ILKOM Jurnal Ilmiah Vol. 17, No. 3, December 2025, pp. 302-311
Accredited 2nd by KEMDIKTISAINTEK No. 10/C/C3/DT.05.00/2025; E-ISSN 2548-7779 | P-ISSN 2087-1716

Optimizing K-Means Using Greylag Goose Optimization Algorithm for Household Energy Consumption Pattern Segmentation

Florentina Yuni Arini[a,1,*]; Ahmad Rozaq Heryansyah[a,2]; Rahima Ratna Dewanti[a,3]; Rizky Aulia Adi Saputro[a,4]; Awan Saputra Romadhoni[a,5]; Muhammad Lutfi Wibowo[a,6]; Ikhsan Ardiansyah[a,7]; Poomin Duankhan[b,8]

[a] *Universitas Negeri Semarang, Sekaran, Gunung Pati, Semarang, Central Java, Indonesia*
[b] *Khon Kaen University, Khon Kaen, Thailand*
[1]*floyuna@mail.unnes.ac.id,* [2]*ahmadrozaq45@students.unnes.ac.id,* [3]*rahimaratna@students.unnes.ac.id,* [4]*awansaputrar@students.unnes.ac.id,* [5]*rizky1243@students.unnes.ac.id,* [6]*lutfey33@students.unnes.ac.id,* [7]*notanyone172@students.unnes.ac.id,* [8]*mini@kkumail.com*
** Corresponding author*

Article history: Received July 16, 2025; Revised August 26, 2025; Accepted December 04, 2025; Available online December 15, 2025

Abstract

Electricity is a crucial resource in everyday life, and rising household energy demand requires smarter monitoring and management approaches. Analyzing consumption data enables the discovery of typical energy usage behaviors that support efficient resource planning. Clustering techniques are widely used to group usage profiles without predefined categories, with K-Means being one of the most popular methods because of its speed and practical implementation. However, this algorithm is highly dependent on the initial centroid selection and may generate inaccurate grouping results if trapped in local optima. To overcome these drawbacks, this research combines K-Means with the Greylag Goose Optimization (GGO) algorithm, a nature-inspired metaheuristic that simulates the adaptive navigation and social coordination of migratory grey geese. By enhancing both exploration and exploitation, GGO improves the accuracy of centroid placement and overall clustering performance. The research utilized Individual Household Electric Power Consumption dataset, which consists of minute-by-minute measurements of several electrical attributes. After preprocessing and exploratory analysis, clustering was executed using three approaches: conventional K-Means, GGO, and a hybrid K-Means–GGO model. Based on the Silhouette Score evaluation, clustering performance improved significantly from 0.6236 with standard K-Means to 0.9675 using the hybrid approach. The resulting segmentation provides deeper insights into household consumption behaviors.

Keywords: Electric Consumption; Clustering; K-Means; Household Energy; Optimization

Introduction

Electrical energy is a primary need in modern life and plays a vital role in daily household activities [1]. As residential electricity demand increases, energy providers and policymakers face the challenge of ensuring efficient energy distribution and consumption. A critical step in addressing this challenge is understanding consumer behavior through consumer usage data [2]. Recent advances in data analytics enable the extraction of meaningful patterns from large data sets. In particular, clustering algorithms are often used to classify energy consumption patterns without prior indicators [3].

Among clustering algorithms, K-Means is a common and efficient method for identifying similar usage behaviors within a population [4], [5], [6]. However, K-Means has weaknesses such as sensitivity to the initial centroid selection and the possibility of getting stuck in local solutions [7], [8]. Therefore, additional approaches are needed to optimize the clustering results. One approach is the optimization algorithm, a systematic procedure for finding the best solution to a problem based on given criteria. In this context, metaheuristic algorithms are widely used because they can handle complex optimization problems, are flexible, and do not require information on the derivatives of the objective function [9]. Metaheuristics combine the principles of exploration (searching for solutions broadly) and exploitation (sharpening the best solution) to find a solution that is close to optimal in a limited time [10]. More specifically, nature-inspired algorithms are also developing, namely, algorithms inspired by phenomena or behaviors in nature, such as biological evolution [11], animal social behavior [12], and natural physical mechanisms [13]. One of the nature-inspired algorithms used in this study is Greyhold Goose Optimization (GGO), which mimics the migration patterns and social structure of grey geese to find the best position in the environment. GGO has a dynamic leadership mechanism and adaptive exploration so that it does not quickly get stuck in local minima and can explore a more

http://dx.doi.org/10.33096/ilkom.v17i3.2851.302-311

expansive solution space [14]. The competitive ability of GGO allows it to improve the performance of the K-means clustering algorithm.

