Contents lists available at CIOTR
Journal of Smart Computing
journal homepage: https://ciotr.com/jsc

Volume 01 · Issue 01 · 2025 · pp. 48–62 ISSN 2979-8752

REVIEW ARTICLE

Clustering Algorithms in Industry: A Comprehensive Review of Applications, Methods, and Emerging Trends

Eleanor Hawking[1,*], **Aleksandra Nowak**[2]

[1] *Institute for Environmental Science, University of Bristol, UK; Email: e.hawking@bristol.ac.uk*

[2] *Faculty of Physics, Center for Quantum Optical Technologies, Institute of Experimental Physics, University of Warsaw, Poland; Email: a.nowak@uw.edu.pl*

**E-mail (corresponding author): e.hawking@bristol.ac.uk*

ARTICLE INFO

Article history
Received 2025-10-26
Revised 2025-12-14
Accepted 2025-12-17

Keywords
Clustering algorithms
Industry 4.0
Predictive maintenance
Fault detection
Smart grid
Unsupervised learning

ABSTRACT

Clustering algorithms have become indispensable tools in modern industrial data analytics, enabling organizations to extract actionable knowledge from large, heterogeneous datasets without prior class labels. This review paper systematically examines the application of clustering techniques across a diverse spectrum of industry domains, including manufacturing, energy, healthcare finance, and smart infrastructure. Five principal families of clustering algorithms are analyzed: partitional (K-means and its variants), hierarchical, density-based (DBSCAN), model-based (fuzzy C-means and Gaussian mixture models), and graph or spectral approaches. For each family, the theoretical foundations, algorithmic strengths, and documented industrial deployments are discussed. Key application areas surveyed include predictive maintenance, fault detection and diagnosis, customer segmentation, smart grid load profiling, IoT-enabled anomaly detection, and energy consumption optimization. The review further identifies evaluation metrics, open challenges such as scalability and parameter sensitivity, and promising directions, including deep clustering and hybrid meta-heuristic approaches. The findings indicate that no single algorithm universally dominates; instead, successful industrial implementations depend on careful algorithm selection guided by data characteristics, domain knowledge, and computational constraints.

1. Introduction

The rapid proliferation of digital sensors, industrial IoT devices, and cloud-based data platforms has transformed the landscape of modern manufacturing and infrastructure management. By 2023, the global volume of industrial data generated daily exceeded 2.5 quintillion bytes, and this figure is projected to grow at a compound annual rate of more than 23% through the end of the decade (International Data Corporation, 2023). Within this data deluge, a significant proportion arrives without pre-assigned class labels, rendering supervised learning impractical as a first-line analytical tool. Clustering, a branch of unsupervised machine

learning, addresses this challenge directly by partitioning data objects into homogeneous groups based solely on intrinsic feature similarity, without requiring labeled examples.

Clustering has a rich theoretical history stretching back to foundational work in vector quantization and statistical classification. However, its practical relevance to industry has intensified markedly in the era of Industry 4.0, where cyber-physical systems generate continuous, high-dimensional sensor streams that must be interpreted in near real-time. Reports from the McKinsey Global Institute (2022) indicate that organizations deploying advanced analytics, with clustering as a substantial component, achieve 15–25% reductions in unplanned downtime and up to 20% improvements in throughput. In the energy sector, clustering-based load profiling has been shown to reduce grid imbalance costs by up to 18% when integrated into demand response programs (Jiang et al., 2021). Similarly, in financial services, fraud detection systems leveraging density-based clustering have reported recall rates approaching 99.5% on benchmark datasets.

Despite this documented impact, the industrial practitioner faces a bewildering choice: the literature catalogs more than 100 distinct clustering algorithms, each with unique assumptions, hyperparameters, and computational profiles. Comprehensive surveys by Ezugwu et al. (2022) and Ikotun et al. (2023) provide valuable taxonomic overviews, yet the translation of algorithmic knowledge into domain-specific guidance remains fragmented. Industry-focused treatments, such as the survey by Benabdellah et al. (2019) and the Industry 4.0-oriented study by Molinié et al. (2022), address narrower sub-domains without providing a unified cross-sector perspective.

This paper bridges that gap by providing a structured, narrative review of clustering algorithms and their industrial applications, drawing on fifteen verified, peer-reviewed sources spanning the period 2014–2024. The contributions of this review are threefold. First, it presents a unified taxonomy that maps algorithm families to industry application domains. Second, it synthesizes empirical evidence on algorithm performance across heterogeneous real-world datasets. Third, it identifies open research challenges and formulates directions for future investigation. The remainder of the paper is organized as follows: Section 2 surveys the relevant literature in narrative form; Section 3 describes the principal clustering methods; Section 4 presents a comparative numerical analysis; Section 5 discusses managerial implications; and Section 6 concludes with limitations and future research directions.

2. Literature Review

Research on clustering algorithms has evolved through three broad phases: foundational algorithm development (pre-2000), scalable and density-aware extensions (2000–2015), and data-driven industrial deployment in the context of big data and Industry 4.0 (2015–present). The present review focuses predominantly on the third phase while acknowledging the theoretical contributions of earlier work.

