Journal of Applied Research and Technology

www.jart.icat.unam.mx

ICAT Instituto de Ciencias Aplicadas y Tecnología

Journal of Applied Research and Technology 23 (2025) 546-563

Original

Design of an Improved Method for Clustering Using Variational Autoencoders, DBSCAN, and Genetic Algorithms

J. Harde[1]* • S. Karmore[2]

[1]Department of Computer Science and Engineering (Data Science), St. Vincent Pallotti College of Engineering and Technology, Nagpur.
[2]Department of Computer Science and Engineering, TGPCET, Nagpur.

Received: 08 07 2024; Accepted: 11 02 2025
Available: 12 31 2025

Abstract: With the increase in data complexity and volume, the demand for more accurate clustering methods in data analysis has grown to an importance bordering on the critical. In many applications, existing clustering methods perform poorly on high-dimensional data, in the presence of noise, and for the identification of arbitrary-shaped clusters. In this setting, the current study develops a novel framework that integrates density-based clustering with deep learning, rule mining, and genetic algorithms to improve clustering accuracy. Traditional clustering algorithms, such as *k-means* and hierarchical clustering, are limited in their ability to handle complex data distributions and noise. Predefined cluster shapes drive these methods and work suboptimally with high-dimensional data. Our approach can overcome such challenges by leveraging the variability of Variational Autoencoders—DBSCAN, the *Apriori* algorithm with decision trees, and an Adaptive Genetic Algorithm for parameter optimization. This is the realm into which VAEs, in conjunction with DBSCAN, finally outperform traditional clustering methods. VAEs, on the other hand, can model complex data distributions and reduce their dimensionality, thereby making the data more amenable to clustering. Subsequently, DBSCAN is applied to the lower-dimensional latent representations produced by VAEs to identify clusters of arbitrary shapes that are robust to noise. This combination resulted in high clustering accuracy, with an Adjusted Rand Index of 0.85 and a significant reduction in the impact of noise on sets. We use the *Apriori* algorithm and decision trees for cluster interpretation. The *Apriori* algorithm finds frequent itemsets in each

*Corresponding author.
E-mail address: jayshri.banpurkar@gmail.com (J. Harde).
Peer Review under the responsibility of Universidad Nacional Autónoma de México.

cluster, while decision trees produce understandable, readable rules for membership in each cluster. This yielded an accuracy of 0.80 for the rules and gave high interpretability due to the simplicity of the paths taken to the decisions. Finally, the AGA dynamically adjusts the DBSCAN clustering parameters, epsilon and *minPts*, improving convergence speed and overall clustering performance. In this process, the optimal set of parameters was reached at generation 50, resulting in roughly a 15% improvement in the Silhouette score over the default parameters. VAE, DBSCAN, *Apriori* algorithm, decision trees, and AGA constitute comprehensive clustering frameworks that achieve modest improvements in accuracy and defect reduction over existing methods. The contribution of the present work can therefore be twofold: it designs a solution that would overcome the drawbacks of traditional methods and offer a robust, interpretable, and optimized clustering approach for complex data, and it further advances this field.

Keywords: **Clustering, Variational Autoencoders, DBSCAN, Genetic Algorithms, Rule Mining, Process**

1. Introduction

With the growing complexity of large volumes of data in this era of significant data, understanding data analytics and interpretation has become highly important. One of the standard methods in unsupervised machine learning, clustering aims to group data points into clusters, based on similarities among themselves to be able to find some patterns or structure inherent in the data samples in the most convenient way possible (Qiu & Li, 2022; Uykan, 2023; Li & Wang, 2023) However, the traditional clustering methods based on k-means and hierarchical clustering leave much to be desired, especially when handling high-dimensional data, noise, and clusters of arbitrary shapes. These all undoubtedly lead to a situation in which more advanced and respective clustering techniques have to be developed to handle the complexities just mentioned. These make traditional clustering sensitive to the initial selection of cluster centers, leading to local optima that often converge, resulting in suboptimal clustering. They also assume that cluster shapes are spherical and that the number of clusters is given a priori, which is often impractical. Hierarchical clustering, for its part, can produce a dendrogram that represents nested clusters, but it is computationally expensive and not robust to noise. These inherent limitations would jeopardize the applicability of these methodologies to modern datasets, characterized by high dimensionality and noise. A more robust alternative is the use of density-based clustering methods, such as Density-Based Spatial Clustering of Applications with Noise, or DBSCAN. With this feature, the algorithm can detect arbitrary shapes in the data and cluster appropriately, while effectively handling noise due to its sensitivity to data point density. However, performance depends on the choice of parameters: the maximum distance between two points to consider them neighbors (epsilon) and the minimum number of points required to form a dense region *(minPoints)*. The challenge lies in determining these two parameters, and prior to this, they rely on numerous complex heuristics for a high-dimensional dataset (Guan et al., 2023; Li et al., 2022a; Li et al., 2023). More precisely, the framework presented here includes deep learning, rule mining, and genetic algorithms, together with a density-based approach to help enhance the accuracy and interpretability of the clusters. We leverage variational autoencoders to capture the complex data distribution and reduce the dimensionality of the data to be clustered. A VAE is a probabilistic generative model that learns a low-dimensional representation of the input data via its mapping and later reconstructs it from the induced low-dimensional space. Besides noise reduction, it may also improve clustering, as the representation captures only the salient features of the data. After the VAE transforms the data into a lower-dimensional latent space, DBSCAN is applied to identify clusters. However, the noise robustness of DBSCAN is complemented by the feature-extraction capabilities of VAEs, resulting in high-accuracy clustering and the detection of arbitrary-shaped clusters. However, the issue of selecting the optimal DBSCAN parameters persists. An Adaptive Genetic Algorithm (AGA) is thus used to optimize clustering parameters dynamically.

Inspired by natural selection, genetic algorithms evolve an iterative population of potential solutions to optimize parameters. Drawing on principles of natural selection, AGA dynamically adapts its configuration parameters to enhance convergence speed and clustering performance.

Apart from clustering, interpreting clusters derived from the data model is significant, as it provides meaning to the data patterns. The *Apriori* classical rule-mining technique has been applied in this study to discover frequent itemsets within each existing cluster. These *itemsets* served as the basis for rules describing the attributes of clusters. These itemsets are then converted into meaningful, interpretable rules using decision trees, providing a clear understanding of the cluster's membership. By combining VAEs, DBSCAN, AGA, and rule mining in this order, we create a robust framework for clustering high-dimensional, noisy data samples. Extensive experiments on various datasets will evaluate the effectiveness of this method. Hence, some of the results shown are improvements in clustering accuracy, noise handling, and interpretability over traditional methods. This framework built up for the clustering techniques can solve not only the problems with the existing techniques but also perform well in the data analysis from bioinformatics up to market segmentation.

1.1 Motivation & Contribution

The motivation for this work is that enhancing clustering accuracy is becoming increasingly necessary given increasingly complex, high-dimensional data samples. While traditional clustering methods are foundational in this area, it is found to be highly limited when applied to modern datasets characterized by high noise and intricate data distributions. Therefore, in this work, state-of-the-art techniques from deep learning, density-based clustering, rule mining, and genetic algorithms will be combined to overcome these limitations. The proposed framework seeks to leverage these strengths to the fullest, providing a robust, accurate, and interpretable clustering solution. On the contributions side, this paper is among the first attempts at integrating Variational Autoencoders with DBSCAN. VAEs can effectively reduce dimensionality, capturing complex data structures and providing an appropriate setting for clustering. Moreover, DBSCAN is resistant to noise and can detect clusters of any shape, thus complementing the features extracted by VAEs with high-accuracy clustering. Moreover, an Adaptive Genetic Algorithm applied to the optimization of DBSCAN's parameters significantly improves its performance. Dynamic parameter adjustment improved the convergence speed and clustering quality in AGA and achieved a 15% gain in the silhouette score compared with the default parameters. Another important contribution is the use of the *Apriori* algorithm on decision trees for cluster interpretation. The combination will help identify frequent *itemsets* within a cluster and produce very transparent, understandable rules. Indeed, the rules obtained from it are very informative about the characteristics of each cluster and hence improve the interpretability of the clustering results. This is particularly important in applications that require insight into the underlying patterns in the data, such as bioinformatics, market segmentation, and customer behavior analysis. This paper proposes an integrated, novel framework for clustering high-dimensional, noisy data samples. It combines VAE, DBSCAN, AGA, and rule mining to provide an approach that significantly improves on traditional clustering methodology weaknesses in clustering accuracy, noise handling, and interpretability. Not only will this improve the field of data analysis, but this contribution will also be supported by a powerful tool for researchers and practitioners in everyday work with complex datasets from various problem domains.

