Data-driven Raw Materials Segmentation Using Unsupervised Learning: A Case Study in Semiconductor Industry

Piyathida Somwong[1] and Naragain Phumchusri[2]

[1]Department of Industrial Engineering, Faculty of Engineering, Chulalongkorn University, Thailand
[2]Department of Industrial Engineering, Faculty of Engineering, Chulalongkorn University, Thailand
6770115921@student.chula.ac.th; naragain.p@chula.ac.th

Keywords: Data-Driven Inventory Segmentation, Raw Material Inventory, Unsupervised Learning, Hierarchical Clustering, Semiconductor Industry

Abstract

This study presents a comparative analysis of three unsupervised learning techniques, namely K-means, K-medoids, and hierarchical clustering, to segment 92 high-value substrate SKUs in a semiconductor back-end manufacturing facility. The research addresses the limitations of traditional single-criterion ABC analysis by integrating six operational and risk-based criteria: demand volume, demand variability, intermittency, lead time, lead time variability, and shelf life. The performance of each method is evaluated using the Silhouette Coefficient and the Davies-Bouldin Index (DBI). The results show that hierarchical clustering achieves the strongest performance among the evaluated methods, with the highest silhouette score (0.440) and the lowest DBI value (0.6742), indicating superior cluster separation and compactness. The findings demonstrate that hierarchical clustering provides a more effective and interpretable segmentation of raw materials, supporting improved raw materials inventory classification in practice.

1 Introduction

The semiconductor manufacturing industry operates under highly uncertain and dynamic conditions, characterized by volatile demand, long and variable lead times, and high-value raw materials with limited shelf life. The production is typically make-to-order and relies on specialized material. In such environments, ineffective raw materials inventory management can result in excessive stock accumulation, obsolescence, and frequent stockouts [1]. The case study considered in this research involves a semiconductor back-end manufacturing company that manages raw materials using a uniform raw material inventory control policy. Despite substantial differences in demand patterns, lead times, and shelf-life constraints across materials, all items are planned to use the same replenishment strategy without systematic segmentation. As a result, the company experiences significant imbalance in raw material inventory performance, with some materials accumulating extremely high days-on-hand levels, while others suffer from recurring shortages. These issues highlight the limitations of applying a single control logic to heterogeneous raw material inventory items. Traditional inventory classification approaches, such as ABC analysis, prioritize items based on usage value. While simple and widely adopted, such single-criterion methods fail to capture important operational characteristics, including demand variability, intermittency, and supply uncertainty [2, 3]. To address these limitations, multi-criteria inventory classification methods have been proposed, incorporating additional demand- and supply-related attributes. For example, Flores et al. [4] introduced the Analytic Hierarchy Process (AHP), while Ng [5] proposed simplified linear models to improve practical applicability. Recent advances in data analytics have enabled the integration of machine learning techniques into inventory classification. In particular, unsupervised learning methods, such as clustering, allow inventory items to be grouped based on similarity across multiple attributes without requiring predefined labels. Among these methods, K-means is widely used due to its computational efficiency and interpretability [6], while K-medoids provides improved robustness to outliers commonly observed in industrial demand data [7]. Hierarchical clustering has also been applied to support structural interpretation through dendrograms [8]. However, different clustering algorithms may produce substantially different segmentation outcomes when applied to the same dataset, which can affect both interpretability and practical decision-making [9].

Motivated by this issue, this study conducts a comparative analysis of three unsupervised clustering techniques applied to raw material inventory data. The objective is to identify a clustering approach that better captures material heterogeneity. This study contributes to the existing literature in three main aspects. First, it provides a comparative evaluation of multiple clustering algorithms using real industrial data, highlighting differences in clustering structure and validation performance. Second, it incorporates multiple operational and risk-related features, including demand variability, intermittency, and lead-time uncertainty, to better reflect real-world raw material inventory characteristics. Third, the study evaluates clustering performance using

multiple internal validation metrics, including the Silhouette Coefficient and the Davies-Bouldin Index (DBI).

2 Methodology

This study employs a data-driven case study approach to examine raw material inventory segmentation for raw materials in semiconductor back-end manufacturing.

