Review Article

Ridhwan Dewoprabowo*, Lim Yohanes Stefanus, and Ari Saptawijaya

Explainable clustering: Methods, challenges, and future opportunities

https://doi.org/10.1515/jisys-2024-0477
received November 22, 2024; accepted July 31, 2025

Abstract: In recent years, artificial intelligence (AI) has increasingly relied on subsymbolic techniques like machine learning (ML). Despite their widespread use, these techniques often lack transparency, leading to potential distrust. The field of eXplainable artificial intelligence (XAI) addresses this issue by making intelligent systems observable, explainable, and accountable. While much research has focused on explainability in supervised learning, there is a growing need to explore it in an unsupervised setting, especially given the challenges of unlabeled data in high volume. Clustering is an unsupervised ML strategy that groups data based on similarity. However, its reasoning often lacks transparency. This article reviews state-of-the-art explainable and/or interpretable clustering methods, categorizing them based on explanation generation techniques and highlighting the importance of making clustering results interpretable. We also discuss the challenges and opportunities in this domain and suggest future research directions, particularly the interpretability of advanced AI techniques like neural networks and large language models in the context of clustering. Our contributions include a comprehensive categorization of explainable clustering research and potential future research avenues to enhance the transparency and trustworthiness of clustering methods.

Keywords: eXplainable artificial intelligence (XAI), unsupervised learning, clustering, survey

1 Introduction

Artificial intelligence (AI) has recently seen increasing reliance on subsymbolic techniques, particularly machine learning (ML), which allows intelligent systems or agents to learn and improve using data and experience. ML techniques have been widely used in domains such as computer vision, pattern recognition, and classification. However, many of these techniques suffer from a lack of transparency, as they often generate outputs without providing accessible reasoning behind those decisions. This opacity can lead to confusion, fear, and distrust in AI systems [1], especially in critical decision-making contexts like healthcare and finance. eXplainable artificial intelligence (XAI) seeks to mitigate this issue by ensuring that intelligent systems are not only powerful and accurate but also observable, explainable, and accountable to users [1]. XAI makes AI more trustworthy by enabling users to understand how decisions are made, which factors are most influential, and why certain results are reached. For instance, in autonomous vehicles, XAI can improve user experience and perceived safety by providing real-time explanations for the vehicle's actions. Studies show that users feel more in control and have a positive experience when informed about the vehicle's reasoning, such as during unexpected maneuvers, through augmented reality displays and textual overlays [2]. By

*** Corresponding author: Ridhwan Dewoprabowo**, Faculty of Computer Science, Universitas Indonesia, Depok 16424, Indonesia, e-mail: ridhwan.dewoprabowo31@ui.ac.id
Lim Yohanes Stefanus: Faculty of Computer Science, Universitas Indonesia, Depok 16424, Indonesia, e-mail: yohanes@cs.ui.ac.id
Ari Saptawijaya: Faculty of Computer Science, Universitas Indonesia, Depok 16424, Indonesia, e-mail: saptawijaya@cs.ui.ac.id

bridging the gap between performance and transparency, XAI fosters trust and promotes responsible AI adoption in various real-world applications.

Research on explainability for ML mainly focused on supervised learning [3,4], while unsupervised learning and reinforcement learning have yet to be as widely explored. In recent years, unsupervised ML has gained increased importance, particularly in real-world applications that deal with large volumes of unlabeled data. These applications, such as healthcare and finance, often generate massive amounts of data where manually labeling every instance is impractical due to time, cost, and expertise constraints [4]. In healthcare, for instance, patient data include clinical measurements and other vital information that are not prelabeled with diagnoses or outcomes. Similarly, financial institutions process vast amounts of transactional data daily to detect frauds, and much of these data are unlabeled. In such scenarios, relying solely on supervised learning is inefficient, as it requires a substantial amount of labeled data and can introduce biases from these labels [4]. Unsupervised learning offers a solution by autonomously discovering patterns and relationships in unlabeled datasets, making it crucial for handling the growing scale and complexity of data in these domains.

Given the abundance of real-world unlabeled data, developing explainable unsupervised ML methods is essential to ensure that the resulting clusters or patterns are not only accurate but also interpretable. Explainable unsupervised learning is particularly valuable because it allows decision-makers to understand the discovered clusters, enabling better, more informed decisions [4]. For example, in healthcare, understanding how certain medical and demographic factors group patients can lead to improved treatment strategies [5], while in finance, detecting new, unknown fraud patterns through clustering can reduce risks [6]. Therefore, this work focuses on explainable unsupervised learning, specifically clustering, which is one of the key methods used to explore unlabeled data.

Clustering comprises a large class of methods aiming at discovering a number of meaningful groups (or clusters) in given data. Specifically, given a dataset with n entities, the goal of (crisp) clustering is to identify k groups such that (1) each group contains at least one entity, and (2) each entity belongs to one and only one group, i.e., the groups are disjoint [7]. In other words, clustering defines a *partition* of n entities in k clusters. Since there can be many partitions, we need a way to explain how such partitions can be meaningful. We say a group is meaningful if the entities in that group share similarities. Moreover, they also have to be different from entities in other groups. Currently, there are already many approaches to cluster data, and a comprehensive review of clustering methods is available [7].

While clustering algorithms are capable of grouping data points based on their similarity, these algorithms often lack transparency in the reasoning behind their cluster assignments. In the real-world setting, it is important to understand the rationale behind cluster formation to ensure the trustworthiness and interpretability of clustering results, especially in domains where decision-making relies heavily on these insights, such as patient segmentation in healthcare [5,8] and fraud detection [9]. Clustering is useful for fraud detection because it can uncover new, unseen fraud patterns by identifying outliers and risks in both approved and declined transactions, unlike supervised models that only recognize known patterns [6]. For example, in the health insurance domain, they are often implemented to detect false insurance claims, and experts are then notified if a decision needs to be made. However, to fully trust the outcomes of the AI system, clustering explanation techniques are highly desirable to avoid incorrect decisions, which may cause harm to a person who may need healthcare but get wrongfully rejected, or financial loss for an insurance company in case the fraud is actual [6,10].

Explanation generation techniques help users validate, refine, and derive actionable insights from clustering results by providing human-understandable explanations for clustering decisions. In this paper, we present a review of state-of-the-art explainable clustering (ExClust) methods. We explore several techniques employed to explain the rationale behind cluster assignments. Moreover, we assess the challenges and opportunities that arise with respect to explainable clustering and identify several promising avenues for future research. The contributions of this work can be summarized as follows:

- We explore and categorize existing research on explainable clustering, organizing it based on the underlying principles of the explanation generation methods. This categorization offers insights into the diverse approaches that make clustering results more interpretable.

- We outline potential avenues for future research. This includes advancements not only in explainable clustering but also within the field of XAI. We emphasize the current development of AI techniques, e.g., neural networks and large language models (LLMs), and suggest directions for utilizing them in the field of explainable clustering and XAI in general.

The remainder of this article is organized as follows. In Section 2, we explain the foundational concepts of XAI that are used in the subsequent discussions. Section 3 comprehensively explores existing methods for explainable clustering and categorizes them. In Section 4, we discuss the challenges encountered in this domain, which can be explored in future research. Next, we explore several related works on XAI survey in Section 5. Finally, Section 6 concludes this work by summarizing some key insights and outlines for future exploration.

2 Preliminaries

This section introduces fundamental concepts and algorithms relevant to the study of explainable AI (XAI) and clustering methods. We begin by defining key terms in XAI. We then provide an overview of basic clustering algorithms frequently used in the context of explainable clustering.

2.1 Explainable AI

In the current literature of XAI, one issue that persists is that there seems to be no agreement on the definition of the term interpretability and explainability [1,11]. These concepts are notably different, and here we state the definition used by Arrieta et al. [1].

- **Interpretability**. In general, the term interpretability is used to describe a passive characteristic of an ML model, denoting the level at which a human observer can understand how the model works.
- **Explainability**. In contrast, the term explainability usually refers to an active characteristic of an ML model, denoting any action or procedure that the model performs to clarify or detail its internal function.

Moreover, explainable AI is already a diverse research area that consists of many components. In the following, we present a general taxonomy of XAI methods [4,12,13].

- **Intrinsic** or **Extrinsic** (post hoc): The first distinction is whether the ML model itself is already interpretable or requires other methods to analyze the model after training to achieve interpretability [4]. The former is called an *intrinsic* model, with examples like short decision trees [14] that are naturally interpretable. The latter is referred to as *extrinsic* or *post hoc*, which involves applying interpretability methods like local interpretable model-agnostic explanations (LIME) [15] or class activation mapping [16] after the model is trained to explain its predictions.
- **Model-specific** or **Model-Agnostic**: The key difference between these approaches is whether the interpretation method is tailored to a specific type of model or can be applied universally across different models [4]. Model-specific methods, such as neural additive models [17], which blend the interpretability of generalized additive models with deep neural networks, are designed for a specific class of ML models and leverage access to the model's internal structure or parameters. In contrast, model-agnostic methods, like LIME [15] or SHapley Additive exPlanations (SHAP) [18], can be applied to any model, as they do not rely on internal model details. Instead, they generate explanations by analyzing input-output behavior, making them flexible for a wide range of models.
- **Local** or **Global**: The difference here lies in the scope of the explanation provided. *Local* XAI methods focus on explaining individual predictions or specific outputs of an ML model. These methods, such as LIME [15], help users understand why a particular decision was made by analyzing the features that contributed most to a single prediction. On the other hand, *global* XAI methods provide a comprehensive understanding of how the entire ML model operates. These methods, such as decision trees [14] and partial dependence plots

[19] give insights into the overall decision-making process, revealing dependency patterns or rules that apply to the model as a whole, rather than only a single output.

In this survey, we distinguish between interpretable and explainable clustering models based on their inherent transparency to a human audience. An **interpretable model** is a white-box clustering model, meaning its clustering process and mechanism are directly understandable by the audience without the need for additional methods. These clustering models, such as decision trees [20–22] or rule-based systems [23], provide transparency in how clusters are formed, making them easily interpretable by humans. In contrast, an **explainable model** refers to a black-box model where the clustering mechanism is not immediately understandable by the audience. These clustering models require additional post hoc or approximation methods, such as threshold trees [24–26] or pattern mining [27], to provide insight into how the clustering results are derived.

We use the following notations for tables throughout this article: interpretability versus explainability (**IN**/**EX**); intrinsic versus post hoc explanations (**I**/**P**); model-agnostic versus model-specific approaches (**A**/**S**), where * denotes model-agnostic in a restricted class only; global versus local explanations (**G**/**L**); and whether the method deals with numeric or categorical data (**N**/**C**).

2.2 Clustering algorithms

In this subsection, we provide an overview of basic clustering algorithms that are frequently referenced in the context of explainable clustering.

- **k-means** [28]: The k-means algorithm starts by selecting k initial centroids and assigns each data point to the nearest centroids based on a distance metric. It then updates the centroids to the mean of the points in each cluster and iterates this process until a convergence criterion is met.
- **k-median** [28]: The k-medians clustering method calculates the median along each dimension for each cluster, unlike k-means, which uses the mean. The k-medians algorithm seeks k centroids with the goal of minimizing the sum of distances between each data point and its nearest cluster center.
- **k-center** [27]: The k-center clustering method seeks to select k nodes as centers and allocate each node to the nearest center, ensuring that the maximum distance any node has from its center is minimized.

