International Journal on Advanced Science Engineering Information Technology

Vol.14 (2024) No. 2
ISSN: 2088-5334

Performance Comparison of Dimensional Reduction using Principal Component Analysis with Alternating Least Squares in Modified Fuzzy Possibilistic C-Means and Fuzzy Possibilistic C-Means

Edi Satriyanto [a,c,*], Ni Wayan Surya Wardhani [a], Syaiful Anam [a], Wayan Firdaus Mahmudy [b]

[a] Faculty of Mathematics and Natural Science, Universitas Brawijaya, Malang, Indonesia
[b] Faculty of Computer Science, Universitas Brawijaya, Malang, Indonesia
[c] Department of Informatics and Computer Engineering, Politeknik Elektronika Negeri Surabaya, Indonesia
*Corresponding author: *edi@pens.ac.id*

***Abstract*—The clustering method is said to be good if it has resistance to outlier data. One cluster method resistant to outlier data is Fuzzy Possibilistic C-Means (FPCM). FPCM performance on outlier data still has the potential for overlap between cluster members in different clusters, resulting in decreased cluster quality. The Modified Fuzzy Possibilistic C-Means (MFPCM) method is used to modify FPCM in its objective function by inserting updated weight values to increase FPCM performance. In this research, improving the quality of FPCM and MFPCM clusters was carried out by reducing data dimensions through Principal Component Analysis using Alternating Least Squares (PRINCALS) so that members of each cluster do not overlap in the right cluster. The PRINCALS results of the FPCM method have better performance with silhouette values and BSS/TSS ratios of 0.4108 and 60% compared to values without PRINCALS of 0.355 and 43%. The MFPCM method with PRINCALS also performs better, namely 0.4299 and 61%, compared to 0.368 and 42% without PRINCALS. In this study, the performance of MFPCM with PRINCALS or without PRINCALS was better than that of the FPCM method. Overall, PRINCALS can improve the performance of the MFPCM and FPCM methods, resulting in better clusters. PRINCALS in this cluster produce an average silhouette value greater than 0.3 and an average BSS/TSS ratio greater than 50% so that each cluster member is in the right cluster and does not overlap.**

***Keywords*—Between sum of square; fuzzy possibilistic C-Means; modified fuzzy possibilistic C-Means; silhouette coefficient.**

Manuscript received 21 Jan. 2024; revised 23 Mar. 2024; accepted 3 Apr. 2024. Date of publication 30 Apr. 2024.

I. INTRODUCTION

Clustering methods are vital in data analysis, as they can parse and understand complex data for better decision-making in many fields. Clustering methods widely used in previous research include Fuzzy C-means (FCM) and K-means. The FCM and K-Means Clustering methods are used for object image segmentation, which is much needed in processing and analyzing image data. FCM performance shows better image segmentation results when compared to K-Means [1]. FCM is also used in the health sector to detect and classify bone tumors efficiently using the FCM clustering algorithm [2]. The cluster method is also used to predict attitudes toward acceptance of internet health information through a clustering method based on health data management in the younger generation. [3]. In the agricultural sector, the FCM cluster method is used for the detection of several non-leafy vegetables such as Brinjal, Chilly, Bitter Gourd, Onion and Tomato [4].

Cluster methods are usually able to produce optimal performance if the entities or objects in the cluster have high similarities and significant differences between clusters [5]. In reality, we often perform clustering on datasets that have outlier data. Outlier data can move the cluster center (centroid) from the actual data center, thereby shifting the cluster as a whole and causing the cluster to appear not to represent the majority of the data [6]. A good cluster must be relatively resistant to outlier data. Clusters must be stable and not significantly affected by small changes in the data or the presence of inappropriate entities [7] .

A clustering method that has better resistance to outlier data is fuzzy-possibilistic c-means clustering (FPCM). The FPCM method is an improvement on FCM, which is still sensitive to outlier data, and Possibilistic C-Means (PCM), which already has resistance to outlier data [8]. Even though

PCM is no longer sensitive to outlier data, cluster members still overlap because it depends on good initial cluster center initialization [9], [10]. Previous research used FPCM to carry out multi-resolution segmentation of stacked images into objects from coarse to scale. Experimental results demonstrate the effectiveness and stability of the proposed approach [11]. FPCM is also used to investigate data in the real world, namely to break down extensive data sets into meaningful clusters. FPCM can cluster data in the database effectively [12]. In the health sector, FPCM is used in knee osteoarthritis analysis with kernel functions to handle the problem of data that cannot be separated. The results of FPCM performance in clustering knee osteoarthritis disease have an accuracy value of 85.5% [13]. FPCM, in other research, is used to predict groups of workers with low performance in software companies. FPCM can produce high efficiency in completing the clustering of unlabeled data and outlier data [14]. The application of FPCM to cluster changing object information from sensor nodes in a large area uses a wireless sensor network. FPCM can better group objects from wireless sensor networks than the K-Medoid method [15].

In its development, the FPCM method was modified by using weight values in the objective function of the FPCM method, Modified Fuzzy Possibilistic C-Means (MFPCM). MFPCM is considered an improvement over FPCM in many aspects, especially in terms of flexibility in membership adjustment, handling of outliers, robustness to cluster center initialization, ability to handle complex cluster structures, and ability to handle multimodal data [16]. In previous research, MFPCM was used to group students with low competency in helping academics provide appropriate training. The application of the MFPCM algorithm can identify low performers with high accuracy when compared to the Fuzzy Possibilistic Product Partition C-Means Clustering algorithm [17].

The problem of cluster performance results depends not only on the cluster method used. Even though FPCM and MFPCM are resistant to outlier data, determining the number of variables in the research dataset can potentially affect the quality of the clusters. Variables in a dataset do not necessarily contribute to cluster data analysis, let alone mixed data scales. These conditions complicate the clustering process and can lead to suboptimal results. Hence, it is necessary to consider outliers in the clustering process and take appropriate steps to overcome their impact. [18]. One step is understanding the dataset's characteristics using appropriate data preprocessing techniques. Datasets with high dimensions can be preprocessed by reducing their dimensions to simpler ones [19].

The method that is often used to reduce high-dimensional datasets is the Principal Component Analysis (PCA) method [20], [21]. The PCA method in previous research was used to evaluate data anomaly detection by carrying out dimension reduction on extensive manufacturing data so that the data analysis results are faster and more efficient [22], [23], [24]. The data measurement scale is important in the Principal Component Analysis (PCA) method. PCA is sensitive to data scale, meaning that if the variables in the data have different scales, then the resulting main components can also be influenced by these differences so that the data becomes non-linear [25].

