Int. J. Appl. Math. Comput. Sci., 2025, Vol. 35, No. 1, 7–18
DOI: 10.61822/amcs-2025-0001

DETECTION OF POTENTIALLY ANOMALOUS COSMIC PARTICLE TRACKS ACQUIRED WITH CMOS SENSORS: VALIDATION OF ROUGH k–MEANS CLUSTERING WITH PCA FEATURE EXTRACTION

TOMASZ HACHAJ [a,*], MARCIN PIEKARCZYK [a], JAROSŁAW WĄS [a]

[a]Faculty of Electrical Engineering, Automatics, Computer Science and Biomedical Engineering
AGH University of Krakow
Al. Mickiewicza 30, 30-059 Krakow, Poland
e-mail: tomasz.hachaj@agh.edu.pl

We present a method capable of detecting potentially anomalous cosmic particle tracks acquired with complementary metal-oxide-semiconductor (CMOS) sensors. We apply a principal components analysis-based feature extraction method and rough k-means clustering for outlier detection. We evaluated our approach on more than 10^4 images acquired by the Cosmic Ray Extremely Distributed Observatory (CREDO). The method presented in this work proved to be an effective solution. The analysis of the behavior of the rough k-means clustering-based algorithm presented here and the method of selecting its parameters showed that the algorithm performs as expected and demonstrates efficiency, stability, and repeatability of results for the test data set. The results included in this work are very relevant to the international CREDO project and the broader problem of anomaly analysis in image data sets. We plan to deploy the presented methodology in the image processing pipeline of the large data set we are working on in the CREDO project. The results can be reproduced using our source code, which is published in an open repository.

Keywords: cosmic-ray particle, rough sets, rough k-means, anomalies detection, principal components analysis, complementary metal-oxide-semiconductor sensors.

1. Introduction

It has been proven that low-cost commercial metal-oxide semiconductor (CMOS) sensors can be used to register various types of radiation successfully, understood here as detected energy that moves from different sources (Javan, 2002; Johary *et al.*, 2021). One of the types of radiation that can be observed with CMOS sensors is cosmic particles (Whiteson *et al.*, 2016). Particles of this type are recorded on a small part of the CMOS matrix in the form of bright flares of various shapes (different morphology), clearly visible against a nearly black background.

Thanks to the development of microelectronics and mobile telecommunications, the possibility of creating distributed cosmic-ray observatories that use the citizen science paradigm has emerged in recent years.

Among such projects are the following: the Distributed Electronic Cosmic-ray Observatory (DECO) (Vandenbroucke *et al.*, 2015; 2016), CRAYFIS (Albin and Whiteson, 2023) and the Cosmic Ray Extremely Distributed Observatory (CREDO) (Homola *et al.*, 2020; Karbowiak *et al.*, 2021). The vast amount of data produced by those observatories, especially by the CREDO project, which supplies scientific society with open access to measurement data, requires advanced algorithms for acquired image analysis. Among the very interesting problems that emerged in the CREDO data set is the task of potentially anomalous signal detection. By anomalous signals, we mean particle tracks left on CMOS detectors that differ significantly from those typically observed. Anomalous signals might represent various physical phenomena (also unknown ones) and are often explored by various unsupervised-based approaches (Kuusela *et al.*, 2012; Stein *et al.*, 2020; Crispim Romão *et al.*, 2021). In such approaches, anomaly detection problems are usually modeled as outlier detection machine learning problems.

Nowadays, many computer methods that operate

*Corresponding author

on sets are being extended through the use of rough sets (Pawlak, 1982). Among those algorithms are unsupervised learning algorithms (Skowron and Dutta, 2018), for example, clustering algorithms (Peters *et al*., 2002; Wang and Yao, 2018). Application of rough sets in unsupervised data set agglomeration can significantly expand the algorithm's ability to model relationships between data (Afridi *et al*., 2018; Pięta and Szmuc, 2021; Riza *et al*., 2014; Skowron and Ślęzak, 2022).

This paper presents a method capable of detecting potentially anomalous cosmic particle tracks acquired with CMOS sensors. We apply a principal components analysis-based features extraction method and rough k-means clustering for outlier detection. This paper is an extended version of our previous conference article (Hachaj *et al*., 2023). Compared with that paper, we extended the proposed methodology to include the ability to estimate the preferred number of clusters and the thresholding factor for rough k-means clustering. We also conducted appropriate experiments to validate the effectiveness of this extension. Thus, in practice, this work presents a consistent methodology that allows complete detection of potential anomalies in a data set of cosmic rays particle tracks acquired with CMOS sensors and algorithm parameters selection. Only Sections 2.1, 2.2, and 2.3 contain material that has been published previously, while the rest of this work has been drafted from scratch. Our results can be reproduced, and the source codes and data set can be found in an online repository at `https://github.com/browarsoftware/rough-k-means-particles` (accessed on 10 January 2025). It should be emphasized that the methodology and results presented in this paper relate to novel and significant research subjects in particle physics and astronomy, and we utilize the latest available data set of this modality.

2. Material and methods

In our earlier work (Hachaj *et al*., 2023), we showed that using rough sets in the anomaly detection process allows for a broader analysis of the data set by modeling the uncertainty of cluster membership. The evaluation of the proposed method was done by comparing the results obtained with the application of the methodology that uses rough k-means clustering (Lingras and Peters, 2012) and classical k-means clustering. Compared with the algorithm that used only crisp sets (generated by k-means clustering), the solution based on rough k-means clustering allowed better filtering of the resulting objects, which resulted in not including objects with the most common morphological characteristics of particle traces in the result set. In Sections 2.2 and 2.3 we will recall the most crucial methodology discussed earlier (Hachaj *et al*., 2023) and in Section 2.4 we will propose new methods to evaluate the effectiveness of different configurations of the anomaly detection algorithm.

2.1. Data set. We used a representative subset of observations from the CREDO project as the basis for the experimental verification of the hypotheses considered in this paper and for presenting the computational results. The data set we used was selected in such a way as to reflect the internal diversity of the observed signals. From the pool of data available for analysis registered under CREDO infrastructure, containing approximately 10^7 of events, a set of 10^4 of samples was randomly selected. We used only image data and omitted all other metadata. The obtained data set contains all known types of observed signals recorded by the CREDO detectors (Bibrzycki *et al*., 2020), i.e., dots, tracks, worms, as well as other atypical observations that meet the criteria for recognizing them as potential cosmic-ray particle tracks.

