Information Sciences 614 (2022) 374–399

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

PLENARY: Explaining black-box models in natural language through fuzzy linguistic summaries

Katarzyna Kaczmarek-Majer [a,*], Gabriella Casalino [b], Giovanna Castellano [b], Monika Dominiak [c], Olgierd Hryniewicz [a], Olga Kamińska [a], Gennaro Vessio [b], Natalia Díaz-Rodríguez [d]

[a] Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland
[b] University of Bari Aldo Moro, Bari, Italy
[c] Institute of Psychiatry and Neurology, Warsaw, Poland
[d] University of Granada, Spain

ARTICLE INFO

Article history:
Received 10 March 2022
Received in revised form 17 August 2022
Accepted 3 October 2022
Available online 8 October 2022

Keywords:
eXplainable Artificial Intelligence
Linguistic summaries
Granular computing
Fuzzy linguistic descriptions
Machine Learning
Neural networks
Bipolar disorder

ABSTRACT

We introduce an approach called PLENARY (exPlaining bLack-box modEls in Natural lAnguage thRough fuzzY linguistic summaries), which is an explainable classifier based on a data-driven predictive model. Neural learning is exploited to derive a predictive model based on two levels of labels associated with the data. Then, model explanations are derived through the popular SHapley Additive exPlanations (SHAP) tool and conveyed in a linguistic form via fuzzy linguistic summaries. The linguistic summarization allows translating the explanations of the model outputs provided by SHAP into statements expressed in natural language. PLENARY accounts for the imprecision related to model outputs by summarizing them into simple linguistic statements and for the imprecision related to the data labeling process by including additional domain knowledge in the form of middle-layer labels. PLENARY is validated on preprocessed speech signals collected from smartphones from patients with bipolar disorder and on publicly available mental health survey data. The experiments confirm that fuzzy linguistic summarization is an effective technique to support meta-analyses of the outputs of AI models. Also, PLENARY improves explainability by aggregating low-level attributes into high-level information granules, and by incorporating vague domain knowledge into a multi-task sequential and compositional multilayer perceptron. SHAP explanations translated into fuzzy linguistic summaries significantly improve understanding of the predictive modelling process and its outputs.

1. Introduction

eXplainable Artificial Intelligence (XAI) embraces a plethora of intelligent methods that are enriched by the ability to describe and explain the decision made by the model in a way that can be easily understood by end-users. Unfortunately, common XAI techniques provide explanations in the form of visual descriptions (plots, heatmaps, etc.) that are hardly interpreted by common users. Accessibility to such explanations is limited to a technical audience which includes AI experts (developers and data/research scientists) and domain experts who have no technical background but can validate model

* Corresponding author at: Polish Academy of Sciences, Warsaw, Poland.
E-mail address: k.kaczmarek@ibspan.waw.pl (K. Kaczmarek-Majer).

https://doi.org/10.1016/j.ins.2022.10.010
0020-0255/

output and explanation techniques. Most XAI techniques do not use natural language as a universal means of communication to convey model explanations. Indeed, there is an abundance of works that derive natural language descriptions of image classifiers or rather image captioning models, but few attempts have been made to derive linguistic explanations for models learned from tabular data and/or derived from model-agnostic XAI techniques. Furthermore, although some works aim to visualize a saliency map over words [1] to render an explanation in natural language, most approaches centered on explainability focus on local explanations for question answering tasks or saliency analysis to highlight the importance of words based on attribution scores.

Another important aspect is that, in most XAI techniques, the features present are primarily taken into account when explaining their importance, while the missing features are discarded [2]. However, in real-world applications, the lack of a certain feature is also informative in determining a type of outcome class. For example, in supporting medical diagnosis, such as bipolar disorder prediction, a lack of a symptom, such as *anxiety*, is informative in determining the healthy class. Therefore, XAI methods should be equipped with mechanisms to account for the lack of features and symptoms and present explanations at different levels of granularity so that psychiatrists understand the comprehensive process of arriving at the prediction of healthy or pathological classes and symptoms.

The goal of this work is to increase the explainability of a classification model learned from tabular data by providing descriptions in natural language and addressing the uncertainty identified at various stages of the entire modeling process leveraging the expert knowledge available in the form of various levels of annotations associated with the data. To this end, we propose a three-step approach called PLENARY (exPlaining bLack-box modEls in Natural lAnguage thRough fuzzY linguistic summaries)[1] which manages to create an accurate and explainable classifier equipped with linguistic summaries handling imprecision at different levels. In the first step, we apply a neural network model to learn how to classify data by exploiting full supervision in the form of standard class labels, intended as a primary level of labels, as well as additional annotations that may come from domain or expert knowledge, which represent an intermediate labels. Taking advantage of the availability of these mid-level annotations mitigates some of the uncertainty related to the prediction outcome of the final classes. In the second step, we derive model explanations by applying the well-known SHapley Additive exPlanations (SHAP) technique as a versatile and model-agnostic XAI technique based on game theory that is well suited to providing model explanations from tabular data [2]. Since SHAP results are visual and imprecise, 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. Then, through linguistic summaries, it is possible to directly compare different SHAP outputs. It is worth noting that what we are proposing is not a hybrid fuzzy neural approach, but a multi-step pipeline for XAI, where the output of each step is the input to the next one in a cascade fashion. To validate our approach, we consider medical application domain where decision support AI needs to be validated by experts. Specifically, we consider a real-life use case of predicting bipolar disorder states using acoustic attributes of phone calls and a publicly available benchmark dataset on mental health surveys.

To summarize, the major contribution of this work is PLENARY: an explainable classification system enriched by fuzzy linguistic summarization that represents a complete classification framework for the explanation and summarization of outputs. Other major contributions incorporated into the proposed framework are the following:

- A compositional neural network architecture that learns a multi-output classification model via supervised learning based on a two-level hierarchy of labels associated with the data;
- The use of fuzzy linguistic summaries to convey imprecise explanations provided by SHAP in natural language expressions;
- The validation of the proposed framework in a medical diagnosis use case which is the monitoring of bipolar disorder to assess and explain classes related to mental health. The performance of PLENARY is also illustrated in the use case of treatment prediction based on mental health survey data.

The rest of this paper is structured as follows. Section 2 presents the bipolar disorder use case as the main motivating example. Section 3 reviews related literature. Section 4 presents the proposed method. Section 5 describes the experimental setup and discusses the results obtained. Section 6 concludes the paper and outlines the future directions of our research.

2. Motivation: a use case in bipolar disorder classification as a decision support system

In this section, we motivate our research by illustrating a typical case study in the context of psychiatric care. Specifically, we consider the problem of supporting the diagnosis of bipolar disorder (BD) state through the analysis of acoustic data from phone calls. Some progress has been made in the treatment of BD over the past decade; nevertheless, the diagnosis and monitoring of this disorder remains challenging. This is probably due to the still limited understanding of the nature of the disease and, consequently, the difficulty in predicting relapses. One issue that has received attention recently is a fundamental one: the classification of BD episodes.

[1] PLENARY repository:https://github.com/ITPsychiatry/plenary

2.1. The needs for explaining relations between attributes, symptoms, and states

In the classification of BD episodes, the interpretability of an AI model is desirable to gain the user's trust in the model thanks to the understanding of the working mechanism underlying the model supporting the psychiatrist's diagnosis. Moreover, the interpretability of the model can help the user to better identify any criticality in the data related to the specific subject case and improve the model. In particular, in psychiatry, many clinical decisions are made based on both objective factors and subjective considerations. Insights into the AI model decisions give the possibility to improve further human preferences and judgments, which are not readily incorporated into empirical data.

Another motivation for this research is the improvement of the management of uncertainty that comes at different levels when analyzing real-world data. In particular, for the BD use case, it is observed that the data annotations that are commonly applied in state prediction (*euthymia*, *depression*, *mania*, *mixed state*) are uncertain. Recent reviews of existing mental disorder classification systems, e.g. [3], have highlighted the need to revise the criteria for diagnosing affective states that have been promoted for many years. The current guidelines for mood disorders focus on a model that places much more emphasis on symptom groups [4]. These guidelines explicitly recommend a descriptive approach to the diagnosis of BD, based on particular symptoms. Grouping commonly observed symptoms into domains allows for a more accurate reflection of real-world observations. One of these proposed grouping strategies was based on distinct categories based on activity (i.e., general psychomotor activity, speech), cognition (including thought disorder and disorganization), and emotion (i.e., mood, irritability, disruptive behavior). A clear advantage of this approach is the appropriate description of mixed states (states that mix depressive and manic symptoms) and a better understanding of the sequence of symptoms.

Clinicians can obtain a certain amount of information about a patient's affective state by observing activity, mood, speed of thought and speech, and general appearance. However, these conclusions are difficult to quantify and difficult to compare objectively, for example in a subsequent visit. Common rating scales used in psychiatry, such as Hamilton Depression Rating Scale (HDRS) and Young Mania Rating Scale (YMRS), are based on a disease classification (ICD-11/DSM-V) which needs to be revised [4]. In particular, these rating scales are insufficient for motor symptoms, including activity, energy, and speech [5]. Another problem with scales is that these measures also depend on the interviewer's skill, as well as the patients' ability to describe their emotions. Nevertheless, changes in general activity, mood, and speed of speech are key signs of mood states. Therefore, real-time assessment of various objective parameters that reflect activity, mood, cognition, and speech may be an alternative to clinical assessment in BD patients. However, there is a lack of objective biomarkers, especially those that are easy to use in clinical practice. Thus, further research on BD should consider symptom groups rather than a set of strict criteria [4].

2.2. The needs for explaining high-level acoustic attributes of speech

Voice analysis promises to support the monitoring of mental illnesses [6]. Acoustic features such as energy in different bands or Mel-frequency parameters can be extracted from the voice common libraries, e.g., openSMILE[2]. Although the acoustic features of the voice extracted from smartphones, the so-called low-level attributes, allow for accurate prediction of BD states, they are difficult for medical experts to interpret. A typical doctor using a decision support system does not have specific domain knowledge of low-level acoustic descriptors. Psychiatrists are more inclined to make statements at the highest level of granularity so that they can link it to their prior knowledge from the study of the state of the art and observations made in clinical practice on the high-level features of speech. For example, [7] concludes that in the depressive state, the speech activity is reduced and pause-related voice features are intensified.

Finally, the extraction of acoustic attributes from speech is accompanied by several uncertainties. In particular, patients use devices with different microphone quality. The background noises recorded during the patient's phone calls are changing. Furthermore, some phone calls are not processed if the patient has disabled the smartphone application responsible for data collection. The proposed PLENARY approach takes into account high-level features of speech and annotations at different levels. It explains not only the states but also symptoms, thanks to the use of fuzzy logic and a multitask sequential and compositional multilayer perceptron.

3. Related work

In this section, we review relevant literature related to the three main components of our method, namely machine learning for tabular data, XAI techniques and linguistic summarization.

3.1. Machine learning for tabular data

Deep learning models have become the first choice solutions for image, video, and text data [8]. However, when it comes to more standard tabular data, their predominance is no longer maintained and tree-based methods are usually preferred,

[2] https://www.audeering.com/research/opensmile/.

especially Gradient Boosting [9]. In fact, these are generally better suited to mixed categorical and numerical data and are much less sensitive to feature scaling, and anomalies or noise in the data.

In the present study, we experiment with neural networks for the following main reasons. First of all, especially in the presence of large and/or highly dimensional continuous data, such as those collected by sensors or instruments, even the most classic multilayer perceptrons (MLPs) are still competitive in terms of predictive accuracy compared to standard machine learning models, and sometimes they are able to provide better results. For example, in [10] simple neural networks with few hidden layers and neurons have been successfully used for cataract detection, providing superior performance over many other machine learning algorithms. Similarly, in [11] some authors of this work have found that simple MLPs can outperform tree-based methods when asked to classify pediatric patients with multiple sclerosis, based on their miRNA expressions. In these cases, neural networks can directly learn new intermediate representations of the input data, effectively solving classification or regression tasks. To this end, in this paper, we have deliberately chosen a model architecture that is relatively simple and less prone to overfitting.

The second motivation for choosing neural networks is their high flexibility [12] and their ability to perform lifelong or continual learning. In fact, they allow for a non-linear topology, shared layers, and even multiple inputs or outputs. In this work, we will investigate a multi-output model with shared representation to exploit the "entanglement" between labels at different levels concerning the same patients.

3.2. eXplainable AI for tabular data

Using the data a model has been trained with can tell a lot about a model's behavior. This is a strategy often exploited by many of the existing XAI techniques [13]. These are intended to make it easier to trace and understand the logic of a model when resulting in a particular output for a given input. The XAI literature can be divided using several taxonomies [2], but mainly it could be split into *ante-hoc* methods, which inherently provide some level of interpretability, and *post hoc* methods, which, after having trained a black-box model, apply the XAI technique to display an explanation interface on top. Typically, in models where the input data provided to the model is tabular, explanations come in the form of an association of how the presence or absence of a feature affects model performance. The ante-hoc XAI methods inherently reach a certain level of interpretability during model building. These include white-box or gray-box models (e.g., decision trees and random forests) which typically exhibit greater interpretability at the cost of losing performance. To guarantee a proper balance between model accuracy and transparency, in this work we focus on rendering post hoc methods in natural language.

