Assessment of a Trustworthy AI System in Healthcare: Smartphone-based Monitoring of Bipolar Disorder

Katarzyna Kaczmarek-Majer, Ph.D., D.Sc.

Computational Intelligence for Healthcare, Trustworthy AI Lab for Heatlhcare

Systems Research Institute Polish Academy of Sciences, Warsaw, Poland

E-mail: **k.kaczmarek@ibspan.waw.pl**

Workshop 'Social Impact of AI'

Madrid, 30-31 May 2024

Agenda

Assessment of a Trustworthy AI System in Healthcare: Smartphone-based Monitoring of Bipolar Disorder

1. Trustworthiness
2. Explainability

Agenda

Assessment of a Trustworthy AI System in Healthcare: Smartphone-based Monitoring of Bipolar Disorder

1. Trustworthiness
2. Explainability

Trustworthy Artificial Intelligence[1]:

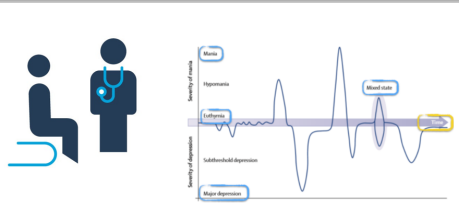
1. robust (technical and social perspective)
2. lawful
3. ethical
   - Respect for human autonomy
   - Prevention of harm,
   - Fairness
   - **Explicability**

Policy-makers pledged to develop a **'human-centric'** approach to AI[2]

[1] High-Level Expert Group on Artificial Intelligence, "Ethics guidelines for trustworthy AI," European Commission, Text

[2] "EU AI Act: first regulation on artificial intelligence". European Parliament News. Archived from the original on 10 January 2024. Retrieved 22 January 2024

Use case: Smartphone-based Monitoring of Bipolar Disorder



Participants of BDMon study[6] (100 patients diagnosed with BD) received a dedicated mobile application, called BDMon, able to collected acoustic data.

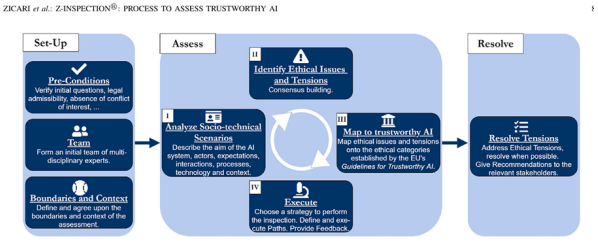
[5] A. Z. Antosik-Wojcinska, M. Dominiak, M. Chojnacka, K. Kaczmarek-Majer, K. R. Opara, W. Radziszewska, A. Olwert, and L. Swiecicki, Smartphone as a monitoring tool for bipolar disorder: a systematic review including data analysis, machine learning algorithms and predictive modelling, Int J Med Inform,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

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

Post Hoc Assessment of Trustworthy AI: Z-Inspection process

- The fundament of the Z-Inspection process is the **identification and discussion of ethical issues and tensions** through the elaboration of socio-technical scenarios. This can only be achieved with interdisciplinary team

- Assessment of trustworthy AI systems with the Z-Inspection **process** (https://z-inspection.org/).



[7] Zicari, R. V., Brodersen, J., Brusseau, J., Dudder, B., Eichhorn, T., Ivanov, T., et al. (2021). Z-inspection: A Process to Assess Trustworthy AI. IEEE Trans. Technol. Soc.

Illustrative example. Socio-Technical Scenario #1

Doctor using the BDMon system based on Model's Predictions and Alarms

Overview of the BDMon pipeline can be summarized as follows:

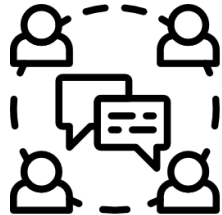
1. measurements and data collection (data acquisition, data annotation with ground truth, and signal processing)
2. feature extraction and feature selection
3. training of the classifier using the annotated examples
4. once the model is trained, actions are taken for new data (alarm or no alarm), based on the model's prediction and interpreted by the doctor and discussed with the patient
   - (1) The BDMon system predicts an alarm for a person
   - (2) The BDMon system predicts no alarm for a person

Illustrative example.