The shape morphology of cosmic-ray particle tracks is a factor that is considered when determining the signal type. The subset does not contain incorrect signals resulting from various measurement errors, referred to in the nomenclature as artifacts (Bibrzycki *et al*., 2020; Piekarczyk *et al*., 2021). Examples of signal types appearing in the CREDO data set are shown in Fig. 1. The sample selection procedure was preceded by filtering out hardware and acquisition artifacts (Bar *et al*., 2021; Piekarczyk *et al*., 2021). Due to the specificity of recording traces of high-energy cosmic ray particles (Hachaj and Piekarczyk, 2023), and the limitations of both automatic and human classification, unusual observations (anomalies) may occur in basically every class of recorded signals. In other words, there is no certainty which objects in a given class may constitute unusual observations (anomalies) from the point of view of morphological and physical interpretation (Homola *et al*., 2020). Therefore, the problem we are dealing with is a classical analysis of an unlabelled data set. RGB images represent each particle observation with a resolution of 60×60 pixels. The subset of the CREDO data set we use in this work consists of 13804 instances.

2.2. Image processing. In order to effectively compare image data sets, it is necessary to generate an appropriate embedding that preserves the interrelationships between the elements of the data set. In this case, the embedding should allow us to search for objects that differ in some way from other typical cosmic-ray particle tracks in terms of morphology. Knowing this, we should base embedding on statistical relationships in the data set, such as the analysis of variance. Effective methods for generating embedding using the analysis of variance are algorithms similar to Eigenfaces (Hachaj *et al*., 2021; Turk and Pentland, 1991). In this approach, the feature

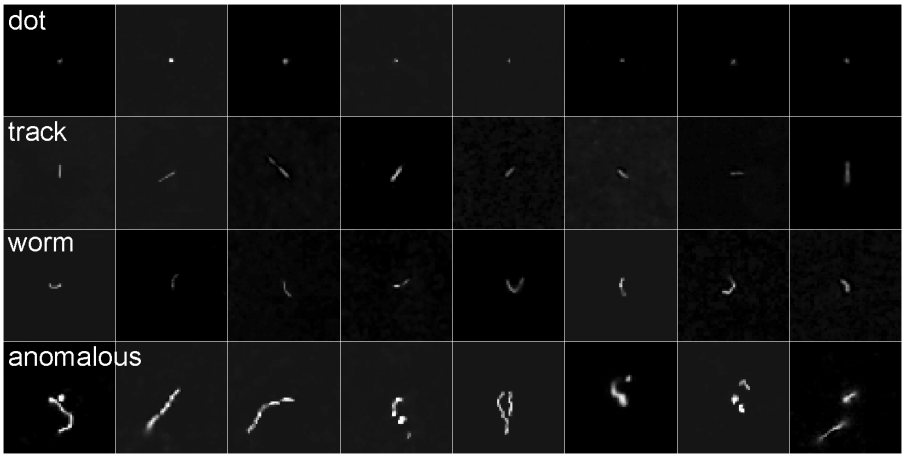

Fig. 1. Examples of types of signals in the CREDO observation data set. The illustration shows the basic classes of useful signals acquired using CMOS sensors. Artifacts resulting from measurement errors and incorrect calibration were omitted as irrelevant to the research issues of the article. Signal types are described in the first column.

vector of images is generated as a linear combination of image coordinates after projecting them onto a space, the coordinate system of which is calculated based on the variance analysis of the entire data set. The axes of the new system follow the principal components of the covariance matrix COV created from the individual vectors of the original data set.

$$COV = \frac{1}{s} D^T D, \tag{1}$$

where D is a matrix in which columns are created from flattened images; an averaged image value M calculated from the entire data set is subtracted from each image.

$$D = [I_1 - M, \ldots, I_s - M], \tag{2}$$

where I_1 is the first image from the data set and there are s images.

The analysis uses the well-known principal components analysis (PCA) approach. An important fact is that image analysis based on eigendecomposition with PCA is very sensitive to even minor variance distortions, so before applying this approach to an image data set, it is necessary to normalize the data set (to perform the so-called image aligning). In the case of our data set, aligning consisted of translating the center of mass of the image so that its newly calculated center of mass is at the center of the image, and rotating the image so that the axis relative to which the variance of nonzero pixels has the most significant value becomes the axis parallel to the horizontal axis. This is also done using PCA, which is calculated on each image from the data set separately.

As a result of applying the eigendecomposition of the COV matrix, we get a new coordinate system in which our embedding will be expressed. Each of the axes of this system can be interpreted in a similar way to the eigenfaces approach. The axes responsible for the higher variance of the data have the characteristics of components responsible for the low-frequency deviation from the average image, and further axes are responsible for high-frequency deviations. The images in our data set have a resolution of 60×60, so the corresponding vectors have 3600 elements. After PCA analysis, we limited the number of dimensions to express 95% of the variance. In the case of our data set, these were the first 62 dimensions. Thus, in the rest of the paper, we will work on the 62-dimensional embedding of our image data set.

It is also possible to perform feature extraction and simultaneously dimensionality reduction using a deep learning approach using a deep encoder-decoder (E-D) architecture (Pang *et al.*, 2021; Wei and Mahmood, 2021). To perform this, an E-D network is trained as an autoencoder. The latent space of such a trained network is used to generate a low dimensional embedding similar to the one calculated by PCA. The application of PCA for feature generation might have advantages over methods using the E-D. Although PCA is a linear method (which might be a disadvantage), it enables direct calculation of the variance explained by the features used for embedding. Also, it is easy to change the size of embedding just by omitting certain parts of matrix D (see Eqn. (2)), without the necessity of retraining the whole method. Due to this fact, in many cases the PCA-based approach for dimensionality reduction is more convenient, and its results are easier to explain, which is very important, especially while tuning a new data analysis method. Due to this, we have utilized PCA in our algorithm.

Algorithm 1. Algorithm for finding anomalous objects using rough k-means clustering.

Require: X—input data set, $k, \epsilon, w_{\text{upper}}, w_{\text{lower}}$—parameters of rough k-means algorithm described in Section 2.3, p—number of potentially anomalous objects to be returned

{perform rough k-means}

C $\leftarrow$ km$(X, k, \epsilon, w_{\text{upper}}, w_{\text{lower}})$ {calculate distances from centroid, assign a pair: element and its distance to centroid to set P}
$P \leftarrow \emptyset$
for $x_i \in X$ **do**
  for $c_j \in C$ **do**
    if $x_i \in \overline{C_j}$ **then**
      $P \leftarrow P \cup \{(x_i, E(x_i, c_j)\}$
    end if
  end for
end for{sort P by distances in descending order}
$P_{\text{sorted}} \leftarrow sort(P)$ {get first p unique objects x_i from P_{sorted}}
R $\leftarrow \emptyset$
s $\leftarrow 0$
while $|R| < p$ **do**
  if $P_{\text{sorted}}[s] \notin R$ **then**
    $R \leftarrow \{P_{\text{sorted}}[s]\}$
  end if
end while
return R {set of p potentially anomalous objects}

2.3. Potential anomaly detection. Using the embedding described in Section 2.1, anomalous images can be defined as those that are relatively far, given the Euclidean metric, from the other images. In other words, we want to find images whose embedding will have the maximum distance from the other objects in the set. We know that the image data set contains several classes of objects, morphologically different from each other (see Fig. 1). Similar objects will form clusters. An additional issue that we discussed before in Section 2.1 is that it is difficult to unambiguously determine precisely where in space the boundary between objects that belong to the classes dots, tracks, and worms should be defined and which objects might be counted as anomalous images. This has already been pointed out by preparing manual annotations for the CREDO data set, in which a group of annotators, through a blind voting process, determined to which class each object belongs (Piekarczyk *et al*., 2021). As a result of this process, the experiment described by Piekarczyk *et al*. (2021) eliminated those objects to which the annotators were uncertain about the class to which they belonged.

