Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Subspace Clustering for High-Dimensional Data: A Survey of Methods, Challenges, and Conceptual Frameworks for Future Research

Vadicherla Raju, Dr. K. P. Supreethi
Research Scholar,
Department of Computer Science & Engineering
Jawaharlal Nehru Technological University, Hyderabad, India
vadicherlaraju@gmail.com
Professor and Head
Department of Computer Science & Engineering
Jawaharlal Nehru Technological University, Hyderabad,India
supreethi.pujari@jntuh.ac.in

ARTICLE INFO

Received: 18 Dec 2024
Revised: 15 Feb 2025
Accepted: 28 Feb 2025

ABSTRACT

Subspace clustering has emerged as a powerful paradigm for analyzing high-dimensional data, where traditional clustering methods struggle due to the curse of dimensionality. By identifying clusters in relevant subsets of dimensions, subspace clustering enhances interpretability, scalability, and robustness in various applications such as bioinformatics, image processing, and IoT. This paper presents a comprehensive survey of existing subspace clustering methods, categorizing them into grid-based, model-based, spectral, and hybrid approaches. We introduce a new taxonomy framework for classifying subspace clustering techniques based on scalability, noise tolerance, and application domains. Additionally, we highlight recent advancements, including deep learning-based subspace clustering, fairness-aware clustering, and real-time streaming data applications. The paper also discusses key challenges such as interpretability, computational complexity, and lack of standardized evaluation metrics, providing insights into future research directions. This survey aims to serve as a roadmap for researchers by consolidating the latest developments and identifying open challenges in subspace clustering.

Keywords: Subspace Clustering, High Dimensional data, Clustering

INTRODUCTION

High-dimensional datasets are present in most fields: bioinformatics and social networks, financial markets, and image processing are only some. Traditional clustering methods[1] usually fail to analyze such data very well due to the curse of dimensionality, and therefore subspace clustering emerges as an important tool for meaningful pattern extraction. This paper discusses subspace clustering's contemporary techniques, applications, challenges, and future directions. The discussion proceeds in the successive sections with a detailed discussion on these aspects, starting with the motivation behind subspace clustering.

Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

1.1 Motivation

In the field of bioinformatics, finance, social network analysis, and image processing, clustering constitutes an important technique of unsupervised learning. As the dimension of data increases, the traditional methods of clustering fail because of the curse of dimensionality In high-dimensional spaces, the distances between points lose significance and result in clustering with very high inaccuracies and noise and unnecessarily high computation costs. These difficulties pose major challenges to clustering large-scale, high-dimensional datasets. Subspace clustering came to be considered an alternative that identifies clusters in some feature subspaces, improving its accuracy and efficiency. This method is useful where clusters exist only in subsets of dimensions, The ever-increasing applications of IoT, cybersecurity, and personalized medicine to generate high-dimensional data have rendered subspace clustering an indispensable tool for uncovering hidden patterns in massive datasets[2].

1.2 Overview of Clustering in High-Dimensional Data

Clustering, one of the essential unsupervised learning methods, is applied for grouping the similar data points on certain similarity measures. Clustering algorithms work wonders in low-dimensional spaces, and they easily cluster data points into separate clusters. However, as the number of dimensions increases, conventional clustering
techniques suffer from the curse of dimensionality: distances between data points become increasingly meaningless, and usage costs grow exponentially[3].High-dimensional data is common in almost all fields, such as bioinformatics
(gene expression data), text mining, image processing, and IoT, thus requiring different approaches of clustering to manage complex structures.

1.3 Traditional Clustering Concepts and Their Disadvantages

Many classical clustering techniques have been applied to various applications, such as:
k-means Clustering: Partitions data points into k clusters by minimizing the variance within clusters. This procedure is time-efficient but invalid for non-spherical clusters or high-dimensional datasets .
Hierarchical Clustering: Constructs a tree-like structure (dendrogram) based on similarity measures. It gives a flexible structure, but it suffers a lot from huge computation complexity on large-scale data sets[4].
DBSCAN (Density-Based Spatial Clustering of Applications with Noise): Discovers density regions, thereby giving added robustness to noise inputs but becomes incapable to work effectively in high-dimensional spaces because of sparse data distribution, affecting density estimation.These types of conventional clustering work well in low dimensions; as dimensions increase, the effectiveness is lost, prompting migration to different approaches such as subspace clustering.

Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

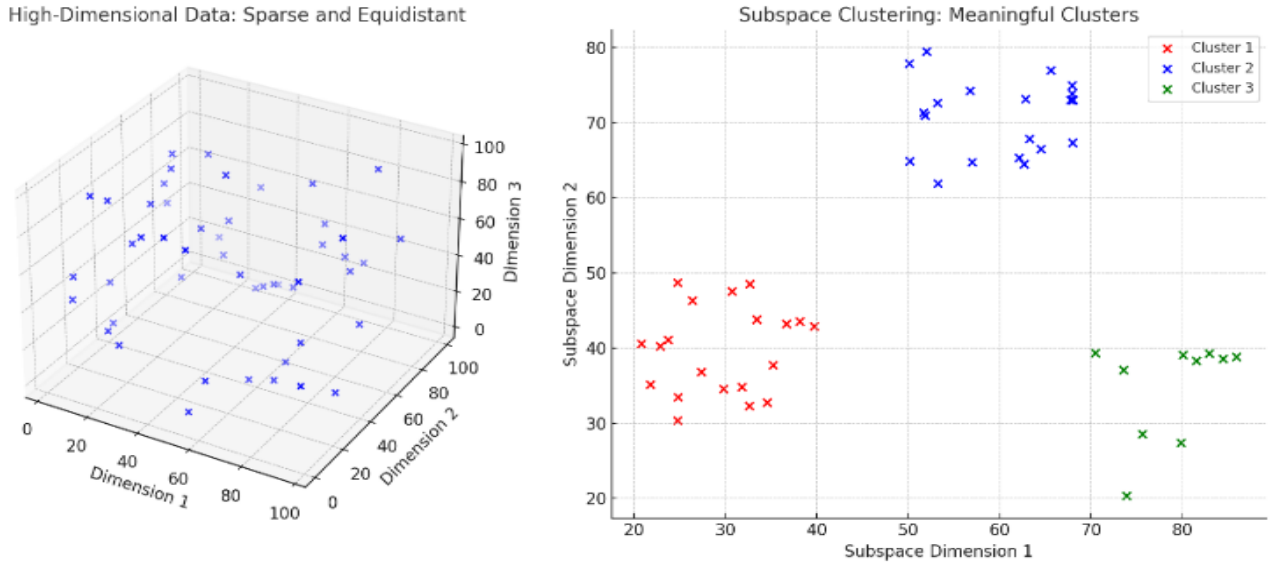

Figure 1: visualization of Curse of Dimensionality and Subspace Clustering

The above figure 1 Left Pane: Illustrates where denser and equidistant points in higher dimensions would turn out when traditional cluster techniques fail. Right Pane: Demonstrates that meaningful clusters in relevant dimensions are taken forward by subspace clustering against the dimensionality curse.

1.4 Introduction to Subspace Clustering

The purpose of subspace clustering is to identify clusters in certain feature subspaces rather than considering the entire feature space and thus alleviating the deficiencies of traditional clustering. This way, only those significant dimensions which can give rise to meaningful clusters shown in figure 2 are included for evaluation purposes, thus improving the clustering accuracy.

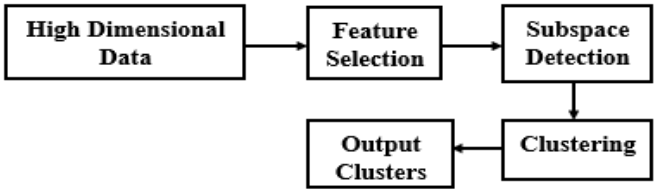

Figure 2: Overview of Subspace Clustering Workflow

The beneficial elements of subspace clustering are:

Improved interpretability: With respect to real-world patterns, the derived clusters are now more valuable, since the relevant subspaces are selected.

Computational efficiency: Reducing the number of dimensions helps clustering in a more efficient manner while handling high-dimensional datasets.

Better noise handling: Subspace clustering methods are generally more robust to irrelevant feature and noisy dimensions .

Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Feature	Traditional Clustering	Subspace Clustering
Feature Selection	Uses all dimensions	Selects relevant subspaces
Noise Handling	Sensitive to noise	Robust to irrelevant dimensions
Scalability	Struggles with high-dimensional data	More efficient with selected subspaces

Table 1: Traditional Clustering vs Subspace Clustering

1.5 Challenges in Subspace Clustering

Despite the several advantages that subspace clustering approaches offer, they come with certain disadvantages shown in below figure 3

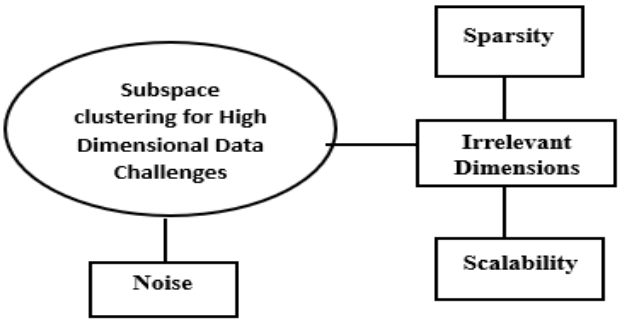

Figure 3: Subspace clustering for high dimensional data challenges

scalability issues are among the major weaknesses of existing subspace clustering approaches as Sparse Subspace Clustering (SSC) and Low-Rank Representation (LRR) for instance use convex optimization and matrix decomposition as their backbone techniques making them severely computational intensive and thus do not scale to large data
Noise Sensitivity: High-dimensional data often involves a fair amount of irrelevant or noisy features, which can detrimental to clustering performances .
Parameter Sensitivity: A good many subspace clustering algorithms do require some kind of manual parameter tinkering, which keeps them from being adaptable to a varied set of datasets.
Computational Complexity: Certain approaches may involve an iterative computation of the similarity matrix, which suffices to become time-consuming for large datasets [5].

