Bulletin of Informatics and Data Science
Vol. 5 No. 1, May 2026, Page 77−85
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v5i1.155
https://ejurnal.pdsi.or.id/index.php/bids/index

Analysis of Electricity Consumption Patterns a Hybrid Algorithm K-Means Clustering and Support Vector Machine

Fika Saputri*, Asminar, Tambi, Mustarum Musarudddin, Muhammad Nadzirin Anshari Nur, Adhi Setiawan Samsul

Faculty of Engineering, Electrical Engineering, Halu Oleo University, Kendari, Indonesia
Email: [1,*]fikasaputri29@gmail.com, [2]asminar.ft@uho.ac.id, [3]tambi@uho.ac.id, [4]mustarum@uho.ac.id, [5]nadzirin@uho.ac.id, [6]adhisetiawansamsul@uho.ac.id

ARTICLE INFORMATION

ARTICLE HISTORY:
Submited : April 28, 2026
Revised : April 30, 2026
Accept : April 30, 2026
Publish : April 30, 2026

KEYWORD

K-Means Clustering;
Support Vector Machine;
Load Pattern;
Energy Consumption;
Machine Learning

CORRESPONDENCE AUTHOR

Email: fikasaputri29@gmail.com

A B S T R A C T

Fluctuations in electricity consumption make it difficult to manage and control energy use efficiently. Therefore, an analytical method capable of accurately identifying electricity consumption patterns is needed. Previous studies have generally applied clustering and classification separately and relied mainly on historical data, limiting their ability to capture dynamic electricity consumption patterns and classify new observations. This study aims to analyze electricity consumption patterns using a machine learning approach based on K-Means Clustering and Support Vector Machine (SVM). The data used consist of historical data and 24-hour real-time data obtained from PT PLN (Persero) UP3 Kendari. The research stages include data preprocessing, feature engineering, standardization using Z-scores, the clustering process using K-Means, and classification using SVM in a hybrid approach. The novelty of this study lies in using K-Means cluster labels as target classes for SVM while combining historical and real-time data. The results show the formation of four clusters: Cluster 0 representing stable moderate consumption, Cluster 1 representing high consumption, Cluster 2 representing low consumption, and Cluster 3 representing fluctuating consumption. Evaluation yielded a Silhouette Score of 0.5095 and a Davies-Bouldin Index of 0.6868, indicating fairly good cluster quality. The SVM model achieved an accuracy of 99.23% with high precision, recall, and F1-score values. These results demonstrate that the hybrid approach is effective in improving analysis performance and producing a reliable model for classifying new electricity consumption data.

1. INTRODUCTION

Rapid technological advancements over the past few decades have driven an increase in the use of electrical appliances in everyday life. This increased use of electronic and electrical devices has resulted in a significant and continuous rise in energy consumption. Consequently, several challenges have emerged in energy management, including rising operational costs, imbalances in electricity load distribution, and potential disruptions to the electrical system that can lead to instability in energy supply. Dynamic and fluctuating electricity consumption patterns have the potential to add to the complexity of energy management. These fluctuations often result in high peak loads, which not only disrupt the stability of the power grid but can also degrade the quality of electricity distribution services provided to consumers[1], [2]. Therefore, it is crucial to adopt an analytical approach capable of analyzing electricity consumption patterns more accurately and efficiently to optimize the electricity distribution system.

Currently, machine learning-based approaches have been widely applied to analyze electricity consumption data. Machine learning is capable of processing large and complex datasets more quickly and accurately than traditional methods. One of the most commonly used methods in electricity consumption analysis is K-Means Clustering, which is an unsupervised learning technique for grouping data based on specific common characteristics [3], [4]. The main objective of K-Means is to partition data into a predetermined number of clusters by minimizing the within-cluster sum of squared distances between each data point and its assigned centroid. The algorithm works iteratively by assigning each data point to the nearest centroid and then recalculating the centroid based on the mean of the assigned data points. Because K-Means relies on distance calculations, data standardization is important when variables have different numerical scales to prevent variables with larger numerical scales from dominating the clustering process [3], [4]. K-Means can help identify electricity consumption patterns based on user behavior, enabling the segmentation of customer groups based on their consumption patterns. Therefore, K-Means is highly useful for electricity customer segmentation as well as load analysis, which is essential for planning more efficient energy distribution. However, K-Means does not inherently learn a supervised decision function for classifying previously unseen observations. Therefore, an additional classifier is required to learn the generated cluster labels and assign new data to the identified consumption patterns [5].

To address the limitations of K-Means, another approach using the Support Vector Machine (SVM) method was applied in this context. SVM is a supervised learning technique that is highly effective in classifying data with high accuracy. SVM works by constructing an optimal hyperplane that can separate data into several classes based on relevant

Bulletin of Informatics and Data Science
Vol. 5 No. 1, May 2026, Page 77−85
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v5i1.155
https://ejurnal.pdsi.or.id/index.php/bids/index

characteristics. The optimal hyperplane is determined by maximizing the margin between different classes, while the observations closest to the decision boundary are referred to as support vectors. For data that cannot be separated linearly, SVM can use kernel functions to map the input data into a higher-dimensional feature space. This mechanism enables SVM to produce a classification model with good generalization capability for previously unseen data [6]. In this study, the cluster labels generated by K-Means are used as target classes for training the SVM model, enabling new electricity consumption data to be classified into the identified consumption patterns [5]. This hybrid approach, which combines clustering and classification, has shown potential for improving data analysis performance, particularly in electrical energy data processing [5], [7], [8]. With this hybrid method, the clustering process can be performed first to identify the data structure, followed by a classification process aimed at improving prediction accuracy on new data, as well as enhancing the model's ability to handle the dynamics of electricity consumption that change over time.

