Nurdin Khoirurizka (諾丁)
312540025

Report Clustering Methods

Clustering the Battery State of Health in Electric Vehicles

Final Project

NYCU

The rapid growth of electric vehicles (EVs) is revolutionizing the automotive industry, driven by the urgent need for sustainable transportation solutions that cut greenhouse gas emissions and reduce our dependence on fossil fuels [1]. However, this shift introduces significant challenges, particularly concerning the batteries that power these vehicles. The increasing adoption of EVs is transforming the transportation sector and is a key solution to minimizing reliance on non-renewable energy sources. The high cost of batteries and the need for regular maintenance to ensure their longevity and safety remain significant concerns [2].

Predicting the state of health (SoH) of EV batteries is crucial for anticipating maintenance needs and preventing potential hazards, thereby extending battery lifespan and enhancing overall vehicle safety. Advanced clustering techniques offer a promising solution for analyzing and predicting battery SoH with high precision. The benefits of implementing these techniques include predictive maintenance, which allows for proactive identification of batteries likely to fail sooner, thereby reducing unexpected downtime. Additionally, optimized charging protocols can be developed based on cluster analysis, improving efficiency and battery longevity. Enhanced Battery Management Systems (BMS) can utilize the valuable data provided by clustering to monitor and control charging processes more effectively, maintaining optimal battery health. This, in turn, minimizes financial losses associated with battery replacement by extending battery lifespan through better maintenance and management [3].

On this research, clustering method was proposed to determine the cluster of SoH from battery feature. With feature that will clustered are voltage, current, and temperature. This feature was chosen because it has strong correlation with SoH [4]. This feature is from 48,717 battery state. The feature distribution is showing on Figure 1. Then to find the result that can provide accurate information about the condition of the battery. The feature will be clustered using K-Medoid, Agglomerative, Mean Shift, and DBSCAN. This method was chosen because their specific benefits:

- K-Medoid: This method is robust to outliers, making it suitable for data with noise or outliers, which is common in real-world battery datasets. Additionally, it offers interpretability, as the clusters are more meaningful in the context of battery health analysis [5].
- Agglomerative: Known for its hierarchical structure, this method allows visualization of different levels of granularity in the data, providing insights into the hierarchy of battery health conditions, from broader categories to specific clusters. It is also flexible in cluster shape, as it does not assume clusters to be spherical [6].
- Mean Shift: This method can automatically adapt the bandwidth parameter based on the density of data points, making it effective in identifying clusters of varying densities. This adaptability is beneficial in detecting different levels of battery degradation. Additionally,

Mean Shift does not require specifying the number of clusters beforehand, making it suitable for datasets where the number of battery health states is unknown [7].

- DBSCAN: Effective in handling noise and outliers, DBSCAN is particularly useful for battery datasets that contain sensor inaccuracies or sporadic anomalies in battery behavior [8].

By comparing and optimizing these clustering methods, it was aim to identify the most effective approach for clustering battery SoH data. This will ultimately improve the reliability and efficiency of battery SoH prediction, contributing to a more sustainable future for the EV industry.

