Improving Individual Fairness in Gaussian Mixture Clustering

Chuan Qian[1], Zhijing Yang[2], Hui Zhang[1*]
[1]Southwest University of Science and Technology, Mianyang, China
[2]Southeast University, Nanjing, China
qianchuan@mails.swust.edu.cn, yangzj@seu.edu.cn, zhanghui@swust.edu.cn

***Abstract*—Clustering is an unsupervised machine learning algorithm. Among the many clustering algorithms, the Gaussian Mixture Clustering (GMC) is widely used for its effectiveness and flexibility in modeling complex data distributions. Fair clustering has received a lot of attention in recent years. Fair clustering ensures fairness between data points or clusters and is essential for reducing bias due to inherent distributional differences. Fairness in clustering is usually categorized into individual level and group level. Individual fairness implies that similar individuals should be treated similarly, but traditional GMC usually ignores this principle, which makes the clustering results inevitably unfair. This paper presents an innovative fair version of GMC, Individual Fair Gaussian Mixture Clustering (IFGMC). By accounting for individual fairness during each iteration, IFGMC ensures the final clustering results consistently respect the principle of equity. Experimental results demonstrate that IFGMC achieves over 30% higher fairness performance than GMC, significantly enhancing individual fairness. This makes it a highly promising solution for achieving fair and reliable clustering.**

***Index Terms*—clustering, fair clustering, Gaussian Mixture Clustering, Individual Fair Gaussian Mixture Clustering.**

I. Introduction

Clustering is a key area in machine learning for identifying hidden patterns in data without the need for labeled samples during training [1]. The goal of clustering is to divide the data into clusters where members of the same cluster are compact and the different clusters are distinct. It has been widely used in many real-world applications, such as multi-view learning [2], social network analysis [3], bioinformatics [4], and image segmentation [5], [6].

Due to the wide applicability and effectiveness of clustering, various clustering techniques have been developed over time, including center-based methods such as K-means [7], density-based methods such as DBSCAN [8], hierarchical methods [9], spectral clustering [10] and so on. Unlike the techniques mentioned above, GMC is a clustering method based on a probabilistic model approach. Specifically, it assumes that the data is a mixture of several Gaussian components. Each Gaussian component has a different mean, variance, and weight. In practical application, GMC is able to capture underlying patterns and perform clustering analysis.

In recent years, issues of fairness, such as discrimination or unequal treatment in the process of machine learning algorithms, have received increasing attention. Researchers have proposed a number of methods and theoretical frameworks to improve fairness in machine learning [11]–[15]. Fairness in machine learning is typically framed within two key concepts. One is group fairness [16], which emphasizes demographic parity and the protection of sensitive groups. In certain clustering tasks, the ultimate goal is often to ensure that the proportions of certain subgroups defined by certain attributes (e.g., males and females for gender) are as similar as possible across all clusters. The other concept is individual fairness [17], whose core idea is that similar individuals should be treated similarly. This means that during the clustering process, if two data points are very similar in the feature space, they should be assigned to the same cluster or have a similar membership affiliation [18].

Numerous studies have focused on clustering analysis or algorithmic improvements with the goal of achieving individual fairness. One prominent approach involves adding distance-based constraints during the iterative clustering process, thereby improving individual fairness without significantly compromising clustering quality [17]. Another notable method uses fairness-aware linkage criteria to construct dendrograms that satisfy individual fairness requirements [19]. However, these approaches often appear rigid due to the lack of a solid theoretical foundation for incorporating individual fairness adjustments into the clustering process. In addition, some methods tend to oversimplify the tradeoffs between fairness and clustering performance, leading to suboptimal results in real-world applications.

GMC not only flexibly adapts to data clusters of various shapes through the covariance matrix, but also employs probabilistic distributions to achieve probabilistic clustering, enabling it to handle sample overlap and membership uncertainty. Although GMC demonstrates exceptional performance in clustering, it does not account for individual fairness. GMC heavily relies on probabilistic membership to assign data points, which may result in highly similar data points with only slight probability differences being assigned to different clusters.