Previous studies have shown that clustering methods are highly effective in identifying electricity consumption patterns based on customer load characteristics [9], [10]. For example, by grouping consumers with similar consumption patterns, electricity providers can more easily predict energy demand and reduce peak loads, which are often a problem in energy distribution. The quality of these clustering results can be evaluated using metrics such as the Silhouette Score and the Davies-Bouldin Index, which measure the extent to which the formed clusters exhibit clear and structured separation. These metrics help ensure that the resulting clusters do not overlap and can be effectively utilized in subsequent classification processes. On the other hand, classification methods such as SVM have also proven effective in various data analysis applications, including energy analysis and electrical systems. SVM has been used to predict load demand, optimize energy allocation, and improve the reliability of electrical systems [11]. Although hybrid clustering-classification approaches have been reported, their application to feeder-level electricity consumption data integrating historical records with 24-hour real-time observations remains limited. In addition, one of the main gaps in existing research is the lack of optimal utilization of real-time data. Most previous studies have relied solely on historical data, which is insufficient to capture the dynamics of changes in electricity consumption patterns occurring in real time. In fact, electricity consumption is heavily influenced by various external factors such as weather changes, holidays, or even special events that can affect consumption behavior. Therefore, it is important to combine historical data with real-time data in an analytical model to better capture these dynamics and make more accurate predictions about electricity consumption patterns.

This study aims to develop and test a hybrid approach that combines K-Means Clustering and Support Vector Machine (SVM) to analyze electricity consumption patterns. This approach will utilize historical data as well as 24-hour real-time data obtained from PT PLN (Persero) UP3 Kendari. With this hybrid approach, the study is expected to improve the identification of representative electricity consumption patterns and produce an effective classification model for assigning new observations to the identified consumption-pattern classes. Additionally, this study is expected to make a greater contribution to more efficient energy management, support more optimal power distribution planning, and provide a reference for future research in the field of machine learning-based energy analysis.

Overall, this study aims to address existing gaps in the literature by integrating clustering and classification methods into a more comprehensive approach. The results of this study are expected to help improve the efficiency of the electricity distribution system in a manner that is more adaptive and responsive to changing consumption patterns, as well as provide a stronger foundation for decision-making in overall energy management.

2. RESEARCH METHODOLOGY

2.1 Research Phases

This study employs a machine learning-based data analysis approach to identify electricity consumption patterns [3], [5]. A hybrid clustering-classification framework is adopted to integrate unsupervised pattern identification with supervised classification [12]. The research stages used in this study are shown in Figure 1. The dataset consists of electricity consumption records obtained directly from PT PLN (Persero) UP3 Kendari.

The research process begins with dataset input, followed by preprocessing, which includes cleaning the data of invalid values, handling missing values, and removing duplicate data. Next, feature engineering is performed to create derived variables relevant to improving the quality of the analysis [13]. The data is then divided into training data (80%) and test data (20%). This division aims to ensure that the built model can be tested using data that has never been used before. After that, data standardization is performed using the Z-score method to normalize the scale across variables, thereby preventing bias in the distance calculation process within the clustering algorithm. The next step is applying K-Means Clustering to the training data to group the data based on the similarity of electricity consumption characteristics. The optimal number of clusters is determined using the Elbow method and Silhouette Score. The clustering results are then used as labels for the classification stage. Next, a Support Vector Machine (SVM) model is trained using the training data and the resulting cluster labels. This sequential clustering-classification procedure follows the hybrid K-Means–SVM approach [14]. K-Means was selected because it can identify groups with similar electricity consumption characteristics in unlabeled data [15], whereas SVM was selected because it can learn the resulting cluster labels and classify previously unseen observations [6]. The electrical variables used in the analysis also represent power-utilization characteristics that are relevant to electricity-load assessment [16]. The resulting SVM model is used to predict labels on

Bulletin of Informatics and Data Science
Vol. 5 No. 1, May 2026, Page 77−85
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v5i1.155
https://ejurnal.pdsi.or.id/index.php/bids/index

the test data. The final stage is model evaluation, where clustering is evaluated using the Silhouette Score and Davies-Bouldin Index (DBI), while classification is evaluated using accuracy, precision, recall, and F1-score.

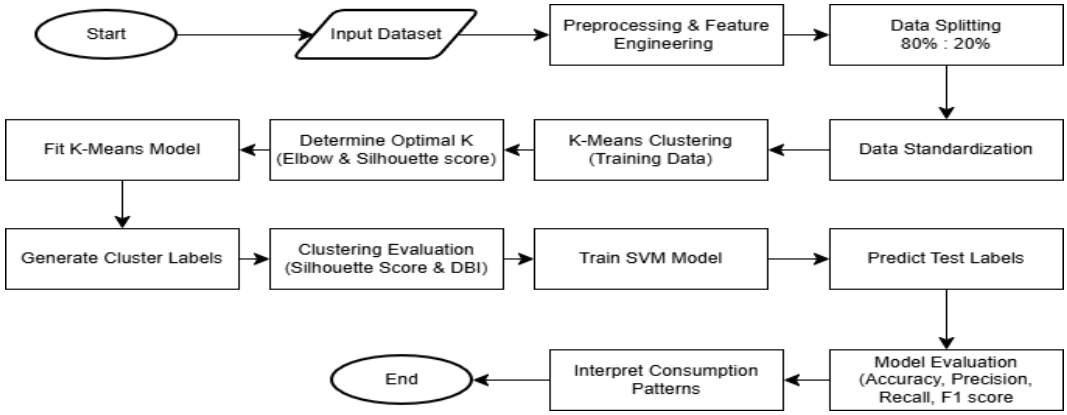

Figure 1. Model Exploration Flowchart

2.2 K-Means Clustering

K-Means Clustering is an unsupervised learning method used to group data based on the similarity of their characteristics and has been applied to identify electricity consumption patterns among customer groups [15]. The clustering process involves determining initial centroids and calculating the distance between each data point and each centroid using the Euclidean distance, as shown in Equation (1):