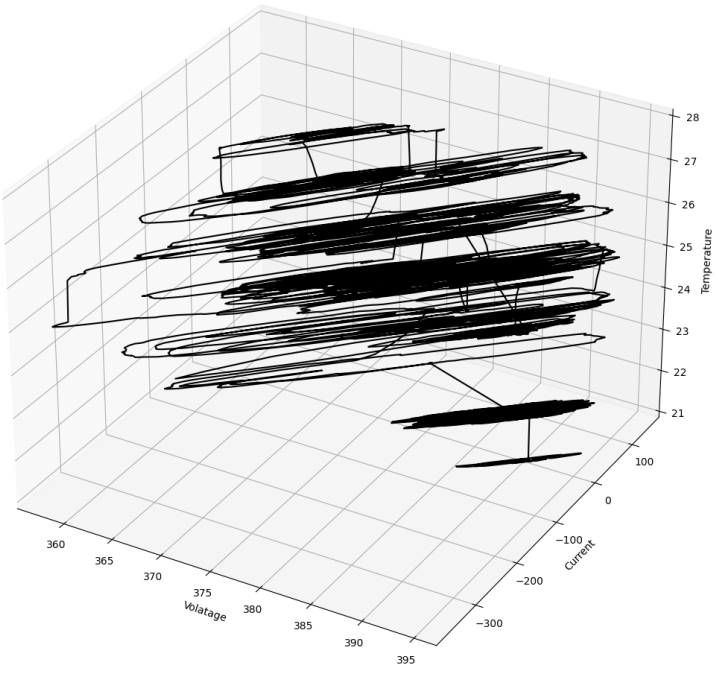

Figure 1. The battery feature distribution

In the experiment, K-Medoid clustering produced the results shown in Figure 2, where the algorithm grouped the data into 20 distinct clusters. The visualization indicates a strong correlation between the clusters and the variables Voltage and Current, suggesting that these variables play a significant role in defining the clusters. The clusters are visually represented with different colors, showing clear groupings of data points that share similar Voltage and Current values. This correlation is further illustrated in Figure 3, which provides detailed 2D projections of the clusters.

The K-Medoid clustering performance was evaluated using several metrics. The Dunn Index, which measures the ratio of the minimum inter-cluster distance to the maximum intra-cluster distance, was 1.058. This value suggests a decent separation between the clusters, indicating that the clusters are well-separated from each other. The Davies-Bouldin Index, which assesses the average similarity ratio of each cluster with the cluster most similar to it, was 0.7588. A lower Davies-Bouldin Index indicates better clustering, implying moderate cluster compactness and separation in this case.

Additionally, the Xi-Beni Index, which evaluates the compactness and separation of the clusters, was 0.0632. This relatively low value suggests that the clusters are compact and well-

separated. The Silhouette Score, which measures how similar an object is to its own cluster compared to other clusters, was 0.4053. A score closer to 1 indicates better-defined clusters, while a score closer to -1 indicates overlapping clusters. In this case, the score suggests that the clusters are somewhat well-defined but have room for improvement. Overall, these metrics indicate that the K-Medoid algorithm provided a reasonable clustering performance for the given data, capturing the underlying patterns effectively.