2.1. Data description and feature construction

The dataset is obtained from the enterprise resource planning (ERP) system. To characterise raw material inventory behaviour, a set of operational features is constructed for each SKU using historical data. These features include average demand over the most recent 24-month, demand variability measured by the coefficient of variation, demand intermittency represented by the average demand interval (ADI), average replenishment lead time, lead-time variability, and shelf life. These variables capture key dimensions of raw material inventory risk related to demand uncertainty, supply uncertainty, and obsolescence. The descriptive statistics of features are summarized in Table 1.

Table 1 Descriptive statistics of raw material inventory features.

Variable	Mean	Std. Dev.	Min	Max
Mean demand (Monthly)	183,178	396,652	1,387	3,306,246
Demand Variability	1.0187	0.5114	0.2881	2.7543
ADI	1.3040	0.4431	1.0000	3.0000
Lead time (months)	1.9250	0.3415	1.3000	4.0000
Lead time Variability	0.2100	0.2151	0.0000	0.7667
Shelf life (months)	17.1	3.2	6	18

2.2. Data preprocessing

Prior to clustering analysis, the data extracted from the company's ERP and planning systems are reviewed to ensure consistency. To reduce the influence of extreme values commonly observed in raw material inventory data, demand-related variables exhibiting highly skewed distributions are transformed using logarithmic scaling. All variables are then aligned to a common monthly time horizon and standardized using z-score normalization to ensure comparability across features in distance-based clustering analysis[8, 10].

2.3. Clustering algorithms

To investigate alternative inventory segmentation structures, three unsupervised machine learning models are applied to the same dataset which are K-means, K-medoids, and Hierarchical clustering with Ward's linkage.

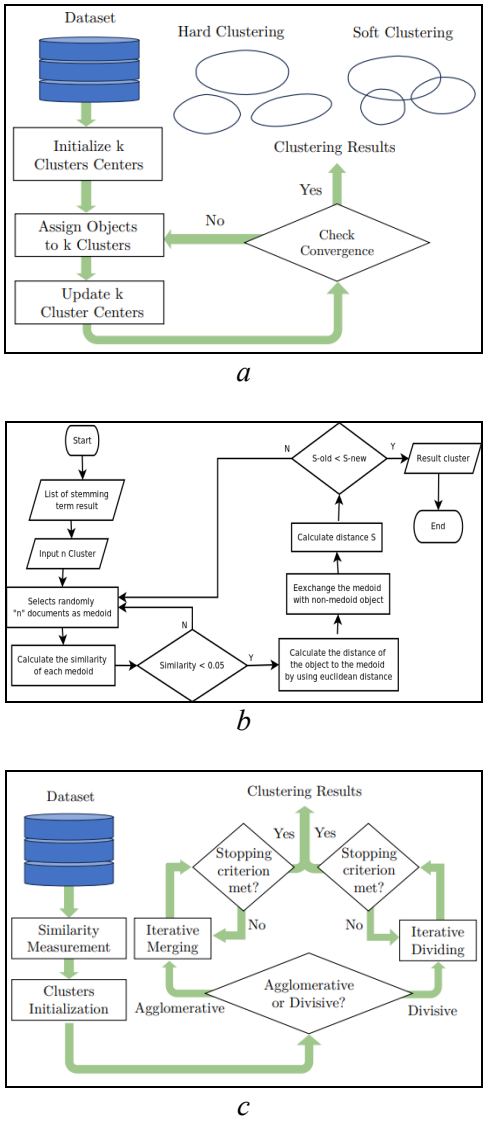

Fig.1 (a) K-means clustering [11], (b) K-medoids clustering [12] and (c) Hierarchical clustering [11].

2.3.1 K-means Clustering Algorithm: in this study, the K-means clustering algorithm is employed to segment raw material SKUs. As illustrated in Fig. 1(a), the algorithm begins by initializing k cluster centroids and iteratively assigns each SKU to the nearest centroid using a distance-based similarity measure. The centroid locations are then updated based on the assigned members, and this process is repeated until convergence is achieved. To determine an appropriate number of clusters, the K-means algorithm is executed for multiple candidate values of k. The Elbow method is employed to evaluate changes in clustering cost, measured by within-cluster inertia, across different values of k. Figure 2(a) shows the elbow curve of within-cluster inertia versus the number of clusters (k), where the point at which the reduction in inertia begins to diminish indicates the selected number of clusters.

