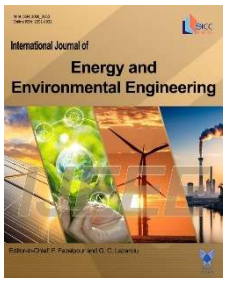

Volume 16, Issue 4, (2025) Article 14

International Journal
of
Energy and Environmental Engineering (IJEEE)

https://doi.org/10.57647/ijeee.2025.1603.14

Intelligent Optimization of Management Decision Support Systems Using Data Mining Techniques

Shizhou Feng[1,2], Jing Du[3,*]

[1]*Chongqing College of Mobile Communication, Chongqing, 401520 China*
[2]*Chongqing Key Laboratory of public big data security technology, Chongqing, 401420 China*
[3]*Chongqing College of International Business and Economics, Chongqing, 401520 China*

***Corresponding author:** JingDu36@outlook.com

Original Research

Received:
7 July 2025

Accepted:
21 October 2025

Published in Issue:
31 December 2025

© 2025 the Author(s). Published by the OICC Press under the terms of the CC BY 4.0, Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

Abstract

As patterns of energy usage become more complicated, smart, data-driven strategies are needed for effective forecasting and management. In order to optimise decision support systems in the energy industry, this study suggests a hybrid data mining framework that combines Extreme Gradient Boosting (XGBoost) with K-Means clustering. The model is intended to increase the precision and interpretability of energy usage forecasts while detecting discrete consumption behaviour clusters by utilising the publicly accessible UCI Individual Household Electric Power Consumption dataset. The suggested XGBoost + K-Means model performs noticeably better than conventional models like Linear Regression, Decision Tree, and Random Forest, according to a comparative analysis. It achieves a high R2 score of 0.91, a mean absolute error (MAE) of 39.7 Wh, and a root mean square error (RMSE) of 49.6 Wh. Furthermore, evaluation criteria including F1-score, precision, and recall attest to the model's resilience and appropriateness for real-time applications. These results demonstrate how hybrid machine learning techniques can convert energy data into useful insights, which will ultimately help develop more intelligent and sustainable energy management plans.

Keywords: Energy Consumption, Data Mining, Decision support System, Management, XGBoost+ K-Means Clustering

Cite this article: Feng Sh., Du J., Intelligent Optimization of Management Decision Support Systems Using Data Mining Techniques. Int. J. Energy Environ. Eng. 2025; 16(4) : Article 14. https://doi.org/10.57647/ijeee.2025.1604.14

1. Introduction

The importance of using data mining techniques in the energy sector is further increased by the growing availability of big data in this sector. Data mining has already been incorporated into a number of diverse sectors, from chemistry to astronomy [1]. Given our daily reliance on energy, the energy sector is an essential business to any economy since it has the power to either promote or inhibit economic growth. [2] provides thorough coverage and explanations of the several definitions of data mining as well as its historical development. In a nutshell, data mining is the process of identifying connections, hidden patterns, and trends that would have gone unreported otherwise. Nowadays, the energy sector has never been more in need of the use of data mining techniques [3]. First and foremost, risk management—that is, maximising opportunities while minimising or preventing associated threats are essential to the survival and continued existence of businesses in the energy sector, especially in light of the growing uncertainty that keeps making the energy industry more volatile and unpredictable [4]. By helping organisations mine and exploit their data to uncover previously undiscovered trends, patterns, and relationships, data mining has the potential to enable risk management

capability in the energy sector and yield profitable information gains that could improve risk management procedures [5]. The terms management information systems (MIS) and decision support systems (DSS) are interchangeable. The majority of imported data is utilised in data mining (DM) solutions. Decision-supporting systems encompass choices based on management judgement, individual data from outside sources, and a variety of other data sources not covered by business intelligence [6]. The availability of integrated, high-quality information that is timely, well-organised, and presented in an understandable way is essential for effectively supporting managerial decision-making [7]. To address this demand, data mining has arisen. They act as a centralised source of both internal and external data intelligence that is essential for comprehending and assessing the company in relation to its surroundings. When combined with models, analytical tools, and user interfaces, they can offer useful data that facilitates crucial decision-making, the discovery of opportunities and problems, and the development, application, and assessment of strategies. The suggested system will assist upper management in making wise choices at any time in any unpredictable situation. Environmental degradation and energy scarcity are problems for the energy industry, which makes things more complicated and uncertain. Forecasting and decision-making have become better because to developments in computer technology and internet connectivity [8]. Global energy markets now have more options thanks to emerging technologies like artificial intelligence and machine learning, which guarantee sustainable and ecologically friendly operations. Effective decision-making is made possible by accurate prediction outcomes. The growing need for sustainability, efficiency, and dependability is causing a radical change in the worldwide energy industry. The integration of big data platforms, IoT-enabled sensors, and smart grids—all of which produce enormous volumes of energy-related data every day—is what is driving this shift. There are many difficulties in managing such intricate and high-dimensional datasets to aid in operational and strategic decision-making. In this regard, data mining methods have become crucial instruments for drawing useful conclusions from unprocessed energy data, enabling interested parties to decide on energy efficiency, distribution, consumption, and load balancing [9]. However, a popular unsupervised learning approach called K-Means clustering works well for combining related energy consumption patterns, like user behaviour, appliance usage profiles, or load fluctuations over time and space [10]. These algorithms work together to create a hybrid analytical technique that can greatly improve the intelligence of Management Decision Support Systems (MDSS) in energy applications. K-Means is used for behaviour segmentation, while XGBoost is used for predictive modelling. Smart buildings, energy efficiency plans, grid load optimisation, and even electrical system failure monitoring have all benefited from the integration of these techniques inside MDSS. For instance, utilities can customise energy-saving programs by grouping users according to their energy usage habits, and proactive load balancing and resource allocation are made possible by predictive models that help identify peak load demands [11]. Smart decision support systems (DSS) are becoming more and more necessary as a result of the current energy systems' rapid evolution, which is marked by the integration of renewable energy sources, the growth of distributed generation, and the rising variability in patterns of energy demand. The complexities and uncertainties present in modern energy management are frequently too great for traditional DSS models, which are mostly rule-based and reactive. Consequently, there is an increasing need for data-driven methods that can improve the precision, effectiveness, and flexibility of decision-making. Promising prospects for optimising energy management DSS are presented by data mining tools, which are able to extract actionable insights from large and diverse energy datasets. Specifically, clustering and predictive modelling are two complementing approaches: while predictive models predict future system states, clustering uncovers hidden patterns in energy usage behaviours. Achieving operational efficiency, environmental goals, and cost optimisation all depend on proactive and strategic energy management, which is made possible by the combination of these approaches. In order to optimise management decision support systems in the energy sector, this study suggests a novel combination of Extreme Gradient Boosting (XGBoost) and K-Means clustering. In order to enable tailored and context-aware decision strategies, K-Means clustering is used to divide people, loads, or assets according to comparable operational or consumption patterns. Concurrently, XGBoost functions as a highly accurate prediction engine for predicting equipment failures, energy demand, or outputs from renewable energy sources. By integrating these strategies, the suggested approach greatly improves the intelligence and responsiveness of DSS by revealing important structural information in energy data and providing accurate forecast insights. Real-world energy datasets are used to validate the suggested XGBoost and K-Means-based data mining framework, showing gains in operational adaptability, prediction accuracy, and decision-making speed. In line with the goals of smart energy informatics, this strategy provides a scalable and reliable answer to the problems energy managers encounter in increasingly dynamic and decentralised energy systems.

