SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V12I12P109

Volume 12 Issue 12, 98-113, December 2025


*Original Article*

# Hadoop Mapreduce-Based Multi-Modal Data Clustering Using CGWQRLO With Spatio-Temporal Analysis Using RBKDE

Shailendra Singh Yadav[1], Kamal Sutaria[2]

[1,2]*Department of Computer Science and Engineering, Parul Institute of Engineering and Technology, Faculty of Engineering and Technology, Parul University, Waghodia, Vadodara, Gujarat, India.*

[1]*Corresponding Author : 2123004136009@paruluniversity.ac.in*

Received: 09 October 2025 Revised: 11 November 2025 Accepted: 10 December 2025 Published: 27 December 2025

***Abstract*** *- Data analysis has been significantly affected by the exponential growth of digital information on the internet in today's digital era, thus leading to the development of Data Clustering (DC) techniques. Yet, none of the prevailing DC approaches captured the temporal and spatial dynamics of the data to produce coherent and temporally/spatially consistent clusters. Therefore, the proposed framework introduces the Hadoop MapReduce-based multi-modal DC methodology using Circle Grey Wolf Quasi-Reflexive Learning Optimizer (CGWQ-RLO) with spatial-temporal analysis using Radial Basis Kernel Density Estimation (RBKDE) for effective clustering. Primarily, the data containing numeric values and images is collected. Next, regarding Missing Value Imputation (MVI) and Normalization, the numeric values are preprocessed. Similarly, regarding noise removal, contrast enhancement, and edge preservation, the images are preprocessed. Now, the dimensionality of the preprocessed numeric and image data is reduced using the Principal Fuzzy Correlated Component Analysis (PF-CCA), followed by feature extraction. Next, the features are aggregated by the computation of Weighted Average (WA). Then, to perform Hadoop MapReduce, the Round Robin Heap (RRH) is employed, and RBKDE is applied for spatio-temporal analysis. Further, to cluster the data, the CGWQ-RLO is applied. Lastly, to optimize the number of clusters, the SSA is done on the clustered data. Therefore, when analogized to the prevailing techniques, the proposed framework clustered the multi-modal data more efficiently by attaining higher accuracy (97.79%).*

***Keywords*** *- Multi-Modal Data Clustering, Spatio-Temporal Analysis, Big data, Radial Basis Kernel Density Estimation (RBKDE), Circle Grey Wolf Quasi-Reflexive Learning Optimizer (CGWQ-RLO), Round Robin Heap (RRH), Weighted Covariance K-Nearest Neighbors (WCKNN), and Principal Fuzzy Correlated Component Analysis (PF-CCA).*

## 1. Introduction

Recently, the data analysis process has been affected by the growing amount of digital information on Internet web pages and modern applications [1]. The process of collecting, processing, and interpreting data is referred to as data analysis [2, 3]. One of the fundamental processes in data analysis is DC, which groups similar data points based on specific characteristics into a predetermined number of clusters [4, 5]. Likewise, the goal of clustering is to identify hidden patterns and structures within data, thus enabling more meaningful insights and efficient decision-making [6, 7]. Furthermore, in various Machine Learning (ML) and data mining techniques like image segmentation, information retrieval, pattern recognition, classification, and network analysis, clustering plays an important role [8, 9]. Yet, it is important to enhance the DC process for multimodal data like text, images, audio, and so on due to the rapid growth of diverse digital data [10].

To integrate and analyze heterogeneous data sources, various traditional multimodal DC methodologies like K-Means, Gaussian Mixture Model (GMM), and so on have been implemented [11]. These techniques offer simplicity, scalability, and the ability to uncover hidden structures within multimodal data [12]. Also, they help to improve data organization, enhance pattern recognition, and improve decision-making across various domains [13]. Yet, owing to the unreliability of similar metrics and the curse of dimensionality, the clustering becomes ineffective and the computation time becomes longer in high-dimensional feature spaces [14, 15]. None of the existing methodologies performed effective multimodal DC due to the lack of focus on its temporal and spatial dynamics, thus leading to inaccurate clustering. Thus, the proposed work introduces a practical multimodal big DC approach with the Hadoop MapReduce approach by analyzing the spatio-temporal dynamics of the multimodal data using the CGWQRLO and RBKED.

### *1.1. Problem Statement*

The limitations of several traditional works have been depicted as follows,

- None of the existing methodologies implemented effective clustering techniques by analyzing the temporal and spatial dynamics of the data.
- The computation time of the [16] was longer owing to high-dimensional or complex optimization problems.
- Due to the ever-increasing size and complexity of the datasets, some traditional works struggled to develop scalable clustering.
- [17] had the curse of dimensionality, thus making the model struggle with high-dimensional data.
- [18] struggled to handle various types of data, like numerical, categorical, text, and mixed data.

### *1.2. Objective*

The primary objectives of the proposed framework for effective multi-modal DC are demonstrated as follows,

- The RBKDF is employed to capture and model the temporal, as well as the spatial, dynamics of data effectively.
- The DC is performed using the CGWQ-RLO, thus improving the convergence rate with less computation time.
- The Hadoop MapReduce process is utilized to develop a scalable clustering model using the RRH.
- The PF-CCA is applied to reduce the data dimensionality.
- For handling various types of missing inputs effectively, by using the WCKNN.

The remainder is organized as follows: Section 2 shows the literature review; Section 3 shows the suggested technique; Section 4 shows the findings and discussion; and Section 5 wraps up the suggested work with suggestions for the future.

## 2. Literature Survey

The section describes the literature overview of some existing methods with their pros and cons, such as [16], which established a global optimization and DC framework using the enhanced GWO with improved explorative searching capacity.

This framework used identical values for the control parameter of GWO for each leader wolf when updating their position. Furthermore, to perform accurate DC over a standard dataset. As per the results, this model improved explorative search capability. This model achieved a slow convergence rate owing to the high-dimensional or complex optimization problem.

Similarly, [17] presented an automatic DC framework based on the nature-inspired binary optimization algorithm. This model used binary optimization. Likewise, for the particle representation, the binary encoding scheme along with the principle of correlation-aware was employed. From the results, it was proven that this model reached a proper number of well-separated clusters. This model struggled with the high-dimensional data due to the curse of dimensionality.

[18] applied a DC methodology called the Water Wave Optimization (WWO)-based metaheuristic approach. Here, by employing the updated search mechanism, the absentia component named global best information was handled in WWO. Similarly, by using the decay operator, the premature convergence was eliminated. As per the results, this model used 13 benchmark clustering datasets and attained 96.51% accuracy. Nevertheless, this model failed to handle numerous types of data owing to the presence of noise and outliers.

[19] showed a model for text feature selection along with clustering by hybridizing GWO with the Grasshopper Optimization Algorithm (GOA). Primarily, the text data were collected and preprocessed. Next, by using the GWO-GOA, the text features were selected. Lastly, by using the Fuzzy C-Means Algorithm (FCMA), the selected features were clustered. From the results, it was proven that this model improved the reliability and minimized the computation time with 87.6% efficiency. Yet, FCMA was affected by long operational time and initial bias, thus impacting its clustering performance.

[20] exemplified a clustering optimization framework by introducing the hybrid algorithm. Initially, to generate Simplified GWO along with "Differential Perturbation (SDPGWO)", the strategy was embedded for improving global search ability. Next, the stochastic mean example learning strategy was used in PSO to prevent the model from getting stuck in the local optimum.

Lastly, to generate a hybrid algorithm," the poor for change strategy" was implemented. As per the results, this model had better optimization performance with a 6.2794s average run time. Yet, the utilization of this strategy is to combine the algorithms, which reduces efficiency and leads to potential errors and increased costs.

[21] demonstrated a DC framework centered on the enhanced metaheuristic algorithm. Here, to enhance the GWO, this model implemented a weight factor along with the K-means algorithm by increasing variety and avoiding premature convergence.

Likewise, this model was evaluated on ten numerical functions using the partitional clustering. According to the results, this model demonstrated the robustness and dependability of the K-means-based GWO in overcoming DC problems. However, noise and outliers were not removed, which affected the model's robustness.

[22] applied an adaptive approach for cluster task scheduling. This framework incorporated the crossover along with mutation strategies with the GWO to enhance DC. Likewise, the coefficients used for specified problem sets were analyzed to improve convergence as well as solution accuracy. According to the results, this model achieved an 87.2% success rate. However, the GWO algorithm did not yield highly accurate solutions, thereby limiting its effectiveness in certain applications.

[23] presented an automatic clustering methodology to distribute the non-uniform data." Here, to slice the raw data into a number of optimal ellipsoidal clusters automatically, this work implemented a Simulated-Annealing-centric Ellipsoidal Clustering (SAELLC) technique". Further, the ellipsoidal clusters were transformed into varying global cluster data points to evaluate the objective function. As per the validation, this model handled the diverse data distribution effectively and identified the optimal clusters automatically. Nevertheless, simulated annealing faced the drawback of producing varied outcomes across different runs.

[24] established a correlation-guided clustering framework for a fast hybrid feature selection of high-dimensional data using PSO. Primarily, by calculating the relevance value, the irrelevant features were removed. Next, by using a "Fast Correlation-guided Feature Clustering (FCFC)", the relevant features were clustered. Lastly, by using the PSO, the respective features were selected from the clusters. As per the result, this framework achieved the best classification accuracy. However, the model faced the challenge of high computational cost when handling large datasets.

[25] presented an effective and straightforward method for DC. This model implemented a deep K-means for learning the hidden representations associated with various implicitly lower-level attributes. Besides, by using the data samples, the clustering task was performed for the same class. As per the results, this model attained a better clustering performance with faster convergence. Still, K-Means was sensitive to initial conditions, which affected the stability and accuracy of clustering.

By evaluating the literature, it was depicted that the enhancement of current optimization is required for specific problems and for clustering purposes. This was done here.

## 3. The Proposed Methodology for Multi-Model Data Clustering based on Hadoop MapReduce with Spatio-Temporal Analysis

Here, the proposed framework and methodology for Hadoop MapReduce-based multi-modal DC and spatiotemporal analysis using CGWQRLO and RBKDE, respectively, are demonstrated. In Figure 1, the block diagram is depicted for the same.