In our case, we do not make such a selection but use the entire data set. Thus, it is natural that if we want to perform an unsupervised analysis of data sets containing sets of objects against which even human annotators cannot make an unambiguous decision, it is reasonable to use an approach that allows modeling uncertainty in the decision-making process. An approach that enables uncertainty modeling in the clustering process is rough k-means clustering. The algorithm, described by Lingras and Peters (2012), adds several improvements to the classic k-means algorithm introduced by Forgey (1965) and Lloyd (1982). The object's cluster membership is defined using rough set methodology.

Assume that objects are represented by n-dimensional vectors and are contained in the set X. In the classical approach, finding the nearest centroid for an object $x_j \in X$ is done by optimizing the following expression:

$$d(x_j)_{\min} = \min_{i \in k} E(x_j, c_i), \tag{3}$$

where k is the number of clusters represented by centroids, c_i is the centroid of cluster C_i with index i and E is the Euclidean metric.

Assume that $\overline{C_i}$ and $\underline{C_i}$ are upper and C_i is a lower approximations of cluster $\overline{C_i}$. In rough k-means, the object belongs not only to the closest cluster in terms of distance to the centroid, but also to all other clusters to whose centroids the distance satisfies the condition

$$\frac{d(x_j)_{\min}}{E(x_j, c_l)} \leq \epsilon, \quad l \neq i, \tag{4}$$

where ϵ is the threshold of the method and l is the index of the centroid that does not minimize (3). If $\epsilon \leq 1$, then rough k-means is performed like a classical k-means. If $\epsilon > 1$ then

- if (4) is satisfied then $x_j \in \overline{C_i}, x_j \in \overline{C_l}$, which means

that x_j belongs to at least two upper approximations of clusters: ($\overline{C_i}$ and all $\overline{C_i}$ that satisfies (4)),

- if (4) is not satisfied then $x_j \in \overline{C_i}, x_j \in \underline{C_i}$

In rough k-means, it is also necessary to modify the centroid updating algorithm, which takes the form of a weighted sum:

$$c_m = w_{\text{upper}} \frac{\sum v \in \overline{C_m}}{\left|\overline{C_m}\right|} + w_{\text{lower}} \frac{\sum v \in \underline{C_m}}{\left|\underline{C_m}\right|}, \quad (5)$$

where $\left|\overline{C_m}\right|$ is cardinal number of set $\overline{C_m}$ and $w_{\text{upper}} + w_{\text{lower}} = 1$. If $\left|\overline{C_m}\right| = 0$ or $\left|\underline{C_m}\right| = 0$ then only the component in which the denominator is nonzero is taken into the sum.

To use rough k-means to find anomalies, we need to calculate the distances of each element in the set to the centroid of the cluster to which this element has been assigned. Since we want to consider the assignment uncertainty to a cluster, we use the *upper approximation of each cluster*. For this reason, each object can be assigned to more than one cluster. Suppose we want to identify p anomalous elements in our data set. To do this, we order the objects belonging to the upper approximation of each cluster by their distance from the centroid of the cluster they belong to and take p unique outermost elements.

There are three operations in the proposed solution that have significant computational complexity. In particular, these are calculating the covariance matrix, PCA solving with singular value decomposition (SVD), and rough k-means. These algorithms are performed sequentially one by one. Assuming no parallel computation, the computational complexity is $\mathcal{O}(n^3)$ where n is proportional to the number of elements in the data set and the resolution of the images.

The main goal of our research was to solve the task of detecting potential anomalous particle tracks in large data sets. Our previous study (Hachaj *et al.*, 2023) showed that we can benefit by applying a soft clustering approach, namely rough k-means, by getting more reliable results than hard clustering. Besides the rough k-means approach, there are many other methods with which one can obtain soft clusters; among them is three-way clustering (Yao, 2010; Wang *et al.*, 2024; Yu, 2017; 2018). In three-way clustering, each cluster is represented by three regions: objects in the core region belong to the cluster definitely, objects in the trivial region do not belong to the cluster definitely, and objects in the fringe region are the boundary elements of the cluster. The clustering process is governed by at least two parameters (besides the number of clusters k), which are thresholds for assigning objects to one of those three regions. We decided to apply rough k-means over other approaches because it is easier to tune and evaluate (it has only one threshold parameter besides the number of clusters k) and, as it has been shown, is sufficient for our needs.

2.4. Evaluation of results and optimizing parameters of the algorithm. Our proposed algorithm for detecting asymmetries works on unlabeled data sets. We neither have labels describing the shapes (morphology) of the particle traces, nor a numerical measure describing the degree of anomalous shapes. Thus, the only way to evaluate the algorithm's effectiveness is based on the statistical properties of the results obtained, and the potential anomalies found cannot be typical shapes such as dots, tracks, or worms (of which there are most in the data set). This is further complicated by the fact that there is no gold standard for describing the morphology of these shapes, for example, the diameter of a dot, the length, and linearity of a track, or the degree of curvature of a worm. For example, a sufficiently heavily curved track can be considered a worm. For this reason, anomaly detection using only the shape of a particle trace has a threshold character, in which the method user decides when he or she is still dealing with an anomalous signal and when he or she is already dealing with a typical signal. Accordingly, the performance of the algorithm for anomaly detection can be evaluated by the following features:

1. The ability of an algorithm to produce an embedding in which atypical elements are at a relatively large distance from typical elements and morphologically similar elements are close together. This assumption is realized by PCA-based embedding, which generates an orthogonal coordinate system that optimally describes variance in a single data set. Due to this fact, the quality of the embedding will not be taken into account in algorithm evaluation. However, we must remember that PCA is a linear transform (see the discussion in Section 2.2).

