BAREKENG: Journal of Mathematics and Its Applications
June 2026 Volume 20 Issue 2 Page 1263–1282
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol20iss2pp1263-1282

A NOVEL APPROACH TO SYMBOLIC DATA CLUSTERING USING ENHANCED K-MEANS ALGORITHM

Husty Serviana Husain[1], **Sapto Wahyu Indratno**[2*], **Sandy Vantika**[3]

[1,2,3]*Department of Mathematics, Faculty of Mathematics and Natural Sciences, Institut Teknologi Bandung*
Jln. Ganesa No.10, Lb. Siliwangi, Kec. Coblong, Kota Bandung, Jawa Barat, 40132, Indonesia

[1]*Study Program of Mathematics, Faculty of Mathematics and Sciences Education,*
Universitas Pendidikan Indonesia
Jln. Dr. Setiabudi No.229, Isola, Sukasari, Kota Bandung, Jawa Barat, 40154, Indonesia

*Corresponding author's e-mail: * **saptowi@itb.ac.id***

Article Info

Article History:
Received: 20th May 2025
Revised: 10th July 2025
Accepted: 23rd September 2025
Available online: 26th January 2026

Keywords:
Cluster image;
K-Means algorithm;
Represent features;
Symbolic data.

ABSTRACT

Clustering is a crucial technique in image analysis, yet traditional methods such as K-Means often struggle when dealing with complex, high-dimensional, or uncertain data. This limitation reduces their effectiveness in accurately grouping images, particularly when variability and overlapping features exist across categories. To address this problem, this paper introduces a novel approach that integrates symbolic data with the K-Means algorithm to cluster image data more effectively. By symbolically representing both color intensity and spatial features, we enhance the algorithm's ability to handle variability and uncertainty. We test our method on the CIFAR-10 dataset, where it achieves a clustering accuracy of 94.0% with an Adjusted Rand Index of 0.7, outperforming traditional methods such as K-Means (82.5%), DBSCAN (78.1%), and Hierarchical clustering (81.3%). Our results demonstrate that symbolic data analysis offers a more flexible and accurate solution for image clustering, with potential applications in fields such as medical image processing and environmental monitoring. Limitations and directions for future research are also discussed.

How to cite this article:

H. S, Husain, S. W. Indratno, and S. Vantika, "A NOVEL APPROACH TO SYMBOLIC DATA CLUSTERING USING ENHANCED K-MEANS ALGORITHM", *BAREKENG: J. Math. & App.*, vol. 20, no. 2, pp. 1263-1282, Jun, 2026.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · **Open Access**

1. INTRODUCTION

Clustering is widely used in various fields, including image processing, environmental monitoring, and medical data analysis. Traditional clustering methods, such as K-Means, focus on numerical data, which limits their effectiveness when handling complex data types, such as distributions, intervals, or sets of values [1], [2]. Symbolic Data Analysis (SDA) extends classical data analysis by allowing more complex data types to be analyzed [3], [4].

Symbolic data represent variability and uncertainty, often present in real-world applications. This makes SDA particularly useful in fields where data are noisy or incomplete, such as environmental monitoring and healthcare [3]. For example, symbolic objects encapsulating multiple values or distributions can represent pollution levels at different locations, capturing the variability in measurements over time [4]. In clustering, symbolic data have been shown to improve the accuracy and interpretability of the results by incorporating these complex data structures [5]. Unlike traditional methods that treat data points as fixed numerical values, SDA allows a more flexible representation, enabling the integration of numerical and categorical variables [3]. As a result, symbolic clustering methods have been applied in various domains, including medical image analysis, where the complexity of data is particularly pronounced [6].

Combining image processing and machine learning has led to many healthcare, security, and autonomous systems innovations. Advanced algorithms that accurately classify and group images are essential for building intelligent systems. However, working with image data is still challenging due to its complexity and high dimensionality [6], [7], [8]. As technology progresses, new techniques for extracting essential features from images are becoming more critical to make the data easier to understand. Methods like edge detection and spatial feature analysis have proven effective in revealing images' content [8], [9]. These methods improve image classification accuracy and help improve machine-learning models. Clustering algorithms, like K-Means, are also crucial in grouping images. Choosing the initial centroids is a key step, as it greatly affects the clustering result by guiding the algorithm towards more meaningful groups [1], [10]. In this research, we introduce symbolic data analysis methods. These methods use symbolic mathematical notations to handle complex data effectively [3]. Symbolic data analysis offers an intense way to manage variations within the data, providing deeper insights into the characteristics of the data.

The convergence of image processing and machine learning has led to significant innovations in healthcare, security, and autonomous systems, where accurate classification and clustering of images are critical. However, image data remain challenging due to their complexity and high dimensionality, which require advanced feature extraction techniques such as edge detection and spatial analysis. To address these challenges, this paper introduces a novel clustering approach that integrates symbolic data representation with the K-Means algorithm. By combining color intensity and spatial features within a symbolic framework, our method enhances the ability to manage variability and uncertainty in image data. Furthermore, a modified distance metric tailored for symbolic objects improves cluster separation and interpretability. This study, therefore, aims to provide a more robust and flexible solution for image clustering compared to traditional methods. The remainder of this paper is organized as follows: Section 2 reviews related works and highlights the research gap; Section 3 presents the proposed methodology, including feature extraction, symbolic representation, and clustering; Section 4 discusses the experimental setup and results; and Section 5 concludes the paper and suggests future research directions.

2. RESEARCH METHODS

2.1 Classical Clustering Algorithms and Their Limitations

Clustering algorithms such as K-Means, DBSCAN, and hierarchical clustering remain popular due to their simplicity and efficiency. However, they largely assume data in fixed numerical form, limiting their effectiveness when facing variability, uncertainty, or complex data structures. For example, enhancements like better initialization [11] or custom distance functions [2] help but still rely on strict numerical representations.

Symbolic Data Analysis (SDA) addresses these limitations by enabling data representation as intervals, distributions, or sets [3]. While promising, symbolic clustering has rarely been applied to image data, mostly focusing on fields like environmental monitoring or aggregated statistics. This leaves a research gap in leveraging symbolic representation coupled with image feature extraction.

Recent advances in image clustering further demonstrate evolving approaches that explore multi-modal inputs and supervision from captions to external knowledge. Below is a comparison of significant works:

Table 1. Summary of Prior Studies

Author & Year	Dataset / Domain	Method & Highlights	Strengths	Limitations / Gap in the Context of Our Study
Hartigan & Wong (1979)	Numerical data	Standard K-Means algorithm [1]	Simple, efficient	Limited to purely numeric data
Celebi (2011)	Color images	Improved centroid initialization [2]	Better convergence in color quantization	No symbolic data usage; numeric only
Fränti & Sieranoja (2019)	Synthetic, real-world	Enhanced initialization K-Means [11]	Higher accuracy, more robust	Still assumes numerical representations
Billard & Diday (2007)	Various	Symbolic Data Analysis foundational [3]	Supports intervals, distributions	Not tailored to image data or symbolic image features
Peng & Li (2023)	General image datasets	Deep clustering with mutual information across views [12]	Reduces intra-class variability via augmented views	Focused on deep learning, not symbolic representation
Li *et al.* (2023) – TAC	ImageNet etc.	Text-Aided Clustering (TAC): uses WordNet semantics as external guidance [13]	Leverages external semantic info for better clustering	Language-based external knowledge, not symbolic feature structure
Stephan *et al.* (2024)	Diverse image datasets	Text-Guided Image Clustering: uses generated captions & VQA prompts [14]	Multimodal embeddings improve interpretability and clustering results	Captions-based, not using symbolic feature representation
Raya *et al.* (2024)	Multimodal clustering	Deep learning for multi-modal data clustering [15]	Integrates multimodal embeddings	Leaned toward deep learning, not symbolic data representation
Hu *et al.* (2024) – ICBPL	CIFAR-10, STL-10, others	Self-supervised clustering with pretrained models and latent distribution optimization [16]	Strong performance using latent features	Not symbolic; more focused on deep feature distribution dynamics
Wu (2024) – RD-FKC	Various image datasets	Robust Deep Fuzzy K-Means Clustering [17]	Embedding robustness and fuzzy logic for clustering	Deep learning; not symbolic
Our Work (2025)	CIFAR-10, Flower datasets	Symbolic K-Means with modified distance metric	Handles variability, uncertainty, interpretable	New in image domain symbolic clustering; bridges a gap

