

## **Monitoring Acoustic Data Streams Using Possibilistic Aggregation and Multi-label Learning**

\***Katarzyna Kaczmarek-Majer<sup>a</sup>** and **Olgierd Hryniewicz<sup>a</sup>** and **Dominika Kornobis<sup>b</sup>**  
and Anna Stachowiak<sup>b</sup>

<sup>a</sup>Systems Research Institute Polish Academy of Sciences, Newelska 6, 01-147 Warsaw, Poland  
katarzyna.kaczmarek@ibspan.waw.pl

<sup>b</sup>Adam Mickiewicz University, Faculty of Mathematics and Computer Science,  
Uniwersytetu Poznańskiego 4, 61-614 Poznan, Poland

### **Abstract**

This paper introduces a novel procedure for statistical monitoring of acoustic data streams supported by possibilistic aggregation and multi-label learning. The primary goal is to learn improved control limits for the residual control charts due to the objective removal of measurements from the non-healthy state of a patient determined through multi-label learning. The proposed procedure is illustrated with real-life acoustic data collected from smartphones of Bipolar Disorder patient. Multi-label classification enabled to distinguish between different degrees of severity of manic and depressive symptoms, and especially for the mixed state - their simultaneous occurrence.

**Keywords:** Possibilistic Aggregation, Residual Control Chart, Acoustic data streams, Multi-label classification, Bipolar Disorder Phase Change Detection.

### **1 Introduction**

Early detection of a new episode in Bipolar Disorder (BD) is essential for more efficient treatment, improved quality of life and prevention of suicides. Acoustic features about patients voice extracted from smartphones are confirmed as reliable markers in the assessment of depressive and manic symptoms in BD, see e.g., [1, 3]. At the same time, applications of statistical process control are rich and already include smartphone-based monitoring of mental illnesses. For example, in [11, 12], the authors show the usefulness of the weighted model averaging in the residual control charts for early detection of change in the state of Bipolar Disorder (BD) basing on the behavioural data about smartphone usage and the limited amount of diagnostic data. However, the problem arises for

non-stationary and imprecise processes such as acoustic data streams. In particular, if there is no initial stable phase that can be considered as in-control and applied to learn the control limits. Acoustic data streams are usually not only non-stationary, but also subject to various sources of uncertainty that should be appropriately managed. In particular, the smartphone data are collected in the patient's naturalistic setting and impacted by various technical and contextual issues, e.g., the background noises are dynamically changing depending on the patient's location, the topic being discussed and the interlocutor speech are not tracked due to privacy issues. Furthermore, there is an additional uncertainty related to the patient's mental state that is assessed during psychiatric interviews with limited frequency and according to partially subjective rating scales. To reflect these uncertainties, we adapt the approximate representation of smartphone data following [10] and multi-label learning.

The main novelty of this paper is the proposed procedure for the construction of control limits of the residual control charts using multi-label filtering of measurements and approximate data representation. The primary goal is to exclude observations from the non-healthy state when learning the control limits of the residual processes. The proposed approach is illustrated with a real-life example of acoustic data collected from smartphones of Bipolar Disorder patients.

The structure of the paper is as follows. Section 2 briefly describes the considered procedure for the statistical process control using possibility distributions and their fuzzy characteristics. In Section 3, multi-label classification applied to identify non-healthy days is explained. Section 4 introduces the application scenario of bipolar disorder episode detection. An illustrative example is presented in Section 5. In Section 6, main conclusions are discussed and future work is outlined.

## 2 Control charts for the possibilistic aggregates

In the case of monitoring acoustic features, we are confronted with non-stationary data streams. Probability distributions that describe measurements of such features for one recorded call may be strongly varying from call to call, even during the same day. Within this research, we assume that each measurement of a certain voice feature is assigned to one of  $m$  classes defined by the class limits  $(c_0, c_1, \dots, c_m)$ . Therefore, the stream of measurements may be represented by the stream of histograms  $k_{i,j,l}, i = 1, \dots, n, j = 1, \dots, n_i, l = 1, \dots, m$ , where  $k_{i,j,l}$  is the number of observations from the  $i$ -th day of observations ( $i = 1, \dots, n$ ) and its  $j$ -th call ( $j = 1, \dots, n_i$ ). Let  $k_{i,j}$  be the number of observations in the  $(i, j)$ -th call. Then, these histograms are transformed into empirical probability density functions (epdf's)  $p_{i,j,l}, i = 1, \dots, n, j = 1, \dots, n_i, l = 1, \dots, m$ , where  $p_{i,j,l} = k_{i,j,l}/k_{i,j}$  represent the estimates of probability distributions with piecewise constant density functions. For example, consider the case that we have a stream of two histograms with  $m=3$  classes: (50; 50; 50) and (10; 40; 10), respectively. These histograms can be transformed into the following epdf's: (1/3; 1/3; 1/3) and (1/6; 2/3; 1/6).

