Explaining machine learning models in natural language with fuzzy linguistic summaries

Katarzyna Kaczmarek-Majer

[1]Systems Research Institute Polish Academy of Sciences, Warsaw, Poland

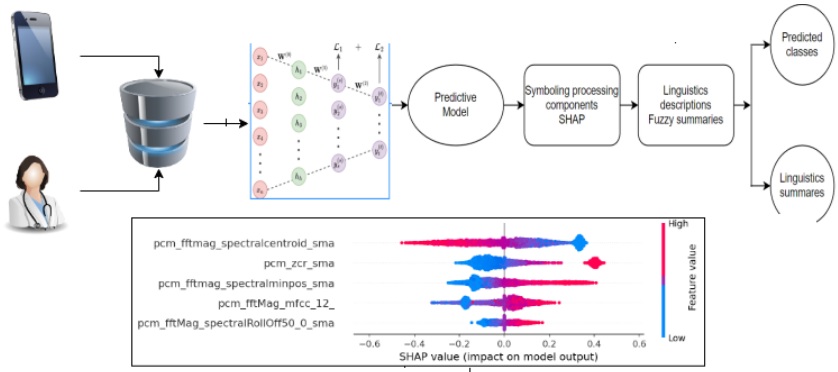
[2] University of Ostrava, Institute for Research and Applications of Fuzzy Modeling, Ostrava, Czech Republic

k.kaczmarek@ibspan.waw.pl

V European Summer School on Fuzzy Logic and Applications

Toledo, Spain, September 2nd - 6th 2024

Materials available here: **https://github.com/kasiakaczmarek/SFLA24-LS**



Among records that contribute positively to predicting the depression class, most of them have high spectral features

[1] Katarzyna Kaczmarek-Majer, Gabriella Casalino, Giovanna Castellano, Monika Dominiak, Olgierd Hryniewicz, Olga Kamińska, Gennaro Vessio, Natalia Diaz-Rodriguez, PLENARY: Explaining black-box models in natural language through fuzzy linguistic summaries, Information Sciences, Volume 614, 2022, Pages 374-399

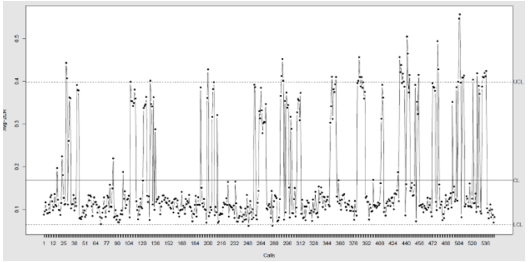
Outline

1. Motivations
2. Fuzzy linguistic summaries
3. **Hands-on data**: Describing dataset collected from smartphone-based mental health monitoring
4. PLENARY: Explaining black-box models in natural language through fuzzy linguistic summaries
5. **Hands-on data + Homework**: Explaining models for classification of bipolar disorder
6. Conclusions and Ongoing work

Outline

1. Motivations
2. Fuzzy linguistic summaries
3. **Hands-on data**: Describing dataset collected from smartphone-based mental health monitoring
4. PLENARY: Explaining black-box models in natural language through fuzzy linguistic summaries
5. **Hands-on data** + ~~**Homework**~~**Challenge**: Explaining models for classification of bipolar disorder
6. Conclusions and Ongoing work

Motivations



https://www.healthline.com/health-news/does-insulin-resistance-cause-fibromyalgia

- **Intelligent systems for the medical domain** are expected to explain why and how outcomes were generated.
- **The need for explanations** is of utmost importance not only due to various regulations, but also to increase trust among users.

As extracted from the Cambridge Dictionary of English Language:

an explanation - *the details or reasons that someone gives to make something clear or easy to understand.*

Definition

Given a certain audience, explainability refers to the details and reasons a model gives to make its functioning clear or easy to understand.

[2] E. Walter, Cambridge advanced learner's dictionary, Cambridge University Press, 2008

[3] A.B. Arrieta et al, Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, Inform. Fusion 58 (2020) 82–115.

- Interpretability: level of understanding how the underlying technology works

  pause

- Explainability: level of understanding how the AI-based system came up

  with a given result

Explanation is textual or visual artifact that provide a qualitative understanding of the relationship between the instance's components (e.g. words in text, patches in an image) and the model's prediction[5]

---

[4] ISO/IEC TR 29119-11:2020, Software and systems engineering, Software testing, Part 11: Guidelines on the testing of AI-based systems. ISO. 2020. Retrieved 25 January 2024

[5] MT Ribeiro, S Singh, C Guestrin (2016) "Why should I trust you?"Explaining the predictions of any classifier Proceedings of the 22nd ACM Knowledge Discovery and Data Mining (ACM KDD)

- Interpretability: level of understanding how the underlying technology works

  pause
- Explainability: level of understanding how the AI-based system came up

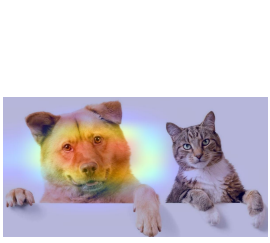
  with a given result

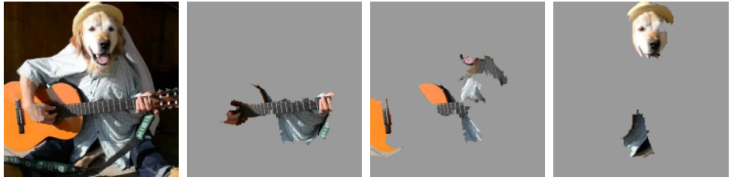
Explanation is textual or visual artifact that provide a qualitative understanding of the relationship between the instance's components (e.g. words in text, patches in an image) and the model's prediction[5]

[4] ISO/IEC TR 29119-11:2020, Software and systems engineering, Software testing, Part 11: Guidelines on the testing of AI-based systems. ISO. 2020. Retrieved 25 January 2024

[5] MT Ribeiro, S Singh, C Guestrin (2016) "Why should I trust you?"Explaining the predictions of any classifier Proceedings of the 22nd ACM Knowledge Discovery and Data Mining (ACM KDD)

- For images, various XAI techniques provide explanations in the form of visual descriptions (plots, heatmaps, etc.), e.g., GRAD-CAM [6].





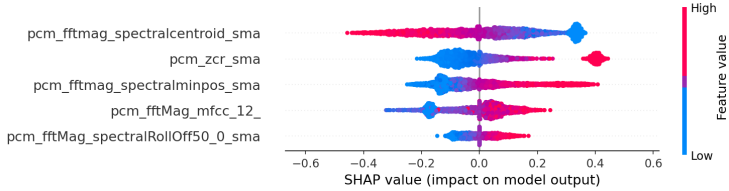
(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception neural network. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

[6] Selvaraju et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization.â€ International Journal of Computer Vision 2019

[7] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In Computer Vision and Pattern Recognition (CVPR), 2015.

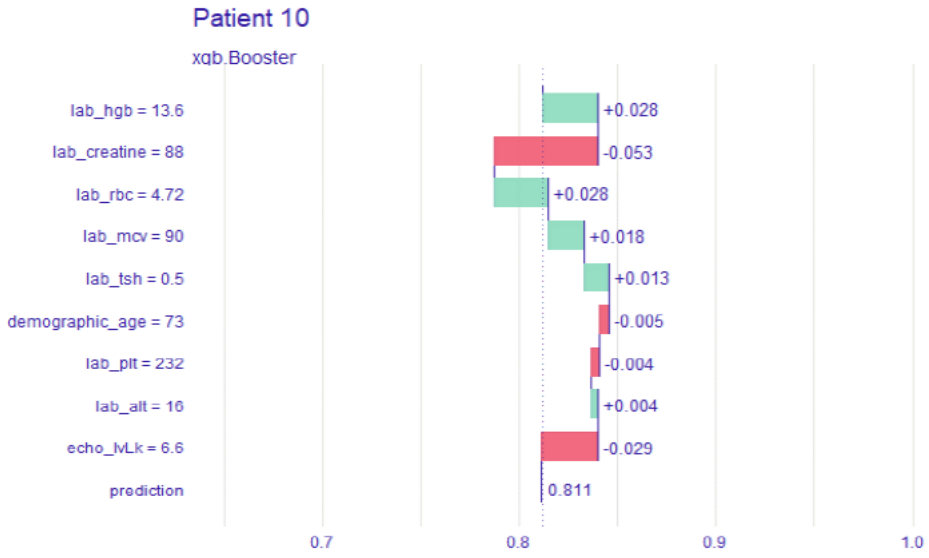
- For tabular data, **various post hoc XAI techniques** aim to explain the outputs of models that are not interpretable by design, e.g.:
  1. **LIME** (Local Interpretable Model-agnostic Explanations) which explains the predictions of any classifier by computing importance scores of features
  2. SHapley Additive exPlanations (**SHAP**)



[8] M.T. Ribeiro, S. Singh, C. Guestrin, why should I trust you? explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135-1144.

[9] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances in neural information processing systems 30.

[10] A.B. Arrieta et al, Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, Inform. Fusion 58 (2020) 82-115.


Patient 10
xgb.Booster
lab_hgb = 13.6
lab_creatine = 88
lab_rbc = 4.72
lab_mcv = 90
lab_tsh = 0.5
demographic_age = 73
lab_plt = 232
lab_alt = 16
echo_lvLk = 6.6
prediction
+0.028
-0.053
+0.028
+0.018
+0.013
-0.005
-0.004
+0.004
-0.029
0.811
0.7
0.8
0.9
1.0

Primary goal: **explanations as statements in natural language**

- Different explanations may be deduced depending on the background users' knowledge
- Post-hoc plots and low-level attributes are often **difficult and not intuitive for interpretation for users, especially for the non-technicians**.
- Our proposal:

  PLENARY for explaining black-box models in natural language through fuzzy linguistic summaries.