Data collection in this study consists of time-series measurements of household power consumption, including total power usage and sub-measurements for specific devices or rooms. The data are sourced from the Individual Household Electric Power Consumption dataset [15] and will be grouped based on general usage patterns. This study aims to identify key consumption patterns that can support energy-saving and load-balancing strategies. The expected result is a segmentation of daily power consumption into groups that reflect various household energy behavior profiles. These findings are expected to contribute to the design of smart home systems, to direct energy-saving campaigns, and to support more accurate electricity demand forecasting mechanisms.

Method

The workflow for this study begins with data collection from appropriate sources, as shown in **Figure 1**. Once acquired, the dataset undergoes a cleaning process [16] to remove errors [17], duplicates [18], and irrelevant values [19]. Following this, Exploratory Data Analysis (EDA) [20] is performed to examine data characteristics, reveal underlying patterns, and derive initial insights. The K-Means algorithm is then applied as the initial clustering technique. Subsequently, an optimization stage is carried out by integrating K-Means with the Great Green Optimization (GGO) method to enhance clustering performance. The outcome of this workflow is an optimized clustering result that aligns with the research objectives.

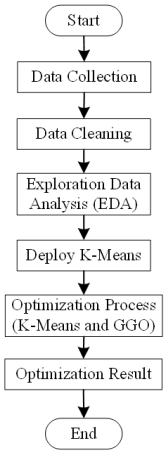

Figure 1. Flowchart Proposed Research Method

A. Data Collection

The data used in this study comes from the UC Irvine Machine Learning Repository. The dataset can be accessed publicly through an online repository at the UC Irvine Machine Learning Repository with the link https://archive.ics.uci.edu/dataset/235/individual+household+electric+power+consumption [15]. This dataset contains household electricity consumption data, including energy consumption and usage timeframes.

B. Data Cleaning

This stage aims to ensure that the data is clean, consistent, and valid before use [21]. Missing values and duplications are removed to prevent distortion of the analysis results. The process of normalizing numerical features is also used to ensure each feature has a balanced scale and that no feature dominates the clustering process.

C. Exploratory Data Analysis

The Exploratory Data Analysis (EDA) stage aims [22], [23] to understand the fundamental characteristics of energy consumption data before clustering. The primary focus is on identifying temporal patterns of consumption. The initial step involves converting the data to a time format [24]. Next, key metrics, including total energy consumption and daily/weekly/monthly consumption patterns (visualized through time-series plots and boxplots), are explored, along with consumption variability (measured by standard deviation or interquartile range). The EDA results provide initial insights to guide the segmentation process and segment interpretation [25].

D. K-Means Clustering Algorithm

Clustering algorithms are used to group data into several groups based on their similarity or characteristics [26], [27]. K-Means Clustering is one of the most frequently used clustering techniques in data analysis. This algorithm clusters or segments the data into K groups, where K is the number of groups of interest. This process begins with the random selection of group centroids in the data space, then groups each data point in the group with the following focus: The center of each group is then calculated based on the average data point of the group, and this process is repeated until there is no change in the arrangement of data points in the group and the specified iteration limit is reached. Due to its speed and scalability, K-means clustering is often used for customer segmentation, document clustering, image analysis, and other applications, but the results can be affected by the choice of initial centroids. In addition to its speed and scalability, K-means clustering can handle large datasets effectively. It is suitable for analyzing large datasets, as is common in business and scientific applications. In practice, K-Means is often used after the data is normalized so that each feature has the same scale, so that no feature dominates the clustering process [28]. In the initial stage, the K-Means algorithm randomly initializes K centroids (cluster centers). The distance from each data point to each centroid is then calculated using the Euclidean distance formula [29], [30] as denoted in (1).

$$d(Z,W) = \sqrt{\sum_{K=1}^{d} (Zk - Zw)^2} \tag{1}$$

where d is the number of dimensions, Z is the p-th data point in the k-th dimension, and W is the centroid of the j-th cluster. After all data is clustered to the nearest centroid, the centroid position [31] is updated by calculating the average of all data in that cluster by (2).

$$M = \frac{Z1 + Z2 + \cdots + Zn}{n} \tag{2}$$

where M is the new centroid, n is the number of data in the cluster, and Z is the nth data value in the cluster. These steps are repeated iteratively: the data are assigned to the closest cluster, and the centroids are recalculated until the centroid positions change little or until the maximum iteration limit is reached. This process aims to minimize variation within clusters and maximize differences between clusters. These steps are shown in pseudocode in Algorithm 1.

Algorithm 1. K-Means Algorithm

Initialize random centroids

Select k randomly as initial centroids

Iterasi=0

while iterasi < MaxIterasion and CriteriaStop = false **do**

  for each data point x_i d**o**

    Calculate distance x_i from each centroid.