The psychiatric state of BD patients is assessed on the level of days of observations (not individual calls). Therefore, Hryniewicz and Kaczmarek-Majer [10] have proposed to aggregate acoustic data measured during the same day into a possibility distribution. The concept of the possibility distribution has many different interpretations. According to one of them, the possibility distribution can be understood as an upper envelope for all ordinary discrete probability distributions compatible with a value imprecisely described by a fuzzy number. In our case, the epdf's describing data collected from separate calls can be viewed upon as possibly different discrete probability distributions, so the attempt to aggregate them using a respective possibility distribution seems to be fully justified.

Hryniewicz and Kaczmarek-Majer in [10] propose several methods of the possibilistic aggregation. Within this research, we will use the simplest one, based on the maximum t-conorm aggregation operator. When this aggregation operator is used, the upper envelope  $(\tilde{p}_{i,1}, \dots, \tilde{p}_{i,m}), i = 1, \dots, n$  of the aggregated epdf's can be computed using the following formula

$$\tilde{p}_{i,l} = \max(p_{i,1,l}, \dots, p_{i,n_i,l}), i = 1, \dots, n, l = 1, \dots, m \quad (1)$$

The upper possibilistic envelope, defined by (1), is not a proper possibility distribution, as it is not represented

as a fuzzy number whose membership function must be convex and normalized (i.e., having the maximum equal to one). Let us find an interval (class) for whom the upper envelope  $(\tilde{p}_{i,1}, \dots, \tilde{p}_{i,m}), i = 1, \dots, n$  attains its maximum. Denote the index of this interval by

$$L_i = \arg \max_{l=1, \dots, m} \tilde{p}_{i,l}, i = 1, \dots, n. \quad (2)$$

Note, that this maximum may be attained in more than one interval, and in this case the choice of  $L_i$  may be arbitrary. For our toy example of two histograms, the upper possibilistic envelope calculated with the maximum t-conorm (AG1) is (1/3; 2/3; 1/3).

Let  $\tilde{p}_i^M = \tilde{p}_{i,L_i}, i = 1, \dots, n$ . The upper possibilistic envelope may be transformed to the proper possibility distribution  $(\pi_{i,1}, \dots, \pi_{i,m}), i = 1, \dots, n$  defined on the set of intervals defined by class limits  $(c_0, c_1, \dots, c_m)$  using the following transformation

$$\pi_{i,l} = \begin{cases} \max(\tilde{p}_{i,1}, \dots, \tilde{p}_{i,l}) / \tilde{p}_i^M & \text{for } l = 1, \dots, L_{i,M} \\ \max(\tilde{p}_{i,l}, \dots, \tilde{p}_{i,m}) / \tilde{p}_i^M & \text{for } l = L_{i,M}, \dots, m \end{cases} \quad (3)$$

for  $i = 1, \dots, n$ .

The fuzzy numbers that represent the possibility distributions for the  $n$  considered segments are defined by the following membership functions

$$\mu_i(x) = \begin{cases} \pi_{i,1} & \text{for } c_0 < x \leq c_1 \\ \dots & \dots \\ \pi_{i,j} & \text{for } c_{j-1} < x \leq c_j \\ \dots & \dots \\ \pi_{i,m} & \text{for } c_{m-1} < x \leq c_m \end{cases} \quad (4)$$

for  $i = 1, \dots, n$ .

There are many methods for the crisp representation of fuzzy numbers, and the center of gravity (COG) is the most popular one.

$$COG_i = \int_{c_0}^{c_m} x \mu_i(x) dx / \int_{c_0}^{c_m} \mu_i(x) dx, i = 1, \dots, m, \quad (5)$$

where  $\mu_i(x)$  is defined by (4).

Other important characteristics of the possibility distribution defined by (3) are computed using the concept of fuzzy intervals introduced by Dubois and Prade [4] and defined using the lower and upper limits or the weighted lower and upper fuzzy limits, calculated according to Delgado et al. [2] that are defined in the following way

$$A_{i,L}^1 = \sum_{i=1}^{L_i} c_i (\pi_{i,L}^2 - \pi_{i-1,L}^2), i = 1, \dots, n, \quad (6)$$

where  $\pi_{0,i} = 0, i = 1, \dots, n$ , and

$$A_{i,U}^1 = \sum_{i=L_i+1}^{m+1} c_i(\pi_{i,t}^2 - \pi_{i-1,t}^2), i = 1, \dots, n, \quad (7)$$

where  $\pi_{m+1,i} = 0, i = 1, \dots, n$ .

