

Article

QGKM: A Quantum Fidelity-Based Graph Clustering Framework for Robust Data Pattern Recognition in Educational Social Networks

Neal N. Xiong ^{1,*}, Weiqing Long ¹, Dacheng He ¹, Xiangwei Meng ¹ , Zulong Diao ¹, Sergey M. Avdoshin ²  and Yevgeni Koucheryavy ²

¹ Department of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan 411201, China; weiqing.long@mail.hnust.edu.cn (W.L.); dachenghe@hnust.edu.cn (D.H.); xiangwei_meng@hnust.edu.cn (X.M.); diaozulong@hnust.edu.cn (Z.D.)

² School of Computer Engineering, National Research University Higher School of Economics, Moscow 123458, Russia; savdoshin@hse.ru (S.M.A.); ykoucheryavy@hse.ru (Y.K.)

* Correspondence: n.xiong@snhu.edu or xionгнаixue@gmail.com

Abstract

In the era of data-driven education, educational social networks generate large volumes of high-dimensional and complex-structured data through learner interactions, collaborative activities, and resource-sharing behaviors, posing significant challenges to traditional unsupervised learning methods. Such data often exhibit non-convex distributions, heterogeneity, and noise sensitivity, making conventional clustering approaches insufficient for capturing their intrinsic structural relationships. To address this issue, this paper proposes Quantum Fidelity-Based Graph K-Means (QGKM), a clustering framework for robust pattern recognition in educational social networks. Specifically, QGKM employs quantum state encoding to map complex educational data into a quantum state space and utilizes quantum fidelity as a similarity metric to uncover latent correlations that Euclidean distance cannot effectively capture. In addition, the incorporation of k-nearest neighbor graphs preserves the local geometric structure of learner interaction networks, while a deterministic greedy hierarchical merging strategy eliminates the instability caused by random initialization. Experimental results on seven real-world datasets demonstrate that QGKM consistently outperforms classical K-Means in clustering accuracy. The proposed framework provides an effective solution for learning pattern discovery, learner profiling, and intelligent recommendation in digital education environments.

Keywords: educational social networks; quantum-inspired computing; quantum fidelity; robust clustering; learning pattern analysis; graph structure analysis



Academic Editor: Igor Aizenberg

Received: 26 March 2026

Revised: 25 April 2026

Accepted: 30 April 2026

Published: 13 May 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Clustering analysis, as a fundamental task in unsupervised learning, plays a vital role in fields such as machine learning, pattern recognition, and data mining [1]. In modern educational environments, the scale and complexity of data continue to increase, leading to learner-generated data becoming increasingly high-dimensional, nonlinear, and structurally diverse for clustering algorithms to process. For instance, in tasks such as learner behavior analysis, collaborative learning assessment, and community detection in Educational Social Networks (ESNs) [2,3], data is often embedded in complex low-dimensional manifolds and

exhibits non-convex or multi-level structures. These characteristics impose higher demands on clustering methods [4].

The classic K-Means algorithm is widely adopted due to its simple implementation and high computational efficiency. However, this algorithm relies on Euclidean distance to measure sample similarity, which can weaken distance discrimination capabilities in high-dimensional spaces. It also suffers from limitations such as sensitivity to outliers and the assumption that clusters are spherical, potentially leading to degraded clustering performance [5]. Although improvements like K-Means++ partially mitigate issues such as initialization sensitivity, their core approach remains rooted in Euclidean distance and the iterative center optimization framework. This makes it difficult to fundamentally address performance challenges in complex educational data structures. These limitations have been extensively discussed and explored for potential enhancements in recent research [6].

To address the limitations of Euclidean distance in complex data structures, graph-based clustering methods construct adjacency graphs to capture the local geometric properties of data. This enables them to identify complex non-convex cluster shapes and local connectivity relationships. This approach allows for the identification of non-convex cluster structures, while also enhancing robustness to local connectivity patterns [7]. In ESNs, such graph structures are especially meaningful because learners are naturally connected through communication, collaboration, and shared resources. By focusing on these local structures, graph-based methods are better equipped to handle the intricate nature of educational data, making them more effective in situations where traditional distance measures may fall short. Spectral clustering, as a representative example, performs clustering by applying spectral decomposition to the graph Laplacian, embedding data into a low-dimensional space where clusters are more easily separable, and has demonstrated strong adaptability to nonlinear and non-convex structures [8]. However, such methods typically depend on the construction of similarity matrices and suffer from high computational costs associated with eigenvalue decomposition, while their performance is also sensitive to the choice of similarity measures and graph construction strategies [9]. Similarly, methods such as Graph K-Means, which incorporate graph structures into the K-Means framework, partially balance local structure modeling and computational efficiency, but still rely on restrictive assumptions. These limitations have been thoroughly discussed in surveys of spectral and graph-based clustering methods [10].

In recent years, quantum computing and quantum information theory have provided new perspectives for machine learning [11]. The superposition property and inner-product structure of quantum states make quantum fidelity a similarity measure with favorable mathematical properties [12,13]. Unlike Euclidean distance, quantum fidelity is defined through inner products in a normalized representation space and is closely related in form to cosine similarity and kernel-based methods, while being grounded in quantum information theory with clear physical and geometric interpretations [14]. Previous studies have shown that, for medium-dimensional data with complex correlations or nonlinear structures, inner-product-based similarity measures often exhibit greater stability and discriminative capability than Euclidean distance. These properties make quantum-inspired similarity measures particularly promising for ESNs, where latent associations among learners, resources, and interaction patterns are often difficult to capture using conventional distance metrics.

Under the constraints of current noisy intermediate-scale quantum (NISQ) devices, implementing large-scale clustering algorithms directly on quantum hardware remains challenging. Consequently, simulating key concepts from quantum information theory on classical computers and developing quantum-inspired algorithms has become an important research direction in quantum machine learning [15]. The focus of this work is

not on achieving computational speedups through quantum hardware, but on exploring the potential advantages of quantum state representations and their induced similarity measures in classical clustering tasks [16].

The main contributions in this paper are as follows:

1. **A Quantum-Inspired Similarity Framework for ESN Data:** We propose QGKM (Quantum Fidelity Graph K-Means), a novel clustering framework specifically tailored for the complex data structures inherent in Educational Social Networks. By mapping high-dimensional learner interaction data into a quantum state space, the algorithm utilizes quantum fidelity to capture subtle, non-linear correlations among learners and educational behaviors that traditional Euclidean metrics frequently overlook.
2. **Enhanced Robustness via Deterministic Optimization:** QGKM incorporates a k-nearest neighbor (k-NN) graph structure and a greedy hierarchical merging strategy. This approach effectively eliminates the performance fluctuations caused by random initialization in traditional K-Means, providing deterministic and stable clustering results for educational data analysis.
3. **Adaptability to Non-Convex Educational Patterns:** Through systematic evaluation on diverse datasets, we demonstrate QGKM's superior capability in recognizing complex, non-convex data distributions common in learner interaction networks, collaborative communities, and resource-sharing environments. The framework achieves an average accuracy improvement of 4.5% overall and a significant 7.9% gain on medium-dimensional datasets.
4. **Empirical Validation of Quantum Encoding Strategies:** This paper provides a comparative analysis of different quantum encoding schemes (Amplitude vs. Angle coding) in the context of educational data. Our findings offer practical guidelines for selecting optimal encoding methods based on the dimensionality and structural complexity of ESN data, bridging the gap between quantum information theory and learning analytics applications.

The rest of this paper is organized as follows: Section 2 is about the related work. In Section 3, we propose our QGKM framework. Section 4 explores the performance analysis. Finally, Section 5 presents the conclusion and future work.

2. The Related Work

Clustering is a fundamental task in unsupervised learning, aiming to discover latent group structures from unlabeled data. Among existing clustering methods, K-Means remains one of the most widely used algorithms because of its simplicity, efficiency, and interpretability. Recent studies have systematically reviewed the development of K-Means and its variants, showing that K-Means is still widely adopted in large-scale data analysis but remains sensitive to initialization, distance metrics, predefined cluster numbers, and data distribution assumptions [17,18]. In particular, the standard K-Means objective relies on Euclidean distance and implicitly assumes that clusters are compact and approximately spherical. As a result, it often performs poorly when handling high-dimensional data, nonconvex clusters, imbalanced distributions, or samples distributed on complex manifolds. Recent deep clustering studies further indicate that representation quality and similarity measurement are crucial factors affecting clustering performance, especially for complex and heterogeneous data [19,20]. K-means++ improves the standard K-means algorithm mainly by selecting initial centroids in a more careful and distance-aware manner, which helps reduce sensitivity to random initialization. However, it still relies on Euclidean distance during the clustering process and therefore cannot fundamentally overcome the limitation of Euclidean-distance-based similarity measurement. Therefore, improving

distance measurement and incorporating structural information have become important directions for extending classical K-Means.

To overcome the limitations of purely centroid-based partitioning methods, graph-based clustering methods have been widely investigated in recent years. These methods usually construct neighborhood graphs, such as k-nearest-neighbor graphs, to model local relationships among samples. Compared with direct Euclidean partitioning, graph-based approaches can better capture non-convex structures and manifold-like data distributions. FINCH, for example, performs efficient parameter-free clustering using first-neighbor relations and does not require the number of clusters to be specified in advance [21]. Other recent graph-based clustering methods improve robustness and scalability by introducing landmark graphs, sparse graph construction, structure-aware graph learning, or large-scale graph clustering mechanisms [22–25]. These studies demonstrate that graph structures are effective for preserving local data geometry and improving clustering performance. However, the quality of graph-based clustering still depends heavily on the similarity measure used to construct the graph. If the adopted distance metric cannot accurately reflect intrinsic sample relationships, the constructed graph may contain noisy edges or lose important neighborhood connections. This limitation motivates the development of more expressive similarity measures for graph construction.

