# A New Adaptive Mixture Distance-Based Improved Density Peaks Clustering for Gearbox Fault Diagnosis

Krishna Kumar Sharma, Ayan Seal, *Senior Member, IEEE*, Anis Yazidi, *Senior Member, IEEE*, and Ondrej Krejcar

***Abstract*— With the rapid development of sensors and mechanical systems, we produce an exponentially large amount of data daily. Usually, faults are prevalent in these sensory systems due to harsh operational conditions. Thus, detecting and diagnosing faults in the gearbox of mechanical systems are done by analyzing an exponentially large amount of data in the form of vibration signals and categorical features. However, the automatic fault detection method can match the increasing requirement for high-quality products in the course of intelligent manufacture. Thus, to acquire more distinguishable fault features under varied conditions, a new adaptive mixture distance-based simple and efficient density peaks clustering algorithm is proposed for handling mixed data as real-world datasets encompassing both numerical and categorical attributes. Our approach revolves around the concept of a sequence of the weighted exponential kernel using a symmetry-favored *c*-nearest neighbor to estimate the global parameter and the local density of each data point. Then, the initial clusters are extracted from a decision graph using an adaptive threshold parameter. The final step is to allocate the remaining data objects, if they are density reachable, to either of the initial groups. Thirteen UCI datasets and one real-world dataset from a mechanical system for gearbox defect diagnosis are employed to validate the proposed method. Five external and two internal evaluation criteria are considered to gauge how well the strategies are working. All of the findings indicate that the proposed method outperforms 13 other approaches.**

Manuscript received 10 July 2022; revised 16 September 2022; accepted 9 October 2022. Date of publication 21 October 2022; date of current version 9 November 2022. This work was supported in part by the SPEV project "Smart Solutions in Ubiquitous Computing Environments" (under ID: UHK-FIMSPEV-2022-2102), University of Hradec Kralove, Faculty of Informatics and Management, Czech Republic. The Associate Editor coordinating the review process was Dr. Xiaofeng Yuan. *(Corresponding author: Ayan Seal.)*

Krishna Kumar Sharma is with the Department of Computer Science and Informatics, University of Kota, Kota, Rajasthan 324005, India (e-mail: krisshna.sharma@gmail.com).

Ayan Seal is with the Department of Computer Science and Engineering, PDPM Indian Institute of Information Technology, Design and Manufacturing Jabalpur, Jabalpur, Madhya Pradesh 482005, India (e-mail: ayanseal30@ieee.org).

Anis Yazidi is with the Department of Computer Science, Oslo Metropolitan University (OsloMet), 0166 Oslo, Norway, also with the Department of Computer Science, Norwegian University of Science and Technology (NTNU), 7034 Trondheim, Norway, and also with the Department of Plastic and Reconstructive Surgery, Oslo University Hospital (OuS), 460167 Oslo, Norway (e-mail: anis.yazidi@oslomet.no).

Ondrej Krejcar is with the Center for Basic and Applied Science, Faculty of Informatics and Management, University of Hradec Kralove, 500 03 Hradec Kralove, Czech Republic, and also with the Malaysia-Japan International Institute of Technology (MJIIT), Universiti Teknologi Malaysia, Kuala Lumpur 54100 , Malaysia (e-mail: ondrej.krejcar@uhk.cz).

Digital Object Identifier 10.1109/TIM.2022.3216366

***Index Terms*— Density peaks clustering (DPC), mixed data (MD), S-distance, symmetric favored *c*-nearest neighbors (*c*-NN).**

## I. Introduction

APPROPRIATE fault detection of the gearbox and bearing will be advantageous for the rotary machine, as they are indispensable components of it. Typically, rotary machines operate under harsh conditions, for example, uncertain or driving loads, up/variable speeds, and material fatigue, which generate possibilities for faults in the gearbox and bearings [1], [2], [22]. Thus, it creates improper situations for the machines and may cause downtime, economic loss, and maintenance costs to the organizations [3], [4], [5]. Therefore, effective signal processing techniques can protect the gearbox and bearing from the unforeseen situations mentioned above. Generally, vibration signal analysis is an efficient and viable approach for detecting faults, as they have a high correlation with the states of machine parts and organizations [6], [7]. There are various learning methods, such as supervised classification and unsupervised clustering, for fault identification. In addition, the most fundamental exploratory, meta-learning data analysis method, is clustering, which splits a set of data objects, denoted as a feature or observation vector, into nonempty, mutually exclusive subsets, groups, or clusters, such that elements of the same group are similar to one another based on some similarity metrics, whereas members of different subsets are dissimilar [8]. Therefore, much consideration must be paid to finding the obscure but imperative information in the data, for example, insights, patterns, and rules. These primitive data have no class information that represents the type of unsupervised learning.

Some frequently employed clustering techniques in machine fault detection are *k*-means, hierarchical, affinity propagation, fuzzy *c*-means (FCM), and kernel spectral. Shuqing et al. [9] and Ramos et al. [10] used FCM for fault detection. However, it adopts the spherical distance data together with the specification and is only effective for homogeneous data distribution. An improved method built on FCM was Gustafson–Kessel. After combining the adaptive distance rule and the covariance matrix, it can handle data with subspace dispersion in any direction [11]. Gustafson–Kessel was applied in the fault detection of the roller bearing by Wang et al. [12].

Liu et al. [13] presented a $k$-means clustering-based fault identification technique for wind turbines. A fuzzy rule-based clustering approach was employed for the detection of anomalies in wind turbines [14]. It was also adopted for bearing fault detection [15]. An improved FCM clustering approach was applied for dissolved gas analysis-data-based transformer fault detection [16]. A hierarchical-based clustering method was adopted for the evaluation of the vibration level and interior noise of vehicles [17]. Only a spherical pattern dataset is suitable for fuzzy and Gustafson–Kessel algorithms, but data obtained from practical systems have a variety of structures and shapes. Consequently, a Gath–Geva was developed to enhance the results. It follows the fuzzy maximum likelihood estimator. It is appropriate for data from variant orientations [18], [19]. In [20], affinity propagation clustering was implemented with adaptive feature selection on vibration signals for the detection of bearing faults. Langone et al. [21] introduced a spectral clustering-based method to identify the normal and erroneous states of a machine. A sparse subspace clustering technique using a composite graph with a new distance was presented to diagnose faults in machines [22]. Fong et al. [23] designed a mean shift-based clustering approach for machinery diagnostics on vibration signals. Hou et al. [24] presented the fuzzy Gath–Geva clustering technique, linear discriminant analysis, and ensemble empirical mode decomposition to diagnose rolling bearing faults. Although the majority of the studies mentioned above on intelligent defect identification have shown useful findings, they still have some obvious flaws, which are given as follows.

1) Most clustering algorithms for fault diagnosis rely on the hypothesis that data comprise only numerical values.
2) Most previous works exploit the FCM to diagnose machine faults. It means that, in the case of high-dimensional data, the selection of the fuzzifier will be crucial, and it may be trapped in local minima. Generally, FCM-based fault diagnosis algorithms prefer overlapping and spherical-shaped vibration signal datasets.
3) Existing clustering-based fault diagnosis methods are not sufficiently general and rely on the input parameters and the number of clusters for fault types. Thus, there are situations where they fail due to quite complicated actual working conditions of the systems.

In this study, we explore the possibilities of using density-based clustering approaches, such as DENCLUE, OPTICS, and DBSCAN, for fault diagnosis to handle the limitations stated above because they are suitable for arbitrary-shaped clusters. Moreover, these algorithms can filter out noise from data [25]. However, they are parameter-dependent. In particular, DBSCAN relies on two parameters, i.e., the minimum number of data objects in a neighborhood (MinPts) and the radius of the neighborhood for a data object. The values of MinPts and the radius of the neighborhood are determined by users manually, which is intrinsically hard to fix [25]. Rodriguez and Laio [26] presented a density peaks clustering (DPC) algorithm to detect arbitrary-shaped clusters. Since then, DPC has received increased research attention over the past few years. Generally, the DPC algorithm assumes that the cluster's center is farther from other cluster centers and is in a zone with a higher local density than its neighbors. For each data object or point, the DPC algorithm computes the local density and the distance from locations of higher density. Cluster centers are positioned in the top-right corner of a decision graph that has been created. Finally, all the data objects are assigned to one of the cluster centers. However, reliable density estimation is a complex problem. In their seminal paper, Rodriguez and Laio [26] suggested estimating density irrespective of the dataset size. However, small datasets are affected by the cutoff distance while estimating local density [26]. The DPC algorithm's significant benefit is its capacity to locate nonspherical clusters without prior knowledge of the number of classes. The DPC does not involve an iterative process. However, DPC might not automatically determine the correct number of clusters.

A density reachable concept and a divide-and-conquer-based 3DC clustering algorithm were given in [27] and [28], respectively, to address the aforementioned problem. Yaohui et al. [28] and Du et al. [29] investigated the concept of $c$-nearest neighbors ($c$-NN) in DPC for estimating local density. However, asymmetric edges are given the same weight as symmetric ones in $c$-NN-based DPC. Moreover, evidence from the literature has shown that data objects with asymmetric edges may end up in different clusters [30]. Furthermore, cluster representatives are selected based on the decision graph using a parameter cutoff distance. Thus, inappropriate selection of parameters may lead to an inaccurate decision graph and, consequently, incorrect cluster representatives. Furthermore, most of the clustering algorithms work under the assumption of numerical or categorical attributes [25], [31], [32], [33], [34], [35], [36]. In reality, all datasets have categorical and numerical attributes, which is known as a mixed-attribute dataset. As we will explain later, clustering unlabeled-mixed datasets is thus a tedious task. To deal with the latter issue, some clustering algorithms transform the categorical attributes into numeric attributes by performing binary encoding. Then, the similarity between the transformed data objects is then determined using the Euclidean distance. However, the obtained distances cannot capture the original structures of categorical attributes.

Moreover, when it comes to categories with two possible values, an associated binary representation of them is meaningless and hard to interpret [37]. Hsu [38] presented a weighted distance tree structure with a distance hierarchy. However, domain knowledge is required for both the creation of distance hierarchies and the assignment of weights. The interested readers are referred to [26], [29], [37], [39], [40], [41], [42], [43], [44] for more information related to several similarity measures for mixed data (MD) clustering. However, there is a trend to introduce nonlinearity into similarity measures for clustering [31]. Thus, clustering a dataset involving mixed attributes is still a challenging task.

This article suggests a novel adaptive mixture distance (AD)-based DPC technique to diagnose mechanical system gearbox failures. The following are some imperative contributions made by the proposed approach.

1) The inherent pattern of categorical characteristics is squashed by most existing techniques, which convert categorical attributes into sets of binary features. In other words, transformed binary features have no use. Moreover, their values are difficult to comprehend [37]. Thus, an entropy-based distance is presented to categorical features of the 13 UCI datasets and one real-world dataset, which keeps the original pattern of categorical attributes without transforming their representation. The real-world dataset consists of nonstationary vibration signals from a mechanical system for gearbox fault diagnosis.
2) A novel AD metric is introduced in this study that utilizes a weight parameter to merge the two similarity metrics, S-distance and similarity index. The former is defined in the open cone of positive definite matrices and is based on the concept of S-divergence [8], [31]. It is considered to calculate the separation between two numerical properties of the datasets examined. The latter is employed to determine how far apart categorized features are from one another. If there are more categorical features than numerical attributes, the similarity index is given a higher weight and vice versa.
3) A relatively new local density metric is specially adopted to deal with the noise that may be produced in real time while recording nonstationary vibration signals from a mechanical system. The local density metric relies on a sequence of the weighted exponential kernel using a symmetry-favored $c$-NN (SFCNN). It is capable of overcoming the limitations of fixed $c$-NN. Moreover, it characterizes the implicit geometrical structures. Furthermore, it increases the space in the density between outliers and core objects, which helps in generating efficient and correct cluster representatives.
4) A new method for the selection of initial cluster centers is presented, which assures correct cluster centers even in the case of an unbalanced dataset and nonuniform distribution of classes.

The modified DPC based on AD (MDPC-AD) is implemented on a total of 13 UCI datasets and a real-world dataset for gearbox fault diagnosis of a mechanical machine. Five clustering validation indices, namely, accuracy (A), precision (P), recall (R), F-Score (F), and the Jaccard index (JI), are used to show the superiority of the MDPC-AD. However, abbreviated forms of the validation indices mentioned above will be used only in figures for better accommodation and presentation. Moreover, two internal validation indices, for example, average clustering error and ratio of separation and compactness, are also adopted in this study. According to the results, the MDPC-AD outranks 13 state-of-the-art (SOTA) approaches.

The remaining work consists of the following. Section II discusses pertinent related studies. In Section III, the proposed distance metric definition is discussed, followed by MDPC-AD. In Section IV, all experimental findings are presented. The work is finally concluded in Section V.

## II. Theoretical Foundation of DPC

### A. Notations

Let $O = \{O_1, O_2, \ldots, O_i, \ldots, O_n\}$ be a dataset of $n$ MD objects. Each data object $O_i \in \Re^{d=|\psi|+|\phi|}$, where $1 \leq i \leq n$, has $d$ number of features or attributes in total. However, each $O_i$ has $|\psi|$ and $|\phi|$ number of numerical $\psi$ and categorical $\phi$ attributes, respectively. Thus, $O_{i,l}^{\psi}$ is the $l$th numerical feature of $O_i^{\psi}$. Similarly, $O_{i,l}^{\phi}$ is the $l$th categorical attribute of $O_i^{\phi}$. The domain of $l$th categorical feature $\mathrm{dm}(H_l^{\phi}) = \{h_{l,1}, h_{l,2}, \ldots, h_{l,s_l}\}$ has $s_l$ discrete values, whereas domain of the $l$th numerical attribute $\mathrm{dm}(H_l^{\psi})$ is continuous. Therefore, each $O_i$ is a combination of categorical and numerical values and it is denoted by $[O_i^{\psi}, O_i^{\phi}] = [O_{i,1}^{\psi}, O_{i,2}^{\psi}, \ldots, O_{i,|\psi|}^{\psi}, O_{i,|\psi|+1}^{\phi}, \ldots, O_{i,d=\{|\psi|+|\phi|\}}^{\phi}]$.

### B. Density Peaks Clustering

Fundamentally, DPC identifies cluster representatives with a higher density in comparison to their neighbors, and cluster representatives are located at a relatively large distance from each other. The two main parameters of this method are the local density $\beta_i$ of each data object $O_i$ and the distance $\gamma_i$ from objects with greater densities. Furthermore, two hypotheses correspond to the cluster representatives: 1) cluster representatives are located in higher density areas and their neighbors have lower densities and 2) cluster representatives are in relatively distant positions from each other or at a relatively higher distance to the data objects of higher density. The detailed discussion of the computation of $\beta_i$ and $\gamma_i$ is given as follows.