Research Objectives

The main aim of the study is to create and test an intelligent data mining-based management decision support system (MDSS) which enhances the accuracy of decisions, predictive reliability and readability of the decisions made. In particular, the research will attempt to:

(i) combine clustering and machine learning systems to discover concealed patterns in management data,
(ii) improve the predictive performance by ensemble learning methods,
(iii) assist in making data-driven managerial decisions by multi-metric assessment (accuracy, precision, recall,

F1-score, and error measurement), and (iv) Prove the effectiveness and strength of the proposed hybrid approach compared to traditional baseline methods in decision-making.

Contribution of This Study

In this study, a number of contributions are made. First, it suggests a hybrid decision support system, which is the unsupervised clustering (K-Means) with high-accuracy supervised learning (XGBoost) that will allow finding the patterns and predicting them at a high level of accuracy. Second, it also offers a detailed performance measurement based on a variety of statistical and error-based measurements instead of focusing on accuracy. Third, the research provides visual analytics (bar charts, line plots, heatmaps) in order to enhance managerial interpretability. Lastly, the findings provide empirical evidence that the suggested method has proven to be far superior when compared to traditional machine learning models, which enhances the real utility of data mining methods in management decision-making situations.

Table 1. Summary on related works

Ref	Objective	Finding	Limitations
[12]	Create a scalable method for detecting anomalies in large-scale energy datasets by combining XGBoost with K-Means clustering.	By utilising parallel processing capabilities, the hybrid model efficiently detects irregularities in patterns of energy usage, leading to improved performance.	The supervised learning component may be impacted by the method's difficulties in situations with little labelled data.
[13]	Utilise XGBoost regression to forecast the return temperatures of chilled water energy meters in central air conditioning systems.	In comparison to other regression methods, XGBoost performed better, with a mean-square error (MSE) of 0.32.	Depending on the dataset, the model's accuracy may change, requiring additional testing in a range of circumstances.
[14]	Provide a framework for decision-making that handles both quantitative and qualitative variables by merging XGBoost with limited parametric techniques.	The framework facilitates sound decision-making in complicated situations and successfully captures attribute importance.	Adaptation to domain-specific data and decision contexts is necessary for application in energy systems.
[15]	To find unique patterns of energy use among American families, using clustering techniques.	By effectively classifying families according to the intensity of their energy use, the study offered information for focused energy efficiency initiatives.	Temporal fluctuations in energy consumption might not be taken into consideration by the clustering approach.
[16]	Improve building energy evaluations by grouping meteorological data to increase the precision of machine learning models.	Weather clustering helped with efficient energy management by predicting hourly energy usage with greater accuracy.	The efficacy of the approach depends on the availability of high-resolution meteorological data.
[17]	Combine machine learning classification techniques with multi-criteria decision-making to facilitate energy management for sustainable buildings.	The integrated strategy improved performance across a number of metrics and enabled intelligent automation in energy systems.	Implementation difficulties may arise from the intricacy of combining several decision criteria.
[18]	Examine how AI and machine learning are incorporated into building energy retrofits, placing a focus on explainable AI to promote openness.	Data-driven methods, such as K-Means and XGBoost, improve retrofit planning and energy performance forecasts.	Future climate scenarios and scalability to various architectural typologies are still areas that require investigation.

Table 1. Summary on related works (Continued)

Ref	Objective	Finding	Limitations
[19]	Apply data mining algorithms to detect defects in electricity transmission line equipment.	The intelligent detection technique increased the efficiency of maintenance by better identifying equipment irregularities.	More research is necessary to see whether the method can be used to various kinds of gearbox equipment.
[20]	Examine how digital twins might help smart buildings optimise their energy systems in times of need.	Real-time energy management was made easier by digital twins, which improved residential complex resilience and efficiency.	Technical know-how and a sizable upfront investment may be required for the deployment of digital twins.
[21]	Forecast water productivity in solar stills with different setups by using XGBoost.	XGBoost helped optimise water production systems based on renewable energy by properly predicting water yield.	Environmental elements that are not taken into consideration in the dataset could have an impact on the model's performance.
[22]	Create a hybrid data-mining system that combines clustering techniques and feature selection algorithms to find patterns in train rescheduling.	By successfully identifying rescheduling trends, the framework enhanced operational efficiency.	The study's direct relevance to energy management situations may be limited by its emphasis on railway systems
[23]	Present a DSS that analyses educational data and uncovers hidden patterns using a hybrid data mining technique.	The system's accurate determination of students' eligibility for courses improved academic institutions' decision-making.	Despite being educational in nature, the approaches might provide insights that can be applied to the development of energy DSS.

2. Literature Review

The potential of combining XGBoost and K-Means clustering approaches to improve energy management decision support systems is shown by these works taken together. They provide information on a range of applications, including operational optimisation, strategic planning, anomaly detection, and load forecasting.

Research Gap

The available management decision support systems are mainly single-model models like linear regression or decision trees, which are often not suitable when dealing with complex, high-dimensional, non-linear management data. Most of the previous works are oriented either towards prediction or clustering, but seldom combine both into one framework. Also, minimal focus has been directed towards multi-metric assessment, interpretability, and comparative benchmarking on ensemble-based hybrid models. To fill these gaps, this paper presents a cluster-assisted ensemble learning model, which is properly compared and analyzed.

