Acta Polytechnica **59**(5):498–509, 2019
DOI:10.14311/AP.2019.59.0498

available online at https://ojs.cvut.cz/ojs/index.php/ap

AUTOMATIC EEG CLASSIFICATION USING DENSITY BASED ALGORITHMS DBSCAN AND DENCLUE

Marek Piorecký[a,b,*], Jan Štrobl[a,b], Vladimír Krajča[a,b]

[a] *Czech Technical University in Prague, Faculty of Biomedical Engineering, Department of Biomedical Technology, Nám. Sítná 3105, 27201 Kladno, Czech Republic*

[b] *National Institute of Mental Health, Topolova 748, 25067 Klecany, Czech Republic*

[*] corresponding author: marek.piorecky@fbmi.cvut.cz

Abstract. Electroencephalograph (EEG) is a commonly used method in neurological practice. Automatic classifiers (algorithms) highlight signal sections with interesting activity and assist an expert with record scoring. Algorithm K-means is one of the most commonly used methods for EEG inspection.

In this paper, we propose/apply a method based on density-oriented algorithms DBSCAN and DENCLUE. DBSCAN and DENCLUE separate the nested clusters against K-means. All three algorithms were validated on a testing dataset and after that adapted for a real EEG records classification. 24 dimensions EEG feature space were classified into 5 classes (physiological, epileptic, EOG, electrode, and EMG artefact). Modified DBSCAN and DENCLUE create more than two homogeneous classes of the epileptic EEG data. The results offer an opportunity for the EEG scoring in clinical practice. The big advantage of the proposed algorithms is the high homogeneity of the epileptic class.

Keywords: EEG, DBSCAN, DENCLUE, automatic classification, epilepsy.

1. Introduction

Electroencephalograph (EEG) represents the electric activity of the brain. Brain activity is most often recorded on the scalp. EEG is a commonly used method in clinical practice (for example, for detection of epilepsy, schizophrenia, etc.) [1, 2]. The measured signal depends on the type of the electrode used, their number, location, and on many other influences. The raw signal represents a changing voltage on each electrode, in time. We are not able to measure the half-cell potential, therefore, the resulting voltage is given by the difference of the potentials of the two electrodes. Standard pre-processing consists of montage, filtration, and segmentation. Segments are characterized by features, with each feature describing a mathematical characteristic (amplitude, frequency, etc.). Unsupervised methods do not require any user input, so they should be more objective and less time-consuming. The density-based clustering algorithm is used to find non-linear shapes structure based on the density. Density-based spatial clustering of applications with noise (DBSCAN) [3] is not used to classify an EEG record in common practice, although it is an algorithm that is used in many software applications in many modifications. DENsity-based CLUstEring (DENCLUE) [4] is a younger density-based method working on statistical principle, see section2, which is not generally used to classify an EEG record yet. [3, 4] Epilepsy is a serious neurological disorder that is manifested by seizures. In the EEG curve, the epileptic attack is observable with the spike - wave complex. An EEG record in a regular clinical examination contains thousands of segments. Automatic classification methods are often binary aimed at detecting only epileptogenic activity, for example [5, 6]. [7]

Many studies processed non-epileptic EEGs (see for example [8]). Study [9], processes the epileptic EEG records using unsupervised Kohonen's Self-Organizing Maps, although the authors divide the signal into only two classes (epileptic and non-epileptic segments). Only two classes are made, for example, in studies [10] and [11]. The study [12] used five unsupervised algorithms (among other things K-means and K-medoid) for the automatic classification of childhood epileptic's EEGs. The results of this study show that K-means is suitable for clinical practice, although the number of searched classes in this study was also two (seizer and non-seizer class). K-means is the commonly used method in practice [13]. In the study [14], Support Vector Machine and K-means with Multi-Scale K-means (MSK-means) were compared. Two classes were detected (epileptic and non-epileptic) with the best results for MSK-means. Three classes were searched in the study [15] using four algorithms (one algorithm was unsupervised K-means). However, classes were only healthy, ictal, and interictal parts of the signal, where EEG graphoelements (for example EMG artifect) were not detected. Unsupervised K-means algorithm and supervised K-NN algorithm were compared for classifying the EEG graphoelement in the study [16]. Here the K-NN algorithm showed better results than K-means.

Our aim is to classify all artefacts (parts of the EEG signal, which do not have a source in the brain), physiological and pathological segments of the EEG record. We create a plugin compatible with the WaveFinder

software. This software is used to describe and display data at the National Institute of Mental Health. The plugin should help the physician to more accurately estimate the interesting parts of long-term signals and possibly serve as a tool for the creation of ethalons.

2. Methods

2.1. Data

Data simulation was the first step of the algorithms testing. Our simulated data are numerical values that characterize segments of signal in the feature space. We created 2D training data in MATLAB R2015a. This data consists of nested and separated clusters. Figure 1 shows four examples of our simulated data. The labels are made optically in this data - sets are visually separable. We assumed that the EEG space contains nested clusters [17], therefore, we tested the ability of algorithms to separate such spatial clusters. Training sets demonstrate the disadvantages of k-means classification and their compensation using DBSCAN and DENCLUE methods.

Real EEG records were tested by the selected algorithms in the second part of our study. The data were obtained from patients from Bulovka Hospital in Prague. They were obtained on the basis of the project proposal, which was approved by the ethics committee of Bulovka Hospital on the day 28. 6. 2011. These are clinical examinations ranging from 15 to 30 minutes (the dataset was not targeted for this study). Test data were measured on patients who were diagnosed with suspected epilepsy disease (epileptic attack did not have to be present during the recording). 12 whole datasets were tested. The tested patients were men and women aged between 26 and 60 years. The localization of epilepsy and its characteristics differed between patients, we expected automatic detection based on the occurrence of a different kinds of spike-wave complex [18]. The data were analysed anonymously without any assumptions about the nature of the failures.

2.2. Preprocessing

The EEG signals were recorded in a 10-20 system with 19 investigated channels with a uni-polar connection (using an average reference montage of all channels) on the Brainquick system. The signal was filtered by the conventional analogue filter of 0 - 70 Hz. The data were sampled at 128 Hz and converted using a 12 bit converter. [19]

We used the program Wave-Finder (WF) [20] to segment a signal, compute the features for individual segments, and for the visualization of results. The Wave-Finder program uses adaptive segmentation (for more information about the method, see this study [21]). The adaptive segmentation creates segments (parts of the EEG signal) with different segment lengths. Every segment should contain only part of the EEG signal with the same characteristic (for example epileptic activity). If one segment contains parts of the EEG signal with different characteristic, it is marked as a wrongly segmented part of the signal. The settings of the adaptive segmentation used in this study can be seen in the table 1. The segments were made up of whole records and each record entered the classification separately. Segments were evaluated by an expert.

