Fouladgar, Nazanin; Alirezaie, Marjan; Framling, Kary

Metrics and Evaluations of Time Series Explanations: An Application in Affect Computing

Published in:
IEEE Access

DOI:
10.1109/ACCESS.2022.3155115

Published: 01/01/2022

Document Version
Publisher's PDF, also known as Version of record

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 9, 2022, accepted February 22, 2022, date of publication February 28, 2022, date of current version March 8, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3155115

Metrics and Evaluations of Time Series Explanations: An Application in Affect Computing

NAZANIN FOULADGAR[1], MARJAN ALIREZAIE[2], AND KARY FRÄMLING[1,3], (Member, IEEE)

[1]Department of Computing Science, Umeå University, 901 87 Umeå, Sweden
[2]Centre for Applied Autonomous Sensor Systems (AASS), Örebro University, 702 81 Örebro, Sweden
[3]School of Science and Technology, Aalto University, 02150 Espoo, Finland

Corresponding author: Nazanin Fouladgar (nazanin@cs.umu.se)

This work was supported by the Umeå University.

ABSTRACT Explainable artificial intelligence (XAI) has shed light on enormous applications by clarifying why neural models make specific decisions. However, it remains challenging to measure how sensitive XAI solutions are to the explanations of neural models. Although different evaluation metrics have been proposed to measure sensitivity, the main focus has been on the visual and textual data. There is insufficient attention devoted to the sensitivity metrics tailored for time series data. In this paper, we formulate several metrics, including *max short-term sensitivity (MSS)*, *max long-term sensitivity (MLS)*, *average short-term sensitivity (ASS)* and *average long-term sensitivity (ALS)*, that target the sensitivity of XAI models with respect to the generated and real time series. Our hypothesis is that for close series with the same labels, we obtain similar explanations. We evaluate three XAI models, LIME, integrated gradient (IG), and SmoothGrad (SG), on CN-Waterfall, a deep convolutional network. This network is a highly accurate time series classifier in affect computing. Our experiments rely on *data-*, *metric-* and *XAI hyperparameter*-related settings on the WESAD and MAHNOB-HCI datasets. The results reveal that (i) IG and LIME provide a lower sensitivity scale than SG in all the metrics and settings, potentially due to the lower scale of important scores generated by IG and LIME, (ii) the XAI models show higher sensitivities for a smaller window of data, (iii) the sensitivities of XAI models fluctuate when the network parameters and data properties change, and (iv) the XAI models provide unstable sensitivities under different settings of hyperparameters.

INDEX TERMS Explainable AI, metrics, time series data, deep convolutional neural network.

I. INTRODUCTION

Since artificial intelligence (AI)-based applications are increasingly becoming an integral part of our world, the role of XAI in explaining decisions made by neural-based AI models (known as black-boxes) is becoming more critical in different areas [1]–[6]. Specifically, due to the complex structures of machine learning (ML) models in processing time-series data, a lack of explanation may result in isolation of such models in critical decision making, despite their high performance and accuracy. Assume that an ML model, designed for affect computing and fed with time-series sensor data, can detect when a specific individual has been in *stress* as an affective state. Without an explanation of why such a state is recognized, one may not be able to fully rely on the decision made by the system. This limitation could also mislead an expert in charge and cause irreparable medical decisions.

Different XAI solutions are studied in the literature [7]–[10], and are categorized into two classes of *gradient-* and *perturbation-based* solutions [11], [12]. In the former class, the gradient of the output with respect to a specific instance [7], [8] or all instances along a path to a baseline [9] is considered as an explanation. In the latter class, the explanation consists of the output change after replacing the features with randomly permuted values [10]. It is also possible to approximate an interpretable model on a local neighborhood of data of interest in perturbation-based approaches [13], [14].

To evaluate the effectiveness of both categories of explanations, researchers proposed different taxonomies [15], [16]. There are two types of measurements, *qualitative (subjective)* [17], [18] and *quantitative (objective)* [19] that have been classified explicitly in [16]. The qualitative metrics rely

The associate editor coordinating the review of this manuscript and approving it for publication was Yiqi Liu.

on whether humans are satisfied with the explanation and able to understand the model [16]. On the other hand, the quantitative metrics support the theoretically sound foundations and allow a scarce assessment of the state-of-the-art explanation models. We argue that the latter metrics are more convenient for time series data due to the complex nature of such data in both the time and feature domains.

Despite the efforts to introduce and formalize quantitative metrics on different data types [16], [19]–[24], applications of such metrics tailored explicitly for time series data are still missing [25]. More specifically, a number of works [16], [19], [21], [24] have formulated *sensitivity* metrics applicable to image data and examined the stability of the explainable models against perturbations. However, to the best of our knowledge, there are no formal counterpart definitions of these metrics applicable to time series data. As such data (e.g., sensor-based data) are usually noisy in real scenarios, it is highly important to explore how sensitive the XAI models are with respect to such data under different settings. In this paper, we generate temporal-based perturbations upon the series of interest with similar class labels as the latter series. We hypothesize that the XAI models should provide similar explanations for the perturbed and original series. Therefore, one may expect low explanation sensitivities with respect to fluctuations. We also consider the same hypothesis for the clean data, as there are usually training neighbors around the series of interest with the same labels as the latter series.

Incorporating these ideas, we formulate four sensitivity metrics, following comprehensive evaluations on three folds of *data*-, *metric*- and *XAI hyperparameter*-related settings. We employ different explainable methods, namely LIME [13] as a perturbation-based approach, and IG [9] and SG [8] as gradient-based approaches. These XAI models are model-agnostic with respect to any deep learning method. We perform our experiments on a specific black-box, called CN-Waterfall [26], a highly accurate deep neural network. CN-Waterfall is applicable to affect computing time series datasets, namely, WESAD [27] and MAHNOB-HCI [28].

We outline our contributions as follows:

- proposing temporal-based *sensitivity* metrics, namely *max short-term sensitivity (MSS)*, *max long-term sensitivity (MLS)*, *average short-term sensitivity (ASS)* and *average long-term sensitivity (ALS)* tailored for time series data. The metrics aim at evaluating the sensitivity of explanations with respect to series fluctuations under different settings.
- conducting comprehensive experiments and finding out that (i) IG and LIME provide a lower scale of sensitivity than SG in all the metrics and settings, potentially due to the lower scale of important scores generated by IG and LIME, (ii) the XAI models are more sensitive to a smaller window of data, (iii) the sensitivities of XAI models fluctuate by the change in network parameters and data properties, and (iv) the XAI models vary in sensitivities with respect to different settings of the hyperparameters.

The rest of the paper is structured as follows: Section II reviews the literature. Section III overviews the examined neural-based model and datasets. In Section IV, we present the proposed metrics, followed by Section V, which describes the conducted experiments. In Section VI, we discuss some obstacles in the practice of the applied XAI models, and lastly we conclude the paper in Section VII where we also discuss future research directions.

II. RELATED WORKS

In this paper, we focus on the evaluation categories of [16] and review recent literature concerned with the *quantitative* metrics of explanations. We first go through the evaluation metrics applied to data types other than time series data and then explore the measurements used on time series. An overview of these metrics with respect to their tailored data type is shown in Table 1.

