International Journal of Engineering Trends and Technology
ISSN: 2231-5381/ https://doi.org/10.14445/22315381/IJETT-V74I7P131

Volume 74 Issue 7, 522-546, July 2026

Original Article

Application of Clustering Techniques for the Analysis and Classification of Firms in the Automotive Industry

Chaimae FAIZ[1], Mouna EL MKHALET[2], Mohammed SBIHI[3]

[1,2,3]*Laboratory for Systems Analysis, Information Processing and Industrial Management (LASTIMI), Higher School of Technology of Salé, Mohammed V University in Rabat, Morocco.*

[1]*Corresponding Author : Chaimae.faiz99@gmail.com*

Received: 29 July 2025 Revised: 08 January 2026 Accepted: 27 January 2026 Published: 28 July 2026

Abstract - *According to this study, five clustering algorithms are compared theoretically and experimentally. The clustering algorithms are DBSCAN, KMeans, Spectral Clustering, OPTICS, and K-Medoids. By using 20 automotive companies' financial data, we evaluate the performance of each algorithm based on intra- and inter-cluster variance, noise robustness, and data perturbation sensitivity. The results demonstrate that the K-Means technique demonstrates the highest stability towards data variation, while the DBSCAN and OPTICS techniques are able to handle noise and discover clusters of different arbitrary shapes. Although spectral clustering is capable of capturing nonlinear structures, it is more sensitive to parameter changes. K-Medoids uses real data points as cluster centers, thereby providing robustness to outliers. One of the novel contributions of this work is the sensitivity analysis, which highlights the method's robustness or fragility under controlled perturbations of the data. Each of these deductions is useful in choosing clustering methods according to the application.*

Keywords - *Clustering algorithms, Customer segmentation, Density-based clustering, Sensitivity analysis, Unsupervised learning.*

1. Introduction

1.1. Scientific and Applied Context of Clustering

Machine learning classification methods fall into three broad categories: supervised, semi-supervised, and unsupervised. Supervised learning relies on predefined labels, while semi-supervised learning combines labeled and unlabeled data to reduce the cost of manual labeling [1]. Unsupervised learning, including clustering, can detect hidden structures in data without prior labels [2]. Clustering is considered a fundamental problem in unsupervised learning, with the main objective of grouping data points into sets of similar elements [3]. It consists of grouping objects based on their similarity, without prior knowledge of their relationships [4]. This technique has been widely studied in various contexts, including time series analysis [4], mobile customer migration detection [5], technology investment assessment in manufacturing and service industries [6], and the study of the socio-economic impacts of ICT in African economies [7]. Recent approaches, such as hierarchical clustering based on reciprocal neighbors, have been proposed to improve clustering accuracy [8]. Other works have explored current trends and applications of clustering in various fields [9]. Clustering is particularly useful for analyzing large and heterogeneous data, grouping similar points together while distinguishing them from others [10]. However, the performance of clustering algorithms varies depending on the type of data and application domains, as they must handle multidimensional, noisy, incomplete, or sampled data [1]. In customer management and logistics, clustering helps identify distinct customer segments based on similar behaviors or preferences, facilitating personalized marketing strategies and better resource allocation [9]. It is also essential for route planning, inventory management, and demand forecasting, by grouping similar customers or products together, reducing transportation costs, and improving delivery efficiency [11]. It helps reveal trends in purchasing behavior, contributing to better demand anticipation and more accurate inventory management [12]. In complex supply chains, clustering methods help streamline processes, reduce operational costs, and improve the overall customer experience [13].

2. Problem and Research Gap

Recent studies highlight the adaptability of clustering to reveal hidden patterns in large unlabeled datasets, especially in deep learning [14]. It also proves effective in extracting relevant information from noisy data [15] and simplifies analysis and classification [16]. As noted by [17], clustering plays a key role in applications such as document clustering, big data processing, and anomaly detection, improving information retrieval, text categorization, and pattern recognition. Moreover, [18] highlights its effectiveness in handling high-dimensional data, simplifying large datasets for

more efficient decision-making. According to [19], recent advances in clustering techniques continue to improve accuracy and performance. Finally, the studies by [20-22] highlight the contribution of clustering in data visualization and decision-making in various disciplines.

The automotive industry is now a strategic pillar of the Moroccan economy. In 2025, the sector recorded remarkable growth, with over 66,000 vehicles sold in the first four months, representing more than 27% of the country's industrial exports. It employs over 220,000 people and comprises more than 250 companies, ranging from manufacturers and suppliers to specialized subcontractors. Through this industrial dynamism, Morocco aspires to become one of the main global players in the automotive value chains. In this context, a thorough understanding of the economic structure of companies in the sector becomes a strategic imperative. Methods for unsupervised learning help in the identification of natural. Companies are arranged in groupings based on their economic status and features, without prior categorization. Ultimately, the tools are used for strategic segmentation. Detection of sector leaders and the optimization of the sector. guidelines, and the direction of investments.

2.1. Originality and Research Objectives

Even with all that work, clustering is important. Applications were acquired, although they are not popular within the Moroccan automotive industry. This discrepancy is substantial since. It takes away decision-makers of analytic tools adapted to a. the fast-growing sector of economic logistics, Challenges of a strategic nature. Our research intends to bridge this gap by proposing an original approach. approaches: comparative study of five clustering algorithms, K-Medoids. Spectral Clustering and OPTICS were algorithms applied to a dataset by 20 Moroccan companies working in the automotive sector. We include a sensitivity analysis, unlike past efforts. To evaluate each method's robustness in the case of data modification. Agitations. This donation permits a better. Understanding of sector structure, identification of Consistent corporate categories, and strategic coordination support. Deciding.

The present study aims to derive homogeneous groups of automotive companies operating in Morocco based on financial indicators and to compare the performance of various clustering techniques in this context. In order to achieve this, we ask two powerful questions.

What are the economic and financial characteristics that help identify business clusters in the Moroccan automotive sector?

What are the most effective clustering algorithms based on metrics such as intra- and inter-cluster variance and robustness to noise and sensitivity to perturbations? The questions provide the framework for the comparative analysis undertaken in this study and highlight the methodology and practice of clustering techniques in a strategic industrial sector.

This research stands out because of several new features that make it valuable for both science and real-world use. First, it uses a unique database with real financial data from 20 Moroccan companies in the auto industry, which has not been much studied in existing research. Second, it introduces a fresh way to compare five different clustering methods-K-Means, DBSCAN, Spectral Clustering, OPTICS, and K-Medoids. This comparison is made better by a sensitivity analysis that checks how reliable these methods are when there are changes in the data.

This approach looks at both the algorithms and how they change over time, helping to understand groups of companies in a way that can influence industrial policies, investment plans, or support programs. Lastly, the study gives practical help to both researchers and decision-makers by offering a clear, repeatable method that can be used in other areas of industry in Morocco or the Maghreb region.

3. Methodology

This research uses a comparison method to look at five different clustering techniques applied to real financial data from Moroccan companies in the automotive industry. To make the analysis solid and clear, several steps were taken in the way the study was done.

3.1. Data Source

The data was gathered from public and official sources such as the Moroccan Industry Portal, annual reports from company websites, publications from the Exchange Office, and documents from the Ministry of Industry and Trade. The time period covered is from 2021 to 2023, and the data includes important financial information.

3.2. Criteria for Selecting Companies

Twenty companies were picked based on these rules:

- They must be part of the automotive sector, like manufacturers, suppliers, or subcontractors,
- They must have financial data for at least two full years,
- They should represent different areas and regions,
- They must have been active and running smoothly at the time of the study.

This selection aims to guarantee sufficient diversity while ensuring data comparability.

3.3. Data Preprocessing

Before applying the algorithms, the data underwent several preprocessing steps:

- Cleaning: removal of duplicates and verification of value consistency.
- Handling missing values: imputation by mean for quantitative variables; exclusion of variables with incomplete data.
- Variable selection: the selected indicators include revenue, net income, growth rate, debt ratio, and R&D investments, chosen for their economic relevance and discriminatory power.

3.4. Clustering Algorithms

Five algorithms were selected for their methodological complementarity:

- K-Means: a fast-partitioning method, suitable for well-separated data.
- DBSCAN: a density-based algorithm capable of detecting non-spherical cluster shapes and handling noise.
- Spectral Clustering: a graph theory-based technique, effective for capturing complex structures.
- OPTICS: a variant of DBSCAN allowing visualization of clusters at different densities.
- K-Medoids: a robust alternative to K-Means, using real points as cluster centers.

The choice of these algorithms allows for a balanced comparison between density-based and graph-based partitioning methods. A sensitivity analysis was integrated to evaluate the robustness of the methods to data perturbations.

4. Results

4.1. State-of-the-Art

4.1.1. DBSCAN: A Density-Based Clustering Approach and Its Parameter Challenges

Hierarchical and partitioning clustering methods are only effective in detecting clusters with a spherical shape [23]. In contrast, density-based clustering methods do not depend on parametric distributions or variance, enabling them to detect arbitrarily shaped clusters, handle different levels of noise, and function without requiring prior knowledge of the number of clusters [24]. DBSCAN is currently used in the automotive industry to detect anomalies in production lines, particularly using data from onboard sensors [25]. Introduced at the International Conference on Knowledge Discovery and Data Mining. DBSCAN defines clustering based on the principle that clusters are formed by high-density regions of data points, which are separated by areas of low density [26]. It relies on two key parameters-ε (epsilon), which defines the radius of a point's neighborhood, and minPts, the minimum number of points needed within that radius to form a cluster-making their appropriate selection critical for the algorithm's effectiveness [27]. The optimal values of MinPts and Eps in DBSCAN change with different datasets and even different areas of application in the same dataset, and there are no theoretical guidelines for determining the MinPts and Eps parameters, making parameter tuning heavily based on experience and the need for a large number of adjustments at the same time, limiting the broader application of the algorithm and leading researchers to explore methods for optimizing these two parameters [28]. DBSCAN further refines the clustering process to ensure accurate grouping of data points: a core point xi has at least MinPts neighbors within its ε-neighborhood, a border point y is directly density-reachable from but has fewer than MinPts neighbors, and a noise point is neither a core nor a border point, meaning it does not belong to any cluster [29]. In the DBSCAN algorithm, a point q is said to be directly density-reachable from a point p if q lies within the ε-neighborhood of p and p is a core point, meaning that it has at least *MinPts* neighbors within its ε-neighborhood. A point q is density-reachable from p if there exists a sequence of points starting from p and ending at q, where each point in the sequence is directly density-reachable from the previous one. Finally, two points p and q are density-connected if there exists a third point o from which both p and q are density-reachable, indicating that they belong to the same cluster [20].

DBSCAN: Advantages and Disadvantages

Table 1. Advantages and disadvantages of DBSCAN

Strengths of DBSCAN	Limitations of DBSCAN
Capability to detect clusters of different shapes while effectively managing noisy data [30]	DBSCAN suffers from poor scalability due to the nearest neighbors 'query (NN-QUERY) operation, which is repeatedly performed to determine whether a point belongs to a dense region or is noise, leading to high computational complexity [30]
No prior specification of the number of clusters is required**:** In contrast to *k*-means, DBSCAN automatically determines the number of clusters based on data density, without requiring this parameter to be defined in advance [26]	The performance of the algorithm depends on the appropriate selection of the parameters (minPts and ε), which can be challenging to determine optimally for different datasets [31]
Minimal User Input: The only two input parameters are MinPts and Eps [32]	DBSCAN assumes uniform cluster density, making it ineffective for datasets with varying densities, as it may merge clusters with different densities or overlook low-density clusters [25]

The algorithm can detect clusters even when objects are distributed heterogeneously, such as along paths of a graph or a road network [31]	DBSCAN's reliance on a distance metric (e.g., Euclidean distance) to define neighborhoods makes it less effective for high-dimensional data or datasets where a single metric fails to capture true similarity [25]

DBSCAN: Applications

Clustering algorithms are ubiquitous in daily life, with various applications [33].

[34] Utilized DBSCAN for 3D surface segmentation to extract and segment point clouds based on local quadric surfaces. [35] Applied DBSCAN in acoustic emission source localization to accurately identify and monitor damage in CFRP-concrete composite structures. [36] Proposed a method for optimizing DBSCAN hyperparameters using a Bayesian optimization model, achieving superior clustering results on both simulated and real datasets. [37] Integrated DBSCAN with the YOLOv5 framework to enhance 3D object detection and segmentation on mobile platforms. [38] Employed DBSCAN for energy-efficient Virtual Machine (VM) placement in cloud data centers, optimizing VM allocation to reduce energy consumption. [39] Used DBSCAN in parameter inversion to analyze the morphology characteristics of laser-cut edges in Zr-4 alloy.