3 Explainable clustering methods

To illustrate the challenge of interpretability in clustering, consider the task of assigning n points in $\mathbf{R}^d$ into k clusters using k-means or k-median algorithms. In these methods, the clustering is determined by a set of k centers, $c_1, c_2, \ldots, c_k$, where each center represents the mean (in k-means) or median (in k-median) of the points assigned to that cluster. A data point is assigned to a cluster based on its proximity to the nearest center. Specifically, for any given point, the algorithm computes the distances from that point to each center, and the point is assigned to the cluster whose center minimizes the distance. In other words, if a point belongs to cluster i, its distance to center c_i is the smallest when compared to the distances to other centers c_j ($j \neq i$).

While this distance-based assignment is computationally effective, it poses significant challenges for human interpretation, particularly in high-dimensional spaces. Since the assignment depends on calculating distances across all features of the data, it becomes difficult to understand which specific features contribute most to a point's membership in a particular cluster. For example, a point could be closer to one cluster due to certain features being more heavily weighted in the distance calculation, but the algorithm does not explicitly reveal which features are most influential. This "black-box" nature of distance-based methods like k-means and k-median obscures the underlying factors driving cluster assignments, making it hard for users to interpret the clustering results.

A common approach for cluster interpretation is visualizing clusters using two-dimensional projections from principal component analysis (PCA) [29]. PCA reduces the dimensionality of data by projecting it onto new axes that capture the most variance, making high-dimensional data easier to visualize. However, this process often obscures the relationship between clusters and the original feature variables [30], as PCA focuses on maximizing variance rather than feature relevance. As a result, it can be difficult to identify which specific features are responsible for driving the separation between clusters. Moreover, PCA can distort the true structure of clusters by collapsing multiple dimensions into just two, sometimes leading to misleading overlaps or exaggerating separations that are not present in the full-dimensional space. This limitation hinders the interpretability of the clustering results, as the visual representation may not fully reflect the complexity of the data.

In this section, we explore several alternative approaches that have been proposed for explainable clustering, grouping them based on their characteristics into distinct categories. Figure 1 provides a high-level infographic of the five main categories. We primarily focuses on theoretical frameworks and does not include empirical evaluations of the discussed methods. This omission is intentional, as an extensive empirical comparison falls beyond the scope of this study. However, we acknowledge that different explainability methods may require distinct evaluation metrics depending on their characteristics and intended applications. As highlighted by Nauta et al. [31], explainability is a multifaceted concept, and various quantitative evaluation methods exist to assess different aspects of explanation quality, such as correctness, completeness, and coherence. Future work could integrate such empirical analyses to complement the theoretical perspectives presented here.

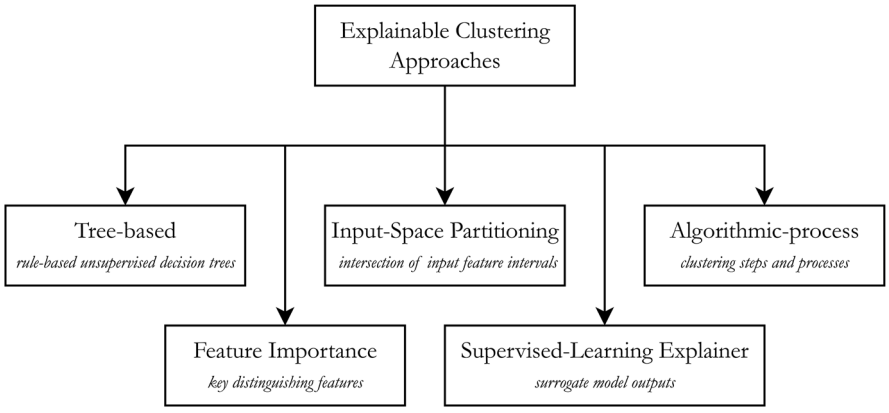

Figure 1: Five main categories of explainable clustering methods. Source: Created by the authors.

3.1 Tree-based approaches

One popular method for achieving interpretability and explainability in clustering is through the use of *unsupervised decision trees*. These tree-based approaches offer several advantages over traditional distance-based clustering methods, as they directly enhance interpretability by providing clear, rule-based explanations for cluster formation. Unlike distance-based methods, which rely on similarity metrics like Euclidean distance to group data points, unsupervised decision trees focus on creating hierarchical, rule-based structures that are more transparent to users. These methods typically possess three main characteristics [22]: (a) they do not require predefined distance measures to assess similarity between data points, thus making them more flexible in handling different types of data; (b) they utilize split evaluation measures to divide data points at each internal node by searching for highly separated and compact groups, optimizing the partitioning process for interpretability; and (c) they describe the resulting clusters using decision trees, where each leaf represents a cluster, and the path from the root to the leaf forms an interpretable rule. This ability to create interpretable rules at each decision node makes unsupervised decision trees highly effective for providing insights into how clusters are formed, offering a level of explainability that traditional distance-based methods often lack.

Liu et al. proposed CLTree [32], a method that leverages decision tree construction to partition the data space into cluster (dense) regions and empty (sparse) regions, which helps identify outliers and anomalies. CLTree introduces virtual nonexisting data points, called "**N**" points, uniformly distributed across the space, contrasting with actual data points labeled as "**Y**." This transforms the clustering task into a classification problem where the decision tree separates **Y** points from **N** points, partitioning the space into interpretable hyper-rectangular regions. A key feature of CLTree is its dynamic introduction of **N** points at each node during the tree-building process, ensuring a balance between dense and sparse regions even in high-dimensional spaces. These **N** points are computed on-the-fly to increase scalability. The final clusters are expressed as hyper-rectangular regions, offering a clear and interpretable description of the clustering structure in the data.

Basak et al. proposed an interpretable hierarchical clustering method that constructs an unsupervised decision tree [33], where each leaf node represents a cluster, and the path from the root to the leaf forms a rule for direct interpretability. The method selects the most significant attribute at each node by minimizing inhomogeneity and uses two algorithms, i.e., valley detection for numerical attributes or binary splitting – to partition the data. However, focusing on a single attribute at each split can overlook interactions between multiple attributes, potentially limiting the complexity and richness of the resulting clusters. This approach emphasizes simplicity and interpretability, but the reliance on individual attributes may prevent it from capturing multidimensional relationships that could offer deeper insights into the clustering structure.

Fraiman et al. proposed a hierarchical top-down method using an unsupervised binary tree called **CUBT** [20] that first grows a maximal tree by applying a recursive partitioning algorithm and minimizes deviance as a heterogeneity criterion. Then, prune the tree using a criterion of minimal dissimilarity based on the Euclidean distance. Finally, using either measure, similar clusters are joined together, even if they do not share the same parent. However, this approach is limited to continuous data. To mitigate this issue, Ghattas et al. [34] present an extension of CUBT [20] to nominal data. To this extent, they propose a new criterion based on mutual information or entropy during the growing and pruning step.

Gutierrez-Rodríguez et al. introduced a pattern-based clustering algorithm for numerical datasets, called **UD3** [21]. This method does not require all numerical features to be a priori discretized, which may cause information loss. However, UD3 evaluates a feature split as the separation of the means, which may cause very different point distributions to produce the same mean. UD3 also requires a parameter that controls the number of objects in the leaf nodes, but the authors do not provide any automatic algorithm to compute this parameter. To mitigate these issues, Loyola-Gonzalez et al. aim to improve UD3 by creating a new split evaluation criterion, which considers both compactness and separation, named **UD3.5** [22]. It does not require a parameter that controls the number of objects in the leaf nodes, and it automatically stops expanding the branches if the new child nodes are evaluated worse than the best evaluation computed in that branch. In addition, to create diversity, they build 100 different unsupervised decision trees using the same feature selection strategy of Random Forest and UD3.5 as an ensemble (**eUD3.5**). However, this method is constrained to numerical data, as its decision tree construction relies on mathematical measures like means and compactness, which are not easily transferable to categorical data.

Another decision tree-based clustering approach is ICOT (interpretable clustering via optimal trees) [30], which builds upon the globally optimal framework of optimal classification trees. ICOT distinguishes itself from other decision tree-based clustering algorithms by employing mixed integer optimization (MIO) to construct the entire decision tree in a single optimization step, rather than relying on the traditional greedy, locally optimal approach where splits are made incrementally. This global optimization ensures that each split is made with full knowledge of the other splits, resulting in a more cohesive and interpretable tree structure. A notable advantage of ICOT is its ability to handle mixed numerical and categorical data by the use of a re-weighted distance measure, which balances the influence of different types of variables, making it more flexible than other decision tree-based algorithms that may struggle with mixed data types. In addition, ICOT does not require manual tuning of the tree's complexity, as the clustering validation criterion naturally controls the tree's depth and complexity, ensuring a balance between interpretability and accuracy without the need for external parameters. However, ICOT may face challenges with nonseparable clusters when using parallel splits, as it relies on strict feature constraints, which may not always capture the true underlying structure of the data.

Moreover, note that all prior work on interpretable clustering is evaluated empirically, without any theoretical analysis of cluster quality compared to the optimal clustering. Therefore, Dasgupta et al. study this problem from a theoretical viewpoint, measuring cluster quality by the k-means and k-medians objectives. They define the **cost of explainability** or the competitive ratio of an explainable k-means clustering as the ratio between the cost of that clustering and the optimal unconstrained k-means clustering for the same dataset [26]. They adopt an approximation strategy in which a given reference clustering is explained by the means of a more interpretable *threshold tree*, and show that top-down decision tree algorithms, such as UD3 [21], may lead to clusterings with arbitrarily high costs. Thus, they designed an efficient greedy algorithm that leads to an $O(k)$ approximation to the optimal k-medians and an $O(k^2)$ approximation to the optimal k-means.

A *threshold tree* is a binary tree with k leaves, where each internal node splits the data into two sets by comparing a data point's feature i_u with a threshold value θ_u. The first set includes points where $x_{i_u} \leq \theta_u$, and the second includes points where $x_{i_u} > \theta_u$. This process continues recursively until each point is assigned to one of the k leaves, effectively partitioning the dataset into k clusters. The algorithm, called iterative mistake minimization (IMM) [26], minimizes mistakes at each internal node, where a mistake occurs when a threshold separates a data point from its original cluster. IMM improves the clustering process by iteratively adjusting the thresholds at each node to minimize these mistakes, ensuring that the clustering cost remains close to the optimal unconstrained clustering. IMM was further developed into a practical algorithm called ExKMC [35], which uses an approximation approach. ExKMC requires a reference clustering, typically derived from the output of an unconstrained clustering method like k-means, and simulates this clustering using a threshold tree.