Generally, the DPC algorithm works on numerical values and adopts the linear Euclidean distance function as a similarity measure for numerical attributes in the clustering analysis. The Euclidean distance $\lambda_e$ between two data objects $O_i$ and $O_j$ is defined by (1) with the assumption that data consist only of numerical attributes

$$\lambda_e^2(O_i, O_j) = \sum_{l=1}^{l=d} (O_{i,l} - O_{j,l})^2. \tag{1}$$

The local density of a data object, $O_i$, is represented by $\beta_i$ and is defined by the following equation:

$$\beta_i = \sum_j \exp\left(-\frac{\lambda_e^2(O_i, O_j)}{\alpha_t}\right) \tag{2}$$

where $\alpha_t$ denotes an adjustable variable, which controls the weight decrease rate. $\alpha_t$ is the single variable in (2), and it relies on choosing the average number of neighbors of all data objects in the dataset. Rodriguez and Laio [26] defined $\alpha_t$ as given in the following equation:

$$\alpha_t = \alpha_{\lceil \tau \rceil} \tag{3}$$

where $\alpha_{\lceil \tau \rceil} \in \{\alpha_1, \alpha_2, \ldots, \alpha_{\binom{n}{2}}\}$ and this set contains distances between every pair of two data objects in the dataset, arranged

ascendingly. Another parameter $\gamma_i$ is also computed using (4), which shows the minimum distance between the data object $O_i$ and other data objects with larger density

$$\gamma_i = \begin{cases} \min\limits_j \left(\lambda_e\left(O_i, O_j\right)\right), & \text{if } \exists \text{ s.t. } \beta_i < \beta_j \\ \max\limits_j \left(\lambda_e\left(O_i, O_j\right)\right), & \text{else.} \end{cases} \tag{4}$$

When $\beta_i$ and $\gamma_i$ for each data object have been computed, large values of $\beta_i$ and $\gamma_i$ are explored anomalously in this method to identify the cluster representatives. Based on this concept, cluster representatives are always located on the decision graph's top right side. Once cluster representatives are identified, the remaining data objects are assigned to the nearest cluster with a higher density.

## III. Proposed Method

It is clear from the previous discussion that there are still some limitations to DPC and its peer methods. Hence, the DPC algorithm is improved in this study by introducing a novel adaptive mixture similarity measure. A new method for estimating the density of data points is introduced. Moreover, we present a novel way to construct a decision graph. In this section, we provide the details of the proposed clustering algorithm, MDPC-AD, and its theoretical complexity analysis.

### A. Similarity Measure of Numerical Attributes

For revealing the natural cluster structure in a given dataset, which is a topic of active research, selecting an appropriate similarity/dissimilarity metric is essential. Since its inception, the proper selection of a similarity/difference metric has been a challenge. Recently, there has been an upsurge of interest in divergence-based nonlinear similarity measure [8], [31] for clustering analysis as this type of distance is susceptible to finding more appropriate complex cluster boundaries. Thus, nonlinear S-distance $\lambda_s$ is considered here for computing the distance between two numerical data objects $O_i^{\psi}$ and $O_j^{\psi}$ in the $|\psi|$-dimensional Euclidean space $\Re_+^{|\psi|}$ using (5) [31].

*Definition 1:* Define $\lambda_s : \Re_+^{|\psi|} \times \Re_+^{|\psi|} \to \Re_+ \cup \{0\}$ as

$$\lambda_s^2\left(O_i^{\psi}, O_j^{\psi}\right) = \sum_{l=1}^{|\psi|} \Big[\log\left(\left(O_{i,l}^{\psi} + O_{j,l}^{\psi}\right)/2\right) - \left(\log\left(O_{i,l}^{\psi}\right) + \log\left(O_{j,l}^{\psi}\right)\right)/2\Big]. \tag{5}$$

The fact that $f$ is an injective function with the definition $f : \Re_+^{|\psi|} \to \mathscr{M}_{|\psi|}$ ensures that the S-distance is well-defined. In particular, $\boldsymbol{O_i^{\psi}} = f(O_i^{\psi}) = \text{diag}((O_{i,1}^{\psi}, O_{i,2}^{\psi}, \ldots, O_{i,|\psi|}^{\psi}))$. In this case, $\mathscr{M}_{|\psi|}$ is a positive definite matrix with the dimensions $|\psi| \times |\psi|$. The notion of S-divergence [8], which is described mathematically by (6), is used to derive the S-distance

$$\lambda_s^2\left(\boldsymbol{O_i^{\psi}}, \boldsymbol{O_j^{\psi}}\right) = \log\left(\left|\frac{\boldsymbol{O_i^{\psi}} + \boldsymbol{O_j^{\psi}}}{2}\right|\right) - \frac{\log\left(\left|\boldsymbol{O_i^{\psi}}\right|\right) + \log\left(\left|\boldsymbol{O_j^{\psi}}\right|\right)}{2} \tag{6}$$

where $|\cdot|$ is a determinant of a matrix and $\lambda_s^2(\boldsymbol{O_i^{\psi}}, \boldsymbol{O_j^{\psi}}) = \lambda_s^2(f(O_i^{\psi}), f(O_j^{\psi}))$.

The S-distance satisfies all the metric properties. Moreover, it also obeys the property of Hadamard product. It is also neither Bregman divergence nor f-divergence. However, it is a Burbea–Rao divergence. Thus, it is convex on $\Re_+^{|\psi|}$. According to a prior study [8], when two data objects are close to the origin and have the same Euclidean distance, their S-distance is bigger than when they are far from the origin. The scope of this study does not include the various S-distance characteristics. To learn more about these features, interested readers are encouraged to explore [8], [31].

Now, the similarity between two data objects is computed using a monotonically decreasing spatial generalization exponential function [45]. Mathematically, the exponential function is defined by the following equation:

$$\chi_{\psi}\left(O_i^{\psi}, O_j^{\psi}\right) = \exp\left(\frac{-\left(\lambda_s\left(O_i^{\psi}, O_j^{\psi}\right)\right)^2}{2}\right) \tag{7}$$

where $\chi_{\psi} \in [0, 1]$. A value of $\chi_{\psi}$ close to 1 indicates that two data objects $O_i^{\psi}$ and $O_j^{\psi}$ are similar. On the other hand, a value of $\chi_{\psi}$ close to 0 indicates that two data objects $O_i^{\psi}$ and $O_j^{\psi}$ are highly dissimilar.

### B. Similarity Measure of Categorical Attributes

Now, it is time to calculate the similarity $\lambda_{\phi}$ between two data objects, namely, $O_i^{\phi}$ and $O_j^{\phi}$ having categorical features on $H_l^{\phi}$. Most of the existing approaches [39], [41], [43], [44] transform categorical attributes into sets of binary attributes, which squashes the native pattern of categorical features. In other words, converted binary features are purposeless, and their values are difficult to understand. Thus, an entropy-based distance is applied to categorical features, which keeps the original pattern of categorical features without transforming their representation in this study. First, the similarity between the $l$th feature of $O_i^{\phi}$ and $O_j^{\phi}$ is computed by the following equation:

$$\lambda_{\phi}\left(O_{i,l}^{\phi}, O_{j,l}^{\phi}\right) = \begin{cases} 1, & \text{if } O_{i,l}^{\phi} = O_{j,l}^{\phi} \\ 0, & \text{if } O_{i,l}^{\phi} \neq O_{j,l}^{\phi}. \end{cases} \tag{8}$$

Thus, the similarity between two data objects is estimated by summing the significance of each categorical attribute. Mathematically, it is defined by the following equation:

$$\chi_{\phi}\left(O_i^{\phi}, O_j^{\phi}\right) = \sum_{l=1}^{|\phi|} \omega_l \lambda_{\phi}\left(O_{i,l}^{\phi}, O_{j,l}^{\phi}\right) \tag{9}$$

where $\omega_l$ is known as the significance of the $l$th feature. The value of $\omega_l$ varies from 0 and 1 and $\sum_{l=1}^{|\phi|} \omega_l = 1$. The significance of the $l$th attribute is computed with the help of entropy in information theory by the following equation:

$$G_l^{\phi} = -\sum_{h_{l,q} \in \text{dm}\left(H_l^{\phi}\right)} p\left(h_{l,q}\right) \log\left(p\left(h_{l,q}\right)\right) \tag{10}$$

where $p(h_{l,q})$ represents the probability of $h_{l,q}$ feature and is estimated as $(\sum_{i=1}^{n} \lambda_\phi(O_{i,l}^{\phi}, h_{l,q})/n)$. In other words, it is a ratio of number of objects whose value is equal to $h_{l,q}$ of categorical feature $H_l^{\phi}$ to the total number of objects $n$ in a given dataset. It is clear from (10) that if the number of $s_l$ is very large, then the entropy of feature $H_l^{\phi}$ will also be large. However, this is not how things actually are. The entropy of a categorical feature is reformulated by (11) to lessen the impact of categorical characteristics having numerous unique or distinct values, such as an ID number

$$\mathscr{G}_l^{\phi} = -\frac{1}{s_l} \sum_{q=1}^{s_l} p\left(h_{l,q}\right) \log\left(p\left(h_{l,q}\right)\right). \tag{11}$$

Thus, the weight assigned to each categorical feature $H_l^{\phi}$ is computed by

$$\omega_l = \frac{\mathscr{G}_l^{\phi}}{\sum_{l=1}^{|\phi|} \mathscr{G}_l^{\phi}}. \tag{12}$$

The similarity measure of categorical attributes can be computed by combining (9) and (12), which is shown in the following equation:

$$\chi_\phi\left(O_i^{\phi}, O_j^{\phi}\right) = \sum_{l=1}^{|\phi|} \frac{\mathscr{G}_l^{\phi}}{\sum_{l=1}^{|\phi|} \mathscr{G}_l^{\phi}} \lambda_\phi\left(O_{i,l}^{\phi}, O_{j,l}^{\phi}\right). \tag{13}$$

### C. Similarity Measure for MD

The similarity between two data points $O_i$ and $O_j$ having $|\psi|$ number of numerical attributes and $|\phi|$ number of categorical features is computed by merging (7) and (13) with the help of the more importance concept of information theory, and the new equation is given by

$$\chi\left(O_i, O_j\right) = \frac{|\psi|}{|\psi| + |\phi|} \exp\left(\frac{-\lambda_s\left(O_i^{\psi}, O_j^{\psi}\right)^2}{2}\right) + \frac{|\phi|}{|\psi| + |\phi|} \sum_{l=1}^{|\phi|} \frac{\mathscr{G}_l^{\phi}}{\sum_{l=1}^{|\phi|} \mathscr{G}_l^{\phi}} \lambda_\phi\left(O_{i,l}^{\phi}, O_{j,l}^{\phi}\right). \tag{14}$$

The value of the similarity measure lies between 0 and 1 due to normalized coefficients. Generally, a DPC algorithm requires a distance function instead of a similarity measure. Hence, a logarithmic function is applied to the negative exponent of (14) as shown in (15). If two data objects are similar, then the distance would be smaller

$$\lambda_m\left(O_i, O_j\right) = \log\left(\chi\left(O_i, O_j\right)^{-1}\right). \tag{15}$$

### D. Local Density Metric

In this section, we present: 1) a new local density metric based on SFCNN; 2) a new method to initialize the cluster centers; and 3) a way to group density-reachable clusters.

For estimating the local density of a data object $X_i$ in a set of data, an SFCNN graph is built in this study instead of a conventional $c$-NN graph since it is more resistant to noise and outliers. Fig. 1 is used to explain the distinction between a standard $c$-NN graph and one that favors symmetry. The graph's symmetric edges have heavier weights than its asymmetric edges because the locations they connect are located in the same submanifold [30], [46]. The underlying manifold characteristics of the data space can also be used to explain the SFCNN graph. It can describe implicit geometrical structures. Moreover, it also increases the space in the density between outliers and core objects, which helps in generating efficient and correct cluster representatives. Generally, density metrics consider Gaussian kernels to estimate the local density values. Data objects in DPC are represented as points in a space, where cluster representatives are always on the top-right part of the decision graph. Once the local density $\beta_i$ and minimum distance $\gamma_i$ from data points of higher density are calculated for each data object, cluster representatives are identified by searching the large parameters $\beta_i$ and $\gamma_i$ for anomalies. The parameter $\alpha_t$ determines the average number of neighbors of all data objects in a dataset. The value of $\alpha_t$ is computed by (16), which depends on the value of "$c$." Based on this idea, cluster centers for the Statlog Heart [47] sample dataset is estimated, and they can be seen in the top-right quadrant of the decision graph in Fig. 2. Here, circles with red color are the initial cluster representatives that are characterized relatively by the higher distance, $\gamma_i$, and larger density, $\beta_i$. They are computed based on the threshold value, $\alpha_t = 0.54$ (dashed line). After the selection of the cluster representatives, the remaining data objects are assigned to the nearest cluster with a higher density. A decision graph assists in taking a decision. The decision graph is a plot of $\gamma_i$ as a function of $\beta_i$ for each data object

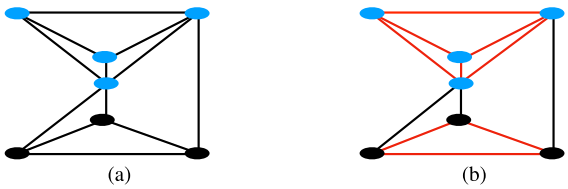


Fig. 1. (a) 3-NN graph's differences from (b) symmetry-favored 3-NN graphs (red edges show higher edge weights).