1.6 Research Gaps in Subspace Clustering

Algorithms for subspace clustering have developed substantially, and yet various research gaps still remain, inhibiting the real-world applicability of the techniques. Research gaps in this area include:
Scalability Issues: Many subspace clustering algorithms, especially spectral ones, face heavy computational demands, such as with eigenvalue decomposition and convex optimization; for example, Sparse Subspace Clustering (SSC) and Low-Rank Representation (LRR) . These

Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

methods therefore have great difficulty scaling to larger datasets and need to be complemented with scalable alternatives.
Sensitivity to Noise and Robustness: Clustering performance is heavily impacted by noise and irrelevant features present in high-dimensional data. This means that existing methods may not effectively denoise the data or enhance the robustness against feature redundancy and inconsistencies[6].
Lack of Interpretability: Many of these deployed deep learning subspace clustering models are black boxes; hence, there are great difficulties explaining why certain subspaces were recommended. By their very nature, these limitations of explainability direct such models unheeded toward application in significantly high-stakes domains, especially within health and finance domains.
Equality and Bias Problems: Recent work has shown that bias exists with respect to subspace selection, possibly resulting in unfair clustering outputs in domains such as medical diagnostics and hiring. In order to ensure fairness, bias-mitigation approaches must be implemented inside subspace clustering mode
No Standardized Evaluation Metrics: Unlike in conventional clustering, there are no universally accepted benchmark datasets and evaluation metrics for subspace clustering. Researchers are often utilizing inconsistent datasets, such as MNIST, 20 Newsgroups, and gene expression data, thereby severely hampering standardized comparisons among different methods
Limited Techniques for Real-Time and Adaptive: Most of the existing subspace clustering methods presume the static nature of the data. However, areas of application such as fraud detection, IoT, and financial market analysis require adaptive and real time clustering to be able to interpret continuously streaming data[7] .
The importance of focusing on subspace clustering uncovered by research is necessary to improve understandings, ensure that these methods scale up to big data online, real-time analytics and interpretable AI systems.

1.7 Objectives and Contributions of This Study

This exhaustive survey deals with state-of-the-art subspace clustering techniques, main challenges, and future research directions. The primary aims and contributions of this study are:
This study enunciates the taxonomy and subspace clustering techniques. The techniques presented so far are classified into grid-based, model-based, spectral, and hybrid techniques wherein a comprehensive discussion is made on the advantages and limitations of those techniques.
Comparative Evaluation of Existing Techniques: We evaluate the various techniques of subspace clustering in terms of scalability, noise robustness, computational complexity, and clustering accuracy so as to make a structured performance comparison.
Recognition of Research Gaps and Challenges: This study addresses some basic limitations in the existing subspace clustering techniques, more especially with respect to scalability, fairness, real-time adaptability, and interpretability.
Emerging Trends and Future Directions of Research: This paper discusses the latest trends presently emerging such as deep learning-based clustering, fairness-aware clustering, and

Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

self-supervised learning to outline how these trends are changing the face of subspace clustering.
It allows identifying possible research opportunities like self-supervised learning, fairness-aware clustering, and stable real-time adaptive clustering frameworks to better the efficiency and then applicability of the subspace clustering model as per key challenges of the framework. This work can serve as a useful resource for both researchers and practitioners who wish to gain insight into subspace clustering techniques, challenges, and future directions.

1.8 Organization of the Paper

The structure of the paper is as such: Section 2 (Background) describes the associated high-dimensional data challenges, relevant subspace learning methods, adaptability measures, elementary concepts of subspace clustering, and computational complexity issues. Section 3 (Related Work) sheds light on some previous research contributions in subspace clustering, pinpointing some of the key break-throughs, strategies and limitations of current techniques. Section 4 (Subspace Clustering Techniques) describes and classifies different methods in subspace clustering such as grid-based, model-based, spectral, hybrid, among others, and their working mechanisms, strengths and weaknesses. Section 5 (Comparative Analysis of Subspace Clustering Methods: Presents Comparative Evaluation of Different Subspace Cluster Methods based on Scalability, Noise Tolerance, Accuracy and Computational Efficiency. During the discussion of Section 6 (Applications of Subspace Clustering), some real-world applications are found within the categories of bioinformatics, image processing, cybersecurity, IoT, and financial analytics. The challenges and research gaps in subspace clustering are further analyzed in Section 7, and limitations of existing methods come into play here, such as issues of scalability, fairness, and interpretability as well as the adaptability in real time." Challenges conceptualized in Section 8. Proposed Conceptual Frameworks to Address Research Gaps presents suggested solutions and conceptual frameworks that could potentially strengthen subspace clustering techniques from the viewpoint of challenges identified. In Section 9 (Emerging Trends and Future Directions), a massive number of current research trends and developments, including deep learning-based clustering, explainable AI, fairness-aware clustering, and real-time adaptations, are provided. Synthesis mission of the paper in Section 10 (Conclusion) ends the paper, discussing some possibilities for the future concerning subspace clustering-the last Section 11 (References) consists of cited literature and research papers in support of the discussions and findings presented in this study

2.BACKGROUND

2.1 Curse of Dimensionality and Its Impact on Clustering

High-dimensional data exposes some very basic challenges and thus aspects that are highly damaging to clustering performance. Thus, the increase in dimension brings with it a physical constraint whereby the feature space becomes increasingly and inversely sparse and renders distance measures less and less meaningful between points [8].

Key effects include: Distance Concentration: The difference between the farthest separated points and a close pair becomes minor with ever-increasing dimensionality, thus making distance-based similarity lower in discriminative power.

Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Sparsity Problems: On the other hand, very sparse data with respect to high dimension bump may be available from dense regions, where some are incorporated in very small or much-less visible clusters
Computational Inefficiency: Usually clustering techniques, such as k-means and others like DBSCAN, in a high-dimensional space become too computationally expensive to be used in practice with respect to targeted datasets .
All these factors highly promoted the need for developing methods of subspace clustering, which selectively give consideration to only some pertinent dimensions instead of the entire feature space.

2.2 Subspace Learning and Feature Selection

A feature selection subspace learning involves understanding some attributes of the data due to its high dimensionality
Dimensionality Reduction Techniques:
In these cases, the components that are orthogonal and account for maximum variance in the data are extracted
t-SNE: Maps high-dimensional data into a low-dimensional space for its visualization[9] .
Autoencoders: Use deep neural networks for encoding data to a compressed latent representation
Feature Selection versus Feature Extraction:
Feature Selection: identifies the relevant dimensions and suppresses redundant or irrelevant features
Feature Extraction: Transforms original features to a new reduced representation usually through an algorithm like PCA or deep learning
Subspace clustering is a way to use these techniques to cluster data into lower dimensional representations, which yield better accuracy and efficiency of operation.

2.3 Similarity Measures for High-Dimensional Data

The selection of similarity measures is fundamental for the effectiveness of these clustering algorithms in high-dimensional space:
Euclidean Distance: Most Common Measure; considered to lose effectiveness in higher dimensions because of concentration of distances
Cosine Similarity: Measures the cosine of the angle between two vectors, making it more appropriate for sparse and high-dimensional
Mahalanobis Distance: The distance is also correlated with the variables rather than absolute distances, thus being effective in cluster formation in correlated subspaces [10]
Correlation-Based Measures: Capture relationships between variables rather than by absolute distances and thus improve the clustering accuracy of feature-rich datasets.

2.4 Foundations of Subspace Clustering

Subspace clustering projects the power of traditional cluster identification from the unitary space to different possible subspaces. The mathematical formulation of subspace clustering is as follows:

Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Cluster Representation: A dataset $X \in R^{nimssed}$ is partitioned into the k clusters, where each cluster exists in a subset of dimensions $Si \subseteq \{ 1,2,...d \}$
Objective function:

$$\min_{\{Ci,Si\}} \sum_{i=1}^{k} \sum_{xi \in C} \sum_{l \in S} ||xj,l - \mu i,l||2,$$

where C_i represents clusters and S_i represents relevant subspaces.

2.5 Computational Complexity of Subspace Clustering Algorithms

Computational complexity is a major challenge in subspace clustering:
Sparse Subspace Clustering (SSC) is an incredibly expensive method in computation as it has to solve a number of convex optimization problems which make it fall into the complexity of $O(n^3)$
Low-Rank Representation (LRR) is based on matrix decomposition techniques and has a complexity $O(n^3)$ that prevents it from scaling [11]
Model- and Grid-Based Methods generally have a lower complexity but this benefit comes at the expense of accuracy.
Efficiency in computation must be enunciated to produce any scaled results in subspace clustering when dealing with large data sets.

3 RELATED WORK

The subspace clustering study of high-dimensional data has progressed extensively over the last few decades. This section constitutes a chronological review of some of the key contributions highlighting the developments and new trends in this area.

3.1 Early Developments (1998 – 2005)

CLIQUE is one of the earliest grid-based subspace clustering algorithms. It partitions the data space into grid cells of equal size and declares dense regions as clusters. CLIQUE can efficiently find clusters in differing subspaces, but it comes with a heavy computational cost[12].
ENCLUS adopted an entropy-based feature selection mechanism to improve the clustering quality of CLIQUE. Instead of using subspace evaluation scores to infer the quality of various subspaces, ENCLUS integrates with the evaluation of entropy as effective means for aggregation. It evaluates the quality of subspaces for clustering so that relevant subspaces apply in clustering. Still, it is not scalable for highly dimensional datasets[13].
MAFIA brought improvements to grid-based clustering through adaptive grid refinement that is able to discover clusters in a more efficient manner. MAFIA, on the other hand, varies the grid size from one region in opposition to another based on the density of the attributed data unlike CLIQUE. It would have worked better in discovering clusters of varying densities. However, it has a disadvantage of depending on pre-defined parameters, and it is sensitive to noise[14]
A model-based approach that extends k-medoid clustering is PROCLUS. A step towards improving dimensionality selection was achieved with this. PROCLUS works by iteratively

Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

optimizing cluster medoids while choosing a subset of dimensions that are the most pertinent for each cluster. Although PROCLUS offers better scalability compared with grid-based approaches[15], it requires the prior specification of an optimal number of clusters and dimensions, thus limiting it in unsupervised situations.
ORCLUS improved PROCLUS using PCA-based feature selection for better cluster formation. ORCLUS basically projects the data points dynamically into lower dimensional subspaces on the fly thus increasing cluster separation and robustness to noise. The multiple PCA operations involved are very costly in computation, thus it is not appropriate for huge datasets.

