2023 IEEE 17th International Conference on Compatibility, Power Electronics and Power Engineering (CPE-POWERENG) | 979-8-3503-0004-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/CPE-POWERENG58103.2023.10227417

Clustering and Outlier Analysis for Key Performance Indicators in Battery Energy Storage Systems applications

Rolando Antonio Gilbert Zequera
Department of Electrical Power Engineering and Mechatronics
Tallinn University of Technology
Tallinn, Estonia
rogilb@ttu.ee

Anton Rassõlkin
Department of Electrical Power Engineering and Mechatronics
Tallinn University of Technology
Tallinn, Estonia
anton.rassolkin@taltech.ee

Toomas Vaimann
Department of Electrical Power Engineering and Mechatronics
Tallinn University of Technology
Tallinn, Estonia
toomas.vaimann@taltech.ee

Ants Kallaste
Department of Electrical Power Engineering and Mechatronics
Tallinn University of Technology
Tallinn, Estonia
ants.kallaste@taltech.ee

***Abstract*— This research work focuses on implementing outlier analysis and clustering to provide an assessment of the charging and discharging processes of Battery Energy Storage Systems (BESSs). K-Means, Density-based spatial clustering of applications with noise (DBSCAN), and Local Outlier Factor (LOF) are the main algorithms executed to illustrate Key Performance Indicators (KPIs) and their corresponding critical values during battery operation. Additional Data Mining methods are implemented to provide feature selection, correlation analysis, parameter estimation, and validation. Implemented algorithms show that there is a strong correlation between specific variables at certain operation stages, which is complemented by the lifetime period of BESSs.**

Keywords—outliers, clusters, charge, discharge, battery.

I. INTRODUCTION

Battery Energy Storage Systems (BESSs) are used for a variety of applications in the energy industry, with mobility systems being one of the most promising fields to reduce CO2 emissions by deploying various categories of electric vehicles around the world. Especially in Europe, where in October 2022 the European Parliament and Council agreed to ensure all new cars registered in Europe will be zero-emission by 2035.

Different charging and discharging strategies have been implemented to monitor and ensure the reliability of BEESs measurements, however, some of the parameters are not controlled all the time by the battery's user, which could lead to battery degradation. Several studies [1,2] have been conducted to demonstrate that charging and discharging processes can be optimized by monitoring Key Performance Indicators (KPIs) and increasing the lifetime period of BEEs, such as the voltage of the cell, the temperature measured, and cell capacity. The charging technique is considered as the most crucial factor that affects the stability of a BESS, all due to its ability to control time, temperature, and assure protection, on the other hand, the discharging process is essential to determine the capacity of a BESS based on cycles, which plays a key role in the lifetime estimation.

In the field of Data Mining, clustering and outlier analysis is the point of discussion to develop new and efficient methodologies to optimize the solution of complex problems, not only limited to the computer science community but also to the energy industry. Remarkable Data Mining studies have also contributed to the development of new insights in the field of BESSs and industry mobility. In 2020, Zhou et al. [3] developed a methodology for second-life batteries usages by implementing a bisecting K-Means algorithm, which demonstrated fast clustering of retired lithium-ion batteries. In addition to the previous study, Ran et al. used a pulse clustering model that was embedded with improved bisecting K-Means, all to effectively sort retired batteries with specific life cycles [4]. Finally, in 2022 Chang et al. [5] implemented and compared the efficiency of Hierarchical clustering, K-Means, and Hybrid clustering to sort pouch cell capacity in battery packs.

The main goal of current research work is to provide clustering and outlier analysis for charge and discharge in a BESS, all to assess operating mechanisms based on the identification and explanation of KPIs through Data Mining algorithms.

The rest of the paper is organized as follows, in Section II, the problem statement and the motivation of this research are explained. Section III describes and implements the methodologies, focusing on feature selection, correlation analysis, and parameter estimation. Section IV provides the results based on the KPIs and their interpretation within the framework of the BESS. Finally, in Section V, a conclusion is provided to promote new areas of opportunity based on fault diagnostics and lifetime estimation of BESSs.