Question to the doctors/team members

to describe problems they see—or expect to see—arising from

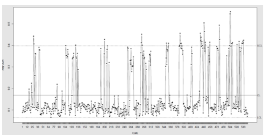
the BDMon system's use



- false positives, false negatives, true positives
- extensive trust into alarms linked to the perception issues of BD patients
- low trust in alarms (even though approx. 90% sensitivity of the system!)
- occasional lack of communication with the patient
- various groups of variables result differently efficient for various types of patient (acoustic, locomotion sensor, armband, etc.)
- delays in data transmissions
- database disconnected (technical watchdogs added)
- doctors overloaded with IT systems (need for extremely simple interface)
- no answer when the voice bot is calling (goodwill of patient needed)
- ...

Consulting the requirements listed in the Ethics Guidelines for Trustworthy AI for individual perspectives for Illustrative Socio-Technical Scenario #1

- **Ethical Issue #1**: **Description**: Using a black-box algorithm might impair the trust of the psychiatrists in the BDMon app, especially if the app has not been verified by multiple prospective studies. This complicates the accountability. Neither the reasoning behind alarms not the responsibility of alarms are not clearly defined for clinicians, healthcare institutions, etc.
- **Map to Ethical Pillars and Requirements**: Explicability > Transparency > Traceability, Explainability
- **Ethical Tensions**: **Accuracy versus Transparency/Explainability**



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

Assessment of a Trustworthy AI System in Healthcare: Smartphone-based Monitoring of Bipolar Disorder:

1. Trustworthiness
2. Explainability

Explanations

As extracted from the Cambridge Dictionary of English Language[3]:

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[4].

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

[4] 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[8]
- Explainability: level of understanding how the AI-based system came up with a given result[8]

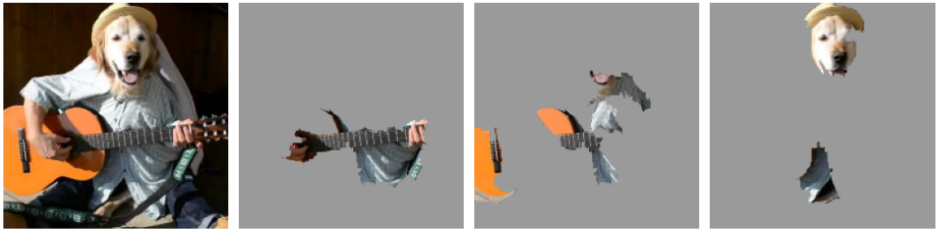
Explanation is textual or visual artifact that provide a qualitative understanding of the relationship between the instance's components (e.g. words in text, patches in an image) and the model's prediction[10]

---

[8] 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

[9] D. A. Broniatowski, "Psychological Foundations of Explainability and Interpretability in Artificial Intelligence," National Institute of Standards and Technology, Tech. Rep., Apr. 2021.

[10] 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)



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

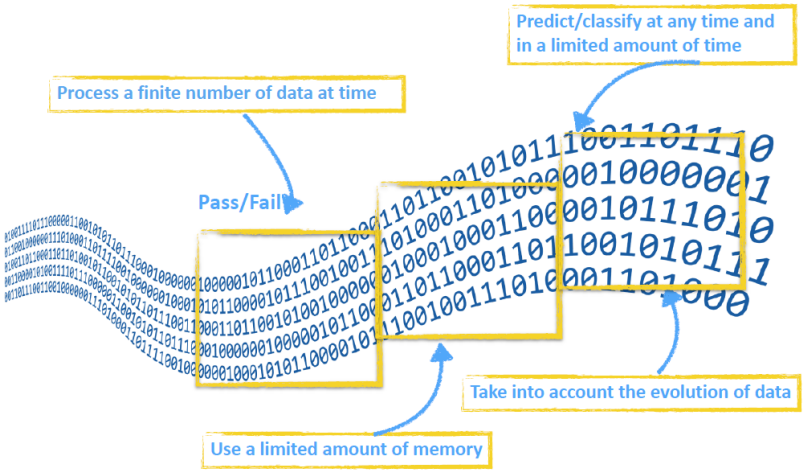
Figure 4: Explaining an image classification prediction made by Google's Inception neural network. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