The lower and upper limits calculated according to (6) - (7) are used for the calculation of such characteristics of the considered possibility distributions as Fuzzy Value (FV)

$$FV_i = A_{i,L}^1 + A_{i,U}^1, i = 1, \dots, n, \quad (8)$$

and Ambiguity (Amb)

$$Amb_i = A_{i,U}^1 - A_{i,L}^1, i = 1, \dots, n. \quad (9)$$

Within this research, we monitor the aforementioned characteristics with statistical process control. Statistical process control is a well-established field that enables to monitor processes and to early detect its deterioration. In 1924 W. Shewhart introduced a simple tool for monitoring stable processes - a control chart. In its initial stage, which is assumed to be in-control (stable), monitored process characteristics are measured, and their mean value and standard deviation are recorded. These recorded values are used for the design of a control chart, known as the Shewhart control chart, which consists of control lines: central, and two (or one, when only deviation of a process level in one direction is interesting) control. The central line represents the mean value of the process level (or a certain target value for this process), and control lines are usually located at three standard deviations from a central line. The process is considered stable when its future observations are located inside control lines (limits). In the classical Shewhart control chart it is assumed that consecutive observations (measurement) are not autocorrelated.

Unfortunately, it is often not the case for data streams of the BD patients. Therefore, within this research, a chart for monitoring residuals is constructed. To obtain the residuals, predictions are commonly calculated according to the autoregressive model AR of the most appropriate order  $p$

$$X_t - \mu - a_1 X_{t-1} - \dots - a_p X_{t-p} = Z_t, \quad (10)$$

where  $X_i, i = t, t-1, \dots$  are random observations of a process,  $Z_i, i = t, t-1, \dots$  is a series of independent random variables with constant variance and zero expectation,  $a_1, \dots, a_p$  are process parameters describing its correlation structure, and  $\mu$  is a constant describing a process level. Thus, instead of the raw results of measurements, we chart the differences between the values of measurements and their predictions.

Finally, the key point in monitoring BD is the lack of a stable period that could drive the calculation of the Upper and Lower Control Limits (UCL and LCL). One solution would be to calculate the control limits basing only on the measurements from the healthy period. However, such an approach would significantly reduce the size of the dataset used to learn the control limits. Another approach is learning the control limits using all measurements. However, the influential observations shall be removed. The use of machine learning allows deleting data that substantially mismatch. Actual outliers are not always identifiable by strictly statistical methods. Removal of such data should enhance the capabilities of the control chart when it comes to future process change detection. Within this paper, we apply a multi-label classification to identify most non-healthy observations.

### 3 Multi-label classification

Multi-label learning considers multi-dimensional space of instances  $\underline{x}_i \in \mathcal{X}$ , where every example is associated with a set of possible class labels  $\mathcal{Y} = \{y_1, y_2, \dots, y_q\}$ . Multi-label classifier is a function from the features space to the power-set of labels  $h : \mathcal{X} \rightarrow 2^{\mathcal{Y}}$  [16]. Majority of the multi-label algorithms use traditional classification algorithms by adapting them to a multi-label approach, the so called Algorithm Adaptation Methods, or transforming a multi-label problem into several subproblems of traditional classification - methods named Problem Transformation Methods [16].

One of the first multi-label problem transformation methods is Binary Relevance (BR), where multi-labels learning problem is turned into several independent binary classification problems. For the feature space associated with  $q$  possible class labels, BR algorithm splits the input data set into  $q$  subsets, where for each of the subsets only one label -  $y_j$  - is considered [9].

For every training set

$$\mathcal{D}_j = \{(\underline{x}_i, \phi(Y_i, y_j)) \mid 1 \leq i \leq m\} \quad (11)$$

where

$$\phi(Y_i, y_j) = \begin{cases} +1 & \text{if } y_j \in Y_i \\ -1 & \text{otherwise} \end{cases} \quad (12)$$

a classifier  $g_j : \mathcal{X} \rightarrow \mathbb{R}$  is constructed. Therefore,  $q$  binary classifiers are built based on each set created. BR algorithm predicts associated label set for unseen instance  $\underline{x}$ :

$$Y = \{y_j \mid g_j(\underline{x}) > 0, 1 \leq j \leq q\} \quad (13)$$

In this paper, we applied the Binary Relevance based classifiers with either the k-nearest neighbours (KNN)

or Random Forest (RF) as a binary classifier within it using the *utiml* [14] package from the CRAN repository for R language. The main advantage of BR is its simplicity and directness. It is possible to improve its performance by using an Ensemble of Binary Relevance (EBR) [9].

