Enhancing Machine Learning Models for Vertical Farm Energy Forecasting: Impact of Data Smoothing and Feature Selection

Shaobo Zhang

Department of Mechanical Engineering

School of Engineering

Cardiff University, Cardiff, UK

ZhangS115@cardiff.ac.uk

Xiao Guo *

Department of Mechanical Engineering

School of Engineering

Cardiff University, Cardiff, UK

GuoX27@cardiff.ac.uk

*Corresponding author

Abstract:

The widespread adoption of vertical farming is constrained by excessive energy consumption, highlighting the need for accurate energy consumption forecasting to develop effective energy-saving strategies. Data-driven models have become increasingly important for this purpose, yet prediction accuracy depends heavily on both data smoothing and feature selection. However, their effects on vertical farm energy estimation remain underexplored.

This study examines how different data preprocessing methods and feature selection techniques influence energy cost prediction in vertical farming using data-driven models. Specifically, it compares two data smoothing techniques—Gaussian Kernel Density Estimation and the Savitzky-Golay filter—in the preprocessing stage. Additionally, it evaluates three feature selection methods: backward elimination, PCA-based backward elimination, and genetic algorithms (GA), assessing their impact on model performance.

The results indicate that the Savitzky-Golay filter and PCA-based backward elimination significantly enhance both prediction accuracy and computational efficiency. These findings provide valuable insights for optimizing energy efficiency in vertical farming.

Keywords: energy prediction, machine learning, vertical farming, data smoothing, feature selection

1. Introduction:

1.1 Background

While vertical farming offers a sustainable solution to modern agricultural challenges, its high energy consumption remains a significant concern. This demand primarily stems from illumination, cooling, ventilation, and heating systems. Harbrick et al. reported that vertical farms consume 11 to 13 times more energy than traditional greenhouses [1]. Similarly, Graamans et al. compared lettuce production in vertical farms and greenhouses, finding that vertical farms require substantially more electricity [2].

Studies have shown that optimizing building operations and control strategies can significantly reduce energy consumption in energy-intensive buildings [3][4][5]. Accurate energy consumption forecasting is crucial for developing effective energy-saving policies. Advances in the Internet of Things (IoT) have simplified large-scale data collection, enabling the widespread adoption of data-driven models for energy prediction [6][7][8]. Implementing high-efficiency management strategies based on these predictions can further reduce energy consumption.

1.2 State of the art

Energy consumption in high-demand buildings is typically estimated using two primary methodologies: physical modeling (white-box) and data-driven approaches (black-box) [9].

The white-box method [10] relies on detailed building physics and simulation software such as EnergyPlus, OpenStudio, and TRNSYS. These tools require comprehensive building attributes, including both geometric and non-geometric data [11]. However, Guo et al. noted that crop transpiration models are often overlooked in vertical farm energy simulations, with only 24% of 72 studies incorporating them. Instead, most rely on CFD simulations for temperature and ventilation analysis [12]. Graamans et al. introduced a plant energy submodule, later integrated into EnergyPlus and TRNSYS [13]. However, these models assume a constant leaf area index (LAI) and overlook key energy transfer processes, leading to inaccuracies in machine learning-based energy forecasting [14].

With advancements in computing power and the Internet of Things (IoT), machine learning (black-box) approaches have become increasingly viable. Various methods, including Artificial Neural Networks (ANN), Random Forest (RF), Support Vector Regression (SVR), and Decision Trees (DT), have been explored for energy prediction [15][16][17]. As a type of high-density, energy-intensive building, vertical farms have gained attention, with growing efforts to predict their internal microclimate using machine learning algorithms [18][19][20].

Data smoothing is a critical preprocessing step in machine learning, ensuring the quality and relevance of input data. Common techniques include Moving Averages, Exponential Smoothing, Low-Pass Filtering, Spline Smoothing, the Savitzky-Golay Filter, Data Aggregation, and Gaussian Kernel Density Estimation (Gaussian KDE). Choosing the appropriate smoothing technique is challenging, as an unsuitable method can lead to information loss, reduced predictive accuracy, and decreased computational efficiency. Despite its importance, many studies use only a single smoothing method or omit this step entirely before training machine learning models.

Beyond data smoothing, feature selection also significantly impacts training time and prediction accuracy. Due to the difficulty of preselecting the most relevant features, some studies incorporate all available variables, often introducing irrelevant data that reduces model interpretability and increases computational cost [21]. However, most research employs only a single feature selection method, lacking a comparative analysis of how different techniques affect model performance.

1.3 Problem statement

Based on the available literature, several key shortcomings have been identified:

- Impact of Data Smoothing: Data smoothing techniques play a crucial role in shaping machine learning model performance. However, their specific impact on energy cost estimation in vertical farming remains largely unexplored.
- Feature Selection Comparisons: While feature selection enhances model accuracy, computational efficiency, and interpretability, most studies in building energy prediction focus on a single feature selection method, lacking comprehensive comparative analyses.
- Limited Training Data Duration: Existing studies primarily use weather data spanning one week to several months for training machine learning models. However, such short-term datasets fail to capture the full relationship between weather variation and energy costs over an entire year, as factors like outdoor temperature and solar radiation fluctuate significantly across seasons.

To address these challenges, this study investigates the influence of different data smoothing and feature selection techniques on energy consumption prediction models. Specifically, two widely used smoothing approaches—Gaussian Kernel Density Estimation (KDE) and the Savitzky-Golay filter—are analyzed to assess their impact on predictive accuracy.

Additionally, the study evaluates the most significant variables influencing energy costs using three feature selection methods: backward elimination, PCA-based backward elimination, and genetic algorithms (GA). These methods aim to prevent inefficient training and performance degradation. The results from these feature selection techniques are compared to assess their computational cost and impact on predictive accuracy.

This paper is structured as follows: Section 2 presents the proposed methodology, detailing data sources, smoothing techniques, feature selection strategies, and the energy prediction model. Section 3 analyzes and compares the findings, examining the influence of smoothing techniques and feature selection strategies. Section 4 discusses the study's limitations and outlines directions for future research.

2. Methods

2.1 Data acquisition

This study utilizes data from three Venlo glass greenhouses in Yangling, China, dedicated to the production of cherry tomatoes, lettuce, and flowers (see Figure 1). Each greenhouse is equipped with a ventilation system, supplementary lighting, and an HVAC system. Environmental data is collected in real-time through sensors at one-hour intervals, covering indoor and outdoor temperatures, indoor humidity, indoor CO_2 concentration, solar intensity, outdoor wind speed, and wind direction. The data collection period spans March 2021 to February 2022. The dataset, obtained from Cao et al. [24], provides detailed greenhouse environmental parameters for energy consumption prediction.

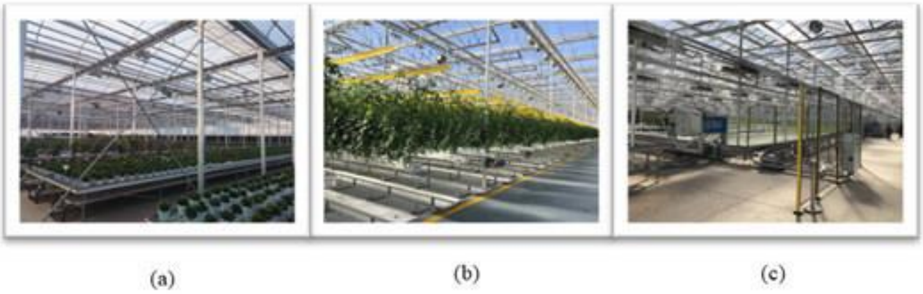


Figure 1. Experimental greenhouse: (a) Greenhouse for flower. (b) Greenhouse for cherry tomato. (c) Greenhouse for lettuce

2.2 Data preprocessing

Data collected from researchers or sensors inherently contains anomalies, missing values, and noise, which can negatively impact the accuracy of machine learning models. Additionally, the dataset comprises various sensor readings, each with a unique feature scale. Since many machine learning models and feature selection methods are sensitive to feature scaling, data normalization is essential to ensure efficient model convergence and improved predictive performance [22].