2.3.2 K-medoid Clustering Algorithm (PAM): the K-medoids clustering algorithm, implemented using the Partitioning Around Medoids (PAM) approach, is applied as a robust alternative to K-means for raw material SKU segmentation. The overall procedure, shown in Fig. 1(b), follows a similar iterative structure to K-means but differs in that each cluster is represented by an actual data point (medoid) rather than a computed centroid. This medoid-based representation reduces sensitivity to extreme values and enhances clustering stability, particularly for SKUs exhibiting irregular demand or supply patterns. The algorithm iteratively evaluates candidate medoids by minimizing the total dissimilarity between objects and their assigned medoids. Similar to the K-means analysis, multiple values of *k* are examined as shown in Fig. 2(a), and the final number of clusters is selected based on the Elbow method.

2.3.3 Hierarchical Clustering with Ward's Linkage: hierarchical clustering with Ward's linkage is employed to further explore the similarity structure among raw material SKUs without requiring a predefined number of clusters. As depicted in Fig. 1(c), the analysis begins by treating each SKU as an individual cluster and iteratively merges clusters based on minimizing the increase in within-cluster variance. The hierarchical structure is represented by using a dendrogram to visualize the clustering process. Figure 2(b) shows the hierarchical clustering dendrogram, where the vertical axis represents linkage distance based on Ward's method. Meaningful cluster partitions are identified by examining changes in linkage distance, and the final number of clusters is determined by selecting an appropriate cut level in the dendrogram. This hierarchical approach complements the partitional clustering methods by providing additional insight into cluster similarity and separation, supporting the interpretation and validation of the final clustering structure.

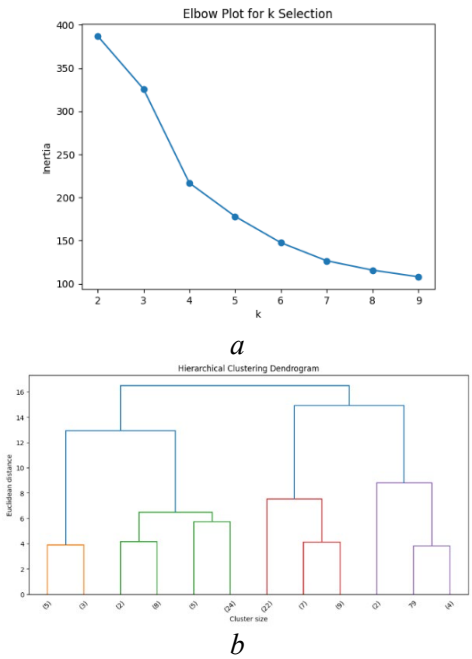

Fig. 2 (a) Elbow curve of within-cluster inertia versus the number of clusters k and (b) Hierarchical clustering dendrogram based on Ward's linkage.

2.4. Model validation and cluster interpretation

To evaluate the quality of clustering results, multiple internal validation metrics are employed. The Silhouette Coefficient is used as a primary measure to assess cluster cohesion and separation based on the relative distance between data points within the same cluster and those in neighbouring clusters [13]. Higher silhouette values indicate more compact and well-separated clusters. In addition to the silhouette coefficient, the Davies-Bouldin Index (DBI) is used to further evaluate clustering performance. The DBI measures the average similarity between each cluster and its most similar counterpart, considering both within-cluster dispersion and between-cluster separation [14]. Lower DBI values indicate better clustering performance, reflecting more compact and well-separated clusters. The use of multiple validation metrics provides a more comprehensive assessment of clustering quality, as different indices capture different aspects of cluster structure. While the silhouette coefficient emphasizes the relative positioning of individual data points, the DBI evaluates cluster-level compactness and separation. However, clustering solutions are not selected solely based on numerical validation metrics. In this study, quantitative results are interpreted together with the consistency and interpretability of cluster characteristics, following established practices in clustering-based analysis[10]. For each clustering model, descriptive statistics of demand, supply, and shelf-life features are examined across clusters to identify meaningful and distinguishable raw material inventory groups. This combined evaluation approach ensures that the selected clustering solution is not only statistically valid but also practically interpretable and relevant for raw material inventory segmentation in real-world applications.

3 Results

This section presents and compares the clustering results obtained from the three unsupervised machine learning models which are K-means, K-medoids, and hierarchical clustering with Ward's linkage, applied to the same raw material dataset. The comparison focuses on clustering structure, validation results, and the interpretability of cluster characteristics in the context of raw material inventory segmentation.