Quantum machine learning has recently emerged as a promising research direction by combining quantum computing principles with machine learning models. Existing studies have shown that quantum states, quantum kernels, and quantum feature maps can provide new perspectives for representing and comparing classical data [26,27]. In particular, quantum feature mapping allows classical samples to be embedded into high-dimensional Hilbert spaces, where the inner product or overlap between quantum states can be used to measure sample similarity [28,29]. This mechanism is conceptually related to kernel methods, but it introduces a quantum-inspired way to define similarity in a potentially richer representation space. Quantum fidelity, which measures the similarity between quantum states, has therefore become a meaningful tool for designing quantum-inspired distance metrics. Recent studies have further explored the potential of quantum-inspired learning and hybrid quantum-classical models in high-dimensional data analysis and clustering-related tasks [30–32]. Nevertheless, most existing quantum or quantum-inspired clustering methods mainly focus on theoretical quantum speedup, hardware-oriented implementation, or isolated similarity computation. They rarely integrate quantum-inspired similarity measures into classical clustering frameworks in a systematic and interpretable manner.

Hierarchical clustering and bottom-up merging strategies provide another important perspective for handling complex cluster structures. Unlike centroid-based methods that directly update cluster centers, agglomerative strategies gradually merge samples or subclusters according to certain similarity or linkage criteria. Recent scalable hierarchical and graph-based clustering methods show that neighborhood-driven merging can reduce dependence on random initialization and better preserve local consistency [33,34]. However, conventional hierarchical clustering methods usually rely on fixed linkage rules and may suffer from high computational cost when a large number of pairwise comparisons are required. Moreover, when the similarity measure is still based on Euclidean distance, hierarchical merging may fail to capture nonlinear and high-dimensional relationships. Therefore, combining graph constraints, robust similarity measurement, and deterministic merging strategies is a promising way to improve clustering stability and interpretability, as recent hierarchical graph-based clustering studies have also demonstrated [35].

In summary, existing studies have made significant progress in K-Means optimization, graph structure modeling, quantum-inspired similarity learning, and hierarchical merging. However, there is still a lack of a unified framework that simultaneously integrates

quantum-inspired similarity measurement, local graph structure, and goal-function-driven clustering optimization. To address this gap, this paper proposes QGKM, a quantum fidelity-based graph K-Means framework. By encoding classical samples into quantum state representations and measuring sample relationships through quantum fidelity distance, QGKM aims to construct a more expressive similarity graph. Furthermore, the proposed bottom-up merging mechanism incorporates local neighborhood information into the clustering process, thereby improving the ability to identify complex, non-convex, and high-dimensional data patterns.

3. QGKM Framework

3.1. Problem Definition and Overall Approach

Given a dataset $X = \{x_1, x_2, \dots, x_n\}$ containing n d -dimensional samples, the clustering task aims to partition it into K mutually exclusive and exhaustive clusters $\{C_1, C_2, \dots, C_K\}$ such that samples within each cluster are as similar as possible, while samples across different clusters are as dissimilar as possible.

The core idea of QGKM is to introduce a similarity measure from quantum information theory—namely, quantum fidelity—into a classical computational framework and to systematically integrate it with a graph-based K-Means clustering paradigm. It should be emphasized that QGKM is not strictly equivalent to minimizing the conventional K-Means objective defined in Euclidean space. Instead, it performs a heuristic approximation of a K-Means-type objective in a similarity space induced by quantum fidelity.

QGKM does not explicitly compute cluster centroids in the classical Euclidean sense. In conventional K-Means, each cluster is represented by the arithmetic mean of its samples, and the algorithm alternates between centroid updating and sample assignment. In contrast, QGKM operates in a quantum fidelity-induced similarity space and performs graph-constrained bottom-up merging. Therefore, no explicit Euclidean centroid or mean quantum state is required.

Moreover, QGKM is not a strict medoid-based method, because it does not select an existing sample as the representative center of each cluster. Instead, a cluster is represented implicitly by its member samples, intra-cluster tightness, and graph-based neighborhood relationships. The term “Graph K-Means” in this paper refers to a K-Means-inspired compactness objective rather than a classical centroid-update procedure.

3.2. Quantum State Encoding and Fidelity Distance

3.2.1. Quantum State Encoding

To map classical data into a quantum state representation space, this work employs quantum state encoding methods, including amplitude encoding and angle encoding [36]. By normalizing the data, classical samples can be represented as pure quantum states, enabling the similarity between samples to be characterized via quantum state inner products. For each data point $x_i \in \mathbb{R}^d$, we encode it as a quantum state $|\psi_i\rangle$.

Amplitude encoding directly maps the components of a data vector to the amplitudes of a quantum state. For a normalized vector $\tilde{x}_i = x_i / \|x_i\|$, the corresponding quantum state is defined as:

$$|\psi_i\rangle = \sum_{j=1}^{d'} \tilde{x}_{i,j} |j\rangle, \quad (1)$$

where $d' = 2^{\lceil \log_2 d \rceil}$ is the dimension padded to the nearest power of two.

To ensure normalization $\langle \psi_i | \psi_i \rangle = 1$, insufficient dimensions are padded with zeros or small values. Specifically, if the original vector dimension $d < d'$, we construct a padded vector:

$$\tilde{x}_i^{\text{pad}} = [\tilde{x}_{i,1}, \tilde{x}_{i,2}, \dots, \tilde{x}_{i,d}, \epsilon, \epsilon, \dots, \epsilon], \quad (2)$$

where $\epsilon = 10^{-10}$ is a small padding value, followed by re-normalization of the entire vector. Amplitude encoding theoretically offers exponential compression of the representation while preserving the inner product relationships between classical vectors. In this work, we focus on the similarity structure induced by this encoding under classical simulation, rather than the storage or computational speedup advantages in quantum computing. In practical implementation, the quantum states and their fidelities are computed explicitly on classical computers.

Angle encoding maps data components to the phase angles of a quantum state. For vector components normalized to the $[0, \pi]$ range and $\theta_{i,j}$, the quantum state is defined as:

$$|\psi_i\rangle = \bigotimes_{j=1}^{n_q} \left(\cos \theta_{i,j} |0\rangle + e^{i\phi_{i,j}} \sin \theta_{i,j} |1\rangle \right), \quad (3)$$

where $n_q = \lceil \log_2 d \rceil$ is the number of qubits, $\theta_{i,j} = \pi \cdot x_{i,j} / \max_j |x_{i,j}|$ is the normalized angle parameter, and $\phi_{i,j}$ is the phase parameter. In this work, we set $\phi_{i,j} = 0$ to simplify to real-valued angle encoding. Angle encoding maps classical data onto the Bloch sphere of each qubit via trigonometric functions, encoding one data dimension per qubit. Angle encoding provides rich phase information and enables nonlinear feature transformation via the combined use of amplitude and phase encoding, showing robustness to noise and perturbations [37].

In the context of educational social networks, amplitude encoding is generally more suitable for relatively dense and continuous feature vectors, such as learner interaction frequency, learning activity duration, resource access counts, and aggregated behavioral statistics. In these cases, most feature dimensions contain meaningful numerical information, and amplitude encoding can preserve the global distribution of normalized feature magnitudes.

Angle encoding may be more suitable for sparse, binary, or nonlinear features, such as whether a learner participates in a specific interaction type, whether a learning resource is accessed, or whether a particular behavioral event occurs. Since angle encoding maps features through trigonometric functions, it introduces a nonlinear transformation before fidelity computation. Therefore, the choice between amplitude encoding and angle encoding should be regarded as data-dependent rather than universal.

It should be noted that the quantum-state-based representation used in this work is still representation-dependent. In particular, the fidelity distance is not intended to be a completely parameterization-free metric. Instead, it is defined after a classical sample has been mapped into a normalized quantum state through a chosen encoding scheme. For amplitude encoding, the fidelity between two pure states is mathematically related to the squared inner product between normalized vectors, and therefore has a close connection to cosine-type similarity. In this sense, the proposed method does not eliminate metric dependence, but changes the geometry in which similarity is evaluated: from the original Euclidean space to a normalized Hilbert-space representation induced by quantum state encoding.

This distinction is important for understanding the role of QGKM. Euclidean distance is sensitive to vector magnitude and often becomes less discriminative in high-dimensional spaces where pairwise distances tend to concentrate. By contrast, fidelity emphasizes the overlap between normalized state vectors and focuses more on directional and relational similarity. When angle encoding is used, the trigonometric mapping further introduces a nonlinear feature transformation before fidelity computation, allowing the similarity measure to capture relationships that are not directly expressed by Euclidean distance in the original feature space. Therefore, the potential advantage of QGKM arises from the combined effect of data normalization, quantum-state encoding, and fidelity-based

similarity measurement, rather than from the claim that quantum fidelity is universally superior to Euclidean distance.

Nevertheless, the effectiveness of fidelity distance depends on the suitability of the selected encoding strategy. If the encoding does not preserve meaningful structural information of the original data, the resulting fidelity distance may not improve clustering performance. This representation dependence is also one reason why this paper compares amplitude encoding and angle encoding experimentally.

3.2.2. Quantum Fidelity Distance

In the quantum space, we use quantum fidelity instead of traditional Euclidean distance to measure sample similarity. For two quantum states $|\psi_i\rangle$ and $|\psi_j\rangle$, the fidelity is defined as:

$$F(|\psi_i\rangle, |\psi_j\rangle) = |\langle\psi_i|\psi_j\rangle|^2, \quad (4)$$

with $F \in [0, 1]$, where $F = 1$ indicates identical states and $F = 0$ indicates orthogonal states. Quantum fidelity is a fundamental concept in quantum information theory for quantifying the similarity between quantum states. Unlike Euclidean distance, fidelity accounts for the superposition and phase information of quantum states, capturing the structure of high-dimensional data more effectively.

