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

D2.4 – Initial features selected

Deliverable No.	D2.4	**Due date**	31-DEC-2022
Type	Report	**Dissemination Level**	Public
Version	1.0	**WP**	WP2
Description	Selection 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
30-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 selection.

This deliverable uses results from D2.3. and provides the main outcomes about the selected features.

In the future, these features will be used for semi-supervised prediction 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 selection is one of the important concepts of advanced data analysis and models building process. It is a way of automatically or manually selecting the subset of the most appropriate and relevant features from the original dataset by removing the redundant, irrelevant, or noisy features.

Generally, the dataset consists of noisy data, irrelevant data, and some part of useful data. Moreover, the huge amount of data also slows down the training process of the model, and with noise and irrelevant data, the model may not predict and perform well. So, it is very necessary to remove such noises and less-important data from the dataset and to do this, and Feature selection techniques are used. Fewer features can allow models to run more efficiently (less space or time complexity) and be more effective.

2.Related work about features selection methods for BD

Current researches indicate that there are several significant feature selection methods. Those methods could be divided into some groups: Wrapper Methods(a), Filter Methods(b) and Embedded Methods(c).[3][5]

a. In wrapper methodology, selection of features is done by considering it as a search problem, in which different combinations are made, evaluated, and compared with other combinations. It trains the algorithm by using the subset of features iteratively.
b. In Filter Method, features are selected on the basis of statistics measures. This method does not depend on the learning algorithm and chooses the features as a pre-processing step. The filter method filters out the irrelevant feature and redundant columns from the model by using different metrics through ranking. T
c. Embedded methods combined the advantages of both filter and wrapper methods by considering the interaction of features along with low computational cost. These are fast processing methods similar to the filter method but more accurate than the filter method.

Besides above methods there are some methods which explain the feature importance using mostly black-box models[1]. An example of that method is SHAP[4] - (SHapley Additive

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

exPlanations) which is a game theoretic approach to explain the output of any machine learning model. It connects optimal credit allocation with local explanations using the classic Shapley values from game theory and their related extensions.

3.Methodology applied

First Feature Selection Method that were applied to BIPOLAR project was RFE (Recursive Feature Selection). 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. The algorithm can wrap around any model (in our case we used Random Forest algorithm), and it produces the best possible set of features that gives the highest performance. To find the optimal number of features cross-validation is used with RFE algorithm to obtain the best scoring collection of features.

RFE were used on aggregated into one call labelled data with all of possible acoustic features. Results are presented in next chapter.

On the other hand, we used SHAP to explain acoustic features while using black box models. In that case, we used only few patients with labelled but not aggregated dataset. [1]

4.Results for acoustic data

Features selection method were tested with different approaches. First approach assumes that, we are developing that method for each patients separately. The reason of it is to verify that different patients have similar or not important features.

Next approach was implemented to compare different types of aggregation – due to aggregating all frames into one mobile recording we are implemented 6 types of aggregation (mean, standard deviation and 3 quartiles).

Last approach assume that label is transformed into 2 classes (healthy and unhealthy).

The obtained results are as follows:

a) Top 10 parameters for 3 different patients

ID	**Patient 1**	**Patient 2**	**Patient 3**
1	"f0_sma"	"pcm_fftmag_spectralharmonicity_sma_compare"	"f0env_sma"
2	"pcm_LOGenergy_sma"	"pcm_fftMag_mfcc_11_"	"pcm_fftMag_fband0_650_sma"
3	"audSpec_Rfilt_sma_compare_25_"	"loghnr_sma"	"pcm_fftMag_fband1000_4000_sma_compare"
4	"pcm_zcr_sma"	"pcm_fftMag_mfcc_4_"	"slope0_500_sma3"

