Fuzzy linguistic summaries for partially-labeled data streams and their application in monitoring of mental health

Katarzyna Kaczmarek-Majer, Olgierd Hryniewicz

Systems Research Institute Polish Academy of Sciences, Warsaw, Poland

k.kaczmarek@ibspan.waw.pl

http://bipolar.ibspan.waw.pl

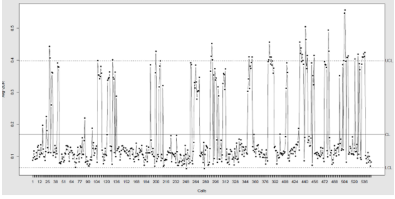
Seminarium Sekcji „Inteligentnych Systemów Wspomagania Decyzji oraz Obliczeń Granularnych" Komitetu Informatyki PAN

Agenda

1. Introduction and related work
2. Fuzzy linguistic summarization
3. Motivating example in smartphone-based monitoring of bipolar disorder
4. **PLENARY: Explaining black-box models in natural language through fuzzy linguistic summaries**
5. **LS-FC: Linguistic Summaries with Fuzzy Clustering**
6. Experimental results
7. Conclusions

Motivation: explaining medical systems

- Intelligent systems for the medical domain often require processing **data streams that evolve over time** and are only **partially labeled**.
- At the same time, **the need for explanations** is of utmost importance not only due to various regulations, but also to increase trust among users.
- Autonomous systems are expected to explain why and how outcomes were generated.



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

Related work: visual explanations

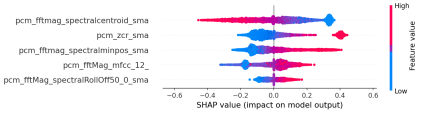
- Recently, various XAI techniques provide explanations in the form of visual descriptions (plots, heatmaps, etc.), e.g., GRAD-CAM [1].
- Saliency analysis [2] is also used to highlight the importance of words based on attribution scores.

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

[2] J. Su, J. Chen, H. Jiang, C. Zhou, H. Lin, Y. Ge, Q. Wu, Y. Lai, Multi-modal neural machine translation with deep semantic interactions, Inf. Sci. 554 (2021)

Related work: global/local explanations

- Post hoc XAI techniques aim to explain the outputs of models that are not interpretable by design. Examples: **LIME** (Local Interpretable Model-agnostic Explanations) [4] which explains the predictions of any classifier by computing importance scores of features; SHapley Additive exPlanations (**SHAP**) [5] is also aimed at providing model explanations from tabular data; and more.



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

[4] 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.

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

Related work: Fuzzy linguistic summarization

- **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.

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.

[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

[7] 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.

Fuzzy linguistic summaries: definitions

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 [9] is defined as:

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

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

---

[9] 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

[10] 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

Fuzzy linguistic summaries: example of extended protoform

One of the earliest and most popular measures of validity of a linguistic summary $\mathcal{LS}$ (Eq. 1) is the **degree of truth** (DoT), defined as:

$$DoT(Q, R, P) = \mu_Q \left(\frac{\sum_{i=1}^{n} (\mu_R(x_i) * \mu_P(x_i))}{\sum_{i=1}^{n} \mu_R(x_i)} \right), \tag{2}$$

where $*$: [0,1] $\times$ [0,1] $\rightarrow$ [0,1] is a triangular norm (t-norm for short) and $\mu_Q, \mu_R, \mu_P : \mathbb{R} \rightarrow [0, 1]$ are the membership functions of the fuzzy numbers representing the quantifier Q, qualifier R, and the summarizer P, respectively.

Fuzzy linguistic summaries: quality criteria

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:

$$DoF(R) = \frac{1}{n} \sum_{i=1}^{n} \mu_R(x_i), \tag{3}$$

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

Fuzzy linguistic summaries: quality criteria

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:

$$DoS(P,R) = \frac{1}{n}\sum_{i=1}^{n}\{x_i : \mu_P(x_i) > 0 \wedge \mu_R(x_i) > 0\}, \tag{4}$$

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.

Fuzzy linguistic summaries: evaluating group of summaries

- However, **low-level sentences have not always proved sufficient without exposing a clarification of the overall rationale of the complete behaviour of the intelligent system**[11]. Thus, apart from sentence-level measures, we investigate the evaluation of groups of summaries.
- The set of summaries is assumed to be consistent when it satisfies the non-contradiction and double negation properties[12]. Non-contradiction implies that linguistic summaries made up of contradicting terms have a complementary degree of truth.

