This project has received funding from SMALL GRANT SCHEME Call under grant agreement NOR/SGS/BIPOLAR/0239/2020-00.

Bipolar disorder prediction with sensor-based semi-supervised learning

D2.3 – Guidelines to BD feature preprocessing

Deliverable No.	D2.3	**Due date**	31-DEC-2022
Type	Report	**Dissemination Level**	Public
Version	1.0	**WP**	WP2
Description	Retrieval and preprocessing of features used in BD prediction		

.

Systems Research Institute Polish Academy of Sciences, Newelska 6, 01-447 Warsaw, Poland

Authors

Name	Email
Katarzyna Kaczmarek-Majer	k.kaczmarek@ibspan.waw.pl
Olga Kamińska	o.kaminska@ibspan.waw.pl
Kamil Kmita	k.kmita@ibspan.waw.pl
Jakub Małecki	j.malecki@ibspan.waw.pl
Izabella Zadrożna	izabella.zadrozna@ibspan.waw.pl

History

Date	Version	Change
30-SEP-2022	0.1	Task assignments and integrated version of the document
16-DEC-2022	0.2	Description of datasets added
22-DEC-2022	0.3	Version for internal review
31-DEC-2022	1.0	Version ready for submission

This project has received funding from SMALL GRANT SCHEME Call under grant agreement NOR/SGS/BIPOLAR/0239/2020-00.

Executive summary

This deliverable outlines the results of Task 2.1 activities dedicated to the feature retrieval and preprocessing. Scenarios defined as D1.1. describe the context for the features considered for this particular BD prediction problem.

First, we provide the main characteristics of the retrieved features. Next, we describe the preprocessing methods applied including noise reduction and aggregation of features (e.g., to a daily level). Different methods to aggregation are considered. This deliverable is closely related with the D.2.1. and D.2.2. (preliminary and final software component for preprocessing of features).

In the future, these features will be used for the feature selection and semi-supervised prediction, and tested in pilots.

This project has received funding from SMALL GRANT SCHEME Call under grant agreement NOR/SGS/BIPOLAR/0239/2020-00.

Table of Contents

This project has received funding from SMALL GRANT SCHEME Call under grant agreement NOR/SGS/BIPOLAR/0239/2020-00.

List of acronyms

Acronym	Explanation
BIPOLAR	Bipolar disorder prediction with sensor-based semi-supervised learning project
BD	Bipolar disorder

1.Introduction

Feature preprocessing turns raw data into a one that is usable by any models. It is a significant step that prepare mostly raw data into more readable, containing a dose of knowledge features. The quality of the model largely depends on the data that is fed into the model. Different types of feature preprocessing is required for different data types and different machine learning models. Some methods are used for numeric data and others for categorical. In that report we presenting first steps for cleaning huge dataset mostly containing acoustic features of voice.

2.Characteristics about features

Planned objective features consists of :

a. physical descriptors of voice

Our database consists of 86 descriptors called acoustic features with extracted values from ~20ms frames of voice signal derived with an openSMILE library. That set describes voice using different aspects such as time-domain descriptors (zero crossing rate), amplitude statistics, signal energy), spectral features (distribution of energy, mel-cepstral coefficients, fundamental frequency and its harmonics), voice quality (jitter, shimmer, harmonics to noise ratio) and prosodic features (voicing probability, normalized loudness);

Full list of acoustic features is attached at the end of that document in appendix A.