  Among records that contribute positively to predicting the depression class, most of them have spectral features at a high level

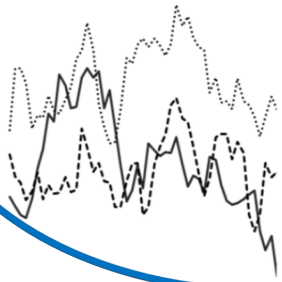
Fuzzy linguistic summaries

Index	my	chainage	interpolatedProfile	zs1	zu1
2018-09-05 15:40:46.2640…	104.3468	0.02445777778	-0.01296	-0.01296	-0.01296
2018-09-05 15:40:46.2680…	104.2044	0.04892888889	-0.01241131886	-0.01295988898	-0.01291856
2018-09-05 15:40:46.2720…	104.062	0.0734133			
2018-09-05 15:40:46.2760…	104.1746	0.0979111			
2018-09-05 15:40:46.2840…	104.3998	0.1469466			
2018-09-05 15:40:46.2880…	104.5124	0.1714844444			
2018-09-05 15:40:46.2920…	104.625	0.1960355556	-0.00880979466	-0.01282968272	
2018-09-05 15:40:46.2960…	104.6774	0.2206	-0.00848898815…	-0.01277240723	
2018-09-05 15:40:46.3040…	104.7822	0.2697688889	-0.00792	-0.01263691187	
2018-09-05 15:40:46.3080…	104.8345	0.2943733333	-0.00785950472…	-0.01256140579	
2018-09-05 15:40:46.3120…	104.887	0.3189911111	-0.00791084954…	-0.01248210238	
2018-09-05 15:40:46.3160…	104.992	0.3436222222	-0.00805341934…	-0.01239933024	-0.00
2018-09-05 15:40:46.3240…	105.202	0.3929244444	-0.00853	-0.01222656868	-0.00853
2018-09-05 15:40:46.3280…	105.307	0.4175955556	-0.00882696682…	-0.01213880964	-0.00867077
2018-09-05 15:40:46.3320…	105.412	0.44228	-0.00915792715…	-0.01205132401	-0.00900094293

SERIES A: 949009, 898828, 900
1 572 756, 1 486 238, 2 227 76

RAW DATA

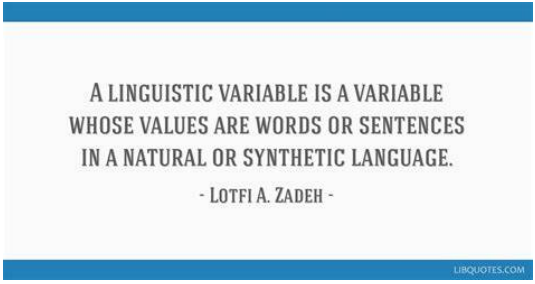
HUMAN-CONSISTENT INFORMATION



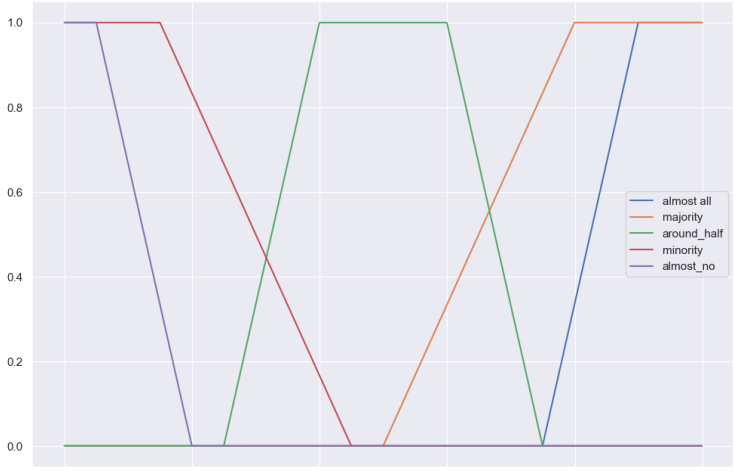
Most weekly segments of series A have small variance.

[Recently,] most time series have decreasing trend.

Zadeh's computing with words (CWW) paradigm as a tool to represent knowledge represent and reason in intelligent systems.



[11] Computing with Words: Principal Concepts and Ideas January 2012 Studies in Fuzziness and Soft Computing 277


1.0
0.8
0.6
0.4
0.2
0.0
almost all
majority
around_half
minority
almost_no

- **Fuzzy linguistic summaries** (LSs) are statements in natural language that describe numerical datasets.
- LSs based on Yager's protoform 'Q R y's are P' have been confirmed as **human-consistent** information granules.

Most young people are tall.

Few young people are tall.

Most young people are short

Most calls with high loudness in mania have low spectrum

The jump height achieved is lower since phase 1 is extended in time. The jump height achieved is lower since the first maximum is much greater than the second one in phase 3. It represents an excessive lowering of the centerof gravity.

[11] 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

[12] J. Moreno-Garcia, J. Abian - Vicen, L. Jimenez-Linares, L. Rodriguez-Benitez, Description of multivariate time series by means of trends characterization in the fuzzy domain, Fuzzy Sets and Systems 285 (2016) 118 - 139.

Related work: Information granules

- **Frequent patterns** e.g., If A happens before B and in the meantime we do not observe C, then it is a failure of class X[13].
- **Linguistic descriptions** e.g., Before the knee lesion, the gait quality is high because the gait symmetry is medium and the gait homogeneity is high[14].

[13] F. Hoeppner, S. Peter, and M. Berthold, Enriching multivariate temporal patterns with context information to support classification,Computational Intelligence in Intelligent Data Analysis. vol. 445, pp. 195-206, 2013.

[14] Gonzalez-Villanueva, L., Alvarez-Alvarez, A., Ascari, L., and Trivino, G. (2014). A tool for linguistic assessment of rehabilitation exercises. Applied Soft Computing Journal, 14:120 131.

- **Fuzzy Association Rules** e.g., IF Strength of Seasonality is Small AND Coefficient of Variation is Roughly Small THEN Weight of the j-th method is Big[15].
- **with Intermediate Quantifiers**[16] e.g., If tone of central banks news is medium and central banks are medium experienced in inflation targeting then customers expect small inflation support 0.24 and confidence 0.78.[17].

[15] M. Burda, M. Stepnicka, and L. Stepnicková, Fuzzy rule-based ensamble for time series prediction: Progresses with associations mining, in Strength. Links Between Data Analysis and Soft Computing, vol. 315, pp. 261-271, Springer, 2014

[16] Novák V. and Murinová P.: A formal model of the intermediate quantifiers "A few, Several, A little" 2019

[17] Association Rules on Data using Intermediate Quantifiers, Murinová Petra, Karel Fiala, Katarzyna Kaczmarek-Majer and Aleksandra Rutkowska, FSTA 2024)

Let $O = \{o_1, o_2, \ldots, o_{N_t}\}$ denote a set of objects, $\mathcal{A} = \{a_1, a_2, \ldots, a_r\}$ is a set of attributes that describe the characteristics of objects. The linguistic term set $l_{a_i} = \{l_1^{a_i}, \ldots, l_{k_{a_i}}^{a_i}\}$ is defined for each attribute from $\mathcal{A}$. A **linguistic summary** ($\mathcal{LS}$) based on an extended protoform in the sense of Yager and Kacprzyk is defined as:

$$\textbf{LS = LS(Q,R,P) = Q R objects O are P [T]} \tag{1}$$

having the quantifier Q, the qualifier R, the summarizer P , and $T \in [0, 1]$ measuring the validity of the sentence.

---

[18] J. Kacprzyk, R. R. Yager, and S. Zadrozny (2000) A fuzzy logic based approach to linguistic summaries of databases, Journal of Applied Mathematics and Computer Science

[19] J. Kacprzyk, R. R. Yager, J. M. Merigo, Towards human-centric aggregation via ordered weighted aggregation op- erators and linguistic data summaries: A new perspective on zadeh's inspirations, IEEE Computational Intelligence Magazine 14 (1) (2019) 16-30

Notation

Each attribute A_i is a function $A_i \colon Y \to X_i,\ i = 1, \ldots, m$, where $X_i \neq \emptyset$ of real numbers (e.g. observation x, ages, salaries).

The set $L(Y, A_i)$ of linguistic term sets for a given $i \in \{1, \ldots, m\}$ and $L(Y, A_i) = \{l_1^{A_i}, \ldots, l_{k_i}^{A_i}\}$ (enables the formulation of summaries in natural language).

Let $k_i = |L(Y, A_i)|$ for given $i \in \{1, \ldots, m\}$.

For example, for A_1 ='observation x',$L(Y, A_1)$ ={'low', 'medium', 'high'}.

Let us define a family $\mathcal{V}$ in the following way

$$\mathcal{V} = \{V_{i,k} \mid V_{i,k} \colon A_i(Y) \to [0, 1], i = 1, \ldots, m,\ k = 1, \ldots, k_i\} \tag{2}$$

Notation

Each attribute A_i is a function $A_i \colon Y \to X_i,\ i = 1, \ldots, m$, where $X_i \neq \emptyset$ of real numbers (e.g. observation x, ages, salaries).

The set $L(Y, A_i)$ of linguistic term sets for a given $i \in \{1, \ldots, m\}$ and $L(Y, A_i) = \{l_1^{A_i}, \ldots, l_{k_i}^{A_i}\}$ (enables the formulation of summaries in natural language).