Afterward, much research focused on finding better bounds for the approximation algorithm. The comparison of upper and lower bounds for each approximation algorithm is summarized in Table 1. Laber et al. improved over the upper bounds in a low-dimensional regime $d \leq \log k$, giving an $O(d \log k)$-approximation algorithm for explainable k-medians and an $O(dk \log k)$-approximation algorithm for explainable k-means [37,41]. In addition, they show an $O(\sqrt{d}k^{1-\frac{1}{d}})$ bound for k-center and an $\Theta(n-k)$ bound for maximum-spacing problem. Gamlath et al. then showed an improved clustering cost of $O(\log^2 k)$ for k-medians and $O(k \log^2 k)$ for k-means [36]. Their algorithm samples threshold cuts uniformly at random for k-medians and k-means until all centers are separated from each other. This helps in reducing the overfitting of the threshold tree to the data points. Makarychev et al. then showed an $O(\log k \log\log k)$ and $O(k \log k \log\log k)$ upper bounds for

Table 1: Methods, upper, and lower bounds for explainable k-clustering in $\mathbf{R}^d$

k-means	k-median	Ref.
Upper bound		
$O(k^2)$	$O(k)$	Dasgupta et al. [26]
$O(k \log^2 k)$	$O(\log^2 k)$	Gamlath et al. [36]
$O(kd \log k)$	$O(d \log k)$	Laber et al. [37]
$O(k \log k \log\log k)$	$O(\log k \log\log k)$	Makarychev et al. [38]
$O(k \log k)$	$O(\log k \log\log k)$	Esfandiari et al. [39]
$O(k^{1-\frac{2}{d}}\text{polylog}(k))$		Charikar et al. [25]
$O(\frac{1}{\delta}\log^2 k)$		Makarychev et al. [40]
Lower bound		
$\Omega(\log k)$	$\Omega(\log k)$	Dasgupta et al. [26]
$\Omega(k)$		Gamlath et al. [36]
$\Omega(\frac{k}{\log k})$		Makarychev et al. [38]
$\Omega(k)$	$\Omega(\min(d, \log k))$	Esfandiari et al. [39]
$\Omega(\frac{k^{1-\frac{2}{d}}}{polylog(k)})$		Charikar et al. [25]
$\Omega(\frac{1}{\delta}\log^2 k)$		Makarychev et al. [40]

k-medians and k-means, respectively, with similar algorithms as Gamlath et al. [36] but with tighter analysis [38]. In their further study, they show that the competitive ratio of their algorithm for k-median has the optimal competitive ratio [42], which matches the lower bound of IMM [26].

Esfandiari et al. gave an $O(\log k \log\log k)$ upper bound for k-medians with similar algorithm as mentioned earlier [39]. In addition, they give an $O(k \log k)$ upper bound for k-means with an algorithm similar to the study by Gamlath et al. [36] but samples cut from a different distribution. Charikar et al then present an $O(k^{1-\frac{2}{d}}\text{poly}(d \log k))$-approximation algorithm for k-means when $d = O\left(\frac{\log k}{\log\log k}\right)$, which is better than any previous algorithm [25]. Makarychev et al. provide a randomized algorithm for finding bi-criteria explainable k-means [40]. This algorithm finds a decision tree with $(1 + \delta)k$ leaves and has a competitive ratio of $O(\frac{1}{\delta}\log^2 k \log\log k)$, where δ is a parameter between 0 and 1.

Finally, Bandyapadhyay et al. introduced a new model of explainable clustering by finding a subset of data points S and a threshold tree T such that the explainable clustering induced by the tree T is exactly the same as the given clustering after removing the data points in S [24]. They measure the "cost of explainability" as the number of outliers whose removal turns the given clustering into an explainable clustering. Given a clustering, they define an optimal explainable clustering as one that minimizes the size of S. However, removing data for the sake of explainability may result in incorrect clustering if the data do not have many outliers.

While these methods focus on the cost of explainability based on the cluster quality, another critical aspect to evaluate is the interpretability of the decision tree, which greatly depends on the depths of its leaves, as previously empirically demonstrated by Piltaver et al. [43]. Explaining leaves far from the root involves many tests, making it harder for a human to grasp the model's logic. To mitigate this issue, Laber et al. propose **ExShallow** [44] with a penalty term in its loss function to favor the construction of shallow decision trees, which translate to clusters that are defined by a small number of attributes and are easier to interpret. ExShallow constructs the tree in a top-down manner, starting with an initial partition (such as from k-means). The algorithm has been shown to produce significantly shallower trees without compromising clustering quality, though it relies on heuristics to find the best cuts and does not provide approximation guarantees to the optimal reference clustering partition.

In addition, many of these algorithms assume that clusters are compact and isotropic, which can result in suboptimal clustering when the actual data contain clusters of varying shapes and densities. To address this, Gioria et al. introduced the explainable likelihood clustering algorithm (ELiCA) [45], which improves explainability and performance by using Gaussian mixture models (GMMs) to model each cluster as a multivariate Gaussian distribution. GMMs allow ELiCA to capture both the mean and covariance of data points in each cluster, making it possible to handle clusters that are not isotropic. ELiCA leverages this flexibility into its decision tree construction using a likelihood-based objective function that evaluates how well each data point fits the Gaussian distribution of a cluster and splits the data accordingly. As the tree is built iteratively, ELiCA refines the clustering by adjusting the splits to account for the full probabilistic model of each cluster. This makes ELiCA particularly well suited for datasets where clusters are not isotropic.

The methods discussed earlier are post hoc, i.e., they use fixed reference clustering, performing the clustering process first and then explaining the resulting clusters using interpretable decision trees. However, some believe that interpretability and/or explainability must be incorporated into the clustering process rather than using a single reference clustering for a truly explainable clustering. Gabidolla et al. establish a joint learning task that includes both clustering and explainer mapping [46]. They propose an algorithm that alternates between clustering and classification and employ an oblique decision tree for the classification step, which provides a good trade-off between interpretability and accuracy and adds a hierarchical structure to the clustering. Hwang et al. propose XClusters [47], which integrates clustering and decision tree training simultaneously, allowing the performance and size of the decision tree to influence the clustering process. Unlike post hoc methods, XClusters incorporates explainability during clustering, balancing cluster distortion and decision tree size using a branch-and-bound algorithm. This joint optimization produces interpretable clusters without relying on a fixed reference clustering, outperforming traditional approaches by ensuring both clustering quality and explainability are considered from the start. Table 2 summarizes the tree-based approaches for explainable clustering.

Table 2: Tree-based approaches for explainable clustering

Ref.	Clustering methods	Explanation methods	IN/EX	I/P	A/S	G/L	N/C
[32]	Decision tree	Hyper-rectangle regions	IN	I	S	G	N
[33]	Unsupervised binary trees	Binary tree	IN	I	S	G	N/C
[20]	Unsupervised binary trees	Binary tree	IN	I	S	G	N
[21,22]	Unsupervised binary trees	Pattern descriptions	IN	I	S	G	N
[34]	Unsupervised binary trees	Binary tree	IN	I	S	G	C
[30]	MIO-induced binary trees	Binary tree	IN	I	S	G	N/C
[24–26,35–40,42,44,45]	Centroid-based clustering	Threshold tree	EX	P	S	G	N
[45]	Probabilistic clustering	Threshold tree	EX	P	A*	G	N
[46]	Any clustering	Oblique decision tree	IN	I	S	G	N/C
[47]	Distance-based clustering	Decision tree	EX	I	A*	G	N/C

3.2 Input-space partitioning approach

Another line of work obtains cluster explanation via *input-space partitioning* by representing the resulting clusters as the intersection of input feature intervals. This approach provides a more geometric interpretation of the clusters by dividing the input space into distinct regions based on feature values. We summarize the methods in this approach in Table 3, together with its XAI taxonomy categorization. Pelleg et al. proposed a probabilistic generative model that assumes a mixture of tailed rectangular distributions [48]. They represent the resulting clusters as an intersection of feature intervals. However, this may be less interpretable if the data dimensionality is very large.

Table 3: Input-space partitioning approaches for explainable clustering

Ref.	Clustering methods	Explanation methods	IN/EX	I/P	A/S	G/L	N/C
[32]	Decision tree	Hyper-rectangle regions	IN	I	S	G	N
[48,49]	Rectangle mixture model	Hyper-rectangle regions	IN	I	S	G	N
[50]	MINLP	Multi polytopes	IN	I	S	G	N/C
[51]	Integer programming (IP)	Polyhedral description	EX	P	A*	G	N/C
[52,53]	GMMs and DBSCAN	Hyper-cubic regions	IN	I	S	G	N/C

* denotes model-agnostic in a restricted class only; global versus local explanations (G/L); and whether the method deals with numeric or categorical data (N/C).

Chen et al. proposed the discriminative rectangle mixture model (DReaM) [49], which integrates domain experts' knowledge by incorporating prior distributions into the decision boundaries for clustering. DReaM facilitates the specification of two types of features: rule-generating features used to construct interpretable decision rules for clustering and cluster-preserving features that define the underlying cluster structure. By using prior knowledge, DReaM allows domain experts to provide initial rules for separating data, which can be refined based on the data itself. These prior distributions enforce the decision boundaries to align closely with the expert-provided rules but are also flexible enough to adjust to the observed data. However, DReaM may combine multiple input features to form composite features in the output clustering description, which can complicate interpretation for human users [52]. While the rectangular decision rules are inherently interpretable, the formation of new composite features may reduce clarity by obscuring the direct relationships between individual features and the clusters they help form.

Lawless et al. proposed an interpretable clustering method that can cluster data points and interpret the resulting clusters by constructing polytopes around them [50]. A polytope is a geometric object with flat sides, which exists in any general number of dimensions. They define the cluster construction problem with

polytopes as a mixed-integer nonlinear program (MINLP), where clusters and polytopes are initialized using alternating minimization, and then coordinate descent is performed to enhance clustering performance. However, using MINLP makes the problem less scalable to larger datasets. To mitigate this issue, they proposed another method for explaining clusters by constructing a polyhedron around each cluster [51], aiming to either minimize the complexity of these polyhedra or reduce the number of features used in the descriptions. They frame the cluster description problem as an integer programming (IP) task instead of MINLP and introduce a column generation approach to explore an exponential number of potential half-spaces for building the polyhedra.

Finally, Sabbatini et al. proposed **ExACT** (EXplainable Automated Clustering Technique) [52], which, like tree-based clustering methods, considers all the input features to create a density-based clustering. They can provide an interpretable approximation of the identified clusters via hypercubic regions. ExACT relies on GMMs to detect clusters inside an input space region and DBSCAN to remove possible outliers before performing hypercubic approximation. In addition, they proposed **CREAM** [53] (Clustering-based Rule Extraction Advanced Method), which is inspired by ExACT but uses a different splitting criteria when performing the hypercubic approximation. In addition to clustering, they can be utilized for classification and regression tasks. However, as a drawback, both ExACT and CREAM require more computational time to be executed. The process of refining the hypercubic regions, especially when dealing with complex datasets and recursive steps, significantly increases execution time compared to other clustering algorithms.

3.3 Algorithmic process-based approaches

Another method for interpretable clustering, which may be less understandable to nonexpert users, is through *algorithmic process-based explanations*. This involves methods where the interpretability and explainability of the clustering results are derived from the algorithmic steps and processes used in the clustering method itself. These types of explanations focus on how the clustering is achieved through the algorithm's operations rather than on the characteristics of the input data or the relationships between input features. We summarize the methods in this approach in Table 4, together with its XAI taxonomy categorization.

Table 4: Algorithmic process-based approaches for explainable clustering

Ref.	Clustering methods	Explanation methods	IN/EX	I/P	A/S	G/L	N/C
[54]	Sorting	Tracing and summarization	EX	I	S	G/L	N
[55]	Any clustering	Distance-based prototype optimization	EX	P	A	G/L	N
[56,57]	k-means based neural network	Interpretable neural network components	IN	I	S	G	N

Chen et al. proposed a density-based clustering method. The proposed method, **CLASSIX** (CLustering by Aggregation with Sorting-based Indexing) [54], comprises two phases: aggregation and merging. Data points are sorted along their first principal component during the aggregation phase and then grouped using a greedy aggregation technique. Afterward, the resulting groups are merged into clusters using a distance or density-based criterion. Unfortunately, they still cannot provide a good theoretical understanding of why sorting helps clustering, even for high-dimensional datasets. The resulting explanation only describes the clustering based on its processing steps with no further insights into the input data characteristics.