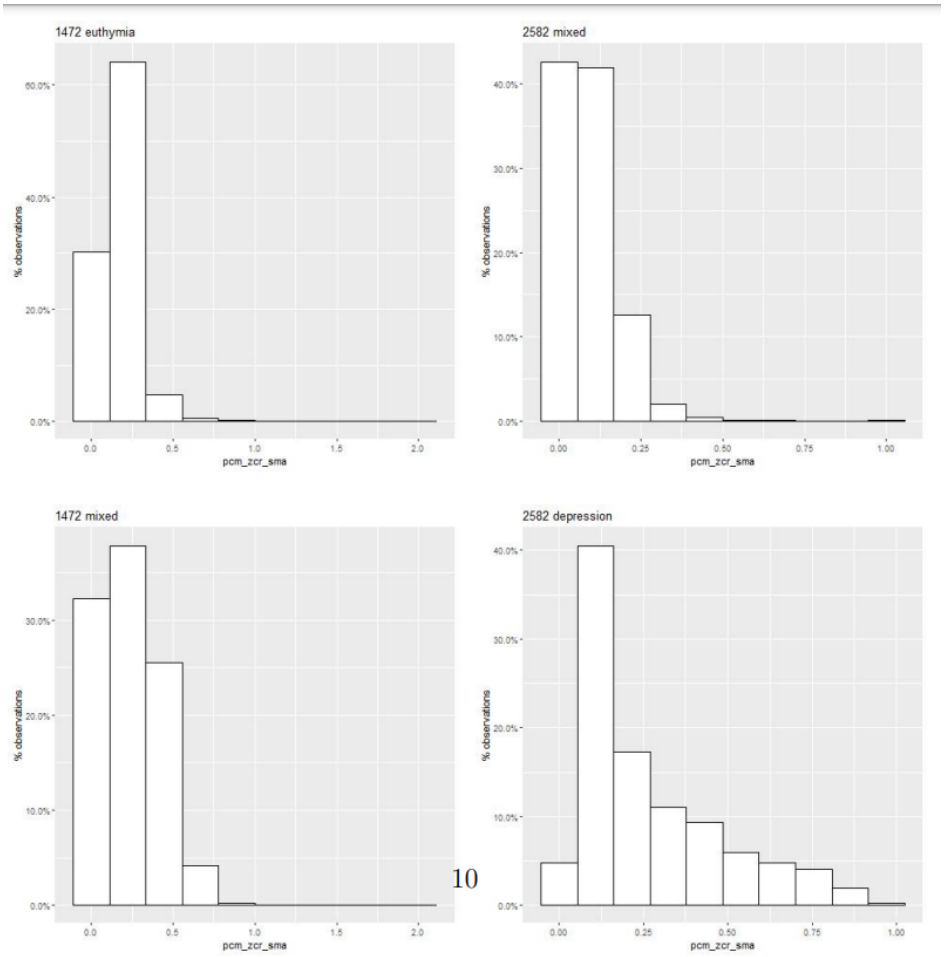
This project has received funding from SMALL GRANT SCHEME Call under grant agreement NOR/SGS/BIPOLAR/0239/2020-00.

Picture below presents a piece of database with raw data

dwh_mobilerecording_id bigint	chunk_number double precision	frame_nr double precision	pcm_LOGenergy_sma double precision	pcm_zcr_sma double precision	voiceprob_sma double precision	f0_sma double precision	f0env_sma double precision	pcm_fftMag_fband0-250_sma double precision	pcm_fftMag_fband0-650_sma double precision
4522	0	82	-22.25742	0.05	0.23492	0	241.406	4.84496e-10	1.598362e-09
4522	0	83	-22.04618	0.08833333	0.2350778	0	241.406	1.055203e-09	2.544545e-09
4522	0	84	-21.57874	0.135	0.3198095	0	241.406	1.006863e-08	4.284379e-08
4522	0	85	-20.64141	0.1466667	0.4988257	0	241.406	3.705783e-08	3.696641e-07
4522	0	86	-19.38575	0.1166667	0.7221381	0	241.406	5.458563e-08	1.090804e-06
4522	0	87	-18.50783	0.08166667	0.8415241	111.1111	249.0666	5.831678e-08	1.896679e-06
4522	0	88	-18.27293	0.075	0.8711537	222.2222	262.4727	4.765461e-08	2.248098e-06
4522	0	89	-18.44129	0.075	0.8582436	222.2222	275.8788	4.829306e-08	1.948288e-06
4522	0	90	-18.93485	0.07666667	0.8001075	111.1111	281.6242	5.066491e-08	1.274511e-06
4522	0	91	-19.66054	0.08833333	0.6828547	0	281.6242	4.422001e-08	6.560079e-07
4522	0	92	-20.56006	0.11	0.4999267	0	281.6242	2.860896e-08	2.451938e-07
4522	0	93	-21.30429	0.1483333	0.3234425	0	281.6242	1.384517e-08	7.381605e-08
4522	0	94	-21.6393	0.1666667	0.2117049	0	281.6242	4.16519e-09	1.538344e-08
4522	0	95	-21.76815	0.165	0.1721521	0	281.6242	1.114963e-09	4.653269e-09
4522	0	96	-21.65616	0.13	0.1822863	0	281.6242	8.351007e-10	3.91295e-09
4522	0	97	-21.75868	0.1033333	0.2123708	0	281.6242	6.24436e-10	2.734814e-09
4522	0	98	-21.80887	0.08333334	0.2415811	0	281.6242	5.382375e-10	2.784551e-09

One of the first steps in each exploratory data analysis process is preparing data visualization and data distribution.