II. DATASET AND PROBLEM STATEMENT

The field of Data Mining has been growing during the past two decades, providing new opportunity areas not only in the computer science community but also in the engineering industry, being renewable energy integration a promising topic that contributes to climate change mitigation. Foundations of Data Mining consider four "super problems", which are clustering, association pattern mining, outlier analysis, and classification. The relevance of these problems relies on the broad use as building blocks in a variety of data mining applications [6], complementing more advanced fields

The research has been supported by the Estonian Research Council under grant PSG453 "Digital twin for propulsion drive of autonomous electric vehicle".

such as Machine Learning, Deep Learning, Natural Language Processing, etc.

In this research, two of the four "super problems" in the field of Data Mining are discussed under the framework of an energy perspective, all to illustrate the KPIs of BESSs in charging and discharging processes.

This research work uses an aging dataset collected by Macintosh in 2010 for a Li-ion battery that ran through two operational profiles at room temperature [7]. The charge was carried out in a constant current (CC) mode at 1.5 A until the battery voltage reached 4.2 V, and then continued in a constant voltage (CV) mode until the charging current load dropped to 20 mA. The discharge was performed at a constant current level of 2.0 A until the battery voltage dropped to 2.7 V. The dataset contains several features that explain the behavior of the BESS: voltage measured, current measured, temperature measured, current charge, voltage charge, time vector for the cycles, capacity for discharging, operation type, ambient temperature, and start time. It is fundamental to point out that only a few variables from the entire dataset will be considered as KPIs when implementing the corresponding methods described in the following sections.

During the operation of a BESS, crucial steps are based on the State of Charge (SOC) and State of Health (SOC), however, identifying potential values that can lead to degradation mechanisms at certain periods of time is an essential task to accomplish. Implementing clustering algorithms ensures the reliability of experimental battery measurements and identifies patterns at various stages.

Regarding outlier analysis, predictive maintenance is an essential task to monitor the lifetime of a BESS. To validate not only stability in battery operation but also battery modeling, some critical points at specific intervals are the main concern to avoid early deterioration, and these points are outliers found in anomaly ranges of battery processes.

III. METHODS

Initially, the Correlation test and Principal Component Analysis (PCA) are implemented to perform data preprocessing, feature selection, and illustrate variance importance. Subsequently, some parameter estimation techniques are applied to optimize the mechanism of selected algorithms and evaluate clustering quality. K-Means, Local Outlier Factor (LOF), and Density-based spatial clustering of applications with noise (DBSCAN) are described and executed.

A. PCA and feature selection

The dataset is processed and sorted based on the Start time to understand problem dimensionality represented by charge and discharge. Null values are searched and not found, which ensures the reliability of the measurements. Additionally, a correlation test is executed to analyze the relevance of all variables.

Ambient temperature and start time are dropped to continue with the following steps, the first because of the constant value during all the measurements and the second because of the negligible numerical correlation with the input features. Similarly, the "Type" variable is separated from the matrix of features based on its clustering properties According to the results of the Correlation test, the temperature measured, time, voltage measured, current measured, and capacity are the most correlated variables in the entire dataset, these being considered as the KPIs.

Before implementing the correlation test, it was expected to have a mutual correlation between the voltage measured vs capacity, all because these variables explain the charge-discharge curves of a BESS. However, in this case, the results indicate an anomaly behavior in the dataset. On the other hand, voltage measured vs. time and temperature measured vs. capacity show the highest correlation expected because of KPIs on aging and degradation in a BESS. A visual representation of the Correlation test can be appreciated in Fig. 1.

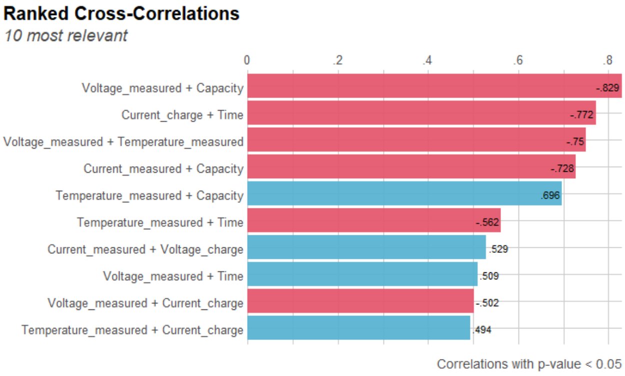

Fig. 1. Correlations with p<0.05 for charge and discharge, in which "p" denotes the level of significance. Negative correlations are represented in red and positive correlations are in blue.