The goal of post hoc XAI techniques is to explain the outputs of models that are not interpretable by design. Hence, a post hoc method is typically applied as an additional step after building a model to derive some form of explanation. Most of these methods are model-agnostic, meaning they can be applied to any underlying machine learning model. One of the most widely used post hoc XAI methods is LIME (Local Interpretable Model-agnostic Explanations) [14] which explains the predictions of any classifier by computing importance scores of features based on a local approximation of the model around a given prediction. Another recent post hoc method is DeepLIFT (Deep Learning Important FeaTures) which has been proposed as an explanation method for deep learning models [15]. DeepLIFT decomposes the output of a neural model by back-propagating all neurons' contributions to each input feature. This is done by evaluating the difference between each neuron's output and a "reference" output and calculating the importance score based on this difference. Another framework is SHAP (SHapley Additive exPlanations) [16] which emerged from game theory and uses the notion of a fair payout; in other words, how to fairly assign responsibility to each feature for a model given the output. Given a model or data point prediction, the SHAP value computed for each feature represents the positive or negative contribution of that particular feature value to the final model result. In [16], SHAP has shown stronger agreement with human explanations than explanations extracted using other XAI methods (LIME and DeepLIFT).

Many explainability techniques such as SHAP normally have a technical audience as the target of the explanation (e.g., data scientists or developers). However, it is often difficult to translate graphical analysis into simple terms for a non-technical audience such as domain experts, decision-makers, or end-users (such as patients). Indeed, explanations of an ML model are diverse, can be complex to interpret, and are often difficult to align with the language spoken by domain experts or decision-makers. The "linguistic misalignment" that exists between algorithm explanations and human expert explanations makes evident the need for simplified ways to communicate the result of XAI techniques. The use of natural language can be a way to make explanations more compatible with those of domain experts. This is because ultimately models and decision support systems will need the verification and validation of these experts.

Other ways to help the compatibility of machine explanations with those of human experts are the use of explainable neuro-symbolic methodologies such as aligning model training with expert knowledge graphs [17]. Backing up AI models with an established knowledge base or scientific model is critical for implementing responsible AI systems, especially for critical applications in the medical domain [18]. Given the high flexibility of the SHAP method (see for example [19]), in this work we exploit SHAP to explain the output of the classifier developed to predict the final class. Since SHAP is suitable to be applied to intermediate features, the model outcome can be compositional [17]. This means we have two levels of explainable features that serve as the logic of the final class explanation. Using SHAP we can verify which features are the most relevant to predict the class (e.g., the BD state of a patient) and intermediate features (e.g., symptoms).

3.3. Linguistic summary of a model outcome

Applications in various domains have demonstrated improved understanding of large datasets through the use of natural language statements, e.g. [20,21]. The main objective of this paper is to mine the outputs of the XAI system to support non-technicians (e.g., medical experts) with linguistic summaries on the relationship between the importance of features and their impact on the model outputs. Within this research, the methodology of *protoforms* [22,23] and computations on fuzzy sets are adapted to describe the relation between groups of attributes and global explanations. Various extensions have been proposed in the literature for the form of linguistic descriptions and for the reasoning process itself; see, for example, the review by Boran et al. [24]. In [25], the authors introduce composite protoforms and ground them on the theory of natural language processing to allow for the identification of discourse relations in texts. In [26], multi-subject linguistic summaries for graph databases are introduced, instead of relational ones. Recently, in [27] the authors introduce and discuss the complex problem of plural referring expressions. There is also a group of works that focus on summarizing time-series data (e.g., [28]) or sequential data (e.g., [29]) to reflect their changing nature. Other methods use rules as an approach to get close to natural language explanations, for example LORE [30]. Some others aim to produce natural language processing of tabular-data predictive models [31]. These normally translate threshold-based conditions into if-then rules that put into context the key features that make a data assigned to given model output. For example, Generalized Linear Rule Model (GLRM) [32] is an XAI technique for producing global explanations of models that are weighted combinations of rules. Once the data scientist has chosen the variables to study, GLRM displays their linear relationship with the model and allows combining rules with linear terms to transform the contribution into rules of the type "*if months since most recent request* $<= 21$*, the model reduces the risk score of repayment by 0.3*". These rules can reflect whether the variable is positively or negatively affecting the prediction, the location and size of the non-linear changes, the relation with the prediction score, the significance of the variable relative to the prediction, and the direction of the correlation of the chosen variables. Regardless of the particular approach to deriving sentences in natural language, a powerful way to represent and process linguistic terms while preserving their human consistency is the definition of information granules [33]. Information granules can be informally defined as a collection of objects linked together by some closeness (resemblance) relation that makes the objects indistinguishable at a higher level [34]. As such, they can be regarded as abstract constructs that are well suited to represent linguistic term sets that describe phenomena in an easily understandable way. Therefore, the idea of information granulation aligns very well with the need to explain artificial models [35].

Being abstract entities, information granules are fuzzy rather than crisp, thus Fuzzy Set Theory (FST) can be a powerful tool for representing granules. FST offers a suitable mathematical framework for defining information granules that can be used within the "computing with words" paradigm [36]. According to this paradigm, propositions in natural language are translated into fuzzy constraints on the variables involved. FST provides a good theoretical background to represent perception-based information granules, which are easily associated with linguistic terms drawn from natural language. Offering an intelligible view of concepts through the use of a simplified natural language, FST is a natural candidate for designing explainable decision support models and linguistic terms described through fuzzy sets naturally lend themselves to XAI techniques. This is the main motivation for using fuzzy granules in our proposed approach.

4. The proposed PLENARY approach

Formally, 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 $\left\{\mathbf{y}_i^{(t)} \in \{0,1\}^t : \sum_{j=1}^{t} y_j^{(t)} = 1\right\}$. The information about this class represents the first (main) level of labels associated with the data that enable the application of any supervised learning method to derive a single-task classification model. 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 $\left\{\mathbf{y}_i^{(s)} \in \{0,1\}^s : \sum_{j=1}^{s} y_j^{(s)} = 1\right\}$. For example, in the motivating BD case study described in Section 2, each patient can be associated with two levels of labels. At first, the patient is associated with s symptoms whose values can range between 0 and a maximum depending on a clinical scale. Based on an estimate of these symptoms, a doctor typically assigns a diagnosis in one of t possible mental states. These mental states represent the main level of labels associated with a patient. Both symptoms and mental states can be jointly exploited to derive a multi-task predictive model.

The key idea of our method is therefore to exploit first-level and second-level labels together to allow the creation of a multi-output model capable of simultaneously predicting both targets $\mathbf{y}^{(s)}$ and $\mathbf{y}^{(t)}$. Given training data annotated with a two-level hierarchy of labels (symptoms and mental states), we develop a modeling methodology to create explainable classification models. Once the classification model has been learned from the data, the classifier result is explained via the model-agnostic SHAP technique and then linguistic summaries are created using fuzzy quantified sentences. In summary, the proposed PLENARY approach consists of three main sequential steps:

1. Creation of a compositional classification model via supervised learning based on a two-level hierarchy of labels associated with 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.

We now describe each step in detail. Then, we explain the evaluation procedure of the proposed PLENARY approach.

4.1. Multi-task sequential and compositional multilayer perceptron

We propose a *multi-output* sequential and compositional MLP, which is trained to simultaneously predict two different levels of labels (symptoms and mental states in our case study) associated with the same data. To this end, the network architecture sequentially combines two output layers to simultaneously predict both targets $\mathbf{y}^{(s)}$ and $\mathbf{y}^{(t)}$. We denote by $\widehat{\mathbf{Y}}^{(s)}$ and $\widehat{\mathbf{Y}}^{(t)}$ the two outputs of the network that approximate the two levels of targets, respectively, for all n training examples.

Formally, the *final* output of the model is:

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

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. $\widehat{\mathbf{Y}}^{(s)} \in \mathbb{R}^{n\times s}$ is the *intermediate* output of the network defined as:

$$\widehat{\mathbf{Y}}^{(s)} = \mathbf{H}\mathbf{W}^{(1)} + \mathbf{b}^{(1)},$$

being $\mathbf{W}^{(1)} \in \mathbb{R}^{h\times t}$ and $\mathbf{b}^{(1)} \in \mathbb{R}^{t}$ the intermediate output layer weights and biases, and $\mathbf{H} \in \mathbb{R}^{n\times h}$ being the output of a hidden layer made of h hidden units:

$$\mathbf{H} = \text{ReLU}\left(\mathbf{X}\mathbf{W}^{(0)} + \mathbf{b}^{(0)}\right),$$

where $\mathbf{W}^{(0)} \in \mathbb{R}^{d\times h}$ and $\mathbf{b}^{(0)} \in \mathbb{R}^{h}$ are the hidden layer weights and biases. Also, to reduce overfitting, a dropout layer (with a dropout rate of $p = 0.2$) is added next to the hidden layer. In this way, a random fraction of hidden activations is dropped out during training with probability p. The common ReLU activation function is used for the hidden layer, whereas a classic softmax activation function is used for the final output layer as the goal is to perform a *multi-class classification*. Hence, for the final output the network minimizes the cross-entropy loss function:

$$\mathcal{L}_2 = \mathcal{H}\left(\mathbf{y}^{(t)}, \hat{\mathbf{y}}^{(t)}\right) = -\sum_{j=1}^{t} y_j^{(t)} \log \hat{y}_j^{(t)},$$

where $\hat{\mathbf{y}}^{(t)}$ are the predicted class probabilities for each training example, and $\mathbf{y}^{(t)}$ is the one-hot ground truth vector of length t.

For the intermediate output, the activation function may vary depending on the nature of the mid-level annotations. If these annotations are available in the form of numeric values (e.g., symptoms are represented as ordinal values in our case study) no activation function is used: in such a case, this layer performs a *multi-output regression* and the associated loss function is the classic mean absolute error element-wise:

$$\mathcal{L}_1 = \text{MAE}\left(\mathbf{y}^{(s)}, \hat{\mathbf{y}}^{(s)}\right) = \left|\mathbf{y}^{(s)} - \hat{\mathbf{y}}^{(s)}\right|,$$

where $\mathbf{y}^{(s)}$ and $\hat{\mathbf{y}}^{(s)}$ are the ground truth and predicted values for each training example, respectively. If mid-level annotations are available in the form of class labels, then $\mathcal{L}_1$ is the cross-entropy as $\mathcal{L}_2$. Overall, using backpropagation, the sequential and compositional MLP network minimizes the composite loss function:

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2.$$

Fig. 1 schematizes this model architecture. In practice, the predictions provided by the intermediate output layer are used directly as new features for final classification. Since the two levels of labels are highly interconnected, this multi-task strategy aims to improve the accuracy and/or agreement of the explanations for the two different outputs. The optimization of the first output, in fact, is intended as an auxiliary regression task to support the main classification task.

4.2. Explaining the model output using SHAP

SHAP [16] is one of the most used model-agnostic methods and one of the most used for tabular data. This game theory-based XAI method computes SHAP values to appropriately allocate the payout associated with a prediction among features based on their contribution. To perform this assignment, the feature attribution is decomposed additively, as a linear model, to obtain the following explanation model g:

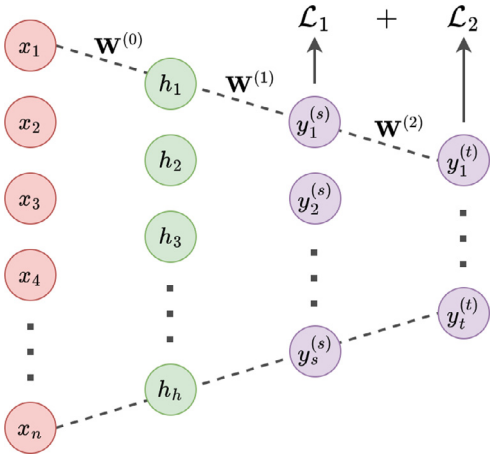

Fig. 1. Architectural diagram of the multi-task neural network.

$$g(x\prime) = \phi_0 + \sum_{i=1}^{M} \phi_i x\prime_i, \tag{1}$$

where M is the size of the coalition in terms of input features, $x\prime \in \{0,1\}^M$ indicates the presence (1) or not (0) of that feature in the given coalition, and $\phi_i \in \mathbb{R}$. This contribution is called the SHAP value of the feature i and in its definition [16] it considers the difference between a prediction model using feature $i, f_{S\cup\{i\}}(x)$, and another that does not use it, $f_S(x)$, where S is a coalition or possible subset of features.

A disadvantage of computing SHAP values is that they can change their value depending on the order of coalition selection [18]. However, obtaining exact SHAP values can be computationally very expensive because of the combinatorial explosion of every possible coalition of subsets of features. This is why in practice some approximation is computed instead, e.g. by performing random samplings of the possible sets S or Monte Carlo approximations of a set of coalitions to compute the average.

An illustrative example of SHAP analysis in a global explanation (summary) plot is in Fig. 2. Each point in the figure represents a classified data point and the color code represents its range of feature values. SHAP presents the model output for a given class (here depression diagnosis) as an inverted pyramid of the most contributing features to that class. This means that the features with the highest contribution, i.e. with the highest absolute value, will be placed at the top, and those with a lower absolute value will be placed from top to bottom. The positive contribution towards that class is shown on the positive side of the X axis (representing positive SHAP values); while the negative side of the axis represents a negative contribution or the contribution of those features against the prediction of that class.

4.3. Creation of linguistic summaries based on fuzzy quantified sentences

Linguistic summaries are descriptions in natural language that summarize large numeric datasets. Within this paper, the main purpose of constructed linguistic descriptions is to linguistically summarize:

1. The relation between the input features and the impact on the prediction of the classes, i.e. the main level of labels (BD classes in our case study). For example, *Among records that contribute positively to predicting depression class, most of them have voice quality-related features at low level* [$DoT = 1.0$], where DoT stands for the degree of truth and measures the validity of this sentence. In this work, linguistic summarization based on Shapley values describes not only the positive contribution to the prediction of a class but also whether an attribute is against predicting a particular class or if it remains unclear.
2. The relation between the input features and the impact on the prediction of the intermediate level of labels (e.g., symptoms in BD case study). For example, *Among records that contribute positively to predicting decreased activity symptom, most of them have quality-related features at high level* [$DoT = 0.65$].

