



# Acoustic Feature Selection with Fuzzy Clustering, Self Organizing Maps and Psychiatric Assessments

Olga Kamińska<sup>(✉)</sup> , Katarzyna Kaczmarek-Majer<sup>(✉)</sup> ,  
and Olgierd Hryniewicz<sup>(✉)</sup>

Systems Research Institute, Polish Academy of Sciences,  
Newelska 6, 01-447 Warsaw, Poland  
{o.kaminska,k.kaczmarek,olgierd.hryniewicz}@ibspan.waw.pl  
<http://www.ibspan.waw.pl/>

**Abstract.** Acoustic features about phone calls are promising markers for prediction of bipolar disorder episodes. Smartphones enable collection of voice signal on a daily basis, and thus, the amount of data available for analysis is quickly growing. At the same time, even though the collected data are crisp, there is a lot of imprecision related to the extraction of acoustic features, as well as to the assessment of patients' mental state. In this paper, we address this problem and perform an advanced approach to feature selection. We start from the recursive feature elimination, then two alternative approaches to clustering (fuzzy clustering and self organizing maps) are performed. Finally, taking advantage of the partially assumed labels about the state of a patient derived from psychiatric assessments, we calculate the degree of agreement between clusters and labels aiming at selection of most adequate subset of acoustic parameters. The proposed method is preliminary validated on the real-life data gathered from smartphones of bipolar disorder patients.

**Keywords:** Self organizing maps · Fuzzy C-Means · Recursive feature selection · Cluster agreement · Bipolar disorder episode prediction

## 1 Introduction

Bipolar disorder (BD) is a chronic mental illness characterized with changing episodes from euthymia (state of health) through depression to mania (euphoric state) and the mixed states (depressive and manic symptoms present). BD affects more than 2% of the world's population [1]. The risk of a new episode can be reduced significantly by an early detection and an appropriate treatment. However, the frequency of visits with the psychiatrist is usually insufficient to provide early intervention, and patients by themselves are usually not aware of the need of treatment if a new episode starts. Therefore, in the recent years,

smartphone becomes an increasingly important tool in the early prediction of a starting episode and smartphone-based objective data become a valid markers in predicting BD episode recurrence [2].

Although, the acoustic data collection can be performed during the everyday life of a patient, labeled data are limited only to days around the psychiatric assessments. The mental state of a patient between psychiatric assessments is often unknown. In majority of the related work, see e., [2,3], the problem is stated as a supervised learning task. Recently, to alleviate the problems of uncertainty about patients state and limited data, Kamińska et al. [4] applied unsupervised learning technique (self-organizing maps) to find groups (clusters) in acoustic data for each patient without taking into account the psychiatric assessments. As a consequence, the whole dataset was used for learning rather than constraining it only to a few days before and after the visit to the psychiatrist. Then, the relation between the learned clusters and the labels from psychiatric assessment was investigated. Kamińska et al. [4] have noted that the degree of agreement between the results of unsupervised learning (clusters learned on acoustic data) and the results of the psychiatric assessments is related to the type of the BD phases recognized during the psychiatric evaluation.

The extraction of acoustic features is accompanied by several uncertainties. First of all, the device used by a patient and the quality of its microphone are unknown. Secondly, voices in the background have some influence on the quality of the collected parameters. Also, due to unexpected technical issues, some phone calls were simply not recorded without clear understanding of this situation. Other phone calls are not recorded because BDapp on patient's smartphone was off due to unknown reason (patients declared that they will use it). Finally, due to memory shortages, not all frames of a phone call could be processed and are simply omitted. Only some of them were selected. At the same time, the process of assigning labels under the psychiatric evaluations is also accompanied by several uncertainties. Its outcomes are subject to the condition during the visit. The mental state and mood of a patient could change quickly after it. Nonetheless, the BD phase of a patient assessed by a doctor during the interview is usually assigned as labels to the surrounding days assuming a specific ground-truth for the analyses. Often authors apply 7 days before the psychiatric assessment and 2 days after [3]. However, there might not be one common ground-truth that describes adequately all patients. Therefore, it is essential to explore the structure of the acoustic data and investigate what is the quality of learned clusters.

