Astronomical Techniques and Instruments, Vol. 3, January 2026, 10–25

Review Open Access

Progress in clustering algorithms for astronomical spectra over a decade

Jianing Tian[1], Haifeng Yang[1*], Jianghui Cai[1,2*], Yuqing Yang[1], Xiangru Li[3], Zhenping Yi[4,5], Lili Wang[6]

[1]*School of Computer Science and Technology*, *Taiyuan University of Science and Technology*, *Taiyuan* 030024, *China*

[2]*School of Computer Science and Technology*, *North University of China*, *Taiyuan* 030051, *China*

[3]*School of Computer Science*, *South China Normal University*, *Guangzhou* 510631, *China*

[4]*School of Mechanical*, *Electrical and Information Engineering*, *Shandong University*, *Weihai* 264209, *China*

[5]*Shandong Key Laboratory of Intelligent Electronic Packaging Testing and Application*, *Shandong University*, *Weihai* 264209, *China*

[6]*School of Computer and Information*, *Dezhou University*, *Dezhou* 253023, *China*

*Correspondences: hfyang@tyust.edu.cn; jianghui@tyust.edu.cn

Received: March 11, 2025; Accepted: May 30, 2025; Published Online: May 31, 2025; https://doi.org/10.61977/ati2025030; https://cstr.cn/32083.14.ati2025030

Citation: Tian, J. N., Yang, H. F., Cai, J. H., et al. 2026. Progress in clustering algorithms for astronomical spectra over a decade. *Astronomical Techniques and Instruments*, **3**(1): 10−25. https://doi.org/10.61977/ati2025030.

Abstract: As large-scale astronomical surveys, such as the Sloan Digital Sky Survey (SDSS) and the Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST), generate increasingly complex datasets, clustering algorithms have become vital for identifying patterns and classifying celestial objects. This paper systematically investigates the application of five main categories of clustering techniques—partition-based, density-based, model-based, hierarchical, and "others"—across a range of astronomical research over the past decade. This review focuses on the six key application areas of stellar classification, galaxy structure analysis, detection of galactic and interstellar features, high-energy astrophysics, exoplanet studies, and anomaly detection. This paper provides an in-depth analysis of the performance and results of each method, considering their respective suitabilities for different data types. Additionally, it presents clustering algorithm selection strategies based on the characteristics of the spectroscopic data being analyzed. We highlight challenges such as handling large datasets, the need for more efficient computational tools, and the lack of labeled data. We also underscore the potential of unsupervised and semi-supervised clustering approaches to overcome these challenges, offering insight into their practical applications, performance, and results in astronomical research.

Keywords: Clustering; Stellar types; Astronomical techniques; Classification; Galaxies

1. INTRODUCTION

As astronomy moves into a data-intensive era, the SDSS[1] and LAMOST [2] are continuously recording huge amounts of observational data. The LAMOST survey alone released over 2.2 million spectra in its first data release[3], presenting both opportunities and computational challenges for traditional methods. Up until now, LAMOST has obtained on the order of 20 million spectral datasets. The latest data release, LAMOST DR10[1] (Data Release 10), is now publicly available. Clustering algorithms are essential in this field, intelligently grouping objects based on their physical properties. For example, the Apache Point Observatory Galactic Evolution Experiment (APOGEE) and LAMOST[2,4] have successfully used clustering algorithms to accurately identify stellar populations. Unsupervised approaches, like K-means, applied to 144340 A-type stars in LAMOST data[5] have proven effective in detecting rare objects that supervised template methods might miss; meanwhile, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Gaussian Mixture Models (GMMs)[6] have also provided strong support for galaxy research, in particular demonstrating advantages in handling high-redshift unlabeled datasets. Here, the "labels" refer to predefined classifications used to characterize astrophysical properties, serving as input criteria for clustering algorithms or benchmarks for result validation.

1. *http://www.lamost.org/dr10*

In the field of stellar and galaxy research, algorithms such as K-means and DBSCAN[7] have demonstrated significant advantages, while algorithms such as Self-Organizing Mapping (SOM)[8] and GMM[9] have also provided strong support for galaxy research. For label-scarce scenarios, innovative methods for chemical tagging have emerged, including the data-driven model Cannon[10] and Conditional Abundance Matching (CAM)[11], to reconstruct missing labels by correlating observable parameters with underlying physical properties.

Clustering algorithms are valuable in the field of high-energy astrophysics, as confirmed by Carlson et al.[12] in the gamma ray wavelength range, and in exoplanet detection, clustering algorithms significantly improve the efficiency of the identification of candidate planets. Additionally, emerging methods such as Topological Anomaly Detection (TAD)[13] and factor-adjusted spectral clustering[14] perform particularly well when dealing with complex astronomical data.

This review gives a comprehensive summary of the current role of spectral clustering as a powerful tool in astronomy, in a range of areas from object classification to high-dimensional data processing. Connections between clustering methods and the regions to which they are applied are given in Table 1. With the ever-increasing complexity and volume of astronomical data, spectral clustering provides a means to extract meaningful patterns and relationships among celestial objects. In recent years, there has been a growing interest in applying and improving clustering algorithms in various areas of astronomical research to better understand the structure and evolution of the universe while navigating the practical constraints of large data volumes[3,5] and sparse labeling[6,10,11]. By exploring the role of clustering in different applications of astronomy—such as stellar classification, galaxy structure analysis, high-energy astrophysics, exoplanet detection, and large-scale survey data processing—we summarize current achievements and point out challenges, with the aim of stimulating further research on the use of clustering algorithms in astronomical data analysis.

Table 1. Astrophysical applications and corresponding clustering methods

Application categories	Clustering methods
Stellar and cluster classification	Partition-based clustering, density-based clustering, model-based clustering, others clustering
Galaxy and large-scale structure analysis	Partition-based clustering, density-based clustering, model-based clustering, hierarchical clustering, others clustering
High-energy and exoplanetary studies	Partition-based clustering, density-based clustering, model-based clustering, hierarchical clustering, others clustering
Clustering methodology and applications	Partition-based clustering, density-based clustering, model-based clustering, others clustering
Anomaly and outlier detection	Density-based clustering, model-based clustering

2. CLUSTERING METHODS

In this section, we summarize studies that have applied a variety of clustering techniques addressing specific challenges in astronomy. Each method offers characteristic strengths suited to different data types and structures.

2.1. Density-Based Clustering

DBSCAN is able to efficiently recognize star clusters in dense regions, handling irregular shapes and noisy data without the need to pre-set the number of clusters[15,16], and has been successfully applied in astronomy for the identification of features such as stellar streams and galaxy halos. As an improved version of the DBSCAN algorithm, Ordering Points to Identify the Clustering Structure (OPTICS) is able to handle datasets with different densities and reveal hierarchical structure information, providing important support for complex galaxy analyses. Yang et al.[17] later proposed nonparametric density clustering algorithm (NAPC), which is used to automatically select clustering centers and reduce the influence of human-related factors in the analysis process.

2.2. Model-Based Clustering

GMM[18] uses Gaussian distributions to model data and identify overlapping sub-populations in astronomy, and was used by Gao[19] to classify tidal tails in star clusters. Similarly, SOMs[20], which preserve data topologies while reducing dimensionality, are used in galaxy spectral classification to reveal complex patterns in galaxy spectra. Both GMMs and SOMs offer deeper insights than methods like K-means, enhancing the analysis of complex astronomical data.

2.3. Hierarchical Clustering

Hierarchical clustering forms nested clusters, either by merging small clusters (agglomerative) or splitting larger ones (divisive), without needing predefined cluster numbers. This method is ideal for detecting complex structures such as tidal streams and stellar halos[21] because it logically combines or separates the data[22], revealing hierarchical relationships in large datasets.

2.4. Partition-Based Clustering

The K-means clustering method is optimized by dividing the dataset into a predetermined number of clusters with the goal of minimizing variance within the cluster. This algorithm is known for its computational efficiency and simplicity of implementation, and is particularly suitable for dealing with the rapid clustering needs of large

quantities of astronomical data. K-means has been successfully applied to key areas such as the classification of features in galaxy spectra[23] and stellar spectral classification[24], proving its wide applicability and practical value.

2.5. Others Clustering

In recent years, as astronomical data analysis needs have increased in complexity, clustering algorithms have provided a means to provide innovative solutions to key challenges in astrophysical research. Córdova Rosado et al.[25] pioneered the combination of clustering algorithms with machine learning techniques to allow accurate characterization of Active Galactic Nuclei (AGN). In cosmology, Balaguera-Antolínez and Montero-Dorta[26] applied clustering algorithms to dark-matter density field analysis, providing a new method for reconstructing the physical properties of the dark-matter halo. The Blanco-Cuaresma et al.[27] used a different approach, using the chemical labeling method to apply the clustering technique to stellar chemical abundance analysis. Euclid team[28] significantly improved the accuracy of the distribution model of galaxies by combining the clustering algorithm with the redshift correction technique. In addition, the Structure, Excitation, and Dynamics of the Inner Galactic InterStellar Medium (SEDIGISM) survey project[29] has systematically classified molecular clouds using clustering algorithms, providing an important observational basis for studying the star formation process.