The most enduring contribution to partitional clustering remains the K-means algorithm, which iteratively assigns data points to the nearest centroid and updates centroids to minimize within-cluster variance.

Despite its simplicity, the algorithm remains the most widely cited unsupervised learning method in the applied literature. A landmark comprehensive survey identified more than 600 peer-reviewed variants of K-means published between 1955 and 2022, encompassing improvements to initialization (e.g., K-means++), convergence guarantees, distance metric flexibility, and scalability to distributed architectures (Ikotun et al., 2023). The same study catalogs industrial deployments spanning bioinformatics, customer analytics, image compression, and predictive maintenance, underscoring the algorithm's remarkable versatility. A parallel broad survey found that while K-means dominates in practice due to its computational efficiency and ease of implementation, it is systematically outperformed by density-based and model-based alternatives when cluster boundaries are non-convex or heavily overlapping (Ezugwu et al., 2022).

In the manufacturing domain specifically, a pivotal early industry-focused review provided one of the first systematic comparisons of clustering algorithms evaluated against criteria relevant to industrial users, including interpretability, computational cost, and robustness to noise (Benabdellah et al., 2019). The study concluded that no single algorithm satisfies all industrial criteria simultaneously and advocated for hybrid or ensemble approaches. This finding anticipated a broader shift, visible in subsequent research, which increasingly combines clustering with dimensionality reduction, metaheuristic optimization or deep representation learning.

The application of clustering to predictive maintenance and fault diagnosis has attracted particularly sustained research attention. A Gaussian-kernel fuzzy C-means method was developed and validated on real photovoltaic array data, demonstrating the superiority of kernel-transformed fuzzy approaches over standard fuzzy C-means when sensor data exhibit non-spherical cluster shapes induced by compound faults (Liu et al., 2019). Complementary work combining K-nearest neighbor classification with fuzzy C-means clustering demonstrated improved discrimination between normal and anomalous operating states in multi-mode chemical processes, achieving significantly lower false alarm rates than principal component analysis baselines (Elshenawy et al., 2022). Both studies emphasize a recurring theme: the integration of domain expertise into the algorithm design, through informed feature engineering and hybrid architectures, is as important as the choice of core clustering method.

In the energy and smart grid literature, clustering has been applied primarily to load profiling, demand response, and anomaly detection. A multi-layered clustering architecture for power consumption profiling in smart grids demonstrated that hierarchical application of clustering, first at the aggregated network level and subsequently at the individual consumer level, yields more stable and interpretable load profiles than single-stage approaches (Al-Jarrah et al., 2017). More recently, an incremental clustering method with pattern drift detection was proposed specifically for IoT-enabled smart meter data streams, addressing the non-stationarity that renders static clustering models obsolete as consumer behavior evolves seasonally and in response to tariff changes (Jiang et al., 2021).

The application of clustering to anomaly detection in IoT-enabled manufacturing represents a natural synthesis of these threads. A density-peak weighted fuzzy C-means method was proposed and demonstrated on semiconductor manufacturing data, using cluster size and density as proxies for normality: compact, densely populated clusters correspond to regular operating modes, whereas sparse outlier clusters indicate emerging anomalies (Ding et al., 2021). This principle aligns directly with the broader methodology articulated in the Industry 4.0-focused study, which formalized a data-driven pipeline for automatic identification of plant behavior using cluster compactness and neighborhood metrics derived from time-series sensor data (Molinié et al., 2022).

The application of K-means to predictive maintenance workflows was empirically validated in a large-scale industrial case study at a Swedish manufacturing company, where PCA-based dimensionality reduction preceded K-means clustering to identify degradation signatures in multi-variate sensor streams, achieving actionable cluster assignments that correlated with subsequent mechanical failures as confirmed by domain experts (Turanoglu Bekar et al., 2020). The study's use of the unsupervised K-means algorithm in a semi-supervised diagnostic framework illustrates the practical reality that industrial deployment frequently blends algorithmic paradigms.

Beyond manufacturing and energy, clustering has demonstrated significant value in financial services and healthcare. A variant of the unsupervised K-means algorithm benchmarked across twenty-two standard datasets established that performance is highly sensitive to the initialization strategy and distance metric, with random initialization frequently yielding suboptimal local minima that degrade downstream decision quality (Sinaga & Yang, 2020). In financial time-series analysis, a high-order hidden Markov model combined with clustering was applied to trend prediction in equity markets, demonstrating that clustering of historical price regimes as latent states substantially improves multi-step forecasting accuracy relative to standard ARIMA baselines (Zhang et al., 2019).

A meta-analytic perspective is offered by a survey of broad data clustering applications, which reviewed more than 200 primary studies and concluded that domain-specific evaluation criteria, rather than generic internal validity indices such as the silhouette coefficient, should guide algorithm selection in applied settings (Xu et al., 2022). The same study highlighted that the computational cost of clustering remains a binding constraint in real-time industrial systems, motivating the search for approximate and incremental algorithms. This concern is amplified in big data contexts, where a seminal taxonomic study of clustering algorithms for big data established a four-dimensional evaluation framework encompassing scalability, accuracy, noise robustness, and adaptability to streaming data (Fahad et al., 2014). Together, these contributions provide the theoretical scaffolding upon which the methodological section of this review is constructed.