Regarding the contribution of the KPIs, the PCA algorithm is executed to illustrate the variance importance for each principal component. It is important to mention that not only is the feature matrix standardized, but also the corresponding eigenvalues of each principal component are used to draw a boundary for the explanatory variables that retain the highest cumulative variance. To understand the contribution of each input variable in the entire dataset, the importance of components is explained by the standard deviation, proportion of variance, and cumulative proportion, the first because of the eigenvalue's representation, the second due to the amount of variance that each principal component accounts for in the dataset, and the third indicate the accumulative amount of explained variance. Fig. 2 shows the percentage of explained variance by each principal component.

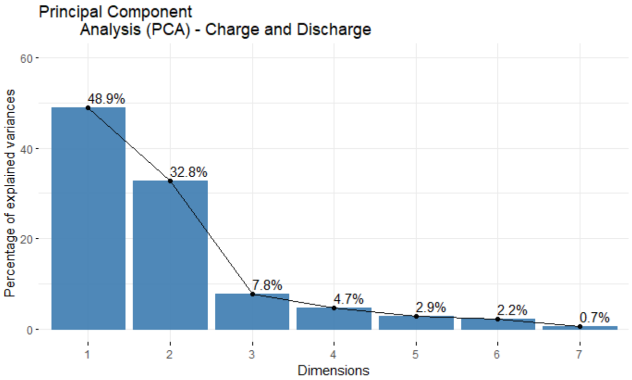

Fig. 2. PCA representation for cumulative variance. Explained variance and contribution of each principal component.

Since the data is standardized and the corresponding eigenvalues of each principal component are obtained, the boundary is implemented to those eigenvalues <1, which means that the component explains less than a single explanatory variable. PCA results show that the first two components explain more than 70% of the total variance in a 7-feature dataset, therefore it is possible to represent the distribution of points by considering the charging and discharging processes, which are defined as the initial labels. Fig. 3 illustrates the distribution of the dataset considering the first two principal components and operation type.

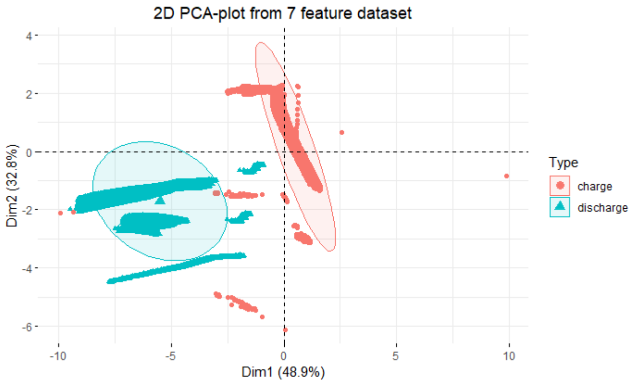

Fig. 3. PCA representation for variables and individuals in the dataset. Points outside the ellipses show an anomaly pattern.

Finally, the matrix of features is separated into two different arrays of [50,285 *7] for discharge and [541, 173 *6] for a charge, which explains the longer duration of the charging process. Before explaining the following sections, it is remarkable to state that PCA is a powerful algorithm that provides support to accomplish dimensionality reduction, however, this does not imply that feature removal is a mandatory task in feature engineering steps, therefore, familiarity with the dataset is highly recommended when implementing Data Mining algorithms, specifically for clustering and classification problems.

B. *Parameter estimation techniques*

In this subsection, the most optimal methods to perform parameter estimation for K-Means, LOF, and DBSCAN are described. It is necessary to mention that there is no completely accurate technique, but there are some that provide more efficient results based on validation and testing.