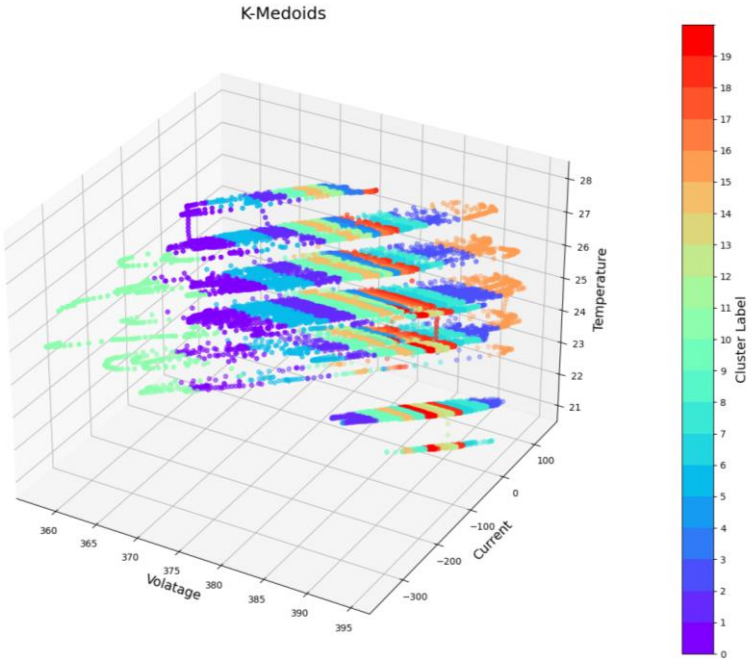

Figure 2. 3D scatter plot of K-Medoid

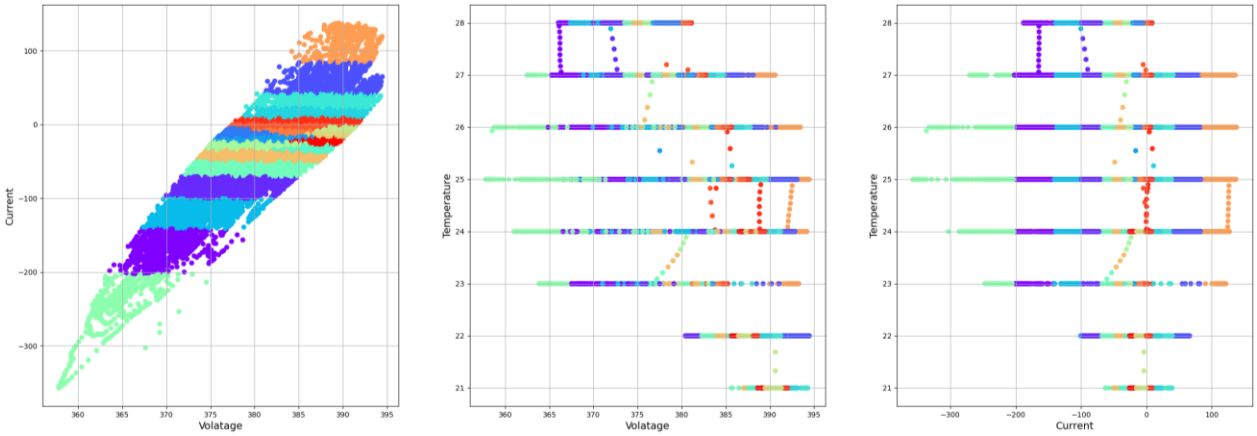

Figure 3. Comparison K-Medoid for each feature

The Agglomerative clustering analysis resulted in the identification of 20 distinct clusters, as shown in Figure 4. These clusters demonstrate a strong correlation with the variables Voltage and Current, suggesting that these variables play a significant role in defining the clusters. The evaluation of the clustering quality, using internal evaluation metrics, yielded a Davies Bouldin Index of 0.6985 and a Silhouette Score of 0.3484. The Davies Bouldin Index indicates that the clusters are reasonably well-separated, with a lower value signifying better clustering. The Silhouette Score, which ranges from -1 to 1, indicates that the clustering structure is moderately

well-defined, with the score of 0.3484 suggesting that some clusters are better defined than others. The internal evaluation only showing that metrics because Agglomerative clustering does not utilize centroids.

Further illustration of this correlation is provided in Figure 5, which presents detailed 2D projections of the clusters. The 2D projections include plots of Current versus Voltage, Temperature versus Voltage, and Temperature versus Current. These projections help to visualize the relationships between the variables and how they contribute to the clustering. For example, the plot of Current versus Voltage shows a linear correlation within each cluster, with distinct groupings based on the Voltage and Current values. The Temperature versus Voltage and Temperature versus Current plots illustrate how temperature varies within and between clusters, providing additional context to the clustering structure.