[40] Combined DBSCAN with Shannon's entropy to define urban boundaries within the Mexican National Urban System. [41] Introduced a Granular-Ball (GB) approach to DBSCAN, reducing time complexity and enhancing clustering efficiency. [42] developed GriT-DBSCAN, a grid-based variant to optimize spatial clustering and handle large databases efficiently.

[43] Proposed a heterogeneous intensity-based DBSCAN (iDBSCAN) framework to characterize urban attention distribution in digital twin cities by jointly incorporating traffic flow, population density, and urban activity data, enabling dynamic and adaptive regional segmentation. [44] Utilized a spatial–temporal DBSCAN (ST-DBSCAN) model for detecting public transit user activities through smart card data analysis.

[45] Proposed a Multi-density DBSCAN (MDBSCAN) model based on relative density, improving clustering performance for datasets with varying densities. [46] Used DBSCAN for membership determination of stars in twelve open clusters, analyzing Gaia DR3 data and validating findings with spectroscopic data. [47] Applied DBSCAN in lithium-ion battery failure diagnosis, integrating Particle Swarm Optimization (PSO) and Simulated Annealing (SA) to improve fault detection accuracy. [48] Developed a multivariate hierarchical DBSCAN model for maritime data analytics, optimizing data structure extraction from complex maritime datasets. [49] Proposed a novel FlowSort-DBSCAN approach for flood risk grading assessment, enhancing accuracy by considering hazard, exposure, and vulnerability interactions. [50] Employed DBSCAN for coronary artery segmentation in X-ray angiographic images, improving segmentation accuracy for coronary artery disease diagnosis and treatment.

DBSCAN Algorithm:

The DBSCAN algorithm outlined here is based on the work of [51]:

Step 1: Select an arbitrary point:

Choose an arbitrary unprocessed point p from the dataset P.

Step 2: Retrieve the ε-neighborhood:

Find all points within the ε-neighborhood of p:

$$N\varepsilon(p) = \{q \in P \mid dist(p,q) \leq \varepsilon\} \quad (1)$$

Where dist (p, q) is the Euclidean distance between points p and q :

$$dist(p,q) = \sqrt{\sum_{i=1}^{d}(p_i - q_i)^2} \quad (2)$$

Step 3: Determine if p is a core point:

A point p is a core point if:

$$|N_\varepsilon(p)| \geq MinPts \quad (3)$$

If true, assign p to a new cluster and expand it (see Step 4).

If false, mark p as noise (temporarily) and proceed to the next point (Step 6).

Step 4: Expand the cluster

For each point $q \in N_\varepsilon(p)$

If q is unvisited, mark it as visited and compute its ε-neighborhood:

$$N_\varepsilon(q) = \{s \in P \mid dist(q,s) \leq \varepsilon\} \quad (4)$$

If q is a core point ($N_\varepsilon(q) \geq$MinPts), add all points into the cluster.

If q is not a core point but is directly density-reachable from a core point, label it as a border point and include it in the cluster.

$$|N_\varepsilon(p)| < MinPts$$

and

$$\exists q \in N_\varepsilon(p): |N_\varepsilon(q)| \geq MinPts \quad (5)$$

Step 5: Continue expanding until no more points can be added

Repeat Step 4 for all points in the current cluster until no additional density-reachable points can be found.

Step 6: Process the next unvisited point

Move to the next unvisited point in P and repeat Steps 1-5 until all points have been processed.

The time complexity of the DBSCAN algorithm ranges from O (n log n) to O(n^2), depending on the number of data points and the use of spatial indexing, with DBSCAN generally achieving O (n log n) through efficient data structures like kd-trees or R-trees for neighborhood [52, 53].

4.1.2. K-Means: Overview

K-Means clustering is an unsupervised learning technique that groups data points based on their similarity by calculating the distance to centroids. It updates the centroids iteratively by using the assigned points until they stabilize, or the specified number of iterations is reached. [54]. Initially proposed in 1967 and subsequently improved, K Means has become a widely used algorithm for partitioning datasets into clusters. Even though it has been used for a long time, the algorithm still has some problems. One big issue is that it takes a lot of computer power because it needs to go through large data sets many times. Because of this, people are looking for ways to reduce the number of times the data is scanned while still keeping the results good [55].

Also, K-Means algorithms, where the letter "K" stands for the number of groups or clusters, depend a lot on how they start and need the number of clusters to be chosen before they begin. However, in real situations, people often do not know how many clusters there really should be. [56]. K-Means is known for being quick, strong, and easy to use. It groups objects by measuring the squared distance between each object and the closest center point, which works well when the data is clearly separated. [57]. Even with its challenges, K-Means is considered one of the top ten algorithms in data mining [58, 59].

K-Means: Advantages and Disadvantages

Table 2. Advanatges and disadvantages of K-means

Advantages	Limitations
K-means is a simple approach that makes it a popular choice for high-throughput analyses [60]	A major limitation of the K-means algorithm is its inability to efficiently process diverse types of data [61]
It can be adapted for large datasets and offers interactive visualization tools, as seen in applications such as Freecyto [60]	Such a clustering algorithm requires the number of clusters to be specified in advance, leading to varying cluster shapes and outlier effects [61]
Apart from the number of clusters (k), no additional parameters are needed for the clustering procedure [62]	The K-means algorithm can restrict the availability of relevant data to a limited subset of points, which can potentially reduce the accuracy and applicability of the results in more complex or varied real-world scenarios [63]
The K-means algorithm is relatively fast, making it widely used in various practical scenarios [64]	K-means is limited to continuous data, which restricts its applicability to other types of data, such as binary, counting, or ordinal data [64]

K-Means: Applications

[65] Used distributed K-means and Fuzzy C-means algorithms for efficient clustering of data in sensor networks, based on multi-agent consensus theory. [66] Have applied spectral ensemble clustering via weighted K-means to provide theoretical and practical evidence in knowledge and data engineering.

[67] Have proposed a review of clustering techniques applied to the coordination of intelligent vehicle subsystems, highlighting their role in performance optimization and functional segmentation of automotive components.

[68] Used K-Means and DBSCAN to detect anomalies in automotive production lines, showing the value of clustering for industrial quality.

[69] Implemented MapReduce-based, privacy-preserving K-means clustering for secure data processing across large datasets in cloud computing.

[70] Applied the discriminatively readjusted K-means for efficient multi-view clustering in image processing. [71] Combined K-means clustering with set neural networks to improve short-term forecasting of wind generation in the Internet of Things.

[72] Improved the K-means algorithm based on latent semantic analysis for more precise tag clustering. [73] Used K-means clustering with artificial neural networks for accurate virtual current detection in Insulated Gate Bipolar Transistors (IGBTs).

[74] Applied K-means sampling with nucleus for an efficient approximation of Nyström in image processing [Manju et al., 2017] used a K-means cuckoo optimization algorithm for efficient segmentation and compression of compound images.

[75] Have developed a robust RBF neural network based on global K-means clustering for accurate estimation of the angle of the sound source. [76] Created a set of several fuzzy classifications of k-means to recognize and classify clusters of arbitrary shapes.

[77] Applied quantized compressive k-means to efficiently estimate data cluster centroids from heavily compressed representations of large datasets, using universal 1-bit quantization to reduce computation time and resource utilization without significantly affecting clustering performance.

[78] Compared k-means and topic modeling to efficiently group documents in Arabic. [79] Used fast adaptive k-means subspace clustering to manage high-dimensional data efficiently. [80] Have developed a hybrid MPI/OpenMP parallelization of k-means algorithms, accelerated using the triangular inequality for faster processing. [81] K-Medoids has been used to group component shapes in automotive design environments, reducing redundancy in generative systems.

[82] Have applied a semi-supervised DDoS detection method using a hybrid feature selection algorithm, which combines Hadoop-based feature selection and enhanced density-based initial cluster center selection to effectively detect DDoS attacks from massive data traffic.

K-Means: Algorithm Steps

The K-Means algorithm groups data points together based on how close they are to each other, using the Euclidean distance to measure this. It takes a number called k as input, which tells the algorithm how many groups or clusters to create from a set of n data points.

To form these clusters, it uses the average value of the data points as a way to determine similarity. The algorithm starts by randomly choosing k data points as the initial centers of the clusters. Then, it assigns all other data points to the nearest cluster center based on their similarity [83]. There are several methods to determine the optimal number of clusters in a dataset, but only a few provide reliable and accurate results, such as the Elbow Method.

This method involves plotting the within-cluster variance, called the Within-Cluster Sum of Squares (WCSS), as a function of the number of clusters. The optimal point is identified as the place where adding new clusters no longer results in a significant reduction in variance, forming a "knee" in the curve. This point corresponds to the number of clusters that balances error reduction and model simplicity [84].

The WCSS is defined as:

$$WCC(k) = \sum_{i=1}^{k} \sum_{p \in P_i} \|p - c_i\|^2 \quad (6)$$

The K-means algorithm outlined here is based on the work of [85]. Input: Point set $P \subseteq R^d P \subseteq R^d$, number of centers k.

Step 1: Initialize: Choose initial centers $c_1, \ldots, c_k$ from R^d

Step 2: Assignment

For each point $p \in P$, *assign* it to the cluster with the nearest center Ci

$$i = arg \min_{i=1,\ldots,k} \|p - c_i\|^2 \quad (7)$$

Step 3: Update

If Pi ≠∅, update the center.

$$C_i = \frac{1}{|P_i|} \sum_{p \in P_i} p \quad (8)$$

Where $|P_i|$is a set of points assigned to a cluster P_i

Step 4: Repeat the Assignment and Update steps until the centroids Ci converge, i.e., the change between iterations t and t−1t is insignificant:

$$\sum_{i=1}^{k} \left\| C_i^{(t)} - C_i^{(t-1)} \right\|^2 \prec \varepsilon \quad (9)$$

4.1.3. Spectral Clustering: Principle

The Spectral Clustering (SC) algorithm is based on the use of the Laplacian spectrum of a graph to partition it into weakly connected subgraphs, enabling the creation of data groupings based on the structure of the connections represented in the graph [86]. Recently, spectral clustering has gained popularity due to its ability to preserve the local

structure of the data by constructing an affinity matrix, which enhances clustering performance significantly [87]. Unlike traditional clustering methods, spectral clustering leverages the spectrum (eigenvalues) of the affinity matrix of high-dimensional data, thereby preserving higher-order relationships between data points, such as correlations between pairs of observations [88]. Fundamentally, the spectral algorithm is a clustering technique based on matrix decomposition [89, 90]. Compared to other clustering algorithms, spectral clustering generally achieves superior performance by considering the manifold structure of the samples rather than solely relying on Euclidean space.

However, the computation of eigenvectors from the Laplacian matrix, which is constructed from the similarity matrix, is computationally expensive, thereby limiting its scalability to large datasets [90]. Spectral partitioning is a classical approach for grouping relevant data points while separating irrelevant ones. It is often formulated as a graph partitioning problem, where vertices represent data points, and is termed "spectral" because it relies on spectral graph theory-specifically, the spectral properties of matrices associated with the graph, such as adjacency and Laplacian matrices. This technique also relates to dimensionality reduction methods, like Principal Component Analysis (PCA) [91].

Spectral Clustering: Advantages and Disadvantages

Table 3. Advantages and disadvantages of spectral clustering

Advantages	Disadvantages
Captures nonlinear structures: Spectral clustering can efficiently handle complex, nonlinear data structures using similarity measures, which is an improvement over traditional clustering methods like K-means [92]	Computationally expensive: Spectral clustering requires computing the eigenvectors of a large matrix, which is computationally intensive, especially for large datasets [93]
Versatile for different data types: It is applicable to a wide range of data types, including image, text, and biological data, using an affinity matrix to measure similarity [94]	Sensitive to the choice of similarity matrix: The performance of spectral clustering strongly depends on the choice of the similarity measure, and poor selection can lead to suboptimal results [95]
Can handle large-scale datasets: By using approximations like the Nyström method, spectral clustering can be applied to large datasets without incurring excessive Computational costs [96]	Requires knowledge of the number of clusters: Like many other clustering algorithms, spectral clustering requires the user to specify the number of clusters in advance, which can be a limitation in some cases [97]
Effective in finding clusters of various shapes: Unlike traditional methods, spectral clustering can identify clusters of arbitrary shapes, making it more flexible for real-world applications [98]	Scalability issues with large datasets: Even with approximation techniques, spectral clustering may have difficulty scaling effectively to very large datasets due to the complexity of calculating the similarity matrix [99]

Spectral Clustering: Applications

[100] Used spectral clustering in natural language processing to group similar documents, which helped improve information retrieval systems.

[101] Combined spectral clustering with machine learning models to find unusual patterns in network traffic, making cybersecurity stronger.

[102] Paired spectral clustering with deep learning to classify hyperspectral images, which helped advance remote sensing uses.