Due to all these uncertainties, in this research we incorporate fuzzy clustering for alternative subsets of acoustic features. This paper is a continuation of our previous works [4]. However, now we aggregate data to a single phone call, whereas in [4] the aggregation has been done to one day. Aggregation process rely on collecting all acoustic parameters [5] for each phonecall for each patient and then calculate the quartiles for received values. We perform an extensive comparative analysis aiming at selection of: (1) smaller subset of acoustic parameters that will require less computational efforts; (2) fuzzy clustering algorithm for the considered smartphone-based acoustic data to reflect the

related imprecision. This research is a step forward the superior goal that is an adequate prediction of BD episode recurrence using smartphone-based acoustic features.

The main novelty of this paper consists in application of Fuzzy C-means and Self organizing map algorithms for truncated datasets from the RFE algorithm. The unsupervised algorithm is selected over the supervised ones to alleviate the problem of limited labeled data and aiming at exploration of the whole data and investigating whether they can be grouped into clusters. The proposed approach is validated on the real-life dataset coming from the voice calls of patients suffering from bipolar disorder and the degree of agreement between learned clusters and psychiatric labels is evaluated.

That paper is organized as follows. In Sect. 2, methodology applied in this research is described, starting from the observational study on bipolar disorder to the brief description of the unsupervised approaches. Then, results of experiments are presented in Sect. 3. In the last Section, main conclusions are discussed.

## 2 Methodology

### 2.1 Observational Study and Acoustic Feature Extraction

Motivation for this research comes from analyzing real-world data collected in a recent observation study<sup>1</sup>. The study included patients diagnosed with bipolar disorder (F31 according to ICD-10 classification). In total, 33 patients were enrolled and used a dedicated smartphone application in everyday life for up to 15 months (starting in September 2017 and ending in December 2018). The study was conducted in the Department of Affective Disorders, Institute of Psychiatry and Neurology in Warsaw, Poland. Each patient was associated to a psychiatrist and control visits were scheduled. The evaluation of the mental state was performed by psychiatrists using both - the standardized measures of depressive and manic symptoms: Hamilton Depression Rating Scale (HAMD) and Young Mania Rating Scale (YMRS), as well as clinician's own assessment based on his experience with BD patients. The interviews were performed with various frequency depending on the need identified by the doctor or a patient.

Participants of the study received a dedicated mobile application, called *BDMon* able to collect acoustic features about phone calls. Patient's voice signal was divided into short 10–20 ms frames (withing a frame it is approximately stationary). With the use of an adopted version of a common library: openSMILE [5], the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) for voice research was extracted from each frame. The considered set contains 73 parameters connected with energy, spectral parameters and cepstral

---

<sup>1</sup> Data considered in this paper come from CHAD project – entitled “Smartphone-based diagnostics of phase changes in the course of bipolar disorder” (RPMA.01.02.00-14-5706/16-00) that was financed from EU funds (Regional Operational Program for Mazovia) in 2017–2018.

parameters (e.g., bandwidth energy, energy ration in different bands, relative volume) and 13 parameters connected with sound source (e.g., intonation contour). The meaning of that parameters is mostly connected with loudness in particular bands, voice energy, pitch etc.

The current state of the art lacks a clear indication which of the acoustic features are the best predictors of BD phase. In [4], the authors use the following 12 parameters: spectral slope in the ranges 0–500 Hz and 500–1500 Hz, energies in bands 0–650 Hz and 1000–4000 Hz, alpha ratio, ratio of the energy in band 50–100 Hz to the energy in band 1000–5000 Hz, spectral roll-off point the frequency below which 25% of the spectrum energy is concentrated, harmonicity of the spectrum, maximal position of the FFT spectrum, Hammarberg index, entropy of the spectrum, modulated loudness (RASTA), and zero-crossing rate. Different subset of parameters was selected with filter feature selection by [3], who use the following parameters: kurtosis energy, mean second and mean third MFCC, mean fourth delta MFCC, max ZCR and mean HNR, std and range F0. Within this research, recursive feature elimination are used as first step to select a subset of predictors.

## 2.2 Recursive Feature Elimination (RFE)

To obtain significant voice parameters we apply one of the automatic feature selection methods called Recursive Feature Elimination (RFE) [6]. The idea of the RFE technique is to build a model with all variables and after that the algorithm removes one by one the weakest variables until there will be achieved established number of variables. To find the optimal number of features cross-validation is used with RFE algorithm to obtain the best scoring collection of features.