1) K-Means

To select the optimal quantity of "k" clusters, the Elbow Method was implemented, which is described in the next steps [8]:

- Compute the clustering algorithm for different values of "k", for instance, by varying "k" from 1 to the maximum and desired clusters.
- For each value of "k" calculate the total Within-Cluster Sum-of-Squares (WCSS).
- Plot the curve of WCSS according to the number of clusters specified in the previous step.
- The inclination point (knee) in the plot is considered as an indicator of the optimal number of clusters.

For comparison purposes, the Average Silhouette method is also implemented and is summarized below [9]:

- Compute the clustering algorithm for different values of "k", from the initial to the maximum desired value.
- For each corresponding "k", calculate the average silhouette of the observations.
- The curve of average silhouette observations is plotted, and location of the maximum point is considered as the optimal number of clusters.

To determine the condition for outlier's detection, the following algorithm is implemented using reference and divided into three stages [10]:

- Stage 1: Calculate the pairwise distance for the whole dataset, considering the quantity of observations in the matrix of features and cluster centers for the K-Means algorithm. Take the maximum and minimum value of the calculated pairwise distances. Threshold value= (maximum distance + minimum distance)/2
- Stage 2: If the distance > Threshold value, this point is considered as outlier, otherwise, is a non-outlier.
- Stage 3: Finding out all outliers for a particular dataset based on the previous conditions.

In Fig. 3 and Fig. 4, the Elbow and Average Silhouette methods are illustrated to represent the cluster's selection and clustering quality; it can be appreciated that according to both methodologies, the optimal value of "k" equals three for this example.

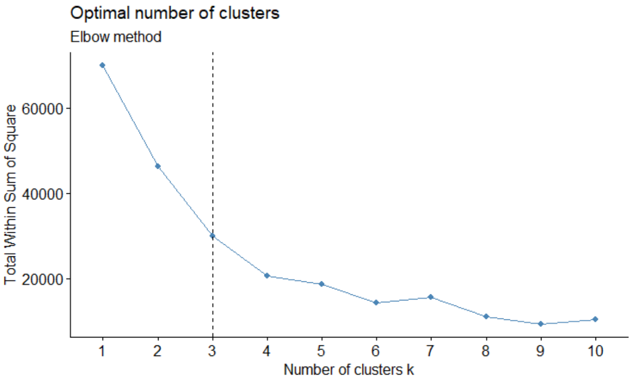

Fig. 4. Elbow method to select the optimal number of clusters in K-Means.

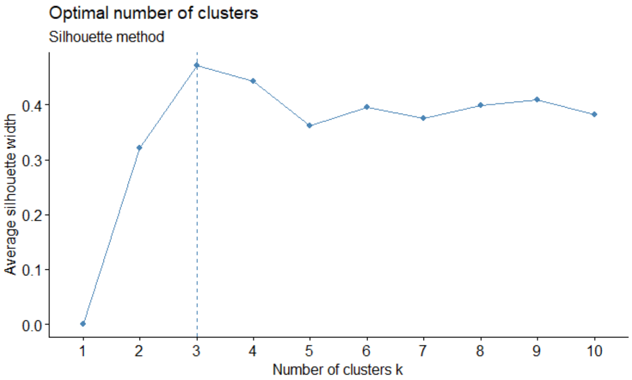

Fig. 5. Average Silhouette method to select the optimal number of clusters in K-Means

2) DBSCAN

DBSCAN is an algorithm implemented for density base clustering that contains a huge amount of data noise and outliers, having Eps and MinPts as the main parameters of performance, denoting the maximum distance between two points and minimum quantity of points, respectively.

To determine the optimal value of Eps, a single-level density algorithm that calculates the slope between the points of the k nearest neighbor distance is implemented, selecting the slope of 1% difference as the optimal Eps value [11]. In Fig. 6, the previous explanation is exemplified. Regarding the selection of MinPts, reference [12] explains a simple but effective heuristic approach that is based on the k-th nearest neighbor distance, a "k-dist" function that maps each point to the k-th nearest neighbor, and a density distribution in the dataset, however, domain knowledge and familiarity are also important to consider when selecting MinPts in the DBSCAN algorithm.