This paper proposes a novel fair Gaussian mixture clustering framework. By introducing a similarity-based fairness penalty, IFGMC ensures that the clustering process prioritizes both accuracy and individual fairness.

* Corresponding author

The main contributions of this paper are as follows:

- This paper proposes a novel fair Gaussian mixture clustering framework.
- This paper innovatively proposes a similarity-based fairness penalty to optimize individual fairness.
- Experiments using real-world datasets validated the effectiveness of IFGMC, demonstrating over a 30% improvement in individual fairness performance compared to GMC.

II. Related Work

A. Gaussian Mixture Clustering

GMC [20] is a weighted combination of multiple single Gaussian components, where each single Gaussian component represents a cluster in the data and has its own mean, variance, and weight, which represents the contribution of that single Gaussian component to the overall data. In theory, the GMC can be fitted to any type of distribution. The probability distribution of the GMC is denoted as:

$$P(x_i;\theta) = \sum_{k=1}^{K} \pi_k N(x_i|\mu_k, \Sigma_k) \tag{1}$$

where $N(\cdot)$ is the probability density function of the kth Gaussian component, π_k is the weight of each Gaussian component, μ_k and Σ_k are the mean vector and covariance matrix of the kth Gaussian component. Equation (1) can be used to derive the log-likelihood function of the GMC:

$$L(\theta) = \sum_{i=1}^{N} \log P(x_i;\theta) \tag{2}$$

The simple approach to solving for the parameters is large likelihood estimation.

$$\theta = arg\max_{\theta} L(\theta) = arg\max_{\theta} \sum_{i=1}^{N} \log P(x_i;\theta) \tag{3}$$

However, there are hidden variables in GMC. For each data point, it is not known in advance which Gaussian component it belongs to, which makes it difficult to maximize the likelihood function directly to find the parameters. The Expectation Maximization (EM) algorithm [21] is usually used to solve for the parameters. The EM algorithm derives the lower bound of Equation (2) through Jensen inequality [22], and then obtains the parameters of each Gaussian component by maximizing this lower bound containing the hidden variables. In GMC, this heuristic iterative process used by the EM algorithm to optimize the parameters is divided into two steps. In the E-step, based on the current estimates of the model parameters, the probability distribution of the hidden variables is computed using Bayes rule [23], i.e., the affiliation of each data point in each Gaussian component is computed:

$$\gamma_{ik} = \frac{\pi_k N(x_i;\mu_k,\Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x_i;\mu_j,\Sigma_j)} \tag{4}$$

The expected log-likelihood function with respect to the hidden variables can now be written as:

$$Q(\theta) = \sum_{i=1}^{N}\sum_{k=1}^{K} \gamma_{ik} \log P(x_i;\theta) \tag{5}$$

The function $Q(\theta)$ is a weighted sum of the log-likelihoods of all data points at each Gaussian component, representing the expected log-likelihood of the observed data and the estimated latent variable distribution. In step M, the parameters θ are updated by maximizing the expected likelihood $Q(\theta)$:

$$\mu_k = \frac{\sum_{i=1}^{N} \gamma_{ik} x_i}{\sum_{i=1}^{N} \gamma_{ik}} \tag{6}$$

$$\Sigma_k = \frac{\sum_{i=1}^{N} \gamma_{ik}(x_i-\mu_k)^T(x_i-\mu_k)}{\sum_{i=1}^{N} \gamma_{ik}} \tag{7}$$