Table 2 provide an overview of clustering methods categorized by their application domains, subcategories, specific techniques, and advantages, with some notable applications in astronomy. The table highlight the adaptability of

Table 2. Clustering applications in astronomy

Application category	Subcategories	Clustering metdods	Precision	Efficiency
Stellar and cluster classification[3,5,10,24,27,30–52]	Enhancing detection and classification efficiency	Partition-based clustering, model-based clustering	√	√
	Identification of stellar populations	Partition-based clustering, model-based clustering	√	√
	Unsupervised classification of stellar populations	Hierarchical clustering, density-based clustering	√	
	Identification and analysis of new stellar populations	Model-based clustering, partition-based clustering	√	
	Advanced clustering in large-scale astronomical surveys	Model-based clustering, density-based clustering	√	√
	Automated stellar spectral classification	Spectral clustering, model-based clustering	√	
	Structural analysis of star clusters	Hierarchical clustering, density-based clustering	√	
	Large-scale stellar population analysis	Partition-based clustering, density-based clustering	√	
	Discovery and analysis of new star clusters	Model-based clustering, density-based clustering	√	
Galaxy and large-scale structure analysis[6, 11,17, 20, 22, 23, 25-29, 34, 53-70]	Galaxy structure and large-scale cosmology	Density-based clustering, hierarchical clustering	√	√
	Galaxy morphology and dynamics	Partition-based clustering, others clustering		
	Galaxy clustering and halo mass function	Density-based clustering, model-based clustering	√	√
	Cosmological structure and redshift studies	Others clustering, density-based clustering	√	
Detection of galactic and interstellar features[15–16,19,21,24,29,71–76]	Detection of tidal debris and galactic halo studies	Hierarchical clustering, density-based clustering		√
	Hierarchical clustering in galactic and dust cloud structures	Density-based clustering, model-based clustering, hierarchical clustering	√	
High-energy and exoplanetary studies[34,61,72,77–84]	Gamma-ray and high-energy studies	Density-based clustering, model-based clustering	√	√
	Exoplanet detection and classification	Partition-based clustering, others clustering		√
	Planetary, exoplanet, and solar studies	Hierarchical clustering, others clustering	√	
	Supernova detection and classification	Density-based clustering, model-based clustering	√	
Clustering methodology and applications[13–14,22,81,85]	Stellar and galactic research	Density-based clustering, partition-based clustering	√	
	Overview of clustering techniques	All methods	√	√
Anomaly and outlier detection[86–88]	Detection of outliers and anomalies	Density-based clustering, others clustering	√	√

clustering algorithms to tackle various challenges in astronomical data analysis. For example, density-based methods like DBSCAN and OPTICS are adept at identifying non-convex and noisy structures, making them invaluable for studying stellar streams and galactic halos. Model-based methods, such as GMM, excel in scenarios involving overlapping populations, aiding in quasar and tidal tail identification. Partition-based approaches, like K-means and SOM, are efficient for large datasets, facilitating stellar and galaxy classification.

Advanced techniques, such as deep learning and trajectory-based clustering, address high-dimensional and temporal data, pushing the boundaries of astronomical research. The distribution of clustering methods used in astronomical studies is shown in Fig. 1, highlighting the range of techniques available.

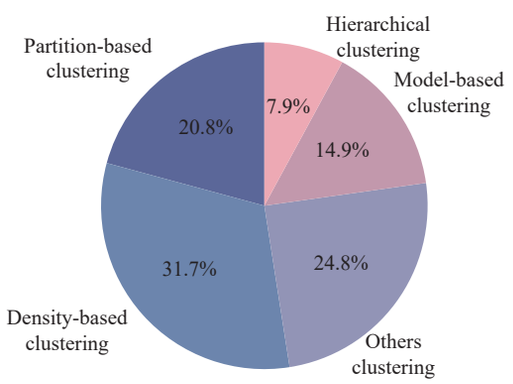

Fig. 1. Distribution of clustering methods used in astronomy.

3. A FEATURE-GUIDED STRATEGY FOR ALGORITHM SELECTION

Astronomical spectral data, because of their high-dimensional features, are exponentially sparsely distributed in high-dimensional space, bringing about the "curse of dimensionality", which makes distance metrics ineffective. Additionally, the complex noise in such data leads to non-stationary noise problems, reducing clustering stability. A third problem is that heterogeneous physical attributes result in violations of intra-cluster consistency, disrupting consistency across physical types. In real-world scenarios, these three issues frequently arise, posing a significant challenge to clustering.

In this section, we focus on the above three core challenges in astronomical spectral clustering and identify three key issues: high-dimensional feature compression[14,17,71,73,80]; noise robustness[13,36,39,64,70,85,87]; and sample and computational constraints[26,33,48,82]. We examine the adaptation of clustering strategies to different data structures, proposing an algorithm selection framework that emphasizes data structure adaptability and is driven by scientific task demands. This approach provides theoretical and methodological guidance for automated processing and discovery in complex astronomical spectral data.

3.1. Spectral Dimensionality and Feature Complexity

The high-dimensional nature of astronomical spectroscopic data leads to increased computational costs and weakened similarity measures, which both impact clustering performance. Distance-based methods struggle with inter-class distinctions, density-based methods face instability, and graph-based methods suffer from structural distortion due to sparse similarity matrices. These challenges affect various algorithms, with common issues like distance metric breakdown and algorithm-specific sensitivities.

To address these challenges, existing studies have explored solutions from two main directions: representation learning and dimensionality reduction. (1) In representation learning, Yang et al.[17] proposed the NAPC algorithm, which addresses the issue of ineffective distance metrics in high-dimensional data, using divergence-based distance measures to enhance local discriminability and incorporating an adaptive thresholding strategy. This makes the density peak clustering (DPC) algorithm more applicable to the LAMOST spectra, improving the clustering performance in high-dimensional spaces. Iwasaki et al.[71] combined Variational Autoencoders (VAE) with GMM. By leveraging nonlinear embeddings, they captured latent structures in complex spectral patterns, which effectively addressed the problem of similarity degradation from high dimensionality and achieved a 30% improvement in classification robustness. (2) Regarding dimensionality reduction strategies, for stellar continuum spectra dominated by linear features, the Principal Component Analysis (PCA)+K-means++ pipeline is effective in reducing problems arising from higher dimensionality and enabling more efficient clustering. The Factor Adjusted Spectral Clustering (FASC) method used by Tang et al.[14] further reduces redundant information through factor modeling. It helps maintain consistency across physical types by handling the heterogeneous physical attributes of data, keeping the misclassification rate below 1% provided the signal-to-noise ratio (SNR) is below 10. Probabilistic models such as GMM and the Normalized Gaussian Mixture Model (NGMM)[73,80] show strong adaptability to multimodal Gaussian distributions, such as gamma-ray burst (GRB) parameters, and can explicitly model covariance structures and cluster overlap, which is useful for dealing with the breakdown of intra-cluster consistency due to physical heterogeneity. Therefore, in high-dimensional spectral data, the performance of clustering strategies is highly dependent on the expressive power of the feature structure in the data being analyzed and the suitability of the dimensionality reduction methods. Effectively integrating dimensionality reduction, modeling approaches, and physical structure assumptions is key to enhancing clustering performance in high-dimensional spectral analysis.

Current research widely uses linear dimensionality reduction and deep embedding to address high-dimensional challenges; however, several approaches remain under-explored. Subspace clustering can uncover local

low-dimensional structures, addressing spectral overlap and multiple physical components. Sparse representation learning aids in feature compression and robust modeling, while graph-based methods, like graph neural networks, excel at modeling non-Euclidean relationships. Multi-scale embedding balances local and global feature extraction. Integrating these strategies offers more adaptive and scalable solutions for complex spectral data. Future research should focus on the structural adaptability of algorithms, analyzing the performance of different clustering methods, and promoting the integration of subspace modeling and deep embedding in astronomy.

3.2. Noise Characteristics and Robust Clustering

Noise is one of the most prevalent and complex problems with astronomical spectral data, especially in large-scale sky surveys such as SDSS and Gaia, where low SNR spectra are widespread. These noisy spectra significantly interfere with the recovery of clustering structures and the delineation of cluster boundaries. Spectral noise can be broadly categorized into static noise, high-frequency structured noise, and temporally evolving noise, and each type poses distinct challenges to the robustness of clustering algorithms.

In astronomical spectroscopic data, static noise obscures the overall signal, leading to blurred structural boundaries.

Fustes et al.[87] proposed a multi-scale strategy that combines Fast Fourier Transform (FFT), wavelet transformation, and Self-Organizing Maps (SOMs), enhancing feature representation and improving the recall rate of anomaly detection in Gaia by 25%. Wu et al.[39] combined PCA denoising with Clustering by Fast Search and Find of Density Peak (CFSFDP) clustering, significantly increasing the accuracy of rare object identification and improving recognition efficiency by 40%. These methods are well-suited for scenarios where signals are compressible and noise is evenly distributed.

High-frequency interference noise, characterized by locality and structural patterns, requires signal separation strategies. Seo et al.[85] working with Stratospheric Terahertz Observatory 2 (STO2) data, combined Asymmetric Least Squares (ALS) and Independent Component Analysis (ICA), using DBSCAN to eliminate outlier samples, achieving a recovery precision of 1–10 K. Lövdal et al.[36] used single-linkage hierarchical clustering and Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) to build a density tree structure, effectively identifying multi-density orbital debris layers.

Non-stationary, temporally evolving noise demands clustering models with dynamic adaptation capabilities. Yang et al.[13] proposed the TAD algorithm, which incorporates a noise tolerance factor and local density monitoring, and uses a sliding window mechanism to achieve incremental clustering updates of light curves, improving the F1-score to 0.93. The TimeTubesX system used by Sawada et al.[70], which integrates Dynamic Time Warping (DTW) and interactive visualization, enhanced the efficiency of detecting sudden events and anomalous trajectories by approximately 30%. Lapi et al.[64] introduced a time-dependent Fokker-Planck equation in cosmic evolution modeling, highlighting the importance of dynamic modeling in identifying dark matter halos at high redshifts. These studies, based on different types of noise, systematically demonstrate the strengths and applicable scenarios of specific algorithms and strategies in handling static, structured, and evolving noise. They not only expand the adaptation boundaries of clustering methods in astronomy but also provide support for building more robust clustering frameworks.

Different types of noise impact clustering robustness in distinct ways. Current methods target specific noise types, but several generalizable strategies remain underexplored. For static noise, robust feature learning methods like low-rank reconstruction and sparse autoencoders can filter background interference while preserving the data structure. High-frequency structured noise can be addressed using multi-resolution techniques like tensor decomposition and spectral filtering. Non-stationary, evolving noise requires dynamic clustering approaches, potentially enhanced by Markov modeling, graph-evolution clustering, and time-driven frameworks. Multimodal anomaly learning and joint reconstruction optimization are also promising methods. Overall, static noise benefits from dimensionality reduction and density estimation, high-frequency noise from signal separation, and evolving noise from dynamic modeling. Techniques like density landscape modeling and multi-scale kernel methods should be further explored.