2. Stability of the algorithm understood as the repeatability of the results if the input data set to be analyzed differs slightly from each other. This means that in a data set that contains a certain number of common elements and a certain number of different elements, the algorithm should be able to search for anomalies that will be found in both sets, especially those anomalies that are a common part of both sets of objects. This means that among the available configurations of the algorithm for detecting potential anomalies, differing in parameters such as k and ϵ, one will be chosen that maximizes the measure of similarity of the algorithm's results obtained for different subsets of the test set.

3. The degree of similarity of the results obtained

Table 1. Values of IOU_{score} (10) calculated on the data set presented in Section 2.1. We set the number of subsets in (7) to $s = 10$.

$k \setminus \epsilon$	1.05	1.25	1.45	1.65	1.85	2.05	2.25	2.45	2.65	2.85
2	0.74	0.76	0.76	0.76	0.76	0.75	0.72	0.72	0.72	0.72
4	0.72	0.73	0.69	0.71	0.71	0.70	0.71	0.71	0.71	0.71
6	0.65	0.68	0.68	0.69	0.69	0.70	0.70	0.70	0.70	0.70
8	0.72	0.64	0.68	0.68	0.69	0.70	0.70	0.70	0.70	0.70
10	0.69	0.61	0.69	0.69	0.69	0.70	0.70	0.70	0.70	0.70

Table 2. Values of $IOU2_{score}$ (11) calculated on the data set presented in Section 2.1. We set the number of subsets in (7) to $s = 10$.

$k \setminus \epsilon$	1.05	1.25	1.45	1.65	1.85	2.05	2.25	2.45	2.65	2.85
2	0.92	0.94	0.94	0.95	0.94	0.94	0.89	0.89	0.9	0.9
4	0.89	0.91	0.86	0.88	0.88	0.87	0.89	0.89	0.89	0.89
6	0.80	0.84	0.85	0.86	0.86	0.87	0.87	0.87	0.87	0.86
8	0.89	0.78	0.85	0.85	0.86	0.87	0.87	0.86	0.86	0.86
10	0.85	0.74	0.85	0.85	0.86	0.87	0.87	0.87	0.87	0.86

Table 3. Values of IOU_{comp} (12) calculated on the data set presented in Section 2.1. We set the number of subsets in (7) to $s = 10$.

$k \setminus \epsilon$	1.05	1.25	1.45	1.65	1.85	2.05	2.25	2.45	2.65	2.85
2	0.71	0.97	0.99	1.0	0.99	0.94	0.92	0.91	0.91	0.91
4	0.75	0.97	0.94	0.95	0.95	0.94	0.95	0.94	0.95	0.95
6	0.69	0.89	0.91	0.91	0.91	0.92	0.92	0.92	0.92	0.92
8	0.60	0.76	0.91	0.90	0.90	0.93	0.93	0.92	0.92	0.92
10	0.59	0.73	0.9	0.91	0.91	0.93	0.93	0.93	0.93	0.92

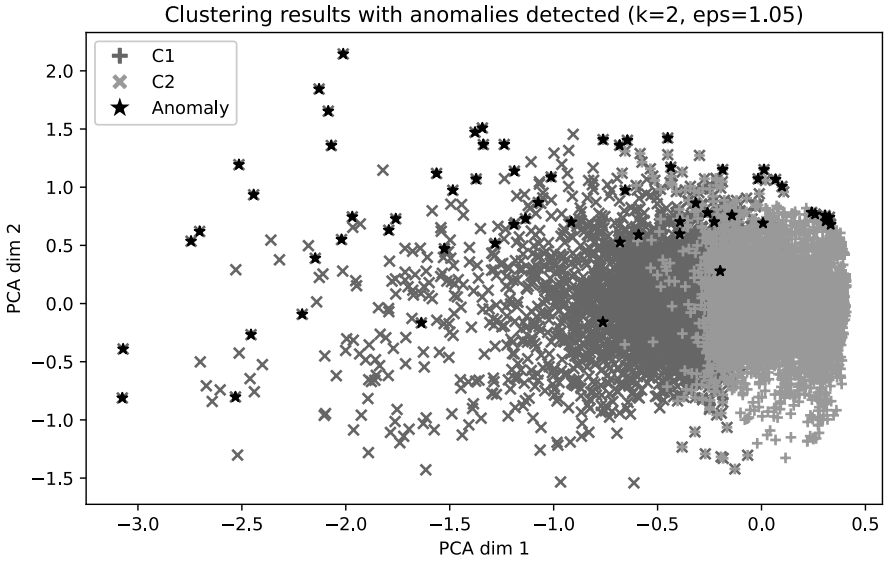

Fig. 2. Example clustering results for Algorithm 1 with $k = 2$ and $\epsilon = 1.05$. Objects assigned to the cluster are indicated by a marker ('+' or '×') and a certain shade of gray. Detected potential anomalies are marked as black stars. We plot upper approximations of clusters, so one object may be assigned to one or two clusters.

(set of anomalies) between the configuration of the algorithm for which the highest result was obtained and other configurations of the algorithm. If different configurations of the algorithm, differing in the values of k and ϵ, give entirely different results, that is, the measure of similarity between the results is relatively low, this means that the algorithm works in a chaotic and unstable way and is probably not suitable for solving the problem of anomaly analysis, because it is too sensitive to the choice of parameters.

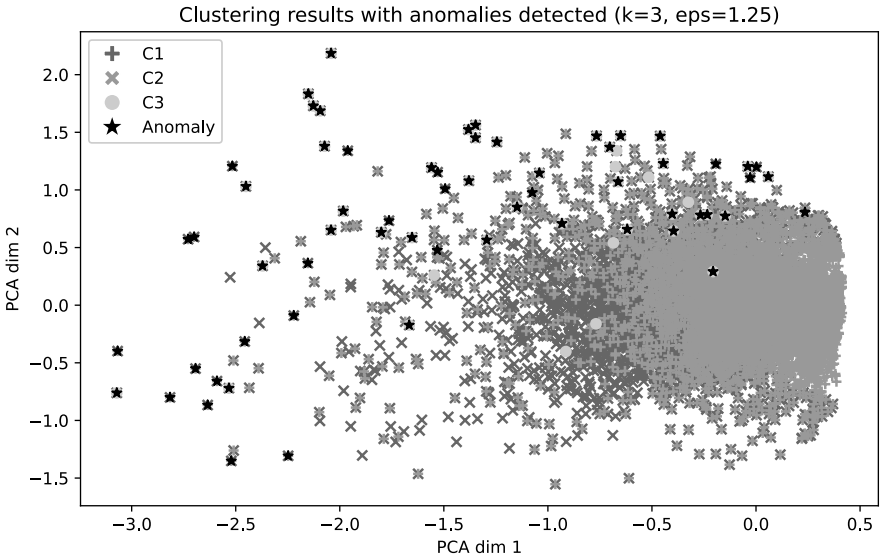

Fig. 3. Example clustering results for Algorithm 1 with $k = 3$ and $\epsilon = 1.25$. Objects assigned to the cluster are indicated by a marker ('+,' '×' or circle) and a certain shade of gray. Detected potential anomalies are marked as black stars. We plot upper approximations of clusters, so one object may be assigned to one, two, or three clusters.