[11] 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

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

Primary goal: explanations as statements in natural language

1. **PLENARY**: Explaining black-box models in natural language through fuzzy linguistic summaries [13]

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

2. **LS-FC**: Linguistic Summaries with Fuzzy Clustering [13]

   Most calls with high loudness in mania have low spectrum compared to the state of euthymia

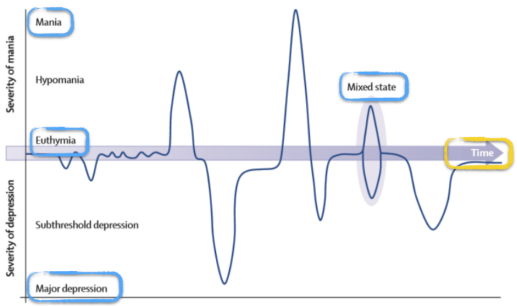
[13] Katarzyna Kaczmarek-Majer, Gabriella Casalino, Giovanna Castellano, Monika Dominiak, Olgierd Hryniewicz, Olga Kamińska, Gennaro Vessio, Natalia Díaz-Rodríguez, PLENARY: Explaining black-box models in natural language through fuzzy linguistic summaries, Information Sciences, Volume 614, 2022, Pages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

[14] 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.

Motivating example: smartphone-based monitoring bipolar disorder

Motivation for this research comes from a **prospective observation study** conducted in the Department of Affective Disorders, Institute of Psychiatry and Neurology in Warsaw, Poland that included patients diagnosed with bipolar disorder (F31 according to ICD-10 classification).

Bipolar disorder (BD) is a serious disease characterized by **mood fluctuations from euthymia through depression and mania to mixed states.**



[15] 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

Motivating example: smartphone-based monitoring bipolar disorder

- Participants of the considered study received a dedicated mobile application, called BDMon, able to collected acoustic data about phone calls. 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.

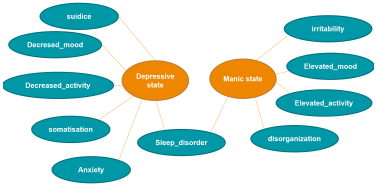
[16] M. Dominiak, K. Kaczmarek-Majer, A.Z. Antosik-Wojcinska, K.R. Opara, M. Wojnar, A. Olwert, W. Radziszewska, O. Hryniewicz, L. Swiecicki, P. Mierzejewski, Behavioural data collected from smartphones in the assessment of depressive and manic symptoms for bipolar disorder patients: Prospective observational study, J. Med. Internet Res. 24.

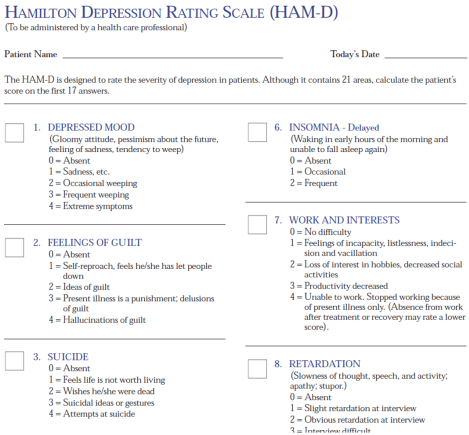
Motivating example: voice is a promising marker of mental state

- **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 (compared to euthymia).
- **Spectral-related features** (spectral flux, 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,

Motivating example: two levels of labels

- Common rating scales used in psychiatry, such as Hamilton Depression Rating Scale and Young Mania Rating Scale, are based on a disease classification (ICD-11).





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

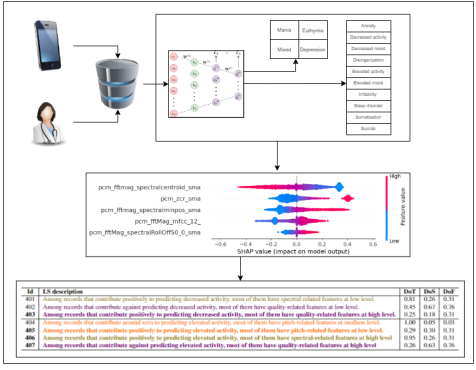
7. WORK AND INTERESTS
0 = No difficulty
1 = Feelings of incapacity, listlessness, indecision and vacillation
2 = Loss of interest in hobbies, decreased social activities
3 = Productivity decreased
4 = Unable to work. Stopped working because of present illness only. (Absence from work after treatment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

[17] A.Z. Antosik-Wojcinska, M. Dominiak, M. Chojnacka, K. Kaczmarek-Majer, K.R. Opara, W. Radziszewska, A. Olwert, 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. 138:104131.