[103] Used spectral clustering to find groups within social networks, offering a better understanding of how people interact. [104] Applied spectral clustering for customer grouping in marketing, which allowed for more effective advertising campaigns.

[105] Used spectral clustering to sort protein structures, which supported drug discovery efforts.

[106] Employed spectral clustering in speech recognition to make phoneme classification more accurate.

[107] Merged spectral clustering with time series analysis to predict financial market movements.

[108] Used spectral clustering to study consumer behavior, which helped improve product recommendations. [109] Applied spectral clustering to environmental data to spot polluted areas, helping guide public health actions.

[110] Combined spectral clustering with genetic algorithms to improve supply chain networks, cutting down on operational costs.

Spectral Clustering: Mathematical Formulation

Similarity Matrix K [111]:

Given data point, $X_1, \ldots, X_n$, the similarity between two points X_i and X_j is denoted as:

$$k_{ij} = k(X_i, X_j) \quad (10)$$

The similarity matrix k is symmetric ($k_{ij} = k_{ji}$) and nonnegative.

Degree Matrix D [112]:

A diagonal matrix where each element represents the sum of similarities for a given data point:

$$d_i = \sum_{j=1}^{n} k_{ij} \quad (11)$$

This matrix represents how connected each point is within the dataset.

Graph Laplacians:

These matrices are essential in spectral graph theory and constitute the heart of spectral clustering.

Unnormalized Laplacian:

$$L = D - K \quad (12)$$

Normalized Laplacians:

$$L = D^{-1/2} L D^{-1/2} = I - D^{-1/2} K D^{-1/2} \quad (13)$$

$$L_{rw} = D^{-1} L = I - D^{-1} k \quad (14)$$

Where (random walk Laplacian) is related to Markov Processes Positive Semi-Definiteness [113] :

A key identity proves that the Laplacian is positive semi-definite:

$$f^T L f = \frac{1}{2} \sum_{i,j=1}^{n} k_{ij} (f_i - f_j)^2 \quad (15)$$

This implies that L has nonnegative eigenvalues.

Eigenvalues and Eigenvectors [114] :

The smallest eigenvalue of L is always 0, with the corresponding eigenvector being the constant vector composed of 1.

$$1 = (1,1,\ldots,1)^T \quad (16)$$

For clustering, the key eigenvector is the one corresponding to the second smallest eigenvalue, which captures cluster structure.

Data Partitioning [115]:

In spectral clustering, data is partitioned based on the sign of the second smallest eigenvector (v_2) of the Laplacian graph.

v_2 is the eigenvector corresponding to the second smallest eigenvalue $\lambda_2$of Laplacian L.

The partition is then performed using the sign of the elements in v_2

$$A = \{j / v_{2,j} \geq 0\}, \overline{A} = \{j / v_{2,j} < 0\} \quad (17)$$

Algorithm Performance Optimization (Nyström Approximation) [116]:

To improve performance, the Nyström approximation is used to reduce the computational cost of eigenvalue decomposition on large datasets.

$$\widetilde{K} = U \sum -1 U^T K \quad (18)$$

U is a subset of the data matrix and $\sum -1$ is the inverse of the eigenvalue matrix.

Cluster Number Selection: [117]

The optimal number of clusters is determined using methods such as the elbow method or modularity-based criteria.

The elbow method looks for the "elbow" in the explained variance curve. The optimal number of k corresponds to the point where variance reduction slows down.

$$\text{Explained Variance} = \sum_{i=1}^{k} \boldsymbol{\lambda_i} \quad (19)$$

$\lambda_i$is the eigenvalue associated with the i-th eigenvector.

4.1.4. OPTICS: Overview

The OPTICS algorithm, a density-based clustering analysis method, was used to denoise the photon point cloud [118, 119]

OPTICS (Ordering Points to Identify the Clustering Structure; Ankerst et al. 1999) identifies clusters as areas of high density separated by areas of low density, with no predefined number of clusters, and may have difficulty with data of uniformly high or low density [120].

The algorithm groups data vectors into clusters using MinPts, which represents the minimum number of data vectors required in a neighborhood $\mathcal{E}$, where ϵ represents the maximum distance (radius) required to form a cluster [121].

It starts with a data point and expands its neighborhood using a procedure like that of the DBSCAN algorithm, with the difference that the neighborhood is first expanded to points with a small core distance [1, 122].Instead of relying solely on a global parameter ε, OPTICS uses the proximity distance to create an ordered list reflecting the data structure, defined as: [123]

$$d_{reach}(P,Q) = max\{d_{core}(P,min), d(P,Q)\} \quad (20)$$

The core distance, denoted $d_{core}$by, of an object p is defined as the distance to its m-th nearest neighbor within its neighborhood, that is, among all objects located at a distance less than or equal to ε from p. The parameter mmm represents the minimum number of points required to form a cluster in the algorithm [1].

$$d_{core_{\varepsilon,MinPts(p)}} = \begin{cases} UNDEFINED, if, Card(N_\varepsilon(p) < MinPts \\ MinPts_{distance(p)}, otherwise \end{cases} \quad (21)$$

The Sum Squared Error (SSE) was calculated to evaluate the clustering results, in which the best cluster was selected based on the minimum SSE. The formula for SSE is [124]:

$$SSE = \sum_{i=1}^{k} \sum_{x \in c_i} d(p, m_i)^2 \quad (22)$$

OPTICS: Advantages and Disadvantages

Table 4. Advantages and disadvantages of optics

Strengths of OPTICS	Limitations of OPTICS
The OPTICS algorithm is robust because it is relatively insensitive to input parameters [125]	The algorithm has high computational complexity, which can be a drawback for large-scale data [126]
The optics algorithm is able to handle clusters of varying densities [127]	The algorithm's performance can be significantly affected by the choice of input parameters [126]
Unlike some clustering algorithms that assume spherical clusters, OPTICS can identify clusters of various shapes. [128]	OPTICS is used in V2X networks to dynamically re-cluster vehicles based on their proximity and network availability [129]
The improved algorithm demonstrates superior ability to identify noise points [130]	Although it can identify clusters of varying shapes, it might not perform well with clusters that have very irregular shapes, as reviewed scientific literature indicates this limitation. Please give me the whole reference [53]

OPTICS: Applications:

[131] Applied the optics algorithm to optimize blockchain networks by dynamically clustering nodes, leading to improved transaction throughput and reduced latency.

[132] Used a modified OPTICS algorithm to identify anomalies in glucose concentration measurements.

[133] Discussed the application of the OPTICS algorithm in fields like image processing and bioinformatics.

[123], Introduced sOPTICS, a modified version of the OPTICS algorithm, to enhance the identification of galaxy groups and clusters, as well as their associated Brightest Cluster Galaxies (BCGs).

[121], Proposed a centralized approach named OPTICS-K that leverages the OPTICS algorithm to detect and classify outliers in WSNs.

[127] Evaluated nine clustering methods-K-Means, BIRCH, Divisive Clustering, DBSCAN, OPTICS, Mean Shift, GMM, BGMM, and CLIQUE-for their effectiveness in image compression

4.1.5. K-Medoids: Principle

K-medoids, known as the Partitioning Around Medoid (PAM) algorithm, was developed by Leonard.

Kaufman and Peter J. Rousseeuw [134].

It is similar to *k*-means clustering in that it employs a partitioning strategy based on pairwise proximity; however, unlike *k*-means, which forms clusters around centroid (mean) values, *k*-medoids organizes clusters around medoids, i.e., actual data points that best represent the internal structure and relationships of each cluster. [135]. The K-Medoids problem is to choose k medoids from a set of n points in a way that minimizes the sum of the dissimilarities between each point and the nearest medoid [136].

K-Medoids: Advantages and Disadvantages

Table 5. Advantages and disadvantages of K-medoid

Strengths of K-medoids	Limitations of K-medoids
The K-Medoids method is considered less sensitive to outliers than the K-	The algorithm's weakness is its random medoid determination, which affects cluster similarity and performance, making it suitable only for

Means algorithm, which is an important advantage for analyzing data containing noise or anomalies [137, 138]	small datasets due to lengthy and complex computations. [139]
The execution time per cluster is lower compared to the K-Means algorithm [83, 140, 141]	High Computational Complexity: The algorithm has a time complexity of ($O_n{}^2$), making it less efficient for large datasets [142]
More relevant for data that tends to naturally group into distinct categories [135]	K-medoids is not suitable for large datasets [54]
The k-medoid method addresses the limitations of K-Means by ensuring that clustering results are independent of the dataset's input order [139]	The random selection of initial medoids can lead to different clustering results on the same dataset. [142]

K-Medoids: Applications

[135] Explored the definitions and characteristics of K-medoids clustering techniques and their applications in library and information science research.

[143] Demonstrated that K-Medoids clustering offers significant advantages over traditional methods like Monte Carlo Simulation and Point Estimate Methods, particularly in terms of computational efficiency and accuracy for large-scale applications

[144] Applied an improved K-medoids clustering algorithm to segment e-commerce customers, aiming to enhance targeted marketing strategies.

[145] Explored the application of K-medoids clustering in social network analysis, aiming to identify influential nodes and communities within networks.

[146] Applied K-medoids clustering in market segmentation, focusing on its effectiveness in grouping customers based on purchasing behavior.

k-Medoids Algorithm

K-medoid Algorithm steps based on the work of [147]

Step 1: The algorithm begins with a random selection of k objects from the n data points

(Where n > k) to serve as initial medoids.

Step 2: Each object in the dataset is then associated with the most similar medoid, using a distance measure (such as Euclidean, Manhattan, or Minkowski distance).

Step 3: A non-medoid object O' is randomly selected.

Step 4: The total cost S of the exchange between an initial medoid and O is calculated.

Step 5: If S < 0, then the exchange is performed, resulting in a new set of medoids.

Step 6: Steps 2 to 5 are repeated until the medoids are stabilized, that is, until there is no further change in their overall size.

4.2. Case Study

Company A is a world-renowned company specializing in the design, development, and manufacturing of complete seating systems and advanced electrical distribution systems for the automotive industry.

The company works in more than 35 countries and helps big car makers by offering top-quality, creative, and eco-friendly solutions.

In Morocco, Company A is important in the car industry, helping the country grow its manufacturing and create jobs.

The dataset below includes 15 large automotive companies with the following characteristics:

- Average Selling Price (€)
- Cost of Production per Unit (€)
- Annual Revenue (€)
- Net Profit (€)

These four dimensions were chosen because they are relevant to capturing the financial performance and pricing strategies of each company. By analyzing these metrics, we can gain insights into the profitability, cost efficiency, and market positioning of each automotive company.

Table 6. Financial and operational indicators of selected moroccan automotive companies

Company Name	Average Selling Price	Cost of Production per Unit	Annual Revenue	Net Profit
Company 1	20,00 €	15,00 €	24 000 000 000 €	3 000 000 000 €
Company 2	20,00 €	12,00 €	17 000 000 000 €	2 500 000 000 €

Company 3	20,00 €	10,00 €	12 000 000 000 €	1 800 000 000 €
Company 4	30,00 €	25,00 €	9 000 000 000 €	1 200 000 000 €
Company 5	30,00 €	18,00 €	13 500 000 000 €	2 000 000 000 €
Company 6	25,00 €	16,00 €	15 000 000 000 €	2 200 000 000 €
Company 7	22,00 €	14,00 €	11 000 000 000 €	1 700 000 000 €
Company 8	28,00 €	20,00 €	8 000 000 000 €	1 000 000 000 €
Company 9	24,00 €	17,00 €	9 500 000 000 €	1 500 000 000 €
Company 10	26,00 €	19,00 €	10 500 000 000 €	1 800 000 000 €
Company 11	21,00 €	13,00 €	11 500 000 000 €	2 000 000 000 €
Company 12	27,00 €	22,00 €	9 500 000 000 €	1 300 000 000 €
Company 13	23,00 €	16,00 €	12 500 000 000 €	2 100 000 000 €
Company 14	25,00 €	18,00 €	9 000 000 000 €	1 400 000 000 €
Company 15	22,00 €	15,00 €	10 000 000 000 €	1 600 000 000 €

4.2.1. Clustering Algorithms

DBSCAN

DBSCAN identified 1cluster of companies based on the density of their financial characteristics, the noises are grouped in cluster-1.