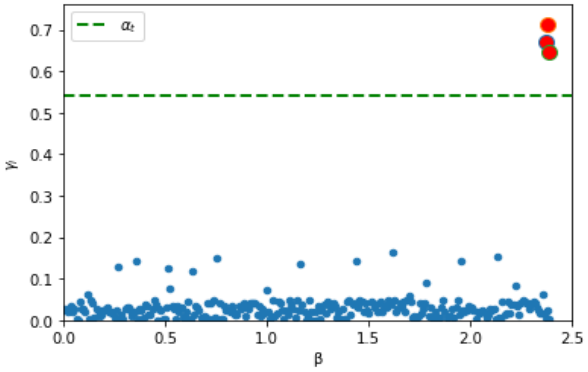


Fig. 2. Decision graph of Statlog Heart dataset with $\alpha_t = 0.54$.

$$\alpha_t = \upsilon^c + \sqrt{\frac{\sum_{i=1}^{n} \left(\gamma_i^c - \upsilon^c\right)^2}{n-1}} \tag{16}$$

where $\gamma_i^c$ is the distance between SFCNN and a data object $i$, defined as $\gamma_i^c = \max_{j \in \text{SFCNN}_i}(\lambda_m^{i,j})$, and $\upsilon^c$ is the average

value of $\gamma_i^c$ and is computed by (17). $\text{SFCNN}_i$ is a set of data objects in an SFCNN to data object $i$

$$v^c = \frac{\sum_{i=1}^{n} \gamma_i^c}{n}. \tag{17}$$

The second part in the right-hand side of (16) represents the standard deviation of distance calculated between each data object and its corresponding SFCNN. The local densities can be estimated by the following equation:

$$\beta_i = \sum_{j \in \text{SFCNN}_i} \exp\left(-\frac{\left(\lambda_m^{i,j}\right)^2}{(\alpha_t)^2}\right). \tag{18}$$

Equation (18) illustrates the distribution information of the SFCNN of a data object $i$ and uses $\alpha_t$ to estimate the local density $\beta_i$. Equation (18) considers the sum of all distances using an exponential kernel. The previous studies [28] reveal that the value of $c$ in an SFCNN graph has a significant impact while estimating density, and it was fixed to 5 because 2-NN, 3-NN, and 4-NN may be close to normal data objects. Thus, an enhanced local density is proposed by combining a fixed SFCNN and a weighted sequence as shown in the following equation:

$$\beta_i = \sum_{j \in \text{SFCNN}_i} \exp\left(-\frac{\left(\lambda_m^{i,j}\right)^2}{(\alpha_t)^2}\right) + \sum_{j \notin \text{SFCNN}_i \text{ and } j \neq i} \frac{\exp\left(-\left(\lambda_m^{i,j}\right)^2\right)}{\max\limits_{j' \in \text{SFCNN}_i}\left(\lambda_m^{j',j}\right)}. \tag{19}$$

The first part of (19) takes care of the symmetry-favored 5-NN estimation, which is inherited from (18). On the other hand, the second part of (19) sums the weighted Gaussian kernel sequence. This second part is a complement to the first part and compensates for the clustering performance by overcoming the limitations of fixed $c$-NN in density estimation. The weights in the second part have a lesser value in the case of data objects away from the $c$-NN and a higher weightage near the $c$-NN.

In this work, the DPC algorithm is enhanced by considering some of the concepts of DBSCAN and OPTICS, which are given as follows.

*Definition 2 (Core Distance $\gamma^e$ of a Cluster $C^e$):* $\gamma^e$ of a cluster $C^e$ is computed by the following equation:

$$\gamma^e = \frac{\sum_{O_i \in C^e} \lambda_m(\text{CP}^e, O_i)}{|C^e|} \tag{20}$$

where $|C^e|$ represents the cardinality of a cluster set $C^e$ and $\text{CP}^e$ is the cluster center of $C^e$. $\gamma^e$ of a $C^e$ is the average of distances between all the data points belonging to $C^e$ and $\text{CP}^e$.

*Definition 3 (Boundary-Data-Object-Pair Set $\rho^{x,y}$ Between Two Clusters, Namely, $C^x$ and $C^y$):* $\rho^{x,y}$ between $C^x$ and $C^y$ is expressed as follows:

$$\rho^{x,y} = \left\{\left(O_i, O_j\right) | \lambda_m\left(O_i, O_j\right) < \min\left(\gamma^x, \gamma^y\right), O_i \in C^x, O_j \in C^y\right\} \tag{21}$$

where $\rho^{x,y}$ is symmetric in nature.

*Definition 4 (Border Density $\beta_\rho^e$ of Cluster $C^e$):* $\beta_\rho^e$ of $C^e$ is computed by the following equation:

$$\beta_\rho^e = \max_{(O_i, O_j) \in \rho^x} \frac{\left(\beta_i + \beta_j\right)}{2} \tag{22}$$

where $\rho^x$ consists all boundary-data-objects-pairs between $C^x$ and other clusters such that $\rho^x = \cup_{y \neq x} \rho^{x,y}$.

*Definition 5 (Density Directly Reachable):* In terms of border density, a cluster $C^x$ is density directly reachable from another cluster $C^y$ if the following conditions hold.

1) $\rho^{x,y} \neq \{\text{Null}\}$.
2) $\exists (O_i, O_j) \in \rho^{x,y}$, $\beta_i < \beta_\rho^x$ and $\beta_j < \beta_\rho^y$

It also satisfies the symmetric property.

*Definition 6 (Density Reachable):* If there is a path connecting two clusters $C^x$ and $C^y$ such that $C^x = C^1, C^2, \ldots, C^n = C^y$, each $C^i$ is directly reachable to $C^{i+1}$, then the two clusters are said to be density reachable to one another. Moreover, it obeys the symmetric as well as transitive properties.

### *E. Improved DPC Algorithm and Its Complexity*

In this section, the essential details of the MDPC-AD are discussed with an analysis of its complexity. Algorithm 1 is a logical step-by-step analysis of the MDPC-AD. The algorithm of the MDPC-AD is presented to make it easy for the reader to identify the process, major decision points, and variables necessary to implement MDPC-AD.

Fig. 3 is employed to illustrate the detailed processes of the MDPC-AD on a particular dataset named Wine [47] consisting of numerical attributes only. The first two principal components of each data object of the Wine dataset are obtained using PCA and are shown on a 2-D plane using pink color in Fig. 3(a). Here, a dataset consisting of numerical features is considered because PCA can only work on numerical attributes to generate principal components. Three density peaks in the top-right corner are automatically recognized as cluster representatives in Fig. 3(b), which shows the decision graph of $\gamma_i$ as a function of $\beta_i$ for each data object and data objects in red are initial cluster representatives above the threshold $\alpha_t$ and threshold $\alpha_t$ is displayed as green dashed line. The remaining data objects are then assigned to the closest clusters to obtain the corresponding cluster, as shown in Fig. 3(c). Finally, a grouping of the density reachable clusters is performed and the final obtained clusters are shown in Fig. 3(d).

The following factors are used to discuss how time-consuming the MDPC-AD is. First, the distance between data objects is calculated with complexity $O(n^2 E)$, where $E$ is the time required to compute $\lambda_m()$ between two data objects and $n$ represents the number of data objects in the dataset. Later, sorting of distance vector will require $O(n^2 \log(n))$ complexity. An SFCNN graph will take $O(cn)$ times for calculation of $\beta_i$, where $c$ is smaller than $n$. The calculation of distance $\gamma_i$ for each data object requires $O(n^2)$ steps. Furthermore, initial representatives for clusters are selected and the assignment of data objects to clusters is completed in $O(n^2)$ times. The calculation of core distance $\gamma^e$ and border density $\beta_\rho^e$ will take only $O(n)$. The estimation

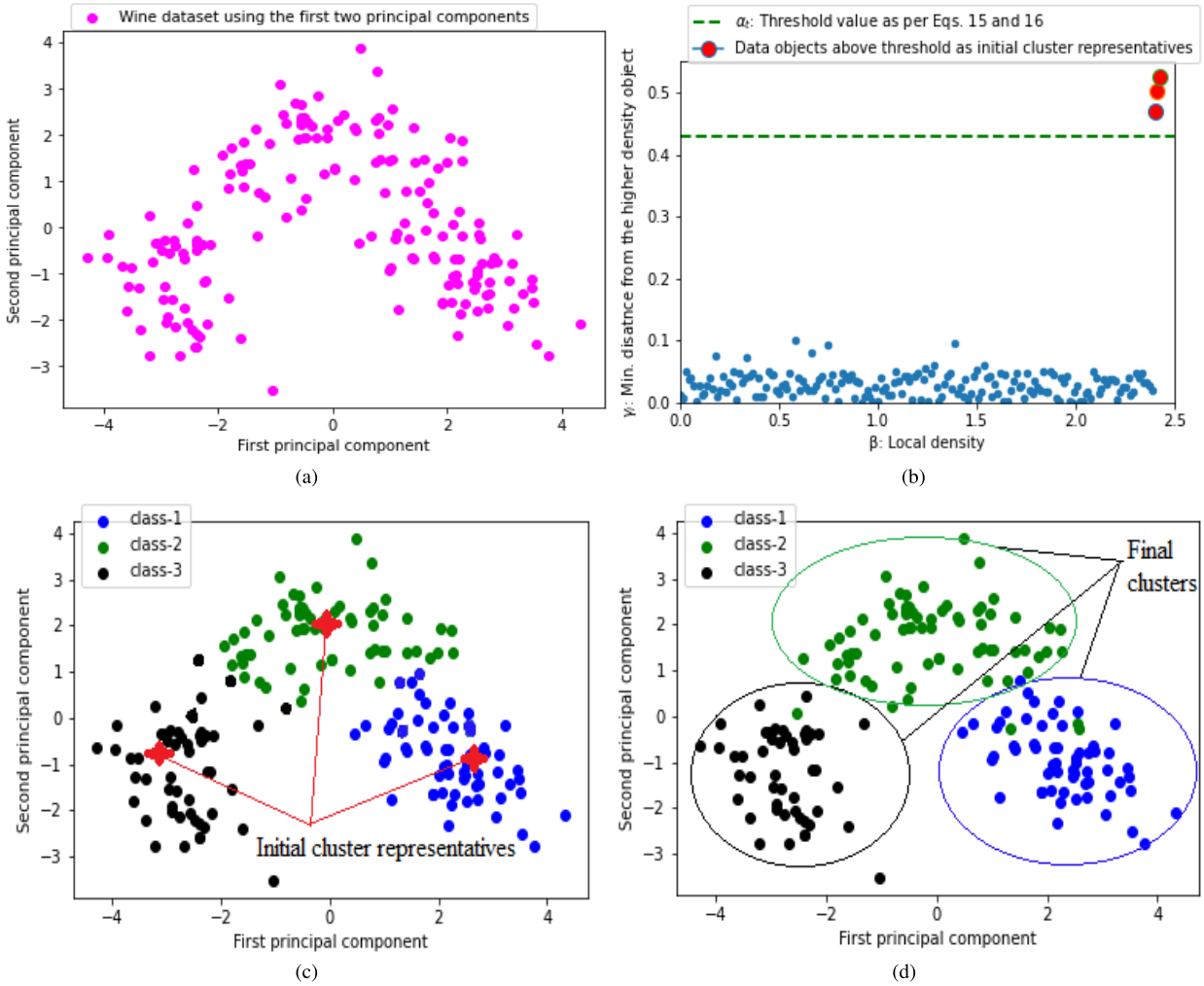


Fig. 3. Illustration of the MDPC AD that has been proposed. (a) Visualization of the wine dataset using the first two principal components, showing the first and second corresponding vectors of the data matrix along the axes. (b) Decision graph for the wine dataset in (a). (c) Clustering result after nearest cluster assignment. (d) Clustering result after grouping of density reachable clusters.

of boundary-data-object-pair sets will require approximately $O(n^2)$ steps. In conclusion, the time complexity of the MDPC-AD is $O(n^2 \log(n))$.

## IV. Experimental Results and Discussion

This work is done on a laptop running Windows 10 with an Intel[1] Core[2] i7-2620M CPU clocked at 2.70 GHz and 8 GB of RAM using the Spyder 3.2.8 Python development environment. This study does not include I/O costs.

### A. Experimental Setup and Dataset Description

*1) Dataset #1:* Thirteen well-known real-world datasets from the UCI repository are considered in this study. Table I contains some statistical information, such as name, type, the number of clusters ($k$), the total number of features ($d$), the number of numerical features ($F_\psi$), the number of categorical features ($F_\phi$), and the total number of samples ($n$) from these datasets. Interested readers may discover more details regarding these datasets in [47].

[1]Registered trademark.
[2]Trademarked.

TABLE I
Statistics of UCI Datasets Used in This Study

| Dataset | Type | k | d | $F_\psi$ | $F_\phi$ | n |
|---|---|---|---|---|---|---|
| Congressional Voting (D1) | | 2 | 16 | 0 | 16 | 232 |
| LED Display Domain (D2) | Categorical | 10 | 7 | 0 | 7 | 500 |
| Soybean (D3) | | 4 | 35 | 0 | 35 | 47 |
| Breast cancer dataset (D4) | | 2 | 30 | 30 | 0 | 569 |
| Iris dataset (D5) | Numerical | 3 | 4 | 4 | 0 | 150 |
| Shuttle dataset (D6) | | 7 | 9 | 9 | 0 | 43500 |
| Wine dataset (D7) | | 3 | 13 | 13 | 0 | 178 |
| Zoo (D8) | | 7 | 15 | 1 | 14 | 101 |
| Acute (D9) | | 2 | 7 | 1 | 6 | 120 |
| Statlog Heart (D10) | | 2 | 13 | 5 | 8 | 270 |
| KDDCUP-99 (D11) | Mixed | 4 | 41 | 26 | 15 | 2000 |
| Lymphography (D112) | | 4 | 18 | 3 | 15 | 148 |
| Australian Credit Approval (D13) | | 2 | 14 | 6 | 8 | 690 |

*2) Dataset #2:* The proposed method's efficiency is also tested on machine fault diagnostics, with data containing three categories of gearbox problems, such as missing teeth, tooth wear, and root faults. As shown in Fig. 5, a test rig is used to identify the faults in a gearbox that are shown in Fig. 4 [22], [23]. The three defective gears, as well as one healthy gear, are installed in the test setup. Then, using a controller, it is powered at its rotational speed by a regulated motor. A brake is used at the shaft's end location to provide

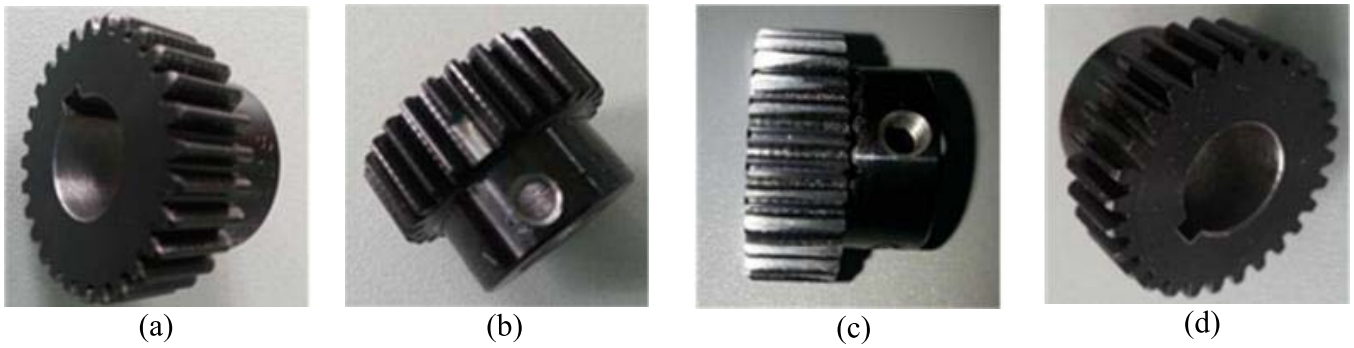


Fig. 4. Faults of the gears: (a) normal, (b) missing teeth, (c) wear in teeth, and (d) fault in root.