3.2Advances in Model-Based and Spectral Clustering (2005 – 2015)

Sparse Subspace Clustering (SSC), introduced by Elhamifar and Vidal in 2013, is a sparse representation that is used to group data into low-dimensional subspaces and improve cluster accuracy in datasets with noise. SSC assumes sparse representations with respect to each other for data points belonging to the same subspace. Even though it is one of the ideal methodologies for effective and accurate capturing of complex subspace structures, it is computationally expensive, particularly for large datasets. This is primarily because of its heavy reliance on convex optimization techniques for creating effects[12]
Low-Rank Representation, or LRR for short was basically an extension of SSC to invoke low-rank constraints combined with the strength of anything that it offered in order to deepen the performance of clustering for larger datasets. High dimensional data can have low rank under certain assumptions, which makes LRR particularly suitable for structured data such as images and videos. So it makes data more robust to noise and outlier effects, but at the same time, it has a high memory and computational cost, thus maximizing the limited applications possible under this paradigm for large scale scenarios.
Combining Density-Based Clustering with Subspace clustering. meant capturing density-based clustering and subspace features, thus making the technique more
robust to noisy data. HDBSCAN is built on DBSCAN but allows for the modelling of varying density clusters and better promotes hierarchical clustering. This subspace clustering integration is intended to find clusters in relevant factor subspaces while providing flexibility across the data distributions. Parameter tuning remains a little bit of a headache to make it suitable for different datasets.

3.3 Deep Learning and Hybrid Approaches (2015 – Present

Deep Subspace Clustering Networks (DSC-Net): Use of autoencoders has successfully enhanced high-dimensional performance with the application of spectral clustering after the latent feature extraction. Deep Subspace Clustering Networks (DSC-Net) using those characteristics may minimize the drawbacks associated with traditional spectral clustering methods by jointly optimizing feature learning and clustering. But it makes hyperparameter tuning complex and requires large labeled training datasets for good outcomes, as it is a deep neural network based system

Fairness-aware clustering used bias-mitigation technique in subspace clustering for equal feature selection. This element is one example of entering algorithmic bias in high-dimensional clustering approaches by embedding fairness to objective clustering. While obtaining sugar reductions depends on extra costly computational workload and a tougher definition of fairness from case to case, these remain open challenges[16].

Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Real-time Subspace Clustering for IoT Data Streams: It's proposed that dynamic subspace clustering techniques be applied such that one can utilize them in the efficient processing of real-time streaming data. Incremental learning and adaptive subspace selection techniques would make it possible to cluster evolving data streams in this dynamic mode of operation. All these could be made possible, but constraint implementation and drift detection were the greatest hurdles towards the accomplishment of real time requirement in IoT[17].

Transformer-Based Subspace Clustering: Using the Transformer networks for adaptive feature selection, the method improves interpretability and scalability in clustering high-dimensional data. In such a case, the self-attention mechanisms are employed to identify the relevant, salient subspaces dynamically, thus reducing feature redundancy. However, there are high computational expenses in transformer-based models which require further optimization for large-scale data processing[18].Compared to earlier times, new advancements have been made in subspace clustering; however, several problems still exist, barring its wide acceptance in large-scale applications. The foremost areas of research need to advance, and limitations exist in:

4.SUBSPACE CLUSTERING TECHNIQUES

Subspace clustering approaches are classified and analyzed in four broad categories in terms of their methodologies: Grid-Based Methods, Model-Based Methods, Spectral Methods, and Hybrid Methods are shown in figure 4 and Each category is analyzed for its strengths, weaknesses, and practical applications

.

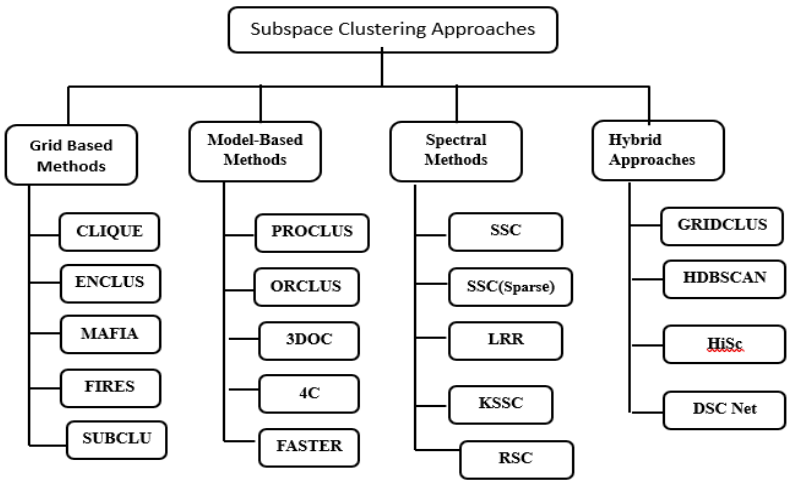

Figure 4: Subspace Clustering Methods

4.1 Grid-Based Methods

The entire point of these techniques is to bundle the data space into a multi-dimensional grid and identify clusters there within denser parts, discretize features into non-overlapping bins, and finally find clusters based on density thresholds. The basic operation is by dividing the data space into fixed-size grid-cells; where grid-cells hold a certain number of points, those grid-cells are termed dense regions. Then clusters are formed by merging the adjacent dense

Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

cells. Figure 5 shows working mechanism of each of the Grid-Based Subspace Clustering methods.

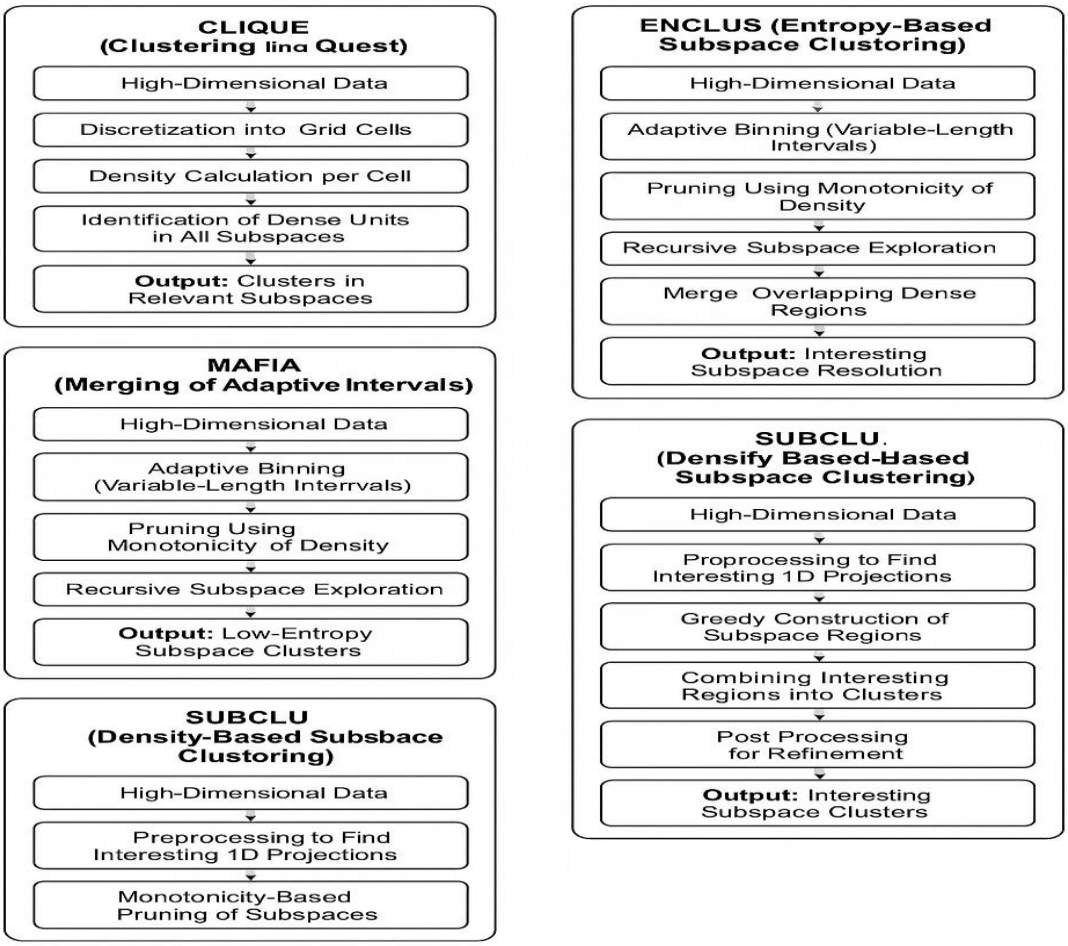

Figure 5: working flow of each of the Grid-Based Subspace Clustering methods

4.1.1 CLIQUE (Clustering in Quest)

CLIQUE (Clustering in Quest) is a grid-based clustering algorithm that discretizes data space into non-overlapping cells, detecting clusters in every subspace on the basis of point density , this approach uses the grid-based partitioning strategy, it proves to be very efficient on large datasets and has moreover the advantage of automatically determining relevant subspaces for clustering. However, the major drawback of CLIQUE is its fixed grid size with which it has difficulty catching the irregular shaped clusters, and it is also sensitive to noise in the data.

4.1.2 ENCLUS (Entropy-Based Subspace Clustering)

Enhance entropy as a criterion for subspace quality in comparison to CLIQUE, where the lower entropy value indicates a greater potential for clustering. Thus, ENCLUS can easily adapt to discover subspaces with meaningful clusters, and hence this is particularly relevant for high-dimensional data. However, parameter tuning is of utmost importance, and secondly, since

Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

ENCLUS is very expensive in its computations with increasing dimension, typical tests are run on various data sets to check its performance on clusters with varying densities and shapes

4.1.3 MAFIA (Merging of Adaptive Intervals)

MAFIA-Merging of Adaptive Intervals-is an enhanced version of CLIQUE which dynamically changes grid size within the intervals according to the density of the data, hence aiding in cluster identification more accurately by merging dense regions across dimensions and intervals [14]. The adaptive nature makes MAFIA quite robust in the discovery of clusters, especially clusters of varying densities, whereas CLIQUE mostly uses the same search strategy for all clusters irrespective of their sizes and densities. This becomes particularly useful for complex data sets containing uneven distributions. Too much overlap among clusters and too much noise may hamper cluster separations. Mafia remains the main contender now being examined for various data sets to test its scalability and robustness for clusterability of different densities and structures[12].