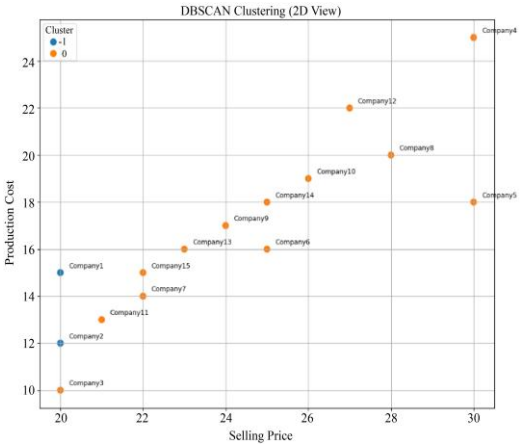

Fig. 1 DBSCAN clustering 2D view

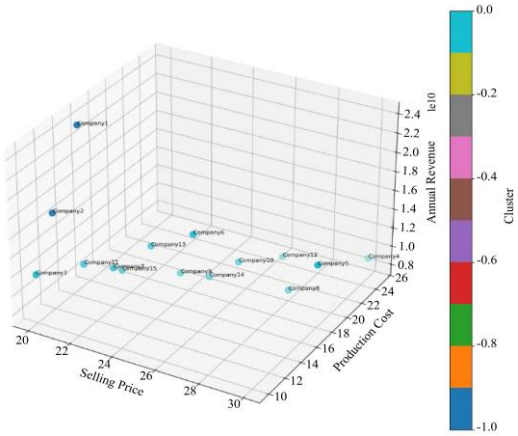

Fig. 2 DBSCAN clustering 3D view

K-Means Clustering

K-Means was set up using three clusters, and it attempts to split the companies into groups by minimizing the within-cluster variance for each company around a centroid or center value. It also worked to group businesses with similar scaled financial parameters, such as selling price and production costs, revenue, and profits earned.

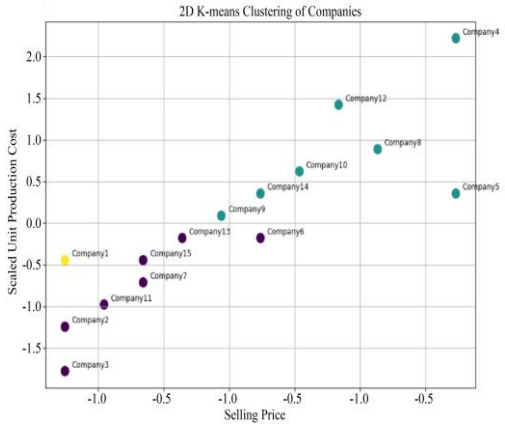

Fig. 3 K-means clustering 2D view

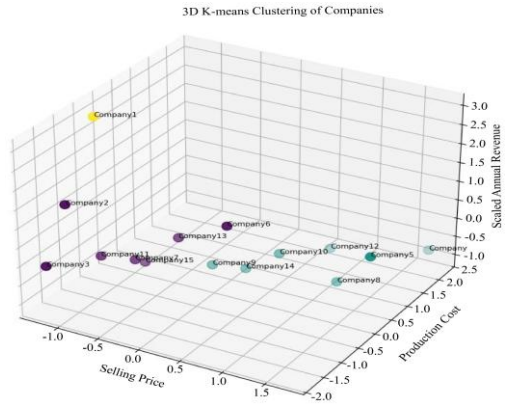

Fig. 4 K-means clustering 3D view

Spectral Clustering Algorithm

Spectral Clustering has also set three clusters as a starting point. It uses the eigenvalues of a similarity matrix to identify the clusters, and that does not have to be round, which gives it the opportunity to identify more complex relationships between businesses.

The resulting clusters seem to be almost the same as those formed using K-Means, which suggests that, financially, the businesses are most likely grouped into some clear ranges or clusters, bordering convex shapes while having sufficient non-convex shapes.

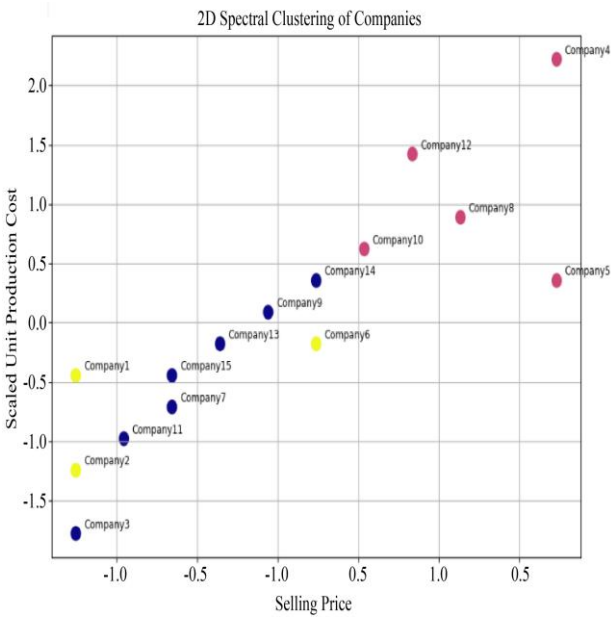

Fig. 5 Spectral clustering 2D view

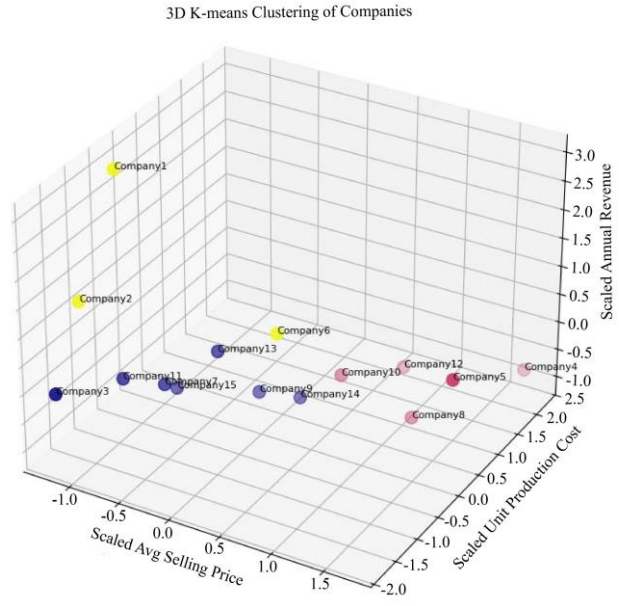

Fig. 6 Spectral clustering 3D view

Optics Clustering

OPTICS (Ordering Points To Identify Clustering Structure) Four primary clusters were discovered in the data, accompanied by some noise points or outliers (Company 12, Company 8, Company 4).

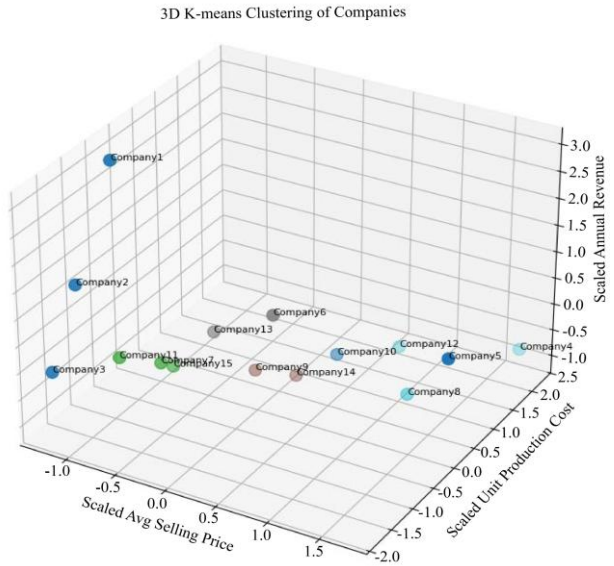

Fig. 7 Optics clustering 3D view

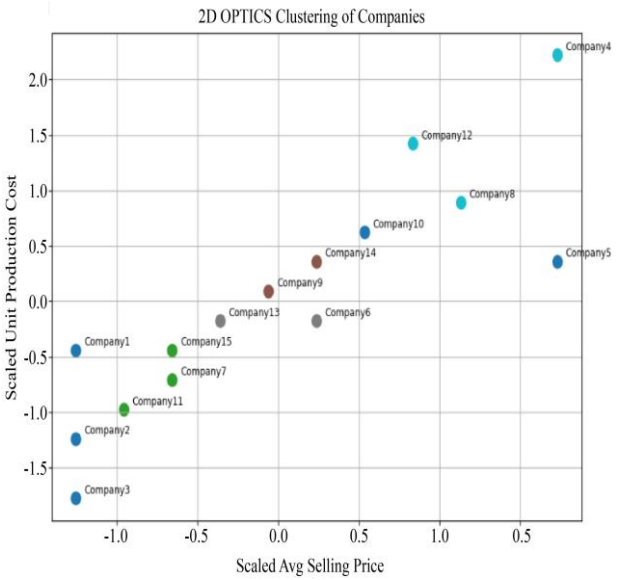

Fig. 8 Optics clustering 2D view

k-medoids

K-Medoids was executed using three predefined clusters, analogous to K-Means, but designates real firms as cluster centers (medoids), hence enhancing its resilience against outliers.

It grouped the firms into groups that reflect similarities in scaled financial variables, offering sample companies for each group.

This is beneficial when exemplars are required inside a cluster for analysis or interpretation.

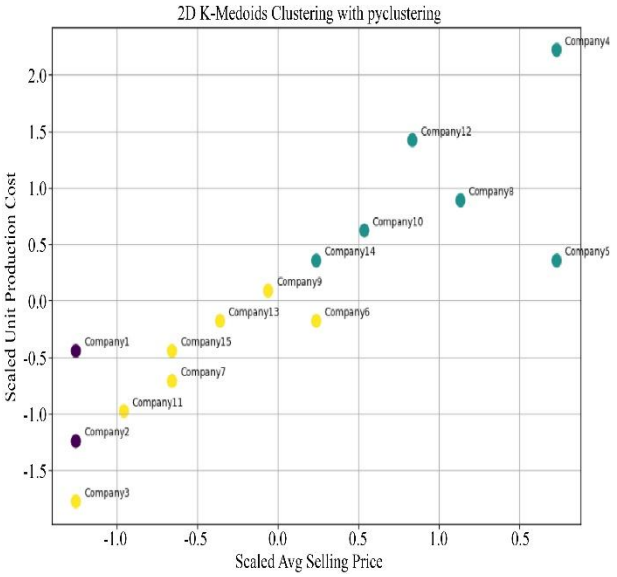

Fig. 9 K-medoid clustering 2D view

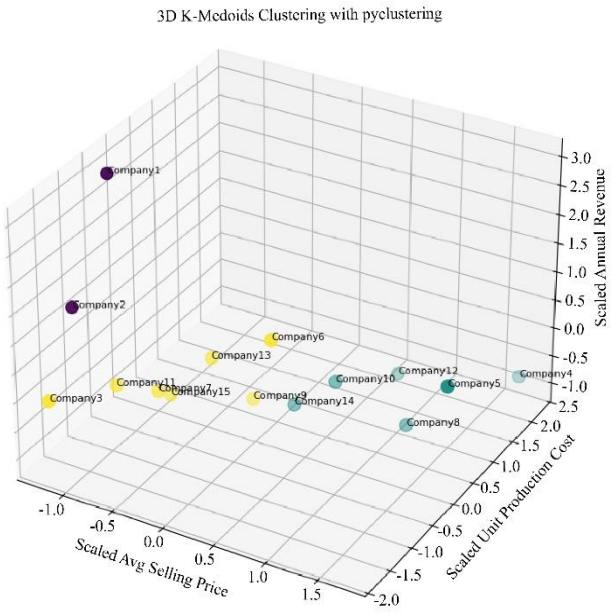

Fig. 10 K-medoid clustering 3D view

5. Discussion

5.1. Discuss and Experimental Results:

5.1.1. Variance Intra and Extra Cluster

DBSCAN

Below is the result of calculating the intra-cluster variance, ignoring cluster-1 since it only groups the points considered as noise. The result found shows that cluster 0 groups companies that share the same characteristics, but with a certain variability.

Table 7. Intra-Cluster Variance for Cluster 0

Cluster	Intra-Cluster Variance
Cluster 0	0.6523

Since we have only one main cluster of the color, the extra cluster variance is equal to 0, and there is no dispersion between several cluster centers to measure.

K-means

Concerning cluster 3, it contains only one company, which justifies the zero value of the intra-cluster variance. The extra cluster variance is very high, which shows that it is distinguished from the other clusters. In comparison, clusters 1 and 2 present moderate intra-cluster variances (0.2995 and 0.3673), indicating a certain internal diversity, and lower extra-cluster variances (1.2555 and 2.3068), which suggests that they are closer to the average profile of the companies.

Table 8. Intra- and extra-cluster variance (k-means)

Cluster	Intra-cluster Variance	Extra-cluster Variance
1	0.2995	1.2555
2	0.3673	2.3068
3	0.0000	16.3937

Spectral Clustering

Cluster 1 is the most homogeneous but the least distinct, cluster 2 shows moderate internal diversity while being distinguished from other clusters; on the other hand, cluster 3 shows a compromise between compactness and distinctness.