2.2.1 Missing Value and normalization

(1) Missing value:

The raw dataset contains a small number of missing sensor readings. However, given the large dataset size, removing these missing values is unlikely to affect subsequent analysis. Therefore, this study excludes records with missing values.

(2) Data normalization

Min-max normalization was applied to standardize wind-related variables.

$$\min - \max = \frac{x - X_{min}}{X_{max} - X_{min}} \tag{1}$$

Where: x is the target eigenvalue; X_{max} is the max value in wind direction; X_{min} is the minimum value in wind direction.

2.2.2 Data smoothing methods

Selecting an appropriate data smoothing technique requires consideration of data characteristics, smoothing objectives, and algorithmic complexity. This study employs Gaussian KDE and the Savitzky-Golay filter as smoothing methods. To determine the optimal approach for energy prediction in vertical farming, the effectiveness of models trained on raw data, Gaussian KDE-smoothed data, and Savitzky-Golay-smoothed data is assessed and compared. The underlying principles and rationale for these two smoothing techniques are outlined below.

(1) Gaussian KDE

Gaussian KDE is primarily used for continuous and multi-dimensional data, including structured data such as time series. It is also commonly applied in outlier detection, as it estimates high-probability regions and identifies low-probability anomalies based on probability thresholds.

The Gaussian KDE is represented by the following equation [23]:

$$K(x,h) = \frac{1}{h\sqrt{\pi}} e^{-\frac{x^2}{2}} \tag{2}$$

$$f(x) = \frac{1}{nh} \sum_{k=1}^{n} K(\frac{x-x_k}{h}) \tag{3}$$

Where: n is the number of data samples, h denotes the bandwidth, K refers to the Gaussian kernel formulated in one-dimensional space.

(2) Savitzky-Golay Filter

The Savitzky-Golay Filter is particularly suitable for continuous time series and experimental measurement data obtained by experimental measurement, especially in scenarios where there is a certain amount of noise, and the data features need to be retained.

$$f(x_i) = \sum_{k=2} c_n x_i^k \tag{4}$$

Where: $f(x_i)$ represents the smoothed value of x_i, k denotes the poly order, c_n refers to the coefficients derived from the least square method, estimated using 2n + 1 points, which are determined by the window length.

2.3 Feature selection

Feature selection improves the efficiency and effectiveness of predictive models by identifying and prioritizing the most relevant variables. Features with low correlation can add unnecessary complexity and increase computational burden, ultimately reducing model accuracy.

2.3.1 Backward elimination based on principal component analysis (PCA):

A backward elimination approach based on PCA is proposed to identify the most significant variable.

Step (1) Data preparation:

Define energy as the target variable, while treating all other variables as independent variables.

Step (2) Data standardization:

Normalize all independent variables to achieve a mean of 0 and a standard deviation of 1.

Step (3) Apply PCA to reduce dimensionality:

Form the covariance matrix.

Calculate eigenvalues and eigenvectors and extract the principal components.

Select the principal components that contribute to over 95% of the cumulative explained variance.

Step (4) Apply backward elimination:

Remove the smallest eigenvalues each time.

Reapply PCA after each removal to evaluate changes in cumulative explained variance and decide whether to retain the variable.

Step (5) Outputs the most important features.

2.3.2 Backward elimination:

Backward elimination is an efficient and straightforward technique for selecting a subset of variables in a linear regression model. It is easily implementable and can be automated.

Step (1) Data preparation:

Set Energy as the dependent variable (Y), while the independent variables X comprise all columns except Energy.

Step (2) Define the backward elimination function:

Define the significance level as 0.01 and construct the backward elimination functions.

Step (3) Apply the backward elimination:

Calculate P-values for all features and evaluate each variable's P-value. Eliminate the feature with the highest P-value each time.

Step (4) Output significant variables.

2.3.3 Genetic algorithm:

GA is a computer science and operations research method for solving optimization problems that use principles of natural selection and evolution. Figure 2. illustrates the flowchart of the GA.

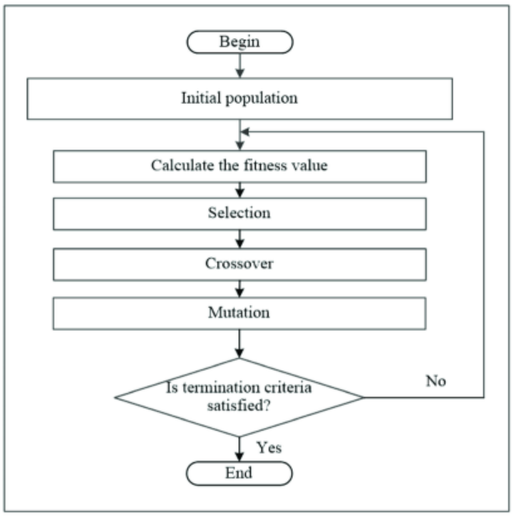

Begin
Initial population
Calculate the fitness value
Selection
Crossover
Mutation
Is termination criteria satisfied?
No
Yes
End

Figure 2. Flow diagram for GA

Step (1) Data preparation:

Assign energy as the target variable and designate all other variables as independent variables.

Step (2) Train/test sets split:

Divide the dataset into a training set and a testing set, allocating 70% for training and 30% for testing.

Step (3) Apply genetic algorithm:

Initialization: Randomly generate initial subsets of features.

Fitness function: Utilize mean squared error (MSE) to evaluate the performance of each subset.

Step (4) Model selection:

Random forest was selected to evaluate the ability of prediction of feature subsets.

Step (5) Outputs the most important features.

2.4 Training set volumes

Prior to training the machine learning model, determining the optimal size of the training set is crucial. The amount of training data significantly influences the prevention of underfitting and overfitting. In this study, the dataset is split into a 3:1 training-test ratio (75% training, 25% testing) to ensure that the model has sufficient data for learning while reserving enough unseen data for reliable evaluation. This ratio provides a balance between model training efficiency and performance assessment, reducing variance in test results while maintaining computational feasibility.

2.5 Prediction algorithm

A previous study successfully used this dataset to predict energy load. After comparing the accuracy and generalization capability of eight widely used machine learning models, random forest (RF) was identified as the optimal choice [24]. Moreover, this study focuses on analyzing factors that influence prediction accuracy rather than the machine learning model itself. Therefore, RF is used as the predictive model.

2.6 Hyperparameter tune

In this study, the Random Search Cross-Validation optimizer was employed to enhance the accuracy of the random forest.

Step (1): Defining the search space, specifying key hyperparameters such as the number of estimators, maximum depth, minimum samples split, and maximum features.

Step (2): Random sampling from this space to select candidate hyperparameter sets.