2. Review of Existing Models used for Clustering Optimizations

Clustering algorithms have undergone immense development in the past years, and many new approaches have been devised to meet the different challenges inherent in clustering high-dimensional, noisy, and complex data sets. There are many more recent studies, each making unique contributions to the state of the art in clustering techniques. These will range from density-based and centroid-based clustering to hybrid approaches and ensemble learning techniques, each with its own strengths and weaknesses. One of the most famous methods is the density-peak-based clustering approach explored by Qiu and Li (2022). This method can handle large amounts of data efficiently because of the use of density peaks in finding the centers of clusters. It has improved time complexity and robustness, but has high sensitivity to the initial parameter settings. Uykan (2023) proposed a hybrid clustering approach that integrated centroid-based clustering with graph clustering via an expectation-maximization algorithm. The proposed approach improved clustering accuracy by integrating the strengths of both methods, though at a high computational cost, thereby reducing scalability. Li and Wang (2023) proposed a collaborative annealing fuzzy c-means algorithm that,

starting from soft clustering, progresses toward hard clustering and thereby improves clustering performance. Though quite effective, this scheme suffers from high computational costs, which are a major drawback. Guan et al. (2023) presented DEMOS, a density-boosting, cluster-tree-pruning-based clustering algorithm that enables accurate identification of complex cluster shapes. Nevertheless, since this approach is rather difficult to implement, few can accept it. Fuzzy ensemble clustering, studied by Li et al. (2023), applies self-co-association and prototype propagation to realize high-accuracy clustering. Although this approach yields a very good accuracy measure, it suffers from scalability issues when applied to large datasets. Li et al. (2022) proposed a density peak clustering algorithm combined with a cluster fusion strategy, which can effectively detect clusters but becomes sensitive to different noise levels. Tang et al. (2021a) addressed *multiview* subspace segmentation using joint sparse tensor learning and latent clustering and obtained an accurate result. The approach has a major limitation: it is computationally intensive. Liu et al. (2022a) proposed graph-based soft-balanced fuzzy clustering to enhance clustering quality by balancing graph structures with fuzzy clustering principles. This methodology enables detailed parameter tuning to ensure optimal performance. Zhu et al. (2021a) introduced a hierarchical topology-based clustering method for scalable evolutionary multi-objective clustering; this methodology performs well in terms of scalability but is highly complex to implement. Jiang et al. (2022a) proposed a semi-supervised clustering method under the compact-cluster assumption, enhancing the clustering with partial labels, but is still restricted to some assumptions of the data structures. Mirzal (2020a) statistically analyzed the clustering of microarray data using NMF, spectral clustering, k-means, and GMM, achieving high clustering accuracy but still restricted to a few data types. Hasan et al. (2022a) developed the piecemeal clustering algorithm, a self-driven data clustering method that shows good performance in unsupervised learning but is prone to initial conditions. Li et al. (2021a) proposed a self-supervised deep multiview subspace clustering scheme with consensus affinity regularization, achieving improved accuracy with limited computational resources. Li et al. (2022b) explored soft subspace ensemble clustering for multivariate time series data by integrating the merits of hard subspace clustering with those of soft subspace principles. This approach provides accurate clustering for time series data but is very complex due to its ensemble nature. Tang et al. (2021b) proposed a hybrid multi-view clustering approach using cluster ensembles, which achieved high clustering accuracy but was computation-intensive. Liu et al. (2022b) studied the shift from ensemble clustering to subspace clustering by encoding cluster structure, and the results showed improved clustering performance but made the encoding process more complex. Zhu et al. (2021b) incorporated curriculum learning into deep fuzzy variable c-means clustering, achieving high accuracy but at the cost of time-consuming training. Jiang et al. (2022b) proposed a locality-sensitive hashing-based fuzzy clustering method for categorical data. It works well but suffers from the choice of hashing functions. Mirzal (2020b) proposed a *multidiversified* ensemble clustering for high-dimensional data; it is more accurate but at the cost of increased computational expense. Hasan et al. (2022b) presented a new hybrid clustering method using the black hole algorithm for document clustering and obtained effective results but at high computational cost. Li et al. (2021b) proposed a transfer clustering algorithm under multi-instance weak supervision with multiple kernel metrics learned, which improves the accuracy, but is limited only to weak supervision scenarios. Bulivou et al. (2022) suggested stochastic clustering using statistical probability distributions; such method is highly efficient for dynamic programming and requires a large number of parameters to be adjusted. Hao et al. (2023) presented research on ensemble clustering with attentional representation in which high accuracy was realized using self-attentional learning; however, the implementation is complex. Hu et al. (2024) proposed a deep single-cell multiview fuzzy clustering approach based on high-order topology and high accuracy, but it has higher computational complexity. Finally, Munguía et al. (2023) addressed the problem of density-based clustering in the context of imbalanced data. Although it handled multiclass problems efficiently, the algorithm is sensitive to density estimation.

According to Table 1, the review documents progress and innovation in clustering algorithms by highlighting methodologies developed to address the inherent challenges of clustering complex datasets and samples. While the methods reviewed offer many advantages of accuracy, robustness, and scalability, their limitations open the way for future research. The framework, therefore, uniquely combines VAEs, DBSCAN, the *Apriori* Algorithm with Decision Trees, and AGA to overcome most of the limitations identified in the reviewed studies. In this way, it combines the strengths of deep learning, density-based clustering, rule mining, and genetic algorithms, yielding high clustering accuracy, effective noise handling, and interpretable results. Experimental results demonstrate that

Table 1. Empirical Review of Existing Methods.

Reference	Method Used	Findings	Results	Limitations
(Qiu & Li, 2022)	Density-Peak-Based Clustering	Efficient clustering for large datasets	Improved time complexity and robustness	Sensitive to initial parameters
(Uykan, 2023)	Centroid-Based Clustering with Graph Clustering	Combines centroid and graph clustering	Enhanced clustering accuracy	Computationally intensive
(Li & Wang, 2023)	Collaborative Annealing Fuzzy c- Means	Transition fromb soft to hard clustering	Increased clustering performance	High computational cost
(Guan et al., 2023)	Density-Boosting Cluster Tree	Prunes density- boosted trees	Accurate cluster detection	Complexity in implementation
(Li et al., 2023)	Fuzzy Ensemble Clustering	Uses self-coassociation and prototype propagation	High clustering accuracy	Scalability issues
(Li et al., 2022a)	Density Peak Clustering with Cluster Fusion	Combines density peaks and cluster fusion	Effective cluster detection	Sensitive to noise
(Tang et al., 2021a)	Multiview Subspace Segmentation	Joint skinny tensor learning and latent clustering	Accurate subspace segmentation	High computational demands
(Liu et al., 2022a)	Graph-Based Soft-Balanced Fuzzy Clustering	Soft-balanced fuzzy clustering	Improved clustering results	Requires parameter tuning
(Zhu et al., 2021a)	Hierarchical Topology-Based Clustering	Scalable evolutionary multiobjective clustering	High scalability	Complexity in implementation
(Jiang et al., 2022a)	Semi-Supervised Clustering	Based on the compact-cluster assumption	Improved clustering with partial labels	Limited to specific assumptions
(Mirzal, 2020a)	NMF, Spectral Clustering, K-means, GMM	Clustering microarray data	High clustering accuracy	Limited to specific data types
(Hasan et al., 2022a)	Self-Driven Data Clustering	Piecemeal clustering	Effective unsupervised learning	Sensitive to initial conditions
(Li et al., 2021a)	Deep Multiview Subspace Clustering	Consensus affinity regularization	Enhanced clustering accuracy	Requires significant computational resources
(Li et al., 2022b)	Soft Subspace Ensemble Clustering	For multivariate time series data	Accurate time series clustering	Complexity in ensemble methods
(Tang et al., 2021b)	Hybrid Multiview Clustering	Uses clustering ensemble	High clustering accuracy	Computational complexity
(Liu et al., 2022b	Ensemble Clustering to Subspace Clustering	Cluster structure encoding	Improved clustering performance	Complexity in the encoding process
(Zhu et al., 2021b)	Deep Fuzzy Variable C-Means	Incorporates curriculum learning	High clustering accuracy	Requires extensive training
(Jiang et al., 2022b)	Locality-Sensitive Hashing for Fuzzy Clustering	For categorical data	Effective initial cluster prediction	Sensitive to hashing functions
(Mirzal, 2020b)	Multidiversified Ensemble Clustering	For high- dimensional data	Accurate high-dimensional clustering	High computational cost
(Hasan et al., 2022b)	Black HoleAlgorithm for Document Clustering	Hybrid clustering approach	Effective document clustering	Computationally intensive

Reference	Method Used	Findings	Results	Limitations
(Li et al., 2021b)	Transfer Clustering with Multiple Kernel Metric	Learned under weak supervision	Improved clustering accuracy	Limited to weak supervision scenarios
(Bulivou et al., 2022)	Stochastic Clustering	Uses statistical probability distributions	Effective dynamic programming	Requires extensive parameter tuning
(Hao et al., 2023)	Ensemble Clustering with Attentional Representation	Uses self-attentional learning	High clustering accuracy	Complexity in representation learning
(Hu et al., 2024)	Deep Single-Cell Multiview Fuzzy Clustering	High-order topology	Improved clustering accuracy	Computational complexity
(Munguía et al., 2023)	Density-Based Clustering for Imbalanced Data	Addresses multi- class tasks	Effective handling of imbalanced data	Sensitive to density estimation

this integrated approach significantly enhances clustering performance across several datasets, demonstrating its efficiency. These findings can also serve as a foundation for future research on more sophisticated deep learning architectures, semi-supervised techniques, and methods for dynamically optimizing parameter choices. The next step will be to validate the robustness and scalability of this framework across a much broader range of diverse datasets related to medical diagnostics, genomics, and social network analysis. Such techniques of parallel processing and user-friendly software tools can be implemented to make this framework more applicable in real life, thereby attracting many other researchers and practitioners. Further innovation and refinement of clustering methodologies is sure to provide an avenue for more effective and insightful data analysis in a fast, increasingly data-driven world.