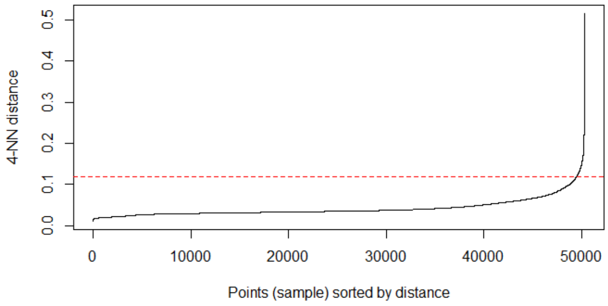

Fig. 6. Determination of the optimal Eps based on a single-level density algorithm. Horizontal red line denotes the selected value.

3) LOF

LOF is a normalized density-based approach. This algorithm detects outliers by comparing the density of each point with the density of its k-nearest neighbor, moreover, its mechanism is integrated by the minimum number of points "q" and the threshold "p". To obtain the optimal value of "q", it is necessary to select a minimum and maximum number of points, after that, for each point take the maximum value over each "q" in the previous specified range; detailed methodology is explained in reference [13].

Regarding the threshold value for outlier detection, the density of the LOF score distribution is considered and visualized, subsequently, quantile point is calculated in accordance with the density distribution, finally, the threshold is adjusted according to the selected quantile and user criteria. Exemplification of the methodology for the threshold selection is shown in Fig. 7.

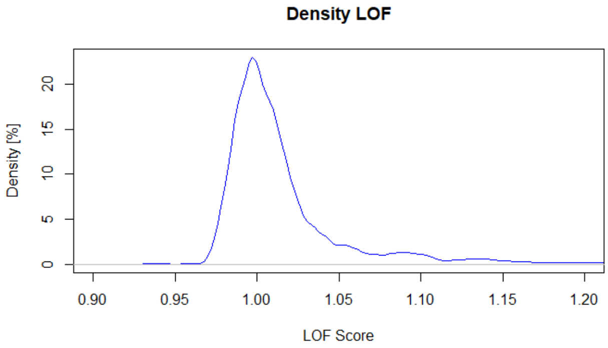

Fig. 7. Determination of the threshold value based on the LOF score distribution.

IV. Results

In this section, K-Means, LOF, and DBSCAN are implemented in the charging and discharging processes. It is fundamental to specify not only that the capacity is restricted to the discharge but also that is the most crucial KPI due to the End-of-Life (EOL) criteria of a BESS.

Usually, the EOL criteria is reached when the capacity of a BESS is lower than 70%-80% of the total rated capacity. It is important to clarify that a battery pack consists of a set of battery modules, in which each module has 12 battery cells. In this dataset, the BESS refers to a battery cell whose total rated capacity is 2 Ah, so the EOL at 70% is reached at approximately 1.4 Ah in cycle number 125. To illustrate the EOL criteria, Fig. 8 shows the capacity of the BESS through the total quantity of cycles, in which each cycle is updated based on different capacity values for all time intervals.

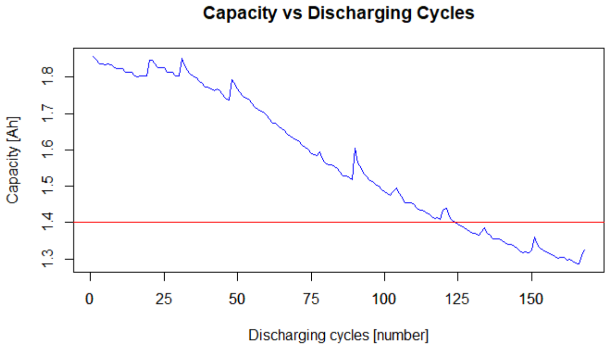

Fig. 8. Capacity vs Discharge cycles. Blue curve denotes the different capacity through the discharge. The horizontal red line denotes the EOL value for the BESS.

Taking KPIs and initial results into account, capacity and cycles will be the focus of the discharge, while the temperature of the charge, all to discuss the results of the clustering and outlier analysis.