3.3. Sample Size and Computational Constraints

The sparse samples in astronomical spectra (e.g., boundary spectra, rare celestial objects) are often insufficient for analysis, which limits the stability and interpretability of unsupervised clustering methods. The two main ways to address these challenges are to use semi-supervised clustering to introduce a small amount of labeled information, to guide the initial clustering structure, and to use transfer learning to transfer feature knowledge from external datasets to the target domain to enhance expressive capability.

Cai et al.[33] proposed an influence space-based clustering method that preprocesses boundary spectra by combining K-Nearest Neighbors (KNN) and Recurrent Neural Network (RNN) intersections to reduce noise and dimensionality. This approach improves initial cluster center selection and addresses ambiguous boundary clustering. Experiments on 20 000 LAMOST spectra achieved efficient computation via feature line extraction (for lines such as Hα and Hβ) and merged sorting with a time complexity of $O(n\log n)$. Pantoja et al.[82] employed a Uniform Manifold Approximation and Projection (UMAP) + hierarchical clustering strategy, achieving a 90% classification accuracy and an 80% subclass purity with only 5% labeled data.

These methods emphasize guiding optimization based on manifold structures and are suitable for spectral classification tasks with ambiguous structural boundaries or significant inter-class overlap.

By contrast, transfer learning focuses more on feature transfer rather than label guidance. Castro-Ginard et al.[48] pre-trained a deep neural network (DNN) on simulated data and applied it to the Gaia dataset. After applying DBSCAN clustering, they successfully identified 628 new star clusters, with an overall detection efficiency improvement of 50%. Balaguera-Antolínez transferred hierarchical features of dark matter halo mass functions to low-redshift observations[26], successfully reconstructing bias signals and reducing parameter errors by 30%. These studies highlight two approaches to improving clustering robustness in small-sample spectral data: label constraint guidance and cross-domain feature transfer. Semi-supervised methods optimize complex data structures with minimal labels, while transfer learning enables quick adaptation in sparse target domains through knowledge transfer. Both approaches complement each other, with semi-supervised methods enhancing similarity learning in labeled domains and transfer learning facilitating generalization across domains with shared structures. Future work can potentially combine both strategies to form a more robust joint modeling approach.

Small sample sizes and limited computational resources are key challenges in astronomical spectral clustering. Current methods primarily focus on semi-supervised and transfer learning approaches for sparse sample distributions, but many underused strategies remain worth exploring. For example, graph contrastive learning can build robust, structure-aware representations without labels, naturally adapting to sparse sample scenarios. Meta-learning methods, by learning fast adaptation capabilities, are suitable for efficient generalization in cases with few category samples. Additionally, low-rank approximation clustering and core set compression strategies can reduce computational overhead without sacrificing clustering accuracy, making them particularly useful for real-time and resource-sensitive observational tasks.

The evolutionary path of spectral clustering methods should not stop at algorithm stacking, but should be systematically designed, based on a three-dimensional collaborative system of feature structure-expression model-physical constraints. This approach not only provides methodological support for the intelligent processing of current large-scale spectroscopic databases but also lays the theoretical foundation for future astronomy data-driven research aimed at knowledge discovery.

4. APPLICATIONS OF CLUSTERING ALGORITHMS IN ASTRONOMY AND RELATED FIELDS

Different clustering algorithms have been applied to tackle key astronomical challenges such as stellar classification, galaxy structure, high-energy astrophysics, and exoplanet detection. Although some of these methods are not directly based on spectroscopic data, they use complementary approaches, such as physical property analysis, to provide a more comprehensive understanding of astronomical phenomena. By integrating these methods with spectral analysis, they further enhance the accuracy of clustering results and the potential for scientific discovery. These studies highlight the strengths and limitations of various techniques, emphasizing the need for more efficient and scalable solutions to handle large and complex datasets in astronomy.

Fig. 2 shows the distribution of clustering methods in six areas of astronomy. It highlights their diverse use in the handling of astronomical data analysis challenges.

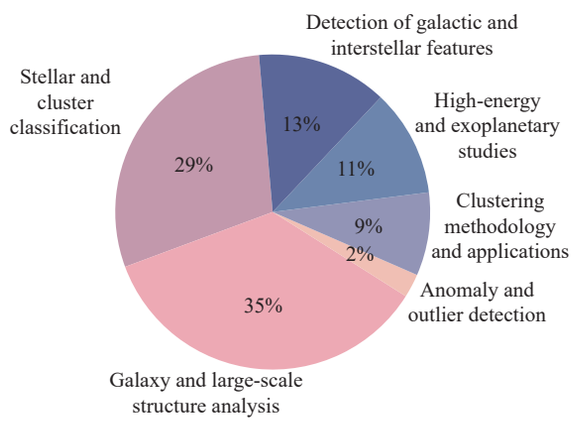

Fig. 2. Proportion of publications applying clustering methods to astronomical applications.

The data in Fig. 1 and Fig. 2 were gathered from the astronomical literature using the Astrophysics Data System (ADS). Based on our selection criteria, we focused on studies that used clustering published from 2014 to 2024. A total of 84 papers were selected; broken down by year: in 2014, there were 4 papers; in 2015, there were 5 papers; in 2016, there were 7 papers; in 2017, there were 5 papers; in 2018, there were 7 papers; in 2019, there were 6 papers; in 2020, there were 8 papers; in 2021, there were 4 papers; in 2022, there were 22 papers; in 2023, there were 4 papers; and, in 2024, there were 12 papers.

4.1. Stellar and Cluster Classification

In the field of star and cluster classification, current research trends increasingly favor the use of advanced algorithms and machine learning techniques, as well as data-driven exploration methods. These approaches can reveal hidden structures and evolutionary relationships, improve classification accuracy, and expand the overall body of knowledge. However, the field still faces many challenges, including ensuring correlation with spectroscopic data, dealing with data complexity, and rationalizing the scientific validity of algorithmic choices.

4.1.1. *Enhancing detection and classification efficiency*

Clustering algorithms play a crucial role in classifying stars and clusters within large datasets. For example, Ma et al.[47] proposed a parallel hybrid clustering algorithm for the screening of pulsar candidates, which analyzes the spectral features of pulsars. When applied to the High Time Resolution Universe Survey (HTRU) 2 and Actual Observation Data from FAST (AOD-FAST) datasets, the algorithm achieved precision and recall values of 0.946 and 0.905, and 0.787 and 0.994, respectively, demonstrating high accuracy. Sasdelli et al.[34] used deep learning to uncover diversity in supernova spectra, demonstrating the versatility of clustering and its contribution to stellar studies.

4.1.2. *Identification of stellar populations*

Clustering techniques have been remarkably effective in identifying stellar populations based on chemical abundances and morphological features. In the analysis of APOGEE data[30], Garcia-Dias et al. employed an unsupervised clustering algorithm to distinguish stellar populations according to their chemical abundance. It was found that DBSCAN achieved the best homogeneity score of 0.85. Ordovás-Pascual et al.[32] and Meusinger et al.[44] systematically classified star clusters using the K-means and Kohonen SOM algorithms. The former was tested on the SDSS-DR7 spectral database, and the results showed that the single-pass K-means algorithm is 20% to 40% faster than the conventional K-means algorithm, while the classifications are statistically equivalent. The latter demonstrated the effectiveness of SOM in identifying galaxies with similar spectral features. Moreover, Moranta et al.[37] used the HDBSCAN algorithm to cluster the Gaia Early Data Release 3 (Gaia EDR3) data to identify new stellar kinematic groups and cluster coronae. Lövdal et al.[36] conducted data-driven clustering using a single-link hierarchical clustering algorithm in the kinematic integral space of the stars. They identified 67 clusters and 232 sub-clusters. Price-Jones and Bovy[40] applied the DBSCAN algorithm to cluster stars. Their results indicated that DBSCAN could recover over 40% of the clusters with high homogeneity and completeness.

4.1.3. *Unsupervised classification of stellar populations*

Clustering algorithms are of great significance in the unsupervised classification of stellar populations, as demonstrated by various studies applying them to analyze different aspects such as astronomical object classification, star formation histories, and globular cluster analysis. Similarly, Thoresen et al.[49] not only proposed an unsupervised clustering method for lunar mineralogy research but also used a convolutional VAE to reduce the dimensionality of spectral data and then applied the K-means algorithm to cluster the latent variables into five different groups, which corresponded to the main mineral components on the lunar surface. They used hyperspectral data from the Moon Mineralogy Mapper (M3) to analyze the distribution of minerals. Logan and Fotopoulou[38] had previously used the HDBSCAN algorithm to classify stars, galaxies, and quasars. Using a dataset of approximately 50000 spectrally labeled objects, they achieved F1 scores of 98.9, 98.9, and 93.13 for stars, galaxies, and quasars, respectively. In the study of the orion region[42], clustering was applied to explore the star formation histories within the Orion OB Association. Chen et al.[41] used chemical dynamics clustering to analyze globular clusters in the APOGEE data. The results showed a recovery rate of 95.8% for M13, which demonstrated the effectiveness of this method in identifying members of globular clusters, deepening understanding of stellar population dynamics. Fig. 3 illustrates the stars identified by a new Shared Nearest Neighbors (SNN) clustering algorithm.

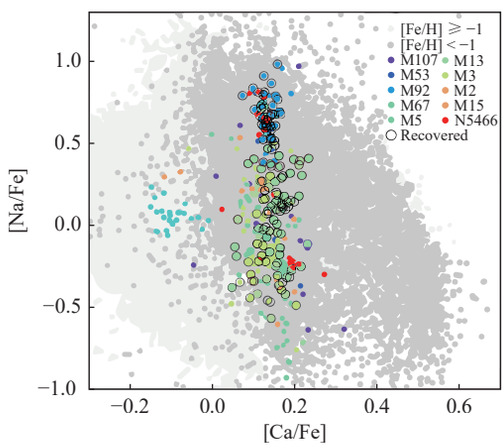

Fig. 3. Two-dimensional projections of abundance distribution using the SNN algorithm in chemical and radial velocity space. APOGEE background stars and known globular cluster members are shown, with SNN recovered members marked in black circles[41].

