Decision Explanation: Applying Contextual Importance and Contextual Utility in Affect Detection

Nazanin Fouladgar[1], Marjan Alirezaie[2], and Kary Främling[1,3]

[1] Department of Computing Science, Umeå University, Sweden,
nazanin@cs.umu.se,
[2] Center for Applied Autonomous Sensor Systems, Örebro University, Örebro, Sweden,
marjan.alirezaie@oru.se,
[3] Aalto University, School of Science and Technology, Finland,
kary.framling@umu.se

Abstract. Explainable AI has recently paved the way to justify decisions made by black-box models in various areas. However, a mature body of work in the field of affect detection is still limited. In this work, we evaluate a black-box outcome explanation for understanding humans' affective states. We employ two concepts of Contextual Importance (CI) and Contextual Utility (CU), emphasizing on a context-aware decision explanation of a non-linear model, mainly a neural network. The neural model is designed to detect the individual mental states measured by wearable sensors to monitor the human user's well-being. We conduct our experiments and outcome explanation on WESAD and MAHNOB-HCI, as multimodal affect computing datasets. The results reveal that in the first experiment the electrodermal activity, respiration as well as accelorometer and in the second experiment the electrocardiogram and respiration signals contribute significantly in the classification task of mental states for a specific participant. To the best of our knowledge, this is the first study leveraging the CI and CU concepts in outcome explanation of an affect detection model.

Keywords: Explainable AI· Affect detection · Black-Box decision · Contextual Importance and Utility.

1 Introduction[4]

Due to the exponential growth in producing wearable sensors and also the success of machine learning methods in providing accurate analysis of such sensors' outputs, monitoring patients and in general humans' well-being have been facilitated to a considerable degree [2]. However, since advanced artificial intelligence (AI) methods such as deep learning models lack transparency, solely relying on such methods in critical decision-making processes is not recommended [14]. Health practitioners finalize their decisions more confidently if they are also provided with a concrete outcome explanation of such AI models. Explainable AI (XAI) can furthermore enable end-users to follow their own health track. XAI has recently attracted a great attention among research communities as well as industries [4, 9]. Some scholars have theoretically scrutinized the XAI potentiality [14, 23], while others made efforts to unveil the practical aspects [5, 10]. The main concern in both aspects lies on the ground of intelligent systems transparency and thereby appealing the experts or end-users trust. The role of XAI in addressing the aforementioned issues has justified its applicability in a vast body of works such as tutoring [18], fault diagnosis [3] and healthcare [16].

Despite some research efforts in associating deep learning models with XAI techniques, the intersection of XAI and affective computing (e.g., affect detection) is still immature and there are open rooms for researchers of this area. In this paper, we study the outcome explanation of a neural network model designed to classify humans' state-of-mind. We employ two datasets including WESAD [20], and MAHNOB-HCI [22], as publicly and academically available datasets respectively, in the domain of multi-modal affective computing (see Section 4). Our main focus is on signal-level explanation relying on the two concepts of Contextual Importance (CI) and Contextual Utility (CU) proposed by Främling [8]. By involving the two aforementioned concepts, we represent how important and favorable different criteria (features) are. Both CI and CU represent numerical values applicable in textual and visual representations and thereby understandable to professionals and end-users.

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

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. [21] 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 [19] and also breast imaging in estrogen receptor classification task [17].

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 [21]. 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 [21]. Testing Concept Activation Vectors (TCAV) [24] 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. [12] 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 [15] 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. [11] 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 [13] 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 Främling [8] 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 [7]. 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:

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

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

Where C_i is the ith context (specific input of black-box referring as 'Case' in Section 5), y_{ij} is the value of jth 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 jth 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 multimodal sensory analysis as well as detecting multiple affective states [20]. 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 700HZ, however the sampling rates are different among Empatica E4 signals. BVP, EDA and TEMP data have been recorded in 64Hz, 4Hz, and 32Hz 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 downsampling the signals into 10HZ 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 [22], 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 abdomn 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 256HZ 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 10HZ sampling rate.