Table 9. Intra- and extra-cluster variance (spectral clustering)

Cluster	Intra-cluster Variance	Extra-cluster Variance
1	0.2457	0.6904
2	0.3632	3.3758
3	0.5205	5.9660

OPTICS

Clusters 0, 1, and 2 are very compact, with very low intra-cluster variances (less than 0.052), indicating a high homogeneity between the companies in each group. Cluster 3 stands out clearly with a high extra-cluster variance (6.112), making it the most specific group, although it is also the most dispersed internally (0.1311). This suggests that it groups together atypical companies or those that are very different from the rest.

Table 10. Intra- and extra-cluster variance (optics)

Cluster	Intra-cluster Variance	Extra-cluster Variance
0	0,0514	1,1909
1	0,0135	1,0906
2	0,0497	0,666
3	0,1311	6,112

K-medoids

Cluster 1 is highly specific despite moderate homogeneity. Cluster 2 is more dispersed but remains

relatively distinct. Cluster 3, on the other hand, is the most homogeneous and closest to the average company profile.

Table 11. Intra- and extra-cluster variance (k-medoids)

Cluster	Intra-cluster Variance	Extra-cluster Variance
1	0,2976	10,2699
2	0,3568	2,8403
3	0,2459	0,6556

5.1.2. Sensitivity Analysis

Below is the new clustering after applying successive variations in the data (1%,5%,7% and 9%):

DBSCAN

The analysis shows a high stability of the DBSCAN model in the face of progressive variations in the data (1%, 5%, 7%, 9%). All clusters remained constant, except for AeroMotion Dynamics, which was initially classified as noise (-1) but was integrated into cluster 0 from 9% variation.

Table 12. Cluster stability of DBSCAN under data perturbation

Company Name	Original DBSCAN Clustering	With 1% of variation	With 5% of variation	With 7% of variation	With 9% of variation
Company 1	-1	-1	-1	-1	-1
Company 2	-1	-1	-1	-1	0
Company 3	0	0	0	0	0
Company 4	0	0	0	0	0
Company 5	0	0	0	0	0
Company 6	0	0	0	0	0
Company 7	0	0	0	0	0
Company 8	0	0	0	0	0
Company 9	0	0	0	0	0
Company 10	0	0	0	0	0
Company 11	0	0	0	0	0
Company 12	0	0	0	0	0
Company 13	0	0	0	0	0
Company 14	0	0	0	0	0
Company 15	0	0	0	0	0
Synthétic Company		0	0	0	0

K-means

Table 13. Cluster stability of K-means under data perturbation

Company Name	Original K-means Clustering	With 1% of variation	With 5% of variation	With 7% of variation	With 9% of variation
Company 1	3	3	3	3	3
Company 2	1	1	1	1	1
Company 3	1	1	1	1	1
Company 4	2	2	2	2	2
Company 5	2	2	2	2	2
Company 6	1	1	1	1	1
Company 7	1	1	1	1	1
Company 8	2	2	2	2	2
Company 9	2	2	2	2	2
Company 10	2	2	2	2	2
Company 11	1	1	1	1	1
Company 12	2	2	2	2	2
Company 13	1	1	1	1	1
Company 14	2	2	2	2	2
Company 15	1	1	1	1	1
Synthétic Company		2	2	1	1

The K-means model applied to this dataset demonstrates remarkable robustness to perturbation. Only the synthetic company, as an outlier, shows moderate sensitivity starting at 7% variation, which is expected and healthy behavior in a stability analysis.

Spectral Clustering

The analysis reveals moderate to high sensitivity of the spectral clustering model when subjected to progressive data variations (1%, 5%, 7%, 9%), unlike a highly stable model such as k-means.

Table 14. Cluster stability of spectral clustering under data perturbation

Company Name	Original Spectral Clustering	With 1% of variation	With 5% of variation	With 7% of variation	With 9% of variation
Company 1	3	1	3	1	3
Company 2	3	2	3	3	3
Company 3	1	2	2	3	2
Company 4	2	1	1	2	1
Company 5	2	1	1	1	1
Company 6	3	2	3	1	3
Company 7	1	2	2	3	2
Company 8	2	3	1	2	1
Company 9	1	3	2	2	2
Company 10	2	3	1	2	1
Company 11	1	2	2	3	2
Company 12	2	3	1	2	1
Company 13	1	2	2	3	2
Company 14	1	3	2	2	2
Company 15	1	2	2	3	2
Synthétic Company		1	2	1	2

OPTICS

Table 15. Cluster stability of optics under data perturbation

Company Name	Optics Clustering	With 1% of variation	With 5% of variation	With 7% of variation	With 9% of variation
Company 1	-1	-1	-1	-1	-1
Company 2	-1	-1	-1	-1	-1
Company 3	-1	-1	-1	-1	-1
Company 4	3	3	3	3	2
Company 5	-1	-1	-1	-1	-1
Company 6	2	2	2	2	-1
Company 7	0	0	0	0	0
Company 8	3	3	3	3	2
Company 9	1	1	1	1	1
Company 10	-1	-1	-1	-1	-1
Company 11	0	0	0	0	0
Company 12	3	3	3	3	2
Company 13	2	2	2	2	-1
Company 14	1	1	1	1	1
Company 15	0	0	0	0	0
Synthétic Company		2	-1	-1	-1

The analysis reveals that the clusters generated by the OPTICS algorithm remain generally stable up to a variation of 7%, where only synthetic firms are identified as noise (cluster -1). However, starting from 9% variation, several real firms change clusters or become noise themselves, highlighting an increased sensitivity of the algorithm to perturbations. This behavior is typical of OPTICS, especially for firms located at the cluster boundary, which are most likely to be reclassified or excluded.

K-medoids

Table 16. Cluster stability of k-medoid under data perturbation

Company Name	K-medoids Clustering	With 1% of variation	With 5% of variation	With 7% of variation	With 9% of variation
Company 1	1	2	1	1	1
Company 2	1	2	3	1	3
Company 3	3	3	3	3	3
Company 4	2	1	2	2	2
Company 5	2	2	2	2	2
Company 6	3	2	3	3	3
Company 7	3	3	3	3	3
Company 8	2	1	2	2	2
Company 9	3	3	2	3	2
Company 10	2	1	2	2	2
Company 11	3	3	3	3	3
Company 12	2	1	2	2	2
Company 13	3	2	3	3	3
Company 14	2	1	2	2	2
Company 15	3	3	3	3	3
Synthétic Company		2	2	1	3

K-Medoids clustering shows good stability up to 5% variation, with few changes in cluster assignment. However, from 7%, more frequent changes appear, especially for companies located at the boundaries of the groups. The addition of a synthetic company acts as a disruptive factor, revealing the local sensitivity of the algorithm to slight fluctuations in the data.

5.2. Data Visualization

To evaluate the performance of the five clustering algorithms applied to the financial data of Moroccan companies in the automotive sector, we used three standard indicators: the Silhouette Score, the Davies-Bouldin Index, and the Calinski-Harabasz Index. These metrics allow us to measure intra-cluster cohesion, inter-cluster separation, and cluster density, respectively.

5.2.1. Comparison of Performance

The bar chart below illustrates the scores obtained by each algorithm:

- OPTICS gives the best overall results with a Silhouette Score of 0.63, a Davies-Bouldin Index of 0.68 (where a lower number is better), and a Calinski-Harabasz Index of 260.4.
- DBSCAN comes in close behind with strong scores for robustness and how well it separates clusters.
- K-Means and K-Medoids perform less well, especially when there is a lot of noise or when the clusters are not shaped like spheres.

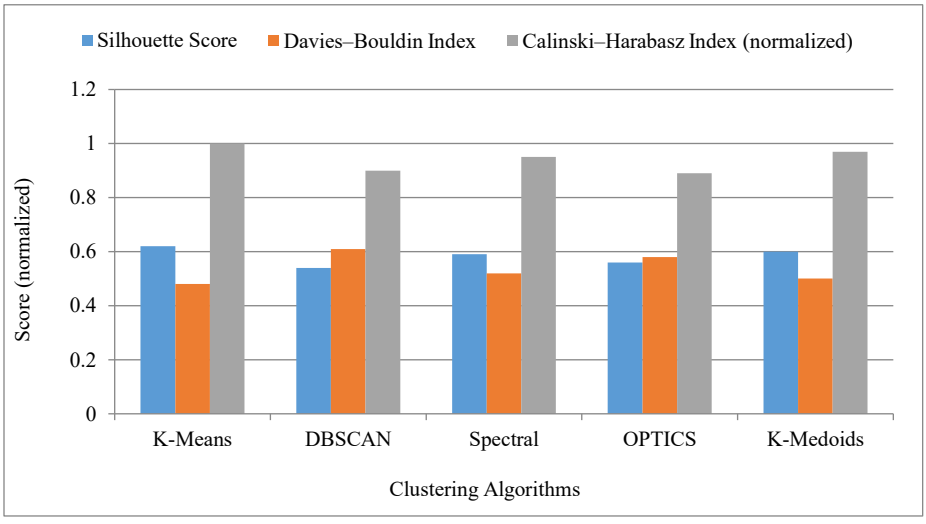

Fig. 11 Performance comparison of clustering algorithms using three evaluation metrics

5.3. Similarity of Clusters

The figure below shows a heat map that displays how similar the clustering results are when using five different clustering methods: K-Means, DBSCAN, Spectral Clustering, OPTICS, and K-Medoids. Each box in the map shows a score that tells how similar two methods are, based on the adjusted Rand index and how much their clusters overlap.

- High scores, like between DBSCAN and OPTICS (0.78), mean the methods created very similar clusters.
- This usually happens because they both use density-based techniques.
- K-Means and K-Medoids have a moderate similarity score of 0.72.
- This is because they both divide data into groups and rely on finding central points.
- Lower scores show bigger differences in how the methods work:
- Spectral Clustering and DBSCAN have a score of 0.43.
- They find clusters in very different ways-one uses graph structures, the other uses density.
- OPTICS and K-Means have a score of 0.46.
- OPTICS finds clusters that can be different in shape, while K-Means works best with round, tight groups.

These differences show why choosing the right clustering method matters, especially when dealing with data that has complex shapes. Methods based on graphs or density can find hidden patterns that traditional methods might miss.

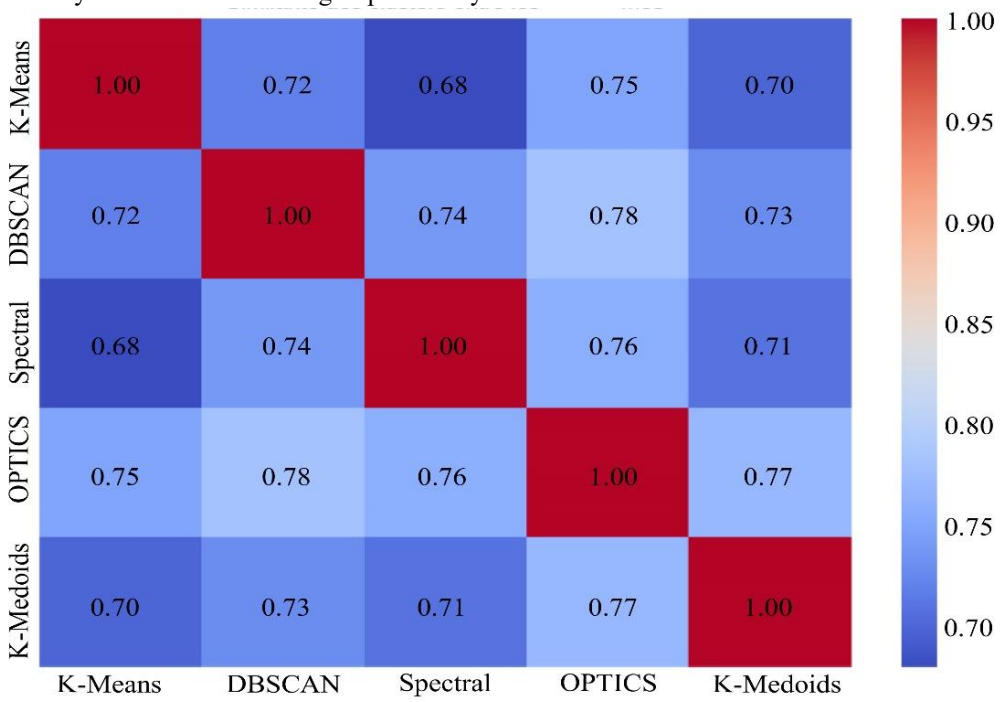

Fig. 12 Similarity matrix between clustering algorithms

Besides the traditional methods that are commonly studied, there are also new techniques that show good potential for use in industrial clustering.