Research datasets often have mixed-scale data; reducing the dimensionality of data sets on mixed scales causes the data to become non-linear. Non-linear PCA on mixed-scale data can be completed using the Alternating Least Squares method [26]. Using Alternating Least Squares (PRINCALS), principal component analysis reduces high dimensions of the mixed scale by transforming high dimensions into low ones using the Alternating Least Squares method [27]. This research also uses a mixed scale. The dataset used in this research is data on new students' economic abilities, consisting of nominal, ordinal, and ratio scales. This research aims to improve the cluster performance of the FPCM and MFPCM methods through data preprocessing using PRINCALS. The dimensions of the research dataset were reduced using PRINCALS, and the dimension reduction transformation results were used in cluster analysis. The performance of FPCM and MFPCM clusters is expected to improve cluster quality according to the characteristics of the research data used.

The best cluster performance results in this research will then be used to classify the Single Tuition Fee (STF). The economic data characteristics of the new students greatly influence the STF amount. The best clustering results from this research will then be used as a source of information in determining the STF level, starting from the lowest STF value to the highest.

II. Materials and Method

A. Dataset

This research uses a mixed-scale dataset that has outlier data. The dataset is from Politeknik Elektronika Negeri Surabaya (PENS) student selection. This dataset has 3 scales, namely ratio, nominal, and ordinal scales, which relate to new student economic data as follows:

X_1: parents' income (ratio scale)
X_2: number of family dependents (ratio scale)
X_3: home ownership (ordinal scale)
X_4: number of houses (ratio scale)
X_5: land ownership (nominal scale)
X_6: ownership of ponds/rice fields (nominal scale)
X_7: apartment ownership (nominal scale)
X_8: building ownership (nominal scale)
X_9: number of cars owned (ratio scale)
X_{10}: number of motorbikes owned (ratio scale)
X_{11}: electric power (ordinal scale: 450 Watt, 900 Watt, 1300 Watt, and over 1300 Watt)
X_{12}: Institutional Development Contribution (IDC) (ratio scale).

B. Methodology

This research was carried out through several stages, as shown in Fig. 1. The dataset before cluster analysis is carried out, and new student economic data is preprocessed to select the required variables. The dataset must be converted into data categories because the data is on a mixed scale. The following process is dimensional reduction using PRINCALS. The PRINCALS algorithm will cause the distance between data points to tend not to get bigger so that cluster analysis becomes more straightforward. The number of dimension reductions is carried out by looking at the scree plot of the eigenvalues. Next, the dataset was transformed using a linear combination through the score component of the PRINCALS dimension reduction algorithm. The PRINCALS transformation data will be used in the cluster analysis process. Determining the number of clusters in the study uses the Silhouetted Coefficient before using the FPCM and MFPCM algorithms. Next, the FPCM and MFPCM performance analysis results are carried out by measuring the precision of cluster members so that the best cluster quality is obtained from each cluster.

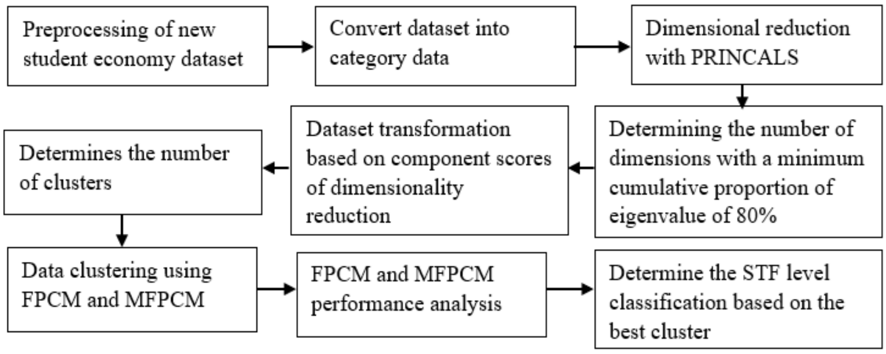

Preprocessing of new student economy dataset
Convert dataset into category data
Dimensional reduction with PRINCALS
Determining the number of dimensions with a minimum cumulative proportion of eigenvalue of 80%
Dataset transformation based on component scores of dimensionality reduction
Determines the number of clusters
Data clustering using FPCM and MFPCM
FPCM and MFPCM performance analysis
Determine the STF level classification based on the best cluster

Fig. 1 Methodology

C. Principal Component Analysis with Alternating Least Squares (PRINCALS)

Principal Component Analysis with Alternating Least Squares (PRINCALS) is one method used to carry out principal component analysis (PCA). This method reduces non-linear datasets due to mixed data scales. PRINCALS uses a least squares-based iterative approach to estimate principal components from the data. This approach minimizes the squared difference between the original and reconstructed data from the extracted principal components [27].

The PRINCALS algorithm steps are as follows:

1. Convert the quantitative datasets into the qualitative data (categories).
2. Perform quantification into an $\boldsymbol{H}$ matrix of $n \times m$ size. The $\boldsymbol{h_{ij}}$ vector is then transformed using quantification as in equation (1).

$$\boldsymbol{H} = (\boldsymbol{h_{ij}}) = \begin{pmatrix} h_{11} & \boldsymbol{h_{12}} & \cdots & h_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ h_{n1} & h_{n2} & \cdots & h_{nm} \end{pmatrix}, \boldsymbol{h_{i1}} = \begin{pmatrix} h_{11} \\ \vdots \\ h_{n1} \end{pmatrix}, \qquad (1)$$

where:

$\boldsymbol{H}$: qualitative data matrix

$\boldsymbol{h_{ij}}$: vector of the i-th object in the k-th category in the j-th variable

n: many observations (objects); i =1,2,…,n

m: many variables; $j = 1,2,\dots,m$

k: number of categories in the j-th variable;

$r = 1,2,\dots,k_j$.

3. Perform quantification technique for the $\boldsymbol{H}$ matrix into $\boldsymbol{G}_j$ matrix of size $n \times k_j$ can be calculated using equation (2).

$$\boldsymbol{G}_j = (g_{ijk}) = \begin{pmatrix} g_{j11} & \cdots & g_{j1k_j} \\ \vdots & \vdots & \vdots \\ g_{jn1} & \cdots & g_{jnk_j} \end{pmatrix} = (g_{j1} \quad \cdots \quad g_{jk_j}), \qquad (2)$$

$$g_{ijk=}\begin{cases} 1, \text{if object } i \text{ is belongs to category } k \text{ of } j \text{ variable} \\ 0, \text{if object } i \text{ not belongs to category } k \text{ of } j \text{ variable} \end{cases}$$

where $\boldsymbol{G}_j$ *is* indicator matrix of $\boldsymbol{h_{ij}}$ and g_{ijk} is matrix column.