3.1. Clustering structure and results validation

All three clustering models divide the raw material SKUs into four clusters with similar overall patterns, as shown in Table 2.

Table 2 Summary of clustering results, cluster sizes, average silhouette score and Davies-Bouldin Index (DBI).

Model	No. of Clusters	Cluster Size	Average Silhouette Score	Average DBI Score
K-Means	4	38 / 37 /10 / 7	0.314	1.026
K-Medoids	4	46 / 30 / 9 / 7	0.306	1.005
Hierarchical	4	39 / 38 / 8 /7	0.440	0.674

All three clustering models divide the raw material SKUs into four clusters with similar overall patterns, as shown in Table 2. K-means and hierarchical clustering produce relatively balanced sizes for the two large clusters, indicating that these methods tend to distribute SKUs more evenly based on overall similarity. In contrast, K-medoids assigns a larger proportion of SKUs to one dominant cluster, while fewer SKUs are placed in the remaining clusters. This reflects the robustness of K-medoids in handling extreme or irregular raw material inventory patterns by focusing on representative data points (medoids). Hierarchical clustering shows a cluster size pattern similar to that of K-means. However, its grouping depends on how the dendrogram is cut. As a result, SKUs near cluster boundaries may be grouped differently, even though the total number of clusters remains the same. Across all models, the cluster size distributions consist of two large clusters and two smaller clusters. This consistent structure suggests the presence of an inherent grouping pattern in the raw material inventory data, rather than clusters being driven solely by algorithmic differences. The two large clusters represent materials with broadly similar raw material inventory behaviour, while the smaller clusters capture more distinct and extreme patterns. Clustering quality is evaluated using both the Silhouette Coefficient and the Davies-Bouldin Index (DBI), as summarized in Table 2. The silhouette score measures how well SKUs are grouped within clusters while remaining distinct from other clusters, with higher values indicating better cohesion and separation. The DBI evaluates cluster compactness and separation at the cluster level, where lower values indicate better clustering performance. Hierarchical clustering achieves the highest silhouette score (0.440), which is noticeably higher than those of K-means (0.314) and K-medoids (0.306). In addition, hierarchical clustering yields the lowest DBI value (0.6742), compared with K-means (1.0258) and K-medoids (1.0046). These results consistently indicate that hierarchical clustering provides superior clustering performance in terms of both cluster cohesion and separation. In contrast, K-means and K-medoids produce lower silhouette scores and higher DBI values, suggesting greater overlap between clusters and less distinct cluster boundaries. Although this overlap reduces internal validation performance, it does not necessarily imply that the resulting clusters lack practical value. Instead, these findings highlight an important trade-off between clustering quality, as measured by internal validation metrics, and the interpretability of clustering results in practical applications.

3.2. Comparison of cluster characteristics

The cluster characteristics obtained from three algorithms are summarized in Tables 3–5. Across all three models, four clusters with similar raw material inventory patterns are consistently identified, indicating that the underlying segmentation structure is stable and not dependent on a specific clustering technique.

Table 3 Summary of cluster characteristics based on K-means

Cluster	A	B	C	D
No. of SKU	38	37	10	7
Mean Demand	35,616	374,620	5,703	225,860
Demand Variability	1.022	0.725	2.0836	1.0317
ADI	1.25	1.08	2.34	1.3
Lead Time (month)	1.86	1.85	1.85	2.5
Lead Time Variability	0.0733	0.3853	0.0863	0.2017
Shelf Life (Month)	18	18	18	6

Table 4 Summary of cluster characteristics based on K-medoids

Cluster	KM0	KM1	KM2	KM3
No. of SKU	30	46	9	7
Mean Demand	26,363	313,674	5,719	225,860
Demand Variability	1.12	0.7314	2.1392	1.0317
ADI	1.31	1.09	2.41	1.3
Lead Time (month)	1.96	1.79	1.83	2.5
Lead Time Variability	0.0732	0.3335	0.0414	0.2017
Shelf Life (Month)	18	18	18	6