$$\pi_k = \frac{1}{N}\sum_{i=1}^{N} \gamma_{ik} \tag{8}$$

B. Individual Fairness

Individual fairness focuses on the fair treatment of similarities between individuals and is concerned with similar decisions between homogeneous individuals. The core idea of individual fairness is that samples that are similar in the feature space should receive similar prediction results or labels. Kar et al. [24] introduced a new concept of individual fairness, which is based on features, under which a randomized k-means algorithm was designed to ensure both the minimization of the clustering distance objective and the approximation of individual fairness, subject to the natural constraints of the distance metric and the fairness constraints. Dwork et al. [25] used the Lipschitz condition as a theoretical basis and applied it to the optimization of algorithmic fairness and the definition of fairness metrics, which ensures that similar data points are treated similarly during the clustering process and improves the fairness and reliability of the algorithm. In k-clustering, Mahabadi et al. [18] design a bicriteria approximation algorithm based on local search to improve individual fairness. They define the fair radius of each point and then ensure that each point has a clustering center within a constant multiple of its fair radius by constructing a keyball and a local search strategy. Kar et al. [24] use linear programming (LP) to relax the individual fair clustering problem. Their proposed LP-FAIR algorithm assigns the point v to the clustering center f_k with probability $x*_{v,fk}$ based on the solution of the LP. This step introduces randomness to ensure that the individual fairness constraint is satisfied with high probability.

III. Methodology

A. Individual Fairness in IFGMC

To achieve individual fairness in probabilistic clustering, this paper innovatively defines the following concepts.

Definition 1 (Similar Individuals)**.** *Given a multidimensional dataset $X = \{x_1, x_2, ..., x_n\}$, if x_j is one of x_i's c nearest*

neighbors in the feature space [24], then x_j is a similar individual to x_i, denoted $x_j \in rd(x_i)$.

Similar individuals discussed in Definition 1 are individuals with similar features, in order to quantify the similarity between individuals, a metric is designed in Definition 2.

Definition 2 (Individual Similarity)**.** *For the individual similarity between data points x_i and x_j, the formula is given below:*

$$s(x_i, x_j) = \left(1 - \frac{d(x_i, x_j)}{m}\right)^2 \tag{9}$$

where $d(x_i, x_j)$ is the distance between x_i and x_j based on the feature space, and m is the number of features in the data points. The $s(\cdot)$ is limited to 0 to 1. The closer $s(\cdot)$ is to 1 means that the two points x_i and x_j are more similar, and the closer it is to 0 means that the two points are less similar.

Individual similarity is only quantitative for similar individuals and does not take into account irrelevant data points. With a valid quantification of fairness, this paper designs a fairness penalty as a mechanism to ensure individual fairness. For x_i and its similar individual x_j, if the individual similarity between them is closer to 1, it means they are more similar, and the closer to 0, the less similar they are. The more similar x_i and x_j become, the smaller their difference in feature space.

B. Fairness Penalty

In order to solve the possible fairness problem of traditional GMC, this paper introduces a fairness penalty, which aims to reduce the difference in affiliation of similar individuals. The fairness penalty is designed based on the strict version of the similarity matrix S in Definition 3 and affiliation differences. The penalty takes into account the difference in affiliation between similar individuals, for each pair of similar individuals, if the difference in their affiliation to each cluster is greater, then the value of FP will be greater. The mathematical expression is:

$$FP_i = \sum_{j=1}^{N} \sum_{k=1}^{K} S_{ij} ||\gamma_{ik}^{(old)} - \gamma_{jk}^{(old)}||^2 \tag{10}$$

The purpose of the fairness penalty is to minimise differences in affiliation between similar individuals within the clustering results. Fairness penalty is calculated by determining the difference in affiliation of each pair of similar individuals in each clustering, and then weighting and summing these differences. This encourages similar individuals to show consistency in the clustering results.

Definition 3 (Strict Version of the Similarity Matrix)**.** *A similarity matrix can be constructed using similar individuals and individual similarity:*

$$S_{ij} = s(x_i, x_j) \tag{11}$$

This constructed similarity matrix may be asymmetric. Suppose that x_j belongs to one of the c nearest neighbors of x_i in the feature space, and that x_i may not belong to one of the c nearest neighbors of x_j in the feature space due to a general difference in the densities of the two sample points. We define the strict version of the similarity matrix:

$$S_{ij} = \begin{cases} 0, & x_i \notin rd(x_j) \ \text{ or } \ x_j \notin rd(x_i) \\ s(x_i, x_j), & x_i \in rd(x_j) \ \text{ and } \ x_j \in rd(x_i) \end{cases} \tag{12}$$

where $rd(\cdot)$ is explained in Definition 1, $S_{ij} = 0$ if and only if x_i is not a similar individual of x_j, or x_j is not a similar individual of x_i.