A. QUANTITATIVE METRICS ON NON-TIME SERIES DATA

Adebayo *et al.* [20] shed light on the inadequacy of saliency maps as a *sanity check* on image data. A sanity check explores whether the explanation of the network changes when the network properties and data labels are randomly perturbed. Passing the sanity check means that the concerned saliency maps were different and thereby faithful to the network and data.

Ghorbani *et al.* [21] also raised awareness of the fragility of neural networks interpretation with respect to adversarial attacks. Characterizing *robustness*, the authors showed that a systematic perturbation to the input data could result in dramatically different interpretations, while the class label remained the same as the clean data. Similarly, some works [16], [19] examined the degree of explanation *sensitivity* and the work in [24] defined *attributional robustness* with respect to perturbations and/or close data.

In [16], [19], [22], the authors quantified *faithfulness (fidelity)* to show that a change in the output should be proportional to the sum of attribution scores of features that are set as baseline. This metric was also presented under the name of *sensitivity-n* in [29] and generalized as *infidelity* in [16]. In a different definition, some *fidelity* metrics were introduced in [14] to compare the prediction of an interpretable model and a black-box. We cite these metrics as *faithful to black-box*.

Stepping further, the authors in [19] proposed several other quantified metrics, such as *complexity*, to address the problem when all the features are used in the explanation, *identity*, to favour non-stochastic explanation, *separability*, to indicate how surprising the explanation of an instance is compared to its counterpart on training data, *conviction*, to emphasize the expected amount of surprise that can predictably occur, *deletion* and *addition*, to show how confidently a model predicts if a subset of important features are deleted and added to the baseline, respectively, and *ROAR* and *KAR*, to denote the difference in accuracy between the original and modified

predictors when removing the most and least important features, respectively.

Monotonicity is another metric proposed in [22], [23]. Using this metric, one can measure whether adding more positive evidence can increase the decision probability [22] or measure how correlated a feature importance is with the impreciseness of prediction emerging from an unknown value of the feature [23]. In [23], the authors also discussed explanation robustness to nonimportant features under the flag of the *non-sensitivity* metric. Another metric introduced by Nguyen and Martínez [23] was *mutual information*, showing how much features and prediction information were lost after the feature extraction process. In the context of example-based explanations, the two metrics of *non-representativeness* and *diversity* were proposed [23]. According to these metrics, one can specify how much the selected examples are faithful to the prediction and how broad these examples are. In this work [23], the authors also considered feature interactions to explain complexity by a metric called *effective complexity*. The main motivation relied on ignoring some features even if they influenced the prediction, for the sake of explanation complexity.

Finally, Guidotti and Ruggieri [30] highlighted *stability* analysis of some interpretable models with respect to different design choices. In an analysis, *stability* was quantified through the deviation of a measure (e.g., number of features used) distribution over the models, learnt from different samples of population. In another analysis, *stability* was quantified by the mean value of similarity (in the number of shared features) over all pairs of the models.

B. QUANTITATIVE METRICS ON TIME SERIES DATA

Following the quantifiable metrics on time series data, the *sanity check* was also performed on the filter influences of a CNN-based anomaly detection system [31]. In this work, the last convolutional layer filters were pruned (i.e., their values were set to zero) to check whether the removal of the most influential filters or the least influential filters had more impact on the final performance. The authors later applied the *sanity check* upon the input data in two levels of point-wise and sequence-wise [32] checks. At the point-wise level, a masking process suppressed the anomalous data point for which the explanation was provided textually. However, in the sequence-wise level, three points, including the data point and its preceding and following points, were masked to explore the most salient region presented by the explanation. Moreover, the work in [33] asserted the quality of a counterfactual-based explanation by the *sanity checks* on data and model parameters. The authors of this literature [33] showed a significant deterioration on the proposed explanation performance, thus passing the check.

In [25], two techniques, called swap time points and mean time points, were presented to *verify* XAI methods on CNN and RNN models. In the former technique, the values of a subsequence were swapped with respect to their time order. The start point of the sub-sequence was assigned to the time point whose relevance score was higher than a specific threshold. In the latter technique, the same process was applied; however, instead of swapping the time points values, the points were set to the mean of the values. Another explanation evaluation on a convolutional-based network was provided in [6]. In this work, the *validity* of a proposed explanation approach was certified by the recall and F1-score quantities. More precisely, the network was retrained with the most contributed features, and later, its performance was compared with the trained network on the full set of features. The generated explanation is valid if the former network achieved similar recall and F1-score as the latter one.

In addition to *validity*, Delaney *et al.* [34] presented *goodness* of explanations quantitatively in the context of counterfactuals. The authors formulated this metric in the form of *relative counterfactual distance (RCF)* and *out of distribution (OOD)* computations. In detail, *RCF* compared the distance of the to-be-explained sequence from a training time series of a different class with the distance of the to-be-explained sequence from a generated counterfactual. A *good* explanation was expected to assign a smaller value to the latter distance than the former. In the case of *OOD*, the aim was to avoid selecting a counterfactual out of the distribution of the to-be-explained sequence. This task was accomplished by a local outlier factors (LOF) algorithm [35], which measures a local deviation of a given data point from its neighbors. Relying on a set of latent exemplars and counter-exemplars, Guidotti *et al.* [36] certified the *usefullness* of explanations by training two 1-NN classifiers. The first classifier was trained on n random (counter-) exemplars while the second classifier learned n random real time series excluding the to-be-explained sequence. It was argued that if the former network showed a higher accuracy than the latter network in classifying the to-be-explained sequence, the explanation would be *usefull*. Reference [36] also quantified the *faithful to black-box* metric by comparing the output of a shapelet-based decision tree with the output of a black-box on the to-be-explained sequence. Moreover, [36] designed the *coherency* property of explanation as a similarity between the explanation of the furthest and closest sequence to the to-be-explained sequence.

Although different metrics have been proposed in the literature, many of them have not been applied to time series data. In our work, we establish the *sensitivity* metrics based on temporal perturbations and real-data consideration, following a comprehensive evaluation.

III. THE BLACK-BOX MODEL AND DATA

In this paper, we apply a specific black-box model, called CN-Waterfall, proposed by Fouladgar *et al.* in [26]. The CN-Waterfall model was designed to detect a set of human affective states with 99% accuracy, superior to several traditional and deep learning models. CN-Waterfall was examined by two publicly and academically available time series datasets, WESAD and MAHNOB-HCI, respectively.

TABLE 1. An overview of recent quantitative metrics with respect to the tailored data types. The red check mark and texts show our contribution in this paper.