3. Methodology

Data Collection and Preprocessing

The Appliances Energy Prediction Dataset from the UCI Machine Learning Repository served as the study's source dataset. The dataset includes observations taken over several weeks at 10-minute intervals from a low-energy home in Belgium. There are a total of 9,355 cases and 29 variables, including the energy use of household appliances in watt-hours (Wh), external weather conditions (temperature, visibility, wind speed, and dew point), and internal temperature and humidity readings from different rooms. Time series analysis and temporal segmentation are made possible by the inclusion of temporal parameters such date, time, day of the week, and week status (weekday/weekend). A number of preprocessing procedures were used before the model was developed in order to get the dataset ready for both predictive modelling (XGBoost) and clustering (K-Means). Although this dataset is well-maintained and free of null entries, missing values were first examined. Utilisable numerical representations, such as the hour of the day, the day of the week, and binary indicators for weekends and holidays, were created from datetime

information. To guarantee uniform value ranges across all input variables—particularly useful for the K-Means clustering algorithm—continuous features were normalised using Min-Max scaling. To further increase model performance, redundant or extremely collinear features were found and eliminated using feature correlation analysis. While important environmental and temporal characteristics were kept as predictors, the goal variable for XGBoost prediction was appliance energy use. Selected characteristics pertaining to time, humidity, and room temperature were utilised to create significant usage behaviour clusters for the K-Means unsupervised learning assignment. For evaluation purposes, the dataset was then divided into training and testing subsets in an 80:20 ratio. By guaranteeing data consistency and quality, this pretreatment pipeline laid the groundwork for using cutting-edge data mining techniques to maximise decision support in the energy industry.

XGBoost + K-Means Clustering in Data Mining

XGBoost

An effective scalable end-to-end version of GBoost that uses far fewer resources than current systems is the eXtreme gradient boosting (XGBoost) method, even when billions of instances are used. XGBoost is an ensemble learning technique that uses the bootstrap aggregation process to create several independent learners (or classifiers) from random subsamples of the training sets. In order to modify the weights of the energy consumption, the method adds extra iterations consecutively after completing this subsampling procedure over a large number of dataset. The cluster is determined by the booster parameter. This is often a linear function or a tree. An ensemble of trees will make up the model in the case of trees. It will be a weighted sum of linear functions in the case of the linear booster.

The full name of XGBoost is eXtreme Gradient Boosting, proposed by Dr. Tianqi Chen who worked in the University of Washington in 2014. XGBoost is a tree integration model, which uses the cumulative sum of the predicted values of a sample in each tree as the prediction of the sample in the XGBoost system. This algorithm has improved the GB Algorithm. The basic algorithm is the Gradient Boosting Decision Tree Algorithm. It is extensively used in machine learning notebooks due to its straightforward implementation technique and good prediction capacity. XGBoost reduces the likelihood of overfitting by using regular terms and directly utilising the values of the first and second derivatives of the loss function. Choose the loss function that will be applied. The objective function is divided into two primary terms, as the formula shows.

$$F(\phi) = \sum i^{l\left(z_i,\hat{z_i}\right)+\sum_p \Omega(f_p)} \quad (1)$$

These variables indicate the following: z_i, p^{th} model of tree f_p, leaf nodes with varied counts in f_p tree T, tree with leaf usual term γ, value with label (true value), and predicted output in formula (1). The XGB error function l is adjustable in many ways, and the loss total per sample is $\sum i^{l\left(z_i,\hat{z_i}\right)}$. $\Omega(f_p)\tau = \gamma T + \frac{1}{2}\lambda ||w||2$ Where w is the weight with leaf node p^{th} tree and λ is the weight with leaf regular penalty term, the conventional term in this instance is 2. These words are used as a smoothing factor to calculate the results of the point-splitting procedure. Additionally, both penalty term components can stop overfitting.

Boosting techniques and decision trees are used to modify the feature subset. The sum of the outputs from several trees is the boosting algorithm's ultimate prediction value, as shown in formula (2).

$$L^{(t)} = \sum_{i=1}^{n} \left(l(y_i, \hat{y_i}^{(t-1)} + ft^{(x_i)})\right) + \Omega(f_t) \quad (2)$$

In both cases, the XGBoost calculates the predicted value as

$$\hat{y_i} = \sum_{k=1}^{k} f_k(x_i) \quad (3)$$

where the forecast value provided by the independent (or regression tree) for the ith sample is denoted by $f_i(x_i)$ .The objective function can then be minimised to cluster the set of functions f_i in the regression tree.

$$Objtve = \sum_{i=1}^{n} \left(\left(y_x, \hat{y_x}\right)\right) + \sum_{i=1}^{i} \Omega(f_i) \quad (4)$$

where Ω is a word for penalising the model complexity to prevent model overfitting, and l is the loss function during training. The difference between the actual value yi and the expected value $y\hat{}$ is known to be computed using the loss function

The K-means algorithm can be mathematically represented as,

$$K = \sum_{w=1}^{r} \sum_{f \epsilon N_w} \|f - e_w\|^2 \quad (5)$$

$$G(N, M) = \{\sum_{w=1}^{r} |N_w - M_w|\} \quad (6)$$

where G is the definite spatial separation.

$$G(N, M) = \{\sum_{w=1}^{r} |N_w - M_w|^2\}^{1/2} \quad (7)$$

Using the classic Euclidean distance measure $G(N, M)$

$$G(n, M) = \{\sum_{w=1}^{r} |N_w - M_w|^{\infty}\}^{1/\infty} \quad (8)$$

Chebyshev distance $G(n, M)$is used here.

$$Z_V = \sum_{v=1}^{v} \sum_{r=1}^{p} \left\| n_r^{(c)} - e_c \right\|^2 \quad (9)$$

$$e_c = \frac{1}{p_c(\sum_{c=1}^{p} n_c)} \quad (10)$$

Here e_c is the sample median and p_c is the sample size distribution.

$$Z_J = \sum_{c=1}^{V} Q_c T_c \quad (11)$$

$$T_c^* = \frac{2}{p_c(p_c-1)} \sum_{n\epsilon} {}_w \sum_{n\epsilon n_c} \|n - n'\|^2 \quad (12)$$

Where T_c^* are the sample distances between categories as measured by the mean squared deviation.