Following charts present histogram of selected acoustic parameters for 2 different patients in different BD phase.



This project has received funding from SMALL GRANT SCHEME Call under grant agreement NOR/SGS/BIPOLAR/0239/2020-00.

Systems Research Institute Polish Academy of Sciences, Newelska 6, 01-447 Warsaw, Poland

Received distributions of values presented above indicate at first step of analysis that there could be some difference during patients BD phase moreover values differ for each of patient.

b. Statistics about text messages such as length of sent text messages, number of text messages that are retrieved from smartphone's keyboard;

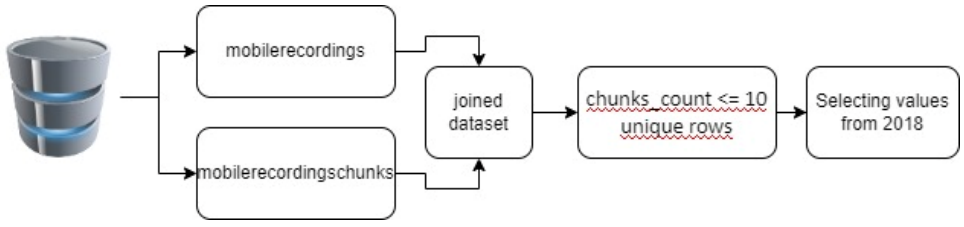
Picture below presents a piece of database with raw data with sms information

dwh_patient_id bigint	create_date timestamp without time zone	sent_date timestamp without time zone	sync_date timestamp without time zone	is_synced boolean	length double precision
11	2017-08-29 11:00:07.671	2017-08-28 19:05:25.037	2017-08-29 09:28:06.385	true	2
26	2017-08-31 17:22:55.079	2017-08-31 13:47:47.892	2017-08-31 15:20:39.867	true	63
26	2017-08-31 17:22:55.081	2017-08-31 13:36:23.248	2017-08-31 15:20:39.9	true	71
28	2017-08-31 22:13:38.592	2017-08-31 21:55:03.838	2017-08-31 20:11:21.051	true	17
26	2017-09-01 10:31:40.998	2017-08-31 23:40:15.045	2017-09-01 08:30:39.852	true	117
28	2017-09-01 18:31:03.198	2017-09-01 13:18:00.222	2017-09-01 16:28:38.162	true	9
33	2017-09-01 22:32:02.962	2017-09-01 21:30:38.689	2017-09-01 20:58:07.032	true	9
33	2017-09-01 22:32:02.964	2017-09-01 21:13:55.774	2017-09-01 20:58:07.069	true	55
33	2017-09-01 22:32:02.965	2017-09-01 21:12:56.755	2017-09-01 22:27:57.892	true	87
33	2017-09-01 22:32:02.965	2017-09-01 21:07:19.424	2017-09-01 22:27:57.906	true	6
33	2017-09-01 22:32:02.965	2017-09-01 21:06:47.95	2017-09-01 22:27:57.922	true	63
33	2017-09-02 04:00:01.504	2017-09-02 00:08:36.087	2017-09-02 02:27:44.413	true	3

3.Preprocessing

a. Acoustic data preprocessing

First step is to connect table **mobilerecordings** that consist of information about established connections with **mobilerecordingschunks** that consists of acoustic parameters in relation to previous one. It could be done using dwh_mobilerecording_id column. Preprocessing starts with selecting only those recordings that come from 2018. Regarding data could have missing's or incorrect parameters. Next step is to select acoustics parameters having number of chunks_count equal 10. This column explains number of divisions of the recording. Assumption of data collection were to record only first 5 minutes of calls – and these records could be included int that number of chunks. Then is important to select only unique rows - some repeated rows have been appeared.