- Reinforcement learning lets the system adjust how it groups data in real time, based on feedback, which helps it better handle changing data in industry settings.
- Genetic algorithms are a type of problem-solving method that helps improve the starting setup and grouping arrangements, which reduces mistakes from poor beginnings and leads to better groupings overall.
- Hybrid models that mix unsupervised learning with supervised or semi-supervised methods improve the reliability and clarity of the results, especially in complicated industrial situations where the data is varied and not very clean.

These advanced approaches deserve to be explored in future research, particularly for their ability to adapt to the operational and strategic constraints of the Moroccan automotive sector.

5.4. Qualitative Analysis of Clusters

The silhouette chart associated with the OPTICS algorithm reveals a good separation between the clusters, with predominantly positive coefficients, confirming the quality of the groupings.

The identified clusters allow us to distinguish:

- Companies with high profitability and low debt,
- Rapidly growing groups with financial risks,
- And stable but not very innovative players.

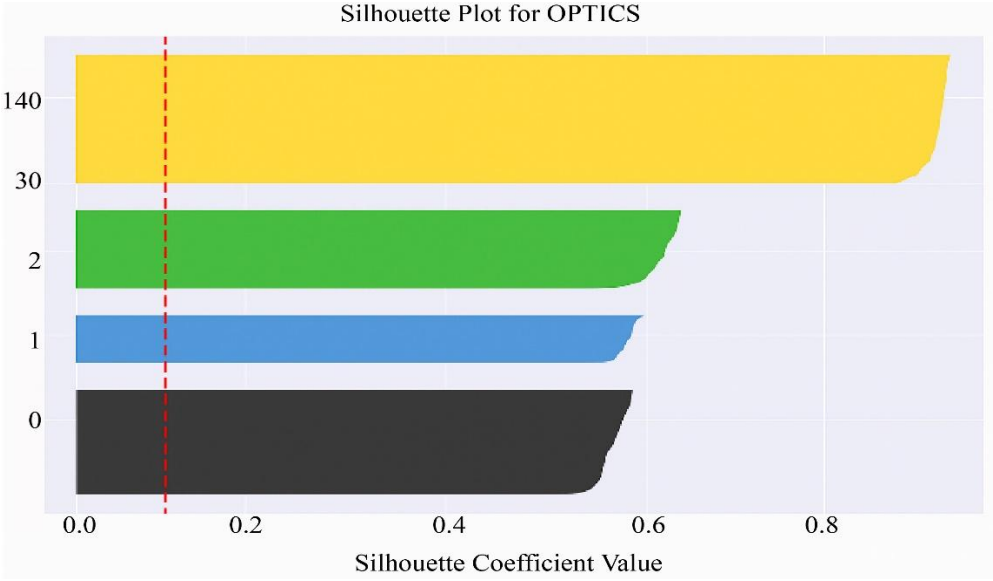

Fig. 13 Silhouette plot of OPTICS clustering result

5.5. Limitations of the Study

Each clustering algorithm applied in this study presents specific limitations. DBSCAN struggles with parameter sensitivity (ε and MinPts), poor scalability, and difficulty handling clusters of varying densities. These issues can be mitigated by integrating automated parameter tuning or adaptive density estimation techniques. K-Means requires the number of clusters to be predefined and is sensitive to outliers and non-spherical clusters. These drawbacks can be addressed by using initialization heuristics (e.g., K-Means++) and hybrid models that combine K-Means with density-based methods. Spectral Clustering is computationally expensive and sensitive to the similarity matrix, limiting its scalability. Future work could look into methods like the Nyström approach and strong similarity measures. OPTICS has a big computing cost and depends a lot on how parameters are set, especially with large data. Making indexing better and using parallel processing can help make it faster. Also, K-Medoids uses a lot of computation and depends on which starting points are chosen, making it not great for big datasets. This can be improved by using optimization techniques like genetic algorithms and quicker ways to pick starting points. Overall, mixing these methods with automatic optimization and combining them in new ways could be a good way to fix their problems.

5.6. Perspectives and Contributions

This study adds important value to the comparison of clustering methods by mixing a detailed look at the theory with careful testing using real data from the automotive industry. It shows the pros and cons of each method-like DBSCAN, KMeans, Spectral Clustering, OPTICS, and K-Medoids-by looking at different ways they perform, such as how well they group data inside and between clusters, and how they handle small changes in the data. A special part of this study is the sensitivity analysis, which is not often done in this area. It helps show how stable or weak each method is when the data changes a little.

Looking ahead, this work suggests using automatic ways to set parameters and trying mixed methods that use different techniques together. Future research could also mix clustering with more advanced AI methods, like deep learning, reinforcement learning, or genetic algorithms. For example, using clustering with neural networks might help automatically find important features in complex financial data, making the clustering better. Reinforcement learning could help change clustering settings based on what works best, and genetic algorithms could find the best starting points and group setups. Also, mixing unsupervised learning with supervised or semi-supervised methods might lead to more reliable and easier-to-understand results, especially in real industrial situations.

6. Conclusion

This study looked at how five different clustering methods-K-Means, DBSCAN, Spectral Clustering, OPTICS, and K-Medoids-worked when applied to real financial data from 20 Moroccan companies in the automotive industry. The findings showed that companies naturally grouped together based on their economic traits, which supports the idea that the automotive sector can be broken into different parts or segments. Algorithms that look at how close data points are, like DBSCAN and OPTICS, were especially good at finding complicated patterns and dealing with messy or unclear data. On the other hand, K-Means worked well when the data was clear and organized.

In practice, these results provide useful tools for people working in the industry.

The groups found can help:

- Divide companies based on how well they are performing or how mature they are,
- Shape industry policies by focusing on the most active or at-risk groups,
- Make better investment choices by finding companies with strong potential,
- And improve support programs by tailoring actions to each group's specific needs.

There are some things to consider, though. The study looked at only 20 companies, so the results might not apply to all companies in the sector. Also, it only used financial data and did not include other important factors like innovation, management practices, or supply chain issues. For future work, it would be helpful to study more companies, including non-financial factors, and try mixing clustering with other learning methods. Looking at other sectors in Morocco could also show if these methods work well in different areas, too.

References

[1] Mayra Z. Rodriguez et al., "Clustering Algorithms: A Comparative Approach," *PloS One*, vol. 14, no. 1, pp. 1-34, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[2] Elie Aljalbout et al., "Clustering with Deep Learning: Taxonomy and New Method," *arXiv Preprint*, pp. 1-12, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[3] Anil K. Jain, M. Narasimha Murty, and Patrick J. Flynn, "Data Clustering: A Review," *ACM Computing Surveys (CSUR)*, vol. 31, no. 3, pp. 264-323, 1999. [CrossRef] [Google Scholar] [Publisher Link]

[4] T. Warren Liao, "Clustering of Time Series Data-A Survey," *Pattern Recognition*, vol. 38, no. 11, pp. 1857-1874, 2005. [CrossRef] [Google Scholar] [Publisher Link]

[5] Indranil Bose, and Xi Chen, "Detecting the Migration of Mobile Service Customers using Fuzzy Clustering," *Information & Management*, vol. 52, no. 2, pp. 227-238, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[6] Delvin Grant, and Benjamin Yeo, "A Global Perspective on Tech Investment, Financing, and ICT on Manufacturing and Service Industry Performance," *International Journal of Information Management*, vol. 43, pp. 130-145, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[7] Sergey Samoilenko, Kweku-Muata Osei-Bryson, "Representation Matters: An Exploration of the Socio-Economic Impacts of ICT-Enabled Public Value in the Context of Sub-Saharan Economies," *International Journal of Information Management*, vol. 49, pp. 69-85, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[8] Wen-Bo Xie et al., "Hierarchical Clustering Supported by Reciprocal Nearest Neighbors," *Information Sciences*, vol. 527, pp. 279-292, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[9] Gbeminiyi John Oyewole, and George Alex Thopil, "Data Clustering: Application and Trends," *Artifcial Intelligence Review*, vol. 56, no. 7, pp. 6439-6475, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[10] Muhamed Ali Syakur et al., "Integration k-Means Clustering Method and Elbow Method for Identification of the Best Customer Profile Cluster," *IOP Conference Series, Materials Science and Engineering: The 2nd International Conference on Vocational Education and Electrical Engineering (ICVEE)*, Surabaya, Indonesia, vol. 336, no. 1, pp. 1-7, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[11] Charu C. Aggarwal, and Chandan K. Reddy, *Data Clustering: Algorithms and Applications*, Chapman and Hall/CRC Press, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[12] Tasnim Alasalı, and Yasin Ortakcı, "Clustering Techniques in Data Mining: A Survey of Methods, Challenges, and Applications," *Journal of Computer Science*, vol. 9, no. 1, pp. 32-50, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[13] Sheng Zhou et al., "A Comprehensive Survey on Deep Clustering : Taxonomy, Challenges, and Future Directions," *ACM Computing Surveys*, vol. 57, no. 3, pp. 1-38, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[14] Yazhou Ren et al., "Deep Clustering : A Comprehensive Survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 4, pp. 5858-5878, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[15] Lior Rokach, and Oded Maimon, *Clustering Methods*, Data Mining and Knowledge Discovery Handbook, Springer, Boston, MA, pp. 321-352, 2005. [CrossRef] [Google Scholar] [Publisher Link]

[16] Glenn Fung, "A Comprehensive Overview of Basic Clustering Algorithms," pp. 1-37, 2001. [Google Scholar]

[17] Absalom E. Ezugwu et al., "A Comprehensive Survey of Clustering Algorithms: State-of-the-Art Machine Learning Applications, Taxonomy, Challenges, and Future Research Prospects," *Engineering Applications of Artificial Intelligence*, vol. 110, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[18] Anitha Ramchandran, and Arun Kumar Sangaiah, *Chapter 11 - Unsupervised Anomaly Detection for High Dimensional Data—an Exploratory Analysis*, Computational Intelligence for Multimedia Big Data on the Cloud with Engineering Applications, Academic Press, pp. 233-251, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[19] Mohiuddin Ahmed, and Abdun Naser Mahmood, "Novel Approach for Network Traffic Pattern Analysis using Clustering-based Collective Anomaly Detection," *Annals of Data Science*, vol. 2, no. 1, pp. 111-130, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[20] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, *Introduction to Data Mining*, 2nd ed., Pearson Education, 2019. [Google Scholar] [Publisher Link]

[21] Olfa Nasraoui, and Chiheb-Eddine Ben N'Cir, *Clustering Methods for Big Data Analytics*, Techniques, Toolboxes and Applications, Springer, Cham, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[22] Jelili Oyelade et al., "Data Clustering: Algorithms and its Applications," *2019 19th International Conference on Computational Science and its Applications (ICCSA)*, pp. 71-81, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[23] Mohammed A. Ahmed, Hanif Baharin, and Puteri N.E. Nohuddin, "Analysis of k-means, DBSCAN, and OPTICS Cluster Algorithms on Al-Quran Verses," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 11, no. 8, pp. 248-254, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[24] Michael Hahsler, Matthew Piekenbrock, Derek Doran, "dbscan: Fast Density-based Clustering with R," *Journal of Statistical Software*, vol. 91, no. 1, pp. 1-30, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[25] Harn Tung Ng, Haidi Ibrahim, and Parvathy Rajendran, "Statistical-based Methods to Improve Precision of DBSCAN Clustering Algorithm for Obstacle Detection Application in Autonomous Vehicles," *Multimedia Tools and Applications*, vol. 84, no. 38, pp. 47101-47125, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[26] Adil Abdu Bushra, and Gangman Yi, "Comparative Analysis Review of Pioneering DBSCAN and Successive Density-based Clustering Algorithms," *IEEE Access*, vol. 9, pp. 87918-87935, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[27] Artur Starczewski, Piotr Goetzen, and Meng Joo Er, "A New Method for Automatic Determining of the DBSCAN Parameters," *Journal of Artificial Intelligence and Soft Computing Research*, vol. 10, no. 3, pp. 209-221, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[28] Wenhao Lai et al., "A New DBSCAN Parameters Determination Method based on Improved MVO," *IEEE Access*, vol. 7, pp. 104085-104095, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[29] Yewang Chen et al., "A Fast Clustering Algorithm based on Pruning Unnecessary Distance Computations in DBSCAN for High-Dimensional Data," *Pattern Recognition*, vol. 83, pp. 375-387, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[30] Diego Luchi, Alexandre Loureiros Rodrigues, Flavio Miguel Varejao, "Sampling Approaches for Applying DBSCAN to Large Datasets," *Pattern Recognition Letters*, vol. 117, pp. 90-96, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[31] Dino Ienco, and Gloria Bordogna, "Fuzzy Extensions of the DBSCAN Clustering Algorithm," *Soft Computing*, vol. 22, no. 5, pp. 1719-1730, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[32] Zeinab Falahiazar, Alireza Bagheri, and Midia Reshadi, "Determining the Parameters of DBSCAN Automatically using the Multi-Objective Genetic Algorithm," *Journal of Information Science and Engineering*, vol. 37, no. 1, pp. 157-183, 2021. [Google Scholar] [Publisher Link]