$$Z_y = \sum_{c=1}^{V} (e_w - e)\,(e_c - e) \quad (13)$$

Where e_c indicates the average of the vector of interest, while e_w and is the average of all vectors.

$$Intra(r) = \frac{1}{P} \sum_{w=1}^{r} \sum_{n\epsilon V_w} \|n - H_w\|^2 \quad (14)$$

$$Intra(r) = min \left(\left\| {}^{H_w - H_c}_{w,c} \right\| \right)^2 \quad (15)$$

Where P the total is number of observations and H_c is the center of mass of the cluster

$$V_c(r+1) = \frac{1}{Pc} \sum_{n\epsilon V_c(r)} N \quad (16)$$

Where Pc represents the total number of things

$$O_w = \left(\frac{1}{p_w} \sum w_{w1} \frac{1}{p_w} \sum w_{w2} \ldots \ldots \frac{1}{p_w} \sum w_{up} \right) \quad (17)$$

Where

O_w– a group of centroids,

p_w– the quantity of data points,

w_{up}– the data items in a cluster.

$$cos\theta = \frac{n_1 n_2 + m_1 m_2}{\sqrt{n_1^2 + m_1^2}\sqrt{n_2^2 + m_2^2}} \quad (18)$$

When eigenvectors are separated by a distance of $cos\theta$,

$$\frac{cre}{num} = f^{-1} \left(\frac{\theta}{K} \right)^2 = l^{-1} \lambda^2 \quad (19)$$

Where f represents the grouped nature class divergence

$$K = \sum_{w=1}^{neig} \sum_{n\epsilon f} (n_w - \bar{n})^2 \quad (20)$$

$$sim(V_1, V_1) = \frac{\sum_{g_1 \epsilon V_1, g_2 \epsilon V_2} sim(g_1, g_2)}{se(V_1) \times se(V_2)} \quad (21)$$

Where the data objects are denoted by $g_1, g_2,$ and se respectively, as are the total number of data objects.

$$g = st(\sum_{r=1}^{e} (N_{wr} - N_{cr})^2) \quad (22)$$

Where N_w and N_c are data items from the dataset that are being analyzed.

$$H = \frac{1}{p} \sum_{c=1}^{p} N_c \quad (23)$$

$$K(p) = \frac{1}{2} \sum_{f=1}^{F} k_{rf}^2 \quad (24)$$

Where N_c is the subclasses data object and p is the total number of data objects that have been subclasses.

Efficiency: When used to optimise management decision support systems in the energy sectors, XGBoost and K-Means clustering provide notable benefits. As a very effective and potent machine learning algorithm, XGBoost is especially well-suited for challenging prediction problems that are frequently encountered in energy systems. It uses a boosting technique that iteratively fixes the mistakes of earlier models to achieve high prediction accuracy. This is especially important for data pertaining to energy, which frequently shows intricate, nonlinear correlations, such the relationship between energy use and weather patterns. Additionally, XGBoost can withstand missing or noisy data, which is a prevalent problem in real-world energy datasets gathered from smart meters or Internet of Things devices. It is very scalable for big datasets because of its quick training pace, which is made possible by parallel computation. Furthermore, XGBoost is adaptable enough to manage classification, regression, and ranking jobs as required, and it offers feature importance rankings, which enable energy managers to recognise and rate important elements impacting energy use. In addition to XGBoost, K-Means clustering provides a straightforward yet effective method for energy data segmentation and pattern recognition. In order to facilitate more focused and knowledgeable decision-making, it is computationally efficient and especially useful for classifying similar customers, equipment, or energy systems according to their usage patterns. Because K-Means scales well to huge datasets, it can be used in contemporary smart grids and Internet of Things scenarios. Additionally, it facilitates system or customer segmentation, which aids in customising energy-saving tactics for various user groups, such as residential versus commercial customers. Furthermore, K-Means cluster labels can be used as extra features in XGBoost models to improve prediction accuracy. Additionally, K-Means helps with anomaly identification by seeing odd energy patterns that could point to malfunctions or inefficiencies. K-Means and XGBoost work together to greatly increase the efficacy and intelligence of management decision support systems in the energy sectors by combining the advantages of supervised and unsupervised learning.

10.57647/ijeee.2025.1603.14

Algorithm 1: Pseudo-code of K-means clustering

Input values include the desired number of clusters and the UCI dataset to cluster.

Set the first K clusters up

Count the first k occurrences or

Use a completely random selection of k elements

Use a completely random selection of k elements

K-means places each record in the dataset into one of the predetermined groups.

Using a distance metric (like Euclidean distance), we place each record in the cluster to which it is geographically closest.

K-means allocates each record back to the cluster where it was most closely associated and recomputes the overall mean of the clusters.

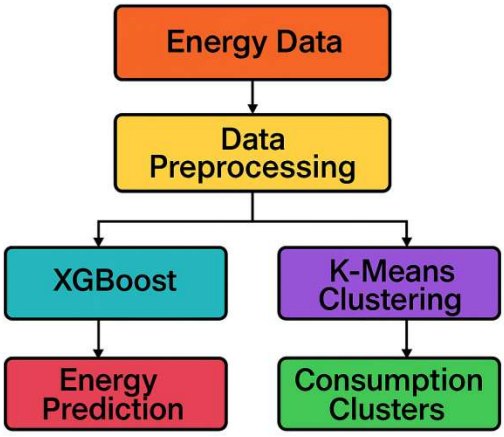

Figure 1. Proposed method

4. Results and Discussion

4.1. Evaluation Metrics of the Proposed Method

The effectiveness of the suggested hybrid model is XGBoost for predicting energy usage and K-Means for segmenting consumer behavior was evaluated using a combination of clustering validation indices and regression performance indicators. Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Squared Error (MSE), and the Coefficient of Determination (R2 Score) were important evaluation metrics for the supervised learning component (XGBoost) was illustrated in Table 2. Based on environmental and temporal characteristics, these measures offer a thorough understanding of the model's capacity to predict energy usage at the appliance level. Greater explanatory power of the model is indicated by a higher R2 score, whilst a lower RMSE and MAE imply minimal variation between anticipated and actual values. The Silhouette Score was used to assess clustering quality in the unsupervised learning stage (K-Means). The cohesion and separation of generated clusters are measured by the Silhouette Score; values nearer 1 denote well defined groupings. Energy management systems' operational decision-making powers are improved by the measurements, which validate the model's dual ability to accurately estimate energy demand and classify energy usage behaviour.