To convert fidelity into a distance measure, we define the quantum fidelity distance as:

$$d_Q(|\psi_i\rangle, |\psi_j\rangle) = 1 - F(|\psi_i\rangle, |\psi_j\rangle), \quad (5)$$

A more precise comparison between Euclidean distance and fidelity distance can be made for normalized real-valued vectors. Let x and y be two normalized vectors with $\|x\| = \|y\| = 1$. Their squared Euclidean distance is given by $\|x - y\|^2 = 2 - 2\langle x, y \rangle$. Under amplitude encoding, the fidelity between the corresponding pure states is $F(x, y) = |\langle x, y \rangle|^2$. This shows that fidelity is closely related to normalized inner-product similarity. The key difference is that Euclidean distance measures coordinate displacement, while fidelity measures state overlap after normalization. Therefore, fidelity reduces the direct influence of vector magnitude and emphasizes directional or relational similarity. This can be useful when the clustering structure is mainly reflected by the relative distribution of features rather than their absolute scale.

However, fidelity distance should not be interpreted as a universal solution to high-dimensional distance concentration. In high-dimensional spaces, Euclidean distances may concentrate around their expected value, while inner-product-based similarities may also become less discriminative if the data contain strong noise or lack meaningful structure. Therefore, the advantage of fidelity is data-dependent. It is expected to be more useful when the quantum-state encoding preserves meaningful structural relationships, which is consistent with the stronger performance gains observed on low- and medium-dimensional datasets in our experiments.

Quantum fidelity is one of several possible quantum-state similarity measures. Another widely used measure is trace distance, which quantifies the distinguishability between two quantum states. For two density matrices ρ and σ , the trace distance is defined as $D_{tr}(\rho, \sigma) = 1/2 \|\rho - \sigma\|_1$. While fidelity measures the degree of overlap or closeness between two states, trace distance measures how well two states can be distinguished. These two quantities are closely related through known inequalities in quantum information theory and provide complementary perspectives on quantum-state similarity.

The reason such quantum-inspired measures may help alleviate some limitations of Euclidean distance is that they compare samples after normalization and quantum-state encoding, rather than directly measuring absolute coordinate differences in the original feature

space. Euclidean distance can be strongly affected by scale, noise, and distance concentration in high-dimensional spaces. Fidelity-based similarity, by contrast, focuses on the overlap between normalized states and therefore emphasizes directional, relational, and structural similarity. For pure states, fidelity has the simple form $F(|\psi_i\rangle, |\psi_j\rangle) = |\langle\psi_i|\psi_j\rangle|^2$, which makes it computationally convenient and directly compatible with k-nearest neighbor graph construction.

In this work, fidelity is selected instead of trace distance for three reasons. First, the proposed framework mainly uses pure-state representations generated by amplitude or angle encoding, for which fidelity has a simple closed-form expression based on the inner product. Second, fidelity can be efficiently computed during classical simulation and directly transformed into a distance-like quantity for graph construction. Third, fidelity naturally reflects state overlap, which is consistent with the goal of identifying samples with similar latent patterns in educational social network data. Trace distance remains an important alternative measure, and comparing fidelity distance with trace-distance-based clustering will be considered in future work.

Overall, the quantum fidelity distance should be regarded as a representation-induced similarity measure that provides a geometric perspective distinct from the Euclidean distance, rather than as a universal replacement for it. Its effectiveness depends on whether the chosen quantum encoding can preserve meaningful structural relationships in the data [38].

3.3. Hierarchical Clustering Based on k-Nearest Neighbor Graphs

For the encoded quantum state set $\{|\psi_1\rangle, |\psi_2\rangle, \dots, |\psi_n\rangle\}$, we construct a k-nearest neighbor (k-NN) graph $G = (V, E)$, where the vertex set $V = \{1, 2, \dots, n\}$ corresponds to all samples, and the edge set E contains connections from each sample to its k-nearest neighbors. The k-NN graph is a sparse graph structure that effectively captures the local geometric structure of the data while avoiding the high computational cost of evaluating distances between all pairs of samples.

For each sample i , its k-nearest neighbor set is defined as $\mathcal{N}_k(i) = j_1, j_2, \dots, j_k$ such that $d_Q(|\psi_i\rangle, |\psi_{j_1}\rangle) \leq \dots \leq d_Q(|\psi_i\rangle, |\psi_{j_k}\rangle)$ with the corresponding distance set:

$$\mathcal{D}_k(i) = \{d_Q(|\psi_i\rangle, |\psi_{j_1}\rangle), \dots, d_Q(|\psi_i\rangle, |\psi_{j_k}\rangle)\}, \quad (6)$$

In practice, we first compute the quantum fidelity distances from each sample to all other samples, and then select the k-nearest neighbors with the smallest distances. This step has a time complexity of $O(n^2 d')$, where d' is the dimension of the quantum states. To improve efficiency, approximate nearest neighbor search algorithms, such as Locality-Sensitive Hashing or Hierarchical Navigable Small World, can be employed to reduce the complexity to $O(n \log n)$ [39].

To ensure clustering consistency and graph connectivity, we perform symmetrization on the k-NN graph. If sample i is among the k-nearest neighbors of sample j , but j is not among the k-nearest neighbors of i , the edge $j \rightarrow i$ is added to form a bidirectional connection. The symmetrized neighbor set is then defined as:

$$\mathcal{N}_k^{\text{sym}}(i) = \mathcal{N}_k(i) \cup \{j \mid i \in \mathcal{N}_k(j)\}, \quad (7)$$

This symmetrization ensures the graph is undirected, i.e., if an edge exists from i to j , there is necessarily an edge from j to i . This property is crucial for the subsequent hierarchical clustering process, as the cluster merging operations depend on a symmetric graph structure. After symmetrization, the number of neighbors per sample may exceed k , with the average number of neighbors being approximately $2k$. The symmetrization process

itself has a time complexity of $O(nk)$ and a space complexity of $O(nk)$. The choice of k affects the neighborhood scale of the graph. A small k may lead to graph fragmentation, whereas a large k may introduce noisy or cross-cluster edges. After symmetrization, the effective number of neighbors may exceed k because an edge is retained if either endpoint selects the other as a k -nearest neighbor. Therefore, k should be selected to balance graph connectivity and neighborhood purity. Although QGKM avoids the random initialization problem of conventional K-Means, it should not be regarded as parameter-free. Under a fixed k and fixed preprocessing procedure, QGKM produces deterministic and reproducible results. However, different k values may lead to different neighborhood graphs and consequently different clustering results. This sensitivity reflects the dependence of graph-based clustering methods on neighborhood scale, which is different from the run-to-run instability caused by random centroid initialization.

At the beginning of the algorithm, each sample is initialized as its own individual cluster: $C_i^{(0)} = i, i = 1, 2, \dots, n$ with the initial number of clusters $c^{(0)} = n$. To efficiently perform hierarchical clustering, we introduce the concept of tightness for each cluster, which quantifies the cohesion of samples within the cluster. Initially, the tightness of each cluster is set to: $T_i^{(0)} = 0, i = 1, 2, \dots, n$. Tightness is continuously updated throughout the cluster merging process, reflecting the average distance between samples within a cluster. This dynamic update allows the tightness measure to adapt as clusters evolve, providing a more accurate representation of cluster cohesion during the merging steps.

The algorithm employs a greedy strategy to iteratively merge the most similar pair of clusters. At the t -th iteration, the cluster pair C_i and C_j with the smallest distance is selected for merging. This approach ensures that at each step, the clusters that are most similar to each other are combined, progressively refining the overall clustering structure.

$$(i^*, j^*) = \arg \min_{(i,j) \in \mathcal{E}^{(t)}} d_{ij}^{(t)}, \quad (8)$$

where $\mathcal{E}^{(t)}$ denotes the current set of edges between clusters, and $d_{ij}^{(t)}$ represents the distance between clusters C_i and C_j . Under this greedy merging strategy, at each step, the cluster pair with the minimal distance in the current similarity space is merged. This approach empirically drives the overall clustering structure toward more compact clusters, as each merge reduces the average distance within the resulting clusters. Although this process does not guarantee a globally optimal solution, the deterministic merging order effectively avoids instability caused by random initialization and exhibits good stability and reproducibility in practice.

Unlike traditional K-Means algorithms that require multiple iterative optimizations, the QGKM algorithm obtains k clusters through hierarchical merging with only $n - k$ merge operations. The computation of inter-cluster distances is the core of the hierarchical clustering algorithm. For two clusters C_i and C_j , their distance is defined as:

$$d_{ij} = \frac{1}{|C_i| \cdot |C_j|} \sum_{p \in C_i} \sum_{q \in C_j} d_Q(|\psi_p\rangle, |\psi_q\rangle) - T_i - T_j, \quad (9)$$

where $|C_i|$ and $|C_j|$ denote the number of samples in clusters C_i and C_j , respectively, and T_i and T_j are the tightness of the clusters. The first term calculates the average distance over all sample pairs between the two clusters, while the second term subtracts the tightness of each cluster to obtain the net inter-cluster distance. This definition is formally inspired by the concept of within-cluster variance change in Ward's linkage method, but it is not strictly limited to Euclidean space or squared error forms. Its purpose is to consider both cluster size and intra-cluster cohesion in the quantum fidelity-induced similarity space,

thereby avoiding the chaining effect or unstable merges often encountered in traditional single-linkage or complete-linkage methods.

When clusters C_i and C_j are merged into a new cluster C_{new} , the tightness is updated as:

$$T_{\text{new}} = d_{ij} + T_i + T_j, \quad (10)$$

This update formula ensures the cumulative property of tightness: the tightness of the new cluster equals the merging distance plus the sum of the tightness of the original clusters. Intuitively, tightness reflects the cumulative average distance of all sample pairs within a cluster: a larger tightness indicates more dispersed samples, while a smaller tightness indicates a more compact cluster.