Ref.	Year	Quantitative Metric	Non-Time Series	Time Series
[20], [31]–[33]	2018,2019,2019,2020	Sanity Check	✓	✓
[21]	2019	Robustness	✓	×
[16], [19], [24], our work	2019, 2020, 2019, 2022	Sensitivity/Attributional Robustness	✓	✓
[16], [19], [22], [29]	2019, 2020, 2019, 2018	Faithfulness/Fidelity/Sensitivity-n	✓	×
[16]	2019	Infidelity	✓	×
[19]	2020	complexity	✓	×
[19]	2020	Identity	✓	×
[19]	2020	Separability	✓	×
[19]	2020	Conviction	✓	×
[19]	2020	Deletion	✓	×
[19]	2020	Addition	✓	×
[19]	2020	Roar	✓	×
[19]	2020	Kar	✓	×
[22], [23]	2019, 2020	Monotonocity	✓	×
[23]	2020	Non-Sensitivity	✓	×
[23]	2020	Mutual Information	✓	×
[23]	2020	Non-Representativeness	✓	×
[23]	2020	Diversity	✓	×
[23]	2020	Effective Complexity	✓	×
[30]	2019	Stability	✓	×
[25]	2019	Verification	×	✓
[6]	2019	Validity	×	✓
[34]	2020	Goodness	×	✓
[36]	2020	Usefulness	×	✓
[14], [36]	2018, 2020	Faithful to black-box	✓	✓
[36]	2020	Coherency	×	✓

Looking into CN-Waterfall, shown in Figure 1, two main components -*Base* and *General*- were presented [26]. First, the low-level data representation of each modality was learned by the *Base* module, providing a signal representation (SR), and later the intermediate- and high-level representations of correlated and non-correlated series were learned by the *General* module. The *Base* module consisted of two convolutional blocks, while the *General* module consisted of four fusions, satisfying the learning strategies. The first and second convolutional blocks of the *Base* module were equipped with 32 and 128 filter sizes, and the dense layers of all fusions (except the last one) in the *General* module consisted of 64 neurons. The dense layer of the last fusion was assigned the same number of neurons as the number of class labels in the problem space.

Briefly speaking about the datasets, WESAD was introduced by Schmidt *et al.* and is a collection of human emotional and stress states by means of several wearable sensors. In [26], the records of eight signals (modalities) of chest-worn sensors were selected from the collected data. The modalities consisted of a 3-axis accelerometer (ACC0, ACC1, ACC2), respiration (RESP), electrocardiogram (ECG), electrodermal activity (EDA), electromyogram (EMG) and skin temperature (TEMP) of 15 participants. Aligned with the laboratory protocols of data collection in WESAD, [26] also employed four emotional states: *neutral*, *amusement*, *stress* and *meditation*.

MAHNOB-HCI was introduced by Soleymani *et al.* in 2012 and uses various physiological sensors for emotion recognition. Considering the collected data of 7 out of 27 participants in [26], seven sensors were chosen as follows: three ECG electrodes (ECG1, ECG2 and ECG3), two GSRs (GSR1 and GSR2), TEMP and RESP. Reference [26] also retrieved three affective states of *amusement*, *happiness* and *surprised* among other states from MAHNOB-HCI.

Both datasets were downsampled to 10 Hz and separately unified in terms of series length. The unification process resulted in a balanced dataset for MAHNOB-HCI and an imbalanced dataset for WESAD. Finally, all data were normalized and segmented into windows of 30 time steps with 10 overlapping instances. In total, 43290 windows of 8 series with 30 time steps were obtained for WESAD, and 1323 windows of 7 series with 30 time steps were obtained for MAHNOB-HCI.

For further details of the logic behind the CN-Waterfall architecture as well as the prepossessing steps on the two datasets, we refer the readers to these studies [5], [26], [37].

IV. EVALUATION METRICS

In this section, we introduce four different *sensitivity* metrics, including *max short-term sensitivity (MSS)*, *max long-term sensitivity (MLS)*, *average short-term sensitivity (ASS)* and *average long-term sensitivity (ALS)*, adapted explicitly for the evaluation of XAI models on time series data. Using these metrics, we measure how sensitive the XAI models are with respect to *close series* of the same class. We hypothesize that these models should provide similar explanations and thereby lower sensitivities for such series. To this end, we use temporal-based perturbations and training neighbors of the series of interest as two sets of *close series*.

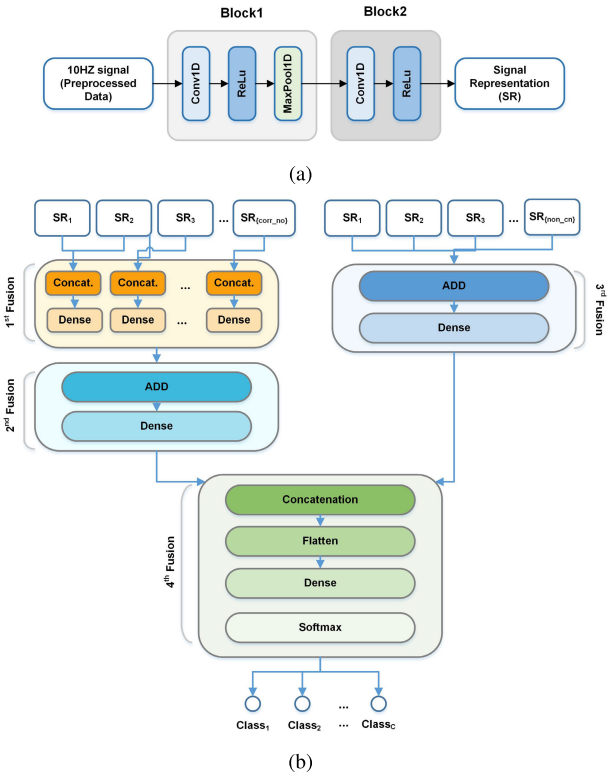


FIGURE 1. CN-Waterfall: (a) The *Base* and (b) *General* modules architecture [26]. The *Base* module learns the low-level data representation of modalities and outputs their signal representation (SR). The representations are then employed by the *General* module, allowing the network to learn intermediate- and high-level features. The convolutional networks in Block1 and Block2 of the *Base* module consist of 32 and 128 filters, respectively. With respect to the *General* module, all the dense layers are designed by 64 neurons, except the dense layer of the fourth fusion. This layer consists of the same number of neurons as the number of classes.

To generate the perturbations, we first transform the time series data (x) into its vectorized representation (x'), called the to-be-explained series (see Figure 2). Such transformation is provided by sequentially attaching all modalities of each time step in x to those of the previous time step. Next, we consider two index lists of $L = \{0, .., d/2 - 1\}$ and $S = \{d/2, .., d\}$, where d denotes the size of x'. Given the lists, the features of x' indexed at L are randomly perturbed with a normal distribution within the radius r, and the rest of the features are kept unchanged. We denote the generated data as a long-term perturbed series (l). The same process is applied in the case of generating short-term perturbed series (s); however, the set of perturbed features is indexed at S.

Considering the training neighbors as another set of *close series*, we select the training data in the radius r of the to-be-explained series.

To measure the similarity between the explanations of short-/long-term perturbed and to-be-explained series (D), we use the Euclidean distance. The same similarity measurement is applied between the explanations of the training neighbours and to-be-explained series. In general, the higher the distance is, the lower the similarity of explanations.

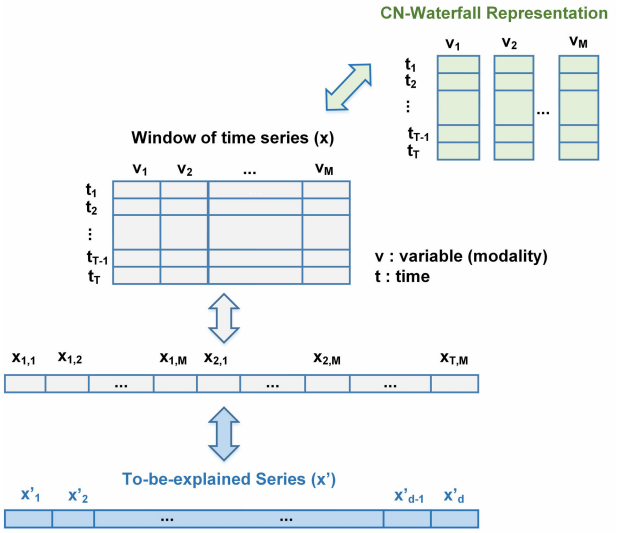


FIGURE 2. Two representations of a window of time series (x). To predict the series by CN-Waterfall, the modalities are transformed into parallel representations (shown in green), in accordance with the input structure of CN-Waterfall. To explain the series by the XAI models, generate temporal-based perturbations and extract training neighbors, a vectorized representation of x, called the to-be-explained series (x'), is provided (shown in blue).