    Determine the nearest centroid.

    Assign point x_i to the cluster.

  end for

  updatePoint = 0

  for each cluster j = 1 to k **do**

    Update the centroid position as the average of all data points assigned to the cluster.

    Calculate the updated centroid position.

    if update position centroid > updatePoint then

      updatePoint = update position centroid

    end if

end for

iterasi = iterasi + 1

end while

The K-Means algorithm, as implemented in Algorithm 1, initializes the clustering process by randomly selecting centroids from the dataset. The iterative phase then involves calculating the Euclidean distance between each data point and all centroids. Based on this distance metric, each data point is associated with the nearest cluster centroid. Centroid positions are updated at each iteration by averaging the vector positions of all clustered data points. Algorithm convergence is evaluated based on changes in centroid positions between iterations; iterations continue until the shift falls below a specified tolerance or the maximum number of iterations is reached.

E. Silhouette Score

Silhouette Score [32] is an evaluation metric used to assess the quality of clustering algorithm results, such as K-Means. This metric measures how well each data point fits within its own cluster compared to other nearby clusters.

F. Determine K Value in Clusters Using the Elbow Method

Clustering requires a k value with a function to determine the best number of clusters. One method for determining the number of clusters is the elbow method [33], a heuristic method that relies on visual representation. This method uses an elbow-based criterion, the point closest to 90°, to determine the optimal number of clusters. Using clustering requires a k-value and a function to determine the best number of clusters. One method for determining the number of clusters is the elbow method, a heuristic method that relies on visual representation. This method uses an elbow-based criterion, the point closest to 90°, to determine the optimal number of clusters.

G. Greylag Goose Optimization (GGO) Algorithm

Greylag Goose Optimization (GGO) algorithm [14] is the name given to the optimization strategy proposed in this study. The Greylag Goose Optimization (GGO) algorithm is a metaheuristic inspired by the migratory behavior and social patterns of grey goose flocks, combining exploration and exploitation to find optimal solutions in complex search spaces. GGO begins by randomly initializing a population of solutions. Iteratively updates the solution positions based on the flock's social interactions and navigation, aiming to maximize global search while refining the best solution found. The GGO algorithm begins by generating many individuals at random. Each individual offers a potential solution to be considered for inclusion in the solution pool for the problem. In each iteration, the number of possible solutions in each group is dynamically adjusted based on the best solution found so far. The goose explorers will hunt for new location, interesting places to investigate in the area around their current position as present in (3).

$$X_{new} = X_{best} - (2ar_1 - a)|2r_2(X_{best} - X_t)| \quad (3)$$

where X_{new} denote as new position while X_{best} and X_t exhibit best position and current position of the geese. The values of r_1 and r_2 are randomly changing within the boundaries of $[0,1]$. The parameter a linearly decrease from 2 to zero. The responsibility for improving the current solution rests with the exploitation team. At the end of each cycle, GGO determines which participants have reached the highest fitness level and rewards them accordingly.

Results and Discussion

This research analyzed household energy consumption patterns. The analysis was conducted by grouping energy consumption patterns into specific segments with similar characteristics.

A. Data Collection

The *Individual Household Electric Power Consumption* dataset contains minute-by-minute measurements of residential electricity usage collected from a single household. The dataset records several electrical quantities, including global active and reactive power, voltage, current intensity, and three sub-metering variables that monitor specific appliance groups, such as kitchen equipment, laundry machines, and water-heating or air-conditioning systems. The dataset contains 9 (nine) features, as shown in **Table 1**.

Table 1. Feature Dataset

No.	Attribute	Data Type	Description/Units
1	Date	Time	Date/format dd/mm/yyyy

No.	Attribute	Data Type	Description/Units
2	Time	Time	Time/format hh:mm:ss
3	Global_active_power	Real	Household global minute-averaged *active power*/**kilowatts (kW)**
4	Global_reactive_power	Real	Household global minute-averaged *reactive power*/in **kilowatts (kW)**
5	Voltage	Real	Minute-averaged voltage/**volts (V)**
6	Global_intensity	Real	Household global minute-averaged current intensity/**amperes (A)**
7	Sub_metering_1	Real	Energy sub-metering No. 1 / **watt-hours of active energy (Wh)**
8	Sub_metering_2	Real	Energy sub-metering No. 2 / Wh. Corresponds to the laundry room (refrigerator, washing machine, light, tumble-drier).
9	Sub_metering_3	Real	Energy sub-metering No. 3 / Wh. Corresponds to an air-conditioner electric water heater and an electric water heater.