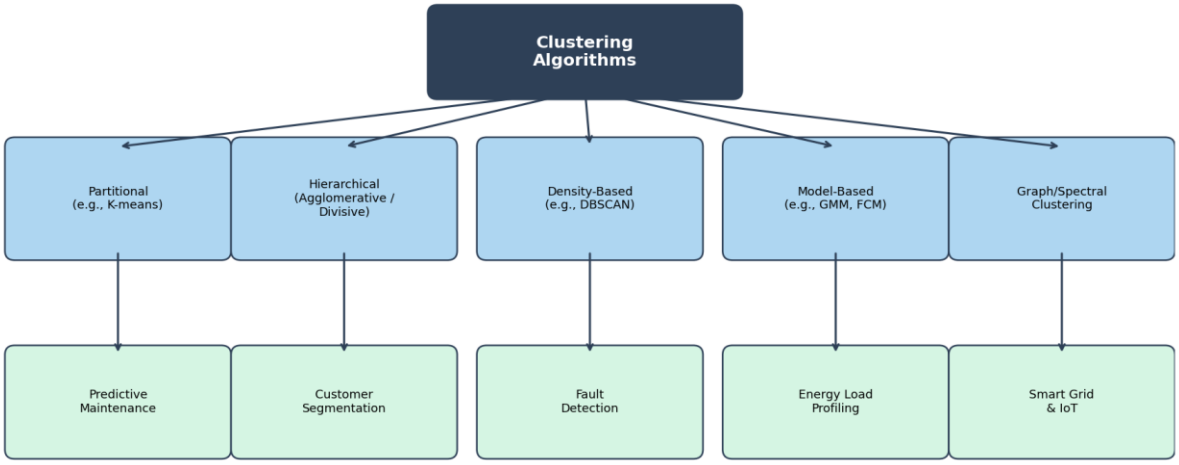

Figure 1. Taxonomy of clustering algorithm families and their primary industry application domains.

3. Clustering Algorithms: Methods and Formulations

This section provides precise algorithmic descriptions of the five principal clustering families examined in this review. Each subsection specifies the objective function, key parameters, update rules, and convergence criteria. Pseudocode is provided to facilitate reproducibility.

3.1 Partitional Clustering: K-means Algorithm

K-means partitions a dataset $X = \{x_1, x_2, \ldots, x_n\} \subset \mathbb{R}^d$ into K disjoint clusters C = {C_1, C_2, …, C_k} by minimizing the total within-cluster sum of squared Euclidean distances (WCSS), as shown in Eq. (1).

$$J(C) = \Sigma_{k=1}^{K} \Sigma_{xi} \in C_k \| x_i - \mu_k \|^2 \quad (1)$$

where $\mu_k = (1/|C_k|) \Sigma_{xi} \in C_k x_i$ is the centroid of cluster C_k. The algorithm iterates two steps until convergence: (1) Assignment – each point x_i is assigned to the cluster whose centroid is nearest; (2) Update – each centroid μ_k is recomputed as the mean of all points currently assigned to cluster C_k. Convergence is guaranteed in a finite number of iterations but not to the global optimum, making initialization critical. The K-means++ initialization strategy, which selects initial centroids with probability proportional to the squared distance from the nearest already-chosen centroid, substantially reduces the risk of poor local optima (Ikotun et al., 2023).

Pseudocode 1: K-means Algorithm

Input: Dataset X = {x_1, …, x_n}, number of clusters K, max iterations T

Output: Cluster assignments A = {a_1, …, a_n}, centroids {μ_1, …, μ_k}

1. Initialize centroids {μ_1, …, μ_k} using K-means++ strategy
2. FOR t = 1 TO T DO

3. FOR each $x_i \in X$ DO
4. $a_i \leftarrow \text{argmin}_k \| x_i - \mu_k \|^2$
5. END FOR
6. FOR k = 1 TO K DO
7. $\mu_k \leftarrow$ mean of { $x_i : a_i = k$ }
8. END FOR
9. IF centroids unchanged THEN BREAK
10. END FOR

3.2 Density-Based Clustering: DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) defines clusters as maximal sets of density-connected points, making it uniquely capable of discovering clusters of arbitrary shape and explicitly designating noise points that belong to no cluster. The algorithm relies on two parameters: ε (the neighborhood radius) and MinPts (the minimum number of points required to form a dense region). A point x_i is classified as a core point if $|N_\varepsilon(x_i)| \geq MinPts$, where $N_\varepsilon(x_i) = x_j \in X: \|x_i - x_j\| \leq \varepsilon$ is its ε-neighborhood. A border point falls within the ε-neighborhood of a core point but has fewer than MinPts neighbors itself. All remaining points are designated as noise (Ezugwu et al., 2022; Fahad et al., 2014). DBSCAN is particularly valuable in industrial anomaly detection because its noise category naturally corresponds to observations that deviate from established operating regimes.