Let $k_i = |L(Y, A_i)|$ for given $i \in \{1, \ldots, m\}$.

For example, for A_1 ='observation x',$L(Y, A_1)$ ={'low', 'medium', 'high'}.

Let us define a family $\mathcal{V}$ in the following way

$$\mathcal{V} = \{V_{i,k} \mid V_{i,k} \colon A_i(Y) \to [0, 1], i = 1, \ldots, m,\ k = 1, \ldots, k_i\} \tag{2}$$

Notation

Each attribute A_i is a function $A_i \colon Y \to X_i,\ i = 1, \ldots, m$, where $X_i \neq \emptyset$ of real numbers (e.g. observation x, ages, salaries).

The set $L(Y, A_i)$ of linguistic term sets for a given $i \in \{1, \ldots, m\}$ and $L(Y, A_i) = \{l_1^{A_i}, \ldots, l_{k_i}^{A_i}\}$ (enables the formulation of summaries in natural language).

Let $k_i = |L(Y, A_i)|$ for given $i \in \{1, \ldots, m\}$.

For example, for A_1 ='observation x',$L(Y, A_1)$ ={'low', 'medium', 'high'}.

Let us define a family $\mathcal{V}$ in the following way

$$\mathcal{V} = \{V_{i,k} \mid V_{i,k} \colon A_i(Y) \to [0, 1], i = 1, \ldots, m,\ k = 1, \ldots, k_i\} \tag{2}$$

Notation

Each attribute A_i is a function $A_i \colon Y \to X_i,\ i = 1, \ldots, m$, where $X_i \neq \emptyset$ of real numbers (e.g. observation x, ages, salaries).

The set $L(Y, A_i)$ of linguistic term sets for a given $i \in \{1, \ldots, m\}$ and $L(Y, A_i) = \{l_1^{A_i}, \ldots, l_{k_i}^{A_i}\}$ (enables the formulation of summaries in natural language).

Let $k_i = |L(Y, A_i)|$ for given $i \in \{1, \ldots, m\}$.

For example, for A_1 ='observation x',$L(Y, A_1)$ ={'low', 'medium', 'high'}.

Let us define a family $\mathcal{V}$ in the following way

$$\mathcal{V} = \{V_{i,k} \mid V_{i,k} \colon A_i(Y) \to [0,1], i = 1, \ldots, m,\ k = 1, \ldots, k_i\} \tag{2}$$

Notation

Each attribute A_i is a function $A_i \colon Y \to X_i,\ i = 1, \ldots, m$, where $X_i \neq \emptyset$ of real numbers (e.g. observation x, ages, salaries).

The set $L(Y, A_i)$ of linguistic term sets for a given $i \in \{1, \ldots, m\}$ and $L(Y, A_i) = \{l_1^{A_i}, \ldots, l_{k_i}^{A_i}\}$ (enables the formulation of summaries in natural language).

Let $k_i = |L(Y, A_i)|$ for given $i \in \{1, \ldots, m\}$.

For example, for A_1 ='observation x',$L(Y, A_1)$ ={'low', 'medium', 'high'}.

Let us define a family $\mathcal{V}$ in the following way

$$\mathcal{V} = \{V_{i,k} \mid V_{i,k} \colon A_i(Y) \to [0, 1], i = 1, \ldots, m,\ k = 1, \ldots, k_i\} \tag{2}$$

Definition

Fuzzy linguistic summary $S = (Y, A, \mathcal{P}, \mathcal{R}, Q)$ of the form 'Q R y's are P' is a 5-tuple such that

- Y is a set of objects,
- A is a set of attributes,
- $\mathcal{P} \subset \mathcal{V}$ is a family of summarizers, $\mathcal{R} \subset \mathcal{V}$ is a family of qualifiers, and they satisfy

$$\exists_{i \in \{1,\ldots,m\}} \exists_{k \in \{1,\ldots,k_i\}} V_{i,k} \in \mathcal{P} \Rightarrow \forall_{k \in \{1,\ldots,k_i\}} V_{i,k} \notin \mathcal{R} \tag{3}$$

$$\exists_{i \in \{1,\ldots,m\}} \exists_{k \in \{1,\ldots,k_i\}} V_{i,k} \in \mathcal{R} \Rightarrow \forall_{k \in \{1,\ldots,k_i\}} V_{i,k} \notin \mathcal{P} \tag{4}$$

- $Q\colon B \to [0,1]$ is a linguistic quantifier and $B \in \{\mathbb{R}^+, [0,1]\}$.

Let $\mathcal{T} = \{T_j | T_j : \mathcal{S} \to [0,1]\}$ be the family of functions, where $\mathcal{S}$ is a family of fuzzy summaries S.

$\mathcal{T}$ is the **j-tuple of evaluative criteria** of linguistic summary S.

Truth of the summary

$$T_1(S) = \mathcal{Q}\left(\frac{\sum_{i=1}^{n} \mathcal{P}(x_i) \star \mathcal{R}(x_i)}{\sum_{i=1}^{n} \mathcal{R}(x_i)}\right), \tag{5}$$

where S is a fuzzy linguistic summary, $\star$ is a t-norm.

[12] L. Zadeh, A computational approach to fuzzy quantifiers in natural languages, Comput. Math. Appl. 9(1) (1983) 149 - 184.

The functions used for intersection and union of fuzzy sets are called respectively: t-norms and t-conorms.

The function $\mathbf{t} : [0,1] \times [0,1] \to [0,1]$ is called a *triangular norm*, or shortly *t-norm* if for every $a, b, c, d \in [0,1]$ it has a following properties:

(T1) $a \; \mathbf{t} 1 = a$ (***Identity Law***)

(T2) $a \; \mathbf{t} \; b = b \; \mathbf{t} \; a$ (***Commutativity***)

(T3) $a \; \mathbf{t} \; b \leq c \; \mathbf{t} \; d$, when $a \leq c$ and $b \leq d$ (***Monotonicity***)

(T4) $a \; \mathbf{t} \; (b \; \mathbf{t} \; c) = (a \; \mathbf{t} \; b) \; \mathbf{t} \; c$ (***Associativity***)

If function **s** : $[0,1] \times [0,1] \to [0,1]$ meets the properties **(T2)** - **(T4)** from definition of t-norm and additionally for $a \in [0,1]$:

(T5) a **s** $0 = a$ (***Identity Law***),

it is called a *triangular conorm*, or shortly: *t-conorm*.

Based on properties **(T5)**, **(T2)** and **(T3)** we have:

1 **s** $1 = 0$ **s** $1 = 1$ **s** $0 = 1$, 0 **s** $0 = 0$.

For $a, b \in [0, 1]$ the following t-norms and t-conorms examples are in use:

- t-norm minimum $\wedge$, called also a standard intersection: $a \wedge b = min(a, b)$,
- t-conorm maximum: $\vee$, called also a standard union: $a \wedge b = max(a, b)$,
- algebraic product $a \ \mathbf{t}_a \ b = ab$,
- algebraic sum: $a \ \mathbf{s}_a \ b = a + b - ab$,
- Łukasiewicz t-norm, called also a bounded product:

  $a \ \mathbf{t}_{\mathcal{L}} \ b = 0 \ \vee \ (a + b - 1)$,
- Łukasiewicz t-conorm, called also a bounded sum:

  $a \ \mathbf{s}_{\mathcal{L}} \ b = 1 \ \wedge \ (a + b)$,

Most young students have high grades [T=?]

Student id	Age	Grade
1	29	3
2	25	2
3	30	3
4	20	4
5	21	5

Most young students have high grades $[\mathrm{T} = \mathrm{Most}(1.6/2) = 1]$

Student id	Age	Grade	young	high grade
1	29	3	0	0.4
2	25	2	0.2	0
3	30	3	0	0.4
4	20	4	1	0.8
5	21	5	0.8	1

Another quality criterion at the sentence level is the **degree of focus** of a linguistic summary $\mathcal{LS}$ that informs about coverage of objects that meet the condition expressed by the qualifier R. It is defined as follows:

$$T_2(S) = \frac{1}{n} \sum_{i=1}^{n} \mu_R(x_i), \tag{6}$$

where $\mu_R : \mathbb{R} \to [0, 1]$ is the membership function of the fuzzy number representing R.

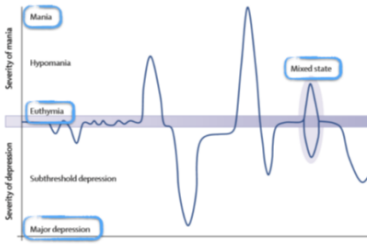
The **degree of support** of a linguistic summary $\mathcal{LS}$ indicates how many objects in the dataset are covered by the particular summary, and it is defined as:

$$T_3(S) = \frac{1}{n} \sum_{i=1}^{n} \{x_i : \mu_P(x_i) > 0 \wedge \mu_R(x_i) > 0\}, \tag{7}$$

where $\mu_R, \mu_P : \mathbb{R} \to [0, 1]$ are membership functions of the fuzzy numbers representing the qualifier R and the summarizer P, respectively.

Hands-on data: describing acoustic data collected from smartphones of bipolar disorder patients

Clinical practice confirms that speech can support the diagnosis of **mental disorders**. In particular, the **smartphone-based voice analysis** has great potential for monitoring mental states in **bipolar disorder (BD)**. For example, reduced speech activity, changes in specific voice features, and pause-related measures were found to be **sensitive markers of depressive symptoms** [21].