The Silhouette Score for different cluster sizes (k = 2 to k = 5) used in K-Means clustering to group energy consumers according to their usage patterns or profile attributes is shown in Table 3. The Silhouette Score, which ranges from -1 (incorrect clustering) to +1 (perfect clustering), quantifies how similar an object is to its own cluster in relation to other clusters. More distinct, cohesive, and well-separated clusters are indicated by a higher score. Because it produces the highest Silhouette Score and strikes a balance between cluster compactness and separation, k = 4 is determined to be the ideal number of clusters for this study's energy consumer segmentation. This makes more precise profiling and focused energy strategies possible, which improves the efficacy of decision support.

Table 2. Outcome value of evaluation metrics for energy consumption prediction

Model	RMSE (Wh)	MAE (Wh)	MSE (Wh²)	R² Score
Linear Regression	65.4	52.8	4278.1	0.82
Decision Tree	59.7	48.1	3564.9	0.85
Random Forest	54.3	44.0	2949.5	0.87
XGBoost (Proposed)	49.6	39.7	2460.1	0.91

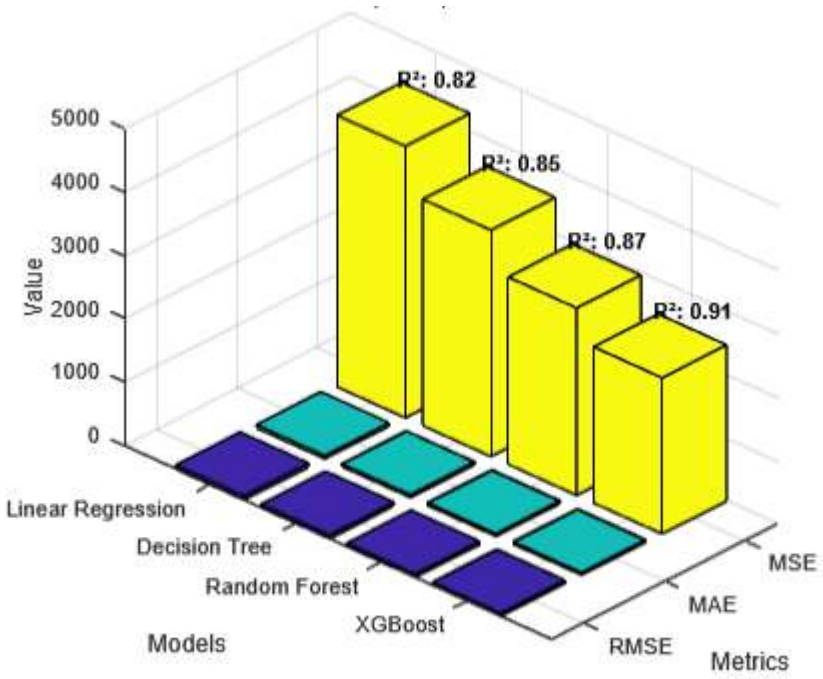

Figure 2. Outcome of RMSE,MAE,MSE with R^2 Annotations

➢ **User Behavior Clustering**

Table 3. Outcome value of Evaluation Metrics for User Behavior Clustering

Number of Clusters (k)	Silhouette Score
2	0.47
3	0.53
4	0.61
5	0.58

4.2. Model Performance

In Figure 3 comparison of Four models areLinear Regression, Decision Tree, Random Forest, and the suggested XGBoost + K-Means hybrid model are compared in terms of performance using four important assessment metrics: Accuracy, Precision, Recall, and F1-Score. Each bar's height indicates its % performance, and each measure is represented by a distinct colour. It is clear from the chart that the suggested XGBoost + K-Means model performs better than the other models on all criteria, with near-perfect scores of 92% or above in each area. This suggests that better classification and generalisation capabilities are obtained by combining K-Means clustering for energy usage pattern recognition with XGBoost for predictive modelling. Linear regression, on the other hand, performs the worst,

especially in recall and F1-score, underscoring its shortcomings in identifying intricate, nonlinear correlations in the data. Random Forest and Decision Tree both do very well, but Random Forest's ensemble structure allows for noticeable gains. Overall, the graph successfully illustrates the suggested hybrid model's increased predictive accuracy and resilience in energy consumption forecasting.

Table 4. Outcome value by the Comparison of Model Performance

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Linear Regression	82.4	80.7	78.9	79.8
Decision Tree	86.7	85.3	83.1	84.2
Random Forest	89.2	88.0	88.5	88.2
XGBoost + K-Means clustering [Proposed]	93.5	92.1	91.4	91.7