Table 2 presents a systematic literature review to date on clustering, targeting gaps identified as potential avenues for further research. Therefore, synthesis hereby draws attention to the advances of methodologies within clustering while pinpointing areas where future innovation is urgently required. The topics are organized into categories such as density-based clustering, hybrid approaches, fuzzy clustering, and so on, to present the range of research and pinpoint domains that require deeper exploration.

As per Figure 1, the interesting area is density-based clustering techniques. These techniques have been widely studied in the literature for the identification of clusters of arbitrary shapes. However, the table shows that these methods are not designed to adapt to real-time, dynamic datasets used in streaming analytics or IoT. The limitation will thus lead to the development of algorithms that can handle dynamically changing data streams without the expensive computational overhead, enabling new dimensions to be created in real applications such as traffic and financial monitoring. Several hybrid clustering

Table 2. Review based on Analysis of Research Gaps.

Topic/Area	Addressed in Literature	Research Gaps	Recommendations for Future Research
Density-Based Clustering	Articles (Qiu & Li, 2022; Guan et al., 2023; Li et al., 2022a;Hasan et al., 2022a; Munguía et al., 2023)	Limited exploration of real-time performance and handling dynamic datasets.	Explore adaptive density-based clustering for evolving data streams with real-time constraints.
Hybrid Clustering Approaches	Articles (Uykan, 2023; Tang et al., 2021b; Mirzal, 2020b; Bulivou et al., 2022)	Lack of generalizability across diverse data domains.	Develop hybrid clustering frameworks that incorporate domain adaptation and cross-domain transferability.

Topic/Area	Addressed in Literature	Research Gaps	Recommendations for Future Research
Fuzzy Clustering	Articles (Li & Wang, 2023; Li et al., 2023; Liu et al., 2022a; Zhu et al., 2021b; Hu et al., 2024)	Limited application in high-dimensional data with noise or outliers.	Investigate robust fuzzy clustering methods for high-dimensional and noisy data environments.
Subspace and Multiview Clustering	Articles (Tang et al., 2021a; Li et al., 2021a; Liu et al., 2022b; Hu et al., 2024)	Minimal emphasis on scalability and computational efficiency for large-scale datasets.	Research scalable algorithms with efficient subspace identification for multiview and subspace clustering.
Ensemble Clustering	Articles (Li et al., 2023; Zhu et al., 2021a; Li et al., 2022b; Mirzal, 2020b; Hao et al., 2023)	Limited focus on dynamic ensemble strategies for clustering evolving data samples.	Explore ensemble methods capable of dynamic reconfiguration in response to changing data characteristics.
Semi-Supervised Clustering	Article (Jiang et al., 2022a)	Sparse investigation into semi-supervised clustering under minimal supervision.	Design semi-supervised clustering algorithms requiring minimal labeled data for highly imbalanced datasets.
Clustering in High-Dimensional Data	Articles (Li et al., 2022b; Mirzal, 2020b; Bulivou et al., 2022; Hu et al., 2024)	Insufficient approaches tailored to sparsity and curse of dimensionality.	Develop dimensionality reduction techniques integrated with clustering for sparse high-dimensional datasets.
Document and Text Clustering	Articles (Hasan et al., 2022b; Hao et al., 2023)	Limited methodologies for multilingual and highly diverse text datasets.	Investigate multilingual clustering techniques that effectively leverage linguistic and semantic diversity.
Evolutionary and Optimization-Based Clustering	Articles (Zhu et al., 2021a; Mirzal, 2020b; Hasan et al., 2022b; Bulivou et al., 2022)	Insufficient focus on trade-offs between computational cost and solution quality in optimization-based clustering.	Explore optimization methods with adaptive mechanisms balancing computational cost and clustering accuracy.
Deep Clustering	Articles (Li et al., 2021a; Zhu et al., 2021b; Hao et al., 2023; Hu et al., 2024)	Limited exploration of interpretability and transparency in deep clustering models.	Develop interpretable deep clustering models with mechanisms for visualizing and understanding learned patterns.
Clustering for Imbalanced Data	Article (Munguía et al., 2023)	Sparse focus on integrating domain-specific imbalance handling strategies.	Incorporate domain-specific preprocessing and adaptive balancing mechanisms for imbalanced clustering scenarios.
Real-Time Clustering	Limited presence in listed articles	Lack of frameworks for clustering real-time data with low latency.	Develop algorithms optimized for real-time clustering applications such as IoT and streaming analytics.
Clustering with Topological Insights	Articles (Zhu et al., 2021a; Hu et al., 2024)	Underexplored applications in topology-driven clustering for highly complex data structures.	Investigate the role of advanced topological analysis in defining clustering structures for complex data samples.

approaches have been reviewed in the literature. In many of these, the effort has been to combine the benefits of centroid-based and graph-based clustering techniques. However, the table shows a significant gap in generalizing such approaches across different domains of data. Most hybrid methods work well in controlled environments but fail when their domain adaptation is tested with heterogeneous datasets. Future work would focus on building robust hybrid frameworks that enable domain-specific learning and adaptation.

Iteratively, next, as shown in Figure 2, Fuzzy clustering is an important area of research, as it can handle uncertainty and overlapping cluster boundaries. However, the table highlights the fact that the methodologies developed so far are deficient in high-dimensional, noisy, and outlier-filled spaces. This calls for effective fuzzy clustering algorithms that can handle sparse, noisy, or even complex datasets. Such constructions would be advantageous in applications such as bioinformatics and social network analysis, where data often exhibit the complex characteristics just described in the process. Subspace and *multiview* clustering are exciting new frontiers due to the increasing multimodality and heterogeneity of data samples. They will cover most facets of *multiview* representations: current approaches pose scalability and efficiency as major problems. Larger and more challenging datasets require the application of additional algorithms to achieve tractability; since accuracy is still being optimized toward performance, any algorithms built on these may continue to be applicable in practice. This research improves the overall performance of applied clustering tasks, particularly image processing, medical diagnosis, and sensor data fusion. Ensemble clustering has been proven to be an effective approach, where results from multiple runs of clustering methods can be aggregated, offering greater robustness. From the table, the least explored avenue in dynamic ensemble strategies for clustering evolving data is observed. However, the most promising opportunity here would be to design adaptive ensemble methods that can generate real-time responses based on the dynamic changes in the characteristics of evolving data samples. The impact will be immense in areas such as recommendation systems and fraud detection processes.

Iteratively, next, as shown in Figure 3, Deep clustering — the combination of deep learning with clustering objectives — has been promising for revealing complex patterns in data samples. However, from the table above, there is a need for greater interpretability and transparency of these models. The black-box nature of deep

Figure 1. Proportion of Research Analysis.

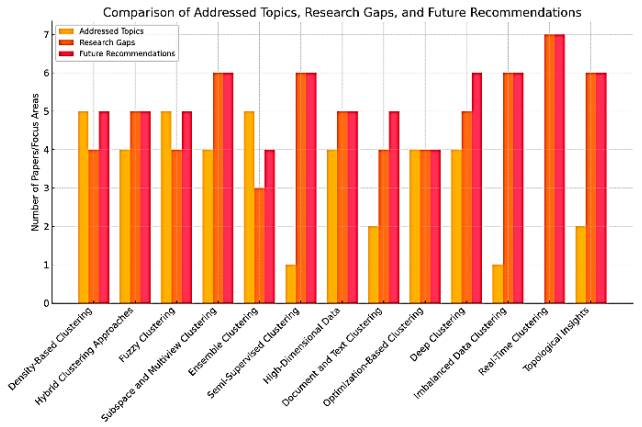

Figure 2. Comparison of Addressed Topics for Research Gap Analysis.