Following these processes, we formulate the *MSS* metric by calculating the maximum distance of the least similar explanations in short-term perturbed series and training neighbors with respect to the to-be-explained series. We perform the same calculations for the *MLS* metric, except that we use the explanations of long-term perturbed series rather than short-term ones.

Formulating the *ASS* and *ALS* metrics, we first take an average of sensitivities for each set of *close series*. By such calculation, the variation center of each set (μ_{A^n}, μ_{A^s} and μ_{A^s} in Section IV-A) is explored. Then, we calculate the average of centers as the final values for *ASS* and *ALS*. In the case of the former metric, the set of short-term perturbed series is employed while in the latter metric, the set of long-term perturbed series is used.

A. SENSITIVITY METRICS

Applying the notations of Table 2, in this section, we mathematically define four *sensitivity* metrics.

Definition 1 (Max Short-Term Sensitivity (MSS)): Given $ST = \{s \in \mathcal{D}_s, \mid \rho(x', s) <= r;\ f(x') = f(s)\}$ as a set of short-term perturbed sequences, $N = \{n \in \mathcal{D}_x, \mid \rho(x', n) <= r;\ f(x') = f(n)\}$ as a set of training neighbors, the black-box f, the explainer g, the distance metric over explanations D and the to-be-explained series x', we define *MSS* of g at x' as:

$$\mu_{MSS}(f, g, r, x') = \max(\max_{n \in N} \{D(g(f, x'), g(f, n))\}, \max_{s \in ST} \{D(g(f, x'), g(f, s))\}) \quad (1)$$

where $D(g(f, x'), g(f, n)) = 0$ if $N = \emptyset$ and $D(g(f, x'), g(f, s)) = 0$ if $ST = \emptyset$.

TABLE 2. Notations and their description.

Symbol/Functions	Description	Symbol/Functions	Description
T	time steps	M	variables (modalities)
d	size of a vector ($= T \times M$)	r	a user-defined radius
X	set of windows	G	set of explainable models
$x \in X$	a matrix of T by M (a window of multivariate time series)	$x' \in \mathbb{R}^d$	a vectorized representation of x, also called to-be-explained series
$s \in \mathbb{R}^d$	short-term perturbed series	$l \in \mathbb{R}^d$	long-term perturbed series
n	training neighbor of x' within the radius r	$f(.) = y$	a black-box (e.g. CN-Waterfall), taking any series and predicting an affective state y
dn	total number of input-output pairs of training data	$\mathcal{D}_x \left\{(x^{(i)}, y^{(i)})\right\}_{i=1}^{dn}$	set of input-output pairs as training data
ln	total number of long-term perturbed series	$\mathcal{D}_l \left\{(l^{(i)}, y^{(i)})\right\}_{i=1}^{ln}$	set of input-output pairs as long-term perturbed series.
sn	total number of short-term perturbed series	$\mathcal{D}_s \left\{(s^{(i)}, y^{(i)})\right\}_{i=1}^{sn}$	set of input-output pairs as short-term perturbed series
$g(f, .) \in \mathbb{R}^d, g \in G$	an explainable model, taking the black-box f and any series, and returning the importance score of each feature in the series	$\mu(f, g, r, x')$	a function returning a scalar value for given inputs of the black-box f, explainer g, radius r and to-be-explained series x'
$\rho : \mathbb{R}^d \times \mathbb{R}^d$	a closeness metric of series. Here, we use the Euclidean distance	$D : \mathbb{R}^d \times \mathbb{R}^d$	a similarity metric of explanations (Here, we use the Euclidean distance).

Definition 2 (Max Long-Term Sensitivity (MLS)): Given $LT = \left\{l \in \mathcal{D}_l, \mid \rho(x', l) <= r;\ f(x') = f(l)\right\}$ as a set of long-term perturbed sequences, $N = \left\{n \in \mathcal{D}_x, \mid \rho(x', n) <= r;\ f(x') = f(n)\right\}$ as a set of training neighbors, the black-box f, the explainer g, the distance metric over explanations D and the to-be-explained series x', we define *MLS* of g at x' as:

$$\mu_{MLS}(f, g, r, x') = \max(\max_{n \in N} \left\{D(g(f, x'), g(f, n))\right\}, \max_{l \in LT} \left\{D(g(f, x'), g(f, l))\right\}) \quad (2)$$

where $D(g(f, x'), g(f, n)) = 0$ if $N = \emptyset$ and $D(g(f, x'), g(f, l)) = 0$ if $LT = \emptyset$.

Definition 3 (Average Short-Term Sensitivity (ASS)): Given the black-box f, the explainer g, the distance metric over explanations D, the radius r around x', the short-term generated data s and the to-be-explained series x', we define *ASS* of g at x' as the average of centers μ_{A^n} and μ_{A^s}:

$$\mu_{A^n}(f, g, r, x') = \frac{1}{|N|} \sum_{n \in N} D(g(f, x'), g(f, n)), \quad (3)$$

$$\mu_{A^s}(f, g, r, x') = \frac{1}{|ST|} \sum_{s \in ST} D(g(f, x'), g(f, s)), \quad (4)$$

$$\mu_{ASS}(f, g, r, x') = \begin{cases} \mu_{A^s}(f, g, r, x'), & \text{if } N = \emptyset \\ \mu_{A^n}(f, g, r, x'), & \text{if } ST = \emptyset \\ \frac{1}{2}(\mu_{A^n}(f, g, r, x') + \mu_{A^s}(f, g, r, x')) & \text{otherwise} \end{cases} \quad (5)$$

Definition 4 (Average Long-Term Sensitivity (ALS)): Given the black-box f, the explainer g, the distance metric over explanations D, the radius r around x', long-term generated data l and the to-be-explained series x, we define *ALS* of g at x' as the average of centers μ_{A^n} (Equation 3) and μ_{A^l}:

$$\mu_{A^l}(f, g, r, x') = \frac{1}{|LT|} \sum_{l \in LT} D(g(f, x'), g(f, l)), \quad (6)$$

$$\mu_{ALS}(f, g, r, x') = \begin{cases} \mu_{A^l}(f, g, r, x'), & \text{if } N = \emptyset \\ \mu_{A^n}(f, g, r, x'), & \text{if } LT = \emptyset \\ \frac{1}{2}(\mu_{A^n}(f, g, r, x') + \mu_{A^l}(f, g, r, x')) & \text{otherwise} \end{cases} \quad (7)$$

B. WORKFLOW

Algorithm 1 represents the process of calculating the values of the evaluation metrics. In an iterative process, the preprocessed data (see Section III) are split into training and test sets with a ratio of 80-20. The training set is then fed into CN-Waterfall, and the best fitted model is extracted as the *Black-Box*. At each iteration, a number of windows (X) are randomly selected from the test set. As discussed earlier (Section IV), we change the representation of each window ($X[j] = x$) to a vector (x'), providing a unified representation for all the XAI models. Sets of short- and long-term perturbations (ST_j and LT_j) as well as training neighbors (N_j) of each vector are then generated and extracted, respectively. We provide the explanations (ex) of all data by each XAI model and evaluate the explanation sensitivities of each vector by the four metrics in Section IV. Finally, we take the average over the sensitivities of all vectors for each metric. To further clarify, Figure 3 illustrates the procedure.