Parameter	Setting
Window Length	128 Samples
G Window Length	15 Samples
STEP	1 Sample
Optim	1 [-]
MINLENGTH	64 Samples
Number of Scan pts	15 Samples
Treshold	81 [-]
Max Segm Length	1024 Samples

Table 1. Setting segmentation parameters: The parameter *Window length* specifies the length of the associated window used in the adaptive segmentation. *G Window length* is the length of the window in which the exact position of the maximum (minimum is 3 points) is searched for. The parameter *Optim* turns on and off is the optimization of the segment boundary. It is at the lowest point nearby. *Number of Scan pts* is the number of points we look at on each side to obtain the minimum. *Threshold* specifies the threshold for segmentation boundary detection. *MINLENGTH* is the smallest possible length of the segment.

Features, which we are using, are described in the book [22]. The calculation of these features is implemented in WF and they are used in clinical practice. There are 24 features which are normalized in the interval $\langle 0;1\rangle$. See table 2, which shows the features that are specified below. Features create a multidimensional space that is counted for each EEG segment.

$APOS$ and $ANEG$ are the extremes of an amplitude for a specific segment. They give a value of real voltage after subtracting the DC component (A_{DC}), shown in equation 1 [22]:

$$A_{DC} = \frac{\sum_{i=1}^{L} y_i}{L}, \tag{1}$$

where L is the length of the segment and y_i is an i-th amplitude sample in the segment.

$MAX1D$ (equation no. 2) and $MD1$ (equation no. 3) determine maximum and average slope of the signal curve [22]:

$$MAX1D = max(y_{i+1} - y_i), \tag{2}$$

$$MD1 = \frac{\sum_{i=1}^{n} (y_{i+1} - y_i)}{n}, \tag{3}$$

where y_i is an i-th amplitude sample in the segment and n is a number of segments.

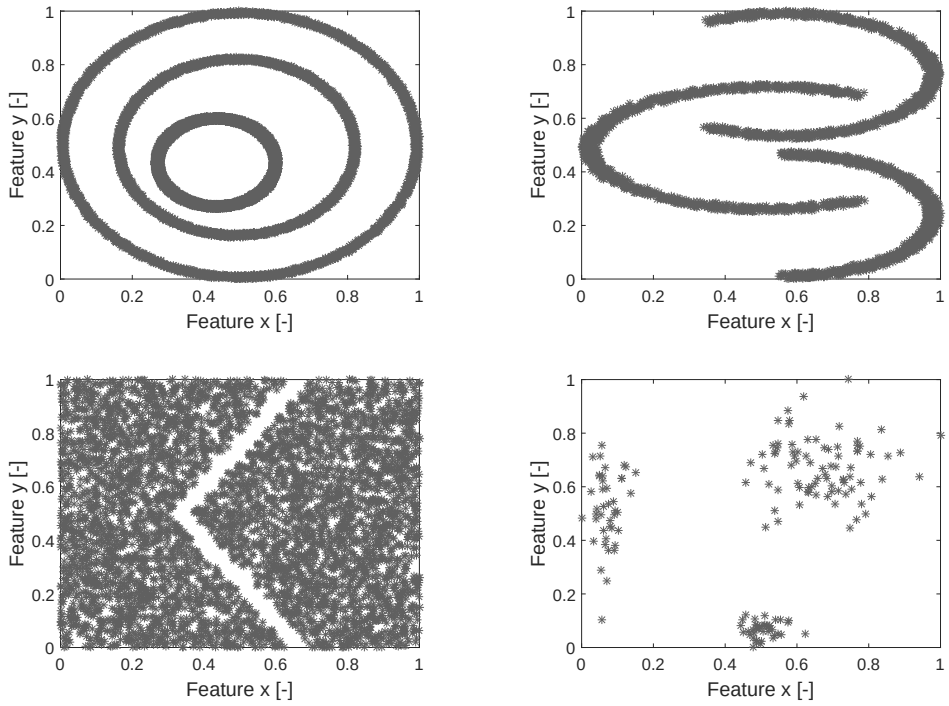

Figure 1. Examples of 2D simulated data in the feature space used in this study to test an algorithms.

Features	Description
SIGM	signal variability
APOS	maximal positive value
ANEG	maximal negative value
DELT1	part of the delta (0.5 Hz − 1.5 Hz)
DELT2	part of the delta (2.0 Hz − 3.5 Hz)
THET1	part of the theta (4.0 Hz − 5.5 Hz)
THET2	part of the theta (6.0 Hz − 7.5 Hz)
ALPH1	part of the alpha (8.0 Hz − 10.0 Hz)
ALPH2	part of the alpha (10.5 Hz − 13.0 Hz)
SIGMA	part of the sigma (18.0 Hz − 29.0 Hz)
BETA	part of the beta (13.5 Hz − 29.0 Hz)
MAX1D	maximum of the first derivation
MAX2D	maximum of the second derivation
mf	medium frequency
MD1	medium of the first derivation
MD2	medium of the second derivation
mob	Hjorths parameter mobility
comp	Hjorths parameter complexity
act	Hjorths parameter activity
LOfC	length of the curve
NLinE	a nonlinear energy
ZC	number of the passes by zero
Peaks	the maximum peak frequency in the spectrum
InfP	inflex point

Table 2. Features used for classification of EEG segments in this study.

$MAX2D$ (equation no. 4) and $MD2$ (equation no. 5) determine the curvature of the curve [22]:

$$MAX2D = max(x_{i+4} - 2x_{i+2} + x_i), \tag{4}$$

$$MD2 = \frac{\sum_{i=1}^{n} x_{i+4} - 2x_{i+2} + x_i}{n}, \tag{5}$$

where y_i is an i-th amplitude sample in the segment and n is a number of segments.

Hjorth parameters are indicators of statistical properties used to process signals from the time domain. We use three Hjorth parameters, *Activity*, *Mobility*, and *Complexity*. *Activity* represents a signal strength, scatter of a time function [23]:

$$Activity = var(y(t)), \tag{6}$$

where $(y(t))$ represents the signal.

Mobility represents the mean frequency - the share of the standard deviation of the power spectrum [23]:

$$Mobility = \sqrt{\frac{var\left(y(t) \cdot \frac{dy}{dt}\right)}{var\left(y(t)\right)}}, \tag{7}$$

where $var(y(t))$ is Hjorth parametr of *Activity*, $\frac{dy}{dt}$ is the derivation of the amplitude of a segment in time and $y(t)$ is the size of the amplitude.

Complexity represents a change in frequency compared to a pure sine wave [23]:

$$Complexity = \frac{Mobility\left(y(t) \cdot \frac{dy}{dt}\right)}{Mobility\left(y(t)\right)}, \tag{8}$$

where $Mobility$ is Hjorth parametr, $\frac{dy(t)}{dt}$ is the derivation of amplitudes of a segment in time and $y(t)$ is the amplitude.