## 2.3 Self Organizing Maps (SOM)

Results of the fuzzy c-mean clustering are compared to the clustering using self-organizing map algorithm known also as Kohonen network [7]. For each patient, we performed clustering of the call aggregates (quartiles) of theirs voice features using the *kohonen* package from the CRAN repository for R language, [8]. Important feature of the self-organizing maps is the preservation of neighborhood between the clusters in the two-dimensional space. Similarly, as in [4], we apply rectangular map topology of dimensions  $3 \times 1$ , which is intended to identify the two most distant affective states mania and depression and euthymia in between. The fourth affective state, mixed, is in its nature a combination of both, manic and depressive symptoms, and as confirmed with our preliminary experiments, it is more adequately represented as a mixture of depression and mania within the  $3 \times 1$  Kohonen network, than as an additional dimension of a map, e.g.,  $4 \times 1$ .

**2.4 Fuzzy C-Means Algorithm (FCM)**

Fuzzy C-means [9] another cluster algorithm is applied for the acoustic data and compared with the SOM algorithm. The specificity of this algorithm is that one value could be clustered as a cluster A with some membership, and the same value could be clustered as a cluster B with another membership. It might seem as thoughest examples to identify mixed phase. In mixed phase patients could be for some time in depression and for some time mania. For the comparative purposes, the number of clusters was predefined and assumed as 3. Package e1071 from CRAN repository has been used.

**2.5 Evaluation Metric**

To compare the degree of agreement between the learned clusters in acoustic data and the labels from the psychiatric assessment, we apply the clustering agreement metric as applied in [4] according to the following formulas, and similar to the Rand Index.

$$f_{t_v,i} = \frac{\sum_{t=t_v-7}^{t_v+2} I\{c_t = i\}}{\sum_{t=t_v-7}^{t_v+2} I\{c_t \text{ is not missing}\}} \tag{1}$$

- $f_{t_v,i}$  frequencies of each cluster
- $I$  indicator function taking value:
  - 1 - if the predicates are true
  - 0 - otherwise
- $c_t$  denote a cluster which was assigned to on day to a particular patient

where  $I$  is the indicator function taking value one if the predicate in curly brackets is true and zero otherwise.

We extract the data around every pair of visits and assign them to clusters trained on the remaining data. Therefore, we aim at comparing two groupings of the same data, one done by the clustering algorithm and the other by psychiatric assessments extrapolated to 7 days before and 2 after the visit.

Then, we compare the distributions  $f_{t_A}$  and  $f_{t_B}$  for two visits A and B with the normalized absolute difference

$$a_{t_A,t_B} = 1 - \frac{1}{2} \sum_{i=1}^3 |f_{t_A,i} - f_{t_B,i}| \tag{2}$$

$a_{t_A,t_B}$  normalized absolute difference

**2.6 Diagram Representation**

In order to easily visualize the received results, they were presented using a heatmaps diagrams for each pair of patient visits. On the X axis there are labels (received by psychiatrifer for the first visit and on the Y axis there are labels for

the second visit. The value presented in the graph is the degree of agreement (2) described above and calculated for each pair of visits to available patients and then averaged for the same pair of visits in reverse order. Values close to 0 mean that the clusters received on two different visits are different from each other, while values close to 1 mean that the clusters obtained on two different visits are similar to each other. The expected values for this chart are as follows. We strive for the highest possible values on the diagonal of the matrix - which means that the received clusters for visits with the same label are similar to each other. However, we strive to keep the remaining values as close to 0 as possible, which means that during two visits with different labels, the received clusters are different.

### 3 Experimental Results

Two set of experiments have been conducted. In the first one, the RFE method has been applied with various parameters for each patient individually due to the high variability between patients. In the second set of experiments, we apply fuzzy clustering vs. self organizing maps and evaluate the degree of agreement between learned clusters and psychiatric assessments.

#### 3.1 RFE on Acoustic Data

RFE calculation has been prepared using *caret* package coming from CRAN repository and 10-fold cross validation. We present and discuss detailed RFE results for 2 exemplary patients. Both patients have 3 visits for which the patient used BDmon application in the surrounding days. At each visit the mental state of a patient was assessed by the doctor (e.g., euthymia, depression). The ground-truth for the analysis is considered as in [4] and [3] and all phone calls conducted in period starting from 7 days before visit, the day of visit and 2 days after visit received the label (which was given during that visit). The number of total labeled phone calls for considered patients is summarized in Table 1. For each phone call, 86 acoustic parameters are extracted for all its frames (frame length: 10–20 ms), so usually there are thousands of frames used as training data for the RFE algorithm for one phone call.