To avoid recomputing distances for all cluster pairs after each merge, an incremental update strategy is employed. After merging C_i and C_j into C_{new} , only the distances between C_{new} and the other clusters need to be updated. For a cluster C_p (where $p \neq i, j$), the distance to C_{new} is computed as follows:

If C_p is a neighbor of both C_i and C_j :

$$d_{\text{new},p} = \frac{(|C_i| + |C_p|)(d_{ip} + T_i + T_p) - T_i|C_i| - T_p|C_p|}{|C_{\text{new}}| + |C_p|} + \frac{(|C_j| + |C_p|)(d_{jp} + T_j + T_p) - T_j|C_j| - T_p|C_p|}{|C_{\text{new}}| + |C_p|} - T_{\text{new}} - T_p, \quad (11)$$

If C_p is a neighbor of C_i only (not C_j):

$$d_{\text{new},p} = \frac{(|C_i| + |C_p|)(d_{ip} + T_i + T_p) - T_i|C_i| - T_p|C_p| + |C_j| \cdot |C_p| \cdot \gamma}{|C_{\text{new}}| + |C_p|} - T_{\text{new}} - T_p, \quad (12)$$

where $\gamma = \max_{i,j} \max_k d_Q(|\psi_i\rangle, |\psi_k\rangle)$ is the maximum distance in the k-NN graph, which is used to estimate the distance between unconnected cluster pairs.

This incremental update strategy significantly reduces the update cost after each merge. Instead of recomputing distances among all remaining cluster pairs, QGKM only updates distances associated with the neighboring clusters of the newly formed cluster. Thus, the number of distance updates is reduced to approximately $O(k)$, and the corresponding heap-based update cost is approximately $O(k \log(nk))$.

3.4. Complete Algorithm Procedure

Integrating the aforementioned components, the complete workflow of the QGKM algorithm is presented in Algorithm 1.

The algorithm first maps classical data into a quantum space via quantum state encoding, and then constructs a k-nearest neighbor graph based on quantum fidelity distances, which is subsequently symmetrized. Each sample is then initialized as an individual cluster, and a min-heap containing all candidate merge pairs is constructed. The core iterative process of QGKM employs a greedy strategy, selecting the pair of clusters with the smallest distance for merging at each iteration, updating the relevant data structures, and finally performing continuous processing on the cluster labels. It is important to clarify that QGKM does not minimize the K-Means objective in classical Euclidean space. Instead, it performs heuristic optimization on a similarity-based objective function weighted by graph structure within a non-Euclidean similarity space defined by quantum fidelity. Its bottom-up merging strategy drives the overall partition toward this optimization goal through locally optimal selections at each step. This approach preserves the intuitive framework of K-Means while enhancing the ability to characterize complex data distributions by integrating graph structures with quantum metrics.

Algorithm 1 QGKM Clustering Algorithm**Input:** Dataset $\mathcal{D} = \{x_1, \dots, x_n\}$, target number of clusters k , number of neighbors k_{nn} **Output:** Cluster labels l , clustering loss $\mathcal{L}$

```

1:  for  $i = 1$  to  $n$  do
2:       $|\psi_i\rangle \leftarrow \text{QuantumEncode}(x_i)$ 
3:  end for
4:  for  $i = 1$  to  $n$  do
5:      Compute  $d_{ij} = 1 - |\langle \psi_i | \psi_j \rangle|^2, \forall j \neq i$ 
6:       $\mathcal{N}_{k_{nn}}(i) \leftarrow k_{nn}$  nearest neighbors by distance
7:  end for
8:  Symmetrize the k-NN graph
9:   $c_{now} \leftarrow n, C_i \leftarrow i, T_i \leftarrow 0, \forall i$ 
10:  $\mathcal{H} \leftarrow \emptyset$  (min-heap)
11: for each edge  $(i, j)$  do
12:     Insert  $(d_{ij}/2, i, j)$  into  $\mathcal{H}$ 
13: end for
14:  $\mathcal{L} \leftarrow 0$ 
15: while  $c_{now} > k$  and  $\mathcal{H} \neq \emptyset$  do
16:      $(d_{min}, i, j) \leftarrow \text{heappop}(\mathcal{H})$ 
17:     if edge  $(i, j)$  is valid then
18:          $\mathcal{L} \leftarrow \mathcal{L} + d_{min}$ 
19:          $C_i \leftarrow C_i \cup C_j, T_i \leftarrow d_{min} + T_i + T_j$ 
20:         Update nearest neighbor list of  $C_i$  and insert new edges into  $\mathcal{H}$ 
21:          $c_{now} \leftarrow c_{now} - 1$ 
22:     end if
23: end while
24: Map cluster labels to  $0, 1, \dots, k - 1$ 
25: return  $l, 2\mathcal{L}$ 

```

The overall computational complexity of QGKM can be summarized as follows. Let n denote the number of samples, d' denote the padded quantum-state dimension, k denote the number of nearest neighbors, and K denote the target number of clusters. Under classical simulation, the initial pairwise fidelity distance computation requires $O(n^2 d')$ time. Based on the computed distance matrix, constructing the k -nearest neighbor graph requires $O(n^2 \log k)$ time when partial sorting or heap-based neighbor selection is used, or $O(n^2 \log n)$ time if full sorting is performed for each sample. The sparse k -nearest neighbor graph contains approximately $O(nk)$ edges in the average case.

During the hierarchical merging stage, QGKM performs approximately $n - K$ merge operations. Candidate merge pairs are maintained in a min-heap, and only local graph-related distances are updated after each merge. If each merge updates approximately $O(k)$ neighboring clusters, the merging stage requires approximately $O((n - K)k \log(nk))$ time. Therefore, using partial neighbor selection, the overall time complexity is approximately $O(n^2 d' + n^2 \log k + (n - K)k \log(nk))$.

If full sorting is used during k -nearest neighbor construction, the second term becomes $O(n^2 \log n)$. The dominant term is usually the initial pairwise fidelity distance computation $O(n^2 d')$, especially for large-scale or high-dimensional datasets.

The space complexity is $O(n^2)$ if the full pairwise distance matrix is stored. If only the encoded states and sparse k -nearest neighbor distances are retained, the space complexity can be reduced to $O(nd' + nk)$.

Although the current manuscript presents the procedure in pseudocode form, the overall workflow of QGKM can be summarized as a sequence of modular operations: data normalization, quantum state encoding, fidelity distance computation, k-nearest neighbor graph construction, graph symmetrization, singleton-cluster initialization, heap-based greedy hierarchical merging, and final cluster label generation.

4. Performance Analysis

To comprehensively validate the effectiveness of the QGKM algorithm, this section presents a multi-faceted experimental evaluation, including visual analysis on synthetic datasets and quantitative comparisons on seven real-world benchmark datasets. The visual experiments demonstrate the ability of QGKM to characterize non-convex structures, while the benchmark experiments compare QGKM with four representative classical clustering algorithms. Statistical significance is further examined using paired *t*-tests to verify the reliability of the observed performance improvements.

4.1. Experiment Configuration

The experiments were conducted on both synthetic and real-world datasets. The synthetic datasets were used for visual analysis, while the real-world benchmark datasets were used for quantitative performance evaluation. These datasets cover diverse data types, dimensionalities, and structural complexities, enabling a comprehensive assessment of the proposed QGKM algorithm.

To comprehensively evaluate the performance of QGKM, four representative clustering algorithms were selected as baselines: K-Means, Spectral Clustering, Gaussian Mixture Model (GMM), and Agglomerative Clustering. The number of clusters was set to the ground-truth number of classes for each dataset. K-Means was adopted as the primary centroid-based baseline and serves as the most direct reference for evaluating improvements over Euclidean-distance-based clustering. Spectral Clustering was included as a representative graph-based method capable of modeling non-convex structures through graph Laplacian embedding, making it closely related to the graph-structured component of QGKM. GMM was selected as a probabilistic model-based clustering method, providing a comparison from the perspective of distribution estimation and soft assignment. Agglomerative Clustering was included as a representative hierarchical method, since QGKM also adopts a bottom-up merging strategy. Therefore, the selected baselines cover centroid-based, graph-based, probabilistic, and hierarchical clustering paradigms, enabling a balanced evaluation of the proposed QGKM framework.

The evaluation metrics include Accuracy (ACC), Normalized Mutual Information (NMI), and Adjusted Rand Index (ARI), which assess clustering quality from different perspectives. For randomized algorithms, each experiment was repeated 10 independent times on each dataset, and the mean and standard deviation were reported. Deterministic algorithms were evaluated once under the same experimental protocol.

4.1.1. Experimental Environment

All experiments were implemented in Python 3.10 using the PyTorch 2.10.0 framework and related scientific computing libraries, including NumPy 1.26.4, SciPy 1.14.0, scikit-learn 1.5.1, and Matplotlib 3.10.0. The experiments were conducted on a server equipped with an NVIDIA GeForce RTX 4090 GPU (NVIDIA Corporation, Santa Clara, CA, USA), an Intel(R) Xeon(R) Gold 6330 CPU (Intel Corporation, Santa Clara, CA, USA), 128 GB RAM (Kingston Technology Company, Inc., Fountain Valley, CA, USA), and a 4 TB SSD (Samsung Electronics Company, Limited, Suwon, Republic of Korea), running Ubuntu 20.04.6 LTS (Canonical Limited, London, UK). The baseline algorithms were implemented using scikit-

learn under consistent parameter settings, and the same data preprocessing procedure and evaluation protocol were applied to all methods to ensure a fair comparison.

4.1.2. Dataset Description

The datasets used in this study cover synthetic, low-dimensional, medium-dimensional, and high-dimensional clustering scenarios. The synthetic datasets include Blobs, Moons, Circles, Anisotropic, and Varied, each containing 1000 samples. These datasets were selected to evaluate the ability of QGKM to identify simple convex structures, non-convex crescent-shaped structures, concentric circular structures, anisotropic distributions, and clusters with different variances.

For real-world datasets, Iris contains 150 samples with 4 features and 3 classes, representing a low-dimensional structured dataset. Wine contains 178 samples with 13 chemical features and 3 classes. Digits contains 1797 handwritten digit images represented by 64-dimensional features and 10 classes, serving as a medium-dimensional image-feature dataset. MNIST and Fashion-MNIST each contain grayscale images with 784-dimensional pixel features and 10 classes. CIFAR-10 consists of color images represented by 3072-dimensional raw pixel features and 10 classes. Olivetti Faces contains face images with 4096-dimensional features and 40 subjects. These datasets provide a broad evaluation spectrum from low-dimensional tabular data to high-dimensional image data, allowing the robustness and scalability of QGKM to be systematically assessed.