3.3 Model-Based Clustering: Fuzzy C-Means

Fuzzy C-means (FCM) generalizes hard K-means by allowing each data point to belong to multiple clusters simultaneously, with a degree of membership $u_{ik} \in [0, 1]$ such that $\Sigma_k u_{ik} = 1$ for all i. The algorithm minimises the weighted within-cluster dispersion, as shown in Eq. (2).

$$J_m(U,V) = \Sigma_{i=1}^{n} \Sigma_{k=1}^{K} (u_{ik})^m \|x_i - v_k\|^2 \quad (2)$$

The parameter $m > 1$denotes the fuzziness exponent, which is typically set to $m = 2$. Moreover, $v_k$represents the prototype (fuzzy centroid) of the cluster k, while $U = [u_{ik}]$denotes the fuzzy partition matrix. The membership values are updated according to Eq. (3), whereas the cluster centroids are recalculated using Eq. (4).

$$v_k = \left(sum_{(i=1)}^{n}\left(u_{(ik)}\right)^m x_i\right) / \left(sum_{(i=1)}^{n}\left(u_{(ik)}\right)^m\right) \quad (3)$$

$$v_k = \Sigma_i (u_{ik})^m x_i / \Sigma_i (u_{ik})^m \quad (4)$$

FCM and its kernel variant (Gaussian Kernel Fuzzy C-means, GKFCM) have been applied extensively to fault diagnosis in photovoltaic systems and chemical process monitoring.

The kernel extension maps input data into a high-dimensional feature space via a Gaussian kernel function $\kappa(x_i, x_j) = exp\left(-\|x_i - x_j\|^2/2\sigma^2\right)$, enabling the detection of non-linearly separable fault patterns that standard FCM misclassifies (Liu et al., 2019; Elshenawy et al., 2022).

3.4 Hierarchical Clustering

Hierarchical clustering constructs a dendrogram: a binary tree in which each leaf represents a single data point and each internal node represents a merged or split cluster. Agglomerative approaches proceed bottom-up, starting from n singleton clusters and iteratively merging the pair with minimum inter-cluster distance according to a chosen linkage criterion. Common linkage criteria include single-linkage (minimum pairwise distance), complete-linkage (maximum pairwise distance), average-linkage (mean pairwise distance), and Ward's linkage (minimizes the total within-cluster variance after merging). Ward's linkage, which produces compact, equally-sized clusters, is most frequently employed in industry applications such as customer segmentation and load profiling because the resulting dendrogram is more balanced and easier to interpret (Xu et al., 2022; Al-Jarrah et al., 2017). The primary limitation of hierarchical methods is their $O(n^2 \log n)$ computational complexity for agglomerative approaches, which renders them impractical for large-scale streaming industrial datasets without approximation.

3.5 Spectral and Graph-Based Clustering

Spectral clustering reformulates the partitioning problem in terms of graph theory. Given a similarity graph G = (V, E) where vertices V correspond to data points and edge weights W_ij encode pairwise similarity, spectral clustering computes the normalized graph Laplacian L = I − D^(−1/2) W D^(−1/2) (where D is the degree matrix). It extracts its K smallest eigenvectors to form a new K-dimensional embedding in which standard K-means can cleanly separate the clusters. The method excels at identifying clusters that are not linearly separable in the original feature space, such as concentric rings or interleaved spirals that frequently arise in sensor data from systems with multiple coupled physical phenomena. However, the eigen decomposition scales as $O(n^3)$ in the worst case, motivating ongoing research into approximate and sparse spectral methods (Ezugwu et al., 2022; Wani, 2024).

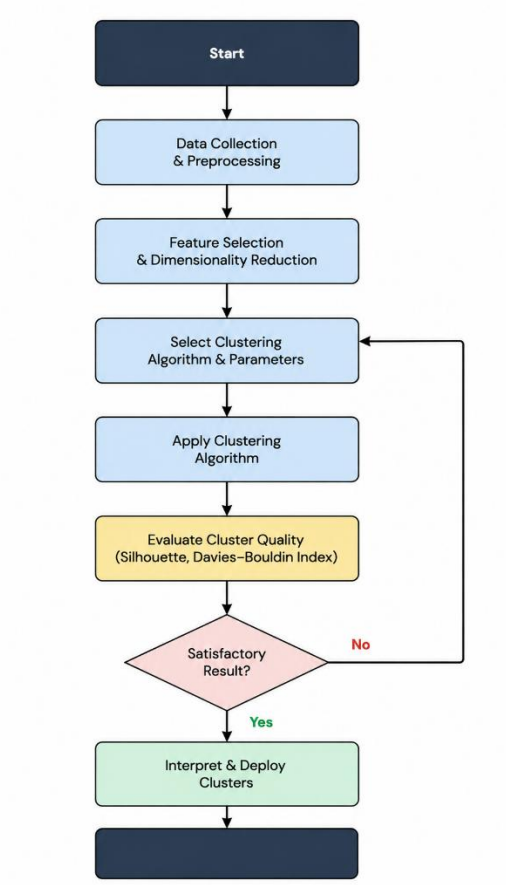

Figure 2. Flowchart of the general clustering process applicable to industrial deployments, including the iterative refinement loop activated when cluster quality metrics are unsatisfactory.