Table 5 Summary of cluster characteristics based on Hierarchical

Cluster	HC1	HC2	HC3	HC4
No. of SKU	39	38	8	7
Mean Demand	30,309	369,777	4,725	225,860
Demand Variability	1.0777	0.7188	2.1441	1.0317
ADI	1.29	1.08	2.48	1.3
Lead Time (month)	1.85	1.87	1.81	2.5
Lead Time Variability	0.0869	0.3722	0.0465	0.2017
Shelf Life (Month)	18	18	18	6

Despite this overall consistency, differences are observed in how clearly raw material inventory risk characteristics are separated. Hierarchical clustering provides the clearest distinction for intermittent and high-risk items. As shown in Table 5, cluster HC3 characterized by very low demand combined with the highest demand variability and intermittency. These profiles indicate different sources of raw material inventory risk that are clearly separated under hierarchical clustering. K-means and K-medoids produce comparable raw material inventory groupings but show greater overlap in variability-related characteristics, particularly among low-demand items. As reported in Tables 3 and 4, clusters with different risk profiles exhibit similar demand variability or intermittency values, which may reduce the clarity of segmentation. This overlap suggests that

partitional methods are more influenced by demand magnitude, while hierarchical clustering better captures the interaction between demand uncertainty and supply-related risk.

3.3. Discussion and model selection

The comparison results in Sections 3.1–3.2 indicate that all three clustering models identify a consistent underlying segmentation structure in the raw material data. However, the key differences among the models lie in how effectively they separate demand uncertainty and supply-related risks across clusters. Hierarchical clustering demonstrates the strongest performance among the evaluated methods. This is supported by both its highest silhouette score and lowest Davies-Bouldin Index (DBI), indicating superior cluster cohesion and separation. In particular, hierarchical clustering more clearly distinguishes materials characterized by high intermittency and variability, which represent critical sources of raw material inventory risk. The dendrogram structure further enhances interpretability by revealing hierarchical relationships among SKUs and enabling transparent identification of cluster boundaries. In addition, the PCA visualization in Fig. 3 provides supporting evidence of cluster separation in a reduced-dimensional space. While PCA is used here solely for visualization purposes, the projected results show that clusters, especially those associated with high-risk and intermittent demand patterns, are reasonably well separated. However, the observed overlap in variability-related characteristics, together with their lower silhouette scores and higher DBI values, suggests that these methods are less effective in distinguishing raw material inventory risk driven by intermittency and supply uncertainty. Based on the combined evaluation of clustering structure, and validation metrics, hierarchical clustering is selected as the most appropriate method for the case study.

From a practical perspective, the findings suggest that clustering-based inventory segmentation can support more differentiated inventory control strategies compared with traditional single-policy approaches. In particular, materials grouped within high-intermittency and high-variability clusters may require more conservative replenishment policies, higher safety stock levels, or closer monitoring to reduce stockout risks. In contrast, clusters characterized by stable demand patterns may be managed using more cost-efficient replenishment strategies. The ability of hierarchical clustering to distinguish these operational characteristics more clearly may therefore improve the alignment between inventory policies and actual material behaviour.

The results also highlight the importance of selecting clustering algorithms based not only on numerical validation metrics but also on interpretability and operational relevance. Although K-means and K-medoids are computationally efficient and widely adopted in industrial applications, their partitioning mechanisms may be less effective in capturing complex interactions among demand uncertainty, intermittency, and supply-related factors. Hierarchical clustering, on the other hand, preserves structural relationships among SKUs through the dendrogram representation, enabling more transparent interpretation of cluster boundaries and similarity structures. This characteristic is particularly valuable in industrial environments where explainability and managerial interpretability are essential for practical implementation.