$O = \{o_1, o_2, \ldots, o_b\}$ is a set of objects (e.g., speech signal extracted from phone calls). The attributes $\mathcal{A} = \{a_1, a_2, \ldots, a_r\}$ (e.g., loudness of speech) measure their properties. Next, the linguistic term set $l_{a_i} = \left\{l_1^{a_i}, \ldots, l_{k_{a_i}}^{a_i}\right\}$ (e.g., *high loudness*) is defined for each attribute from $\mathcal{A}$. We use type-I fuzzy sets to describe linguistic terms, and algorithm [37] for heuristic tree-based search across all linguistic term sets and attributes.

Following [22], a linguistic summary ($\mathcal{LS}$) based on an extended protoform is defined as:

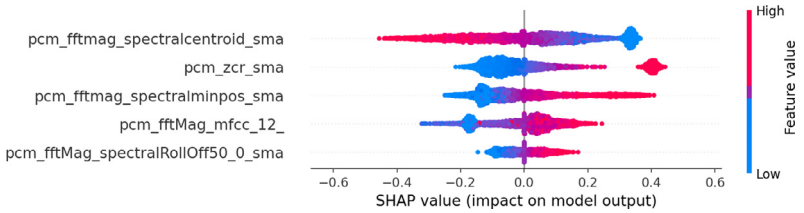

Fig. 2. Illustrative example: summary plot of Shapley values explaining the output of a simple model in our case study with top 5 most contributing features.

$$\mathcal{LS} = LS(Q, R, P) = \textit{Among } R \textit{ objects from } O,\ Q \textit{ have } P\ [DoT] \tag{2}$$

having the quantifier Q (e.g., *most recordings*), the qualifier R (e.g., *high level of the loudness feature of speech*), the summarizer P (e.g., *high level of the Shapley values*), and $DoT \in [0, 1]$ that is the degree of truth of the sentence. Thus, the attribute properties (e.g., *low* loudness, *most* records) are linguistic terms modeled as information granules which are represented by fuzzy numbers. We build triangular fuzzy numbers based on quartiles derived from the data as depicted in Table 1. For example, the *medium* terms are expressed as triangular fuzzy numbers $[Q_1, Q_2, Q_3]$.

As an illustrative example, let us consider we aim to create the linguistic summary based on the following protoform: *Among records that contribute against predicting depression class, most of them have spectral centroid feature at high level.* A fuzzy number describing the quantifier *most* needs to be created. Next, the respective fuzzy numbers describing the *low*, *medium*, and *high* levels of the spectral centroid acoustic feature have to be created. Also, we create fuzzy numbers to describe the *positively contribute to prediction, around zero* and *against predicting* based on the Shapley values. Fig. 3 shows an illustrative example of such linguistic variables. Finally, the number of final linguistic summaries can be very large or difficult to parse as some may contain alternative sentence terms. A common way to address this issue is to apply deletion-based sentence compression techniques [38]. Along with this idea, we include in PLENARY an automatic post-processing filtering step that selects the top-most certain sentences (i.e., those with the *DoT* greater than a threshold) to reduce the number of sentences presented to an expert. This filtering function can be applied according to a quality criterion or multiple criteria.

4.4. Evaluation of SHAP-based linguistic summary explanations

In this section, we describe the evaluation approach starting from the single sentences and, subsequently, we assess the quality of linguistic summaries made up of groups of sentences.

4.4.1. Evaluating individual sentences

One of the earliest and most popular measures of quality of a linguistic summary $\mathcal{LS}$ (Eq. 2) is the degree of truth (DoT), defined as:

$$DoT(Q, R, P) = \mu_Q\left(\frac{\sum_{i=1}^{n}\left(\mu_R(x_i) * \mu_P(x_i)\right)}{\sum_{i=1}^{n}\mu_R(x_i)}\right), \tag{3}$$

where $*$: [0,1] × [0,1] →[0,1] is a triangular norm (t-norm for short) and $\mu_Q, \mu_R, \mu_P : \mathbb{R} \to [0, 1]$ are the membership functions of the fuzzy numbers representing the quantifier Q, qualifier R, and the summarizer P, respectively. We also adopt the degree of support (DoS) and the degree of focus (DoF) [39]. 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. The degree of focus of a linguistic summary $\mathcal{LS}$ informs about coverage of objects that meet the condition expressed by the qualifier R. It is defined as follows:

$$DoF(R) = \frac{1}{n}\sum_{i=1}^{n}\mu_R(x_i), \tag{5}$$

where $\mu_R : \mathbb{R} \to [0, 1]$ is the membership function of the fuzzy number representing R. Furthermore, to complement the objective quality measures, we introduce expert-based evaluation at the sentence level. AI systems often report desirable qualitative properties such as satisfaction, confidence, and/or trust in the explanation (usually involving the use of question-

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*

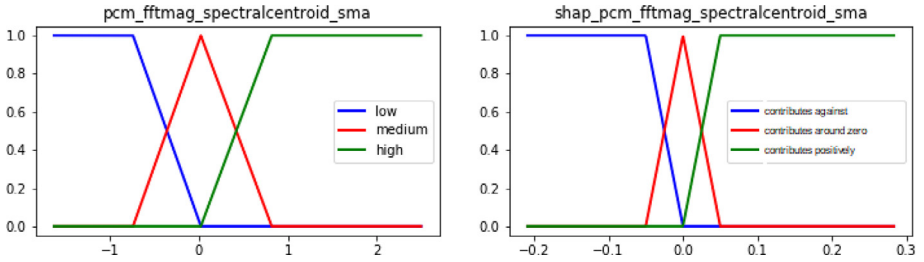

Fig. 3. Illustrative example of linguistic variables describing the spectral centroid acoustic feature and the SHAP values describing its importance.

naires for the explanation demanding audience [40]). Among the various properties, consistency, reliability, relevance, and usefulness are some of the most common XAI metrics [41,42,40]. However, only the latter resulted to be intuitive for the domain experts for the psychiatric use case considered. Indeed, the expert-based criteria proposed to evaluate explanations are often strongly conditioned by the use case to which they refer [43,2,44]. For the sake of simplicity, we have limited ourselves to the degree of usefulness (*DoU*) with the aim of quantifying how useful the sentence explanation is from the perspective of domain experts, e.g. in our case, psychiatrists. Ideally, such expert-based *DoU* scores should be provided by multiple domain experts. Therefore, another metric to consider is the *reliability* defined as the degree of weighted agreement among the raters' ratings [42]. Finally, atomic or low-level explanations (e.g., evaluating contrastive rule-based and example-based explanations) have not always proved sufficient without exposing a clarification of the overall rationale of the complete AI system behaviour [45]. Thus, apart from sentence-level measures, we discuss the evaluation of groups of summaries. First, objective measures such as consistency are recalled; second, we discuss expert-based measures for the group of summaries.

4.4.2. Evaluating groups of sentences

The set of summaries is assumed to be consistent when it satisfies the non-contradiction and double negation properties [46]. Non-contradiction implies that linguistic summaries made up of contradicting terms have a complementary degree of truth. Formally, contradictory forms of a summary based on extended protoform $\mathcal{LS}$ are defined as follows:

$C1(Q, R, P)$ = Among R objects from O, $\neg$ Q have P
$C2(Q,R,P)$ = Among R objects from O, Q have $\neg$ P.

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

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

The double negation property states that $DoT(D(LS)) = DoT(LS)$. For more details on the constraints on the definition of quantifiers and qualifiers, we refer the reader to [46]. Let us now consider the following sentence as an example

*LS*1 = *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:

*C*1 = *Among records that contribute positively to predicting euthymia class,* ***a few*** *of them have energy-related features at* ***low*** *level.*
*C*2 = *Among records that contribute positively to predicting euthymia class,* ***most*** *of them have energy-related features at* ***high*** *level.*

Also, 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.*

Finally, the objective measures are confronted with the expert-based evaluation at the group of summaries level. In particular, to fulfill the needs for causability [47] in model explanations [44], we measure the quality of the group of explanations via the system causability scale (SCS) [43]. We assess the quality of the PLENARY system' outcomes with the following quesitons:

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.*

Furthermore, sentences must be well understood by users, and therefore must comply with standard measures of communication such as Grice's maxims [48]. These were proposed by linguist Paul Grice as an attempt to ensure that textual communicative efforts are effective. Grice's maxims are important criteria to comply with because they are general principles that effective communications should exhibit. We introduce the following questionnaire inspired by Grice's maxims to further assess the quality of PLENARY system's outcomes:

1. *The group of sentences provides all the information we need, and no more* (maxim of *quantity*).
2. *The group of sentences provides truthful statements and avoids providing information not supported by evidence* (maxim of *quality*).
3. *The group of sentences is relevant to the discussion objective of explaining the model* (maxim of *relation*).
4. *The group of sentences is clear, and as brief and orderly as possible, avoiding obscurity and ambiguity* (maxim of *manner*).

Both questionnaires have been assessed with Likert scale ratings (1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, 5 = strongly agree). Finally, we apply statistical tests to compare whether there are significant differences between the characteristics obtained by the various predictive models. The statistics used for the comparison of the considered characteristics are calculated using partially the dependent raw input data. Moreover, these data are not normally distributed. Therefore, for the comparison of the considered approaches, we must use non-parametric (distribution-free) statistical tests for possibly dependent data. For this reason, we have chosen the Wilcoxon signed-rank test. The p-values of these tests have been calculated from the asymptotic distributions of the test statistics. Their values greater than 0.05 indicate that there is insufficient statistical evidence to reject the hypothesis that the characteristics being compared have the same distributions.

5. Experiment

The proposed PLENARY approach was validated in two real-world use cases: explaining the classification of bipolar disorder and mental health survey. The experiments were devoted to investigating whether: (1) SHAP explanations translated into fuzzy linguistic summaries improve understanding of the model outputs and the modeling process itself; (2) a compositional neural network architecture that learns a multi-task classification model via supervised learning based on a two-level hierarchy of labels can effectively incorporate domain knowledge into the predictive model; (3) the introduction of specialist knowledge in the form of middle-layer labels affects performance in terms of prediction accuracy or explainability of model outcomes.

5.1. Case study to explain bipolar disorder classification

The first set of experiments was performed on the BDMON dataset collected from four patients affected by bipolar disorder and monitored for a period of 9 months between February and October 2018 within a prospective study (see [49] for the protocol of this study). BDMON concerns a 4-class classification corresponding to the BD states:

- Euthymia (class 0);
- Depression (class 1) characterized by depressive symptoms, such as decreased mood and energy, anhedonia, anxiety;
- Mania (class 2) characterized by manic symptoms, such as unusually increased energy, decreased need for sleep, euphoria, excessive talking;
- Mixed state (class 3) characterized by the coexistence of manic and depressive symptoms.

The BDMON dataset consists of 86 speech acoustic parameters extracted from voice data using openSMILE [50] and 29 attributes collected from the psychiatric assessments of the intensity of depressive and manic symptoms. The acoustic attributes were divided into short frames of 20 ms, then calculated by omitting the interlocutor's speech. The BDMON dataset used contains only visit data, thanks to which the dataset has valid labels. Table 2 presents a short quantitative summary of the BDMON dataset with the labels obtained along with an overview of the data size. It can be seen that the number of instances characterizing the maniac class is under-represented and is over 20 times smaller than the instances representing depression. Moreover, mania was only observed in one patient. On the other hand, mania is included in the mixed state, which is already present in two patients in a larger number of rows.

In addition to the class labels corresponding to mental states, domain knowledge was also available in the form of psychiatric assessments collected during the patients' visits. During one visit, the doctor interviewed the patient and assessed her/his state using 18 questions concerning depressive symptoms derived from the Hamilton Depression Rating Scale (HAMD) and 17 questions associated with the intensity of manic symptoms derived from the Young Mania Rating Scale (YMRS). Next, questions of both scales were grouped into symptoms (as described in Table 3) and these symptoms were then treated as middle-layer labels. The first column describes the symptom group, the middle column further describes the symptom group and related questions to be evaluated, and the last column indicates the maximum number of data points from that scale. The idea of introducing symptoms as an intermediate level of annotations for BD data is one of the main novelties of this work as it has not been applied before in the context of AI decision support systems for monitoring mental disorders.

Furthermore, it is observed that low-level acoustic data are difficult for psychiatrists to interpret. For this reason, we have offered assistance by grouping the low-level acoustic descriptors into more interpretable objects, such as loudness and pitch of speech. Thus, similar to [29], in this paper we group the acoustic features into more interpretable high-level features, i.e. energy-related features, spectral-related features, pitch-related features, and voice quality-related features. Table 4 shows the amount of low-level acoustic attributes included in each group and Table 13 and Table 14 in the Supplementary Material show detailed grouping and statistical characteristics of the acoustic dataset.

5.1.1. Accuracy evaluation

In this section, we present the evaluation of the proposed approach against baseline approaches in terms of accuracy. The following methods were considered:

- XGBoost, which is trained to classify the bipolar state (primary level of labels) and is considered the basis for benchmarking performance;
- Single-task MLP, which is trained to classify the bipolar state (primary level of labels);
- The proposed multi-task sequential and compositional MLP, which is trained to perform two recognition tasks in sequence, namely symptom prediction and BD state classification.

Table 5 reports the results obtained for the BD state classification by the three methods. All methods have been tested using the same data. It can be seen that both MLP models achieved similar accuracy, which at the same time is much higher than the XGBoost baseline. Notably, the XGBoost model is not able to identify three of the four bipolar disorder states. The healthy class (euthymia) has a recall of 0.69, so the model can identify nearly a third of healthy patients. However, the metrics are very low, especially in a medical context, where we are more interested in correctly identifying the presence of unhealthy states.