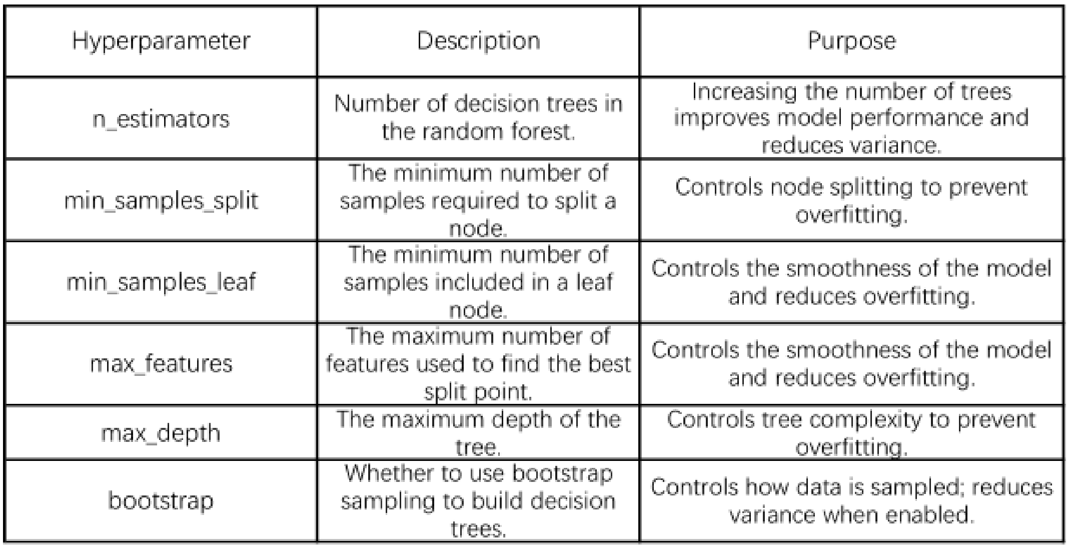
Step (3): Applying RandomizedSearchCV with 3-fold cross-validation and 100 iterations, balancing computational efficiency and performance estimation.

Step (4): Identifying the best hyperparameter set and comparing it with the baseline model.

Random Search Cross-Validation was chosen for its balance between efficiency and accuracy. Unlike Grid Search, which exhaustively tests all hyperparameter combinations and is computationally expensive, Random Search explores a subset of possibilities, reducing cost while still finding near-optimal solutions.

The 3-fold cross-validation was chosen to reduce computation time while maintaining a reasonable performance estimate. Although 5-fold or 10-fold cross-validation could improve stability, the added cost was not justified for hyperparameter tuning in this study. Future work may explore different validation strategies. This study uses R-squared and variance as performance metrics, as they assess model fit and generalization ability, respectively. The summary of various random forest hyperparameters along with their typical default values is presented below:

Table 1. Hyperparameter and description



Hyperparameter	Description	Purpose
n_estimators	Number of decision trees in the random forest.	Increasing the number of trees improves model performance and reduces variance.
min_samples_split	The minimum number of samples required to split a node.	Controls node splitting to prevent overfitting.
min_samples_leaf	The minimum number of samples included in a leaf node.	Controls the smoothness of the model and reduces overfitting.
max_features	The maximum number of features used to find the best split point.	Controls the smoothness of the model and reduces overfitting.
max_depth	The maximum depth of the tree.	Controls tree complexity to prevent overfitting.
bootstrap	Whether to use bootstrap sampling to build decision trees.	Controls how data is sampled; reduces variance when enabled.

2.6 Model evaluation indicators

This study utilizes multiple evaluation metrics to measure the performance of trained models, including root mean square error (RMSE), R-squared (R^2), and training time.

R – squared (R^2) quantifies the goodness of fit of the regression model to the dataset. The formula for R^2 is expressed as follows:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \qquad (5)$$

Where:

The residual Sum of Squares (SS_{res}) quantifies the portion of variance in the dependent variable that remains unexplained by the model. The total Sum of Squares (SS_{tot}) measures the overall variance in the dependent variable, serving as a baseline for variability.

The Root Mean Squared Error (RMSE) quantifies the average deviation between the model's predicted values and the actual values, indicating accuracy. RMSE can be obtained by Equation 6.

$$RMSE = \sqrt{\frac{1}{M}\sum_{t=1}^{M}(\log(y_t + 1) - \log(\hat{y}_t + 1))^2} \qquad (6)$$

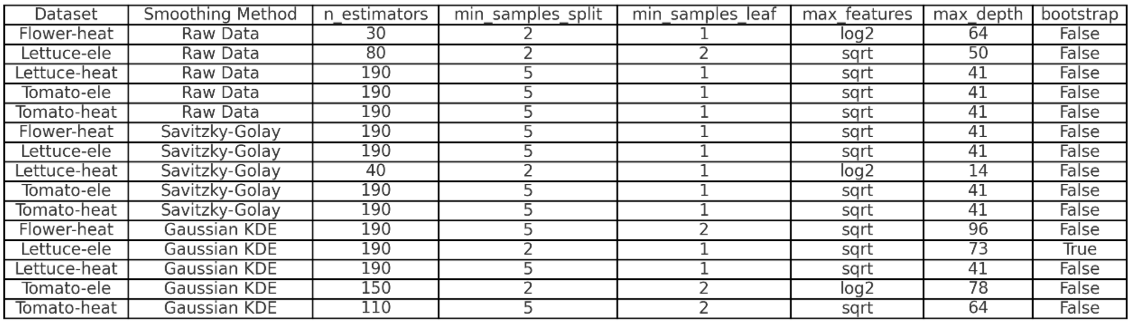
Where: M represents the total number of data points, y_t denotes the actual value, and $\hat{y}_t$ refers to the forecasted value.

3. Results

3.1 Random forest hyperparameter tuning results

Table 2. presents the optimal hyperparameters identified through the Random Search Cross-Validation technique, using R² as the evaluation criterion.

Table 2. Hyperparameter tune



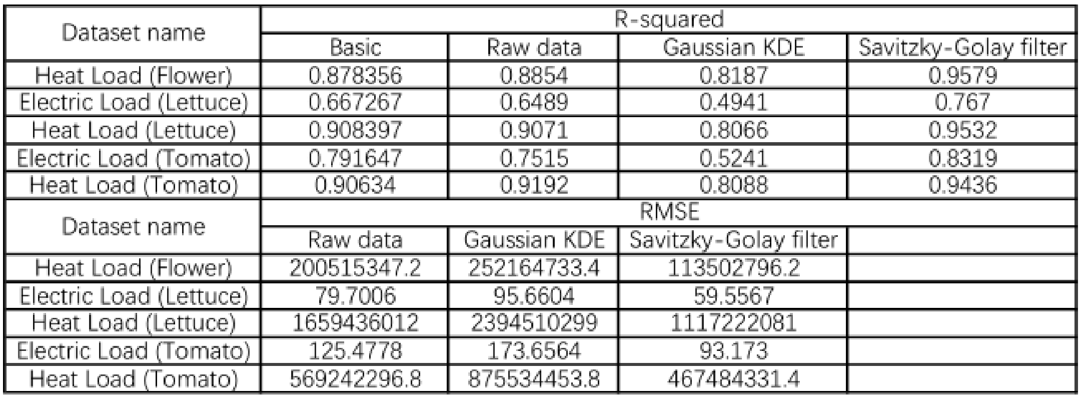
Dataset	Smoothing Method	n_estimators	min_samples_split	min_samples_leaf	max_features	max_depth	bootstrap
Flower-heat	Raw Data	30	2	1	log2	64	False
Lettuce-ele	Raw Data	80	2	2	sqrt	50	False
Lettuce-heat	Raw Data	190	5	1	sqrt	41	False
Tomato-ele	Raw Data	190	5	1	sqrt	41	False
Tomato-heat	Raw Data	190	5	1	sqrt	41	False
Flower-heat	Savitzky-Golay	190	5	1	sqrt	41	False
Lettuce-ele	Savitzky-Golay	190	5	1	sqrt	41	False
Lettuce-heat	Savitzky-Golay	40	2	1	log2	14	False
Tomato-ele	Savitzky-Golay	190	5	1	sqrt	41	False
Tomato-heat	Savitzky-Golay	190	5	1	sqrt	41	False
Flower-heat	Gaussian KDE	190	5	2	sqrt	96	False
Lettuce-ele	Gaussian KDE	190	2	1	sqrt	73	True
Lettuce-heat	Gaussian KDE	190	5	1	sqrt	41	False
Tomato-ele	Gaussian KDE	150	2	2	log2	78	False
Tomato-heat	Gaussian KDE	110	5	2	sqrt	64	False

3.2 Effects of data smoothing on predictive model

This section examines how data smoothing techniques influence machine learning-based energy prediction, utilizing two approaches: Gaussian KD and the Savitzky-Golay Filter.

