1} & |\mathbf{x}_k|^1 & |\mathbf{x}_k|^2 \end{bmatrix} \tag{15}$$

The next sub-section presents the dimension reduction approaches for the raw input data.

Raw input data dimension reduction

This section reviews the Pearson correlation and LASSO approaches utilized for the raw input data dimension reduction.

Pearson correlation-based dimension reduction

Dimension reduction approaches are designed to select the most informative input features to optimize machine learning performance. These techniques primarily focus on:

- **Mitigating high dimensionality:** Reducing the complexity of the raw input space to enhance model generalizability.
- **Optimizing computational efficiency:** Decreasing both the processing time and the overall computational burden during training and inference.
- **Enhancing numerical stability:** Preventing singularities and ill-conditioned matrices associated with rank deficiency in raw datasets.

Pearson correlation $C_{1,y}$ identifies the relationship between the input datasets $\mathbf{x}^1$ and the target output $\mathbf{y}$ as

$$C_{1,y} = \sum_{i=1}^{N} \left(\mathbf{x}_i^1 - \overline{x}^1\right)\left(\mathbf{y}_i - \overline{y}\right) / \sqrt{\sum_{i=1}^{N} \left(\mathbf{x}_i^1 - \overline{x}^1\right)^2 \left(\mathbf{y}_i - \overline{y}\right)^2} \tag{16}$$

where $\overline{x}^1$ and $\overline{y}$ are the mean values of the $\mathbf{x}^1$ and $\mathbf{y}$ input and target datasets, respectively. The corresponding mean, for example $\overline{x}^1$ with data length of N, is determined as

$$\overline{x}^1 = \left(\sum_{i=1}^{N} \mathbf{x}_i^1\right) / N \tag{17}$$

Algorithm 2 presents the Pearson correlation-based dimensionality reduction.

Input:
M number of raw input dataset $\mathbf{x}$ and target output $\mathbf{y}$,
Output:
Correlation matrix $\mathbf{C}$, correlated pair indices $\mathbf{I}^c$,
Main Loop:
1. Determine the mean $\overline{y}$ with Equation (17),
for $i = 2$ ***to*** M
2.Calculate the mean $\overline{x}^i$ with Equation (17),
3.Calculate the correlation $\mathbf{C}_{i,y}$ with Equation (16),
end for i
4.Sort the $\mathbf{C}$ from maximum to minimum value,
5.Assign the corresponding sorted pair indices $\mathbf{I}^c$.

Algorithm 2 Dimension reduction with Pearson correlation.

Next sub-section formulates the LASSO-based dimension reduction.

LASSO-based dimension reduction

The LASSO performs parameter optimization and regularization to prevent overfitting through imposing L_1-norm penalty on the parameter vector $\mathbf{w}_k$ as

$$\min\left(\left\|y_k - \mathbf{w}_k^T \mathbf{x}_k\right\|_2^2 + \lambda \left\|\mathbf{w}_k\right\|_1\right) \tag{18}$$

where λ is the Lagrange multiplier or the costate which regularizes the parameter $\mathbf{w}_k$. It is optimally selected as

$$\lambda = \max\left(\left|\sum_{i=1}^{N}\left(\mathbf{x}_i^1 - \overline{x}^1\right) .* \mathbf{w}_i .* \left(\mathbf{y}_i - \overline{y}\right)\right|\right) \tag{19}$$

The corresponding parameter learning law for $\mathbf{w}_k$ is

$$\mathbf{w}_k = \left(\mathbf{x}_k^T \mathbf{x}_k\right)^{-1} \mathbf{x}_k^T y_k \lambda \tag{20}$$

The next sub-section reviews the machine learning algorithms that will optimize the prediction model in Eq. (3).

Machine learning algorithms for optimization

This section reviews the BLS, NN and ABC machine learning algorithms.

Batch least squares (BLS) algorithm

The BLS utilizes the basis function $\mathbf{b}$ including the whole input data samples and the real output $\mathbf{y}$. Since it uses the complete data, detrimental effects of the random uncertainties are lessened. The BLS follows the $\mathbf{w}$ parameter update rule in Eq. (8) and modifies it as

$$\mathbf{w} = \left(\mathbf{b}^T \mathbf{b}\right)^{-1} \mathbf{b}^T \mathbf{y} \tag{21}$$

In contrast to the BLS, the NN, reviewed next, iteratively learns the unknown parameter vector $\mathbf{w}$ from each input–output data sample.