Another attempt at producing cluster explanations is proposed by Carrizosa et al. [55]. They suggest employing clustering methods based on distance, aiming to divide a group of individuals so that those who are close to each other belong to the same cluster while those who are distant are assigned to different clusters. Consequently, explaining a cluster c is simply stating that it comprises individuals close to a specific individual i, termed the *prototype* of c. They provide algorithms to select a cluster prototype by min-maxing the number of true positive and false positive cases. However, merely stating that a cluster comprises individuals close to a

specific prototype does not provide much context or meaningful insight into the characteristics or behavior of the cluster members.

Finally, in the realm of neural clustering, Peng et al. introduced inTerpretable nEuraL cLustering (**TELL**) [56], a differentiable alternative to the standard k-means algorithm, by reworking the k-means objective into a neural network layer. Concretely, consider a dataset $\{X_i\}_{i=1}^N$ with each data point $X_i \in \mathbb{R}^d$. We seek to cluster these points into k groups, with each cluster j centered at $\Omega_j \in \mathbb{R}^d$. Let S_j denote the set of points assigned to cluster j. The standard k-means objective $\min_\Omega \sum_{j=1}^k \sum_{X_i \in S_j} ||X_i - \Omega_j||_2^2$ is then reworked into a neural layer, i.e., by expanding $||X_i - \Omega_j||_2^2 = ||X_i||_2^2 - 2\Omega_j^\top X_i + ||\Omega_j||_2^2$ and defining $W_j = 2\Omega_j$ and $b_j = -||\Omega_j||_2^2$, leading to the equivalent formulation $||X_i - \Omega_j||_2^2 = \beta_i - W_j^T X_i - b_j$ where $\beta_i = ||X_i||_2^2$ is a constant dependent only on the input and does not need to be optimized. This allows the clustering process to be implemented as a neural layer, where W acts as the weight matrix for the clustering layer and represent the cluster centers Ω, which can be optimized using gradient-based methods. Moreover, rather than clustering directly in the original input space, TELL leverages an autoencoder to first learn a lower-dimensional representation of each data point, which serves as the actual input to the clustering layer. Thus, the network's backpropagation steps simultaneously adjust the cluster centers and refine the data representation. TELL is interpretable with respect to model decomposability. However, this means that the interpretability of TELL depends on the interpretability of k-means and still provides limited insight into deeper, feature-level explanations for cluster memberships.

TELL still suffers from the problem of k-means, where the clustering outcomes depend on the initial placement of cluster centers. Moreover, dynamically determining the appropriate number of cluster centers is a challenging problem. To mitigate these issues, Xie et al. proposed **K-meaNet** [57], an interpretable neural clustering network that is able to determine the location and number of cluster centers adaptively. Unlike TELL, which requires a predefined number of clusters k, K-meaNet initializes with an upper bound K on the number of clusters and leverages learnable gates [58] in its cluster layer. The gates technique works by multiplying the hidden layer corresponding to the K initial centers by an attenuator. During training, the multipliers corresponding to important cluster centers increase in magnitude, while the multipliers for redundant centers shrink and are eventually pruned if it is less than some pruning threshold θ, allowing the network to dynamically refine the number of active clusters. However, similar to TELL, the interpretability of K-meaNet is also limited to model decomposability, meaning that while individual components such as inputs, learned weights, and distance formulations remain transparent, deeper feature-level explanations for cluster membership remain challenging.

3.4 Feature importance approaches

Another approach for cluster explainability is by *finding important, distinguishing features* in the input data. These methods aim to identify and highlight the key features or attributes that define and differentiate each cluster, making it easier to understand what sets one cluster apart from the others. An approach by Kim et al. tried to obtain an interpretable clustering model that describes each cluster by its distinguishing features as a logical formula [59]. However, this approach focuses on binary-valued data and is unable to determine the thresholds for continuous and categorical features. Greene et al. introduced the refined soft spectral co-clustering (RSSC) algorithm [60], a spectral co-clustering approach that enhances the clustering of text documents by generating soft, interpretable membership weights for both terms and documents. RSSC leverages an iterative matrix factorization scheme to refine initial clusters, improving accuracy while maintaining interpretability. This method allows each term and document to have partial memberships across clusters, with membership weights reflecting their relative association strengths, which help generate human-readable labels. The highest-weighted terms in each term-cluster matrix column represent each cluster's primary themes.

Plant et al. proposed **INCONCO** (INterpretable Clustering of Numerical and Categorical Objects) [61], which is designed for clustering mixed-type data by detecting cluster-specific dependency patterns between

numerical and categorical attributes. The method integrates these attributes into super-attributes that represent a cluster's semantic concepts. During the clustering process, INCONCO uses an extended Cholesky decomposition to uncover attribute dependencies, formalized by conditional probabilities, and thus enhancing interpretability. This model-based approach not only identifies clusters but also elucidates why certain data points are grouped together. One of the challenges addressed by INCONCO is avoiding the manual selection of weighting factors for numerical and categorical data, which can bias clustering results. Instead, the method dynamically learns these factors, ensuring they vary between clusters based on the underlying data dependencies.

Chen, in their dissertation [23], proposed a generative interpretable clustering model with feature selection (**GICMFS**) that automatically generates a set of rules that may involve different features for different clusters. They assume that if a data point belongs to a particular cluster, it is likely to satisfy the rules associated with that cluster. The model uses axis-parallel-cut rules, which partition features into segments using one-dimensional Gaussian mixture models. The reason for using axis-parallel-cut rules is that they provide a straightforward mechanism for generating interpretable cluster-specific rules, as each feature is considered independently along its axis, simplifying the explanation of how features contribute to cluster formation. This contrasts with more complex models that use diagonal or curved boundaries, which are harder to interpret. The selection of a feature for a cluster's rule list is based on its ability to characterize the cluster. The trained model can then provide a set of rules for each cluster by analyzing the posterior probability distributions of the Gaussian components. Additionally, they introduced a model for Interpretable clustering with similarity matrices (**ICSM**), whose generated rules are represented by lower and upper bounds on a subset of selected features from the feature matrix. A sample's cluster membership is determined by checking if it satisfies the rules for each cluster through soft thresholding. ICSM ensures that the clustering results are consistent with the similarity matrix.

Other work by Saisubramanian et al. introduced the concept of β-**interpretability** for clusters [27]. A clustering is said to be β-interpretable given features F if each cluster is composed of at least β fraction of nodes that share the same feature value. They then devised a β-interpretable clustering algorithm using k-center that works as multiobjective clustering by optimizing cluster quality and explainability. Cluster explanations are then generated as logical combinations of features of interest values related to the nodes in each cluster. Ellis et al. [62] proposed two methods to obtain explanation for clusters: global permutation percent change (G2PC) and local perturbation percent change (L2PC). These techniques quantify feature importance both globally and locally by measuring changes in clustering assignments when specific features are permuted or perturbed. These methods are applicable across various clustering algorithms (e.g., k-means, DBSCAN, Gaussian mixture models, hierarchical clustering).

Recent advances in interpretable deep clustering have addressed the critical need for explainability and interpretability in unsupervised learning scenarios. Huang et al. [63] developed feature-cluster association clustering (FCAC), an inherently interpretable deep clustering model employing a K-parallel autoreconstructive network structure. Utilizing low-rank graph Laplacians, FCAC explicitly identifies subsets of features relevant to each cluster by learning nonlinear associations within these subsets. The model automatically weights features, effectively reducing the influence of noise and enhancing clustering accuracy. Gai et al. proposed the new interpretable neural network (NINN) [64], which integrates sparse orthogonal nonnegative matrix factorization (NMF) constraints into neural network architectures. By imposing sparse constraints, NINN identifies the most informative features contributing to each cluster, while orthogonal constraints maintain clear and distinct boundaries among clusters. This structured approach yields parts-based, interpretable representations that explicitly reflect meaningful subsets of features for each cluster.

Finally, Svirsky et al. [65] introduced interpretable deep clustering (IDC), designed specifically for general tabular datasets. IDC incorporates a two-stage self-supervised learning process to achieve interpretability at both instance-level and cluster-level. It employs a gating mechanism, consisting of local gates to select informative features on a per-sample basis, and global gates providing cluster-level feature importance. This dual gating mechanism not only improves clustering accuracy but also offers detailed interpretations at both the

Table 5: Feature importance approaches for explainable clustering

Ref.	Clustering methods	Explanation methods	IN/EX	I/P	A/S	G/L	N/C/B
[59]	Mind the gap model	Distinguishing features via logical formulas	IN	I	S	G	B
[60]	RSSC	Term-cluster membership function matrix	IN	I	S	G	Text
[61]	INCONCO	Attribute dependency patterns	IN	I	S	G	N/C
[23]	Generative clustering	Distinguishing features as rules	IN	I	S	G	N/C
[23]	Clustering with similarity matrices	Distinguishing features as rules	IN	I	S	G	N/C
[27]	Centroid-based clustering	Frequent pattern mining	EX	P	S	G	N/C
[62]	Any clustering	Permutation and pertubation	EX	P	A	G/L	N
[63]	FCAC	Relevant subset of features	IN	I	S	G	N
[64]	Interpretable neural network	Non-negative matrix factorization	IN	I	S	G	N
[65]	IDC	Relevant subset of features	IN	I	S	G/L	N

sample and cluster levels, facilitating comprehensive interpretability. We summarize the methods in this approach in Table 5, together with its XAI taxonomy categorization.

3.5 Supervised learning explainability approaches

Finally, *explanation generation methods designed for supervised learning* can be applied to explain clusters. This approach involves utilizing the obtained clusters to train a classifier. Subsequently, various interpretability and/or explainability methods from supervised learning can be employed on this classifier to provide insights into the clusters. We summarize the methods in this approach in Table 6, together with its XAI taxonomy categorization. Some approaches use the resulting clusters from methods such as k-means and agglomerative clustering to train a classifier using supervised methods [66]. Afterward, LIME [15], a model-agnostic method that relies on local model linearization, is utilized to identify the most relevant features that lead to a particular cluster decision. The work by Horel et al. [67] used a similar approach but instead used SFIT [68] to generate the statistically important features that characterize the resulting clusters. These approaches are still considered suboptimal since the clustering result and the explainer is not jointly learned.

Table 6: Supervised learning explainability approaches for explainable clustering

Ref.	Clustering methods	Explanation methods	IN/EX	I/P	A/S	G/L	N/C
[66]	Any clustering	LIME [15]	EX	P	A	L	N
[67]	Any clustering	SFIT [68]	EX	P	A	G	N/C
[69]	k-means, kernel k-means, deep clustering, and related	LRP [70]	EX	P	A	L	N
[71]	k-means++, GMM, agglomerative clustering	Tree SHAP [72]	EX	P	A*	G/L	N/C

* denotes model-agnostic in a restricted class only; global versus local explanations (G/L); and whether the method deals with numeric or categorical data (N/C).

Another method for explainable clustering is by rewriting the clustering models into functionally equivalent neural networks without the need for retraining (neuralization-propagation, **NEON**) [69]. First, they transform the cluster model into a neural network with standardized detection and pooling layers that functionally mirror the original model. Then, the clustering assignments made at the neural network's output are propagated backward until they reach the input variables using a layer-wise relevance propagation (LRP)-type procedure [70]. This method can identify cluster-relevant input features as explanations precisely and systematically for k-means to some recent deep clustering model.