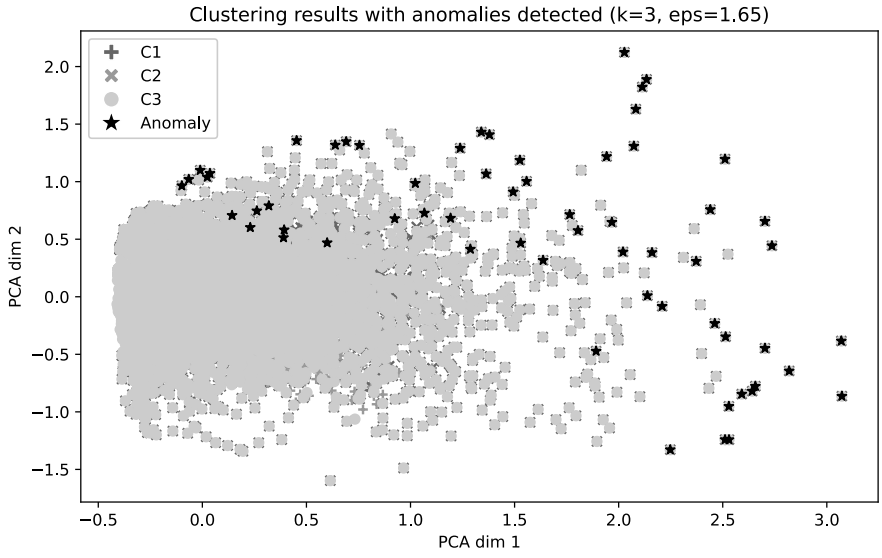

Fig. 4. Example clustering results for Algorithm 1 with $k = 3$ and $\epsilon = 1.65$. Objects assigned to the cluster are indicated by a marker ('+,' '×' or circle) and a certain shade of gray. Potential detected anomalies are marked as black stars. We plot upper approximations of clusters, so one object may be assigned to one, two, or three clusters.

In practice, to measure the stability of the algorithm understood according to the second point above, we can apply a leave-one-out cross-validation test. The training data set should be divided into s equal parts, that is, the sets X_i should have similar counts,

$$X = \bigcup_{i=1}^{s} X_i. \tag{6}$$

Then we from s subsets of the set X defined as follows:

$$\widehat{X_i} = X \setminus X_i, \quad i \in [1, s] \tag{7}$$

and execute Algorithm 1 on each $\widehat{X_i}$ from (7).

Suppose that for sets A and B, the anomalies detected by Algorithm 1 are in sets $R(A)$ and $R(B)$. The following two metrics can determine the similarity between the sets $R(A)$ and $R(B)$. The first one is the

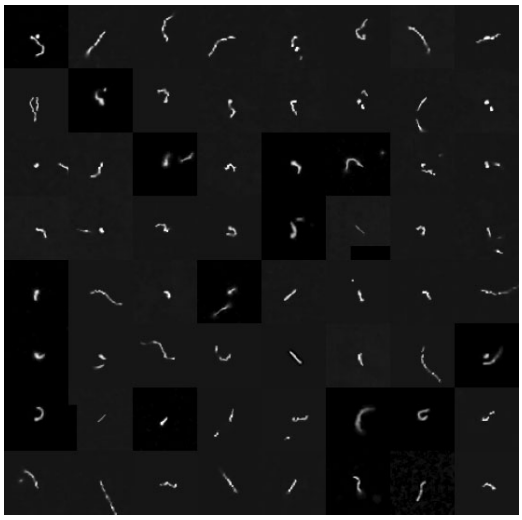

Fig. 5. Set of first 64 potential anomalies returned by Algorithm 1 when $k = 2$ and $\epsilon = 1.65$.

intersection over union (IOU):

$$\text{IOU}(R(A), R(B)) = \frac{R(A) \cap R(B)}{R(A) \cup R(B)}. \quad (8)$$

The second is the modified intersection over union (IOU2):

$$\text{IOU2}(R(A), R(B)) = \frac{(R(A) \cap B) \cap (R(B) \cap A)}{(R(A) \cap B) \cup (R(B) \cap A)}. \quad (9)$$

In the case of Eqn. (9), we take into account only those potential anomalies that are present both in A and B sets. The value returned by (9) is not less than the value returned by (8). Here (9) measures the similarity between potential anomalies, taking into account only objects that are present in both input object sets for Algorithm 1. For this reason, the values returned by (9) will more meaningfully count the similarity between the results obtained by the algorithm for different $\widehat{X_i}$. To estimate the configuration of the algorithm that works most stably, a minimum should be found:

$$\text{IOU}_{\text{score}}(k, \epsilon) = avg(\text{IOU}(R_{(k,\epsilon)}(\widehat{X_i}), R_{(k,\epsilon)}(\widehat{X_j}))), \quad i < j, \quad i, j \in [1, s], \quad (10)$$

and/or

$$\text{IOU2}_{\text{score}}(k, \epsilon) = avg(\text{IOU2}(R_{(k,\epsilon)}(\widehat{X_i}), R_{(k,\epsilon)}(\widehat{X_j}))), \quad i < j, \quad i, j \in [1, s], \quad (11)$$

where avg is the averaging function, $R_{(k,\epsilon)}(\widehat{X_i})$ is potential anomaly dateset returned by Algorithm 1 with certain values of k and ϵ. As can be seen, we calculate the average values of IOU and IOU2 between all possible pairs of subsets defined by (7).

In estimating the degree of similarity of the obtained results (that is, sets of anomalies) between the different algorithm configurations and the best performing algorithm due to (10), we can use the following estimate:

$$\text{IOU}_{\text{comp}}(R_{(k_1,\epsilon_1)}, R_{(k_2,\epsilon_2)}) = avg(\text{IOU}(R_{(k_1,\epsilon_1)}(\widehat{X_i}), R_{(k_2,\epsilon 2)}(\widehat{X_i}))), \quad i \in [1, s], \quad (12)$$

where $(k_1, \epsilon 1)$ are parameters of the first algorithm and $(k_2, \epsilon 2)$ are parameters of the second algorithm.