Iterative neural network (NN) algorithm

The NN evolves step by step to learn the unknown parameter vector $\mathbf{w}_k$ and adds up to the gained parameter learning experience. It essentially optimizes the squared error in Eq. (5) as

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \frac{\eta_k}{2}\frac{\partial J_k^2}{\partial J_k}\frac{\partial J_k}{\partial \mathbf{w}_k} \tag{22}$$

where η_k is an adaptive learning rate which is formed with maximum learning rate $\eta^{\max}$ and the basis $\mathbf{b}_k$ as

$$\eta_k = \eta^{\max}/\left(\mathbf{b}_k^T\mathbf{b}_k\right) \tag{23}$$

Solving Eq. (22) yields

$$\mathbf{w}_{k+1} = \mathbf{w}_k + \eta_k e_k \mathbf{b}_k \tag{24}$$

Since BLS and NN are gradient-driven optimization algorithms, their convergence speed and the quality of the resulting optimal solutions are intrinsically linked to the availability and precision of gradient information. Consequently, any redundancies or singularities in the input data can adversely affect the optimization surface. To provide a robust alternative to these gradient-based approaches, the subsequent section develops the search-based ABC machine learning algorithm.

Artificial bee colony (ABC) algorithm

The ABC algorithm iteratively searches for optimal solutions within a defined search space by minimizing the constrained cost function defined in Eq. (5) modified as

$$\begin{aligned} &\text{minimize } J_k = \|y_k - \hat{y}_k\|_2^2 \\ &\text{subject to } \mathbf{w}_k^{\min} \le \mathbf{w}_k \le \mathbf{w}_k^{\max} \end{aligned} \tag{25}$$

where $\mathbf{w}^{\min}$ and $\mathbf{w}^{\max}$ are the minimum and maximum values of the unknown parameter vector, respectively. The parameter update rule is

$$\mathbf{w}_{k+1} = \mathbf{w}_k + \eta^a \bullet \left(\mathbf{w}_k - \mathbf{w}_k^r\right) \tag{26}$$

In Eq. (26), η^a is the vector of the random numbers from continuous uniform distribution on the interval -1 and 1, which acts as an accelerator or learning rate. The parameter vector $\mathbf{w}_k^r$ is the randomly selected parameters among the initial onlookers described in Fig. 2. The parameter constraints are imposed as

$$\begin{aligned} \mathbf{w}_{k+1} &= \max(\mathbf{w}_{k+1}, \mathbf{w}^{\min}) \\ \mathbf{w}_{k+1} &= \min(\mathbf{w}_{k+1}, \mathbf{w}^{\max}) \end{aligned} \tag{27}$$

Figure 2 illustrates the ABC algorithm.

Next sub-section provides the performance measurement criteria that quantify the effectiveness of the machine learning algorithms.

Performance measurement criteria

To evaluate the machine learning prediction accuracies MAE, RMSE and MRAE are utilized. The MAE is represented as

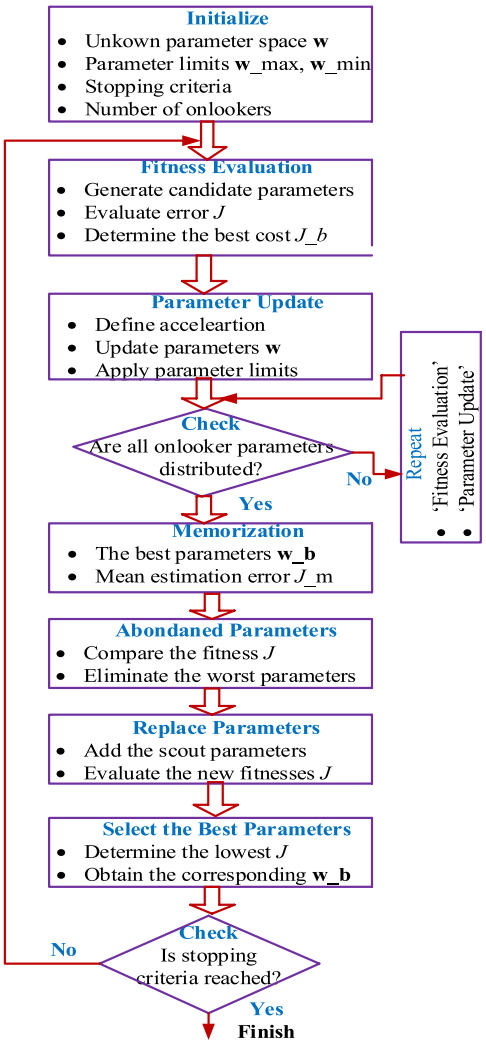


Fig. 2 ABC algorithm for the parameter optimization.