PLENARY: Motivation

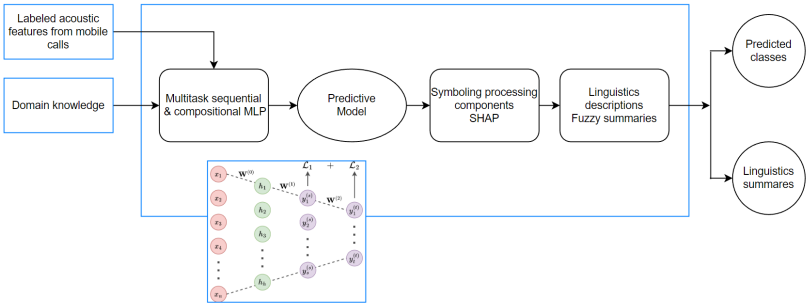
- **SHAP results are visual, imprecise amd might be hard for interpretation by non-technicians**, we address this imprecision in the third step of our approach by applying linguistic summarization to derive natural language sentences that support understanding of such graphical explanations.



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	0.36
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	0.36

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.



[13] K. Kaczmarek-Majer, G. Casalino, G. Castellano, M. Dominiak, O. Hryniewicz, O. Kami nska, G. Vessio, N. Diaz- Rodriguez, Plenary: Explaining black-box models in natural language through fuzzy linguistic summaries, Infor- mation Sciences (2022

PLENARY: Explaining black-box models with fuzzy 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: Explaining black-box models with fuzzy linguistic summaries

1. **Creation of a compositional classification model via supervised learning based 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.
2. **Explanation of the outcomes of the predictive model using SHAP;**
3. **Creation of linguistic summaries on global model explanations using fuzzy quantified sentences.**

PLENARY: Illustrative example of constructing linguistic summaries

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*

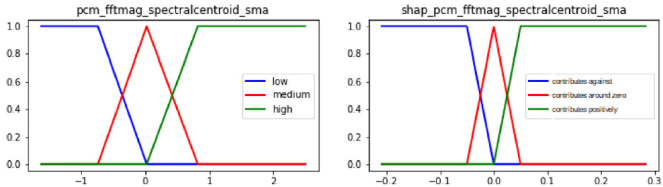


Fig. 3. Illustrative example of linguistic variables describing the spectral centroid acoustic feature and the SHAP values describing its importance.

PLENARY: Expert-based 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) [18], 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.*

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

PLENARY: objective evaluation of group of summaries

Formally, contradictory forms of a summary based on extended protoform $\mathcal{LS}$ are defined as follows:

$$C1(Q, R, P) = \text{Among } R \text{ objects from } O, \neg\ Q \text{ have } P,$$

$$C2(Q, R, P) = \text{Among } R \text{ objects from } O, Q \text{ have } \neg\ P.$$

The double negation D of a sentence $\mathcal{LS}$ is defined as

$$D(LS) = C1(C2(LS)) = C2(C1(LS)) = \text{Among } R \text{ objects from } O, \neg\ Q \text{ have } \neg\ P.$$

The double negation property states that $DoT(D(LS)) = DoT(LS)$.

PLENARY: objective evaluation of group of summaries

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.*

PLENARY: objective evaluation of group of summaries

Let us still 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.*

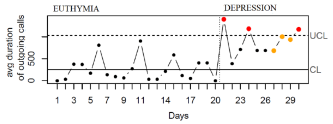
The following sentence is an example of double negation to *LS1*:

LS2 = Among records that contribute positively to predicting euthymia class, ***a few*** *of them have energy-related features at* ***high*** *level.*

LS-FC: Explaining partially-labeled data in natural language

- We construct fuzzy linguistic summaries using **evolving membership functions** based on prototypes from semi-supervised learning following the idea of **evolving fuzzy systems** [19].
- The main purpose of LS-FC is to **generate linguistic summaries to incrementally explain** in natural language the evolution of data in a stream.