4. *Determine* the object score matrix ($\boldsymbol{X}$) and quantify the $\boldsymbol{Y}_j$ category by minimizing meet loss as in equation (3).

$$\sigma_M(\boldsymbol{X};\boldsymbol{Y}) = \frac{1}{m}\sum_{j-1}^{m} tr(\boldsymbol{X}-\boldsymbol{G}_j\boldsymbol{Y}_j)'\,\boldsymbol{M}_j(\boldsymbol{X}-\boldsymbol{G}_j\boldsymbol{Y}_j)\ , \qquad (3)$$

where:

tr: *trace* (summation of main diagonal elements),

$\boldsymbol{X}$: ordered object component score matrix $n \times p$,

$\boldsymbol{Y}$: a collection of multiple and single category quantification,

$\boldsymbol{G}_j$: indicator matrix for the jth variable of size $n \times k_j$,

$\boldsymbol{Y}_j$: quantification of multiple ordered categories $k_j \times p$,

$\boldsymbol{M}_j$: sized identity matrix $n \times n$.

5. Determine the number of primary components selected using total diversity, eigenvalue, scree plot, and hypothesis testing. Eigenvalues (λ) can be searched using PRINCALS from the correlation matrix $m^{-1}\boldsymbol{R}(\boldsymbol{Q})$ with equation (3),

$$|m^{-1}\boldsymbol{R}(\boldsymbol{Q}) - \lambda\boldsymbol{I}| = 0, \qquad (4)$$

where:

m: the number of variables used,

$\boldsymbol{R}(\boldsymbol{Q})$: correlation matrix between the combined linear scores of all sets of $\boldsymbol{Q}$ matrices,

$\boldsymbol{Q}$: transformation data matrix of order $n \times m$ with column q_j where $q_j = \boldsymbol{G}_j y_j$,

q_j : transformation data.

6. Determine the main component score or loading component using equation (5):

$$\boldsymbol{a_j} = (\boldsymbol{y}_j'\boldsymbol{D}_j\boldsymbol{y}_j)^{-1}(\boldsymbol{Y}_j'\boldsymbol{D}_j\boldsymbol{y}_j), \qquad (5)$$

where:

$\boldsymbol{a}_j$: component weight (component loading) of the order $p \times 1$,

$\boldsymbol{D}_j$: diagonal matrix $k_j \times k_j$ with the relative frequency of the jth variable on the main diagonal,

$\boldsymbol{y}_j$: Single category quantification.

7. Calculate component scores according to the following equation (6),

$$K_{ij} = \sum_{j=1}^{m}\boldsymbol{a}_j\boldsymbol{G}_j\boldsymbol{Y}_j\, , K_{ij} = \sum_{j=1}^{m}\boldsymbol{a}_j\boldsymbol{q}_j, \qquad (6)$$

where :

K_{ij} : is the main component score on the ith object of the jth variable with i=1,2,3,..., n and j=1,2,3,...,m

$\boldsymbol{a}_j$: component weights of order $p \times 1$

$\boldsymbol{q}_j$: j th transformation data.

The variance that the new variable can explain i depends on the contribution of the Main Component proportion in percent or p_i for each eigien value which is calculated using the equation (7),

$$p_i = \frac{\lambda_i}{\sum_{j=1}^{D} \lambda_j} \times 100\%, \tag{7}$$

where:

p_i: principal component proportion in percent (%)

λ_i: the i-th eigenvalue and λ_j eigenvalue of each diagonal.

D: the number of principal components.

D. Fuzzy Possibilistic C Means (FPCM) Algorithm

FPCM is a development algorithm from FCM and PCM. The $\boldsymbol{\mu_{ik}}$ value in the FCM algorithm is influenced by $\boldsymbol{x_k}$ and all cluster centers [9]. Meanwhile, in the PCM algorithm, the value of $\boldsymbol{t_{ik}}$ is only affected by $\boldsymbol{x_k}$, the center of the ith cluster and $\boldsymbol{\gamma_i}$. FPCM has the advantage of ignoring outlier data sensitivity deficiencies such as FCM and overcoming the problem of identical cluster such as PCM [14].

The FPCM method has the following objective functions:

$$J_{FPCM}(\boldsymbol{X}; \boldsymbol{V}, \boldsymbol{U}, \boldsymbol{T}) = \sum_{i=1}^{c} \sum_{j=1}^{n} (u_{ij}^{m} + t_{ij}^{\eta})\, d^2(\boldsymbol{x}_i, \boldsymbol{v}_j), \tag{8}$$

where :

$\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n\} \subseteq \mathbb{R}^{\boldsymbol{p}}$ is a dataset of n records with dimension in the p -dimensional data space $\mathbb{R}$,

$\boldsymbol{V} = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_n\} \subseteq \mathbb{R}^{\boldsymbol{p}}$ is the matrix of cluster centers,

$\boldsymbol{U} = \{u_{ij}\}$ is a uniqueness matrix (u_{ij})**,**

$\boldsymbol{T} = \{t_{ij}\}$ is the absolute uniqueness matrix (t_{ij}),

m is the value determined for rank in a cluster $(m > 1)$ or $m = 2$ and η is the typicality exponent $(\eta > 1)$ or $\eta = 2$.

Determining the similarity of data in clusters using the distance method as in the equation (9),

$$d^2(\boldsymbol{x}_j, \boldsymbol{v}_i) = \|\boldsymbol{x}_j - \boldsymbol{v}_i\|^2 = (\boldsymbol{x}_j - \boldsymbol{v}_i)'(\boldsymbol{x}_j - \boldsymbol{v}_i), \tag{9}$$

where:

$d^2(\boldsymbol{x}_j, \boldsymbol{v}_i)$ is the squared euclidean distance between data $(\boldsymbol{x}_j)$ and cluster centers $(\boldsymbol{v}_i)$. Each element in the membership matrix $\boldsymbol{U}$ and $\boldsymbol{T}$, namely u_{ij} and t_{ij}, can be expressed in equations:

$$\sum_{i=1}^{c} u_{ij} = 1; \forall\, j \in \{1, \dots, n\}, \tag{10}$$

$$\sum_{j=1}^{c} t_{ij} = 1; \forall\, i \in \{1, \dots, c\}. \tag{11}$$

The cluster results are obtained by minimizing the objective function $J_{FPCM}(\boldsymbol{X}; \boldsymbol{V}, \boldsymbol{U}, \boldsymbol{T})$ based on the updated of u_{ij}, t_{ij} and $\boldsymbol{v}_i$ in the following equations:

$$u_{ij} = \left[\sum_{j=1}^{k} \left(\frac{d^2(\boldsymbol{x}_j, \boldsymbol{v}_i)}{d^2(\boldsymbol{x}_j, \boldsymbol{v}_k)}\right)^{\frac{1}{m-1}}\right]^{-1} ; 1 \le i \le c, 1 \le j \le n, \tag{12}$$

$$t_{ij} = \left[\sum_{k=1}^{n} \left(\frac{d^2(\boldsymbol{x}_i, \boldsymbol{v}_j)}{d^2(\boldsymbol{x}_i, \boldsymbol{v}_l)}\right)^{\frac{1}{(\eta-1)}}\right]^{-1} ; 1 \le i \le c, 1 \le j \le n, \tag{13}$$

$$\boldsymbol{v}_i = \frac{\sum_{j=1}^{n} (u_{ij}^{m} + t_{ij}^{\eta}) \boldsymbol{x}_j}{\sum_{j=1}^{n} (u_{ij}^{m} + t_{ij}^{\eta})} \; ; \; 1 \le i \le c \tag{14}$$

E. E. Modified Fuzzy Possibilistic C Means (MFPCM) Algorithm

The MFPCM algorithm is a modification of the FPCM algorithm by including weight parameters in the objective function. The MFPCM method will improve cluster performance so that it can minimize the distance between points in the cluster and maximize the distance between clusters. The MFPCM method as a whole has almost the same algorithm as the FPCM method. The difference between the FPCM and MFPCM methods is that the objective function of the MFPCM method is added to the weight parameter values as follows [17]:

$$J_{MFPCM}(\boldsymbol{X}; \boldsymbol{V}, \boldsymbol{U}, \boldsymbol{T}) = \sum_{i=1}^{c} \sum_{j=1}^{n} \Big(u_{ij}^{m}\, w_{ji}^{m}\, d^{2m}(\boldsymbol{x}_j, \boldsymbol{v}_i) + t_{ij}^{\eta} w_{ji}^{\eta}\, d^{2\eta}(\boldsymbol{x}_j, \boldsymbol{v}_i) \Big) \tag{15}$$

In the MFPCM objective function each cluster is given such a weight so that it will produce better cluster classification even though it has outlier data. Determining the weight for each cluster is obtained by the equation (16).

$$w_{ji} = exp\left[- \frac{d^2(\boldsymbol{x}_j, \boldsymbol{v}_i)}{\left(\sum_{j=1}^{n} d^2(\boldsymbol{x}_j, \boldsymbol{v}_i)\right)\frac{c}{n}}\right] \tag{16}$$

The cluster results from MPFCM are obtained by minimizing the objective function using the following update equation[24]:

$$u_{ij} = \left[\sum_{k=1}^{c} \left(\frac{d^2(\boldsymbol{x}_j, \boldsymbol{v}_i)}{d^2(\boldsymbol{x}_j, \boldsymbol{v}_k)}\right)^{\frac{2m}{(m-1)}}\right]^{-1} , 1 \le i \le c, 1 \le j \le n, \tag{17}$$

$$t_{ij} = \left[\sum_{k=1}^{n} \left(\frac{d^2(\boldsymbol{x}_j, \boldsymbol{v}_i)}{d^2(\boldsymbol{x}_j, \boldsymbol{v}_k)}\right)^{\frac{2\eta}{(\eta-1)}}\right]^{-1} , 1 \le i \le c, 1 \le j \le n, \tag{18}$$

$$\boldsymbol{v}_i = \frac{\sum_{j=1}^{n} (u_{ij}^{m} w_{ji}^{m} + t_{ij}^{\eta} w_{ji}^{\eta}) \boldsymbol{x}_j}{\sum_{j=1}^{n} (u_{ij}^{m} w_{ji}^{m} + t_{ij}^{\eta} w_{ji}^{\eta})} \; ; \; 1 \le i \le c \tag{19}$$

F. Cluster Quality Performance

Some previous research to measure cluster (C_k) quality performance can be done by looking for the Between Sum of Squares, Total Sum of Squares (BSS/TSS) ratio and Silhouette Coefficient values. The BSS/TSS ratio is a comparison of the values between the sum of the distances (BSS) between pairs of data in different clusters and the sum of the distances of all pairs in the cluster (TSS) according to the following equation [29]:

$$\text{BSS/TSS} = \frac{\sum_{k=1}^{c} \sum_{i \in C_k} \sum_{j \notin C_k} d(i,j)}{\sum_{i,j} d(i,j)} \tag{20}$$

The BSS/TSS ratio value will measure the separation between clusters with a percentage value between 0% and 100%. Cluster performance if the BSS/TSS ratio value is close to 100% means the cluster has good quality because there is less data overlap between different clusters [30]. Cluster quality can also be seen from evaluation metrics based on the Silhouette Coefficient value. This metric provides a measure of how well an object fits the cluster assigned to it and how different it is from other clusters [30]. The formula for calculating the Silhouette Coefficient is as follows (21):

$$S_i = \frac{(b_i - a_i)}{\max(a_i, b_i)} \tag{21}$$

where S_i is silhouette Coefficient, a_i is the average distance from the object i to all objects that are in the same cluster, and b_i is the smallest value of the average distance of the object i to other objects in different clusters. The Silhouetted coefficient has a minimum value of -1 and a maximum value of 1. The cluster results are said to be in the right cluster if the Silhouetted coefficient is positive and vice versa. The cluster method used has valid cluster results if the average value of the Silhouetted coefficient is close to 0.5, and the cluster results are invalid if the average value of the Silhouetted coefficient is less than 0.3.

III. Results and Discussion

A. Dimension Reduction

The dataset in this study is on a mixed scale, so it must be converted into categorical data before reducing its dimensions using PRINCALS. The PRINCALS computing results obtained a cumulative proportion eigenvalue, as in Table 1. In the 8th dimension, the cumulative proportion eigenvalue was 81.64%, so the proportion of variance was considered sufficient to represent the total cumulative variance of 70% to 80%, as in equation (7).