Table 3 presents the performance of different smoothing techniques in energy prediction.

Table 3. Data smoothing methods comparison



Dataset name	R-squared			
	Basic	Raw data	Gaussian KDE	Savitzky-Golay filter
Heat Load (Flower)	0.878356	0.8854	0.8187	0.9579
Electric Load (Lettuce)	0.667267	0.6489	0.4941	0.767
Heat Load (Lettuce)	0.908397	0.9071	0.8066	0.9532
Electric Load (Tomato)	0.791647	0.7515	0.5241	0.8319
Heat Load (Tomato)	0.90634	0.9192	0.8088	0.9436
Dataset name	RMSE			
	Raw data	Gaussian KDE	Savitzky-Golay filter	
Heat Load (Flower)	200515347.2	252164733.4	113502796.2	
Electric Load (Lettuce)	79.7006	95.6604	59.5567	
Heat Load (Lettuce)	1659436012	2394510299	1117222081	
Electric Load (Tomato)	125.4778	173.6564	93.173	
Heat Load (Tomato)	569242296.8	875534453.8	467484331.4	

The "Basic" column in Table 3 represents the R² value of the RF model from prior research, while "Raw Data" reflects the performance of the model trained on unprocessed data.

For R², the Gaussian KDE model performs worse than the raw data model, indicating its unsuitability for energy prediction in vertical farming. In contrast, the Savitzky-Golay filter significantly improves model performance, yielding a notably higher R².

As shown in Table 3, thermal load predictions exhibit a better model fit than electrical load predictions, likely due to a stronger correlation between input variables and thermal load. In contrast, the dataset lacks sufficient variables to explain variations in electrical load.

For RMSE, models trained on Gaussian KDE-smoothed data demonstrate significantly lower accuracy than those using the Savitzky-Golay filter, confirming that the latter is more suitable for energy consumption prediction in vertical farming.

The superior performance of the Savitzky-Golay filter over Gaussian KDE stems from their differing smoothing mechanisms. Gaussian KDE relies on bandwidth selection, which may over-smooth data, reducing accuracy. In contrast, the Savitzky-Golay filter preserves local trends while reducing noise, leading to improved predictive performance.

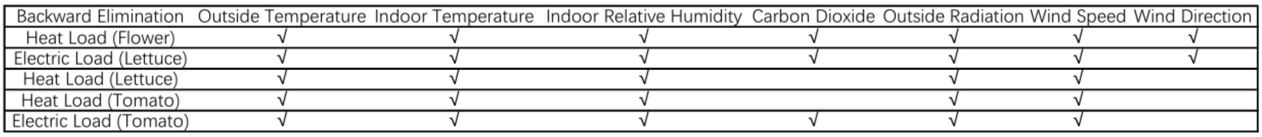
These findings highlight the significant impact of data smoothing on model performance in vertical farm energy prediction. Future research could explore alternative techniques, such as wavelet denoising or adaptive filtering, to further enhance prediction accuracy while preserving critical data patterns.

3.3 Impacts of Feature selection on the prediction model

This section applies three feature selection methods: backward elimination, PCA-based backward elimination, and genetic algorithms (GA).

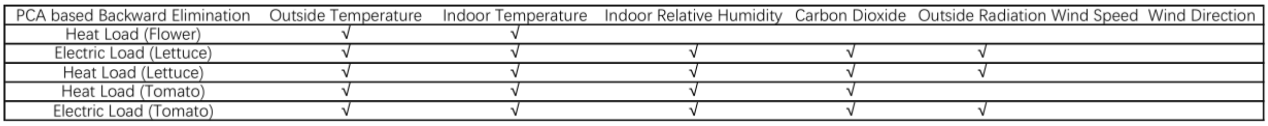
Table 4 presents the key features identified through backward elimination for different greenhouses and target loads. Table 5 shows the results of PCA-based backward elimination under similar conditions, while Table 6 highlights the significant features selected using genetic algorithms (GA) for each greenhouse and target load.

Table 4. Backward elimination



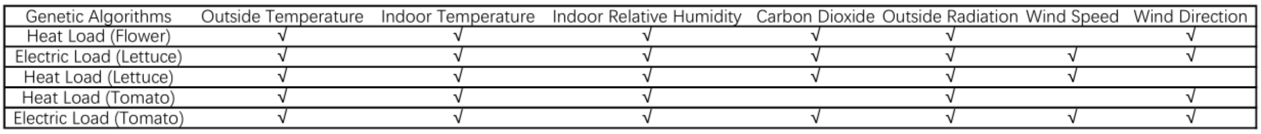
Backward Elimination	Outside Temperature	Indoor Temperature	Indoor Relative Humidity	Carbon Dioxide	Outside Radiation	Wind Speed	Wind Direction
Heat Load (Flower)	√	√	√	√	√	√	√
Electric Load (Lettuce)	√	√	√	√	√	√	√
Heat Load (Lettuce)	√	√	√		√	√	
Heat Load (Tomato)	√	√	√		√	√	
Electric Load (Tomato)	√	√	√	√	√	√	

Table 5. PCA based on backward eliminations



PCA based Backward Elimination	Outside Temperature	Indoor Temperature	Indoor Relative Humidity	Carbon Dioxide	Outside Radiation	Wind Speed	Wind Direction
Heat Load (Flower)	√	√					
Electric Load (Lettuce)	√	√	√	√	√		
Heat Load (Lettuce)	√	√	√	√	√		
Heat Load (Tomato)	√	√	√	√			
Electric Load (Tomato)	√	√	√	√	√		

Table 6. Genetic Algorithms



Genetic Algorithms	Outside Temperature	Indoor Temperature	Indoor Relative Humidity	Carbon Dioxide	Outside Radiation	Wind Speed	Wind Direction
Heat Load (Flower)	√	√	√	√	√		√
Electric Load (Lettuce)	√	√	√	√	√	√	√
Heat Load (Lettuce)	√	√	√	√	√	√	
Heat Load (Tomato)	√	√	√		√		√
Electric Load (Tomato)	√	√	√	√	√	√	√

Tables 4, 5, and 6 show that different feature selection methods identify varying key features. However, all three methods consistently highlight indoor and outdoor temperatures, humidity, and CO2 concentration as closely related to energy load.

PCA-based backward elimination suggests that wind speed and wind direction have minimal impact on heat and electric load, favoring variables that directly influence the internal greenhouse environment. In contrast, the GA method tends to retain all features, indicating weaker feature elimination, which may be better suited for capturing complex nonlinear relationships.

For heat load, all three feature selection methods identify indoor and outdoor temperatures as key factors, aligning with fundamental heat load dynamics in natural environments. Additionally, relative humidity and carbon dioxide concentration are repeatedly selected, suggesting their strong influence on crop transpiration and overall thermal balance within the greenhouse.

