# Explaining black-box models in natural language through fuzzy linguistic summaries - Bipolar Disorder case study


OLGA KAMIŃSKA [1]

Katarzyna Kaczmarek-Majer[1], Gabriella Casalino[2], Giovanna Castellano[2],Monika Dominiak[3], Olgierd Hryniewicz[1], Gennaro Vessio[2], Natalia Diaz Rodriguez[4]

[1] Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland
[2] University of Bari Aldo Moro, Bari, Italy
[3] Institute of Psychiatry and Neurology, Warsaw, Poland
[4] University of Granada, Spain


BIPOLAR

http://bipolar.ibspan.waw.pl/

https://github.com/ITPsychiatry/plenary

## MOTIVATION

We consider the problem of supporting the diagnosis of bipolar disorder (BD) state through the analysis of acoustic data from phone calls. Some progress has been made in the treatment of BD over the past decade; nevertheless, the diagnosis and monitoring of this disorder remains challenging. This is probably due to the still limited understanding of the nature of the disease and, consequently, the difficulty in predicting relapses. One issue that has received attention recently is a fundamental one: the classification of BD episodes [5]

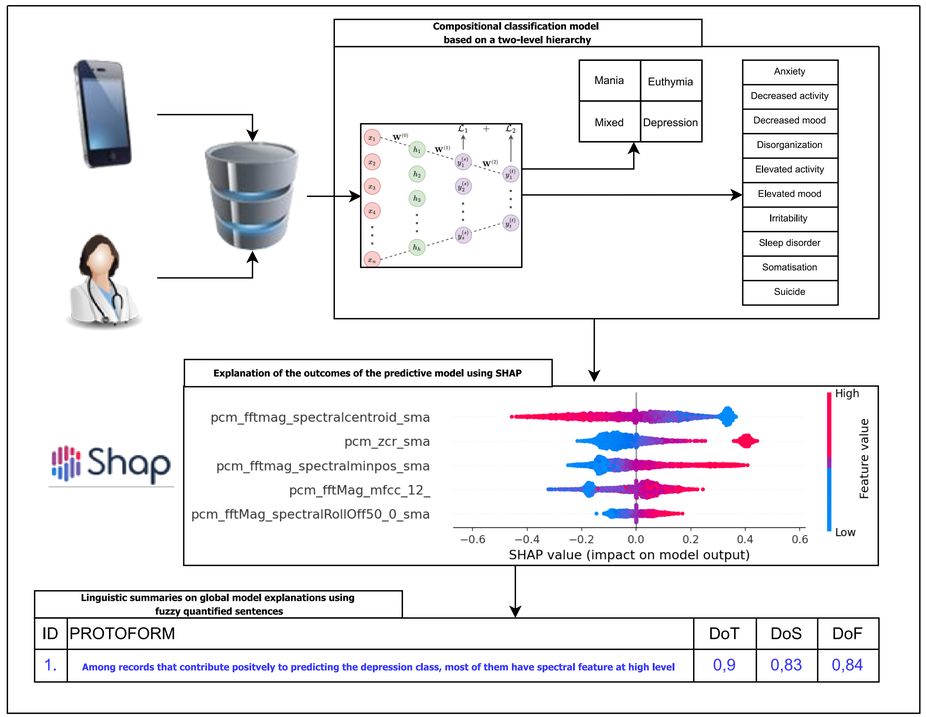
- The needs for explaining relations between attributes, symptoms, and states
- The needs for explaining high-level acoustic attributes of speech

DO BIPOLAR PATIENTS SPEAK LOUDER WHEN IN DEPRESSIVE STATE?

## LINGUISTIC SUMMARIES

- Fuzzy linguistic summaries (LSs) are statements in natural language that describe numerical datasets[6]
- LSs have been confirmed as human-consistent information granules with applications in various domains.
- The main purpose of summarization is usually to improve comprehension of large datasets

## METHODOLOGY




## CONCLUSIONS

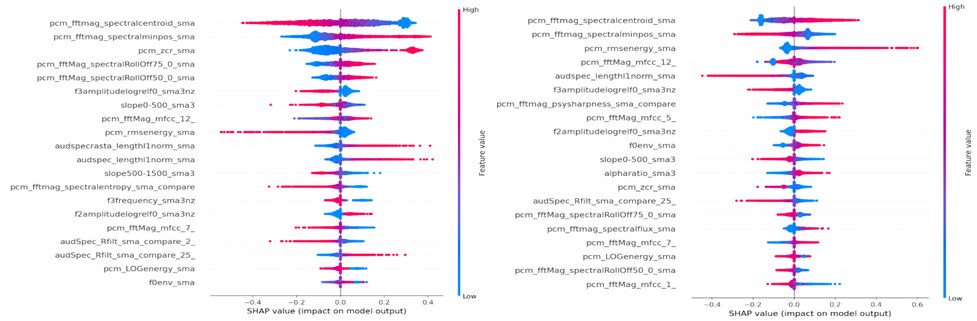
- Grouping of low-level attributes into high-level information granules using linguistic summarization **improves the over all explain ability of the model results**.
- Experimental evaluations confirmed that fuzzy linguistic summarization complements global model explanations derived from the popular SHAP tool.
- Furthermore, the results demonstrate that improves understanding of model outputs by appropriate incorporation of the domain knowledge.
- The introduction of specialist knowledge in the form of middle-layer labels does not affect performance in terms of prediction accuracy (it remains at a comparable level); however, the inclusion of this knowledge improves the understanding of the model outputs.