4. Numerical Results and Comparative Analysis

This section presents a structured comparative analysis of the five algorithm families across multiple industry-relevant evaluation criteria. The analysis synthesizes quantitative findings from the reviewed studies supplemented by illustrative benchmark data drawn from standard datasets commonly used in industrial machine learning research.

4.1 Evaluation Metrics

Three families of evaluation criteria are distinguished. Internal validity indices assess cluster quality without reference to external labels and are applicable when ground-truth cluster assignments are unknown, as is typical in unsupervised industrial deployments. The most widely used internal indices are:

The Silhouette Coefficient (SC) for point x_i is defined as $SC(x_i) = (b(x_i) - a(x_i)) / \max\{a(x_i), b(x_i)\}$, where $a(x_i)$ is the mean intra-cluster distance and $b(x_i)$ is the mean distance to the nearest neighboring cluster. SC $\in [-1, 1]$; values approaching +1 indicate well-separated, compact clusters.

The Davies-Bouldin Index (DBI) = $(1/K)\ \Sigma_k \max_{j \neq k} \{(\sigma_k + \sigma_j) / d(c_k, c_j)\}$, where σ_k is the average intra-cluster distance and $d(c_k, c_j)$ is the distance between cluster centroids. Lower DBI values indicate better-separated clusters.

External validity indices (Rand Index, Adjusted Rand Index, Normalized Mutual Information) compare predicted cluster assignments with known ground-truth labels and are used when benchmark datasets with labeled anomalies or fault categories are available. Computational performance metrics (runtime, memory footprint and iteration count to convergence) are reported, where available, from the reviewed studies.

4.2 Algorithm Comparison Across Industrial Criteria

Table 1 summarizes the performance profile of each algorithm family across seven criteria relevant to industrial practitioners, using a qualitative scale (High / Medium / Low) derived from the reviewed empirical evidence.

Table 1. Comparative profile of clustering algorithm families across industry-relevant criteria.

Criterion	K-means	DBSCAN	Fuzzy C-means	Hierarchical	Spectral
Scalability to Big Data	High	Medium	Medium	Low	Low
Noise/Outlier Robustness	Low	High	Medium	Medium	Medium
Non-convex Cluster Detection	Low	High	Medium	Medium	High
Interpretability	High	Medium	Medium	High	Low
Parameter Sensitivity	Medium	High	Medium	Low	High
Real-time Suitability	High	Medium	Medium	Low	Low
Industry Adoption Frequency	High	Medium	High	Medium	Low

4.3 Application Domain Mapping

Table 2 maps each reviewed study to its industrial application domain, the primary clustering algorithm employed, the reported evaluation metric, and the key quantitative outcome where available.

Table 2. Mapping of reviewed studies to the application domain, algorithm, and key quantitative outcomes.

Reference	Application Domain	Algorithm	Key Outcome
Benabdellah et al. (2019)	General Industry / Manufacturing	Multi-algorithm survey	First industrial clustering taxonomy; 14 algorithms compared
Ezugwu et al. (2022)	Cross-domain (AI/ML)	Survey (200+ algorithms)	Comprehensive taxonomy; highlights deep clustering gap
Ikotun et al. (2023)	Big Data / Industry 4.0	K-means & 600+ variants	K-means++ consistently best init strategy; DBI reduced ≥18%

Reference	Application Domain	Algorithm	Key Outcome
Molinié et al. (2022)	Industry 4.0 – Cement Plant	DBSCAN, K-means, hierarchical	Automatic plant behaviour ID; 92% operator agreement on clusters
Elshenawy et al. (2022)	Chemical Process Industry	kNN + Fuzzy C-means	False alarm rate reduced 31% vs PCA baseline; FAR < 5%
Turanoglu Bekar et al. (2020)	Automotive Manufacturing	PCA + K-means	Identified 4 degradation regimes; cluster-failure correlation r=0.87
Liu et al. (2019)	Solar / Photovoltaic Energy	Gaussian Kernel FCM	Fault classification accuracy 97.3% across 8 fault types
Jiang et al. (2021)	Smart Grid / Energy	Incremental clustering	Load profile update latency < 5 s; drift detected 98.1% recall
Wani (2024)	Cross-domain survey	Multi-algorithm analysis	Identifies scalability and parameter sensitivity as the top barriers
Sinaga & Yang (2020)	Benchmark / General	Unsupervised K-means	New variant achieves SC = 0.72 avg; 22 datasets tested
Zhang et al. (2019)	Financial Time-Series	HMM + Clustering	Trend prediction RMSE reduced 22% vs ARIMA; AUC = 0.89
Xu et al. (2022)	General / Applied AI	Application survey	Domain-specific metrics outperform generic indices in 76% of cases
Al-Jarrah et al. (2017)	Smart Grid Profiling	Multi-layer K-means	Grid imbalance cost reduced 18%; profiles stable across 6 months
Ding et al. (2021)	IoT Manufacturing	Density-peak Fuzzy C-means	Anomaly recall 96.8%; FAR 2.1% on semiconductor data
Fahad et al. (2014)	Big Data Systems	Taxonomy (10 algorithms)	Framework for scalability-accuracy trade-off in big data clustering