On the other hand, while the MLP models return higher results than the ensemble method, quantitative evaluations do not suggest improvements from the injection of expert knowledge in the form of an additional layer of labels. We can observe that the best precision and recall values are obtained for the euthymia state. The MLP models are also able to detect quite well the depression and mixed states (precision 0.60 and recall 0.70). These are quite high values if we consider that this is a multi-class classification problem. The models are unable to identify samples belonging to the under-represented class (corresponding to the mania state). This suggests that the neural networks need more samples to learn how to discriminate this class.

5.1.2. Explanation of predictive model results using SHAP

Fig. 4 shows the global explanation of SHAP in terms of the four bipolar states. The color bars represent the average impact of the twenty most important features on the magnitude of the model output, obtained with the baseline and the sequential and compositional MLP model. The SHAP global plot for all classes in Fig. 4 shows that the class to which most of the features contribute is the mixed state. This shows that the more crisp disease states (other than the mixed state) are responsible for accounting for a smaller subset of features contributing to these classes (euthymia, mania, and depression). However, the purple bar shows an exception, and this is for the mania state, for which, in many cases, less globally relevant features become important for the model to diagnose the case as maniac (see the second topmost quarter of features where purple is the predominant color in the bars). This correlates with our expert explanations which pointed out that although energy is normally more important than pitch features when diagnosing manic states, loud speaking is indeed the key element, something that is not the case or not considered a critical feature when diagnosing states such as depression, mixed, or

Table 2
Summary of the BDMON dataset presenting patient characteristics, their BD states, psychiatric ratings (HAMD and YMRS), and phone calls recorded in the ground-truth considered.

Patient id	Visit	BD state	HAMD	YMRS	# instances	# calls
[0.5ex] 1	2	mixed	12	35	3.64 M	188
1	3	euthymia	1	4	1.95 M	142
1	4	mixed	8	10	0.76 M	73
2	1	euthymia	4	1	0.15 M	20
2	2	euthymia	2	0	2.63 M	147
2	3	depression	13	0	2.84 M	71
3	2	euthymia	1	5	1.06 M	57
3	3	mixed	9	18	1.11 M	75
3	4	euthymia	2	1	2.17 M	91
3	5	depression	11	3	1.48 M	69
4	1	depression	13	4	0.71 M	44
4	2	depression	10	2	0.65 M	84
4	3	depression	12	3	1.88 M	115
4	4	depression	13	3	1.42 M	82
4	5	mania	2	8	0.32 M	32
				TOTAL:	22.77 M	1290

Table 3
Description of symptoms (middle-layer labels). The max # points reflect the possible maximum points describing the maximum intensity of the symptom, based on the rating scales of the psychiatric assessment.

Symptom	Description	Max # points
Anxiety	Anxiety, fear mental symptoms	6
Decreased activity	Inhibition, work, and interests	8
Decreased mood	Criticism, depressing mood, feeling guilty	10
Disorganization	Appearance, formal thoughts disorder, thoughts disorder, view	16
Elevated activity	Increased activity, speech	12
Elevated mood	Elevated mood	4
Irritability	Irritability, destructive behavior	16
Sleep disorder	Early awakening, intermittent sleep, sleep disorder, sleep	10
Somatisation	Fear somatic symptoms, generic somatic symptoms, hypochondria	10
Suicide	Suicide tendencies	4

Table 4
High-level acoustic features applied to group low-level parameters for fuzzy linguistic summarization.

High-level acoustic features	# low-level features
Energy/loudness-related features	36
Spectral features	33
Pitch-related features	12
Voice quality-related features	5

euthymia. Moreover, unlike the quantitative results in which the mania state was not identified, these plots suggest that for the sequential and compositional MLP model (Fig. 4 (b)) several features contribute significantly to modeling the output for the mania state (purple one).

Comparing top and bottom plots in Fig. 4 we can see that a bimodal-like distribution in terms of feature relevance is observed in the sequential compositional MLP model. This might better capture the importance of certain types of acoustic features that contribute to two different classes in a different manner depending on the class being predicted, as experts corroborate, with regards to features related to high pitch value and quality of voice. For some classes, energy is more important than pitch, but for other classes, quality of voice and loud speaking becomes the most important (e.g., in maniac state). This may be explained by the smoothness of this bimodal-looking distribution in the sequential compositional MLP model. This smoothness becomes a little less obvious if we look at the baseline model, which has less information to perform the classification task.

Figs. 5 and 6 show global SHAP explanations, in terms of single data points, for the four classes euthymia (a), depression (b), mania (c), and mixed state (d), respectively for the baseline and the sequential and compositional MLP model. If we compare the graphs obtained with the two models, we can see how the sequential and compositional MLP model provides a smoother and easier to interpret SHAP distribution of the data points. The sequential and compositional MLP shows in the first plot (class 0) that the top 4 features changed in terms of feature value (color reversed in the same top features, except for the top 3 features contributing to class 0). It is also interesting to see that the rankings are not necessarily pre-

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			

served, e.g. alpharatio_sma3 in the baseline model at position 20 in the feature contribution ranking disappears from the ranking in the sequential and compositional MLP model. Also, there appears to be better feature disentanglement in terms of less color mix-up at each side of the x-axis, i.e. a more homogeneous color coding transition appears to occur in the sequential and compositional MLP against the baseline model, which seems natural, as the baseline does not perform any intermediate prediction steps to aid in symptom detection before deciding the class of the disease state.

Fig. 7 shows the global model SHAP analysis obtained from the first output of the sequential and compositional MLP model, i.e. the first labels pertaining to symptoms from low-level speech features. As discussed above, clinicians use ten different symptoms to derive patient states (see Table 3). Unlike the other methods, the sequential and compositional MLP model is also able to predict these symptoms from the collected acoustic features, which are more significant than the low-level features for clinicians. It can be observed that clear color distinctions are preserved, which means that high values of those speech features are positively correlated with/against the detection of that particular symptom. It would have been disturbing to have mixed colors on one side of the SHAP plot, which is not the case in our SHAP plots obtained for classes (anxiety, irritability, somatization, suicide, sleep disorder, etc.).

Finally, for each symptom in Table 3, a global SHAP explanation has been generated with the sequential and compositional MLP. Fig. 8 shows the SHAP plots for the symptoms of elevated activity (a) and decreased activity (b). The presence of several outliers makes the graphs flattened. However, it is interesting to note that the low-level pcm_LOGenergy_sma is the most significant feature for the decreased activity symptom and negatively contributes to modeling it. Medical experts have included this feature in the energy/loudness-related features, so the higher the value of this feature, the lower its contribution to the symptom. To confirm this, we can find the same feature with inverted colors for the elevated activity symptom. However, even though we have used symptoms as high-level labels since they are close to the way clinicians think in diagnosing BD, the SHAP graphs are still not very intuitive, especially for them. Thus, in the next subsection, we summarize these groups of explanations with the use of fuzzy linguistic summarization.

5.1.3. Linguistic summaries of global model explanations

We now present linguistic summaries derived from PLENARY. Table 6 collects the linguistic summaries for the prediction of euthymia (healthy class, formerly referred to as class 0), depression (class 1), mania and the mixed state based on the sequential and compositional MLP.

Summaries describe either a *positive* contribution to predicting a particular class, a contribution *against* predicting a class, or an unclear relation between the attribute and the model output (expressed as the *around zero* contribution to the prediction of a particular class). In addition, summaries inform about the direction of these relations, and in particular whether *low* or *high* values of the attributes are meaningful for the prediction. For example, from Table 6 we see that the following summary is valid: *Among records that contribute positively to predicting the depression class, most of them have spectral features at a high level* $[DoT = 1.0, DoS = 0.29, DoF = 0.31]$ (*Id102*). The degree of support informs how many objects in the dataset are covered by this summary, so the condition of *contributing positively to predicting depression* and *spectral features at high-level* is satisfied. The degree of support is complemented by the degree of focus which informs about the overall coverage of the qualifier feature, that is *contributing positively to predicting depression class*. Let us now look at the summary related

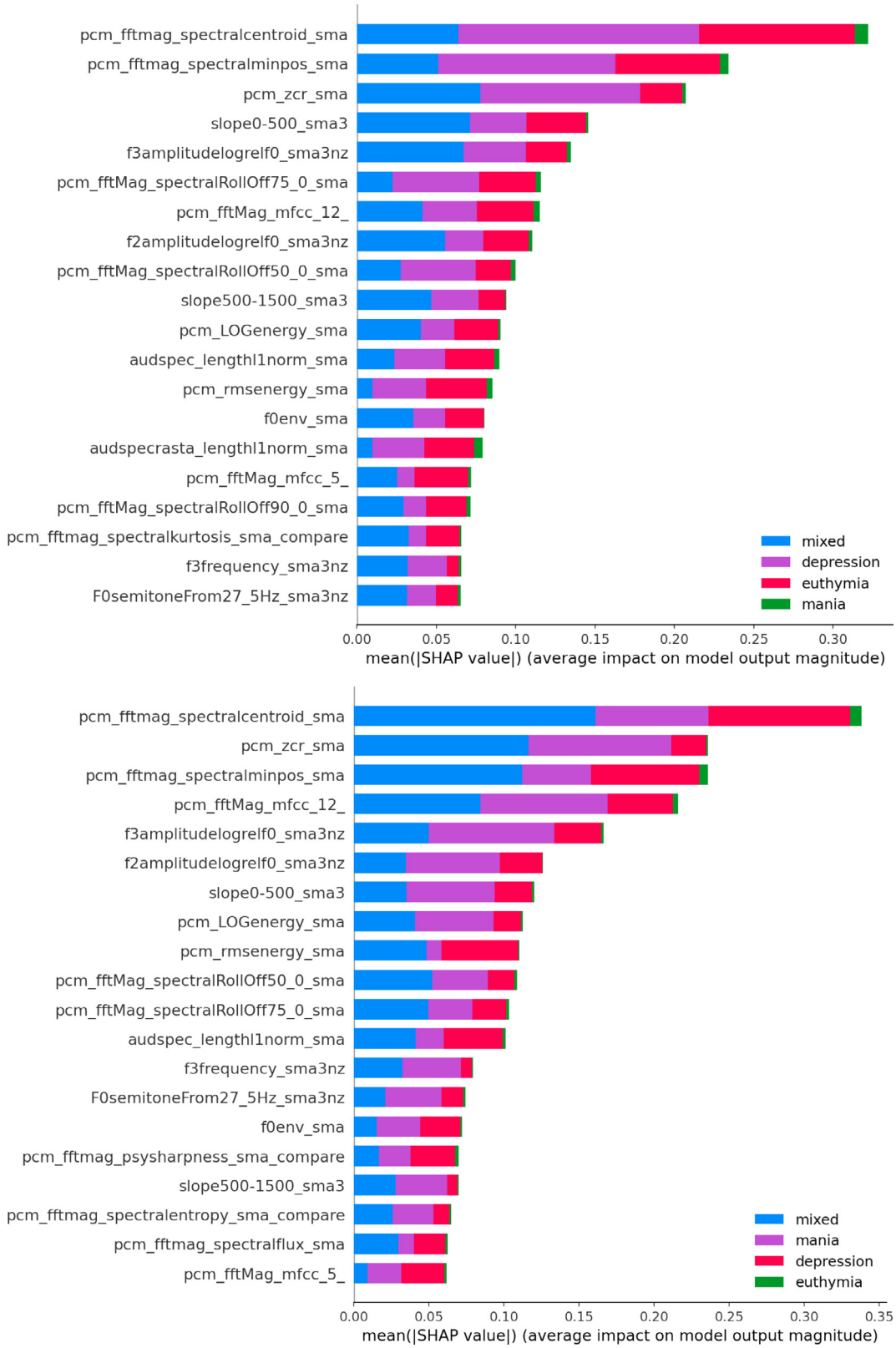

Fig. 4. Global model SHAP analysis for disease state prediction with a) the baseline model, and b) the sequential and compositional MLP model. The bi-modality of feature contributions is more smoothly assessed and appreciated with our sequential and compositional MLP, demonstrating the usefulness for the interpretability of the results when having a two-step compositional approach to classification based on recognizing symptoms first.