**Algorithm 1** Proposed DPC Algorithm

**Require:** $O = \{O_1, \ldots, O_i, \ldots, O_n\}$ ▷ where $O_i \in \Re^{d=|\psi|+|\phi|}$

**Ensure:** $C = \{C^1, C^2, \ldots, C^k\}$ ▷ a set of resultant clusters

1: Calculate distance matrix and parameter $\alpha_t$ using Eqs. 15 and16, respectively
2: Calculate $\beta_i$ and $\gamma_i$ for all data objects $O_i \in O$ by Eqs. 16 to 19.
3: Choose all data objects whose $\gamma_i$ is larger than the $\alpha_t$ cutoff distance in the decision graph and set $k'$ initial representatives of clusters as $CP = \{CP^i | 1 \leq i \leq k'\}$ and remaining data object will be $O' = O - CP$.
4: **for all** $O_i \in O'$ **do**
label = $\min_{CP^j \in C^j \& 1 \leq j \leq k'} \{\lambda_m(O_i, CP^j)\}$; ▷ nearest cluster
$C^{label} \leftarrow C^{label} \cup O_i$
5: **end for**
6: Compute the core distance $\gamma^e$ and boundary density $\beta_\rho^e$ of each cluster $e$ using Eqs. 20 to 22.
7: **repeat** ▷ Group all density-reachable clusters
8: **for all** $C^i \in C$ **do**
9: **for all** $C^j \in C - C^i$ **do**
10: **if** $C^i$ and $C^j$ satisfy Defs. 5 and 6 **then**
11: $C^i \leftarrow C^i \cup C^j$ and update set $C$
12: **end if**
13: **end for**
14: **end for**
15: **until** Grouping of density reachable clusters
16: Return $C = \{C^1, \ldots, C^k\}$ as the set of the clusters.

a load to the system. As shown in Fig. 5, an experiment for intelligent fault detection was carried out. The gearbox was not loaded, and the motor's speed was set to 1800 r/min. Three acceleration sensors that were fixed in the housing's vertical, horizontal, and axial directions and connected to its right end were employed to collect vibration data at a sampling rate of 12.8 kHz. However, categorical data, such as the number of cylinders, the forwarding gear values, and the number of carburetors, are discontinuous parameters. Four separate conditions, namely, tooth wear, root defect, missing teeth, and healthy, were used to collect vibration signals. As shown in Fig. 6, the original vibration signal is split into 90 segments, each of which contains 5023 samples.

The MDPC-AD method is presented to diagnose machine faults via vibration signals, and categorical features are obtained to determine the state variation due to faults [22], [23]. As discussed in Section II-A, $O = \{O_1, O_2, \ldots, O_i, \ldots, O_n\}$ is a dataset of $n$ MD vectors. Each data vector $O_i \in \Re^{d=|\psi|+|\phi|}$, where $1 \leq i \leq n$, has a total of $d$ features or attributes. However, each $O_i$ has a $|\psi|$ and $|\phi|$ number of signal features $\psi$ and categorical $\phi$ attributes, respectively. One assumption is made in this application that the number of data objects in each cluster is equal. Let $k$ be the number of clusters, and the data objects from each cluster are $n/k$. The MDPC-AD method is performed for the diagnosis of a faulty gearbox, as discussed in Algorithm 1.

### B. Evaluation Metrics

Accuracy is one of the most commonly reported evaluation measures. It describes the percentage of accurate clustering outcomes among all the outcomes a machine learning algorithm produces. It is an intuitive and straightforward evaluation metric. A machine learning algorithm is better and more preferable if its percentage accuracy is near 100. On the other hand, depending just on accuracy for unbalanced data can be deceptive. In this situation, in addition to accuracy, other assessment measures, including precision, recall, F-Score, and JI, may be considered to determine how effective a model is [25]. Moreover, two internal validation indices, for example, the average clustering error and the ratio of separation to compactness, are also adopted in this study.

### C. Computational Protocol

This study compares the performance of the proposed approach, MDPC-AD, with 13 SOTA methods.

1) *K-PC:* $k$-prototypes clustering algorithm for mixed datasets [39].
2) *EK-PC:* An evolutionary $k$-prototypes clustering algorithm for mixed-type datasets [41].
3) *KL-FCM-GM:* An FCM-type clustering algorithm for mixed datasets with a probabilistic dissimilarity function [42].
4) *FK-PC:* A fuzzy $k$-prototypes clustering algorithm for MD [43].
5) *IK-PC:* An improved $k$-prototypes clustering algorithm for mixed numerical and categorical data [44].

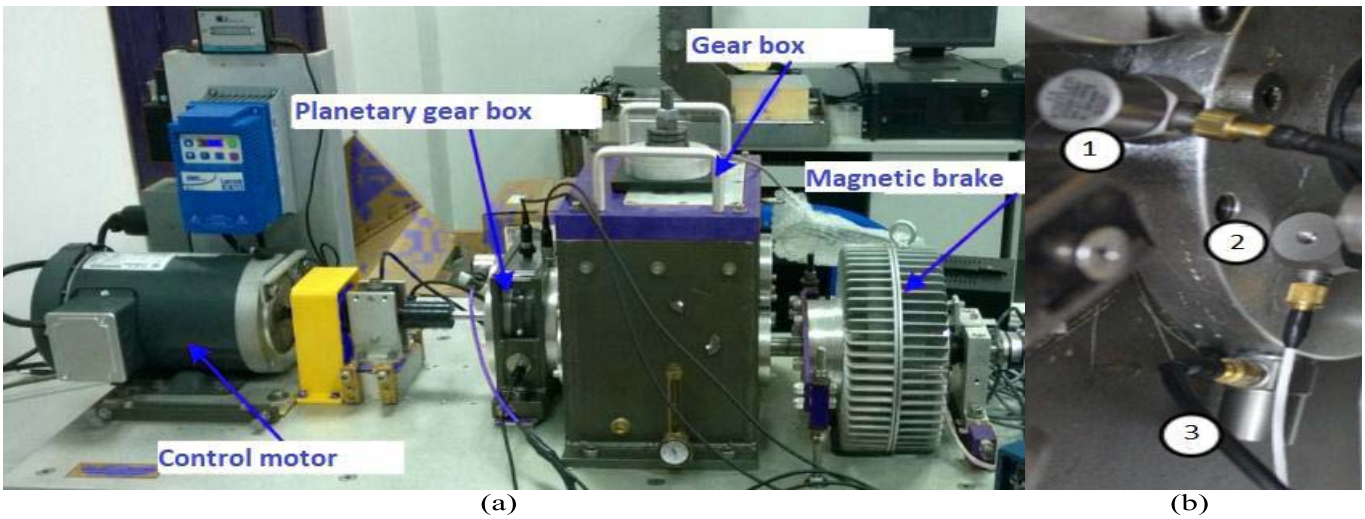


Fig. 5. (a) Gear test rig and (b) accelerometers fixed in the vertical, axial, and horizontal directions.

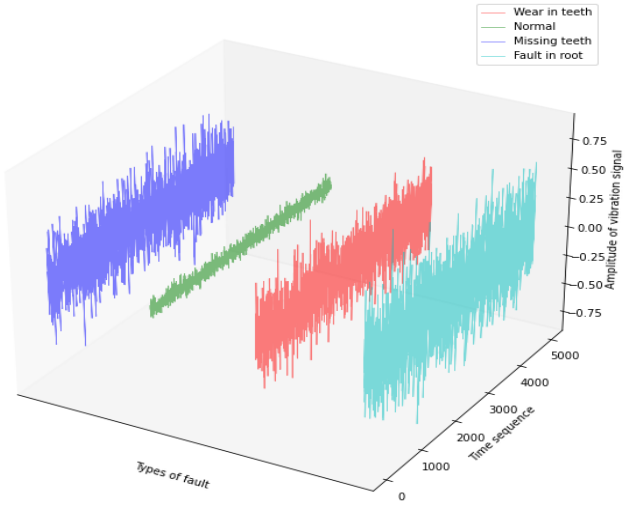


Fig. 6. Vibration signal in four different conditions of the gear from the experiment.

6) *CAVE:* A clustering algorithm based on variance and entropy for mixed datasets [40].
7) *SBAC:* A similarity-based agglomerative clustering algorithm for data with mixed features [37].
8) *SpectralCAT:* Categorical spectral clustering for numerical and nominal data [48].
9) *DKFCM:* A density-oriented kernel FCM algorithm for fault diagnosis [10].
10) *DPC-MD:* A novel DPC method for MD using a distance for MD [29].
11) *PE-EEMD-GG:* A method based on permutation entropy, ensemble empirical mode decomposition, and the Gath–Geva clustering method for bearing fault diagnosis [24].
12) *CG-SSC:* A composite graph-based sparse subspace clustering method for machine fault diagnosis [22].
13) *MSC:* A mean shift clustering-based approach with a spectral preprocessing technique for machinery diagnostics [23].

Since the researchers have not given their works a name, appropriate nomenclatures for these methodologies are employed. The scope of this study does not include a thorough description of these techniques. However, we use the precise procedures outlined in the original papers. As a result, interested readers are directed to the source works for more information.

### D. Results and Comparison

In this study, a total of nine experiments are conducted to validate the MDPC-AD. The first eight experiments are carried out on the UCI datasets using five external validation indices. The last experiment is performed to identify the faults in the gearbox of a rotary mechanical machine with the help of two internal validation measures.

*1) Experiment on Categorical Datasets:* In the first experiment, the MDPC-AD is executed on datasets, namely, D1, D2, and D3, having categorical attributes only. These datasets do not possess numerical features. The second part of (14), followed by (15), is considered while computing distance. The clustering report obtained by the MDPC-AD is noted in the last column of Table II. The clustering reports generated by existing methods are also included in Table II. The best clustering report produced by a method is marked by bold characters. All the results of Table II demonstrate that the MDPC-AD outperforms all the above-discussed 13 SOTA methods. However, the performance of DPC-MD and MSC on D3 is the same as that of the MDPC-AD.

*2) Experiment on Numerical Datasets:* In the second experiment, the MDPC-AD is implemented on datasets, namely, D4, D5, D6, and D7, having only numerical attributes. These datasets do not have categorical features. Fig. 7 shows the first two principal components of D4, D5, D6, and D7 plotted in 2D planes. It is clear from Table I and Fig. 7 that each dataset consists of a varying number of data points. Moreover, they have arbitrary shape clusters. Now, the first part of (14), followed by (15), is employed while computing distance. The clustering reports of all the 14 methods, including the MDPC-AD, are reported in Table III. The best performance

TABLE II

CLUSTERING REPORTS ON D1, D2, AND D3 USING FIVE VALIDATION INDICES

| Database | Validation Index | $K-PC$ | $EK-PC$ | $KL-FCM-GM$ | $FK-PC$ | $IK-PC$ | $CAVE$ | DKFCM | PE-EEMD | SBAC | CG-SSC | MSC | DPC-MD | SpectralCAT | MDPC-AD |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| D1 | Precision | 0.8604 | 0.8703 | 0.4795 | 0.8852 | 0.9078 | 0.7689 | 0.8660 | 0.8616 | 0.5899 | 0.9177 | 0.9248 | 0.9149 | 0.8444 | **0.9321** |
| | Recall | 0.8563 | 0.8646 | 0.4771 | 0.8788 | 0.8971 | 0.7543 | 0.8587 | 0.8506 | 0.5917 | 0.9009 | 0.9086 | 0.9236 | 0.8487 | **0.9290** |
| | F-Score | 0.8583 | 0.8675 | 0.4783 | 0.8820 | 0.9024 | 0.7615 | 0.8624 | 0.8561 | 0.5908 | 0.9092 | 0.9167 | 0.9193 | 0.8465 | **0.9306** |
| | Jaccard Index | 0.7499 | 0.7630 | 0.3188 | 0.7838 | 0.8176 | 0.6021 | 0.7561 | 0.7324 | 0.4174 | 0.8243 | 0.8387 | 0.8471 | 0.7306 | **0.8697** |
| | Accuracy | 0.8578 | 0.8664 | 0.5043 | 0.8793 | 0.9009 | 0.7543 | 0.8630 | 0.8462 | 0.5905 | 0.9052 | 0.9138 | 0.9181 | 0.8448 | **0.9310** |
| D2 | Precision | 0.6538 | 0.6453 | 0.5817 | 0.5764 | 0.5279 | 0.6588 | 0.6840 | 0.6379 | 0.5181 | 0.6583 | 0.7187 | 0.7039 | 0.7029 | **0.7217** |
| | Recall | 0.6392 | 0.6333 | 0.5937 | 0.5649 | 0.5471 | 0.6400 | 0.6867 | 0.6230 | 0.5414 | 0.6478 | 0.7107 | 0.6918 | 0.6971 | **0.7151** |
| | F-Score | 0.6464 | 0.6392 | 0.5876 | 0.5706 | 0.5373 | 0.6493 | 0.6854 | 0.6304 | 0.5295 | 0.6530 | 0.7147 | 0.6978 | 0.7000 | **0.7184** |
| | Jaccard Index | 0.4677 | 0.4587 | 0.4045 | 0.3970 | 0.3597 | 0.4688 | 0.5125 | 0.4499 | 0.3471 | 0.4757 | 0.5447 | 0.5306 | 0.5309 | **0.5565** |
| | Accuracy | 0.6340 | 0.6240 | 0.5640 | 0.5540 | 0.5140 | 0.6340 | 0.6780 | 0.6168 | 0.4940 | 0.6420 | 0.6960 | 0.6920 | 0.6840 | **0.7140** |
| D3 | Precision | 0.7829 | 0.8681 | 0.8681 | 0.8341 | 0.9101 | 0.8410 | 0.9152 | 0.9118 | 0.6575 | 0.9417 | **1.0000** | **1.0000** | 0.8902 | **1.0000** |
| | Recall | 0.8222 | 0.8881 | 0.8881 | 0.8503 | 0.9330 | 0.8036 | 0.9201 | 0.8849 | 0.6914 | 0.9476 | **1.0000** | **1.0000** | 0.9048 | **1.0000** |
| | F-Score | 0.8021 | 0.8780 | 0.8780 | 0.8421 | 0.9214 | 0.8219 | 0.9176 | 0.8982 | 0.6740 | 0.9446 | **1.0000** | **1.0000** | 0.8974 | **1.0000** |
| | Jaccard Index | 0.6597 | 0.7728 | 0.7728 | 0.7175 | 0.8469 | 0.6774 | 0.8452 | 0.7967 | 0.5144 | 0.8914 | **1.0000** | **1.0000** | 0.8015 | **1.0000** |
| | Accuracy | 0.7872 | 0.8723 | 0.8723 | 0.8298 | 0.9149 | 0.8085 | 0.9153 | 0.8723 | 0.6596 | 0.9362 | **1.0000** | **1.0000** | 0.8936 | **1.0000** |