For electric load, indoor and outdoor temperatures remain the primary features. However, unlike heat load, indoor humidity and carbon dioxide concentration also play a significant role, likely due to their indirect impact on cooling and ventilation system operation.

In this study, random forest was also used to evaluate feature importance across different greenhouses. The results are as follows:

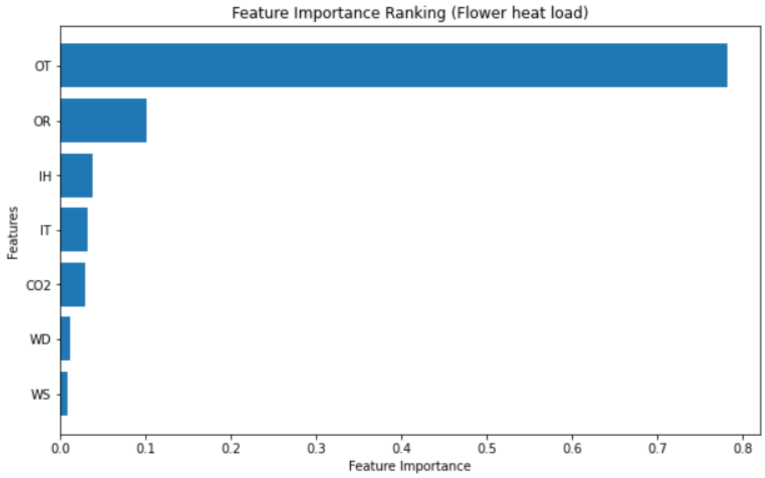


Figure 3. Feature important ranking (Flower heat load)

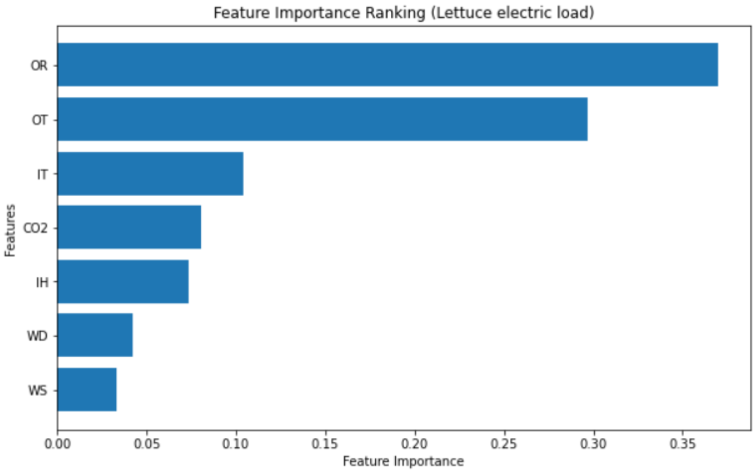


Figure 4. Feature important ranking (Lettuce electric load)

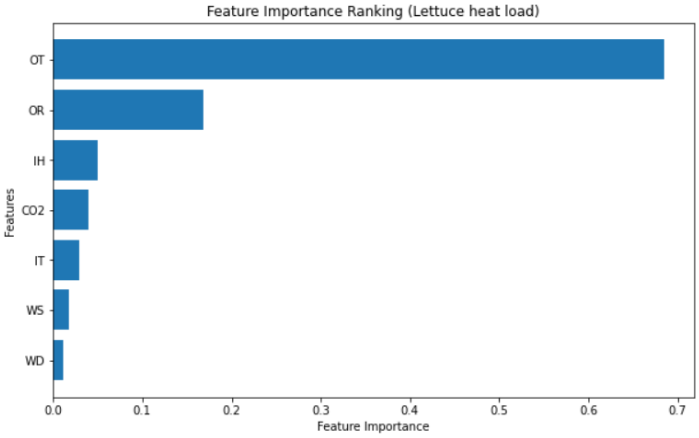


Figure 5. Feature important ranking (Lettuce heat load)

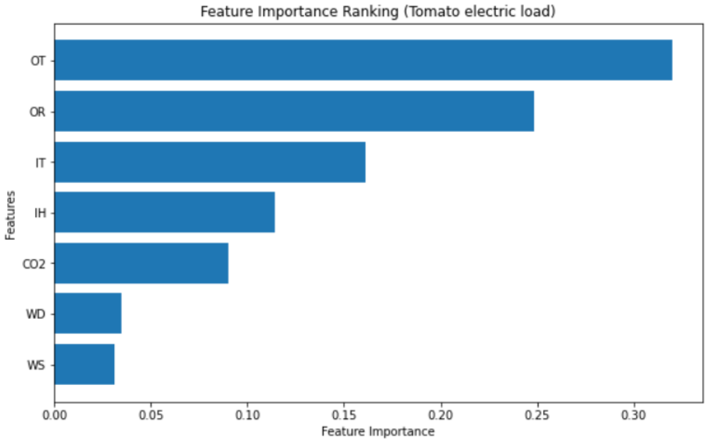


Figure 6. Feature important ranking (Tomato electric load)

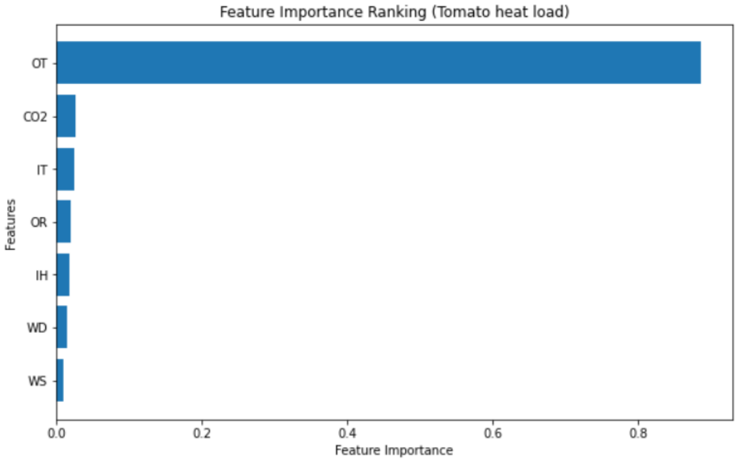


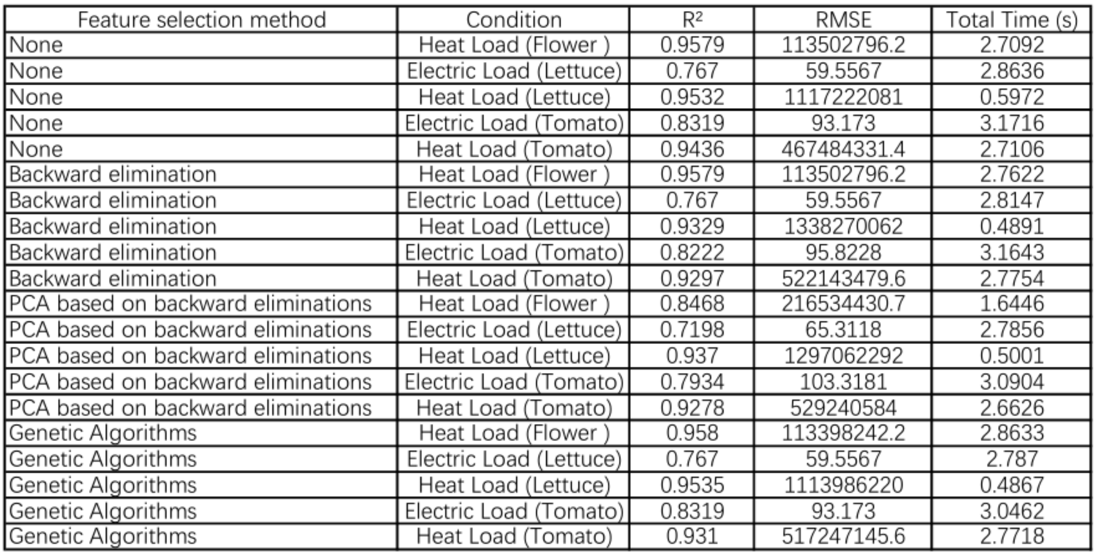
Figure 7. Feature important ranking (Tomato heat load)