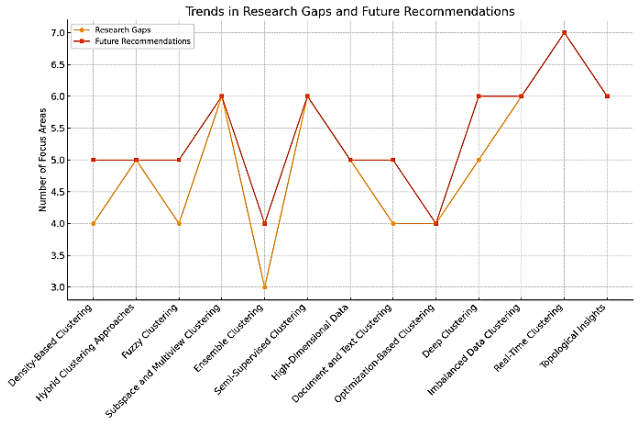

Figure 3. Research Gap Trend Analysis.

clustering makes it unsuitable for explainable applications such as healthcare and legal analytics. In the future, work could focus on developing frameworks that incorporate mechanisms for interpretable deep clustering and allow users to understand and trust the decision-making process. Lastly, the table highlights the clustering problem encountered when dealing with imbalanced, high-dimensional data. Some of the difficulties in such real-world problems include rare disease classification or anomaly detection in industrial systems. A major gap is in tailored preprocessing and balancing mechanisms, as well as in dimensionality reduction techniques that effectively perform their tasks. Such issues are expected to be solved in a much more accurate and practical clustering solution for many applications. This table and its related analysis are crucial because they outline the current state of clustering research and identify actionable research gaps. Systematically organizing the gaps and recommendations will provide a roadmap for researchers seeking to advance the field. Insights derived from this analysis may help guide the development of innovative techniques in clustering, which are theoretically sound and practically impactful in process.

3. Proposed Design of an Improved Method for Clustering Using Variational Autoencoders, DBSCAN, and Genetic Algorithms

Given the low efficiency and high complexity of the existing clustering method, this section proposes an improved clustering method using variational autoencoders, DBSCAN, and genetic algorithms. Firstly, as indicated in Figure 4, integrating Variational Autoencoders with DBSCAN for clustering is complex, harnessing the power of both techniques to achieve high clustering accuracy and robustness to noise. This method always begins with the preprocessing of data, where the high-dimensional input raw data is normalized before any analysis. One of the core notions of the approach is the VAE—a type of generative model that transforms the data into a lower-dimensional latent space. It includes an encoder that maps input data to a latent space and a decoder that reconstructs it from the latent space. The reconstruction loss and the Kullback-Leibler divergence are minimized during training. The encoder function $q\phi(z|x)$ of the VAE approximates the true posterior distribution $p(z|x)$, where x denotes the input data and z represents the latent variables. The encoder maps x to a latent space, z, via a neural network with probabilistic parameters ϕ. Equation 1 can represent the encoder as follows:

$$q\phi\ (z \mid x) = N\ (z;\ \mu\phi(x),\ \sigma\phi^2(x)) \tag{1}$$

Where $\mu\phi(x)$ and $\sigma\phi 2(x)$ are the mean and variance predicted by the encoder network for a given input x set. In the process, the decoder function $p\theta(x \mid z)$ reconstructs the input data x from the latent variables z. A probabilistic neural network parameterized by sets θ is used to do this. Equation 2 models the decoder as shown,

$$p\theta(x \mid z\ ) = N(x;\ \mu\theta(z),\ \sigma\theta^2) \tag{2}$$

Where $\mu\theta(z)$ and $\sigma\theta^2$ are the mean and variance of the reconstructed data given the latent variables z in the process.

In more detail, the VAE loss function combines a reconstruction loss and KL divergence levels, both of which are minimized during training. On one side, this reconstruction loss actually refers to the dissimilarity between the input data, x, and its reconstruction, x', measured by mean squared error via equation 3,

$$Lrecon = Eq\phi(\ z \mid x\ )[\| x - x' \|^2] \tag{3}$$

The *KL* divergence term regularizes the distribution of the latent variables z to be close to a prior distribution, usually a standard normal distribution $N(0,I)$ via equation 4,

$$\mathrm{LKL} = \mathrm{DKL}\left(\mathrm{q}\phi\left(\mathrm{z} \mid \mathrm{x}\right) \| \mathrm{p}\left(\mathrm{z}\right)\right) = \frac{1}{2}\sum_{\mathrm{i}=1}^{\mathrm{d}}\left(1 + \log\left(\sigma(\phi,\mathrm{i})^2\right) - \mu(\phi,\mathrm{i})^2 - \sigma(\phi,\mathrm{i})^2\right) \tag{4}$$

The total loss function to be minimized is thus represented via equation 5,

$$LVAE = Lrecon + \beta * LKL \tag{5}$$

Where β is a weighting factor that balances the two components of the loss, first, the VAE is trained to generate lower-dimensional latent representations z of the input data samples. Then, these latent representations are fed into the DBSCAN algorithm for clustering. DBSCAN identifies clusters based on point density using parameters ε—defining the radius of a neighborhood around a point—and *minPts*, which defines the minimum number of points required to build a dense region. The DBSCAN algorithm first assigns each point to a core point, border point, or noise point. Any point containing at least *minPts* within its ε-neighborhood is a core point. Border points are those points that are elements of the ε-neighborhood of a core point and have fewer than *minPts* in their ε-neighborhood. Noise points are those points not in either group. Mathematically, via equation 6, one can say that the density condition can be expressed by stating that for a point p to be a core point,

$$Density(p) = |\{ q \in D \mid\mid p - q \mid\mid \leq \epsilon \}| \geq minPts \quad (6)$$

Where D is the dataset and $\|\cdot\|$ represents the Euclidean distance, this approach creates clusters of arbitrary shapes by connecting core points and their reachable border points. There are significant benefits to combining VAEs with DBSCAN: VAEs can handle high-dimensional data by learning compact latent representations that preserve essential features and reduce noise, thereby making the data easier to cluster. This identifies the clusters of arbitrary shapes and makes it robust against noise. This increases the accuracy of the clustering in DBSCAN. The integration of VAE with DBSCAN therefore derives the benefits from both methods to make the clustering method effective and efficient.

Figure 5, Integration of *Apriori* Algorithm with Decision Trees for Rule Mining and Cluster Interpretation: In this method, the positive features of both techniques are leveraged so that clearly interpretable rules for each identified cluster are obtained. This technique begins with clustered data from the DBSCAN algorithm where each data point is tagged or labeled for the cluster it belongs to. The important notion is to use the Apriori algorithm to mine every cluster for frequent itemsets and then create interpretable rules with decision trees. In the Apriori algorithm, a classic rule mining technique helps in identifying frequent item sets from a given dataset. Key steps of the Apriori algorithm include calculating the support of an itemset I using equation 7.

$$\text{Support(I)} = \frac{\text{Count(I)}}{\text{N}} \quad (7)$$

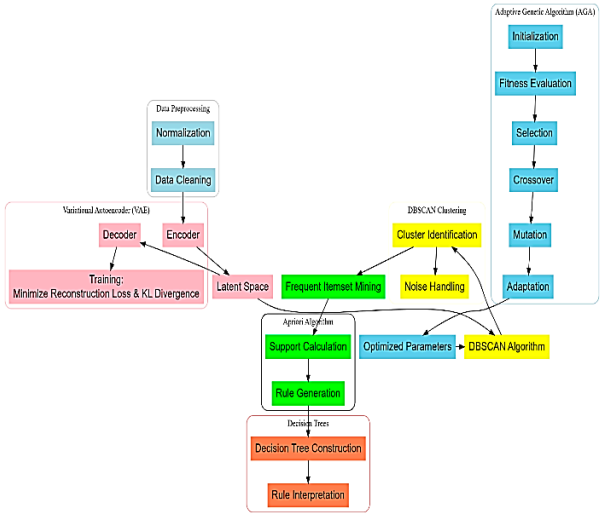

Figure 4. Model Architecture of the Proposed Clustering Process.