4.2. Experiments on Synthetic Datasets

We evaluated the performance of QGKM on five synthetic datasets (Blobs, Moons, Circles, Anisotropic, Varied) and visualized the results. These datasets feature complex geometric structures and distribution characteristics, enabling assessment of the algorithm's ability to handle diverse data distribution patterns. Table 1 demonstrates QGKM's clustering performance across these datasets, where QGKM achieved outstanding clustering results on all datasets.

Table 1. Clustering performance on synthetic datasets.

Dataset	Clusters	ACC	NMI	ARI
Blobs	4	0.999	0.995	0.997
Moons	2	0.999	0.990	0.996
Circles	2	1.000	1.000	1.000
Anisotropic	3	1.000	1.000	1.000
Varied	3	0.992	0.964	0.974

As shown in Table 1, QGKM demonstrates outstanding performance in handling non-convex structures. On the Moons and Circles datasets, the algorithm achieved accuracy rates of 99.90% and 100%, respectively, successfully identifying two interleaved crescent-shaped clusters and two concentric circles. This demonstrates the algorithm's ability to effectively capture non-linearly separable data structures while preserving local geometric properties. On datasets such as Blobs, Anisotropic, and Varied, QGKM also demonstrated outstanding performance with accuracy rates exceeding 99%, indicating the algorithm's excellence in handling data with simple structures and high inter-cluster separability. Particularly on the Anisotropic and Varied datasets, despite significant variations in cluster shapes and variances, QGKM accurately identified all clusters, showcasing robust performance. Figure 1 presents the clustering visualization results of QGKM across five synthetic datasets. It clearly shows that QGKM correctly identifies crescent-shaped

and concentric circular structures, validating the algorithm's effectiveness in handling non-convex data structures.

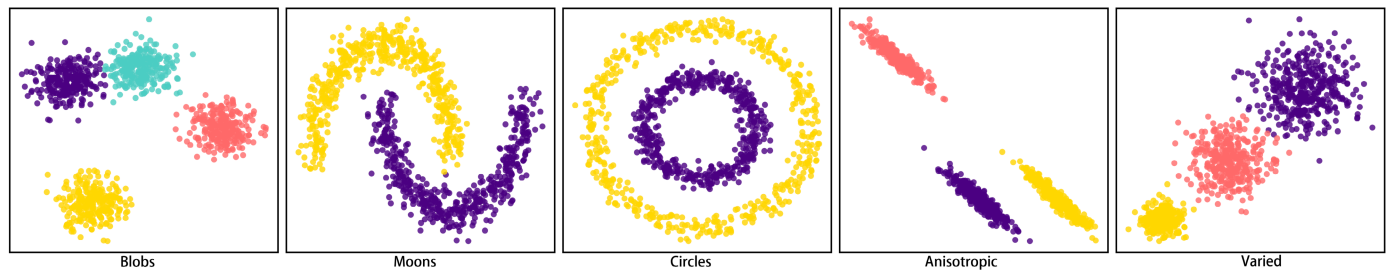


Figure 1. Cluster visualization of the QGKM algorithm on synthetic datasets, each containing 1000 samples. The Moons dataset features two interlaced crescent-shaped structures, while the Circles dataset consists of two concentric circular structures. The QGKM algorithm successfully identified these non-convex clusters, demonstrating its capability to handle complex geometric configurations. Different colors represent different clusters.

The synthetic datasets used in this study mainly cover several representative geometric structures, including convex clusters, non-convex crescent-shaped clusters, concentric circular clusters, anisotropic distributions, and clusters with varying variances. More challenging pseudo-random datasets with partially overlapping clusters and ambiguous decision boundaries will be included in future work to further evaluate the robustness of QGKM under less clearly separated conditions.

4.3. Experiments on Real Datasets

Seven real-world datasets were employed to evaluate the clustering performance of the proposed QGKM algorithm. These datasets encompassed diverse types, including low-dimensional feature data such as Iris and Wine, medium-dimensional handwritten digit recognition data like Digits, high-dimensional image data including MNIST, Fashion-MNIST, and CIFAR-10, as well as the face recognition dataset Olivetti-Faces. Each algorithm was run independently 10 times, recording the mean and standard deviation of each metric. Table 2 presents the performance results of the five clustering algorithms.

K-Means, as the most classic and widely used clustering algorithm, is renowned for its simplicity, efficiency, and ease of implementation. It serves as a crucial benchmark for evaluating the performance of new clustering algorithms. As shown in the experimental results of Table 2, QGKM outperforms K-Means on all seven real-world datasets, demonstrating significant advantages on multiple datasets. The improvement is most pronounced on the Digits dataset, where the ACC increases by 13.1 percentage points. On four datasets—Iris, Olivetti-Faces, Wine, and MNIST—the ACC improvements were 7.7 percentage points, 3.7 percentage points, 2.9 percentage points, and 2.2 percentage points, respectively. QGKM also maintained its lead on the Fashion-MNIST and CIFAR-10 datasets. Overall, QGKM demonstrates comprehensive advantages over K-Means, achieving an average ACC improvement of 4.5 percentage points across seven datasets. The gains are especially pronounced on moderately dimensional datasets (Iris, Wine, Digits), with an average improvement of 7.9 percentage points. This validates QGKM's superiority in capturing the intrinsic structure of data with complex distributions or nonlinear structures. Figures 2 and 3 visually illustrate QGKM's performance improvements over K-Means across various datasets.

Table 2. Comparison of clustering performance on seven real-world datasets. The reported values are the mean $\pm$ standard deviation over 10 independent runs. Bold values indicate the best performance for each dataset and evaluation metric among all compared methods.

Dataset	Metric	K-Means	Spectral	GMM	Agglomerative	QGKM
Iris	ACC	0.890 $\pm$ 0.004	0.907 $\pm$ 0.000	0.967 $\pm$ 0.000	0.893 $\pm$ 0.000	0.967 $\pm$ 0.000
	NMI	0.744 $\pm$ 0.008	0.796 $\pm$ 0.000	0.898 $\pm$ 0.000	0.761 $\pm$ 0.000	0.885 $\pm$ 0.000
	ARI	0.723 $\pm$ 0.007	0.759 $\pm$ 0.000	0.904 $\pm$ 0.000	0.731 $\pm$ 0.000	0.904 $\pm$ 0.000
Wine	ACC	0.960 $\pm$ 0.009	0.961 $\pm$ 0.000	0.959 $\pm$ 0.008	0.927 $\pm$ 0.000	0.989 $\pm$ 0.000
	NMI	0.862 $\pm$ 0.018	0.858 $\pm$ 0.000	0.859 $\pm$ 0.013	0.783 $\pm$ 0.000	0.946 $\pm$ 0.000
	ARI	0.879 $\pm$ 0.028	0.880 $\pm$ 0.000	0.875 $\pm$ 0.023	0.790 $\pm$ 0.000	0.965 $\pm$ 0.000
Digits	ACC	0.755 $\pm$ 0.066	0.808 $\pm$ 0.000	0.789 $\pm$ 0.068	0.831 $\pm$ 0.000	0.886 $\pm$ 0.000
	NMI	0.722 $\pm$ 0.030	0.833 $\pm$ 0.000	0.783 $\pm$ 0.034	0.838 $\pm$ 0.000	0.809 $\pm$ 0.000
	ARI	0.631 $\pm$ 0.060	0.758 $\pm$ 0.000	0.689 $\pm$ 0.062	0.768 $\pm$ 0.000	0.771 $\pm$ 0.000
MNIST	ACC	0.515 $\pm$ 0.043	0.581 $\pm$ 0.036	0.516 $\pm$ 0.042	0.582 $\pm$ 0.048	0.537 $\pm$ 0.062
	NMI	0.480 $\pm$ 0.026	0.602 $\pm$ 0.028	0.481 $\pm$ 0.026	0.578 $\pm$ 0.039	0.540 $\pm$ 0.042
	ARI	0.341 $\pm$ 0.038	0.439 $\pm$ 0.036	0.342 $\pm$ 0.038	0.419 $\pm$ 0.053	0.371 $\pm$ 0.075
Fashion-MNIST	ACC	0.544 $\pm$ 0.055	0.538 $\pm$ 0.023	0.550 $\pm$ 0.053	0.505 $\pm$ 0.031	0.571 $\pm$ 0.057
	NMI	0.526 $\pm$ 0.018	0.575 $\pm$ 0.017	0.541 $\pm$ 0.020	0.532 $\pm$ 0.023	0.555 $\pm$ 0.025
	ARI	0.376 $\pm$ 0.032	0.400 $\pm$ 0.024	0.397 $\pm$ 0.030	0.371 $\pm$ 0.029	0.401 $\pm$ 0.040
CIFAR-10	ACC	0.219 $\pm$ 0.011	0.213 $\pm$ 0.010	0.221 $\pm$ 0.011	0.209 $\pm$ 0.012	0.224 $\pm$ 0.012
	NMI	0.096 $\pm$ 0.008	0.102 $\pm$ 0.012	0.098 $\pm$ 0.007	0.094 $\pm$ 0.011	0.101 $\pm$ 0.012
	ARI	0.041 $\pm$ 0.004	0.039 $\pm$ 0.006	0.042 $\pm$ 0.003	0.037 $\pm$ 0.006	0.043 $\pm$ 0.006
Olivetti-Faces	ACC	0.588 $\pm$ 0.023	0.644 $\pm$ 0.011	0.588 $\pm$ 0.023	0.647 $\pm$ 0.000	0.627 $\pm$ 0.005
	NMI	0.761 $\pm$ 0.012	0.794 $\pm$ 0.006	0.761 $\pm$ 0.012	0.807 $\pm$ 0.000	0.799 $\pm$ 0.001
	ARI	0.443 $\pm$ 0.028	0.494 $\pm$ 0.014	0.443 $\pm$ 0.028	0.518 $\pm$ 0.000	0.520 $\pm$ 0.004