$LOfC$ is length of the curve if we unpack it

$$L = \sum_{i=1}^{N_s} abs[y(i) - y(i+1)], \tag{9}$$

where N_s is a number of samples in the segment and $y(i)$ is i-th amplitude sample in the segment.

Number of passes by zero indicates how many times the curve passed from positive to negative and vice versa. Peaks indicates the number of peaks in the specific segment. $NLinE$ characterizes the signal in terms of energy. It indicates the average power in the band where the 80 % of the total energy of the spectrum is concentrated [24]:

$$NLinE(i) = y^2(i) - y(i-1)y(i+1), \tag{10}$$

where $y(i)$ is the amplitude in the i-th sample in a segment.

All of the features were normalized to create a single feature space for the classification. Normalization was realized by the minimum and maximum by the following equation:

$$Y(i) = \frac{X(i) - min(X)}{max(X) - min(X)}, \tag{11}$$

where min and max are the minimum and maximum values in X dataset.

2.3. K-means

We used the K-means algorithm like a commonly used (on the EEG data classification in clinical practice) unsupervised method for a comparison with testing algorithms. The unsupervised algorithm was chosen because DBSCAN and DENCLUE are also unsupervised. The K-means algorithm separate segments to the classes using distance computing.

You can see, for example, [25] for more details about the K-means principle. We used the K-means MATLAB R2015a function for simulated data in this study. The K-means from WF program was used for the automatic classification of a real EEG record in this study. The program WF exploits K-means in clinical practice and we can use them to compare K-means with density based algorithms.

2.4. DBSCAN

Density based algorithms take advantage of different density distributions of classified objects in space to separate individual datasets. Objects are, in our case, the segments of EEG signals. DBSCAN classifies objects based on density, so a range of objects with a similar density distribution is classified in the same class. Density means the number of points in the unit area of the feature space. In order to avoid classifying object regions from each other in a very distant space, the DBSCAN defines a cluster as a high dot density region that is separated by a low density location. The input parameters of the algorithm are the radius (Eps) and the number of objects in it ($nPts$). Compared to the K-means algorithm, the number of clusters is the default rather than the input parameter. It also depends on the initiation site, in a similar way as we perform a count with the random centre of clusters in the case of K-means. DBSCAN searches in the specified radius of the objects that fall into it. The number of object in its radius makes it possible to classify an object into three groups: a noisy, a marginal, a centre object. [3, 26, 27]

The algorithm starts with any data object (initialization object). We find all the objects that fall within its radius. On the basis of the number of them, the object is the initial object classified into one of the three classes: centre, marginal, or placed in the class of distant - noise objects. Gradually, for each of these objects (in the initial objects radius), we look for the number of objects in its radius, and by its number we classify an object. If it is a centre object (it has a sufficient number of objects in its radius), we assign this object and all the objects that fall within its radius to the centre of the object class. We repeat the previous steps for all objects that belonged to the radius of the starting object. If we exhaust all of the objects from the vicinity of the initiation object, we automatically go to the next non-classified object in the algorithm. The entire process continues until every single object is assigned to one of the three classes.

The formula for the automatic calculation of the Eps value can be found in Article [28]:

$$Eps = \left(\frac{\left(\Pi_{i=1}^{max(x)-min(x)} i\right) \cdot minPts \cdot \gamma \cdot (0{,}5 \cdot n + 1)}{m \cdot \sqrt{\pi^n}} \right)^{\frac{1}{n}}, \tag{12}$$

where x are the data in the form of the matrix (m, n), where m represents the number of segments and n the number of flags, $nPts$ is the number of objects in radius and γ is the interpolation coefficient.

According to Ali's Touhk's [29] design, we obtain the number of objects in the radius by sequentially testing $nPts$ values (from 1 to N - number of all objects) and compiling the graphs from these values. Each graph describes a set of test data, for example, see 2.

The average value of the pooled objects, which are in one cluster, is on the x axis is . The y axis is the $nPts$ value for which we get the number of clusters. We get a $nPts$ span from each graph, which provides the required number of classes. We choose the ideal value across all charts. Based on the results from all tested graphs, we chose $nPts = 15$.

The above calculation of Eps is adequate for 2D data. It does not give adequate results on multidimensional data. We assume an uneven layout of data in the space for DBSCAN modification. Therefore,

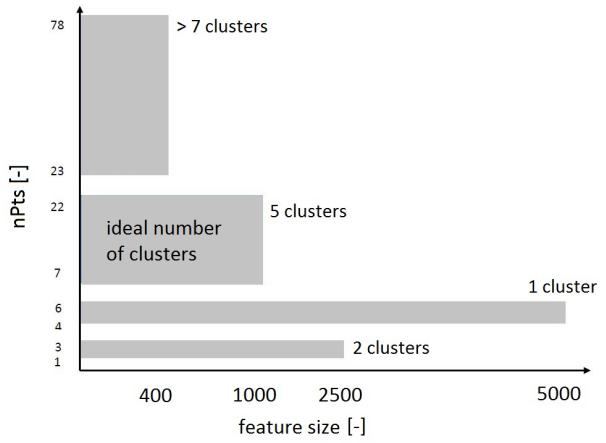

Figure 2. Optimal number of nPts.

we used the Dynamic Method for Discovering Density Varied Clusters (DMDBSCAN) principle and calculate the radius from the curve of the nearest neighbours. We count the distance of all objects to each object in the file. These values are ranked ascending. For each object, we select its first three neighbours and make an average of them [30]. The study showed that from the 4th neighbour, the results are no different [31]. These values are ranked in an ascendant order according to size and we create a curve. In the elbow of this curve, the values are appropriate for the radius of the given data set. The evaluation was made visually here. [26, 32]

For the tested EEG records, the curve never contained more than one knee, as well as in a study [32]. Inflection objects on the curve can only be obtained at $nPts > 50$ (in our 24-D space). If we reduce the number of dimensions, we get inflection objects at $nPts = 30$.

Another solution to this situation is offered by other modifications: GRIDBSCAN. GRIDBSCAN is a merger of several algorithms. The basic idea is that it is possible to divide space into cells (see figure 3) that evenly distributes space [33, 34]. In the first step, the cells are divided into over-limit (objects further classified) and under-limit (noise). If an oversized cell is adjacent to an already identified cell, it is assigned to the same class. If a gradual shift across the cell gets to one that is not adjacent to any classified, a new class is created [34]. In this case, there is a problem in dividing the multidimensional EEG space, because, with a higher number of cells, the filling becomes smaller and the classification quality decreases.