As observed in Table 1, data are incomplete and for 3 out of 6 visits records from some days are missing (2 days for visit from 20.06, 6 days for 07.08 and 1 day for 19.06).

Results obtained by the RFE for both patients are presented in Table 2 and Table 3. It turned out that for patient 1472 the best results are received when all 86 variables are taken into account and then accuracy of that model is slightly above 80%. Similarly, for patient 2582 the best model is the one that uses 86 parameters and its accuracy equals to 65%. However, the difference in accuracy for smaller number of parameters is relatively small and for example, the accuracy with 8 parameters (reduction by over 90%) amounts to 78.2% and 61.4%, respectively. These results are very promising.

**Table 1.** Summary of considered available data: psychiatric assessments and recorded phones calls for 2 exemplary patients in the days surrounding the visit to the doctor (labeled data).

Patient	Visit date	Psychiatric assignments	Nb. of phone calls	Nb of surrounding days with active BDmon app
1472	28.03	Mixed	188	10
1472	20.06	Euthymia	142	8
1472	07.08	Mixed	73	4
2582	19.06	Euthymia	57	9
2582	17.07	Mixed	75	10
2582	09.10	Depression	69	10

Another coefficient called Kappa presented in Table 2 and Table 3 points to classification accuracy because is useful during class imbalance. Classification is normalized at the baseline of random chance on dataset. Received values oscillate around 0.3 which is interpreted as fair agreement.

It is also important to mention that the RFE method is rather time-consuming. Calculations for one patient lasted more than 17h<sup>2</sup> when it was conducted for 5% of randomly selected frames from each voice call.

**Table 2.** Results received by RFE methods for patient 1472

Patient 1472 (binary classification)				
Variables	Accuracy	Kappa	AccuracySD	KappaSD
4	0.761	0.192	0.003	0.011
8	<b>0.782</b>	0.250	0.006	0.027
16	0.793	0.291	0.006	0.010
<b>86</b>	<b>0.802</b>	<b>0.323</b>	<b>0.002</b>	<b>0.007</b>
Time calculation: 17.34 h				

The applied RFE methods returned the subset of ordered variables, that achieved the best results for classification (according to the random forest algorithm).

The final subsets of first 10 parameters learned separately on data of both patients are presented in Table 4. Selected first 10 most relevant parameters in received order because of that received accuracy between 86 parameters and 8 parameters published in Table 2 and Table 3 are small.

Analysis of the received parameters shows that majority of the parameters coming from Mel-Frequency Cepstral Coefficient Fourier transformate (group of

<sup>2</sup> 3,1 GHz Intel Core i7 500GB SSD, 16 GB Ram.

**Table 3.** Results received by RFE methods for patient 2582

Patient 2582 (3-class classification)				
Variables	Accuracy	Kappa	AccuracySD	KappaSD
4	0.538	0.138	0.011	0.026
8	<b>0.614</b>	0.293	0.005	0.010
16	0.636	0.328	0.004	0.008
<b>86</b>	<b>0.650</b>	<b>0.345</b>	<b>0.003</b>	<b>0.006</b>
Time calculation: 17.17 h				

**Table 4.** Parameters selected by RFE for patients 1472 (left) and 2582 (right)

No.	Parameter
1	<b>f0env_sma</b>
2	<b>slope0500_sma3</b>
3.	<b>pcm_fftMag_mfcc_1_</b>
4	pcm_fftMag_mfcc_3_
5	<b>pcm_fftMag_mfcc_4_</b>
6	loudness_sma3
7	<b>pcm_fftMag_mfcc_6_</b>
8	pcm_fftMag_mfcc_9_
9	slope5001500_sma3
10	pcm_fftMag_mfcc_8_

No.	Parameter
1	<b>pcm_fftMag_mfcc_4_</b>
2	<b>slope0500_sma3</b>
3.	<b>f0env_sma</b>
4	pcm_fftMag_mfcc_2_
5	<b>pcm_fftMag_mfcc_1_</b>
6	pcm_fftMag_fband0250_sma
7	pcm_fftMag_spectralentropy-_sma_compare
8	pcm_fftMag_mfcc_0_
9	<b>pcm_fftMag_mfcc_6_</b>
10	pcm_zcr_sma

variables: pcm\_fftMag\_mfcc\_ $n$ ) which indicates range of pitch. We conclude that 5 (out of 10) parameters (marked in bold in Table 4) are present in both of the subsets and these 5 parameters are considered as *RFE subset* of parameters for the clustering algorithms in next Sections.