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

5	"pcm_fftMag_fband1000_4000_sma_compare"	"jitterddp_sma"	"f2amplitudelogrelf0_sma3nz"
6	"audSpec_Rfilt_sma_compare_0_"	"voicingfinalunclipped_sma"	"f1frequency_sma3nz"
7	"audspec_lengthl1norm_sma"	"pcm_fftMag_mfcc_5_"	"alpharatio_sma3"
8	"pcm_fftMag_mfcc_11_"	"pcm_fftmag_spectralvariance_sma_compare"	"f1amplitudelogrelf0_sma3nz"
9	"f2amplitudelogrelf0_sma3nz "	"audSpec_Rfilt_sma_compare_25_"	"logRelF0_H1_A3_sma3nz"
10	"f1frequency_sma3nz"	"shimmerlocal_sma"	"pcm_fftmag_spectralflux_sma"

b) Top 10 parameters for mean and standard deviation

ID	**Mean**	**Standard Deviation**	**Skewness**
1	pcm_LOGenergy_sma	pcm_LOGenergy_sma	shimmerlocal_sma
2	f1amplitudelogrelf0_sma3nz	shimmerlocal_sma	pcm_fftMag_fband0-650_sma
3	f2amplitudelogrelf0_sma3nz	f3frequency_sma3nz	pcm_fftMag_mfcc_11_
4	f3frequency_sma3nz	pcm_fftMag_mfcc_11_	pcm_fftMag_spectralRollOff90_0_sma
5	hammarbergindex_sma3	pcm_fftMag_mfcc_3_	jitterlocal_sma
6	audspec_lengthl1norm_sma	pcm_fftmag_spectralharmonicity_sma_compare	loudness_sma3
7	f2frequency_sma3nz	audSpec_Rfilt_sma_compare_2_	pcm_fftmag_spectralflux_sma
8	pcm_fftMag_mfcc_3_	f1bandwidth_sma3nz	pcm_fftmag_spectralkurtosis_sma_compare
9	shimmerlocal_sma	f1frequency_sma3nz	pcm_fftmag_spectralminpos_sma
10	alpharatio_sma3	f2frequency_sma3nz	f0env_sma

c) Top 10 parameters depend on label (4class CGI vs 2class - healthy/unhealthy)

ID	**4 class CGI**	**2 class Healthy/unhealthy**
1	pcm_fftMag_mfcc_11_	audSpec_Rfilt_sma_compare_25_
2	f2frequency_sma3nz	loghnr_sma
3	f3frequency_sma3nz	pcm_fftMag_fband1000-4000_sma_compare
4	pcm_fftMag_spectralRollOff90_0_sma	pcm_fftMag_mfcc_11_
5	audSpec_Rfilt_sma_compare_25_	pcm_fftMag_mfcc_4_
6	hammarbergindex_sma3	pcm_fftmag_spectralharmonicity_sma_compare
7	pcm_fftMag_fband0-650_sma	audspec_lengthl1norm_sma
8	pcm_fftmag_spectralkurtosis_sma_compare	f1frequency_sma3nz
9	pcm_fftmag_spectralminpos_sma	f2amplitudelogrelf0_sma3nz
10	shimmerlocal_sma	pcm_fftMag_fband0-250_sma

All 3 tables above indicate that each time we get a different set of features. Some of them are often repeated.

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

Results received from RFE method are as follows:

Patient 1				
Variables	Accuracy	Kappa	AccuracySD	KappaSD
4	0,761	0,192	0,003	0,011
8	0,782	0,250	0,006	0,027
16	0,793	0,291	0,006	0,010
86	0,802	0,323	0,002	0,007

Patient 2				
Variables	Accuracy	Kappa	AccuracySD	KappaSD
4	0,538	0,138	0,011	0,026
8	0,614	0,293	0,005	0,010
16	0,636	0,328	0,004	0,008
86	0,650	0,345	0,003	0,006