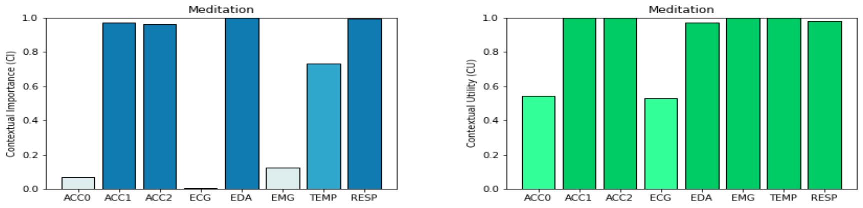
5 Outcome Explanation

The data-driven method employed to classify data into four affective states is a neural network consisting of one hidden layer with 100 units. The basic idea behind these 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 feature explainable. Suppose we consider the following data as an input instance of first dataset (henceforth is referred to as the 'Case'): 0.898 (ACC0), -0.003(ACC1), -0.179 (ACC2), -0.003 (ECG), 7.078 (EDA), 0.001 (EMG), 32.97 (TEMP), -0.865 (RESP). Given the 'Case', the trained neural model results in ''meditation'' state (class) as the classification output with the following probability: meditation class 97%, baseline class 0.4%, stress class 0.1% and amusement class 2%. The same procedure could verify the state of 'Case' in the second dataset. We examine: -849000 (ECG1), -544000 (ECG2), -777000 (ECG3), 2900000 (GSR1), 90 (GSR2), -1560000 (RESP), 26078 (TEMP) as the 'Case' of MAHNOB-HCI dataset. The classification output of our network on this specific instance yields to ''surprised'' state with the highest probability (100%) and to amusement and happiness states with the lowest probability (0%).

According to the CI and CU formulas, the values of $Cmax_x$ and $Cmin_x$ are required to examine the explanation. 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 signals 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 signal by one of the 100 generated samples while keeping the data of other signals unchanged. Later, we calculate the class probability of each sample by our neural network model. This provides the knowledge about minimum and maximum class probability, 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 experiment. In other words, eight different $Cmin_x$, $Cmax_x$, CI and CU values are generated in total. The same procedure is dominated on the second experiment, yet generating seven $Cmin_x$, $Cmax_x$, CI and CU values in accordance with seven signals of MAHNOB-HCI dataset. In all the iterations, the *absmin* and *absmax* values in Equation 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.

Table 1: Numerical results of 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
Cmin	0.933	0.0	0.0	0.965	0.0	0.845	0.0	0.0
Cmax	0.999	0.969	0.959	0.972	0.999	0.969	0.732	0.991
CI%	7%	97%	96%	0.63%	100%	12%	73%	99%
CU%	54%	100%	100%	53%	97%	100%	100%	98%



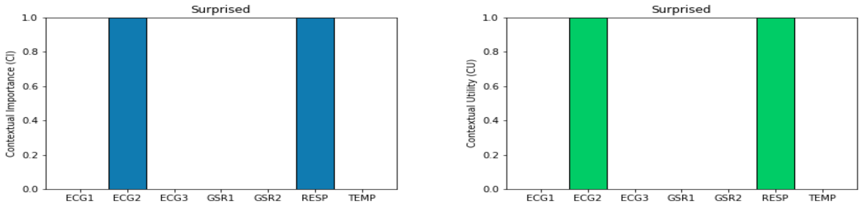
(a) (b)

Fig. 1: (a) CI and (b) CU values of all signals in meditation state

Table 1 demonstrates the numerical results of the aforementioned process regarding the first dataset. In addition, Figure 1 shows the visual representation of how important and favorable the signals of the first dataset are to choose ‘‘meditation’’ state as the predicted class of our ‘Case’. The results reveal that ACC1, ACC2, EDA and RESP are highly important and favorable signals contributing in the outcome class, while the other signals except TEMP could be ignored within the decision making process. More specifically, in theory, the former signals produce CI and CU values around/equal to 100%, whereas the latter signals provide CI values around zero in spite of (highly) favorable utilities. In practice, the importance of EDA, TEMP and RESP signals could be considered as the meditation state had been designed to de-excite participants after exciting them in the stress and amusement states. This result 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.