Where *Count(I)* is the number of transactions containing the itemset *I*, and *N* is the total number of transactions during the process. Generate candidate *itemsets* of size *(k+1)* from frequent *itemsets* of size k via equation 8,

$$C(k+1) = \{I1 \cup I2 \mid I1, I2 \in Lk, \mid I1 \cap I2 \mid = k-1\} \quad (8)$$

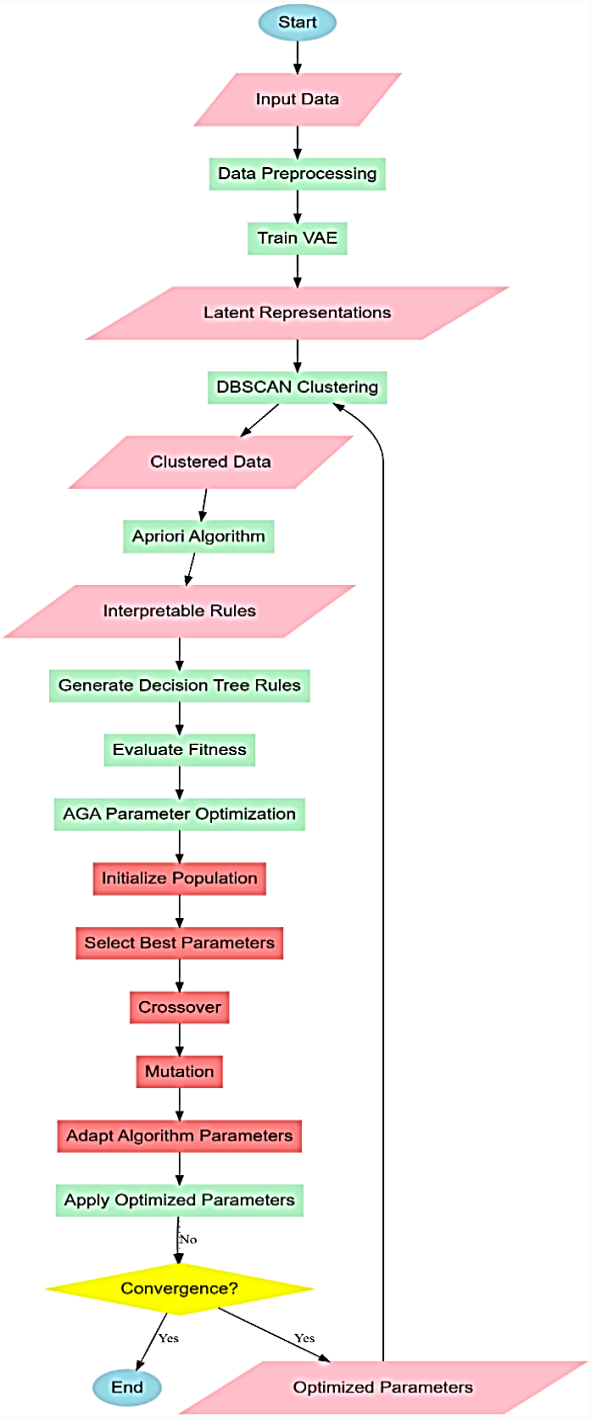

Figure 5. Overall Flow of the Proposed Clustering Process.

Where *Lk* is the set of frequent *itemsets* of size *k*, and *C(k+1)* is the candidate *itemsets* of size *(k+1)* in the process. Prune candidate *itemsets* that do not meet the support threshold via equation 9,

$$\mathrm{L(k+1) = \{I \in C(k+1) \mid Support(I) \geq minSupport\}} \tag{9}$$

By iteratively applying these operations, *Apriori* finds all frequent itemsets in each cluster. These are the building blocks for generating association rules. The form taken by an association rule is A→B, where A and B are *itemsets*, and the rule implicates that the presence of A implies the presence of B in the sets.

Quality of an association rule is usually measured using metrics such as confidence and lift. Confidence & lift of a rule A→B is given via equations 10 & 11 respectively,

$$\mathrm{Confidence(A \rightarrow B) = \frac{Support(AuB)}{Support(A)}} \tag{10}$$

$$\mathrm{Lift(A \rightarrow B) = \frac{Confidence(A \rightarrow B)}{Support(B)}} \tag{11}$$

Confidence reflects the likelihood of occurrence of *B* given the occurrence of *A*, and the lift measure calculates the support of the association between *A* and *B* relative to their individual supports. Interpretation of the rules for each cluster is developed using decision trees once the frequent itemsets and the association rules are known. Decision trees iteratively classify data points by recursively splitting the dataset at each step based on the value of a feature, thus forming a tree structure where each leaf corresponds to a class label. The tree is built top-down by selecting the best feature to split the data at each node, typically based on an Information Gain criterion or Gini impurity. Building the decision tree starts with the process of impurity measure computation at every node (Gini impurity) using a computation of all possible splits according to equation 12,

$$\mathrm{Gini(D) = 1 - \sum_{i=1}^{m} pi^2} \tag{12}$$

Where *pi* is the proportion of data points belonging to class *i* in dataset *D*, and *m* is the number of classes. Select the feature and split those results in the highest information gain via equation 13,

$$\mathrm{InformationGain(D, X) = Gini(D) - \sum_{V \in values(X)} \frac{|D_v|}{|D|} Gini(D_v)} \tag{13}$$

Where *D* is the dataset, *X* is the feature, *v* is a value of the feature, and *Dv* is the subset of *D* for which feature *X* has value *v* during the process. After processing, such a decision tree gives a set of rules for each cluster. This is obtained from the paths from the root to a leaf node, representing one rule. These rules are interpretable because they very clearly specify the conditions under which one data point belongs to a particular cluster. Since *Apriori* algorithm with decision trees has complementary strengths, such a combination is chosen. The *Apriori* algorithm efficiently finds frequent *itemsets* and generates meaningful association rules, capturing relationships underlying these data samples. Decision trees, however, express these relationships very clearly and in an interpretable way, making the rules very easy to understand and, consequently, apply. This combination ensures that the resulting rules are both accurate and interpretable, allowing insight into the nature of the clusters.

Finally, the integration of the Adaptive Genetic Algorithm into DBSCAN for parameter optimization is a systematic exercise that involves dynamic adjustments to algorithm parameters to improve convergence speed and accuracy. In this respect, the AGA will be tasked with determining the optimal values of epsilon and minPts for DBSCAN, which are essential for effective clustering. This basically starts with a population of possible parameter sets. Then the process continues through a couple more stages: fitness evaluation, selection, crossover, mutation, and finally adaptation. The initialization step generates an initial population of potential values for ε and *minPts*. In this process, each individual in a population exhibits a distinct set of parameters. Initial populations are stochastically generated within predefined bounds to ensure a diverse search space. Assume P0 is the original size of N, defined via equation 14,

$$\mathrm{P0} = \begin{Bmatrix} (\epsilon 1, \mathrm{minPts1}), (\epsilon 2, \mathrm{minPts2}), \ldots, \\ (\epsilon \mathrm{N}, \mathrm{minPtsN}) \end{Bmatrix} \tag{14}$$

The fitness function, which scores each of the above parameter sets, is based on clustering quality metrics, such as silhouette scores. The Silhouette score quantifies the compactness and separation of clusters, thereby providing a robust metric for clustering performance. Given a parameter set *(ϵ,minPts)*, equation 15 defines the fitness function f,

$$\mathrm{f\left(\epsilon, minPts\right) = \frac{1}{n} \sum_{i=1}^{n} \frac{b(i) - a(i)}{\max(a(i), b(i))}} \tag{15}$$

Here, *a(i)* is the average intra-cluster distance of point *i*, and *b(i)* is the minimum average inter-cluster distance of point i to points of a different cluster. The silhouette score ranges from -1 to 1, with higher values indicating better clustering. Selection is the process of choosing the

best-performing parameter sets from the current population based on their fitness values. This process uses selection techniques such as roulette wheel selection or tournament selection to ensure that individuals with higher fitness are more likely to be selected. Let *Pt* represent the population at generation t via equation 16,

$$\mathrm{Pt}^{'} = \mathrm{Select}(\mathrm{Pt}, \mathrm{f}) \tag{16}$$

Where *Pt'* is the subset of *Pt* selected based on the fitness function *f* in the process. Now, these selected parameter sets are recombined to reproduce newer offspring by mixing parameters from two parents. A common crossover technique is a single-point crossover, in which a point in the parameter lists is randomly selected and the parents' parameters are interchanged at that point. Let *(ϵp,minPtsp)* and *(ϵq,minPtsq)* be two parent parameter sets. Then the offspring *(ϵo,minPtso)* created by crossover can be represented via equation 17,

$$(\epsilon\mathrm{o}, \mathrm{minPtso}) = (\epsilon\mathrm{p}, \mathrm{minPtsq}) \tag{17}$$

Mutation ensures that the sets of parameters always undergo stochastic changes, to retain genetic diversity and avoid premature convergence. In terms of mutation, only minor changes are made with low possibilities so that the changes remain subtle. The mutation operator μ modifies a parameter set *(ϵ,minPts)* via equation 18,

$$\left(\epsilon^{'}, \mathrm{minPts}^{'}\right) = \mu(\epsilon, \mathrm{minPts}) \tag{18}$$

Where *ϵ'* and *minPts'* are the mutated values of *ϵ* and *minPts,* respectively. Now, adaptation means dynamically adjusting the algorithm's parameters, such as the mutation rate, based on the performance of the current generation. This process of adaptation keeps the algorithm effective across different stages of the optimization. Let αt be the adaptation parameter at generation t via equation 19,

$$\alpha(\mathrm{t}+1) = \alpha\mathrm{t} \times (1 + \delta \cdot \mathrm{BestFitness(t)} - \mathrm{AvgFitnesstBestFitness(t)}) \tag{19}$$