$$MAE = \frac{1}{N}\sum_{k=1}^{N}|y_k - \hat{y}_k| \tag{28}$$

The RMSE is defined as

$$RMSE = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(y_k - \hat{y}_k)^2} \tag{29}$$

Finally, the MRAE is expressed as

$$MRAE = \frac{1}{N}\sum_{k=1}^{N}\left|\frac{y_k - \hat{y}_k}{y_k}\right| \tag{30}$$

Next section analyses the obtained results extensively.

Table 2 Parameters of the machine learning algorithms.

Parameter	Value	Description
p^1 and p^2	1 and 2	Algorithm 1
N and M	8619 and 14	Algorithm 1
η^{max} and η^a	0.9 and 100	Equations 21, 24
$\mathbf{w}^{min}$, $\mathbf{w}^{max}$	-0.5 and 0.5	Equation 23
Onlookers	150	Figure 2
Iterations	100	Figure 2

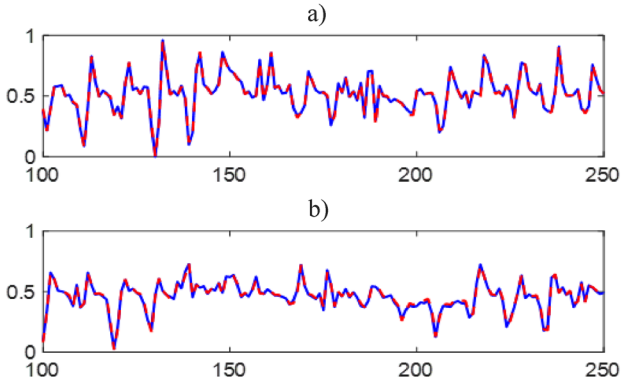


Fig. 3 BLS machine learning algorithm results, (**a**) Training, (**b**) Testing (partial view for peak visibility). The MAE, RMSE and MRAE on the full set are reported in Tables 3 and 4.

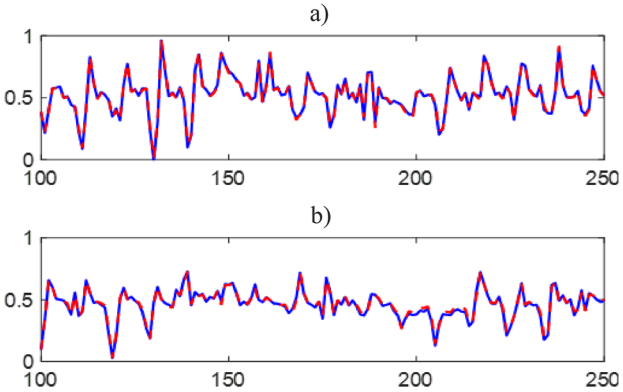


Fig. 4 NN machine learning algorithm results, (**a**) Training, (**b**) Testing (partial view for peak visibility). The MAE, RMSE and MRAE on the full set are reported in Tables 3 and 4.

Results

This section first presents the optimal parameter configurations for the implemented machine learning algorithms and subsequently evaluates their performance using the selected input subsets.

Parameters of the algorithms

Parameters of the machine learning algorithms are illustrated in Table 2.

The next section discusses the major finding of the paper.

Solar estimates with training and testing data

In the following figures, the x-axis represents the data samples, while the solid blue and dashed red lines denote the target and estimated outputs, respectively. Figure 3 shows and compares the estimated and target outputs with the BLS machine learning algorithm.