In many cases, it is also important to consider dependencies between labels [7]. Such feature is provided by the Classifier Chains (CC) algorithm and its extension to Ensemble of Classifier Chains (ECC). The purpose of applying the ECC algorithm is also, as was the case with the BR, to transform the multi-label problem into  $q$  simpler binary classification problems. The difference is that classifiers are trained sequentially using a fixed order of labels [15]. Permutation function is used to define ordering:  $\tau : \{1, \dots, q\} \rightarrow \{1, \dots, q\}$ .

Each subsequent input set contains all of the attributes and is extended with the prediction of the previous label defined in order:

$$\mathcal{D}_{\tau(j)} = \{([\underline{x}_i, \underline{pre}_{\tau(j)}^i], \phi(Y_i, y_{\tau(j)})) \mid 1 \leq i \leq m\} \quad (14)$$

where:

$$\underline{pre}_{\tau(j)}^i = (\phi(Y_i, y_{\tau(1)}), \dots, \phi(Y_i, y_{\tau(j-1)})) \quad (15)$$

Having constructed training sets, classifiers are built  $g_{\tau(j)} : \mathcal{X} \times \{-1, +1\}^{j-1} \rightarrow \mathbb{R}$ . Let us use the notation of the binary prediction label  $y_{\tau(j)}$  on the instance  $\underline{x}$ :  $\lambda_{\tau(j)}^{\underline{x}} \in \{-1, +1\}$  in accordance with [16]. Then we present the set of labels as follows

$$Y = \{y_{\tau(j)} \mid \lambda_{\tau(j)}^{\underline{x}} = +1, 1 \leq j \leq q\} \quad (16)$$

The idea of the ECC is that several classifier chains are trained using different label sequences and we get the final prediction using collective voting [15].

#### 4 Application context: Bipolar disorder episode prediction

Motivation for this research comes from a study conducted in the Department of Affective Disorders, Institute of Psychiatry and Neurology in Warsaw, Poland that included patients diagnosed with bipolar disorder (F31 according to ICD-10 classification). Bipolar disorder is a serious disease characterized by mood fluctuations from euthymia through depression and mania to mixed states.

Participants of the considered study received a dedicated mobile application, called *BDMon*, able to collect acoustic data about phone calls. The patient's

voice signal was divided into 20 ms frames (within one frame it is approximately stationary). The extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) for voice research [6] was extracted from each frame. This set comprises 86 features including (1) time-domain descriptors, such as e.g., zero-crossing rate (ZCR), amplitude statistics, (2) spectral features, such as e.g., mel-cepstral coefficients (MFCC), fundamental frequency (F0) and its harmonics. The selection of the best subset of acoustic features for BD episode prediction is still subject to research. Hopefully, the considered standardized feature extraction procedures (eGeMAPS) enable to compare results across studies [3]. In [8], the authors perform the filter feature selection and select the following parameters: kurtosis energy, mean second and mean third MFCC, mean fourth delta MFCC, max ZCR, std and range F0. In [13], the authors perform the recursive feature elimination and fuzzy clustering algorithm to establish the degree of agreement between concordant visits. The following parameters are selected: mean first, fourth and sixth MFCC, fundamental frequency (F0) envelope and the spectrum slope in sub-band (0-500 Hz) [13]. In this paper, we illustrate the performance of the proposed procedure with data streams of one of the Mel-frequency cepstral coefficients, that is sixth MFCC (MFCC6) calculated with the openSMILE library [5]. In sound processing, the mel-frequency cepstral coefficients collectively make mel-frequency cepstrum, which is a representation of the short-term power spectrum of a sound. These features are regarded as quite promising also in speech recognition.

Apart from the collection of the acoustic data from smartphones, evaluation of the patients' mental state was performed by psychiatrists with a various frequency depending on the need identified by the doctor or a patient. Rating scales of depressive and manic symptoms Hamilton Depression Rating Scale (HAMD) and Young Mania Rating Scale (YMRS) were adapted (see Table 1).

BD state	HAMD	YMRS
euthymia	less than 8	less than 6
weak depression	from 8 to 12	less than 6
depression	13 and more	less than 6
hipomania	less than 8	from 6 to 13
mania	less than 8	14 and more
mixed	8 and more	6 and more

Table 1: Characteristics of BD states

In the literature, the bipolar disorder predictive modelling is most commonly stated as a supervised learning task to distinguish between the following four

states: euthymia, mania, depression and the mixed states [1]. In [8], the authors consider 7 days before and 2 days after the examination (and phone interview) as ground-truth for analyses. However, the current state of the art seems to lack clear guidelines about the ground-truth before and after the psychiatric assessments [1]. Furthermore, bipolar disorder patients suffer from manic or depressive symptoms, which may also occur simultaneously, and thus, multi-label classification is very appealing to support BD episode prediction. To the best of the authors' knowledge, this is the first paper that applies multi-label classification to support monitoring of bipolar disorder patients. However, due to the very limited availability of psychiatric labels, statistical process control which is a kind of unsupervised learning seem more promising and in this research, the results of the multi-label classification are applied to support the statistical process control.