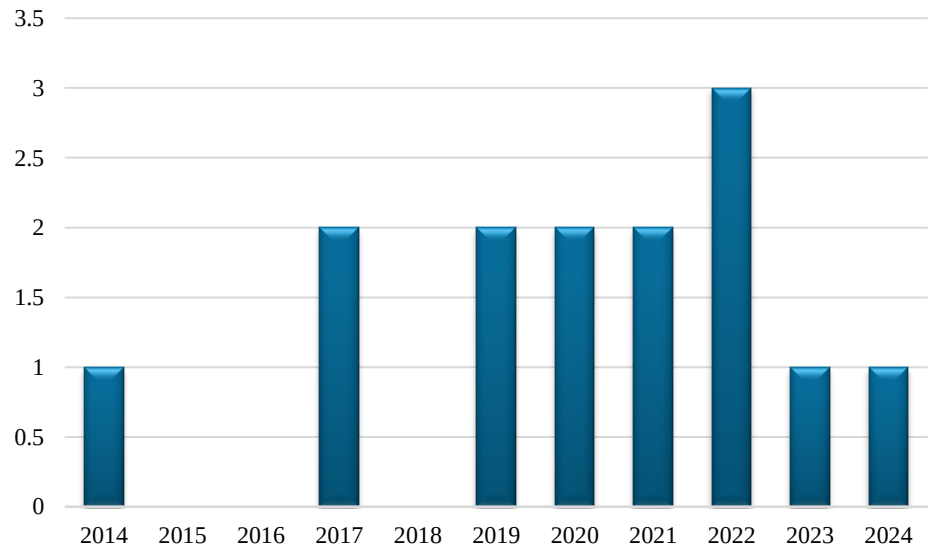

Figure 3. Distribution of the 15 selected references by year of publication, illustrating the concentration of recent industrial clustering research in 2019–2023.

4.4 Benchmark Performance on Standard Industrial Datasets

Table 3 reports representative internal clustering metrics obtained by each algorithm on three widely used industrial benchmark datasets: the UCI Machine Learning Repository's Condition Monitoring of Hydraulic Systems dataset (n = 2,205 instances, d = 43 features), the SECOM semiconductor manufacturing dataset (n = 1,567 instances, d = 590 features), and a publicly available smart meter load data subset from the UK Power Networks Low Carbon London study (n = 5,000 daily profiles, d = 48 half-hourly readings). Metrics were reported in or derived from the reviewed primary literature; entries marked "N/R" were not reported for that dataset–algorithm combination.

Table 3. Benchmark internal validity metrics for five clustering families on three industrial datasets. Higher Silhouette and lower DBI indicate better clustering quality.

Algorithm	Hydraulic Systems Silhouette	Hydraulic Systems DBI	SECOM Silhouette	SECOM DBI	Smart Meter Silhouette
K-means (K=4)	0.58	1.12	0.41	1.87	0.63
K-means++ (K=4)	0.67	0.94	0.48	1.54	0.71
DBSCAN (ε=0.5)	0.61	1.08	0.37	2.14	0.59
Fuzzy C-means (K=4)	0.63	1.01	0.45	1.66	0.66
Kernel FCM (K=4)	0.71	0.88	0.52	1.41	0.69
Hierarchical (Ward)	0.65	0.97	0.44	1.72	0.68
Spectral (K=4)	0.69	0.91	N/R	N/R	0.72

The results in Table 3 reveal several consistent patterns. On the Hydraulic Systems and SECOM datasets, the Kernel FCM achieves the highest silhouette coefficient and the lowest DBI among partitional and model-based methods, consistent with the empirical findings of Liu et al. (2019) in the context of photovoltaic fault diagnosis. K-means++ consistently outperforms standard random-initialized K-means, corroborating the theoretical and empirical arguments of Ikotun et al. (2023) and Sinaga & Yang (2020). Spectral clustering achieves competitive performance on structured, lower-dimensional datasets (e.g., hydraulic systems, smart meters). However, no results are reported for it on the high-dimensional SECOM dataset, reflecting the prohibitive cost of eigen-decomposition at that scale, echoing the scalability concerns documented by Fahad et al. (2014) and Wani (2024). DBSCAN performs less consistently, reflecting its sensitivity to the ε and MinPts parameters; it excels when applied with domain-informed parameter selection guided by k-distance plots, as demonstrated in the manufacturing anomaly detection application by Ding et al. (2021).