TABLE I
CUMULATIVE PROPORTION OF EIGENVALUE

Dimension	Eigenvalue	Cumulative Proportion
1	0.2186	21.86
2	0.1191	33.77
3	0.0949	43.26
4	0.0919	52.45
5	0.0846	60.91
6	0.0773	68.64
7	0.0687	75.51
8	0.0613	81.64
9	0.0529	86.93
10	0.0509	92.02
11	0.0434	96.36
12	0.0364	100

The PRINCALS algorithm produces component loadings as in Table 2, obtained from equation (4). Based on Table 2, the PRINCALS algorithm then transforms a dataset using equation (6) to produce a reduced dataset with dimensions of 8, as in Table 3.

TABLE II
COMPONENT LOADINGS

Variabel	Dimensions							
	1	2	3	4	5	6	7	8
X_1	0.66	-0.24	0.14	-0.06	-0.12	-0.01	0.04	-0.24
X_2	0.01	0.07	0.64	0.39	0.47	0.22	0.24	-0.25
X_3	0.51	0.34	-0.15	0.58	-0.14	-0.17	-0.01	0.21
X_4	0.62	0.39	-0.17	0.38	-0.05	-0.03	0.18	0.05
X_5	0.39	0.08	-0.23	-0.11	0.28	0.76	-0.23	0.15
X_6	0.26	0.57	0.06	-0.37	-0.32	0.06	0.14	-0.49
X_7	0.20	0.39	-0.17	-0.48	0.46	-0.19	0.42	0.25
X_8	0.07	0.28	0.68	-0.21	-0.41	0.12	-0.09	0.46
X_9	0.36	0.31	0.17	-0.11	0.38	-0.36	-0.67	-0.11
X_{10}	0.70	-0.28	0.17	-0.08	0.04	0.07	0.09	0.05
X_{11}	0.51	-0.53	0.17	-0.13	0.13	-0.29	0.12	0.13
X_{12}	0.64	-0.29	-0.18	-0.12	-0.19	-0.07	-0.07	-0.08

TABLE III
DATASET DIMENSION REDUCTION TRANSFORMATION

No	K1	K2	K3	K4	K5	K6	K7	K8
1	20.8	-1.9	3.5	1.5	1.1	0.3	0.6	-1.6
2	15.0	3.2	5.2	2.2	4.8	-1.7	-2.4	-0.9
3	16.0	2.2	5.5	1.9	5.0	-2.3	-2.1	-0.7
….	….	….	….	….	….	….	….	….
432	11.2	4.8	2.7	1.7	4.1	-1.7	-3.6	-0.3

The dataset resulting from the dimension reduction transformation in Table 3 is used in cluster analysis using the MFPCM and MFPCM methods so that it can improve cluster quality.

B. Cluster Analysis

The quality of the cluster can be seen using the Silhouetted and Ratio BSS/TSS methods. Good cluster performance is influenced by a Silhouetted value that is close to 1 by considering the distance between clusters based on the high percentage value of distribution between clusters via the BSS/TSS ratio. Based on Fig. 2. The performance of FPCM with PRINCALS has better performance when compared to without PRINCALS. FPCM with PRINCALS in clusters of 2 to 8 clusters all Silhouetted values and the BSS/TSS Ratio with higher performance values than without PRINCALS.In cluster 2 the Silhouetted value increased from 0.4093 to 0.6226 and the BSS/TSS ratio percentage increased from 20.25% to 44.11%. Likewise, clusters 3 to cluster 8 show better performance values compared to FPCM without using PRNCALS.

In the MFPCM method with PRINCALS in 2 clusters there was also an increase in the Silhouetted value from 0.4098 to 0.623 and the BSS/TSS Ratio by 20.46% to 44.5%. Even though the Silhouetted value has increased, the BSS/TSS percentage ratio value shows that the performance is still small, so it is considered that the cluster members still have overlapping data in the members of each cluster. The performance results of MFPCM always increase, seen from cluster 3 to cluster 8, all silhouette values and BSS/TSS ratios always get better.

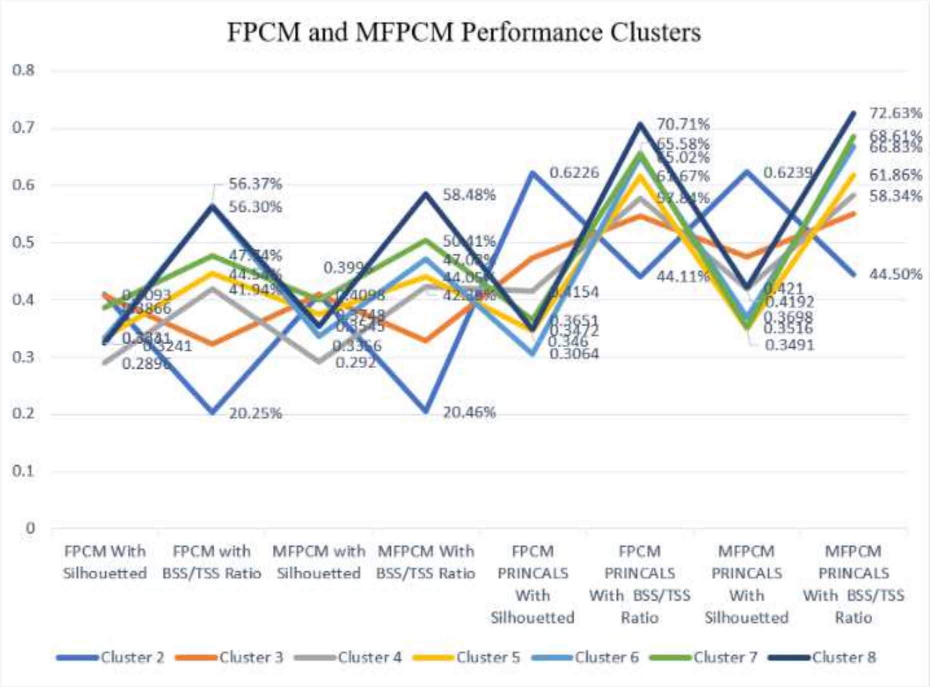


Fig. 2 FPCM and MFPCM performance clusters

In cluster 8, the Silhouette value and BSS/TSS ratio have a relatively better BSS/TSS ratio, namely FPCM of 70.71% and MFPCM of 72.63%, so there is less overlapping of members of each cluster. Overall, the performance graph compares cluster performance between the FPCM and MFPCM methods using the PRINCALS dimension reduction method and without carrying out dimension reduction. The overall performance shows that dimension reduction using PRINCALS causes the Silhouetted and BSS/TSS Ratio values to be higher. Hence, PRINCALS can significantly improve cluster performance.