### 5 Experimental results

We now present illustrative examples using real-life data collected from smartphones and the psychiatric assessments. The main goal of these experiments was to evaluate if the proposed procedure is able to detect a starting episode.

#### 5.1 About data

Experiments are performed for acoustic data streams of an exemplary patient, who experienced two transitions from euthymia (healthy state) into the disease states during the study period and actively used the smartphone. Datasets contain partially labeled data. They are split into a training set (data streams collected in the transition from euthymia to the mixed state: 28 days between psychiatric assessments) and test set (data streams collected during the transition from euthymia to weak depression: 13 days between psychiatric assessments). Table 2 presents characteristics of datasets.

Characteristic	Training set	Test set
1st visit psych. state	E	E
2nd visit psych. state	WD	WD, M
no of valid patient-days	22	30
no of valid calls	160 (43%)	216 (57%)

Table 2: Dataset characteristics. Meaning of abbreviations: *E* - euthymia; *WD* - weak depression; *M* - mania; *T* - transition period, *D* - disease episode.

Training data are used to train the multi-label classifier. Ground-truth of 7 days before the psychiatric assessment and 2 days after is applied for classification. The

classifier is then run on test set to identify non-healthy days. Finally, residual control charts are constructed for the time frame of the test set.

Performance of the statistical process control for episode detection is evaluated with an assessment of the occurrence of alarms in euthymia, the transition or the disease period. It is expected that the alarm signal is generated in the disease period.

#### 5.2 Control chart for daily possibilistic aggregates

We start from illustrating a standard Shewart control chart as a reference starting point for the mean value of the considered Mel-cepstral coefficient (MFCC6) data stream of the test set (see Figure 1). It needs to

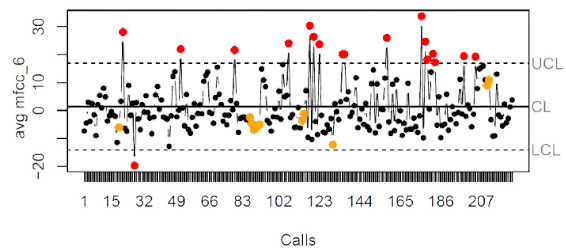


Figure 1: Shewart control chart of mean value of MFCC6 parameter for individual calls

be noted that call measurements were taken at various timestamps (uneven time intervals). Furthermore, it is observed that there are multiple alarms generated, including alarms in the beginning (euthymia) and there is no clear shift of the mean value observed for this process in the transition or the disease period.

The size of the input data is relatively large (e.g., 503685 frames for 33 days of 1 patient), but the aggregation enables to significantly decrease it. Let us now illustrate the outcomes of aggregation and construction of possibility distribution for data streams of one exemplary day. Table 3 gathers streams of histograms for the considered MFCC6 parameter from an exemplary day. There were 5 phone calls, so 5 histograms are constructed.

call	l=1	l=2	l=3	l=4	l=5	l=6	l=7	l=8	l=9	l=10
1	2	16	123	955	3074	2029	324	61	1	0
2	2	18	209	1306	3520	2640	618	43	0	0
3	0	0	18	150	496	709	286	82	5	0
4	0	0	45	651	1619	952	149	12	1	0
5	0	0	12	438	2325	1527	212	20	0	0

Table 3: Streams of histograms for MFCC6 data from day 30, no of calls=5, m=10

Each measurement is assigned to one of the 10 classes

( $m=10$ ). Next, maximum conorm is applied (upper envelope) to aggregate the daily data. Finally, possibility distribution is created. Figure 2 illustrates the upper envelope and the possibility distribution calculated for the considered exemplary day number 30.

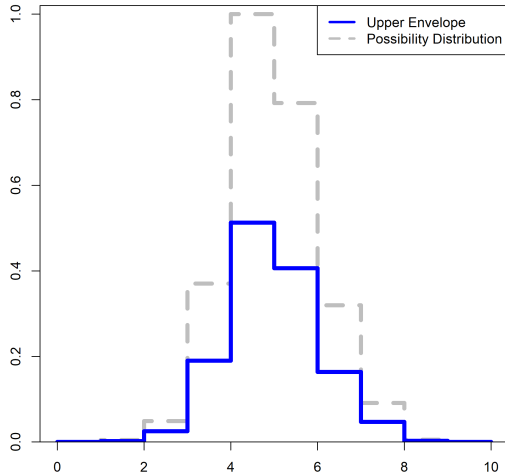


Figure 2: Upper envelope aggregation and possibility distribution for streams of MFCC6 data from one day

Finally, several characteristics starting from COG, see Eg. (5), are calculated and time series of such daily aggregates are monitored.