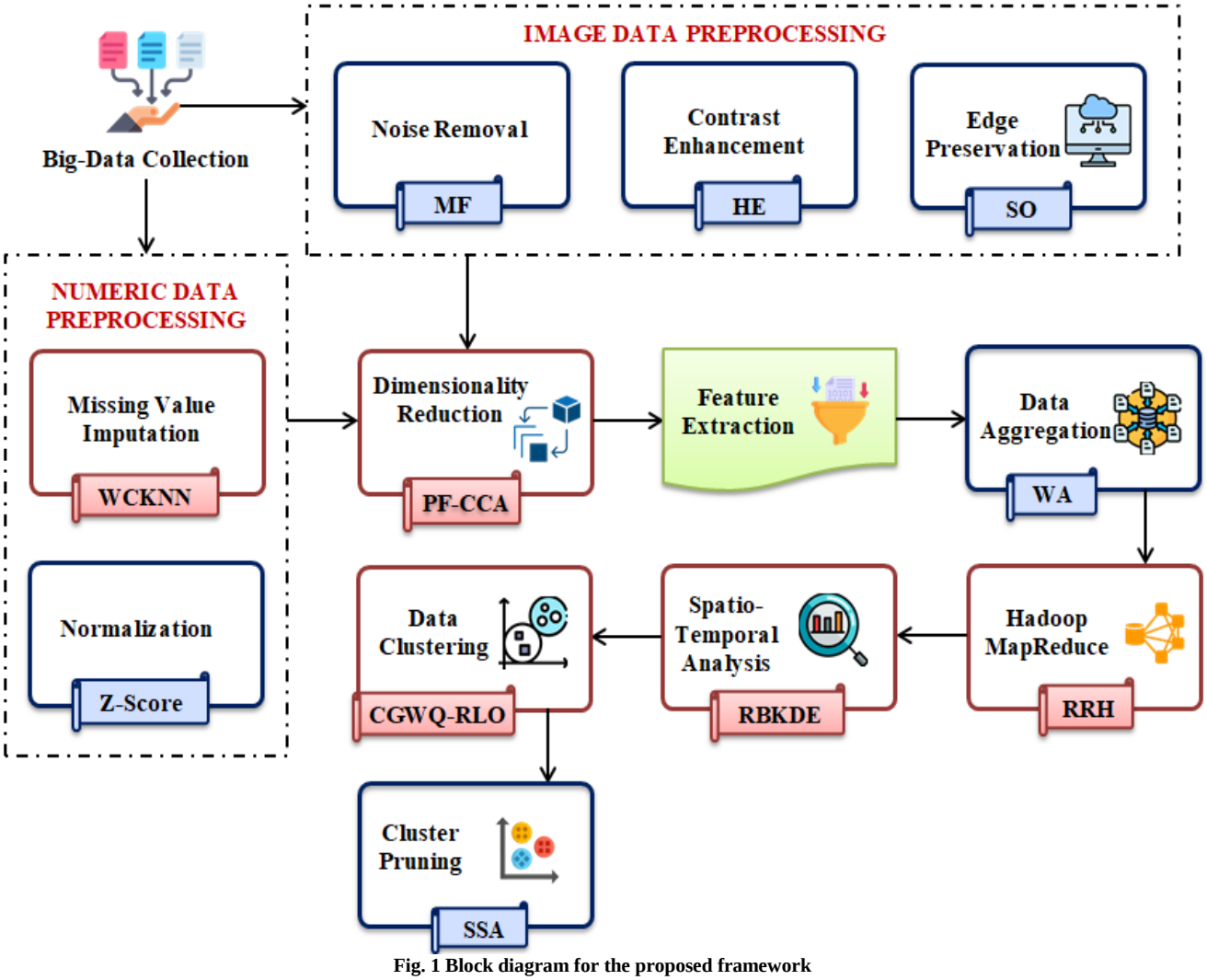


**Fig. 1 Block diagram for the proposed framework**

### *3.1. Data Collection*

Initially, the proposed framework collects multi-model data, including numeric values and images, from the Kaggle source for efficient DC.

$$\Im_{num}^{a'} = \{\Im_{num}^{1}, \Im_{num}^{2}, \Im_{num}^{3}, \cdots, \Im_{num}^{a''}\}; a' = 1 \rightarrow a'' \quad (1)$$

$$\Re_{img}^{b'} = \{\Re_{img}^{1}, \Re_{img}^{2}, \Re_{img}^{3}, \cdots, \Re_{img}^{b''}\}; b' = 1 \rightarrow b'' \quad (2)$$

Here, $\Im_{num}^{a'}$and $\Re_{img}^{b'}$Portray the collected numeric value and image data, respectively, and $a''$and $b''$exhibit the maximum number of$\Im_{num}^{a'}$ and$\Re_{img}^{b'}$, correspondingly.

### *3.2. Numeric Data Preprocessing*

Now, $\Im_{num}^{a'}$Often contains missing values, outliers, varying scales, and so on, thus affecting clustering performance. $\Im_{num}^{a'}$They are preprocessed regarding Missing Value Imputation (MVI) and Normalization in order to enhance the clustering performance.

#### *3.2.1. MVI*

Primarily, by using the WCKNN, the missing values in $\Im_{num}^{a'}$Are imputed. The conventional K-Nearest Neighbor (KNN) algorithm helps in predicting missing values based on the values of the nearest neighbors in the feature space. The KNN can lead to imputation bias for very few neighbors and excessively smooth out imputed values for large numbers of neighbors owing to the reliance on the choice of the number of nearest neighbors.

Hence, the proposed framework incorporates the Weighted Covariance (WC) measure to select nearest neighbors. The WC improves the imputation process efficiency by quantifying the degree of linear relationship between two variables centered on the weights assigned to each data point.

Initially, the WC $(\delta_{WC})A$ measure is employed to find the nearest neighbors $(\eta_k)$from$\Im_{num}^{a'}$.

$$\delta_{WC}\left(\Im_{num}^{a'}, \Im_{num}^{j}\right) = \frac{\sum \varpi^{wei}\left(\Im_{num}^{a'} - \bar{\Im}_{num}^{a'}\right)\left(\Im_{num}^{j} - \bar{\Im}_{num}^{j}\right)}{\sum \varpi^{wei}} \quad (3)$$

Here, $\varpi^{wei}$,$\bar{\Im}_{num}^{a'}$ , and $\bar{\Im}_{num}^{j}$represent the weight value, the weighted mean of$\Im_{num}^{a'}$, and weighted means of $j_{th}$data in$\Im_{num}^{a'}$, respectively, and $\Im_{num}^{j}$demonstrates $j_{th}$data in$\Im_{num}^{a'}$. Based on$\delta_{WC}$, the $K$-number of $\eta_k$Is found.

$$\eta_k = \{\eta_1, \eta_2, \eta_3, \cdots, \eta_K\}; k = 1 \rightarrow K \quad (4)$$

Lastly, to impute the missing values, the weighted mean of the selected $\eta_k$ Is calculated.

$$\Im^{mvi} = \frac{\sum \varpi^{wei} \cdot \eta_k}{\sum \varpi^{wei}} \quad (5)$$

Where, $\Im^{mvi}$Demonstrates the missing values in the imputed data.

#### *3.2.2. Normalization*

After MVI, by using the Z-score, the normalization is applied for each $\Im^{mvi}$ To enhance its quality for efficient DC. The Z-score helps standardize the data by transforming it into a normal distribution with a mean of zero and a variance of one.

$$\Im_{nor} = \frac{\Im^{mvi} - \mu_M(\Im^{mvi})}{\sigma_S(\Im^{mvi})} \quad (6)$$

Here, $\mu_M$and $\sigma_S$portray the mean and standard deviation of$\Im^{mvi}$, respectively, and $\Im_{nor}$Specifies the normalized or preprocessed numeric data.

### *3.3. Image Data Preprocessing*

Now, $\Re_{img}^{b'}$Often contains noise, low contrast, and flat edges, which can reduce the image quality, thus leading to lower clustering performance. Thus, the proposed framework performs preprocessing steps, such as noise removal and contrast enhancement, followed by edge preservation steps to enhance image quality.

- Noise Removal: Primarily, by using the Median Filter (MF), the noises like Gaussian and salt-and-pepper noises from$\Re_{img}^{b'}$ Are removed. The MF effectively removes the noise while preserving edges and fine details.

$$\Re_{noi}(\rho_x, \rho_y) = \upsilon^M \cdot \Re_{img}^{b'}(\rho_x - \ell_x, \rho_y - \ell_y) \quad (7)$$

Where, $\Re_{noi}$and$\upsilon^M$ Specify the noise-removed image and median value, respectively, and $(\rho_x, \rho_y), (\ell_x, \ell_y)$ exemplify the pixel and coordinates of $\Re_{img}^{b'}$.

- Contrast Enhancement: Now, to enhance the contrast of each$\Re_{noi}$The Histogram Equalization (HE) is employed. The HE helps to increase the visual quality and improve the appearance of the image by redistributing the intensity values of the image's pixels.

$$\Re_{con} = (I_m - 1)\sum \frac{\Re_{noi}(\rho_I)}{k'} \quad (8)$$

Here,$\Re_{con}$ ,$\rho_I$, and $k'$Demonstrate the contrast-enhanced image, pixel at intensity.$(I)$, and the maximum number of$I$, respectively.

- Edge Preservation: Lastly, the preprocessed images are attained by preserving the image edges of $\Re_{con}$Using the Sobel Operator (SO). The SO enhances the edge features by highlighting the areas of rapid intensity level.

$$\Re^{edg} = \sqrt{[G_{hor}(\Re_{con})]^2 + [G_{ver}(\Re_{con})]^2} \quad (9)$$

Where, $G_{hor}$and$G_{ver}$ Exemplify the gradients in horizontal and vertical directions, respectively, and $\Re^{edg}$Represents the edge preserved or preprocessed image.

### *3.4. Dimensionality Reduction*

Then, the high-dimensional data can cause computational complexity, redundancy, and overfitting in big DC and analysis tasks. To address this, the PF-CCA is employed to reduce the dimensionality of $\Re^{edg}$and$\Im_{nor}$. The conventional Principal Component Analysis (PCA) identifies the principal component of data with the highest variance, thus preserving essential information while reducing data dimensionality. However, the PCA assumes a Gaussian distribution for the data, thus providing suboptimal dimensionality reduction results when the data distribution deviates from this assumption. Therefore, the Fuzzy Correlation Matrix (FCM) is used instead of the covariance matrix. The FCM effectively captures the nonlinear relationships between variables and allows for more flexible modeling of complex relationships, accommodating nonlinear patterns within the data.