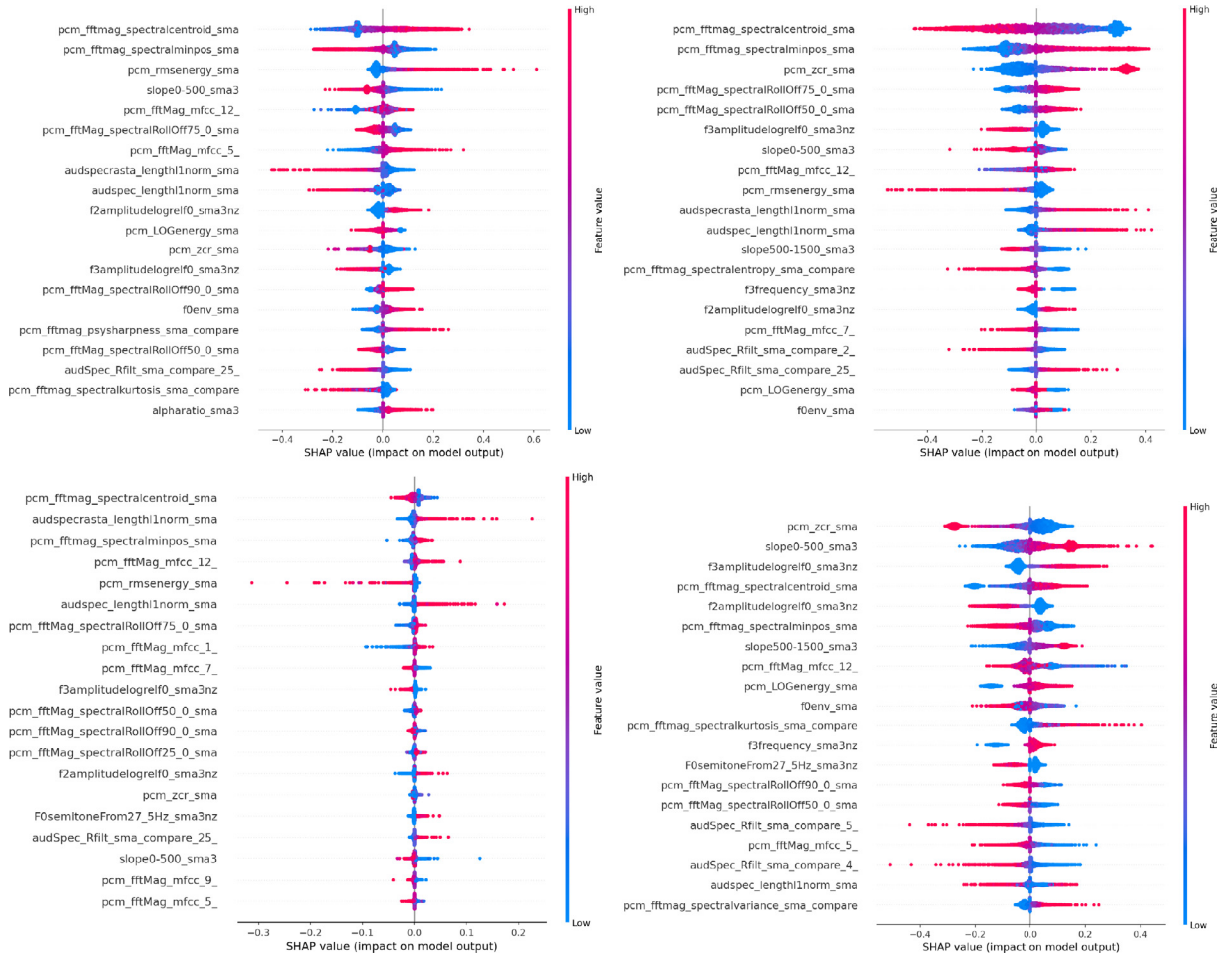

Fig. 5. From left to right and top to bottom: 20 most contributing features to the baseline MLP model for a) class 0 (euthymia), b) class 1 (depression), c) class 2 (mania), d) class 3 (mixed state). The mixed state corresponds to symptoms of depression and mania together.

to another group of parameters which is the voice quality. We observe that *Among records that contribute positively to predicting depression class, most of them have voice quality-related features at a low level* $[DoT = 1.0, DoS = 0.18, DoF = 0.31]$ (*Id104*). While the support of this summary is small (and amounts to 0.18), this summary is also true.

Linguistic summaries allow us to reason about the overall contribution of the parameter groups and, in doing so, better understand how the models work at a global level. For example, the following linguistic summary *Among records that contribute around zero to predicting the depression class, most of them have energy-related features at high level* $[DoT = 0.12, DoS = 0.17, DoF = 0.06]$ (*Id101*) is true only to some extent. We see that it is the only summary in this group about the energy-related features and its relation with the depression class (class 1). The analysis of the SHAP values from the previous section revealed that observing the low-level attributes in Fig. 6 (top 20 features), the following five belong to the considered energy/loudness high-level group: *pcm rmsenergy sma*, *audspec lengthl1norm sma*, *audSpec Rfilt sma compare 0–25*, *pcm LOGenergy sma* and *pcm zcr sma*. And, out of those five, only for the *rmsenergy* attribute the Shapley values confirm that the levels of *rmsenergy* contribute positively to predicting the depression class. Considering that there are 36 low-level acoustic attributes in the energy-related features group and only one has a significant impact, it is not surprising that the summary *Id101* with the considered quantifier *most* is fairly true. PLENARY allows the generation of summaries on single sentences; however, these summaries need to be further analyzed to draw meaningful conclusions. Further research will address this issue and extend the PLENARY approach by generating summaries of the behavior or patterns identified in the non–homogeneous high-level parameter groups. Furthermore, the summaries provided by PLENARY allow us to effectively compare the impact of various high-level groups on class predictions. For example, from Table 6 it can easily be concluded that pitch-related features are more significant for the prediction of mania or mixed state than for euthymia and depression. We also identify summaries that complement each other. For example, *Among records that contribute around zero to predicting mania, most of them have energy-related features at low level* $[DoT = 1.0, DoS = 0.19, DoF = 0.03]$ (*Id001*). At the same time, we see that *Among records that contribute against predicting the mania class, most of them have energy-related*

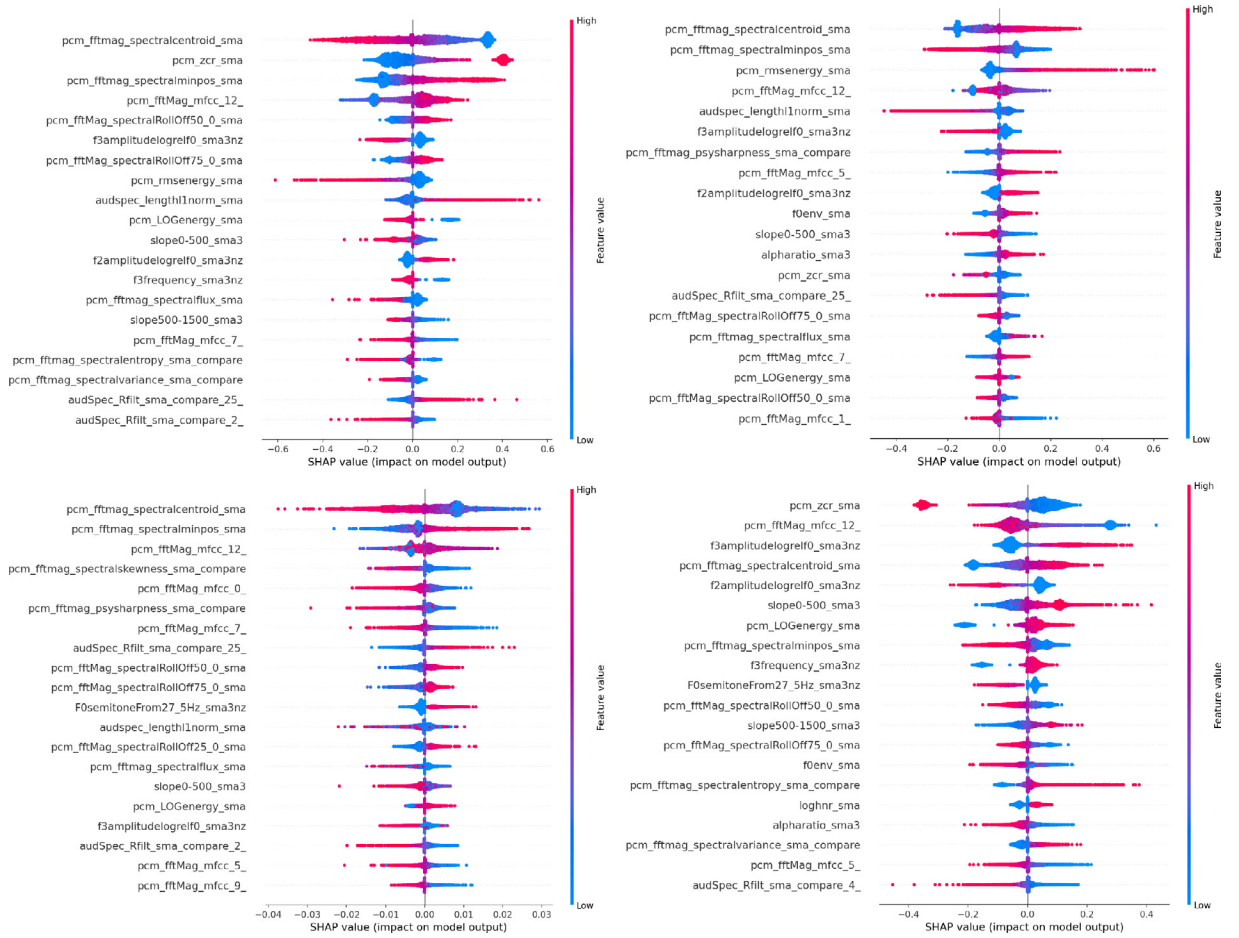

Fig. 6. From left to right and top to bottom: 20 most contributing features to the sequential and compositional MLP model for a) class 0 (euthymia), b) class 1 (depression), c) class 2 (mania), d) class 3 (mixed state). The mixed state corresponds to symptoms of depression and mania together, and this plot accordingly shows the consistent and clean contribution of each feature. The low values of the top two speech features contribute most to this state, while high values of the top 3 and 4 features contribute positively to this class.

features at low level $[DoT = 0.33, DoS = 0.68, DoF = 0.73]$. The aforementioned summaries, while valid only to a certain extent, complement each other.

Table 7 shows validated linguistic summaries of the relation between high-level acoustic attributes and the prediction of two exemplary symptoms, namely elevated activity and decreased activity. We can observe in Table 7 that the spectral-related features are considered in the summaries related to the elevated activity symptom. At the same time, the spectral-related features are not valid for predicting the decreased activity symptom. This result is consistent with the domain knowledge and the fact that the elevated activity is usually directly related to the mania state. The main purpose of symptom-based summaries is a better understanding of the predictive modeling process. For example, we can conclude from this table that *Among records that contribute positively to predicting decreased activity symptom, most of them have spectral-related features at low level* $[DoT = 0.81, DoS = 0.26, DoF = 0.31]$ (*Id401*). This summary is surprising in view of the summary *Among records that contribute positively to predicting euthymia, most of them have spectral-related features at low level* $[DoT = 1.0, DoS = 0.30, DoF = 0.21]$ (*Id005*). The linguistic summaries for symptoms generated by PLENARY clearly extend understanding of the modeling process and the importance of features.

5.1.4. Medical perspective on the results of PLENARY in the BD use case

In this section, we describe and discuss the expert-based evaluation of the results of PLENARY at the group of sentences level. First, from a medical point of view, the information about symptoms (Table 7) significantly extends approaches focused on the classification of mental states. For example, mood and activity do not always go hand in a given illness episode, and such knowledge could help predict the most varied types of deterioration in BD, not just the classic forms of depression or mania, as in the label-based approach. This is particularly helpful in the prediction of mixed states and subsyndromal conditions, and, consequently, in the selection of a pharmacological strategy appropriate to the given symptoms.

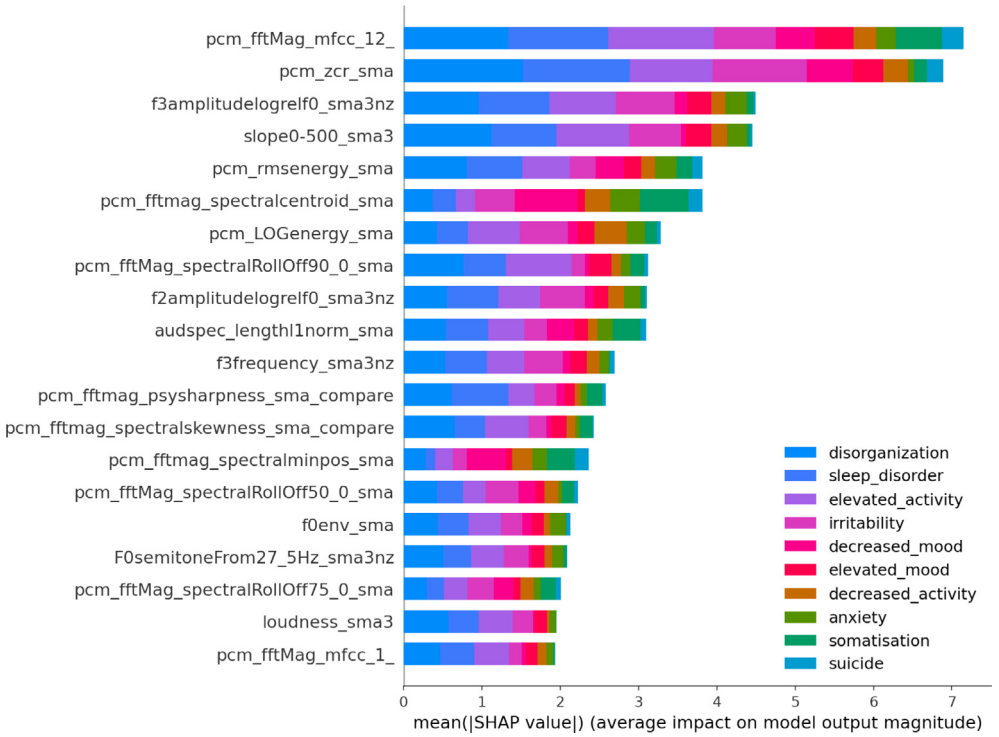

Fig. 7. Global model SHAP analysis for symptom prediction with the sequential and compositional MLP model.

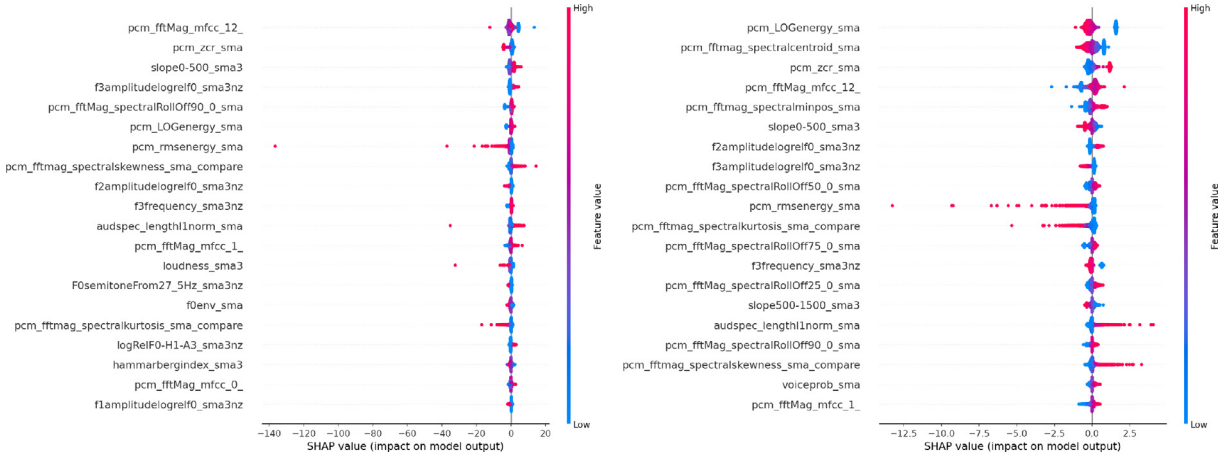

Fig. 8. From left to right: 20 top contributing features to the sequential and compositional MLP model global SHAP values for a) elevated activity symptom and b) decreased activity symptom.