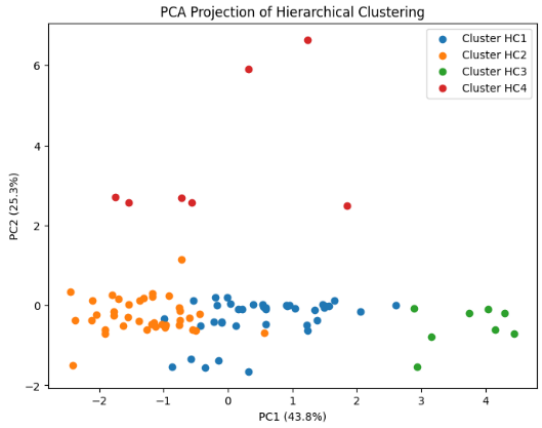

Fig. 3 PCA visualization of hierarchical clustering results

4 Conclusion

This study compares three clustering methods for raw material inventory segmentation and finds that all models identify a consistent four-cluster structure, indicating the presence of an inherent grouping pattern in the raw material data. Among the evaluated approaches, hierarchical clustering demonstrates the strongest performance, achieving both the highest silhouette score and the lowest Davies-Bouldin Index (DBI), which indicates superior cluster cohesion and separation. The results show that hierarchical clustering provides clearer differentiation of raw material inventory risk, particularly in capturing demand intermittency and supply uncertainty in addition to demand magnitude. This leads to more interpretable cluster structures that align closely with the company's practical understanding of raw material behaviour. Compared with partitional methods, hierarchical clustering more effectively distinguishes high-risk and intermittent demand items. Although this study does not explicitly quantify operational performance improvements, the clustering results provide a structured basis for differentiated raw material inventory segmentation and decision-making in practice. The findings highlight the importance of incorporating multiple operational and risk-related factors in raw material inventory classification beyond traditional single-criterion approaches. Future research will extend this clustering framework to raw material inventory policy design and quantitatively evaluate its impact on key performance indicators such as inventory cost, service level, and stockout reduction.

5 Acknowledgements

The authors would like to acknowledge the participating semiconductor manufacturing company for providing the industrial data, operational information, and practical insights that supported this research. The authors also appreciate the valuable guidance and support received throughout the study.

6 References

[1] B. Li, "Procurement and Inventory Control Mechanisms for Critical Raw Materials in Semiconductor Supply Chains," *Journal Européen des Systèmes Automatisés,* vol. 58, no. 4, pp. 863–875, April, 2025.

[2] B. E. Flores, and D. C. Whybark, "Implementing multiple criteria ABC analysis," *Journal of Operations Management,* vol. 7, no. 1, pp. 79–85, Oct, 1987.

[3] R. Ramanathan, "ABC inventory classification with multiple-criteria using weighted linear optimization," *Computers & Operations Research,* vol. 33, no. 3, pp. 695–700, March, 2006.

[4] B. E. Flores, D. L. Olson, and V. Dorai, "Management of multicriteria inventory classification," *Mathematical and Computer modelling,* vol. 16, no. 12, pp. 71–82, 1992.

[5] W. L. Ng, "A simple classifier for multiple criteria ABC analysis," *European Journal of Operational Research,* vol. 177, no. 1, pp. 344–353, February, 2007.

[6] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in Proc. 5th Berkeley Symp. Mathematical Statistics and Probability, 1967, pp. 281–297.

[7] C. C. Campo, K. P. Jimenez, L. V. Peñaloza, L. A. Cristobal, C. L. Vargas, and J. S. Herrera, "Optimization Strategies for Industrial Spare Parts Inventory Based on Advanced Data Analysis Techniques," *Procedia Computer Science,* vol. 257, pp. 1160–1165, January, 2025.

[8] E. Balugani, F. Lolli, R. Gamberini, B. Rimini, and A. Regattieri, "Clustering for inventory control systems," *IFAC-PapersOnLine,* vol. 51, no. 11, pp. 1174–1179, January, 2018.

[9] A. A. Qaffas, M. A. B. Hajkacem, C. E. B. Ncir, and O. Nasraoui, "Interpretable Multi-Criteria ABC Analysis Based on Semi-Supervised Clustering and Explainable Artificial Intelligence," *IEEE Access,* vol. 11, pp. 43778–43792, 2023.

[10] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognition Letters,* vol. 31, no. 8, pp. 651–666, June, 2010.

[11] T. Dinh, H. Wong, D. Lisik, M. Koren, D. Tran, P. S. Yu, and J. Torres-Sospedra, "Data clustering: a fundamental method in data science and management," *Data Science and Management*, August, 2025.

[12] M. Milkhatun, A. Rizal, N. Asthiningsih, and A. Latipah, "Performance Assessment of University Lecturers: A Data Mining Approach," *Khazanah Informatika : Jurnal Ilmu Komputer dan Informatika,* vol. 6, August, 2020.

[13] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics,* vol. 20, pp. 53–65, Nov ember, 1987.

[14] D. L. Davies, and D. W. Bouldin, "A Cluster Separation Measure," *IEEE Transactions on Pattern Analysis & Machine Intelligence,* vol. 1, no. 02, 1979.