Most recent work on this approach by Alvarez-Garcia et al. [71] proposed a comprehensive methodological framework for explainable clustering, which involves steps from data preprocessing, dimensionality

reduction, clustering, and classification. To this effect, they developed a new data imputation method and improved existing dimensionality reduction frameworks for mixed-type data to enhance interpretability. Their framework supports comparing clustering methods and a classification pipeline incorporating local and global feature importances using Tree SHAP [72].

3.6 Comparison of explainable clustering approaches

Across the five main categories of explainable clustering methods, we highlight their key strengths and limitations in Table 7. When choosing an approach, practitioners should weigh and understand these trade-offs for selecting the right explainable clustering method in any given application.

Table 7: Strengths and limitations of explainable clustering categories

Category	Key strengths	Key limitations
Tree-based	1. Exact, rule-based explanations (global scope)	1. Deep trees become hard to follow
	2. White-box interpretability via decision paths	2. Axis-parallel splits limit shape flexibility
		3. Difficult to handle handle mixed-type data
Input-space partitioning	1. Geometric clarity via hyper-rectangles/ polyhedra	1. Scalability degrades in high dimensions
	2. Direct mapping from feature intervals to clusters	2. Difficult to handle mixed-type and very large datasets
		3. Often computationally expensive
Algorithmic process-based	1. Leverages clustering steps for insight	1. Explanations remain at procedural level, not feature-centric
		Hard for nonexperts to understand data-level rationale
Feature importance	1. Highlights the most discriminative features	1. Can oversimplify by focusing on a small subset of features
	2. Supports both global and local interpretability	2. Structural relationships among features may be lost
	3. Often handles mixed-type data naturally	3. Evaluation of “importance” is itself subjective
Supervised-learning explainers	1. Model-agnostic and broadly reusable	1. Surrogate models introduce approximation error
	2. Benefits from mature XAI tools (LIME, SHAP, LRP)	2. Fidelity to original clustering depends on classifier quality
		3. Often requires training an additional model

4 Challenges and future opportunities

In this segment, we explore various opportunities for research and highlight unresolved issues gleaned from the literature reviewed in the previous section. The challenges articulated in this section are outlined and summarized in Figure 2. This figure presents approaches to explainable clustering, where ungrouped data serves as input, along with optional reference clusters generated by conventional clustering methods. These approaches aim to produce interpretable clusters or offer distinct explanations for each cluster, enhancing clarity and understanding of the clustering results. The challenges, highlighted in color-coded boxes, indicate the specific processes from which they originate, providing a visual link between each challenge and its source process. The ordering reflects the relative significance of each challenge, as determined by their frequency of identification as key issues in previous systematic reviews on XAI [1,7,11,31,73], and their relevance to the context of explainable clustering. Some parts of the figure are taken from several prior works [26,53,59].

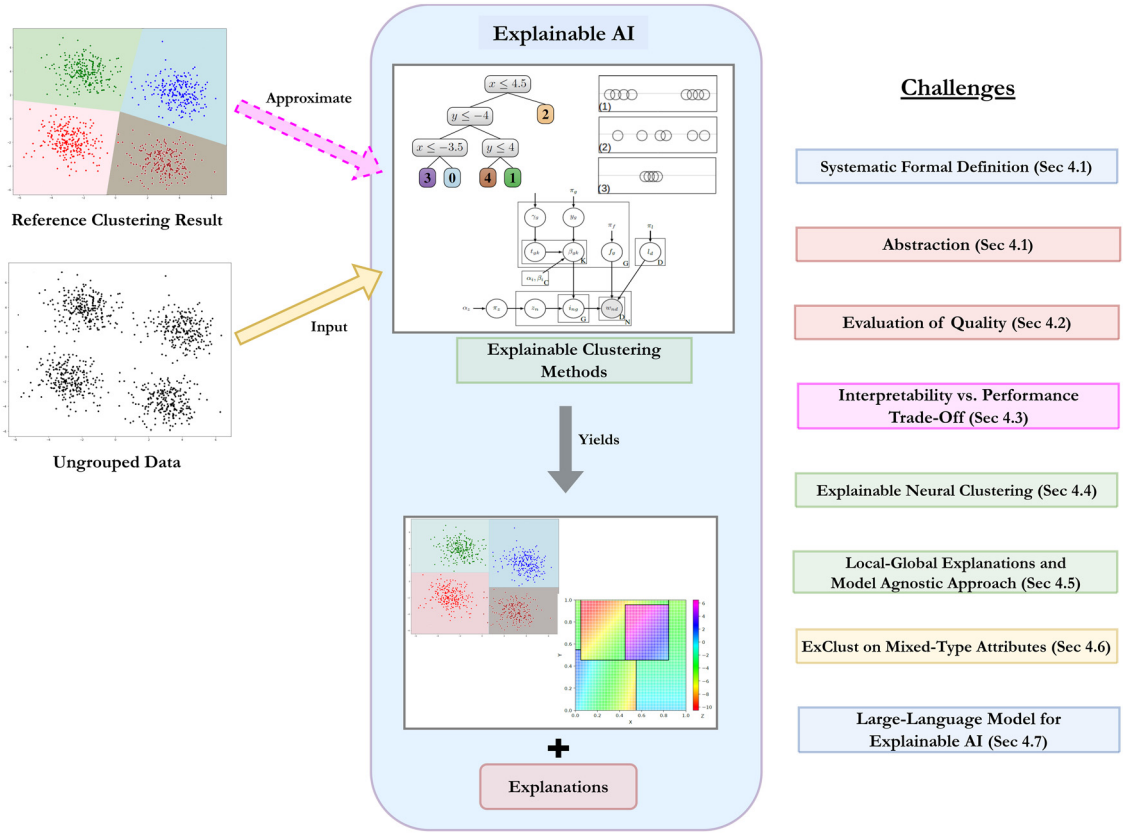

Figure 2: Current challenges in the field of explainable clustering (ExClust). Source: Created by the authors.

4.1 Abstraction and formalism

The concept of XAI, and explainable clustering in particular, is complex and cannot be tackled by a single field of study. It requires a collaborative approach that coherently combines different research methods.

(1) *Formalism*: Formalism involves providing precise definitions and a structured approach to explaining concepts. One of the issues in the field of explainable clustering (and for XAI in general) is the interchangeable use of the terms “interpretability" and “explainability" in various academic works [1,11]. Many researchers use these terms as synonyms [20,24,34,45], while others conclude that they have different meanings and implications [27]. This lack of clarity also raises another issue: the concept of explainability is very subjective and complex to formalize. An intelligent system that seems interpretable for a specific expert group may not be understandable by others, and thus, this problem is very domain specific [74]. Therefore, to advance this field, a dedicated research community should focus on establishing formal definitions for explainability. This formalization is important in XAI because it enables researchers to develop consistent and unambiguous definitions of key concepts, facilitating communication and knowledge transfer among researchers from different backgrounds [1]. Furthermore, an agreed-upon definition supports the comparison of different methods for explainable clustering, such as the formalism of “cost of explainability” [26], which provides a way to evaluate the trade-offs between clustering quality and explainability.

(2) *Abstraction*: Here, abstractions deal with how experienced the user is in the task. The user’s expertise influences the level of information they expect in their explanations [75]. In the field of explainable clustering, abstraction is a crucial component for generating interpretable explanations for clusters. By selecting and focusing only on the most important and relevant features, we can simplify complex clustering algorithms, making them more transparent and understandable for non experts.

For example, if a clustering algorithm identifies clusters based on numerous features, an abstraction layer might summarize these features into broader categories (e.g., "user engagement" instead of specific metrics like "click-through rate" and "session duration"). This makes the explanation more accessible to users who are not familiar with the technical details. In other words, the methods must be able to generate different levels of abstraction, i.e., the level of terminology used in the explanation, to suit the level of user expertise [76].

Moreover, abstraction can be used to generate explanations that go beyond technical details and provide insights into the data-generating processes that are relevant for future AI systems [12]. By understanding the underlying data-generating processes, we can identify biases, improve model performance, and provide more transparent and trustworthy AI systems.

Therefore, in explainable clustering, abstraction is an important tool that can help bridge the gap between humans and machines and advance the field of XAI. It is one of the most raised challenges in the literature of XAI, where it has been suggested that more formalism should be considered in terms of systematic definitions, abstraction, and formalizing and quantifying [11]. By establishing systematic and formal definitions for explainability and leveraging abstraction, we can generate more interpretable and trustworthy explanations that nonexperts can easily understand.

4.2 Evaluation of explanation quality

Building on the significance of abstraction in enhancing the interpretability of clustering methods, it is important to establish robust frameworks for evaluating the quality of explanations generated by these methods. As abstraction facilitates the generation of more interpretable explanations, evaluating the efficacy of these explanations becomes crucial in assessing the overall performance and utility of explainable clustering techniques.

Explainable clustering methods aim to provide insight into the underlying structure of complex datasets by producing meaningful and human-understandable clusters. However, despite the growing interest in this field, there is a lack of consensus on how to evaluate the quality of cluster explanation. Hence, it is uncertain which method constitutes the most suitable form of explanation for a particular clustering technique within a specific context, task, and domain expertise [74]. Currently, many researchers rely on subjective and anecdotal evaluations [30,49,59], which can lead to unreliable and inconsistent results. Moreover, Nauta et al. [31] found that one in three XAI papers rely only on anecdotal evidence when validating explanations, and one in five papers evaluate with users. This lack of consensus is a significant challenge for developing and adopting explainable clustering methods, as the ability to accurately assess their explainability is crucial for their success.

To ensure that these clustering models provide understandable and relevant explanations, employing human-centered evaluations involving end-users is essential. These evaluations often rely on subjective measures such as trust and confidence to assess the quality of the explanations. However, despite the growing interest in human-centered evaluations, there has yet to be a consensus on the criteria for evaluating these measures or the experimental designs to be used. One approach by Laber et al. attempted to evaluate decision tree explanation with respect to its depth [44] since humans usually found shallow trees to be more understandable [43]. However, such insight to evaluate explanation quality is rare in the current literature on explainable clustering.

To address this challenge, future research should investigate practical approaches to collect subjective measures for explanation evaluations. These measures can then be used to develop agreed criteria for human-centered evaluations, making it easier to compare the quality of different explanations [74]. While these evaluations may vary depending on the application domain and target users, identifying common components can help establish a foundation for effective evaluations. Moreover, for users to understand the quality of explanations, some properties, such as clarity, broadness of interpretability, soundness, and completeness of

fidelity, are important [11]. Comparing explanations based on these properties is essential to suggest the most appropriate explanation for users. Currently, the challenge of evaluating explanation quality is widely recognized as a major bottleneck in XAI research, with many studies noting that the lack of rigorous evaluation standards hinders progress in the field [31,77].

4.3 Interpretability vs performance trade-off

Currently, there are various methods for data clustering (see [7] for a thorough survey of clustering methods). However, many of these methods tend to be complex and lack transparency, which presents some challenges in interpreting and explaining their outcomes. While the complexity of models does not always guarantee superior clustering results, there are scenarios where complex models excel, particularly when the dataset exhibits a well-defined structure and contains highly meaningful features [1]. In such cases, a trade-off exists between model interpretability and its performance metrics [11]. The strategy to minimize this trade-off between interpretability and performance lies in the adoption of explainability techniques. However, implementing such techniques prompts questions regarding how to decide the optimal balance between interpretability and performance and what factors influence this decision [11]. The trade-off between interpretability and performance is one of the most persistent and fundamental challenges in XAI, with researchers emphasizing the difficulty of maintaining model accuracy while ensuring explanations remain understandable [1].

