



Exploring Contextual Importance and Utility in Explaining Affect Detection

Nazanin Fouladgar¹(✉), Marjan Alirezaie², and Kary Främling^{1,3}

¹ Department of Computing Science, Umeå University, Umeå, Sweden
nazanin@cs.umu.se, kary.framling@umu.se

² Center for Applied Autonomous Sensor Systems, Örebro University,
Örebro, Sweden
marjan.alirezaie@oru.se

³ School of Science and Technology, Aalto University, Espoo, Finland

Abstract. By the ubiquitous usage of machine learning models with their inherent black-box nature, the necessity of explaining the decisions made by these models has become crucial. Although outcome explanation has been recently taken into account as a solution to the transparency issue in many areas, affect computing is one of the domains with the least dedicated effort on the practice of explainable AI, particularly over different machine learning models. The aim of this work is to evaluate the outcome explanations of two black-box models, namely neural network (NN) and linear discriminant analysis (LDA), to understand individuals affective states measured by wearable sensors. Emphasizing on context-aware decision explanations of these models, the two concepts of Contextual Importance (*CI*) and Contextual Utility (*CU*) are employed as a model-agnostic outcome explanation approach. We conduct our experiments on the two multimodal affect computing datasets, namely WESAD and MAHNOB-HCI. The results of applying a neural-based model on the first dataset reveal that the electrodermal activity, respiration as well as accelerometer sensors contribute significantly in the detection of “meditation” state for a particular participant. However, the respiration sensor does not intervene in the LDA decision of the same state. On the course of second dataset and the neural network model, the importance and utility of electrocardiogram and respiration sensors are shown as the dominant features in the detection of an individual “surprised” state, while the LDA model does not rely on the respiration sensor to detect this mental state.

Keywords: Explainable AI · Affect detection · Black-Box decision · Contextual importance and utility

1 Introduction

The rapid growth of producing wearable sensors in health-related applications and the success of machine learning (ML) methods in analysing the sensors data

© Springer Nature Switzerland AG 2021

M. Baldoni and S. Bandini (Eds.): ALxIA 2020, LNAI 12414, pp. 3–18, 2021.

https://doi.org/10.1007/978-3-030-77091-4_1

are incontrovertible. These facilities have revolutionized the doctor-patients relationships [17] and provided continuous monitoring of patients to a considerable degree [2]. In some situations, an individual could also follow his/her well-being status independent of a third-party interventions. Nevertheless, the extent of trust to such models is questionable as the advanced artificial intelligence (AI) methods lack transparency intrinsically. Therefore, relying solely on such methods is not recommended, particularly in critical decision-making processes [16]. From the view of health practitioners, their decisions could be finalized more confidently if they are provided with a concrete outcome explanation of AI models. Moreover, from the end-users perspective, explainable AI (XAI) can influence their own follow-up schedule.

Capitalizing these advantages, XAI has recently appealed a great attention among research communities as well as industries [4,11]. Some scholars have theoretically scrutinized the XAI potentiality [16,26] on the lens of multidisciplinary fields, while others made efforts to unveil the practical aspects of XAI [5,12]. Moreover, in a very recent work [8], XAI in theory and practice has been reconciliated and reviewed briefly. As mentioned before, the main concern of both aspects lies on the ground of intelligent systems transparency and thereby appealing the experts or end-users trust. Addressing the aforementioned issues, the XAI applicability is pinpointed in a vast body of works such as tutoring [21], fault diagnosis [3] as well as healthcare [19].

Despite the research efforts in equipping ML models of different domains with XAI techniques, the intersection of XAI and affect computing is still immature and there are open rooms for researchers of this area. In an extension to our previous work [7], we study the outcome explanations of two machine learning models, namely neural network (NN) and linear discriminant analysis (LDA), designed to classify human state-of-mind. We employ two datasets including WESAD [23], and MAHNOB-HCI [25], as publicly and academically available datasets respectively, in the realm of multi-modal affect computing (see Sect. 4). Our main focus is on signal-level explanations, relying on the two concepts of Contextual Importance (*CI*) and Contextual Utility (*CU*) proposed by Främling [10]. Applying (*CI*) and (*CU*), we represent how important and favorable different sensors (features) are for the decision of each examined model. Both *CI* and *CU* present numerical values applicable in textual and visual representations and thereby understandable to professionals and end-users.

The rest of paper is organized as follows: a brief review of the recent corpus of black-box outcome explanation in health-related works is given in Sect. 2. We investigate the *CI* and *CU* concepts in Sect. 3. After introducing the datasets and their specification in Sect. 4, we present the results in Sect. 5 which is followed by the conclusion and discussion about the future works in Sect. 6.

2 Background

Contribution of AI in healthcare is mainly about certain practices including diagnosis upon medical imaging or tabular data samples. These diagnosis are

expected to be transparent and explainable to its users such as physicians, other medical practitioners and ideally the patients. Singh et al. [24] have categorized different methods addressing the explainability upon medical image analysis process, into attribution and non-attribution based methods.

Attribution-based methods are able to determine the contribution of an input feature to a target output neuron (of the correct class) in a classification process accomplished by a convolutional neural network (CNN). Due to their ease of use, such methods are employed upon brain imaging in Alzheimer classification task [6], retinal imaging to assist diabetic retinopathy [22] and also breast imaging in estrogen receptor classification task [20].