2.2 Foundations of Symbolic Data Clustering

Symbolic Data Analysis (SDA) extends classical data analysis techniques by representing complex data structures, such as intervals, distributions, or sets, instead of single numerical values [3]. This approach allows SDA to manage uncertainty, variability, and multivalued attributes, which are common in real-world applications. Several studies have proposed methodologies for clustering symbolic data, including the extension of traditional algorithms like K-Means and DBSCAN [2], [4].

In SDA, the centroid of a symbolic object, such as an interval or distribution, differs from the classical centroid used in traditional clustering. For an interval-valued object, the centroid is defined as the middle of the interval:

$$C = \frac{l+u}{2}, \tag{1}$$

where l and u are the lower and upper bounds of the interval, respectively [3]. This centroid calculation allows symbolic clustering algorithms to capture the variability within the data.

Recent advances have focused on improving the efficiency and scalability of SDA-based clustering methods. These include modifications to the distance metrics used for symbolic objects, such as the Hausdorff distance, which measures the dissimilarity between two sets or intervals [3]. The Hausdorff distance between two intervals $[l_1, u_1]$ and $[l_2, u_2]$ is given by:

$$d_H\,([l_1, u_1], [l_2, u_2]) = max(|l_1 - l_2|, |u_1 - u_2|). \tag{2}$$

This metric is critical in comparing symbolic objects, considering the entire range of possible values, not just individual points.

2.3 Overview of the Framework

The proposed framework consists of several stages:

1. Image Pre-processing: Resize images and extract RGB color intensities.
2. Feature Extraction: Compute symbolic representation of color intensity (R, G, B) and spatial features (edges, positions, lengths).
3. Symbolic Data Conversion: Transform features into symbolic objects (intervals/distributions).
4. Initial Centroid Selection: Initialize centroids using predefined categories.
5. Clustering with Modified Minkowski Distance: Perform K-Means adapted for symbolic data.
6. Evaluation: Measure clustering performance using Accuracy, Adjusted Rand Index, and runtime.

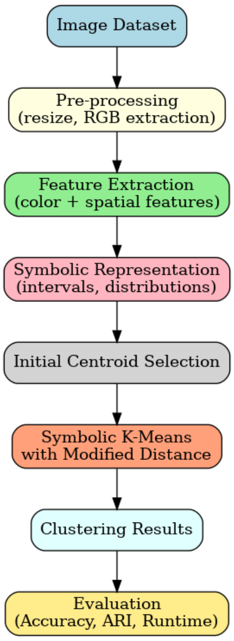


Image Dataset
Pre-processing (resize, RGB extraction)
Feature Extraction (color + spatial features)
Symbolic Representation (intervals, distributions)
Initial Centroid Selection
Symbolic K-Means with Modified Distance
Clustering Results
Evaluation (Accuracy, ARI, Runtime)

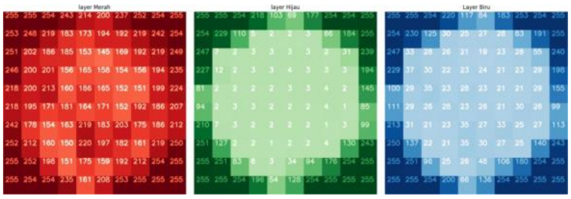
Figure 1. Flowchart of the Proposed Methodology Integrating Symbolic Data Representation with the K-Means Algorithm

2.4 Pre-processing Images

In this research, we collected random digital images of flowers from different online sources. Each color image was created by combining red, green, and blue images, each with 8-bit depth, giving 24 bits. The color data is stored in three matrices: one for red, one for green, and one for blue, known as the RGB (Red, Green, Blue) matrix. Let O represent the images $o_1, o_2, \dots, o_n$. First, we resized all images to the same size, for example, $p \times q$ pixels (such as 50×50). Then, we wrote the image in a matrix with p rows and q columns as follows:

$$J_{p\times q} = \begin{bmatrix} j_{11} & j_{12} & \cdots & j_{1p} \\ j_{21} & j_{22} & \cdots & j_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ j_{p1} & j_{p2} & \cdots & j_{pq} \end{bmatrix} \tag{3}$$

Here, J represents the red, green, and blue channels, each with dimensions $p \times q$.



layer Merah

255	255	254	243	214	200	237	252	254	255
253	248	219	183	173	194	192	219	242	254
251	202	186	185	153	145	169	192	219	249
246	200	201	156	165	158	154	156	194	235
218	200	213	160	186	165	152	151	199	224
218	195	171	181	164	171	152	192	186	207
242	178	154	163	219	183	203	175	186	212
252	212	160	150	220	197	182	161	219	250
255	252	198	151	175	159	192	212	254	255
255	254	254	235	161	208	253	255	255	255

layer Hijau

255	255	254	218	103	69	177	254	254	255
254	229	110	6	2	2	3	66	184	255
247	7	2	3	3	3	3	2	31	239
227	12	2	3	3	4	3	3	3	194
81	2	2	2	2	3	3	4	2	145
94	2	3	3	2	2	2	4	1	85
210	7	3	2	2	2	2	1	3	99
251	127	2	2	1	2	2	4	130	243
255	251	83	6	3	34	94	176	254	255
255	255	254	198	54	128	254	255	255	255

Layer Biru

255	255	254	220	117	84	183	253	254	255
254	230	125	30	25	27	28	83	191	255
247	33	28	26	21	19	23	28	55	240
229	37	30	22	23	24	21	23	29	198
100	29	35	23	20	23	21	21	29	155
111	29	26	28	23	26	21	30	26	99
213	31	21	23	35	27	33	25	27	113
250	137	22	21	35	30	27	25	140	243
255	251	96	25	28	48	106	180	254	255
255	255	254	200	66	136	254	255	255	255

Figure 2. Illustration of the RGB Color Channels Feature Extraction in the image clustering

As shown in Fig. 2, the RGB color channels are utilized in the feature extraction process to represent image pixels symbolically. This representation aids in clustering by capturing color intensities.

2.5 Definition of Symbolic Data

Symbolic data analysis (SDA) extends traditional data analysis that handles more complex data types, such as intervals, multiple values, and distributions. These data types allow for representing variability and uncertainty, often found in real-world situations [3].

Symbolic data are generalized data units capable of encapsulating multiple values, ranges, or distributions for a single variable. This extension allows for representing uncertainty and variability inherent in real-world applications [3]. For example, a symbolic object can represent pollution levels as an interval, capturing the variability over time. Fig. 3 illustrates a symbolic object representation.