### 3.2 Fuzzy C-Means vs. Self Organizing Maps

The second set of experiments consists of application of fuzzy clustering and self organizing maps to two alternative subsets variables:

- *RFE subset* (described in Sect. 3.1.);
- 12 features subjectively selected by medical experts and data analysts as introduced in [4] denoted as *Kam20*.

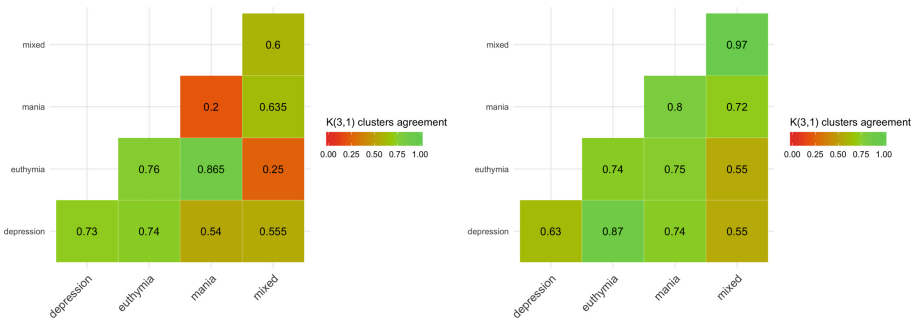
For every phone call, the selected acoustic features were extracted from the 10–20 ms frames and aggregated by five-number summary consisting of quartiles (0, 0.25, 0.5, 0.75, 1). The resulting two datasets were used as input for the unsupervised learning algorithms. This experiment was performed for all patients available for the *BDmon* study who have at least one pair of assessments with at



least 2 data points available during the surrounding days of the assumed ground-truth (-7 to +2 days). As a result, 17 patients and 62 pairs were considered in this experiment. Degree of agreement (2) was calculated for each pair of assessments. Next, we averaged results for the same (concordant) and different (incompatible) types of BD episodes assessed during visits.

It also needs to be noted, that in [4], which was the inspiration for the *Kam20* subset, different level of aggregation was applied and all of frames coming from mobile calls from the last 3 days were aggregated into quartiles. In this research, the aggregates are calculated for all individual frames from one phone calls and this procedure is implemented for each voice parameter.

**Self Organizing Maps.** Results received from SOMs for both subsets of parameters are depicted in Fig. 1. As observed, there are notable differences between the two heatmaps.



**Fig. 1.** Degree of agreement for SOM for (left) *Kam20* parameters (right RFE subset of acoustic parameters).

On the left diagram, there are results of the degree of agreement where SOM algorithm is used on *Kam20* parameters. On the diagonal where we strive to achieve values aiming to 1, we received following values: 0.73 for agreement between depression-depression and 0.76 for agreement between euthymia-euthymia which are quite satisfying, 0.2 for agreement between mania and mania seems insufficient due to specificity of that phase, and 0.6 for agreement between mixed and mixed - that value is rather high considering the overall difficulty to identify the mixed state (depressive and manic symptoms are present).

The results obtained on the remaining positions strive for the lowest possible values. As observed, the degree of agreement between states euthymia-mania is high and equal to 0.865, and this result is contrary to the knowledge of medical experts and their intuitions. Euthymia is the state of health and mania is the state of BD disease. We'd rather expect that clusters learned for data around these two types of labeled visits does not agree to a high degree. The remaining

results are quite satisfactory like in case mixed-euthymia where receive 0.25 and between mania-depression where receive 0.54.

On the right diagram of Fig. 1, there are results of the degree of agreement where SOM algorithm used parameters selected by RFE methods.