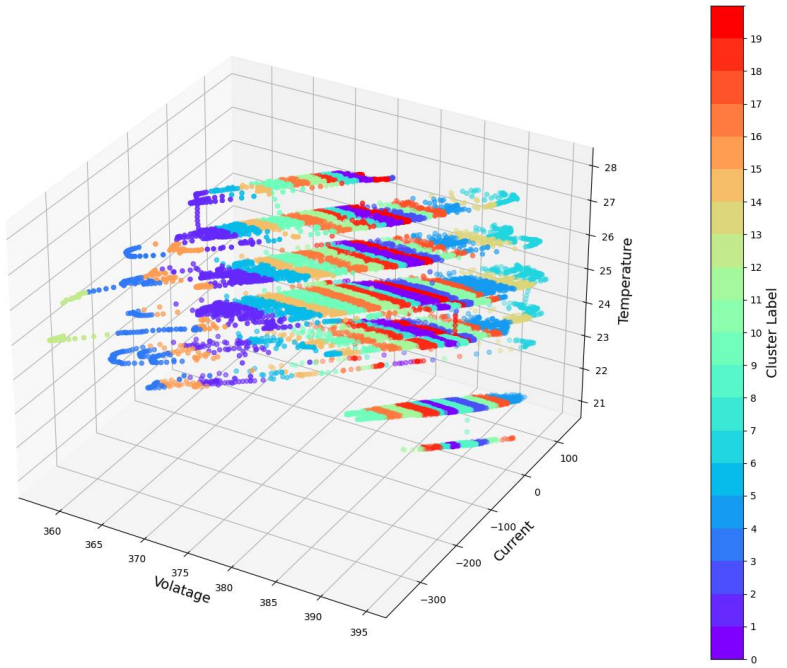

Figure 4. 3D scatter plot of Agglomerative

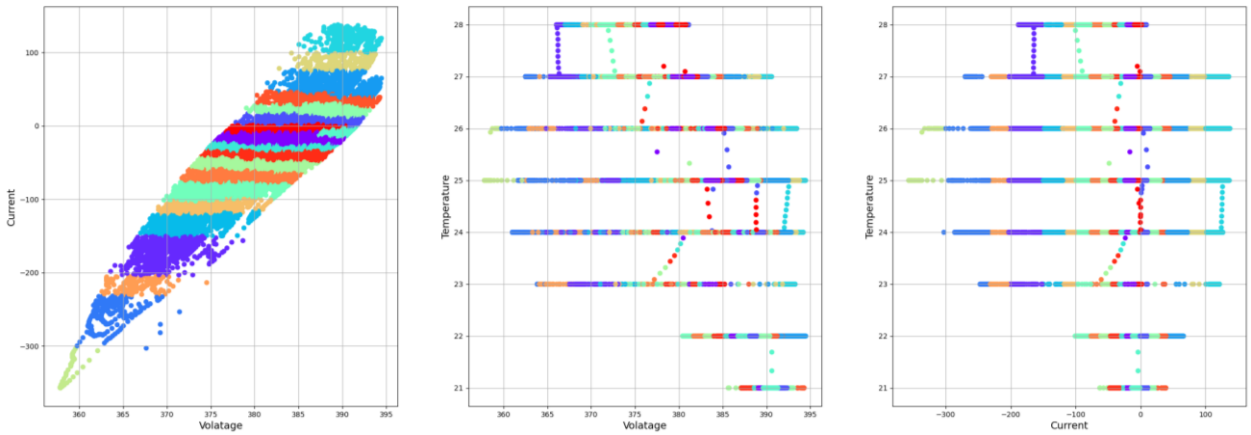

Figure 5. Comparison Agglomerative for each feature

Nurdin Khoirurizka (諾丁)
312540025

The Mean Shift clustering results, shown in Figure 6, reveal four clusters automatically determined by the algorithm. These clusters exhibit a strong correlation with the variables Voltage and Current, indicating that these variables significantly define the clusters. The clustering quality is quantified by several performance metrics: The Dunn Index of 0.2881, Davies-Bouldin Index of 0.4195, Xi-Beni Index of 0.0991, and Silhouette Score of 0.6340.

The Dunn Index of 0.2881 indicates a moderate level of separation and compactness among the clusters. The Davies-Bouldin Index of 0.4195 suggests that the clusters are distinct from each other, with low similarity between them. The Xi-Beni Index of 0.0991 reflects that the clusters are compact and well-separated. The Silhouette Score of 0.6340 indicates that the data points are well-matched to their own clusters and poorly matched to neighboring clusters, showing clear and distinct clustering.

The visualization indicates a strong correlation between the clusters and the variables Voltage and Current, suggesting that these variables play a significant role in defining the clusters. The clusters are visually represented with different colors, showing clear groupings of data points that share similar Voltage and Current values. This correlation is further illustrated in Figure 7, which provides detailed 2D projections of the clusters. These projections show the distribution and separation of clusters across the Voltage-Current, Voltage-Temperature, and Current-Temperature planes, reinforcing the significant role of Voltage and Current in defining the clusters.