The cluster performance in the FPCM and MFPCM methods has an average value as in Fig. 2. FPCM performance using PRINCALS has better performance, namely an average Silhouetted value of 0.4108 with an average BSS/TSS ratio of 60%. FPCM without PRINCALS has an average Silhouetted value of 0.355 with a BSS/TSS ratio of 43%. FPCM with PRINCALS increased Silhouetted performance on average by 0.0558 with a BSS/TSS Ratio of 17%. The MFPCM with the PRINCALS method performs better, namely with an average Silhouetted value of 0.4299 with a BSS/TSS ratio of 61%.

MFPCM without PRINCALS has an average Silhouetted value of 0.3680 with a BSS/TSS ratio of 42%. MFPCM with PRINCALS increased Silhouetted performance on average by 0.0619 with a BSS/TSS Ratio of 19%. PRINCALS dimension reduction using the MFPCM method has better performance with a Silhouetted value of 0.4299 and a BSS/TSS ratio of 61% when compared to the PRINCALS FPCM algorithm, which has an average Silhouetted value of 0.4108 with a BSS/TSS ratio of 60%.

The performance of MFPCM and FPCM in Fig. 3 shows that, on average, the MFPCM method using PRINCALS experienced a significant increase in cluster quality. FPCM performance using PRINCALS has better performance, namely an average Silhouetted value of 0.4108 with an average BSS/TSS ratio value of 60% when compared to FPCM without using PRINCALS, which has an average Silhouetted value of 0.355 and an average value BSS/TSS ratio of 43%. The MFPCM method using PRINCALS also performs better than MFPCM without PRINCALS. The average value of silhouetted MFPCM using PRINCALS is 0.4299, and the average value of the BSS/TSS ratio is 61%, while without using PRINCALS, the average value of Silhouetted and the average value of the BSS/TSS ratio is 0.3680 and 42%—comparison of PRINCALS cluster results for MFPCM and FPCM methods. Overall, the Silhouetted value and BSS/TSS ratio for the MFPCM method is higher than FPCM, as shown in Fig. 3.

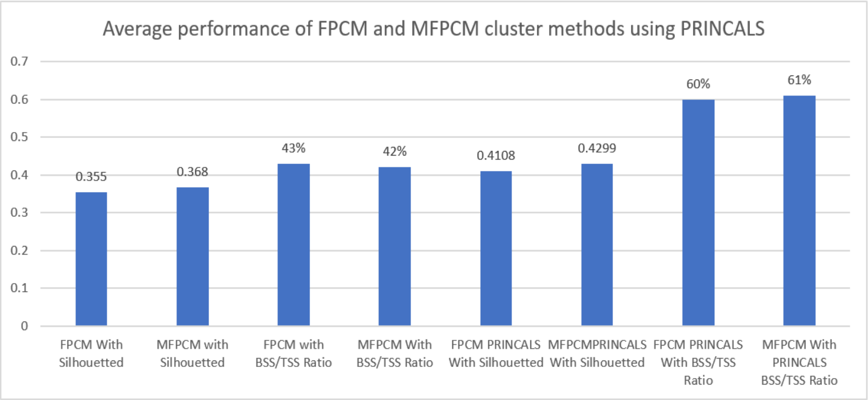
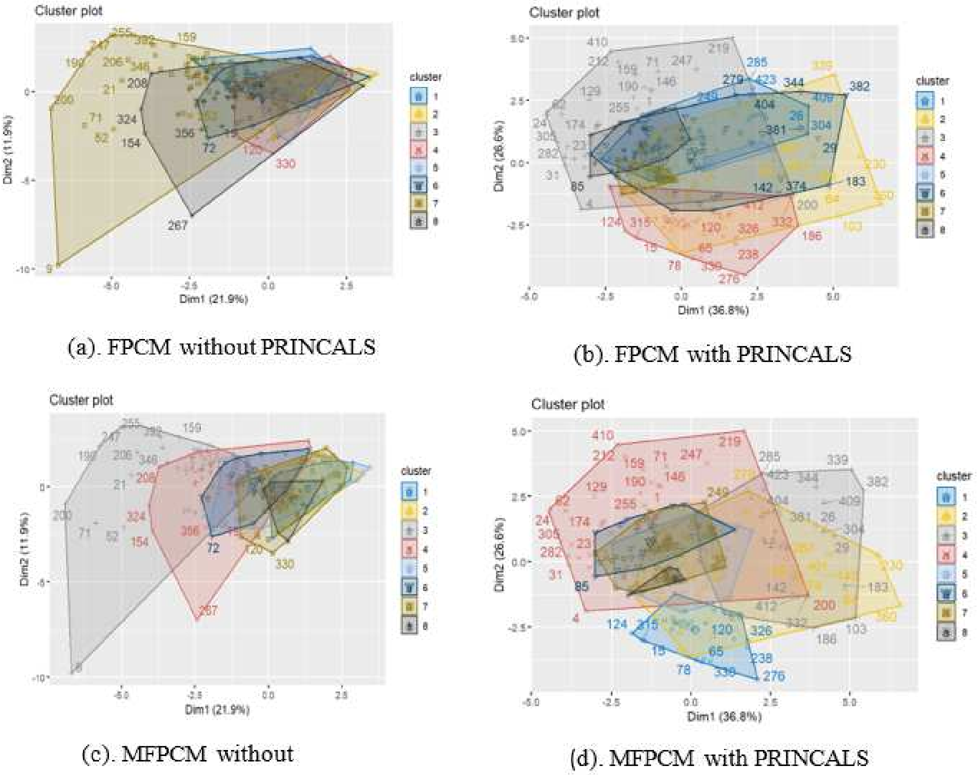


Fig. 3 Average performance of FPCM and MFPCM cluster methods using PRINCALS

The visualization performance of PRINCALS in the cluster method is shown in Fig.4. Based on Figure 4, using PRINCALS in the FPCM and MFPCM methods can produce better clusters because each cluster member appears to have less overlap. FPCM and MFPCM, without using PRINCALS visually, have more overlap between cluster members. The MFPCM cluster method using PRINCALS produces the best performance compared to other cluster methods because the cluster members appear to have the most minor overlap.