Algorithm 1 Extract Explanation Sensitivities

Inputs: *PreporocessedDataset*
Output: μ_{MSS}, μ_{MLS}, μ_{ASS} and μ_{ALS}
Begin:

for $i \in Iterations$ **do**
  $X_{train_i}, X_{test_i} \leftarrow Split(PreprocessedDataset)$
  $Black\_Box_i \leftarrow CN - Waterfall(X_{train_i})$
  $X_i \leftarrow RandomSelection(X_{test_i})$
  for $j \in size(X_i)$ **do**
    $x' \leftarrow ChangeRepresentation(X_i[j])$
    $ST_j \leftarrow GenerateShortTermPurturbations(x')$
    $LT_j \leftarrow GenerateLongTermPurturbations(x')$
    $N_j \leftarrow ExtractTrainingNeighbors(x', X_{train_i})$
    $ex_{x'}, ex_{ST_j}, ex_{LT_j}, ex_{N_j} \leftarrow XAI\_Model(x', ST_j, LT_j, N_j, Black\_Box_i)$
    $\mu_{MSS_j} \leftarrow MSS(ex_{x'}, ex_{ST_j}, ex_{N_j})$ [*Eq.* 1]
    $\mu_{MLS_j} \leftarrow MLS(ex_{x'}, ex_{LT_j}, ex_{N_j})$ [*Eq.* 2]
    $\mu_{ASS_j} \leftarrow ASS(ex_{x'}, ex_{ST_j}, ex_{N_j})$ [*Eq.* 5]
    $\mu_{ALS_j} \leftarrow ALS(ex_{x'}, ex_{LT_j}, ex_{N_j})$ [*Eq.* 7]
  end for
  $\mu_{MSS_i} \leftarrow Average(\mu_{MSS_0}, .., \mu_{MSS_{size(X_i)}})$
  $\mu_{MLS_i} \leftarrow Average(\mu_{MLS_0}, .., \mu_{MLS_{size(X_i)}})$
  $\mu_{ASS_i} \leftarrow Average(\mu_{ASS_0}, .., \mu_{ASS_{size(X_i)}})$
  $\mu_{ALS_i} \leftarrow Average(\mu_{ALS_0}, .., \mu_{ALS_{size(X_i)}})$

end for

V. EXPERIMENTS AND RESULTS

In this section, we discuss the results of our evaluation metrics examined on the XAI models. We focus on both gradient-based (IG and SG) and perturbation-based (LIME) approaches.

According to our standard settings, each XAI model runs over 10 iterations. At each iteration, 50 windows of test data are selected and represented as vectors for explanation (see Algorithm 1). We also generate 20 temporal-based perturbed series, and extract 20 training neighbors in the radius of $r = 1$ per vector. Following these settings, for IG,[1] we calculate the average over the training data and consider it as the IG baseline (reference). We also set $no\_steps = 10$, referring to the number of steps in which the gradients of series are computed along a straight line path from the baseline. For SG,[2] 20 noisy samples are generated by a Gaussian noise kernel with a mean of 0 and standard deviation (STD) of 1.0. In case of LIME,[3] we restrict ourselves to 50 samples, by which a linear model is approximated. The restriction is mainly due to the time complexity issue, which will be discussed in Section VI. More precisely, we sample 50 vectors in the neighborhood of the vectorized representation of each series. Similar to SG, the applied sampling kernel in LIME also relies on a Gaussian distribution with a mean of 0 and standard deviation of 1.0.

[1] https://github.com/hiranumn/IntegratedGradients
[2] https://github.com/sicara/tf-explain
[3] https://github.com/marcotcr/lime

In the following, we present our evaluations on the three-fold setting of *data*-, *metric*- and *XAI hyperparameters*- in which the results are averaged over the 50 selected data in each iteration. Figure 4 shows, as an example, a window of ACC0 series in WESAD as well as one of its short-term perturbed series. Figure 5 also shows explanations of each XAI model over ACC0 and its perturbation. The explanations are provided with respect to the importance score at each time step. As observed, the scores of the original ACC0 and the perturbed version vary in each XAI model, implying the sensitivity of explanation in ACC0. We also infer that the scale of importance scores in IG and LIME are much lower than in SG (nearly close to 0). Such scaling could result in lower sensitivities in the former models than in the latter.

A. DATA

In this section, we investigate the effect of two parameters, the window size and the overlapping stride, on the sensitivity of XAI models. In addition to our standard setting discussed earlier, we partition both datasets into windows with the two following settings: the window size of 30 time steps together with 20 overlapping strides, as well as the window size of 60 time steps together with 20 overlapping strides. For ease of use, we indicate our settings as 30_10 (standard setting), 30_20 and 60_20. The reported results of all the settings are the average of the outputs over 10 iterations.

Figure 6 shows the results on WESAD. As we observe, in all the XAI models, the standard setting of 30_10 achieves