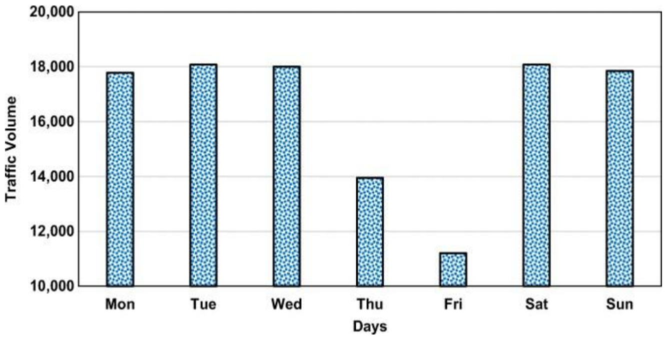


Figure 3. Diagram of a Symbolic Object Representing Pollution Levels.

A symbolic object is a structured data unit that characterizes a collection of individuals or entities using a combination of intervals, sets, or distributions across one or more variables. Symbolic objects facilitate encapsulating relationships, hierarchies, and other complex data structures [3]. To illustrate the practical application of symbolic data analysis, consider the case study of environmental monitoring:

Consider the monitoring of pollution levels across a city. The data collected from various stations might include pollutant concentrations such as NO2, SO2, PM2.5, and CO. These pollutants are measured at irregular intervals and can be affected by factors such as weather conditions, station malfunction, or local environmental changes [18]. Symbolic clustering can represent these data points as intervals, capturing the variability of pollutant levels over time and across different stations.

For example, the concentration of NO2 at a particular station may range from 40 to 80 µg/m3, representing an interval [40, 80] in SDA. This allows the clustering algorithm to account for the uncertainty and variability in the data, resulting in more accurate identification of pollution hotspots [19]. Moreover, symbolic data can better model the spatial-temporal distribution of pollutants, which is crucial for environmental policymakers when deciding on mitigation strategies [20]. This approach can thus improve the understanding of how pollution evolves in space, leading to more informed decisions in urban planning and public health.

2.6 Comparison to Other Clustering Methods

Symbolic data-based clustering offers several advantages over traditional clustering methods, particularly in handling variability and complex data types. Below, we compare symbolic clustering and other well-known methods such as K-Means, DBSCAN, and hierarchical clustering.

2.6.1 K-Means Clustering

K-Means clustering is widely used in many fields due to its simplicity and efficiency [1]. However, K-Means operates on fixed numerical values, which limits its flexibility when dealing with complex or variable data. Variable data refers to data that can take on different values. In the context of data analysis, “variable data” can be categorized into:

1. Multivariable data: This refers to datasets that contain more than one variable or feature, each representing a different aspect of the observed phenomenon. For instance, in a medical dataset, variables could include patient age, weight, height, and blood pressure. Each variable contributes to the overall analysis, and their interactions may reveal deeper insights.
2. Multidimensional data: This refers to data spread across multiple dimensions, often visualized as points in a high-dimensional space. For example, image data can be represented in multiple dimensions, where each dimension corresponds to different features such as color channels (Red, Green, Blue), spatial location, and texture. Clustering multidimensional data often requires advanced techniques that can capture the complexity of the data across various dimensions.

These types of data require specialized analysis methods, as traditional approaches may struggle to capture the full complexity of the interactions between variables or dimensions. In contrast, symbolic data analysis (SDA) allows data representation as intervals or distributions, providing a more nuanced approach in environments with high variability, such as medical or environmental data [3]. For example, in symbolic K-Means, data points are represented as intervals, making it more adaptable to noise and uncertainty compared to traditional K-Means [4].

2.6.2 DBSCAN

DBSCAN is effective in handling noise and discovering clusters of arbitrary shape [21]. However, DBSCAN is limited in its ability to process symbolic data, which can include intervals or distributions. Symbolic K-Means, on the other hand, can represent and cluster more complex data types by capturing the variability and uncertainty inherent in real-world datasets. This makes symbolic clustering more versatile in applications where the data exhibit heterogeneity [5].

2.6.3 Hierarchical Clustering

Hierarchical clustering is particularly useful when the number of clusters is unknown, as it does not require a predefined number of clusters [22]. However, like K-Means, it typically operates on fixed numerical data. Symbolic clustering extends this by handling complex data structures, such as multi-valued or interval data, providing a more flexible framework for clustering in uncertain environments [3]. While hierarchical clustering excels at visualizing relationships between clusters, symbolic data analysis improves interpretability by allowing for richer data representations [5].

2.7 Mathematical Model for Symbolic Data

Symbolic data allows for the representation of more complex data structures than traditional statistical data. This section introduces a simplified mathematical model to describe symbolic data, focusing on interval data, one of the most common symbolic data types.

Interval data for a variable includes the lower and upper limits, indicating the range in which the values of this variable for a specific entity are situated. Formally, for a variable X, an interval is expressed as $[X_{min}, X_{max}]$, where X_{min} and X_{max} represent the minimum and maximum values of X, respectively [3]. This approach can be extended to represent more complex symbolic data types, such as multi- valued variables and distributions. For instance, a multi-valued variable could be represented as a set of discrete values or categories, and a distribution could be described by its parameters, such as mean and variance for normal distributions. The representation and manipulation of symbolic data require specific mathematical and computational models. For interval data, operations such as the computation of the interval mean or the interval distance can be defined to facilitate analysis. For example, the mean of an interval $[X_{min}, X_{max}]$ is given by $\frac{X_{min}+X_{max}}{2}$. Symbolic data analysis provides a rich framework for dealing with heterogeneous and complex data, enabling more nuanced and comprehensive analyses in various fields, from environmental science to marketing research [16].

Cartesian Join is a method to merge features from two distinct categories of images; the Cartesian join operation, as defined in [3], is utilized. The Cartesian join $A \oplus B$ between two sets A and B is their componentwise union, defined as:

$$A \oplus B = (A_1 \oplus B_1, \dots, A_p \oplus B_p), \tag{4}$$

where $A_j \oplus B_j =$ '$A_j \cup B_j$'. When A and B are multi-valued objects with $A_j = \{a_{j1}, \dots, a_{js_j}\}$ and $B_j = \{b_{j1}, \dots, b_{t_j}\}$, then

$$A_j \oplus B_j = \left\{a_{j1}, \dots, a_{js_j}, b_{j1}, \dots, b_{t_j}\right\} \tag{5}$$

is the set of values in A_j, B_j, or both. When A and B are interval-valued objects with

$A_j = [a_j^A, b_j^A]$ and $B_j = [a_j^B, b_j^B]$, then

$$A_j \oplus B_j = [min(a_j^A, a_j^B), max(b_j^A, b_j^B)]. \tag{6}$$

This operation is particularly useful when the features of the images are interval-valued, allowing us to construct a symbolic object that combines the features of two image categories in a more informative way than simple aggregation.

2.8 Feature Extraction

Feature extraction in image processing entails obtaining essential image attributes, including color intensity and edge properties. These features are vital for numerous applications such as image classification and pattern recognition.

2.8.1 Average Intensity Colour Image

The average intensity color image values for each image in a dataset offer insights into the dominant color tones. For the k-th image in a collection of n images, the average values for the Red (R), Green (G), and Blue (B) channels are calculated as follows:

$$Avg_{R_o} = \frac{1}{p_o \times q_o} \sum_{i=1}^{p_o} \sum_{j=1}^{q_o} R_{ijo}, \tag{7}$$

$$Avg_{G_o} = \frac{1}{p_o \times q_o} \sum_{i=1}^{p_o} \sum_{j=1}^{q_o} G_{ijo}, \tag{8}$$

$$Avg_{B_o} = \frac{1}{p_o \times q_o} \sum_{i=1}^{p_o} \sum_{j=1}^{q_o} B_{ijo}, \tag{9}$$

where p_o and q_o denote the dimensions of the o-th image, and R_{ijo}, G_{ijo}, and B_{ijo} represent the intensity colour image values at the pixel located at (i, j) [23].