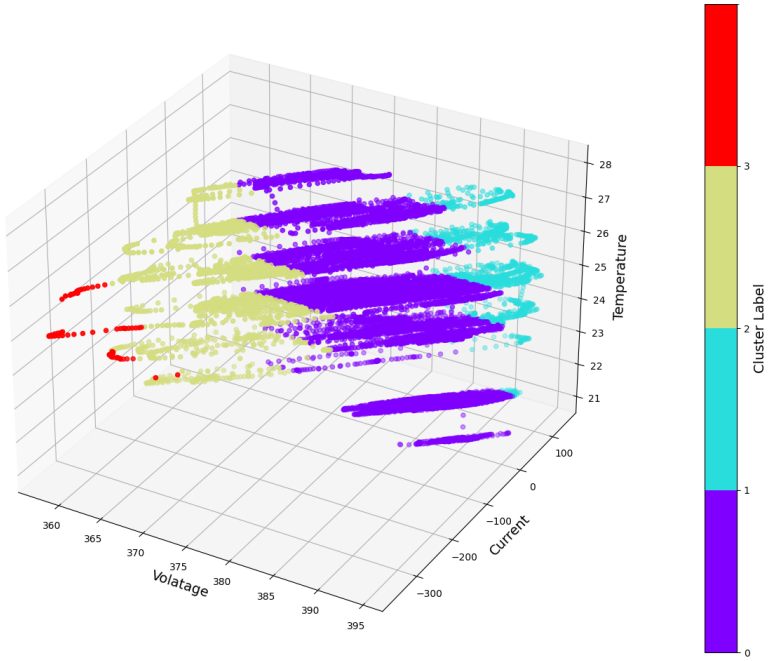

Figure 6. 3D scatter plot of Mean Shift

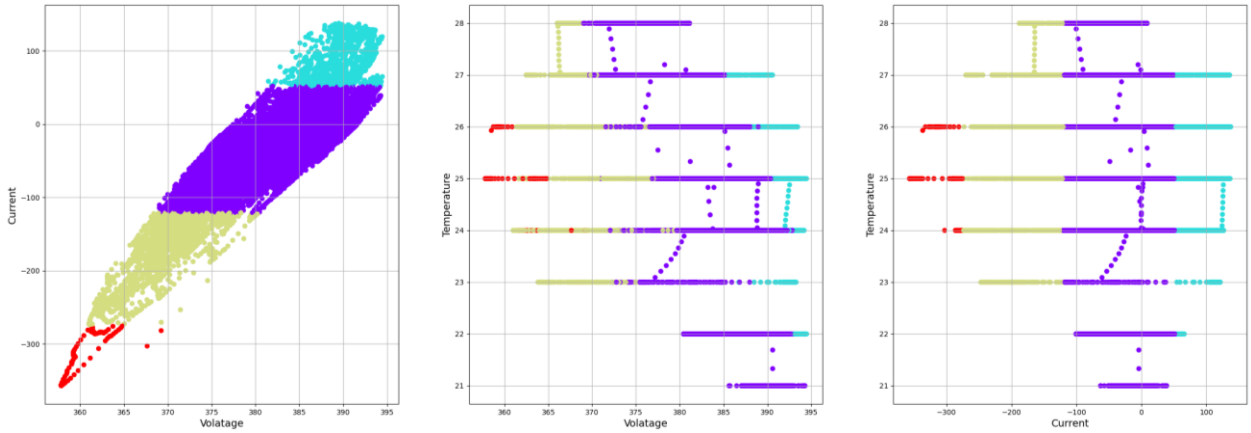

Figure 7. Comparison Mean Shift for each feature

The results of the experiment with DBSCAN clustering are illustrated in Figure 8. The clustering process resulted in a single cluster, indicating that the method failed to distinguish distinct clusters. This failure can be attributed to the data points being too close to each other, making it difficult to effectively separate them into different clusters.

The evaluation metrics for the DBSCAN clustering showed a Davies-Bouldin Index of 2.2532 and a Xi-Beni Index of 0.6291, suggesting poor clustering performance. The high Davies-Bouldin Index value indicates that the clusters are not well-separated and have significant overlap, while the Xi-Beni Index, although lower, further confirms the inadequacy of the clustering performance in this scenario.

Figure 9 provides detailed visualizations of the clustering results, offering additional insights. The first plot, showing the relationship between Current and Voltage, indicates that the data points are densely packed, contributing to the clustering failure. The second plot, illustrating the relationship between Temperature and Voltage, shows horizontal lines representing cluster boundaries, highlighting the algorithm's inability to separate distinct temperature ranges effectively. The third plot, displaying the relationship between Temperature and Current, similarly shows poor separation, leading to the formation of a single cluster.