Secondly, for the medical expert, linguistic summaries describing records that *contribute positively to predicting* a class or symptom are considered more informative than those that contribute negatively or those with an uncertain contribution. The key point of Table 6 is that most groups of voice parameters (reflecting pitch, voice signal spectrum, and voice quality) differ in episodes of illness (depression, mania) from the state of euthymia. As shown in Table 6, among the groups of voice parameters that contribute positively to predicting the euthymic state, most have low-level values. Among the parameters that contribute positively to predicting a depressive state, most of the features related to the voice spectrum have values at a high level. The exact opposite association was found for mania: most of the spectral features that contribute positively to predicting mania have values at a low level. What is also clinically relevant is that most of the pitch-related features that

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

contribute positively to predicting mania have values at a high level. For the mixed state, it is not possible to draw clear clinical conclusions about the prediction of this state. Most of the results are in line with previous studies, as well as with clinical observations. In particular, the increased pitch has been correlated with mania [51]. As for loudness (energy-related features), although this voice parameter appears to be important in the clinic (manic patients often speak louder, in depression quieter), here the results obtained are inconclusive. One reason could be that this is an individually variable trait, depending on gender as well as the type of depression (depression with psychomotor retardation, depression with anxiety and agitation). Therefore, considering symptoms rather than specific labels can help improve prediction. In addition, linguistic summaries from Table 6 have been evaluated by the domain expert in terms their usefulness. Fig. 9 compares the degree of usefulness with *DoT* and *DoS* of the linguistic summaries for the prediction of euthymia, depression, mania, and mixed state, from the sequential and compositional MLP model. Summary Ids are spanned in a two-dimensional space to verify whether these pairs of measures are in accordance. It can be observed that the expert found useful summaries with low *DoT* (graphs in the first column, e.g. summaries *Id002*, *Id205*, and *Id301*), and vice versa summaries with a high degree of truth have been deemed not useful (e.g., summaries *Id204* and *Id202* referring to *contribution around zero to predict mania*). Similar results can be observed for DoS (graph on the right), where summaries with low covering (DoS) are considered useful (e.g., *Id002*, *Id104*, *Id206*, and *Id301*). Overall, summaries with positive contributions are more understandable and easier for the clinician to interpret than those with around zero contributions, especially when the sentence specifies an extreme value such as *low, high* of certain voice features, e.g. *Among records that contribute positively to predicting mania class, most of them have pitch-related features at high level*. Next, we ran a statistical test for *DoT* and *DoS* of the set of linguistic summaries from the sequential and compositional model and the baseline model. The results are presented in Table 8. The characteristics are compared within all four classes, respectively. As can be seen in the table, the distributions underlying our samples are not statistically significantly different. For example, when the Wilcoxon signed-rank test is used to compare degrees of truth for linguistic summaries on the prediction of the euthymia class, the computed p-value is 0.5, supporting the statistical hypothesis that the median of the pairwise differences is zero. We conclude that in terms of objective quality mea-

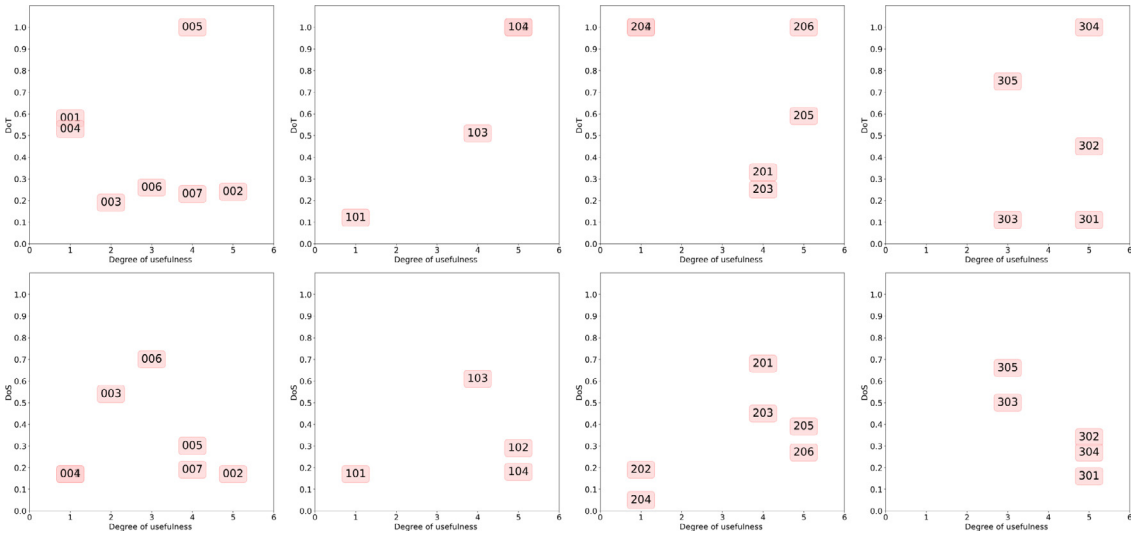

Fig. 9. Top row from left to right: degree of usefulness and degree of truth for linguistic summaries on euthymia, depression, mania, and mixed state from the sequential and compositional MLP model. Bottom row: degree of usefulness and degree of support for linguistic summaries for prediction of euthymia,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

Table 8
Results of the Wilcoxon signed-rank test (W denoted the test statistics) that compare the quality of linguistic summaries produced by the proposed sequential and compositional MLP vs. the MLP baseline. The degree of truth and degree of support are considered quality measures for predicting the four BD classes.

Criterion	Euthymia	Depression	Mania	Mixed state
DoT	W = 30.5; p-value = 0.50	W = 17.5; p-value = 0.94	W = 19.5; p-value = 0.12	W = 30.0; p-value = 0.78
DoS	W = 305.0; p-value = 0.86	W = 249.0; p-value = 0.27	W = 215.0; p-value = 0.10	W = 301.0; p-value = 0.81

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**

Table 10
Mental health survey classification results. The best hyperparameter configuration is also reported under the results of each model, obtained by grid-searching over the following sets: *# estimators* $\in \{100, 200, 300\}$; *max depth* $\in \{3, 5, 7\}$; *optimizer* $\in \{Adam, SGD\}$; *learning rate* $\in \{0.01, 0.001, 0.001\}$; *batch size* $\in \{16, 32, 64\}$; *epochs* $\in \{30, 50, 100\}$; *# HF* $\in [2, 10]$; *NMF solver* $\in \{multiplicative\ update, coordinate\ descent\}$.

Method	Class	Precision	Recall	F1-score
XGBoost	N (Control)	0.70	0.68	0.69
	Y (Treatment)	0.81	0.83	0.82
	Accuracy			0.77
	# estimators = 200, max depth = 3, objective = logistic			
Single-task MLP	N (Control)	0.76	0.52	0.61
	Y (Treatment)	0.76	0.90	0.82
	Accuracy			0.76
	optimizer = Adam, learning rate = 0.001, batch size = 32, epochs = 50			
Multi-task MLP	N (Control)	0.87	0.38	0.53
	Y (Treatment)	0.72	0.97	0.83
	Accuracy			0.75
	optimizer = Adam, learning rate = 0.001, batch size = 32, epochs = 50			
	# HF = 5, NMF solver = coordinate descent, max iter = 1000, tol = 1e–6			

sures, the linguistic summaries obtained from the baseline and the proposed MLP are comparable. Although the tests did not reveal statistically significant differences for the considered characteristics, the inclusion of domain knowledge allowed the generation of additional information in the form of linguistic summaries on the middle-layer labels, and, therefore, the understanding of the modeling process has been improved. Finally, we present and discuss the outcomes of the expert-based evaluation of the group of linguistic summaries generated for the BD use case; the questionnaire to assess LS in terms of Grice's maxims is also reported in Table 9. As the table shows, the psychiatry domain expert granted a high-quality mean SCS (0.7/1) to PLENARY explanations based on LS rendering the SHAP analysis. When it comes to Grice's maxims, a high score of 4.25/5 is obtained regarding the relevance, quality, quantity, and manner of communicating the predictive process. However, the ability to change the degree of detail on demand, in a future interactive explanation system, could improve the inconsistency among explanations found (SCS3, SCS7). Likewise, the *manner* in which LSs are presented could be studied to improve ambiguity (GM4), since the precision and granularity of the explanations may at times be hard to digest by experts (SCS1).

5.2. Use case in mental health survey

As a second case study, we considered the Mental Health in Tech Survey data[3] collected by the Open Sourcing Mental Illness non-profit corporation. It aims to measure how mental health is viewed by employers and employees of technology companies and the occurrence of these disorders. The original data contain more than 1200 surveys, described by 27 attributes. Preprocessing step has been performed to remove unnecessary information (such as timestamp, country, state, and comments) and to standardize the answers (e.g., "Male", "M", and "Male-ish" refer to the same concept). In addition, rows containing missing values have been removed, categorical data have been encoded into ordinal features where appropriate, and only features with at least three categories have been included in this analysis. Thus, the final data consist of 972 rows described by 22 features corresponding to the questions in the survey.[4] Two target classes indicate whether (Y) or not (N) a subject has sought a cure for mental health. Interestingly, more subjects sought treatment (619) than those who did not (353).

5.2.1. Accuracy evaluation

We start with the presentation of the accuracy evaluation. Similarly to the previous use case, the following methods have been considered: XGBoost, the baseline MLP, and the sequential and compositional MLP, which is trained to perform two recognition tasks in sequence: middle-level labels and target classes. In this case, the middle-level labels are simulated with a non-negative matrix factorization which enables to group the data while assigning them to hidden factors, and thus allowing a certain level of interpretability. Table 10 reports classification results for these three methods. Overall, the results are comparable. XGBoost performs slightly better than the MLP models, suggesting that perhaps the problem does not need complex models to be solved. Indeed, we are using this simple dataset on purpose, to easily show the three-step process used by PLENARY to explain the classification results of black-box models. It is worth pointing out that although the sequential and compositional MLP model has lower accuracy, the use of the intermediate labels identified by non-negative matrix factorization has increased the recall of class Y to 0.97 (i.e., the model is able to identify nearly everyone who has sought a cure).

[3] https://osmihelp.org/research (OSMH/OSMI Mental Health in Tech Survey).
[4] Please refer to the GitHub repository for more details on the data.

5.2.2. Explanation of predictive model results using SHAP

Fig. 10 compares SHAP global explanations for the two target classes, N and Y, with the single-task MLP and the sequential and compositional MLP model. It can be clearly observed that the most important feature to discriminate between the two classes is *family_history*. Indeed, if a previous mental illness has occurred in the family, it positively influences the prediction (looking for a cure). The same result can be derived from both algorithms. However, looking at the two classes, the two models disagree on the feature rankings and their influence on the predictive models.

Let us dwell on the treatment class (Y), to understand what are the factors that most influence the need for mental health care, according to the two algorithms and their visual explanations. For the sequential and compositional MLP model, the top contributing features are *benefits*, *anonymity*, and *work_interfere*, which positively influence the choice of seeking cures. It means that benefits for mental health care are available in the company, anonymity is protected, and the individual suffers from a mental health condition that interferes with her/his work. For the baseline model, the top influencing factors for class Y are the presence of a previous mental illness in the family (*family_history*), the interference between the subject's mental illness and her/his regular work activities, (*work_interfere*), the guarantee of anonymity (*anonymity*), and the availability of benefits for mental health cure (*benefits*). The number of employees (*no_employees*) negatively affects the prediction, sug-