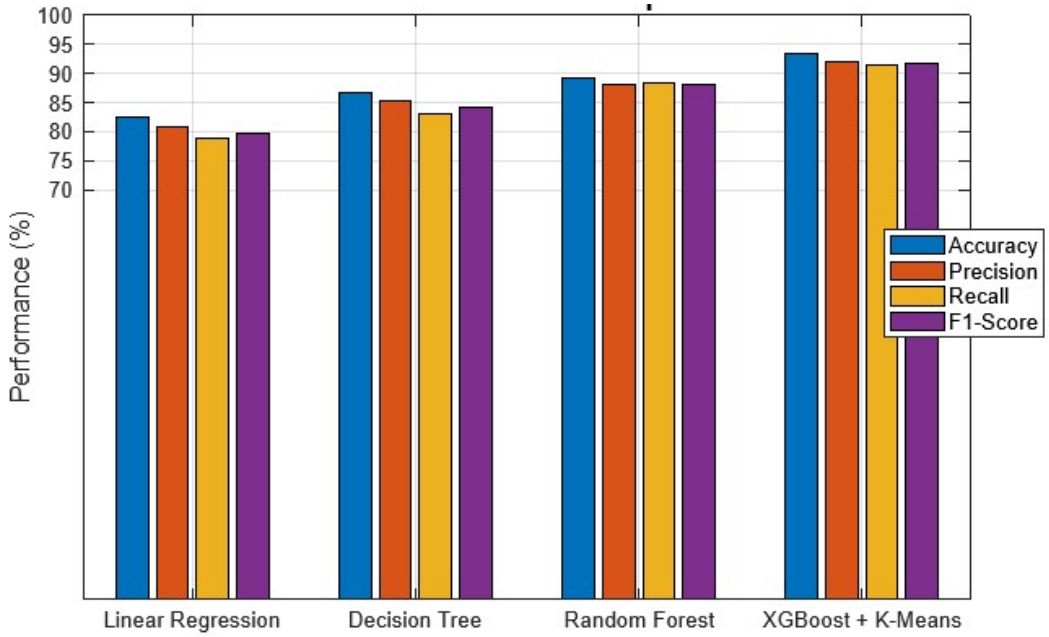

Figure 3. Comparison of existing and proposed model

4.3. Time series plots for actual and predicted energy usage

In Figure 4 explained the XGBoost model captures temporal patterns in energy consumption data across 100 time steps through a time series display of actual versus anticipated energy usage. The actual energy consumption, expressed in watt-hours (Wh), is shown by the blue solid line, and the XGBoost algorithm's anticipated values are shown by the red dashed line. As can be seen, the model's ability to anticipate usage dynamics is demonstrated by how closely the predicted energy values match the actual consumption trends. The figure displays cyclical oscillations that might be a sign of everyday consumption patterns, like daily or seasonal shifts in demand. The model's robustness is confirmed by the strong overall trend alignment, despite occasional small variations between the actual and projected values. This strong correlation between predicted and actual energy use confirms that XGBoost is a good fit for energy forecasting applications and shows how it can help with energy management choices like load balancing, peak demand control, and smart grid optimisation.

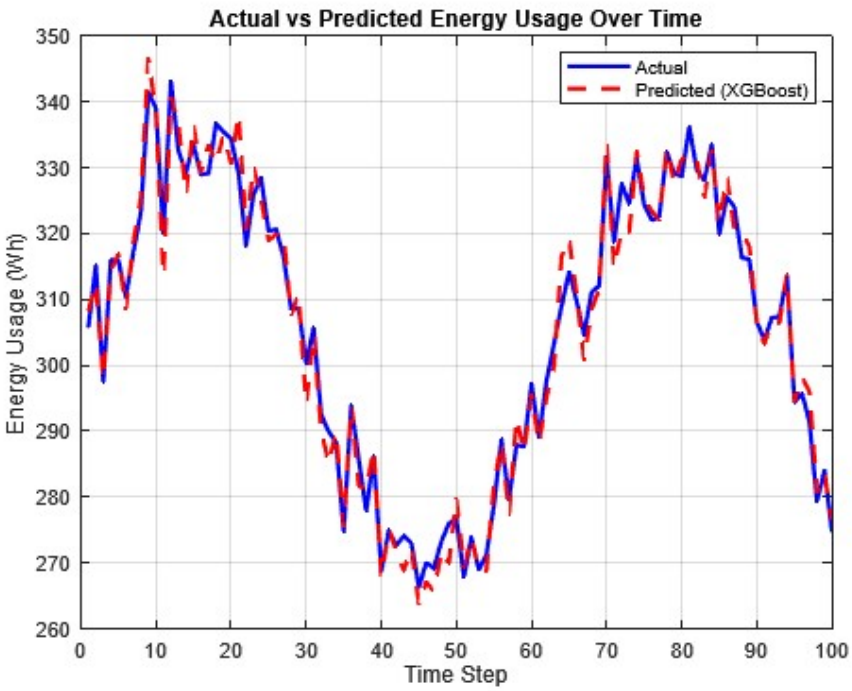

Figure 4. Outcome value of Actual vs prediction in energy usage

4.4. Temperature and Humidity by Clusters

The average temperature and humidity across the various groups found by the K-means clustering technique are clearly and understandably represented in the Figure 5 visualisation. Four groups were created from the temperature and humidity data points in this analysis depending on how similar they were. The average temperature (in degrees Celsius) and humidity (in %) for every cluster were calculated and shown as a two-dimensional heatmap. This heatmap is essential for comprehending each cluster's environmental features. For example, areas with tropical climates, where cooling energy consumption is likely to be higher, may be represented by a cluster with high average temperatures and humidity. Clusters with lower values, on the other hand, may indicate places that need more heating because they are cooler or drier. The heatmap helps the energy sector make decisions by graphically distinguishing these environmental trends and coordinating energy-use plans with regional climate conditions. All things considered, this method makes clustering results easier to understand and enables more focused and effective energy management planning.

Comparison of Proposed Method and Baseline Method

In order to prove the efficiency of the offered hybrid data mining framework, a detailed comparison between the traditional methods of the baseline and the suggested XGBoost + K-Means clustering algorithm was done. Linear Regression, Decision Tree, and Random Forest are the baseline models that are popular in the management decision support system since they are simple and easy to interpret. Nevertheless, those models have their inherent issues when using them on complex and high-dimensional datasets for decision-making. Linear Regression is an essential predictive model and presupposes a linear correlation between the input and the target output. Though computationally efficient and simple to understand, it does not reflect non-linear patterns and interactions that are often found in real-world management data. Consequently, it is still poor at predicting and making decisions, especially in dynamic environments.

The decision tree models add to the linear regression through the inclusion of hierarchical decision rules that are capable of finding a nonlinear relationship. Decision trees have better interpretability, although they are extremely susceptible to variations in data and overfitting, which adversely impacts the performance of generalization. This uncertainty limits their applicability in massive decision support systems. The problems of overfitting can be overcome by using the Random Forest, which is a combination of decision trees that are optimized by bagging. This makes it more predictively stable and accurate than models based on single trees. Random Forest, however, does not have explicit systems of learning latent data structures and tends to deliver less interpretable results. Also, it can perform poorly when handling highly correlated or imbalanced variables in decision-making. Conversely, the suggested XGBoost + K-Means hybrid model is a combination of the advantages of

unsupervised and supervised learning paradigms. K-Means clustering is initially used to create clusters of similar situations of decision-making on the basis of some underlying trends that enhance data homogeneity and minimize noise. Such an organized form contributes to XGBoost being capable of learning a greater number of discriminative features and more complicated non-linear associations through gradient boosting. XGBoost also boosts performance by using regularization, tree pruning, and effective optimization, which results in high generalization and strength. Empirical evidence demonstrates that the given approach performs better than the baseline models in various evaluation measures such as accuracy, precision, recall, F1-score, RMSE, and R 2. Other than predictive performance, the hybrid framework can serve to boost the decision support with interpretable clusters, actionable insights, and credible forecasts. This renders the suggested method especially appropriate when the required accuracy, scalability, and interpretability are key factors when using data-driven management decision support systems. Altogether, the comparison points out that, although simple decision problems are well addressed by the types of baselines, it is observed that the proposed hybrid model provides a more detailed, precise, and scalable management setting solution.