C. Individual Fair Gaussian Mixture Clustering

GMC does not consider fairness when updating affiliation using Equation (4). IFGMC adopts Equation (13) as a key step in affiliation updates. By constructing a penalty based on the similarity matrix and affiliation differences, this penalty is incorporated when updating each data point's affiliation. This minimizes assignment discrepancies among similar individuals during clustering, thereby yielding fairer clustering results. Additionally, the fairness penalty employs a weighting mechanism that prevents over-penalization by constraining the FP_i value within the range of 0 to 1. The updated formula is as follows:

$$\gamma_{ik}^{(new)} = \gamma_{ik}^{(old)} \cdot \left(1 - \frac{FP_i}{FP_i + \gamma_{ik}^{(old)}}\right) \tag{13}$$

In the traditional EM algorithm, the goal of the M-step is to update the parameters of the GMC based on the affiliation in the E-step. The modified M-step formulation remains unchanged because the fairness penalty mainly affects the updating of affiliation and does not directly affect the estimation of the parameters. S is a similarity matrix based on similar individuals, ensuring that irrelevant data points are not affected when adjusting the affiliation assignment. For each pair of similar individuals (x_i, x_j), S_{ij} tends to 1 meaning that the similarity is high, at which point the value of the fairness penalty is higher if the difference in their affiliation to each cluster is large, i.e., the higher the value of $||\gamma_{ik} - \gamma_{jk}||$. During the IFGMC iteration, such pairs of points should not be assigned to different clusters, and a larger penalty is imposed on such unreasonable pairs of points to ensure that the difference in affiliation between them becomes smaller and smaller, and after many iterations, they are assigned to the same cluster so that similar individuals are treated similarly. Since the fairness penalty is unique to each pair of similar individuals, each iteration is globally updated and no local optimization takes place.

D. Summary

IFGMC achieves fairer clustering results by introducing a fairness penalty based on feature similarity and adjusting the affiliation updating method to reduce the differences in affiliation of similar individuals. The core steps of IFGMC are shown in Algorithm 1.

Algorithm 1 IFGMC

Input: Dataset X, number of clusters K.
Initialize the mean μ, covariance matrix Σ, and weight π with random values for all Gaussian components.
repeat
Calculate the probability density function and update γ according to Equation (4).
Use Equation (13) to correct for unfair data point allocation.
Normalize the γ matrix.
Update μ, Σ, π by $\gamma^{(new)}$.
until The likelihood function changes below the threshold or reaches the maximum number of iterations.
Output: Labels of data points.

TABLE I
DETAILS OF ALL DATASETS USED IN THIS PAPER.

Dataset	Size	Features	Description
Wine [26]	4898	4	Contains data about white wines from Portugal.
Rice [27]	3810	7	Images of two types of rice taken by Turkish scientists, processed and feature inferred to obtain morphological characteristics of rice.
Bike [28]	8760	5	A dataset that predicts the number of rental bikes needed per hour to ensure a steady supply of rental bikes.
Bean [29]	13611	4	Images of seven different dry beans were captured using a high-resolution camera, and these images were segmented and feature extracted.
Churn [30]	3150	4	Randomly collected data from the database of an Iranian telecommunications company.
Abalone [31]	4177	7	Abalone species classification based on physical measurements.
Card [32]	30000	4	Customers and their credit card transaction information with a bank in Taiwan.

IV. EXPERIMENT

A. Dataset

This paper uses seven real-world datasets that are commonly used to evaluate clustering tasks and fairness improvements. In them, the number of feature selections is appropriately reduced considering the impact on performance. Table I shows the details of these datasets, including name, provenance, number of rows and number of features.