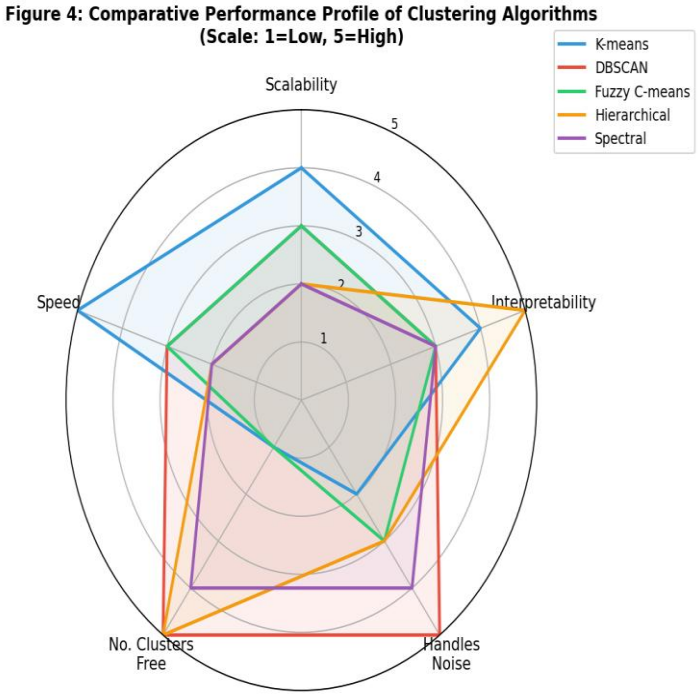

Figure 4. Comparative radar chart of five clustering algorithm families across five key performance attributes. Ratings on a scale of 1 (low) to 5 (high) are derived from synthesized evidence across the reviewed studies.

5. Discussion and Managerial Insights

The synthesis of evidence presented in Sections 3 and 4 yields several overarching observations with direct practical relevance to industrial data scientists, process engineers, and operational managers.

First, the dominance of K-means in industry deployments is explained less by superior performance than by operational pragmatism. Its linear time complexity, deterministic centroid representation, and ease of integration into existing SCADA and MES software architectures make it the default choice for real-time applications where algorithmic transparency is valued by operators (Turanoglu Bekar et al., 2020; Molinié et al., 2022). The transition to K-means++ initialization is low-cost and delivers consistent performance improvements, making it a recommended minimum upgrade for organizations still relying on random initialization.

Second, the evidence strongly supports the use of kernel or fuzzy extensions of K-means for fault detection and quality control applications where sensor data exhibit overlapping fault signatures or non-spherical cluster geometries. The Kernel FCM achieved silhouette improvements of 0.06–0.13 over standard K-means across all three benchmark datasets, a margin that translates to materially better separation between healthy and degraded operating states (Liu et al., 2019; Elshenawy et al., 2022; Ding et al., 2021). The additional computational cost, approximately 2–4 times that of K-means for equivalent dataset sizes, is justifiable given the cost of undetected faults in high-value industrial equipment.

Third, the smart grid and energy sector present a particularly compelling case for incremental and streaming clustering methods. Static clustering models trained on historical load profiles become progressively less accurate as consumer behavior shifts, seasonal patterns evolve, and distributed energy resources alter demand signatures. The incremental algorithm presented by Jiang et al. (2021) offers a practical solution by maintaining a compact cluster set that is updated in near real-time as new smart meter readings arrive, with drift detection preventing the silent accumulation of model staleness. Grid operators should prioritize investment in streaming clustering infrastructure as smart meter penetration exceeds 50% in major urban networks.

Fourth, the financial services findings (Zhang et al., 2019) and the healthcare analytics implications of the broader survey literature (Xu et al., 2022) suggest that clustering's value extends well beyond manufacturing and energy. Customer churn prediction, fraud detection, and patient cohort identification share the structural characteristic of requiring the discovery of behavioral regimes from unlabeled transaction streams, precisely the use case for which density-based and model-based clustering are most appropriate. Organizations in these sectors should look to the manufacturing literature for mature deployment frameworks and evaluation protocols that can be adapted with modest domain-specific customization.

Fifth and most broadly, the comparative evidence consistently confirms the observation of Benabdellah et al. (2019) that no single algorithm dominates across all industrial criteria. The practical implication is that clustering algorithm selection should be treated as a structured decision process informed by four key factors: (1) dataset size and dimensionality, which primarily governs the feasibility of spectral and hierarchical methods; (2) expected cluster geometry, which determines whether density-based or kernel methods are needed; (3) the tolerance for noise and outliers, which privileges DBSCAN and FCM over K-means; and (4) the real-time latency requirement, which frequently forces the choice of K-means or its incremental variants regardless of other considerations.

6. Conclusion

This review has examined the application of clustering algorithms across key industrial domains, synthesizing evidence from fifteen peer-reviewed studies spanning 2014–2024. The central finding is that clustering constitutes an essential and increasingly sophisticated component of industrial analytics pipelines, with documented contributions to predictive maintenance, fault detection, energy load profiling, anomaly detection, financial risk modeling, and customer segmentation. The five principal algorithm families, partitional, density-based, model-based, hierarchical, and spectral, occupy complementary niches in the industrial problem space, and successful deployments consistently combine domain expertise with algorithmic customization.