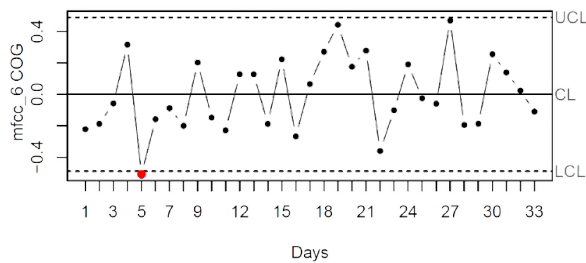


Figure 3: Shewart control chart of residuals estimated from AR(1) of COG process for the MFCC6 parameter for days

Since the time series of the daily aggregates are weakly autocorrelated (e.g.,  $r_1 = 0.101$ ), residual control charts are applied using AR(1). Two-sigma control limits are applied since, as in our case, the main objective of the monitoring is to signal possible changes in the patient's health as quickly as possible. Figure 3 illustrates the resulting control chart for residual for COG.

Contrary to the individual calls, when monitoring a process of daily COG aggregates, there are no alarms generated in the transition period and no alarms generated in the disease period. It is observed though that an

alarm signal is generated already on day 5, when the patient was expected to be still in euthymia (healthy state). This might be a consequence of too wide control limits. In the next Subsection, we thus propose a different method for constructing control limits.

### 5.3 Multi-label classification

Multi-label classification (MLC) is performed to improve the construction of the control limits. First, the performance of multiple classifiers is compared. Classifiers are trained on the training set and tested on the test set with the purpose to predict non-healthy days (see Table 4). Table 4 shows the effectiveness of the algorithms. For each day from the test set, we can observe what percentage of calls made by the patient on that day was properly classified. "Properly" means here that the algorithm indicated euthymia when the patient was healthy or a non-euthymic state when the patient was currently in the mixed state. The algorithm was considered effective when most of the days in the test set have reached the accuracy threshold of at least 0.5

day	BD state	calls per day	EBR RF	ECC KNN	RF	KNN
1	euthymia	2	<b>0.50</b>	<b>1.00</b>	<b>0.50</b>	0.00
2	euthymia	6	0.17	<b>0.50</b>	<b>0.50</b>	0.00
3	euthymia	7	0.43	<b>0.57</b>	<b>0.71</b>	0.43
4	euthymia	8	0.38	0.38	0.38	0.00
5	euthymia	6	0.17	0.17	<b>0.83</b>	0.00
6	euthymia	10	0.40	<b>0.60</b>	0.40	0.10
7	euthymia	10	0.40	<b>0.80</b>	<b>0.60</b>	0.00
8	euthymia	2	<b>0.50</b>	<b>1.00</b>	<b>1.00</b>	0.00
9	euthymia	6	0.17	<b>0.50</b>	<b>0.67</b>	0.00
24	mixed	11	<b>1.00</b>	<b>0.73</b>	<b>0.64</b>	<b>1.00</b>
25	mixed	2	<b>0.50</b>	0.00	<b>0.50</b>	<b>1.00</b>
26	mixed	6	<b>0.83</b>	0.33	0.33	<b>1.00</b>
27	mixed	10	<b>0.50</b>	0.30	0.10	<b>0.80</b>
28	mixed	7	<b>0.86</b>	0.29	0.43	<b>1.00</b>
29	mixed	7	<b>0.71</b>	<b>0.71</b>	0.29	<b>0.71</b>
30	mixed	11	<b>0.55</b>	0.45	0.27	<b>0.82</b>
31	mixed	2	<b>1.00</b>	<b>1.00</b>	<b>0.50</b>	<b>1.00</b>
32	mixed	5	<b>0.80</b>	<b>0.80</b>	<b>0.60</b>	<b>1.00</b>
33	mixed	14	<b>0.50</b>	<b>0.79</b>	0.14	<b>0.64</b>

Table 4: Performance of classifiers for the test set evaluated with the accuracy in detection of euthymia (days 1-9) and accuracy in detection the mixed state (days 24-33).

As depicted in Table 4, when the classifier was trained only on the train data and tested on the test set, it enabled to detect the disease episode, which is the mixed state in this context. By using the multi-label approach, it was possible to distinguish many degrees of severity of symptoms and especially for a mixed state - their simultaneous occurrence. It helped in a proper indication of observations predicted as non-healthy. It is also observed that both variants of the multi-label algorithm perform on a comparable level.