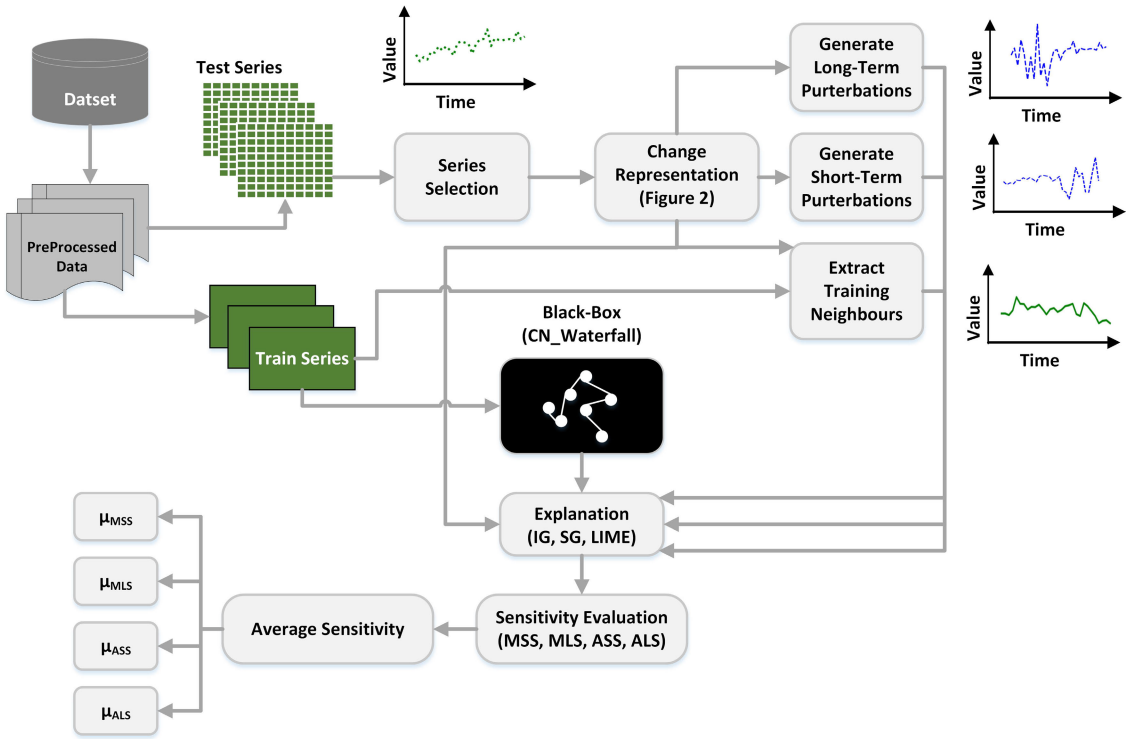


FIGURE 3. The workflow of measuring the explanation sensitivities in each iteration. First, the data are preprocessed (see Section III) and split into the training and test sets. Our black-box, CN-Waterfall, is learnt from the training set, and the best model is extracted. From the test set, 50 windows of series are randomly selected and transformed into the vectorized representation. For each vector, 20 temporal-based perturbations and 20 training neighbors are generated and extracted, respectively (see Section IV). We, then, provide the explanations of all the data by the IG, SG and LIME models. Finally, the explanation sensitivities of the vectorized windows of series are calculated and averaged over 50 in each iteration.

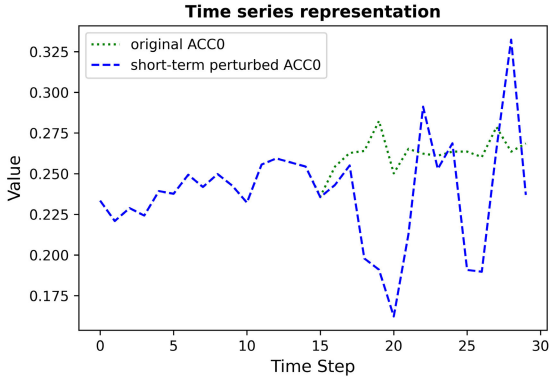


FIGURE 4. The representation of the original and short-term perturbed ACC0. The first 15 time steps of the two series are identical, while the second 15 time steps of the original ACC0 are manipulated. The values of the latter steps are randomly perturbed with a normal distribution in a radius $r = 1$.

a higher sensitivity in all metrics compared to 60_20. In other words, a larger window size (60_20) shows less maximum and average sensitivity than a smaller window size (30_10) over the IG, SG and LIME models. Moreover, a higher sensitivity of 30_10 with respect to 30_20 is observed for all the XAI models and metrics. Regardless of the window size and the overlap stride values, we find similarities between the results of the sensitivity metrics in both short- and long-term settings. The latter argument motivates further effort in future work to examine a flexible range of temporal perturbations rather than a fixed equal size. In addition, we infer a lower scale of sensitivities in IG and LIME than SG, which could potentially refer to the lower scale of generated scores by the former models (see Figure 5). In the following, we further analyze each model individually.

As shown in Figure 6(a), in IG, sensitivity differences of approximately 0.14 and 0.13 are seen between the 60_20 and 30_10 settings in *MSS* and *MLS*, respectively, while a lower difference of approximately 0.07 is observed between the same metrics for the 30_20 and 30_10 settings. In the case of *ASS* and *ALS*, all the settings show tightly close sensitivities to each other.

We find that SG assigns high scale importance scores to the features. Correspondingly, high scale sensitivities of the metrics are observed in Figure 6(b). This figure shows a difference around 1.5 and 3.0 for the maximum sensitivity metrics between the 30_20 and 30_10, and 60_20 and 30_10 settings, respectively. We also see quite similar results for 30_20 and 60_20 in terms of *ASS* and *ALS* metrics, with a difference around 1.5 between the latter settings and 30_10 in the aforementioned metrics.

In LIME (Figure 6(c)), we observe a decrease around 0.1 between pairwise metrics in 60_20 and standard settings. However, a lower decrease, around 0.04, is seen between these metrics in 30_20 and standard settings.