(a). FPCM without PRINCALS

(b). FPCM with PRINCALS

(c). MFPCM without

(d). MFPCM with PRINCALS

Fig. 4 Visualization of PRINCALS performance on cluster members using The FPCM and MFPCM methods

Based on the best performance of the cluster method, the MFPCM method with PRINCALS with 8 clusters is then used as cluster analysis to determine the STF classification. Table IV shows the results of MFPCM cluster analysis with PRINCALS on 8 clusters in descriptive analysis. The results of the MFPCM cluster with PRINCALS then calculated the average value for each variable of X_1 to X_{12}. Based on student economic data from each cluster member, STF levels 1 to 8 show the order of tuition fees from lowest to highest.

The STF level is determined by sorting the STF scores by calculating the total result by multiplying the average value of X_1 to X_{12} with the weighted percentage of the determined student economic data. Based on Table IV, the lowest single tuition fees are at STF Level 1, with 79 students in cluster 2 and 43 students in cluster 3, with the highest single tuition fees at STF Level 8.

TABLE IV
THE AVERAGE VALUE OF EACH VARIABLE FROM EACH CLUSTER

Cluster	N	Average												Score	Level
		X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	STF	STF
		20%	5%	10%	10%	2.5%	2.5%	2.5%	2.5%	10%	10%	10%	15%		
1	35	17.0	2.4	2.9	1.2	1.4	1.1	1.0	1.0	1.9	1.2	3.2	33.7	9.73	7
2	79	2.9	2.5	2.5	0.8	1.1	1.0	1.0	1.0	1.2	0.0	1.8	7.5	2.56	1
3	43	55.1	2.3	2.9	1.0	1.1	1.1	1.0	1.0	3.3	0.1	2.2	9.8	13.7	8
4	49	7.7	2.1	3.0	1.1	1.3	1.1	1.0	1.0	1.9	0.7	2.7	1.9	2.98	2
5	87	3.7	2.1	2.5	0.9	1.1	1.1	1.0	1.0	2.0	0.0	2.2	8.9	3.05	4
6	39	10.8	2.9	3.0	1.2	1.4	1.2	1.0	1.0	3.1	1.0	3.0	19.9	6.54	6
7	43	4.0	4.3	2.4	0.8	1.1	1.0	1.0	1.0	1.7	0.1	2.1	8.1	3.04	3
8	56	5.7	2.6	2.7	0.9	1.1	1.1	1.0	1.0	2.1	0.8	2.6	10.3	3.83	5

IV. CONCLUSION

Based on the cluster analysis, it can be concluded that the FPCM method using PRINCALS has better cluster performance than without PRINCALS. FPCM using PRINCALS can increase the average cluster silhouette value by 13.58%, and the BSS/TSS ratio can increase cluster quality by 28.33%. The performance results of the MFPCM method without using PRINCALS have a higher average silhouetted value of 3.53% compared to FPCM, but the average BSS/TSS ratio decreases by 2.38%. The performance results of the MFPCM method using PRINCALS are better than those of FPCM using PRINCALS, with an average silhouette value of 4.44% and an average BSS/TSS ratio increasing by 1.64%. The cluster performance is considered valid because the average silhouette value is more significant than 0.3. The dimension reduction method using PRINCALS can improve the performance of FPCM and MFPCM clusters. The performance results of MFPCM using PRINCALS have better performance because the distance between the clusters is not close together, and there is no overlapping of the BSS/TSS ratio values, which increases significantly.

REFERENCES

[1] A. Kumar and S. S. Sodhi, "Comparative analysis of fuzzy C-Means and K-Means clustering in the case of image segmentation," *Proc. 2021 8th Int. Conf. Comput. Sustain. Glob. Dev. INDIACom 2021*, pp. 194–200, 2021, doi: 10.1109/INDIACom51348. 2021.00035.

[2] D. Mansoor Hussain, M. Anuroopa, A. Dharshini, and G. Durganandini, "Efficient Bone Tumor Detection and Classification using Fuzzy C Means Clustering Algorithm," *Proc. 5th Int. Conf. Trends Electron. Informatics, ICOEI 2021*, pp. 1422–1428, 2021, doi:10.1109/ICOEI51242.2021.9452876.

[3] D. Czerwinski, M. Czerwinska, P. Karczmarek, and A. Kiersztyn, "Influence of the Fuzzy Robust Gamma Rank Correlation, Fuzzy C-Means, and Fuzzy Cognitive Maps to Predict the Z Generation's Acceptance Attitudes towards Internet Health Information," *IEEE Int. Conf. Fuzzy Syst.*, vol. 2021-July, 2021, doi:10.1109/FUZZ45933.2021.9494596.

[4] D. G. Savakar and A. K. Talawar, "Fuzzy C-Means clustering based identification of indian common non-leafy vegetables," *Proc. 2021 8th Int. Conf. Comput. Sustain. Glob. Dev. INDIACom 2021*, vol. 591156, no. c, pp. 858–863, 2021, doi: 10.1109/INDIACom51348.2021.00154.

[5] L. Ge, S. Member, and K. K. Parhi, "Robust Clustering using Hyperdimensional Computing," *IEEE Open J. Circuits Syst.*, vol. PP, no. Xx, p. 1, 2024, doi: 10.1109/OJCAS.2024.3381508.

[6] Z. Li and Y. Li, "COCL: Cluster Combined with Outlier Contrast Learning for Unsupervised Person Re-Identification," *Proc. - 2023 IEEE SmartWorld, Ubiquitous Intell. Comput. Auton. Trust. Veh. Scalable Comput. Commun. Digit. Twin, Priv. Comput. Data Secur. Metaverse, SmartWorld/UIC/ATC/ScalCom/DigitalTwin/PCDS/Metaverse 2023*, pp. 1–8, 2023, doi: 10.1109/SWC57546.2023.10449321.

[7] C. Wu and X. Guo, "A Novel Single Fuzzifier Interval Type-2 Fuzzy C-Means Clustering with Local Information for Land-Cover Segmentation," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 14, pp. 5903–5917, 2021, doi:10.1109/JSTARS.2021.3085606.

[8] H. Yadav, J. Singh, and A. Gosain, "Experimental Analysis of Fuzzy Clustering Techniques for Outlier Detection," *Procedia Comput. Sci.*, vol. 218, pp. 959–968, 2022, doi: 10.1016/j.procs.2023.01.076.

[9] M. H. F. Zarandi, S. Sotudian, and O. Castillo, "A new validity index for fuzzy-possibilistic c-means clustering," *Sci. Iran.*, vol. 28, no. 4, pp. 2277–2293, 2021, doi: 10.24200/SCI.2021.50287.1614.

[10] A. Saha and S. Das, "On the unification of possibilistic fuzzy clustering: Axiomatic development and convergence analysis," *Fuzzy Sets Syst.*, vol. 340, pp. 73–90, 2018, doi:10.1016/j.fss.2017.07.005.