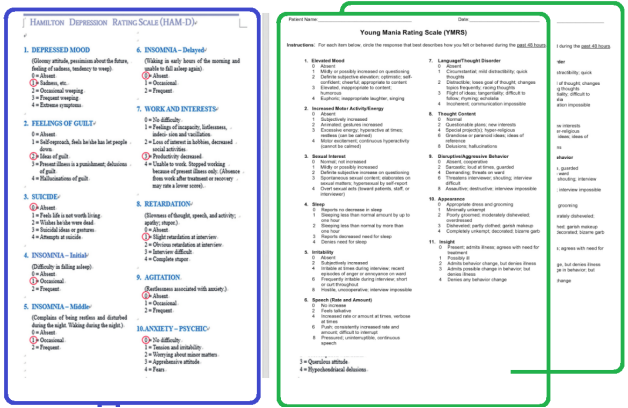
Predictions of Google's Pre-trained Inception neural network[1]

explained by a post-hoc method [2]

[1] 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.

[2] 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.


Predict/classify at any time and in a limited amount of time
Process a finite number of data at time
Pass/Fail
Take into account the evolution of data
Use a limited amount of memory



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 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, indecisiveness 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.

Young Mania Rating Scale (YMRS)

Patient Name: _____ Date: _____

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 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
1 Subjectively increased
2 Irritable at times during interview; recent irritability on the ward; irritable with peers
3 Frequently irritable during interview; short or curt throughout
4 Hostile, uncooperative; interview impossible

6. Speech (Rate and Amount)
0 No increase
1 Feels talkative
2 Increased rate or amount at times, verbose at times
3 Push; consistently increased rate and amount, difficult to interrupt
4 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
1 Questionable psych. new interests
4 Special project(s); hyper-religious
6 Grandiose or paranoid ideas; ideas of reference
8 Delusions; hallucinations

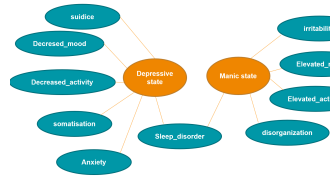
9. Disruptive/Aggressive Behavior
0 Cooperative; appropriate
2 Sarcastic; loud at times; guarded
4 Demanding; threats on ward
6 Threatens interviewer; shouting; interview difficult
8 Assaultive; destructive; interview impossible

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 = Questions attitude.
4 = Hypochondriacal delusions.

- Symptoms in the common rating scales used in psychiatry, such as Hamilton Depression Rating Scale and Young Mania Rating Scale



Fuzzy linguistic summarization

- **Fuzzy linguistic summaries** (LSs) are statements in natural language that describe numerical datasets[11]. LSs have been confirmed as **human-consistent** information granules with applications in various domains.

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.

- **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[14].
- **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].

---

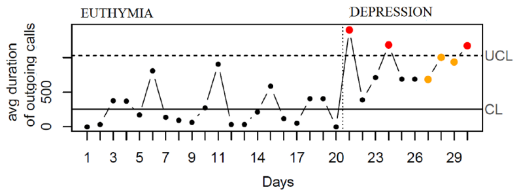
[14] F. Höppner, S. Peter, and M. Berthold, Enriching multivariate temporal patterns with context information to support classification,Computational Intelligence in Intelligent Data Analysis. vol. 445, pp. 195-206, 2013.

[15] M. Burda, M. Stepnicka, and L. Stepnická, 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)

How to efficiently communicate with the psychiatrist about an alarming situation basing on the data collected from smartphone ?