Nurdin Khoirurizka (諾丁)
312540025

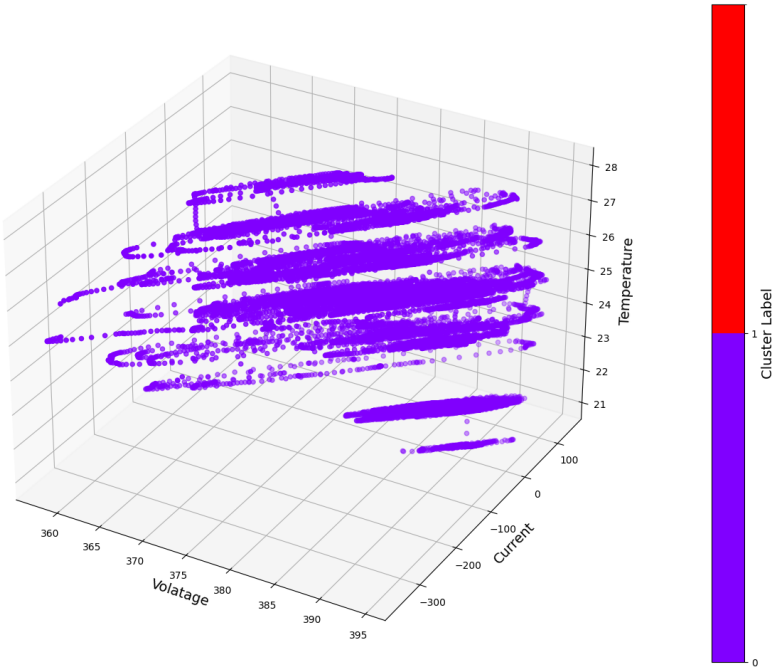

Figure 8. 3D scatter plot of DBSCAN

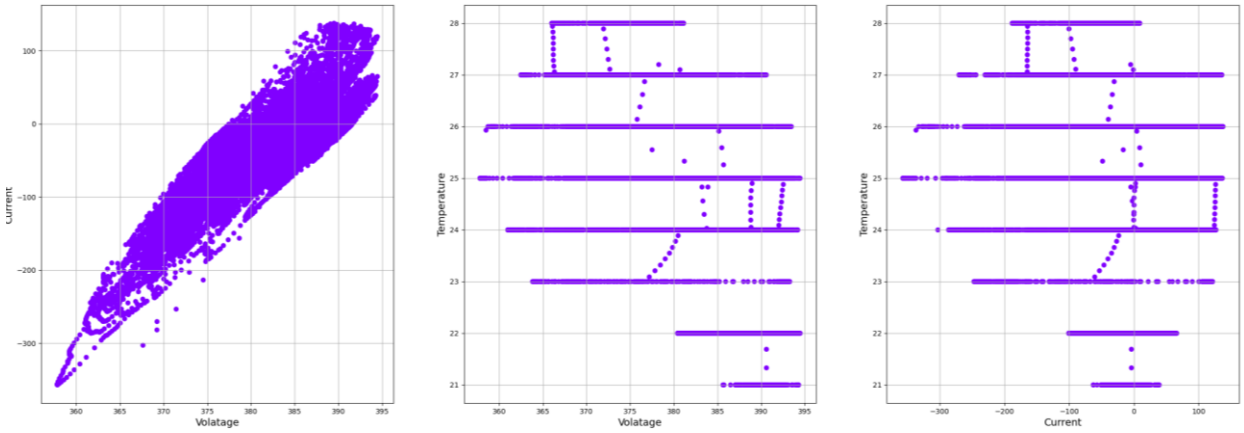

Figure 9. Comparison DBSCAN for each feature

The Table 1 presents a comparative internal evaluation of each cluster using four metrics: Dunn Index, Davies Bouldin Index, Xi-Beni Index, and Silhouette Score. The results indicate that K-Medoid performs the best based on the Dunn Index (1.058) and Xi-Beni Index (0.0632), signifying superior compactness and separation. Mean Shift best in the Davies Bouldin Index (0.4195) and Silhouette Score (0.6340), highlighting its effectiveness in average cluster similarity and object-cluster similarity. Agglomerative and DBSCAN show less favorable results, with DBSCAN having the highest Davies Bouldin Index (2.2532), indicating poorer clustering. The highlighted values in the table mark the best performance for each metric among the clustering methods.