Primarily, $\Re^{edg}$and $\Im_{nor}$are standardized to ensure their mean $\mu_M$and standard deviation$(\sigma_S')$.

$$\aleph^{sta} = \left[\frac{1}{\sigma_S}(\{\Re^{edg}; \Im_{nor}\} - \mu_M)\right] \quad (10)$$

Here, $\aleph^{sta}$ Validates the standardized data. To capture the nonlinear and complex relationships, the FCM is employed. $(M_{FCM})$Is generated.

$$M_{FCM}(\aleph^{sta}) = exp\left[-\left|\aleph_p^{sta} - \aleph_q^{sta}\right|^2\right] \quad (11)$$

Where, $\aleph_p^{sta}$and $\aleph_q^{sta}$exhibit $p_{th}$and $q_{th}$standardized data, respectively. Lastly, the dimensionality-reduced data $(\aleph^{dim}())$are achieved by computing the eigenvectors$(\varsigma_{vec})$ and eigenvalues$(\tau_{val})$ from$M_{FCM}$.

$$\varsigma_{vec} = \frac{\varsigma_{vec} \cdot \tau_{val}}{M_{FCM}} \quad (12)$$

$$\aleph^{dim^{sta}{}_{vec}} \quad (13)$$

Therefore, the proposed framework efficiently reduces the high-dimensional data in$\Im_{nor}$ and $\Re^{edg}$Using the PF-CCA approach thereby reduces computational complexity and converges faster. The proposed PF-CCA's pseudo-code is illustrated as follows,

Pseudo-code for the PF-CCA technique

Input: Preprocessed image$\Re^{edg}$ and numeric data$\Im_{nor}$
Output: Dimensionality-reduced data$\aleph^{dim}$

Begin
  Initialize$\mu_M'$,$\sigma_S'$,$\varsigma_{vec}$, iteration$(\Gamma)$, and maximum iteration$(\Gamma_{max}())$
  Set$(\Gamma = 1)$
  While$(\Gamma_{max}())$
    For each $\Re^{edg}$and $\Im_{nor}$do
      Ensure $\mu_M$and$\sigma_S$
      #Standardization
      $$\aleph^{sta} = \left[\frac{1}{\sigma_S'}(\{\Re^{edg}; \Im_{nor}\} - \mu_M')\right]$$
      Compute FCM
      $$M_{FCM}(\aleph^{sta}) = exp\left[-\left|\aleph_p^{sta} - \aleph_q^{sta}\right|^2\right]$$
      Find eigenvectors and eigenvalues.
      $$\varsigma_{vec} = \frac{\varsigma_{vec} \cdot \tau_{val}}{M_{FCM}}$$
    End For
  End While
  Return: $\aleph^{dim^{sta}{}_{vec}}$
End

Next, the features are extracted from $\aleph^{dim}$ For effective multimodal DC, as depicted below.

### *3.5. Feature Extraction*

Now, the proposed framework initiates feature extraction to identify the most discriminative information from$\aleph^{dim}$ For an effective big DC. Likewise, this process is important for improving clustering performance, as it eliminates redundant or irrelevant features that are not related to the large DC. Therefore, the features like mean, variance, entropy, Gray-Level Co-occurrence Matrix (GLCM), and Homogeneity are extracted from $\aleph^{dim}$. Here, the GLCM features can be extracted using the image processing toolbox, which provides efficient methods for computing GLCM-related features. Next, the extracted features$\left(\aleph_{c'}^{fea}\right)$ are depicted as,

$$\aleph_{c'}^{fea} = \left\{\aleph_1^{fea}, \aleph_2^{fea}, \aleph_3^{fea}, \cdots, \aleph_{c''}^{fea}\right\}; c' = 1 \to c'' \quad (14)$$

Where, $c''$validates the maximum number of$\aleph_{c'}^{fea}$. Therefore, the proposed framework effectively extracts prominent data-related features for the most accurate and efficient clustering.

### *3.6. Data Aggregation*

Next, to enhance the clustering process, $\aleph_{c'}^{fea}$ They are aggregated using the Weighted Average (WA) technique. The WA allows the model to assign different levels of importance to individual data points centered on their relevance or specific characteristics, thus resulting in a more precise representation of overall data.

- Initially, by assigning $\varpi^{wei}$to each$\aleph_{c'}^{fea}$, the weighted sum $\left(\overset{\varpi^{wei}}{\Sigma}\right)$Is computed.

$$\gamma^{\aleph_{c'}^{fea}} = \left\{\gamma_1^{\aleph_{c'}^{fea}}, \gamma_2^{\aleph_{c'}^{fea}}, \gamma_3^{\aleph_{c'}^{fea}}, \cdots \gamma_{k''}^{\aleph_{c'}^{fea}}\right\} \quad (15)$$

$$\overset{\varpi^{wei}}{\Sigma} = \sum \varpi^{wei} \cdot \gamma^{\aleph_{c'}^{fea}} \quad (16)$$

Here, $k''$ establishes the maximum number of weight-assigned features$\left(\gamma^{\aleph_{c'}^{fea}}\right)$.

- Now, by calculating the WA of$\aleph_{c'}^{fea}$, the aggregated data ($\aleph^{agg}$)Are obtained.

$$\aleph^{agg} = \frac{1}{\sum \varpi^{wei}} \left( \overset{\varpi^{wei}}{\Sigma} \right) \tag{17}$$

Therefore, the proposed framework effectively aggregates the data using the WA, thereby facilitating better clustering accuracy and efficiency.

### *3.7. Hadoop MapReduce*

After data aggregation, the Hadoop MapReduce process is employed for each$\aleph^{agg}$ To improve the scalability of effective DC using the RRH technique. The Hadoop MapReduce is a software framework that helps to process large amounts of raw data in parallel effectively and is used to store and process big data. Here, the round-robin partitioning is employed in the mapping phase of Hadoop to split the data, and then Heap sorting is applied to optimize the partitions. In Figure 2, the structural diagram for RRH is depicted.

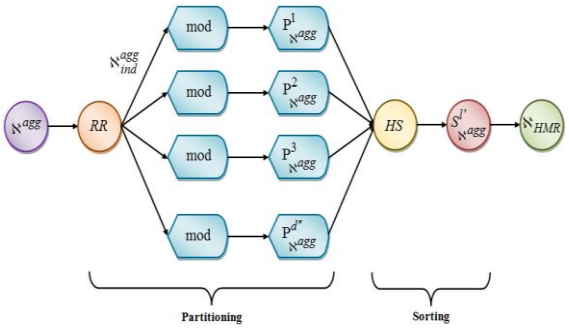


**Fig. 2 Structural diagram for the RRH**

Primarily, the Round-Robin($RR$) The approach is applied to each$\aleph^{agg}$ to split it into a number of partitions$\left(P_{\aleph^{agg}}^{d'}\right)$.

$$RR = \aleph_{ind}^{agg}\ mod(d'') \Rightarrow P_{\aleph^{agg}}^{d'} \tag{18}$$

$$P_{\aleph^{agg}}^{d'} = \left\{P_{\aleph^{agg}}^{1}, P_{\aleph^{agg}}^{2}, P_{\aleph^{agg}}^{3}, \cdots, P_{\aleph^{agg}}^{d''}\right\}; d' = 1 \rightarrow d'' \tag{19}$$

Where, $\aleph_{ind}^{agg}$and $d''$designate the index of$\aleph^{agg}$ and the maximum number of$P_{\aleph^{agg}}^{d'}$, respectively. Next, to optimize the order of elements within each$P_{\aleph^{agg}}^{d'}$Heap sorting is applied, thus attaining sorted data.$\left(S_{\aleph^{agg}}^{l'}\right)$. If $P_{\aleph^{agg}}^{d'}\left[\aleph_{ind}^{agg}\right] \neq \left[P_{\aleph^{agg}}^{d'}\right]_{max}$ Then the value is justified in equation 20, otherwise. $P_{\aleph^{agg}}^{d'}$

$$S_{\aleph^{agg}}^{l'} = \lambda_{SW}\left(P_{\aleph^{agg}}^{d'}\left[\aleph_{ind}^{agg}\right], \left[P_{\aleph^{agg}}^{d'}\right]_{max}\right) + \lambda_{heap}\left(P_{\aleph^{agg}}^{d'}\left[P_{\aleph^{agg}}^{d'}\right]_{max}\right); \tag{20}$$

Here, $\ell_{sw}$ and$\ell_{heap}$ exemplify the swapping and heap function, respectively, and $\left[P_{\aleph^{agg}}^{d'}\right]_{max}$ illuminates the largest value of$P_{\aleph^{agg}}^{d'}$. The Hadoop MapReduce data ($\aleph_{HMR}$) is obtained based on$S_{\aleph^{agg}}^{l'}$. The clustering process becomes more scalable and efficient by using the effective RRH technique, thus ensuring that large-scale multi-modal data is optimally portioned and sorted.

### *3.8. Spatio-Temporal Analysis*

Now, by using the RBKDE, spatiotemporal analysis is carried out between $\aleph_{HMR}$ For effective cluster formation. The traditional Kernel Density Estimation (KDE) estimates the spatial and temporal distribution of events or phenomena, thus allowing for visualization and analysis of their density over both space and time.

Nonetheless, owing to the requirement of kernel function calculation at each data point, the KDE increases the computational complexity. To overcome this issue, the Radial Basis Function (RBF) kernel, which captures the complex, nonlinear relationships between data points, is employed.

Additionally, the RBF kernel learns decision boundaries that are not linear in the original input space by mapping the raw input data into a high-dimensional feature vector space.

- Primarily, for accurate density estimation, the optimal bandwidth parameter ($\theta_{BW}$)is selected from $\aleph_{HMR}$.

$$\theta_{BW} = \tau_{sca} \cdot \vec{\sigma}_S \cdot h''^{-\frac{1}{6}} \tag{21}$$

Where, $\tau_{sca}$,$\vec{\sigma}_S$, and $h''$exhibit the scaling constant, standard deviation, and maximum number of$\aleph_{HMR}$, respectively.

- Next, the RBF kernel ($\Delta_{RBF}$) is used to determine each $\aleph_{HMR}$That contributes to the overall density estimation for analyzing the spatio-temporal characteristics.

$$\Delta_{RBF}\left[\aleph_{HMR}, \aleph_{HMR}^{i}\right] = exp\left(-\frac{\left\|\aleph_{HMR} - \aleph_{HMR}^{i}\right\|^2}{2(\theta_{BW})^2}\right) \tag{22}$$

Where, $\aleph_{HMR}^{i}$shows the $i_{th}$data in$\aleph_{HMR}$.