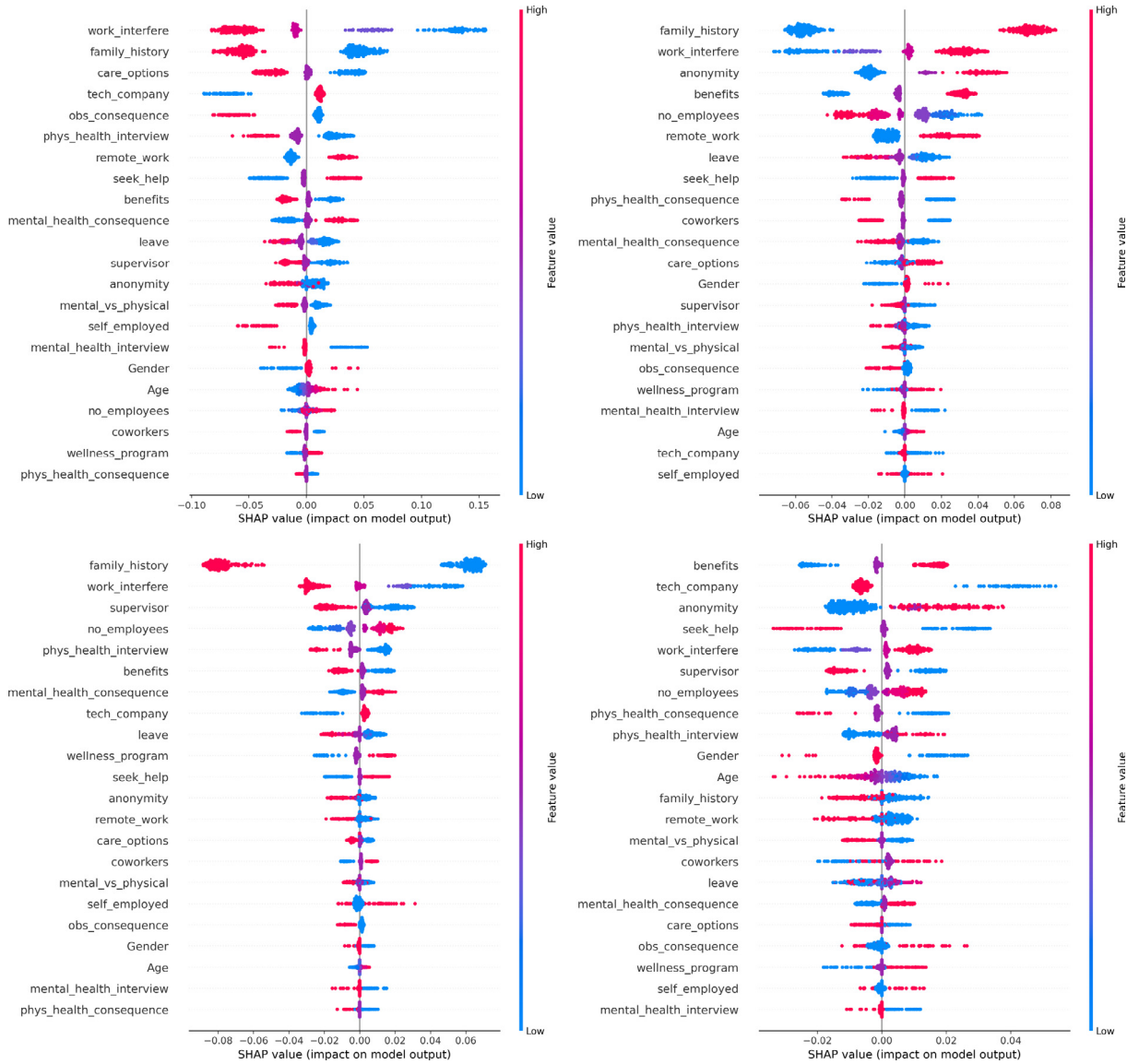

Fig. 10. From left to right and top to bottom: a) baseline global SHAP analysis for class N (a subject did not seek treatment for a mental health condition); b) baseline global SHAP analysis for class Y (a subject sought treatment for a mental health condition); c) sequential and compositional MLP global SHAP analysis for class N; d) sequential and compositional MLP global SHAP analysis for class Y, for the mental health survey data.

gesting that the larger the company, the less an employee would be willing to seek cures. Conversely, fewer employees may encourage them to seek help. Surprisingly, this observation concerning the distinction between the complementarity of each class explanations for *no_employees* is not observable for the sequential and compositional MLP model.

It should be noted that the two classes are complementary: indeed, with quite a few exceptions at the tail of the ranking, the other class should often find the opposite explanation. While rankings are not preserved in this binary classification model, the variation in ranking positions is nearly preserved (except for *physical_health_consequence* and a few others). However, this is not that dramatic, considering that the overall SHAP values in absolute value for the top contributing features are never very high for any of the features. A curiosity is the fact that some feature contributions are not reversed in terms of positive vs. negative contribution to the model (*gender*, *age*, *obs_consequence*, *mental_vs_physical*, *mental_health_interview*), as one would normally expect in a binary classification problem. This may indicate their irrelevance or their correlation with some other feature. In fact, for example, the *mental_health_interview* feature corresponds to the question: *"Would you bring up a mental health issue with a potential employer in an interview?*". The answer to this question is somewhat controversial, despite the subjects' willingness to access mental healing therapy.

Fig. 10 shows with both SHAP global plots that the sequential and compositional MLP model has on average, for all features, lower SHAP values than the baseline, providing less explanatory power to the same features. However, the spread of the points does not seem to vary much. The lack of change in the density of feature contributions in both classes shows similar rankings among the top contributing features to the disease state outcome.

5.2.3. Linguistic summaries of global model explanations

We now present linguistic summaries derived from the PLENARY approach. Table 11 collects the linguistic summaries for the prediction of treatment based on the sequential and compositional MLP model. We filtered them based on the quality criteria and only summaries with $DoT > 0.5$ are presented. We can see in Table 11 that most of the selected protoforms have a degree of truth of 1.0, suggesting that the selected explanations are trustable. DoS and DoF measure the coverage of a given protoform and the coverage of the condition expressed by the qualifier, respectively. Protoforms suggest that features pos-

Table 11
Linguistic summaries for prediction of the treatment class with $DoT > 0.5$ from the sequential and compositional MLP model. Summaries that contribute positively to predicting a class are presented in bold.

Id	Protoform	DoT	DoS	DoF
001	Among records that contribute against predicting treatment class, most of them have Age-related features at low level	1	0.57	0.11
002	Among records that contribute around zero to predicting treatment class, most of them have Age-related features at low level	1	0.83	0.84
003	**Among records that contribute positively to predicting treatment class, most of them have age feature at low level**	1	0.42	0.05
004	Among records that contribute against predicting treatment class, most of them have wellness_program feature at medium level	1	0.46	0.10
005	Among records that contribute around zero to predicting treatment class, most of them have wellness_program feature at medium level	1	0.68	0.82
006	**Among records that contribute positively to predicting treatment class, most of them have wellness_program feature at medium level**	1	0.54	0.08
007	Among records that contribute against predicting treatment class, most of them have anonymity feature at low level	1	0.41	0.38
008	Among records that contribute around zero to predicting treatment class, most of them have anonymity feature at low level	1	0.63	0.42
009	**Among records that contribute positively to predicting treatment class, most of them have anonymity feature at low level**	1	0.22	0.19
010	**Among records that contribute positively to predicting treatment class, most of them have leave feature at low level**	1	0.25	0.11
011	Among records that contribute against predicting treatment class, most of them have phys_health_consequence feature at medium level	1	0.51	0.08
012	Among records that contribute around zero to predicting treatment class, most of them have phys_health_consequence feature at medium level	1	0.68	0.80
013	**Among records that contribute positively to predicting treatment class, most of them have phys_health_consequence feature at medium level**	1	0.38	0.12
014	Among records that contribute against predicting treatment class, most of them have coworkers feature at medium level	1	0.49	0.12
015	Among records that contribute around zero to predicting treatment class, most of them have coworkers feature at medium level	1	0.63	0.77
016	**Among records that contribute positively to predicting treatment class, most of them have coworkers feature at medium level**	1	0.51	0.01
017	Among records that contribute against predicting treatment class, most of them have mental_health_interview feature at medium level	1	0.72	0.05
018	Among records that contribute around zero to predicting treatment class, most of them have mental_health_interview feature at medium level	1	0.83	0.85
019	**Among records that contribute positively to predicting treatment class, most of them have mental_health_interview feature at medium level**	1	0.78	0.09
020	**Among records that contribute positively to predicting treatment class, most of them have seek_help feature at medium level**	0.77	0.39	0.17
021	Among records that contribute around zero to predicting treatment class, most of them have leave feature at low level	0.65	0.56	0.67

itively contributing to the treatment class are *age* at low level (i.e., younger employees are more willing to see for treatments), *wellness program*, *anonymity*, *leave*, *phys health consequence*, *coworkers*, *mental health interview*, and *seek for help*, at different levels. However, some of these features, such as *age*, at the same time give a positive, negative, and zero contribution to the treatment class prediction. This behavior derives from different subsets of data and expresses in natural language what also the SHAP graphs have shown, that is the impact of the model output, even if positive or negative, has very low values (Fig. 10).

It is also observed that some summaries are true to some extent, e.g. *Among records that contribute positively to predicting treatment class, most of them have sought help feature at medium level* [$DoT = 0.77, DoS = 0.39, DoF = 0.17$] (*Id020*). We have then compared the measures used to quantitatively evaluate the linguistic summaries (DoT and DoS), with the expert-based degree of usefulness (DoU). For this purpose, we used a two-dimensional graph that represents the protoforms in the space defined by the degree of usefulness and by each of the measures (Fig. 11). As observed in Fig. 11, and similarly to the bipolar disorder case study, there is no agreement between *useful* protoforms and the quantitative measures used to describe the quality of the explanations. Indeed, some protoforms with high *DoT* are considered not useful (e.g., *Id018* and *Id017* in the first graph), protoforms with low coverage, such as *Id009* and *Id010*, are considered very useful, and, on the contrary, protoforms with high coverage, such as *Id018* and *Id017*, are assessed as not useful. However, these results do not suggest that the qualitative measures or the expert evaluation are necessarily wrong: they only point out that different criteria are considered.

Finally, we ran statistical tests for various quality measures to assess whether the quality of the resulting set of linguistic summaries is significantly different than summaries from the baseline model (MLP). We considered *DoT* and *DoS* as quality measures and compared the prediction of the treatment class vs no treatment class.

The results of the Wilcoxon signed-rank test are reported in Table 12. For example, when this test is used to compare degrees of truth for linguistic summaries on treatment prediction from the baseline and the proposed MLP model, the computed p-value is 0.75, supporting the statistical hypothesis that the median of pairwise differences is zero. We conclude that in terms of the objective quality measures, the linguistic summaries obtained from the two predictive models are comparable.

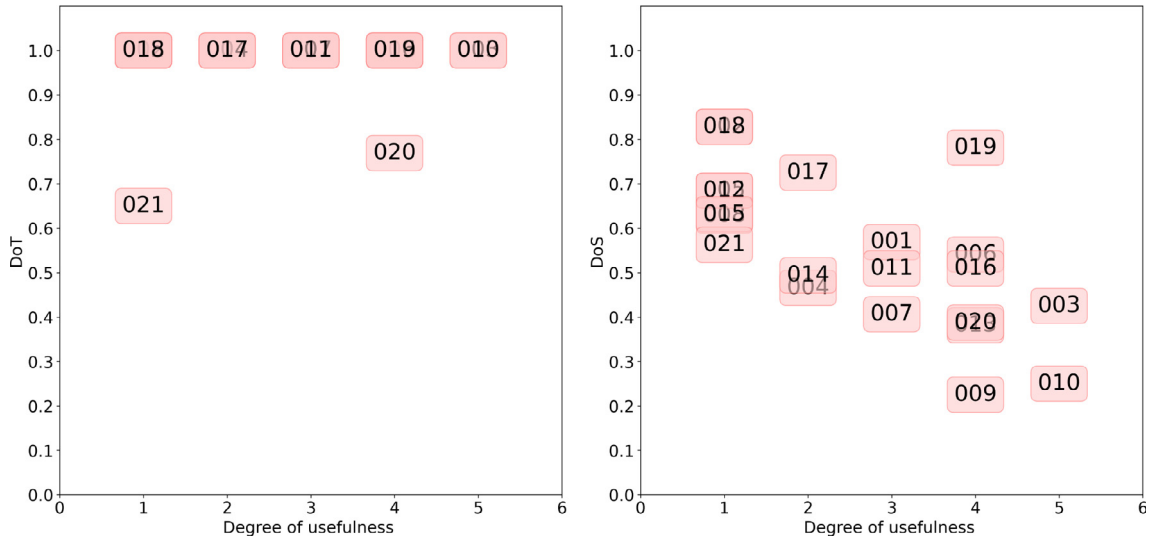

Fig. 11. From left to right: comparisons between degree of usefulness and degree of truth or degree of support for linguistic summaries (with $DoT > 0.5$) to predict the treatment class, from the sequential and compositional MLP model. Descriptions of LS Ids are provided in Table 11.

Table 12
Results of the Wilcoxon signed-rank test (W denotes the test statistic) that compares the quality of linguistic summaries produced by the proposed sequential and compositional MLP vs. the MLP baseline. The degree of truth and degree of support are considered quality measures for predicting the treatment and no treatment class, respectively.

Criterion	Treatment	No treatment
DoT	W = 9.0; p-value = 0.75	W = 9.0; p-value = 0.75
DoS	W = 1387.5; p-value = 0.69	W = 1012.0; p-value = 0.03

6. Conclusions and future work

In this paper, we have proposed a new framework, PLENARY, to explain black-box models in natural language through the use of fuzzy linguistic summaries. Several key features make PLENARY a suitable tool for XAI. A key feature is the introduction of an intermediate layer of annotations to mitigate some uncertainties related to classes. A second key feature is that the grouping of low-level attributes into high-level information granules using linguistic summarization improves the overall explainability of the model results. Experimental evaluations confirmed that fuzzy linguistic summarization complements global model explanations derived from the popular SHAP tool. Furthermore, the results demonstrate that PLENARY improves understanding of model outputs by appropriate incorporation of the domain knowledge. In particular, the proposed sequential and compositional neural network architecture can effectively incorporate domain knowledge into the predictive model. 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.