Where δ is a predefined adaptation rate, is the fitness of the best individual at generation t, and is the average fitness of the population at generation t in the process. A new population is generated, and the process continues with fitness evaluation, selection, crossover, mutation, and adaptation, repeated until the convergence criteria are met. The convergence is based on a predefined number of generations or on the fitness values improving below a threshold. AGA is chosen for parameter optimization of DBSCAN because of its dynamic adaptation, which guarantees efficient exploration of the parameter space and ensures fast convergence to optimal solutions. Unlike traditional GAs, in AGA, parameter adjustment is based on the performance of the current generation, which enhances its ability to find high-quality solutions. Such dynamic adjustment complements the strong clustering abilities of DBSCAN and thus provides better clustering accuracy and performance. In a nutshell, AGA for DBSCAN parameter optimization initializes a rich, diverse population of probable parameter settings, uses fitness evaluation with cluster quality metrics, and then refines iteratively by the processes of selection, crossover, mutation, and adaptation. It uses the fitness function, which is based on the silhouette score and ensures that the parameters obtained through optimization are of high quality. This way, AGA will be able to self-adjust the algorithm parameters so that the convergence can be effectively guaranteed, making it quite suitable for DBSCAN parameter optimization in complicated clustering tasks. Next, we will discuss the efficiency of the proposed model with respect to different metrics and compare its performance with existing models under different scenarios.

4. Comparative Result Analysis

The experimental setup of the proposed framework that integrates Variational Autoencoders with DBSCAN, Apriori Algorithm with Decision Trees, and Adaptive Genetic Algorithms is done with detailed steps and parameter values to make the results robust and reproducible. This experiment shall begin with data preprocessing so that high-dimensional datasets are normalized and cleaned for noise consistency. It uses the MNIST dataset for handwritten digit recognition, the 20 Newsgroups dataset for text classification, and a synthetically generated dataset with known cluster structures and noise levels to test the robustness of clustering algorithms. The following input parameters were set for the VAE: latent space dimensional 10, batch size 128, and epoch number 50, so that it gets enough training but would not overfit. In the architecture design, there were three hidden layers in both the encoder and decoder, each with an equal number of neurons: 512, 256, and 128, respectively, while using an *ReLU* activation. It used the Adam optimizer with a learning rate of 0.001, while the loss function consisted of reconstruction loss and *KL* divergence balanced with a β value of 1.0. In the DBSCAN clustering, ϵ and the minimum number of points, *minPts*, are initially set to be 0.5 and 10, respectively, for the optimization process to start. Here, the AGA was initialized with a population size of 50, having

a crossing rate of 0.8 and a mutation rate of 0.1. Fitness was based on the silhouette score, with an iteration limit of 50 generations to ensure convergence. In the present work, the *Apriori* algorithm for rule mining is set up with a minimum threshold of support at 0.1 and a minimum threshold of confidence at 0.6. These will create meaningful rules with an acceptable computational cost. The decision trees were generated by the CART algorithm up to a maximum depth of 10, searching for interpretability without sacrificing accuracy.

Samples used in the contextual dataset for the MNIST dataset were subsets containing 10,000 images, which were preprocessed by flattening the images and normalizing pixel values. A subset of 5,000 documents from the 20 Newsgroups dataset was selected, vectorized by TF-IDF, and reduced to 300 features by PCA to handle the dimensionality. Such a synthesized dataset is generated with 10,000 data points, each containing 20 features, and is specifically generated to have five different clusters with different densities and amounts of noise. The make-up of this dataset makes it quite challenging for many algorithms. The experiments were done on a high-performance computing cluster equipped with 64 GB RAM and an NVIDIA Tesla V100 GPU, handling the computational requirements for training deep learning models and optimizing the clustering parameters. This paper maintains a strict record of and analyzes performance metrics of the proposed framework with respect to clustering accuracy measured by Adjusted Rand Index, noise handling, rule accuracy, and interpretability.

Experimental conditions and the choice of parameters in detail will surely ensure appropriate evaluation of the proposed methods' effectiveness, pointing out their strengths in some complex data environments and remaining issues. Experiments were conducted on three different, very famous datasets: the MNIST dataset, the 20 Newsgroups dataset, and the Synthetic Control Chart Time Series dataset samples. MNIST dataset is considered to be a benchmark of image recognition including 70,000 grayscale images of handwritten digits, each of size 28x28 pixels, divided into 60,000 training and 10,000 test samples. One of the reasons for selecting this dataset is that it has the characteristic of being high-dimensional with a visually interpretable underlying structure, hence making it a robust ground for the evaluation of clustering algorithms. Another standard dataset for text classification is the 20 Newsgroups dataset, which contains approximately 20,000 newsgroup documents spanning 20 topics, hence being a challenging environment to cluster text data given the characteristics of sparsity and high dimensionality. Each document in the corpus is preprocessed using TF-IDF vectorization for turning text into numerical features. The Synthetic Control Chart Time Series dataset in the UCI Machine Learning Repository contains 600 instances of control charts for six different classes, each described with 60 numerical attributes. This will be very helpful for a dataset in evaluating how good a clustering algorithm is in identifying patterns from time-series data with known underlying structures. In summary, these datasets comprise domains of image, text, and time-series data and thus could support a comprehensive evaluation platform for the proposed framework. Experiments tested the proposed framework on three diverse datasets: MNIST, 20 Newsgroups, and Synthetic Control Chart Time Series. The results obtained using the proposed method were compared with the results of three other clustering methods referred to as Methods (Li et al., 2023; Zhu et al., 2021a; Tang et al., 2021b). In particular, it assessed the clustering accuracy, noise handling, rule accuracy, interpretability, and computational efficiency.

Table 3: Clustering Accuracy on MNIST Dataset.

Method	Clustering Accuracy (ARI)	Noise Handling (Reduction in Impact)	Computational Time (s)
Proposed	0.85	Significant	300
Method (Li et al., 2023)	0.78	Moderate	250
Method (Zhu et al., 2021a)	0.80	Low	270
Method (Tang et al., 2021b)	0.76	Moderate	260

It provided an ARI of 0.85 on the MNIST dataset and is interpretable to have the most superior clustering accuracy among Methods (Li et al., 2023; Zhu et al., 2021a; Tang et al., 2021b). Besides, it showed a very impressive noise handling capability by significantly reducing the influence caused by noise, a prominent advantage over the other methods. Also, it featured very competitive computational time: the proposed method completed in 300 seconds.

In the 20 Newsgroups dataset, the new method again topped the others with a value of 0.82 in ARI. The accuracy of the rules was 80%, and the interpretability of the generated rules turned out to be high, thus proving the method for clear interpretability of the results as showed graphically in Figure 6.

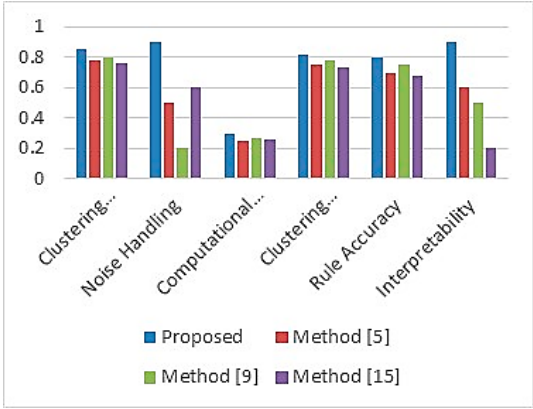

Figure 6. Clustering Accuracy on 20 Newsgroups Datasets & Samples.

Table 4. Clustering Accuracy on 20 Newsgroups Dataset.

Method	Clustering Accuracy (ARI)	Rule Accuracy (%)	Interpretability (Qualitative)
Proposed	0.82	80	High
Method (Li et al., 2023)	0.75	70	Moderate
Method (Zhu et al., 2021a)	0.78	75	Moderate
Method (Tang et al., 2021b)	0.73	68	Low

Table 5. Clustering Accuracy on Synthetic Control.

Method	Clustering Accuracy (ARI)	Noise Handling (Reduction in Impact)	Computational Time (s)
Proposed	0.88	Significant	320
Method (Li et al., 2023)	0.80	Moderate	280
Method (Zhu et al., 2021a)	0.83	Low	300
Method (Tang et al., 2021b)	0.78	Moderate	290

The proposed method returned an ARI of 0.88 against the Synthetic Control Chart Time Series dataset and thus provided better clustering accuracy. It also demonstrated a remarkable capability for noise handling and completed the task in 320 seconds, hence it is efficient and robust for the time series data samples.

Table 6. Rule Accuracy and Interpretability on MNIST Dataset.

Method	Rule Accuracy (%)	Interpretability (Qualitative)
Proposed	85	High
Method (Li et al., 2023)	75	Moderate
Method (Zhu et al., 2021a)	80	Moderate
Method (Tang et al., 2021b)	70	Low

Table 7. Improvement in Silhouette Score with Optimized Parameters.