- Lastly, the kernel contribution is computed on each$\aleph_{HMR}$ to estimate density by $\Delta_{RBF}$and$\theta_{BW}$.

$$\aleph^{STd} = \frac{1}{h'' \cdot \theta_{BW}} \sum \Delta_{RBF}\left[\aleph_{HMR}, \aleph_{HMR}^{i}\right] \tag{23}$$

Here, the estimated density illustrates the spatio-temporal data analyzed$\aleph^{STd}$. Hence, the proposed framework efficiently analyzes the spatio-temporal characteristics by using the RBKDE approach, providing a continuous representation of event distribution over space and time.

### *3.9. Data Clustering*

Afterward, by using the CGWQ-RLO,$\aleph^{STd}$ they are clustered based on similar data points to enhance the clustering efficacy and achieve accurate multimodal cluster data. The conventional Grey Wolf Optimizer (GWO) is a metaheuristic optimization algorithm that employs a combination of global exploration and local exploitation mechanisms, allowing the algorithm to effectively explore the search space and find the nearest optimal solutions while balancing exploration and exploitation. Likewise, it is computationally efficient and suitable for large-scale optimization problems with moderate resources, thus making it a practical choice for clustering large datasets. But the GWO has a slower convergence rate and limited ability to escape from the local optimum and converge to the global optimum, especially in the case of high-dimensional or multi-modal optimization problems and complex search landscapes. Furthermore, the GWO leads to suboptimal convergence or premature convergence to local optima due to poor initialization. Hence, to overcome these limitations, the Circular Chaotic (CC) function is used to initialize GWO's population. Also, the proposed work uses Quasi-Reflexive Learning (QRL) to enhance the convergence rate by characterizing the search mechanism of grey wolves. Hence, the DC is efficiently estimated while ensuring optimal cluster formation by using these techniques.

#### *3.9.1. Initialization*

Primarily, using the CC function.($\gamma_{CC}$), the position of the grey wolf ($\Omega$) with its dimension ($f'$)The search space is initialized. Here, $\aleph^{STd}$ They are considered grey wolves.

$$\gamma_{CC}[\aleph^{STd}] \Rightarrow \Omega_{e\prime}^{f'} = sin\left(\varsigma_a^{Sca} \cdot \aleph_i^{STd}\right) + cos\left(\varsigma_b^{Sca} \cdot \aleph_j^{STd}\right) \quad (24)$$

$$\Omega_{e\prime}^{f'} = \left\{\Omega_{e\prime}^{1}, \Omega_{e\prime}^{2}, \Omega_{e\prime}^{3}, \cdots, \Omega_{e\prime}^{f''}\right\}; f' = 1 \rightarrow f'' \quad (25)$$

Here, $\varsigma_a^{Sca}$and$\varsigma_b^{Sca}$ exhibit the scaling parameter, $\aleph_i^{STd}$ and $\aleph_j^{STd}$portray$i_{th}$and$j_{th}$ grey wolf, correspondingly, and$f''$ , $e'$Exemplifies the maximum number of$f'$and$\aleph^{STd}$, respectively. Now, the fitness $\left(\Psi_{fit}\right)$is calculated for each $\Omega_{e\prime}^{f'}$based on the minimum distance ($X_{min}()$)between the cluster point($\partial_{CP}$) and$\aleph^{STd}$.

$$\Psi_{fit}\left[\Omega_{e\prime}^{f'}\right] = X[\partial_{CP}, \aleph^{STd}]_{min} \quad (26)$$

The best($^{\alpha}\Omega$), along with the second best$\left(^{\beta}\Omega\right)$, along with the third best($^{\chi}\Omega$), and the worst$\left(^{\delta}\Omega\right)$ Grey wolves are considered to begin the hunting process based on$\Psi_{fit}$.

#### *3.9.2. Prey Encircling*

Now, based on the behavior of $\aleph^{STd}$, the position of $\aleph^{STd}$ is updated $\left(\hat{\Omega}_{e\prime}^{f'}\right)$ to encircle the prey using QRL($\Lambda^{QRL}$) To avoid premature convergence.

$$\hat{\Omega}_{e\prime}^{f'} = \left[\Xi_{prey} - \xi_A \cdot \xi_D\right] + \Lambda^{QRL} \quad (27)$$

$$\Lambda^{QRL}\left(\Omega_{e\prime}^{f'}\right) = \Omega_{e\prime}^{f'} + \tau_{lear} \cdot \lambda_{lea}^{inf\left(\Xi_{prey}, \Omega_{e\prime}^{f'}\right) exp\left(\Omega_{e\prime}^{f'}, \Omega_{e\prime}^{f'^{-1}}\right)_{inf}} \quad (28)$$

$$\xi_D = \left|\xi_C \cdot \Xi_{prey} - \Omega_{e\prime}^{f'}\right| \quad (29)$$

Where, $\Xi_{prey}$and $\Omega_{e\prime}^{f'^{-1}}$Specify the prey position and the previous position of$\aleph^{STd}$, respectively. Likewise, $\xi_A, \xi_C$illustrates the coefficient vectors, $\tau_{lear}, \tau_{inf}$demonstrates the learning rate and influence factor, correspondingly, $\lambda_{lea}^{inf^{exp}}$portrays the leader influence and exploration factors, respectively, and $\xi_D$represents the vector to specify$\hat{\Omega}_{e\prime}^{f'}$.

$$\xi_A = \infty_{cf}[2r' - 1] \quad (30)$$

$$\xi_C = 2r' \quad (31)$$

$$\infty_{cf} = 2\left[\frac{\ell_{max}}{\ell_{max}}\right] \quad (32)$$

Here, $\infty_{cf}, r', \ell$, and $\ell_{max}$Portray the convergence factor, random vectors, iteration, and maximum iteration, respectively.

#### *3.9.3. Hunting*

Next, three random wolves($\Omega_1^{ran}, \Omega_2^{ran}, \Omega_3^{ran}$) are generated around the three leading wolves$\left(^{\alpha}\Omega, ^{\beta}\Omega, ^{\chi}\Omega\right)$and $^{\delta}\Omega$Adjust their positions based on$^{\alpha}\Omega$, $^{\beta}\Omega$, and$^{\chi}\Omega$.

$$\Omega_1^{ran} = {}^{\alpha}\Omega - \xi_A^1 \cdot X_{\alpha_{\Omega}} \quad (33)$$

$$\Omega_2^{ran} = {}^{\beta}\Omega - \xi_A^2 \cdot X_{\beta_{\Omega}} \quad (34)$$

$$\Omega_3^{ran} = {}^{\chi}\Omega - \xi_A^3 \cdot X_{\chi_{\Omega}} \quad (35)$$

Where, $\{\xi_A^1, \xi_A^2, \xi_A^3\} \in \xi_A$ and $X_{\alpha_{\Omega}}, X_{\beta_{\Omega}}$, and$X_{\chi_{\Omega}}$ demonstrates the distance between $\Omega_{e\prime}^{f'}$and $^{\alpha}\Omega$, $^{\beta}\Omega$, and$^{\chi}\Omega$, respectively.

$$X_{\alpha_{\Omega}} = \left|\xi_C^1 \times {}^{\alpha}\Omega - \Omega_{e\prime}^{f'}\right| \quad (36)$$

$$X_{\beta_{\Omega}} = \left|\xi_C^2 \times {}^{\beta}\Omega - \Omega_{e\prime}^{f'}\right| \quad (37)$$

$$X_{\chi_{\Omega}} = \left|\xi_C^3 \times {}^{\chi}\Omega - \Omega_{e\prime}^{f'}\right| \quad (38)$$

Here, $\{\xi_C^1, \xi_C^2, \xi_C^3\} \in \xi_C$. Now, the position of $\aleph^{STd}$is updated based on$^{\alpha}\Omega$, $^{\beta}\Omega$, and$^{\chi}\Omega$.

$$\hat{\Omega}_{e\prime}^{f'} = \frac{\Omega_1^{ran} + \Omega_2^{ran} + \Omega_3^{ran}}{3} \quad (39)$$

Until the solution converges, the process repeats. Therefore, the position of the best wolves is updated by moving towards the average position of the top three best solutions. Next, from the positions of the best grey wolf, the final clustered data$\left(\Phi_{Clu}^{g'}\right)$ Are obtained.

$$\Phi_{Clu}^{g'} = \Phi_{Clu}^{1}, \Phi_{Clu}^{2}, \Phi_{Clu}^{3}, \cdots, \Phi_{Clu}^{g''}; g' = 1 \rightarrow g'' \qquad (40)$$

Where, $g''$specifies the maximum number of$\Phi_{Clu}^{g'}$. Thus, the proposed framework attains 25 cluster points. The proposed CGWQ-RLO's pseudo-code is depicted as follows,

Pseudo-code of the proposed CGWQ-RLO

Input: Spatio-temporal analyzed data$\aleph^{STd}$

Output: Clustered data$\Phi_{Clu}^{g'}$

Begin

Initialize$\Omega_{e'}^{f'}$,$^{\alpha}\Omega$,$^{\beta}\Omega$,$^{\chi}\Omega$,$^{\delta}\Omega$, iteration($\Gamma$), and maximum iteration($\Gamma_{max}()$)

Set($\Gamma = 1$)

While($\Gamma_{max}()$)

For each $\aleph^{STd}$do

Initialize positions #using the CC function.

$$\gamma_{CC}[\aleph^{STd}] \Rightarrow \Omega_{e'}^{f'} = sin\left(\varsigma_a^{Sca} \cdot \aleph_i^{STd}\right) + cos\left(\varsigma_b^{Sca} \cdot \aleph_j^{STd}\right)$$

Calculate fitness

$$\Psi_{fit}\left[\Omega_{e'}^{f'}\right] = X[\partial_{CP}, \aleph^{STd}]_{min}$$

Find$^{\alpha}\Omega$, $^{\beta}\Omega$, $^{\chi}\Omega$ and $^{\delta}\Omega$

Update position #using QRL

$$\hat{\Omega}_{e'}^{f'} = \left[\Xi_{prey} - \xi_A \cdot \xi_D\right] + \Lambda^{QRL}$$

#Prey Encircling

Calculate vectors towards$^{\alpha}\Omega$, $^{\beta}\Omega$, and$^{\chi}\Omega$

$$X_{^{\alpha}\Omega} = \left|\xi_C^1 \times {}^{\alpha}\Omega - \Omega_{e'}^{f'}\right|$$

$$X_{^{\beta}\Omega} = \left|\xi_C^2 \times {}^{\beta}\Omega - \Omega_{e'}^{f'}\right|$$

$$X_{^{\chi}\Omega} = \left|\xi_C^3 \times {}^{\chi}\Omega - \Omega_{e'}^{f'}\right|$$