Figures 3–7 and the analysis of five tables lead to the following conclusions:

- Heat Load Prediction: Outdoor temperature (OT) is the most influential factor across all three greenhouses, with its importance significantly surpassing other features. This suggests that heat load is primarily driven by external temperature fluctuations, as greenhouses regulate their internal climate accordingly.
- Electric Load Prediction: While outdoor temperature (OT) and outdoor solar radiation intensity (OR) remain critical, other factors—such as indoor temperature (IT), indoor relative humidity (IH), and CO_2 concentration (CO_2)—also play significant roles, varying across different crops. This indicates that electric load is influenced by a broader range of factors.

By comparing the evaluation metrics of different feature selection methods, this study analyzes their impact on the performance and efficiency of the machine learning prediction model. The results are as follows:

Table 7. Comparison of feature selection methods



Feature selection method	Condition	R²	RMSE	Total Time (s)
None	Heat Load (Flower)	0.9579	113502796.2	2.7092
None	Electric Load (Lettuce)	0.767	59.5567	2.8636
None	Heat Load (Lettuce)	0.9532	1117222081	0.5972
None	Electric Load (Tomato)	0.8319	93.173	3.1716
None	Heat Load (Tomato)	0.9436	467484331.4	2.7106
Backward elimination	Heat Load (Flower)	0.9579	113502796.2	2.7622
Backward elimination	Electric Load (Lettuce)	0.767	59.5567	2.8147
Backward elimination	Heat Load (Lettuce)	0.9329	1338270062	0.4891
Backward elimination	Electric Load (Tomato)	0.8222	95.8228	3.1643
Backward elimination	Heat Load (Tomato)	0.9297	522143479.6	2.7754
PCA based on backward eliminations	Heat Load (Flower)	0.8468	216534430.7	1.6446
PCA based on backward eliminations	Electric Load (Lettuce)	0.7198	65.3118	2.7856
PCA based on backward eliminations	Heat Load (Lettuce)	0.937	1297062292	0.5001
PCA based on backward eliminations	Electric Load (Tomato)	0.7934	103.3181	3.0904
PCA based on backward eliminations	Heat Load (Tomato)	0.9278	529240584	2.6626
Genetic Algorithms	Heat Load (Flower)	0.958	113398242.2	2.8633
Genetic Algorithms	Electric Load (Lettuce)	0.767	59.5567	2.787
Genetic Algorithms	Heat Load (Lettuce)	0.9535	1113986220	0.4867
Genetic Algorithms	Electric Load (Tomato)	0.8319	93.173	3.0462
Genetic Algorithms	Heat Load (Tomato)	0.931	517247145.6	2.7718

The subsequent analysis will evaluate the strengths and limitations of the three feature selection methods concerning different greenhouse loads.

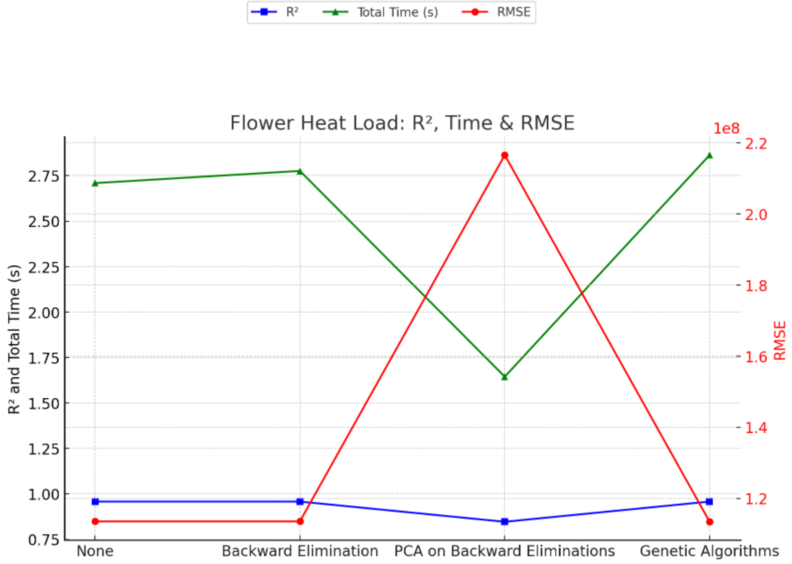


Figure 8. Comparison of feature selection methods in heat load (Flower)

Figure 3 illustrates that outdoor temperature (OT) is the primary factor influencing heat load in flower greenhouses, followed by outdoor solar radiation.

Analysis of Tables 4–6 indicates that the backward elimination method retains all input features, while PCA-based backward elimination selects only outdoor and indoor temperatures. In contrast, GA excludes outdoor wind speed.

As shown in Figure 8, PCA-based backward elimination significantly improves computational efficiency, reducing prediction time by approximately 40%. In contrast, the other two methods offer negligible runtime improvements compared to the original dataset without feature selection.

Regarding model fit, PCA-based backward elimination results in a 12% reduction in R², while backward elimination and GA maintain performance comparable to the original dataset. However, PCA-based backward elimination substantially increases RMSE, indicating higher prediction errors.

These findings suggest that for heat load prediction in flower greenhouses, PCA-based backward elimination effectively identifies critical features and enhances computational efficiency. However, this comes at the cost of reduced prediction accuracy. If accuracy is the priority, backward elimination or GA are more suitable. If reducing runtime is the main concern, PCA-based backward elimination provides a favorable trade-off.

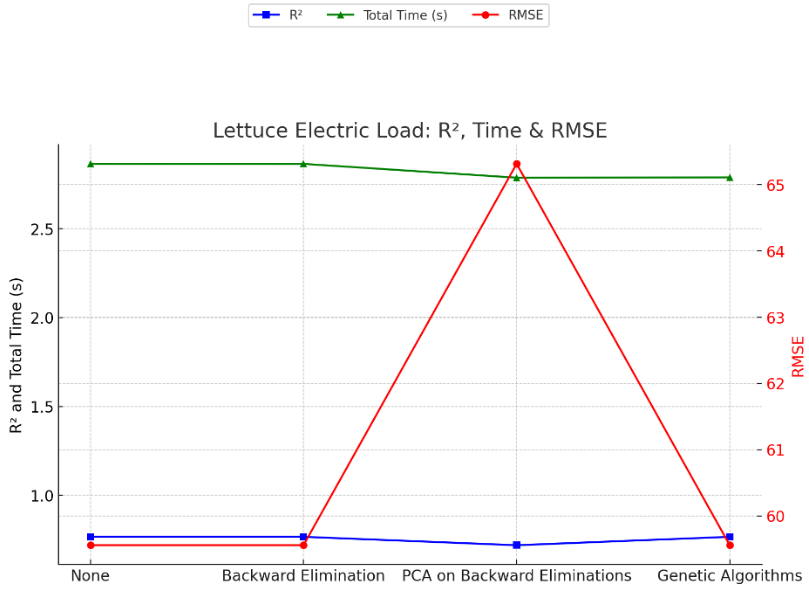


Figure 9. Comparison of feature selection methods in electric load (Lettuce)

In Figure 4, outdoor temperature and solar radiation are important factors influencing electrical load in lettuce greenhouses, while outdoor wind speed and wind direction have minimal impact.