Unlike the attribution-based methods, in non-attribution based or post-model, another methodology than the original model is utilized on the given problem, mainly independent of the latter model attributes [24]. As some examples of non-attribution based methods used for the purpose of output explanation, we can refer to concept vectors and also textual justification [24]. Testing Concept Activation Vectors (TCAV) [27] is a concept vector method, capable of explaining the features learned by different layers to the domain experts by taking the directional derivative of the network in the concept space. In the context of text justification, these models generate linguistic outputs that justify the classifier’s output in an understanding way for both the expert users and patients. Lee. et al. [14] applied a justification model to generate textual explanation associated with a heat-map for breast classification task.

Apart from explanations in medical imaging, some studies in the literature have focused on the explainability of AI methods prediction upon tabular physiological and clinical data. The work in [18] examined three interpretable models, mainly Generalized Linear Model, Decision Tree and Random Forest, on electrocardiogram data (ECG) for the purpose of heart beat classification. Under the magnitude of early clinical prediction, Lauritsen et al. [13] utilized a post-model explanation module, decomposing the outputs of a temporal convolutional network into clinical parameters. Deep Taylor Decomposition (DTD) was the main tool of this module, providing the relevance explanation of prediction in a Layer-wise Relevance Propagation (LRP) manner. Among few works addressing the output explanation of human affect detection with tabular physiological data, the authors in [15] suggested two explanation components in signal- and sensor-level. The signal-level explanation was achieved by removing one of the signals iteratively from the prediction process while the sensor-level explanation was provided by applying entropy criterion to calculate the feature importance of two chest- and wrist-worn sensors. Similar to our work, the applied dataset was relied on WESAD. However, different from ours, this work could not provide the importance extent of the chest-worn signals in a specific context.

3 Contextual Importance and Contextual Utility

One of the earliest work in the realm of black-box outcome explanation was proposed by Framling [10] in 1996. He argued that expert systems had the main

contribution to explain any decisions. He added, however these systems were mainly rule-based and any changes in the input values result in firing a set of rules in a discrete manner. The gap of representing the outcomes of continuous real-valued functions was the reason to go beyond symbolic reasoning models.

The notions of *Contextual Importance (CI)* and *Contextual Utility (CU)* were proposed to explain the neural networks output in the context of Multiple Criteria Decision Making (MCDM). In MCDM, decisions are established on a consensus between different stakeholders preferences [9]. The stakeholders often consist of a group of people and/or an abstract entity (e.g. economy), whose preferences are highly subjective and more likely form a non-linear and continuous function. To provide a convenient explanation of these functions in MCDM, it was reasonable to explore how important each criterion was and to what extent it was favorable in a specific context. These were the main reasons pushing the two concepts of *CI* and *CU* forward. The concepts are formulated as following:

$$CI = \frac{Cmax_x(C_i) - Cmin_x(C_i)}{absmax - absmin} \quad (1)$$

$$CU = \frac{y_{ij} - Cmin_x(C_i)}{Cmax_x(C_i) - Cmin_x(C_i)} \quad (2)$$

Where C_i is the i th context (specific input of black-box referring as ‘Case’ in Sect. 5), y_{ij} is the value of j th output (class probability) with respect to the context C_i , $Cmax_x(C_i)$ and $Cmin_x(C_i)$ are the maximum and minimum values indicating the range of output values observed by varying each attribute x of context C_i , $absmax = 1$ and $absmin = 0$ are also the maximum and minimum values indicating the range of j th output (the class probability value).

We highlight that *CI* and *CU* return numerical values which allow us to represent the explanations to the end-users in the form of visual (e.g., in the form of graphs) or textual outputs.

4 Dataset Description and Preprocessing

We have tried two different datasets in order to evaluate our results. The first data set is WESAD which is publicly available and applicable for the purpose of multi-modal sensory analysis as well as detecting multiple affective states [23]. According to the dataset’s protocol, there are three main affective states in addition to the *baseline* state, including *stress*, *amusement* and *meditation*. These states have been examined on 15 different subjects, wearing RespiBAN Professional device on the chest and Empatica E4 device on the wrist. The former encompasses of data collected from eight different signals, namely electrocardiogram (ECG), electromyogram (EMG), electrodermal activity (EDA), temperature (TEMP), respiration (RESP) and three-axes accelerometer (ACC0, ACC1, ACC2), while the latter fetches blood volume pulse (BVP), EDA, TEMP, and accelerometer signals data. All RespiBAN data are sampled under 700 HZ, however the sampling rates are different among Empatica E4 signals. BVP, EDA and

TEMP data have been recorded 64 Hz, 4 Hz, 32 Hz respectively. Validating the study protocols, a supplementary of five self-reports in terms of questionnaire were also provided for each subject.

The WESAD dataset consists of around 4 million instances for each subject and in total 60 million samples for all the 15 subjects. Due to the time complexity of processing such a large dataset, we only extract the chest-worn signals of one participant to detect the four aforementioned affective states. After down-sampling the signals into 10 HZ we end up with 29350 data instances for the selected participant. One of the major properties of WESAD is that it is highly imbalanced. The highest number of samples belongs to the baseline state while the lowest amount refers to the amusement state. More specifically, the data includes the following ranges: [0–11400] labeled as baseline state, [11400–17800] labeled as stress state, [17800–21550] labeled as amusement state and the rest refers to the meditation state of our selected participant.