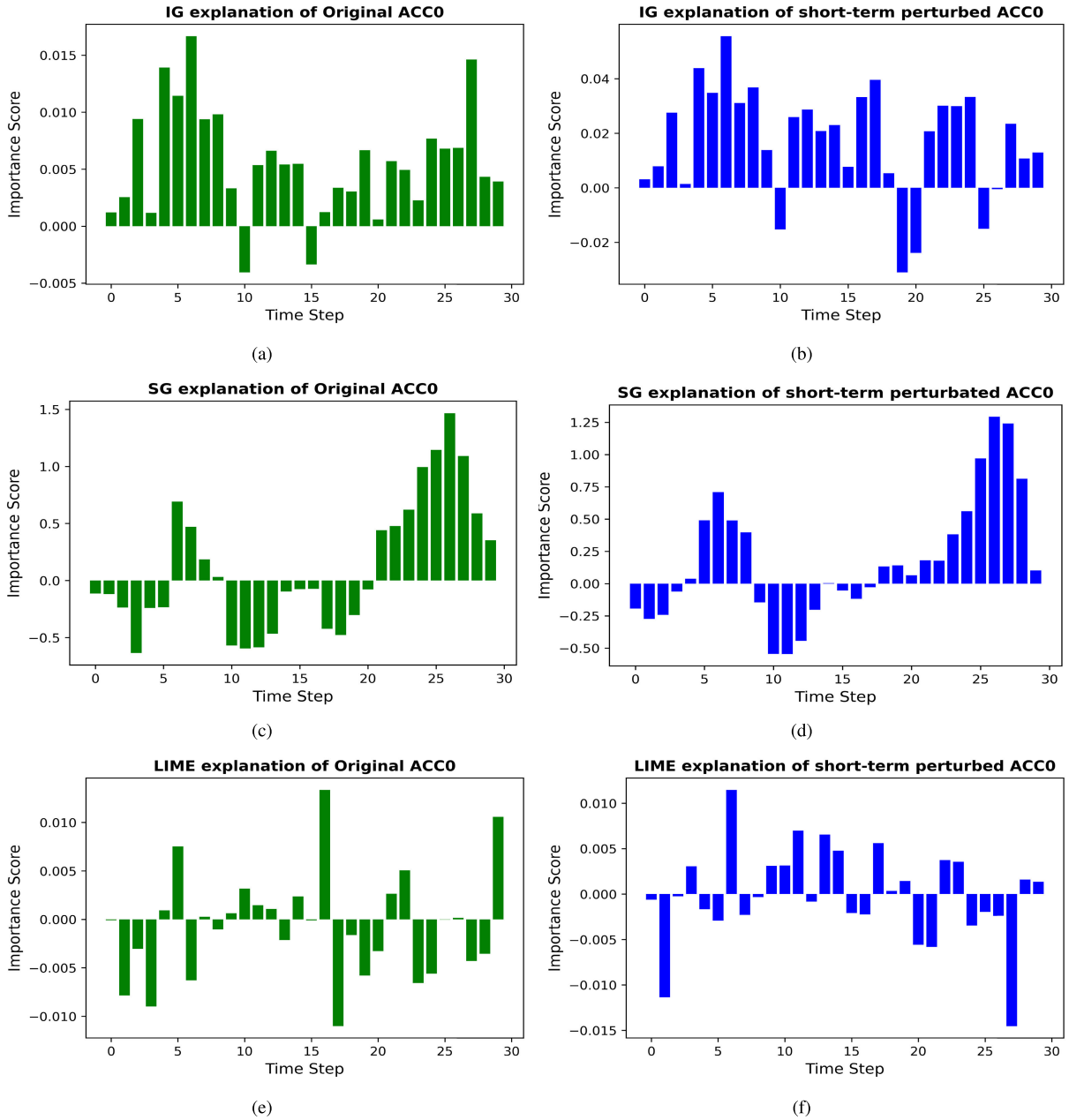


FIGURE 5. The explanations of the original and short-term perturbed ACC0 in terms of importance scores. Each row corresponds to the results of one XAI model as follows: IG in the first row, (a) and (b), SG in the second row, (c) and (d), and LIME in the last row, (e) and (f). The importance score of each time step in the original (the left column) and perturbed ACC0 (the right column) are indicated by "green" and "blue" colors, respectively. It could be inferred that the scores of the original ACC0 and its perturbed version are not identical, implying the sensitivity of explanation in ACC0. In addition, lower scales of scores in IG and LIME than SG could lead to lower sensitivities in the former models.

Regarding MAHNOB-HCI as a balanced dataset, Figure 9 shows that in IG, there are slight sensitivity differences between the 60_20 and 30_10 settings, while there are not considerable differences between 30_20 and 30_10 over all metrics. In SG, a larger window/overlap size (60_20 and 30_20) provides less sensitivity than a smaller window/overlap size (30_10). In the case of LIME, one may argue that only the impact of a larger window size (60_20) decreases the sensitivities in all metrics. Similar to the WESAD results, the sensitivities of short- and long-term related metrics are fairly the same in the IG and LIME models. However, in SG, *ALS* provides lower sensitivities than *ASS* in all settings. We also find a lower scale of sensitivities in IG and LIME than SG.

In detail, the maximum and average sensitivity metrics between the 60_20 and 30_10 settings in IG show differences around 0.04 and 0.02, respectively (Figure 7(a)). However, such differences are less than 0.01 between the metrics of the 30_20 and 30_10 settings in most cases.

In SG (Figure 7(b)), the differences are approximately 0.4 between the highest sensitivity setting (30_10) and 30_20 in *MSS* and *ASS*; however, there are approximately 0.1 lower differences in the *MLS* and *ALS* metrics. In the case of a larger window size (60_20) in SG, there is a higher variation in sensitivities between 60_20 and 30_10 than between 30_20 and 30_10. A difference of 1.2 in *MSS*, *MLS* and *ASS*, and 0.8 in *ALS* are seen between 60_20 and 30_10.

In LIME (Figure 7(c)), we observe variations around 0.1 between 30_20 and the standard setting in *MSS*, *MLS* metrics, and around 0.05 in *ASS* and *ALS*. This is while the sensitivity variations between 60_20 and the standard setting are approximately 0.1 for all metrics.

B. METRIC

In this section, we investigate the impact of different standard deviations for generating the temporal-based perturbations on the sensitivities of the XAI models. In particular, we examine the standard deviations of 0.001, 0.01, 0.05 (the standard setting) and 0.1 in all the experiments. Since these settings are designed based on the definition of metrics, we name the experiments in this section metric-related experiments.

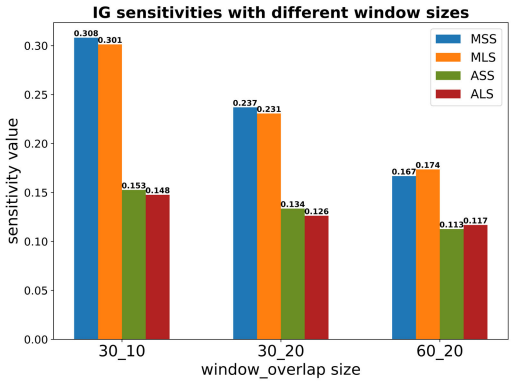
Figures 8 (a), (b) and (c) show the experimental results of IG, SG and LIME, on the WESAD dataset, respectively. Due to the similarities found in the results of the four metrics, we only show our analysis on *MSS*. The same arguments are applied for the MAHNOB-HCI dataset, and the results of the *MSS* metric are shown in Figure 9.

The results shown in Figures 8 and 9 indicate that none of the XAI models achieve a steady sensitivity trend within the iterations. In other words, the XAI models provide different sensitivity results in each epoch. Since the black-box parameters and the to-be-explained series differ in each epoch, one may argue that the provided explanations are independent of the black-box dynamism and test data properties.

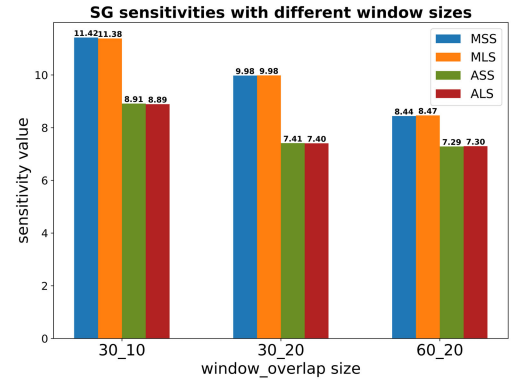
Regarding the WESAD dataset, we see that higher standard deviations worsen the sensitivity of the IG model (Figure 8(a)). Additionally, with STD $= 0.001$, we observe the same pattern as with STD $= 0.01$. In the case of SG (Figure 8(b)) and LIME (Figure 8(c)), the models are found to be insensitive to different STDs but highly fluctuate within iterations.

Regarding the MAHNOB-HCI dataset, as seen in Figure 9, we can assert the same evaluations as WESAD. However, in MAHNOB-HCI, we observe lower scales of sensitivities for IG and SG in all settings. Moreover, all the STDs in IG follow quite similar patterns of sensitivities.

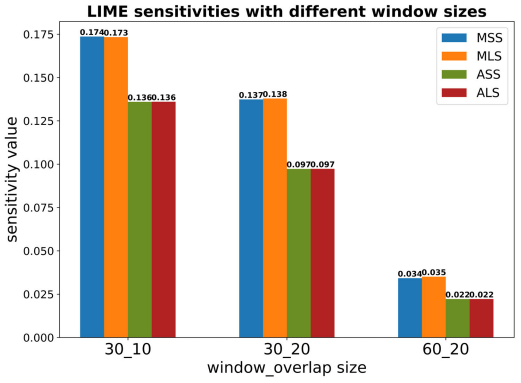
Comparing the XAI models over all STDs, we first normalize the previously achieved results and then take an average over the sensitivities in each iteration. Due to the similarities found in the results of the four metrics, we only show the outputs of the *MSS* metric on each dataset (Figure 10). Given the results, we can infer that IG and LIME provide much lower *MSS* values than SG for both datasets. More specifically, the former models provide an average value below 1.0 on both datasets, while in the case of SG, the average is approximately



(a)



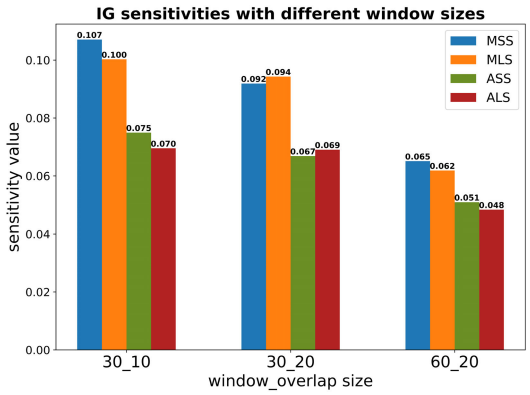
(b)