$$d(\mathbf{x}_i, \boldsymbol{\mu}_k) = \sqrt{\sum_{j=1}^{p} \left(x_{ij} - \mu_{kj}\right)^2} \tag{1}$$

where $d(\mathbf{x}_i, \boldsymbol{\mu}_k)$ is the distance between observation $i$and the centroid of cluster k, $x_{ij}$is the value of feature $j$for observation i, $\mu_{kj}$is the centroid value of feature $j$in cluster k, and $p$is the total number of features. Each observation is assigned to the nearest centroid, after which the centroid is recalculated based on the mean of the observations assigned to the corresponding cluster. This process is performed iteratively until the centroid values no longer change significantly.

2.3 Support Vector Machine (SVM)

Machine learning methods have been applied to smart-grid and electricity-load analysis because of their ability to process complex consumption data [17]. Support Vector Machine is a supervised learning method used for data classification [6], [18]. SVM works by finding an optimal hyperplane that separates data into different classes, as shown in Equation (2):

$$f(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}) + b \tag{2}$$

where $f(\mathbf{x})$is the decision function for input vector $\mathbf{x}$, $\mathbf{w}$is the weight vector that determines the orientation of the hyperplane, $\phi(\mathbf{x})$represents the transformation of the input data into a higher-dimensional feature space induced by the kernel function, and $b$is the bias. The optimal hyperplane is determined by maximizing the margin between different classes, while the observations closest to the decision boundary are referred to as support vectors [6]. In this study, the cluster labels generated by K-Means are used as target classes for training the SVM model [14]. This method is suitable for handling high-dimensional data and has good classification capability [18].

2.4 Model Evaluation

Model evaluation was conducted to measure the performance of the clustering and classification methods. During the clustering stage, the Silhouette Score was used to assess the quality of data grouping. A value closer to 1 indicates that an observation is appropriately assigned to its cluster. The Silhouette Score is calculated using Equation (3):

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{3}$$

where $a(i)$ is the average distance between observation $i$and all other observations within the same cluster, while $b(i)$ is the minimum average distance between observation $i$and observations in the nearest different cluster. In addition, the Davies-Bouldin Index was used to measure cluster compactness and separation. The DBI is calculated using Equation (4):

$$\text{DBI} = \frac{1}{K} \sum_{i=1}^{K} \max_{j \neq i} \left(\frac{S_i + S_j}{M_{ij}} \right) \tag{4}$$

where K is the total number of clusters, S_i and S_j represent the within-cluster dispersion values of clusters i and j, and M_{ij} is the distance between the centroids of clusters i and j. A smaller DBI value indicates more compact and better-

Bulletin of Informatics and Data Science
Vol. 5 No. 1, May 2026, Page 77−85
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v5i1.155
https://ejurnal.pdsi.or.id/index.php/bids/index

separated clusters [15]. During the classification stage, evaluation was performed using accuracy, precision, recall, and F1-score [19]. Accuracy measures the proportion of correctly classified observations and is calculated using Equation (5):

$$\text{Accuracy} = \frac{\sum_{k=1}^{K} TP_k}{N} \tag{5}$$

Precision for class $k$is calculated using Equation (6):

$$\text{Precision}_k = \frac{TP_k}{TP_k + FP_k} \tag{6}$$

Recall for class $k$is calculated using Equation (7):

$$\text{Recall}_k = \frac{TP_k}{TP_k + FN_k} \tag{7}$$

The F1-score for class k is calculated using Equation (8):

$$F1_k = 2\left(\frac{\text{Precision}_k \cdot \text{Recall}_k}{\text{Precision}_k + \text{Recall}_k}\right) \tag{8}$$

where $K$is the total number of classes, $N$is the total number of observations in the test data, $TP_k$is the number of true-positive predictions for class k, $FP_k$is the number of false-positive predictions for class k, and $FN_k$is the number of false-negative predictions for class k.

3. RESULT AND DISCUSSION

The electricity consumption data used in this study were obtained from PT PLN (Persero) UP3 Kendari for the period from 2023 to 2025. The dataset compilation was carried out in stages, starting with the use of historical data as the main basis, followed by calibration using 24-hour real-time data so that the derived parameters refer to the characteristics of a single daily cycle (24 hours). The result of this stage is a final daily dataset per feeder, which is then used in the modeling process. The dataset consists of several key variables, such as current, active power, active energy, power factor, and peak load, which are used to identify electricity consumption patterns. Before analysis is conducted, the data first undergoes a preprocessing stage which includes format standardization and column validation, data cleaning, removal of incomplete observations in key variables, handling of missing values, and removal of duplicate data. This process aims to ensure that the data used in modeling has good quality and minimizes noise that could affect the analysis results.

After the preprocessing stage, a total of 5,305 data points were obtained and ready for use in the modeling process. This relatively large dataset offers an advantage in the model training process, as the more data used, the better the model's ability to recognize electricity consumption patterns. Additionally, the diversity of the data allows the model to capture more complex variations in electricity consumption patterns.1800 words min.