2.8.2 Min Max Intensity Colour Image

Beyond average intensity color image values, the minimum and maximum values for each color channel provide additional insight into the color range and contrast within each image:

$$Min_{R_o} = \min_{i,j} R_{ijo} \ , Max_{R_o} = \max_{i,j} R_{ijo} \ , \tag{10}$$

$$Min_{G_o} = \min_{i,j} G_{ijo} \ , Max_{G_o} = \max_{i,j} G_{ijo} \ , \tag{11}$$

$$Min_{B_o} = \min_{i,j} B_{ijo} \ , Max_{B_o} = \max_{i,j} B_{ijo} \ . \tag{12}$$

These calculations help understand the dynamic range and variability of colors present in the image dataset.

2.8.3 Definition of an Edge

Edges are significant changes in intensity in an image and are crucial for understanding the structure and features within an image. The Canny edge detection algorithm is commonly used for identifying edges due to its effectiveness in minimizing error rates and noise. The Canny Edge detection algorithm was chosen due to its ability to minimize error rates and its robustness against noise. It offers an optimal balance between detecting edges and preserving important structural information, crucial for accurate clustering in image analysis [3].

$$E = \{(x,y) \big| Canny(I_{gray})(x,y) > \ threshold\}. \tag{13}$$

Here, E represents the set of edge points detected in a grayscale image I_{gray}, with (x,y) denoting the coordinates of an edge point [3].

2.8.4 Average Position

The spatial distribution of edges within an image is captured by calculating the average position of detected edges, which indicates the structural alignment and composition of the image's features.

$$Avg_{Position} = \frac{1}{L} \sum_{o=1}^{L} Position(o) \ . \tag{14}$$

This equation averages the positions of L detected edges within an image, providing a single vector that represents the central tendency of edge locations within the image.

2.8.5 Min Max

For each image I_k, after applying the Canny edge detector, the minimum and maximum positions of the detected edges in both x and y directions are calculated by:

$$Min_x^o = \min_{(\mathrm{x},\_)\in \mathrm{E_o}} x \ , Max_x^o = \max_{(\mathrm{x},\_)\in \mathrm{E_o}} x \ , \tag{15}$$

$$Min_y^o = \min_{(\mathrm{y},\_)\in \mathrm{E_o}} y \ , Max_y^o = \max_{(\mathrm{y},\_)\in \mathrm{E_o}} y \ , \tag{16}$$

where E_o is the set of edge points detected in the o-th image using the Canny edge detector, with each edge point represented as a coordinate pair (x,y).

2.8.6 Length

Given an image, the Canny edge detection algorithm identifies its edges. For any detected edge, its positions along the X and Y axes are denoted by xpositions and ypositions, respectively. The length of these edges along the X and Y axes can be defined as:

$$Length_x^o = \max(x_{positions})^o \ - \min(x_{positions})^o \ , \tag{17}$$

$$Length_y^o = \max(y_{positions})^o \ - \min(y_{positions})^o \ , \tag{18}$$

where:

1. $x_{positions}$ and $y_{positions}$ are the sets of X and Y coordinates of the edge points detected by the Canny algorithm.
2. $max(\cdot)$ and $min(\cdot)$ represent the maximum and minimum functions, respectively.

The $Length_X$ and $Length_Y$ represent the span of the detected edges along the X and Y axes, providing a measure of the object's size within the image.

2.8.7 Convert Images to Symbolic Data

Converting images to symbolic data involves representing the image attributes using symbolic objects. This enables the encapsulation of multiple values, ranges, or distributions for a single variable, providing a more comprehensive representation of the image data. The following steps outline the conversion process:

Step 1: Extract Features. Firstly, the relevant features from the image, such as the average color intensities for the Red (R), Green (G), and Blue (B) channels, and spatial features including the positions and lengths of detected edges, are extracted. Let V_o represent the feature vector for image o:

$$Vo = \begin{bmatrix} Avg_R^o\ , Avg_G^o\ , Avg_B^o\ , Length_x^o\ , Length_y^o, Avg_x^o\ , Avg_y^o, \\ Min_R^o\ , Min_G^o\ , Min_B^o, Max_R^o\ , Max_G^o\ , Max_B^o\ , \\ Min_x^o\ , Max_x^o\ , Min_y^o, Max_y^o \end{bmatrix}. \tag{19}$$

Step 2: Define Intervals for Each Feature. For each feature in the vector V_o, define the lower and upper bounds, creating an interval that represents the range of values for that feature. For instance, for the average red channel intensity Avg_R, the interval can be defined as:

$$Interval_R = [Min_R, Max_R]. \tag{20}$$

Step 3: Create Symbolic Object. Using the intervals defined for each feature, construct a symbolic object that encapsulates all the intervals. The symbolic object for image o is denoted as V_o and is defined as:

$$Vo = \left\{ \begin{array}{c} [Min_R, Max_R], [Min_G, Max_G], [Min_B, Max_B], \\ [Length_x^{min}, [Length_x^{max}], [Length_y^{min}, [Length_y^{max}, \\ \left[Avg_x^{min}, Avg_x^{max}\right], \left[Avg_y^{min}, Avg_y^{max}\right] \end{array} \right\}. \tag{21}$$

Step 4: Use Symbolic Data in Clustering. The symbolic objects V_o are then utilized in the clustering algorithm. The distance measure for clustering symbolic data must account for the interval nature of the features. The generalized distance between two symbolic objects V_o and V_p can be defined using a modified Minkowski distance:

$$d_q = \left(V_o, V_p = \sum_{j=1}^{p} w_j . \left| \frac{Min_{o,j} + Max_{o,j}}{2} - \frac{Min_{p,j} + Max_{p,j}}{2} \right|^q \right)^{\frac{1}{q}}, \tag{22}$$

where w_j is the weight for the j-th feature, $Min_{o,j}$ and $Max_{o,j}$ are the interval bounds for the j-th feature of a symbolic object V_o, and similarly for V_p. Following these steps, images are effectively converted to symbolic data, enabling more sophisticated analysis and clustering those accounts for the inherent variability and complexity of image features.

2.9 Initial Centroid

In the K-Means clustering algorithm, the initial centroids C_l) which $l = 1, 2, \ldots, s$, which merges the feature sets from distinct image categories, such as roses and sunflowers, for each image O, the initial centroid is applied to the extracted features to form a composite feature set:

$$C_{rose} = V_{rose}(o)\ , C_{sunflower} = V_{sunflower}(o), \ldots Cl = V_l(o) \tag{23}$$

where $V_{rose(o)}$ and $V_{sunflower(o)}$ are the feature vectors for the rose and sunflower categories, respectively.

Each composite feature vector includes:

1. Composite average intensity color image values $Avg_{intensity}^{(R \oplus S)_o}$.

2. Composite minimum and maximum intensity color image values $Min_{intensity}^{(R\oplus S)_o}$, $Max_{intensity}^{(R\oplus S)_o}$.
3. Composite average, minimum, and maximum spatial positions of edges $Avg_x^{(R\oplus S)_o}$, $Min_x^{(R\oplus S)_o}$, $Max_x^{(R\oplus S)_o}$, and so on for y coordinates.
4. Composite length of edges $Length_x^{(R\oplus S)_o}$, $Length_y^{(R\oplus S)_o}$.

Accordingly, the initial centroid for a category C comprising both roses and sunflowers is computed as the mean of the combined feature vectors:

$$C_c = \frac{1}{n}\sum_{o=1}^{n} V(o)\,. \tag{24}$$

This consolidated centroid C_c provides a robust foundation for the clustering process, capturing the essence of both roses and sunflowers in the feature space.

2.10 Distance Measures in Clustering

In clustering, the choice of distance measure is crucial for determining the similarity between data points. K-Means typically uses the Euclidean distance, a particular case of the generalized Minkowski distance, when the order $q = 2$.