4.1.4 FIRES (Finding Interesting Regions in Subspaces)

FIRES can be described as a subspace clustering algorithm that employs an adaptively constructed grid coupled with entropy-based pruning for cluster discovery. Rather than dividing the space into uniformly sized portions, like CLIQUE it allows for dynamic adaption of the granularity of the grid. Adaptation is particularly recommended for datasets

that do not exhibit homogenous density, as the grid can further adapt to improve its structure based on local distribution of data points. The drawback of this adaptability is heavier computation time because continuous refinement of the different granularity grid requires more processing effort. Typically, FIRES is tested on databases to evaluate whether the approach effectively deals with clusters of different densities and distributions[20].

4.1.5 SUBCLU (Density-Based Subspace Clustering)

SUBCLU uses density-based clustering techniques to extend the DBSCAN algorithm into the subspaces, allowing it to find dense regions inside subsets of dimensions [21]. Using grid-based methods, SUBCLU can thus easily find clusters of arbitrary shapes, which is very useful for heavier and complex high-dimensional data sets. The computational cost, however, is very high since the density estimation needs to be done in multiple subspaces. This makes SUBCLU a good candidate for testing the performance of algorithms in clustering data of different densities and structures.

4.2 Model-Based Methods

Model-based clustering techniques assume that data points within a cluster follow some distribution model-a Gaussian mixture or k-medoids-and have further characterized their application through probabilistic or statistical techniques for more accurate results in clustering. The ways are to formulate some probabilistic model for the underlying data, assigning points to clusters based on an estimation of the likelihood, and iteratively adjusting the parameters to maximize performance for the clustering. working mechanism of each of the Model-Based Subspace Clustering methods shown in figure 6. Some important methods that have been introduced in this frameworks are as follows:

Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

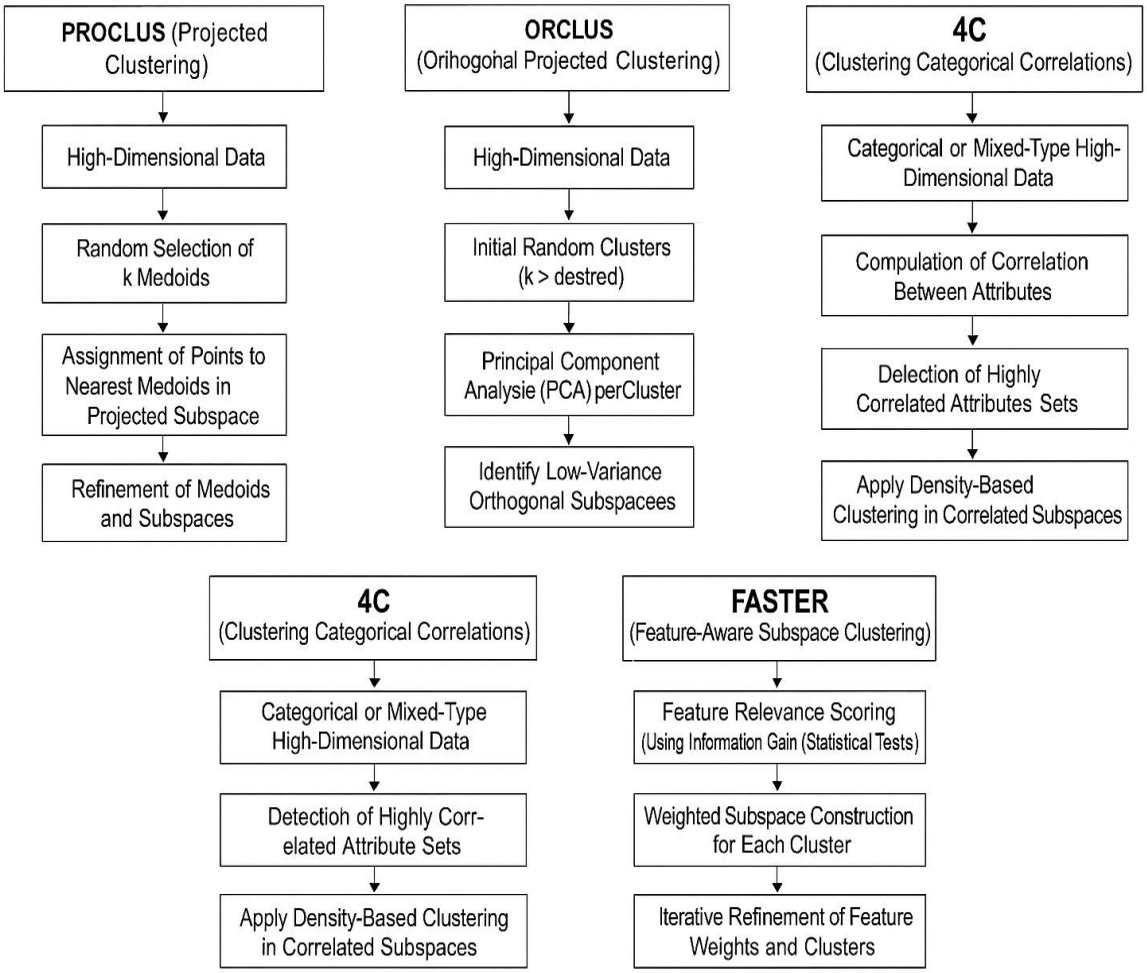

Figure 6: working flow of each of the Model-Based Subspace Clustering methods

4.2.1 PROCLUS (Projected Clustering)

The PROCLUS (Projected Clustering algorithm) is k-medoid-based subspace clustering, which iteratively selects medoids and the most relevant dimensions for each cluster [15]. Its positive features include its ability to model data point distribution in subspaces and thus to handle noisy data sets very well. The algorithm performs of clusters with different densities and with different sizes, providing flexibility for selecting relevant dimensions for each cluster. On the down side, PROCLUS is computationally intensive due to the iterative process of medoid selection and optimization, which makes its application on very large data sets virtually impossible because of the memory-and-time considerations. PROCLUS is also sensitive to parameter setting, mainly concerning the number of clusters and dimensions. In order to assess its efficiency of subspace clustering, PROCLUS is evaluated extensively on synthetic and real datasets

Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

4.2.2 ORCLUS (Orthogonal Projected Clustering)

ORCLUS that means orthogonal projected clustering is an extension of PROCLUS, which is orthogonal transformations for dimensionality reduction clustering, and iteratively selecting dimensions that maximize the compactness of clusters[22]. This makes ORCLUS very capable for discovering overlapping clusters in subspaces, but at the same time makes the selection of the subspace more focused on preserving essential cluster structures. Still, the computational requirement for this algorithm is too heavy for large datasets because of the iterative optimization processing and requires applied predefined parameters like where the number of clusters affects the clustering performances when set improperly. Recently, it was common to use ORCLUS to evaluate the detection of arbitrarily oriented subspace clusters by applying synthetic datasets.

4.2.3 DOC (Density-Based Optimal Projected Clustering)

In DOC (Density-Based Optimal Projected Clustering), clusters are located in the subspaces, maximizing the density of the points projected along user-defined dimensions. The subspace search is achieved through an efficient Monte Carlo sampling-based approach [24].This technique copes well with high noise levels, thereby rendering it appropriate for sparse data sets where classical clustering methodologies fail. Because of its extensive sampling, though, the actual computational cost incurred by DOC becomes substantial for large data sets since the search for optimal projections involves intensive iterations; thus, it is usually expensive. DOC is most often evaluated using synthetic datasets to analyze its capability of detecting clusters under various noise and sparsity conditions.

4.2.4 4C (Clustering Categorical Correlations)

4C extends subspace clustering to categorical data by integrating correlation analysis and clustering. That is, it can determine clusters based on correlations from patterns in sub-dimensions of the data[23]. Moreover, this makes 4C quite effective in processing mixed-type data sets, since it identifies relationships in both numeric and categorical attributes. Albeit, 4C has demonstrated that it can not be scaled up for high-dimensional data because clustering based on correlations will become computationally expensive as the number of attributes increases. It is also a very effective method to handle noise in categorical data and therefore suitable for areas with inconsistencies or missing values. 4C is often tested on synthetic data to determine the capability of detection of correlated clusters by the method.

4.2.5 FASTER (Feature-Aware Subspace Clustering)

FASTER proposes a rapid and model-based solution to the problem of subspace clustering by incorporating feature selection and clustering into a single task; by optimizing the relevant subspaces for clustering, one may balance against the criterion for cluster compactness[25]. The special merit of FASTER is its scalability into large datasets while preserving algorithmic efficiency for dimensionality reduction during clustering. However, parameter tuning is quite indispensable for an enabling performance, while the FASTER model performs poorly where cluster overlap occurs with potentially complicating feature interactions; hence, it is mostly tested always on synthetic datasets for performance evaluation on high-dimensional clustering tasks.

Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

4.3 Spectral Methods

The use of spectral clustering techniques is basically graph-based methods for clustering data such that it will use eigenvalue decomposition of the data to find the subspace most suitable for the clustering. This is done by first constructing a graph of similarities with nodes representing the data points, then computes the Laplacian matrix with its eigenvalues and eigenvectors, and finally partitions the graph into clusters using k-means or hierarchical techniques. The working mechanism of spectral methods shown in figure 7.There are several important methods introduced in this field given below:

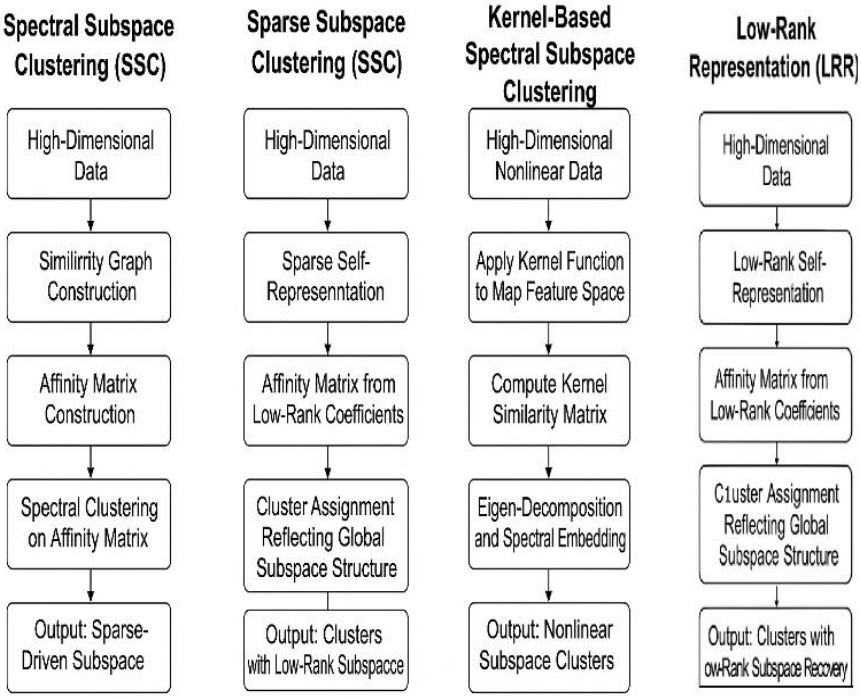

Figure 7: working flow of each of the Spectral Methods

4.3.1 Spectral Subspace Clustering (SSC)

SSC is one such basic spectral method that employs eigenvector analysis in constructing a similarity graph that can capture the data point relationships, making it very useful in discovering global structures in subspace clustering[27].This approach is suitable for capturing the non-linear structure of clusters as well as overlapping clusters. Unfortunately, the amounts of operations needed to perform eigenvector decomposition incur a high computation cost and can be sensitive to the choice of the metric used for similarity, which could affect clustering performance..

4.3.2 Sparse Subspace Clustering (SSC) :

Sparse Subspace Clustering (SSC) is a method to enhance spectral clustering by developing a similarity graph using sparse representation whereby every data point is defined as a sparse linear combination of other data points corresponding to the same subspace [28].Thereafter,

Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

the sparse representation matrix will do the deed of spectral clustering. The method is robust against noise and irrelevant dimensions, which make it very well suited to cluster sparse subspaces. The main crux of SSC is its computational cost. Usually, cost incurred solving sparse optimization problems is more expensive than usual. SSC is popular for performing tests on synthetic datasets to see the efficiency at high-dimensional settings. Application ranges from motion segmentation clustering moving objects within video sequences to hyperspectral image analysis that separates various materials in remote sensing, or in facial recognition clustering images on the grounds of inherent facial features.

4.3.3 Low-Rank Representation (LRR)

Low-Rank Representation (LRR) is a spectral clustering method that constructs a similarity graph by representing data points as a low-rank matrix, ensuring that points belonging to the same subspace are grouped together[29].This method is particularly effective in handling noise and outliers, making it suitable for structured data where relationships between points should be preserved. However, LRR requires solving computationally expensive optimization problems, which can be limiting for large-scale data. LRR is tested on synthetic datasets to analyze its ability to extract low-rank structures.

4.3.4 Kernel-based Spectral Subspace Clustering

Kernel-based Spectral Subspace Clustering(KSSC) extends the standard spectral clustering methods by implementing kernel functions that map the data into higher-dimensional spaces to uncover potential non-linear clusters in the subspaces[30]. This procedure becomes a significant technique to efficiently detect clusters in non-linear and complex structures. Merely, it augments separating power over those that are not linearly separable in the original space. However, the process incurs high costs due to the kernel computations and the graph construction, making it difficult for scaling purposes concerning very large data sets. It is often performed/testing kernel-based spectral clustering on generated artificial data to assess efficiency on complex manifolds.

4.3.5 Robust spectral clustering

Robust spectral clustering(RSC) is concerned about the development of traditional spectral clustering, whereby the formation of the above types of robust similarity or graph construction techniques would ensure noise and outlier resistance making it more applicable in real-world situations [31]. It provides solutions to clean cluster structures while minimizing the influence of outliers. The problem is that this comes at a higher cost in terms of processing because of the extra operations necessary for noise filtering and the construction of robust graphs. Robust spectral clustering is typically tested with synthetic datasets for measuring efficiency under different levels of noise or corruption.

4.4 Hybrid Approaches

Hybrid approaches integrate several clustering paradigms to enhance performance and accuracy, generally combining some type of deep learning with standard clustering. These

Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

approaches use deep-learning models to extract high-quality feature representations, which they then pass on to algorithms of specific clustering paradigms such as spectral clustering, model-based clustering, or density-based clustering. Therefore, a few main streams of work were developed in this area.

4.4.1 GRIDCLUS (Grid-Based Discretization With Density-Based Clustering)

In addition, multiscale grid clusters or GRIDCLUS, . These are regions defined by grids and clustered at density within grid structures; the density connected principles such as DBSCAN will be used to form clusters[32]. GRIDCLUS can analyze not only irregularly shaped dense clusters but can also outstand in noise resistance from traditional grid clustering techniques. However, the expense of computation becomes higher because most dimensional increases have adverse effects on the richness of data represented by grids. The performance of GRIDCLUS has frequently been certified with synthetic datasets on clustering of non-uniform data distributions.

4.4.2 Hierarchical Density-Based Spatial Clustering

HDBSCAN with Subspaces augments the standard Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) algorithm with a new feature of subspace selection, which means that subspaces can be incrementally refined along the self-similar hierarchical clustering process in order to uncover clusters in relevant dimensions This makes it less unlike traditional density-based techniques, in which varied densities of clusters could be detected, and noise and outliers were automatically filtered, rendering it suitable for complex, high-dimensional data[33]. However, HDBSCAN with Subspaces has very high computation costs when applied to large datasets due to the additional burden of processing overhead for the subspace refinement operations.

4.4.3 Hierarchical Subspace Clustering

HiSC is a hierarchical subspace clustering algorithm that integrates density-based clustering and grid partitioning. It refines the subspaces and hierarchically groups the dense regions[34]. It is an effective clustering technique for a hierarchical and nested cluster structure. Its effective for data sets in which clusters are made up of one another. This method is very sensitive to parameter settings such as density thresholds and grid sizes, whose bad settings may significantly affect clustering quality.

4.4.4 Deep Subspace Clustering Networks (DSC-Net)

Integrating deep learning into the subspace clustering with a sparse subspace clustering layer following low-dimensional data processing through autoencoder is what DSC-Net actually uses in real-world applications, which proves a powerful way for dealing with non-linear and complex subspaces, for example, high-dimensionalities in images and videos[35].Automatic feature representation learning for cluster analysis of complex, structured data is among many benefits of DSC-Net. Unfortunately, however, it is computationally expensive and needs large-

Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

scale training datasets for proper learning. Testing performance mostly involves using image and video datasets .

4.5 Discussion on Strengths and Weaknesses of Subspace Clustering Algorithms

The various subspace clustering algorithms present their diverse strengths and weaknesses that subject them to an appropriate use with different dataset types and applications. For instance, while grid-based approaches like CLIQUE, MAFIA, and FIRES are very scalable and automatic in subspace selection, they also have a lot of weaknesses such as noise, fixed grid sizes, and irregular shapes of clusters. On the other hand, density-based methods such as SUBCLU and HDBSCAN with Subspaces are noisy, can detect arbitrary shapes, and are also computationally expensive, especially in higher dimensions. Also, spectral methods such as Sparse SSC, LRR, and Kernel-Based SSC become useful for applications related to extremely complicated subspaces and densely overlapping polytopes but remain quite computationally intensive and sensitive to parameter tuning. Hybrid techniques such as PROCLUS and ORCLUS balance accuracy and efficiency, thus making them more generally applicable for clustering. Deep-learning-based methods fit well with high-dimensional data, especially from images, but they also have disadvantages, such as expensive computation being required and lots of training data. Thus, the final choice trail must be between trade-offs of scale, accuracy, and complexity and will thus depend on the data size, amounts of noise, and computational constraints.

5. COMPARATIVE ANALYSIS OF SUBSPACE CLUSTERING METHODS

To create an improved understanding of the trade-offs among the various subspace clustering methods, we thus present a comparative study across the major performance criteria

5.1 Comparison table

The table below summarizes the comparative evaluation of major subspace clustering techniques:

Method	Scalability	Noise Tolerance	Accuracy	Computational Complexity
CLIQUE [12]	High	Low	Moderate	High (Grid Search Complexity)
ENCLUS [13]	Moderate	Moderate	High	High (Entropy-Based Selection)
MAFIA [14]	High	Low	Moderate	High (Adaptive Grid Partitioning)
pCLIQUE [19]	Very High	Low	Moderate	High (Parallel Grid Search)
FIRES[20]	High	Moderate	High	High (Adaptive Grid Refinement)
SUBCLU [21]	High	High	High	Moderate (Density-Based)

Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

PROCLUS [15]	Moderate	High	High	Moderate (K-Medoids Based)
ORCLUS [22]	Moderate	High	High	High (PCA-Based Feature Selection)
4C [23]	Moderate	High	High	High (Correlation-Based Clustering)
FASTER [25]	High	Moderate	High	High (Feature Selection Optimization)
SSC [27]	Low	High	Very High	Very High (Spectral Decomposition)
Sparse SSC [28]	Low	High	Very High	Very High (Sparse Optimization)
LRR [29]	Low	High	Very High	Very High (Matrix Factorization)
Kernel-Based SSC [30]	Moderate	High	Very High	High (Kernel and Graph Construction)
Robust Spectral Clustering[31]	Moderate	High	High	High (Robust Graph Construction)
GRIDCLUS [32]	High	Moderate	Moderate	High (Grid and Density Computation)
HDBSCAN with Subspaces[33]	High	High	High	High (Hierarchical an d Density-Based Computation)
HiSC [34]	Moderate	Moderate	High	High (Hierarchical Subspace Refinement)
DSC-Net [35]	Low	High	Very High	Very High (Neural Network Training)

Table 2: Comparative Analysis of Subspace Clustering Methods

5.1.1 Discussion of Comparative Findings

A comparative study found that different algorithms of subspace clustering perform differently on key aspects that include scalability, noise robustness, accuracy, and computational complexity. It was found that no one algorithm is good in all situations but rather that they have different scores based on datasets, computational constraints, and application domains.