In the study [35], they used a grid that is composed of cells with small overlaps. The DBSCAN algorithm runs in each cell separately. In the overlapping area, objects are classified multiple times (depending on the cell). The interconnection of clusters of neighbour cells occurs just above these objects. The clusters that have been assigned to their cells are combined in one.

```
Data: border, featureMatrix
Result: Classification of segments
for 1:NumberCells do
    border = border of cell
    objets = objects belonging to the cell
    if objects =! 0 then
        [classpoint,class,type] = dbscan(segments)
        NumberCells = recalculated number of cell
        matrix = [data, NumberCells, class, type]
    end
end
for 1:NumberRecurringObjects do
    if CenterObject then
        NumberCluster = numbers of occurrence
         classes
        FN = min(NumberCluster)
        NumberCells = recalculated number of cell
        for 1:NumberCluster do
            matrix2 = all objects will be
             overwritten to FN
        end
    end
end
for 1:Number of rows of matrix2 do
    repeatedly = repeating objects
    result = stores only one object (with the
     lowest class number)
end
```

Algorithm 1: General algorithm of modified DBSCAN which we used in this study.

2.5. DENCLUE

DENCLUE was the second algorithm that we tested. DENCLUE is also a density based algorithm. The algorithm is created for large datasets in a multidimensional space [36]. The number of clusters is estimated automatically, like with DBSCAN. The initial computing segment is not random, compared to the DBSCAN and K-means algorithms. The DENCLUE algorithm used two initial coefficients, the smoothness coefficient h and the noise coefficient ξ. The main difference between DENCLUE and DBSCAN is that DENCLUE is based on a statistical bases. Specifically, DENCLUE is based on Kernel Density Estimation (KDE). [4, 37]

The basic idea of KDE is that we can describe the influence of each object in its neighbourhoods by kernel function. We can calculate the total density function in an object x by summing this mathematical function, as you can see in the next equation [4]:

$$f^D(x) = \frac{1}{Nh^d} \sum_{i=1}^{N} F_K\left(\frac{1}{h}(x - x_i)\right), \qquad (13)$$

where $f^D(x)$ is the total density function in an object x, h is the smoothness coefficient, N is the number of objects, d is the size dimensions of feature space and F_K is the specific Kernel function.

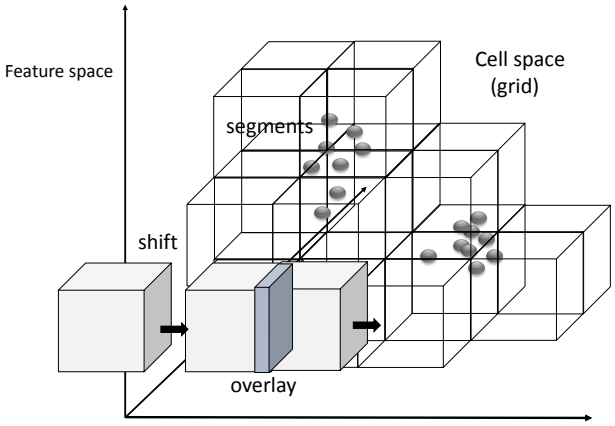

FIGURE 3. Cell space approximation in GRIDBSCAN classification. The dashed lines indicate the shift in the space with the overlap in which the clusters in the individual cells are interconnected.

The smoothness coefficient controls the influence of object neighbourhoods on the output of the total density function. The higher smoothness coefficient leads to a lesser influence of the objects' distances and higher smoothness of the total density function. This means that the higher smoothness coefficient will reduce the number of result clusters. We can use several kernel functions F_K. [4, 37]

We used the triangle kernel in our article, see equation 14 [38]:

$$F_K(x) = 1 - |x| \qquad x \,\epsilon\, (-1, 1)\,, \tag{14}$$

where $F_K(x)$ is Kernel function in an object x.

The disadvantage of KDE is in its high computational complexity. The DENCLUE algorithm reduces the computational complexity by using the Average Shifted Histogram (ASH). Because we compute only with occupied cells in ASH. The histogram is very sensitive to origin and that is a principle of ASH. We have several histograms with different origins (shifted histograms). When a number of shifting histograms are approaching infinity, the result of ASH is similarly to the result of a KDE [39]. For more information, see, for example, [40, 41]. The occupancy of histogram cells decreases with a rising dimension of feature space. This problem also improves ASH. We used $2+d$ shifted histograms in DENCLUE, where d is the number of dimensions in feature space. [4, 40]

The second part of the DENCLUE algorithm is a distribution of objects in clusters. The DENCLUE algorithm is looking for the local maximum of the density function, which was created using ASH. Multicentre definition of the cluster can be used in DENCLUE to distinguish nested clusters. We will go on to describe this definition of the cluster. The DENCLUE algorithm finds an object (cell containing objects) on the position of the local maximum of the density function and determine, if the local maximum is higher than the noise coefficient. If the local maximum is higher than the noise coefficient, the object in the local maximum forms the centre of the new cluster. If the local maximum is lower than the noise coefficient, the object in the local maximum is included in the noise cluster. Every object attracted to this maximum includes to the same cluster as their attractor. The local maximum includes the same cluster, if there is a path higher than the noise coefficient between them. The noise coefficient decreases the computing complexity of the algorithm, allowing creation of the spatially entwined clusters and the located noise objects. [4]

The DENCLUE algorithm was programmed in MATLAB R2015a and had to be modified for the

Data: Objects in feature space, h, ξ
Result: Objects distributed to clusters
loading of objects
for *i=1:NumberOfObjects + 2* **do**
  creation of histograms cells (using the h)
  execution of i-th shifts
  distribution of objects to cell of histograms
end
creation of ASH from histograms
finds local maxima of ASH
for *j=1:NumberOfLocalMaxima* **do**
  if *amplitude of j-th maximum* $> \xi$ **then**
    center of new cluster is j-th maximum
    objects attracted to i-th max. = same cluster
  else
    assignment of j-th max. to the noise cluster
    objects attracted to i-th max. = noise cluster
  end
end
if ϵ *path between 2 local maxima* $> \xi$ **then**
  local maxima are from same cluster
end

Algorithm 2: General of DENCLUE algorithm used in this study with the smoothness coefficient h and the noise coefficient ξ.

EEG record. The problem can be that there are many more segments of physiological activity compared with other segments in the EEG record. Therefore, the DENCLUE algorithm had two parts in this modification. Segments of the physiological activity are separated from other segments in the first part. Separation is done by the noise coefficient where every others segments should be included in a noise cluster. The smoothness coefficient had a value of 0.0083 and the noise coefficient had a value of 750 in the first part of the modified algorithm. Segments from the noise cluster are divided in the second part of the modified algorithm. The smoothness coefficient had a value of 0.0625 and the noise coefficient had a value of 35 during the second run of the DENCLUE algorithm.

2.6. Statistical Analysis

We divided the segments of the tested EEG records into five classes, concretely: physiological activity (PHY), EMG artefacts (EMG), epileptic activity (EPI), slow eye artefacts (SLOW), and artefacts from poor electrode contact - electrode pop (POP).