The initial stage of the analysis involved determining the optimal number of clusters using the Elbow method and the Silhouette Score. The Elbow method was used to examine how the Within-Cluster Sum of Squares (WCSS) value changes with respect to the number of clusters. The elbow point on the graph indicates the optimal number of clusters that is, the point at which increasing the number of clusters no longer results in a significant decrease in the WCSS value. Meanwhile, the Silhouette Score is used to evaluate the quality of separation between clusters, where a value close to 1 indicates that the data are in the correct clusters. The results of these two methods are shown in Figure 2.

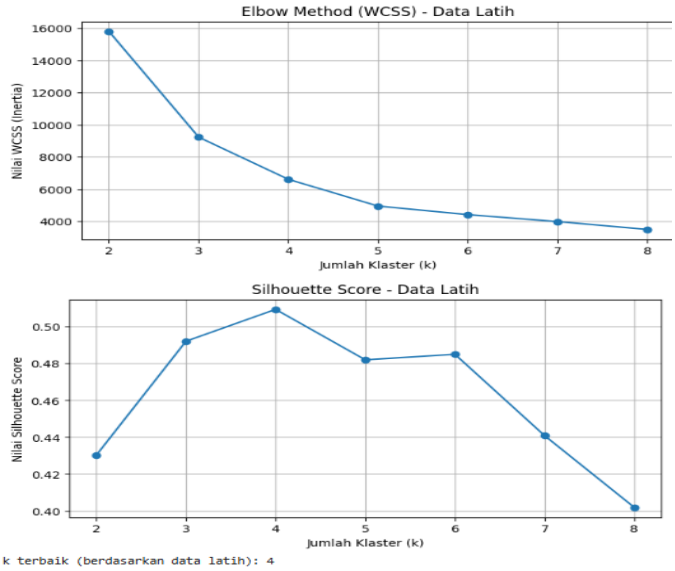

Figure 2. Graph of the Elbow Method and Silhouette Score

Bulletin of Informatics and Data Science
Vol. 5 No. 1, May 2026, Page 77−85
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v5i1.155
https://ejurnal.pdsi.or.id/index.php/bids/index

Figure 2 shows a graph of the Elbow Method based on WCSS values. In this graph, the X-axis represents the number of clusters (k) tested, ranging from k=2 to k=8, while the Y-axis represents the WCSS values, which indicate the degree of data variation relative to the centroid in each cluster. Based on the graph, it can be seen that the WCSS value decreases quite sharply from k=2 to around k=4. After that, the decrease in the WCSS value tends to level off until k=8. This pattern of change forms an elbow point at k=4, indicating that increasing the number of clusters beyond that value does not significantly improve cluster compactness. Thus, k=4 can be considered the optimal number of clusters based on the Elbow Method.

Next is the evaluation graph using the Silhouette Score. In this graph, the X-axis represents the number of clusters (k), while the Y-axis represents the Silhouette Score, which ranges from -1 to 1. The Silhouette Score is used to measure how well data is placed within its cluster compared to other clusters. A value close to 1 indicates that the data is in the correct cluster, while a value close to 0 indicates that the data is located at the boundary between clusters. Based on the graph, the highest Silhouette Score is obtained at k=4 with a value of 0.5095. After that value, the Silhouette Score tends to decrease as the number of clusters increases, indicating that the quality of cluster separation becomes less optimal.

Based on a combination of the results from the Elbow Method and the Silhouette Score, it was determined that the optimal number of clusters is k=4. This value was chosen because it provides the best balance between cluster compactness (based on WCSS) and the quality of separation between clusters based on the Silhouette Score.

3.1 Clustering Results Using K-Means

To interpret the characteristics of each cluster, the mean values of the variables for each cluster were calculated (cluster profiles). A summary of the cluster profiles is shown in Figure 2.

cluster	tegangan_rata2	arus_rata2	daya_aktif	energi_aktif	faktor_daya	frekuensi_sistem	beban_puncak	load_factor	durasi_puncak_jam	jumlah_pelanggan
0	21000.0	116.241	3.555	85.329	0.85	50.0	4.446	80.197	19.247	15338.874
1	21000.0	180.977	5.535	132.850	0.85	50.0	7.590	73.316	17.596	22829.772
2	21000.0	40.991	1.254	30.091	0.85	50.0	1.770	71.592	17.182	9784.242
3	21000.0	91.057	2.785	66.842	0.85	50.0	5.828	47.974	11.514	20018.098

Figure 2. Cluster Profile Summary

Figure 2 shows the differences in characteristics among clusters based on the variable of electricity consumption. In general, each cluster has the following characteristics:

a. Cluster 1 (High Consumption)
   This cluster has the highest average values for current, active power, active energy, and peak load compared to the other clusters. This indicates that this group has a high level of electricity consumption.
b. Cluster 2 (Low Consumption)
   This cluster exhibits the lowest values for electricity consumption variables. This indicates that this group has a low level of energy usage.
c. Cluster 0 (Stable Moderate Consumption)
   This cluster falls into the moderate category with a relatively high load factor. This indicates that electricity usage in this cluster tends to be stable over time, without significant fluctuations.
d. Cluster 3 (Fluctuating Consumption)
   This cluster has an unstable consumption pattern, characterized by a significant difference between the average load and the peak load. This indicates a fairly high level of fluctuation in electricity usage.

To clarify the characteristics of each cluster at the feeder level, one representative feeder was selected from each cluster based on the largest amount of data within that cluster. A visualization of the representative feeder profiles for each cluster is shown in Figure 3.