The Generalized Minkowski distance of order $q \geq 1$ between two sets A and B is defined as:

$$d_q(A,B) = \left(\sum_{j=1}^{p} w_j^* |\phi j(A,B)|^q\right)^{\frac{1}{q}}, \tag{25}$$

where w_j is an appropriate weight for the distance component $\phi_j(A,B)$ on Y_j, for $j = 1,\dots,p$.

2.11 Differences from Existing Methods

While traditional clustering methods like K-Means rely on precise numerical data and Euclidean distances, our approach extends these methods to symbolic data by modifying the centroid and distance calculations. For instance, instead of using the standard Euclidean distance, we implement a modified Minkowski distance for symbolic objects, defined as:

$$d(x,y) = \left(\sum_{j=1}^{n} |x_i - y_i|^p\right)^{\frac{1}{p}}, \tag{26}$$

where x and y are symbolic objects, and p controls the degree of generalization [2]. When $p = 2$, this becomes the symbolic version of the Euclidean distance.

Our method diverges from traditional clustering by considering the inherent uncertainty in symbolic data. This uncertainty is often reflected in the data's multivalued or interval nature, making it crucial to adapt the distance measures and centroid calculations. These adaptations allow our approach to outperform traditional methods, as evidenced by the higher accuracy rates achieved in our CIFAR-10 experiments (discussed in Section 3).

2.12 Experiment Setup and Testing

The proposed method was tested on three datasets:

1. Flowers dataset: 317 images of roses and sunflowers collected online.
2. CIFAR-10: 60,000 color images across 10 categories.

The experiments were implemented in Python, using NumPy [24], OpenCV [25], and scikit-learn [26] for image processing and clustering. Symbolic data functions were implemented manually to handle interval-valued objects.

1. Evaluation setup:

   Each dataset was clustered into its known categories (2 for Flowers and 10 clusters for CIFAR-10).

2. Performance metrics included:

   a. Clustering Accuracy (%): ratio of correctly clustered images.
   b. Adjusted Rand Index (ARI): measures clustering similarity against ground truth.
   c. Computation Time (seconds): runtime efficiency.

The results (detailed in Section 4) demonstrate that the proposed symbolic K-Means consistently outperforms traditional clustering approaches in accuracy and interpretability.

3. RESULTS AND DISCUSSION

This section presents the outcomes of the clustering analysis, focusing on the distribution of intensity colour image and spatial features across the identified clusters.

3.1 Clustering Outcomes

To evaluate the effectiveness of the proposed clustering method, we first applied it to the flower dataset consisting of rose and sunflower images. The confusion matrix in Table 2 summarizes the clustering results for these two categories.

Table 2: Confusion Matrix of Rose and Sunflower

	Predicted Rose	**Predicted Sunflower**
Actual Rose	31	1
Actual Sunflower	3	20

As seen in Table 2, the model accurately distinguishes between rose and sunflower images, with an overall accuracy of 96% for roses and 87% for sunflowers. The misclassifications suggest slight confusion between visually similar classes, such as sunflower images classified as roses.

Integrating symbolic data analysis techniques provided a deeper understanding of the clustering results. Symbolic objects and descriptors allowed for sophisticated mathematical modeling and interpretation, revealing patterns and relationships not immediately evident in the raw data. This approach facilitated a more nuanced analysis, particularly in handling the qualitative aspects of the image data.

3.2 Centroid Analysis

In our study, we have utilized an extensive set of images of sunflowers and roses, which have been meticulously annotated and categorized. This rich dataset serves as a foundational element for our image classification tasks. The sunflower images, consisting of 146 labeled photographs, display a wide range of angles, lighting conditions, and backgrounds, capturing the vibrant aesthetics of these helianthuses. Similarly, the rose dataset comprises 171 annotated images that showcase the nuanced pigmentation and form of rose species.

The image datasets for this study were curated from a niche repository geared toward the classification of floral imagery. The respective datasets for sunflowers and roses can be found online: Sunflower dataset and Rose dataset. The diversity and quality of these datasets are instrumental in developing robust machine-learning models capable of distinguishing between the intricate features of sunflower and rose images.

Figure 4. Sample Image, the Sunflower Dataset

Download **171 Rose** labeled image dataset
This dataset contain **171** images of **annotated Rose images**

Figure 5. Sample Image, the Rose Dataset

The dataset used in this study includes images of sunflowers and roses. As shown in Fig. 4, the sunflower dataset consists of high-resolution images capturing various angles and lighting conditions. Similarly, Fig. 5 presents a sample image from the rose dataset, exhibiting diversity in terms of color and structure.

The following Table 3 presents the initial and final centroids for two categories of objects: roses and sunflowers. The centroids are characterized by their average Red (R), Green (G), and Blue (B) color intensities and their average spatial positions (X, Y).

Table 3. Centroid Values for Rose and Sunflower Categories with Improved Readability

Category	Avg R	Avg G	Avg B	Length X	Length Y	Avg X	Avg Y	Min R	Min G	Min B	Max R	Max G	Max B	Min X	Max X	Min Y
Initial Rose	118.7	91.4	70	449.3	483.6	4.3	3.1	0.1	252.7	232.1	224.5	903.8	1009.1	1.1	904.9	0.8
Initial Sunflower	137.3	729.9	76.9	31.3	32.8	13.2	13.4	0.7	25.1	238.8	211.6	61.5	61.6	0.6	62.1	1.2
Final Rose	94.1	77.3	53.8	31	33.6	4.4	4	0.6	250.5	214.7	196.1	60.1	61	1.1	61.2	1.8
Final Sunflower	160.2	142.4	103	31.7	32.8	21.8	15.2	0.7	252.3	241.5	235.8	60.5	61.6	1.3	61.8	1.1

Category	Max Y
Initial Rose	1009.9
Initial Sunflower	62.8
Final Rose	62.8
Final Sunflower	62.7

The results in Table 3 provide centroid values for the Rose and Sunflower categories, comparing their initial and final states after the clustering process. Below is a detailed interpretation of the observed discrepancies and deviations.

The initial centroid values for the rose category in the red, green, and blue channels (Avg R = 118.7, Avg G = 91.4, Avg B = 70.0) suggest a broader range of color variations. After clustering, the final centroid values (Avg R = 94.1, Avg G = 77.3, Avg B = 53.8) indicate that the algorithm refined the rose cluster to include images with lower color intensity and more consistent shades. For sunflowers, there is an opposite trend where the final centroid values increase (Avg R = 160.2, Avg G = 142.4, Avg B = 103.0), suggesting that the algorithm focused on more vibrant sunflower images, likely separating images with lower intensity from the core sunflower cluster. The spatial dimensions of the rose cluster show a significant reduction from the initial state (Length X = 449.3, Length Y = 483.6) to the final state (Length X = 31.0, Length Y = 33.6). This significant decrease suggests that the initial cluster included a wide variety of rose images in size and orientation, but the final cluster consists of more uniform images.

On the other hand, the sunflower cluster remains relatively stable in terms of its spatial dimensions (Length X = 31.3 and Length Y = 32.8 initially, and Length X = 31.7 and Length Y = 32.8 finally), indicating that the sunflower images were more homogeneous from the beginning, and the algorithm made only minor adjustments. The maximum values for the red, green, and blue channels (Max R, Max G, Max B) also show significant reductions in the rose cluster. Initially, Max R = 224.5, Max G = 903.8, and Max B = 1009.1, which drop to Max R = 196.1, Max G = 60.1, and Max B = 61.0 in the final cluster. This reduction suggests that the algorithm removed extreme outliers or images with unusually high color intensity from the final cluster. The sunflower cluster, however, shows very little change in these metrics, implying that the initial and final clusters were already homogeneous in terms of color intensity.