TABLE III

CLUSTERING REPORTS ON D4, D5, D6, AND D7 USING FIVE VALIDATION INDICES

| Database | Validation Index | $K-PC$ | $EK-PC$ | $KL-FCM-GM$ | $FK-PC$ | $IK-PC$ | $CAVE$ | DKFCM | PE-EEMD | SBAC | CG-SSC | MSC | DPC-MD | SpectralCAT | MDPC-AD |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| D4 | Precision | 0.8488 | 0.8238 | 0.7840 | 0.8426 | 0.8303 | 0.8617 | 0.8009 | 0.7471 | 0.7725 | 0.8633 | 0.9040 | 0.8826 | 0.8072 | **0.9300** |
| | Recall | 0.8424 | 0.8155 | 0.7821 | 0.8361 | 0.8247 | 0.8540 | 0.7970 | 0.7403 | 0.7650 | 0.8715 | 0.8918 | 0.8769 | 0.8134 | **0.9384** |
| | F-Score | 0.8456 | 0.8196 | 0.7831 | 0.8394 | 0.8275 | 0.8578 | 0.7989 | 0.7437 | 0.7687 | 0.8674 | 0.8979 | 0.8797 | 0.8103 | **0.9342** |
| | Jaccard Index | 0.7301 | 0.6908 | 0.6432 | 0.7200 | 0.7040 | 0.7443 | 0.6601 | 0.5872 | 0.6190 | 0.7631 | 0.8074 | 0.7841 | 0.6787 | **0.8751** |
| | Accuracy | 0.8453 | 0.8190 | 0.7838 | 0.8383 | 0.8278 | 0.8541 | 0.7957 | 0.7417 | 0.7663 | 0.8600 | 0.8946 | 0.8805 | 0.8102 | **0.9350** |
| D5 | Precision | 0.8425 | 0.8633 | 0.8770 | 0.9048 | 0.8340 | 0.8879 | 0.8822 | 0.8342 | 0.8179 | 0.9532 | 0.9583 | 0.9407 | 0.9227 | **0.9738** |
| | Recall | 0.8545 | 0.8715 | 0.8840 | 0.8882 | 0.8437 | 0.8932 | 0.8774 | 0.8374 | 0.8246 | 0.9503 | 0.9534 | 0.9322 | 0.9187 | **0.9735** |
| | F-Score | 0.8484 | 0.8674 | 0.8805 | 0.8964 | 0.8388 | 0.8906 | 0.8798 | 0.8358 | 0.8212 | 0.9518 | 0.9558 | 0.9364 | 0.9207 | **0.9736** |
| | Jaccard Index | 0.7066 | 0.7631 | 0.7798 | 0.7929 | 0.7184 | 0.8015 | 0.7786 | 0.7146 | 0.6883 | 0.9035 | 0.9117 | 0.8760 | 0.8521 | **0.9481** |
| | Accuracy | 0.8267 | 0.8600 | 0.8733 | 0.8800 | 0.8333 | 0.8867 | 0.8757 | 0.8362 | 0.8133 | 0.9467 | 0.9533 | 0.9333 | 0.9133 | **0.9733** |
| D6 | Precision | 0.6647 | 0.7266 | 0.6206 | 0.7048 | 0.7139 | 0.7569 | 0.8769 | 0.8488 | 0.5873 | 0.8289 | 0.8328 | 0.8191 | 0.7048 | **0.8800** |
| | Recall | 0.6272 | 0.6617 | 0.5846 | 0.6449 | 0.6511 | 0.6814 | 0.8635 | 0.8424 | 0.5610 | 0.8546 | 0.8549 | 0.8378 | 0.6449 | **0.8639** |
| | F-Score | 0.6454 | 0.6927 | 0.6021 | 0.6735 | 0.6811 | 0.7172 | 0.8702 | 0.8456 | 0.5739 | 0.8416 | 0.8437 | 0.8283 | 0.6735 | **0.8719** |
| | Jaccard Index | 0.4835 | 0.5336 | 0.4454 | 0.5101 | 0.5189 | 0.5519 | 0.7593 | 0.7301 | 0.4201 | 0.7064 | 0.7030 | 0.6837 | 0.5101 | **0.7747** |
| | Accuracy | 0.8514 | 0.8822 | 0.8118 | 0.8627 | 0.8739 | 0.8871 | 0.8644 | 0.8453 | 0.7982 | 0.9147 | 0.9002 | 0.9184 | 0.8627 | **0.9617** |
| D7 | Precision | 0.8198 | 0.8342 | 0.8747 | 0.8611 | 0.9114 | 0.8392 | 0.8611 | 0.8225 | 0.7801 | 0.8901 | 0.9145 | 0.8936 | 0.8384 | **0.9152** |
| | Recall | 0.8326 | 0.8374 | 0.8898 | 0.8763 | 0.8821 | 0.8400 | 0.8763 | 0.8109 | 0.7776 | 0.8758 | 0.8876 | 0.9017 | 0.8641 | **0.9201** |
| | F-Score | 0.8261 | 0.8358 | 0.8822 | 0.8686 | 0.8965 | 0.8396 | 0.8686 | 0.8167 | 0.7788 | 0.8829 | 0.9009 | 0.8976 | 0.8511 | **0.9176** |
| | Jaccard Index | 0.6844 | 0.7146 | 0.7798 | 0.7545 | 0.7935 | 0.7229 | 0.7545 | 0.6891 | 0.6304 | 0.7720 | 0.8021 | 0.8074 | 0.7215 | **0.8452** |
| | Accuracy | 0.8136 | 0.8362 | 0.8757 | 0.8588 | 0.8814 | 0.8418 | 0.8588 | 0.8136 | 0.7740 | 0.8701 | 0.8870 | 0.8927 | 0.8362 | **0.9153** |

TABLE IV

CLUSTERING REPORTS ON D8, D9, D10, D11, D12, AND D13 USING FIVE VALIDATION INDICES

| Database | Validation Index | $K-PC$ | $EK-PC$ | $KL-FCM-GM$ | $FK-PC$ | $IK-PC$ | $CAVE$ | DKFCM | PE-EEMD | SBAC | CG-SSC | MSC | DPC-MD | SpectralCAT | MDPC-AD |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| D8 | Precision | 0.7626 | 0.8661 | 0.8780 | 0.9013 | 0.6276 | 0.8000 | 0.8421 | 0.7757 | 0.5606 | 0.8348 | 0.8249 | **0.9239** | 0.8135 | 0.9179 |
| | Recall | 0.7148 | 0.8227 | 0.8410 | 0.8716 | 0.6114 | 0.7988 | 0.8216 | 0.7812 | 0.5588 | 0.8127 | 0.8692 | 0.9049 | 0.8533 | **0.9239** |
| | F-Score | 0.7380 | 0.8439 | 0.8591 | 0.8862 | 0.6194 | 0.7994 | 0.8317 | 0.7785 | 0.5597 | 0.8236 | 0.8465 | 0.9143 | 0.8329 | **0.9209** |
| | Jaccard Index | 0.5507 | 0.7284 | 0.7473 | 0.7981 | 0.4016 | 0.6660 | 0.7057 | 0.6314 | 0.3301 | 0.6955 | 0.7201 | 0.8289 | 0.7037 | **0.8419** |
| | Accuracy | 0.6931 | 0.8317 | 0.8416 | 0.8713 | 0.5644 | 0.7624 | 0.8000 | 0.7228 | 0.4851 | 0.7900 | 0.8317 | 0.9010 | 0.8218 | **0.9109** |
| D9 | Precision | 0.6163 | 0.7233 | 0.7471 | 0.7715 | 0.5420 | 0.7847 | 0.8294 | 0.8333 | 0.5490 | 0.8449 | 0.8472 | 0.8615 | 0.8214 | **0.8956** |
| | Recall | 0.6249 | 0.7180 | 0.7403 | 0.7655 | 0.5397 | 0.7761 | 0.8346 | 0.8204 | 0.5514 | 0.8576 | 0.8371 | 0.8715 | 0.8274 | **0.8884** |
| | F-Score | 0.6205 | 0.7207 | 0.7437 | 0.7685 | 0.5409 | 0.7804 | 0.8320 | 0.8268 | 0.5502 | 0.8512 | 0.8421 | 0.8665 | 0.8244 | **0.8920** |
| | Jaccard Index | 0.4105 | 0.5632 | 0.5872 | 0.6199 | 0.3668 | 0.6383 | 0.7114 | 0.6886 | 0.3798 | 0.7356 | 0.7234 | 0.7614 | 0.6996 | **0.8026** |
| | Accuracy | 0.5833 | 0.7250 | 0.7417 | 0.7667 | 0.5417 | 0.7833 | 0.8333 | 0.8167 | 0.5583 | 0.8500 | 0.8417 | 0.8667 | 0.8250 | **0.8917** |
| D10 | Precision | 0.5848 | 0.6534 | 0.8075 | 0.8333 | 0.6313 | 0.8351 | 0.7975 | 0.7840 | 0.7684 | 0.8349 | 0.8333 | 0.8525 | 0.8075 | **0.8660** |
| | Recall | 0.5821 | 0.6481 | 0.7991 | 0.8211 | 0.6381 | 0.8321 | 0.8091 | 0.7821 | 0.7778 | 0.8255 | 0.8211 | 0.8351 | 0.7991 | **0.8587** |
| | F-Score | 0.5835 | 0.6508 | 0.8033 | 0.8272 | 0.6347 | 0.8336 | 0.8033 | 0.7831 | 0.7731 | 0.8302 | 0.8272 | 0.8437 | 0.8033 | **0.8624** |
| | Jaccard Index | 0.4078 | 0.4770 | 0.6684 | 0.6998 | 0.4563 | 0.7151 | 0.6406 | 0.6432 | 0.6238 | 0.7058 | 0.6998 | 0.7209 | 0.6684 | **0.7561** |
| | Accuracy | 0.5815 | 0.6481 | 0.8037 | 0.8259 | 0.6296 | 0.8370 | 0.7815 | 0.7838 | 0.7704 | 0.8296 | 0.8259 | 0.8407 | 0.8037 | **0.8630** |
| D11 | Precision | 0.4996 | 0.7809 | 0.8272 | 0.8551 | 0.7439 | 0.8864 | 0.8340 | 0.8425 | 0.8012 | 0.7266 | 0.9028 | 0.9910 | 0.8691 | **0.9985** |
| | Recall | 0.5430 | 0.7657 | 0.8187 | 0.8468 | 0.7326 | 0.8957 | 0.8437 | 0.8545 | 0.7923 | 0.6617 | 0.8983 | 0.9911 | 0.8707 | **0.9985** |
| | F-Score | 0.5204 | 0.7732 | 0.8229 | 0.8509 | 0.7382 | 0.8910 | 0.8388 | 0.8484 | 0.7967 | 0.6927 | 0.9005 | 0.9910 | 0.8699 | **0.9985** |
| | Jaccard Index | 0.3294 | 0.6182 | 0.6932 | 0.7358 | 0.5764 | 0.7988 | 0.7184 | 0.7066 | 0.6538 | 0.5336 | 0.8133 | 0.9822 | 0.7688 | **0.9970** |
| | Accuracy | 0.4820 | 0.7585 | 0.8175 | 0.8455 | 0.7285 | 0.8845 | 0.8333 | 0.8267 | 0.7885 | 0.8822 | 0.8945 | 0.9910 | 0.8595 | **0.9985** |
| D12 | Precision | 0.5714 | 0.4872 | 0.6068 | 0.5358 | 0.4869 | 0.5792 | 0.6313 | 0.6254 | 0.5623 | 0.5958 | 0.6526 | 0.6400 | 0.5919 | **0.6824** |
| | Recall | 0.5830 | 0.5063 | 0.6405 | 0.5489 | 0.5170 | 0.5888 | 0.6381 | 0.6254 | 0.6149 | 0.6215 | 0.6310 | 0.6212 | 0.5717 | **0.6713** |
| | F-Score | 0.5772 | 0.4966 | 0.6232 | 0.5423 | 0.5015 | 0.5839 | 0.6347 | 0.6254 | 0.5875 | 0.6084 | 0.6417 | 0.6304 | 0.5816 | **0.6768** |
| | Jaccard Index | 0.3967 | 0.3153 | 0.4540 | 0.3702 | 0.3228 | 0.4067 | 0.4563 | 0.4302 | 0.4074 | 0.4190 | 0.4648 | 0.4541 | 0.4029 | **0.4993** |
| | Accuracy | 0.5743 | 0.4865 | 0.6081 | 0.5405 | 0.4932 | 0.5811 | 0.6296 | 0.6045 | 0.5676 | 0.5932 | 0.6351 | 0.6149 | 0.5676 | **0.6622** |
| D13 | Precision | 0.5495 | 0.8009 | 0.8248 | 0.7988 | 0.8435 | 0.8194 | 0.8272 | 0.8003 | 0.6103 | 0.8248 | 0.8601 | 0.8715 | 0.8317 | **0.8809** |
| | Recall | 0.5509 | 0.7970 | 0.8189 | 0.7923 | 0.8497 | 0.8167 | 0.8187 | 0.7958 | 0.6071 | 0.8189 | 0.8558 | 0.8645 | 0.8369 | **0.8864** |
| | F-Score | 0.5502 | 0.7989 | 0.8218 | 0.7955 | 0.8466 | 0.8180 | 0.8229 | 0.7980 | 0.6087 | 0.8218 | 0.8579 | 0.8680 | 0.8343 | **0.8837** |
| | Jaccard Index | 0.3782 | 0.6601 | 0.6904 | 0.6500 | 0.7284 | 0.6888 | 0.6932 | 0.6598 | 0.4327 | 0.6904 | 0.7464 | 0.7632 | 0.7143 | **0.7888** |
| | Accuracy | 0.5522 | 0.7957 | 0.8174 | 0.7884 | 0.8435 | 0.8159 | 0.8175 | 0.7957 | 0.6058 | 0.8174 | 0.8551 | 0.8667 | 0.8348 | **0.8826** |

achieved by a method is marked by a bold character. We can conclude after observing all the results of Table III that the MDPC-AD outperforms all 13 SOTA methods even when the data points of D4, D5, D6, and D7 are varying and have arbitrary shape clusters.