Figure 4. We include waves with sinus characteristics among the physiological activity.

Figure 5. Segments from eye movement and blinking are included in the class slow eye artefacts.

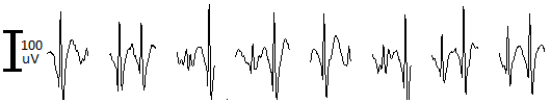

Figure 6. Epileptic activity with high amplitude of the apex.

Figure 7. EMG artefacts is in the classroom, which includes shaded segments with a line noise, when the noise completely distorts the original signal.

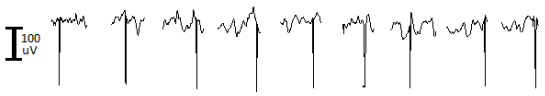

Figure 8. The wrong contact of the electrode is manifested by a narrow positive point with a high amplitude. These artefacts usually occur only in one channel.

We tested the efficiency of algorithms using the ROC analysis, which is suitable for binary data evaluation and was used for an evaluation of EEG [42]. Data is divided into two groups: correct assignment to a given cluster, and bad segment classification. We created five binary confusion matrices for five different classes evaluated in this study. The confusion matrix contains information about the real classification into the class (expertly evaluated) and predicted distribution (classification by algorithms). Segments are labelled on the basis of this information. True positive (TP) are correctly classified segments, False positive (FP) are mismatched segments that do not belong to the cluster. False negative (FN) are segments that belong to the cluster but were mistakenly assigned to another cluster, True negative (TN) are the segments correctly assigned to a different class. Then, we calculated $Specificity$, which shows the likelihood of the segments belonging to the cluster will not be included into another (see equation 15). *Sensitivity* determines the probability of a successful detection, which means finding all the segments belonging to the cluster (see equation 16). *Positive predictive value* (PPV) is the most telling parameter in our case, we can compare it to the homogeneity of the class (see equation 17) [43]:

$$Specificity = \frac{TN}{TN + FP}, \tag{15}$$

where TN is a True negative and FP is a False positive value.

$$Sensitivity = \frac{TP}{TP + FN}, \tag{16}$$

where TP is a True positive and FN is a False negative value.

$$PPV = \frac{TP}{TP + FP}, \tag{17}$$

where PPV is a *Positive predictive value*, TP is a True positive and FP is a False positive value.

3. Results

3.1. Test Data

The tests data were created to verify the accuracy of the proposed algorithms. We verified the good proposítion of the algorithms. Therefore, the test data represent the basic features that these algorithms

Algor.	Sensitivity [-]				
	EPI	EMG	SLOW	PHY	POP
K-means	0.851	0.803	0.725	0.427	0.944
DC	0.827	0.639	-	0.986	0.931
DB	0.896	-	0.094	0.997	0.909

TABLE 3. The sensitivity parameter of the classes epileptic activity (EPI), EMG artefacts (EMG), slow eye artefacts (SLOW), physiological activity (PHY) and artefacts from poor electrode contact (POP) identified in the EEG signal for the tested algorithms DENCLUE (DC), DBSCAN (DB), and K-means.

Algor.	Specificity [-]				
	EPI	EMG	SLOW	PHY	POP
K-means	0.956	0.966	0.782	0.984	0.999
DC	0.983	0.999	-	0.826	1.000
DB	0.997	-	0.999	0.804	0.999

TABLE 4. The specificity parameter of the classes epileptic activity (EPI), EMG artefacts (EMG), slow eye artefacts (SLOW), physiological activity (PHY) and artefacts from poor electrode contact (POP) identified in the EEG signal for the tested algorithms DENCLUE (DC), DBSCAN (DB), and K-means.

Algor.	PPV [-]				
	EPI	EMG	SLOW	PHY	POP
K-means	0.641	0.178	0.010	0.996	0.932
DC	0.813	0.941	-	0.979	0.998
DB	0.969	-	0.333	0.970	0.942

TABLE 5. The PPV parameter of the classes epileptic activity (EPI), EMG artefacts (EMG), slow eye artefacts (SLOW), physiological activity (PHY) and artefacts from poor electrode contact (POP) identified in the EEG signal for the tested algorithms DENCLUE (DC), DBSCAN (DB), and K-means.

(in correct proposition) should be/should not be able to distinguish. Two types of test data were created from nested clusters, one type of test data contained outliers and the last type of test data was formed by two good separated clusters.

All three algorithms (DBSCAN, DENCLUE, and K-means) were verified with test data. Every algorithm assigns different class numbers to the same data. The sorting of the classes by maximum amplitude was then used for real EEG data. In this section, we tested the correct design of algorithms and sorting of the classes is a simple separate step same for all tested algorithms. Therefore, the sorting of the classes is not used for the testing data. We are only viewing a correct separation of the data. Density based algorithms had the same results for the test data. For this reason, examples of test data results for both density based algorithms are shown in figure 9. The results of algorithm K-means for the test data are displayed in figure 10.

3.2. Real EEG Data

The expert classified 49.554 segments from all EEG records. Segments included physiological activity (PHY), epileptic activity (EPI), EMG artefacts (EMG), artefacts from poor electrode contact (POP), slow eye artefacts (SLOW) and wrong segmented parts of the signal. Adaptive segmentation used for signal pre-processing (see section 2.2) sometimes rank part of the signal, where the transition between two classes occurs, into one segment. These wrong segmented parts of the signal were removed from the statistical analysis, which has prevented it from affecting the results. Individual tables show successive results of sensitivity (see table 3), specificity (see table 4), and PPV (see table 5) for all three tested algorithms.

4. Discussion

In this study, we tested the utilization of the density based algorithms DBSCAN and DENCLUE on the EEG signal classification. The aim of this classification was to assist to an expert with a scoring of the EEG signal. Algorithms should be able to identify individual elements occurring in the EEG record. This should make it easier for the expert to find important parts of the EEG signal and then evaluate it. We tested whether the algorithms can identify epileptic activity (specifically spike and wave complex) and physiological activity. We also tested the ability to classify artefacts, because these parts of the EEG assist in the overall view of the signal.

We compared testing algorithms with the K-means algorithm. The K-means algorithm is a relatively old unsupervised algorithm, although it is still being commonly used in recent studies into EEG classification (for example, studies [12] and [44]). Many other supervised algorithms are also used in clinical practice (for example K-NN, Neuronal network, or Support Vector Machine). Supervised algorithms have their advantages (better results in the case of the best learning process) and disadvantages (the need to create a training set). We tested unsupervised algorithms DBSCAN and DENCLUE and it is necessary to compare their results to another unsupervised algorithm. The results are, therefore, not distorted by the difference between unsupervised and supervised algorithms in principle.