As can be seen from Fig. 3a, the BLS algorithm is able to learn the corresponding unknown parameters from the Pearson-selected inputs in Table 1. With three features the test MAE is 0.0088 (Table 4). With five features it falls to 0.0036, which is the best BLS result. It is clear that even though the target output data pattern is non-periodic and has random sharp peaks, the BLS algorithm effectively manages to capture the information from the

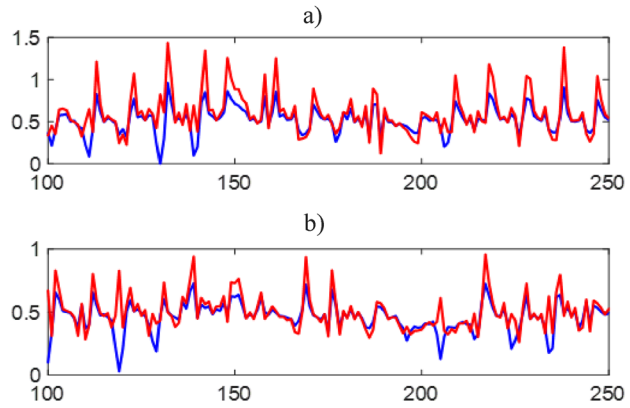


Fig. 5 ABC machine learning algorithm results, (**a**) Training, (**b**) Testing (partial view for peak visibility). The MAE, RMSE and MRAE on the full set are reported in Tables 3, 4, 5 and 6.

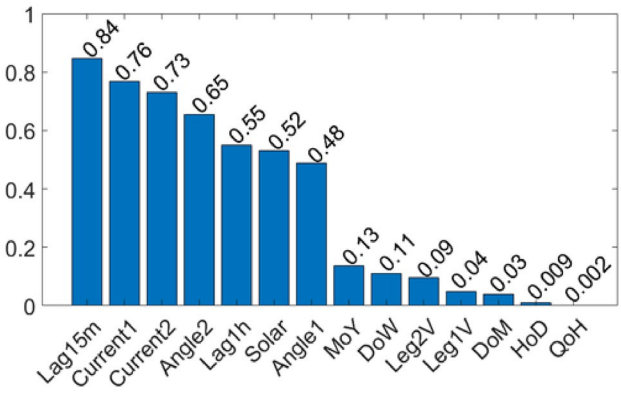


Fig. 6 Quantified input data with Pearson correlation.

input data and learns the associated output data. In order to examine the robustness of the BLS machine learning algorithm, the trained model is tested with the unseen data in Fig. 3b. It is noticeable that the BLS machine learning algorithm is capable of producing precise estimates although the test input data is new to it. Figure 4 illustrates the training and testing estimates with the NN machine learning algorithm.

It is noticeable from Fig. 4 that the training and testing estimates are considerably similar to the ones obtained with the BLS machine learning algorithm. While the BLS utilizes the whole data at a single instant, the NN learns

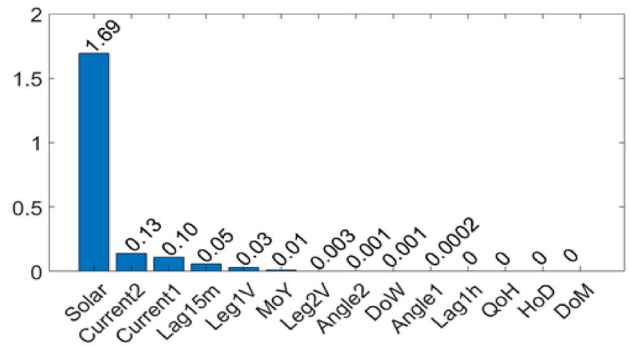


Fig. 7 Quantified input data with LASSO.

Table 3 Training estimates with the Pearson correlation.

Algorithm	Features	MAE	RMSE	MRAE
BLS	3	0.0100	0.0144	0.0295
	4	0.0100	0.0144	0.0294
	5	**0.0037**	**0.0048**	**0.0089**
NN	3	0.0162	0.0251	0.0451
	4	0.0180	0.0280	0.0525
	5	0.0068	0.0090	0.0167
ABC	3	0.1864	0.1941	0.3932
	4	0.1342	0.1868	0.3587
	5	0.1838	0.2043	0.4126

Table 4 Testing estimates with the Pearson correlation.

Algorithm	Features	MAE	RMSE	MRAE
BLS	3	0.0088	0.0127	0.0238
	4	0.0088	0.0126	0.0238
	5	**0.0036**	**0.0046**	**0.0090**
NN	3	0.0122	0.0201	0.0489
	4	0.0141	0.0233	0.0585
	5	0.0074	0.0101	0.0199
ABC	3	0.1902	0.1958	0.4993
	4	0.1128	0.1560	0.4388
	5	0.1574	0.1703	0.4513

Table 5 Training estimates with the LASSO.