[11] T. Xiao, Y. Wan, J. Chen, W. Shi, J. Qin, and D. Li, "Multiresolution-Based Rough Fuzzy Possibilistic," vol. 16, pp. 570–580, 2023.

[12] R. Devi, "Unsupervised Kernel-Induced Fuzzy Possibilistic C-Means Technique in Investigating Real-World Data," *J. Phys. Conf. Ser.*, vol. 2199, no. 1, 2022, doi: 10.1088/1742-6596/2199/1/012033.

[13] Z. Rustam, J. Pandelaki, D. A. Utami, R. Hidayat, and A. A. Ramli, "Comparison support vector machine and fuzzy possibilistic C-Means based on the kernel for knee osteoarthritis data classification," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 9, no. 6, pp. 2142–2146, 2019, doi:10.18517/ijaseit.9.6.9243.

[14] T. Thilagaraj and N. Sengottaiyan, "Implementation of fuzzy C-means and fuzzy possibilistic C-means algorithms to find the low performers using R-tool," *Int. J. Sci. Technol. Res.*, vol. 8, no. 8, pp. 1697–1701, 2019.

[15] S. Kushwaha and K. S. Jadon, "Improved performance using fuzzy possibilistic C-Means clustering algorithm in wireless sensor network," *Proc. - 2020 IEEE 9th Int. Conf. Commun. Syst. Netw. Technol. CSNT 2020*, pp. 134–139, 2020, doi: 10.1109/CSNT48778.2020.9115740.

[16] H. A. Gangadwala and R. M. Gulati, "Analysis of Web Usage Mining Using Various Fuzzy Techniques and Cluster Validity Index," *2022 1st Int. Conf. Electr. Electron. Inf. Commun. Technol. ICEEICT 2022*, pp. 1–7, 2022, doi:10.1109/ICEEICT53079.2022.9768580.

[17] T. Thilagaraj and N. Sengottaiyan, "Implementation of fuzzy possibilistic product partition C-means and modified fuzzy possibilistic C-means clustering to pick the low performers using R-tool," *Int. J. Recent Technol. Eng.*, vol. 8, no. 2, pp. 5942–5946, 2019, doi:10.35940/ijrte.B3580.078219.

[18] N. Tressa, "Customer-Based Market Segmentation using Clustering in Data mining," *2024 2nd Int. Conf. Intell. Data Commun. Technol. Internet Things*, pp. 687–691, 2024, doi:10.1109/IDCIoT59759.2024.10467258.

[19] S. Tokat, K. Karagul, Y. Sahin, and E. Aydemir, "Fuzzy c-means clustering-based key performance indicator design for warehouse loading operations," *J. King Saud Univ.- Comput. Inf. Sci.*,vol.34, no.8, pp. 6377–6384,2022, doi:10.1016/j.jksuci.2021.08.003.

[20] A. Wu, C. Deng, S. Member, and W. Liu, "Unsupervised Out-of-Distribution Object Detection via PCA-Driven Dynamic Prototype Enhancement," *IEEE Trans. Image Process.*, vol. PP, p. 1, 2024, doi:10.1109/TIP.2024.3378464.

[21] A. Hussein and A. Monfared, "Assessing Classical and Evolutionary Preprocessing Approaches for Breast Cancer Diagnosis," *2024 20th CSI Int. Symp. Artif. Intell. Signal Process.*, pp. 1–8, 2024, doi:10.1109/AISP61396.2024.10475310.

[22] H. Lee, Y. Jo, and J. Jeong, "Multivariate PCA-based Composite Criteria Evaluation Method for Anomaly Detection in Manufacturing Data," *2024 26th Int. Conf. Adv. Commun. Technol.*, pp. 1–9, doi:10.23919/ICACT60172.2024.10471960.

[23] A. K. Singh, S. Mitra, D. Chaudhuri, B. B. Chaudhuri, and M. P. Singh, "Optimization of Multi-Class Non-Linear SVM Image Classifier Using A Sobel Operator Based Feature Map and PCA," *3rd Int. Conf. Range Technol. ICORT 2023*, pp. 1–6, 2023, doi:10.1109/ICORT56052.2023.10249196.

[24] P. R. V. Terlapu *et al.*, "Optimizing Chronic Kidney Disease Diagnosis in Uddanam: A Smart Fusion of GA-MLP Hybrid and PCA Dimensionality Reduction," *Procedia Comput. Sci.*, vol. 230, no. 2023, pp. 522–531, 2023, doi:10.1016/j.procs.2023.12.108.

[25] P. Gupta, A. Varshney, and K. Suneetha, "Exploring Non-linear Dimensionality Reduction Methodology for Enhanced Target Identification from Hyper Spectral Data," *2024 Int. Conf. Optim. Comput. Wirel. Commun.*, pp. 1–6, 2024, doi:10.1109/ICOCWC60930.2024.10470745.

[26] D. Ö. Şahin, O. E. Kural, S. Akleylek, and E. Kılıç, "Permission-based Android malware analysis by using dimension reduction with PCA and LDA," *J. Inf. Secur. Appl.*, vol. 63, no. October, p. 102995, 2021, doi:10.1016/j.jisa.2021.102995.

[27] C. Liu *et al.*, "Partial least squares regression and principal component analysis: Similarity and differences between two popular variable reduction approaches," *Gen. Psychiatry*, vol. 35, no. 1, pp. 1–5, 2022, doi: 10.1136/gpsych-2021-100662.

[28] L. Yu and C. Zhou, "Determining the Best Clustering Number of K-Means Based on Bootstrap Sampling," *Proc. - 2nd Int. Conf. Data Sci. Bus. Anal. ICDSBA 2018*, pp. 78–83, 2018, doi:10.1109/ICDSBA.2018.00022.

[29] F. Nie, X. Dong, Z. Hu, R. Wang, and X. Li, "Discriminative Projected Clustering via Unsupervised LDA," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 34, no. 11, pp. 9466–9480, 2023, doi:10.1109/TNNLS.2022.3202719.

[30] A. Bosisio *et al.*, "Performance assessment of load profiles clustering methods based on silhouette analysis," *21st IEEE Int. Conf. Environ. Electr. Eng. 2021 5th IEEE Ind. Commer. Power Syst. Eur. EEEIC /I CPS Eur. 2021 - Proc.*, pp. 1–6, 2021, doi:10.1109/EEEIC/ICPSEurope51590.2021.9584629.