B. Data Cleaning

This stage cleans and processes data in a DataFrame. To produce accurate analyses and predictive models, the data-cleaning stage is crucial for ensuring a clean, consistent, and valid dataset. Removing missing values is essential to prevent analysis errors caused by incomplete data. Removing duplicate rows is necessary to maintain the validity of analysis results by avoiding double-counting observations, which can distort data distributions and model reliability. Data normalization is also performed for numeric columns.

C. Exploratory Data Analysis

In addition to the data cleaning stage, the next step is exploratory data analysis to determine variations in energy consumption patterns. In the exploratory data analysis (EDA) stage, visualization is performed to understand the distribution and relationships between variables in the household energy consumption dataset. The daily electricity consumption trends are shown in **Figure 2.** Daily household electricity consumption data from January 2007 to December 2008 demonstrate variations in energy use. The analyzed data comprises seven main variables: global active power, global reactive power, voltage, current intensity, and energy consumption. This data was collected from three different zones within the home, referred to as submetering zones 1, 2, and 3.

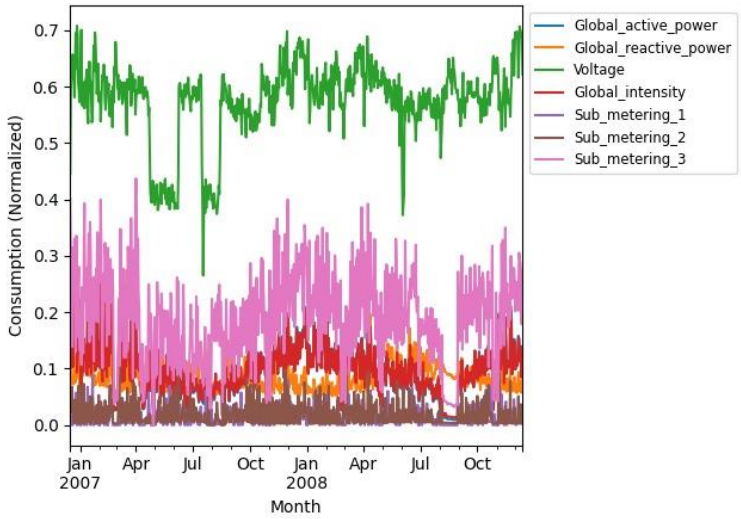

Figure 2. Daily Electricity Consumption Trend Diagram

In general, active power consumption patterns worldwide tend to be consistent over time, indicating relatively constant baseline energy use. Global reactive power changes slightly when devices that generate magnetic fields, such as electric motors and transformers, are used. Because these two variables are technically interconnected, current intensity exhibits a trend in line with active power. Over time, the voltage drops significantly. This drop can be caused by system disruptions or significant load fluctuations in the network. This drop is quite substantial compared to other periods, indicating stability. Submetering zone 3 shows an irregular, sharp increase, likely due to high-power appliances such as water heaters or air conditioners being used irregularly. Submetering 1 and 2 show relatively consistent usage in certain zones, indicating more frequent use, likely for kitchen appliances and the laundry room.

This analysis is important to identify daily electricity consumption patterns. Furthermore, it can be used to conserve energy and design a more efficient household electrical system.

Meanwhile, the distribution of active power worldwide exhibits a right-skewed pattern, with the highest concentration in the low-power range, particularly between 0 and 2 kilowatts, as shown in **Figure 3**. The highest frequency is recorded very close to zero, indicating that low-power conditions dominate daily activities. Above this value, the frequency gradually decreases, punctuated by small spikes between 1 and 2 kilowatts, indicating a consumption pattern. The number of occurrences drops sharply above 3 kilowatts, and there are almost none above 6 kilowatts. This pattern indicates low to moderate energy use, with high loads occurring only occasionally or in specific situations. The use of medium-power household appliances, such as water heaters, electric ovens, or washing machines, can cause secondary peaks. This characteristic indicates a stable level of energy consumption during daily activities, as most of the time is spent in low-power conditions. The uneven pattern can serve as a clue to identify unusual or anomalous behavior in the electrical system. Understanding this distribution helps with load planning, energy savings, and power efficiency evaluation. Visual representations like these are essential for aiding data-driven decision-making without relying on subjective interpretations.