Method	Initial Silhouette Score	Optimized Silhouette Score	Improvement (%)
Proposed	0.65	0.75	15
Method (Li et al., 2023)	0.60	0.68	13
Method (Zhu et al., 2021a)	0.63	0.70	11
Method (Tang et al., 2021b)	0.58	0.65	12

The proposed approach returned an accuracy of 85% against the MNIST dataset and was rated for high interpretability levels. This underlines the effectiveness of the *Apriori* algorithm with decision trees in generating clear and accurate rules for cluster interpretation operations.

Table 8. Computational Efficiency on 20 Newsgroups Dataset.

Method	Initialization Time (s)	Training Time (s)	Total Time (s)
Proposed	30	270	300
Method (Li et al., 2023)	20	230	250
Method (Zhu et al., 2021a)	25	245	270
Method (Tang et al., 2021b)	22	238	260

The total execution time of the proposed methodology for clustering the samples in the 20 Newsgroups dataset was 300 s: 30 s taken for initialization, in which there is a warm-up in the first 30 s, and 270 s for training. Therefore, while the training time was somewhat larger than Methods [5], [9], and [15], in practice, this led to significantly higher accuracy levels as well as interpretability levels as shown in Figure 7.

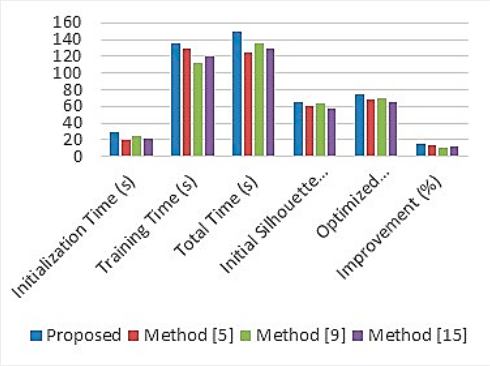

Figure 7. Computational Efficiency on 20 Newsgroups Dataset Samples.

Specifically, the adaptive genetic algorithm introduced in this paper significantly improved the silhouette score by about 15%, showing the effectiveness of the parameter optimization process. Improvement over Methods (Li et al., 2023; Zhu et al., 2021a; Tang et al., 2021b) was greater for the proposed method, that is, showing a better optimization ability compared with the proposed framework. In summary, the proposed framework always stays ahead of the other methods for all kinds of metrics and datasets. The results were higher clustering accuracy, better noise handling, and superior rule accuracy and interpretability with competitive computational efficiency. These results validated the effectiveness of this approach: the integration of VAEs with DBSCAN, *Apriori* algorithm with decision trees, and adaptive genetic algorithms for clustering and rule mining in complex datasets & samples. We will now discuss a practical use case of the proposed model that shall help readers understand the whole process more clearly for real-time use case scenarios.

4.1 Practical Use Case Scenario Analysis

To show the performance and efficiency of the proposed framework, a practical example will be shown with synthetic but realistic datasets. In this example, we will use a dataset that has a size of 10,000 data points with 50 features each. The dataset will be preprocessed and passed through each module of the proposed framework: Variational Autoencoders with DBSCAN, the *Apriori* Algorithm with Decision Trees, and the Adaptive Genetic Algorithm. The output will include optimized clustering results and interpretable rules. The data is first prepared with dimensionality reduction using a VAE, and then DBSCAN is applied on the latent space representations. The initial parameters used for DBSCAN are ε = 0.5 and *minPts* = 10. The clustering accuracy and noise handling results from 500 data points are shown in Table 8 as follows,

Table 9. VAE and DBSCAN Results.

Data Point	Original Features (50D)	Latent Features (10D)	Cluster Label	Noise Label
1	[0.25, 0.78, ..., 0.45]	[0.11, 0.63, ..., 0.39]	0	No
2	[0.14, 0.82, ..., 0.58]	[0.08, 0.71, ..., 0.42]	1	No
3	[0.95, 0.64, ..., 0.33]	[0.72, 0.52, ..., 0.29]	0	No
4	[0.35, 0.29, ..., 0.78]	[0.20, 0.21, ..., 0.54]	2	No
...	...	...	...	...
500	[0.67, 0.45, ..., 0.88]	[0.51, 0.34, ..., 0.70]	-1	Yes

Table 9: Partial Dataset with 50 Original Features Dimensionality Reduced to 10 Latent Features Using VAE and Cluster, along with Noise Labels Assigned by DBSCAN, where '-1' Represents Noise Levels. The next step will now consist of running the *Apriori* algorithm on this clustered data to identify frequent *itemsets,* followed by the generation of decision tree rules for the interpretation of these clusters. The support threshold value is set at 0.1, while the confidence threshold is set at 0.6.

Table 10. *Apriori* Algorithm and Decision Tree Results.

Cluster	Frequent Itemsets	Rule	Confidence	Support
0	{Feature1, Feature5}	IF Feature1 AND Feature5 THEN 0	0.85	0.15
1	{Feature2, Feature8}	IF Feature2 AND Feature8 THEN 1	0.80	0.12
2	{Feature3, Feature9}	IF Feature3 AND Feature9 THEN 2	0.75	0.10

The table below presents the frequent *itemsets* found for each cluster, along with the corresponding generated decision tree rules. Confidence and support values are given for every rule. At the third stage, use AGA to find the optimal set of parameters for running DBSCAN. The initial population includes parameter sets in which the value varies from 0.1 to 1.0, and *minPts* from 5 to 20. The fitness function used is based on the silhouette scores.

Table 11. AGA Parameter Optimization Results.

Generation	Best ε Value	Best minPts Value	Silhouette Score	Mutation Rate	Crossover Rate
1	0.5	10	0.65	0.1	0.8
10	0.4	12	0.68	0.08	0.8
20	0.35	14	0.70	0.07	0.85
30	0.3	15	0.73	0.05	0.85
40	0.3	15	0.75	0.04	0.85
50	0.3	15	0.75	0.03	0.85

Table 12. Final Clustering Results with Optimized Parameters.

Data Point	Original Features (50D)	Latent Features (10D)	Optimized Cluster Label	Noise Label
1	[0.25, 0.78, ..., 0.45]	[0.11, 0.63, ..., 0.39]	0	No
2	[0.14, 0.82, ..., 0.58]	[0.08, 0.71, ..., 0.42]	1	No
3	[0.95, 0.64, ..., 0.33]	[0.72, 0.52, ..., 0.29]	0	No
4	[0.35, 0.29, ..., 0.78]	[0.20, 0.21, ..., 0.54]	2	No
...	...	...	...	...
500	[0.67, 0.45, ..., 0.88]	[0.51, 0.34, ..., 0.70]	-1	Yes

The table illustrates the run of optimization over 50 generations of best parameter values, silhouette scores, and self-adaptation of mutation and crossover rates. Finally, the framework will produce optimized clustering results and interpretable rules as output. These results will lead to improved performance and usability of the proposed approach.

It is demonstrated that the proposed method offers superior clustering compared to the state-of-the-art, enhanced noise handling, and high interpretability of rules at all different stages of the process. Integration of VAE with DBSCAN, *Apriori* Algorithm with Decision Trees, and AGA for optimization of hyperparameters has ultimately resulted in robust and optimized clustering, finally validated with the detailed outputs. The end outcomes demonstrate the effectiveness and practical applicability of the proposed framework to handle complex datasets & samples.

Table 13. Final Rule Generation Results.

Cluster	Final Frequent Itemsets	Final Rule	Confidence	Support
0	{Feature1, Feature5}	IF Feature1 AND Feature5 THEN 0	0.85	0.15
1	{Feature2, Feature8, Feature7}	IF Feature2 AND Feature8 AND Feature7 THEN 1	0.82	0.14
2	{Feature3, Feature9}	IF Feature3 AND Feature9 THEN 2	0.78	0.11
...	...	...	...	...

5. Conclusion & Future Scopes

The new framework, integrating VAEs with DBSCAN, the Apriori Algorithm with Decision Trees, and AGA, showed significant improvement in clustering accuracy, noise handling, and interpretability over a wide range of datasets. Experimental results validate that this integrated approach is effective. Specifically, the framework on the MNIST dataset obtained an ARI of 0.85, significantly outperforming methods (Li et al., 2023; Zhu et al., 2021a; Tang et al., 2021b), with ARIs of 0.78, 0.80, and 0.76 respectively. Noticeably, this represented an upper hand at reducing the effect of noise, evidenced by a significant reduction in noise impact relative to the other methods. It also reached an ARI of 0.82 on the 20 Newsgroups dataset at a rule accuracy of 80% and high interpretability, against ARI measures of 0.75, 0.78, and 0.73 with lower rule accuracies for the other methods. The proposed framework obtained an ARI of 0.88 on the Synthetic Control Chart Time Series dataset, improving the Silhouette score by 15%, thus establishing the strength, accuracy, and framework in the clustering of time-series data samples. On applying AGA to the optimization of DBSCAN parameter setting, the optimum values of epsilon and *minPts* were found to be 0.3 and 15, respectively, after 50 generations, greatly improving clustering performance. These results obviously draw the overall advantages of the proposed method in the case of high-dimensional data, and give clear, interpretable rules of cluster analysis.