1.Scalability Considerations.

When considering large datasets, scalability is one important aspect, pCLIQUE, MAFIA, FIRES, and GRIDCLUS show between high to very high scalability when using the most

Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

efficient grid-based or other forms of parallel processing .Spectral methods, such as Sparse SSC ,LRR or DSC-Net however, show low scalability, as they most commonly require matrix factors or deep learning-based models for sparse
optimization works.
2. Noise Tolerance and Robustness.
Real-world clustering applications require an efficient method for handling noise and outliers. SUBCLU, HDBSCAN with Subspaces, and Robust Spectral Clustering, will be used to show density-based approaches, which are effective in filtering out the noise and, thus, can be useful in areas like anomaly detection, fraud detection, and the likes in the cybersecurity domain. Basically, grid-based algorithms, such as CLIQUE and MAFIA, exhibit low noise tolerance as they are fixed partitions sensitive to noisy variations in data

3. Consistency and Clustering Effectiveness

The performance of these methods, including Sparse SSC, LRR, Kernel-Based SSC, and DSC-Net, is the highest among all spectral clustering methods because of their capability to represent non-linear relationships and complex subspace structures However, these methods have very high computational complexities, which limit their applications in large datasets. Alternatively, grid-based and density-based clustering schemes such as FIRES, SUBCLU, and GRIDCLUS provide moderate-to-high accuracy, which makes them computationally efficient and thus applicable in large datasets.

4. Computational Complexity and Feasibility

One of the most important trade-offs in the subspace clustering would be between cost and clustering modified efficacy: Low-cost: Grid-based clustering algorithms (CLIQUE, MAFIA, FIRES) and density-based clustering (SUBCLU, GRIDCLUS, HDBSCAN) are low-cost; instead, they offer good computational effectiveness,

High-cost: Some algorithms adopt spectral clustering approach (SSC, LRR, DSC-Net, Kernel-Based SSC); these show very high computational cost due to matrix decomposition, kernel operation, and deep learning models

5.2 Performance Metrics:

Different performance metrics are used to check the performance of subspace clustering methods, depending on the characteristics of the dataset and requirements of the application. Different subspace clustering methods would have different advantages based on the characteristics of the datasets where they are applied, such as size, sparsity of features, amount of noise, or density of clusters.

5.2.1 Common Performance Metrics

Usually there are few crucial measures on which the methods of subspace clustering are very much evaluated:
Clustering Accuracy (ACC): Proportion of assigned clusters which were correct. Normalized Mutual Information (NMI): Measures how similar the predicted cluster assignments are to the corresponding ground truth cluster assignments.
Adjusted Rand Index (ARI): Measures clustering agreement after an adjustment for some level of chance agreement.

Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Silhouette Score: Quantifies how well separated are the clusters in terms of distances within and between clusters.
Execution Time: A metric for computational efficiency of the algorithm, very important for applications which are large scale and/or need real-time performance

5.2.2 Best Methods for Different Dataset Types

Dataset Type	Best Performing Methods	Key Considerations
Gene Expression Data	Sparse SSC, SUBCLU, LRR, Robust Spectral Clustering, PROCLUS	High-dimensional, noisy, sparse; clusters may be small, overlapping, or non-linear
Large-Scale Datasets	pCLIQUE, FIRES, GRIDCLUS, HDBSCAN with Subspaces	Handles large -scale data efficiently but may require high computational resources
Noise-Prone Datasets	Robust Spectral Clustering, SUBCLU, HDBSCAN with Subspaces	Robust to noise and outliers; effective for anomaly detection but may have higher complexity
High-Dimensional Structured Data	Sparse SSC, LRR, Kernel-Based SSC, DSC-Net	Captures non-linear structures but is computationally expensive
General-Purpose Subspace Clustering	FIRES, PROCLUS, ORCLUS	Balances accuracy and efficiency; suitable for general clustering tasks
Sparse Data	DOC, Sparse SSC, P3C	Designed for sparse data distributions; effective in high-dimensional space
Overlapping Clusters	SSC, ORCLUS, Kernel-Based SSC	Excels at finding overlapping clusters but can be sensitive to parameter settings
Real-Time Clustering	GRIDCLUS, HDBSCAN, FIRES	Optimized for fast execution but may sacrifice some clustering accuracy
Anomaly Detection	HDBSCAN with Subspaces, Robust Spectral Clustering, LRR	Strong outlier detection capabilities; effective in cybersecurity and fraud detection
Image & Video Data	DSC-Net, Kernel-Based SSC, Sparse SSC	Handles complex image/video data but requires deep learning-based models
Hierarchical Clustering Needs	HiSC, Robust Spectral Clustering, HDBSCAN	Suitable for multi-level clustering structures; sensitive to parameter tuning

Table 3: Comparative Analysis of Subspace Clustering Method for Different Dataset Types

Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

5.2.3 Discussion on Method Suitability

This whole issue varies considerably according to the nature of the target dataset concerning its size, amount of noise, dimensionality, and computational constraints. Grid-specific methods such as pCLIQUE, FIRES, and GRIDCLUS are best suited for clustering very large datasets since they are scalable and efficient. In the presence of noise and outliers, methods like robust spectral clustering, SUBCLU, or clusters with HDBSCAN focus on those datasets that are appropriate for them. Sparse subspace clustering, LRR, and kernel-based methods are given to the high-dimensional setting, which has some useful structure but is heavy on computation. Methods such as FIRES, PROCLUS, and ORCLUS also serve as reasonable compromises with respect to performance and efficiency for other clustering problems. Sparse datasets may also be treated by DOC and P3C. Here SSC and ORCLUS are used. The above scenarios would fit GRIDCLUD and HDBSCAN in situations where clusterization is closer to real-time; learning methods based on deep learning, for instance, DSC-Net, fit very well in terms of image and video data and
essentially require heavy hardware. The optimal method then becomes the balancing act for trade-off between data
complexity and algorithm performance.

5.2.4 Summary of Performance Trade-offs

Translated into other languages, the subspace clustering algorithms will cause trade-offs in scaling and noise tolerance versus accuracy and computational complexity. They are therefore best suited to certain dataset characteristics. Generally, grid-based algorithms (such as CLIQUE, FIRES, GRIDCLUS) are highly scalable but are limited in convergence, because even an irregular cluster shape will tend to noise susceptibility. As such, a density-based method-effective in general case noise and arbitrary cluster structures, e. g. SUBCLU, HDBSCAN with subspaces-is characterized by expensive computation in high-dimensional data. High precision can be obtained from spectral methods (e.g. Sparse SSC, LRR, Kernel-Based SSC) in relatively complex subspaces but are very computationally intensive and "require tuning" of their parameters. Hybrid methods (e. g. PROCLUS, ORCLUS) can achieve a balance between accuracy and efficiency, hence making them generally applicable but still require parameter optimization. DSC-Net, a deep-learning-based clustering scheme, works excellently for image and video data but requires huge training datasets and enormous computing power. In effect, therefore, the choice of that algorithm hinges on efficiency against precision and computational feasibility, among other things, which includes dataset size, noise levels, and dimensional complexity.

6 APPLICATIONS OF SUBSPACE CLUSTERING

Subspace clustering is something that has been widely applied in various fields, especially where high-dimensional data poses a challenge for traditional clustering techniques. Some of the important applications where it has made a huge difference are highlighted below.

6.1 Bioinformatics and Genomics

Gene Expression Analysis- The Subspace Clustering techniques have been popularly used in so many applications aiming to find sets of genes that behave the same under certain experimental conditions from microarray and RNA sequence data.Such techniques include

Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Sparse Subspace Clustering (SSC) and Low-Rank Representation (LRR) which work well for high-dimensional analytical gene-expression profiles ..

6.2 Image and Video Processing

Face Recognition : Subspace clustering can include more fun ways of clustering images of facial features according to illumination, pose, or even facial expression. In fact, there are deep learning-based algorithms such as DSC-Net and Transformer-Based Subspace Clustering that yield best accuracy results .The approaches such as HDBSCAN and SSC were found to work well in exploiting sparse text representations .

6.3 Text Mining and Natural Language Processing (NLP)

Document Clustering: In quite high dimensions from the term-frequency feature space, one may do subspace clustering to combine documents that are somewhat similar in topics together .Topic modelling: Subspace clustering would reveal these invisible topical structures within large corpora for applications such as news categorization and in recommendation systems.

6.4 Internet of Things (IoT) and Sensor Networks

Detection of Anomalies in IoT Systems: Subspace clustering is being applied for identifying anomalous activities in smart homes, industrial monitoring systems, and cybersecurity.

6.5 Financial and Business Analytics

Subspace clustering is indeed one method used in the financial institution to detect some unusual patterns in multi-dimensional records of transactions to identify fraudulent transactions. Among the more well-known model-based clustered techniques for this purpose are PROCLUS and ORCLUS .

6.6 The healthcare and medical imaging level:

Use of subspace clustering includes segmentation of MRI and CT scans and identification of abnormal tissue patterns in radiology images.The spectral clustering techniques like SSC and LRR are particularly useful in this area

6.7 Cyber Security and Network Analysis

Intrusion detection systems or IDS: Subspace clustering techniques cluster traffic patterns in high-dimensional log data and detect intrusion and other malicious activities. Other cyber threat detections include the use of techniques such as SSC, density-based clustering.

6.8 Social Network Analysis

Community Detection: Subspace clustering is highly used for the detection of communities in social networks, with an interest in identifying users who have shown similar patterns of interaction between the users.

7. CHALLENGES AND RESEARCH GAPS IN SUBSPACE CLUSTERING

Despite a significant promise shown by subspace clustering with respect to high-dimensional data handling, some serious challenges remain unresolved, due to which this clustering

Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

remains ineffective in most real-world cases. In this section, we will highlight the main challenges and research gaps in subspace clustering.

7.1 Scalability and Computational Complexity

Subspace clustering algorithms like spectral clustering (SSC, LRR) involve heavy computations via convex optimization and matrix decomposition. Since datasets have grown up to their live use, the need for more real-time scalable approaches arose. Grid-based and model-based methods are scalable but may not capture complicated subspace structures

7.2 Sensitivity to Noise and Outliers A single example, many subspace clustering algorithms contend that data is a set of clean, well-structured high-dimensional vectors with few exceptions. Unfortunately, real-life datasets like financial transactions or IoT sensors carry a lot of noise and much redundant dimensionality.