On the diagonal we received the following values: 0.63 for agreement between depression-depression which is worse than using *Kam20*; 0.74 for the agreement between euthymia-euthymia which is quite satisfying; 0.8 for the agreement between mania and mania is an increase (compared to the previous heatmap) in a positive way, and 0.97 for agreement between mixed and mixed - that value is very impressive.

Results obtained on remaining position strive for the lowest possible values. All of the remaining values are above 0.5 (which could be a border) which is satisfactory only to some extent.

Summarizing this result, overall the *RFE subset* delivered better degrees of agreement than *Kam20*. It is surprising that when we aim to as low value as possible, got the highest degree (comaprison of mania and euthymia on the first heatmap).

**Fuzzy C-Means.** Similarly to SOM, results received from fuzzy clustering differ between alternative subsets of acoustic features. Figure 2 summarizes the degree of agreement for both subsets.

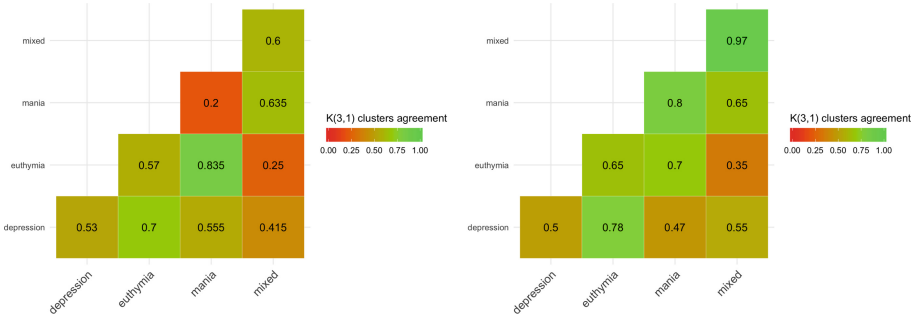
On the left diagram there are results of the degree of agreement where algorithm used *Kam20* parameters. On the diagonal where we strive to achieve values aiming to 1, we received the following values: 0.53 for agreement between depression-depression and 0.57 for agreement between euthymia-euthymia which are sufficient only to some extent; 0.2 for the agreement between mania and mania which is definitely insufficient due to the specificity of phase, and 0.6 for the agreement between the mixed and mixed.

Results obtained on the remaining position strive for the lowest possible values. The worst results were received again between the states of euthymia-mania where the degree of agreement is high and equal to 0.835 which is very high. In case of euthymia-mixed that could be sufficient because degree of agreement is only 0.25

On the right diagram there are results of the degree of agreement where algorithm used parameters selected by RFE methods.

On the diagonal we received the following values: 0.5 for the agreement between depression-depression which is moderately high compared to previous heatmaps; 0.65 for the agreement between euthymia-euthymia which is quite satisfying; 0.8 for the agreement between mania and mania where is increase in a positive way; and 0.97 for agreement between mixed and mixed - that value is very impressive.

Results obtained on remaining position strive for the lowest possible values. Two of the remaining values are below 0.5 which is satisfactory in comparison to the previous heatmaps. Overall others values are lowest then in previous examples.



**Fig. 2.** Degree of agreement for Fuzzy C-Means algorithm (left) Kam20 parameters (right) RFE subset of acoustic parameters

Detailed results of the experiments for the Fuzzy C-mean algorithm using parameters coming from RFE methods are presented in Table 5. Meaning of columns is as follow:

Visit A & Visit B - means visit order number, Assessment A & Assessment B - contain received labels from psychiatrists, Grouping agreement - is the coefficient calculated for each case using (2), Valid days A & Valid days B - contain the number of days that have any of voice parameters for that day.

**Comparative Analysis.** To compare all of the received results, the average degree of agreement for particular pairs of visits were calculated and are presented in Table 6.

We distinguish pairs of visits with the concordant labels of psychiatric assessment, namely: euthymia-euthymia, depression-depression, mania-mania, mixed-mixed. The average degree of agreement for these concordant labels is the highest for the FCM applied in the *RFE* subset of with parameters and amounts to 0.71. For the incompatible labels (e.g. euthymia-depression), the average degree of agreement is expected to be the lowest, and again, the FCM on *RFE* approach outperforms other variants (0.51).