First, we have verified the main advantages and disadvantages of the proposed algorithms and the K-means algorithm on the testing dataset. The tested data were created only for a verification of the basic properties of algorithms. If the algorithms are correctly designed, the expected behaviour of the algorithms is confirmed. According to the assumptions, all three algorithms presented good results for the testing data with two good separated clusters. Both density based algorithms DBSCAN and DENCLUE displayed good results for the testing data with nested

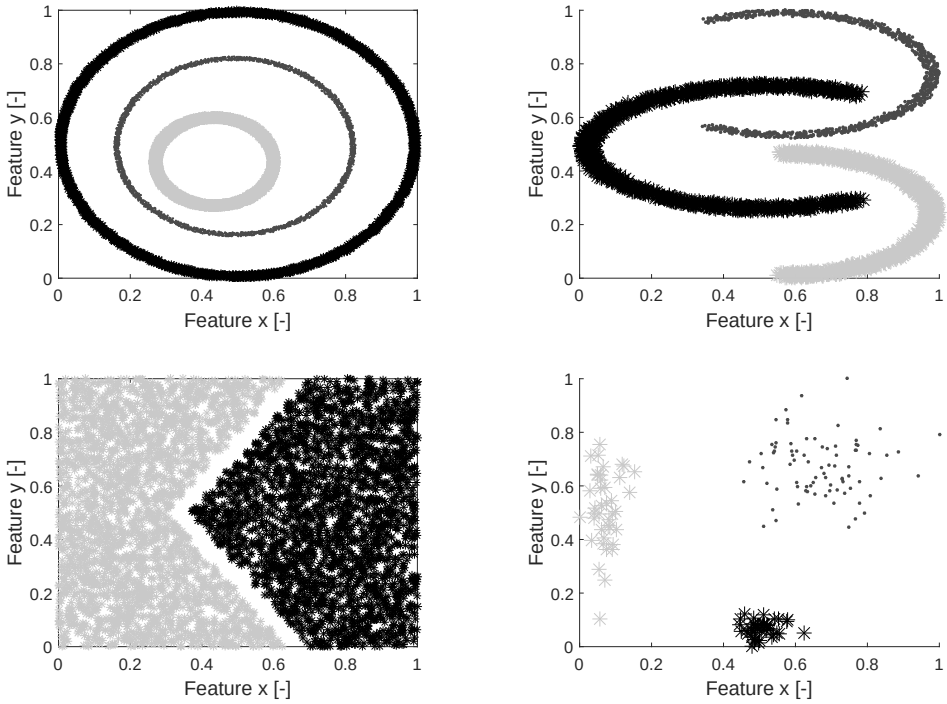

Figure 9. Example of the classification of simulated data by algorithms DBSCAN and DENCLUE. Different classes are represented by different shades of grey.

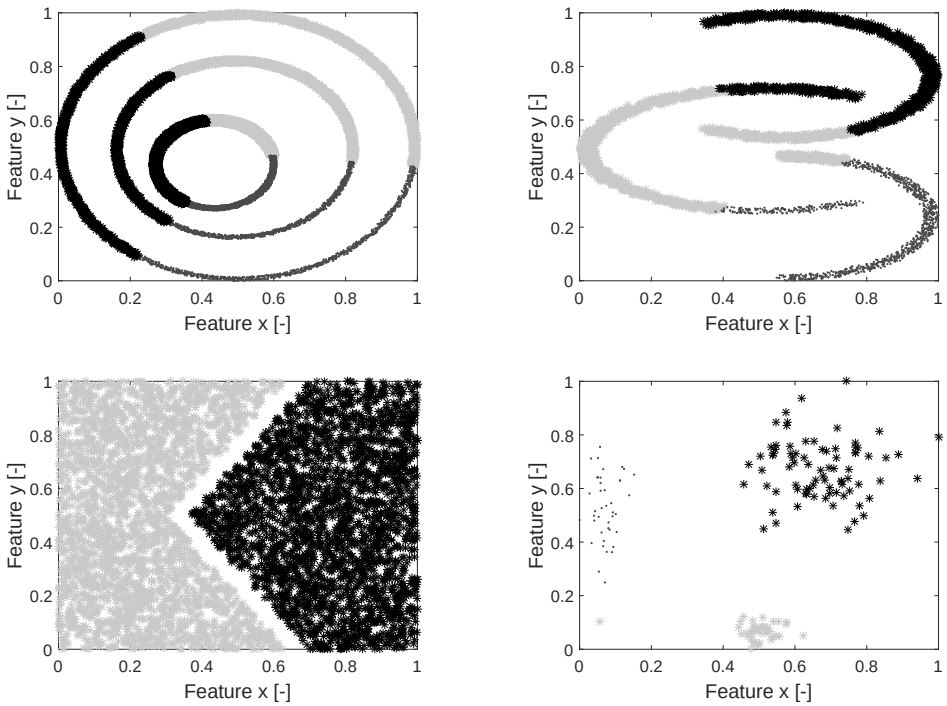

Figure 10. Example of classification of simulated data by algorithm K-means. Different classes are displayed by different shades of grey.

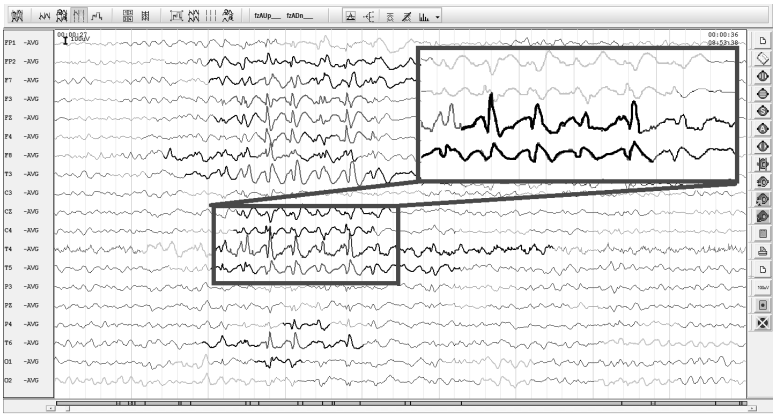

FIGURE 11. Work motivation: highlighting the EEG record sections to which the physician should pay attention.

clusters (see figure 9), which is their main advantage. Conversely, the K-means algorithm was unable to separate nested clusters and also had problems with the classification of the outliers (see figure 10). The results of the testing dataset are in line with the assumptions.

Algorithms were used on the real EEG data after their verifying on the testing data. We divided the EEG data into five clusters according to the clinical significance of the EEG segments. These classes represent the important parts of the epileptic EEG signal (physiological activity, epileptic activity, EMG artefacts, slow eye artefacts, and technical artefacts from a poor electrode contact). For all three tested algorithms, five classes were computed: sensitivity, specificity, and positive predictive value (PPV).