The spatial dimensions, particularly the maximum and minimum coordinates (Max X, Max Y, Min X, Min Y), reflect a similar pattern. For the rose category, the initial values (Max X = 904.9, Max Y = 1009.9) indicate a wide variety of image sizes, while the final values (Max X = 61.2, Max Y = 62.8) suggest that the algorithm refined the cluster to include only smaller, more consistent images. In contrast, the sunflower cluster remains essentially unchanged, with minor deviations in Max X and Max Y, reinforcing the idea that the sunflower cluster was already more consistent in size. The visual representation of these centroids, as depicted in the bar graphs, can provide further insight into the clustering behavior and the distinct characteristics of each category.

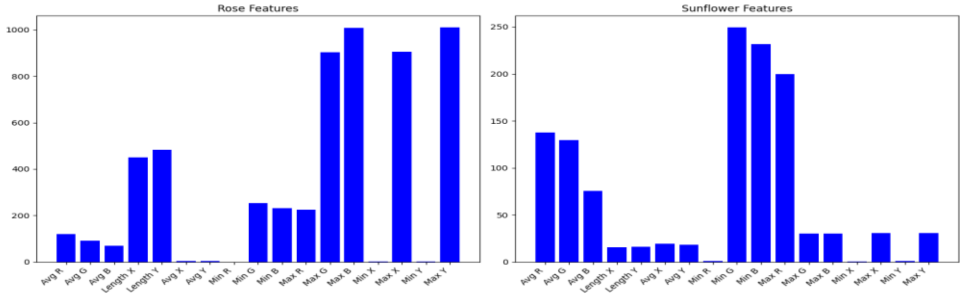


Figure 6. Centroid Comparisons between Initial Stages for Roses and Sunflowers.

The visual representation of these centroids, as depicted in the bar graphs in Fig. 6, provides further insight into the clustering behavior and highlights the distinct characteristics of each category. The differences in average color intensity values and spatial positions between the initial and final stages for roses and sunflowers are clearly visible, illustrating how the clustering algorithm refines each category.

From Table 3, it is evident that the final centroids for roses have higher intensity colour image values and lower spatial values, suggesting denser and more vivid coloration at the core of the cluster. Conversely, sunflowers show lower intensity colour image values in the final centroids, which could indicate a more diverse range of colors and a slight shift in the spatial domain. These variations between the initial and final centroids highlight the adaptive nature of K-Means clustering in response to the data's distribution.

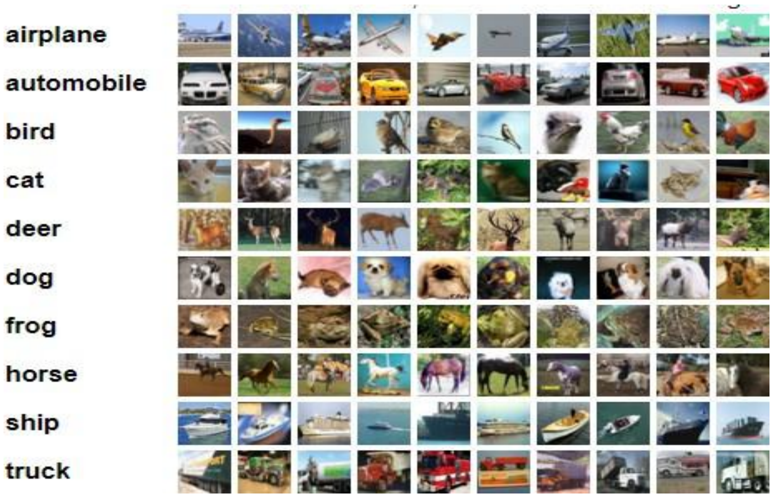


Figure 7. CIFAR 10 class images.

The CIFAR-10 dataset, which contains images from various classes, is visually represented in Fig. 7. This dataset serves as a standard benchmark for evaluating image classification and clustering algorithms, providing a diverse set of images across multiple categories. The CIFAR-10 dataset is renowned for its extensive use in benchmarking image classification algorithms [6]. This dataset, introduced by Krizhevsky *et al.*, comprises 60,000 32x32 color images distributed across 10 distinct classes [13]. In our study, we employed the CIFAR-10 images to verify the accuracy of our clustering methodology. The table below presents the initial and final centroid values for various categories within the CIFAR-10 dataset. CIFAR-10 consists of ten categories: truck, airplane, automobile, bird, cat, deer, dog, frog, horse, and ship. The centroid values are represented by the red (R), green (G), and blue (B) color components as well as spatial coordinates (X and Y).

The Table 4, Table 5, and Table 6 showcase the initial and final centroids for each category, providing insights into the average values for each category before and after the clustering process. The "Avg" columns display the average values of the color components and spatial coordinates, while the "Min" and "Max" columns indicate the minimum and maximum values of these components within the dataset. The observed changes between the initial and final centroids reflect how the average values for each category have shifted throughout the clustering process.

Table 4. Initial and Final Centroids of CIFAR-10

Category	Avg R	Avg G	Avg B	Min R	Max R	Min G	Max G
Initial Centroids							
Truck	127.2	123.8	121.9	10.2	245.1	30.9	245.3
Airplane	134.0	142.9	150.2	24.2	235.9	29.3	232.4
Automobile	120.2	115.9	114.0	8.1	242.0	30.9	244.3
Bird	124.8	125.3	108.1	16.7	216.9	29.2	228.3
Cat	126.4	116.4	106.0	12.8	228.1	30.2	235.9
Deer	120.3	118.6	96.4	17.9	202.6	30.6	226.5
Dog	127.5	118.5	106.2	13.2	229.5	30.2	238.5
Frog	119.9	111.8	88.0	10.2	201.8	30.5	228.9
Horse	128.0	122.4	106.3	14.0	234.9	30.8	239.9
Ship	125.0	134.0	141.4	21.9	237.6	30.7	237.1
Final Centroids							
Truck	120.6	118.1	111.4	9.6	247.3	30.8	248.2
Airplane	188.6	189.8	189.6	28.6	250.9	28.9	249.4

Category	Avg R	Avg G	Avg B	Min R	Max R	Min G	Max G
Final Centroids							
Automobile	90.1	85.6	77.1	5.1	239.2	30.5	244.2
Bird	121.8	143.5	165.1	43.4	225.2	27.9	202.1
Cat	112.4	109.6	98.6	8.8	210.6	30.6	221.5
Deer	112.3	113.5	97.3	20.5	164.3	28.5	180.9
Dog	141.8	129.9	98.0	22.5	192.1	30.3	231.2
Frog	85.4	75.9	55.2	5.8	160.1	30.1	204.7
Horse	146.9	144.7	140	41.5	244.5	30.5	246.6
Ship	149.1	149.0	148.2	9.9	247.8	30.7	247.6