Update positions

If$\left(\Psi_{fit}\left[\Omega_{e'}^{f'}\right] < \Psi_{fit}({}^{\alpha}\Omega)\right)$

{

$$\Omega_1^{ran} = {}^{\alpha}\Omega - \xi_A^1 \cdot X_{^{\alpha}\Omega}$$

}

Else If$\left(\Psi_{fit}\left[\Omega_{e'}^{f'}\right] < \Psi_{fit}({}^{\beta}\Omega)\right)$

{

$$\Omega_2^{ran} = {}^{\beta}\Omega - \xi_A^2 \cdot X_{^{\beta}\Omega}$$

}

Else If$\left(\Psi_{fit}\left[\Omega_{e'}^{f'}\right] < \Psi_{fit}({}^{\chi}\Omega)\right)$

{

$$\Omega_3^{ran} = {}^{\chi}\Omega - \xi_A^3 \cdot X_{^{\chi}\Omega}$$

}

End If

Find a new position

$$\hat{\Omega}_{e'}^{f'} = \frac{\Omega_1^{ran} + \Omega_2^{ran} + \Omega_3^{ran}}{3}$$

End For

End While

Return: $\Phi_{Clu}^{g'} = \Phi_{Clu}^{1}, \Phi_{Clu}^{2}, \Phi_{Clu}^{3}, \cdots, \Phi_{Clu}^{g''}$

End

Therefore, by efficiently optimizing the cluster centers, balancing the exploration along with exploitation, and preventing premature convergence, the CGWQ-RLO performs the big DC process effectively.

### *3.10. Cluster Pruning*

Lastly, to refine the cluster structure and remove the redundant clusters, the cluster pruning process is employed by computing the Silhouette Score Analysis (SSA). The SSA ($\Xi^{SSA}$)This method appraises the quality of cluster points. It optimizes the number of cluster points by ensuring that only the most distinct and well-separated clusters remain, thereby leading to improved overall clustering performance.

$$\Xi^{SSA}\left(\Phi_{Clu}^{g'}\right) = \frac{Z_{avg}^{\text{near}}\left(\Phi_{Clu}^{g'}\right) - Z_{avg}^{\text{int}}\left(\Phi_{Clu}^{g'}\right)}{\max\left[Z_{avg}^{\text{near}}\left(\Phi_{Clu}^{g'}\right) - Z_{avg}^{\text{int}}\left(\Phi_{Clu}^{g'}\right)\right]} \qquad (41)$$

Where, $Z_{avg}^{\text{int}}$ and$Z_{avg}^{\text{near}}$ Specify the average of the intra-cluster distance and the nearest cluster distance within$\Phi_{Clu}^{g'}$, correspondingly. The global silhouette score($\Xi_{Fin}$) is calculated based on$\Xi^{SSA}$To improve the overall clustering performance.

$$\Xi_{Fin} = \frac{1}{g''}\sum_{g'=1}^{g''} \Xi^{SSA}\left[\Phi_{Clu}^{g'}\right] \qquad (42)$$

Therefore, the proposed framework efficiently clusters the big data and optimizes the clustered data by computing the SSA. This computation enhances the overall clustering performance by eliminating redundant clusters, refining the cluster assignments, and optimizing the cluster structures.

## 4. Experimental Result and its Discussion

Here, the proposed framework's overall performance in multi-modal DC is validated using various performance metrics.

### *4.1. Dataset Description*

The proposed CGWQ-RLO-based based framework uses three datasets to analyze its performance in multi-modal DC, namely the Credit Card Dataset, Image Clustering Analysis dataset, and Wine Quality Dataset. The source link to access these datasets is depicted under the reference section. In order to show the inclusion of big data, the proposed work collects the numeric data and image data from the Credit Card Dataset and Image Clustering dataset, respectively. The credit card dataset

includes 8950 instances of customer credit card details by summarizing the usage behavior of credit card holders. Likewise, the Image dataset for clustering purposes contains colored images of 100 objects, with a black background. Moreover, the performance of the proposed work is also evaluated by using the Wine Quality Dataset. Therefore, the proposed framework's methodology is evaluated in terms of its performance in effective multimodal DC using these three datasets. In Table 1, the image result for the proposed framework is depicted.

**Table 1. Image results**

| Phase | Sample Image 1 | Sample Image 2 |
|---|---|---|
| Input Image |  |  |
| Noise Removal |  |  |
| Contrast Enhancement |  |  |
| Edge Preservation |  |  |

Table 1 Depicts the image results obtained after performing each phase of DC, such as input, noise removal, contrast enhancement, and edge preservation, thus illustrating the proposed framework's efficiency in image DC.

### *4.2. Performance Evaluation of the Proposed Framework for Multi-modal Data*

Here, the performance for both image and numeric data regarding DC, spatiotemporal analysis, and missing value imputation is evaluated based on the Credit Card Dataset and Image Clustering dataset. In Figure 3, the CGWQ-RLO's efficiency in DC for preprocessed multi-models is depicted. As per the figure, regarding the accuracy, rand index, Dunn index, and precision, the CGWQ-RLO attains 97.79%, 97.68%, 97.68%, and 97.51%; while, the conventional techniques like Enhanced GWO (EGWO), Cat Swarm Optimization (CSO), GWO, Harris Hawks Optimization (HHO), and Particle Swarm Optimization (PSO) attain 93.67%, 91.92%, 87.32%, 86.94%, and 82.58%, respectively.

Thus, by effectively clustering the multi-modal data, the CGWQ-RLO outperforms the other prevailing approaches. Therefore, this is due to the utilization of the chaotic circle function and QRL strategy, which effectively clusters the data by improving population initialization and enhancing search efficiency.

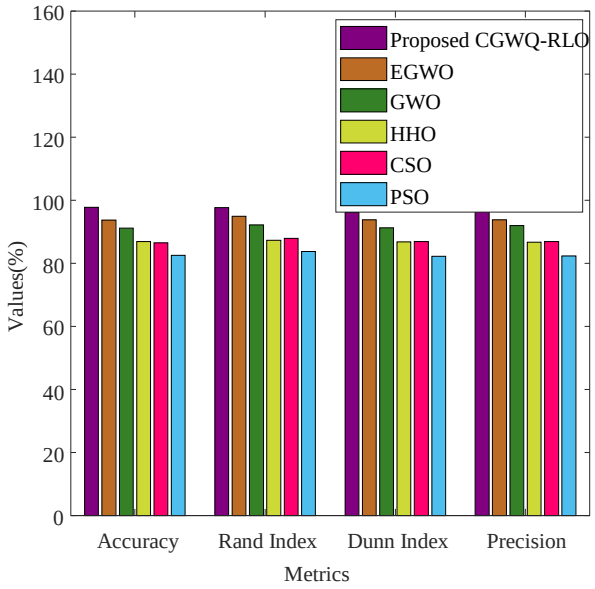


**Fig. 3 Performance validation of the proposed framework in data clustering with preprocessing**

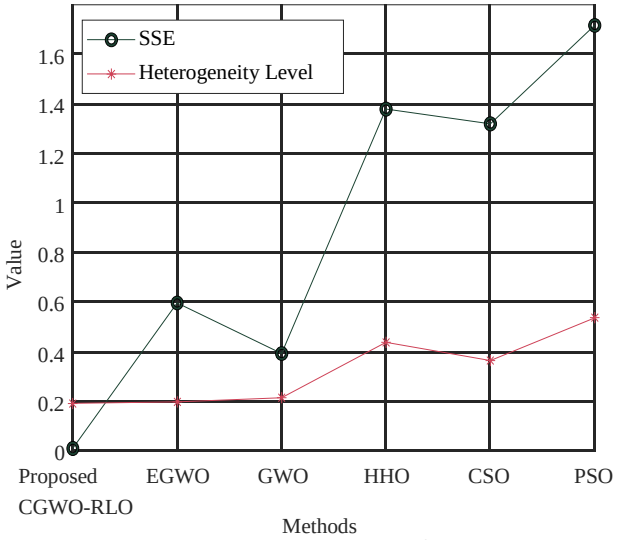


**Fig. 4 SSE and heterogeneity level analysis for the CGWQ-RLO without preprocessing**

The CGWO-RLO's performance in DC without preprocessing the multi-model data regarding the Sum of Squared Error (SSE) and heterogeneity level is depicted in Figure 4. Here, the CGWQ-RLO outperforms the other prevailing approaches by effectively clustering the noisy multimodal data using the circle chaotic function and QRL technique. As per Figure 4, the CGWQ-RLO attains 0.00938 SSE and 0.191341 heterogeneity levels, which are very low when compared to the other conventional methodologies like EGWO (0.59685 SSE), GWO (0.213791 Heterogeneity level), HHO (1.379303 SSE), CSO (0.364576 Heterogeneity level), and PSO (1.717171 SSE).

In Figure 5, the Mean Integrated Square Error (MISE) of the RBKDE in spatiotemporal analysis is demonstrated for both preprocessed and noisy multi-modal data. Here, to enhance density estimation by effectively capturing complex nonlinear spatiotemporal relationships, the RBKDE uses the RBF kernel, thus improving clustering accuracy and decision boundary formation. The RBKDE attains better MISE of 0.067377 and 0.04431 for both preprocessed and noisy multi-modal data, respectively. Thus, the RBKDE outperformed other conventional methodologies, including Gaussian-KDE (GKDE), KDE, Parzen Estimator (PE), and Multivariate-KDE (MKDE).