Soft clustering resulting from rough k-means, depending on its parameter ϵ at a fixed k, can generate different assignments of objects to clusters. In such cases, a single object can be assigned to a different number of clusters. Intuitively, the wider the clusters' border regions, the more clusters the object will be assigned. In such a case, using intuitive and popular similarity measures between two clusterings, considering all pairs of samples, like the Rand Index, cannot be applied. In addition, our interest is to see how different configurations of the proposed algorithm affect the detection of outliers. The problem of detecting outliers differs from the problem of clustering the set because small fluctuations at the border regions of the clusters, practically not measurable when considering all the objects in the large data set, can cause significant differences in detecting outliers. For this reason, we used the methods described by Eqns. (10)–(12) to evaluate our solution.

3. Results

We implemented the methodology presented in Section 2 in the Python the programming language. We utilized packages NumPy 1.22, OpenCV-Python 4.5, and the modified package `https://github.com/geofizx/rough-clustering` (accessed on 10 January 2025) so that it can work with Python 3.X. For evaluation purposes, we have used the data set described in Section 2.1. Results can be reproduced using source code published in an open repository at `https://github.com/browarsoftware/rough-k-means-particles` (accessed on 10 January 2025). We evaluated the results and optimization of parameters of Algorithm 1 using the methodology described in Section 2.4. We have set the number of subsets of (7) to $s = 10$. In Table 1 we present the values of $\text{IOU}_{\text{score}}$ in Table 2 values of $\text{IOU2}_{\text{score}}$ and in Table 3 the values of IOU_{comp}. The parameters w_{upper} and w_{lower} in (5) were set to 0.9 and 0.1, respectively. The

range of parameters was chosen experimentally so that the method would achieve convergence at 10^{-4}. The potential anomalies set size returned by Algorithm 1 was set to 0.5% of the input data set $\widehat{X}_i$.

In Figures 2–4 we present example clustering results for various configuration of Algorithm 1. Objects assigned to the cluster are indicated by a marker ('+,' '$\times$' or circle) and a shade of gray. Detected potential anomalies are marked as black stars. We plot upper approximations of clusters so that one object may be assigned to one, two, or three clusters. In Fig. 5 we present the set of the first 64 anomalies returned by Algorithm 1 ($k = 2, \epsilon = 1.65$). Results in Figs. 2–5 are obtained for various subsets of X; however, as will be discussed in Section 4, the selection of subset (7) for certain configurations of Algorithm 1 does not change the result much.

4. Discussion

The results presented in Section 3 confirm the effectiveness of the proposed approach. The clustering method behaves as expected. An increase in ϵ in (4) increases the size of upper approximations of clusters. As seen in Figs. 2–4, the higher ϵ, the more objects are assigned to more than one cluster. Consequently, higher values of ϵ increase the search space of potential anomalies according to Algorithm 1 because there are more objects that belong to an upper approximation of each cluster.

Rough k-means clustering enhances the search space of potential anomalies (outliers). When we use standard k-means clustering, list P in Algorithm 1 contains only pairs of elements and their distance to the nearest centroid. After applying rough k-means, list P also includes pairs of elements and their distance to all centroids that satisfy (4). Due to this fact, a single object x_i can be present several times in P, with possible various distances from centroids of clusters to which the upper approximation belongs. Parameter ϵ of rough k-means enables enhancing the algorithm's search space by widening cluster border regions. As shown by Hachaj *et al.* (2023), widening the search space allowed better filtering of the resulting objects, which resulted in not including objects with the most common morphological characteristics of particle traces in the result set. In the problem of analyzing cosmic particle tracks, the fact that it belongs to one or many upper approximations of clusters does not directly indicate whether the object should be treated as an anomaly. Certainly, objects that are relatively far from other objects are statistically different from them. This is a consequence of the PCA-based embedding. Belonging to the upper approximation, or a border region in general, means that the object contains features that may be typical of several shape topologies that have been assigned to different clusters. Anomalies can also be located in areas far from the cluster's center but not in the border area and, thus, according to (4), will be included in the lower approximation of the cluster.

As can be seen in Fig. 2, when ϵ is relatively low, most of the objects are in the lower approximation of each cluster, and there is some fraction of objects that belong to the upper approximation of clusters. The higher the ϵ becomes, the more objects that are farther from the cluster center are attached to the upper approximation. This situation can be observed as the increase in the fraction of objects that belong to more than one upper approximation of clusters. As shown in Fig. 3 in our case, when $\epsilon = 1.25$ and $k = 3$, we can observe a group of objects assigned to one, two, or three clusters. In Fig. 3, which presents a situation when $\epsilon = 1.65$ and $k = 3$, most objects belong to an upper approximation of each cluster. Equation (4), which governs the size of the diagonal of the upper approximation of the cluster, is a distance-based approach and creates hyper-spherical clusters.

The appropriate value of ϵ and k for finding potential anomalies in the data set depends on the distribution of objects in the space. Our approach for determining those two parameters described in Section 2.4 seems reasonable. As can be seen in Tables 1 and 2, both scoring methods indicate that very similar of configurations Algorithm 1 have the highest values. In the case of $\text{IOU}_{\text{score}}$ highest averaged IOU was obtained for $k = 2$ and $\epsilon = 1.45, 1.65, 185$ and equals 0.76. In the case of $\text{IOU}_{\text{score}}$ the highest value was obtained for $(k = 2, \epsilon = 1.65)$ and equals 0.95. This means that nearly all potential anomaly data sets except for about 5% of data are common between various subsets (7). As seen in both the tables for all tested configurations, the proposed approach has very nice stability. In the case of $\text{IOU}_{\text{score}}$ its value does not drop below 0.61 and in the case of $\text{IOU2}_{\text{score}}$ below 0.74.

Thus, our method gives a stable solution, where potentially anomalous signals are usually returned among those in the set R (see Algorithm 1). In our case, with a leave-one-out test for (7) $s = 10$ (10 subsets) where $\text{IOU}(\widehat{X}_i, \widehat{X}_j) = 0.8$ these are very satisfactory results.

We can conclude that the sets of potential anomalies were almost identical, and detecting anomalies is not susceptible to fluctuations within the analyzed data sets. The situation is also similar when comparing the set of potential anomalies obtained with Algorithm 1 with $(k = 2, \epsilon = 1.65)$ with other configurations of the algorithm in Table 3. A high value of IOU_{comp} that does not fall below 0.6 indicates that the algorithm does not operate chaotically, and a relatively continuous relationship exists between the resulting set of anomalies and the algorithm parameters.