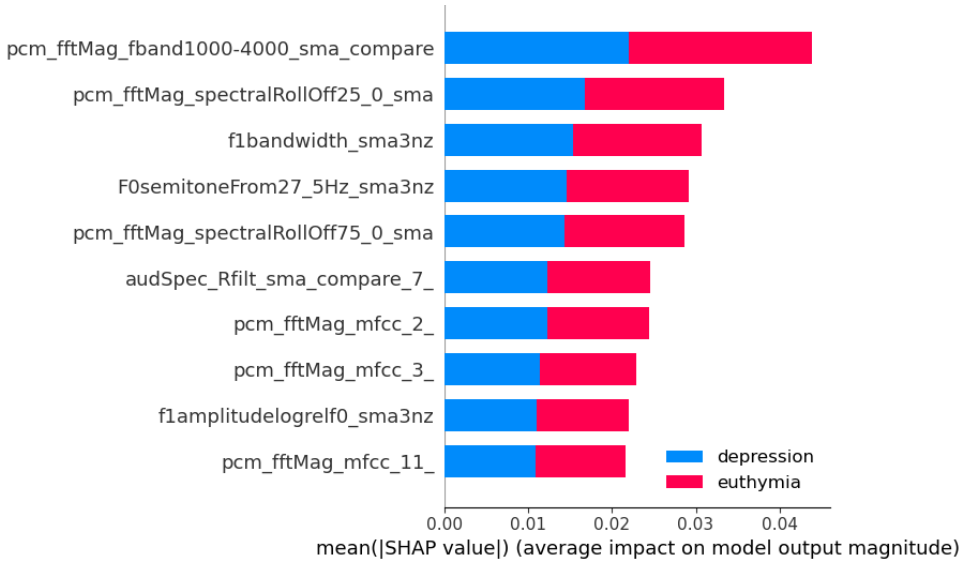
For both patients turned out that the best results are received when all 86 variables are taken into account. However, the difference in accuracy for smaller number of parameters is relatively small and for example, the accuracy with 8 parameters (reduction by over 90% of parameters) amounts to 78.2% and 61.4%, respectively. These results are very promising. Another coefficient called Kappa presented in both tables 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.

5.Results for SHAP

Firstly SHAP method were used during researches described in PLENARY. With interesting results, that method were implemented using aggregated data.
The SHAP value plot can further show the positive and negative relationships of the predictors / attributes with the target variable Red color means an increase while blue show decrease.

First graph ilustrates top 10 features with highest impact of attributes on the model prediction. Ranking following the descending Shapley values the higher the most influential here pcm_fftMag_fband1000-4000_sma_compare is the most influential for both BD states.

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



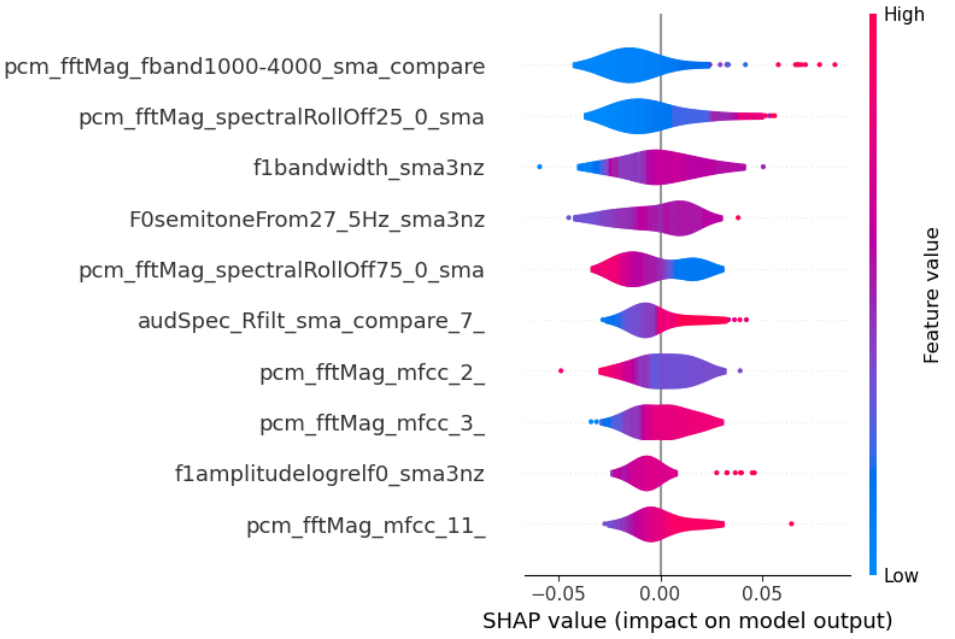
The SHAP value plot can further show the positive and negative relationships of the predictors / attributes with the target variable Red color means an increase while blue show decrease.