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

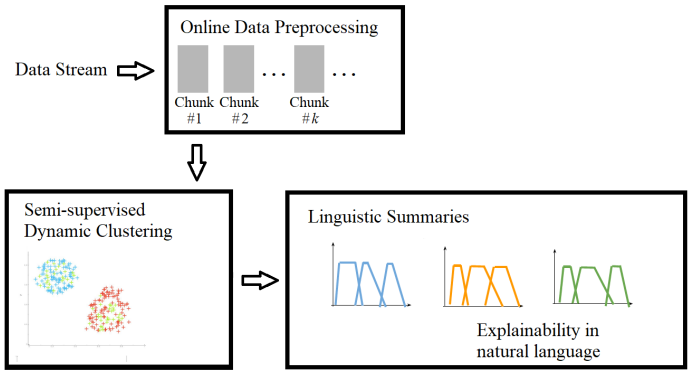


Most outgoing calls of patient P are long [T=0.8]
Depressive episode may have started.

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

[20] 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: Overview of the proposed approach



Overview of the proposed approach that explains data stream by means of Linguistic Summaries and online learning using Semi-Supervised Dynamic Fuzzy C-Means.

LS-FC: Incremental Semi-Supervised Learning

The proposed method builds on the **Dynamic Incremental Semi-Supervised Fuzzy C-Means (DISSFCM)** introduced in [21], inspired by work of Pedrycz et. al[22].

Objective function J

$$J = \sum_{k=1}^{K} \sum_{j=1}^{N_t} u_{jk}^2 d_{jk}^2 + \alpha \sum_{k=1}^{K} \sum_{j=1}^{N_t} \left(u_{jk} - b_j f_{jk}\right)^2 d_{jk}^2$$

- K is the number of clusters
- $N_t = |X_t|$ is the number of samples in the t-th chunk
- $u_{jk} \in [0, 1]$ is the membership degree of a sample $\mathbf{x}_j$ in the k-th cluster
- d_{jk} is the Euclidean distance between a sample $\mathbf{x}_j$ and the center $\mathbf{c}_k$ of the k-th cluster
- $b_j = b(\mathbf{x}_j)$, where $b : X \mapsto \{0, 1\}$ such that $b(\mathbf{x}) = 1$ iff $\mathbf{x}$ is pre-labeled, i.e., its class value is known
- $f_{jk} = 1$ iff the j-th sample has the same class label of the k-th cluster ($f_{jk} = 0$, otherwise)

[21] G. Casalino, G. Castellano, and C. Mencar (2019) Data stream classification by dynamic incremental semi-supervised fuzzy clustering, International Journal on Artificial Intelligence Tools

[22] W. Pedrycz and J. Waletzky (1997) Fuzzy clustering with partial supervision, IEEE Transactions on Systems, Man, and Cybernetics

LS-FC: Incremental Semi-Supervised Learning

Objective function J

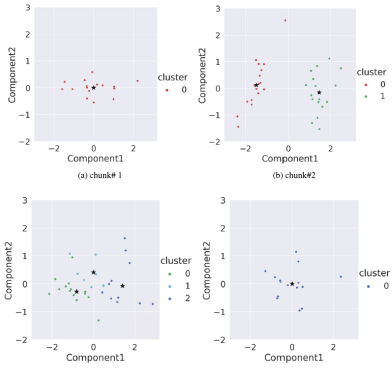
$$J = \sum_{k=1}^{K} \sum_{j=1}^{N_t} u_{jk}^2 d_{jk}^2 + \alpha \sum_{k=1}^{K} \sum_{j=1}^{N_t} \left(u_{jk} - b_j f_{jk}\right)^2 d_{jk}^2$$

- $\alpha \geqslant 0$ is a regularization parameter that weights the second term of the objective function. It exploits the class information.
- For each chunk, DISSFCM computes medoids and returns them as cluster prototypes. Medoids are representative points for the cluster whose sum of dissimilarities to all the points in the cluster is minimal.

[23] K. Kmita, G. Casalino, G. Castellano, O. Hryniewicz, and K. Kaczmarek-Majer (2022) Confidence path regularization for handling label uncertainty in semi-supervised learning: use case in bipolar disorder monitoring, FUZZ-IEEE

LS-FC: Construction of evolving linguistic summaries

1. Fuzzification of acoustic features using clusters' prototypes learned by DISSFCM algorithm.
2. Granulation of the acoustic features.
3. Derivation of linguistic summaries.



g. 1. PCA visualization of four chunks obtained from IRIS data and prototypes (denoted by a ☆) resulting from DISSFCM..