It needs to be noted that the day 31 was classified as

a day with the disease episode by both of the considered variants of the multi-label algorithms. Furthermore, the availability of outcomes of psychiatric assessments is usually very low and although multi-label approach delivered relatively good performance, it could be evaluated only on a relatively small sample limited to the labelled observations. We use the outcomes of the multi-label approach though, to improve the construction of control limits in the statistical process monitoring.

**5.4 Daily possibilistic aggregates filtered with multi-label learning**

We now build residual control charts excluding days indicated by MLC. It was observed that for both versions of the MLC algorithm, improved calculation of the Upper Control Limit (UCL) using limited sample enabled to detect an alarming day in the episode period when monitoring the COG process. The outcomes for the COG are presented in Figure 4 and EBR algorithm.

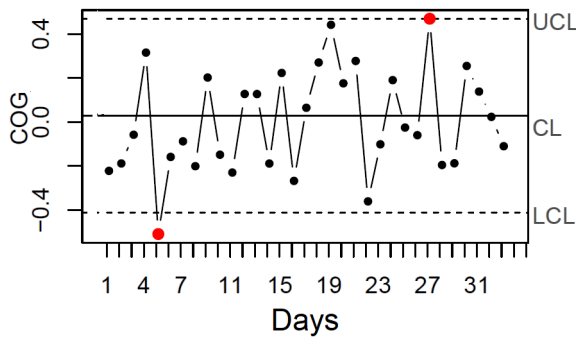


Figure 4: Shewart control chart of residuals estimated from AR(1) of COG process for the MFCC6 parameter for days. Control limits calculated using outcomes of the EBR MLC algorithm.

Similarly as in Figure 3, in Figure 4, we still observe an alarm generated in day no 5. However, an alarm is also observed now on day 27 which is the day of the mixed state due to the new control limits, which are  $LCL=-0.4121$  and  $UCL=0.4686$ . For the filtering with the ECC variant, alarms are generated the same days and the control limits are as follows:  $LCL=-0.4365$  and  $UCL=0.4445$ .

Figure 5 illustrates the residual control chart for the daily calls described with the Fuzzy Value (8). Interestingly, there was only one alarm signal on day 19 (the transition period). Finally, Figure 6 depicts residual control chart for the Ambiguity characteristic (9). Similarly, there was no alarm generated in euthymia. Instead, an alarm is generated on day 27 during the disease episode, so such an outcome was expected.

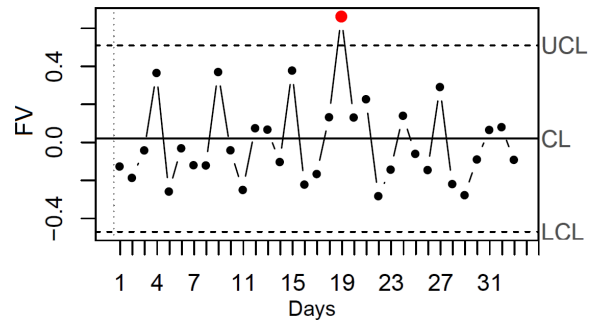


Figure 5: Shewart control chart of residuals estimated from AR(1) of Fuzzy Value process for the MFCC6 parameter for days. Control limits calculated using outcomes of the EBR MLC algorithm.

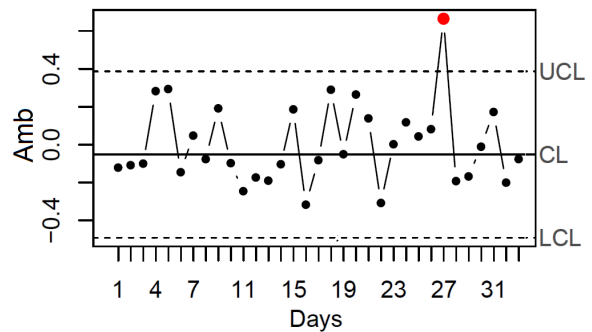


Figure 6: Shewart control chart of residuals estimated from AR(1) of Ambiguity process for the MFCC6 parameter for days. Control limits calculated using outcomes of the EBR MLC algorithm.