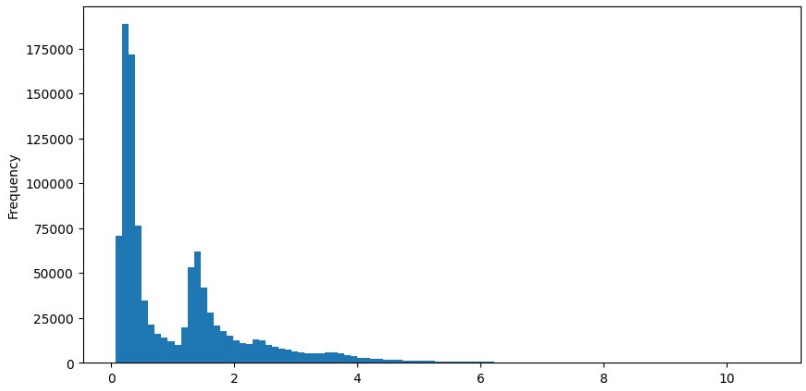

Figure 3. Distribution Global Active Power

For a correlation heatmap of numerical variables, the analysis of electrical attributes and consumption variables reveals a distinct relationship between the two (**Figure 4**). There is a strong correlation between global active power and global intensity, with a value of nearly 1. This indicates a direct relationship between active power and global electric current intensity in the system. Since more power used means more current flows, this relationship makes sense. Small negative correlations between voltage and several other variables, such as Global_intensity and Sub_metering_3, indicate that while a relationship exists, the effect is not very strong. There is a correlation between voltage and Global_active_power and Global_reactive_power, indicating that when the load increases, voltage tends to decrease.

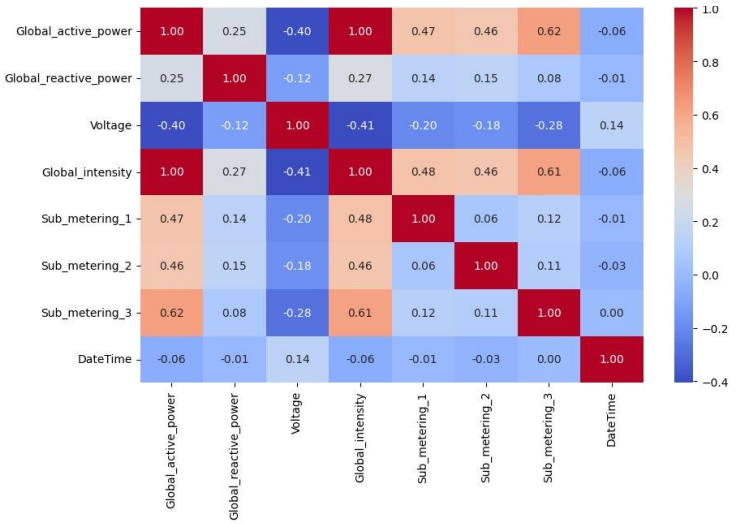

Figure 4. Heatmap of Correlation between Numerical Variables

There is a low correlation between sub-metering, indicating that each measurement point serves a different type of load or is used at various times for other purposes. There is a very weak relationship between the electrical variables and the time variable (DateTime), suggesting that consumption trends are more influenced by daily or weekly patterns than by time. The distribution of these correlations can identify variables that affect each other. They can also be used to create predictive models of energy consumption or to aggregate loads for electrical system efficiency [29].

D. Segmentation (Clusterization) using K-Means, GGO, and K-Means with GGO

The clustering process was performed using the K-Means algorithm after preprocessing and normalization. To determine the optimal number of clusters, the Elbow method was used. Based on the visualization results of the Elbow curve shown in **Figure 5,** where the optimal point was obtained when the value of k = 3 for regular K-Means, while after optimization using Greylag Goose Optimization (GGO), the best k value was found at k = 15.