A comparison of Tables 4–6 reveals that backward elimination and GA retain all input features, whereas PCA-based backward elimination excludes wind speed and wind direction. This aligns with the feature importance ranking from the random forest model. As shown in Figure 9, PCA-based backward elimination slightly improves runtime, reducing computation time by 2.7%. However, it also results in a 6.2% decrease in R² compared to the original dataset. In contrast, backward elimination and GA maintain performance comparable to the original dataset. For RMSE, PCA-based backward elimination increases error by 9.7%.

These findings suggest that PCA-based backward elimination effectively identifies key features while enhancing computational efficiency. For electric load prediction in lettuce greenhouses, it provides a balanced trade-off between feature selection and performance.

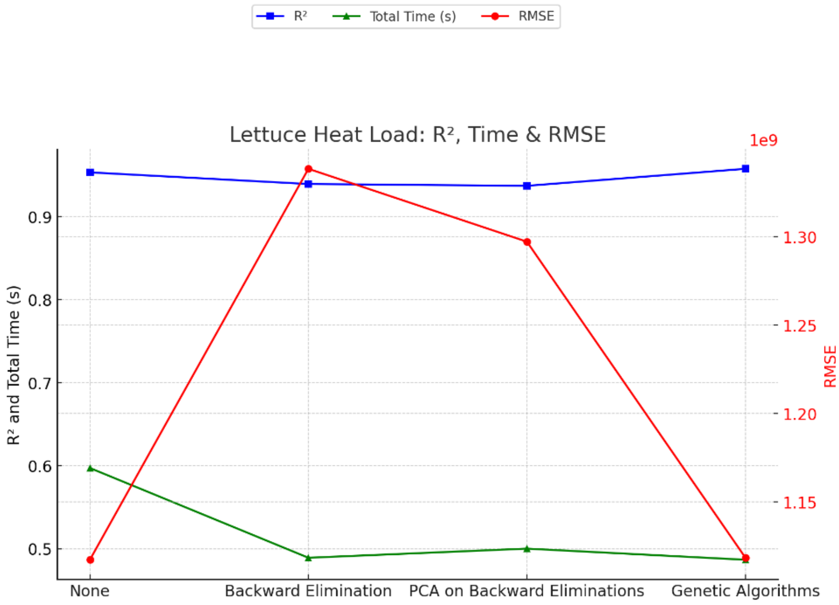


Figure 10. Comparison of feature selection methods in heat load (Lettuce)

As shown in Figure 5, outdoor temperature is considered a key feature influencing electric load, while solar radiation also plays a significant role. Other features have minimal impact on electric load.

A comparison of Tables 4–6 shows that backward elimination removes carbon dioxide concentration and wind direction, PCA-based backward elimination excludes outdoor wind speed and wind direction, while GA removes only outdoor wind direction. All three methods accurately identify the key factors.

As illustrated in Figure 10, all three feature selection methods enhance computational efficiency. Backward elimination reduces computation time by 18%, PCA-based backward elimination by 16%, and GA by 19%, demonstrating similar performance in runtime efficiency. Model fit (R^2) remains comparable across all methods, indicating minimal impact on predictive performance.

For RMSE, GA achieves prediction errors closest to the original dataset. Backward elimination results in the highest RMSE increase (20%), while PCA-based backward elimination increases RMSE by 16%.

These findings suggest that for heat load prediction in lettuce greenhouses, GA is the most favorable feature selection method, offering a balance between computational efficiency and predictive accuracy.

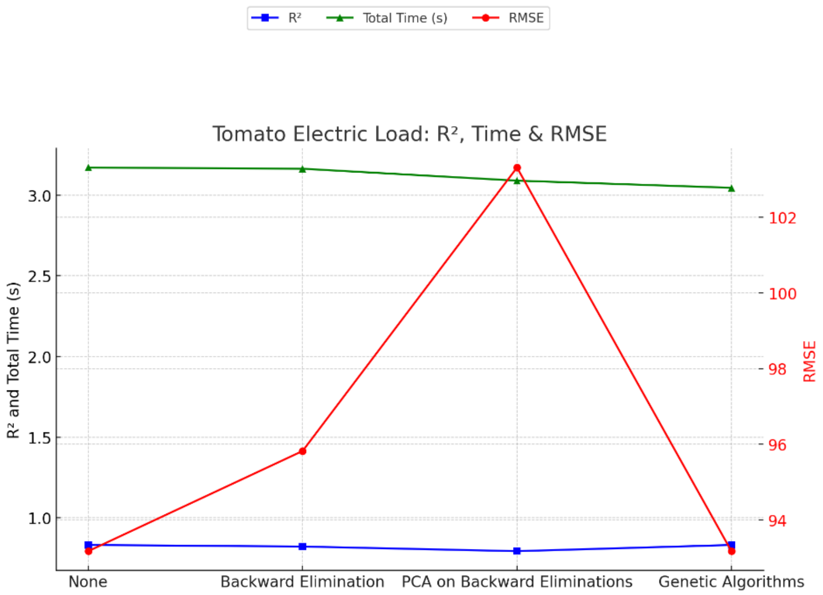


Figure 11. Comparison of feature selection methods in electric load (tomato)

As illustrated in Figure 6, the critical factors influencing electric load in tomato greenhouses include outdoor and indoor temperatures, solar radiation, humidity, and indoor CO2 concentration, while wind speed and direction have minimal impact.

An analysis of Tables 4–6 shows that backward elimination removes outdoor wind direction, PCA-based backward elimination excludes both outdoor wind speed and wind direction, while GA retains all features.

As shown in Figure 11, the three feature selection methods exhibit negligible differences in computational efficiency. Regarding model fit, PCA-based backward elimination reduces R² by 4.6%, while backward elimination and GA maintain performance comparable to the original dataset. In terms of RMSE, PCA-based backward elimination increases error by approximately 11%, while the other two methods show minimal changes.

These findings suggest that PCA-based backward elimination effectively identifies key features but provides limited computational efficiency improvements. However, the slight reduction in model fit and increase in prediction error remain within an acceptable range. Overall, for electric load prediction in tomato greenhouses, all three methods perform similarly, with PCA-based backward elimination being a slightly more favorable option due to its balance between feature selection and performance trade-offs.

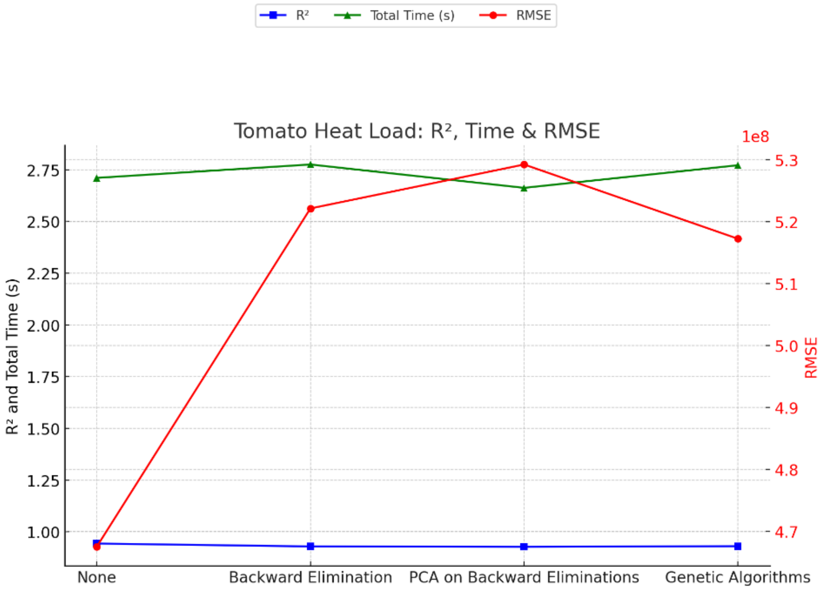


Figure 12. Comparison of feature selection methods in heat load (tomato)

As shown in Figure 7, outdoor temperature is the primary factor affecting heat load in tomato greenhouses, while other features have a relatively minor impact.