At the same time, it needs to be noted that the fact that for SOM we receive in general clusters that are not that well corresponding to the psychiatric assessments does not necessarily mean that the applied clustering approach is making a mistake. It needs to be noted that there is a lot of uncertainty related to the psychiatric assessments itself, including the fact that the episodes are determined depending on the total number of points using to the Hamilton Scale of Depression (HAMD), and e.g., 8 points are classified as depressive episode whereas 7 points are regarded still as a healthy episode (euthymia).

**Table 5.** The degree of agreement between clusters learned by the fuzzy clustering with RFE features vs. labels from psychiatric assessment based on pairs of visits

Patient ID	Visit A	Visit B	Assessment A	Assessment B	Grouping agreement	Valid days A	Valid days B
837	1	2	euthymia	euthymia	0.80	2	10
1472	2	3	mania	mixed	0.70	9	10
1472	2	4	mania	mixed	0.61	9	6
1472	3	4	mixed	mixed	0.96	10	6
2004	1	2	euthymia	euthymia	0.83	3	10
2004	1	3	euthymia	depression	0.96	3	10
2004	2	3	euthymia	depression	0.70	10	10
2582	2	3	mixed	mania	0.80	10	10
2582	2	4	mixed	euthymia	0.50	10	6
2582	2	5	mixed	depression	0.60	10	10
2582	3	4	mania	euthymia	0.40	10	6
2582	3	5	mania	depression	0.40	10	10
2582	4	5	euthymia	depression	0.70	6	10
4248	1	2	depression	depression	0.00	3	3
4248	1	3	depression	depression	0.60	3	10
4248	1	4	depression	depression	1.00	3	10
4248	1	5	depression	mania	0.33	3	3
4248	2	3	depression	depression	0.33	3	10
4248	2	4	depression	depression	0.33	3	10
4248	2	5	depression	mania	0.33	3	3
4248	3	4	depression	depression	0.80	10	10
4248	3	5	depression	mania	0.60	10	3
4248	4	5	depression	mania	1.00	10	3
4953	2	3	mania	mixed	0.70	10	3
4953	2	4	mania	depression	0.70	10	10
4953	3	4	mixed	depression	0.80	3	10
5656	1	2	euthymia	euthymia	0.55	7	9
5656	1	3	euthymia	depression	0.85	7	7
5656	1	4	euthymia	depression	1.00	7	3
5656	2	3	euthymia	depression	0.69	9	7
5656	2	4	euthymia	depression	0.55	9	3
5656	3	4	depression	depression	0.85	7	3
5659	1	2	euthymia	euthymia	0.40	3	10
5736	2	3	depression	mixed	0.40	10	10
5768	2	3	euthymia	euthymia	0.60	10	10
5768	2	4	euthymia	euthymia	0.80	10	10
5768	3	4	euthymia	euthymia	0.70	10	10
6139	3	4	mania	mixed	0.00	3	6
6139	3	5	mania	depression	0.00	3	2
6139	4	5	mixed	depression	0.00	6	2
6601	1	3	euthymia	euthymia	0.50	2	6
6601	1	4	euthymia	depression	1.00	2	6
6601	3	4	euthymia	depression	0.50	6	6
8866	1	2	depression	mixed	0.88	3	9
8866	1	3	depression	depression	0.91	3	8
8866	2	3	mixed	depression	0.36	9	8
9341	2	3	mania	mania	0.80	1	5
9341	2	4	mania	mixed	1.00	1	3
9341	2	5	mania	euthymia	0.20	1	5
9341	3	4	mania	mixed	0.00	5	3
9341	3	5	mania	euthymia	0.60	5	5
9341	4	5	mixed	euthymia	0.20	3	5
9829	1	2	depression	depression	0.10	2	10
9829	1	3	depression	depression	0.50	2	10
9829	1	4	depression	depression	0.00	2	10
9829	1	5	depression	depression	1.00	2	10
9829	2	3	depression	depression	0.50	10	10
9829	2	4	depression	depression	0.60	10	10
9829	2	5	depression	depression	0.20	10	10
9829	3	4	depression	depression	0.50	10	10
9829	3	5	depression	depression	0.80	10	10
9829	4	5	depression	depression	0.10	10	10

**Table 6.** The average degree of agreement for concordant, semi-concordant and incompatible labels. Concordant labels are as follows: E-E, D-D, M-M, X-X; incompatible labels: E-M, E-D, E-X; and semi-concordant labels: M-X, D-X, where E denoted euthymia (healthy state), D stands for depression, M for the mania and X denotes the mixed state