Some researchers have attempted to formalize this trade-off within the realm of explainable clustering, as evidenced by studies by Dasgupta et al. [26] and Bandyapadhyay et al. [78]. These researchers have introduced frameworks for defining the cost associated with explainability with respect to a reference clustering that is considered optimal. However, these frameworks are currently tailored to centroid-based clustering algorithms like k-means or k-median. Consequently, there arises a need to develop a more generalized definition of cost that can be applied across a broader spectrum of clustering methodologies, such as hierarchical, spectral, density-based, and deep clustering techniques. By extending these definitions, researchers can facilitate a more comprehensive understanding of the trade-offs in balancing interpretability and performance across diverse clustering methodologies, thus advancing the field of explainable clustering.

4.4 Explainable neural clustering

Clustering with neural networks presents challenges in achieving interpretability, particularly with complex data structures and model architectures. While neural networks offer powerful tools for capturing intricate patterns and relationships in data, understanding and interpreting the clustering decisions made by these models remain challenging. Existing research on neural clustering, e.g., DeepCluster [79] and AdaDC [80], has only primarily focused on improving clustering performance and is not interpretable since the proposed neural networks are black-box models.

Efforts to enhance interpretability in neural clustering, such as TELL [56] and K-meaNet [57], attempt to create more transparent models by linking network layers with the formulation of the k-means algorithm. However, their interpretability relies on the interpretability of k-means itself, which may not always provide clear explanations, especially in high-dimensional or complex data spaces. On the other hand, NINN [64] attempts to create interpretable neural clustering models using nonnegative matrix factorization. They can produce cluster explanations by finding features that have consistent similarities. However, the interpretability of its explanation has yet to be thoroughly evaluated.

Future research in interpretable neural clustering could focus on developing methodologies that improve clustering performance and provide clear and understandable explanations for the clustering decisions made by neural network models justified by some evaluation of explanation quality. Such methodology could

involve exploring alternative clustering algorithms that offer greater interpretability or devising techniques for visualizing and explaining the clustering process within neural networks and its results.

4.5 Local-global explanations and model agnostic approach

The literature on explainable clustering has mainly focused on providing global explanations and overall patterns and structures of each cluster in the entire dataset. However, there has been limited exploration into providing both local and global explanations for clustering results, e.g., NINN [45]. While global explanations offer valuable insights into the overall clustering behavior, local methods offer explanations for individual instances, catering to a wide range of users with varying requirements and access to resources [81]. For example, model developers, who possess both data access and insight into the model's workings, can directly examine its behavior. Analysts can assess the model's performance using a limited sample of instances. Meanwhile, end-users can obtain explanations for decisions that directly impact them.

Some explainable clustering methods, such as decision tree [26,34,47], can be extended to generate local explanations. However, existing methodologies do not directly tackle this challenge; instead, they often utilize the entire tree structure to explain clusters. This approach may lack clarity for nonexpert users and is challenging to provide, e.g., contrastive explanations (why P and not Q?), particularly when the resulting decision tree is complex. Thus, a notable gap exists in the literature regarding methodologies that integrate local and global explanations to offer a comprehensive understanding of clustering outcomes.

Moreover, the limited number of research on model-agnostic explanation techniques for clustering presents an opportunity for further research. Previous studies on explainable clustering have mainly focused on explaining specific clustering techniques, with only a few methods trying to encompass a broader, yet still restricted, array of clustering techniques [45,47,55]. Unlike model-specific explanations, which are limited to a particular algorithm, model-agnostic explanations offer a broader understanding of clustering outcomes across different algorithms [1]. This broader usage can facilitate comparisons between clustering results obtained from various algorithms. However, note that model-agnostic techniques often depend on surrogate models or alternative approximations. This may compromise the precision of the generated explanations [12]. Meanwhile, model-specific methods typically utilize the model itself for interpretation and potentially yield more precise explanations.

4.6 Explainable clustering on mixed-type attributes

Most existing research on explainable clustering has focused on homogeneous datasets, either numerical or categorical, neglecting the complexities that arise in mixed-type data. Mixed-type datasets, which combine numerical and categorical attributes, are common in real-world applications like healthcare and finance, where both types of information are integral for comprehensive analysis. For example, in healthcare, patient records may include clinical measurements (numerical) alongside demographic or lifestyle information (categorical), and in finance, customer data often involve both transactional histories (numerical) and categorical data like credit ratings or account types [12]. Explainable clustering on mixed-type data is crucial because it enables domain experts to make sense of clusters that arise from both quantitative and qualitative variables, which are often difficult to interpret when combined.

Interpreting clusters in mixed-type datasets poses unique challenges. Numerical and categorical attributes differ fundamentally in their representation, requiring distinct distance measures or transformation techniques to combine them meaningfully in a clustering process. Without careful handling, clusters may emphasize one data type over the other, leading to biased or misleading results. In addition, the challenge lies in generating clear, interpretable explanations that account for the interaction between both attribute types, ensuring that domain experts can trust and understand the results. For example, in healthcare,

doctors need to understand how medical readings and demographic information combine to form patient groups, helping them make better treatment decisions. Similarly, in finance, explainability is crucial to ensure that the clustering results, which might influence credit decisions or fraud detection, are transparent and comprehensible to regulators and stakeholders [1]. Only a few studies have addressed the challenge of providing interpretable explanations for clustering on mixed-type attributes, and further work is needed to make this process both transparent and reliable [27,30,34, 52,53,61]. In addition, extending explainable clustering techniques to other complex data types, such as time series, graphs, and spatiotemporal data, can be beneficial [11].

4.7 LLM and explainable AI

The integration of LLMs with explainable AI (XAI) presents a promising avenue for enhancing the transparency and explainability of ML systems. While XAI has made significant strides in improving the transparency of various models [11], there is growing recognition of the potential for LLMs to further this progress. Currently, there is a gap in effectively utilizing XAI for practical improvements in model evaluation and development [82]. Many existing methods focus on technical aspects of explainability without fully addressing the needs of practitioners and nontechnical users, leading to a misalignment between the goals of explainable methods and the expectations across different application domains [83]. LLMs offer unique capabilities in generating and understanding natural language explanations, which could help bridge this gap and address some of these challenges.

On the one hand, the current development of LLMs reinforces the opacity concerns that are still not fully solved for conventional deep models [84]. Given the influence of LLMs across many domains, guaranteeing their explainability and ethical utilization has become an important requirement in practical settings [82]. On the other hand, the evolving capabilities of LLMs offer opportunities for research in XAI. LLMs demonstrate considerable promise in incorporating a wide range of AI tasks and adjusting to different scenarios. This capability enables them as one possible solution to enhance the adoption of XAI [82,85] by enhancing the comprehensibility of explanations.

Current LLMs are not typically used directly for clustering tasks [86]. However, one way LLMs can contribute to clustering is through techniques like word embeddings or contextual embeddings [86–88] for text clustering tasks. These embeddings capture semantic similarities between words or phrases, which can be utilized in clustering algorithms to group similar text data together. For instance, in the work by Petukhova et al. [87], they show that text embeddings generated by pretrained LLMs can be used as features in traditional clustering algorithms. Moreover, for semi supervised text clustering, LLMs is shown to be useful as a pairwise constraint pseudo-oracle [89]. However, a large number of LLM queries are needed to be effective.

In retrospect, while the explainability of LLMs remains a challenge, they may still be useful for the field of explainable clustering. By leveraging their natural language generation capabilities, LLMs can enhance cluster explanations produced by other explanation generation methods, making them more intuitive and accessible. Furthermore, LLMs can facilitate the integration of domain-specific knowledge and contextual information into the clustering process, enhancing the relevance and usefulness of the generated explanations [82]. Through the combination of LLMs and explainable clustering techniques, users can gain deeper insights into the underlying structure of their data and make more informed decisions based on transparent and interpretable clustering results.

5 Related works

Current surveys on XAI primarily address methods for supervised learning. For instance, recent XAI surveys mainly focus on explaining supervised learning models and do not include interpretable methods for

unsupervised learning [11,13,90,91]. While these surveys comprehensively discuss various techniques to enhance the transparency and interpretability of supervised models, they overlook the specific challenges and methodologies associated with unsupervised learning.

Several surveys discuss explainable methods for unsupervised learning [4,14,92]. These surveys provide insights into various techniques to enhance unsupervised learning models' interpretability. However, despite their coverage, most of these works do not discuss and categorize many of the works on interpretability methods specifically for clustering. Other works discuss explainable anomaly detection methods [93,94]. While some methods involve clustering, this is not their primary focus. This gap indicates a need for more research and surveys that address the distinct challenges and strategies for achieving interpretability in clustering algorithms.

Recent surveys on clustering methods do not discuss in detail the challenges related to cluster interpretability or explainability, only highlighting specifically cluster visualization issues [7,95]. A notable exception is the work by Yang et al. [96], which specifically examines interpretable clustering techniques. This study reviews several methods for making clustering algorithms more interpretable and suggests possible directions for future research. However, it does not address the challenges arising from newer ML paradigms, particularly the necessity for interpretable neural network clustering and the research opportunities regarding the use of the recently developed transformer-based LLMs for explainable AI.

6 Concluding remarks

The field of explainable clustering presents both challenges and promising avenues for future research. As the complexity of ML algorithms continues to rise, there is a pressing need to develop methods that provide understandable explanations for their decisions. Explainable clustering methods offer a way to address this need by generating insights into the underlying structure of complex datasets. However, several challenges must be addressed to realize the full potential of explainable clustering techniques. This review paper summarizes the achievements of explainable clustering (ExClust) to date, categorize them based on their characteristics, and explores the associated challenges and research opportunities.

Current explainable clustering techniques can be organized into five categories: (1) unsupervised decision trees, attempt to split groups based on their feature values using trees and are interpretable by design; (2) input space partitioning, which geometrically represents clusters by partitioning the input space into distinct regions according to feature values; (3) feature importance, which shows which features are important in obtaining and distinguishing the resulting clusters; (4) algorithmic process based, which only elucidates how the clustering processes and steps work as the explanation; and (5) using explanation generation methods designed for supervised techniques, usually by training a classifier on the resulting clusters.

These methods, as with XAI in general, suffer from a lack of abstraction and formalism. This limitation may hinder knowledge transfer between researchers and is less understandable to non expert users. Therefore, solving this issue is an important first step for future research. If a general formalism can be established, then the evaluation of explanation quality across different methods, domains, and user expertise can be performed more systematically and comparably. Moreover, using LLMs in XAI may be beneficial in ensuring that users can understand explanations more easily.

Another important consideration in developing explainable AI methods is the issue of performance and interpretability/explainability trade-off. While this issue has been explored for several centroid-based clustering algorithms, a more general definition is needed to analyze the effect of interpretability/explainability on the resulting clusters quality. This problem can be addressed by building model-agnostic explanation generation methods for clustering. Furthermore, current clustering methods still need to be improved to address more diverse data types that usually occur in the real world. Current methods may need to present local explanations, which is helpful for a wide range of users.