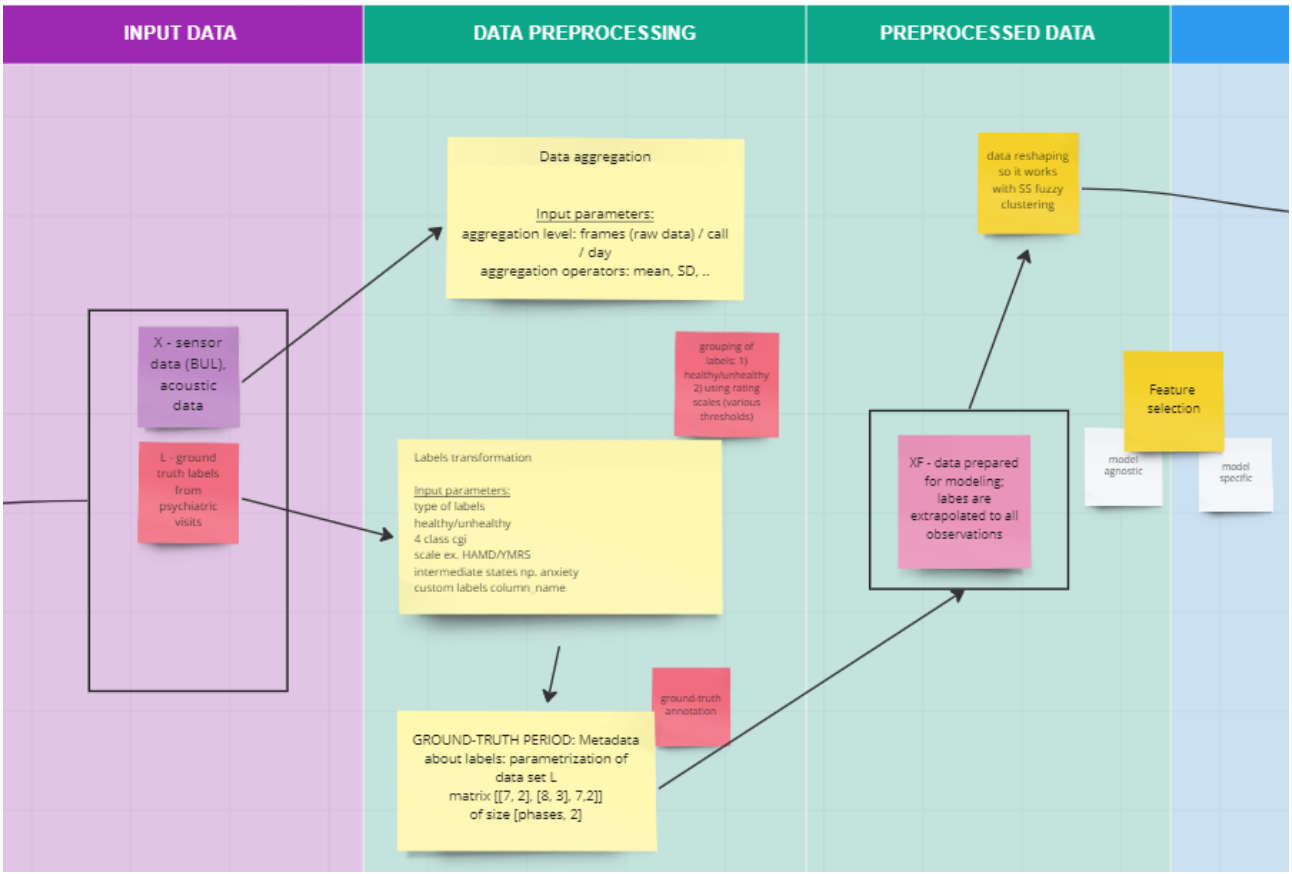
b. Labels preprocessing

This project has received funding from SMALL GRANT SCHEME Call under grant agreement NOR/SGS/BIPOLAR/0239/2020-00.

Only visits from 2018 are considered which have filled values describing hamilton and young scales.

4. Aggregation of features

During many team brainstroms, following ideas has been set as recommended.



Data aggregation can be done in several different ways. Data could be aggregated into one mobile calls, mobile calls in the time of day, mobile calls within one day or mobile calls from last period of time (ex. 3 days). Each of that approaches have some adventages. During that study we selected aggregating acoustic frames into recording coming from each mobile calls (mobilecall level).