Avg degree of agreement	FCM-RFE	FCM-Kam20	SOM-RFE	SOM-Kam20
Concordant	<b>0.71</b>	0.47	0.79	0.57
Incompatible	<b>0.51</b>	0.65	0.73	0.68
Semi-concordant	<b>0.52</b>	0.54	0.63	0.59

### 3.3 Conclusions

Recursive feature elimination enabled to significantly reduce the number of important acoustic features (from 86 to 5 parameters). Furthermore, we conclude that the degree of agreement between clustering results and psychiatric labels vary between the applied fuzzy and SOM clustering methods and the subsets of acoustic features. The highest degree of agreement for concordant BD episodes has been achieved using RFE subset of 5 parameters and the fuzzy clustering algorithm (Fuzzy C-means). The lowest degree of agreement for incompatible BD episodes has been achieved also for the fuzzy clustering algorithm. Thus, the most satisfactory results for the degree of agreement has been achieved by this fuzzy clustering and the subset of acoustic features selected with the RFE method.

In future work, we consider representation of the smartphone data as fuzzy numbers instead of vectors with craps quartiles. Also, we plan to further examine the characteristic of clusters obtained with Fuzzy C-means algorithm and interpret them from the medical perspective. It seems that RFE methods is very promising and should be tested for higher number of patients, to obtain more accurate results. However, that is time consuming, so it needs to be tested in more efficient environment.

**Acknowledgment.** Datasets considered in this paper were collected in the CHAD project – entitled “Smartphone-based diagnostics of phase changes in the course of bipolar disorder” (RPMA.01.02.00-14-5706/16-00) that was financed from EU funds (Regional Operational Program for Mazovia) in 2017–2018. The authors thank psychiatrists and patients that participated in the observational study for their commitment. The authors thank the researchers Karol Opara and Weronika Radziszewska from Systems Research Institute, Polish Academy of Sciences for their support in data preparation and analysis, as well as the researchers Monika Dominiak, Anna Wójcińska and Łukasz Świącicki from Institute of Psychiatry and Neurology for their advice and comments.

## References

1. Grande, I., et al.: Bipolar disorder. In: *The Lancet*, vol. 387, no. 10027, pp. 1561–1572 (2016). [https://doi.org/10.1016/S0140-6736\(15\)00241-X](https://doi.org/10.1016/S0140-6736(15)00241-X), <http://www.sciencedirect.com/science/article/pii/S014067361500241X>, ISSN: 0140-6736
2. Faurholt-Jepsen, M., et al.: Objective smartphone data as a potential diagnostic marker of bipolar disorder. *Aust. New Zealand J. Psychiatry* **53**(2), 119–128 (2019). <https://doi.org/10.1177/0004867418808900>, PMID: 30387368
3. Gruërbl, A., Muaremi, A., Osmani, V.: Smartphone-based recognition of states and state changes in bipolar disorder patients. *IEEE J. Biomed. Health Inform.* **19**(1), 140–148 (2015)
4. Kamińska, O., et al.: Self-organizing maps using acoustic features for prediction of state change in bipolar disorder. In: Marcos, M., et al. (eds.) *KRAHC/TEAAM - 2019. LNCS (LNAI)*, vol. 11979, pp. 148–160. Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-37446-4\\_12](https://doi.org/10.1007/978-3-030-37446-4_12)
5. Wollmer, M., Eyben, F., Schuller, B.: openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor (2010)
6. Guyon, I., et al.: Gene selection for cancer classification using support vector machines. *Mach. Learn.* **46**(1–3), 389–422 (2002). <https://doi.org/10.1023/A:1012487302797>
7. Kohonen, T.: Self-organizing maps (1995). <https://doi.org/10.1007/978-3-642-97610-0>
8. Wehrens, R., Krusselbrink, J.: Flexible self-organizing maps in kohonen 3.0. *J. Stat. Softw.* **87**(7), 1–18 (2018). ISSN: 1548-7660, <https://doi.org/10.18637/jss.v087.i07>, <https://www.jstatsoft.org/v087/i07>
9. Bezdeck, J.C., Ehrlich, R., Full, W.: FCM: fuzzy C-means algorithm. *Comput. Geosci.* **10**(2–3), 191–203 (1984)