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

	ECG1	**ECG2**	**ECG3**	**GSR1**	**GSR2**	**RESP**	**TEMP**
Sample	-849000	-544000	-777000	2900000	90	-1560000	26078
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%



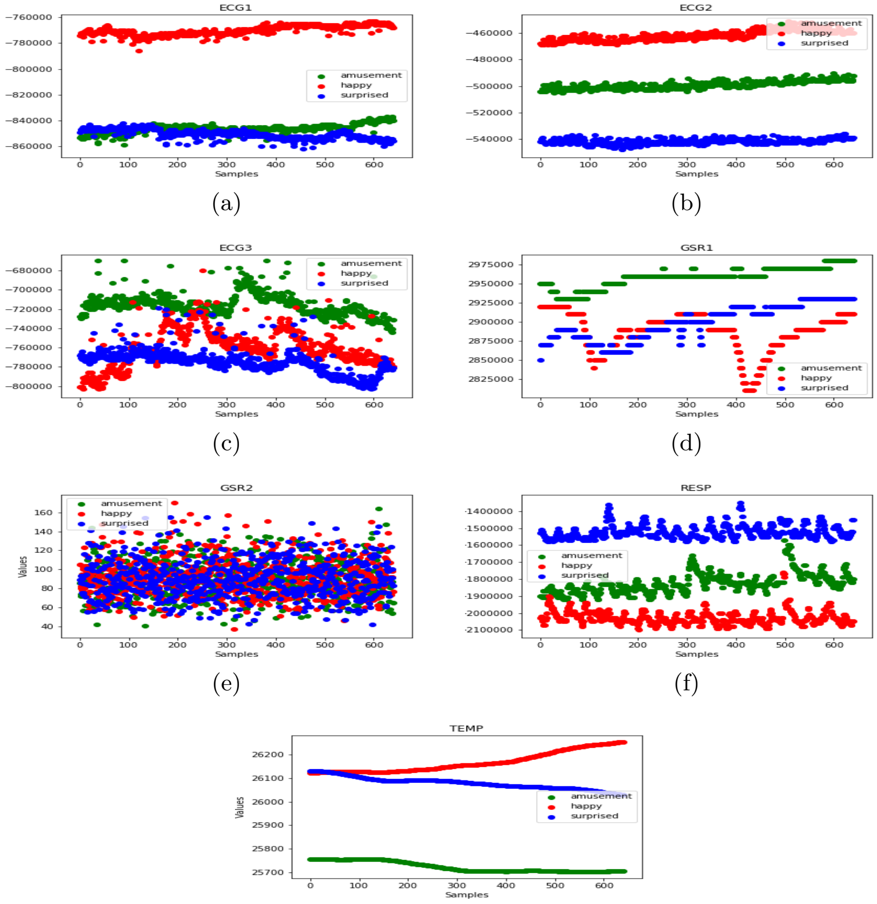
(a)

(b)

Fig. 2: (a) CI and (b) CU values of all signals in surprised state

Following the same procedure in the second dataset, the importance of signals are illustrated in Table 2 and Figure 2. The results unveil that ECG2 and RESP are highly contributing in the ‘‘surprised’’ state (CI and CU values are equal to 100%). However, other physiological responses do not represent their relative contribution (CI and CU values are equal to 0%). Asserting this argument, we found that the statistical specifications of the classes in this database are overlapped on all signals except ECG2 and RESP (see Figure 3). Therefore, distinguishing the ‘‘surprised’’ class from ‘‘amusement’’ and ‘‘happiness’’ are challenging by neural network relying on ECG1, ECG3, GSR1, GSR2 and TEMP in comparison with ECG2 and RESP signals.

It should be noted that we further examined a few other instances of both datasets with the same classes as the ‘Case’ and reached out to (rather) similar results as the Tables 1 and 2.