Next step is to select appropriate methof of aggregation. Usyually most data agre aggregated with mean and standard deviation on selected data set. We are using additionaly following methods: median, skewness, 1st and 3rd quartiles to obtain more specific features dataset.

Moreover, labels transformation are another problem taking into account in that project. Patients received labels on each of visits. It is described using two nominal scales: Hamilton Depression Scale and Young Mania Scale. The sum up of both scale is transformed into label indicating current BD phase. Therefore label could be described as nominal value in particular

This project has received funding from SMALL GRANT SCHEME Call under grant agreement NOR/SGS/BIPOLAR/0239/2020-00.

scales or BD phase. Another solution is binominal labels scales which indicate that patitent is helathy (in stable phase) or unhealthy (any BD symptomps appear).

The ground truth of each labels is set to day of the vists. However that label could be extend into several days before – because most of symptomp occur some days befor visits. In literature there are some assumtions like extending ground truth period into 10 days (7 days before visit, day of visit and 2 days after visit). That soulution increase number of labelled data.

Package consists of following functions:

- transform_label_healthy_unhealthy – function that creates label with 2 possible values of labe: healthy or unhealthy
- transform_label_custom - function that creates label using specified column indicated by user
- transform_label_cgi - function that creates label with 4 possible values from CGI scale (depression, mania, mixed, euthymia)
- transform_label_hy - function that creates numeric label with sum of point from selected Hamilton/Young questionnaire
- transform_label_symptoms - function that creates 10 symptoms numeric labels based on specified points from hamilton/young scales
- extend_ground_truth_period – function that extend ground truth of label with user definition (example 7 days before visit and 2 day after visit)
- calculate_aggregates – function that return aggregated acoustic parameters

4.Future extensions

Features and functions described in this report can be further extended. In particular, imprecision about labeling can be included already at the feature preprocessing step. This task will be further investigated in the next months of the project.

This project has received funding from SMALL GRANT SCHEME Call under grant agreement NOR/SGS/BIPOLAR/0239/2020-00.

Appendix A

Full List of acoustic features:

A.1 Parameters connected with energy, spectral, cepstral

- pcm LOGenergy sma
- pcm fftMag fband0250 sma
- pcm fftMag fband0650 sma
- pcm fftMag spectralRollOff25.0 sma
- pcm fftMag spectralRollOff50.0 sma
- pcm fftMag spectralRollOff75.0 sma
- pcm fftMag spectralRollOff90.0 sma
- pcm fftMag spectralFlux sma
- pcm fftMag spectralCentroid sma
- pcm fftMag spectralMaxPos sma
- pcm fftMag spectralMinPos sma
- audspec lengthL1norm sma
- audspecRasta lengthL1norm sma
- pcm RMSenergy sma
- audSpec Rfilt sma[0] − audSpec Rfilt sma[25] //26 parameters
- pcm fftMag fband250650 sma
- pcm fftMag fband10004000 sma
- pcm fftMag spectralEntropy sma
- pcm fftMag spectralVariance sma
- pcm fftMag spectralSkewness sma
- pcm fftMag spectralKurtosis sma
- pcm fftMag psySharpness sma
- pcm fftMag spectralHarmonicity sma
- Loudness sma3
- alphaRatio sma3
- hammarbergIndex sma3
- slope0500 sma3
- slope5001500 sma3
- F1frequency sma3nz
- F1bandwidth sma3nz
- F1amplitudeLogRelF0 sma3nz
- F2frequency sma3nz
- F2amplitudeLogRelF0 sma3nz
- F3frequency sma3nz
- F3amplitudeLogRelF0 sma3nz
- pcm fftMag mfcc[0] − pcm fftMag mfcc[12] //13 parameters

This project has received funding from SMALL GRANT SCHEME Call under grant agreement NOR/SGS/BIPOLAR/0239/2020-00.

A.2 Parameters connected with sound source

- pcm zcr sma
- voiceProb sma
- F0 sma
- F0env sma
- F0final sma
- voicingFinalUnclipped sma
- jitterLocal sma
- jitterDDP sma
- shimmerLocal sma
- logHNR sma
- F0semitoneFrom27.5Hz sma3nz
- logRelF0-H1-H2 sma3nz
- logRelF0-H1-A3 sma3nz

This project has received funding from SMALL GRANT SCHEME Call under grant agreement NOR/SGS/BIPOLAR/0239/2020-00.