7.3 Lack of Standardized Benchmark Datasets and Evaluation Metrics

Whereas in classic clustering one finds practically universally accepted benchmark datasets and evaluation metrics, such is altogether lacking in subspace clustering. The majority of researchers use MNIST, 20 Newsgroups, and Gene Expression Data, yet nothing is agreed upon when it comes to candidate methods for evaluating clustering effectiveness This gap leads to severe hindrances in objectively comparing the performance of different methods.

7.4 Interpretability and Explainability

Models for subspace clustering, and especially those based on deep learning (for instance, DSC-Net and Transformer-Based Clustering), are severely regarded as black boxes and become highly challenging to explain the choice of any given subspace. Interpretability is crucial in such high-impact applications as medical diagnostics and financial decision-making, thus tied to the model's trustworthiness .

7.5 Parameter Sensitivity and Hyperparameter Tuning

Most of the previously proposed subspace clustering approaches involve manual tuning of hyperparameters such as the number of clusters, dimensionality thresholds, and similarity measures. Such frameworks cannot be conveniently exercised for practical deployment in areas where tuning hyperparameters is not possible manually.

7.6 Fairness and Bias in Clustering Decisions

As shown by recent studies, bias in feature selection in subspace clustering algorithms leads to unintended discrimination in applications such as the social sciences, hiring decisions, and medical diagnostics Fairness issues demand the need for bias-deduction frameworks and fairness-aware clustering algorithms.

7.7 Dynamic and Streaming Data

Most of the traditional subspace clustering approaches cannot support dynamic settings. Application domains such as real-time fraud detection, IoT analytics, and stock market price analysis require somewhat fast adaptive clustering [36].Open research problems remain in the development of incremental subspace clustering approaches.

Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

8. PROPOSED CONCEPTUAL FRAMEWORKS TO ADDRESS RESEARCH GAPS

We advance and propose several conceptual frameworks and methodological advancements to effectively deal with those not only challenges but also the research gaps that have been pin-pointed in subspace clustering. The below figure 8 shows, A structured taxonomy highlighting key challenges and proposed methodological advancements in scalability, robustness, interpretability, and fairness within subspace clustering.

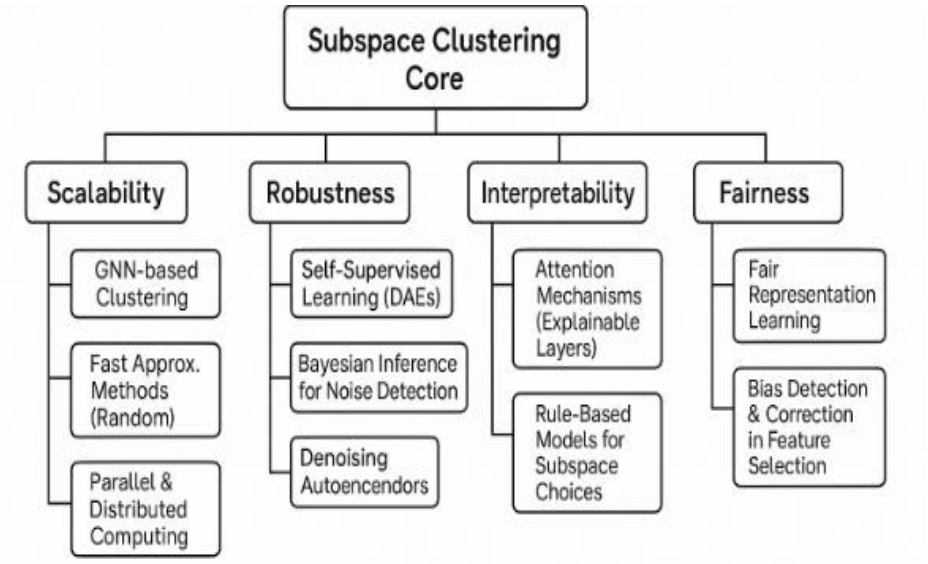

Figure 8: Conceptual Framework for Advancing Subspace Clustering Research

8.1 Scalable and Efficient Subspace Clustering Algorithms

Modern approaches to subspace clustering utilizing spectral methods like Sparse Subspace Clustering (SSC) and Low-Rank Representation (LRR) have been found to become extremely resource demanding as the size of the dataset increases from large to enormous .To further improve scalability, we suggest:
Graph Neural Network (GNN)-Based Clustering: GNNs can exploit feature extraction for similarity learning that enables subspace clustering [45]
Fast Approximation Methods: Uses randomized matrix decomposition and low-rank approximations to speed up eigenvalue computations in spectral clustering .
Understanding of Parallelized and Distributed Computing Approaches: Abstraction from credibility-to-distributed clustering methods would concern the large-scale application areas from IoT and Big Data-but yet again, explains how effective adaptations are made.

8.2 Robust Subspace Clustering for Noisy and High-Dimensional Data

Subspace clustering has been greatly impeded by noise sensitivity in applications such as bioinformatics and financial analytics.To encourage robustness to noise,
we propose:
Self-Supervised Learning for Noise Filtering: The self-supervised deep learning architectures such as the learning of robust representations before the clustering are able to clear noise in the model[37] for Feature Selection: DAEs are trained to remove irrelevant features and noise before subspace clustering

Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Bayesian Subspace Clustering Models: Bayesian inference is used on a model detecting uncertainties and noisy dimensions

8.3 Explainable and Interpretable Subspace Clustering Models

Modeling subspace clusters typically employs black-box techniques, which make interpretation much more difficult when it comes to applications in healthcare, finance, and social sciences. The answer is:
Attention Mechanisms of Subspace: Attention-based Interpretability Layers within Deep Subspace Clustering Frameworks: Eyes and Ears under the Supervision of Visual Attention Rule Based Subspace Selection: Explainable Rule-based Models Give Explanation About Subspace Selection

8.4 Fairness-Aware Subspace Clustering

Social sciences, recruitment, and medical diagnostics may face discriminatory consequences due to selection bias in subspaces. Some possible approaches to fairness are: Fair Representation Learning in Clustering: Whereby clustering algorithms are developed to mitigate bias in feature selection [38].Fair Clustering Metrics and Constraints: Where loss functions for clustering that enforce fairness are proposed.Bias detection and correction mechanisms: Real-time audit of feature selection for fairness to identify and rectify bias

9. EMERGING TRENDS AND FUTURE DIRECTIONS

Emergence of subspace clustering discovery trends for future research directions. These trends try to improve the efficiency, robustness, and domain applicability, among others, of subspace clustering.

9.1 Integration of Deep Learning with Subspace Clustering

By integrating deep neural networks with subspace clustering methods, good results have truly demonstrated improvements in feature representation quality and clustering accuracy. Deep subspace clustering networks (DSC-net), autoencoders, and attention mechanisms make possible improved subspace discovery via transformer-based[43]. Future research shall pay more attention to self-supervised learning to further reduce dependence on labelled data.

9.2 Explainable AI (XAI) for Subspace Clustering

As explainable models for subspace clustering are in high demand, the need for interpretability in the domains of healthcare and finance cannot be overemphasized. Future research agendas should include the following[40]**.**Subspace attention mechanisms for the model to attend to the most important features in its decision process.This includes post-hoc explainability tools such as SHAP and LIME for the interpretation of model outputs

9.3 Fairness and Ethical Considerations in Clustering

Biases in clustering algorithms can result in unfairness and discrimination, especially in domains like social sciences, recruitment, and medical diagnoses. Hence, researchers are working on.Fair clustering loss functions to ensure unbiased grouping .Bias-correction approaches to identify and mitigate skewed feature selection,

Journal of Information Systems Engineering and Management
2025, 10(35s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

9.4 Real-Time and Adaptive Subspace Clustering

Such traditional subspace clustering fails to adapt to real-time demands posed by Internet of Things applications for example, fraud detection and stock market analysis and their future advancements must target:Incremental and online subspace clustering for continuously increasing streams of data Reinforcement learning-based clustering application in order to allow dynamic adaptation to changing [42].

9.5 Hybrid and Multi-Modal Clustering Methodologies

As datasets become more challenging with higher complexity, some forms of hybridization between clustering paradigms may yield better clustering results. As an extension to the future work, one may envisage the following:The ability to perform cross-domain clustering, integrating text, images, and structured data in a multi-modal setting. Hybrid clustering will be a blend of graph, probabilistic, and deep learning models[41]

9.6 Evaluation Metrics and Benchmark Datasets Standardization

More challenging than all is the absence of common datasets and evaluation criteria to benchmark subspace clustering Recommendations for future work should include: A broadened benchmark dataset specifically targeted for subspace clustering.An agreed upon evaluation metric accepted by all practitioners wherein a fair comparison of results could be made.

9.7 Energy-Efficient and Green AI Approaches

That's why energy-efficient clustering algorithms aimed at power optimization are under research. They involve:Low-power hardware acceleration for clustering algorithms and Efficient deep-clustering architectures that have less memory overhead.

CONCLUSION:

Subspace clustering has become an essential technique for analyzing high-dimensional data, enabling more effective clustering by identifying meaningful subspaces. This paper has provided a comprehensive survey of subspace clustering techniques, categorizing them into grid-based, model-based, spectral, and hybrid approaches. We have also presented a comparative analysis of these methods based on scalability, noise tolerance, accuracy, and computational complexity.

Despite significant advancements, several challenges and research gaps remain. These include scalability issues, sensitivity to noise, lack of standardized benchmarks, interpretability concerns, and fairness-related biases. To address these limitations, we proposed conceptual frameworks involving deep learning integration, fairness-aware clustering, real-time adaptive clustering, and explainable AI techniques.

Looking ahead, emerging trends such as self-supervised learning, hybrid clustering models, and reinforcement learning-driven adaptive clustering are expected to shape the future of subspace clustering. Furthermore, standardization of evaluation benchmarks and the development of energy-efficient clustering models will be critical for the field’s growth.