[21] M. Dominiak, K.Kaczmarek-Majer, A. Z. Antosik-Wojcinska, K. R. Opara, M. Wojnar, A. Olwert, W. Radziszewska, O. Hryniewicz, L. Swiecicki, and P. Mierzejewski, Behavioural data collected from smartphones in the assessment of depressive and manic symptoms for bipolar disorder patients: Prospective observational study, Journal of Medical Internet Research, 2021

- Participants of CHAD and MoodMon studies (over 100 patients diagnosed with BD) received a dedicated mobile application, called BDMon, able to collected acoustic data.
- The patient's voice signal was divided into 20ms frames (within one frame it is approximately stationary).
- The extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) for voice research was extracted (www.audeering.com/research/opensmile).

[22] Study was conduced within the CHAD project entitled Smartphone-based diagnostics of phase changes in the course of bipolar disorderâ€ (RPMA.01.02.00-14-5706/16-00) that was financed from EU funds (Regional Operational Program for Mazovia) in 2017-2018

- Participants of CHAD and MoodMon studies (over 100 patients diagnosed with BD) received a dedicated mobile application, called BDMon, able to collected acoustic data.
- The patient's voice signal was divided into 20ms frames (within one frame it is approximately stationary).
- The extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) for voice research was extracted (www.audeering.com/research/opensmile).

[22] Study was conduced within the CHAD project entitled Smartphone-based diagnostics of phase changes in the course of bipolar disorderâ€ (RPMA.01.02.00-14-5706/16-00) that was financed from EU funds (Regional Operational Program for Mazovia) in 2017-2018

- Participants of CHAD and MoodMon studies (over 100 patients diagnosed with BD) received a dedicated mobile application, called BDMon, able to collected acoustic data.
- The patient's voice signal was divided into 20ms frames (within one frame it is approximately stationary).
- The extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) for voice research was extracted (www.audeering.com/research/opensmile).

[22] Study was conduced within the CHAD project entitled Smartphone-based diagnostics of phase changes in the course of bipolar disorderâ€ (RPMA.01.02.00-14-5706/16-00) that was financed from EU funds (Regional Operational Program for Mazovia) in 2017-2018

Exercise I. Define fuzzy linguistic variables about acoustic features (e.g., energy). Next, construct fuzzy linguistic summaries to describe the dataset about smartphone-based acoustic features using fuzzy linguistic summaries.

- Input data: dataMoodMonitoring.csv
- Notebook: LSMoodMon.ipynb
- Example of the resulting summary: *For majority patients in state 0 with energy high we have f0 medium*

Exercise I. Define fuzzy linguistic variables about acoustic features (e.g., energy). Next, construct fuzzy linguistic summaries to describe the dataset about smartphone-based acoustic features using fuzzy linguistic summaries.

- Input data: dataMoodMonitoring.csv
- Notebook: LSMoodMon.ipynb
- Example of the resulting summary: *For majority patients in state 0 with energy high we have f0 medium*

Exercise I. Define fuzzy linguistic variables about acoustic features (e.g., energy). Next, construct fuzzy linguistic summaries to describe the dataset about smartphone-based acoustic features using fuzzy linguistic summaries.

- Input data: dataMoodMonitoring.csv
- Notebook: LSMoodMon.ipynb
- Example of the resulting summary: *For majority patients in state 0 with energy high we have f0 medium*

- **Loudness-related features** (loudness of speech signal and its energy): Patients in affective are expected to state speak louder compared to euthymia.
- **Pitch-related features** (F0final, F0envelope): Patients in an affective state are expected to speak with a higher or lower tone of voice.
- **Spectral-related features** (spectral harmonicity): Patients in the affective state are expected to have lower dynamics of changes in the speech signal spectrum.
- **Voice quality-related features** (jitter, shimmer): Patients with depressive symptoms are expected to speak less clearly, less fluently, more monotonously (chanting less), the intensity of the voice fluctuates more, they have a more asthenic voice. Patients with manic symptoms speak less clearly, more fluently, chant less, the intensity of the voice fluctuates less.

[23] K.Kaczmarek-Majer et al, Acoustic features from speech as markers of depressive and manic symptoms in bipolar disorder: A prospective study, Acta Psychiatrica Scandinavica, 2024

- In the state-of-the art, the accuracy ranging from 67% and higher for the **BD phase classification**.
- Confidential data alleviating to reproduce results.

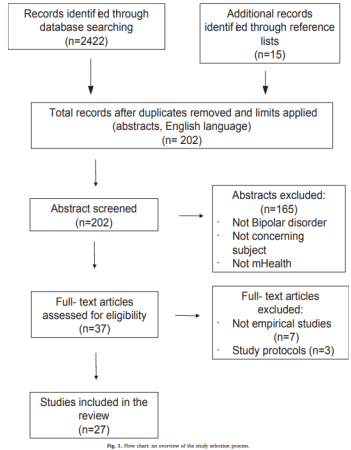
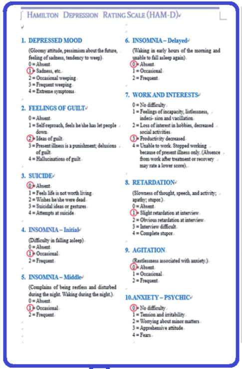


Fig. 1. Flow chart: an overview of the study selection process.

[23] A. Z. Antosik-Wojcinska, M. Dominiak, M. Chojnacka, K. Kaczmarek-Majer, K. R. Opara, W. Radziszewska, A. Olwert, and L. Swiecicki, Smartphone as a monitoring tool for bipolar disorder: a systematic review including data analysis, machine learning algorithms and predictive modelling, Int J Med Inform, vol. 138:104131, 2020.



HAMILTON DEPRESSION RATING SCALE (HAM-D)

1. DEPRESSED MOOD
(Gloomy attitude, pessimism about the future, feeling of sadness, tendency to weep)
0 = Absent
1 = Sadness, etc.
2 = Occasional weeping
3 = Frequent weeping
4 = Extreme symptoms

2. FEELINGS OF GUILT
0 = Absent
1 = Self-reproach, feels he/she has let people down
2 = Ideas of guilt
3 = Present illness is a punishment; delusions of guilt
4 = Hallucinations of guilt

3. SUICIDE
0 = Absent
1 = Feels life is not worth living
2 = Wishes he/she were dead
3 = Suicidal ideas or gestures
4 = Attempts at suicide

4. INSOMNIA – Initial
(Difficulty in falling asleep)
0 = Absent
1 = Occasional
2 = Frequent

5. INSOMNIA – Middle
(Complains of being restless and disturbed during the night. Waking during the night.)
0 = Absent
1 = Occasional
2 = Frequent

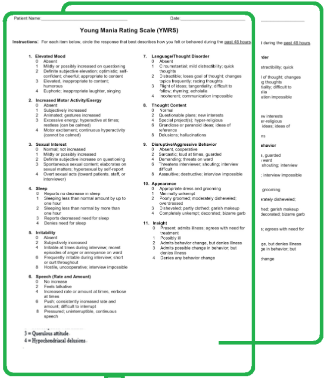
6. INSOMNIA – Delayed
(Waking in early hours of the morning and unable to fall asleep again)
0 = Absent
1 = Occasional
2 = Frequent

7. WORK AND INTERESTS
0 = No difficulty
1 = Feelings of incapacity, listlessness, indeci- sion and vacillation
2 = Loss of interest in hobbies, decreased social activities
3 = Productivity decreased
4 = Unable to work. Stopped working because of present illness only. (Absence from work after treatment or recovery may rate a lower score).

8. RETARDATION
(Slowness of thought, speech, and activity; apathy; stupor.)
0 = Absent
1 = Slight retardation at interview
2 = Obvious retardation at interview
3 = Interview difficult
4 = Complete stupor

9. AGITATION
(Restlessness associated with anxiety.)
0 = Absent
1 = Occasional
2 = Frequent

10. ANXIETY – PSYCHIC
0 = No difficulty
1 = Tension and irritability
2 = Worrying about minor matters
3 = Apprehensive attitude
4 = Fears



Patient Name: Date:

Young Mania Rating Scale (YMRS)

Instructions: For each item below, circle the response that best describes how you felt or behaved during the past 48 hours.