Table 5. Initial and Final Centroids of CIFAR-10

Category	Min B	Max B	Avg X	Avg Y	Min X	Max X
Initial Centroids						
Truck	29.0	244.0	10.8	10.6	0.1	31
Airplane	24.5	233.0	20.9	23.1	0.9	30.2
Automobile	28.7	241.8	8.5	8.2	0.1	30.9
Bird	29.0	223.5	24.5	23.4	0.9	30.1
Cat	30.0	231.5	18.6	16.0	0.4	30.6
Deer	30.0	217.5	26.0	24.4	0.2	30.8
Dog	30.1	233.4	18.3	16.1	0.4	30.6
Frog	30.1	220.4	19.6	16.9	0.3	30.8
Horse	29.9	237.3	19.3	17.2	0.1	30.9
Ship	25.8	236.6	17.8	20.4	0.2	30.8
Final Centroids						
Truck	29.7	247.4	11.1	10.8	0.1	30.9
Airplane	25.3	250.1	33.3	30.2	1.1	30
Automobile	29.7	242.1	6.4	6.6	0.2	30.7
Bird	22.5	210	29.1	37.9	1.5	29.5
Cat	29.3	215.2	11.1	11	0.2	30.8
Deer	26.9	174.7	27.9	27.3	1.2	29.7
Dog	29.3	216.2	42.2	35.5	0.4	30.6
Frog	29.4	184.9	10.9	10.3	0.4	30.5
Horse	28.5	245.3	46.1	44.9	0.3	30.8
Ship	28.1	247.2	10.8	9.9	0.2	30.8

Table 6. Initial and Final Centroids of CIFAR-10

Category	Min Y	Max Y	Length X	Length Y
Initial Centroids				
Truck	1.1	30.1	15.5	16.0
Airplane	3.8	28.3	15.5	16.6
Automobile	1.2	29.8	15.5	15.4
Bird	1.6	30.6	15.5	16.5
Cat	0.5	30.6	15.5	15.8
Deer	0.6	30.6	15.5	15.8
Dog	0.5	30.6	15.5	15.7
Frog	0.6	30.8	15.5	16.0
Horse	0.6	30.5	15.5	15.7

Category	Min Y	Max Y	Length X	Length Y
Ship	3.5	29.3	15.5	17.1
Final Centroids				
Truck	0.8	30.4	15.5	15.7
Airplane	3.6	28.9	15.5	17.0
Automobile	0.8	30.5	15.4	15.7
Bird	5.1	27.6	15.5	16.8
Cat	1.1	30.3	15.6	16.0
Deer	2.6	29.4	15.4	16.4
Dog	1.0	30.4	15.5	16.0
Frog	0.9	30.4	15.4	15.8
Horse	1.7	30.2	15.6	16.2
Ship	1.9	30.0	15.5	16.5

The results of the initial and final centroid calculations for the CIFAR-10 dataset are presented in Table 4, Table 5, and Table 6. These tables provide detailed insights into how the centroids for each class (e.g., truck, airplane, automobile, etc.) shift during the clustering process. Key metrics such as average color intensities (Avg R, Avg G, Avg B), minimum and maximum values (Min R, Max G), and spatial dimensions (Length X, Length Y) highlight the changes between the initial and final cluster centroids. Subsequently, we applied K-Means clustering along with comprehensive feature analysis to these images. The table below details the clustering results, demonstrating the performance of our approach for each class.

Table 7. Results Clustering CIFAR-10

Cluster	Class	Error(n)	Images (N)	Accuracy (%)
0	truck	102	970	89.5
1	airplane	180	1000	82.0
2	automobile	94	1000	90.6
3	bird	75	720	89.6
4	cat	93	848	89.0
5	deer	74	720	89.7
6	dog	65	731	91.1
7	frog	60	669	91.0
8	horse	78	720	89.2
9	ship	108	970	88.9

Subsequently, the performance of K-Means clustering on these classes is summarized in Table 7, where the accuracy of clustering for each class is detailed. The truck and frog classes demonstrate higher accuracy (89.5% and 91.0%, respectively), while the airplane class shows a slightly lower accuracy of 82.0%. These results reflect the effectiveness of the clustering algorithm in distinguishing between different classes of images in the CIFAR-10 dataset.

3.3 Comparison with Related Approaches

This study utilized clustering techniques combining symbolic data analysis with traditional clustering methodologies to address the image data involving both intensity and spatial features. The clustering process was applied to a diverse dataset of images, including real-world images. The use of symbolic data analysis allowed for a nuanced differentiation between similar image categories, achieving an overall accuracy rate of 94%.

To validate the novelty of our proposed symbolic data-based clustering method, we conducted a comparative analysis with traditional clustering approaches, including K-Means, DBSCAN, and hierarchical clustering, using the CIFAR-10 dataset. Table 8 summarizes each method’s clustering accuracy and performance metrics.

Table 8. **Comparison of Clustering Approaches on CIFAR-10 Dataset**

Method	Accuracy (%)	Time (s)	Clusters	Adjusted Rand Index
K-Means	82.5	12.3	10.0	0.7
DBSCAN	78.1	34.8	Varies	0.6
Hierarchical	81.3	28.4	10.0	0.6
Symbolic data-based	94.0	15.6	10.0	0.7

The proposed method outperforms traditional approaches regarding clustering accuracy and interpretability, capturing symbolic data’s inherent complexity.

3.4 Discussion

The results demonstrate that symbolic data clustering performs better in handling complex image datasets like CIFAR-10. The proposed method’s ability to represent variability within symbolic objects provides better clustering accuracy than traditional approaches like K- Means and DBSCAN. Symbolic data clustering excels because it captures both variability and uncertainty inherent in the data, often overlooked by traditional methods. Unlike K-Means, which relies solely on numerical representations and Euclidean distances, the symbolic method incorporates interval and distribution-based features, allowing it to process high-dimensional data more effectively. This flexibility suits it particularly for datasets like CIFAR-10, which contain diverse object categories with overlapping visual characteristics. One key advantage of the symbolic approach lies in its tailored distance metric, which is optimized to handle symbolic objects. This metric enhances cluster separation by considering the variability within feature intervals, which is critical for datasets with overlapping clusters or noisy data points. For instance, the superior accuracy of 94.0% on CIFAR-10 highlights its robustness in distinguishing complex patterns thattraditional methods might misclassify.

Additionally, the symbolic data-based clustering method maintains consistent performance regardless of the dataset’s complexity. For CIFAR-10, the method effectively handles the inherent diversity of object categories, such as animals, vehicles, and landscapes, which often share similar color distributions or spatial features. In contrast, traditional methods like DBSCAN struggle with determining an optimal number of clusters in such varied datasets, leading to reduced accuracy. Hierarchical clustering, while performing better than DBSCAN in some cases, suffers from high computational overhead and difficulty in scaling to large datasets like CIFAR-10. While separating color channels helps extract richer features for datasets with significant color variability (e.g., CIFAR-10 or flower datasets), this flexibility highlights the adaptability of the approach across diverse datasets. However, additional techniques, such as intensity normalization or the inclusion of illumination-invariant features, could further enhance the method's robustness. These aspects will be explored in future work.

4. CONCLUSION

This paper presents a novel approach for clustering image data using symbolic data integrated with the K-Means algorithm. Our method outperformed traditional approaches regarding accuracy and flexibility, particularly in handling complex datasets like CIFAR-10. Despite these promising results, our study has several limitations. First, the computational complexity of symbolic data analysis may limit its scalability to very large datasets. Second, our approach has not been tested on highly heterogeneous datasets, which would be an important area for future research. Future work could focus on optimizing the algorithm’s computational efficiency and exploring its application to more diverse datasets, including real-world medical and environmental data. Furthermore, integrating symbolic data with other clustering algorithms, such as density-based or spectral clustering, could yield interesting insights.

Author Contributions

Husty Serviana Husain: Conceptualization, writing-original draft, software, validation, formal analysis, investigation, resource, data curation, writing-original draft preparation, writing-review and editing, project

administration. Sapto Wahyu Indratno: Conceptualization, methodology, supervision. Sandy Vantika: validation, visualization. All authors discussed the results and contributed to the final manuscript.

Funding Statement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Acknowledgment

Thank you for the LPDP scholarships.

Declarations

The authors declare that no conflicts of interest to report study.