In conclusion, subspace clustering continues to be a high-impact research area with widespread applications in bioinformatics, image processing, cybersecurity, IoT analytics, and social network analysis. By addressing the identified research gaps and leveraging emerging technologies, future advancements will further enhance the scalability, interpretability, and applicability of subspace clustering methods.

REFERENCES:

[1] Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques* (3rd ed.). Elsevier..
[2] Deng, Z., Choi, K.S., Jiang, Y., Wang, J. and Wang, S., 2016. A survey on soft subspace clustering. *Information sciences*, *348*, pp.84-106.
[3] Assent, Ira. "Clustering high dimensional data." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2, no. 4 (2012): 340-350.
[4] Zhao, Ying, George Karypis, and Usama Fayyad. "Hierarchical clustering algorithms for document datasets." *Data mining and knowledge discovery* 10 (2005): 141-168.
[5] Deng, Zhaohong, Kup-Sze Choi, Yizhang Jiang, Jun Wang, and Shitong Wang. "A survey on soft subspace clustering." *Information sciences* 348 (2016): 84-106.
[6] Kaur, A., Datta, A. A novel algorithm for fast and scalable subspace clustering of high-dimensional data. *Journal of Big Data* **2**, 17 (2015).
[7] Kaur, A., Datta, A. A novel algorithm for fast and scalable subspace clustering of high-dimensional data. *Journal of Big Data* **2**, 17 (2015).
[8] Murtagh, Fionn, Jean-Luc Starck, and Michael W. Berry. "Overcoming the curse of dimensionality in clustering by means of the wavelet transform." *The Computer Journal* 43, no. 2 (2000): 107-120.
[9] Wang, Kaiye, Ran He, Liang Wang, Wei Wang, and Tieniu Tan. "Joint feature selection and subspace learning for cross-modal retrieval." *IEEE transactions on pattern analysis and machine intelligence* 38, no. 10 (2015): 2010-2023.
[10] Lee, Sang-hyuk, Sun Yan, Yoon-su Jeong, and Seung-soo Shin. "Similarity measure design for high dimensional data." *Journal of Central South University* 21 (2014): 3534-3540.
[11] Hinojosa, Carlos, Esteban Vera, and Henry Arguello. "A fast and accurate similarity-constrained subspace clustering algorithm for hyperspectral image." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 14 (2021): 10773-10783
[12] Agrawal, R., Gehrke, J., Gunopulos, D., & Raghavan, P. (1998, June). Automatic subspace clustering of high dimensional data for data mining applications. In *Proceedings of the 1998 ACM SIGMOD international conference on Management of data* (pp. 94-105).
[13] Vignesh, T., K. K. Thyagharajan, D. Murugan, M. Sakthivel, and S. Pushparaj. "A novel multiple unsupervised algorithm for land use/land cover classification." *Indian J. Sci. Technol* 9 (2016): 1-12.
[14] Goil, Sanjay, Harsha Nagesh, and Alok Choudhary. "Mafia: E±cient and scalable subspace clustering for very large data sets." In *Proc. 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Citeseer*, pp. 443-452. 1999.
[15] Yip, Kevin Y., David W. Cheung, and Michael K. Ng. "Harp: A practical projected clustering algorithm." *IEEE Transactions on knowledge and data engineering* 16, no. 11 (2004): 1387-1397.
[16] Elhamifar, Ehsan, and René Vidal. "Sparse subspace clustering: Algorithm, theory, and applications." *IEEE transactions on pattern analysis and machine intelligence* 35, no. 11 (2013): 2765-2781.
[17] Zhang, Xiaoyue, Jingfei He, Yatong Zhou, and Yue Chi. "A subspace approach to sparse-sampling-based multi-attribute data aggregation in IoT." *IEEE Internet of Things Journal* 9, no. 18 (2022): 18054-18063.

[18] Zhao, Mingyu, Weidong Yang, and Feiping Nie. "Transformer-Based Contrastive Multi-view Clustering via Ensembles." In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 678-694. Cham: Springer Nature Switzerland, 2023.

[19] Zhai, Hongjie, Makoto Haraguchi, Yoshiaki Okubo, and Etsuji Tomita. "A fast and complete algorithm for enumerating pseudo-cliques in large graphs." *International Journal of Data Science and Analytics* 2, no. 3 (2016): 145-158.

[20] Elhamifar, Ehsan, and René Vidal. "Sparse subspace clustering: Algorithm, theory, and applications." *IEEE transactions on pattern analysis and machine intelligence* 35, no. 11 (2013): 2765-2781.

[21] Prinzbach, Jürgen, Tobias Lauer, and Nicolas Kiefer. "Accelerating density-based subspace clustering in high-dimensional data." In *2021 International Conference on Data Mining Workshops (ICDMW)*, pp. 474-481. IEEE, 2021.

[22] Agrawal, Rakesh, Johannes Gehrke, Dimitrios Gunopulos, and Prabhakar Raghavan. "Automatic subspace clustering of high dimensional data." *Data Mining and knowledge discovery* 11 (2005): 5-33.

[23] Böhm, Christian, Christos Faloutsos, and Claudia Plant. "Outlier-robust clustering using independent components." In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pp. 185-198. 2008.

[24] Yiu, Man Lung, and Nikos Mamoulis. "Iterative projected clustering by subspace mining." *IEEE Transactions on Knowledge and Data Engineering* 17, no. 2 (2005): 176-189.

[25] Peng, Chong, Zhao Kang, Ming Yang, and Qiang Cheng. "Feature selection embedded subspace clustering." *IEEE Signal Processing Letters* 23, no. 7 (2016): 1018-1022.

[26] Moise, Gabriela, Jorg Sander, and Martin Ester. "P3C: A robust projected clustering algorithm." In *Sixth international conference on data mining (ICDM'06)*, pp. 414-425. IEEE, 2006.

[27] Vidal, René. "Subspace clustering." *IEEE Signal Processing Magazine* 28, no. 2 (2011): 52-68.

[28] Elhamifar, Ehsan, and René Vidal. "Sparse subspace clustering: Algorithm, theory, and applications." *IEEE transactions on pattern analysis and machine intelligence* 35, no. 11 (2013): 2765-2781.

[29] Li, Chenyu, Bing Zhang, Danfeng Hong, Jing Yao, and Jocelyn Chanussot. "Low-rank representations meets deep unfolding: A generalized and interpretable network for hyperspectral anomaly detection." *arXiv preprint arXiv:2402.15335* (2024).

[30] Su, Yuanchao, Lianru Gao, Mengying Jiang, Antonio Plaza, Xu Sun, and Bing Zhang. "NSCKL: Normalized spectral clustering with kernel-based learning for semisupervised hyperspectral image classification." *IEEE Transactions on Cybernetics* 53, no. 10 (2022): 6649-6662.

[31] Qin, Yalan, Xinpeng Zhang, Liquan Shen, and Guorui Feng. "Maximum block energy guided robust subspace clustering." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, no. 2 (2022): 2652-2659.

[32] Yan, Yan, and Frederick C. Harris Jr. "A Survey of Data Clustering for Cancer Subtyping." *International Journal for Computers and Their Applications* 28, no. 2 (2021): 1-13.

[33] Sahu, Ramgopal T., Mani Kant Verma, and Ishtiyaq Ahmad. "Density-based spatial clustering of application with noise approach for regionalisation and its effect on hierarchical clustering." *International Journal of Hydrology Science and Technology* 16, no. 3 (2023): 240-269.

[34] Ou, Qiyuan, Siwei Wang, Pei Zhang, Sihang Zhou, and En Zhu. "Anchor-based multi-view subspace clustering with hierarchical feature descent." *Information Fusion* 106 (2024): 102225.

[35] Ji, Pan, Tong Zhang, Hongdong Li, Mathieu Salzmann, and Ian Reid. "Deep subspace clustering networks." *Advances in neural information processing systems* 30 (2017).

[36] Feng, Chunhui, Junhua Fang, Yue Xia, Pingfu Chao, Pengpeng Zhao, Jiajie Xu, and Xiaofang Zhou. "Ocean: online clustering and evolution analysis for dynamic streaming data." In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, pp. 4504-4517. IEEE, 2024.

[37] Tian, Zhiyi, Jiaming Xu, and Jen Tang. "Clustering High-Dimensional Noisy Categorical Data." *Journal of the American Statistical Association* 119, no. 548 (2024): 3008-3019.

[38] Liu, Zeyuan, Xin Zhang, and Benben Jiang. "Active learning with fairness-aware clustering for fair classification considering multiple sensitive attributes." *Information Sciences* 647 (2023): 119521.

[39] Sui, Jinping, Zhen Liu, Li Liu, Alexander Jung, and Xiang Li. "Dynamic sparse subspace clustering for evolving high-dimensional data streams." *IEEE Transactions on Cybernetics* 52, no. 6 (2020): 4173-4186.

[40] Nguyen, Dung, Ariel Vetzler, Sarit Kraus, and Anil Vullikanti. "Contrastive explainable clustering with differential privacy." *arXiv preprint arXiv:2406.04610* (2024).

[41] Raya, Sura, Mariam Orabi, Imad Afyouni, and Zaher Al Aghbari. "Multi-modal data clustering using deep learning: A systematic review." *Neurocomputing* (2024): 128348.

[42] Zhuoyue Ou, Xiuqin Deng, Lei Chen, Jiadi Deng,Adaptive multi-view subspace clustering algorithm based on representative features and redundant instances,Neurocomputing, Volume 620,2025,128839,ISSN 0925-2312

[43] Li, Yanming, Shiye Wang, Changsheng Li, Ye Yuan, and Guoren Wang. "Towards very deep representation learning for subspace clustering." *IEEE Transactions on Knowledge and Data Engineering* 36, no. 7 (2024): 3568-3579.

[44] Arafat, Muhammad Yeasir, Sungbum Pan, and Eunsang Bak. "Distributed energy-efficient clustering and routing for wearable IoT enabled wireless body area networks." *IEEe Access* 11 (2023): 5047-5061.

[45] Singha, Sondip Poul, Md Mamun Hossain, Md Ashiqur Rahman, and Nusrat Sharmin. "Investigation of graph-based clustering approaches along with graph neural networks for modeling armed conflict in Bangladesh." *International Journal of Data Science and Analytics* 18, no. 2 (2024): 187-203.