1. Elevated Mood
0 Absent
1 Mildly or possibly increased on questioning
2 Definite subjective elevation; optimistic; self-confident; cheerful; appropriate to content
3 Elevated, inappropriate to content; humorous
4 Euphoric; inappropriate laughter, singing

2. Increased Motor Activity/Energy
0 Absent
1 Subjectively increased
2 Animated; gestures increased
3 Excessive energy; hyperactive at times; restless (can be calmed)
4 Motor excitement; continuous hyperactivity (cannot be calmed)

3. Sexual Interest
0 Normal; not increased
1 Mildly or possibly increased
2 Definite subjective increase on questioning
3 Spontaneous sexual content; elaborates on sexual matters; hypersexual by self-report
4 Overt sexual acts (toward patients, staff, or interviewer)

4. Sleep
0 Reports no decrease in sleep
1 Sleeping less than normal amount by up to one hour
2 Sleeping less than normal by more than one hour
3 Reports decreased need for sleep
4 Denies need for sleep

5. Irritability
0 Absent
2 Subjectively increased
4 Irritable at times during interview; recent episodes of anger or annoyance on ward
6 Frequently irritable during interview; short or curt throughout
8 Hostile, uncooperative; interview impossible

6. Speech (Rate and Amount)
0 No increase
2 Feels talkative
4 Increased rate or amount at times, verbose at times
6 Push; consistently increased rate and amount; difficult to interrupt
8 Pressured; uninterruptible, continuous speech

7. Language/Thought Disorder
0 Absent
1 Circumstantial; mild distractibility; quick thoughts
2 Distractible; loses goal of thought; changes topics frequently; racing thoughts
3 Flight of ideas; tangentiality; difficult to follow; rhyming; echolalia
4 Incoherent; communication impossible

8. Thought Content
0 Normal
2 Questionable plans; new interests
4 Special project(s); hyper-religious
6 Grandiose or paranoid ideas; ideas of reference
8 Delusions; hallucinations

9. Disruptive/Aggressive Behavior
0 Absent, cooperative
2 Sarcastic; loud at times, guarded
4 Demanding; threats on ward
6 Threatens interviewer; shouting; interview difficult
8 Assaultive; destructive; interview impossible

10. Appearance
0 Appropriate dress and grooming
1 Minimally unkempt
2 Poorly groomed; moderately disheveled; overdressed
3 Disheveled; partly clothed; garish makeup
4 Completely unkempt; decorated; bizarre garb

11. Insight
0 Present; admits illness; agrees with need for treatment
1 Possibly ill
2 Admits behavior change, but denies illness
3 Admits possible change in behavior, but denies illness
4 Denies any behavior change

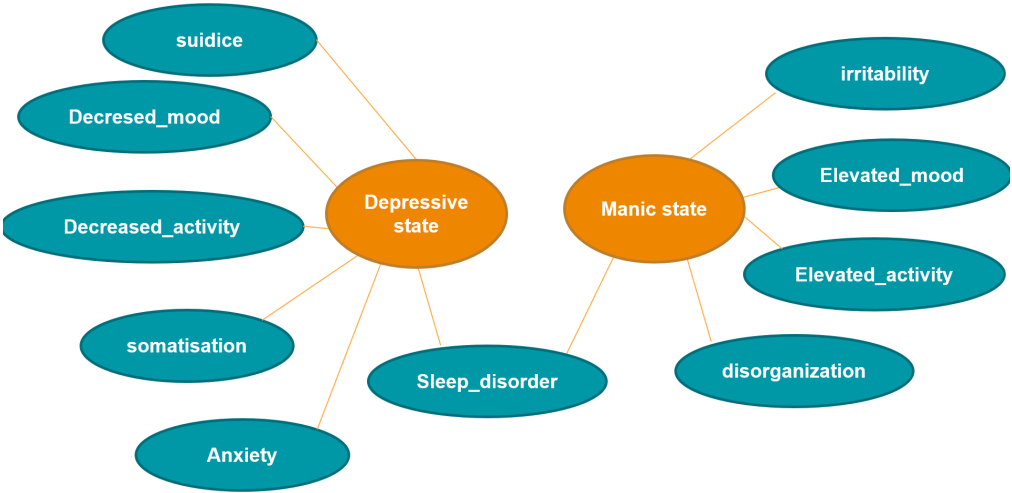
3 = Querulous attitude
4 = Hypochondriacal delusions

The state of euthymia is defined as HDRS<8 and YMRS<6

depression: HDRS≥8 and YMRS<6

hypomania/mania: HDRS<8 and YMRS≥6

mixed state: HDRS≥8 and YMRS≥6


suidice
Decresed_mood
Decreased_activity
somatisation
Anxiety
Depressive state
Sleep_disorder
Manic state
irritability
Elevated_mood
Elevated_activity
disorganization

Hamilton Depression Rating Scale (HAM-D)

(To be administered by a health care professional)

Patient Name _____ Today's Date _____

The HAM-D is designed to rate the severity of depression in patients. Although it contains 21 areas, calculate the patient's score on the first 17 answers.

☐ 1. DEPRESSED MOOD
(Gloomy attitude, pessimism about the future, feeling of sadness, tendency to weep)
0 = Absent
1 = Sadness, etc.
2 = Occasional weeping
3 = Frequent weeping
4 = Extreme symptoms

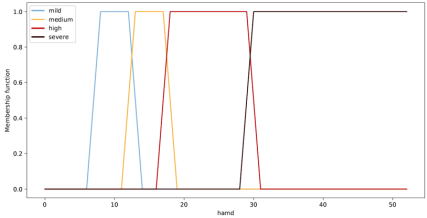
☐ 2. FEELINGS OF GUILT
0 = Absent
1 = Self-reproach, feels he/she has let people down
2 = Ideas of guilt
3 = Present illness is a punishment; delusions of guilt
4 = Hallucinations of guilt

☐ 3. SUICIDE
0 = Absent
1 = Feels life is not worth living
2 = Wishes he/she were dead
3 = Suicidal ideas or gestures
4 = Attempts at suicide

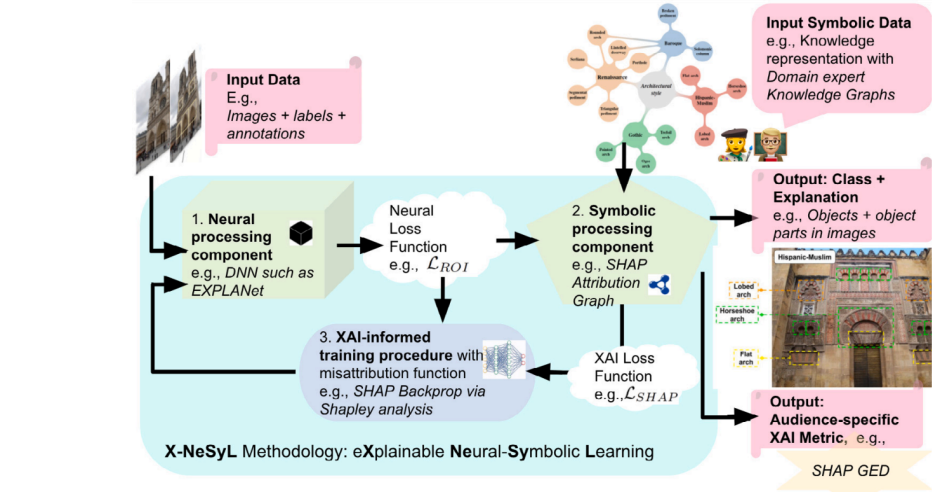
☐ 6. INSOMNIA - Delayed
(Waking in early hours of the morning and unable to fall asleep again)
0 = Absent
1 = Occasional
2 = Frequent

☐ 7. WORK AND INTERESTS
0 = No difficulty
1 = Feelings of incapacity, listlessness, indecision and vacillation
2 = Loss of interest in hobbies, decreased social activities
3 = Productivity decreased
4 = Unable to work. Stopped working because of present illness only. (Absence from work after treatment or recovery may rate a lower score).

☐ 8. RETARDATION
(Slowness of thought, speech, and activity; apathy; stupor.)
0 = Absent
1 = Slight retardation at interview
2 = Obvious retardation at interview
3 = Interview difficult



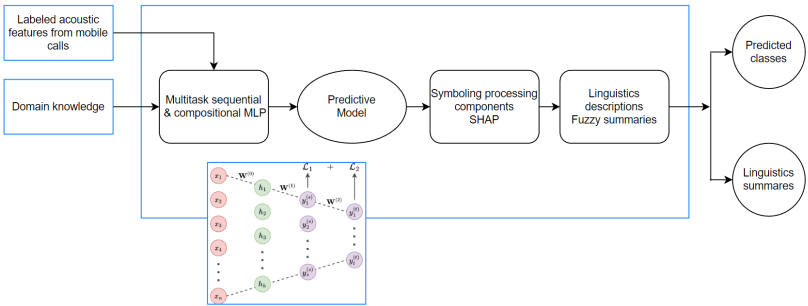
Related work: domain-expert knowledge graphs



[24] N. Diaz-Rodriguez, et al. EXplainable Neural-Symbolic Learning (X-NeSyL) methodology to fuse deep learning representations with expert knowledge graphs: The MonuMAI cultural heritage use case Inform. Fusion, 79 (2022), pp. 58-83

PLENARY: Explaining black-box models with fuzzy linguistic summaries

- **PLENARY (exPlaining bLack-box modEls in Natural lAnguage thRough fuzzY linguistic summaries)** is a three-step approach to create an accurate and explainable classifier equipped with linguistic summaries.



- We assume the availability of a set $\mathbf{X} \subset \mathbb{R}^{n \times d}$ of n training examples represented by d attributes (features) and labeled with one of t classes.
- Thus, each sample $\mathbf{x}_i \in \mathbf{X}$ is associated with a one-hot ground truth vector of length t, here denoted by $\left\{ \mathbf{y}_i^{(t)} \in \{0,1\}^t : \sum_{j=1}^{t} y_j^{(t)} = 1 \right\}$.
- We also assume that a second, intermediate level of s labels (mid-level labels for short), coming from domain knowledge, is associated with the training data.
- Hence, each sample $\mathbf{x}_i \in \mathbf{X}$ is also associated with a one-hot ground truth vector of length s, here denoted by $\left\{ \mathbf{y}_i^{(s)} \in \{0,1\}^s : \sum_{j=1}^{s} y_j^{(s)} = 1 \right\}$.

PLENARY Step 1. Creation of a compositional classification model

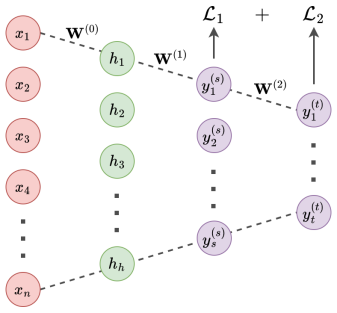
Supervised learning is on a two-level hierarchy of labels associated with data.