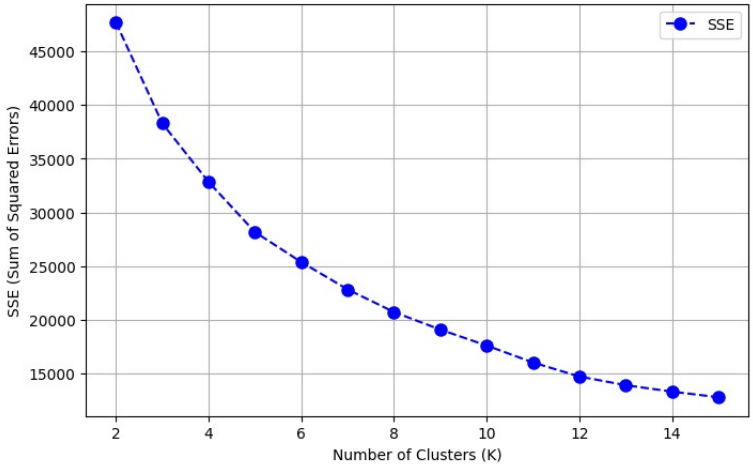

Figure 5. Elbow method for determining Optimal K

The determination of the number of clusters in the K-Means analysis was carried out using the Elbow Strategy approach by observing the SSE (Sum of Squared Errors) value against the variation in the number of clusters (K). In the graph, the decrease in the SSE value appears very significant at the initial number of clusters, then slopes down as K increases. The point of transition from a sharp to a gentler decline is considered the ideal number of clusters. Based on this pattern, the initial ideal number of clusters was identified at K = 3. The results of applying K-Means with 3 clusters yielded the following grouping of energy consumption data. Cluster 1 has low and stable energy consumption, reflecting households with minimal electrical activity. Cluster 2 shows moderate energy consumption, with a spike pattern in the morning and evening, typical in daily routines. Cluster 3 shows high energy consumption, especially during peak hours. Peak hours correspond to households with high electrical device activity or intensive use [19]. This cluster represents households with high activity or intensive electrical devices.

Although this division provides useful initial information, the level of segmentation detail remains limited. Therefore, further optimization was performed using the Greylag Goose Optimization (GGO) method. This method combines two solution-finding mechanisms, namely exploration (searching for new possibilities) and exploitation (improving the best solution), thus allowing the identification of more representative centroid positions and determining a more ideal number of clusters [20]. After applying GGO, the ideal number of clusters increased significantly to K = 15, as indicated by the continued downward trend in SSE thereafter. This clustering with 15 groups resulted in a more detailed and specific understanding of energy consumption behavior. Some clusters reflect spikes in energy use on weekends, indicating more intense household activity during holidays. Other clusters show high consumption at night, while some exhibit erratic energy use patterns that may originate from small- and medium-sized businesses or special activities.

The final evaluation of clustering performance was performed using the Silhouette Scores of the regular K-Means, GGO, and GGO-optimized K-Means methods to assess the quality of cluster separation and the consistency within each cluster. The evaluation results support the conclusion that GGO-based optimization provides sharper, more relevant segmentation of energy consumption behavior. The results obtained are shown in **Figure 6**.

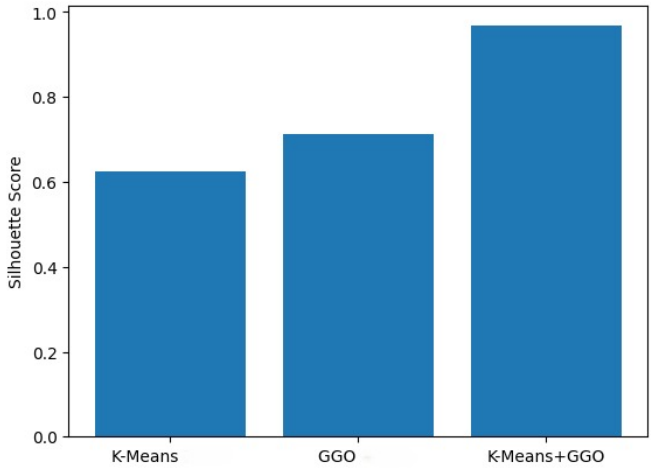

Figure 6. Result Experiment Based on Silhouette Score Values

The regular K-Means with the number of clusters K = 3 produces a Silhouette Score value of 0.6236 or 62.36%. This value indicates that the separation between clusters is quite good, but there is still room for improvement in segmentation accuracy. When the Greylag Goose Optimization (GGO) method is used independently, without integration with K-Means, the Silhouette Score increases to 0.7736 (77.36%), indicating a significant improvement in cluster separation accuracy compared to regular K-Means. Furthermore, when GGO is used to optimize K-Means, resulting in a combination of K-Means + GGO with the ideal number of clusters K = 15, a Silhouette Score of 0.9675 (96.75%) is obtained. This value indicates a very high, clearly defined quality of separation between clusters, reflecting a cluster structure that is highly representative of the energy consumption information pattern. Based on the evaluation results, the cross-breed method of K-Means and Greylag Goose Optimization (GGO) not only produces accurate clusters but also creates meaningful segmentation structures. The resulting clusters provide a deep understanding of energy usage habits, including patterns based on time, frequency, and intensity of use. This understanding has important implications in various contexts.

The first implication is that it facilitates the design of smart home systems; the clustering results can be used to configure home automation systems that adjust device operation based on user energy activity [28]. Second, in energy efficiency campaigns, clustering information can be used to target households with high energy consumption or energy-wasting behavior with more targeted education and interventions. Third, in the context of energy load prediction, the identified cluster patterns can serve as important inputs for more accurate modeling and prediction of spikes or drops in electricity consumption.