Declarations of Generative AI and AI-assisted Technologies

Generative AI (ChatGPT) was used exclusively for language editing and stylistic improvement. The authors take full responsibility for the content and confirm that all analyses, results, and interpretations are their own. The final manuscript was additionally reviewed for linguistic accuracy by an English-language expert.

REFERENCES

[1] J. A. Hartigan and M. A. Wong, "ALGORITHM AS 136: A K-MEANS CLUSTERING ALGORITHM," *Applied Statistics*, vol. 28, no. 1, hlm. 100, 1979, doi: https://doi.org/10.2307/2346830.

[2] M. E. Celebi, "IMPROVING THE PERFORMANCE OF K-MEANS FOR COLOR QUANTIZATION," *Image and Vision Computing*, vol. 29, no. 4, hlm. 260–271, Mar 2011, doi: https://doi.org/10.1016/j.imavis.2010.10.002.

[3] L. Billard and E. Diday, *SYMBOLIC DATA ANALYSIS: CONCEPTUAL STATISTICS AND DATA MINING*, in Wiley series in computational statistics. Hoboken, NJ: Wiley, 2007.

[4] M. Noirhomme-Fraiture and P. Brito, "FAR BEYOND THE CLASSICAL DATA MODELS: SYMBOLIC DATA ANALYSIS," *Statistical Analysis*, vol. 4, no. 2, hlm. 157–170, Apr 2011, doi: https://doi.org/10.1002/sam.10112.

[5] K. Jajuga, A. Sokołowski, and H.-H. Bock, Ed., *CLASSIFICATION, CLUSTERING, AND DATA ANALYSIS: RECENT ADVANCES AND APPLICATIONS*. in Studies in Classification, Data Analysis, and Knowledge Organization. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002. doi: https://doi.org/10.1007/978-3-642-56181-8.

[6] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A.W.M. V. D. Laak, B. V.aGinneken and C. I. Sánchez., "A SURVEY ON DEEP LEARNING IN MEDICAL IMAGE ANALYSIS," *Medical Image Analysis*, vol. 42, hlm. 60–88, Des 2017, doi: https://doi.org/10.1016/j.media.2017.07.005.

[7] Y. J. Zhang, "A SURVEY ON EVALUATION METHODS FOR IMAGE SEGMENTATION," *Pattern Recognition*, vol. 29, no. 8, hlm. 1335–1346, Agu 1996, doi: https://doi.org/10.1016/0031-3203(95)00169-7.

[8] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "DEEP CLUSTERING FOR UNSUPERVISED LEARNING OF VISUAL FEATURES," in *Computer Vision – ECCV 2018*, vol. 11218, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Ed., in Lecture Notes in Computer Science, vol. 11218. , Cham: Springer International Publishing, 2018, hlm. 139–156. doi: https://doi.org/10.1007/978-3-030-01264-9_9.

[9] Y. LeCun, Y. Bengio, and G. Hinton, "DEEP LEARNING," *Nature*, vol. 521, no. 7553, hlm. 436–444, Mei 2015, doi: https://doi.org/10.1038/nature14539.

[10] J. Macqueen, "SOME METHODS FOR CLASSIFICATION AND ANALYSIS OF MULTIVARIATE OBSERVATIONS".

[11] P. Fränti and S. Sieranoja, "HOW MUCH CAN K-MEANS BE IMPROVED BY USING BETTER INITIALIZATION AND REPEATS?," *Pattern Recognition*, vol. 93, hlm. 95–112, Sep 2019, doi: https://doi.org/10.1016/j.patcog.2019.04.014.

[12] F. Peng and K. Li, "DEEP IMAGE CLUSTERING BASED ON LABEL SIMILARITY AND MAXIMIZING MUTUAL INFORMATION ACROSS VIEWS," *Applied Sciences*, vol. 13, no. 1, hlm. 674, Jan 2023, doi: https://doi.org/10.3390/app13010674.

[13] Y. Li, P. Hu, D. Peng, J. Lv, J. Fan, and X. Peng, "IMAGE CLUSTERING WITH EXTERNAL GUIDANCE," 16 Juli 2024, *arXiv*: arXiv:2310.11989. doi: https://doi.org/10.48550/arXiv.2310.11989.

[14] A. Stephan, L. Miklautz, K. Sidak, J. P. Wahle, B. Gipp, C. Plant and B. Roth., "TEXT-GUIDED IMAGE CLUSTERING".

[15] S. Raya, M. Orabi, I. Afyouni, and Z. Al Aghbari, "MULTI-MODAL DATA CLUSTERING USING DEEP LEARNING: A SYSTEMATIC REVIEW," *Neurocomputing*, vol. 607, hlm. 128348, Nov 2024, doi: https://doi.org/10.1016/j.neucom.2024.128348.

[16] H.-Y. Hsu, K. H. Keoy, J.-R. Chen, H.-C. Chao, and C.-F. Lai, "PERSONALIZED FEDERATED LEARNING ALGORITHM WITH ADAPTIVE CLUSTERING FOR NON-IID IOT DATA INCORPORATING MULTI-TASK LEARNING AND NEURAL NETWORK MODEL CHARACTERISTICS," *Sensors*, vol. 23, no. 22, hlm. 9016, Nov 2023, doi: https://doi.org/10.3390/s23229016.

[17] X. Wu, Y.-F. Yu, L. Chen, W. Ding, and Y. Wang, "ROBUST DEEP FUZZY K -MEANS CLUSTERING FOR IMAGE DATA," *Pattern Recognition*, vol. 153, hlm. 110504, Sep 2024, doi: https://doi.org/10.1016/j.patcog.2024.110504.

[18] J. D. Yanosky, C. J. Paciorek, F. Laden, J. E. Hart, R. C. Puett, D. Liao and H. H. Suh., "SPATIO-TEMPORAL MODELING OF PARTICULATE AIR POLLUTION IN THE CONTERMINOUS UNITED STATES USING GEOGRAPHIC AND METEOROLOGICAL PREDICTORS," *Environ Health*, vol. 13, no. 1, hlm. 63, Des 2014, doi: https://doi.org/10.1186/1476-069X-13-63.

[19] D. Hasenfratz, O. Saukh, S. Sturzenegger, and L. Thiele, "PARTICIPATORY AIR POLLUTION MONITORING USING SMARTPHONES".

[20] A. J. Alaran, N. O'Sullivan, L. Tatah, R. Sserunjogi, and G. Okello, "AIR POLLUTION ($PM_{2.5}$) AND ITS METEOROLOGY PREDICTORS IN KAMPALA AND JINJA CITIES, IN UGANDA," *Environ. Sci.: Atmos.*, vol. 4, no. 10, hlm. 1145–1156, 2024, doi: https://doi.org/10.1039/D4EA00074A.

[21] M. Ester, H.-P. Kriegel, and X. Xu, "A DENSITY-BASED ALGORITHM FOR DISCOVERING CLUSTERS IN LARGE SPATIAL DATABASES WITH NOISE".

[22] S. C. Johnson, "HIERARCHICAL CLUSTERING SCHEMES," in *Psychometrika*, vol. 32, hlm. 241–254.

[23] R. C. Gonzalez and R. E. Woods, *DIGITAL IMAGE PROCESSING*, 2nd ed. Upper Saddle River, N.J: Prentice Hall, 2002.

[24] "NUMPY DOCUMENTATION — NUMPY V2.3 MANUAL." Diakses: 23 September 2025. [Daring]. Available online: https://numpy.org/doc/stable/

[25] "PYPI DOCS." Diakses: 23 September 2025. [Daring]. Available online: https://docs.pypi.org/

[26] "USER GUIDE," scikit-learn. Diakses: 23 September 2025. [Daring]. Available online: https://scikit-learn/stable/user_guide.html