As shown in Figure 3, each cluster exhibits distinct load pattern characteristics, with representative feeders illustrating these variations. Cluster 0 is represented by the P_RSUD BAHTERAMAS feeder, with 1,012 data points. Its load profile shows relatively stable values, starting at 3.66 MW at 10:00 AM, increasing to 3.91 MW at 12:00 PM, dropping slightly to 3.84 MW at 6:00 PM, and rising again to 3.94 MW at 7:00 PM. This pattern indicates a stable medium-level load with a peak at 7:00 PM.

Cluster 1 is represented by the P_PUULORO feeder with 886 data points. Its load value increased consistently, starting at 5.51 MW at 10:00 AM, rising to 6.20 MW at 12:00 PM, surging to 7.58 MW at 6:00 PM, and peaking at 8.34 MW at 7:00 PM. This pattern indicates a rise in high power consumption as the day progresses toward evening.

Cluster 2 is represented by the P_BORO-BORO feeder with 1,014 data points. The observed load increased gradually, from 1.32 MW at 10:00 AM, 1.42 MW at 12:00 PM, 1.53 MW at 6:00 PM, to 1.64 MW at 7:00 PM. This pattern illustrates stable, low power consumption compared to the other clusters.

Cluster 3 is represented by P_PUULORO with 78 data points. Its load shows sharp fluctuations, starting at 1.81 MW at 10:00 AM, increasing to 2.60 MW at 12:00 PM, surging to 4.75 MW at 6:00 PM, and peaking at 6.32 MW at 7:00 PM. This pattern indicates a sharp change in load as evening approaches. Interestingly, the P_PUULORO feeder appears

Bulletin of Informatics and Data Science
Vol. 5 No. 1, May 2026, Page 77−85
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v5i1.155
https://ejurnal.pdsi.or.id/index.php/bids/index

as a representative in more than one cluster, indicating that a single feeder can exhibit different load patterns across different observations.

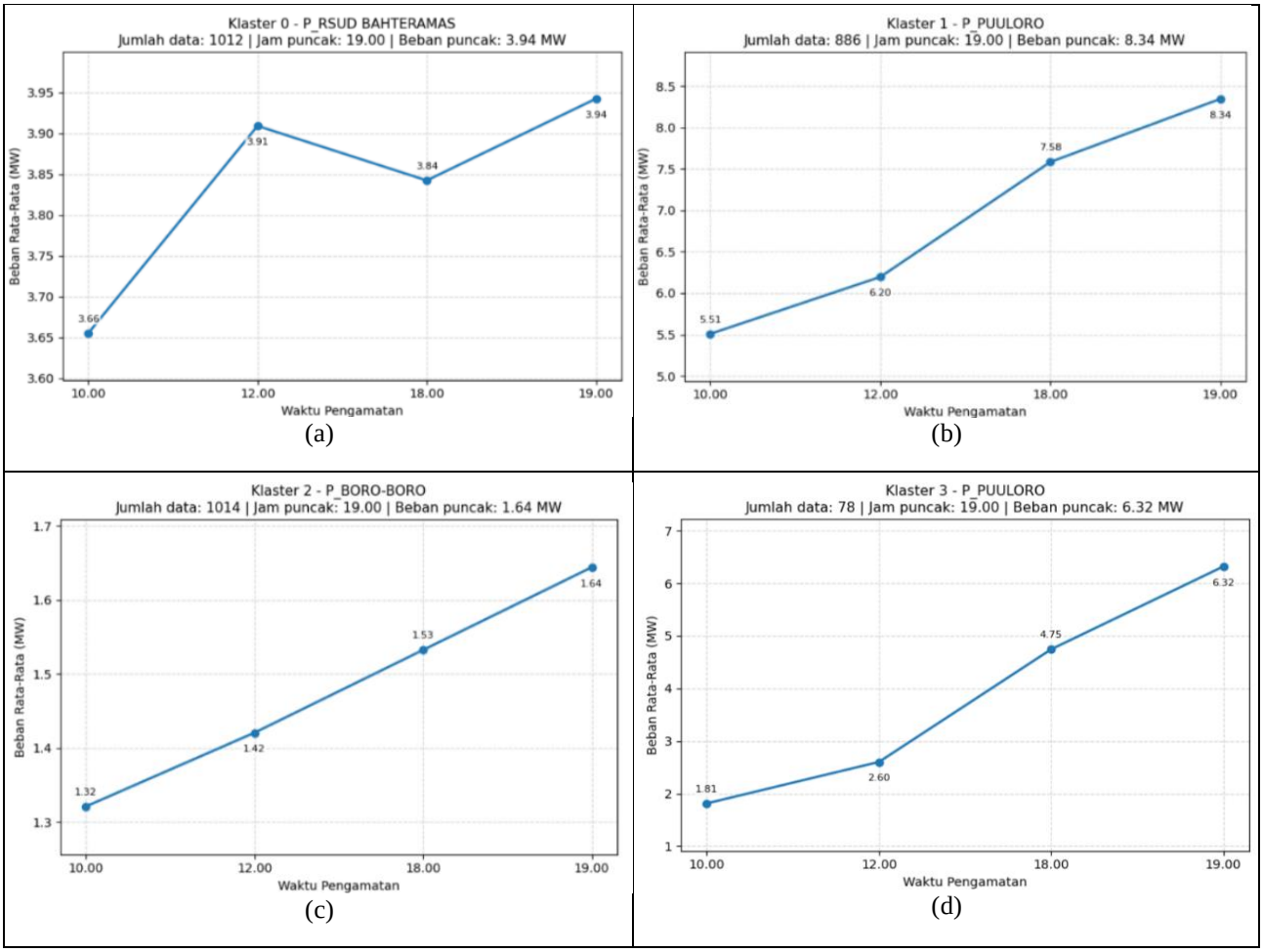

Figure 3. Representative Cluster Feeder Profile

To visually clarify the separation between clusters, PCA is used to provide a visual overview of the clustering results. The visualization results are shown in Figure 4.