B. Evaluation Metrics

In order to evaluate the performance of the IFGMC algorithm and the GMC algorithm in terms of fairness and clustering quality, this paper chooses one metric for each aspect.

1) Silhouette Coefficient (SC): The SC [33] is a metric for evaluating the effectiveness of a clustering algorithm that takes into account both the tightness of the clustering results and the separation between the clusters. If a data point has a small distance from other data points in the cluster to which it belongs, but a large distance from data points in other clusters, it means that the tightness within the cluster to which this data point belongs is high, and the separation between the clusters is large, then the profile coefficient of this data point will be larger. Its value range is [-1,1], the closer it is to 1, the better the clustering results are, and the closer it is to -1, the worse the classification results are. The formula is as follows:

$$SC_i = \frac{b_i - a_i}{max(a_i, b_i)} \tag{14}$$

where a_i is the average distance between x_i and other data points in its cluster, and b_i is the average distance between x_i and data points in other clusters. The total SC is then given for the entire dataset:

$$SC = \frac{\sum_{i=1}^{N} SC_i}{N} \tag{15}$$

TABLE II
FAIRNESS PERFORMANCE COMPARISON BETWEEN GMC AND IFGMC ACROSS DIFFERENT DATASETS. HIGHER NDP VALUES INDICATE SUPERIOR FAIRNESS PERFORMANCE.

Dataset	K=4		K=6		K=8	
	GMC	IFGMC	GMC	IFGMC	GMC	IFGMC
Wine	0.330	0.578	0.428	0.594	0.359	0.508
Rice	0.445	0.673	0.237	0.571	0.218	0.592
Bike	0.427	0.677	0.471	0.592	0.341	0.519
Bean	0.535	0.785	0.505	0.760	0.510	0.782
Churn	0.465	0.767	0.511	0.829	0.614	0.825
Abalone	0.372	0.553	0.230	0.498	0.347	0.570
Card	0.514	0.776	0.390	0.661	0.368	0.642

TABLE III
EXPERIMENTAL COMPARISON OF CLUSTERING QUALITY BETWEEN GMC AND IFGMC ACROSS DIFFERENT DATASETS. HIGHER SC VALUES INDICATE SUPERIOR CLUSTERING QUALITY.

Dataset	K=4		K=6		K=8	
	GMC	IFGMC	GMC	IFGMC	GMC	IFGMC
Wine	0.236	0.231	0.152	0.235	0.208	0.194
Rice	0.370	0.352	0.238	0.230	0.355	0.341
Bike	0.256	0.244	0.266	0.252	0.213	0.225
Bean	0.388	0.380	0.287	0.279	0.265	0.233
Churn	0.376	0.362	0.387	0.363	0.445	0.418
Abalone	0.211	0.248	0.280	0.257	0.229	0.214
Card	0.437	0.426	0.315	0.310	0.423	0.402