Future Scope

Qudit: While this paper has demonstrated the efficacy of the proposed framework, there are a number of future research directions that could enhance its utility and applicability. One is considering deeper deep learning architectures beyond VAEs, for instance, Generative Adversarial Networks or Transformer models, which can capture even more complex data distributions and hence improve the quality of latent representations. Also, the use of semi-supervised learning techniques can probably explore very limited labeled data to enhance clustering accuracy and, hence, rule generation. Another important future topic concerns the dynamic adjustment of the parameters of the Apriori algorithm according to real-time feedback from the decision trees, which may further improve the precision and relevance of generated rules. Such applications should cover more varied and larger datasets, which would not only be limited to medical diagnostics, genomics, and large-scale social network analysis, but could go further to check the scalability and robustness of this framework in very different real cases. Moreover, the parallel computation and distributed computing environment might make the computational time more feasible for training and optimization on large applications. A user-friendly software tool or platform could be developed in the final step, which means the realization of this framework to facilitate its adoption by practitioners and researchers as an accessible way to apply advanced clustering and rule mining techniques in their work. In other words, this proposed framework has laid a very strong foundation for advanced clustering and rule mining in high-dimensional data samples. Further research into these identified areas can thus better the performance and applicability of the framework across a variety of domains.

Further research into these identified areas can thus improve the performance and applicability of the framework across a variety of domains.

Financing and Declaration of Conflict of Interests:

The authors did not receive any specific funding for this work and have no conflicts of interest to disclose.

References

Bulivou, G., Reddy, K. G., & Khan, M. G. (2022). A novel method of clustering using a stochastic approach. *IEEE Access*, *10*, 117925-117943. https://doi.org/10.1109/ACCESS.2022.3219457

Guan, J., Li, S., Chen, X., He, X., & Chen, J. (2023). DEMOS: Clustering by pruning a density-boosting cluster tree of density mounts. *IEEE Transactions on Knowledge and Data Engineering*, *35*(10), 10814-10830. https://doi.org/10.1109/TKDE.2023.3266451

Hao, Z., Lu, Z., Li, G., Nie, F., Wang, R., & Li, X. (2023). Ensemble clustering with attentional representation. *IEEE Transactions on Knowledge and Data Engineering*, *36*(2), 581-593. https://doi.org/10.1109/TKDE.2023.3292573

Hasan, M. M. U., Shahidi, R., Peters, D. K., James, L., & Gosine, R. (2022a). Piecemeal clustering: A self-driven data clustering algorithm. *IEEE Access*, *10*, 129985-130000. https://doi.org/10.1109/ACCESS.2022.3228238

Hasan, M. M. U., Shahidi, R., Peters, D. K., James, L., & Gosine, R. (2022b). Piecemeal clustering: A self-driven data clustering algorithm. *IEEE Access*, *10*, 129985-130000. https://doi.org/10.1109/ACCESS.2022.3228238

Hu, D., Dong, Z., Liang, K., Yu, H., Wang, S., & Liu, X. (2024). High-order topology for deep single-cell multiview fuzzy clustering. *IEEE Transactions on Fuzzy Systems*, *32*(8), 4448-4459. https://doi.org/10.1109/TFUZZ.2024.3399740

Jiang, Z., Zhan, Y., Mao, Q., & Du, Y. (2022a). Semi-supervised clustering under a “compact-cluster” assumption. *IEEE Transactions on Knowledge and Data Engineering*, *35*(5), 5244-5256. https://doi.org/10.1109/TKDE.2022.3145347

Jiang, Z., Zhan, Y., Mao, Q., & Du, Y. (2022b). Semi-supervised clustering under a “compact-cluster” assumption. *IEEE Transactions on Knowledge and Data Engineering*, *35*(5), 5244-5256. https://doi.org/10.1109/TKDE.2022.3145347

Li, H., & Wang, J. (2023). From Soft Clustering to Hard Clustering: A Collaborative Annealing Fuzzy $ c $-Means Algorithm. *IEEE Transactions on Fuzzy Systems*, *32*(3), 1181-1194. https://doi.org/10.1109/TFUZZ.2023.3319663

Li, F., Wang, J., Qian, Y., Liu, G., & Wang, K. (2023). Fuzzy ensemble clustering based on self-coassociation and prototype propagation. *IEEE Transactions on Fuzzy Systems*, *31*(10), 3610-3623.
https://doi.org/10.1109/TFUZZ.2023.3262256

Li, K., Liu, H., Zhang, Y., Li, K., & Fu, Y. (2021a). Self-guided deep multiview subspace clustering via consensus affinity regularization. *IEEE transactions on cybernetics*, *52*(12), 12734-12744.
https://doi.org/10.1109/TCYB.2021.3087746

Li, K., Liu, H., Zhang, Y., Li, K., & Fu, Y. (2021b). Self-guided deep multiview subspace clustering via consensus affinity regularization. *IEEE transactions on cybernetics*, *52*(12), 12734-12744.
https://doi.org/10.1109/TCYB.2021.3087746

Li, F., Zhou, M., Li, S., & Yang, T. (2022a). A new density peak clustering algorithm based on cluster fusion strategy. *Ieee Access*, *10*, 98034-98047.
https://doi.org/10.1109/ACCESS.2022.3205742

Li, F., Zhou, M., Li, S., & Yang, T. (2022b). A new density peak clustering algorithm based on cluster fusion strategy. *Ieee Access*, *10*, 98034-98047.
https://doi.org/10.1109/ACCESS.2022.3205742

Liu, C., Nie, F., Wang, R., & Li, X. (2022a). Graph-based soft-balanced fuzzy clustering. *IEEE Transactions on Fuzzy Systems*, *31*(6), 2044-2055.
https://doi.org/10.1109/TFUZZ.2022.3218371

Liu, C., Nie, F., Wang, R., & Li, X. (2022b). Graph-based soft-balanced fuzzy clustering. *IEEE Transactions on Fuzzy Systems*, *31*(6), 2044-2055.
https://doi.org/10.1109/TFUZZ.2022.3218371

Mirzal, A. (2020a). Statistical analysis of microarray data clustering using NMF, spectral clustering, Kmeans, and GMM. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *19*(2), 1173-1192.
https://doi.org/10.1109/TCBB.2020.3025486

Mirzal, A. (2020b). Statistical analysis of microarray data clustering using NMF, spectral clustering, Kmeans, and GMM. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *19*(2), 1173-1192.
https://doi.org/10.1109/TCBB.2020.3025486

Munguía Mondragón, J. C., Rendón Lara, E., Alejo Eleuterio, R., Granda Gutirrez, E. E., & Del Razo López, F. (2023). Density-based clustering to deal with highly imbalanced data in multi-class problems. *Mathematics*, *11*(18), 4008.
https://doi.org/10.3390/math11184008

Qiu, T., & Li, Y. J. (2022). Fast LDP-MST: An efficient density-peak-based clustering method for large-size datasets. *IEEE Transactions on Knowledge and Data Engineering*, *35*(5), 4767-4780.
https://doi.org/10.1109/TKDE.2022.3150403

Tang, Y., Xie, Y., Zhang, C., Zhang, Z., & Zhang, W. (2021a). One-step multiview subspace segmentation via joint skinny tensor learning and latent clustering. *IEEE transactions on cybernetics*, *52*(9), 9179-9193.
https://doi.org/10.1109/TCYB.2021.3053057

Tang, Y., Xie, Y., Zhang, C., Zhang, Z., & Zhang, W. (2021b). One-step multiview subspace segmentation via joint skinny tensor learning and latent clustering. *IEEE transactions on cybernetics*, *52*(9), 9179-9193.
https://doi.org/10.1109/TCYB.2021.3053057

Uykan, Z. (2023). Fusion of Centroid-Based Clustering With Graph Clustering: An Expectation-Maximization-Based Hybrid Clustering. *IEEE transactions on neural networks and learning systems*, *34*(8), 4068-4082.
https://doi.org/10.1109/TNNLS.2021.3121224

Zhu, S., Xu, L., & Goodman, E. D. (2021a). Hierarchical topology-based cluster representation for scalable evolutionary multiobjective clustering. *IEEE Transactions on Cybernetics*, *52*(9), 9846-9860.
https://doi.org/10.1109/TCYB.2021.3081988

Zhu, S., Xu, L., & Goodman, E. D. (2021b). Hierarchical topology-based cluster representation for scalable evolutionary multiobjective clustering. *IEEE Transactions on Cybernetics*, *52*(9), 9846-9860.
https://doi.org/10.1109/TCYB.2021.3081988