**6 Conclusion and Future Work**

In this paper, we monitor the stability of the aggregated acoustic data streams using the Shewart control chart and possibilistic distributions. We propose a new way of computing the control limits through filtering of observations according to the multi-label classification. Its main strength consists in taking into account the uncertainty about acoustic and psychiatric data through multi-label learning and the possibilistic representation. Illustrative examples have been discussed to show the performance of the proposed procedure in the real-life scenario of monitoring a starting episode in BD patients. Our preliminary results suggest that the proposed approach outperforms procedures for non-aggregated data and is competitive with the state-of-the-art procedures with no filtering. Notably, this approach is very promising also for detecting changes even in case of no psychiatric assessments (similarly to the idea of unsupervised learning). In future work, we plan to validate the proposed approach

with data for other patients and all acoustic features. Moreover, further developments of the methodology are planned, so that the multi-label filtering leads to removal of individual calls (not days) and improved aggregation procedure taking into account additional information about phone calls, such as e.g., type of calls (incoming/outgoing).

### Acknowledgement

Katarzyna Kaczmarek-Majer acknowledges funding from the Foundation for Polish Science through the “NeuroSmog: Determining the impact of air pollution on the developing brain” (Nr. POIR.04.04.00-1763/18-00) project which is implemented as part of the TEAM-NET programme of the Foundation for Polish Science, co-financed from EU resources, obtained from the European Regional Development Fund under the Smart Growth Operational Programme. 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) financed from EU funds (Regional Operational Program for Mazovia). The authors thank Karol Opara and Weronika Radziszewska for their support.

### References

- [1] A. Z. Antosik-Wójcicka, M. Dominiak, M. Chojnacka, K. Kaczmarek-Majer, K. R. Opara, W. Radziszewska, A. Olwert, Łukasz Świecicki, Smartphone as a monitoring tool for bipolar disorder: a systematic review including data analysis, machine learning algorithms and predictive modelling, *Int J Med Inform* 138:104131.
- [2] M. Delgado, M. Vila, W. Voxman, On a canonical representation of a fuzzy number 93 (1998) 125–135.
- [3] L. DM, B. KH, G. S. Kessing, Automated assessment of psychiatric disorders using speech: A systematic review, *Laryngoscope Investig Otolaryngol* 31;5(1) (2020) 96–116.
- [4] D. Dubois, H. Prade, The mean value of a fuzzy number, *Fuzzy sets and systems* 24 (3) (1987) 279–300.
- [5] F. Eyben, F. Weninger, F. Gross, B. Schuller, Recent developments in opensmile, the munich open-source multimedia feature extractor, in: *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 835–838.
- [6] F. Eyben et al., K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, K. P. Truong, The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing, *IEEE Transactions on Affective Computing* 7 (2) (2016) 190–202.
- [7] E. Gibaja, S. Ventura, *Multilabel learning: A review of the state of the art and ongoing research*, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery.
- [8] A. Grünerbl, A. Muaremi, V. Osmani, Smartphone-based recognition of states and state changes in bipolar disorder patients, *IEEE Journal of Biomedical and Health Informatics* 19(1) (2015) 140–148.
- [9] F. Herrera, F. Charte, A. J. Rivera, M. J. del Jesus, *Multilabel Classification - Problem Analysis, Metrics and Techniques*, Springer, 2016.
- [10] O. Hryniewicz, K. Kaczmarek-Majer, Possibilistic aggregation of inhomogeneous streams of data, in: *IEEE International Conf. on Fuzzy Systems FUZZ-IEEE 2021*, 2021.
- [11] O. Hryniewicz, K. Kaczmarek-Majer, K. Opara, Control charts based on fuzzy costs for monitoring short autocorrelated time series 114 (2019) 166–181.
- [12] K. Kaczmarek-Majer, O. Hryniewicz, K. R. Opara, W. Radziszewska, A. Olwert, J. W. Owsinski, S. Zadrozny, Control charts designed using model averaging approach for phase change detection in bipolar disorder, in: *Uncertainty Modelling in Data Science*, Vol. 832 of *Advances in Intelligent Systems and Computing*, Springer International, 2019, pp. 115–123.
- [13] O. Kamińska, K. Kaczmarek-Majer, O. Hryniewicz, Acoustic feature selection with fuzzy clustering, self organizing maps and psychiatric assessments, *Proceedings of Information Processing and Management of Uncertainty in Knowledge-Based Systems, IPMU 2020*, Lisbon.
- [14] A. Rivolli, A. C. P. L. F. de Carvalho, The utiml Package: Multi-label Classification in R, *The R Journal* 10 (2) (2018) 24–37.
- [15] R. Wang, S. Ye, K. Li, S. Kwong, Bayesian network based label correlation analysis for multi-label classifier chain, *Information Sciences* 554 (2021) 256–275.
- [16] M.-L. Zhang, Z.-H. Zhou, A review on multi-label learning algorithms, *Knowledge and Data Engineering, IEEE Transactions on* 26 (2014) 1819–1837.