Algorithm	Features	MAE	RMSE	MRAE
BLS	3	0.0120	0.0167	0.0248
	4	**0.0080**	0.0111	0.0177
	5	**0.0080**	**0.0109**	**0.0170**
NN	3	0.0130	0.0188	0.0221
	4	0.0091	0.0125	0.0234
	5	0.0093	0.0130	0.0359
ABC	3	0.1862	0.1939	0.3927
	4	0.0901	0.1525	0.0239
	5	0.4447	0.4938	1.0472

iteratively and both of them rely on the gradient information of the data. This proves that the input has sound gradient knowledge that can map to the output data. In contrast to the BLS and NN, the ABC machine learning algorithm learns though a search-based process, whose results are provided in Fig. 5.

Figure 5 demonstrates that there is an obvious bias between the estimated and target outputs even though it managed to learn the overall character of the target output. It is also noticeable that both the training and testing estimates have larger peaks and high frequency fluctuations, which possibly occur due to the aggressive search process. The next sub-section presents the input dataset selection results.

Selected input dataset

The dimension of the input data is 14 and the most informative ones are selected for the learning problem. Figure 6 shows the quantified input data with the Pearson correlation.

As can be seen from Fig. 6, the Pearson correlation determines the Lag15m and QoH as the most and least related input data for the grid power, respectively. It is also clear that the Current1, Current2, Angle2, Lag1h, Solar and Angle1 input data are the further most informative ones according to the Pearson correlation. It is

Table 6 Testing estimates with the LASSO.

Algorithm	Features	MAE	RMSE	MRAE
BLS	3	0.0150	0.0204	0.0381
	4	0.0097	0.0128	0.0255
	5	**0.0093**	**0.0124**	**0.0247**
NN	3	0.0140	0.0195	0.0359
	4	0.0103	0.0143	0.0259
	5	0.0104	0.0145	0.0268
ABC	3	0.0190	0.1955	0.4981
	4	0.0605	0.1096	0.2825
	5	0.4301	0.4464	1.2286

noteworthy that there is a sharp drop from Angle1 to MoY input data quantification, where the quantified value reduces suddenly from 0.48 to 0.13. Figure 7 illustrates the quantified raw input data results with the LASSO.

It is straightforward to express that the quantified raw input data values with the LASSO in Fig. 7 significantly differ from the Pearson correlation. This is because the LASSO applies L_1-norm weighting, as defined in Eq. (18), on the unknown parameters and therefore, the input with the maximum absolute sum value overwhelm the others. The next sub-section examines the quality of the estimates with the selected input dataset.

Estimates with Pearson correlation-based features

Table 3 provides the training estimates of the machine learning models with the Pearson correlation-based selected input dataset.

Table 3 demonstrates that the BLS machine learning model yields the lowest MAE, RMSE and MRAE training performances with the 5 input features. It is notable that even though the BLS machine learning model does not get the benefit of increasing the input features from 3 to 4, it significantly reduces the errors with the 5th input feature addition. Similar comments can be made for the NN machine learning model, but the ABC machine learning model differs from them since it yields smaller errors with the 4th input feature addition. Table 4 illustrates the testing estimates with the Pearson correlation.

The comments made for the training performances of the machine learning models in Table 3 are valid for their testing performances provided in Table 4. This similarity implies that the developed machine learning models are robust against the unseen data and also unknown uncertainties in the data. The next sub-section analyses the estimates with the input dataset selected by the LASSO.

Estimates with LASSO-based input dataset

Table 5 presents the training estimates of the machine learning algorithms with the LASSO-based input dataset selection. The numerical study compares BLS, NN and ABC on the same Pearson and LASSO subsets, rather than retraining RF, SVM or LSTM from[11–13], because those models were reported with meteorological inputs, whereas Table 1 contains only local PV measurements. On this setting, Pearson-BLS with five features gives a test MAE of 0.0036 (Table 4), which is the strongest result in the paper.

As can be deduced from Table 5, the BLS machine learning model outperforms the others for all the input datasets. It also benefits insignificantly when the number of the input dataset is increased from 4 to 5. It is also noticeable that the NN machine learning model produces the second-best training accuracies after the BLS machine learning model. It is clear that for all the measurement criteria, the search-based ABC machine learning model produces the far worst accuracy results. Table 6 investigates the testing performances of these trained models.