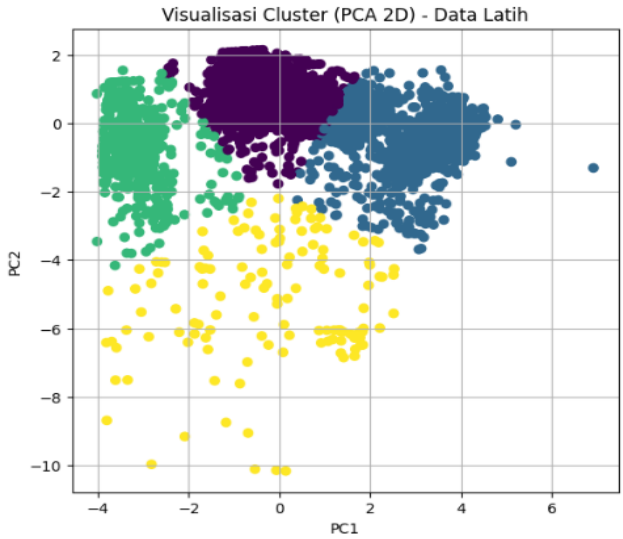

Figure 4. Cluster Visualization Using 2D PCA

Figure 4 shows the distribution of observations in the two-dimensional PCA space colored based on cluster labels. This visualization helps assess whether the formed clusters tend to be separate or overlapping. The visualization results are used to support the interpretation of cluster quality measured through the silhouette score and DBI.

Bulletin of Informatics and Data Science
Vol. 5 No. 1, May 2026, Page 77−85
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v5i1.155
https://ejurnal.pdsi.or.id/index.php/bids/index

Next, the clustering results are used as labels in the classification process using the Support Vector Machine (SVM) method. The SVM model is trained to learn patterns from each cluster so that it can accurately classify new data. To evaluate the model's performance in detail, a classification report is used, which is shown in Figure 5.

```
Akurasi SVM (vs label K-Means train): 0.9923

Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.99      0.99       493
           1       0.99      0.99      0.99       297
           2       1.00      1.00      1.00       209
           3       0.93      1.00      0.96        41

    accuracy                           0.99      1040
   macro avg       0.98      0.99      0.99      1040
weighted avg       0.99      0.99      0.99      1040
```

Figure 5. Classification Report of SVM Classification Results

Based on Figure 5, the performance of the SVM model on the test data shows excellent results across all clusters. Cluster 0 has a precision value of 1.00, a recall of 0.99, and an F1-score of 0.99 with a data count (support) of 493. These results indicate that the model can identify Cluster 0 very well, although there are still a few actual data points that are not perfectly predicted. In Cluster 1, a precision value of 0.99, a recall of 0.99, and an F1-score of 0.99 with a data count of 297 were obtained. These values indicate that the model's performance in Cluster 1 is also very high and stable. Furthermore, in Cluster 2, the model shows the best performance with a precision value of 1.00, a recall of 1.00, and an F1-score of 1.00 on 209 data points. This indicates that all data in Cluster 2 were correctly classified by the SVM model.

In Cluster 3, the model produced a precision value of 0.93, a recall of 1.00, and an F1-score of 0.96 with a total of 41 data points. The recall value of 1.00 indicates that all actual data in Cluster 3 were successfully recognized by the model. However, the slightly lower precision value indicates that there were still a small number of data points from other clusters predicted as Cluster 3. Nevertheless, the F1-score of 0.96 still shows that the model's performance in this cluster is considered very good. Overall, the evaluation results indicate that the SVM model has very high classification capability in predicting the cluster labels resulting from K-Means on the test data. This is shown by an accuracy value of 0.9923, as well as macro average values of 0.98 for precision, 0.99 for recall, and 0.99 for F1-score. In addition, the weighted average values for these three metrics are also around 0.99, which shows that the model's performance remains very good even though the amount of data in each cluster is not the same.

To provide a clearer picture of the distribution of the SVM model's prediction results on the test data, a confusion matrix is used, as shown in Figure 6.

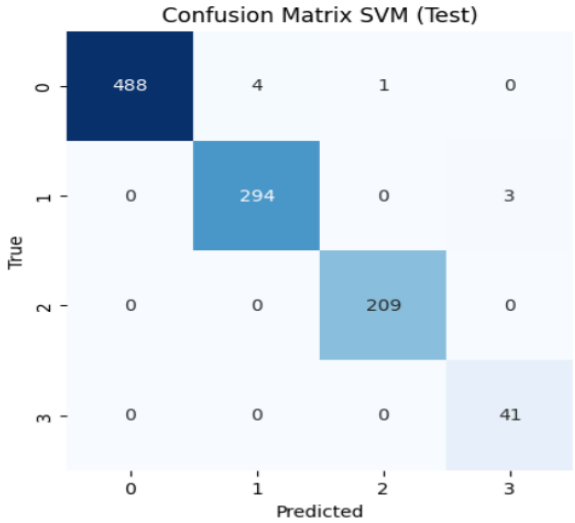

Figure 6. Confusion Matrix Model SVM

Based on Figure 6, it can be seen that most of the data lies on the main diagonal, which indicates that the model's prediction results correspond to the actual labels. The number of correct predictions for each cluster is 488 data points in cluster 0, 294 data points in cluster 1, 209 data points in cluster 2, and 41 data points in cluster 3. The classification errors that occur are relatively small, with only a few data points from cluster 0 being predicted as clusters 1 and 2, as well as a small portion of data from cluster 1 being predicted as cluster 3. This shows that there are similarities in characteristics between some clusters, but overall the model is able to distinguish each cluster very well.