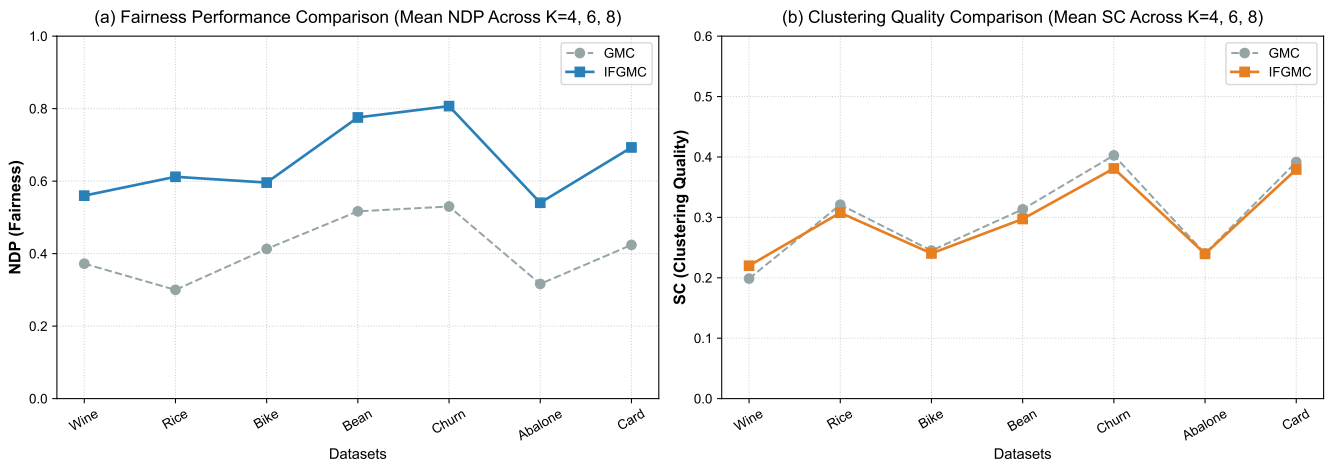

Fig. 1. NDP and SC comparison between IFGMC and GMC.

2) Neighbor Discrepancy Penalty (NDP): The fairness metrics proposed in previous studies are not applicable to probabilistic clustering. In order to visually compare the differences in individual fairness performance between IFGMC and GMC, this paper devises a novel metric.

Definition 4 (NDP)**.** *NDP calculates penalties based on the similarity of the data points and their neighbors, as well as label differences, and is used to measure the fairness of the clustering results, especially in cases where individuals are more similar and it is expected that their clustering labels should be more consistent. The value of NDP is in the range of $[0, 1]$, and the closer it is to 1, the fairer the clustering results are at the individual level.*

The formula is as follows:

$$NDP = 1 - \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} S_{ij} \cdot 1_{\{labels[i] \neq labels[j]\}} \tag{16}$$

where S is the similarity matrix, which can be used here as a penalty term to penalize those similar points that are assigned to different clusters. $1_{\{labels[i] \neq labels[j]\}}$ is an indicator function that indicates whether x_i and x_j belong to different clusters or not (1 if they are different clusters, 0 otherwise). A larger value of NDP indicates a fairer clustering result at the individual level. The similarity matrix only takes into account the similarity of similar individuals, and dissimilar individuals are not included in the calculation, so the only thing that affects the final result is the similar individuals that are assigned to different clusters, and the higher their similarity is, the higher the penalty for NDP. If there are ultimately a large number of similar individuals assigned to different clusters, and the similarity between them is high, then the value of NDP will tend to be zero.

C. Baseline

As the sole probabilistic clustering algorithm, current research on GMC's fairness is scarce within the field. There exists no comparable fair GMC algorithm for evaluation. Furthermore, GMC's implementation differs significantly from other methods (such as spectral clustering and density-based clustering), making it difficult to compare IFGMC with other types of fair clustering algorithms. Therefore, this paper only conducts comparative experiments between IFGMC and traditional GMC.

D. Parameter Settings

This paper adopts the same range of values as other studies on fairness clustering. The number of clusters, K, is set to 4, 6 and 8, and random initialisation is used for the parameter initialisation method.

E. Discussion

Table II and Table III respectively illustrate the differences in fairness and clustering performance between IFGMC and GMC under various K values. Fig. 1 presents the table data as a line chart for intuitive comparison of differences.

1) Fairness Performance: On the NDP metric, IFGMC achieves over 30% improvement compared to traditional GMC. For instance, in the Churn and Bean datasets, when K=6, the NDP values reached 0.829 and 0.760 respectively higher than the traditional GMC values of 0.511 and 0.505. This consistent performance improvement across different datasets demonstrates the robustness of IFGMC in handling various complex data distributions. As the K value increases, the fairness advantage of IFGMC remains robust. This shows that the fairness enhancement mechanism of IFGMC is not dependent on a specific clustering number and is therefore highly universal.