4.1.4. *Identification and analysis of new stellar populations*

Clustering algorithms have played a crucial role in identifying new stellar populations and giving insight into star cluster properties. Castro-Ginard et al.[48] applied the Open cluster (OC) finder algorithm to Gaia EDR3 data, resulting in the identification of 628 new open clusters, verifying the effectiveness of DBSCAN and giving insight into the star formation history of the Milky Way. Additionally, Tarricq et al.[51] employed the HDBSCAN algorithm to analyze the Gaia EDR3 data. By identifying a large number of cluster coronae and tidal tails, they demonstrated that HDBSCAN is highly effective when dealing with large-scale data.

4.1.5. *Advanced clustering in large-scale astronomical surveys*

Logan et al.[38] used the HDBSCAN algorithm to systematically classify 50000 celestial objects, achieving high-precision recognition of stars, galaxies, and quasars and demonstrating the powerful performance of unsupervised learning techniques in celestial object classification.

Similarly, Garcia-Dias et al.[30] effectively distinguished stellar populations based on chemical abundance features using clustering algorithms, while Zari et al.[42] employed clustering analysis of kinematics and age distributions of young stellar objects, revealing a complex star formation history. Blanco-Cuaresma et al.[27] verified the feasibility of this technique by analyzing spectroscopic data from open clusters. The clustering of stars using the K-means algorithm provides important clues for understanding the chemical evolution of the Milky Way (shown in Fig. 4). In addition, Chen et al.[5] used the K-means algorithm to cluster the line indices of 144340 A-type stars in LAMOST to identify anomalous spectra, studying the physical properties of the stars and their evolutionary patterns in depth. Together, these studies highlight the central role of clustering algorithms in stellar population analysis, object classification, and stellar evolution studies.

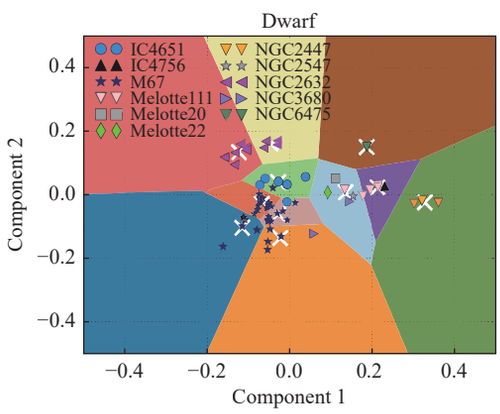

Fig. 4. Dwarf stars represented using the first two components of PCA[27]. Background colors correspond to the clusters found by the K-means algorithm. Centroids are marked with white crosses.

4.1.6. *Automated stellar spectral classification*

Recent advances in stellar and galaxy spectral classification demonstrate the growing significance of machine learning and clustering techniques. The SDSS-DR12 Bulk Stellar Spectral Classification[24] adopted Probabilistic Neural Network (PNN), Support Vector Machines (SVM), and K-means clustering to conduct automated classification of stellar spectra. Experimental results show that PNN outperforms K-means and SVM in automatic classification. After reducing the data dimensions using PCA, the classification errors of PNN, SVM, and K-means are 1.391, 1.529, and 1.654, respectively. PCA has also been used to reduce the data dimensionality for enhanced efficiency, and the PNN exhibited higher accuracy compared with other methods. Mosby and Tremonti et al.[35] introduced a diffusion K-means technique to retrieve the star formation history from galaxy spectra with a low SNR. This technique worked effectively for high-redshift objects and integrated field surveys, enabling better modeling of quasars and their host galaxies. Moreover, Olivares et al.[31] employed deep learning and clustering techniques to group a large number of stellar spectral models, aiming to reduce the search time and complexity. The experimental results show an accuracy of 85% and the algorithm execution time varies depending on the classified clusters, ranging from 6 to 13 minutes.

4.1.7. *Structural analysis of star clusters*

Gao[43] and Gao et al.[52] applied DBSCAN to identify members of NGC 188 (472 stars) and NGC 6819 (537 stars), using high-precision proper motion and radial velocity data. It effectively distinguished cluster members from field stars and assisted in determining physical parameters of the clusters, like those related to NGC 6819 (with the stellar distribution shown in Fig. 5). Meanwhile, Moe et al.[50] used K-means and k-Shape for clustering profiles. These studies display the robustness of DBSCAN in uncovering open cluster membership and structural features for precise galactic mapping.

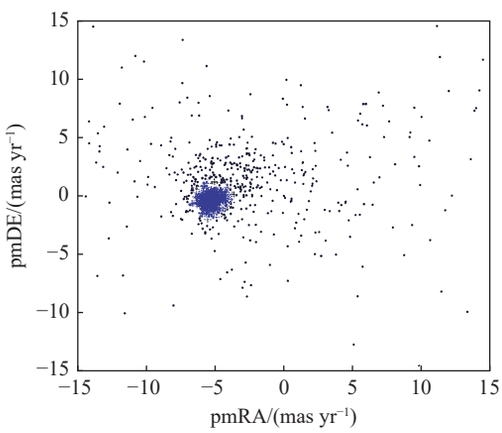

Fig. 5. Vector point diagram showing the proper motion of 472 member stars derived by the DBSCAN clustering algorithm[43]. The plus signs show 472 member stars, and the dots show 574 field stars.

4.1.8. *Large-scale stellar population analysis*

The LAMOST DR1 data release[3] plays a crucial role in stellar classification by providing an expansive dataset of stellar spectra. These spectra have been used in numerous clustering studies to understand stellar populations across the Milky Way. Furthermore, Hogg et al.[10] used a similar clustering method to identify stars, with experimental results showing an accuracy of about 0.04 dex, successfully identifying numerous known structures and identifying others of interest.

4.1.9. *Discovery and analysis of new star clusters*

In recent years, significant progress has been made in the study of evacuated clusters with the release of the Gaia EDR3. Castro-Ginard et al.[48] applied DBSCAN clustering algorithms to identify new open clusters in the Gaia EDR3 data, leading to the discovery of 628 new open clusters. These clusters were analyzed in terms of their age, distance, and extinction properties, with the

method proving especially effective for detecting both young and old star clusters located across the Milky Way.

4.2. Galaxy and Large-Scale Structure Analysis

In the field of galaxy and large-scale structure analysis, trends include multi-technique integration, a focus on dark matter and cosmology, and automation or optimization. This provided insight into the role of clustering in understanding complexity, galaxy properties, and cosmology, while also identifying challenges such as complex interactions, data-related issues, and modeling uncertainties.

4.2.1. *Galaxy structure and large-scale cosmology*

Secondary Halo Bias Through Cosmic Time II[26] and the re-creation of new high-redshift quasar populations[53] use clustering to analyze dark-matter halo properties and high-redshift quasar populations, respectively. Among these, the GMM approach reduces the acceptance rate of contaminants by 86% while retaining a similar number of candidates for quasars.

4.2.2. *Galaxy morphology and dynamics*

Clustering and dimensionality reduction techniques play a pivotal role in the analysis of galaxy morphology and dynamics. Rosito et al.[65] integrated PCA with clustering algorithms, achieving precise classification of galactic motion structures. Karademir et al.[57] used clustering techniques to significantly enhance the accuracy of redshift estimation, while Rahmani et al.[20] effectively classified galaxy spectra using clustering algorithms, with the results illustrated in Fig. 6.

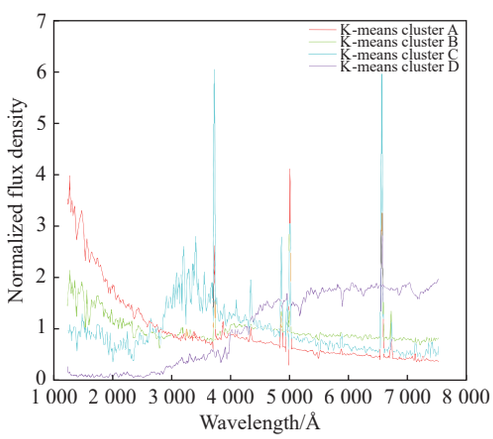

Fig. 6. Clustering the K96 template spectra using K-means clustering[20]. Note that both methods randomly assign the initial values for their analysis.

Fielding et al.[67] and Li et al.[54] enhanced the clustering accuracy using a novel initial cluster center selection algorithm. Experiments conducted on 12 typical benchmark datasets and the LAMOST survey spectral data indicate that this novel algorithm outperforms the other six algorithms in initial cluster center selection. For example, when applied to the Iris dataset, it achieves an accuracy of 0.9533, a precision of 0.9544, and a recall of 0.9533. This algorithm simplifies the data processing flow via automatic galaxy morphology identification and an optimized cluster center selection strategy. Moreover, the ICCS K-means algorithm shows greater advantages than the other algorithms in terms of time efficiency.

Finally, Tramacere et al.[58] applied the DBSCAN and Density-based Clustering (DENCLUE) algorithms to the Galaxy Zoo 2 dataset. The experimental results revealed an accuracy of 93%, with a precision of 92.6% and a recall of 93.5%. This validates the practical application value of these methods in galaxy classification research.

Clustering techniques have been instrumental in enhancing our understanding of galaxy properties and dynamics. For example, Beck et al.[55] used K-means for spectral classification to analyze the correlation between emission lines and stellar continuum spectra of galaxies, and the experimental results showed that only about 6% of the samples were misclassified, which provides valuable information for us to understand the spectral properties of galaxies and their intrinsic dynamics. Similarly, clustering the color distributions of point sources in the Messier 83 galaxy[56] helps to identify distinct stellar populations and star clusters, yielding information into the star formation history and spatial distribution of star populations in this galaxy.

4.2.3. *Galaxy clustering and halo mass function*

Clustering research has significantly advanced our understanding of galaxy clustering properties and halo mass functions. DeRose et al.[11] used the SHAM model for redshift-space clustering of spectroscopic data, significantly improving the efficiency of the analysis.