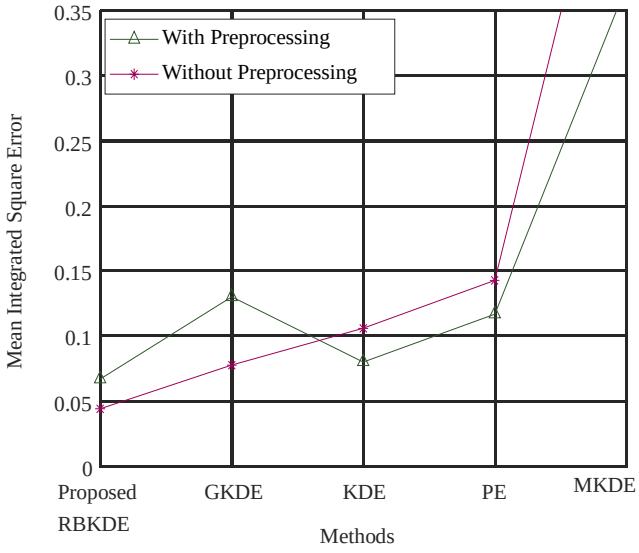


**Fig. 5 MISE validation for the RBKDE with and without preprocessing**

**Table 2. RMSE and MSE validation for the WCKNN**

| Methodology | RMSE | MSE |
|---|---|---|
| Proposed WCKNN | 0.043312514 | 0.208116587 |
| KNN | 1.133438028 | 1.064630466 |
| K-Means | 1.09071445 | 1.044372754 |
| MLE | 1.076859753 | 1.037718533 |
| MI | 1.847040225 | 1.35905858 |

The WCKNN's efficiency in MVI for multi-modal data regarding Root Mean Square Error (RMSE) along with Mean Squared Error (MSE) is depicted in Table 2. From the table, when analogized to other conventional methodologies like KNN, K-Means, Maximum Likelihood Estimation (MLE), and Mean Imputation (MI), the WCKNN performs more effectively in MVI. This is because of the adaptive weighting mechanism in WC, thus preserving the data variability and reducing bias in multi-modal data imputation. Hence, the WCKNN attains a lower RMSE and MSE of 0.043312514 and 0.208116587, respectively.

### *4.3. Performance Validation on Input Image Data*

Here, the proposed framework's performance in efficient multimodal DC regarding the input image data using the Image Clustering dataset is demonstrated. Likewise, the overall proposed framework's efficiency is evaluated by validating the image data through DC, spatiotemporal analysis, dimensionality reduction, and preprocessing (with or without these steps).

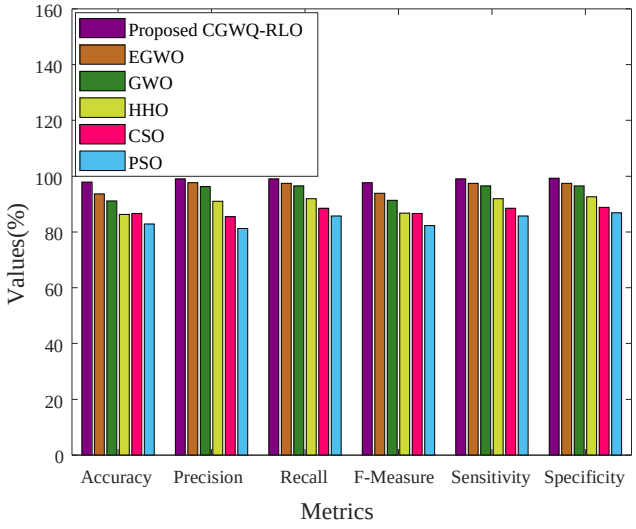


**Fig. 6 Performance analysis of the CGWQ-RLO with preprocessing using image clustering dataset**

The CGWQ-RLO's efficiency in DC for the input image data is demonstrated in Figure 6. As per the figure, the CGWQ-RLO outperforms the other conventional methodologies. Here, the CGWQ-RLO attains 97.888% accuracy, 99.04% precision, 99% recall, 97.65% f-measure, 99% sensitivity, and 99.36% specificity. Therefore, the proposed framework efficiently performs the DC process using the CGWQ-RLO, which includes chaotic initialization and QRL function, thus enhancing the population diversity, preventing premature convergence, and optimizing cluster separation.

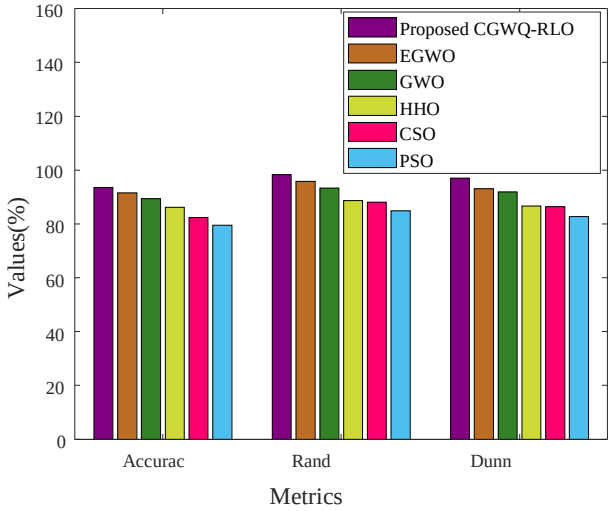


**Fig. 7 Effectiveness validation of the CGWQ-RLO without preprocessing using the image clustering dataset**

In Figure 7, the CGWQ-RLO's effectiveness in DC without including the preprocessing phase is depicted. Here, when analogied to other conventional approaches, the CGWQ-RLO without preprocessing the input image data attains higher accuracy, rand index, and Dunn index. Therefore, the proposed framework also provides an efficient DC methodology without preprocessing by efficiently mitigating the conventional GWO's limitations. Here, the conventional EGWO, GWO, HHO, CSO, and PSO attain 91.48% accuracy, 93.27% rand index, 86.72% Dunn index, 82.37% accuracy, and 84.84% rand index, respectively.

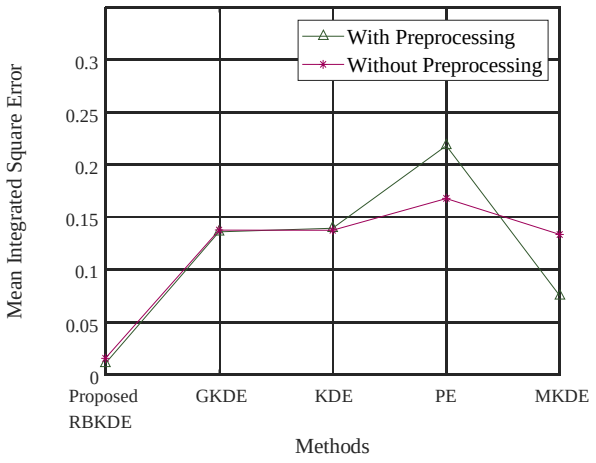


**Fig. 8 MISE analysis of the RBKDE with and without preprocessing using the image clustering dataset**

The RBKDE's performance in the spatiotemporal analysis with and without preprocessing in terms of MISE is depicted in Figure 8. As per Figure 8, the MISE of the RBKDE with and without preprocessing is 0.046716 and 0.015584, respectively, which are very low when weighed against the other conventional approaches. Therefore, the proposed framework efficiently analyzes the data for both preprocessed and noisy input images.

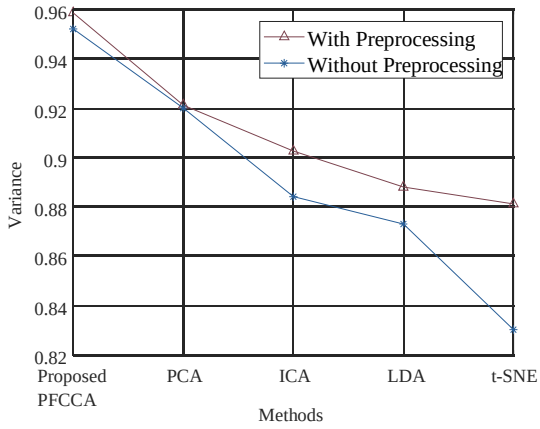


**Fig. 9 Variance validation for the PF-CCA with and without preprocessing**

In Figure 9, the PF-CCA's efficiency in reducing dimensionality for the preprocessed and noisy image data in terms of variance is demonstrated. Here, when compared with the other conventional approaches like PCA, Linear Discriminant Analysis (LDA), Independent Component Analysis (ICA), and t-distributed Stochastic Neighborhood Embedding (t-SNE), the PF-CCA has a higher variance for both preprocessed (0.9587) and noise image (0.9521). This is because of the inclusion of the

FCM approach, which captures the complex nonlinear relationships to achieve superior dimensionality reduction, thus ensuring better feature retention and data representation.

### *4.4. Effectiveness Analysis on Input Numeric Data*

Here, the proposed CGWQ-RLO framework's performance in efficient multi-modal DC regarding DC, spatiotemporal analysis, dimensionality reduction, and missing value imputation using credit card and wine quality datasets is elucidated.

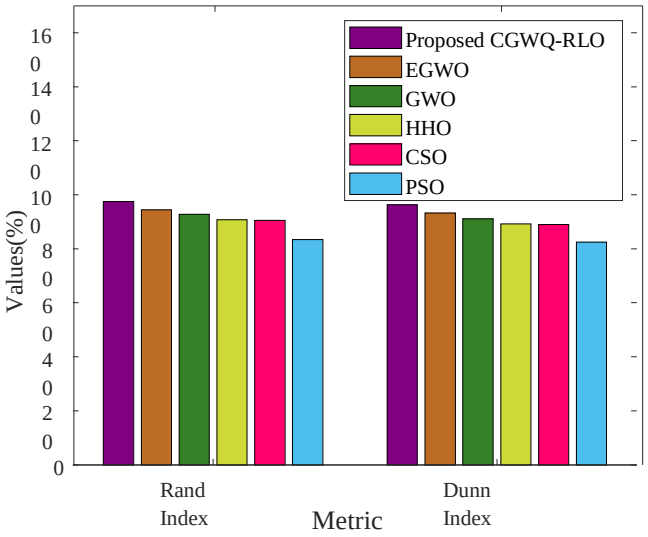


**Fig. 10 Efficiency validation of CGWQ-RLO with preprocessing for the credit card dataset**

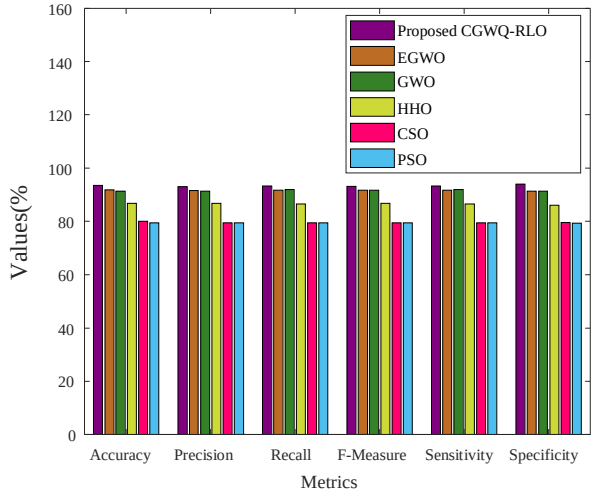


**Fig. 11 Performance evaluation of the CGWQ-RLO without preprocessing for the credit card dataset**

The CGWQ-RLO's performance in DC when processing the numeric data with and without preprocessing is illustrated in Figures 10 and 11. As per Figure 10, the CGWQ-RLO efficiently clustered data by attaining a higher rand index (97.57%) and Dunn index (96.32%). Similarly, for noisy numeric data, the CGWQ-RLO attains good results in DC with 93.55% accuracy, 93.03% precision, 93.26% recall, 93.15% f-measure, 93.26% sensitivity, and 93.98% specificity. This is because of the utilization of the chaotic map function and the QRL approach in the GWO. Nevertheless, for both preprocessed and noisy numeric data, the conventional technique achieves lower performance in DC.