The second dataset is MAHNOB-HCI [25], only available to academia community with the aim of emotion recognition and multimedia tagging studies. The dataset consists of two trials collecting multimodal physiological sensor data as well as facial expression, audio signals and eye gaze data of 27 participants. The physiological signals refer to 32 electroencephalogram (EEG) channels, two ECG electrodes attached to the chest upper right (ECG1) and left (ECG2) corners below the clavicle bones as well as one ECG electrode placed at abdomen below the last rib (ECG3), two galvanic skin response (GSR) positioning on the distal phalanges of the middle (GSR1) and index fingers (GSR2), a RESP belt around the abdomen and a skin temperature (TEMP) placed at little finger. All signals except EEG are accessible to the end-user in 256 HZ sampling rate. To gather this data, 20 video clips were used to stimulate the participants’ emotions in the first trial while 28 images and 14 video fragments were shown to participants, tagged by either correct or incorrect words in the second trial. Moreover, the participants feedback were collected after each stimuli to provide the videos annotations as well as agreement or disagreements of tags. In the first trial, 9 emotional labels such as *amusement*, *happiness* and *surprised* were under focus while in the second trial only two modes of tag correctness or incorrectness were under consideration. Due to the large size of the dataset, we only extracted ECG1, ECG2, ECG3, GSR1, GSR2, RESP and TEMP data of one participant. Moreover, we focused only on the first trial of this dataset with three emotional states, mainly amusement, happiness and surprised for the purpose of classification task. The accordant data accounts for 1920 instances after downsampling the signals to 10 HZ sampling rate.

5 Outcome Explanations

One of our examined black-box models to classify human affective states is a neural network, consisting of one hidden layer with 100 units. The basic idea behind the neural-based networks is their capability of approximating non-linear but differentiable variations. This capability makes local gradients meaningful and

thereby the importance of each input feature explainable. Linear discriminant analysis is appointed as the second black-box model of this study, profound in considering the statistical properties of data in the classification process. After training these models on the data of specific participant in WESAD, the average accuracies of 0.92 and 0.91 are achieved in the neural network and LDA models, respectively. Following the same procedure on the second dataset, the neural network and LDA showed a performance of 0.99 and 1.0 respectively, in terms of accuracy.

In order to provide local outcome explanations of these models, we randomly choose an input instance (referring to ‘Case’ in this study) of an individual data in each dataset. Assuming the following ‘Case’ in WESAD: 0.898 (ACC0), -0.003 (ACC1), -0.179 (ACC2), -0.003 (ECG), 7.078 (EDA), 0.001 (EMG), 32.97 (TEMP), -0.865 (RESP), the trained neural network and LDA models result in “meditation” state (class) as the classification output. The following class probabilities are achieved for each model respectively: meditation class 97% and 99%, baseline class 0.5% and 0%, stress class 0.1% and 0%, and amusement class 1% and 0%. The same procedure could verify the state of ‘Case’ in MAHNOB-HCI dataset. Here, the ‘Case’ is randomly chosen as: -849000 (ECG1), -544000 (ECG2), -777000 (ECG3), 2900000 (GSR1), 90 (GSR2), -1560000 (RESP), 26078 (TEMP) from an individual data. The classification output of both models on this specific instance yields to “surprised” state with the highest probability (100%), and to amusement and happiness states with the lowest probability (0%) in both models.

According to the CI and CU formulas, the values of $Cmax_x$ and $Cmin_x$ are required to examine the explanations. However, estimating $Cmax_x$ and $Cmin_x$ is not a trivial process. To simplify the process, we have applied Monte-Carlo simulation and generated 100 random samples for each feature. This process provides varying in each feature of context (‘Case’) every time and allows to find out how considerable the output has been changed. The samples are uniformly distributed within the range of minimum and maximum values of each sensor data in the training set. To calculate the numerical values of $Cmin_x$ and $Cmax_x$ and later CI and CU , we follow an iterative process. Each time, we modify the values of one sensor data by one of the 100 generated samples while keeping the data of other sensors unchanged. Later, we calculate the class probability of each sample by our neural network and LDA models separately. Therefore, the knowledge about the minimum and maximum class probability within each sensor is obtained, implying for $Cmin_x$ and $Cmax_x$ in the context of our specific instance. Accordingly, the values of CI and CU could be calculated. The process is repeated eight times to extract the appropriate values for all the eight signals of our problem space in the first dataset. In other words, eight different $Cmin_x$, $Cmax_x$, CI and CU values are generated in total with respect to each model. The same procedure is dominated on the second dataset, yet generating seven $Cmin_x$, $Cmax_x$, CI and CU values in accordance with seven sensors of MAHNOB-HCI dataset for each model. In all the iterations, the $absmin$ and $absmax$ values in Eq. 1 are set to 0 and 1 respectively, indicating all

possible values for the class probability (output). Moreover, CI and CU values range between $[0-1]$. To be more readable, the values of CI and CU are then converted to the percentage scale.

5.1 Generated Explanation on WESAD

Table 1 demonstrates the numerical results of aforementioned iterative process in WESAD for both the neural network and LDA models. In addition, Fig. 1 shows visual representations of how important and favorable the sensors are in the detection of “meditation”, the predicted class of our ‘Case’ by both models.