In addition, through spectroscopic classification of the clustering properties of AGN, Córdova et al.[25] showed that unobscured Broad Line AGN (BL AGN) have an average halo mass approximately 5.5 times higher than that of obscured AGN. Additionally, Eltvedt et al.[68] offer insights into quasar halo masses and Li et al.[54] improved clustering accuracy in large datasets. González-Morán et al.[62] used an unsupervised machine learning clustering method, GMM, to classify the spectral energy distributions (SEDs) of galaxies in the PAU Survey. Collectively, these studies demonstrate the role of clustering in understanding galaxy evolution, halo mass, and dark-matter structures, which are essential for modeling large-scale cosmic structures.

4.2.4. *Cosmological structure and redshift studies*

Clustering techniques can be applied effectively to the precise measurement of redshifts and the analysis of large-scale cosmological structures. In the realm of redshift space corrections, research on relativistic redshift space distortions[28] has significantly enhanced correction accuracy. For high redshift quasar studies, Wagenveld et al.[53] effectively screened candidate samples and improved redshift measurement precision through GMM.

4.3. Detection of Galactic and Interstellar Features

In the field of detecting galactic and interstellar features, trends include expanding clustering applications, combining them with other techniques, and using them in large-scale data analysis. Insights are used to uncover hidden structures, understand galactic evolution, and support theoretical models. The challenges involve the complexity of celestial objects, data quality and quantity, and interpretation of results.

4.3.1. *Detection of tidal debris and galactic halo studies*

Density-based clustering algorithms, such as OPTICS and DBSCAN, play a pivotal role in identifying tidal debris and classifying stellar structures, providing essential tools for reconstructing galaxy merger histories. Fuentes et al.[15] and Sapozhnikov et al.[16] applied the OPTICS and DBSCAN algorithms to successfully detect stellar streams and tidal debris, offering unique insight into the formation history of the Milky Way. Furthermore, Iwasaki et al.[71] clustered X-ray spectral data of the Tycho supernova remnant using a VAE in combination with a GMM. Gao[19] combined PCA with GMM, identifying 2301 stars potentially associated with NGC 2506, of which 147 are likely to be on the tidal tail, providing dynamical interactions of clusters and mass-loss mechanisms together with new insights. Oliver et al.[21] effectively identified tidal stream structures by integrating kinematic and metallicity data, advancing our understanding of the formation processes of galactic halos.

4.3.2. *Hierarchical clustering in galactic and dust cloud structures*

Hierarchical clustering techniques provide a critical perspective for understanding the multi-scale structure of the universe. Kheirdastan et al.[24] conducted experiments showing that the classification errors are 1.391 for PNN, 1.529 for SVM, and 1.654 for K-means. Hierarchical clustering methods have validated the hierarchical formation theory of cosmic structures. Yu et al.[74] discussed examples of the application of hierarchical clustering to the processing of astronomical spectroscopic data and systematically summarized wide-ranging applications of this technique on scales from asteroids to galaxies. A clustering analysis of KiDS-DR3 data using the AMICO algorithm[76] has further refined constraints on cosmological parameters.

Hierarchical clustering techniques have also been used to map interstellar dust clouds. The Spectral Clustering for Interstellar Molecular Emission Segmentation (SCIMES) algorithm for clustering molecular clouds[29] also helps characterize their physical properties and star formation potential.

4.4. High-Energy and Exoplanetary Studies

In high-energy and exoplanetary studies, trends show the wide application of clustering. Insights have been revealed into its role in uncovering patterns and distinguishing signals, while challenges identified include accurate classification and handling high-dimensional data.

4.4.1. *Exoplanet detection and classification*

Jin et al.[83] demonstrated how clustering can identify and classify exoplanet candidates, uncovering new patterns among exoplanetary systems. Such studies highlight clustering's role in advancing knowledge in exoplanetary and high-energy astronomy.

4.4.2. *Gamma-ray and high-energy studies*

Clustering methods play a pivotal role in high-energy astrophysics, particularly in areas such as gamma-ray burst identification; for example, Mehta et al.[80] analyzed the spectral properties of gamma-ray bursts (GRBs) using an unsupervised clustering algorithm (Nested Gaussian Mixture Model, NGMM). Currently, Armstrong et al.[61] are applying DBSCAN to data from the Fermi Large Area Telescope (Fermi-LAT), achieving precise identification of AGN and new radiation sources. Additionally, Tóth et al.[72] conducted systematic studies on magnetars and GRBs using clustering techniques, substantially advancing our understanding of cosmic high-energy explosive phenomena and extreme physical processes.

4.4.3. *Planetary, exoplanet, and solar studies*

Clustering techniques play an indispensable role in planetary science and solar system research. Guez and Claire[79] proposed a method for clustering spectra of planetary atmospheres that is independent of specific molecular features. The spectra are clustered using the HDBSCAN algorithm. Experimental results show that Molecule Agnostic Spectral Clustering (MASC) can effectively separate data at low resolution. Pantoja et al.[82] proposed a combination of semi-supervised hierarchical classification and clustering analysis for variable star classification. Furthermore, Hayes et al.[77] used transmission spectra to train a classifier and generate a priori information for atmospheric inversion by combining PCA with the K-means algorithm. At $R = 100$ with a 1% noise level, the number of iterations using the classification method was reduced by 41%, saving approximately 3.3 hours of CPU time.

4.4.4. *Supernova detection and classification*

Sasdelli et al.[34] applied the K-means algorithm for unsupervised learning, to analyze the spectra of type Ia supernovae, with their findings suggesting that subtypes may form a continuous distribution rather than discrete categories. Fig. 7 illustrates the isomap space with K-means clusters, showcasing the effectiveness of combining deep learning with unsupervised algorithms for high-dimensional data visualization. Similarly, Rubin et al.[78] applied K-means clustering to classify supernova light curves, contributing to the understanding of stellar life cycles and energetic processes in high-energy astrophysics.

4.5. Clustering Methodology and Applications

Research trends indicate the continued emergence and development of new methodologies with respect to employing clustering for stellar or galactic research. Insights

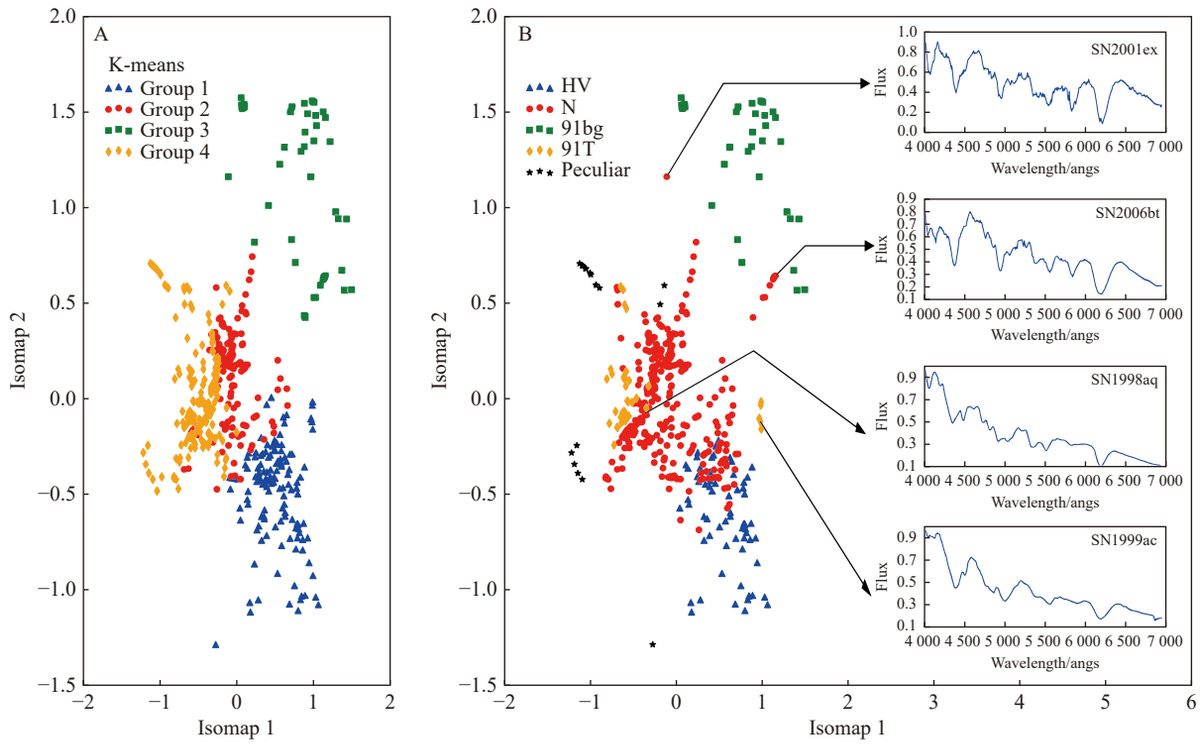

Fig. 7. Four-dimensional feature space from a deep learning model reduced to two dimensions using isomap[34]. (A) Groups found by the K-means algorithm when four groups are imposed. (B) Objects separated according to the classification proposed by Wang et al.[59].

include improved analysis and task support, while the main challenge is choosing the most appropriate technique.

4.5.1. *Stellar and galactic research*

Clustering methods are vital for analyzing multidimensional astronomical data. Chattopadhyay et al.[81] studied mixture models in astronomy. The TAD Algorithm[13] identifies key "stay points" in celestial object trajectories, where stay points refer to specific positions or regions at which celestial objects remain stationary or exhibit prolonged stays. The algorithm uses Neighborhood Move Ability and Stay Time (NMAST) density functions and Noise Tolerance (NT) factors for spatial-temporal modeling. Applied to LAMOST data, it supports efficient observation planning.

4.5.2. *Overview of clustering techniques*

Saxena et al.[22] provides a comprehensive overview of clustering methods, guiding astronomers in selecting the best techniques for tasks such as galaxy classification, cluster identification, spectral analysis, and streamlining research in various astronomical fields. Different clustering methods reportedly show different results when dealing with high-dimensional and noisy data. For example, hierarchical clustering has high computational complexity when dealing with large-scale datasets, while partitional clustering such as K-means performs better when dealing with spherical clusters.