Nurdin Khoirurizka (諾丁)
312540025

Table 1. Comparison internal evaluation for each method

Internal Evaluation	K-Medoid	Agglomerative	Mean Shift	DBSCAN
Dunn Index	1.058	-	0.2881	-
Davies Bouldin Index	0.7588	0.6985	0.4195	2.2532
Xi-Beni Index	0.0632	-	0.0991	-
Silhouette Score	0.4053	0.3484	0.6340	0.6291

The conclusion of this research, The K-Medoid clustering demonstrated a strong performance with a Dunn Index of 1.058, Davies-Bouldin Index of 0.7588, Xi-Beni Index of 0.0632, and Silhouette Score of 0.4053. This method can produce 20 distinct clusters with a clear correlation between voltage and current. Agglomerative clustering, also producing 20 clusters, provided a hierarchical structure that offered insights into different levels of battery health. Its evaluation metrics showed moderate cluster separation, with a Davies-Bouldin Index of 0.6985 and Silhouette Score of 0.3484. Mean Shift clustering identified four clusters, demonstrating good clustering quality with a Dunn Index of 0.2881, Davies-Bouldin Index of 0.4195, Xi-Beni Index of 0.0991, and Silhouette Score of 0.6340. This method was particularly effective in identifying clusters of varying densities. In contrast, DBSCAN clustering failed to effectively separate the data, resulting in a single cluster due to the close proximity of data points, and showed poor performance with a Davies-Bouldin Index of 2.2532 and Xi-Beni Index of 0.6291. Overall, the results indicate that K-Medoid and Mean Shift clustering methods provided the most effective clustering performance for predicting battery SoH. These methods offer promising solutions for improving the reliability and efficiency of battery SoH prediction, contributing to better maintenance and management of EV batteries.

For future work, incorporating additional features such as state of charge (SoC), internal resistance, charge/discharge cycles, and ambient conditions could enhance clustering accuracy. Also this method can combine with machine learning models or real-time data analysis capabilities to improve prediction accuracy and practical applicability.

Reference:

[1] I. Aijaz and A. Ahmad, “Electric vehicles for environmental sustainability,” *Smart Technol. Energy Environ. Sustain.*, pp. 131–145, 2022.
[2] X. Lai *et al.*, “A review of lithium-ion battery failure hazards: Test standards, accident analysis, and safety suggestions,” *Batteries*, vol. 8, no. 11, p. 248, 2022.
[3] J. Yang, B. Xia, W. Huang, Y. Fu, and C. Mi, “Online state-of-health estimation for lithium-ion batteries using constant-voltage charging current analysis,” *Appl. Energy*, vol. 212, pp. 1589–1600, 2018.
[4] C. Zhang *et al.*, “Battery SOH estimation method based on gradual decreasing current, double correlation analysis and GRU,” *Green Energy Intell. Transp.*, vol. 2, no. 5, p. 100108, 2023.
[5] W. Tang, H. Wang, X.-L. Lee, and H.-T. Yang, “Machine learning approach to uncovering residential energy consumption patterns based on socioeconomic and smart meter data,” *Energy*, vol. 240, p. 122500, 2022.

[6] D. S. Lamb, J. Downs, and S. Reader, “Space-time hierarchical clustering for identifying clusters in spatiotemporal point data,” *ISPRS Int. J. Geo-Inf.*, vol. 9, no. 2, p. 85, 2020.
[7] F. Barranco, C. Fermuller, and E. Ros, “Real-time clustering and multi-target tracking using event-based sensors,” presented at the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2018, pp. 5764–5769.
[8] N. Li, M. Peng, B. Cao, K. Li, and K. Li, “Similarity Graph Learning and Non-linear Deep Representations for Spectral Clusterings,” presented at the Journal of Physics: Conference Series, IOP Publishing, 2021, p. 012001.