Experimental results: PLENARY

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			

[24] 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

Experimental results: PLENARY

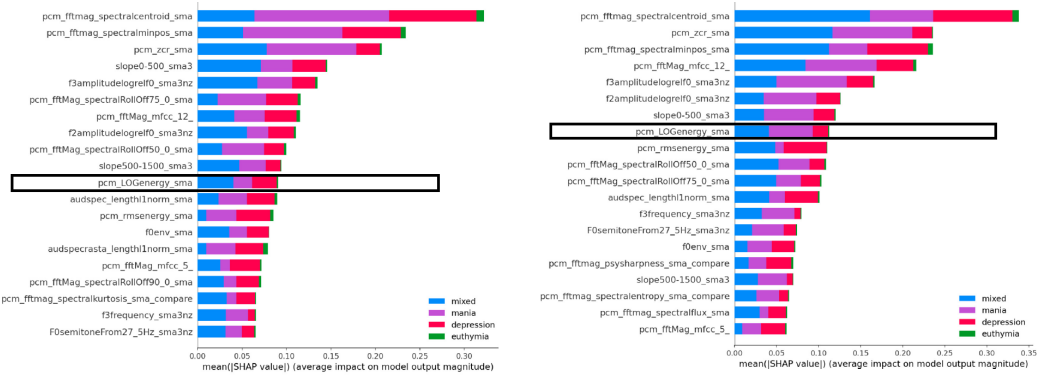
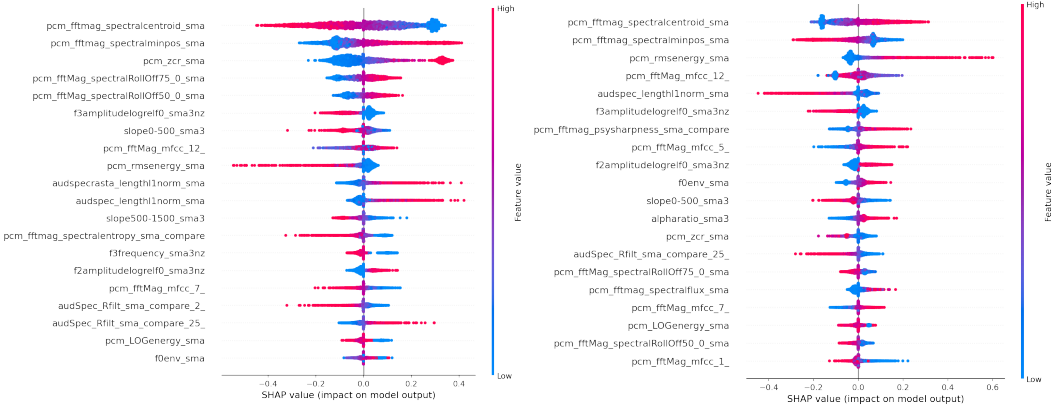


Fig. 4. Global model SHAP analysis for disease state prediction with a) the baseline model, and b) the sequential and compositional MLP model

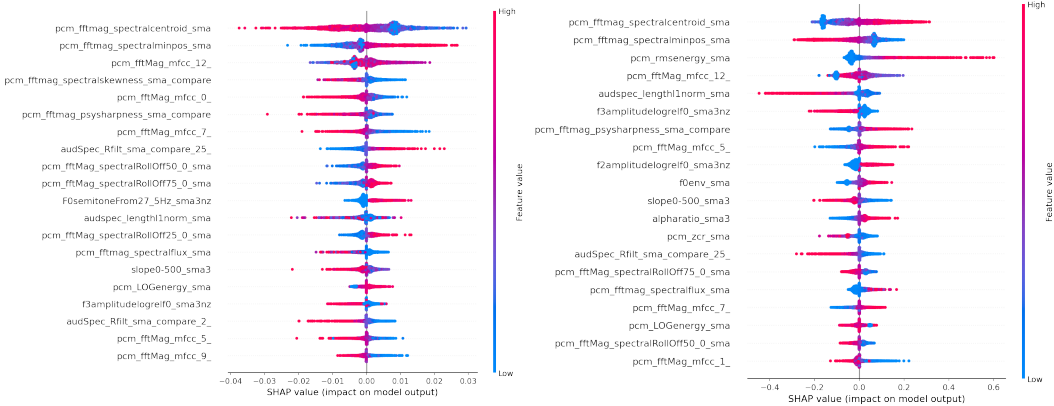
[24] The program code and running examples of are available at the following link: https://github.com/ PLENARY ITPsychiatry/plenary