4.6. Anomaly and Outlier Detection

In the field of anomaly and outlier detection in astronomy, trends show the use of improved algorithms for data refinement. Insights are that outlier detection is crucial for data accuracy and discovery, while challenges include the effective application of clustering algorithms to different areas for the detection of outliers and anomalies.

4.6.1. *Detection of outliers and anomalies*

Fustes et al.[87] used a novel SOM ensemble approach for unsupervised outlier analysis of astronomical spectroscopic data in the Gaia Survey. It improves outlier detection by combining multiple SOM models, enabling the identification of rare objects. Additionally, Tiwari et al.[86] applied the hierarchical K-means clustering method to cluster in the PCA feature space and identify quasar spectral anomalies in SDSS DR16. In a study on carbon stars[88], an outlier detection method based on morphological feature extraction and interval representation was used to identify 88 carbon stars with emission lines from 3546 spectra, providing insights into their stellar activity.

Fig. 8 illustrates the evolution of clustering applications in astronomy from 2014 to 2024 in six distinct areas: stellar and cluster classification, galaxy and large-scale structure analysis, detection of galactic and interstellar features, high-energy and exoplanetary studies, clustering methodology, and anomaly detection. The figure highlights the most highly cited literature in each respective year.

5. DISCUSSION

This section focuses on astronomy clustering algo-

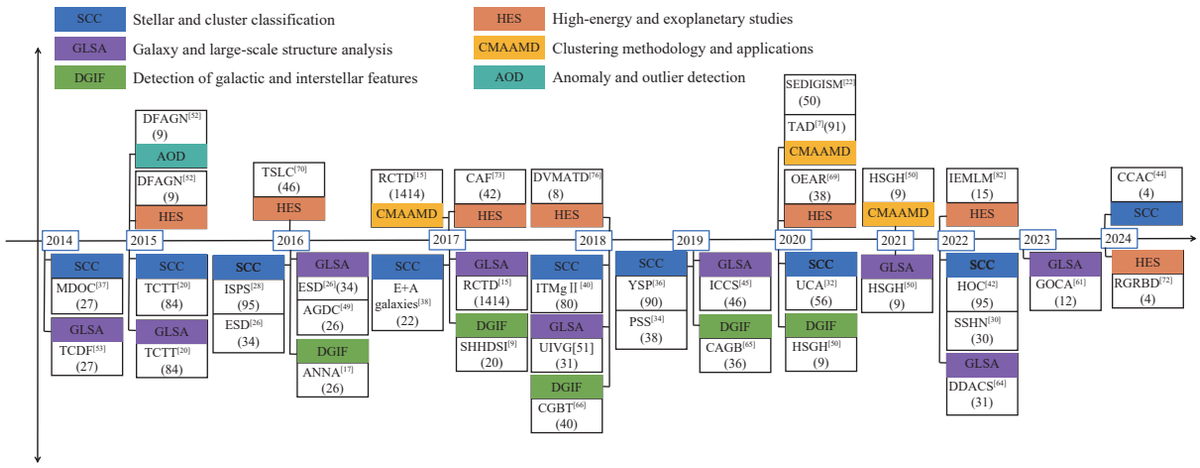

Fig. 8. Progress of clustering methods in astronomy (2014–2024), where square brackets indicate cited literature and parentheses indicate the number of citations. The abbreviations for the titles are derived from the respective paper titles. The cut-off year for the cited papers is 2024, with these data being gathered from the Google Scholar website.

rithms, their role in handling complex data, and methods like hierarchical clustering and DBSCAN. It also addresses data quality and scalability, and calls for better techniques for sparse, unlabeled data. Future research should focus on semi-supervised learning and Large Language Model (LLM) integration, considering current achievements and challenges.

5.1. Research Limitations and Challenges in Future Clustering Applications

Although clustering techniques in astrophysics hold immense promise, they face significant challenges. A major hindrance is that the scarcity of high-quality observational data, such as high-resolution spectra and precise kinematic information, limits the accuracy of clustering analyses. Another concern is that the exponential growth in dataset sizes has led to a dramatic increase in computational complexity, particularly evident in the data processing of large-scale surveys like Euclid and the Large Synoptic Survey Telescope (LSST). Traditional clustering methods, such as K-means, often underperform when handling non-spherical clusters and imbalanced data distributions, falling short of the precision required for modern astrophysical research. Moreover, the severe lack of labeled data for rare celestial objects, such as high-redshift quasars, underscores the urgent need for semi-supervised and unsupervised learning approaches. To address these challenges, the development of novel clustering algorithms is imperative. These algorithms should be able to efficiently process large-scale datasets, adapt to diverse data distributions, integrate multi-source heterogeneous data, and be capable of handling dynamic streaming data. Only through such algorithmic innovations can fully harness the physical insights embedded in astronomical data, driving transformative advancements in astrophysics.

5.2. Future Research Directions and Performance Variations of Clustering Methods

Future astronomical clustering research should focus on improving multimodal data analysis methods in four key areas: scalability, robustness, accuracy, and generalizability. Currently, various clustering methods have distinct characteristics. PCA-based hierarchical clustering offers high precision but is computationally intensive; K-means and DBSCAN are highly efficient but require parameter tuning; density-based methods are fast but may compromise accuracy; and DBSCAN for solar filament detection[89] is highly sensitive to data quality. LLMs have shown great potential in astronomical analysis. They can handle large-scale and diverse astronomical data, assist in feature extraction, and improve the accuracy of astronomical analysis. For example, in the task of galaxy classification, the StarWhisper LightCurve (LC) series models—including three LLM-based models: the Large Language Model (LLM), the Multimodal Large Language Model (MLLM), and the Large Astronomy Language Model (LALM)—significantly reduce the reliance on explicit feature engineering[90]. With their development, they are expected to bring revolutionary changes to astronomical clustering analysis, promote precise classification research, help discover new astronomical phenomena, and can also be combined with gravitational wave detection technology for real-time analysis. In this context, Wu et al.[91] designed an evaluation framework and collected data with the help of a Slack chatbot based on Retrieval-Augmented Generation (RAG) to provide support for the evaluation and improvement of LLMs and promote the development of efficient algorithms. Fouesneau et al.[92] explored various application scenarios of LLMs in astronomical research through experiments and surveys, clarified their advantages and disadvantages, and provided practical guidance. Li et al.[90] used deep learning and LLMs to conduct classification research on variable star light curves, demonstrating the ability of LLMs to handle astronomical time series data. Pan et al.[93] quantitatively evaluated LLMs in the field of astronomy, proposed new models, and analyzed factors affecting performance, provid-

ing a theoretical basis for optimization. Smirnov[94] demonstrated the excellent capabilities of LLMs in tasks such as asteroid classification, providing new ideas for their application in astronomical pattern recognition.

To address these challenges—including limited multimodal data integration, poor scalability, low robustness, insufficient generalization, and the early-stage adoption of LLMs—research efforts should shift toward semi-supervised learning, unsupervised learning, and hybrid deep clustering techniques while integrating active learning, cloud computing, and LLMs to enhance pattern recognition capabilities. Specifically, priority should be given to exploring the deep integration of LLMs with traditional clustering algorithms, leveraging their strengths in natural language processing and knowledge reasoning. Building on their successful applications in astronomical literature analysis and code generation, these approaches can be extended to clustering scenarios such as variable star light curves and gravitational wave data. Furthermore, a systematic evaluation framework should be established to guide methodological optimization and ensure robust performance across diverse astronomical datasets.

6. CONCLUSION

Clustering algorithms play a pivotal role in the analysis of astronomical survey datasets and multidisciplinary research. Here, we systematically review various clustering methods, highlighting their strengths and limitations in practical applications. Traditional approaches, such as DBSCAN's limitations in low-density regions and the high computational costs of hierarchical methods, struggle to meet the demands of large-scale sky surveys, particularly in low-signal scenarios like exoplanet research, because of challenges posed by data dimensionality, noise, and sparsity. When selecting clustering algorithms, it is necessary to fully consider the characteristics of spectral data, such as resolution, noise level and sample size. For high-dimensional data, algorithms like NAPC can be used in combination with dimensionality reduction and initial-center-optimization strategies. For noisy data, algorithms such as SOM integration can be chosen according to their characteristics. For small-sample data, methods like hierarchical clustering with migration can be considered.

Future research should focus on developing more efficient, scalable and adaptive clustering techniques, including kernel spectral clustering, hybrid methods and cloud-based processing platforms. Meanwhile, semi-supervised and unsupervised learning are crucial for solving the sparsely labeled data problem. As astronomical datasets continue to grow, clustering analysis will remain a fundamental approach. Future technological innovations must prioritize the dynamic nature of astronomical data, which will not only deepen our understanding of the universe, but also expand the scope of clustering applications in astronomy.

ACKNOWLEDGEMENTS

We express our gratitude to StarWhisper (Multi-Workflow AI Writing) for its invaluable assistance in the writing process, which greatly contributed to the quality and efficiency of this work.

The work was supported by the National Natural Science Foundation of China (12473105 and 12473106), the central government guides local funds for science and technology development (YDZJSX2024D049), and the Graduate Student Practice and Innovation Program of Shanxi Province (2024SJ313).

AI DISCLOSURE STATEMENT

StarWhisper and Doubao was employed for language polishing, translation, and standardization checks within the article. The authors conducted rigorous human review and gave final approval to all AI modifications, assuming ultimate responsibility for the content of the publication.

AUTHOR CONTRIBUTIONS

Jianing Tian implemented the research and wrote the paper. Haifeng Yang and Jianghui Cai provided guidance on topic selection and paper writing, and reviewed the paper. Yuqing Yang reviewed the paper and improved the English quality. Xiangru Li, Zhenping Yi, and Lili Wang reviewed it and gave guidance suggestions. All authors read and approved the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

[1] York, D. G., Adelman, J., Anderson Jr, J. E., et al. 2000. The sloan digital sky survey: Technical summary. *The Astronomical Journal*, **120**(3): 1579.