Multi-output sequential and compositional MLP is trained to simultaneously predict two different levels of labels (symptoms and mental states in our case study) associated with the same data.

The *final* output of the model is:

$$\hat{\mathbf{Y}}^{(t)} = \text{softmax}\left(\hat{\mathbf{Y}}^{(s)}\mathbf{W}^{(2)} + \mathbf{b}^{(2)}\right),$$

where $\mathbf{W}^{(2)} \in \mathbb{R}^{s\times t}$ and $\mathbf{b}^{(2)} \in \mathbb{R}^{t}$ are the final output layer weights and biases. $\hat{\mathbf{Y}}^{(s)} \in \mathbb{R}^{n\times s}$ is the *intermediate* output of the network.



Architectural diagram

Explanation of the outcomes of the predictive model using SHAPley values.

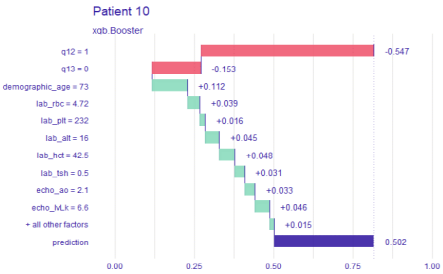
The feature attribution is decomposed additively to obtain the following explanation model g:

$$g(x') = \phi_0 + \sum_{i=1}^{M} \phi_i x'_i, \tag{8}$$

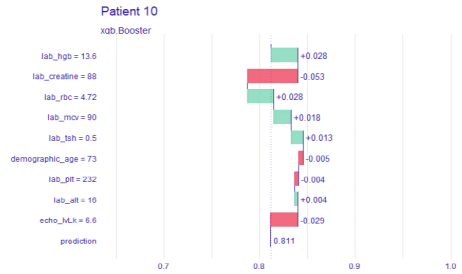
where M is the size of the coalition in terms of input features, $x' \in \{0,1\}^M$ indicates the presence (1) or not (0) of that feature in the given coalition, and $\phi_i \in \mathbb{R}$.

[25] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances in neural information processing systems 30.

Model A

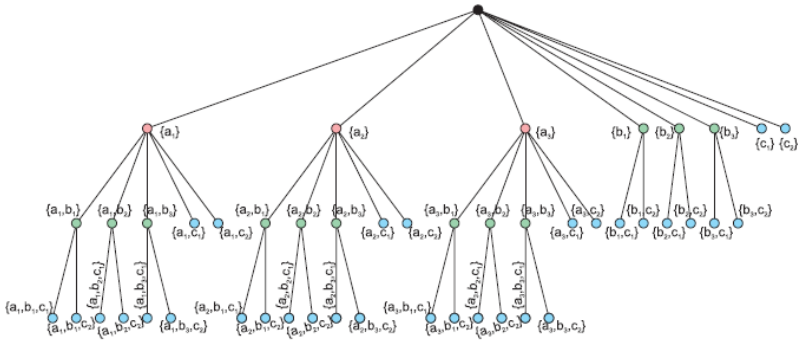


Model B



PLENARY Step 3. Linguistic summarization

Creation of linguistic summaries on global model explanations using fuzzy quantified sentences and tree-search algorithm



[26] Kacprzyk J. Linguistic summarization of time series using a fuzzy quantier driven aggregation. Fuzzy Sets Systems 2011

[27] **Kaczmarek-Majer, K.** and Hryniewicz, O. (2019) **Application of linguistic summarization methods in time series forecasting**

Among records that contribute against predicting depression class, most of them have spectral centroid feature at high level

Table 1
Construction of fuzzy numbers $A = (f_1, f_2, f_3, f_4)$ based on quartiles. Q_1 is the first quartile, Q_2 is median, and Q_3 is the third.

Attribute	Type	f_1	f_2	f_3	f_4
low	z-shape	*min*	*min*	Q_1	Q_2
medium	triangular	Q_1	Q_2	Q_2	Q_3
high	s-shape	Q_2	Q_3	*max*	*max*

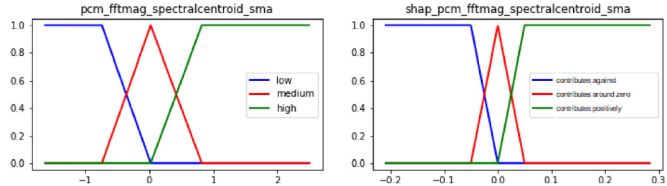


Fig. 3. Illustrative example of linguistic variables describing the spectral centroid acoustic feature and the SHAP values describing its importance.

Evaluation

- **Sentence level**: Degree of usefulness quantifying how useful the sentence explanation is from the perspective of human expert, reliability.
- **Group of summaries level**: system causability scale (SCS) [28], Grice's maxims.

1. *I found that the data included all relevant known causal factors with sufficient precision and granularity.*
2. *I understood the explanations within the context of my work.*
3. *I could change the level of detail on demand.*
4. *I did not need support to understand the explanations.*
5. *I found the explanations helped me to understand causality.*
6. *I was able to use the explanations with my knowledge base.*
7. *I did not find inconsistencies between explanations.*
8. *I think that most people would learn to understand the explanations very quickly.*
9. *I did not need more references in the explanations (e.g., medical guidelines, regulations).*
10. *I received the explanations in a timely and efficient manner.*

[28] A. Holzinger, A. Carrington, H. MÃ¼ller, Measuring the quality of explanations: the system causability scale (SCS), KI-KÃ¼nstliche Intelligenz 34 (2) (2020) 193-198.

Table 5
Comparative results for the BD state classification task. The best hyperparameter configuration is also reported under the results of each model, obtained by grid-searching over the following sets: *# estimators* $\in \{250, 500, 750\}$; *max depth* $\in \{3, 5, 7\}$; *objective* $\in \{softmax, softprob\}$; *optimizer* $\in \{Adam, SGD\}$; *learning rate* $\in \{0.01, 0.001, 0.001\}$; *batch size* $\in \{16, 32, 64\}$; *epochs* $\in \{5, 10, 15\}$.

Method	Class	Precision	Recall	F1-score
XGBoost	0 (Euthymia)	0.34	0.69	0.46
	1 (Depression)	0.00	0.00	0.00
	2 (Mania)	0.3	0.02	0.02
	3 (Mixed state)	0.00	0.00	0.00
	Accuracy			0.29
	# estimators = 500, max depth = 3, objective = softprob			
Single-task MLP	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
	optimizer = Adam, learning rate = 0.001, batch size = 32, epochs = 15			
Multi-task 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
	optimizer = Adam, learning rate = 0.001, batch size = 32, epochs = 15			

[30] BDMON dataset collected from four patients affected by bipolar disorder and between February and October 2018 within a prospective study. The program code and running examples of are available at the following link: https://github.com/ PLENARY ITPsychiatry/plenary

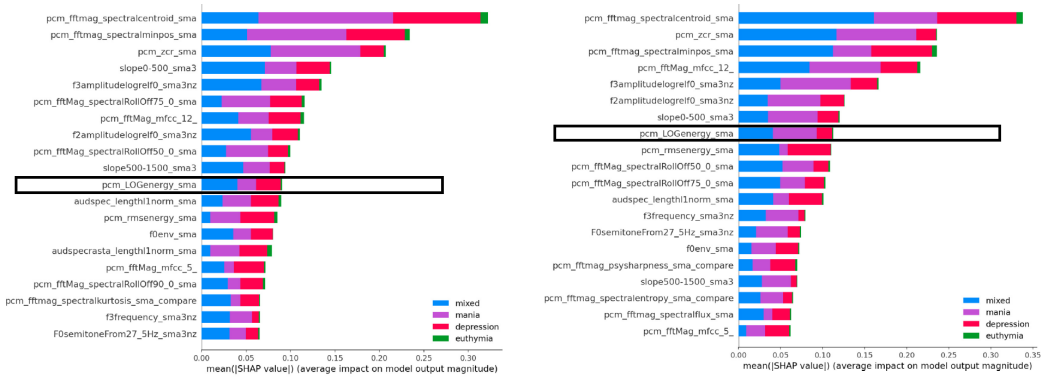
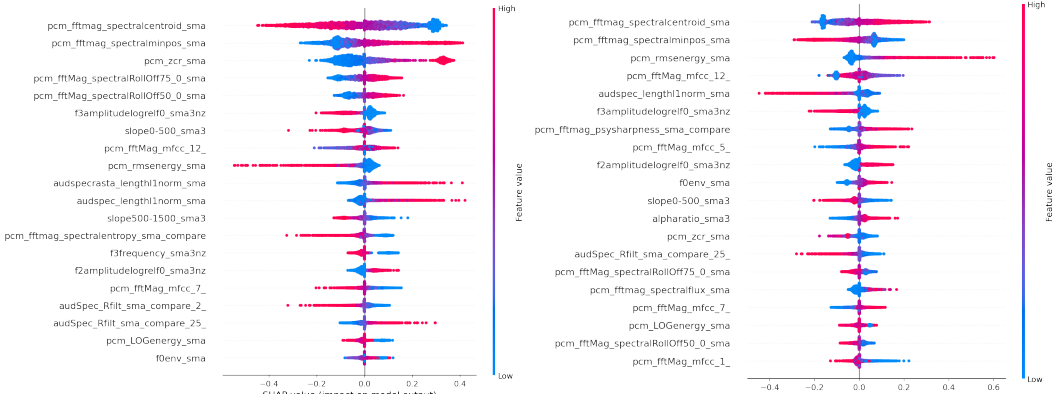
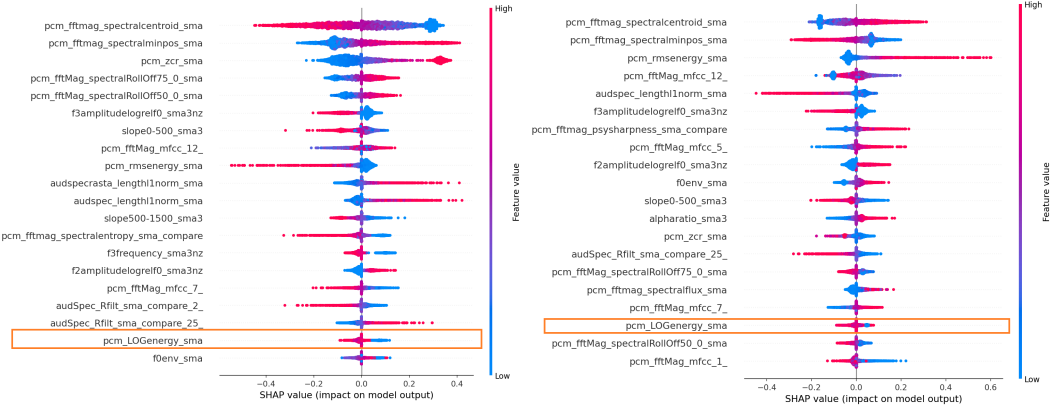