Next plots are made of all the dots in the train data. It demonstrates the following information:

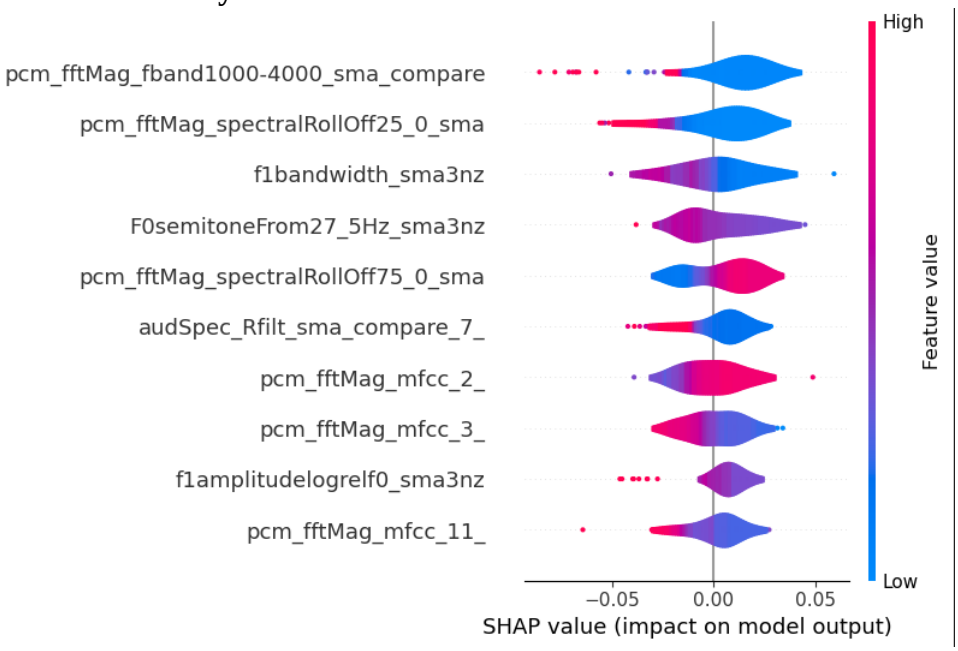
- Feature importance: Variables are ranked in descending order.
- Impact: The horizontal location shows whether the effect of that value is associated with a higher or lower prediction.
- Original value: Color shows whether that variable is high (in red) or low (in blue) for that observation.

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

Results for Depression:



Results for Euthymia:



In compering to previous important parameters, we can see that we will get a set of features that does not repeat in its entirety with the previous ones, but contains features already indicated as essential.

We received very interesting results. It is important that for each of approaches we received different list of best features.

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

6.References

[1] Kaczmarek-Majera K., Casalino G., Castellano G., Dominiak M., Hryniewicz O., Kaminska O., Vessio G., Dıaz-Rodrıguez N., PLENARY Explaining black-box models in natural language through fuzzy linguistic summaries, Information Sciences Volume 614, 374-399

[2] Kamińska O. Kaczmarek-Majer K. and Hryniewicz O. Acoustic features for prediction of state change in bipolar disorder. IPMU, Lisbon 2020

[3] Isabelle Guyon et al. "Gene selection for cancer classification using support vector machines". In: Machine learning 46.1-3 (2002), pp. 389–42

[4] Heuillet et al. "Collective eXplainable AI: Explaining Cooperative Strategies and Agent Contribution in Multiagent Reinforcement Learning with Shapley Values".In:IEEE Computational Intelligence Magazine, 17, 59–71, 2022

[5] B.Venkatesh et al. A review of feature selection and its methods, Cybernetics and Information Technologies, 2019

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