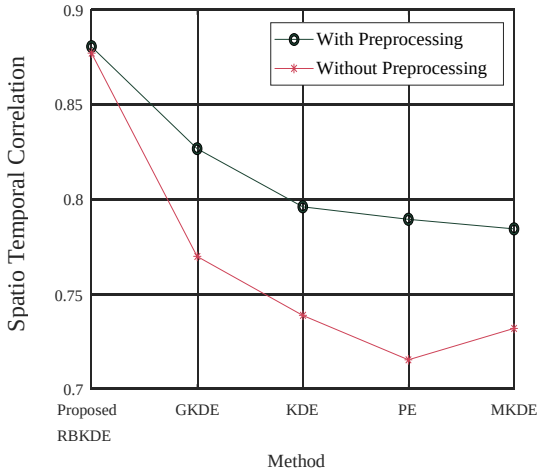


**Fig. 12 Spatial-Temporal correlation analysis with and without preprocessing using the credit card dataset**

The RBKDE's efficiency in spatial correlation and temporal analysis for both preprocessed and noisy numeric data is depicted in Figure 12. Here, the RBKDE's spatial-temporal correlation is compared with that of other conventional methodologies. From these results, the RBKDE attains better spatial-temporal correlation for both preprocessed (0.880553) and noisy data (0.877178). Therefore, by using the RBF kernel, the performance of the RBKDE in spatio-temporal analysis is improved by enhancing the robustness against noise and the ability to model complex data.

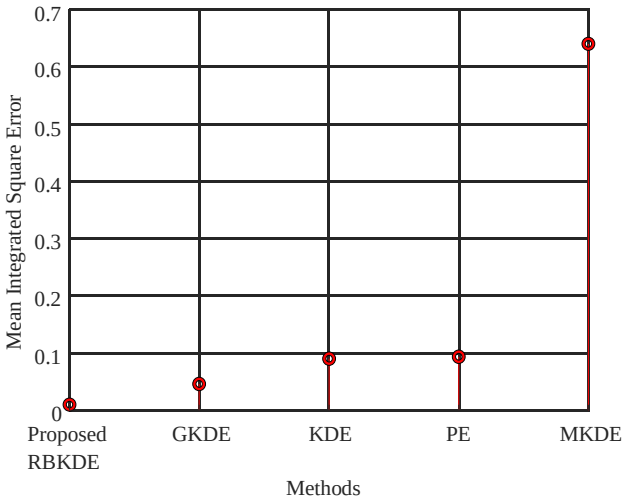


**Fig. 13 MISE evaluation regarding wine quality dataset**

The RBKDE's effectiveness for the numeric data collected from the Wine Quality dataset is depicted in Figure 13. To highlight the RBKDE's efficiency in analyzing spatial-temporal characteristics based on MISE, the proposed work processes the numeric data from the Wine Quality dataset. Here, the RBKDE achieves a higher MISE of 0.01014, whereas the conventional techniques, such as GKDE, KDE, PE, and MKDE, attain lower MISEs

of 0.75665, 0.835842, 0.811441, and 0.754159, respectively. Therefore, the proposed framework generalizes well and attains robust performance on the Wine Quality dataset.

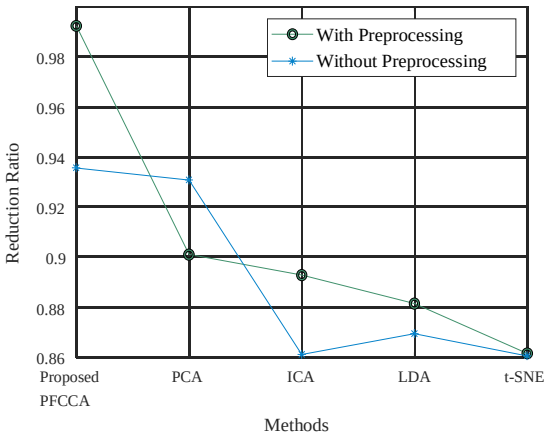


**Fig. 14 Reduction ratio validation with and without preprocessing for credit card dataset**

In Figure 14, the reduction ratio of the PF-CCA technique for both preprocessed and noisy numeric data is demonstrated. Here, the PF-CCA attains a higher reduction ratio. Nevertheless, for both preprocessed and noisy data, conventional approaches like PCA, ICA, LDA, and t-SNE attain lower reduction ratios. Therefore, the proposed framework efficiently reduces the dimensionality by using the FCM methodology with a higher reduction ratio of 0.899536 and 0.935681 for preprocessed and noisy data, respectively.

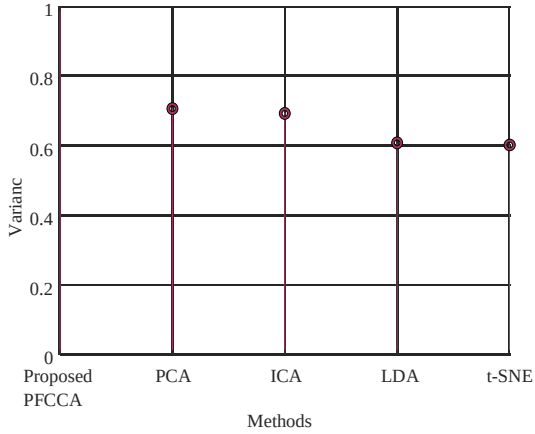


**Fig. 15 Variance analysis using the wine quality dataset**

**Table 3. MAE analysis**

| Techniques | MAE | |
|---|---|---|
| | Credit Card Dataset | Wine Quality Dataset |
| Proposed WCKNN | 0.0054 | 0.001504418 |
| KNN | 0.00697 | 0.014205088 |
| K-Means | 0.00897 | 0.008700398 |
| MLE | 0.029997 | 0.285 |
| MI | 0.069997 | 0.418381274 |

The PF-CCA's variance analysis using the Wine Quality dataset to showcase its performance in dimensionality reduction is depicted in Figure 15. The PF-CCA uses the FCM instead of the covariance matrix, thus reducing the dimensionality of the numeric data from the Wine Quality dataset. The PF-CCA reduces the dimensionality with a variance of 0.999922 by capturing the nonlinear relationships between variables. This indicates that the PF-CCA outperforms the other conventional methodologies.

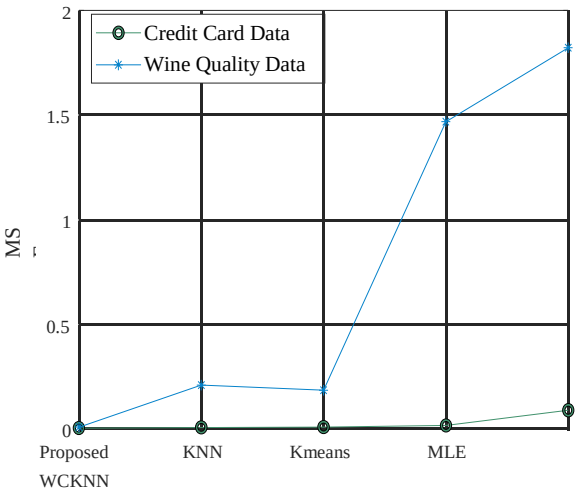


**Fig. 16 MSE validation regarding the credit card and wine quality dataset**

The WCKNN's performance in MVI for both the credit card and wine quality datasets regarding Mean Absolute Error (MAE) and MSE is depicted in Table 3 and Figure 16. The proposed framework efficiently imputes the missing values by integrating the WC with KNN, thus attaining accurate preprocessed data for both the credit and Wine Quality datasets. As per the results, this WCKNN attains better MAE (0.0054 and 0.001504 for the credit card and wine quality dataset, respectively) and MSE (0.004457 and 0.008549 for the credit card and wine quality dataset, respectively). Therefore, the proposed work outperforms other conventional methodologies.

### *4.5. Comprehensive Evaluation*

To highlight the proposed CGWQ-RLO's work efficiency in multimodal DC, the comparative evaluation is carried out by comparing it with several conventional works. A comprehensive evaluation of the proposed CGWQ-RLO's framework's efficiency in multimodal DC compared to conventional approaches is depicted in Table 4. In multimodal big DC, the CGWQ-RLO technique for spatiotemporal analysis-based multimodal clustering achieves high accuracy. Although the existing techniques, such as metaheuristic optimization, HRSA and ROA, hybrid firefly-PSO, modified GWO, and MutaSwarmClus, perform well in handling large datasets, solving clustering problems, ensuring scalability, and enhancing convergence, they suffer from limitations like ineffective parameter tuning, high execution time, and unsuitability for real-world applications [31, 32].

**Table 4. Comparative analysis**

| Authors' Name | Objectives | Methodology | Advantage | Limitation |
|---|---|---|---|---|
| Proposed Framework | Spatio-temporal analysis-based multimodal DC | CGWQ-RLO | Provides an effective multimodal big DC with high accuracy | Failed to learn discriminative features from Hyper Spectral Images (HSI) |
| [26] | Big data text clustering | Metaheuristic optimization | Effectively handles large datasets and complex clustering problems | Failed to tune the parameters effectively, thus affecting the initial cluster centroid. |
| [27] | DC to solve optimization problems | Hybrid Reptile Search Algorithm (HRSA) and Remora Optimization Algorithm (ROA) | Effectively solved the clustering problems with remarkable efficacy | Requires more time for execution |
| [28] | DC efficiency enhancement | Mutation-enhanced Swarm Clustering (MutaSwarmClus) | Provides scalability, efficiency, and accuracy | This model was not suitable for real-world applications. |
| [29] | Automatic DC | Hybrid enhanced firefly and PSO | Provides better convergence speed and diversification | Failed to solve various complex optimization problems. |
| [30] | Metaheuristic-based DC | Modified GWO | Reduces error data by lowering the intra-cluster distance | This model failed to learn various data, thus leading to a high error rate |

## 5. Conclusion

Here, this paper proposed a practical multimodal big DC framework with Hadoop MapReduce and spatio-temporal analysis using RBKDE and CGWQ-RLO techniques. The proposed work effectively clustered the big data than the other prevailing approaches based on the spatio-temporal analysis.Here, from the two datasets, the multimodal data were collected and preprocessed. During preprocessing, the missing values from the numeric data were effectively imputed with an MAE of 0.001504. Next, the PF-CCA reduced the dimensionality of the multimodal data. Furthermore, RRH performed the Hadoop MapReduce, and the RBKDE analyzed the spatio-temporal characteristics of the multimodal data. Lastly, by using the CGWQ-RLO, the DC was carried out, which attained a higher accuracy for numeric (97.93%), image (97.88%), and multimodal (97.79%) datasets. Therefore, the proposed framework generalized well on effective multimodal big DC and outperformed the other existing approaches.