Example results of potential anomalous trajectories presented in Fig. 5 also indicate that the proposed approach works as expected. As images are ordered

row-by-row by decreasing the distance from the nearest cluster center, the top-left image is the most anomalous in the data set, the second in the top row is the second most anomalous, etc. The shapes of detected trajectories agree with our expectations of what type of trajectories we more or less expected. The farthest from cluster centers (most anomalous) are often multipart traces, longer than typical track and worm signals, or consisting of more than one "typical" trajectory, like, for example, the top-left, which is a combination of a very long worm-like shape and the dot. The closer we get to the centers of clusters, the more typical the signals become; for example, in the fourth line and sixth column, there is a track-shaped signal followed by a worm. The last row of Fig. 5 contains, in fact, only typical signal traces.

5. Conclusion

In summary, the method presented in this work proved to be an effective algorithm for the detection of potentially anomalous cosmic particle tracks acquired with CMOS sensors. The analysis of the behavior of the rough k-means clustering-based algorithm presented in this work and the method of selecting its parameters showed that the algorithm performs as expected and demonstrates efficiency, stability and repeatability of results for test data set. The results presented in this work are very relevant to the CREDO project as well as to the wider problem of anomaly analysis in the image data set. We plan to deploy the methodology presented in this work in the image processing pipeline of the large data set we are working on in the CREDO project.

With the method proposed in this work, we find potential anomalies only based on the analysis of the particle tracks recorded on the CMOS array. We do not have direct measurements of the energy values of the particles and we do not use additional metadata that are collected during the detection of an event such as time, the geographical position of the sensor or the orientation of the sensor in space. For this reason, the method proposed in this work can serve as a kind of trigger indicating that certain images, which are in a large image repository, are worth further analysis based on other physical measurements.

A topic worth further investigation is the application of the method proposed in our work to anomaly detection in image collections of other modalities. This will probably require the use of other embedding methods, for example deep encoder-decoder-based approaches (Jewell *et al*., 2022; Fan *et al*., 2020).

Acknowledgment

We sincerely thank the entire CREDO team for providing the data set that was used to validate the algorithm proposed in this work (`https://credo.science` (accessed on 10 January 2025)).

References

Afridi, M.K., Azam, N., Yao, J. and Alanazi, E. (2018). A three-way clustering approach for handling missing data using GTRS, *International Journal of Approximate Reasoning* **98**: 11–24.

Albin, E.K. and Whiteson, D. (2023). Feasibility of correlated extensive air shower detection with a distributed cosmic-ray network, *Astrophysical Journal* **954**(1): 106.

Bar, O., Bibrzycki, Ł., Niedźwiecki, M., Piekarczyk, M., Rzecki, K., Sośnicki, T., Stuglik, S., Frontczak, M., Homola, P., Alvarez-Castillo, D.E., Andersen, T. and Tursunov, A. (2021). Zernike moment based classification of cosmic ray candidate hits from CMOS sensors, *Sensors* **21**(22): 7718.

Bibrzycki, Ł., Burakowski, D., Homola, P., Piekarczyk, M., Niedźwiecki, M., Rzecki, K., Stuglik, S., Tursunov, A., Hnatyk, B., Castillo, D.E.A., Smelcerz, K., Stasielak, J., Duffy, A. R., Chevalier, L., Ali, E., Lakerink, L., Poole, G. B., Wibig, T. and Zamora-Saa, J. (2020). Towards a global cosmic ray sensor network: Credo detector as the first open-source mobile application enabling detection of penetrating radiation, *Symmetry* **12**(11): 1802.

Crispim Romão, M., Castro, N.F. and Pedro, R. (2021). Finding new physics without learning about it: Anomaly detection as a tool for searches at colliders, *European Physical Journal C* **81**(1): 27.

Fan, Z., Li, C., Chen, Y., Wei, J., Loprencipe, G., Chen, X. and Di Mascio, P. (2020). Automatic crack detection on road pavements using encoder-decoder architecture, *Materials* **13**(13): 2960.

Forgey, E. (1965). Cluster analysis of multivariate data: Efficiency vs. interpretability of classification, *Biometrics* **21**(3): 768–769.

Hachaj, T., Koptyra, K. and Ogiela, M.R. (2021). Eigenfaces-based steganography, *Entropy* **23**(3): 273.

Hachaj, T. and Piekarczyk, M. (2023). The practice of detecting potential cosmic rays using CMOS cameras: Hardware and algorithms, *Sensors* **23**(10): 4858.

Hachaj, T., Piekarczyk, M. and Wąs, J. (2023). Searching of potentially anomalous signals in cosmic-ray particle tracks images using rough k-means clustering combined with eigendecomposition-derived embedding, *in* A. Campagner *et al.* (Eds), *Rough Sets*, Springer Nature Switzerland, Cham, pp. 431–445.

Homola, P., Beznosko, D., Bhatta, G., Bibrzycki, Ł., Borczyńska, M., Bratek, Ł., Budnev, N., Burakowski, D., Alvarez-Castillo, D.E., Cheminant, K.A., Ćwikła, A., Dam-o, P., Dhital, N., Duffy, A.R., Głownia, P., Gorzkiewicz, K., Góra, D., Gupta, A.C., Hlávková, Z., Homola, M., Jałocha, J., Kamiński, R., Karbowiak, M., Kasztelan, M., Kierepko, R., Knap, M., Kovács, P., Kuliński, S., Łozowski, B., Magryś, M., Medvedev, M.V., Mędrala, J., Mietelski, J.W., Miszczyk, J., Mozgova, A.,

Napolitano, A., Nazari, V., Ng, Y.J., Niedźwiecki, M., Oancea, C., Ogan, B., Opiła, G., Oziomek, K., Pawlik, M., Piekarczyk, M., Poncyljusz, B., Pryga, J., Rosas, M., Rzecki, K., Zamora-Saa, J., Smelcerz, K., Smolek, K., Stanek, W., Stasielak, J., Stuglik, S., Sulma, J., Sushchov, O., Svanidze, M., Tam, K.M., Tursunov, A., Vaquero, J.M., Wibig, T., and Woźniak, K.W. (2020). Cosmic-ray extremely distributed observatory, *Symmetry* **12**(11): 1835.

Javan, H. (2002). Characterizing sensitive CMOS radiation detector, *Proceedings of IEEE SoutheastCon 2002, Columbia, USA*, pp. 28–32.

Jewell, J.T., Khazaie, V.R. and Mohsenzadeh, Y. (2022). One-class learned encoder-decoder network with adversarial context masking for novelty detection, *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, USA*, pp. 3591–3601.

Johary, Y.H., Trapp, J., Aamry, A., Aamri, H., Tamam, N. and Sulieman, A. (2021). The suitability of smartphone camera sensors for detecting radiation, *Scientific Reports* **11**(1): 12653.

Karbowiak, M., Wibig, T., Alvarez Castillo, D., Beznosko, D., Duffy, A.R., Góra, D., Homola, P., Kasztelan, M. and Niedźwiecki, M. (2021). Determination of zenith angle dependence of incoherent cosmic ray muon flux using smartphones of the credo project, *Applied Sciences* **11**(3): 1185.