## RESULTS

### BLACKBOX MODELS RESULTS

| Method | Class | Precision | Recall | F1-score |
|---|---|---|---|---|
| MLP Baseline | 0 (Euthymia) | 0.83 | 0.80 | 0.82 |
| | 1 (Depression) | 0.60 | 0.67 | 0.63 |
| | 2 (Mania) | 0.79 | 0.01 | 0.03 |
| | 3 (Mixed state) | 0.70 | 0.70 | 0.70 |
| | **Accuracy** | | | 0.72 |
| Multitask Sequential Compositional MLP | 0 (Euthymia) | 0.83 | 0.80 | 0.81 |
| | 1 ( Depression) | 0.59 | 0.68 | 0.63 |
| | 2 (Mania) | 0.78 | 0.02 | 0.03 |
| | 3 (Mixed state) | 0.71 | 0.68 | 0.69 |
| | **Accuracy** | | | 0.72 |

### SHAP RESULTS




(Left): 20 most contributing features to the baseline MLP model for depression, (Right) 20 most contributing features to the sequential and compositional MLP model for depression

### LINGUISTIC SUMMARIES RESULTS

| Id | LS description | DoT | DoS | DoF |
|---|---|---|---|---|
| 401 | Among records that contribute positively to predicting decreased activity, most of them have spectral-related features at low level. | 0.81 | 0.26 | 0.31 |
| 402 | Among records that contribute against predicting decreased activity, most of them have quality-related features at low level. | 0.45 | 0.67 | 0.76 |
| **403** | **Among records that contribute positively to predicting decreased activity, most of them have quality-related features at high level.** | 0.25 | 0.18 | 0.31 |
| 404 | Among records that contribute around zero to predicting elevated activity, most of them have pitch-related features at medium level. | 1.00 | 0.05 | 0.03 |
| **405** | **Among records that contribute positively to predicting elevated activity, most of them have pitch-related features at low level.** | 0.29 | 0.30 | 0.31 |
| **406** | **Among records that contribute positively to predicting elevated activity, most of them have spectral-related features at high level** | 0.95 | 0.26 | 0.31 |
| **407** | **Among records that contribute against predicting elevated activity, most of them have quality-related features at high level** | 0.26 | 0.63 | 0.76 |

### LINGUISTIC SUMMARIES EVALUATION

Evaluation of the quality of the group of LS sentences in terms of explanation quality and causability based on the System Causability Scale (SCS) questionnaire(the mean SCS score is computed as the sum of the avarage values of the 10 questions divided by 50) and Grice's maxims with Likert scale ratings (1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, 5 = strongly agree)

| Questionnaire | Domain expert evaluation |
|---|---|
| **System Causability Scale statement** | |
| SCS1. I found that the data included all relevant known causal factors with sufficient precision and granularity | 2 |
| SCS2. I understood the explanations within the context of my work | 4 |
| SCS3. I could change the level of detail on demand | 1 |
| SCS4. I did not need support to understand the explanations | 4 |
| **Mean SCS score** (on a [0.2, 1] range): | **0.7** |
| **Grice's Maxims** | |
| GM1. The group of sentences provides all the information we need, and no more (maxim of *quantity*) | 4 |
| GM2. The group of sentences provides truthful statements and avoids providing information not supported by evidence (maxim of *quality*) | 5 |
| GM3. The group of sentences is relevant to the discussion objective of explaining the model (maxim of *relation*) | 5 |
| **Mean Grice's maxims rating** (on a 1–5 Likert scale): | **4.25** |

### References


[1] K. Kaczmarek-Majer et al. PLENARY: Explaining black-box models in natural language through fuzzy linguistic summaries 2022 Information Sciences 614, p. 374-399.
[2] F.Eyben et al. - The Munich Versatile and Fast Open-Source Audio Feature Extractor MM'10 - Proceedings of the ACM Multimedia 2010 International Conference. 1459-1462
[3] Alejandro Barredo Arrieta and et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI Information Fusion 58 (2020) 82–115
[4] M. D. Pelaez-Aguilera et al. Fuzzy linguistic protoforms to summarize heart rate streams of patients with is chemic heart disease Complexity 2019
[5] A.Z. Antosik-Wojcinska, et al., 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
[6] J. Kacprzyk, R. R. Yager, and J. M. Merigo (2019) Towards human-centric aggregation via ordered weighted aggregation operators and linguistic data summaries: A new perspective on zadeh's inspirations," IEEE Computational Intelligence Magazine, vol. 14, no. 1, pp. 16–30


WiML Women in Machine Learning

NEURAL INFORMATION PROCESSING SYSTEMS