A. Charge

K-Means and LOF are algorithms that detect a similar quantity of outliers based on the determination of threshold and the number of clusters, on the other hand, DBSCAN is the algorithm that identifies the least quantity of outliers in the dataset. A remarkable insight corresponds to the values of the outliers related to the KPIs of the BESS, which is found in the same range for all the Data Mining algorithms, specifically during the initial and ending period of the EOL criteria.

In Fig. 9 the outlier analysis is represented by K-Means, DBSCAN and LOF, taking a sample of values and showing a similarity due to the distribution of points in the KPIs, specifically in the intervals for Temperature and Time, demonstrating that the parameter estimation validates the optimal performance of the algorithms.

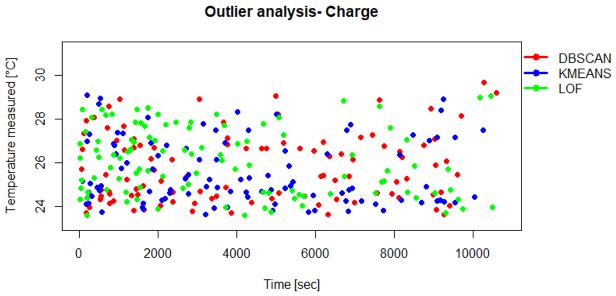

Fig. 9. Outlier analysis representation by DBSCAN, K-Means and LOF in the charging process.

The complexity of the subset in the charging process can be exemplified by the irregular shape of clusters because of the longer duration of the cycles, which in this case is supported by charging the batteries until they are almost at maximum capacity and then starting the discharge. Altering densities can be appreciated in this subset, visually represented by clustering the charging process in Fig. 10.

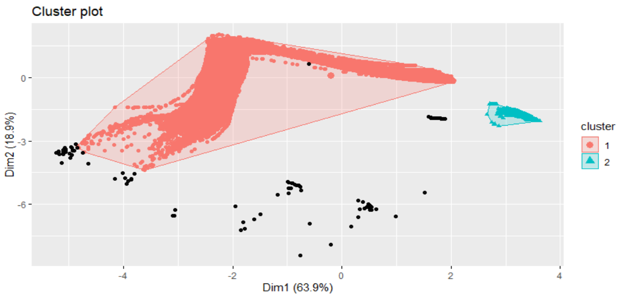

Fig. 10. Clustering representation of the charging process. Two clusters are selected according to parameter estimation techniques. Black points indicate an anomaly behaviour.

The interpretation of the results from an engineering perspective is explained not only by noise data in experimental measurements but also by showing anomaly patterns in intervals outside the range of the temperature distribution, these being contributors to the deterioration of a BESS.

B. Discharge

In this subsection, the capacity and discharging cycles are analyzed to detect outliers before reaching the EOL criteria, specifically to identify the interval of cycles with the major quantity of anomaly patterns.

The results show there are many outliers when the battery capacity is above the mean value that equals 1.5 Ah; furthermore, this result is supported by the duration of the cycles to initiate a degradation in the BESS. It is appreciated that when the EOL criteria is reached, the number of outliers decreases because the degradation previously had an effect, so the KPIs play a key role in sections with the highest density points, which correspond to the 30-75 discharging cycles. A sample of points in Fig. 11 illustrates the outlier analysis, showing a slight similarity in the interval values by implementing DBSCAN and LOF; however, there is a pattern of significant difference in the range of values for the K-Means algorithm

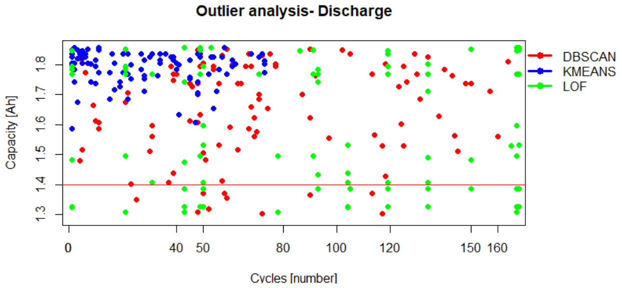

Fig. 11. Outlier analysis representation by DBSCAN, K-Means and LOF in the discharging process. Horizontal red line denotes the EOL value for the BESS

Finally, clustering of the discharge has been successfully implemented and can be appreciated in Fig. 12, in which PCA is executed to illustrate the KPIs based on the first two principal components, complemented by the parameter estimation techniques.