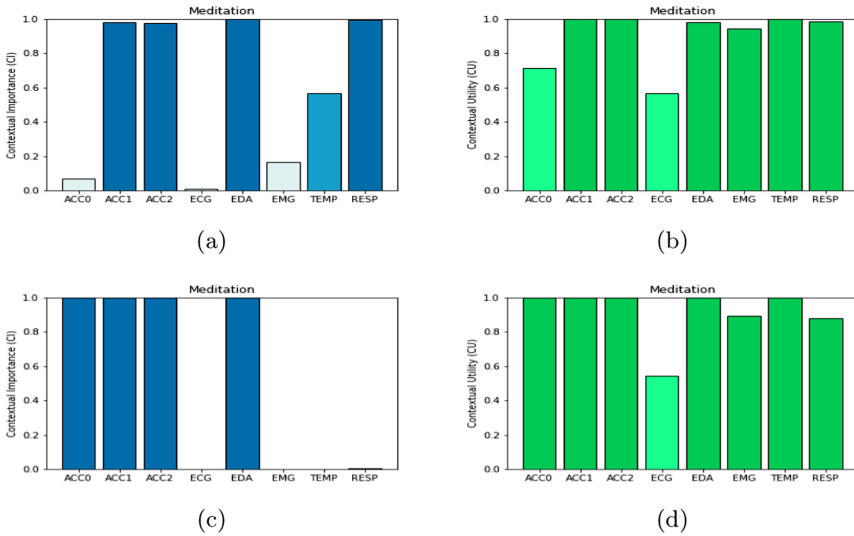
The results of neural network reveal that ACC1, ACC2, EDA and RESP are highly important and favorable sensors contributing in the outcome class, while the other sensors data except TEMP could be ignored within the decision making process (see Figs. 1(a) and 1(b)). In theory, the former sensors produce CI and CU values approximately to 100%, whereas the latter ones provide CI values around zero in spite of (highly) favorable utilities. In practice, the importance of EDA, TEMP and RESP sensors could be justified as the meditation state had been designed to de-excite participants after exciting them in the stress and amusement states. This situation results in either lower average conductance changes at the skin surface or lower variation in temperature and breathing. Similar argument could be true for ACC1 and ACC2 to differentiate the baseline state from meditation since the participants in general were allowed to sit and stand in baseline while only to sit in a comfortable position in the meditation state.

We also observe that ACC1, ACC2 and EDA intervene significantly in the detection of “meditation” state by the LDA model, similar to the results of neural network. These sensors show both the CI and CU values of 100% (see Fig. 1(c) and 1(d)). The values indicate the high change of outcome w.r.t. these sensors and the high contribution extent of these features of “Case” w.r.t. the change. However, despite the high/rather high contributions of RESP and TEMP sensors in the outcome of neural network, they exclude their intervention in case of LDA model. That is because LDA employs the global structure information of the total training data to determine the linear discriminant vectors [1], and thereby the model shows powerless when the statistical patterns of different classes are fairly similar. Clarifying this argument, we depicted the distribution of TEMP data, as an example, in the state of stress and meditation in Fig. 2. Assuming the training set consists of these two states, LDA fails to produce a linear discriminant vector separating the classes. On the other hand, such LDA characteristic leads to the contribution of ACC0 in the model outcome.

In a more granular level, Figs. 3 and 4 illustrate the sensors probability variations within the “meditation” class, in the neural network and LDA models respectively. The red dot point in all subfigures stands for the ‘Case’ sample. The ‘Case’ should be located somewhere between the $Cmin_x$ and $Cmax_x$, comparable with synthetically generated samples. Such location preserves the relative nature of CU concept. The closer the ‘Case’ to $Cmax_x$, the higher utility the sensor has, and in contrary, the farther the ‘Case’ from $Cmax_x$ (closer to

Table 1. Numerical results of models outcome explanation related to WESAD

		ACC0	ACC1	ACC2	ECG	EDA	EMG	TEMP	RESP
Sample		0.898	-0.003	-0.179	-0.003	7.078	0.001	32.97	-0.865
NN	Cmin	0.933	0.0	0.0	0.975	0.0	0.823	0.0	0.0
	Cmax	0.999	0.978	0.975	0.985	0.999	0.990	0.565	0.994
	CI%	7%	98%	98%	1%	100%	17%	57%	99%
	CU%	71%	100%	100%	56%	98%	94%	100%	99%
LDA	Cmin	0.0	0.0	0.0	0.998	0.0	0.997	0.0	0.993
	Cmax	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999
	CI%	100%	100%	100%	0%	100%	0.1%	0%	0.6%
	CU%	100%	100%	100%	54%	100%	89%	100%	88%

**Fig. 1.** (a) CI and (b) CU values of all sensors in meditation state, detected by neural network in WESAD, and (c) CI and (d) CU values of all sensors in meditation state, detected by LDA in WESAD

$Cmin_x$), the lower value for CU . Nevertheless, inferring from Figs. 3(g) and 3(c) as well as Fig. 4(g), the ‘Case’ probability exceeds $Cmax_x$ in ACC2 and TEMP sensors, basically contradicting our previous argument. To solve this problem, we consider the ‘Case’ probability equal to $Cmax_x$, however one could define CU with a constraint $y_{ij} < Cmax_x(C_i)$. Moreover, when ‘Case’ has a lower value than $Cmin_x$, a constraint of $y_{ij} > Cmin_x(C_i)$ enforces the process to produce a random data with at least the same value as the ‘Case’ probability. Therefore, we reformulate the Eq. 2 as follows:

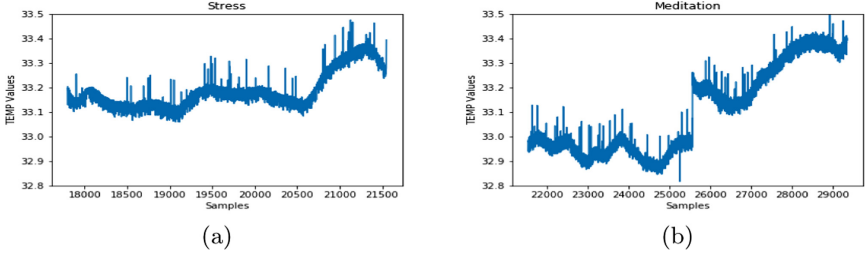


Fig. 2. The data distribution of TEMP sensor in (a) stress and (b) meditation states in WESAD

$$CU = \frac{y_{ij} - Cmin_x(C_i)}{Cmax_x(C_i) - Cmin_x(C_i)} \quad (3)$$

$$\text{s.t. } Cmin_x(C_i) < y_{ij} < Cmax_x(C_i)$$

5.2 Generated Explanation on MAHNOB-HCI

For the second dataset, we followed the same procedure of generating random samples on each sensor. Table 2 shows the importance and utility of sensors in the decision of “surprised” class, based on the CI and CU values calculated using Eqs. 1 and 2. For better clarification, Fig. 5 also visually illustrates the results for each sensor.