The class of the epileptic activity is very important for an expert in clinical practice. All algorithms have high specificity for this class (see table 4). This means that all tested algorithms correctly do not include segments that do not belong in the observed class of the epileptic activity. The class of epilepsy activity also had higher values for all algorithms for the sensitivity parameter (see table 3), so the tested algorithms found most of the epilepsy segments. The DBSCAN algorithm had the highest PPV for the class of an epilepsy activity. The DENCLUE algorithm had PPV 0.813, although the K-means algorithm had PPV value of the class of the epilepsy activity only 0.641 (see table 5). The K-means algorithm creates a class of the epilepsy activity, including segments from other classes.

The class of the physiological activity had a high PPV and specificity for all tested algorithms (see tables 5 and 4). The reason is a much larger number of physiological segments contained in the EEG signals than segments of other classes. The K-means algorithm has a predetermined number of clusters. We used seven clusters in our study. This number of clusters is taken from settings for the clinical practice of the program Wave-Finder and is selected to better detect the clinically significant epilepsy segments. The problem of the K-means algorithm is that segments of physiological activity are divided into more classes, this causes a low sensitivity of the K-means algorithm for segments of the physiological activity. The DBSCAN and DENCLUE algorithms had the sensitivity for these segments higher than 0.98 (see table 3).

The DENCLUE algorithm could not find segments of the slow eye artefacts. The K-means algorithm also had problems with these segments because its PPV is only 0.010. The DBSCAN algorithm formed classes of the slow eye artefacts, although it did not find most of the segments of this class (sensitivity was only 0.094). The DBSCAN algorithm did not find segments of the EMG artefacts. The DENCLUE algorithm had the highest PPV for segments of the EMG artefacts, so it formed a homogeneous class of these segments. However, the DENCLUE algorithm did not find segments of the EMG artefact (the sensitivity for the DENCLUE algorithm was 0.6394). All algorithms were able to identify segments of the artefacts from the poor electrode contact (see tables 3, 4 and 5).

5. Conclusion

We have investigated the efficiency of the density-based DENCLUE and DBSCAN algorithms for a classification of the EEG segments to clinically relevant classes. The K-means algorithm was used as a comparison algorithm used in clinical practice. Density-based algorithms displayed good results for clinically very important classes of the epilepsy activity and physiological activity. All algorithms had problems with the identification of the segments of the slow eye artefacts. The DBSCAN algorithm created the most homogeneous classes with the exception of the class of the EMG artefacts, which could not be identified. Conversely, the DENCLUE algorithm forms the homogeneous class of the EMG artefacts. The results suggest that the use of algorithms, especially for the creation of homogeneous classes, is promising, although there is a need for further testing of the algorithms.

Acknowledgements

This work was supported by the Grant Agency of the Czech Technical University in Prague with the topic: Feature

space analysis using linear and non-linear reduction of EEG space dimensions, grant no. SGS18/159/OHK4/2T/17; and by the Grant Agency of Czech Republic with topic: Temporal context in analysis of long-term non-stationary multidimensional signal, register no. 17-20480S. We thank to MUDr. Svojmil Petranek and Bulovka Hospital in Prague, Department of Neurology, Prague, Czech Republic.

References

[1] Z. Dvey-Aharon, N. Fogelson, A. Peled, et al. Schizophrenia detection and classification by advanced analysis of eeg recordings using a single electrode approach. *PLOS ONE* **10**(4), 2015-4-2. doi:10.1371/journal.pone.0123033.

[2] S. J. M. Smith. Eeg in the diagnosis, classification, and management of patients with epilepsy. *Journal of Neurology, Neurosurgery and Psychiatry* **76**:ii2–ii7, 2005-06-01. doi:10.1136/jnnp.2005.069245.

[3] K. Khan, S. U. Rehman, K. Aziz, et al. Dbscan: Past, present and future. In *Applications of Digital Information and Web Technologies (ICADIWT)*, vol. 5, pp. 232–238. IEEE, 2014. doi:10.1109/ICADIWT.2014.6814687.

[4] A. Hinneburg, D. A. Keim. A general approach to clustering in large databases with noise. *Knowledge and Information Systems* **5**(4):387–415, 2003. doi:10.1007/s10115-003-0086-9.

[5] R. Sharma, R. B. Pachori. Classification of epileptic seizures in eeg signals based on phase space representation of intrinsic mode functions. *Expert Systems with Applications* **42**(3):1106–1117, 2015. doi:10.1016/j.eswa.2014.08.030.

[6] U. R. Acharya, H. Fujita, V. K. Sudarshan, et al. Application of entropies for automated diagnosis of epilepsy using eeg signals. *Knowledge-Based Systems* **88**:85–96, 2015. doi:10.1016/j.knosys.2015.08.004.

[7] F. Lotte, M. Congedo, A. Lecuyer, et al. A review of classification algorithms for eeg based brain computer interfaces. *Journal of Neural Engineering* **4**(2):R1–R13, 2007-06-01. doi:10.1088/1741-2560/4/2/R01.

[8] I. Mporas, A. Efstathiou, V. Megalooikonomou. Sleep stages classification from electroencephalographic signals based on unsupervised feature space clustering. In *Brain Informatics and Health*, pp. 77–85. Springer International Publishing, Cham, 2015. doi:10.1007/978-3-319-23344-4_8.

[9] C. R. Azevedo, C. F. Boos, F. M. de Azevedo. Classification of epileptiform events in eeg signals using neural classifier based on som. In *2015 International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)*, pp. 1–5. IEEE, 2015. doi:10.1109/ICEEICT.2015.7307340.

[10] S. Belhadj, A. Attia, A. B. Adnane, et al. Whole brain epileptic seizure detection using unsupervised classification. In *2016 8th International Conference on Modelling, Identification and Control (ICMIC)*, pp. 977–982. IEEE, 2016. doi:10.1109/ICMIC.2016.7804256.

[11] J. M. del Rincon, M. J. Santofimia, X. del Toro, et al. Non-linear classifiers applied to eeg analysis for epilepsy seizure detection. *Expert Systems with Applications* **86**:99–112, 2017. doi:10.1016/j.eswa.2017.05.052.

[12] O. Smart, M. Chen. Semi-automated patient-specific scalp eeg seizure detection with unsupervised machine learning. In *2015 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pp. 1–7. IEEE, 2015. doi:10.1109/CIBCB.2015.7300286.

[13] H. Rajaguru, S. K. Prabhakar. *KNN Classifier and K-Means Clustering for Robust Classification of Epilepsy from EEG Signals. A Detailed Analysis.* Anchor Academic Publishing, 2017.

[14] G. Zhu, Y. Li, P. Wen, et al. Unsupervised classification of epileptic eeg signals with multi scale k-means algorithm. In *Brain and Health Informatics*, pp. 158–167. Springer International Publishing, Cham, 2013. doi:10.1007/978-3-319-02753-1_16.