### *5.1. Future Recommendation*

This framework failed to reduce clustering time and learn more discriminative features from the HIS, despite providing effective multi-modal DC. Therefore, this work will include the super-pixel clustering task in the future to reduce clustering time and consider the multi-view information of the HIS.

## Dataset

- Credit Card Dataset for Clustering, Kaggle. [Online]. Available: https://www.kaggle.com/datasets/arjunbhasin2013/ccdata
- Image-Clustering-Analysis, Kaggle. [Online]. Available: https://www.kaggle.com/code/marcosd3souza/image-clustering-analysis/input

## References

[1] Ammar Kamal Abasi et al., "An Improved Text Feature Selection for Clustering Using Binary Grey Wolf Optimizer," *Proceedings of the 11th National Technical Seminar on Unmanned System Technology*, Kuala Lumpur, Malaysia, pp. 503-516, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[2] Arvinder Kaur, Yugal Kumar, and Jagpreet Sidhu, "Exploring Meta-Heuristics for Partitional Clustering: Methods, Metrics, Datasets, and Challenges," *Artificial Intelligence Review*, vol. 57, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[3] Chou-Yuan Lee, Wei Wang, and Jian-Qiong Huang, "Clustering and Classification for Dry Bean Feature Imbalanced Data," *Scientific Reports*, vol. 14, 2024. [CrossRef] [Google Scholar] [Publisher Link]
[4] Gbeminiyi John Oyewole, and George Alex Thopil, "Data Clustering: Application and Trends," *Artificial Intelligence Review*, vol. 56, pp. 6439-6475, 2023. [CrossRef] [Google Scholar] [Publisher Link]
[5] Grzegorz Mrukwa, and Joanna Polanska, "DiviK: Divisive Intelligent K-means for Hands-Free Unsupervised Clustering in big Biological Data," *BMC Bioinformatics*, vol. 23, pp. 1-24, 2022. [CrossRef] [Google Scholar] [Publisher Link]
[6] Haitao Ding, Chu Sun, and Jianqiu Zeng, "Fuzzy Weighted Clustering Method for Numerical Attributes of Communication Big Data Based on Cloud Computing," *Symmetry*, vol. 12, no. 4, pp. 1-14, 2020. [CrossRef] [Google Scholar] [Publisher Link]
[7] Papia Ray, S. Surender Reddy, and Tuhina Banerjee, "Various Dimension Reduction Techniques for High Dimensional Data Analysis: A Review," *Artificial Intelligence Review*, vol. 54, pp. 3473-3515, 2021. [CrossRef] [Google Scholar] [Publisher Link]
[8] Parul Agarwal, Shikha Mehta, and Ajith Abraham, "A Meta-Heuristic Density-Based Subspace Clustering Algorithm for High Dimensional Data," *Research Square*, 2021. [CrossRef] [Google Scholar] [Publisher Link]
[9] Seyyed Mohammad Razavi, Mohsen Kahani, and Samad Paydar, "Big Data Fuzzy C-means Algorithm based on Bee Colony Optimization using an Apache Hbase," *Journal of Big Data*, vol. 8, pp. 1-22, 2021. [CrossRef] [Google Scholar] [Publisher Link]
[10] Yusei Ito et al., "Rational Partitioning of Spectral Feature Space for Effective Clustering of Massive Spectral Image Data," *Scientific Reports*, vol. 14, pp. 1-11, 2024. [CrossRef] [Google Scholar] [Publisher Link]
[11] Majjari Sudhakar, and Koteswara Rao Anne, "Optimizing Data Processing for Edge-Enabled IoT Devices using Deep Learning Based Heterogeneous Data Clustering Approach," *Measurement: Sensors*, vol. 31, pp. 1-11, 2024. [CrossRef] [Google Scholar] [Publisher Link]
[12] Mandli Rami Reddy et al., "Energy-Efficient Cluster Head Selection in Wireless Sensor Networks using an Improved Grey Wolf Optimization Algorithm," *Computers*, vol. 12, no. 2, pp. 1-17, 2023. [CrossRef] [Google Scholar] [Publisher Link]
[13] Mohammadhossein Ghahramani et al., "Intelligent Geodemographic Clustering based on Neural Network and Particle Swarm Optimization," *IEEE Transactions on Systems Man and Cybernetics Systems*, vol. 52, no. 6, pp. 3746-3756, 2022. [CrossRef] [Google Scholar] [Publisher Link]
[14] Raj Kumar Paul et al., "Image Compression Scheme based on Optimized K-Means Clustering and Higher-Level Decomposed DWT," *Procedia Computer Science*, vol. 235, pp. 642-655, 2024. [CrossRef] [Google Scholar] [Publisher Link]
[15] Tie Li et al., "An Integrated Cluster Detection, Optimization, and Interpretation Approach for Financial Data," *IEEE Transactions on Cybernetics*, vol. 52, no. 12, pp. 13848-13861, 2022. [CrossRef] [Google Scholar] [Publisher Link]
[16] Gyanaranjan Shial, Sabita Sahoob, and Sibarama Panigrahi, "An Enhanced GWO Algorithm with Improved Explorative Search Capability for Global Optimization and Data Clustering," *Applied Artificial Intelligence*, vol. 37, no. 1, pp. 1-49, 2023. [CrossRef] [Google Scholar] [Publisher Link]
[17] Behnaz Merikhi, and M.R. Soleymani, "Automatic Data Clustering Framework using Nature-Inspired Binary Optimization Algorithms," *IEEE Access*, vol. 9, pp. 93703-93722, 2021. [CrossRef] [Google Scholar] [Publisher Link]
[18] Arvinder Kaur, and Yugal Kumar, "A New Metaheuristic Algorithm based on Water Wave Optimization for Data Clustering," *Evolutionary Intelligence*, vol. 15, pp. 759-783, 2021. [CrossRef] [Google Scholar] [Publisher Link]
[19] R. Purushothaman, S.P. Rajagopalan, and Gopinath Dhandapani, "Hybridizing Gray Wolf Optimization (GWO) with Grasshopper Optimization Algorithm (GOA) for Text Feature Selection and Clustering," *Applied Soft Computing*, vol. 96, 2020. [CrossRef] [Google Scholar] [Publisher Link]
[20] Xinming Zhang et al., "Hybrid Particle Swarm and Grey Wolf Optimizer and its Application to Clustering Optimization," *Applied Soft Computing*, vol. 101, 2021. [CrossRef] [Google Scholar] [Publisher Link]
[21] Manoharan Premkumar et al., "Augmented Weighted K-means Grey Wolf Optimizer: An Enhanced Metaheuristic Algorithm for Data Clustering Problems," *Scientific Reports*, vol. 14, pp. 1-33, 2024. [CrossRef] [Google Scholar] [Publisher Link]
[22] Hongbo Wang et al., "An Improved Grey Wolf Optimizer with Flexible Crossover and Mutation for Cluster Task Scheduling," *Information Sciences*, vol. 704, pp. 1-42, 2025. [CrossRef] [Google Scholar] [Publisher Link]
[23] Alaa E. Abdel-Hakim et al., "Ellipsoidal K-Means: An Automatic Clustering Approach for Non-Uniform Data Distributions," *Algorithms*, vol. 17, no. 12, pp. 1-23, 2024. [CrossRef] [Google Scholar] [Publisher Link]
[24] Xian-Fang Song et al., "A Fast Hybrid Feature Selection based on Correlation-Guided Clustering and Particle Swarm Optimization for High-Dimensional Data," *IEEE Transactions on Cybernetics*, vol. 52, no. 9, pp. 9573-9586, 2022. [CrossRef] [Google Scholar] [Publisher Link]
[25] Shudong Huang et al., "Robust Deepk-Means: An Effective and Simple Method for Data Clustering," *Pattern Recognition*, vol. 117, 2021. [CrossRef] [Google Scholar] [Publisher Link]
[26] Laith Abualigah et al., "Advances in Meta-Heuristic Optimization Algorithms in big Data Text Clustering," *Electronics*, vol. 10, no. 2, pp. 1-29, 2021. [CrossRef] [Google Scholar] [Publisher Link]
[27] Khaled H. Almotairi, and Laith Abualigah, "Hybrid Reptile Search Algorithm and Remora Optimization Algorithm for Optimization Tasks and Data Clustering," *Symmetry*, vol. 14, no. 3, pp. 1-29, 2022. [CrossRef] [Google Scholar] [Publisher Link]
[28] Saleem Malik et al., "MutaSwarmClus: Enhancing Data Clustering Efficiency with Mutation-Enhanced Swarm Algorithm," *Cluster Computing*, vol. 28, pp. 1-24, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[29] Mandakini Behera et al., "Automatic Data Clustering by Hybrid Enhanced Firefly and Particle Swarm Optimization Algorithms," *Mathematics*, vol. 10, no. 19, pp. 1-29, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[30] Ramin Ahmadia, Gholamhossein Ekbatanifardb, and Peyman Baya, "A Modified Grey Wolf Optimizer based Data Clustering Algorithm," *Applied Artificial Intelligence*, vol. 35, no. 1, pp. 63-79, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[31] Zainab Saadoon Naser, Hend Marouane, and Ahmed Fakhfakh, "Multi-Object-Based Efficient Traffic Signal Optimization Framework via Traffic Flow Analysis and Intensity Estimation Using UCB-MRL-CSFL," *Vehicles*, vol. 7, no. 3, pp. 1-27, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[32] Moshiur Rahman Faisal et al., "Context-Aware Data Cleaning: Optimizing Bengali Text for Contextual Text Classification," *SN Computer Science*, vol. 6, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[33] P. Cortez, *Wine Quality*, UCI Machine Learning Repository, 2009. [CrossRef] [Google Scholar] [Publisher Link]