Experimental results: PLENARY



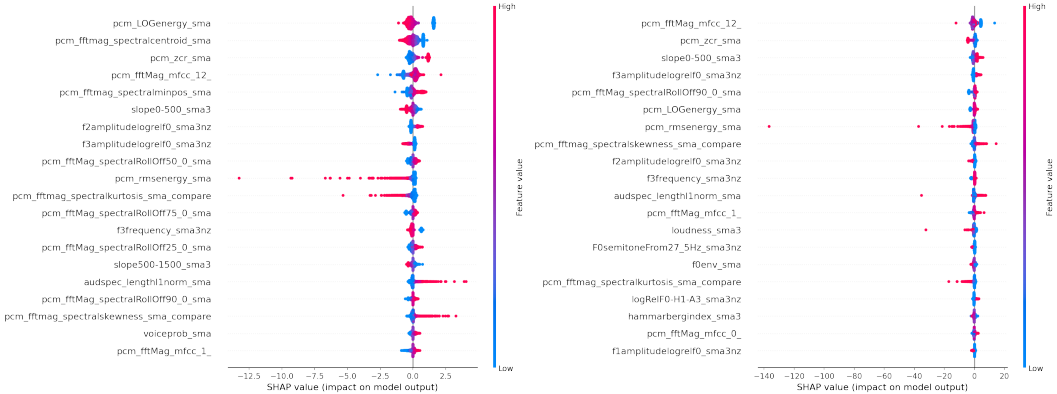
(Left): 20 most contributing features to the baseline MLP model for depression, (Right) 20 most contributing features to the sequential and compositional MLP model for depression

Experimental results: PLENARY



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

Experimental results: PLENARY



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

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

Experimental results: PLENARY

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

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

Experimental results: PLENARY

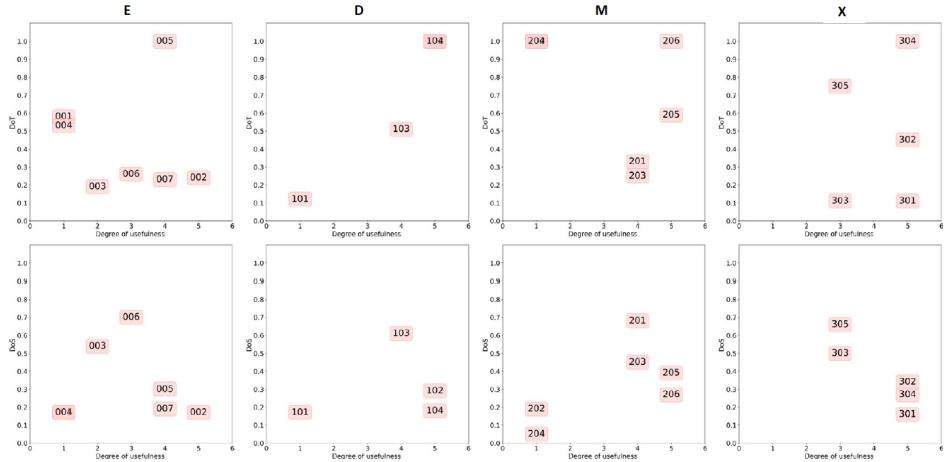


Fig. 9. Top row from left to right: degree of usefulness and degree of truth for linguistic summaries on euthymia, depression, mania, and mixed state from the sequential and compositional MLP model. Bottom row: degree of usefulness and degree of support for linguistic summaries for prediction of euthymia,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

Experimental results: PLENARY

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**

Experimental results: LS-FC

Table 3
Performance evaluation of DISSFCM and WeScatterNet for BIPOLAR and benchmark data streams.