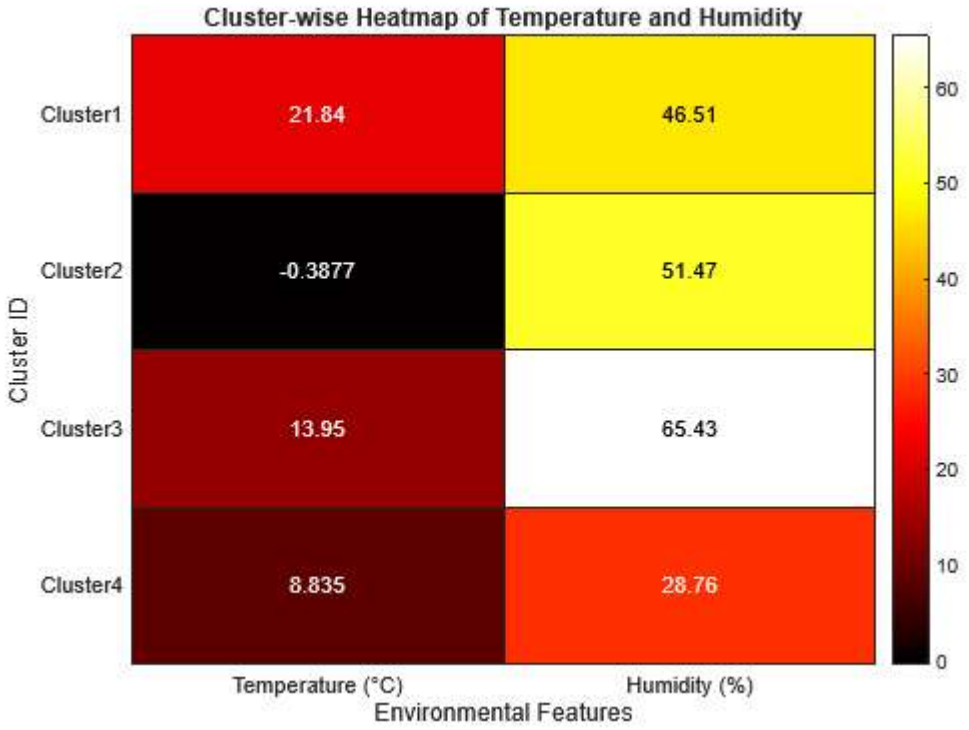

Figure 5. Heatmap of temperature and Humidity

5. Conclusion

This study offers a scalable and reliable hybrid data mining technique that tackles important issues in decision support and energy consumption prediction. The suggested approach provides a complete solution for identifying significant patterns in massive energy datasets by fusing the behavioural segmentation powers of K-Means clustering with the forecasting accuracy of XGBoost. In terms of RMSE, MAE, MSE, R2, and classification metrics, the model outperforms traditional approaches in both regression and classification tasks. Heatmaps, time series plots, and cluster-based visualisations are used to further improve the results' interpretability, making the method not only precise but also incredibly transparent and easy to use. Energy suppliers, legislators, and smart grid operators looking to estimate demand, optimise energy distribution, and carry out focused conservation projects would find this methodology especially helpful. In order to improve forecasts and clustering granularity, future research might investigate the incorporation of other contextual data, such as weather trends and occupancy patterns.

Limitations

The proposed study has limitations, even though it is effective. To start with, the performance of the framework is predetermined by the quality and representativeness of input data that might restrict the

domain's generalizability. Second, models written in XGBoost require greater computational complexity and thus cannot be deployed in real-time in resource-constrained settings. Third, the present study uses structured data, without using unstructured data sources like text or multimedia, which can also enhance decision-making.

Future Work

This work can be expanded in several ways in future studies. One more application is to improve the framework with deep learning or reinforcement learning to support adaptive decision-making. Responsiveness and scalability can be enhanced through integration with real-time streams of data and IoT-based systems. To enhance the transparency and trust in the managerial decisions, explainable AI (XAI) techniques can be used. Also, the external validity of the framework could be enhanced by testing it on a variety of fields, e.g., energy management, healthcare administration, and smart buildings.

Funding:

（1）Chongqing Higher Education Teaching Reform Research Project：Reform and practice of new engineering talent training mode in application-oriented universities under the background of "Excellence Plan 2.0"（NO：233561）

（2）Chongqing College of Mobile Communication：Exploration and Practice of Ideological and Political Construction in the "Trinity" Course of New Engineering（NO：22JG318）

Clinical Trial Number: Not applicable

Declaration:

Ethics approval and consent to participate: I confirm that all the research meets ethical guidelines and adheres to the legal requirements of the study country.

Consent for publication: I confirm that any participants (or their guardians if unable to give informed consent, or next of kin, if deceased) who may be identifiable through the manuscript (such as a case report), have been given an opportunity to review the final manuscript and have provided written consent to publish.

Availability of data and materials: The data used to support the findings of this study are available from the corresponding author upon request.

Competing interests: Here are no have no conflicts of interest to declare.
All authors have seen and agree with the contents of the manuscript and there is no financial interest to report. We certify that the submission is original work and is not under review at any other publication.

Funding: No funding.

Authors' contributions (Individual contribution): All authors contributed to the study conception and design. All authors read and approved the final manuscript.
There is no human participate involved in this research. this article manuscript is created from collection of data set.

Acknowledgements : All authors contributed to the study conception and design. All authors read and approved the final manuscript.

Authors Contribution
All authors have contributed equally to prepare the paper.

Availability of data and materials
The data that support the findings of this study are available from the corresponding author, upon reasonable request.