Bulletin of Informatics and Data Science
Vol. 5 No. 1, May 2026, Page 77−85
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v5i1.155
https://ejurnal.pdsi.or.id/index.php/bids/index

3.2 Discussion

The performance of the proposed hybrid K-Means–SVM approach was compared with previous studies that employed clustering and alternative machine learning algorithms for electricity consumption analysis. Palaniappan et al. applied K-Means Clustering followed by eight classification algorithms, including Random Forest, Decision Tree, Naive Bayes, k-Nearest Neighbors, Logistic Regression, Support Vector Machine, Gradient Boosting Machine, and AdaBoost. Their results showed that SVM achieved the best generalization performance, with accuracies of 99.09% for binary classification and 98.19% for three-class classification [5]. In comparison, the present study achieved an accuracy of 99.23% in classifying four electricity consumption pattern classes. This result indicates that SVM maintains strong classification performance when applied to cluster labels representing a larger number of consumption-pattern classes. However, the numerical differences should be interpreted cautiously because the two studies used different datasets, consumer characteristics, input variables, sample sizes, and numbers of classes.

The clustering results were also compared with the study conducted by Michalakopoulos et al., who evaluated K-Means, K-Medoids, Hierarchical Agglomerative Clustering, and DBSCAN for residential electricity load profiling [3]. Their study identified seven optimal clusters, whereas the present study identified four clusters representing stable moderate consumption, high consumption, low consumption, and fluctuating consumption. This difference indicates that the optimal number and characteristics of clusters are influenced by the structure, variables, temporal resolution, and operational context of the dataset. In the present study, the Silhouette Score of 0.5095 and Davies-Bouldin Index of 0.6868 indicate that the four-cluster solution provides a reasonable balance between within-cluster compactness and between-cluster separation for feeder-level electricity consumption data.

Alternative hybrid approaches involving Neural Networks have also been proposed. Chenghu and Thammano combined K-Means with a Neural Network to classify observations located in overlapping cluster regions and reported improved classification performance compared with standalone K-Means-based classification across several experimental datasets [12]. In addition, Syed et al. integrated clustering, consumption-pattern recognition, and deep learning within a short-term load forecasting framework for smart-grid applications [20]. Unlike these studies, the present research focuses on using K-Means cluster labels as target classes for SVM classification. These findings collectively demonstrate that clustering can be integrated with different supervised learning models for either classification or forecasting tasks.

Random Forest, Decision Tree, and Neural Network models offer different methodological advantages. Random Forest can reduce overfitting by combining predictions from multiple decision trees, while Decision Tree provides easily interpretable classification rules. Neural Networks can model complex nonlinear relationships but generally require larger datasets, more extensive hyperparameter tuning, and greater computational resources. In comparison, SVM is suitable for high-dimensional data and can construct effective decision boundaries using a relatively limited number of training observations. In this study, SVM produced high precision, recall, and F1-score values across all four clusters, although the number of observations in each cluster was unequal.

Overall, the accuracy of 99.23%, together with high precision, recall, and F1-score values, demonstrates that the sequential K-Means–SVM framework is effective for classifying electricity consumption patterns generated from the studied dataset. Nevertheless, this result does not conclusively demonstrate that SVM is superior to Random Forest, Decision Tree, or Neural Network models because these alternative classifiers were not trained and evaluated using the same PT PLN (Persero) UP3 Kendari dataset. A definitive comparison would require identical training and test data, feature sets, preprocessing procedures, cross-validation strategies, evaluation metrics, and hyperparameter optimization. Therefore, future research should conduct a controlled comparison between SVM and these alternative models to determine whether the proposed hybrid approach consistently provides superior classification performance.

4. CONCLUSION

Based on the results of this study, the K-Means Clustering method successfully grouped electricity consumption patterns into four main clusters: Cluster 0 representing moderate and relatively stable consumption, Cluster 1 representing high consumption, Cluster 2 representing low consumption, and Cluster 3 representing fluctuating consumption. The clustering evaluation produced a Silhouette Score of 0.5095 and a Davies-Bouldin Index of 0.6868, indicating fairly good cluster compactness and separation. The Principal Component Analysis visualization showed that the cluster distributions were generally distinguishable, although slight overlap remained between several data groups. Furthermore, the Support Vector Machine model classified the cluster labels generated by K-Means with an accuracy of 99.23%, while the precision, recall, and F1-score values approached 1.00. The confusion matrix also showed that most test observations were correctly classified with minimal errors. These results demonstrate that the hybrid K-Means Clustering and SVM approach is effective in identifying and classifying electricity consumption patterns and can support electricity consumption monitoring and power distribution planning. Nevertheless, this study did not consider external factors such as environmental conditions and electricity-user behavior.Therefore, future research should incorporate these external variabels and evalueate additional machine learning methods to improve and validate the modal’s performance.

REFERENCES

[1] I. M. A. A. Putra and I. B. G. Manuaba, "Literatur Review Tantangan dan Teknologi dalam Pengembangan Advance Metering Infrastructure (AMI)," *MITE*, vol. 24, no. 1, pp. 1–8, Aug. 2025, doi: 10.24843/MITE.205.v24i01.P01.