An analysis of Tables 4–6 reveals that backward elimination removes outdoor wind direction and indoor carbon dioxide concentration, PCA-based backward elimination excludes outdoor wind speed, wind direction, and solar radiation, while GA eliminates carbon dioxide concentration and wind speed.

As shown in Figure 12, the three feature selection methods exhibit minimal differences in computational efficiency, with PCA-based backward elimination offering a slight runtime improvement. In terms of model fit, all three methods perform similarly to the model trained on raw data, showing no substantial differences in fit quality. However, for RMSE, backward elimination increases error by 11.7%, PCA-based backward elimination by 13.2%, and GA by 10.6%.

These findings suggest that none of the feature selection methods significantly improve computational efficiency or accuracy for heat load prediction in tomato greenhouses. Model fit remains unchanged, while prediction errors increase. Based on these results, feature selection methods are not recommended for predicting heat load in tomato greenhouses.

4. Conclusion and future work

This study enhances energy consumption prediction for vertical farms by evaluating data smoothing and feature selection techniques. Results highlight the superior performance of the Savitzky-Golay Filter for data smoothing and the effectiveness of PCA-based backward elimination for feature selection under specific conditions. Key predictive features vary across load types and crop varieties, emphasizing the importance of tailored data processing strategies.

Beyond vertical farms, these findings may generalize to other high-energy-consumption buildings, such as greenhouses, data centers, or smart buildings, where accurate energy forecasting is critical for efficiency and sustainability. The demonstrated impact of smoothing techniques on model accuracy suggests potential applications in real-time energy monitoring systems and adaptive control strategies.

Future research will focus on incorporating high-resolution seasonal weather data to better account for climate-induced variations, exploring k-fold cross-validation for improved model tuning, and investigating alternative smoothing techniques (e.g., wavelet denoising) for further accuracy enhancements. Additionally, integrating advanced feature selection methods with domain-specific knowledge may provide novel insights into energy consumption patterns, enabling more generalizable and scalable predictive models.

Furthermore, future work will explore time-series-specific machine learning models, such as Long Short-Term Memory (LSTM) networks, to better capture temporal dependencies in energy consumption patterns. Combining deep learning approaches with existing feature selection and smoothing techniques could further improve prediction accuracy and adaptability in dynamic environments.

5. Acknowledgements

This research is funded by the Cardiff University-China Scholarship Council joint program.

Reference

1. Harbick, K. and L. Albright. *Comparison of energy consumption: Greenhouses and plant factories*. in *VIII International Symposium on Light in Horticulture 1134*. 2016.
2. Graamans, L., et al., *Plant factories versus greenhouses: Comparison of resource use efficiency.* Agricultural Systems, 2018. **160**: p. 31-43.
3. Cui, X., et al., *Energy consumption prediction and household feature analysis for different residential building types using machine learning and SHAP: Toward energy-efficient buildings.* Energy and Buildings, 2024. **309**: p. 113997.
4. Ouali, N., et al., *Indoor temperature regulation and energy consumption inside a working office in building system using a predictive functional control.* Energy Sources, Part A: Recovery, Utilization, and Environmental Effects, 2024. **46**(1): p. 10473-10493.
5. Ran, J., et al., *Coordinated optimization design of buildings and regional integrated energy systems based on load prediction in future climate conditions.* Applied Thermal Engineering, 2024. **241**: p. 122338.
6. Huo, D., et al., *Mapping smart farming: Addressing agricultural challenges in data-driven era.* Renewable and Sustainable Energy Reviews, 2024. **189**: p. 113858.
7. Ilojianya, V.I., et al., *Data-driven energy management: review of practices in Canada, USA, and Africa.* Engineering Science & Technology Journal, 2024. **5**(1): p. 219-230.

8. Le, T.T., et al., *Harnessing artificial intelligence for data-driven energy predictive analytics: A systematic survey towards enhancing sustainability.* International Journal of Renewable Energy Development, 2024. **13**(2): p. 270-293.

9. Ali, U., et al., *Review of urban building energy modeling (UBEM) approaches, methods and tools using qualitative and quantitative analysis.* Energy and buildings, 2021. **246**: p. 111073.

10. Reinhart, C.F. and C.C. Davila, *Urban building energy modeling–A review of a nascent field.* Building and Environment, 2016. **97**: p. 196-202.

11. Hong, T., et al., *Ten questions on urban building energy modeling.* Building and Environment, 2020. **168**: p. 106508.

12. Guo, Y., et al., *Modeling and optimization of environment in agricultural greenhouses for improving cleaner and sustainable crop production.* Journal of Cleaner Production, 2021. **285**.

13. Graamans, L., et al., *Plant factories; crop transpiration and energy balance.* Agricultural Systems, 2017. **153**: p. 138-147.

14. Liebman-Pelaez, M., et al., *Validation of a building energy model of a hydroponic container farm and its application in urban design.* Energy and buildings, 2021. **250**: p. 111192.

15. Ali, U., et al., *Urban building energy performance prediction and retrofit analysis using data-driven machine learning approach.* Energy and Buildings, 2024. **303**: p. 113768.

16. Afzal, S., et al., *Building energy consumption prediction and optimization using different neural network-assisted models; comparison of different networks and optimization algorithms.* Engineering Applications of Artificial Intelligence, 2024. **127**: p. 107356.

17. Sadaghat, B., S. Afzal, and A.J. Khiavi, *Residential building energy consumption estimation: a novel ensemble and hybrid machine learning approach.* Expert Systems with Applications, 2024. **251**: p. 123934.

18. Cletus, F. and A.E. John, *Comparative Analysis Of Machine Learning Models For Greenhouse Microclimate Prediction.* Brilliance: Research of Artificial Intelligence, 2024. **4**(1): p. 162-175.

19. Jeon, Y.-J., et al., *Machine Learning-Powered Forecasting of Climate Conditions in Smart Greenhouse Containing Netted Melons.* Agronomy, 2024. **14**(5): p. 1070.

20. Liu, G., et al., *A state of art review on time series forecasting with machine learning for environmental parameters in agricultural greenhouses.* Information Processing in Agriculture, 2024. **11**(2): p. 143-162.

21. Van Zyl, C., X. Ye, and R. Naidoo, *Harnessing eXplainable artificial intelligence for feature selection in time series energy forecasting: A comparative analysis of Grad-CAM and SHAP.* Applied Energy, 2024. **353**: p. 122079.

22. Li, G., et al., *Performance assessment of cross office building energy prediction in the same region using the domain adversarial transfer learning strategy.* Applied Thermal Engineering, 2024. **241**: p. 122357.

23. Bullmann, M., et al. *Fast kernel density estimation using Gaussian filter approximation*. in *2018 21st International Conference on Information Fusion (FUSION)*. 2018. IEEE.

24. Cao, Y., et al., *Impact of Derived Features from the Controlled Environment Agriculture Scenarios on Energy Consumption Prediction Model.* Buildings, 2023. **13**(1).