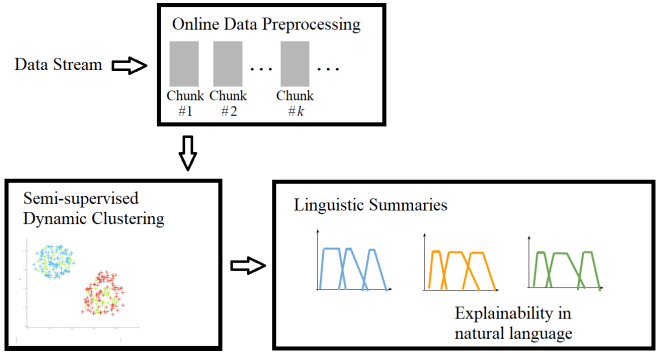
Most outgoing calls of patient P are long [T=0.8]
Depressive episode may have started.

- We construct fuzzy linguistic summaries using **evolving membership functions** based on prototypes from semi-supervised learning following the idea of **evolving fuzzy systems** [18].

[18] P. Angelov, D. P. Filev, N. Kasabov, Evolving Takagi-Sugeno Fuzzy Systems from Streaming Data (eTS+), 2010, pp. 21 - 50

[19] Hryniewicz, O. Kaczmarek-Majer, K. Opara, K. (2019) Control Charts Based on Fuzzy Costs for Monitoring Short Autocorrelated Time Series. International Journal of Approximate Reasoning, p 166-181, 10.1016/j.ijar.2019.08.013

LS-FC: Explaining partially-labeled data in natural language[20]



Rysunek: Overview of the proposed *LS-FC* approach that explains data stream by means of summarization and online learning

[20] 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, Information Sciences 588 (2022) 174–195.

[21] Kaczmarek-Majer, K. and Hryniewicz, O. (2019) Application of linguistic summarization methods in time series forecasting. Information Sciences Vol 478, p 580 – 594

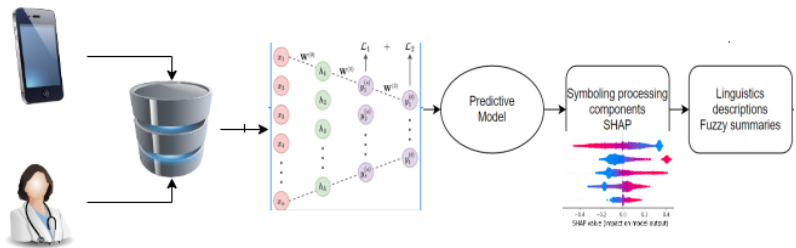
Table 6
Relative linguistic summaries based on short protoforms for mania and hypomania episodes (LS with $T = 1.0$) and extended protoforms for mania and hypomania episodes (LS with $T > 0.5$).

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

PLENARY: Explaining black-box models in natural language using fuzzy linguistic summaries[24]

$LS1$: *Among records that contribute positively to predicting depression class,*

most *of them have energy-related features at* ***low*** *level.*



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

The program code and running examples of are available at the following link: https://github.com/ PLENARY ITPsychiatry/plenary

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

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

The program code and running examples of are available at the following link: https://github.com/ PLENARY ITPsychiatry/plenary

- **Z-Inspection** as a holistic process to assess Trustworthy AI.
- **Fuzzy linguistic summaries** as explanations / human-consistent information granules with applications in various domains. Individual sentences have not always proven sufficient without exposing a clarification for the group of sentences to explain the rationale of the intelligent system's behaviour. Semi-Supervised Fuzzy-C Means as an explainable-by-design method for partially-labeled data.

Thank to all my collaborators!

Thank you for your attention!

Katarzyna Kaczmarek-Majer

k.kaczmarek@ibspan.waw.pl

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

https://z-inspection.org/

http://bipolar.ibspan.waw.pl

https://github.com/ITPsychiatry/ssfclust

https://github.com/ITPsychiatry/bipolar

https://github.com/ITPsychiatry/plenary