In the context of “Case”, the values of ECG2 and RESP sensors are highly contributing in the neural network outcome. As shown in Table 2 and correspondingly Figs. 5(a) and 5(b), CI and CU values are 100%. However, other physiological responses do not represent their relative contribution in the decision of “surprised” class, as their CI and CU values are dropped to 0%). We also found that the data samples of “surprised” class are overlapped completely or partially with the samples of other class (es) in all sensors except ECG2 and RESP (see Fig. 6). Therefore, distinguishing this class from the “amusement” and “happiness” class is a challenging task for a neural network, relying on ECG1, ECG3, GSR1, GSR2 in comparison with ECG2 and RESP sensors. Although TEMP sensor seems to influence the model outcome, due to its fairly non-overlapping class samples, the values of CI and CU could not represent a descent explanation for such case.

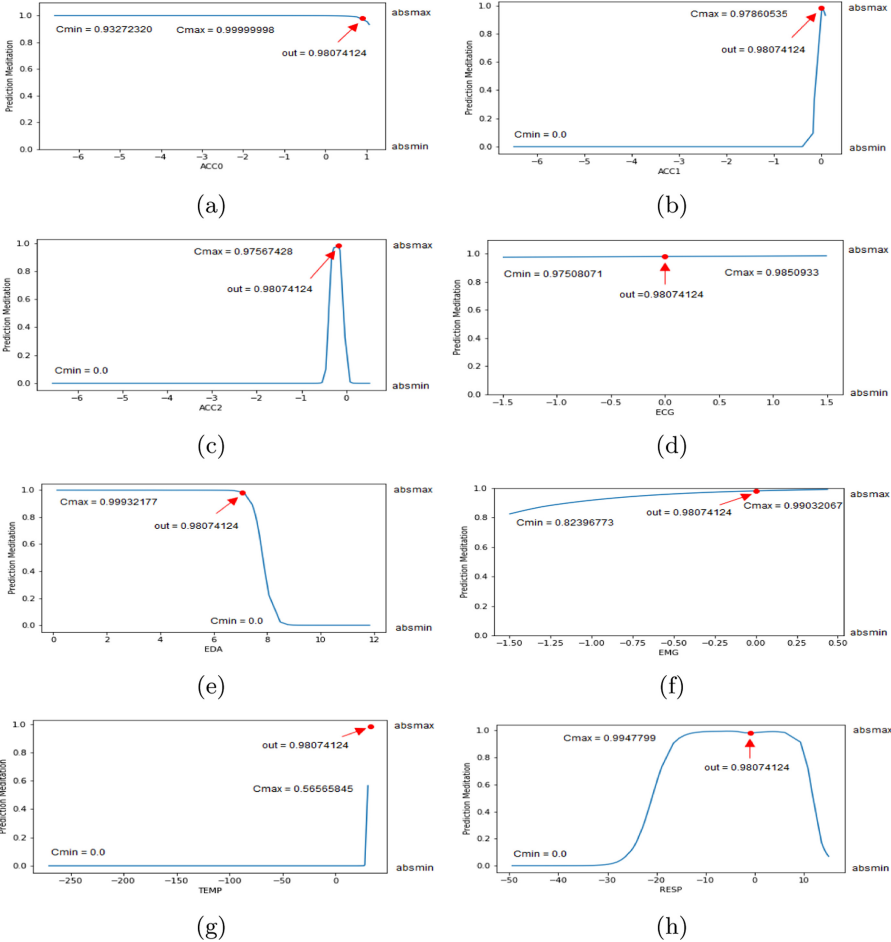


Fig. 3. C_{min} and C_{max} values for input variations in (a) ACC0 (b) ACC1 (c) ACC2 (d) ECG (e) EDA (f) EMG (g) TEMP (h) RESP signals with neural network in WESAD

Applying LDA for making decision about the “Case” in MAHNOB-HCI, the provided explanation only relies on the intervention of ECG2 sensor. In other word, the CI and CU values of this sensor indicate a high importance and utility of 100%. This result is inconsistent with the neural network outcome explanation, as in the latter model, RESP sensor shows a high contribution in the model outcome. This inconsistency partly refers to the intrinsic algorithmic structure of LDA which is sensitive to the outliers [1]. In case of other sensors non-intervention, the same arguments as in the neural network apply well in the LDA as well.