*3) Experiment on Mixed Datasets:* In the third experiment, the MDPC-AD is executed on mixed datasets, namely, D8, D9, D10, D11, D12, and D13. These datasets contain both numerical as well as categorical variables. Here, (15) is used to compute the distance between data objects. The clustering reports produced are presented in Table IV. The best clustering report generated by a method is marked by bold characters. It is clear from Table IV that the MDPC-AD outperforms all the above-discussed 13 SOTA methods.

In summary, we can say that the clustering reports in terms of precision, recall, F-Score, JI, and accuracy obtained by the proposed method named MDPC-AD are higher than the other 13 SOTA existing approaches mentioned in Section IV-C. The primary reason is that K-PC, EK-PC, IK-PC, FK-PC, DKFCM, PE-EEMD, CG-SSC, MSC, and DPC-MD are sensitive to the initialization of cluster representatives and are improper for nonspherically distributed

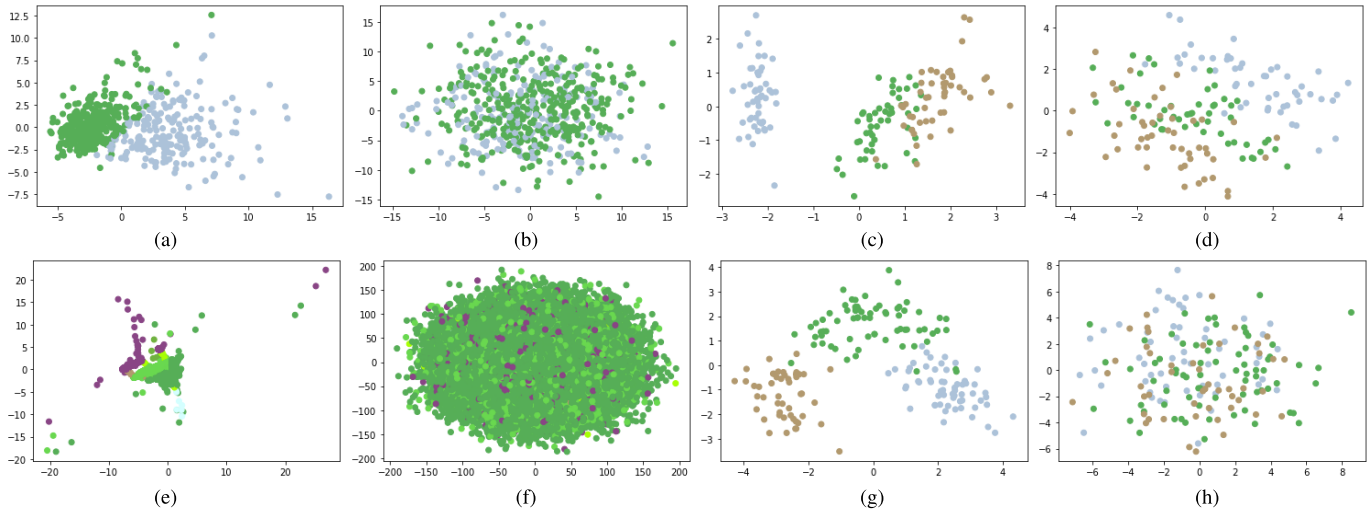


Fig. 7. First two principal components of the dataset on the left and its updated version with noise features as the number of genuine features on the right were shown on a plane to display the first and second matching vectors of the data matrix along the axes. Different colors indicate different data point classifications. (a) D4. (b) D4 with noisy features. (c) D5. (d) D5 with noisy features. (e) D6. (f) D6 with noisy features. (g) D7. (h) D7 with noisy features.

data. Moreover, SBAC addresses similarity measures with the assumption that the unusual matched feature values correspond to higher weights. DKFCM adopts a fuzzy objective function that is designed based on several probabilistic theories regarding the organization of the obtained clusters. CAVE is sensitive to the procedure of sampling. The efficiency of SpectralCAT depends on the selection of kernel function to compute the Markov matrix. On the other hand, MDPC-AD overcomes some of the abovementioned issues. Thus, MDPC-AD achieved good clustering results on the above-discussed numerical, categorical, and mixed datasets.

*4) Experiment on Noisy Datasets:* The fourth experiment is conducted on datasets, namely, D4, D5, D6, and D6, having only numerical variables to know whether the MDPC-AD is robust against the noisy features. After adding noisy features produced by a uniform random distribution in the length and size limit of the original dataset, the influence of noisy features is examined. Therefore, a dataset's number of features would be twice as many as its initial number of attributes. The first two principal components of each of the aforementioned datasets, as plotted on 2-D planes, are shown in Fig. 7 before and after the addition of noisy features. Fig. 7 makes it obvious that for a dataset; practically, all of the mapped locations have significant overlaps with one another. This simply means that the existence of noisy features hurts these datasets. The results obtained by all techniques after including noisy features are shown in Fig. 8 against five assessment metrics, including accuracy, precision, recall, F-Score, and Jaccard index on D4–D7. The results show that the MDPC-AD performs better than any other approach, from K-PC to SpectralCAT. For a small number of datasets, clustering performance is dramatically reduced. Nevertheless, the S-distance, which is utilized to calculate the separations between numerical data items and is invariant to the Hadamard product, makes the MDPC-AD resilient [8].

*5) Experiment for Knowing the Impact of "c" in Symmetric Favored c-NN:* In this work, local density is estimated based on a sequence of the weighted exponential kernel using an SFCNN. In the previous four experiments, the value of "$c$" is considered 5, as suggested by the past studies [28]. However, the fifth experiment is conducted to verify the previous claim. The value of "$c$" varies from 1 to 19 with a step size of 2. The values of accuracy, precision, recall, F-Score, and Jaccard index for each value of "$c$" over all 13 datasets are shown in Fig. 9. The values of "$c$" are shown on the $x$-axis, and the clustering results are shown on the $y$-axis. It is clear from Fig. 9 that the values of clustering metrics are maximum when the value of "$c$" varies from 1 to 5 on datasets, i.e., D3, D8, and D9. However, the performance increases slightly on D6 when the value of "$c$" is beyond 5. On the other hand, the performances remain consistent on D11 and D13. The clustering performances deteriorate on D1, D2, D4, D5, D7, D10, and D12 when the value of "$c$" is beyond 12. It means that it is really difficult to find out the optimum value of "$c$." However, it relies on the characteristics of a dataset.

*6) Experiment on Order Sensitivity:* In the final experiment, the order of the data objects in a dataset is changed while still analyzing the clustering result. This sensitivity analysis measures the stability of the algorithm due to randomness and erroneous assessment. The MDPC-AD is executed ten times on 13 datasets. However, the positions of data objects are changed by shuffling them randomly, and the corresponding clustering results are shown in Fig. 10. The $x$-axis and $y$-axis of each plot in Fig. 10 denote, respectively, the number of iterations and the value of the metric in question. It is clear from Fig. 9 that the MDPC-AD is not sensitive to the position of data objects or order since the performance is the same or constant in all ten runs.

*7) Experiment for Run-Time Comparison:* All the methods mentioned in Section IV-C are not only compared based on their clustering reports but also compared based on their execution time, which is measured and noted in Table V. It is clear from Table V that most of the time MDPC-AD takes


Fig. 8. Comparison of clustering results on numerical datasets with noisy features. (a) Precision. (b) Recall. (c) F-Score. (d) Jaccard index. (e) Accuracy.

TABLE V
METHODS RUN TIME (UNIT: S)

| Method | D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 | D9 | D10 | D11 | D12 | D13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| K-PC | **0.251** | 0.451 | **0.061** | 0.693 | **0.083** | 20.961 | 0.189 | **0.091** | 0.121 | 0.245 | 3.117 | 0.229 | 0.832 |
| EK-PC | 0.268 | 0.462 | 0.069 | 0.710 | 0.095 | 20.034 | 0.186 | 0.139 | 0.118 | 0.236 | 3.215 | 0.235 | 0.799 |
| KL-FCM-GM | 0.279 | 0.519 | 0.073 | 0.742 | 0.115 | 19.088 | 0.217 | 0.105 | 0.106 | 0.261 | 3.514 | 0.242 | 0.823 |
| FK-PC | 0.265 | **0.450** | 0.063 | 0.691 | 0.091 | 19.358 | 0.173 | 0.097 | 0.112 | **0.226** | 3.651 | 0.259 | 0.815 |
| IK-PC | 0.272 | 0.497 | 0.072 | 0.732 | 0.108 | 19.421 | 0.179 | 0.110 | 0.107 | 0.253 | 3.912 | 0.214 | 0.807 |
| CAVE | 0.357 | 0.574 | 0.115 | 0.811 | 0.176 | 23.752 | 0.211 | 0.144 | 0.142 | 0.315 | 4.712 | 0.286 | 0.873 |
| DKFCM | 0.294 | 0.572 | 0.102 | 0.799 | 0.171 | 22.901 | 0.197 | 0.141 | 0.147 | 0.29 | 3.988 | 0.278 | 0.871 |
| PE-EEMD | 0.372 | 0.598 | 0.119 | 0.805 | 0.187 | 24.863 | 0.213 | 0.155 | 0.164 | 0.311 | 4.267 | 0.306 | 0.901 |
| SBAC | 0.381 | 0.601 | 0.128 | 0.813 | 0.181 | 24.895 | 0.208 | 0.121 | 0.116 | 0.309 | 4.271 | 0.265 | 0.889 |
| CG-SSC | 0.321 | 0.579 | 0.107 | 0.801 | 0.179 | 23.264 | 0.199 | 0.146 | 0.152 | 0.297 | 3.996 | 0.287 | 0.873 |
| MSC | 0.297 | 0.566 | 0.115 | 0.802 | 0.174 | 22.913 | 0.196 | 0.138 | 0.15 | 0.293 | 3.972 | 0.281 | 0.869 |
| DPC-MD | 0.287 | 0.523 | 0.106 | 0.788 | 0.134 | 20.567 | 0.187 | 0.117 | **0.104** | 0.285 | 3.453 | 0.227 | 0.848 |
| SpectralCAT | 0.376 | 0.612 | 0.133 | 0.845 | 0.192 | 27.168 | 0.224 | 0.153 | 0.159 | 0.297 | 4.978 | 0.301 | 0.904 |
| *MDPC − AD* | 0.301 | 0.554 | 0.104 | **0.683** | 0.165 | **19.405** | **0.164** | 0.137 | 0.135 | 0.281 | **3.101** | **0.203** | **0.794** |

less time compared to K-PC, EK-PC, KL-FCM-GM, FK-PC, IK-PC, CAVE, DKFCM, DPC-MD, SpectralCAT, PE-EEMD, CG-SSC, MSC, and SBAC. However, in some cases, K-PC has a lesser execution time compared to the MDPC-AD,

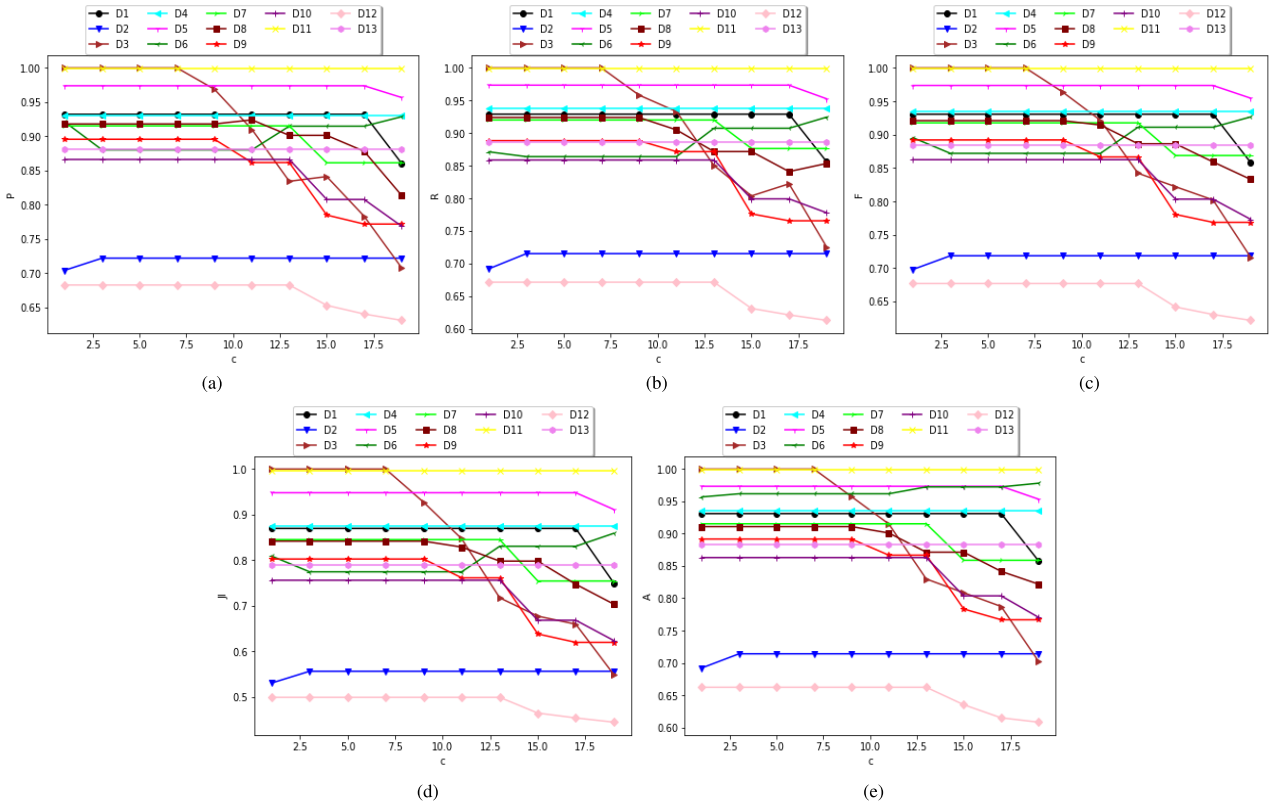


Fig. 9. Evaluation of MDPC-AD for different values of "$c$" values on all the datasets. (a) Precision. (b) Recall. (c) F-Score. (d) Jaccard index. (e) Accuracy.