By addressing the outlined challenges, researchers can develop more comprehensible and versatile ExClust methods. The integration of LLMs and other advanced AI techniques into ExClust can enhance its accessibility

and effectiveness. As the field of ExClust evolves, it will play a crucial role in making complex ML models more transparent and trustworthy, and broadening their applicability and acceptance across various domains.

Funding information: This work was supported by the research Grant No. NKB-9/UN2.F11.D/HKP.05.00/2024 funded by the Faculty of Computer Science, Universitas Indonesia. This study is also supported by the LPDP Scholarship, Ministry of Finance, Indonesia.

Author contributions: Ridhwan Dewoprabowo: conceptualization, methodology, investigation, writing–original draft, review and editing. Lim Yohanes Stefanus, Ari Saptawijaya: conceptualization, methodology, writing–review and editing, project administration, supervision.

Conflict of interest: The authors declare no conflict of interest.

Data availability statement: Data sharing is not applicable to this article as no datasets were generated or analysed during the current study.

References

[1] Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inform Fusion. 2020;58:82–115. doi: 10.1016/j.inffus.2019.12.012.

[2] Schneider T, Hois J, Rosenstein A, Ghellal S, Theofanou-Fülbier D, Gerlicher AR. Explain yourself! Transparency for positive UX in autonomous driving. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems; 2021. p. 1–12. doi: 10.1145/3411764.3446647.

[3] Cocarascu O, Toni F. Argumentation for machine learning: A survey. In: COMMA; 2016. p. 219–30. doi: 10.3233/978-1-61499-686-6-219.

[4] Wickramasinghe CS, Amarasinghe K, Marino DL, Rieger C, Manic M. Explainable unsupervised machine learning for cyber-physical systems. IEEE Access. 2021;9:131824–43. doi: 10.1109/ACCESS.2021.3112397.

[5] Newcomer SR, Steiner JF, Bayliss EA. Identifying subgroups of complex patients with cluster analysis. Am J Manag Care. 2011;17(8):e324–32. Available from: http://europepmc.org/abstract/MED/21851140.

[6] Min W, Liang W, Yin H, Wang Z, Li M, Lal A. Explainable deep behavioral sequence clustering for transaction fraud detection. 2021. arXiv: http://arXiv.org/abs/arXiv:210104285. doi: 10.48550/arXiv.2101.04285.

[7] Singh J, Singh D. A comprehensive review of clustering techniques in artificial intelligence for knowledge discovery: Taxonomy, challenges, applications and future prospects. Adv Eng Inform. 2024;62:102799. doi: 10.1016/j.aei.2024.102799.

[8] Elbattah M, Molloy O. Data-driven patient segmentation using K-means clustering: The case of hip fracture care in Ireland. In: Proceedings of the Australasian Computer Science Week Multiconference; 2017. p. 1–8. doi: 10.1145/3014812.3014874.

[9] Maddila S, Ramasubbareddy S, Govinda K. Crime and fraud detection using clustering techniques. Innovations in Computer Science and Engineering: Proceedings of 7th ICICSE. 2020. p. 135–43. doi: 10.1007/978-981-15-2043-3_17.

[10] Cirqueira D, Helfert M, Bezbradica M. Towards design principles for user-centric explainable AI in fraud detection. In: International Conference on Human-Computer Interaction. Springer; 2021. p. 21–40. doi: 10.1007/978-3-030-77772-2_2.

[11] Saeed W, Omlin C. Explainable AI (XAI): a systematic meta-survey of current challenges and future opportunities. Knowl-Based Syst. 2023;263:110273. doi: 10.1016/j.knosys.2023.110273.

[12] Adadi A, Berrada M. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access. 2018;6:52138–60. doi: 10.1109/ACCESS.2018.2870052.

[13] Dwivedi R, Dave D, Naik H, Singhal S, Omer R, Patel P, et al. Explainable AI (XAI): core ideas, techniques, and solutions. ACM Comput Surveys. 2023;55(9):1–33. doi: 10.1145/3561048.

[14] Míi JX, Li AD, Zhou LF. Review study of interpretation methods for future interpretable machine learning. IEEE Access. 2020;8:191969–85. doi: 10.1109/ACCESS.2020.3032756.

[15] Ribeiro MT, Singh S, Guestrin C. Why should I trust you? Explaining the Predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016. p. 1135–44. doi: 10.1145/2939672.2939778.

[16] Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016. p. 2921–9. doi: 10.1109/CVPR.2016.319.

[17] Agarwal R, Melnick L, Frosst N, Zhang X, Lengerich B, Caruana R, et al. Neural additive models: interpretable machine learning with neural nets. Adv Neural Inform Proces Syst. 2021;34:4699–711.Available from: https://dl.acm.org/doi/abs/10.5555/3540261.3540620.

[18] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inform Proces Syst. 2017;30:4765–74. Available from: https://dl.acm.org/doi/abs/10.5555/3295222.3295230.
[19] Greenwell BM PDP: An R package for constructing partial dependence plots. The R Journal. 2017;9(1):421. Available from: https://digitalcommons.unl.edu/r-journal/438/.
[20] Fraiman R, Ghattas B, Svarc M. Interpretable clustering using unsupervised binary trees. Adv Data Anal Classification. 2013;7:125–45. doi: 10.1007/s11634-013-0129-3.
[21] Gutierrez-Rodríguez AE, Martínez-Trinidad JF, García-Borroto M, Carrasco-Ochoa JA. Mining patterns for clustering on numerical datasets using unsupervised decision trees. Knowl-Based Syst. 2015;82:70–9. doi: 10.1016/j.knosys.2015.02.019.
[22] Loyola-Gonzalez O, Gutierrez-Rodríguez AE, Medina-Pérez MA, Monroy R, Martínez-Trinidad JF, Carrasco-Ochoa JA, et al. An explainable artificial intelligence model for clustering numerical databases. IEEE Access. 2020;8:52370–84. doi: 10.1109/ACCESS.2020.2980581.
[23] Chen J. Interpretable clustering methods. PhD thesis. Northeastern University; 2018. Available from: https://www.proquest.com/docview/2116583667.
[24] Bandyapadhyay S, Fomin FV, Golovach PA, Lochet W, Purohit N, Simonov K. How to find a good explanation for clustering? Artif Intel. 2023;322:103948. doi: 10.1016/j.artint.2023.103948.
[25] Charikar M, Hu L. Near-optimal explainable k-means for all dimensions. In: Proceedings of the 2022 Annual ACM-SIAM symposium on discrete algorithms (SODA). SIAM; 2022. p. 2580–606. doi: 10.1137/1.9781611977073.101.
[26] Dasgupta S, Frost N, Moshkovitz M, Rashtchian C. Explainable k-means and k-medians clustering. In: Proceedings of the 37th International Conference on Machine Learning, Vienna, Austria; 2020. p. 12–8. Available from: https: //dl.acm.org/doi/abs/10.5555/3524938.3525592.
[27] Saisubramanian S, Galhotra S, Zilberstein S. Balancing the tradeoff between clustering value and interpretability. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society; 2020. p. 351–7. doi: 10.1145/3375627.3375843.
[28] Aggarwal CC, Reddy CK. Data clustering: algorithms and applications. 3rd ed. Boca Raton: Chapman and Hall/CRC; 2012.
[29] Jolliffe IT, Cadima J. Principal Component Analysis: A review and recent developments. Philos Trans R Soc A Math Phys Eng Sci. 2016;374(2065):20150202. doi: 10.1098/rsta.2015.0202.
[30] Bertsimas D, Orfanoudaki A, Wiberg H. Interpretable clustering: An optimization approach. Machine Learn. 2021;110:89–138. doi: 10.1007/s10994-020-05896-2.
[31] Nauta M, Trienes J, Pathak S, Nguyen E, Peters M, Schmitt Y, et al. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. ACM Comput Surveys. 2023;55(13s):1–42. doi: 10.1145/3583558.
[32] Liu B, Xia Y, Yu PS. Clustering via decision tree construction. Foundations and advances in data mining. Berlin, Heidelberg: Springer; 2005. p. 97–124. doi: 10.1007/11362197_5.
[33] Basak J, Krishnapuram R. Interpretable hierarchical clustering by constructing an unsupervised decision tree. IEEE Trans Knowl Data Eng. 2005;17(1):121–32. doi: 10.1109/TKDE.2005.11.
[34] Ghattas B, Michel P, Boyer L. Clustering nominal data using unsupervised binary decision trees: Comparisons with the state of the art methods. Pattern Recognit. 2017;67:177–85. doi: 10.1016/j.patcog.2017.01.031.
[35] Dasgupta S, Frost N, Moshkovitz M, Rashtchian C. Explainable k-means clustering: Theory and practice. In: XXAI Workshop. ICML; 2020. Available from: http://interpretable-ml.org/icml2020workshop/pdf/06.pdf.
[36] Gamlath B, Jia X, Polak A, Svensson O. Nearly-tight and oblivious algorithms for explainable clustering. Adv Neural Inform Proces Syst. 2021;34:28929–39. Available from: https://dl.acm.org/doi/abs/10.5555/3540261.3542477.
[37] Laber E, Murtinho L. On the price of explainability for some clustering problems. In: International Conference on Machine Learning. PMLR; 2021. p. 5915–25. Available from: https://proceedings.mlr.press/v139/laber21a.html.
[38] Makarychev K, Shan L. Near-optimal algorithms for explainable k-medians and k-means. In: International Conference on Machine Learning. PMLR; 2021. p. 7358–67. Available from: https://proceedings.mlr.press/v139/makarychev21a.html
[39] Esfandiari H, Mirrokni V, Narayanan S. Almost tight approximation algorithms for explainable clustering. In: Proceedings of the 2022 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA). SIAM; 2022. p. 2641–63. doi: 10.1137/1.9781611977073.103.
[40] Makarychev K, Shan L. Explainable k-means: Don't be greedy, plant bigger trees! In: Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing; 2022. p. 1629–42. doi: 10.1145/3519935.3520056.
[41] Laber E, Murtinho L. Nearly Tight Bounds on the Price of Explainability for the k-center and the Maximum-spacing clustering problems. Theoret Comput Sci. 2023;949:113744. doi: 10.1016/j.tcs.2023.113744.
[42] Makarychev K, Shan L. Random cuts are optimal for explainable k-medians. Adv Neural Inform Proces Syst. 2023;36:66890–901. Available from: https://dl.acm.org/doi/abs/10.5555/3666122.3669043.
[43] Piltaver R, Luštrek M, Gams M, Martinčić-Ipšić S, et al. What makes classification trees comprehensible? Expert Syst Appl. 2016;62:333–46. doi: 10.1016/j.eswa.2016.06.009.
[44] Laber E, Murtinho L, Oliveira F. Shallow decision trees for explainable k-means clustering. Pattern Recognit. 2023;137:109239. doi: 10.1016/j.patcog.2022.109239.
[45] Gioria L. Improving explainable clustering via probabilistic modeling, Master's Thesis. Politecnico di Milano. Milan, Italy; 2023. Available from: https://www.politesi.polimi.it/bitstream/10589/203612/4/2023_05_Gioria.pdf.
[46] Gabidolla M, Carreira-Perpinnnán MA. Optimal interpretable clustering using oblique decision trees. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining; 2022. p. 400–10. doi: 10.1145/3534678.3539361.