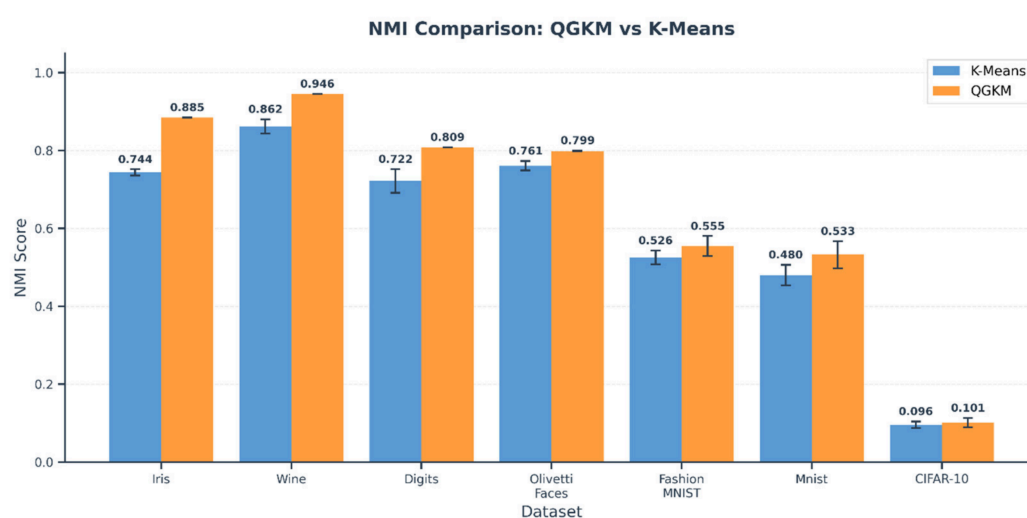


Figure 2. QGKM outperformed K-Means in NMI performance across all datasets, indicating its superior ability to capture the clustering structure of the data.

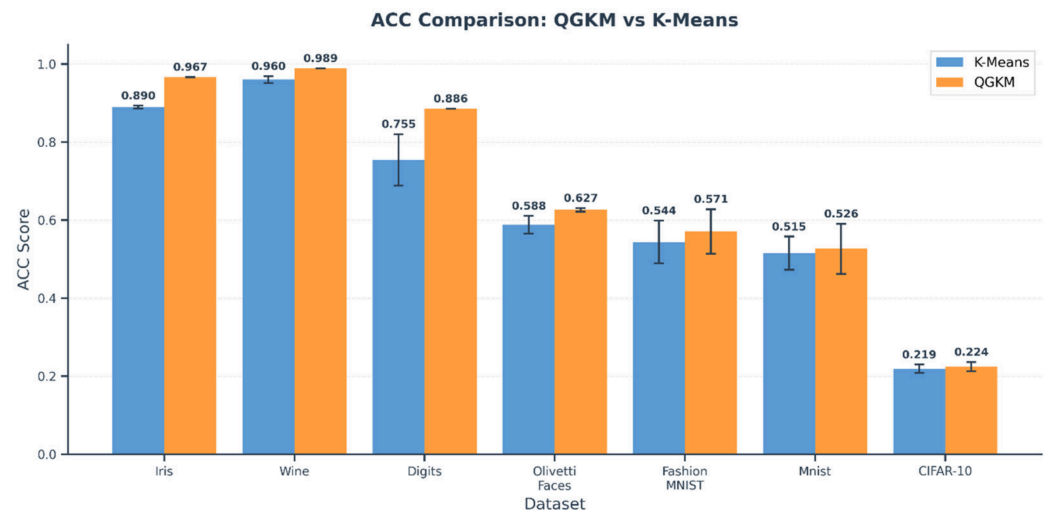


Figure 3. Comparison of clustering accuracy between QGKM and K-Means across seven datasets. QGKM outperforms the traditional K-Means algorithm on all datasets.

Compared to other classical clustering methods, QGKM also demonstrates remarkable performance. On the Iris, Wine, and Digits datasets, QGKM achieved the best performance. Spectral Clustering and Agglomerative outperformed QGKM on the MNIST and Olivetti-Faces datasets, but QGKM held an advantage on the remaining five datasets. Regarding GMM, QGKM tied with it for the best performance on the Iris dataset but outperformed it on the remaining six datasets. In summary, QGKM achieved the best performance on five datasets, only slightly lagging behind Spectral and Agglomerative on the MNIST and Olivetti-Faces datasets. In comparison with other datasets and methods, QGKM consistently demonstrated superior performance.

Experimental results demonstrate that the performance advantages of QGKM are closely tied to the dimensional characteristics of the dataset. Specifically, on moderately dimensional datasets such as Iris (4-dimensional), Wine (13-dimensional), and Digits (64-dimensional), QGKM exhibits the most significant improvements, achieving ACC values of 96.7%, 98.9%, and 88.6%, respectively—all substantially outperforming all baseline methods. This indicates that within this dimensional range, quantum fidelity distance effectively captures the local structure of the data without being significantly disturbed by high-dimensional noise. In contrast, on extremely high-dimensional datasets (e.g., MNIST (784-dimensional), Fashion-MNIST (784-dimensional), and CIFAR-10 (3072-dimensional)), QGKM's performance is roughly comparable to or slightly better than mainstream baseline methods, but the improvement is limited. This phenomenon aligns with the “dimension collapse” issue commonly observed in similarity metrics within high-dimensional spaces, further confirming that QGKM is more suitable for data analysis scenarios featuring clear local structures and dimensions within a moderate range.

4.4. Quantum Encoding Comparison Experiments

The core of the QGKM algorithm lies in mapping classical data to quantum state space and utilizing quantum fidelity distance to embed graph K-Means for clustering. To investigate the impact of quantum state encoding on QGKM performance, this paper implements two primary quantum state encodings—amplitude encoding and angle encoding—and conducts comparative encoding experiments on the same seven real-world datasets.

The encoding comparison experiments employed the same datasets and evaluation metrics as Section 4.3. Each encoding method was independently run 10 times on each dataset, recording the mean and standard deviation of three metrics: ACC, NMI, and ARI. Both encoding approaches utilized an identical QGKM algorithm framework, identical k-

nearest neighbor parameters, and identical random seed sequences to ensure experimental variations stem solely from the differences between the two quantum state encodings. The sole difference lies in the mapping process: amplitude encoding directly maps normalized data vectors to quantum state amplitude components, whereas angle encoding requires an additional step—using trigonometric functions to map data components to quantum state rotation angles.

Figure 4 shows the NMI performance of the two encoding methods across seven datasets. The results reveal that amplitude coding significantly outperforms angle coding on low-dimensional datasets such as Iris, Wine, and Digits. The gap reaches as high as 39.1 percentage points on the Wine dataset. On MNIST and Fashion-MNIST datasets, amplitude coding maintains a slight lead. However, on extremely high-dimensional datasets like CIFAR-10 and Olivetti-Faces, the performance of amplitude coding and angle coding is comparable, with negligible differences.

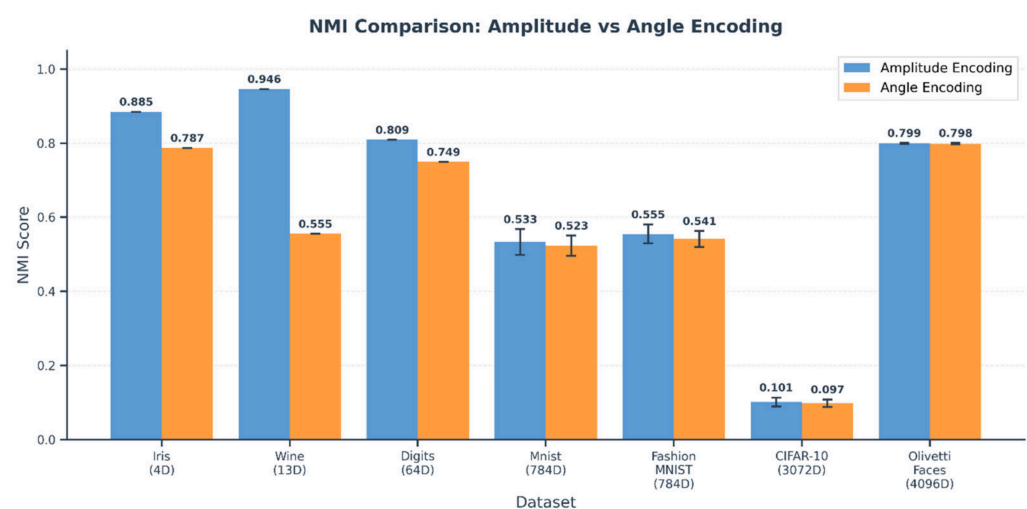


Figure 4. Comparison of normalized mutual information (NMI) between amplitude encoding and angle encoding across seven real-world datasets. Higher NMI values indicate better consistency between the clustering results and the ground-truth class labels. The results show that amplitude encoding generally achieves better performance on low- and medium-dimensional datasets, while the performance gap between the two encoding schemes becomes smaller on high-dimensional datasets.

In high-dimensional spaces, the NMI values of the two encoding schemes tend to converge, reflecting the “curse of dimensionality” effect: as the dimensionality increases, the similarity between samples becomes more uniform, reducing the ability of clustering algorithms to distinguish different clusters and consequently diminishing the differences in NMI. This indicates that the choice of encoding scheme has a greater impact on clustering performance for low- and medium-dimensional data, whereas for high-dimensional data, applying dimensionality reduction or feature extraction may be more effective in improving clustering results than the choice of encoding. Based on these experimental findings, Amplitude Encoding is adopted as the default quantum encoding scheme for QGKM in all subsequent experiments.

4.5. Statistical Significance Test

To evaluate whether the performance improvements of QGKM over baseline methods are statistically significant, paired *t*-tests [40] were conducted on the seven datasets. For each dataset, QGKM was compared with the other four methods (K-Means, Spectral, GMM, and Agglomerative) in terms of ACC, NMI, and ARI across 10 independent runs.