2) Clustering Performance: When introducing fairness constraints, a slight decline in clustering quality typically accompanies the process. However, in this study, this trade-off is kept to an extremely minimal extent. As can be seen in Table III and Fig. 1(b), the SC value of IFGMC is very similar to that of GMC. In certain cases, such as with the Abalone dataset (K = 4), the SC value of IFGMC (0.248) is slightly

higher than that of GMC (0.211). This slight variation suggests that introducing the similarity fairness penalty does not alter the original manifold structure of the data. Instead, it fine-tunes the membership scores while maintaining the clusters' compactness and separability.

V. Conclusion

The traditional GMC lacks attention to individual fairness, to address this problem, this paper proposes IFGMC, which introduces individual fairness into probabilistic models for the first time, and designs a fairness penalty based on the concept of individual fairness, to modify the irrational allocation strategy that may exist in the traditional GMC, so that the algorithm assigns similar data points to the same clusters as much as possible, and this provides an extensible framework for future research on fair clustering. In order to verify the excellent fairness of IFGMC compared to GMC, this paper proposes a new evaluation metric and uses several real datasets to compare the performance difference between IFGMC and GMC in terms of clustering performance and fairness. The experimental results show that in the vast majority of cases, IFGMC demonstrates significantly superior fairness performance compared to GMC, with the loss in clustering performance remaining within an acceptable range.

Acknowledgment

This paper is supported by the Major Project of Sichuan Provincial Natural Science Foundation (2025ZNSFSC0005), the Sichuan Science and Technology Program (2024YFFK0120), the Central Government Guides Local Projects of China (2025ZYDF105) and the National Science Foundation of China (62072419).

References

[1] Ling Ding, Chao Li, Di Jin, and Shifei Ding, "Survey of spectral clustering based on graph theory," *Pattern Recognition*, vol. 151, pp. 110366, 2024.

[2] Zhiwen Yu, Ziyang Dong, Chenchen Yu, Kaixiang Yang, Ziwei Fan, and CL Philip Chen, "A review on multi-view learning," *Frontiers of Computer Science*, vol. 19, no. 7, pp. 197334, 2025.

[3] Ash Mohammad Abbas, "Social network analysis using deep learning: applications and schemes," *Social Network Analysis and Mining*, vol. 11, no. 1, pp. 106, 2021.

[4] Aasim Ayaz Wani, "Comprehensive analysis of clustering algorithms: exploring limitations and innovative solutions," *PeerJ Computer Science*, vol. 10, pp. e2286, 2024.

[5] Yuhang Ding, Liulei Li, Wenguan Wang, and Yi Yang, "Clustering propagation for universal medical image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 3357–3369.

[6] Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 7, pp. 3523–3542, 2021.

[7] Juntao Wang and Xiaolong Su, "An improved k-means clustering algorithm," in *2011 IEEE 3rd international conference on communication software and networks*. IEEE, 2011, pp. 44–46.

[8] Dingsheng Deng, "Dbscan clustering algorithm based on density," in *2020 7th international forum on electrical engineering and automation (IFEEA)*. IEEE, 2020, pp. 949–953.

[9] Qinghua Gu, Yiwen Niu, Zegang Hui, Qian Wang, and Naixue Xiong, "A hierarchical clustering algorithm for addressing multi-modal multi-objective optimization problems," *Expert Systems with Applications*, vol. 264, pp. 125710, 2025.

[10] Jialu Liu and Jiawei Han, "Spectral clustering," in *Data clustering*, pp. 177–200. Chapman and Hall/CRC, 2018.

[11] Dana Pessach and Erez Shmueli, "A review on fairness in machine learning," *ACM Computing Surveys (CSUR)*, vol. 55, no. 3, pp. 1–44, 2022.

[12] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan, "A survey on bias and fairness in machine learning," *ACM computing surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

[13] Zhijing Yang, Hui Zhang, Chunming Yang, Bo Li, Xujian Zhao, and Yin Long, "Fair laplace: A unified framework for fair spectral clustering," *Information Processing & Management*, vol. 62, no. 4, pp. 104124, 2025.