[2] M. A. Dávila R., C. L. Trujillo R., and A. A. Jaramillo M., "Revisión de mecanismos de gestión del lado de la demanda para la gestión de energía en el hogar," *Ingeniare. Rev. chil. ing.*, vol. 30, no. 2, pp. 353–367, Jun. 2022, doi: 10.4067/S0718-33052022000200353.

[3] V. Michalakopoulos, E. Sarmas, I. Papias, P. Skaloumpakas, V. Marinakis, and H. Doukas, "A machine learning-based framework for clustering residential electricity load profiles to enhance demand response programs," *Applied Energy*, vol. 361, p. 122943, May 2024, doi: 10.1016/j.apenergy.2024.122943.

[4] 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, p. 3083, Jun. 2025, doi: 10.3390/en18123083.

[5] S. Palaniappan, S. Karuppannan, and D. Velusamy, "Categorization of Indian residential consumers electrical energy consumption pattern using clustering and classification techniques," *Energy*, vol. 289, p. 129992, Feb. 2024, doi: 10.1016/j.energy.2023.129992.

[6] K.-L. Du, B. Jiang, J. Lu, J. Hua, and M. N. S. Swamy, "Exploring Kernel Machines and Support Vector Machines: Principles, Techniques, and Future Directions," *Mathematics*, vol. 12, no. 24, p. 3935, Dec. 2024, doi: 10.3390/math12243935.

[7] H. Yıldız and M. Balta, "Comparison of Machine Learning Based Anomaly Detection for Energy Consumption Values in SDN-IoT Based Home Area Networks," *SAUCIS*, vol. 8, no. 3, pp. 518–535, Sep. 2025, doi: 10.35377/saucis.8.94717.1641393.

[8] X. Wang, H. Wang, B. Bhandari, and L. Cheng, "AI-Empowered Methods for Smart Energy Consumption: A Review of Load Forecasting, Anomaly Detection and Demand Response," *Int. J. of Precis. Eng. and Manuf.-Green Tech.*, vol. 11, no. 3, pp. 963–993, May 2024, doi: 10.1007/s40684-023-00537-0.

[9] F. Rodríguez-Gómez, J. Del Campo-Ávila, and L. Mora-López, "A novel clustering based method for characterizing household electricity consumption profiles," *Engineering Applications of Artificial Intelligence*, vol. 129, p. 107653, Mar. 2024, doi: 10.1016/j.engappai.2023.107653.

[10] G. E. Okereke, M. C. Bali, C. N. Okwueze, E. C. Ukekwe, S. C. Echezona, and C. I. Ugwu, "K-means clustering of electricity consumers using time-domain features from smart meter data," *Journal of Electrical Systems and Inf Technol*, vol. 10, no. 1, p. 2, Jan. 2023, doi: 10.1186/s43067-023-00068-3.

[11] X. Li, M. Jiang, D. Cai, W. Song, and Y. Sun, "A Hybrid Forecasting Model for Electricity Demand in Sustainable Power Systems Based on Support Vector Machine," *Energies*, vol. 17, no. 17, p. 4377, Sep. 2024, doi: 10.3390/en17174377.

[12] C. Chenghu and A. Thammano, "A Novel Classification Model Based on Hybrid K-Means and Neural Network for Classification Problems," *HighTech. Innov. J.*, vol. 5, no. 3, pp. 716–729, Sep. 2024, doi: 10.28991/HIJ-2024-05-03-012.

[13] A. Forootani, M. Rastegar, and A. Sami, "Short-term individual residential load forecasting using an enhanced machine learning-based approach based on a feature engineering framework: A comparative study with deep learning methods," *Electric Power Systems Research*, vol. 210, p. 108119, Sep. 2022, doi: 10.1016/j.epsr.2022.108119.

[14] F. Topaloglu, "A hybrid approach based on k-means and SVM algorithms in selection of appropriate risk assessment methods for sectors," *PeerJ Computer Science*, vol. 10, p. e2198, Jul. 2024, doi: 10.7717/peerj-cs.2198.

[15] R. Wu, "Behavioral analysis of electricity consumption characteristics for customer groups using the k-means algorithm," *Systems and Soft Computing*, vol. 6, p. 200143, Dec. 2024, doi: 10.1016/j.sasc.2024.200143.

[16] F. Mazidah and S. Anisah, "ANALISIS PEMANFAATAN DAYA LISTRIK BAGI PELANGGAN TEGANGAN MENENGAH," *POLEKTRO*, vol. 13, no. 2, pp. 208–211, May 2024, doi: 10.30591/polektro.v13i2.6699.

[17] M. M. Asiri, G. Aldehim, F. A. Alotaibi, M. M. Alnfiai, M. Assiri, and A. Mahmud, "Short-Term Load Forecasting in Smart Grids Using Hybrid Deep Learning," *IEEE Access*, vol. 12, pp. 23504–23513, 2024, doi: 10.1109/ACCESS.2024.3358182.

[18] G. Gusrianty, F. Fenly, D. Jollyta, E. Erlin, R. N. Putri, and D. Oktariana, "Penerapan Linear Discriminant Analysis Untuk Meningkatkan Kinerja Algoritma Support Vector Machine," *JPIT*, vol. 10, no. 4, pp. 895–906, Sep. 2025, doi: 10.30591/jpit.v10i4.8772.

[19] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Sci Rep*, vol. 14, no. 1, p. 6086, Mar. 2024, doi: 10.1038/s41598-024-56706-x.

[20] D. Syed *et al.*, "Deep Learning-Based Short-Term Load Forecasting Approach in Smart Grid With Clustering and Consumption Pattern Recognition," *IEEE Access*, vol. 9, pp. 54992–55008, 2021, doi: 10.1109/ACCESS.2021.3071654.