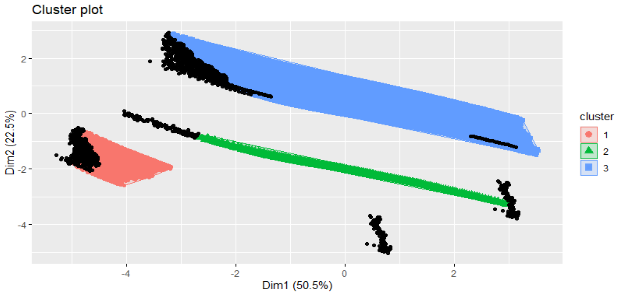

Fig. 12. Clustering representation of the discharging process. Three clusters are selected according to parameter estimation techniques. Black points indicate an anomaly behaviour.

C. Clustering validation and discussion

The implementation of the parameter estimation techniques shows that the number of "k" clusters in the K-Means, MinPts in DBSCAN, and "q" value in LOF play the most important role in increasing or decreasing the quantity of outliers and clusters points, in addition, the threshold will determine the optimal interval values of the of the data points as clusters or outliers for K-Means and LOF, while Eps for DBSCAN.

The validation of the clustering techniques is based on the Silhouette curve, which shows the suitability of the selected clusters according to the Silhouette score. Although the parameter estimation technique for K-Means and the DBSCAN implementation indicate the same number of selected clusters, the Silhouette curve shows a remarkable difference in the shape of the clusters for charging and discharging processes. Fig. 13 shows the lack of efficiency of the K-Means algorithm, which provides an anomaly pattern for groups of clusters 1 and 3, and some negative values for the Silhouette score in cluster 1.

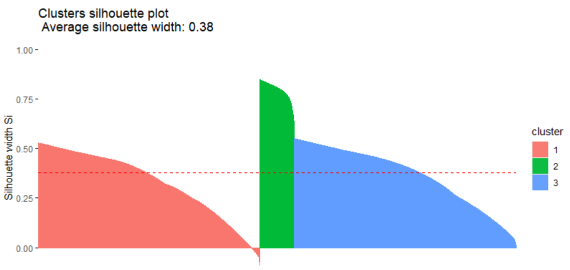

Fig. 13. Silhouette curve for clustering validaton of K-Means algorithm

Due to the assumption of elliptically shaped clusters in the K-Means algorithm, the Silhouette curve shows an anomaly behavior in cluster 1 and 3, while for DBSCAN a different shape of curves and better-quality clustering are achieved. It is necessary to point out that according to the results in the previous subsections, DBSCAN experience difficulties in datasets with major differences in density, so that complications arise when identifying outliers and noise points. A validation result that considers the DBSCAN implementation is appreciated in Fig. 14.

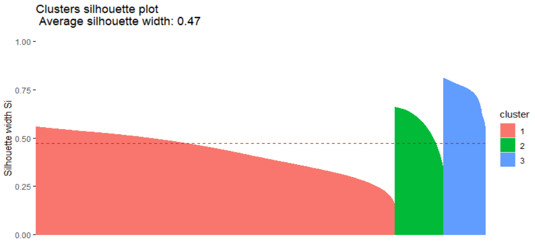

Fig. 14. Silhouette curve for clustering validaton of DBSCAN algorithm

Results show a similarity in the outlier intervals, on the other hand, clustering representation have slight differences, specifically in the shape of the corresponding clusters for charge and discharge, which can be problematic when dealing with more complex datasets during battery measurements. The drawbacks of the proposed methodology are the lack of interpretability of both the clusters and the outliers, so domain knowledge is a key component in the results, all to explain possible bias or noise in the dataset. In addition, sampling and selected input parameters could be potential sources of anomalies in the results; however, these causes may be attributed to the needs and criteria of the user, which are entirely dependent on the tolerance of the algorithm for considering a data point as a cluster or outlier.