It is clear from Tables 5 and 6, the BLS machine learning model produces smaller estimation error than the others with the testing data, which is applied to the trained models. It is also noteworthy that all the developed models generate stable estimates and these estimates do not diverge notably with respect to the number of the

input features. Differently from the trained results in Table 5, the MAE with the BLS model gets a moderate benefit of increasing the number of the input features from 4 to 5.

Discussions, conclusion and future works

This paper investigated the accurate estimation of prosumer net electricity consumption by utilizing BLS, INN, and ABC machine learning algorithms. To enhance model performance and reduce input dimensionality, statistical Pearson correlation and LASSO optimization were integrated into the framework. A polynomial basis function was constructed to expand the unknown parameter space, and two higher-order polynomial powers were determined through matrix conditioning to maximize the information content of the feature set. The optimization process identified 1 and 2 as the optimal polynomial powers, which were subsequently utilized as the basis for all three machine learning algorithms.

In order to select the most informative input dataset among the 14-dimension raw input data, Pearson correlation, which quantifies the relationship between each input data and the target output, is applied. In addition, a L_1-norm constrained LASSO was utilized for the input data dimension reduction. The corresponding results in Figs. 6 and 7 confirmed that the sequence of the selected input data differed notably. While the Pearson correlation quantified the Lag15m data heavily, the LASSO weighted the Solar data considerably. However, both the Pearson correlation and LASSO approaches determined the Current1 and Current2 as the second and third most informative input data. For deployment, the Pearson-based five-feature subset with the BLS model is preferred, because it produces the lowest unseen MAE of 0.0036 in Table 4, compared with 0.0093 for the LASSO-based BLS model in Table 6. Pearson measures linear correlation with the grid power, so Lag15m is ranked first, whereas LASSO applies an L1 penalty and therefore weights the large-magnitude Solar input more heavily.

Estimation accuracies of the BLS and NN gradient-driven and ABC search-based machine learning algorithms were analysed with respect to the selected input data. The testing results in Table 4 verified that the BLS algorithm clearly outperformed the others by producing 0.0036, 0.0046 and 0.0090 MAE, RMSE and MRAE with the Pearson correlation-based selected input data, respectively. The BLS also yielded the best training performance with the LASSO-based selected input data in Table 5, where the MAE, RMSE and MRAE were 0.0080, 0.0109 and 0.0170, respectively. In addition, the NN machine learning algorithm produced the second-best estimates, whereas the ABC only managed to capture the overall character of the target output data. This result verified that the input data is rich in terms of the gradient information and the machine learning algorithms capture the knowledge between the input and output data.

The present study has three main limitations. First, the models are trained and tested on a single Pecan Street household data set (Table 1, 8619 samples), so the errors in Tables 3, 4, 5 and 6 are not yet confirmed on other homes or climates. Second, meteorological variables such as humidity and wind speed are not used; this reduces sensor dependency, but it also means the estimates cannot be compared directly with weather-driven PV forecasts. Third, seasonal and operating-condition splits are not reported. Future work will train the same polynomial-Pearson-LASSO pipeline on meteorological and multi-household data, and will include image-based features and site-specific cloud and shading effects for real-time use.

Author contributions O Tutsoy performed the research, K Zor analysed the data and the research findings, O Celik helped to prepare the methodology of the paper.

Data availability The residential household electricity consumption and photovoltaic generation datasets analyzed during the current study are publicly available in the Pecan Street Dataport repository (https://www.pecanstreet.org/dataport).

Declarations

Competing interests The authors declare that they have no competing interests.

Disclosure During the preparation of this work, the authors used Gemini (Google) in order to refine the technical prose and improve the linguistic clarity of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the final content of the publication.