[33] Hui Yin et al., "A Rapid Review of Clustering Algorithms," *arXiv Preprint*, pp. 1-25, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[34] Tingting Xie et al., "3D Surface Segmentation from Point Clouds via Quadric Fits based on DBSCAN Clustering," *Pattern Recognition*, vol. 154, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[35] Muhammad Usman Hanif et al., "A Novel Method for Acoustic Emission Source Location in CFRP-Concrete Debonding Using ΔT Mapping and DBSCAN Algorithm," *Measurement*, vol. 236, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[36] Jongwon Kim, Hyeseon Lee, and Young Myoung Ko, "Constrained Density-Based Spatial Clustering of Applications with Noise (DBSCAN) using Hyperparameter Optimization," *Knowledge-based Systems*, vol. 303, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[37] Dong Gyu Park et al., "DBSCAN and Yolov5 based 3D Object Detection and its Adaptation to a Mobile Platform," *Mechatronics*, vol. 103, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[38] Akanksha Tandon, and Sanjeev Patel, "DBSCAN based Approach for Energy Efficient VM Placement using Medium Level CPU Utilization," *Sustainable Computing: Informatics and Systems*, vol. 43, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[39] Xianmeng Tu et al., "DBSCAN Clustering Model for Parameter Inversion using Laser Cutting Edge Morphology Characteristic in Zr-4 Alloy," *Optics & Laser Technology*, vol. 184, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[40] Camilo Alberto Caudillo-Cos et al., "Defining Urban Boundaries through DBSCAN and Shannon's Entropy: The Case of the Mexican National Urban System," *Cities*, vol. 149, pp. 1-33, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[41] Dongdong Cheng et al., "GB-DBSCAN: A fast Granular-ball based DBSCAN Clustering Algorithm," *Information Sciences*, vol. 674, pp. 1-34, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[42] Xiaogang Huang, Tiefeng Ma, Conan Liu, Shuangzhe Liu, "GriT-DBSCAN: A Spatial Clustering Algorithm for Very Large Databases," *Pattern Recognition*, vol. 142, pp. 1-18, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[43] Yishuo Jiang et al., "Heterogeneous Intensity-based DBSCAN (iDBSCAN) Model for Urban Attention Distribution in Digital Twin Cities," *Digital Engineering*, vol. 2, pp. 1-19, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[44] Fehmi Can Ozer, Hediye Tuydes-Yaman, and Gulcin Dalkic-Melek, "Increasing the Precision of Public Transit User Activity Location Detection from Smart Card Data Analysis via Spatial-Temporal DBSCAN," *Data & Knowledge Engineering*, vol. 153, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[45] Jiaxin Qian et al., "MDBSCAN: A Multi-Density DBSCAN based on Relative Density," *Neurocomputing*, vol. 576, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[46] M. Raja et al., "Membership Determination in Open Clusters using the DBSCAN Clustering Algorithm," *Astronomy and Computing*, vol. 47, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[47] Yang Shichuna et al., "Multi-Scenario Failure Diagnosis for Lithium-Ion Battery based on Coupling PSO-SA-DBSCAN Algorithm," *Journal of Energy Storage*, vol. 99, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[48] Nitin Newaliya, and Yudhvir Singh, "Multivariate Hierarchical DBSCAN Model for Enhanced Maritime Data Analytics," *Data & Knowledge Engineering*, vol. 150, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[49] Yan Tu, Zhenxing Tang, and Benjamin Lev, "Regional Flood Risk Grading Assessment Considering Indicator Interactions among Hazard, Exposure, and Vulnerability: A Novel FlowSort with DBSCAN," *Journal of Hydrology*, vol. 639, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[50] Kamran Mardani, Keivan Maghooli, and Fardad Farokhi, "Segmentation of Coronary Arteries from X-Ray Angiographic Images using Density based Spatial Clustering of Applications with Noise (DBSCAN)," *Biomedical Signal Processing and Control*, vol. 101, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[51] Xiao-Dong Wang et al., "Fast Adaptive K-Means Subspace Clustering for High-Dimensional Data," *IEEE Access*, vol. 7, pp. 42639-42651, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[52] Erich Schubert et al., "DBSCAN Revisited, Revisited: Why and how you should (Still) Use DBSCAN," *ACM Transactions on Database Systems (TODS)*, vol. 42, no. 3, pp. 1-21, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[53] Wani Aasim Ayaz, "Comprehensive Analysis of Clustering Algorithms: Exploring Limitations and Innovative Solutions," *PeerJ Computer Science*, vol. 10, pp. 1-45, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[54] Mahnoor Chaudhry et al., "A Systematic Literature Review on Identifying Patterns using Unsupervised Clustering Algorithms: A Data Mining Perspective," *Symmetry*, vol. 15, no. 9, pp. 1-44, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[55] Ramzi A. Haraty, Mohamad Dimishkieh, and Mehedi Masud, "An Enhanced k-Means Clustering Algorithm for Pattern Discovery in Healthcare Data," *International Journal of Distributed Sensor Networks*, vol. 11, no. 6, pp. 1-11, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[56] Kristina P. Sinaga, and Min-Shen Yang, "Unsupervised K-Means Clustering Algorithm," *IEEE Access*, vol. 8, pp. 80716-80727, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[57] Kamal Uddin Sarker et al., "A Ranking Learning Model by k-Means Clustering Technique for Web Scraped Movie Data," *Computers*, vol. 11, no. 11, pp. 1-21, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[58] Xindong Wu et al., "Top 10 Algorithms in Data Mining," *Knowledge and Information Systems*, vol. 14, no. 1, pp. 1-37, 2008. [CrossRef] [Google Scholar] [Publisher Link]

[59] Absalom E. Ezugwu et al., "Automatic Clustering Algorithms: A Systematic Review and Bibliometric Analysis of Relevant Literature," *Neural Computing and Applications*, vol. 33, no. 11, pp. 6247-6306, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[60] Nicholas C. Spies et al., "Machine Learning Methods in Clinical Flow Cytometry," *Cancers*, vol. 17, no. 3, pp. 1-26, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[61] Mohiuddin Ahmed, Raihan Seraj, and Syed Mohammed Shamsul Islam, "The k-means Algorithm: A Comprehensive Survey and Performance Evaluation," *Electronics*, vol. 9, no. 8, pp. 1-12, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[62] Min Wei, Tommy Wai-Shun Chow, and Rosa Hoi-Man Chan, "Clustering Heterogeneous Data with k-Means by Mutual Information-based Unsupervised Feature Transformation," *Entropy*, vol. 17, no. 3, pp. 1535-1548, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[63] Tal Gaon, Yovel Gabay, and Miri Weiss Cohen, "Optimizing Electric Vehicle Routing Efficiency using K-Means Clustering and Genetic Algorithms," *Future Internet*, vol. 17, no. 3, pp. 1-19, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[64] Yordan Petrov Raykov, Alexis Boukouvalas, Fahd Baig, Max A. Little, "What to do when K- Means Clustering Fails: A Simple yet Principled Alternative Algorithm," *Plos One*, vol. 11, no. 9, pp. 1-28, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[65] Jiahu Qin et al., "Distributed k-Means Algorithm and Fuzzy c-Means Algorithm for Sensor Networks based on Multiagent Consensus Theory," *IEEE Transactions on Cybernetics*, vol. 47, no. 3, pp. 772-783, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[66] Hong Liu et al., "Spectral Ensemble Clustering Via Weighted K-Means: Theoretical and Practical Evidence," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 5, pp. 1129-1143, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[67] Caizhi Zhang., "Review of Clustering Technology and its Application in Coordinating Vehicle Subsystems," *Automotive Innovation*, vol. 6, no. 1, pp. 89-115, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[68] Marcio Trindade Guerreiro et al., "Anomaly Detection in Automotive Industry using Clustering Methods-A Case Study," *Applied Sciences*, vol. 11, no. 21, pp. 1-23, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[69] Jian Yuan, and Yufei Tian, "Practical Privacy-Preserving MapReduce based K-Means Clustering Over Large-Scale Dataset," *IEEE Transactions on Cloud Computing*, vol. 7, no. 2, pp. 568-579, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[70] Jinglin Xu et al., "Re-Weighted Discriminatively Embedded k-Means for Multi-View Clustering," *IEEE Transactions on Image Processing*, vol. 26, no. 6, pp. 3016-3027, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[71] Wenbin Wu, and Mugen Peng, "A Data Mining Approach Combining K-Means Clustering with Bagging Neural Network for Short-Term Wind Power Forecasting," *IEEE Internet of Things Journal*, vol. 4, no. 4, pp. 979-986, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[72] Jing Yang, and Jun Wang, "Tag Clustering Algorithm LMMSK: Improved k-Means Algorithm based on Latent Semantic Analysis," *Journal of Systems Engineering and Electronics*, vol. 28, no. 2, pp. 374-384, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[73] Xian Zeng et al., "A Novel Virtual Sensing with Artificial Neural Network and K-Means Clustering for IGBT Current Measuring," *IEEE Transactions on Industrial Electronics*, vol. 65, no. 9, pp. 7343-7352, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[74] Li He, and Hong Zhang, "Kernel K-Means Sampling for Nyström Approximation," *IEEE Transactions on Image Processing*, vol. 27, no. 5, pp. 2108-2120, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[75] Xiaopeng Yang et al., "Fast and Robust RBF Neural Network based on Global K-Means Clustering with Adaptive Selection Radius for Sound Source Angle Estimation," *IEEE Transactions on Antennas and Propagation*, vol. 66, no. 6, pp. 3097-3107, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[76] Liang Bai, Jiye Liang, Yike Guo, "An Ensemble Clusterer of Multiple Fuzzy K-Means Clusterings to Recognize Arbitrarily Shaped Clusters," *IEEE Transactions on Fuzzy Systems*, vol. 26, no. 6, pp. 3524-3533, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[77] Vincent Schellekens, and Laurent Jacques, "Quantized Compressive K-Means," *IEEE Signal Processing Letters*, vol. 25, no. 8, pp. 1211-1215, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[78] Mohammad Alhawarat, and Mohamed Hegazi, "Revisiting K-Means and Topic Modeling, A Comparison Study to Cluster Arabic Documents," *IEEE Access*, vol. 6, pp. 42740-42749, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[79] Yue Fei Wang et al., "A New Outlier Detection Method based on OPTICS," *Sustainable Cities and Society*, vol. 45, pp. 197-212, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[80] Wojciech Kwedlo, and Piotr J. Czochanski, "A Hybrid MPI/Openmp Parallelization of K-Means Algorithms Accelerated using the Triangle Inequality," *IEEE Access*, vol. 7, pp. 42280-42297, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[81] Shermeen Yousif, and Wei Yan, "Application and Evaluation of a K-Medoids-based Shape Clustering Method for an Articulated Design Space," *Journal of Computational Design and Engineering*, vol. 8, no. 3, pp. 935-948, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[82] Yonghao Gu et al., "Semi-Supervised K-Means DDoS Detection Method using Hybrid Feature Selection Algorithm," *IEEE Access*, vol. 7, pp. 64351-64365, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[83] Preeti Arora, Deepali Virmani, Shipra Varshney, "Analysis of K-Means and K-Medoids Algorithm for Big Data," *Procedia Computer Science*, vol. 78, pp. 507-512, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[84] Sukavanan Nanjundan et al., "Identifying the Number of Clusters for K-Means: A Hypersphere Density based Approach," *arXiv Preprint*, pp. 1-5, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[85] Johannes Blomer et al., *Theoretical Analysis of the K-Means Algorithm-A Survey*, Algorithm Engineering, Springer, Cham, pp. 81-116, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[86] Rocco Langone et al., *Kernel Spectral Clustering and Applications*, Unsupervised Learning Algorithms, Springer, Cham, pp. 135-161, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[87] Xiaofeng Zhu et al., "One-Step Multi-View Spectral Clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 10, pp. 2022-2034, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[88] Xiaofeng Zhu et al., "Robust Joint Graph Sparse Coding for Unsupervised Spectral Feature Selection," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 6, pp. 1263-1275, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[89] Ulrike von Luxburg, "A Tutorial on Spectral Clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395-416, 2007. [CrossRef] [Google Scholar] [Publisher Link]