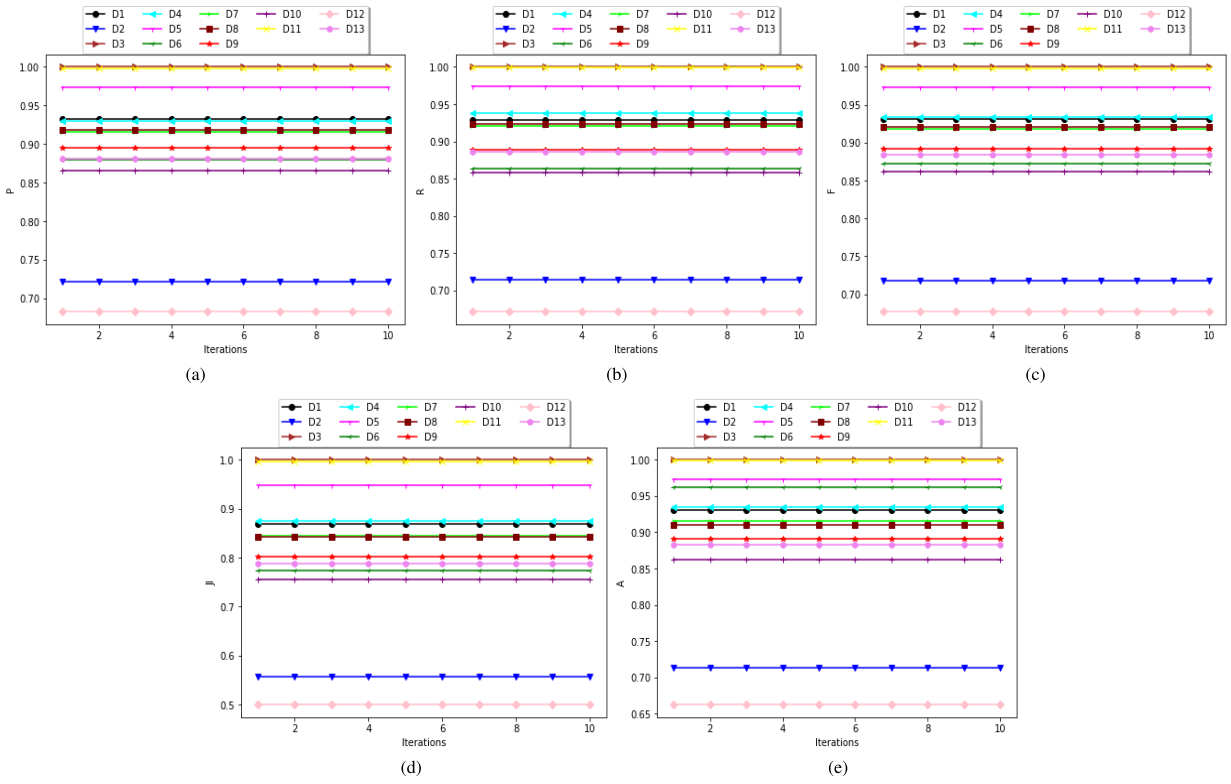


Fig. 10. Evaluation of MDPC-AD for testing sensitivity on 13 datasets. (a) Precision. (b) Recall. (c) F-Score. (d) Jaccard index. (e) Accuracy.

whereas MDPC-AD has a lesser execution time than EK-PC, KL-FCM-GM, IK-PC, CAVE, DKFCM, SpectralCAT, PE-EEMD, CG-SSC, MSC, and SBAC. Overall, the proposed MDPC-AD executes in lesser time and outperforms other methods. Moreover, the best execution times are bold in Table V.

TABLE VI
ABLATION STUDY ON D8, D9, D10, D11, D12, AND D13

| Database | Validation Index | *DPC* | *DPC − AD* | *DPC − AD − SFCNN* | *DPC − AD − SFCNN − CR* | *MDPC − CNN − MD* | *MDPC − CNN − AD* | *MDPC − SFCNN − MD* | *MDPC − AD* |
|---|---|---|---|---|---|---|---|---|---|
| D8 | Precision | 0.799 | 0.846 | 0.878 | 0.914 | 0.854 | 0.875 | 0.903 | **0.918** |
| | Recall | 0.840 | 0.809 | 0.841 | 0.900 | 0.861 | 0.876 | 0.899 | **0.924** |
| | F-Score | 0.819 | 0.827 | 0.859 | 0.907 | 0.858 | 0.875 | 0.901 | **0.921** |
| | Jaccard Index | 0.684 | 0.704 | 0.747 | 0.831 | 0.741 | 0.779 | 0.818 | **0.842** |
| | Accuracy | 0.802 | 0.812 | 0.842 | 0.891 | 0.844 | 0.873 | 0.898 | **0.911** |
| D9 | Precision | 0.777 | 0.813 | 0.854 | 0.872 | 0.846 | 0.864 | 0.874 | **0.896** |
| | Recall | 0.767 | 0.820 | 0.861 | 0.879 | 0.838 | 0.878 | 0.868 | **0.888** |
| | F-Score | 0.772 | 0.816 | 0.858 | 0.875 | 0.842 | 0.871 | 0.871 | **0.892** |
| | Jaccard Index | 0.626 | 0.687 | 0.749 | 0.776 | 0.723 | 0.761 | 0.770 | **0.803** |
| | Accuracy | 0.775 | 0.817 | 0.858 | 0.875 | 0.833 | 0.862 | 0.867 | **0.892** |
| D10 | Precision | 0.816 | 0.843 | 0.863 | 0.857 | 0.840 | 0.844 | 0.861 | **0.866** |
| | Recall | 0.826 | 0.833 | 0.842 | 0.849 | 0.831 | 0.859 | 0.868 | **0.859** |
| | F-Score | 0.821 | 0.838 | 0.852 | 0.853 | 0.835 | 0.851 | 0.864 | **0.862** |
| | Jaccard Index | 0.690 | 0.717 | 0.732 | 0.743 | 0.711 | 0.736 | 0.761 | **0.756** |
| | Accuracy | 0.819 | 0.837 | 0.848 | 0.856 | 0.830 | 0.843 | 0.859 | **0.863** |
| D11 | Precision | 0.954 | 0.975 | 0.990 | 0.991 | 0.989 | 0.991 | 0.993 | **0.999** |
| | Recall | 0.953 | 0.975 | 0.990 | 0.991 | 0.989 | 0.991 | 0.993 | **0.999** |
| | F-Score | 0.954 | 0.975 | 0.990 | 0.991 | 0.989 | 0.991 | 0.993 | **0.999** |
| | Jaccard Index | 0.911 | 0.950 | 0.979 | 0.982 | 0.977 | 0.982 | 0.986 | **0.997** |
| | Accuracy | 0.953 | 0.975 | 0.990 | 0.991 | 0.989 | 0.991 | 0.993 | **0.999** |
| D12 | Precision | 0.599 | 0.600 | 0.661 | 0.684 | 0.675 | 0.697 | 0.672 | **0.682** |
| | Recall | 0.606 | 0.634 | 0.648 | 0.649 | 0.651 | 0.675 | 0.663 | **0.671** |
| | F-Score | 0.602 | 0.616 | 0.654 | 0.666 | 0.663 | 0.686 | 0.667 | **0.677** |
| | Jaccard Index | 0.415 | 0.448 | 0.478 | 0.488 | 0.475 | 0.485 | 0.499 | **0.499** |
| | Accuracy | 0.595 | 0.601 | 0.635 | 0.655 | 0.639 | 0.651 | 0.658 | **0.662** |
| D13 | Precision | 0.830 | 0.845 | 0.873 | 0.887 | 0.853 | 0.871 | 0.879 | **0.881** |
| | Recall | 0.823 | 0.851 | 0.866 | 0.881 | 0.873 | 0.882 | 0.890 | **0.886** |
| | F-Score | 0.827 | 0.848 | 0.869 | 0.884 | 0.863 | 0.877 | 0.885 | **0.884** |
| | Jaccard Index | 0.697 | 0.731 | 0.765 | 0.787 | 0.746 | 0.782 | 0.793 | **0.789** |
| | Accuracy | 0.822 | 0.845 | 0.868 | 0.881 | 0.851 | 0.871 | 0.880 | **0.883** |

TABLE VII
GEARBOX FAULT DIAGNOSIS USING TWO INTERNAL VALIDATION INDICES

| Validation Index | *DPC* | *DPC − AD* | *DPC − AD − SFCNN* | *DPC − AD − SFCNN − CR* | *MDPC − AD* |
|---|---|---|---|---|---|
| $ID_1$ | 5.6 | 11.2 | 17.4 | 22.1 | ∞ |
| $ID_2$ | 3.2 | 1.6 | 0.8 | 0.1 | **0** |

*8) Ablation Study:* The steps of MDPC-AD are determined after conducting an ablation study on mixed datasets only. DPC presented by Rodriguez and Laio [26] is implemented in the first analysis. In the second analysis, the impact of the proposed AD mentioned in (12) is evaluated by incorporating it in DPC and renaming the modified algorithm as DPC-AD. In the third analysis, a new method to estimate local density is adopted, which relies on a sequence of the weighted exponential kernel using an SFCNN to overcome the limitation of fixed $c$-NN and names the method DPC-AD-SFCNN. In the fourth study, a method for automatic selection of the initial cluster representatives is utilized from MDPC-AD and merged with DPC-AD-SFCNN and marked as DPC-AD-SFCNN-CR. The MDPC-CNN-MD considers the $c$-NN in place of an SFCNN for estimating local density. It also adopts the presented distance from [29]. The MDPC-CNN-AD is similar to the MDPC-CNN-MD, except that instead of using the presented distance in [29], it employs the proposed distance. Finally, CNN in MDPC-CNN-MD is substituted with SFCNN, which is now called MDPC-SFCNN-MD. Furthermore, the results of these studies are noted in Table VI to validate the superiority of the proposed method, MDPC-AD. All the results demonstrate that the proposed method, MDPC-AD, is superior to other combinations of the DPC algorithm.

*9) Case Study for Gearbox Fault Diagnosis Using Clustering:* In this section, MDPC-AD is compared with DPC, DPC-AD, DPC-AD-SFCNN, and DPC-AD-SFCNN-CR. This experiment is performed to group the state of the gears of a mechanical machine. For the same, features of the data obtained in 1 s are considered as a sample, and 90 samples/data objects are analyzed. A quantitative clustering analysis is conducted using two internal validation indices. First, a ratio of separation SP and compactness CM is computed, and their ratio $\mathrm{ID}_1 = \mathrm{SP}/\mathrm{CM}$ is used as the metric for comparison. For this metric, a high value of SP is desired, and a low value of CM is suitable for good clusters. Thus, a high value of $\mathrm{ID}_1$ indicates effective clustering results. Second, a clustering error, $\mathrm{ID}_2$, is adopted for comparison and a smaller value of clustering error shows the most effective. A comparison of $\mathrm{ID}_1$ index on the above discussed five methods is shown in Table VII. In the case of MDPC-AD, SP has a high value, and CM obtained the ideal value equal to zero; thus, the value of $\mathrm{ID}_1$ is extremely large or tends to infinity, which shows that MDPC-AD outperforms other SOTA methods. It indicates that MDPC-AD can yield well-separated and most compact clusters. Second, a comparison is shown using $\mathrm{ID}_2$ where all the methods are executed 50 times, and the average error is computed and presented in Table VII. As mentioned in Table VII, MDPC-AD achieved the lowest value equal to zero, meaning that there was nil clustering error in the proposed method while diagnosing the faults in a gearbox. Therefore, the proposed study is an effective method for gearbox fault diagnosis.

## V. CONCLUSION

This study introduces the DPC-based clustering technique named MDPC-AD for gearbox fault diagnostics. An ablation study is conducted to demonstrate the effectiveness of each part of the MDPC-AD. The obtained results illustrate that the novel AD can more clearly reveal the structure of the 13 real-world datasets from UCI under examination. To calculate the global parameter and the local density of each data object, the MDPC-AD employs the idea of a series of weighted Gaussian kernels based on an SFCNN. In addition, the MDPC-AD is easier to control than DBSCAN and DPC since it can choose the initial cluster representatives automatically by establishing

the cutoff distance as a function of parameter $c$. The first cluster representatives' computation, however, ensures the inclusion of genuine initial cluster representatives. The concept of grouping density reachable clusters is employed to solve this problem in subsequent iterations, even though our method may initially choose incorrect cluster representatives. According to experiments on different real-world datasets, the MDPC-AD outperforms the 13 SOTA approaches mentioned in this article. Even though an experiment is run to see how '$c$' affects things, more research is still required. Finding the ideal value for "$c$" and understanding the relationship between the parameters $\alpha_t$ and "$c$" call for more investigation. It would also be intriguing to expand the capabilities of the current algorithm to manage large datasets with mixed attributes. Although MDPC-AD improves clustering efficiency, the decreased efficiency of MDPC-AD compared to DPC is due to the high computational complexity of density estimation. Therefore, it is essential to reduce the computational complexity of density estimation, a subject that merits more study. Finally, in mechanical systems, clustering algorithms have proven to be more reliable, particularly for fault diagnosis. The names of two faults that could affect gearboxes, bearings, and wind turbines, are mentioned in this study. The last two, which merit additional research, have not been covered in this article.

## ACKNOWLEDGMENT

The authors would like to thank the support of Ph.D. student Michal Dobrovolny for consultations.

## REFERENCES

[1] J. Sun, C. Yan, and J. Wen, "Intelligent bearing fault diagnosis method combining compressed data acquisition and deep learning," *IEEE Trans. Instrum. Meas.*, vol. 67, no. 1, pp. 185–195, Jan. 2018.

[2] C. Sun, M. Ma, Z. B. Zhao, and X. Chen, "Sparse deep stacking network for fault diagnosis of motor," *IEEE Trans. Ind. Informat.*, vol. 14, no. 7, pp. 3261–3270, Mar. 2018.