V. Conclusion

K-Means, LOF, and DBSCAN are described, analyzed, and implemented to achieve clustering and outlier analysis in a BESS during charging and discharging operations. Our aim was to familiarize the reader with the importance of BESS assessment according to the capacity and EOL criteria. Regarding the Data Mining algorithms, parameter estimation and feature selection techniques must be considered to identify KPIs and provide an optimal performance. Thus, the first step is to become familiar with the robust computational algorithm that explains the behavior of input and output parameters.

Due to its robustness to irregularly shaped clusters, DBSCAN is the optimal algorithm for this particular problem composed of charging and discharging processes, all because in some datasets, the shape of the underlying clusters is already defined implicitly by the underlying distance function or probability distribution, so that Grid and Density-based clustering explore the idea that clusters are of a different density than space between them. Limitations of K-Means rely on clustering datasets where points have distinct size and density, which can lead not only to clustering outliers, but a convergence of a constant value in distance-based similarity measure as the number of dimensions increases. Considering LOF implementation, parameter estimation is straightforward and less complex to achieve compared to K-Means and DBSCAN, consequently, this algorithm is the most efficient to show the tendency of a point and explore outlier analysis in each dataset.

The novelties discussed in this article consist of implementing Data Mining methods to ensure battery model quality, specifically during feature selection and data exploration. In addition, the proposed methodology associates the correlation between the different KPIs to optimize the SOH of a BESS before reaching the EOL criteria, comparing the advantages and disadvantages of each algorithm. Current research improvements will be based on making experimental measurements of different battery cells to assess their performance according to charge and discharge profiles. The future scope of EOL criteria will be determined through more advanced algorithms such as Regression, Binary Classifiers, and Ensemble Learning, considering the KPIs studied in this article. This will help establish assessment and verification procedures for possible fault diagnostics to support commercial consulting, research, and testing for enterprises based on the digital twin concept.

References

[1] E. Banguero et al., "A Review on Battery Charging and Discharging Control Strategies: Application to Renewable Energy Systems", Energies, volume 11, (2018), pp 1-15.

[2] S. Hany, A. Emad, "Effect of the Different Charging Techniques on Battery Lifetime: Review", 2018 International Conference on Innovative Trends in Computer Engineering (ITCE 2018), 2018, pp 421-426.

[3] Z. Zhou et al., "A fast screening framework for second-life batteries based on improved bisecting K-means algorithms combined with fast pulse test, Energy Storage, volume 31, 2020, pp 1-11.

[4] A. Mohammad, J. Chang and T. Chang, "Machine Learning Algorithms of Battery Packs at Smart Manufacturing Industries", Advanced Production and Industrial Engineering, 2022, pp 19-24.

[5] A. Ran et al., "Data Driven Fast Clustering of Second-Life Lithium-Ion Battery: Mechanism and Algorithm", Advanced Theory and Simulations, volume 3, 2020, pp 1-9.

[6] P. Cichosz. Data Mining Algorithms: Explained Using R. Wiley, 2015.

[7] Macintosh, A. 2010, " Li-ion Battery Aging Datasets", Ames Research Center. Web. https://c3.ndc.nasa.gov/dashlink/members/38/, September 2010.

[8] J. Shandong, "Introduction to the K-Means Clustering Algorithm Based on the Elbow Method", Accounting, Auditing and Finance, volume 1, 2020, pp 5-8.

[9] P. Rousseeuw, "Silhouette: A graphical aid to the interpretation and validation of cluster analysis", Journal of Computational and Applied Mathematics, volume 20, 1987, pp 53-65 .

[10] A. Barai, L. Dey, "Outlier Detection and Removal Algorithm in K-Means and Hierarchical Clustering", World Journal of Computer Application and Technology, volume 5, 2017, pp 24-29.

[11] N. Rahmah, I. S. Sitangaang, "Determination of Optimal Epsilon (Eps) Value on DBSCAN Algorithm to Clustering Data on Peatland Hotspots in Sumatra", IOP Conference Series: Earth and Environmental Science, volume 31, 2016, pp 1-5.

[12] M. Ester et al., "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise", Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining, 1996, pp 1-6.

[13] M. Breunig et al., "LOF: Identifying Density- Based Local Outliers", International Conference On Management of Data, 2000, pp 1-12