[2] Cui, X. Q., Zhao, Y. H., Chu, Y. Q., et al. 2012. The large sky area multi-object fiber spectroscopic telescope (LAMOST). *Research in Astronomy and Astrophysics*, **12**(9): 1197.

[3] Luo, A. L., Zhao, Y. H., Zhao, G., et al. 2015. The first data release (DR1) of the LAMOST regular survey. *Research in Astronomy and Astrophysics*, **15**(8): 1095.

[4] Ahn, C. P., Alexandroff, R., Prieto, C. A., et al. 2014. The tenth data release of the sloan digital sky survey: First spectroscopic data from the sdss-iii apache point observatory galactic evolution experiment. *The Astrophysical Journal Supplement Series*, **211**(2): 17.

[5] Chen, S. X., Sun, W. M., Yan, Q. 2018. Clustering analysis of line indices for LAMOST spectra with AstroStat. *Research in Astronomy and Astrophysics*, **18**(6): 073.

[6] Dehghan, S., Johnston-Hollitt, M. 2014. Clusters, groups, and filaments in the Chandra deep field–South up to redshift 1. *The Astronomical Journal*, **147**(3): 52.

[7] Ester, M., Kriegel, H. P., Sander, J., et al. 1996. A density-

based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of KDD-96.

[8] Kohonen, T. 1982. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, **43**(1): 59−69.

[9] Stauffer, C., Grimson, W. E. L. 1999. Adaptive background mixture models for real-time tracking. In Proceedings of 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition.

[10] Hogg, D. W., Casey, A. R., Ness, M., et al. 2016. Chemical tagging can work: Identification of stellar phase-space structures purely by chemical-abundance similarity. *The Astrophysical Journal*, **833**(2): 262.

[11] DeRose, J., Becker, M. R., Wechsler, R. H. 2022. Modeling redshift-space clustering with abundance matching. *The Astrophysical Journal*, **940**(1): 13.

[12] Carlson, E., Linden, T., Profumo, S., et al. 2013. Clustering analysis of the morphology of the 130 GeV gamma-ray feature. *Physical Review D—Particles, Fields, Gravitation, and Cosmology*, **88**(4): 043006.

[13] Yang, Y. Q., Cai, J. H., Yang, H. F., et al. 2020. TAD: A trajectory clustering algorithm based on spatial-temporal density analysis. *Expert Systems with Applications*, **139**: 112846.

[14] Tang, S. G., Jana, S., Fan, J. Q. 2024. Factor adjusted spectral clustering for mixture models. *arXiv: 2408.12564*.

[15] Fuentes, S. S., De Ridder, J., Debosscher, J. 2017. Stellar halo hierarchical density structure identification using (F) OPTICS. *Astronomy & Astrophysics*, **599**: A143.

[16] Sapozhnikov, S., Kovaleva, D. 2020. Clustering stellar pairs to detect extended stellar structures. *Proceedings of the International Astronomical Union*, **16**(S362): 150−151.

[17] Yang, Y. Q., Cai, J. H., Yang, H. F., et al. 2022. Density clustering with divergence distance and automatic center selection. *Information Sciences*, **596**: 414−438.

[18] Yang, H. F., Shi, C. H., Cai, J. H., et al. 2022. Data mining techniques on astronomical spectra data–I. Clustering analysis. *Monthly Notices of the Royal Astronomical Society*, **517**(4): 5496−5523.

[19] Gao, X. H. 2020. Discovery of tidal tails around the old open cluster NGC 2506. *The Astrophysical Journal*, **894** (1): 48.

[20] Rahmani, S., Teimoorinia, H., Barmby, P. 2018. Classifying galaxy spectra at $0.55 < z < 1$ with self-organizing maps. *Monthly Notices of the Royal Astronomical Society*, **478**(4): 4416−4432.

[21] Oliver, W. H., Elahi, P. J., Lewis, G. F., et al. 2021. The hierarchical structure of galactic haloes: classification and characterization with HALO-OPTICS. *Monthly Notices of the Royal Astronomical Society*, **501**(3): 4420−4437.

[22] Saxena, A., Prasad, M., Gupta, A., et al. 2017. A review of clustering techniques and developments. *Neurocomputing*, **267**: 664−681.

[23] Guo, X. Y., Liu, C. X., Qiu, B., et al. 2022. Unsupervised clustering and analysis of WISE spiral galaxies. *Monthly Notices of the Royal Astronomical Society*, **517**(2): 1837−1848.

[24] Kheirdastan, S., Bazarghan, M. 2016. SDSS-DR12 bulk stellar spectral classification: Artificial neural networks approach. *Astrophysics and Space Science*, **361**(9): 304.

[25] Córdova Rosado, R., Goulding, A. D., Greene, J. E., et al. 2024. Cross-correlation of Luminous Red Galaxies with ML-selected AGN in HSC-SSP II: AGN classification and clustering with DESI spectroscopy. *arXiv: 2410.24020*.

[26] Balaguera-Antolínez, A., Montero-Dorta, A. D. 2024. Secondary halo bias through cosmic time: II. Reconstructing halo properties using clustering information. *Astronomy & Astrophysics*, **692**: A32.

[27] Blanco-Cuaresma, S., Soubiran, C., Heiter, U., et al. 2015. Testing the chemical tagging technique with open clusters. *Astronomy & Astrophysics*, **577**: A47.

[28] Euclid Collaboration, Elkhashab, M. Y., Bertacca, D., et al. 2024. Euclid preparation. The impact of relativistic redshift-space distortions on two-point clustering statistics from the Euclid wide spectroscopic survey. *arXiv: 2410.00956*.

[29] Duarte-Cabral, A., Colombo, D., Urquhart, J., et al. 2021. The SEDIGISM survey: Molecular clouds in the inner Galaxy. *Monthly Notices of the Royal Astronomical Society*, **500**(3): 3027−3049.

[30] Garcia-Dias, R., Prieto, C. A., Almeida, J. S., et al. 2019. Machine learning in APOGEE- Identification of stellar populations through chemical abundances. *Astronomy & Astrophysics*, **629**: A34.

[31] Olivares, E., Curé, M., Araya, I., et al. 2024. Estimation of physical stellar parameters from spectral models using deep learning techniques. *Mathematics*, **12**(20): 3169.

[32] Ordovás-Pascual, I., Sánchez Almeida, J. 2014. A fast version of the k-means classification algorithm for astronomical applications. *Astronomy & Astrophysics*, **565**: A53.

[33] Cai, J. H., Li, Y. T., Yang, H. F. 2020. A new method for clustering of boundary spectra. *Journal of Astrophysics and Astronomy*, **41**: 15.

[34] Sasdelli, M., Ishida, E., Vilalta, R., et al. 2016. Exploring the spectroscopic diversity of Type Ia supernovae with DRACULA: a machine learning approach. *Monthly Notices of the Royal Astronomical Society*, **461**(2): 2044−2059.

[35] Mosby Jr, G., Tremonti, C., Hooper, E., et al. 2015. Simple stellar population modelling of low S/N galaxy spectra and quasar host galaxy applications. *Monthly Notices of the Royal Astronomical Society*, **447**(2): 1638−1660.

[36] Lövdal, S. S., Ruiz-Lara, T., Koppelman, H. H., et al. 2022. Substructure in the stellar halo near the Sun-I. Data-driven clustering in integrals-of-motion space. *Astronomy & Astrophysics*, **665**: A57.

[37] Moranta, L., Gagné, J., Couture, D., et al. 2022. New coronae and stellar associations revealed by a clustering analysis of the solar neighborhood. *The Astrophysical Journal*, **939**(2): 94.

[38] Logan, C., Fotopoulou, S. 2020. Unsupervised star, galaxy, QSO classification-application of HDBSCAN. *Astronomy & Astrophysics*, **633**: A154.

[39] Wu, M. L., Pan, J. C., Yi, Z. P., et al. 2020. Rare object search from Low-S/N stellar spectra in SDSS. *IEEE Access*, **8**: 66475−66488.

[40] Price-Jones, N., Bovy, J. 2019. Blind chemical tagging with DBSCAN: prospects for spectroscopic surveys. *Monthly Notices of the Royal Astronomical Society*, **487**(1): 871−886.

[41] Chen, B. Q., D'Onghia, E., Pardy, S. A., et al. 2018. Chemodynamical clustering applied to APOGEE data: rediscovering globular clusters. *The Astrophysical Journal*, **860**(1): 70.

[42] Zari, E., Brown, A., De Zeeuw, P. 2019. Structure, kinematics, and ages of the young stellar populations in the Orion region. *Astronomy & Astrophysics*, **628**: A123.