[14] Boyang Yan, Zhijing Yang, Junjie Zheng, Yiding Tang, and Hui Zhang, "Individual fair fuzzy c-means clustering via density-adaptive spectral regularization," *Neurocomputing*, vol. 651, pp. 130794, 2025.

[15] Zhijing Yang, Hui Zhang, Chunming Yang, Bo Li, Xujian Zhao, and Yin Long, "Spectral clustering with scale fairness constraints," *Knowledge and Information Systems*, vol. 67, no. 1, pp. 273–300, 2025.

[16] Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, and Sergei Vassilvitskii, "Fair clustering through fairlets," *Advances in neural information processing systems*, vol. 30, 2017.

[17] Matthäus Kleindessner, Samira Samadi, Pranjal Awasthi, and Jamie Morgenstern, "Guarantees for spectral clustering with fairness constraints," in *International conference on machine learning*. PMLR, 2019, pp. 3458–3467.

[18] Sepideh Mahabadi and Ali Vakilian, "Individual fairness for k-clustering," in *International conference on machine learning*. PMLR, 2020, pp. 6586–6596.

[19] Eduardo Fernandes Montesuma, Fred Maurice Ngole Mboula, and Antoine Souloumiac, "Recent advances in optimal transport for machine learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[20] Joachim Giesen, Philipp Lucas, Linda Pfeiffer, Laines Schmalwasser, and Kai Lawonn, "The whole and its parts: Visualizing gaussian mixture models," *Visual Informatics*, vol. 8, no. 2, pp. 67–79, 2024.

[21] Shu Kay Ng, Thriyambakam Krishnan, and Geoffrey J McLachlan, "The em algorithm," *Handbook of computational statistics: concepts and methods*, pp. 139–172, 2012.

[22] Corentin Briat, "Convergence and equivalence results for the jensen's inequality—application to time-delay and sampled-data systems," *IEEE Transactions on Automatic Control*, vol. 56, no. 7, pp. 1660–1665, 2011.

[23] Alicia A Johnson, Miles Q Ott, and Mine Dogucu, *Bayes rules!: An introduction to applied Bayesian modeling*, Chapman and Hall/CRC, 2022.

[24] Debajyoti Kar, Mert Kosan, Debmalya Mandal, Sourav Medya, Arlei Silva, Palash Dey, and Swagato Sanyal, "Feature-based individual fairness in k-clustering," *arXiv preprint arXiv:2109.04554*, 2021.

[25] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel, "Fairness through awareness," in *Proceedings of the 3rd innovations in theoretical computer science conference*, 2012, pp. 214–226.

[26] P. Cortez, Antonio Luíz Cerdeira, Fernando Almeida, Telmo Matos, and José Reis, "Modeling wine preferences by data mining from physicochemical properties," *Decis. Support Syst.*, vol. 47, pp. 547–553, 2009.

[27] Ilkay Cinar and Murat Koklu, "Identification of rice varieties using machine learning algorithms," *Journal of Agricultural Sciences*, pp. 9–9, 2022.

[28] "Seoul Bike Sharing Demand," UCI Machine Learning Repository, 2020, DOI: https://doi.org/10.24432/C5F62R.

[29] Murat Koklu and Ilker Ali Özkan, "Multiclass classification of dry beans using computer vision and machine learning techniques," *Comput. Electron. Agric.*, vol. 174, pp. 105507, 2020.

[30] "Iranian Churn," UCI Machine Learning Repository, 2020, DOI: https://doi.org/10.24432/C5JW3Z.

[31] Nash, Warwick, and et al., "Abalone," UCI Machine Learning Repository, 1994, DOI: https://doi.org/10.24432/C55C7W.

[32] I-Cheng Yeh, "Default of Credit Card Clients," UCI Machine Learning Repository, 2009, DOI: https://doi.org/10.24432/C55S3H.

[33] Ketan Rajshekhar Shahapure and Charles Nicholas, "Cluster quality analysis using silhouette score," in *2020 IEEE 7th international conference on data science and advanced analytics (DSAA)*. IEEE, 2020, pp. 747–748.