Moreover, from the point of view of the application considered, new technologies have great potential to support psychiatrists in understanding the outcomes of disease classification. As our results have shown, in some cases it is not possible to draw clearly interpretable conclusions from a label-based analysis, which relies on rigid classification criteria. A symptom-enhanced approach appears to be more supportive in predicting various combinations of depressive and manic symptoms present in non-classical forms of episodes, including subsyndromal and mixed states. Notably, we present explanations at different levels of granularity so that end-users (e.g., psychiatrists) understand the comprehensive process of arriving at the prediction of healthy or pathological classes and symptoms. This represents a huge opportunity to improve clinical decision-making from early recognition of relapse to personalized and more effective treatment at the individual level. This could be a step forward in creating a personalized approach based on objective real-world biomarkers, such as voice data. This study also showed that the evaluation of SHAP values by domain experts still proves to be very difficult, since most of the speech features have an unclear relationship with clinical symptoms. However, psychiatrists might say more when SHAP feature contributions are organized into class groups (such as Fig. 5 and 6). We have found that in this way it is possible to draw more general conclusions concerning only what is relatively clear, i.e. the mixed state and mania. Another illustrative lesson from this study reflects how different XAI metrics compared to human experts assess quality in model explanations. While we can propose to domain experts to evaluate intuitive XAI metrics based on notions such as consistency, reliability, relevance, and usefulness, only the latter appeared to be intuitive and quantifiable in terms of clinical practicality in the considered psychiatric use cases. Furthermore, this study revealed the need for having an expert-in-the-loop for the whole modeling process. For example, some features presented in the SHAP plots are much more important for the domain expert, and, ideally, such preferences of experts towards features should be wisely included early in the modeling process. Also, the current SHAP library contains very limited features for comparative analyses of the outcomes of various predictive models. In particular, the global SHAP plots should have the same scale to make it easier for domain experts to analyze the differences between alternative models.

The proposed approach is a significant use case but could be extended to other application domains as future work. Future work also assumes an analysis of uncertain labels and domain knowledge. For example, the number of classes depends on the domain but there may be several, such as mild depression, severe depression, etc. The proposed PLENARY approach could serve as a global and human-consistent validation framework for assessing whether global model explanations are robust. However, further experiments are needed to demonstrate all the benefits and limitations in terms of validating the robustness of the system through linguistic summarization. Also, as of now, SHAP is able to explain uncertain (gradual) assignments to classes. For example, our model assigns a 0.3 chance that an instance is of class A and 0.7 that it is of class B. We aim to evaluate that, for example, *"The model is not certain enough to predict neither class A nor class B*". In SHAP, we show features ranked in terms of feature contribution. How we select key features may be different for each class or symptom, as domain experts characterize each class based on a priority of symptoms that can vary when trying to prove diagnosis A versus diagnosis B.

The PLENARY approach proposed for linguistic summarization on global model explanations starts a new and promising research direction with 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. Another extension would be the creation of a dynamic approach to summarize high-level groups that are not homogeneous in terms of impact on the predicted class. Also, further research will consider not only other types of protoforms but also quantifiers and t-norms. 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. Finally, to improve the understandability of the final explanations generated by PLENARY, abstraction summarization approaches could also be investigated to generate paragraph-based linguistic summaries rather than individual sentence-based summaries. Extracting linguistic units larger than sentences, such as paragraphs, could actually make the final summaries easier to read. To this aim, we plan to apply other operations to the sentences, clustering semantically related sentences into groups [52]. Another possibility is to adopt deep learning models for abstractive summarization, as proposed [53]. These further investigations will be the focus of our future work devoted to improving the explainable facet of PLENARY.

Code availability

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

Authors and their contribution

Katarzyna Kaczmarek-Majer (KKM), Giovanna Castellano (GiC), Olga Kaminska (OK), Gabriella Casalino (GaC), Gennaro Vessio (GV), and Natalia Diaz-Rodriquez (NR) designed the study. OK, KKM, GC, and GV performed the experiments with numerical data. Monika Dominiak (MD) contributed to the interpretation and discussion of results from the medical perspective. KKM, GaC, GiC, and GV prepared the design of the experiments. GiC, GV, OK, and GaC contributed to the design and implementation of the neural network components. KKM and Olgierd Hryniewicz (OH) contributed to the design and implementation of linguistic summarization components. GaC and OK contributed to the implementation of the SHAP components. All authors contributed to the drafting and revision of the manuscript. All authors approved the final manuscript.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

Katarzyna Kaczmarek-Majer and Olga Kaminska are supported by the Small Grants Scheme (NOR/SGS/BIPOLAR/0239/2020-00) within the research project "Bipolar disorder prediction with sensor-based semi-supervised Learning (BIPOLAR)". BDMON was collected in the CHAD project entitled "Smartphone-based diagnostics of phase changes in the course of bipolar disorder" (RPMA.01.02.00-14-5706/16-00) funded by EU funds in 2017–2018. The authors thank researchers Karol Opara and Weronika Radziszewska from Systems Research Institute Polish Academy of Sciences for support in data preparation and analysis. Natalia Díaz-Rodríguez was supported through the Juan de la Cierva Incorporación grant IJC2019-039152-I funded by MCIN/AEI/10.13039/501100011033 by "ESF Investing in your future" and Google Research Scholar Program. The authors thank Adrien Bennetot for support in data analysis. Gabriella Casalino acknowledges the funding support of the Italian Ministry of University and Research through the European PON project AIM (Attraction and International Mobility) 1852414, activity 2, line 1.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, athttps://doi.org/10.1016/j.ins.2022.10.010.

References

[1] 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) 47–60.
[2] A.B. Arrieta et al, Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, Inform. Fusion 58 (2020) 82–115.
[3] W.H. Organization, International statistical classification of diseases and related health problems, (11th ed.) doi:https://icd.who.int/.
[4] G.S. Malhi, E. Bell, P. Boyce, D. Bassett, M. Berk, R. Bryant, M. Gitlin, A. Hamilton, P. Hazell, M. Hopwood, et al, The 2020 royal australian and new zealand college of psychiatrists clinical practice guidelines for mood disorders: Bipolar disorder summary, Bipolar disorders 22 (8) (2020) 805–821.
[5] J. Scott, A. Vaaler, O. Fasmer, G. Morken, K. Krane-Gartiser, A pilot study to determine whether combinations of objectively measured activity parameters can be used to differentiate between mixed states, mania, and bipolar depression, Int. J. Bipolar. Disord. 5(1).
[6] 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.
[7] R. Horwitz, T.F. Quatieri, B.S. Helfer, B. Yu, J.R. Williamson, J. Mundt, On the relative importance of vocal source, system, and prosody in human depression, in: 2013 IEEE International Conference on Body Sensor Networks, IEEE, 2013, pp. 1–6.
[8] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.
[9] R. Shwartz-Ziv, A. Armon, Tabular data: Deep learning is not all you need, Inform. Fusion 81 (2022) 84–90.
[10] Y. Zhou, G. Li, H. Li, Automatic cataract classification using deep neural network with discrete state transition, IEEE Trans. Med. Imaging 39 (2) (2019) 436–446.
[11] G. Casalino, G. Castellano, A. Consiglio, N. Nuzziello, G. Vessio, MicroRNA expression classification for pediatric multiple sclerosis identification, J. Ambient Intell. Humaniz. Comput. (2021) 1–10.
[12] D. Berthiaume, R. Paffenroth, L. Guo, Understanding deep learning: Expected spanning dimension and controlling the flexibility of neural networks, Front. Appl. Math. Stat. (2020) 52.
[13] A. Holzinger, C. Biemann, C.S. Pattichis, D.B. Kell, What do we need to build explainable ai systems for the medical domain?, arXiv preprint arXiv:1712.09923.
[14] 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.

[15] A. Shrikumar, P. Greenside, A. Kundaje, Learning important features through propagating activation differences, in: International Conference on Machine Learning, PMLR, 2017, pp. 3145–3153.
[16] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances in neural information processing systems 30.
[17] N. Díaz-Rodríguez 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) 58–83.
[18] A. Bennetot, I. Donadello, A.E. Qadi, M. Dragoni, T. Frossard, B. Wagner, A. Saranti, S. Tulli, M. Trocan, R. Chatila, et al., A practical tutorial on explainable ai techniques, arXiv preprint arXiv:2111.14260.
[19] W. Caicedo-Torres, J. Gutierrez, Iseeu: Visually interpretable deep learning for mortality prediction inside the icu, J. Biomed. Inform. 98 (2019) 103269.
[20] M.D. Peláez-Aguilera, M. Espinilla, M.R. Fernández Olmo, J. Medina, Fuzzy linguistic protoforms to summarize heart rate streams of patients with ischemic heart disease, Complexity.
[21] N. Marín, D. Sánchez, On generating linguistic descriptions of time series, Fuzzy Sets and Systems 285 (2016) 6–30, special Issue on Linguistic Description of Time Series.
[22] J. Kacprzyk, R.R. Yager, S. Zadrozny, A fuzzy logic based approach to linguistic summaries of databases, Journal of, Appl. Math. Comput. Sci. 10 (2000) 813–834.
[23] J. Kacprzyk, R.R. Yager, J.M. Merigo, Towards human-centric aggregation via ordered weighted aggregation operators and linguistic data summaries: A new perspective on zadeh's inspirations, IEEE Comput. Intell. Mag. 14 (1) (2019) 16–30.
[24] F.E. Boran, D. Akay, R.R. Yager, An overview of methods for linguistic summarization with fuzzy sets, Expert Syst. Appl. 61 (2016) 356–377.
[25] A. Ramos-Soto, P. Martin-Rodilla, Enriching linguistic descriptions of data: A framework for composite protoforms, Fuzzy Sets Syst. 407 (2019) 1–26.
[26] M. Bartczak, A. Niewiadomski, Linguistic summaries of graph databases in customer relationship management (crm), J. Appl. Comput. Sci. 27 (1) (2019) 7–26.
[27] N. Marín, G. Rivas-Gervilla, M.D. Ruiz, D. Sánchez, Formal concept analysis for the generation of plural referring expressions, Inf. Sci. 579 (2021) 717–731.
[28] J. Moreno-Garcia, L. Rodriguez-Benitez, L. Jimenez-Linares, G. Triviño, A linguistic extension of petri nets for the description of systems: An application to time series, IEEE Trans. Fuzzy Syst. 27 (9) (2019) 1818–1832.
[29] 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, Inf. Sci. 588 (2022) 174–195.
[30] R. Guidotti, A. Monreale, S. Ruggieri, D. Pedreschi, F. Turini, F. Giannotti, Local rule-based explanations of black box decision systems, arXiv preprint arXiv:1805.10820.
[31] M. Danilevsky, S. Dhanorkar, Y. Li, L. Popa, K. Qian, A. Xu, Explainability for natural language processing, in: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2021, pp. 4033–4034.
[32] D. Wei, S. Dash, T. Gao, O. Gunluk, Generalized linear rule models, International Conference on Machine Learning, PMLR (2019) 6687–6696.
[33] L.A. Zadeh, Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic, Fuzzy Sets Syst. 90 (2) (1997) 111–127.
[34] C. Mencar, J.M. Alonso, Paving the way to explainable artificial intelligence with fuzzy modeling, in: International Workshop on Fuzzy Logic and Applications, Springer, 2018, pp. 215–227.
[35] W. Pedrycz, S. Chen, Interpretable Artificial Intelligence: A Perspective of Granular Computing, Vol. 937, Springer Nature, 2021.
[36] L.A. Zadeh, From computing with numbers to computing with words, From manipulation of measurements to manipulation of perceptions, IEEE Transactions on circuits and systems I: fundamental theory and applications 46 (1) (1999) 105–119.
[37] K. Kaczmarek-Majer, O. Hryniewicz, Application of linguistic summarization methods in time series forecasting, Inf. Sci. 478 (2019) 580–594.
[38] K. Knight, D. Marcu, Summarization beyond sentence extraction: A probabilistic approach to sentence compression, Artif. Intell. 139 (1) (2002) 91–107.
[39] J. Kacprzyk, R.R. Yager, Linguistic summaries of data using fuzzy logic, Int. J. Gener. Syst. 30 (2) (2001) 133–154.
[40] J. Zhou, A.H. Gandomi, F. Chen, A. Holzinger, Evaluating the quality of machine learning explanations: A survey on methods and metrics, Electronics 10 (5) (2021) 593.
[41] R. Wilming, C. Budding, K.-R. Müller, S. Haufe, Scrutinizing XAI using linear ground-truth data with suppressor variables, Mach. Learn. (2022) 1–21.
[42] F. Cabitza, A. Campagner, L.M. Sconfienza, As if sand were stone, New concepts and metrics to probe the ground on which to build trustable AI, BMC Medical Informatics and Decision Making 20 (1) (2020) 1–21.
[43] A. Holzinger, A. Carrington, H. Müller, Measuring the quality of explanations: the system causability scale (SCS), KI-Künstliche Intelligenz 34 (2) (2020) 193–198.
[44] A.T. Holzinger, H. Muller, Toward human–AI interfaces to support explainability and causability in medical AI, Computer 54 (10) (2021) 78–86.
[45] J. van der Waa, E. Nieuwburg, A. Cremers, M. Neerincx, Evaluating XAI: A comparison of rule-based and example-based explanations, Artif. Intell. 291 (2021) 103404.
[46] M.J. Lesot, G. Moyse, B. Bouchon-Meunier, Interpretability of fuzzy linguistic summaries, Fuzzy Sets Syst. 292 (2016) 307–317.
[47] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, H. Müller, Causability and explainability of artificial intelligence in medicine, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 9 (4) (2019) e1312.
[48] P. Grice, Studies in the Way of Words, Harvard University Press, 1989.
[49] 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.
[50] F. Eyben, F. Weninger, F. Gross, B. Schuller, Recent developments in opensmile, the munich open-source multimedia feature extractor, in: Proc. of the 21st ACM Int. Conf. on Multimedia, 2013, pp. 835–838.
[51] J. Zhang, Z. Pan, C. Gui, T. Xue, Y. Lin, J. Zhu, D. Cui, Analysis on speech signal features of manic patients, J. Psychiatr. Res. 98 (2018) 59–63.
[52] M. Cao, H. Zhuge, Grouping sentences as better language unit for extractive text summarization, Future Gener. Comput. Syst. 109 (2020) 331–359.
[53] S. Song, H. Huang, T. Ruan, Abstractive text summarization using LSTM-CNN based deep learning, Multimedia Tools Appl. 78 (1) (2019) 857–875.