(a) (b) (c) (d) (e) (f) (g)

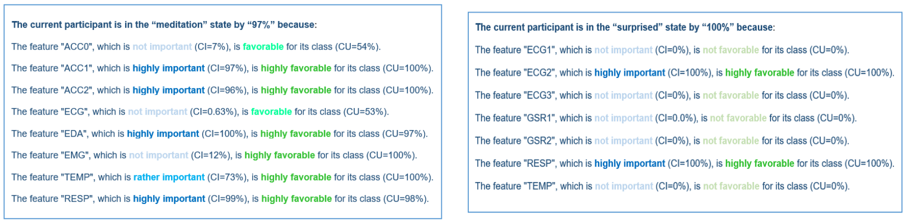
Fig. 3: Three classes of MAHNOB-HCI in (a) ECG1 (b) ECG2 (c) ECG3 (d) GSR1 (e) GSR2 (f) RESP (g) TEMP signals

To better show the intensity of CI and CU values, we have used different colors in Figures 1 and 2 related to the two datasets. The higher the CI and CU values, the darker the colors become. Figures 4 (a) and (b) also represent

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

the textual explanation of CI and CU values related to all the signals. This representation is based on a conversion method (see Table 3) from numerical values to linguistic texts suggested by [1].



The current participant is in the "meditation" state by "97%" because:

The feature "ACC0", which is not important (CI=7%), is favorable for its class (CU=54%).

The feature "ACC1", which is **highly important** (CI=97%), is **highly favorable** for its class (CU=100%).

The feature "ACC2", which is **highly important** (CI=96%), is **highly favorable** for its class (CU=100%).

The feature "ECG", which is not important (CI=0.63%), is favorable for its class (CU=53%).

The feature "EDA", which is **highly important** (CI=100%), is **highly favorable** for its class (CU=97%).

The feature "EMG", which is not important (CI=12%), is **highly favorable** for its class (CU=100%).

The feature "TEMP", which is **rather important** (CI=73%), is **highly favorable** for its class (CU=100%).

The feature "RESP", which is **highly important** (CI=99%), is **highly favorable** for its class (CU=98%).

(a)

The current participant is in the "surprised" state by "100%" because:

The feature "ECG1", which is not important (CI=0%), is not favorable for its class (CU=0%).

The feature "ECG2", which is **highly important** (CI=100%), is **highly favorable** for its class (CU=100%).

The feature "ECG3", which is not important (CI=0%), is not favorable for its class (CU=0%).

The feature "GSR1", which is not important (CI=0.0%), is not favorable for its class (CU=0%).

The feature "GSR2", which is not important (CI=0%), is not favorable for its class (CU=0%).

The feature "RESP", which is **highly important** (CI=100%), is **highly favorable** for its class (CU=100%).

The feature "TEMP", which is not important (CI=0%), is not favorable for its class (CU=0%).

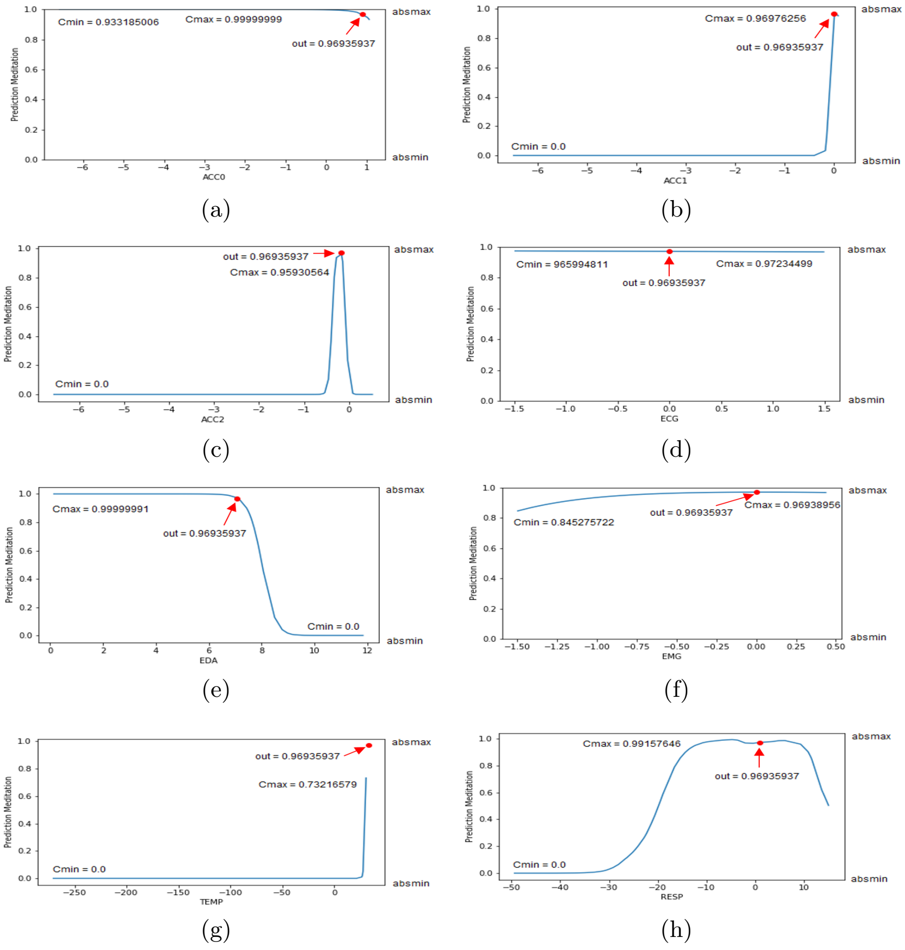
(b)