[15] S. Ghosh-Dastidar, H. Adeli, N. Dadmehr. Mixed-band wavelet-chaos-neural network methodology for epilepsy and epileptic seizure detection. *IEEE Transactions on Biomedical Engineering* **54**(9):1545–1551, 2007. doi:10.1109/TBME.2007.891945.

[16] H. Schaabova, V. Krajca, V. Sedlmajerova, et al. Supervised learning used in automatic eeg graphoelements classification. In *2015 E-Health and Bioengineering Conference (EHB)*, pp. 1–4. IEEE, 2015. doi:10.1109/EHB.2015.7391470.

[17] M. Piorecky, E. Cerna, V. Piorecka, et al. Simulation, modification and dimension reduction of eeg feature space. In *World Congress on Medical Physics and Biomedical Engineering 2018*, pp. 425–429. Springer Singapore, Singapore, 2019. doi:10.1007/978-981-10-9038-7_80.

[18] E. Niedermeyer, F. H. L. da Silva. *Electroencephalography, basic principles, clinical applications, and related fields.* Urban & Schwarzenberg, Baltimore, 1982.

[19] S. R. Sinha, L. Sullivan, D. Sabau, et al. American clinical neurophysiology society guideline 1. *Journal of Clinical Neurophysiology* **33**(4):303–307, 2016. doi:10.1097/WNP.0000000000000308.

[20] V. Krajča, S. Petránek, T. Pietilä, H. Freay. "wavefinder": A new system for automatic processing of long term eeg recording. *Quantitative EEG analysis-clinical utility and new methods* pp. 103–106, 1993.

[21] D. Kala, V. Krajca, H. Schaabova, et al. Optimal parameters of adaptive segmentation for epileptic graphoelements recognition. *Radioengineering* **26**(1):323–329, 2017-04-14. doi:10.13164/re.2017.0323.

[22] V. Krajča, J. Mohylová. *Číslicové zpracování neurofyziologických signálů.* České vysoké učení technické v Praze, 2011.

[23] S. T.-B. Hamida, B. Ahmed, T. Penzel. A novel insomnia identification method based on hjorth parameters. In *2015 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, pp. 548–552. IEEE, 2015. doi:10.1109/ISSPIT.2015.7394397.

[24] H. Qu, J. Gotman. A patient-specific algorithm for the detection of seizure onset in long-term eeg monitoring. *IEEE Transactions on Biomedical Engineering* **44**(2):115–122, 1997. doi:10.1109/10.552241.

[25] M. R. Anderberg. *Cluster analysis for classification*. Academic press, inc. London, 1973.

[26] E. Schubert, J. Sander, M. Ester, et al. Dbscan revisited, revisited. *ACM Transactions on Database Systems* **42**(3):1–21, 2017-08-24. DOI:10.1145/3068335.

[27] V. S. Ware, H. N. Bharathi. Study of density based algorihms. *Journal of Computer Applocations* **32**(8):68–75, 1999. DOI:10.5120/12132-8235.

[28] A. Karami, R. Johansson, R. Choosing. Choosing dbscan parameters automatically using differential evolution. *International Journal of Computer Applications* **91**(7), 2014. DOI:10.5120/15890-5059.

[29] A. Thouka. Choosing parameters of dbsccan algorithm., 2012.

[30] M. T. H. Elbatta, W. M. Ashour. A dynamic method for discovering density varied clusters. *Inf Journal of Signal Processing, Image Processing and Pattern Recognition* **6**(1):123–134, 2013.

[31] M. Ester, H. P. Kriegel, J. Sander, X. Xu. A dentsity based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, vol. 96, pp. 226–231. 1996. DOI:10.5120/739-1038.

[32] N. Rahman, I. S. Sitanggang. Determination of optimal epsilon (eps) value on dbscan algorithm to clustering data on peatland hotspots in sumatra. In *IOP Conference Series: Earth an Environmental Science*, vol. 31. 2016. DOI:10.1088/1755-1315/31/1/012012.

[33] C. J. Pang. Research of grid-similarity-based clustering algorithm. In *WASE International Conference on Information Engineering (ICIE'09)*, vol. 2, pp. 33–36. 2009. DOI:10.109-ICIE.2019.202.

[34] C. F. Tsai, J. H. Zhang. Grid clustering algorithm with simple leaping search technique. In *Internatioanl Symposium on Computer, Consumer and Control (IS3C)*, pp. 938–941. 2012. DOI:10.1109/IS3C.2012.244.

[35] S. Mahran, K. Mahar. Using grid for accelerating density-based clustering. In *8th IEEE International Conference on Computer and Information Technology*, pp. 35–40. 2008. DOI:10.1109/CIT.2018.4594646.

[36] A. Hinneburg, D. A. Keim. An efficient approach to clustering in large multimedia databases with noise. In *Proceedings of the 4th International Conference on Knowledge Discovery and Datamining (KDD'98)*, pp. 58–65. 1998.

[37] A. Hinneburg, H.-H. Gabriel. Fast clustering based on kernel density estimation. In *Advances in Intelligent Data Analysis VII*, pp. 70–80. Springer Berlin Heidelberg, 2007. DOI:10.1007/978-3-540-74825-0_7.

[38] D. W. Scott, S. R. Sain. Multidimensional density estimation. In *Data Mining and Data Visualization*, pp. 229–261. Elsevier, 2005. DOI:10.1016/S0169-7161(04)24009-3.

[39] D. W. Scott. Averaged shifted histograms: Effective nonparametric density estimators in several dimensions. *The Annals of Statistics* **13**(3):1024–1040, 1985. DOI:10.1214/aos/1176349654.

[40] D. W. Scott. *Multivariate Density Estimation: theory, practice, and visualization*. John Wiley & Sons, 2015.

[41] D. W. Scott. Averaged shifted histogram. *Wiley Interdisciplinary Reviews: Computational Statistics* **2**(2):160–164, 2010. DOI:10.1002/wics.54.

[42] M. A. Sovierzoski, F. M. de Azevedo, I. F. M. Arqoud. Performance evaluation of an ann ff classifier of raw eeg data using roc analysis. In *International Conference on BioMedical Engineering and Informatics (BMEI)*, vol. 1, pp. 332–336. 2008. DOI:10.1109/BMEI.2008.220.

[43] C. O'Reilly, T. Nielsen. Revisiting the roc curve for diagnostic applications with an unbalanced class distribution. In *2013 8th International Workshop on Systems, Signal Processing and their Applications (WoSSPA)*, pp. 413–420. IEEE, 2013. DOI:10.1109/WoSSPA.2013.6602401.

[44] K. Rai, V. Bajaj, A. Kumar. Novel feature for identification of focal eeg signals with k-means and fuzzy c-means algorithms. In *2015 IEEE International Conference on Digital Signal Processing (DSP)*, pp. 412–416. IEEE, 2015. DOI:10.1109/ICDSP.2015.7251904.