[3] S. B. Wang, X. F. Chen, C. W. Tong, and Z. B. Zhao, "Matching synchrosqueezing wavelet transform and application to aeroengine vibration monitoring," *IEEE Trans. Instrum. Meas.*, vol. 66, no. 2, pp. 360–372, Feb. 2017.

[4] Y. Chen and M. J. Zuo, "A sparse multivariate time series model-based fault detection method for gearboxes under variable speed condition," *Mech. Syst. Signal Process.*, vol. 167, Mar. 2022, Art. no. 108539.

[5] Y. Liu, B. Liu, X. Zhao, and M. Xie, "A mixture of variational canonical correlation analysis for nonlinear and quality-relevant process monitoring," *IEEE Trans. Ind. Electron.*, vol. 65, no. 8, pp. 6478–6486, Aug. 2018.

[6] W. Teng, Y. Liu, Y. Huang, L. Song, Y. Liu, and Z. Ma, "Fault detection of planetary subassemblies in a wind turbine gearbox using TQWT based sparse representation," *J. Sound Vibrat.*, vol. 490, Jan. 2021, Art. no. 115707.

[7] Y. Liao, L. Zhang, and W. Li, "Regrouping particle swarm optimization based variable neural network for gearbox fault diagnosis," *J. Intell. Fuzzy Syst.*, vol. 34, no. 6, pp. 3671–3680, 2018.

[8] S. Chakraborty and S. Das, "K—Means clustering with a new divergence-based distance metric: Convergence and performance analysis," *Pattern Recognit. Lett.*, vol. 100, pp. 67–73, Dec. 2017.

[9] Z. Shuqing, S. Guoxiu, L. Liang, L. Xinxin, and J. Xiong, "Study on mechanical fault diagnosis method based on LMD approximate entropy and fuzzy C-means clustering," *Chinese J. Sci. Instrum.*, vol. 34, no. 3, pp. 714–720, 2013.

[10] A. R. Ramos et al., "A novel fault diagnosis scheme applying fuzzy clustering algorithms," *Appl. Soft Comput.*, vol. 58, pp. 605–619, Sep. 2017.

[11] D. Gustafson and W. Kessel, "Fuzzy clustering with a fuzzy covariance matrix," in *Proc. IEEE Conf. Decis. Control including 17th Symp. Adapt. Processes*, Jan. 1978, pp. 761–766.

[12] S. Wang, L. Li, S. Zhang, and G. Sun, "Mechanical fault diagnosis method based on EEMD sample entropy and GK fuzzy clustering," *China Mech. Eng.*, vol. 24, no. 22, p. 3036, 2013.

[13] X. Liu, M. Li, S. Qin, X. Ma, and W. Wang, "A predictive fault diagnose method of wind turbine based on K-means clustering and neural networks," *J. Internet Technol.*, vol. 17, no. 7, pp. 1521–1528, 2016.

[14] P. Baraldi, F. D. Maio, M. Rigamonti, E. Zio, and R. Seraoui, "Unsupervised clustering of vibration signals for identifying anomalous conditions in a nuclear turbine," *J. Intell. Fuzzy Syst.*, vol. 28, no. 4, pp. 1723–1731, 2015.

[15] S. Fu, K. Liu, Y. Xu, and Y. Liu, "Rolling bearing diagnosing method based on time domain analysis and adaptive fuzzy C-means clustering," *Shock Vibrat.*, vol. 2016, pp. 1–8, Jan. 2016.

[16] E. Li, L. Wang, B. Song, and S. Jian, "Improved fuzzy C-means clustering for transformer fault diagnosis using dissolved gas analysis data," *Energies*, vol. 11, no. 9, p. 2344, Sep. 2018.

[17] Z. M. Nopiah, A. K. Junoh, and A. K. Ariffin, "Vehicle interior noise and vibration level assessment through the data clustering and hybrid classification model," *Appl. Acoust.*, vol. 87, pp. 9–22, Jan. 2015.

[18] I. Gath and A. B. Geva, "Unsupervised optimal fuzzy clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 7, pp. 773–780, Jul. 1989.

[19] J. C. Bezdek and J. C. Dunn, "Optimal fuzzy partitions: A heuristic for estimating the parameters in a mixture of normal distributions," *IEEE Trans. Comput.*, vol. C-24, no. 8, pp. 835–838, Aug. 1975.

[20] Z. X. Wei, Y. X. Wang, S. L. He, and J. D. Bao, "A novel intelligent method for bearing fault diagnosis based on affinity propagation clustering and adaptive feature selection," *Knowl.-Based Syst.*, vol. 116, pp. 1–12, Jan. 2017.

[21] R. Langone, C. Alzate, B. D. Ketelaere, J. Vlasselaerc, W. Meertc, and J. A. K. Suykensa, "LS-SVM based spectral clustering and regression for predicting maintenance of industrial machines," *Eng. Appl. Artif. Intell.*, vol. 37, pp. 268–278, Jan. 2015.

[22] C. Sun, X. Chen, R. Yan, and R. X. Gao, "Composite-graph-based sparse subspace clustering for machine fault diagnosis," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 5, pp. 1850–1859, May 2020.

[23] S. Fong, J. Harmouche, S. Narasimhan, and J. Antoni, "Mean shift clustering-based analysis of nonstationary vibration signals for machinery diagnostics," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 7, pp. 4056–4066, Jul. 2020.

[24] J. Hou, Y. Wu, H. Gong, A. S. Ahmad, and L. Liu, "A novel intelligent method for bearing fault diagnosis based on EEMD permutation entropy and GG clustering," *Appl. Sci.*, vol. 10, no. 1, p. 386, Jan. 2020.

[25] A. Fahad et al., "A survey of clustering algorithms for big data: Taxonomy and empirical analysis," *IEEE Trans. Emerg. Topics Comput.*, vol. 2, no. 3, pp. 267–279, Sep. 2014.

[26] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, Jun. 2014.

[27] Z. Liang and P. Chen, "Delta-density based clustering with a divide-and-conquer strategy: 3DC clustering," *Pattern Recognit. Lett.*, vol. 73, pp. 52–59, Apr. 2016.

[28] L. Yaohui, M. Zhengming, and Y. Fang, "Adaptive density peak clustering based on K-nearest neighbors with aggregating strategy," *Knowl.-Based Syst.*, vol. 133, pp. 208–220, Oct. 2017.

[29] M. Du, S. Ding, and Y. Xue, "A novel density peaks clustering algorithm for mixed data," *Pattern Recognit. Lett.*, vol. 97, pp. 46–53, Oct. 2017.

[30] L. C. Jiao, F. Shang, F. Wang, and Y. Liu, "Fast semi-supervised clustering with enhanced spectral embedding," *Pattern Recognit.*, vol. 45, no. 12, pp. 4358–4369, 2012.

[31] A. Karlekar, A. Seal, O. Krejcar, and C. Gonzalo-Martín, "Fuzzy K-means using non-linear S-distance," *IEEE Access*, vol. 7, pp. 55121–55131, 2019.

[32] F. Cao, J. Z. Huang, and J. Liang, "A fuzzy SV-K-modes algorithm for clustering categorical data with set-valued attributes," *Appl. Math. Comput.*, vol. 295, pp. 1–15, Feb. 2017.

[33] U. Maulik, S. Bandyopadhyay, and I. Saha, "Integrating clustering and supervised learning for categorical data analysis," *IEEE Trans. Syst., Man, Cybern. A, Syst. Humans*, vol. 40, no. 4, pp. 664–675, Jul. 2010.

[34] I. Saha and U. Maulik, "Incremental learning based multiobjective fuzzy clustering for categorical data," *Inf. Sci.*, vol. 267, pp. 35–57, May 2014.

[35] I. T. R. Yanto, M. A. Ismail, and T. Herawan, "A modified fuzzy K-partition based on indiscernibility relation for categorical data clustering," *Eng. Appl. Artif. Intell.*, vol. 53, pp. 41–52, Aug. 2016.

[36] M. Li, S. Deng, L. Wang, S. Feng, and J. Fan, "Hierarchical clustering algorithm for categorical data using a probabilistic rough set model," *Knowl.-Based Syst.*, vol. 65, pp. 60–71, Jul. 2014.

[37] C. Li and G. Biswas, "Unsupervised learning with mixed numeric and nominal data," *IEEE Trans. Knowl. Data Eng.*, vol. 14, no. 4, pp. 673–690, Jul./Aug. 2002.

[38] C.-C. Hsu, "Generalizing self-organizing map for categorical data," *IEEE Trans. Neural Netw.*, vol. 17, no. 2, pp. 294–304, Mar. 2006.
[39] Z. Huang, "Clustering large data sets with mixed numeric and categorical values," in *Proc. 1st Pacific–Asia Conf. Knowl. Discovery Data Mining*, 1997, pp. 21–34.
[40] C.-C. Hsu and Y.-C. Chen, "Mining of mixed data with application to catalog marketing," *Exp. Syst. Appl.*, vol. 32, no. 1, pp. 12–23, 2007.
[41] Z. Zheng, M. Gong, J. Ma, L. Jiao, and Q. Wu, "Unsupervised evolutionary clustering algorithm for mixed type data," in *Proc. IEEE Congr. Evol. Comput.*, Jul. 2010, pp. 1–8.
[42] S. P. Chatzis, "A fuzzy C-means-type algorithm for clustering of data with mixed numeric and categorical attributes employing a probabilistic dissimilarity functional," *Exp. Syst. Appl.*, vol. 38, no. 7, pp. 8684–8689, 2011.
[43] J. Ji, W. Pang, C. Zhou, X. Han, and Z. Wang, "A fuzzy K-prototype clustering algorithm for mixed numeric and categorical data," *Knowl.-Based Syst.*, vol. 30, pp. 129–135, Jun. 2012.
[44] J. Ji, T. Bai, C. Zhou, C. Ma, and Z. Wang, "An improved K-prototypes clustering algorithm for mixed numeric and categorical data," *Neurocomputing*, vol. 120, pp. 590–596, Nov. 2013.
[45] S. Santini and R. Jain, "Similarity measures," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 9, pp. 871–883, Sep. 1999.
[46] J. Jiang, Y. Chen, X. Meng, L. Wang, and K. Li, "A novel density peaks clustering algorithm based on K nearest neighbors for improving assignment process," *Phys. A, Stat. Mech. Appl.*, vol. 523, pp. 702–713, Jun. 2019.
[47] D. Dheeru and E. K. Taniskidou. (2017). *UCI Machine Learning Repository*. [Online]. Available: http://archive.ics.uci.edu/ml
[48] G. David and A. Averbuch, "SpectralCAT: Categorical spectral clustering of numerical and nominal data," *Pattern Recognit.*, vol. 45, no. 1, pp. 416–433, 2012.

**Krishna Kumar Sharma** received the M.Tech. (Information Technology) degree from IIIT Allahabad, Allahabad, India, in 2011, and the Ph.D. degree from the Department of Computer Science and Engineering, PDPM Indian Institute of Information Technology, Design and Manufacturing Jabalpur, Jabalpur, India, in 2021.

He is currently an Assistant Professor with the Department of Computer Science and Informatics, University of Kota, Kota, India. His current research interests include pattern recognition.

**Ayan Seal** (Senior Member, IEEE) received the Ph.D. degree in engineering from Jadavpur University, Kolkata, India, in 2014.

He is currently an Assistant Professor with the Department of Computer Science and Engineering, PDPM Indian Institute of Information Technology, Design and Manufacturing Jabalpur, Jabalpur, India. He has visited the Universidad Politécnica de Madrid, Madrid, Spain, as a Visiting Research Scholar. He has authored or coauthored several journals, conferences, and book chapters in the area of biometric and medical image processing. His current research interests include image processing and pattern recognition.

Dr. Seal was a recipient of several awards. Recently, he received the Sir Visvesvaraya Young Faculty Research Fellowship from Media Lab Asia, Ministry of Electronics and Information Technology, Government of India.

**Anis Yazidi** (Senior Member, IEEE) received the M.Sc. and Ph.D. degrees from the University of Agder, Grimstad, Norway, in 2008 and 2012, respectively.

He was a Researcher with Teknova AS, Grimstad. From 2014 to 2019, he was an Associate Professor with the Department of Computer Science, Oslo Metropolitan University, Oslo, Norway. He is currently a Full Professor with the Department of Computer Science, Oslo Metropolitan University, where he is leading the research group in applied artificial intelligence. He is also a Professor II with the Norwegian University of Science and Technology (NTNU), Trondheim, Norway, and a Senior Researcher with Oslo University Hospital, Oslo. His current research interests include machine learning, learning automata, stochastic optimization, and autonomous computing.

**Ondrej Krejcar** received the Ph.D. degree in technical cybernetics from the Technical University of Ostrava, Ostrava, Czechia, in 2008.

He is currently a Full Professor in systems engineering and informatics with the University of Hradec Kralove (UHK), Hradec Kralove, Czechia, and the Faculty of Informatics and Management, Center for Basic and Applied Research, UHK; and a Research Fellow with the Malaysia-Japan International Institute of Technology, University of Technology Malaysia, Kuala Lumpur, Malaysia. Since June 2020, he has been a Vice-Rector for science and creative activities of the UHK. He is also the Director of the Center for Basic and Applied Research, UHK. From 2016 to 2020, he was the Vice-Dean for science and research with the Faculty of Informatics and Management, UHK. His H-index is 21, with more than 1800 citations received in the Web of Science, where more than 120 IF journal articles are indexed in JCR index.

Dr. Krejcar has been a Management Committee Member substitute at Project COST CA16226 since 2017. In 2018, he was the 14th top peer Reviewer in Multidisciplinary in the World according to Publons and a Top Reviewer in the Global Peer Review Awards 2019 by Publons. He is currently on the Editorial Board of the MDPI *Sensors* IF journal (Q1/Q2 at JCR) and several other ESCI-indexed journals. Since 2018, he has also been a Vice-Leader and a Management Committee Member at WG4 at project COST CA17136. Since 2019, he has been the Chairperson of the Program Committee of the KAPPA Program, Technological Agency of the Czech Republic as a regulator of the EEA/Norwegian Financial Mechanism in the Czech Republic for the term 2019–2024. Since 2020, he has been the Chairperson of Panel 1 (Computer, Physical and Chemical Sciences) of the ZETA Program, Technological Agency of the Czech Republic. From 2014 to 2019, he has been the Deputy Chairperson of Panel 7 (Processing Industry, Robotics, and Electrical Engineering) of the Epsilon Program, Technological Agency of the Czech Republic. He is also a guarantee of the doctoral study program in applied informatics with UHK, where he is focusing on lecturing on smart approaches to the development of information systems and applications in ubiquitous computing environments.