Fig. 4: Textual explanation of model prediction for each signal in (a) WESAD (b) MAHNOB-HCI

As an example of more granular level, Figure 5 represents each signal's variation within the ‘‘meditation’’ class in WESAD dataset. The red dot point in all subfigures stands for the ‘Case’ sample. As shown in the figure, the ‘Case’ should be located somewhere between the $Cmin_x$ and $Cmax_x$, comparable with synthetically generated samples. This argument preserves the relative nature of CU concept. The closer the ‘Case’ to $Cmax_x$, the higher utility the signal has and in contrary, the further away the ‘Case’ from $Cmax_x$ (closer to $Cmin_x$), the lower CU is generated. However, inferring from TEMP and ACC2 signals in WESAD dataset, (see Figures 5 (g) and (c)) the ‘Case’ probability exceeds $Cmax_x$, 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, in a situation where ‘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 Equation 2 as follows:

$$CU = \frac{y_{ij} - Cmin_x(C_i)}{Cmax_x(C_i) - Cmin_x(C_i)} \qquad (3)$$
$$\text{s.t.} \quad Cmin_x(C_i) < y_{ij} < Cmax_x(C_i)$$

In conclusion, further experiments are required to explore the limitations of CI and CU concepts in the context of black-box outcome explanation and multimodal affective computation. Although these concepts could provide explanations to both the 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. A higher level of explanation should be integrated w.r.t both the skills of experts as well as the affective state of non-expert users. However, as we mentioned previously, CI and CU concepts are theoretically correct from the Decision Theory point of view [7].



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

Fig. 5: $Cmin$ and $Cmax$ values for input variations in (a) ACC0 (b) ACC1 (c) ACC2 (d) ECG (e) EDA (f) EMG (g) TEMP (h) RESP signals in WESAD

6 Conclusion

In this paper, we examined one of the earliest concepts in the black-box outcome explanation. We filled out the gap in explaining the detection of human mental states by utilizing Contextual Importance (CI) and Contextual Utility (CU) concepts. The concepts are assigned as model-agnostic explanations [5], applicable in linear and non-linear black-box models. In this study, we focused on neural network model as a non-linear function approximator. For this purpose, we conducted two experiments on WESAD and MAHNOB-HCI, as publicly and academically available benchmarks in the area of multimodal affective computation. Choosing wearable sensors, different signals were experimented in the process of personalized decision making in the first and second datasets. We further explained the outcome of neural network by CI and CU. The results revealed that mainly electrodermal activity, respiration as well as accelorometer, have significantly contributed in the ‘‘meditation’’ state of the first experiment, in terms of feature importance and utility. Moreover, in case of second experiment, the electrocardiogram and respiration provided intervention in the ‘‘surprised’’ outcome of examined neural network. More 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 as in the first experiment only ACC1 and ACC2 and in the second experiment only ECG2 proved their contribution in the decision making. In conclusion, this work opened a new room of XAI in health domain applications by critically examining affect detection problems. For future work, we will 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. In addition, augmenting the generated explanation with further clarifications can be performed for different types of users.