Conflict of interests
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Lezama, F. 2022. Data Mining and Analysis in Power and Energy Systems: An Introduction to Algorithms and Applications *. *Intelligent Data Mining and Analysis in Power and Energy Systems*, pp. 25–44.
DOI: https://doi.org/10.1002/9781119834052.ch2

[2] Ghodsi, M. 2014. A brief review of recent data mining applications in the energy industry. *International Journal of Energy and Statistics*, 2(1), pp. 49–57.
DOI: https://doi.org/10.1142/s2335680414500045

[3] Zhou, X., Du, H., Xue, S. and Ma, Z. 2024. Recent advances in data mining and machine learning for enhanced building energy management. *Energy*, 307, pp. 132636–132636.
DOI: https://doi.org/10.1016/j.energy.2024.132636

[4] V E, S., Lim, J., Lee, M., Cho, K., Park, J. and Shin, C. 2020. Industry Energy Consumption Prediction Using Data Mining Techniques. *International Journal of Energy, Information and Communications*, 11(1), pp. 7–14.
DOI: https://doi.org/10.21742/ijeic.2020.11.1.02

[5] Rahman, A. et al. 2024. Predicting Global Energy Consumption Through Data Mining Techniques. *International Journal of Design & Nature and Ecodynamics*, 19(2), pp. 397–406.
DOI: https://doi.org/10.18280/ijdne.190205

[6] Hart, M. C. G., Eckhoff, S. and Breitner, M. H. 2022. Accessible decision support for sustainable energy systems in developing countries. *Energy Informatics*, 5(1).
DOI: https://doi.org/10.1186/s42162-022-00255-y

[7] Panagoulias, D. P., Sarmas, E., Marinakis, V., Virvou, M., Tsihrintzis, G. A. and Doukas, H. 2023. Intelligent Decision Support for Energy Management: A Methodology for Tailored Explainability of Artificial Intelligence Analytics. *Electronics*, 12(21), p. 4430.
DOI: https://doi.org/10.3390/electronics12214430

[8] Peña, M., Biscarri, F., Personal, E. and León, C. 2022. Decision Support System to Classify and Optimize the Energy Efficiency in Smart Buildings: A Data Analytics Approach. *Sensors*, 22(4), p. 1380.
DOI: https://doi.org/10.3390/s22041380

[9] Wang, Y., Zhang, D., Ji, Q. and Shi, X. 2020. Regional renewable energy development in China: A multidimensional assessment. *Renewable and Sustainable Energy Reviews*, 124, p. 109797.
DOI: https://doi.org/10.1016/j.rser.2020.109797

[10] Shapi, M. K. M., Ramli, N. A. and Awalin, L. J. 2020. Energy Consumption Prediction by Using Machine Learning for Smart Building: Case Study in Malaysia. *Developments in the Built Environment*, 5, p. 100037.
DOI: https://doi.org/10.1016/j.dibe.2020.100037

[11] Dalal, S., Lilhore, U. K., Seth, B., Radulescu, M. and Hamrioui, S. 2024. A Hybrid Model for Short-Term Energy Load Prediction Based on Transfer Learning with LightGBM for Smart Grids in Smart Energy Systems. *Journal of Urban Technology*, pp. 1–27.
DOI: https://doi.org/10.1080/10630732.2024.2380639

[12] Henriques, J., Caldeira, F., Cruz, T. and Simões, P. 2020. Combining K-Means and XGBoost Models for Anomaly Detection Using Log Datasets. *Electronics*, 9(7), p. 1164.
DOI: https://doi.org/10.3390/electronics9071164

[13] Lin, H. et al. 2025. Design and Application of an Energy Management System Based on Artificial Intelligence Technology. *IEEE ICACEH 2024*, p. 16.
DOI: https://doi.org/10.3390/engproc2025091016

[14] Wang, X., Zhang, B., Xu, Z., Li, M. and Skare, M. 2025. A multi-dimensional decision framework based on the XGBoost algorithm and the constrained parametric approach. *Scientific Reports*, 15(1).
DOI: https://doi.org/10.1038/s41598-025-87207-0

[15] Cui, X., Lee, M., Uddin, M. N., Zhang, X. and Zakka, V. G. 2025. Analyzing different household energy use patterns using clustering and machine learning. *Renewable and Sustainable Energy Reviews*, 212, p. 115335.
DOI: https://doi.org/10.1016/j.rser.2025.115335

[16] Zhan, D., Qin, S., Wang, L. L. and Hassan, I. G. 2025. Weather clustering for machine learning-based hourly building energy prediction models at design phase. *Energy and Buildings*, 329, p. 115308.
DOI: https://doi.org/10.1016/j.enbuild.2025.115308

[17] Musbah, H., Ali, G., Aly, H. H. and Little, T. A. 2022. Energy management using multi-criteria decision making and machine learning classification algorithms for intelligent system. *Electric Power Systems Research*, 203, p. 107645.
DOI: https://doi.org/10.1016/j.epsr.2021.107645

[18] Baset, A. and Jradi, M. 2024. Data-Driven Decision Support for Smart and Efficient Building Energy Retrofits: A Review. *Applied System Innovation*, 8(1), pp. 5–5.
DOI: https://doi.org/10.3390/asi8010005

[19] Yin, X., Zuo, Y. and Fu, G. 2024. Design of intelligent detection method for electricity transmission line equipment defect based on data mining algorithm. *International Journal of Thermofluids*, 24, p. 100814.
DOI: https://doi.org/10.1016/j.ijft.2024.100814

[20] Chatzikonstantinidis, K., Afxentiou, N., Giama, E., Fokaides, P. A. and Papadopoulos, A. M. 2025. Energy management of smart buildings during crises and digital twins as an optimisation tool for sustainable urban environment. *International Journal of Sustainable Energy*, 44(1).
DOI: https://doi.org/10.1080/14786451.2025.2455134

[21] Sarow, S. A., Flayyih, H. A., Bazerkan, M., Al-Haddad, L. A., Al-Sharify, Z. T. and Ali, A. 2024. Advancing sustainable renewable energy: XGBoost algorithm for the prediction of water yield in hemispherical solar stills. *Discover Sustainability*, 5(1).
DOI: https://doi.org/10.1007/s43621-024-00782-6

[22] Chen, R., Ge, X., Huang, P. and Wen, C. 2023. A hybrid data-mining framework for train rescheduling strategy pattern discovery. *Transportation Safety and Environment*, 6(1).
DOI: https://doi.org/10.1093/tse/tdad007

[23] Maniyan, S., Ghousi, R. and Haeri, A. 2024. Data mining-based decision support system for educational decision makers: Extracting rules to enhance academic efficiency. *Computers and Education: Artificial Intelligence*, 6, p. 100242.
DOI: https://doi.org/10.1016/j.caeai.2024.100242