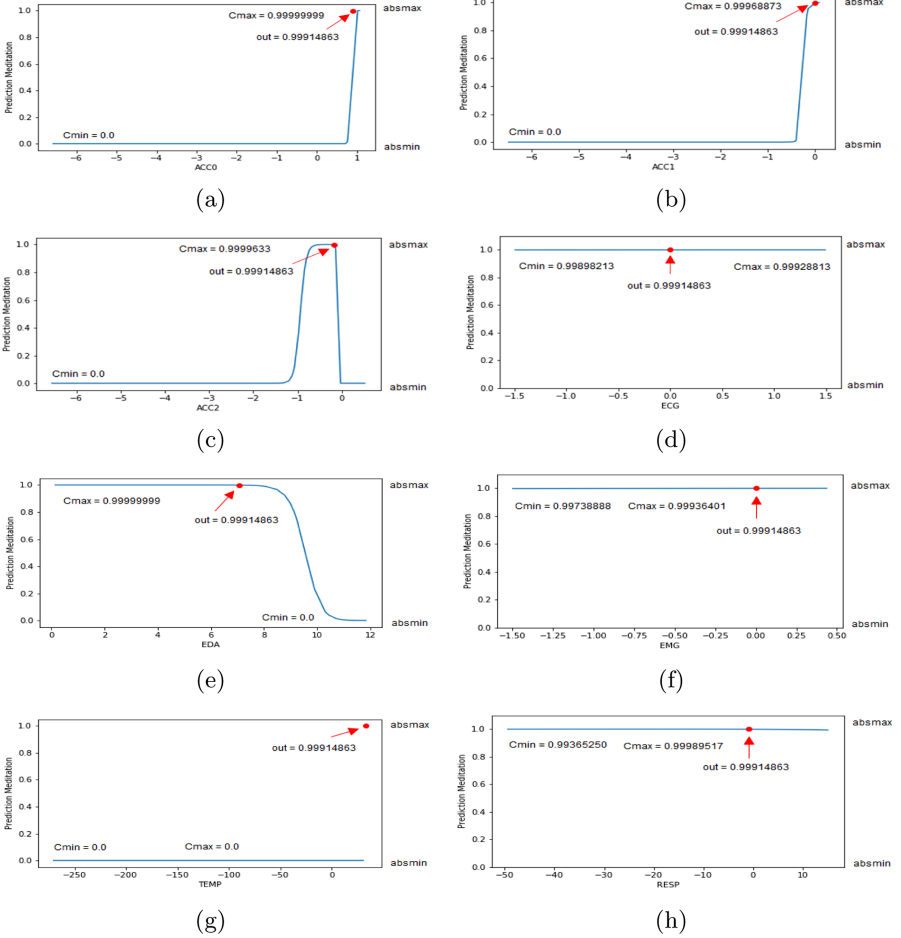


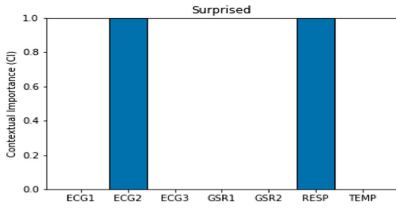
Fig. 4. C_{min} and C_{max} values for input variations in (a) ACC0 (b) ACC1 (c) ACC2 (d) ECG (e) EDA (f) EMG (g) TEMP (h) RESP signals with LDA in WESAD

Complementary Comments on Experiments:

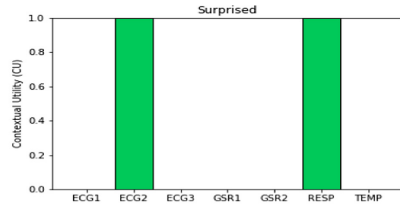
- To better understand the intensity of CI and CU values, we have used different colors in Figs. 1 and 5. The higher the CI and CU values, the darker the colors become.
- Figures 7(a) and 7(b) represent the textual explanation of CI and CU values for all sensors of WESAD and MAHNOB-HCI datasets respectively, in the neural network model. This representation is based on a conversion (see Table 3) from numerical values to linguistic texts [7].
- We have examined other instances of the same class as “Case” in both datasets and ended up with quite similar results shown in the Tables 1 and 2.

Table 2. Numerical results of models outcome explanation related to MAHNOB-HCI

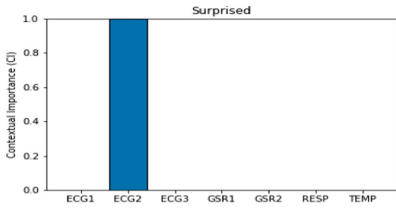
		ECG1	ECG2	ECG3	GSR1	GSR2	RESP	TEMP
Sample		-849000	-544000	-777000	2900000	90	-1560000	26078
NN	Cmin	1.0	0.0	1.0	1.0	1.0	0.0	1.0
	Cmax	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	CI%	0%	100%	0%	0%	0%	100%	0%
	CU%	0%	100%	0%	0%	0%	100%	0%
LDA	Cmin	1.0	0.0	1.0	1.0	1.0	1.0	1.0
	Cmax	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	CI%	0%	100%	0%	0%	0%	0%	0%
	CU%	0%	100%	0%	0%	0%	0%	0%



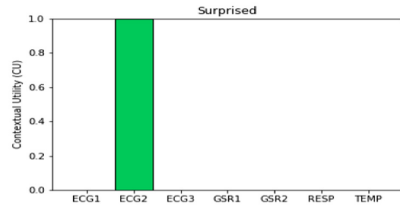
(a)



(b)



(c)



(d)

Fig. 5. (a) *CI* and (b) *CU* values of all signals in surprised state, detected by neural network in MAHNOB-HCI, and (c) *CI* and (d) *CU* values of all signals in surprised state, detected by LDA in MAHNOB-HCI