References

1. Anjomshoae, S., Främling, K., Najjar, A.: Explanations of black-box model predictions by contextual importance and utility. In: Calvaresi, D., Najjar, A., Schumacher, M., Främling, K. (eds.) Explainable, Transparent Autonomous Agents and Multi-Agent Systems. p. 95–109. Springer (2019)
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. Artificial Intelligence in Medicine **105**, 101840 (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: Interpretability

of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support. pp. 3--11. Springer (2019)

7. Främling, K.: Decision theory meets explainable ai. In: Explainable, Transparent Autonomous Agents and Multi-Agent Systems. pp. 57--74. Springer (2020)
8. Främling, K.: Explaining results of neural networks by contextual importance and utility. In: the AISB'96 conf. Citeseer (1996)
9. Grath, R.M., Costabello, L., Van, C.L., Sweeney, P., Kamiab, F., Shen, Z., Lécué, F.: Interpretable credit application predictions with counterfactual explanations. CoRR **abs/1811.05245** (2018)
10. 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)
11. Lauritsen, S.M., Kristensen, M.R.B., Olsen, M.V., Larsen, M.S., Lauritsen, K.M., Jørgensen, M.J., Lange, J., Thiesson, B.: Explainable artificial intelligence model to predict acute critical illness from electronic health records. ArXiv **abs/1912.01266** (2019)
12. 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)
13. Lin, J., Pan, S., Lee, C.S., Oviatt, S.: An explainable deep fusion network for affect recognition using physiological signals. In: Proc. of the 28th ACM Int. Conf. on Information and Knowledge Management. p. 2069–2072. CIKM '19, ACM (2019)
14. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence **267**, 1–38 (2019)
15. Nisha, P., Pawar, U., O'Reilly, R.: Interpretable machine learning models for assisting clinicians in the analysis of physiological data. In: Proc. for the 27th AIAI Irish Conf. on Artificial Intelligence and Cognitive Science, Galway, Ireland, December 5-6, 2019. CEUR, vol. 2563, pp. 434--445. CEUR-WS.org (2019)
16. Panigutti, C., Perotti, A., Pedreschi, D.: Doctor xai: An ontology-based approach to black-box sequential data classification explanations. In: Proc. of the 2020 conf. on Fairness, Accountability, and Transparency. p. 629–639. FAT* '20, ACM (2020)
17. Papanastasopoulos, Z., Samala, R.K., Chan, H.P., Hadjiiski, L., Paramagul, C., Helvie, M.A., Neal, C.H.: 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)
18. Putnam, V., Conati, C.: Exploring the need for explainable artificial intelligence (xai) in intelligent tutoring systems (its). In: IUI Workshops (2019)
19. Sayres, R., Taly, A., Rahimy, E., Blumer, K., Coz, D., Hammel, N., Krause, J., Narayanaswamy, A., Rastegar, Z., Wu, D., Xu, S., Barb, S., Joseph, A., Shumski, M., Smith, J., Sood, A.B., Corrado, G.S., Peng, L., Webster, D.R.: Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy. Ophthalmology **126**(4), 552--564 (2019)
20. Schmidt, P., Reiss, A., Duerichen, R., Marberger, C., Van Laerhoven, K.: Introducing wesad, a multimodal dataset for wearable stress and affect detection. In: Proc. the 20th ACM Int. Conf. on Multimodal Interaction. p. 400–408. ACM (2018)
21. Singh, A., Sengupta, S., Lakshminarayanan, V.: Explainable deep learning models in medical image analysis. Journal of Imaging **6**(6), 52 (2020)
22. Soleymani, M., Lichtenauer, J., Pun, T., Pantic, M.: A multimodal database for affect recognition and implicit tagging. IEEE Trans. on Affective Computing **3**(1), 42--55 (2012)

23. Wang, D., Yang, Q., Abdul, A., Lim, B.Y.: Designing theory-driven user-centric explainable ai. In: Proc. of the 2019 CHI conf. on Human Factors in Computing Systems. p. 1–15. CHI '19, ACM (2019)
24. Zhang, Z., Xie, Y., Xing, F., McGough, M., Yang, L.: Mdnet: A semantically and visually interpretable medical image diagnosis network. IEEE conf. on Computer Vision and Pattern Recognition (CVPR) pp. 3549--3557 (2017)