Ethics The residential household electricity and photovoltaic measurement data utilized in this study were obtained from the Pecan Street Dataport. Because the data were sourced entirely from a pre-existing, fully de-identified public repository, institutional ethical approval and individual informed consent were not required.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Osman, A. I. et al. Cost, environmental impact, and resilience of renewable energy under a changing climate: a review. *Environ. Chem. Lett.* **21**(2), 741–764 (2023).
2. Salkuti, S. R. Modeling of various renewable energy resources for smart electrical power systems. In *Next Gen. Smart Grids*, pp. 29–47 (Springer, 2022).
3. Maka, A. O. M. & Alabid, J. M. Solar energy technology and its roles in sustainable development. *Clean Energy* **6**(3), 476–483 (2022).
4. Victoria, M. et al. Solar photovoltaics is ready to power a sustainable future. *Joule* **5**(5), 1041–1056 (2021).
5. Zhu, X., Liao, B. & Yang, S. An optimal incentive policy for residential prosumers in Chinese distributed photovoltaic market: A Stackelberg game approach. *J. Clean. Prod.* **308**, 127325 (2021).
6. Erdener, B. C., Feng, C., Doubleday, K., Florita, A. & Hodge, B.-M. A review of behind-the-meter solar forecasting. *Renew. Sustain. Energy Rev.* **160**, 112224 (2022).
7. Wu, Y. K. et al. Artificial Intelligence Applications in estimating invisible solar power generation. *Energies* **15**(4), 1312 (2022).
8. Alqarni, M. et al. Utilizing a neutrosophic fuzzy logic system with ANN for short-term estimation of solar energy. *Int. J. Neutrosophic Sci.* **20**(4), 240–259 (2023).
9. Mellit, A., Pavan, A. M. & Lughi, V. Deep learning neural networks for short-term photovoltaic power forecasting. *Renew. Energy* **172**, 276–288 (2021).
10. Bouselham, L., Rabhi, A., Hajji, B. & Mellit, A. Photovoltaic array reconfiguration method based on fuzzy logic and recursive least squares: An experimental validation. *Energy* **232**, 121107 (2021).
11. Hissou, H. et al. A novel machine learning approach for solar radiation estimation. *Sustainability* **15**(13), 10609 (2023).
12. Zazoum, B. Solar PV power prediction using different machine learning methods. *Energy Rep.* **8**, 19–25 (2022).
13. Mahmud, K. et al. Machine learning based PV power generation forecasting. *IEEE Access* **9**, 46117–46128 (2021).
14. Kaur, D. et al. A VAE-Bayesian deep learning scheme for solar power generation forecasting based on dimensionality reduction. *Energy AI* **14**, 100279 (2023).
15. Fariz, T. N. & Basha, S. Enhancing solar radiation predictions through COA optimized neural networks and PCA dimensionality reduction. *Energy Rep.* **12**, 341–359 (2024).
16. Garcia-Cuesta, E., Aler, R., del Pozo-Vázquez, D. & Galván, I. M. A combination of supervised dimensionality reduction and learning methods to forecast solar radiation. *Appl. Intell.* **53**, 13053–13066 (2023).
17. Fu, X. Statistical machine learning model for capacitor planning considering uncertainties in photovoltaic power. *Prot. Control Mod. Power Syst.* **7**(1), 1–13 (2022).
18. Dhaked, D. K. et al. Exploring deep learning methods for solar photovoltaic power output forecasting: a review. *Renew. Energy Focus* **53**, 100682 (2025).
19. Jain, S. & Kanwar, N. Performance comparison of Gaussian and Weibull distribution for next-day solar irradiance and PV power generation forecasting. In *Proc. 2025 IEEE 5th Int. Conf. Sustain. Energy Future Elect. Transp. (SEFET)*, Jaipur, India, pp. 1–6 (2025).
20. Cavus, M., Jiang, J., Allahham, A. & Sun, H. Learning-enhanced stochastic model predictive control with trustworthy forecasting for resilient multi-energy microgrid operation. *Appl. Energy* **422**, 128321 (2026).
21. Cavus, M., Jiang, J., Allahham, A. & Sun, H. A physics-informed machine learning framework for climate-aware digital twins in decentralised energy systems. *Appl. Energy* **416**, 128013 (2026).

22. Nguyen, T.-V.-A., Dao, Q.-T. & Bui, N.-T. Optimized fuzzy logic and sliding mode control for stability and disturbance rejection in rotary inverted pendulum. *Sci. Rep.* **14**, 31116 (2024).
23. Truong, G. M., Dao, Q.-T. & Bui, N.-T. Model predictive control using linearized neural network-based NARX modeling for antagonistic pneumatic artificial muscle systems. *IEEE Access* **14**, 103071–103087 (2026).
24. Pecan Street Dataport. Available: https://www.pecanstreet.org/dataport/ (2025).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Onder Tutsoy[1] · **Kasim Zor**[1] · **Ozgur Celik**[1]

✉ Onder Tutsoy
otutsoy@atu.edu.tr

[1] Adana Science and Technology University, Adana, Turkey