[47] Hwang H, Whang SE. XClusters: Explainability-first clustering. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37; 2023. p. 7962–70. doi: 10.1609/aaai.v37i7.25963.
[48] Pelleg D, Moore AW. Mixtures of rectangles: interpretable soft clustering. In: Proceedings of the Eighteenth International Conference on Machine Learning; 2001. p. 401–8. Available from: https://dl.acm.org/doi/abs/10.5555/645530.658306.
[49] Chen J, Chang Y, Hobbs B, Castaldi P, Cho M, Silverman E, et al. Interpretable clustering via discriminative rectangle mixture model. In: 2016 IEEE 16th International Conference on Data Mining (ICDM). IEEE; 2016. p. 823–8. doi: 10.1109/ICDM.2016.0097.
[50] Lawless C, Kalagnanam J, Nguyen LM, Phan D, Reddy C. Interpretable clustering via multi-polytope machines. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36; 2022. p. 7309–16. doi: 10.1609/aaai.v36i7.20693.
[51] Lawless C, Gunluk O. Cluster explanation via polyhedral descriptions. In: International Conference on Machine Learning. PMLR; 2023. p. 18652–66. Available from: https://proceedings.mlr.press/v202/lawless23a.html.
[52] Sabbatini F, Calegari R. ExACT explainable clustering: Unravelling the intricacies of cluster formation. In: Proceedings of the 2nd International Workshop on Knowledge Diversity, KoDis; 2023. p. 2–8. Available from: https://ceur-ws.org/Vol-3548/paper3.pdf.
[53] Sabbatini F, Calegari R. Explainable clustering with CREAM. In: Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning. Vol. 19. 2023. p. 593–603. doi: 10.24963/kr.2023/58.
[54] Chen X, Güttel S. Fast and explainable clustering based on sorting. arXiv preprint arXiv:220201456. 2022. doi: 10.48550/arXiv.2202.01456.
[55] Carrizosa E, Kurishchenko K, Marín A, Morales DR. Interpreting clusters via prototype optimization. Omega. 2022;107:102543. doi: 10.1016/j.omega.2021.102543.
[56] Peng X, Li Y, Tsang IW, Zhu H, Lv J, Zhou JT. XAI Beyond classification: interpretable neural clustering. J Machine Learn Res. 2022;23(6):1–28. Available from: https://www.jmlr.org/papers/v23/19-497.html.
[57] Xie X, Pu YF, Zhang H, Mańdziuk J, El-Alfy ESM, Wang J. An interpretable neural network for robustly determining the location and number of cluster centers. Int J Machine Learn Cybernet. 2024;15(4):1473–501. doi: 10.1007/s13042-023-01978-4.
[58] Xie X, Zhang H, Wang J, Chang Q, Wang J, Pal NR. Learning optimized structure of neural networks by hidden node pruning with L_1 regularization. IEEE Trans Cybernet. 2019;50(3):1333–46. doi: 10.1109/TCYB.2019.2950105.
[59] Kim B, Shah JA, Doshi-Velez F. Mind the gap: A generative approach to interpretable feature selection and extraction. Adv Neural Inform Proces Syst. 2015;28:2260–8. Available from: https://dl.acm.org/doi/abs/10.5555/2969442.2969492.
[60] Greene D, Cunningham P. Producing accurate interpretable clusters from high-dimensional data. In: European Conference on Principles of Data Mining and Knowledge Discovery. Springer; 2005. p. 486–94. doi: 10.1007/11564126_49.
[61] Plant C, Böhm C. INCONCO: Interpretable clustering of numerical and categorical objects. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2011. p. 1127–35. doi: 10.1145/2020408.2020584.
[62] Ellis CA, Sendi MS, Geenjaar E, Plis SM, Miller RL, Calhoun VD. Algorithm-agnostic explainability for unsupervised clustering. 2021. arXiv:210508053. doi: 10.48550/arXiv.2105.08053.
[63] Huang H, Xue F, Yan W, Wang T, Yoo S, Xu C. Learning associations between features and clusters: an interpretable deep clustering method. In: 2021 International Joint Conference on Neural Networks (IJCNN). IEEE; 2021. p. 1–10. doi: 10.1109/IJCNN52387.2021.9534368.
[64] Gai Y, Liu J. Clustering by sparse orthogonal NMF and interpretable neural network. Multimedia Syst. 2023;29(6):3341–56. doi: 10.1007/s00530-023-01187-7.
[65] Svirsky J, Lindenbaum O. Interpretable deep clustering for tabular data. In: Proceedings of the 41st International Conference on Machine Learning; 2024. p. 47314–30. Available from: https://proceedings.mlr.press/v235/svirsky24a.html.
[66] Morichetta A, Casas P, Mellia M. EXPLAIN-IT: Towards explainable AI for unsupervised network traffic analysis. In: Proceedings of the 3rd ACM CoNEXT Workshop on Big DAta, Machine Learning and Artificial Intelligence for Data Communication Networks. 2019. p. 22–8. doi: 10.1145/3359992.3366639.
[67] Horel E, Giesecke K, Storchan V, Chittar N. Explainable clustering and application to wealth management compliance. In: Proceedings of the First ACM International Conference on AI in Finance; 2020. p. 1–6. doi: 10.1145/3383455.3422530.
[68] Horel E, Giesecke K. Computationally efficient feature significance and importance for machine learning models. 2019. arXiv preprint arXiv:190509849. 2019. doi: 10.48550/arXiv.1905.09849.
[69] Kauffmann J, Esders M, Ruff L, Montavon G, Samek W, Müller KR. From clustering to cluster explanations via neural networks. IEEE Trans Neural Net Learn Syst. 2022;35(2):1926–40. doi: 10.1109/TNNLS.2022.3185901.
[70] Bach S, Binder A, Montavon G, Klauschen F, Müller KR, Samek W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE. 2015;10(7):e0130140. doi: 10.1371/journal.pone.0130140.
[71] Alvarez-Garcia M, Ibar-Alonso R, Arenas-Parra M. A comprehensive framework for explainable cluster analysis. Inform Sci. 2024;663:120282. doi: 10.1016/j.ins.2024.120282.
[72] Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nature Machine Intel. 2020;2(1):56–67. doi: 10.1038/s42256-019-0138-9.
[73] Wei X, Zhang Z, Huang H, Zhou Y. An overview on deep clustering. Neurocomputing. 2024;590:127761. doi: 10.1016/j.neucom.2024.127761.
[74] Zhou J, Gandomi AH, Chen F, Holzinger A. Evaluating the quality of machine learning explanations: a survey on methods and metrics. Electronics. 2021;10(5):593. doi: 10.3390/electronics10050593.
[75] Doshi-Velez F, Kim B. Considerations for evaluation and generalization in interpretable machine learning. In Explainable and interpretable models in computer vision and machine learning. Cham: Springer; 2018. p. 3–17. doi: 10.1007/978-3-319-98131-4_1.

[76] Swartout WR, Moore JD. Explanation in second generation expert systems. In: Proceedings of the Second Generation Expert Systems. Springer; 1993. p. 543–85. doi: 10.1007/978-3-642-77927-5_24.
[77] Das A, Rad P. Opportunities and challenges in explainable artificial intelligence (XAI): A survey. arXiv preprint arXiv:200611371. 2020. doi: 10.48550/arXiv.2006.11371.
[78] Bandyapadhyay S, Fomin F, Golovach PA, Lochet W, Purohit N, Simonov K. How to find a good explanation for clustering? In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36; 2022. p. 3904–12. doi: 10.1016/j.artint.2023.103948.
[79] Caron M, Bojanowski P, Joulin A, Douze M. Deep clustering for unsupervised learning of visual features. In: Proceedings of the European Conference on Computer Vision (ECCV); 2018. p. 132–49. doi: 10.1007/978-3-030-01264-9_9.
[80] Li S, Yuan M, Chen J, Huuu Z. AdaDC: Adaptive deep clustering for unsupervised domain adaptation in person re-identification. IEEE Trans Circuits Syst Video Tech. 2021;32(6):3825–38. doi: 10.1109/TCSVT.2021.3118060.
[81] Setzu M, Guidotti R, Monreale A, Turini F, Pedreschi D, Giannotti F. GlocalX - from local to global explanations of black box AI models. Artificial Intelligence. 2021;294:103457. doi: 10.1016/j.artint.2021.103457.
[82] Wu X, Zhao H, Zhu Y, Shi Y, Yang F, Liu T, et al. Usable XAI: 10 strategies towards exploiting explainability in the LLM era. arXiv preprint arXiv:240308946. 2024. doi: 10.48550/arXiv.2403.08946.
[83] Malizia A, Paternò F. Why is the current XAI not meeting the expectations? Commun ACM. 2023;66(12):20–3. doi: 10.1145/3588313.
[84] Cambria E, Malandri L, Mercorio F, Nobani N, Seveso A. XAI meets LLMs: A Survey of the relation between explainable AI and large language models. arXiv preprint arXiv:240715248. 2024. doi: 10.48550/arXiv.2407.15248.
[85] Ali T, Kostakos P. HuntGPT: Integrating machine learning-based anomaly detection and explainable AI with large language models (LLMs). arXiv preprint arXiv:230916021. 2023. doi: 10.48550/arXiv.2309.16021.
[86] Zhang Y, Wang Z, Shang J. ClusterLLM: Large language models as a guide for text clustering. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; 2023. p. 13903–20. doi: 10.18653/v1/2023.emnlp-main.858.
[87] Petukhova A, Matos-Carvalho JP, Fachada N. Text clustering with LLM embeddings. arXiv preprint arXiv:240315112. 2024. doi: 10.48550/arXiv.2403.15112.
[88] Tipirneni S, Adkathimar R, Choudhary N, Hiranandani G, Amjad RA, Ioannidis VN, et al. Context-aware clustering using large language models. arXiv preprint arXiv:240500988. 2024. doi: 10.48550/arXiv.2405.00988.
[89] Viswanathan V, Gashteovski K, Lawrence C, Wu T, Neubig G. Large language models enable few-shot clustering. Trans Assoc Comput Linguistics. 2024;12:321–33. doi: 10.1162/tacl_a_00648.
[90] Chamola V, Hassija V, Sulthana AR, Ghosh D, Dhingra D, Sikdar B. A review of trustworthy and explainable artificial intelligence (XAI). IEEE Access. 2023;11:78994–9015. doi: 10.1109/ACCESS.2023.3294569.
[91] Hassija V, Chamola V, Mahapatra A, Singal A, Goel D, Huang K, et al. Interpreting black-box models: a review on explainable artificial intelligence. Cognitive Comput. 2024;16(1):45–74. doi: 10.1007/s12559-023-10179-8.
[92] Schwalbe G, Finzel B. A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. Data Mining Knowl Discovery. 2024;38:3043–101. doi: 10.1007/s10618-022-00867-8.
[93] Li Z, Zhu Y, Van Leeuwen M. A survey on explainable anomaly detection. ACM Trans Knowl Discovery Data. 2023;18(1):1–54. doi: 10.1145/3609333.
[94] Yepmo V, Smits G, Pivert O. Anomaly explanation: a review. Data Knowl Eng. 2022;137:101946. doi: 10.1016/j.datak.2021.101946.
[95] Ikotun AM, Ezugwu AE, Abualigah L, Abuhaija B, Heming J. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. Inform Sci. 2023;622:178–210. doi: 10.1016/j.ins.2022.11.139.
[96] Yang H, Jiao L, Pan Q. A survey on interpretable clustering. In:2021 40th Chinese Control Conference (CCC). IEEE; 2021. p. 7384–8. doi: 10.23919/CCC52363.2021.9549986.