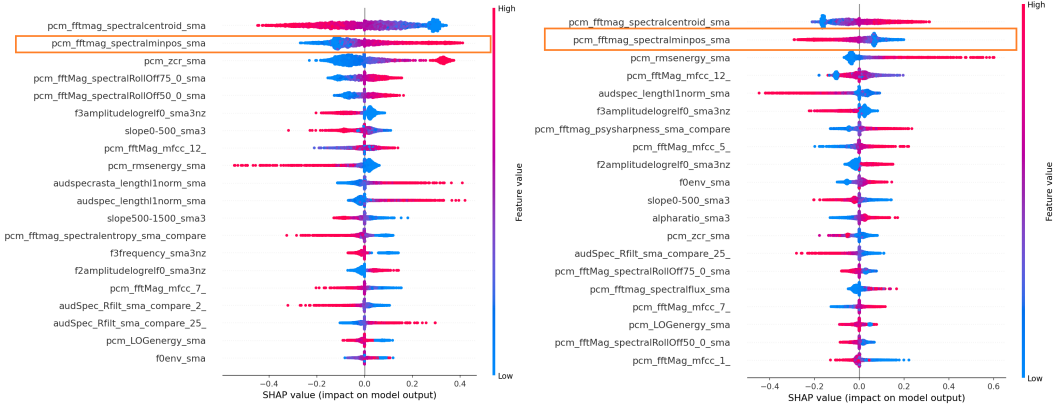
Fig. 4. Global model SHAP analysis for disease state prediction with a) the baseline model, and b) the sequential and compositional MLP model



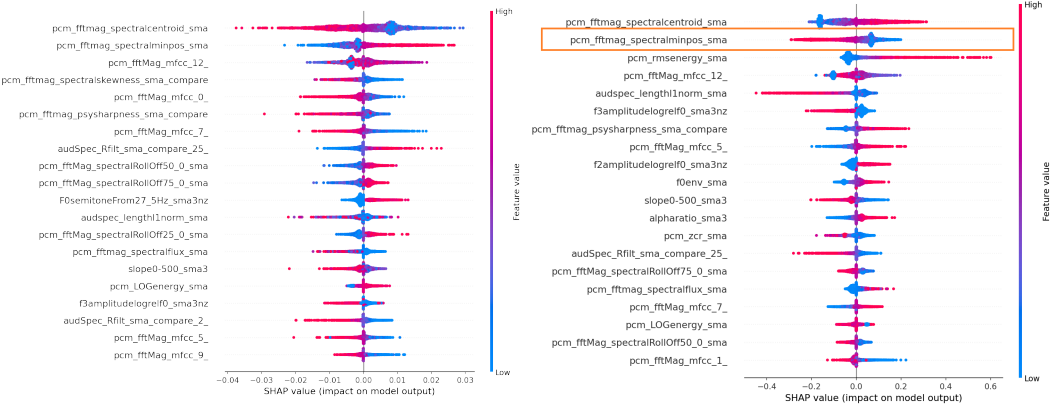
(Left): 20 most contributing features to the baseline MLP model for depression, (Right) 20 most contributing features to the sequential and compositional MLP model for depression



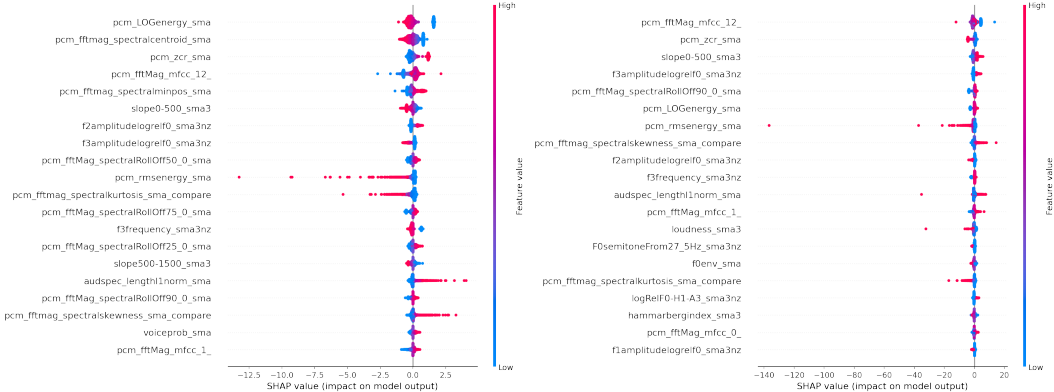
(Left): 20 most contributing features to the baseline MLP model for depression, (Right) 20 most contributing features to the sequential and compositional MLP model for depression



(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



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



20 most contributing features to the sequential and compositional MLP model (Left): for decreased activity, (Right) for elevated activity symptom

Table 6

Evaluation of linguistic summaries from PLENARY for the prediction of BD classes with the sequential and compositional MLP model. Degree of truth, degree of support, degree of focus, and expert-based degree of usefulness are applied as criteria. Post-processing criteria: $DoT > 0.1$ and $DoF > 0.05$. Summaries that contribute positively to predicting a class are presented in bold. The font colors of the LS description indicate the high-level semantic groups of acoustic features. LS related to: the energy-related features are marked in black; the spectral-related features are marked in olive; the pitch-related features in orange; and the quality-related features are marked in purple.

Id	LS description	DoT	DoS	DoF	DoU
001	Among records that contribute around zero to predicting euthymia, most of them have energy-related features at low level.	0.58	0.17	0.06	1
002	**Among records that contribute positively to predicting euthymia, most of them have energy-related features at low level.**	0.24	0.17	0.21	5
003	Among records that contribute against predicting euthymia, most of them have spectral-related features at high level.	0.19	0.54	0.63	2
004	Among records that contribute around zero to predicting euthymia, most of them have spectral-related features at low level.	0.53	0.17	0.06	1
005	**Among records that contribute positively to predicting euthymia, most of them have spectral-related features at low level.**	1.00	0.30	0.21	4
006	Among records that contribute against predicting euthymia, most of them have quality-related features at high level.	0.26	0.70	0.63	3
007	**Among records that contribute positively to predicting euthymia, most of them have quality-related features at low level.**	0.23	0.19	0.21	4
101	Among records that contribute around zero to predicting depression, most of them have energy-related features at high level.	0.12	0.17	0.06	1
102	**Among records that contribute positively to predicting depression, most of them have spectral-related features at high level.**	1.00	0.29	0.31	5
103	Among records that contribute against predicting depression, most of them have quality-related features at low level.	0.51	0.61	0.76	4
104	**Among records that contribute positively to predicting depression, most of them have quality-related features at low level.**	1.00	0.18	0.31	5
201	Among records that contribute against predicting mania, most of them have energy-related features at low level.	0.33	0.68	0.73	4
202	Among records that contribute around zero to predicting mania, most of them have energy-related features at low level.	1.00	0.19	0.03	1
203	Among records that contribute against predicting mania, most of them have pitch-related features at low level.	0.25	0.45	0.73	4
204	Among records that contribute around zero to predicting mania, most of them have pitch-related features at low level.	1.00	0.05	0.03	1
205	**Among records that contribute positively to predicting mania, most of them have pitch-related features at high level.**	0.59	0.39	0.44	5
206	**Among records that contribute positively to predicting mania, most of them have spectral-related features at low level.**	1.00	0.27	0.44	5
301	**Among records that contribute positively to predicting mixed state, most of them have energy-related features at high level.**	0.11	0.16	0.31	5
302	**Among records that contribute positively to predicting mixed state, most of them have pitch-related features at low level.**	0.45	0.34	0.31	5
303	Among records that contribute against predicting mixed state, most of them have spectral-related features at low level.	0.11	0.50	0.63	3
304	**Among records that contribute positively to predicting mixed state, most of them have spectral-related features at high level.**	1.00	0.27	0.31	5
305	Among records that contribute against predicting mixed state, most of them have quality-related features at low level.	0.75	0.66	0.63	3

Table 7
Evaluation of linguistic summaries for the prediction of elevated activity and decreased activity symptoms with $DoT > 0.1$ from the sequential and compositional MLP model. Degree of truth, degree of support, degree of focus, and expert-based degree of usefulness are applied as criteria. Summaries that contribute positively to predicting a class are presented in bold. The font colors of the LS description indicate the high-level semantic groups of acoustic features. LS results for all other symptoms are collected in the GitHub repository.