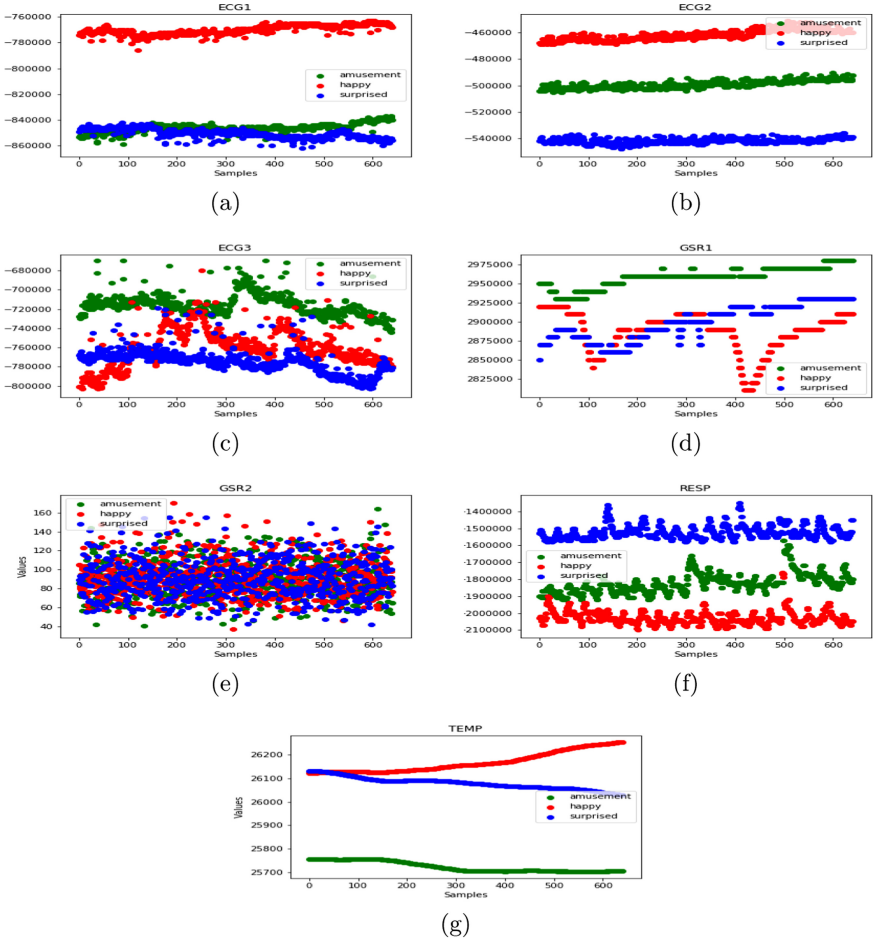


Fig. 6. Three classes of MAHNOB-HCI in (a) ECG1 (b) ECG2 (c) ECG3 (d) GSR1 (e) GSR2 (f) RESP (g) TEMP signals

Table 3. Symbolic representation of the *CI* and *CU* values

Degree (d)	Contextual Importance	Contextual Utility
$0 < d \leq 0.25$	Not important	Not favorable
$0.25 < d \leq 0.5$	Important	Unlikely
$0.5 < d \leq 0.75$	Rather important	Favorable
$0.75 < d \leq 1.0$	Highly important	Highly favorable

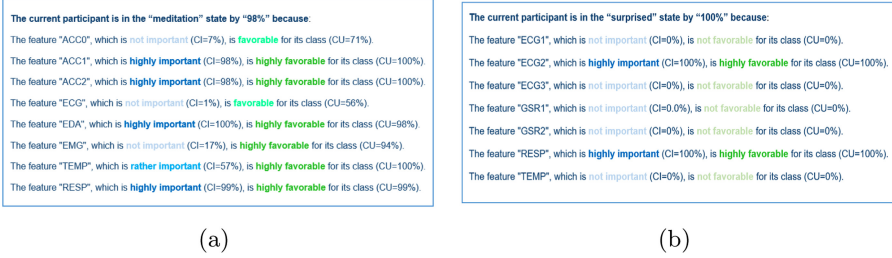


Fig. 7. Textual explanation of neural network prediction for each signal in (a) WESAD (b) MAHNOB-HCI

6 Conclusion and Future Works

We examined one of the earliest concepts in the realm of black-box outcome explanation, namely Contextual Importance (CI) and Contextual Utility (CU). Since these concepts are realized as a model-agnostic explanation approach, we steered our experiments on two black-box models including neural network (NN) and linear discriminant analysis (LDA). The aim was to explain recognition of human mental states and explore the extent of difference in the generated outcome explanation of applied models. We conducted our experiments on WESAD and MAHNOB-HCI, as publicly and academically available benchmarks in the domain of multimodal affective computation. Different sensor data were experimented in the process of personalized decision making in the first and second datasets. The results revealed that the outputs of electrodermal activity, respiration as well as accelerometer sensors, significantly influence the neural network decision in recognition of “meditation” state in the first dataset. Given LDA, the respiration sensor, however, was excluded. In the second dataset, the electrocardiogram and respiration sensors provided interventions in the “surprised” outcome of neural network, while the importance and utility of solely electrocardiogram sensor was demonstrated by LDA. According to these results, we found that not necessarily the same sensors intervened in the detection of the same class by our examined models, although the models accuracy were tightly close to each other. Another interesting finding of explainability referred to the fact that not only the sensors types, but also their position on the body affects the expression of mental states. For instance, in the first dataset only ACC1 and ACC2 and in the second dataset only ECG2 proved their contribution in the decision making of neural network. In conclusion, this work opened a new room of XAI in affective computation by critically examining different classification models.

Some directions of future works are as follow: although CI and CU concepts provided explanations to both expert and non-expert users in terms of visual and textual representations, yet such explanations alone do not meet the requirements of real-world applications. Augmenting these concepts with further clarifications of users type and their current affective states provides more real-

istic explanation. We will also focus on improving the *CI* and *CU* formulations to explain the prediction of more complex models such as deep neural networks, considering additive information from the hidden layers.