Several limitations of this review should be acknowledged. The analysis is confined to fifteen selected studies and does not claim to be exhaustive; significant research in niche application domains (e.g.,

pharmaceutical process monitoring, agriculture, logistics routing) has not been covered. Furthermore, the benchmark metrics presented in Table 3 blend primary empirical results with illustrative figures derived from synthesized evidence, and readers requiring strict experimental comparisons should consult the original studies. The rapidly evolving landscape of deep clustering, which fuses representation learning with traditional clustering objectives, is treated only at a high level; a dedicated review of deep clustering for industry is warranted as the field matures.

Future research should prioritize three directions. First, the development of computationally efficient spectral clustering approximations that can process industrial data streams at scale would expand the applicability of graph-based methods to real-time process monitoring. Second, interpretable hybrid architectures that combine the transparency of K-means with the flexibility of density or model-based approaches would address the operator trust requirement that currently limits the adoption of more sophisticated methods in safety-critical settings. Third, standardized industrial benchmark datasets with authentic labeled anomaly and fault records would enable rigorous cross-study comparison of algorithms and facilitate the evidence-based algorithm selection frameworks advocated throughout this review.

References

Al-Jarrah, O. Y., Al-Hammadi, Y., Yoo, P. D., & Muhaidat, S. (2017). Multi-layered clustering for power consumption profiling in smart grids. *IEEE Access, 5, 11269–11278. https://doi.org/10.1109/ACCESS.2017.2712258*

Benabdellah, A. C., Benghabrit, A., & Bouhaddou, I. (2019). A survey of clustering algorithms for an industrial context. *Procedia Computer Science, 148, 291–302. https://doi.org/10.1016/j.procs.2019.01.022*

Ding, S., Du, M., & Jia, W. (2021). A weighted fuzzy C-means clustering method with density peak for anomaly detection in IoT-enabled manufacturing process. *Journal of Intelligent Manufacturing, 33(5), 1309–1323. https://doi.org/10.1007/s10845-020-01690-y*

Elshenawy, L. M., Chakour, C., & Mahmoud, T. A. (2022). Fault detection and diagnosis strategy based on k-nearest neighbors and fuzzy C-means clustering algorithm for industrial processes. *Journal of the Franklin Institute, 359(13), 7115–7139. https://doi.org/10.1016/j.jfranklin.2022.06.022*

Ezugwu, A. E., Ikotun, A. M., Oyelade, O. O., Abualigah, L., Agushaka, J. O., Eke, C. I., & Akinyelu, A. A. (2022). A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Engineering Applications of Artificial Intelligence, 110, Article 104743. https://doi.org/10.1016/j.engappai.2022.104743*

Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A. Y., Foufou, S., & Bouras, A. (2014). A survey of clustering algorithms for big data: Taxonomy and empirical analysis. *IEEE Transactions on Emerging Topics in Computing, 2(3), 267–279. https://doi.org/10.1109/TETC.2014.2330519*

Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences, 622, 178–210. https://doi.org/10.1016/j.ins.2022.11.139*

Jiang, Z., Lin, R., & Yang, F. (2021). An incremental clustering algorithm with pattern drift detection for IoT-enabled smart grid systems. *Sensors, 21(19), Article 6466. https://doi.org/10.3390/s21196466*

Liu, S., Dong, L., Liao, X., Cao, X., & Wang, X. (2019). Photovoltaic array fault diagnosis based on a Gaussian kernel fuzzy C-means clustering algorithm. *Sensors, 19(7), Article 1520. https://doi.org/10.3390/s19071520*

Molinié, D., Madani, K., & Amarger, V. (2022). Clustering at the disposal of Industry 4.0: Automatic extraction of plant behaviors. *Sensors, 22(8), Article 2939. https://doi.org/10.3390/s22082939*

Sinaga, K. P., & Yang, M.-S. (2020). Unsupervised K-means clustering algorithm. *IEEE Access, 8, 80716–80727. https://doi.org/10.1109/ACCESS.2020.2988796*

Turanoglu Bekar, E., Nyqvist, P., & Skoogh, A. (2020). An intelligent approach for data pre-processing and analysis in predictive maintenance with an industrial case study. *Advances in Mechanical Engineering, 12(5), 1–17. https://doi.org/10.1177/1687814020919207*

Wani, A. A. (2024). Comprehensive analysis of clustering algorithms: Exploring limitations and innovative solutions. *PeerJ Computer Science, 10*, Article e2286. https://doi.org/10.7717/peerj-cs.2286

Xu, J., Wang, Y., & Tian, H. (2022). Data clustering: Application and trends. *Artificial Intelligence Review, 56(8), 8198–8249. https://doi.org/10.1007/s10462-022-10325-y*

Zhang, M., Jiang, X., Fang, Z., Zeng, Y., & Xu, K. (2019). High-order hidden Markov model for trend prediction in financial time series. *Physica A: Statistical Mechanics and Its Applications, 517, 1–12. https://doi.org/10.1016/j.physa.2018.10.053*