[43] Gao, X. H. 2014. Membership determination of open cluster ngc 188 based on the dbscan clustering algorithm. *Research in Astronomy and Astrophysics*, **14**(2): 159.
[44] Meusinger, H., Brünecke, J., Schalldach, P., et al. 2017. A large sample of Kohonen selected E+A (post-starburst) galaxies from the Sloan Digital Sky Survey. *Astronomy & Astrophysics*, **597**: A134.
[45] Merényi. E., Taylor, J., Isella, A. 2016. IEEE symposium series on computational intelligence (SSCI). Greece: IEEE.
[46] Panos, B., Kleint, L., Huwyler, C., et al. 2018. Identifying typical Mg II flare spectra using machine learning. *The Astrophysical Journal*, **861**(1): 62.
[47] Ma, Z., You, Z. Y., Liu, Y., et al. 2022. A preliminary study of large scale pulsar candidate sifting based on parallel hybrid clustering. *Universe*, **8**(9): 461.
[48] Castro-Ginard, A., Jordi, C., Luri, X., et al. 2022. Hunting for open clusters in Gaia EDR3: 628 new open clusters found with OCfinder. *Astronomy & Astrophysics*, **661**: A118.
[49] Thoresen, F., Drozdovskiy, I., Cowley, A., et al. 2024. Insights into lunar mineralogy: An unsupervised approach for clustering of the moon mineral mapper (M3) spectral data. *arXiv: /2411.03186.*
[50] Moe, T. E., Pereira, T. M., van der Voort, L. R., et al. 2024. Comparative clustering analysis of Ca II 854.2 nm spectral profiles from simulations and observations. *Astronomy & Astrophysics*, **682**: A11.
[51] Tarricq, Y., Soubiran, C., Casamiquela, L., et al. 2022. Structural parameters of 389 local open clusters. *Astronomy & Astrophysics*, **659**: A59.
[52] Gao, X. H., Xu, S. K., Chen, L. 2015. 3D cluster members and near-infrared distance of open cluster NGC 6819. *Astronomy & Astrophysics*, **15**(12): 2193.
[53] Wagenveld, J., Saxena, A., Duncan, K., et al. 2022. Revealing new high-redshift quasar populations through Gaussian mixture model selection. *Astronomy & Astrophysics*, **660**: A22.
[54] Li, Y. T., Cai, J. H., Yang, H. F., et al. 2019. A novel algorithm for initial cluster center selection. *IEEE Access*, **7**: 74683−74693.
[55] Beck, R., Dobos, L., Yip, C. W., et al. 2016. Quantifying correlations between galaxy emission lines and stellar continua. *Monthly Notices of the Royal Astronomical Society*, **457**(1): 362−374.
[56] Kiar, A. K., Barmby, P., Hidalgo, A. 2017. Deconstructing a galaxy: Colour distributions of point sources in Messier 83. *Monthly Notices of the Royal Astronomical Society*, **472**(1): 1074−1087.
[57] Karademir, G. 2022. The galaxy luminosity function via clustering based redshift inference: can we find the bottom of the galaxy population. In Proceeding of Hypatia Colloquium.
[58] Tramacere, A., Paraficz, D., Dubath, P., et al. 2016. Asterism: Application of topometric clustering algorithms in automatic galaxy detection and classification. *Monthly Notices of the Royal Astronomical Society*, **463**(3): 2939−2957.
[59] Wang, X., Filippenko, A. V., Ganeshalingam, M., et al. 2009. Improved distances to Type Ia supernovae with two spectroscopic subclasses. *The Astrophysical Journal Letters*, **699**: L139−L143.
[60] Mahajan, S., Singh, A., Shobhana, D. 2018. Ultraviolet and optical view of galaxies in the Coma supercluster. *Monthly Notices of the Royal Astronomical Society*, **478**(4): 4336−4347.
[61] Armstrong, T., Brown, A. M., Chadwick, P. M., et al. 2015. The detection of Fermi AGN above 100 GeV using clustering analysis. *Monthly Notices of the Royal Astronomical Society*, **452**(3): 3159−3166.
[62] González-Morán, A., Arrabal Haro, P., Muñoz-Tuñón, C., et al. 2023. The PAU survey: classifying low-z SEDs using Machine Learning clustering. *Monthly Notices of the Royal Astronomical Society*, **524**(3): 3569−3581.
[63] Chen, L. H., Hartwig, T., Klessen, R. S., et al. 2022. Comparing simulated Milky Way satellite galaxies with observations using unsupervised clustering. *Monthly Notices of the Royal Astronomical Society*, **517**(4): 6140−6149.
[64] Lapi, A., Ronconi, T., Danese, L. 2022. A stochastic theory of the hierarchical clustering. III. the nonuniversality and nonstationarity of the halo mass function. *The Astrophysical Journal*, **941**(1): 14.
[65] Rosito, M., Bignone, L. A., Tissera, P., et al. 2023. Application of dimensionality reduction and clustering algorithms for the classification of kinematic morphologies of galaxies. *Astronomy & Astrophysics*, **671**: A19.
[66] Romanello, M., Marulli, F., Moscardini, L., et al. 2025. Tomographic cluster clustering as a cosmological probe. *Astronomy & Astrophysics*, **693**: A195.
[67] Fielding, E., Nyirenda, C. N., Vaccari, M. 2022. The classification of optical galaxy morphology using unsupervised learning techniques. In Proceedings of 2022 International Conference on Electrical, Computer and Energy Technologies (ICECET).
[68] Eltvedt, A. M., Shanks, T., Metcalfe, N., et al. 2024. The VST ATLAS Quasar Survey-III. Halo mass function via quasar clustering and quasar-CMB lensing cross-clustering. *Monthly Notices of the Royal Astronomical Society*, **535**(3): 2105−2114.
[69] Hattori, K., Okuno, A., Roederer, I. U. 2023. Finding r-II Sibling Stars in the Milky Way with the greedy optimistic clustering algorithm. *The Astrophysical Journal*, **946**(1): 48.
[70] Sawada, N., Uemura, M., Fujishiro, I. 2022. Multi-dimensional time-series subsequence clustering for visual feature analysis of blazar observation datasets. *Astronomy and Computing*, **41**: 100663.
[71] Iwasaki, H., Ichinohe, Y., Uchiyama, Y. 2019. X-ray study of spatial structures in Tycho's supernova remnant using unsupervised deep learning. *Monthly Notices of the Royal Astronomical Society*, **488**(3): 4106−4116.
[72] Tóth, B., Rácz, I., Horváth, I. 2019. Gaussian-mixture-model-based cluster analysis of gamma-ray bursts in the BATSE catalog. *Monthly Notices of the Royal Astronomical Society*, **486**(4): 4823−4828.
[73] Acuner, Z., Ryde, F. 2018. Clustering of gamma-ray burst types in the Fermi GBM catalogue: indications of photosphere and synchrotron emissions during the prompt phase. *Monthly Notices of the Royal Astronomical Society*, **475**(2): 1708−1724.
[74] Yu, H., Hou, X. L. 2022. Hierarchical clustering in astronomy. *Astronomy and Computing*, **41**: 100662.
[75] Yan, Q. Z., Yang, J., Su, Y., et al. 2020. Distances and statistics of local molecular clouds in the first Galactic quadrant. *The Astrophysical Journal*, **898**(1): 80.
[76] Lesci, G., Nanni, L., Marulli, F., et al. 2022. AMICO galaxy clusters in KiDS-DR3: Constraints on cosmological para-

meters and on the normalisation of the mass-richness relation from clustering. *Astronomy & Astrophysics*, **665**: A100.
[77] Hayes, J. J., Kerins, E., Awiphan, S., et al. 2020. Optimizing exoplanet atmosphere retrieval using unsupervised machine-learning classification. *Monthly Notices of the Royal Astronomical Society*, **494**(3): 4492−4508.
[78] Rubin, A., Gal-Yam, A. 2016. Unsupervised clustering of type II supernova light curves. *The Astrophysical Journal*, **828**(2): 111.
[79] Guez, I. A., Claire, M. 2024. Reading between the rainbows: Comparative exoplanet characterisation through molecule agnostic spectral clustering. *arXiv:2410.16986*.
[80] Mehta, N., Iyyani, S. 2024. Exploring gamma-ray burst diversity: Clustering analysis of the emission characteristics of Fermi- and BATSE-detected Gamma-Ray Bursts. *The Astrophysical Journal*, **969**(2): 88.
[81] Chattopadhyay, S., Maitra, R. 2017. Gaussian-mixture-model-based cluster analysis finds five kinds of gamma-ray bursts in the BATSE catalogue. *Monthly Notices of the Royal Astronomical Society*, **469**(3): 3374−3389.
[82] Pantoja, R., Catelan, M., Pichara, K., et al. 2022. Semi-supervised classification and clustering analysis for variable stars. *Monthly Notices of the Royal Astronomical Society*, **517**(3): 3660−3681.
[83] Jin, Y. C., Yang, L. Y., Chiang, C. E. 2022. Identifying exoplanets with machine learning methods: A preliminary study. *arXiv: 2204.00721*.
[84] Shin, M. S., Chang, S. W., Yi, H., et al. 2018. Detecting variability in massive astronomical time-series data. iii. variable candidates in the superwasp dr1 found by multiple clustering algorithms and a consensus clustering method. *The Astronomical Journal*, **156**(5): 201.
[85] Seo, Y. M., Goldsmith, P. F., Tolls, V., et al. 2020. Applications of machine learning algorithms in processing Terahertz Spectroscopic data. *Journal of Astronomical Instrumentation*, **9**(3): 2050011.
[86] Tiwari, A., Vivek, M. 2024. Spectroscopic quasar anomaly detection (SQuAD) I: Rest-frame UV spectra from SDSS DR16. *arXiv: 2411.16858*.
[87] Fustes, D., Dafonte, C., Arcay, B., et al. 2013. SOM ensemble for unsupervised outlier analysis. Application to outlier identification in the Gaia astronomical survey. *Expert Systems with Applications*, **40**(5): 1530−1541.
[88] Zhou, L. C., Cai, J. H., Yang, H. F., et al. 2025. A study of emission lines in carbon stars. *The Astrophysical Journal*, **981**(2): 151.
[89] Liang, B., Cai, J. H., Yang, H. F. 2022. A new cell group clustering algorithm based on validation & correction mechanism. *Expert Systems with Applications*, **193**: 116410.
[90] Li, Y. Y., Bai, Y., Wang, C. S., et al. 2024. Deep Learning and Methods Based on Large Language Models Applied to Stellar Light Curve Classification. *Intelligent Computing*, **4**: 0110.
[91] Wu, J. F., Hyk, A., McCormick, K., et al. 2024. Designing an evaluation framework for large language models in astronomy research. *arXiv: 2405.20389*.
[92] Fouesneau, M., Momcheva, I. G., Chadayammuri, U., et al. 2024. What is the role of large language models in the evolution of astronomy research? *arXiv:2409.20252*.
[93] Pan, R., Nguyen, T. D., Arora, H., et al. 2024. AstroMLab 2: AstroLLaMA-2-70B model and benchmarking specialised LLMs for astronomy. In Proceedings of SC24-W: Workshops of the International Conference for High Performance Computing, Networking, Storage and Analysis.
[94] Smirnov, E. 2024. Effortless and accurate time series analysis in astronomy using Large Language Models. In Proceedings of Europlanet Science Congress.