Overall, the evaluation and analysis results indicate that the combination of the K-Means and GGO algorithms has excellent potential for extracting valuable insights from large-scale energy consumption data. These findings are not only relevant in an academic context but also very useful as a basis for decision-making in the development of innovative energy technology and adaptive automation systems. The regular K-Means (K=3) produces a Silhouette Score of 0.6236 (62.36%), although it shows quite good separation, this score indicates there is room for improvement. This can be improved with optimization algorithms, such as the GGO algorithm, which can achieve up to 96.75% improvement.

Conclusion

Based on the analysis results, the K-Means algorithm successfully identified three distinct household energy consumption patterns: low-stable, medium with temporal spikes, and high during peak hours. However, optimization with the Greylag Goose Optimization (GGO) Algorithm significantly improved segmentation resolution, yielding 15 more specific clusters rich in information on variations in consumption behavior, including weekend, evening, and irregular usage patterns. This improved clustering quality is supported by the Silhouette Score value, which increased sharply from 0.6236 to 0.9675 after GGO optimization, indicating much better separation between clusters. The implications of this more detailed segmentation of energy consumption patterns are very significant for the development of practical applications. The results of this study can be used to design smart home systems that are responsive to user energy habits, target energy efficiency campaigns more effectively based on specific consumption profiles, and improve the accuracy of energy load prediction models.

References

[1] B. Zohuri and P. McDaniel, "The Electricity: An Essential Necessity in Our Life," *Adv. Smaller Modul. React.*, pp. 1–21, 2019, doi: 10.1007/978-3-030-23682-3_1.

[2] K. Zhou and S. Yang, "Understanding household energy consumption behavior: The contribution of energy big data analytics," *Renew. Sustain. Energy Rev.*, vol. 56, pp. 810–819, 2016, doi: 10.1016/j.rser.2015.12.001.

[3] X. Gao and A. Malkawi, "A new methodology for building energy performance benchmarking: An approach based on intelligent clustering algorithm," *Energy Build.*, vol. 84, pp. 607–616, 2014, doi: 10.1016/j.enbuild.2014.08.030.

[4] Stuart~P.~Lloyd, "Least squares quantization in {PCM}," *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129–136, 1982.

[5] M. A. W. J. A. Hartigan, "Algorithm AS 136 A K-Means Clustering Algorithm," *J. R. Stat. Soc. Ser. B Methodol.*, vol. 28, no. 1, pp. 100–108, 2012.

[6] R. Wu, "Behavioral analysis of electricity consumption characteristics for customer groups using the k-means algorithm," *Syst. Soft Comput.*, vol. 6, 2024, doi: 10.1016/j.sasc.2024.200143.

[7] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," *Inf. Sci. (Ny).*, vol. 622, pp. 178–210, 2023, doi: 10.1016/j.ins.2022.11.139.

[8] M. Suyal and S. Sharma, "A Review on Analysis of K-Means Clustering Machine Learning Algorithm based on Unsupervised Learning," *J. Artif. Intell. Syst.*, vol. 6, no. 1, pp. 85–95, 2024, doi: 10.33969/ais.2024060106.

[9] O. E. Turgut, M. S. Turgut, and E. Kırtepe, "A systematic review of the emerging metaheuristic algorithms on solving complex optimization problems," *Neural Comput. Appl.*, vol. 35, no. 19, pp. 14275–14378, 2023, doi: 10.1007/s00521-023-08481-5.

[10] K.-L. Du and M. N. S. Swamy, "Search and Optimization by Metaheuristics," *Search Optim. by Metaheuristics*, 2016, doi: 10.1007/978-3-319-41192-7.

[11] N. Gupta, M. Khosravy, N. Patel, and I. Sethi, "Evolutionary Optimization Based on Biological Evolution in Plants," *Procedia Comput. Sci.*, vol. 126, pp. 146–155, 2018, doi: 10.1016/j.procS.2018.07.218.

[12] F. Y. Arini, F. R. N. Saputra, Liafathra, R. D. R. Artama, and S. Pichai, "Optimizing Heart Disease Prediction: A Collaborative Approach of Support Vector Regression and Grey Wolf Optimizer," *2025 Int. Conf. Smart Comput. IoT Mach. Learn. SIML 2025*, 2025, doi: 10.1109/SIML65326.2025.11080978.

[13] T. Alexandros and D. Georgios, "Nature Inspired Optimization Algorithms Related to Physical Phenomena and Laws of Science: A Survey," *Int. J. Artif. Intell. Tools*, vol. 26, no. 6, 2017, doi: 10.1142/S0218213017500221.

[14] E. S. M. El-kenawy, N. Khodadadi, S. Mirjalili, A. A. Abdelhamid, M. M. Eid, and A. Ibrahim, "Greylag Goose Optimization: Nature-inspired optimization algorithm," *Expert Syst. Appl.*, vol. 238, p. 122147, 2024, doi: 10.1016/j.eswa.2023.122147.