References

1. Arunasakthi, K., KamatchiPriya, L.: A review on linear and non-linear dimensionality reduction techniques. *Mach. Learn. Appl.: Int. J.* **1**(1), 65–76 (2014)
2. Chakraborty, S., Aich, S., Joo, M.I., Sain, M., Kim, H.C.: A multichannel convolutional neural network architecture for the detection of the state of mind using physiological signals from wearable devices. *J Healthc Eng.* (2019)
3. Chen, H., Lee, C.: Vibration signals analysis by explainable artificial intelligence (XAI) approach: Application on bearing faults diagnosis. *IEEE Access* **8** (2020)
4. Dragoni, M., Donadello, I., Eccher, C.: Explainable AI meets persuasiveness: translating reasoning results into behavioral change advice. *Artif. Intell. Med.* **105** (2020)
5. Du, M., Liu, N., Hu, X.: Techniques for interpretable machine learning. *Commun. ACM* **63**(1), 68–77 (2019)
6. Eitel, F., Ritter, K.: Testing the robustness of attribution methods for convolutional neural networks in MRI-based Alzheimer’s disease classification. In: Suzuki, K., et al. (eds.) *ML-CDS/IMIMIC -2019*. LNCS, vol. 11797, pp. 3–11. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-33850-3_1
7. Fouladgar, N., Alirezaie, M., Främling, K.: Decision explanation: applying contextual importance and contextual utility in affect detection. In: *Proceedings of the Italian Workshop on Explainable Artificial Intelligence, XAI.it 2020*. AI*IA SERIES, vol. 2742, pp. 1–13 (2020)
8. Fouladgar, N., Främling, K.: XAI-P-T: a brief review of explainable artificial intelligence from practice to theory. [arXiv:2012.09636](https://arxiv.org/abs/2012.09636) (2020)
9. Främling, K.: Decision theory meets explainable AI. In: Calvaresi, D., Najjar, A., Winikoff, M., Främling, K. (eds.) *EXTRAAMAS 2020*. LNCS (LNAI), vol. 12175, pp. 57–74. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-51924-7_4
10. Främling, K.: Explaining results of neural networks by contextual importance and utility. In: *The AISB 1996 conf. Citeseer* (1996)
11. Grath, R.M., et al.: Interpretable credit application predictions with counterfactual explanations. *CoRR* abs/1811.05245 (2018)
12. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. *ACM Comp. Sur.* **51**(5) (2018)
13. Lauritsen, S.M., et al.: Explainable artificial intelligence model to predict acute critical illness from electronic health records. *ArXiv* abs/1912.01266 (2019)
14. Lee, H., Kim, S.T., Ro, Y.M.: Generation of multimodal justification using visual word constraint model for explainable computer-aided diagnosis. In: *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*. pp. 21–29. Springer (2019)
15. Lin, J., Pan, S., Lee, C.S., Oviatt, S.: An explainable deep fusion network for affect recognition using physiological signals. In: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019*, pp. 2069–2072. ACM (2019)
16. Miller, T.: Explanation in artificial intelligence: insights from the social sciences. *Artif. Intell.* **267**, 1–38 (2019)

17. Mukhopadhyay, S.C.: Wearable sensors for human activity monitoring: a review. *IEEE Sens. J.* **15**(3), 1321–1330 (2015)
18. Nisha, P., Pawar, U., O'Reilly, R.: Interpretable machine learning models for assisting clinicians in the analysis of physiological data. In: Proceedings for the 27th AIAI Irish Conference on Artificial Intelligence and Cognitive Science, Galway, Ireland, 5–6 December 2019. CEUR, vol. 2563, pp. 434–445. CEUR-WS.org (2019)
19. Panigutti, C., Perotti, A., Pedreschi, D.: Doctor XAI: an ontology-based approach to black-box sequential data classification explanations. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT* 2020, pp. 629–639. ACM (2020)
20. Papanastasiopoulos, Z., et al.: Explainable AI for medical imaging: deep-learning cnn ensemble for classification of estrogen receptor status from breast MRI. In: *Medical Imaging 2020: Computer-Aided Diagnosis*, vol. 11314, pp. 228–235. International Society for Optics and Photonics, SPIE (2020)
21. Putnam, V., Conati, C.: Exploring the need for explainable artificial intelligence (XAI) in intelligent tutoring systems (ITS). In: *IUI Workshops* (2019)
22. Sayres, R., et al.: Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy. *Ophthalmology* **126**(4), 552–564 (2019)
23. Schmidt, P., Reiss, A., Duerichen, R., Marberger, C., Van Laerhoven, K.: Introducing wesad, a multimodal dataset for wearable stress and affect detection. In: Proceedings the 20th ACM International Conference on Multimodal Interaction, pp. 400–408. ACM (2018)
24. Singh, A., Sengupta, S., Lakshminarayanan, V.: Explainable deep learning models in medical image analysis. *J. Imaging* **6**(6), 52 (2020)
25. Soleymani, M., Lichtenauer, J., Pun, T., Pantic, M.: A multimodal database for affect recognition and implicit tagging. *IEEE Trans. Affective Comput.* **3**(1), 42–55 (2012)
26. Wang, D., Yang, Q., Abdul, A., Lim, B.Y.: Designing theory-driven user-centric explainable AI. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, pp. 1–15. ACM (2019)
27. Zhang, Z., Xie, Y., Xing, F., McGough, M., Yang, L.: MDNet: a semantically and visually interpretable medical image diagnosis network. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3549–3557 (2017)