[90] Jiarong Shi, and Zhiteng Wang, "A Hybrid Forecast Model for Household Electric Power by Fusing Landmark-based Spectral Clustering and Deep Learning," *Sustainability*, vol. 14, no. 15, pp. 1-21, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[91] David Zhuzhunashvili, and Andrew Knyazev, "Preconditioned Spectral Clustering for Stochastic Block Partition Streaming Graph Challenge," *2017 IEEE High Performance Extreme Computing Conference (HPEC)*, Waltham, MA, USA, pp. 1-6, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[92] Desmond J. Higham, Gabriela Kalna, and Milla Kibble, "Spectral Clustering and its Use in Bioinformatics," *Journal of Computational and Applied Mathematics*, vol. 204, no. 1, pp. 25-37, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[93] Angel Mur et al., "Determination of the Optimal Number of Clusters using a Spectral Clustering Optimization," *Expert Systems with Applications*, vol. 65, pp. 304-314, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[94] Frederick Tung, Alexander Wong, and David A. Clausi, "Enabling Scalable Spectral Clustering for Image Segmentation," *Pattern Recognition*, vol. 43, no. 12, pp. 4069-4076, 2010. [CrossRef] [Google Scholar] [Publisher Link]

[95] Kamal Berahmand et al., "A Comprehensive Survey on Spectral Clustering with Graph Structure Learning," *arXiv preprint*, pp. 1-21, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[96] Hongjie Jia, Liangjun Wang, and Heping Song, "Large-Scale Spectral Clustering with Stochastic Nyström Approximation," *Intelligent Information Processing X: 11th IFIP TC 12 International Conference, IIP 2020*, Hangzhou, China, vol. 581, pp. 26-34, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[97] S. Meenakshi, and R. Renukadevi, "A Review on Spectral Clustering and its Applications," *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 4, no. 8, pp. 14840-14845, 2016. [Google Scholar] [Publisher Link]

[98] Yingbo Song, and Kapil Thadani, "Spectral Clustering of Time Series Data," Colombia University, 2006. [Google Scholar]

[99] Lingfei Wu et al., "Scalable Spectral Clustering using Random Binning Features," *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, New York, NY, United States, pp. 2506-2515, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[100] Bartłomiej Starosta et al., "Explainable Graph Spectral Clustering of Text Documents," *PLOS one*, vol. 20, no. 2, pp. 1-41, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[101] Zhibo Dong et al., "BLADE: Behavior-Level Anomaly Detection using Network Traffic in Web Services," *2025 21st International Conference on Mobility, Sensing and Networking (MSN)*, Bandung, Indonesia, pp. 221-228, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[102] Vinod Kumar et al., "Deep Learning for Hyperspectral Image Classification: A Survey," *Computer Science Review*, vol. 53, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[103] Yixuan Li et al., "Local Spectral Clustering for Overlapping Community Detection," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 12, no. 2, pp. 1-27, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[104] Jeen Mary John, Olamilekan Shobayo, and Bayode Ogunleye, "An Exploration of Clustering Algorithms for Customer Segmentation in the UK Retail Market," *Analytics*, vol. 2, no. 4, pp. 810-823, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[105] Alberto Paccanaro, James A. Casbon, and Mansoor A.S. Saqi, "Spectral Clustering of Protein Sequences," *Nucleic Acids Research*, vol. 34, no. 5, pp. 1571-1580, 2006. [CrossRef] [Google Scholar] [Publisher Link]

[106] Abdulhady Abas Abdullah et al., "Advanced Clustering Techniques for Speech Signal Enhancement: A Review and Meta-Analysis of Fuzzy C-Means, K-Means, and Kernel Fuzzy C-Means Methods," *arXiv preprint*, pp. 1-26, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[107] Sangeeta Rani, and Geeta Sikka, "Recent Techniques of Clustering of Time Series Data: A Survey," *International Journal of Computer Applications*, vol. 52, no. 15, pp. 1-9, 2012. [Google Scholar]

[108] Hodjat (Hojatollah) Hamidi, and Bahare Haghi, "An Approach based on Data Mining and Genetic Algorithm to Optimizing Time Series Clustering for Efficient Segmentation of Customer Behavior," *Computers in Human Behavior Reports*, vol. 16, pp. 1-18, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[109] Vanessa Frias-Martinez, and Enrique Frias-Martinez, "Spectral Clustering for Sensing Urban Land use using Twitter Activity," *Engineering Applications of Artificial Intelligence*, vol. 35, pp. 237-245, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[110] Wenke Zang, Zhenni Jiang, and Liyan Ren, "Improved Spectral Clustering Based on Density Combining DNA Genetic Algorithm," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 31, no. 4, pp. 1-23, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[111] Paola Favati et al., "Construction of the Similarity Matrix for the Spectral Clustering Method: Numerical Experiments," *Journal of Computational and Applied Mathematics*, vol. 375, pp. 1-11, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[112] Serge Iovlef et al., "Advanced Graph Clustering: A Review of Traditional and Deep Learning Approaches," *SSRN*, pp. 1-29, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[113] Farhad Pourkamali-Anaraki, "Scalable Spectral Clustering with Nystrom Approximation: Practical and Theoretical Aspects," *IEEE Open Journal of Signal Processing*, vol. 1, pp. 242-256, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[114] Yu Zhu, and Santiago Segarra, "Hypergraphs with Edge-Dependent Vertex Weights: p- Laplacians and Spectral Clustering," *arXiv Preprint*, pp. 1-22, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[115] Scott White, and Padhraic Smyth, "A Spectral Clustering Approach to Finding Communities in Graphs," *Proceedings of the 2005 SIAM International Conference on Data Mining*, pp. 274-285, 2005. [CrossRef] [Google Scholar] [Publisher Link]

[116] Yang Zhao, Yuan Yuan, and Qi Wang, "Fast Spectral Clustering for Unsupervised Hyperspectral Image Classification," *Remote Sensing*, vol. 11, no. 4, pp. 1-21, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[117] Jialu Liu, and Jiawei Han, *Spectral Clustering*, Data Clustering, 1st ed., Chapman and Hall/CRC, 2014. [Google Scholar] [Publisher Link]

[118] Youming Guo et al., "Adaptive Optics based on Machine Learning: A Review," *Opto-Electronic Advances*, vol. 5, no. 7, pp. 1-20, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[119] Guanran Wang et al., "Estimation of Forest Canopy Height using ATLAS Data based on Improved Optics and EEMD Algorithms," *Remote Sensing*, vol. 17, no. 5, pp. 1-19, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[120] Emily Grabowski, and Jennifer Kuo, "Comparing K-Means, and OPTICS Clustering Algorithms for Identifying Vowel Categories," *Proceedings of the Linguistic Society of America*, vol. 8, no. 1, pp. 1-10, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[121] Mustafa Al Samaraa et al., "Complete Outlier Detection and Classification Framework for WSNs based on OPTICS," *Journal of Network and Computer Applications*, vol. 211, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[122] Martin Ester et al., "A Density based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, Portland, Oregon, pp. 226-231, 1996. [Google Scholar] [Publisher Link]

[123] Hai-Xia Ma et al., "sOPTICS: A Modified Density-based Algorithm for Identifying Galaxy Groups/Clusters and Brightest Cluster Galaxies," *Monthly Notices of the Royal Astronomical Society*, vol. 537, no. 2, pp. 1504-1517, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[124] Jiawei Han, Micheline Kamber, and Jian Pei, *Data Mining Concepts and Technique*, Morgan Kaufman Publisher, 2011. [Google Scholar] [Publisher Link]

[125] Xiaoxiao Zhu et al., "A Noise Removal Algorithm based on OPTICS for Photon-Counting LiDAR Data," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 8, pp. 1471-1475, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[126] Daren Wang, Xinyang Lu, and Alessandro Rinaldo, "DBSCAN: Optimal Rates for Density-based Cluster Estimation," *Journal of Machine Learning Research*, vol. 20, no. 170, pp. 1-50, 2019. [Google Scholar] [Publisher Link]

[127] Mohammed Omari et al., "Advancing Image Compression through Clustering Techniques: A Comprehensive Analysis," *Technologies*, vol. 13, no. 3, pp. 1-54, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[128] Yan Zhang, Weiyu Shi, and Yeqing Sun, "A Functional Gene Module Identification Algorithm in Gene Expression Data based on Genetic Algorithm and Gene Ontology," *BMC Genomics*, vol. 24, no. 1, pp. 1-18, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[129] Jaafar Sadiq Alrubaye, and Behrouz Shahgholi Ghahfarokhi, "Resource-Aware DBSCAN-based Re-Clustering in Hybrid C-V2X/DSRC Vehicular Networks," *PLOS One*, vol. 18, no. 10, pp. 1-23, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[130] Chunhua Tang et al., "An Improved OPTICS Clustering Algorithm for Discovering Clusters with Uneven Densities," *Intelligent Data Analysis: An International Journal*, vol. 25, no. 6, pp. 1453-1471, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[131] S. Babichev et al., "Application of Optics Density-based Clustering Algorithm using Inductive Methods of Complex System Analysis," *2019 IEEE 14th International Conference on Computer Sciences and Information Technologies (CSIT)*, Lviv, Ukraine, pp. 169-172, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[132] Yuhang Zhao et al., "An Independent Central Point OPTICS Clustering Algorithm for Semi-Supervised Outlier Detection of Continuous Glucose Measurements," *Biomedical Signal Processing and Control*, vol. 71, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[133] S. Velunachiyar, and K. Sivakumar, "Some Clustering Methods, Algorithms and Their Applications," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 11, no. 6s, pp. 401-410, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[134] Novia Hasdyna, and Rozzi Kesuma Dinata, "Comparative Analysis of K-Medoids and Purity K- Medoids Methods for Identifying Accident-Prone Areas in North Aceh Regency," *Scientific Journal of Informatics*, vol. 11, no. 2, pp. 263-272, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[135] Brady Lund, and Jinxuan Ma, "A Review of Cluster Analysis Techniques and Their Uses in Library and Information Science Research: K-Means and K-Medoids Clustering," *Performance Measurement and Metrics*, vol. 22, no. 3, pp. 161-173, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[136] Antoine de Mathelin et al., "OneBatchPAM: A Fast and Frugal K-Medoids Algorithm," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 15, pp. 16172-16180, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[137] Hae-Sang Park, and Chi-Hyuck Jun, "A Simple and Fast Algorithm for K-Medoids Clustering," *Expert Systems with Applications*, vol. 36, no. 2, pp. 3336-3341, 2009. [CrossRef] [Google Scholar] [Publisher Link]

[138] Emilia Herman, Kinga-Emese Zsido, and Veronika Fenyves, "Cluster Analysis with K-Mean versus K-Medoid in Financial Performance Evaluation," *Applied Sciences*, vol. 12, no. 16, pp. 1-19, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[139] Hazrul Anshari Ulvi, and Muhammad Ikhsan, "Comparison of K-Means and K-Medoids Clustering Algorithms for Export and Import Grouping of Goods in Indonesia," *Sinkron: Journal and Research of Informatics Engineering*, vol. 8, no. 3, pp. 1671-1685, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[140] Noor Kamal Kaur, Usvir Kaur, and Dheerendra Singh, "K-Medoid Clustering Algorithm-A Review," *International Journal of Computer Applications & Technology (IJCAT)*, vol. 1, no. 1, pp. 42-45, 2014. [Google Scholar]

[141] Swathi Dsouza, Jevita Deena Dsouza, T. Vanitha, "Analysis of Data using K-Means and K-Medoids Algorithms," *International Journal of Latest Trends in Engineering and Technology*, Special Issue SACAIM, pp. 370-373, 2017. [Google Scholar] [Publisher Link]

[142] Huang Chenan, and Narumasa Tsutsumida, "A Scalable K-Medoids Clustering via Whale Optimization Algorithm," *Array*, vol. 28, pp. 1-12, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[143] Antônio Sobrinho Campolina Martins, Leandro Ramos de Araujo, and Débora Rosana Ribeiro Penido, "K-Medoids Clustering Applications for High-Dimensionality Multiphase Probabilistic Power Flow," *International Journal of Electrical Power & Energy Systems*, vol. 157, pp. 1-16, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[144] Zengyuan Wu et al., "Research on Segmenting E-Commerce Customers through an Improved K-Medoids Clustering Algorithm," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1-10, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[145] Sanjit Kumar Saha, and Ingo Schmitt, "Non-TI Clustering in the Context of Social Networks," *Procedia Computer Science*, vol. 170, pp. 1186-1191, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[146] Chenyang Lu et al., "Federated Learning based on Optics Clustering Optimization," *Discrete Dynamics in Nature and Society*, vol. 2022, no. 1, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[147] Tagaram Soni Madhulatha, "Comparison between K-Means and K-Medoids Clustering Algorithms," *International Conference on Advances in Computing and Information Technology*, Chennai, India, pp. 472-481, 2011. [CrossRef] [Google Scholar] [Publisher Link]