[15] G. Hebrail and A. Berard, "Individual Household Electric Power Consumption," *UCI Mach. Learn. Repos.*, 2006.

[16] O. Alotaibi, E. Pardede, and S. Tomy, "Cleaning Big Data Streams: A Systematic Literature Review," *Technologies*, vol. 11, no. 4, 2023, doi: 10.3390/technologies11040101.

[17] P. Martins, F. Cardoso, P. Váz, J. Silva, and M. Abbasi, "Performance and Scalability of Data Cleaning and Preprocessing Tools: A Benchmark on Large Real-World Datasets," *Data*, vol. 10, no. 5, 2025, doi: 10.3390/data10050068.

[18] M. Ravikanth, S. Korra, G. Mamidisetti, M. Goutham, and T. Bhaskar, "An efficient learning based approach for automatic record deduplication with benchmark datasets," *Sci. Rep.*, vol. 14, no. 1, 2024, doi: 10.1038/s41598-024-63242-1.

[19] F. Gröger *et al.*, "SelfClean: A Self-Supervised Data Cleaning Strategy," 2023.

[20] S. Dhummad, "The Imperative of Exploratory Data Analysis in Machine Learning," *Sch. J. Eng. Technol.*, vol. 13, no. 01, pp. 30–44, 2025, doi: 10.36347/sjet.2025.v13i01.005.

[21] A. M. Sharifnia, D. E. Kpormegbey, D. K. Thapa, and M. Cleary, "A Primer of Data Cleaning in Quantitative Research: Handling Missing Values and Outliers," *J. Adv. Nurs.*, 2025, doi: 10.1111/jan.16908.

[22] H. I. Patel and D. Patel, "Exploratory Data Analysis and Feature Selection for Predictive Modeling of Student Academic Performance Using a Proposed Dataset," *Int. J. Eng. Trends Technol.*, vol. 72, no. 11, pp. 131–143, 2024, doi: 10.14445/22315381/IJETT-V72I11P116.

[23] W. A. Prastyabudi and Isa Hafidz, "Energy Consumption Data Analysis: Indonesia Perspective," *J. Comput. Electron. Telecommun.*, vol. 1, no. 1, 2020, doi: 10.52435/complete.v1i1.47.

[24] C. Ragupathi, S. Dhanasekaran, N. Vijayalakshmi, and A. O. Salau, "Prediction of electricity consumption using an innovative deep energy predictor model for enhanced accuracy and efficiency," *Energy Reports*, vol. 12, pp. 5320–5337, 2024, doi: 10.1016/j.egyr.2024.11.018.

[25] D. Muyulema-Masaquiza and M. Ayala-Chauvin, "Segmentation of Energy Consumption Using K-Means: Applications in Tariffing, Outlier Detection, and Demand Prediction in Non-Smart Metering Systems," *Energies*, vol. 18, no. 12, 2025, doi: 10.3390/en18123083.

[26] A. E. Ezugwu *et al.*, "A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects," *Eng. Appl. Artif. Intell.*, vol. 110, 2022, doi: 10.1016/j.engappai.2022.104743.

[27] G. J. Oyewole and G. A. Thopil, "Data clustering: application and trends," *Artif. Intell. Rev.*, vol. 56, no. 7, pp. 6439–6475, 2023, doi: 10.1007/s10462-022-10325-y.

[28] C. Wongoutong, "The impact of neglecting feature scaling in k-means clustering," *PLoS One*, vol. 19, no. 12, 2024, doi: 10.1371/journal.pone.0310839.

[29] A. Zhu, Z. Hua, Y. Shi, Y. Tang, and L. Miao, "An improved k-means algorithm based on evidence distance," *Entropy*, vol. 23, no. 11, 2021, doi: 10.3390/e23111550.

[30] H. Yin, A. Aryani, S. Petrie, A. Nambissan, A. Astudillo, and S. Cao, "A Rapid Review of Clustering Algorithms," 2024.

[31] Z. Deng, Y. Wang, and M. M. Alobaedy, "Federated k-means based on clusters backbone," *PLoS One*, vol. 20, no. 6 June, 2025, doi: 10.1371/journal.pone.0326145.

[32] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, no. C, pp. 53–65, 1987, doi: 10.1016/0377-0427(87)90125-7.

[33] E. Umargono, J. E. Suseno, and S. . Vincensius Gunawan, "K-Means Clustering Optimization Using the Elbow Method and Early Centroid Determination Based on Mean and Median Formula," 2020, doi: 10.2991/assehr.k.201010.019.