Table 3 presents the p -values of QGKM compared with each baseline method across different datasets and metrics, where bold p -values indicate significant differences ($p < 0.05$). These statistical tests provide a reliable measure of the significance of the observed performance improvements, validating the effectiveness of QGKM in comparison to other clustering methods.

Table 3. Results of paired t -tests comparing QGKM with baseline clustering methods across seven real-world datasets. Bold p -values indicate statistically significant differences at the 0.05 significance level.

Comparative Method	Iris	Wine	Digits	MNIST	Fashion	CIFAR	Olivetti
ACC							
K-Means	<0.001	<0.001	<0.001	0.222	0.686	0.830	<0.001
Spectral	<0.001	<0.001	<0.001	0.005	0.739	0.112	<0.001
GMM	1.000	<0.001	<0.001	0.541	0.115	0.226	<0.001
Agglomerative	<0.001	<0.001	<0.001	0.017	0.012	0.087	<0.001
NMI							
K-Means	<0.001	<0.001	<0.001	<0.001	0.005	0.165	<0.001
Spectral	<0.001	<0.001	0.056	<0.001	<0.001	0.191	<0.001
GMM	0.999	<0.001	<0.001	<0.001	<0.001	0.005	<0.001
Agglomerative	<0.001	<0.001	<0.001	<0.001	0.374	0.498	<0.001
ARI							
K-Means	<0.001	<0.001	<0.001	0.128	0.250	0.130	<0.001
Spectral	<0.001	<0.001	0.174	<0.001	0.019	0.276	<0.001
GMM	1.000	<0.001	<0.001	0.035	0.003	0.017	<0.001
Agglomerative	<0.001	<0.001	0.891	0.006	0.499	0.100	0.999

Based on the significance test results, the QGKM method demonstrates significantly better performance than K-Means across multiple datasets. On the Iris, Wine, Digits, and Olivetti-Faces datasets, QGKM significantly outperformed K-Means in all three evaluation metrics—ACC, NMI, and ARI—at a statistical significance level of $p < 0.001$. These datasets feature dimensions ranging from 4 to 4096, indicating that the quantum fidelity-based distance metric markedly improves clustering performance across varying data dimensions. For the MNIST dataset, although the difference in accuracy (ACC) was not significant ($p = 0.222$), QGKM still significantly outperformed K-Means in NMI ($p < 0.001$), indicating that this method has a greater advantage in capturing the internal structure of the data. On Fashion-MNIST, the NMI of QGKM is significantly higher than K-Means ($p = 0.005$), while the ACC difference is not significant. On CIFAR-10, although ACC values are comparable, the difference is not significant, reflecting the extreme challenge of this dataset. Overall, QGKM achieves statistically significant improvements over K-Means in ACC on 5 out of 7 datasets and in NMI on 6 datasets, demonstrating that the quantum fidelity distance provides a statistically meaningful performance gain over traditional Euclidean distance, particularly on medium-dimensional datasets and in terms of mutual information.

Compared with other baseline methods, QGKM also exhibits significant performance advantages. For GMM, QGKM achieves the same best performance as GMM on Iris (ACC = 96.7%, $p = 1.000$), but significantly outperforms GMM on Wine and Digits ($p < 0.001$). Compared with Spectral Clustering and Agglomerative Clustering, QGKM shows very high significance on Wine and Digits ($p < 0.001$), and also significantly outperforms Spectral and GMM on Olivetti-Faces. On extremely high-dimensional datasets

(MNIST, Fashion-MNIST, CIFAR-10), the statistical significance of QGKM is mainly observed in the NMI metric rather than ACC. This indicates that while the quantum fidelity distance is constrained by the curse of dimensionality in terms of accuracy, it still significantly improves the mutual information between clustering results and ground-truth labels, better capturing the intrinsic data structure.

Considering the overall statistical significance, among the 28 valid comparisons (7 datasets $\times$ 4 baseline methods), QGKM shows statistically significant improvements in 17/28 (60.7%) of ACC comparisons, 24/28 (85.7%) of NMI comparisons, and 18/28 (64.3%) of ARI comparisons. The overall trend indicates that QGKM consistently exhibits statistically meaningful advantages over baseline methods on most datasets and evaluation metrics, with the highest consistency observed in the NMI metric. This suggests that the quantum fidelity-induced similarity measure has a potential advantage in capturing the information correlation between clustering results and the true class structure. On medium-dimensional datasets (Iris, Wine, Digits), QGKM shows statistical significance against K-Means and other methods in nearly all comparisons, highlighting its strongest application scenario. On extremely high-dimensional datasets, significance is primarily observed in the NMI metric, indicating that the quantum fidelity distance retains certain advantages in terms of information gain, even in the presence of high-dimensional noise. This suggests that QGKM remains effective at capturing meaningful data structures, particularly in terms of preserving the underlying relationships within the data.

Through experimental validation, the QGKM algorithm demonstrates significant advantages on both synthetic and real-world datasets. Experiments show that the algorithm effectively handles complex non-convex data structures, achieving near-ideal clustering results on synthetic datasets such as Moons. Across all seven real-world datasets, QGKM outperformed traditional K-Means methods, achieving an average accuracy improvement of 4.5 percentage points. This enhancement was particularly pronounced on moderately dimensional datasets (e.g., Iris, Wine, Digits), where the average gain reached 7.9 percentage points. Statistical tests further confirm that QGKM significantly outperforms baseline methods in 60.7% of ACC comparisons and 85.7% of NMI comparisons. The pronounced advantage in NMI metrics particularly indicates that the adopted quantum fidelity distance more accurately captures the intrinsic structure and information correlations within the data.

5. Conclusions and Future Work

The proposed quantum fidelity-based graph K-Means clustering method (QGKM) effectively translates classical clustering problems into the quantum domain by leveraging quantum state encoding and fidelity distance. It achieves clustering optimization through a combination of k-nearest neighbor graphs and a greedy hierarchical merging mechanism. Test results across seven real-world datasets demonstrate that this method outperforms classical K-Means across all metrics, achieving an average accuracy improvement of 4.5%. Statistical tests further confirm its superiority over baseline methods in 60.7% of accuracy comparisons and 85.7% of normalized mutual information comparisons, particularly excelling in mutual information metrics reflecting structural similarity. This validates the quantum fidelity distance's capability to effectively capture intrinsic data structures. Furthermore, on synthetic datasets, QGKM demonstrates excellent recognition capabilities in non-convex distributions such as crescent-shaped and concentric circular patterns, proving its adaptability in complex pattern analysis.

From a practical perspective, the proposed QGKM framework can be applied to several intelligent education and Educational Social Network systems. First, it can support learner profiling by identifying groups of students with similar behavioral patterns, inter-

action habits, or learning preferences. Second, it can be used for community detection in collaborative learning platforms, where learners are connected through communication, discussion, resource sharing, and group activities. Third, QGKM can assist personalized recommendation systems by discovering latent learner groups and recommending learning resources, peer collaborators, or learning paths according to cluster-level characteristics. Fourth, it can be applied to abnormal learning behavior detection, where learners who significantly deviate from normal interaction patterns may be identified as potential at-risk students requiring timely intervention. Compared with traditional K-Means, the use of quantum fidelity distance and k-nearest neighbor graph structure enables QGKM to better capture nonlinear and local relationships in educational data, thereby improving clustering stability, learner group identification accuracy, and the reliability of downstream educational decision-making.

Although QGKM demonstrates outstanding performance across multiple datasets, several limitations remain that require improvement in future work. First, while quantum fidelity distance holds theoretical advantages over traditional Euclidean distance, its computational overhead remains substantial on extremely high-dimensional datasets. Second, the algorithm's performance is sensitive to parameter selection in the k-nearest neighbor graph, and there is currently a lack of theoretical guidance for adaptively selecting the optimal k value. Third, while the greedy merging strategy employed by QGKM avoids the impact of random initialization, it may sometimes get stuck in local optima. Finally, the algorithm assumes the number of clusters k is known, which often requires heuristic methods or domain knowledge to determine in practical applications.

Based on the achievements and limitations of the current work, future research can be pursued in the following directions. First, investigate adaptive parameter selection mechanisms to automatically determine the parameters of k-nearest neighbor graphs and the target number of clusters through data-driven approaches, thereby reducing the need for manual parameter tuning. Information criteria from model selection theory or cross-validation methods can be adopted to evaluate the clustering quality of different parameter configurations. Second, extend QGKM to semi-supervised and active learning scenarios, leveraging sparse labeled data to guide the clustering process and enhance performance on complex datasets. Third, explore integrating QGKM with deep learning by employing neural networks to learn more effective data representations, then applying QGKM for clustering within the learned feature space.

Overall, QGKM is a quantum-inspired graph-structured K-Means clustering method. Its core contribution lies in systematically introducing quantum fidelity distances into the classical clustering framework and enhancing algorithmic stability through a deterministic hierarchical merging strategy. This work demonstrates the potential of quantum fidelity distance in capturing complex data structures in medium-dimensionality scenarios. With the ongoing advancement of quantum computing technology and deepening theoretical research, quantum-inspired clustering methods are poised to play a significant role in more practical applications, offering novel solutions for complex data analysis.

Author Contributions: Conceptualization, N.N.X. and W.L.; method N.N.X., W.L., D.H., X.M. and Z.D.; software, W.L., D.H., X.M. and Z.D.; validation, D.H. and X.M.; formal analysis, S.M.A. and Y.K.; investigation, N.N.X. and W.L.; resources, N.N.X. and W.L.; data curation S.M.A. and Y.K.; writing-original draft preparation, N.N.X. and W.L. writing—review and editing, N.N.X., S.M.A. and Y.K.; visualization, W.L.; supervision, N.N.X.; project administration, N.N.X.; funding acquisition, N.N.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was in part by NSFC Project with No. U25A20428 and the Hunan Provincial Key Research and Development Program under Grant 2024AQ2028.