Kuusela, M., Vatanen, T., Malmi, E., Raiko, T., Aaltonen, T. and Nagai, Y. (2012). Semi-supervised anomaly detection—Towards model-independent searches of new physics, *Journal of Physics: Conference Series* **368**: 012032.

Lingras, P. and Peters, G. (2012). Applying rough set concepts to clustering, *in* G. Peters *et al.* (Eds), *Rough Sets: Selected Methods and Applications in Management and Engineering*, Springer, Berlin/Heidelberg, pp. 23–37.

Lloyd, S. (1982). Least squares quantization in PCM, *IEEE Transactions on Information Theory* **28**(2): 129–137.

Pang, G., Shen, C., Cao, L. and Hengel, A.V.D. (2021). Deep learning for anomaly detection: A review, *ACM Computing Surveys* **54**(2): 1–38.

Pawlak, Z. (1982). Rough sets, *International Journal of Computer & Information Sciences* **11**: 341–356.

Peters, J.F., Skowron, A., Suraj, Z., Rzasa, W. and Borkowski, M. (2002). Clustering: A rough set approach to constructing information granules, *Proceedings of the 6th International Conference on Soft Computing and Distributed Processing, SCDP, Rzeszów, Poland*, pp. 57–61.

Piekarczyk, M., Bar, O., Bibrzycki, Ł., Niedźwiecki, M., Rzecki, K., Stuglik, S., Andersen, T., Budnev, N.M., Alvarez-Castillo, D.E., Cheminant, K.A., Góra, D., Gupta, A.C., Hnatyk, B., Homola, P., Kamiński, R., Kasztelan, M., Knap, M., Kovács, P., Łozowski, B., Miszczyk, J., Mozgova, A., Nazari, V., Pawlik, M., Rosas, M., Sushchov, O., Smelcerz, K., Smolek, K., Stasielak, J., Wibig, T., Woźniak, K.W. and Zamora-Saa, J. (2021). CNN-based classifier as an offline trigger for the credo experiment, *Sensors* **21**(14): 4804.

Pięta, P. and Szmuc, T. (2021). Applications of rough sets in big data analysis: An overview, *International Journal of Applied Mathematics and Computer Science* **31**(4): 659–683, DOI: 10.34768/amcs-2021-0046.

Riza, L. S., Janusz, A., Bergmeir, C., Cornelis, C., Herrera, F., Ślęzak, D. and Benítez, J.M. (2014). Implementing algorithms of rough set theory and fuzzy rough set theory in the R package "RoughSets", *Information Sciences* **287**: 68–89.

Skowron, A. and Dutta, S. (2018). Rough sets: Past, present, and future, *Natural Computing* **17**: 855–876.

Skowron, A. and Ślęzak, D. (2022). Rough sets turn 40: From information systems to intelligent systems, *17th Conference on Computer Science and Intelligence Systems (FedCSIS), Sofia, Bulgaria*, pp. 23–34.

Stein, G., Seljak, U. and Dai, B. (2020). Unsupervised in-distribution anomaly detection of new physics through conditional density estimation, *arXiv:* 2012.11638.

Turk, M. and Pentland, A. (1991). Face recognition using eigenfaces, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Maui, USA*, pp. 586–591.

Vandenbroucke, J., BenZvi, S., Bravo, S., Jensen, K., Karn, P., Meehan, M., Peacock, J., Plewa, M., Ruggles, T., Santander, M., Schultz, D., Simons, A.L. and Tosi, D. (2016). Measurement of cosmic-ray muons with the distributed electronic cosmic-ray observatory, a network of smartphones, *Journal of Instrumentation* **11**(04): P04019.

Vandenbroucke, J., Bravo, S., Karn, P., Meehan, M., Plewa, M., Ruggles, T., Schultz, D., Peacock, J. and Simons, A.L. (2015). Detecting particles with cell phones: The distributed electronic cosmic-ray observatory, *arXiv:* 1510.07665.

Wang, P., Yang, X., Ding, W., Zhan, J. and Yao, Y. (2024). Three-way clustering: Foundations, survey and challenges, *Applied Soft Computing* **151**: 111131.

Wang, P. and Yao, Y. (2018). CE3: A three-way clustering method based on mathematical morphology, *Knowledge-Based Systems* **155**(1): 54–65.

Wei, R. and Mahmood, A. (2021). Recent advances in variational autoencoders with representation learning for biomedical informatics: A survey, *IEEE Access* **9**: 4939–4956.

Whiteson, D., Mulhearn, M., Shimmin, C., Cranmer, K., Brodie, K. and Burns, D. (2016). Searching for ultra-high energy cosmic rays with smartphones, *Astroparticle Physics* **79**: 1–9.

Yao, Y. (2010). Three-way decisions with probabilistic rough sets, *Information Sciences* **180**(3): 341–353.

Yu, H. (2017). A framework of three-way cluster analysis, *in* L. Polkowski *et al.* (Eds), *Rough Sets: International Joint Conference, IJCRS 2017*, Springer, Cham, pp. 300–312.

Yu, H. (2018). Three-way decisions and three-way clustering, *in* H.S. Nguyen *et al.* (Eds), *Rough Sets: International Joint Conference, IJCRS 2018*, Springer, Cham, pp. 13–28.

Tomasz Hachaj received his PhD and DSc degrees in computer science in 2010 and 2017, respectively. He is employed as a professor at the AGH University of Krakow. His research interests include machine learning, deep neural networks, analysis of complex signals and their applications in the analysis of physical phenomena, human motion analysis, sports, medicine, social media issues. He has authored and co-authored more than 100 publications. He has been the principal investigator and co-investigator of scientific and commercial projects.

Marcin Piekarczyk received his PhD degree in computer science in 2011. His research interests include machine learning, graph analysis techniques, deep learning architectures and biometrics, as well as applications of artificial intelligence methods in computer physics, movement pattern analysis for sports and rehabilitation or gesture recognition. He is the author of more than 50 peer-reviewed scientific articles.

Jarosław Wąs received his PhD and DSc degrees in computer science in 2007 and 2015, respectively. In 2024 he was granted the professorial title by the President of Poland. He is the author and a co-author of more than 100 publications. In 2023, he was re-elected to the Computer Science Committee of the Polish Academy of Sciences. He is interested in the topics of modeling and simulation of complex systems, in particular, data-driven modeling and agent-based modeling. He also focuses on applications of advanced algorithms and artificial intelligence in engineering, including the IoT as well as ambient and computational intelligence.

Received: 9 March 2024
Revised: 28 June 2024
Accepted: 9 January 2025