Dataset	#classes	#obs	#batch	Algorithm	Labeling per batch (%)	Average accuracy (%)	Training time per batch (s)	Testing time per batch (s)
Bipolar	3	921 K	25	DISSFCM	100	83.75	188.42	0.06
					75	79.17	121.53	0.04
					50	79.17	114.17	0.04
					25	79.17	121.22	0.04
				WeScatterNet	100	71.95	19.81	1.72
					75	71.95	7.23	1.63
					50	71.95	4.27	1.61
					25	71.95	2.91	4.65
Higgs	2	11,500 K	198	DISSFCM	100	96.48	128.52	0.10
					75	96.43	161.82	0.45
					50	96.40	162.79	0.10
					25	96.38	175.62	0.10
				WeScatterNet	100	63.62	9.92	5.25
					75	63.59	10.2	6.14
					50	63.47	8.69	5.67
					25	63.26	6.41	5.1
Hepmass	2	11,000 K	189	DISSFCM	100	98.86	130.34	0.10
					75	98.77	153.07	0.10
					50	98.73	159.98	0.10
					25	98.68	206.79	0.10
				WeScatterNet	100	83.54	12.12	5.19
					75	83.49	12.68	5.79
					50	83.48	11.28	2.91
					25	83.45	9.21	2.74
RLCPS	2	5,000 K	90	DISSFCM	100	47.09	267.62	0.07
					75	47.06	225.89	0.08
					50	47.06	1269.68	0.35
					25	47.05	1147.14	0.08
				WeScatterNet	100	99.64	11.16	-
					75	99.64	10.41	1.92
					50	99.64	10.23	2.07
					25	99.64	9.11	1.89

[25] M. Pratama, C. Za'in, E. Lughofer, E. Pardede, D.A. Rahayu, Scalable teacher forcing network for semi-supervised large scale data streams, Information Sciences 576 (2021) 407–431

Experimental results: LS-FC

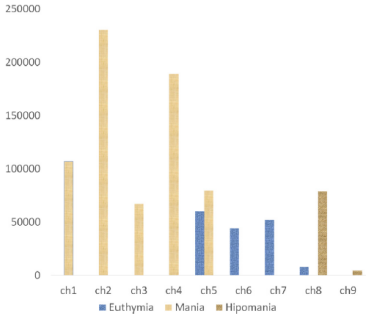


Fig. 3. Classes distribution through chunks.

Table 4
Accuracy values on test sets for chunks #5 and #8, varying the labeling percentages 25%, 50%, 75%, 100%.

Chunk	**Labeling 25%**	**Labeling 50%**	Labeling **75%**	Labeling **100%**
#5	0.45	0.55	0.58	0.58
#8	0.40	0.44	0.77	0.79

Experimental results: LS-FC

Table 5
Recall and *Precision* values on test sets for disease (D) and healthy (H) conditions, for chunks #5 and #8, varying the labeling percentages 25%, 50%, 75%, 100%. Note that for chunk #5: D is mania, whilst for chunk #8 D is hypomania.

Recall

	Labeling **25%**		Labeling **50%**		Labeling **75%**		Labeling **100%**	
Chunk	*H*	*D*	*H*	*D*	*H*	*D*	*H*	*D*
#5	0.64	0.30	0.68	0.44	0.45	0.66	0.40	0.71
#8	0.64	0.37	0.64	0.42	0.56	0.79	0.55	0.81

Precision

	Labeling **25%**		Labeling **50%**		Labeling **75%**		Labeling **100%**	
Chunk	*H*	*D*	*H*	*D*	*H*	*D*	*H*	*D*
#5	0.41	0.52	0.48	0.65	0.50	0.62	0.51	0.61
#8	0.09	0.9	0.09	0.92	0.21	0.95	0.22	0.95

Experimental results: LS-FC

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

Conclusions

- Grouping of low-level attributes into high-level information granules using linguistic summarization improves the over all explainability of the model results.
- Experimental evaluations confirmed that fuzzy linguistic summarization comple ments 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.

Future work

1. PLENARY has potential for further extensions and applications. In addition to summarizing the global model explanations, there is also a need to provide protoforms that allow for linguistic descriptions of **local explanations** in a synthetic way.
2. Creation of a dynamic approach to summarize high-level groups that are not homogeneous in terms of impact on the predicted class.
3. Other types of protoforms but also quantifiers and t-norms.
4. This paper also illustrates the need for more comprehensive **multi-object summaries that allow for effective assessment and comparative analysis of global model explanations from multiple predictive models**.

Thank you!

Katarzyna Kaczmarek-Majer

k.kaczmarek@ibspan.waw.pl

[0]Katarzyna Kaczmarek-Majer received funding from Small Grants Scheme (NOR/SGS/BIPOLAR/0239/2020-00) within the research project: "Bipolar disorder prediction with sensor-based semi-supervised Learning (BIPOLAR)". **http://bipolar.ibspan.waw.pl**