Data Availability Statement: The original data presented in the study are openly available at <https://archive.ics.uci.edu/ml/datasets/iris> (accessed on 25 April 2026), <https://archive.ics.uci.edu/ml/datasets/wine> (accessed on 25 April 2026) and <http://yann.lecun.com/exdb/mnist/> (accessed on 25 April 2026).

Conflicts of Interest: The authors declare no conflict of interest.

References

- Li, T.; Liu, W.; Zeng, Z.; Xiong, N.N. DRLR: A deep-reinforcement-learning-based recruitment scheme for massive data collections in 6G-based IoT networks. *IEEE Internet Things J.* **2021**, *9*, 14595–14609. [\[CrossRef\]](#)
- Guo, W.; Xiong, N.; Vasilakos, A.V.; Chen, G.; Yu, C. Distributed k-connected fault-tolerant topology control algorithms with PSO in future autonomic sensor systems. *Int. J. Sens. Netw.* **2012**, *12*, 53–62. [\[CrossRef\]](#)
- Yu, J.; Liu, Z.; Liang, W.; Han, X.; Xiong, N. The Entropy-Based Hop Scheme for Influence Maximization in Dynamic Social Networks. *Hum.-Centric Comput. Inf. Sci.* **2023**, *13*, 40. [\[CrossRef\]](#)
- Zhang, Y.; Yao, Y.; Lin, C.-C.; Chang, C.-C. Dual-image reversible data hiding based on Sudoku block mapping and adaptive embedding. *J. Internet Technol.* **2025**, *26*, 147–163. [\[CrossRef\]](#)
- Pan, J.; Chen, Y.; Zhao, C.; Liu, Y.; Chu, J. Long text classification model based on Transformer sliding window and threshold optimization. *J. Internet Technol.* **2025**, *26*, 231–240. [\[CrossRef\]](#)
- Abdussamad, A.M.; Inayat, A. Addressing limitations of the K-means clustering algorithm: Outliers, non-spherical data, and optimal cluster selection. *AIMS Math.* **2024**, *9*, 25070–25097. [\[CrossRef\]](#)
- Chu, S.-C.; Li, Z.; Pan, J.-S.; Tseng, K.-K.; Kong, L. The interpretable graph neural network of ECG beat classification. *J. Internet Technol.* **2025**, *26*, 453–461. [\[CrossRef\]](#)
- Wei, X.; Song, H.; Chang, Y.; Jiang, W.; Xu, L. Integrated deep learning and attention mechanisms for accurate license plate recognition. *J. Internet Technol.* **2025**, *26*, 513–522. [\[CrossRef\]](#)
- Chen, X.; Zhang, Z.; Qiu, A.; Xia, Z.; Xiong, N.N. Novel coverless steganography method based on image selection and StarGAN. *IEEE Trans. Netw. Sci. Eng.* **2020**, *9*, 219–230. [\[CrossRef\]](#)
- Berahmand, K.; Saberi-Movahed, F.; Sheikhpour, R.; Li, Y.; Jalili, M. A comprehensive survey on spectral clustering with graph structure learning. *arXiv* **2025**, arXiv:2501.13597. [\[CrossRef\]](#)
- Peng, Y.; Li, D.; Li, X.; Liang, Z.; Ding, Y.; Wang, Y. Quantum-inspired fidelity-based divergence. *arXiv* **2025**, arXiv:2501.19307.
- Xiong, N.N.; Zhang, H.; Chen, Y.; Jin, K.; Liang, W. QHGA: An Effective Quantum Hybrid Genetic Algorithm to Achieve an Industrial Reliable Qubit Allocation. *Hum.-Centric Comput. Inf. Sci.* **2026**, *16*, 8. [\[CrossRef\]](#)
- Xiong, N.; Kuang, C.; Liang, W.; Chen, L.; Shen, M. Efficient Routing Scheme Based on Segmentation in Quantum Networks. *IEEE Internet Things J.* **2025**, *12*, 52401–52412. [\[CrossRef\]](#)
- Devadas, R.M.; Sowmya, S.T. Quantum machine learning: A comprehensive review of integrating AI with quantum computing for computational advancements. *MethodsX* **2025**, *14*, 103318. [\[CrossRef\]](#) [\[PubMed\]](#)
- Varga, M.; Bermejo, P.; Pellicer-Guridi, R.; Orús, R.; Molina-Terriza, G. Quantum-inspired clustering with light. *Sci. Rep.* **2024**, *14*, 21726. [\[CrossRef\]](#)
- Liu, L.; Fan, X. Optimizing Grover’s algorithm for routing in quantum wireless communication networks. *J. Internet Technol.* **2024**, *25*, 851–863. [\[CrossRef\]](#)
- Ikotun, A.M.; Ezugwu, A.E.; Abualigah, L.; Abuhaija, B.; Heming, J. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Inf. Sci.* **2022**, *622*, 178–210. [\[CrossRef\]](#)
- Ahmed, M.; Seraj, R.; Islam, S.M.S. The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics* **2020**, *9*, 1295. [\[CrossRef\]](#)
- Ren, Y.; Pu, J.; Yang, Z.; Xu, J.; Li, G.; Pu, X.; Yu, P.S.; He, L. Deep clustering: A comprehensive survey. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, *36*, 5858–5878. [\[CrossRef\]](#)
- Zhou, S.; Xu, H.; Zheng, Z.; Chen, J.; Li, Z.; Bu, J.; Wu, J.; Wang, X.; Zhu, W.; Ester, M. A comprehensive survey on deep clustering: Taxonomy, challenges, and future directions. *ACM Comput. Surv.* **2024**, *57*, 1–38. [\[CrossRef\]](#)
- Sarfraz, S.; Sharma, V.; Stiefelhofen, R. Efficient parameter-free clustering using first neighbor relations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 8934–8943.
- Yang, B.; Wu, J.; Sun, A.; Gao, N.; Zhang, X. Robust landmark graph-based clustering for high-dimensional data. *Neurocomputing* **2022**, *496*, 72–84. [\[CrossRef\]](#)
- Shen, S.; Li, W.; Zhu, Z.; Huang, G.; Du, D.; Lu, J.; Zhou, J. Structure-aware face clustering on a large-scale graph with 107 nodes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 9085–9094.

24. Li, Y.; Nguyen, J.; Anastasiu, D.C.; Arriaga, E.A. CosTaL: An accurate and scalable graph-based clustering algorithm for high-dimensional single-cell data analysis. *Brief. Bioinform.* **2023**, *24*, bbad157. [[CrossRef](#)]
25. Xue, J.; Xing, L.; Wang, Y.; Fan, X.; Kong, L.; Zhang, Q.; Nie, F.; Li, X. A comprehensive survey of fast graph clustering. *Vicinagearth* **2024**, *1*, 7. [[CrossRef](#)]
26. Biamonte, J.; Wittek, P.; Pancotti, N.; Rebentrost, P.; Wiebe, N.; Lloyd, S. Quantum machine learning. *Nature* **2017**, *549*, 195–202. [[CrossRef](#)] [[PubMed](#)]
27. Cerezo, M.; Verdon, G.; Huang, H.-Y.; Cincio, L.; Coles, P.J. Challenges and opportunities in quantum machine learning. *Nat. Comput. Sci.* **2022**, *2*, 567–576. [[CrossRef](#)] [[PubMed](#)]
28. Schuld, M.; Killoran, N. Quantum Machine Learning in Feature Hilbert Spaces. *Phys. Rev. Lett.* **2019**, *122*, 040504. [[CrossRef](#)]
29. Havlíček, V.; Córcoles, A.D.; Temme, K.; Harrow, A.W.; Kandala, A.; Chow, J.M.; Gambetta, J.M. Supervised learning with quantum-enhanced feature spaces. *Nature* **2019**, *567*, 209–212. [[CrossRef](#)]
30. Felser, T.; Trenti, M.; Sestini, L.; Gianelle, A.; Zuliani, D.; Lucchesi, D.; Montangero, S. Quantum-inspired machine learning on high-energy physics data. *npj Quantum Inf.* **2021**, *7*, 111. [[CrossRef](#)]
31. Lazarev, I.D.; Narozniak, M.; Byrnes, T.; Pyrkov, A.N. Hybrid quantum-classical unsupervised data clustering based on the self-organizing feature map. *Phys. Rev. A* **2025**, *111*, 012416. [[CrossRef](#)]
32. Date, P.; Arthur, D.; Pusey-Nazzaro, L. QUBO formulations for training machine learning models. *Sci. Rep.* **2021**, *11*, 10029. [[CrossRef](#)]
33. Bai, L.; Cheng, X.; Liang, J.; Guo, Y. Fast graph clustering with a new description model for community detection. *Inf. Sci.* **2017**, *388*, 37–47. [[CrossRef](#)]
34. Murtagh, F.; Contreras, P. Algorithms for hierarchical clustering: An overview, II. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2017**, *7*, e1219. [[CrossRef](#)]
35. Chehreghani, M.H.; Chehreghani, M.H. Hierarchical correlation clustering and tree preserving embedding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 23083–23093.
36. Schuld, M.; Bocharov, A.; Svore, K.M.; Wiebe, N. Circuit-centric quantum classifiers. *Phys. Rev. A* **2020**, *101*, 032308. [[CrossRef](#)]
37. Pérez-Salinas, A.; Cervera-Lierta, A.; Gil-Fuster, E.; Latorre, J.I. Data re-uploading for a universal quantum classifier. *Quantum* **2020**, *4*, 226. [[CrossRef](#)]
38. Schuld, M.; Sweke, R.; Meyer, J.J. Effect of data encoding on the expressive power of variational quantum-machine-learning models. *Phys. Rev. A* **2021**, *103*, 032430. [[CrossRef](#)]
39. Malkov, Y.A.; Yashunin, D.A. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *42*, 824–836. [[CrossRef](#)]
40. Rainio, O.; Teuho, J.; Klén, R. Evaluation metrics and statistical tests for machine learning. *Sci. Rep.* **2024**, *14*, 6086. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.