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

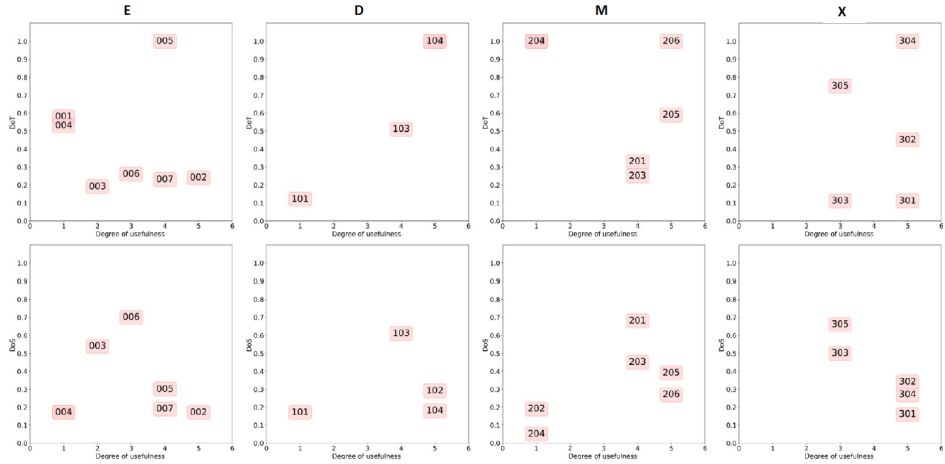


Fig. 9. Top row from left to right: degree of usefulness and degree of truth for linguistic summaries on euthymia, depression, mania, and mixed state from the sequential and compositional MLP model. Bottom row: degree of usefulness and degree of support for linguistic summaries for prediction of euthymia, depression, mania, and mixed state. The descriptions of the Ids are provided in Table 6.

Table 9
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 [43] (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
SCS5. I found the explanations helped me to understand causality	4
SCS6. I was able to use the explanations with my knowledge base	4
SCS7. I did not find inconsistencies between explanations	2
SCS8. I think that most people would learn to understand the explanations very quickly	5
SCS9. I did not need more references in the explanations (e.g., medical guidelines, regulations)	4
SCS10. I received the explanations in a timely and efficient manner	5
Mean SCS score (on a [0, 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
GM4. The group of sentences is clear, and as brief and orderly as possible, avoiding obscurity and ambiguity (maxim of *manner*)	3
Mean Grice's maxims rating (on a 1–5 Likert scale):	**4.25**

Exercise IIa. Build predictive models (XGBoost, NN, etc.) for the classification of bipolar state based on acoustic features.

- Input data: df train noise.csv, df test noise.csv
- Notebook: PLENARY1BuildingModels.ipynb

Exercise IIb. Define fuzzy linguistic variables about acoustic features and corresponding Shapley values, next construct fuzzy linguistic summaries to describe the identified models for classification of bipolar state based on acoustic features.

- Notebook: PLENARY2LS.ipynb
- Example of resulting summaries: *Among records that contribute against predicting 0 class, most of them have energy-related features at high level.*

Exercise IIa. Build predictive models (XGBoost, NN, etc.) for the classification of bipolar state based on acoustic features.

- Input data: df train noise.csv, df test noise.csv
- Notebook: PLENARY1BuildingModels.ipynb

Exercise IIb. Define fuzzy linguistic variables about acoustic features and corresponding Shapley values, next construct fuzzy linguistic summaries to describe the identified models for classification of bipolar state based on acoustic features.

- Notebook: PLENARY2LS.ipynb
- Example of resulting summaries: *Among records that contribute against predicting 0 class, most of them have energy-related features at high level.*

Exercise IIa. Build predictive models (XGBoost, NN, etc.) for the classification of bipolar state based on acoustic features.

- Input data: df train noise.csv, df test noise.csv
- Notebook: PLENARY1BuildingModels.ipynb

Exercise IIb. Define fuzzy linguistic variables about acoustic features and corresponding Shapley values, next construct fuzzy linguistic summaries to describe the identified models for classification of bipolar state based on acoustic features.

- Notebook: PLENARY2LS.ipynb
- Example of resulting summaries: *Among records that contribute against predicting 0 class, most of them have energy-related features at high level*.

Open Challenge. Select two predictive models (XGBoost, NN, etc.) for the classification of bipolar state based on acoustic features. Next, define fuzzy linguistic variables about acoustic features and corresponding Shapley values (or alternatives) to describe the identified models for classification of bipolar state based on acoustic features.

- Input data: MoonMonMonitoring.csv
- Deadline for sending slides (10min) summarizing the outcomes (for psychiatrists) + code: 26.09.2024. E-mail: k.kaczmarek@ibspan.waw.pl

Conclusions

- **Fuzzy linguistic summaries** as human-consistent information granules for describing in natural language the relation between the observations and model predictions
- Introduction of an intermediate layer of annotations to mitigate some uncertainties related to classes.
- Individual sentences have not always proven sufficient without exposing a clarification for the group of sentences to explain the broader context.

Conclusions

- **Fuzzy linguistic summaries** as human-consistent information granules for describing in natural language the relation between the observations and model predictions
- Introduction of an intermediate layer of annotations to mitigate some uncertainties related to classes.
- Individual sentences have not always proven sufficient without exposing a clarification for the group of sentences to explain the broader context.

Future work

1. Other types of protoforms but also quantifiers and t-norms.
2. Need for more comprehensive **multi-object summaries that allow for effective assessment and comparative analysis of global model explanations from multiple predictive models**.
3. Creation of a dynamic approach to summarize high-level groups that are not homogeneous in terms of impact on the predicted class.

Ongoing work

Varying confidence in actual class

Evaluating group of fuzzy linguistic summaries summaries

- The set of summaries is assumed to be consistent when it satisfies the non-contradiction and double negation properties. Non-contradiction implies that linguistic summaries made up of contradicting terms have a complementary degree of truth.

[31] J. van der Waa, E. Nieuwburg, A. Cremers, M. Neerincx, Evaluating XAI: A comparison of rule-based and example-based explanations, Artificial Intelligence 291 (2021) 103404. 3

[32] M. J. Lesot, G. Moyse, B. Bouchon-Meunier, Interpretability of fuzzy linguistic summaries, Fuzzy Sets and Sys- tems 292 (2016) 307-317

Let us now consider the following sentence as an example

$LS1 =$ *Among records that contribute positively to predicting euthymia class,* ***most*** *of them have energy-related features at* ***low*** *level.*

Assuming *high* and *low* are antonyms, as are *most* and *a few*, the following two sentences exemplify contradictory forms:

$C1 =$ *Among records that contribute positively to predicting euthymia class,* ***a few*** *of them have energy-related features at* ***low*** *level.*

$C2 =$ *Among records that contribute positively to predicting euthymia class,* ***most*** *of them have energy-related features at* ***high*** *level.*

Fuzzy Linguistic Summaries for Semi-Supervised Clustering

Table 6
Relative linguistic summaries based on short protoforms for mania and hypomania episodes (LS with $T = 1.0$) and extended protoforms for mania and hypomania episodes (LS with $T > 0.5$).

Relative LS based on short protoform	T
Most calls in the state of mania have low spectrum compared to the state of euthymia.	1.0
Most calls in the state of mania have low quality compared to the state of euthymia.	1.0
Most calls in the state of hypomania have low spectrum compared to the state of euthymia.	1.0
Most calls in the state of hypomania have low loudness compared to the state of euthymia.	1.0
Most calls in the state of hypomania have low qualty compared to the state of euthymia.	1.0
Relative LS based on extended protoform - HYPOMANIA	T
Most calls with low loudness in hypomania have low spectrum compared to the state of euthymia.	1.0
Most calls with low loudness in hypomania have low quality compared to the state of euthymia.	1.0
Most calls with high loudness in hypomania have high spectrum compared to the state of euthymia.	1.0
Most calls with high loudness in hypomania have high quality compared to the state of euthymia.	1.0
Most calls with low pitch in hypomania have low spectrum compared to the state of euthymia.	1.0
Most calls with low pitch in hypomania have low loudness compared to the state of euthymia.	1.0
Most calls with low pitch in hypomania have low quality compared to the state of euthymia.	1.0
Most calls with low spectrum in hypomania have low loudness compared to the state of euthymia.	1.0
Most calls with low spectrum in hypomania have low quality compared to the state of euthymia.	1.0
Most calls with high spectrum in hypomania have high loudness compared to the state of euthymia.	1.0
Most calls with high spectrum in hypomania have high quality compared to the state of euthymia.	1.0
Most calls with low quality in hypomania have low loudness compared to the state of euthymia.	1.0
Most calls with low quality in hypomania have low spectrum compared to the state of euthymia.	1.0
Most calls with high quality in hypomania have high loudness compared to the state of euthymia.	1.0
Most calls with high quality in hypomania have high spectrum compared to the state of euthymia.	1.0

https://github.com/ITPsychiatry/ssfclust

[33] K. Kaczmarek-Majer, G. Casalino, G. Castellano, O. Hryniewicz, M. Dominiak, Explaining smartphone-based acoustic data in bipolar disorder: Semi-supervised fuzzy clustering and relative linguistic summaries

Thank to all my collaborators!

Thank you for your attention!

Katarzyna Kaczmarek-Majer

k.kaczmarek@ibspan.waw.pl

https://www.ibspan.waw.pl/ kaczmar/

SFLA materials: **https://github.com/kasiakaczmarek/SFLA24-LS**