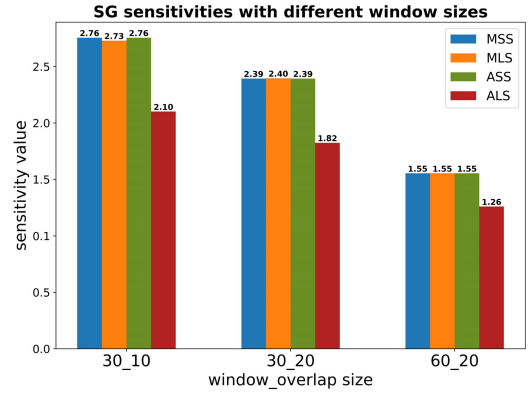
(c)

FIGURE 6. Comparison of *MSS*, *MLS*, *ASS* and *ALS* sensitivities with the time series windows and overlapping of 30_10, 30_20 and 60_20 for (a) IG, (b) SG and (c) LIME on WESAD. In all the XAI models, lower sensitivities are observed in the settings with larger window and overlap sizes (60_20, 30_20) than in the standard setting (30_10).

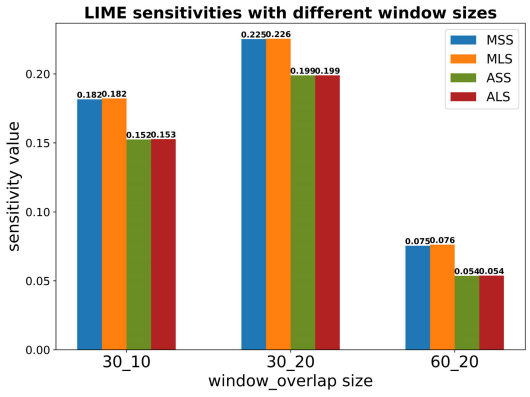
12.5 and 3.0 on WESAD and MAHNOB-HCI, respectively. The results could be justified as the scale of generated explanations by IG and LIME are lower than their SG counterparts (see Figure 5). On WESAD (Figure 10(a)), we also see a constant behavior of IG and LIME within 10 epochs. However, on MAHNOB-HCI (Figure 10(b)), some fluctuations are observed in LIME.



(a)

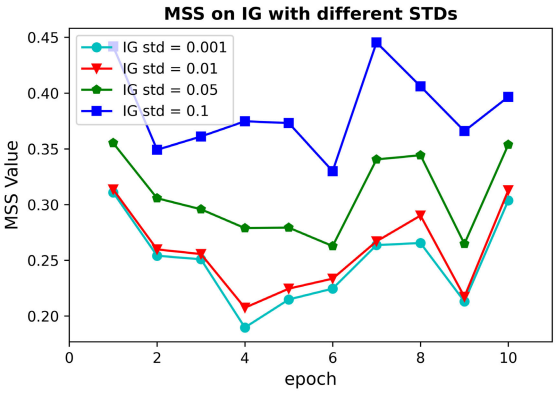


(b)

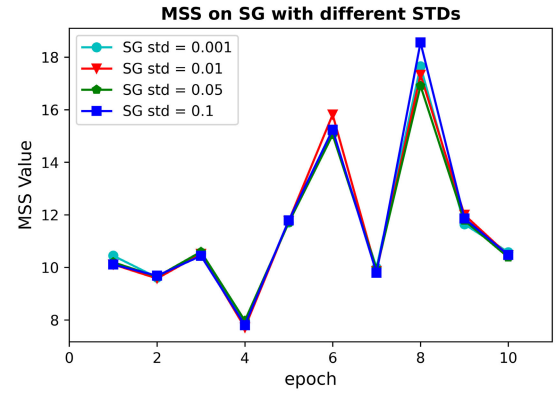


(c)

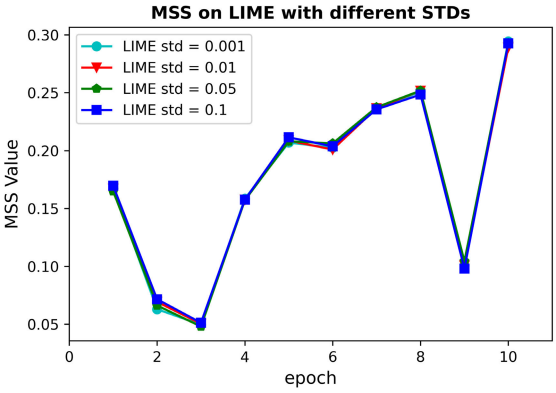
FIGURE 7. Comparison of *MSS*, *MLS*, *ASS* and *ALS* sensitivities with the time series windows and overlapping of 30_10, 30_20 and 60_20 for (a) IG, (b) SG and (c) LIME on MAHNOB-HCI. Fairly, in all the XAI models, the setting with a larger window size (60_20) than the standard setting (30_10) provides lower sensitivities, while this is not always the case for settings (e.g., 30_20) with a larger overlap size (e.g., see the results in LIME (c)).



(a)



(b)



(c)

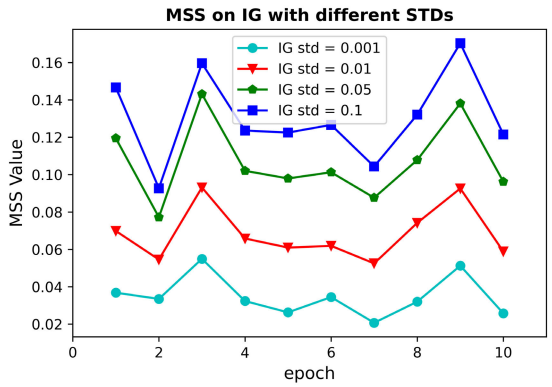
FIGURE 8. Comparison of *MSS* with different STDs of perturbation for IG (a), (b) SG and (c) LIME on WESAD. SG and LIME show similar sensitivity under different settings, while the sensitivity of IG varies by the scale of perturbations. In all the XAI models, different sensitivities are observed within the epochs, implying the independence of these models from the black-box dynamism and data properties.

C. XAI HYPERPARAMETER

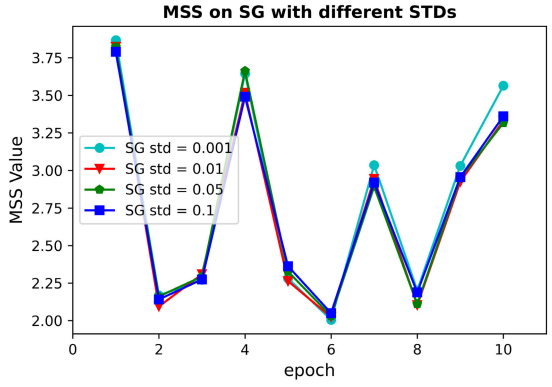
In this section, we explore the impact of the hyperparameters of each XAI model on its sensitivity. To carefully design the experiments, we assume the data- and metric-related settings are unchanged. The reported values are the average of the achieved results over 10 iterations.

1) INTEGRATED GRADIENT

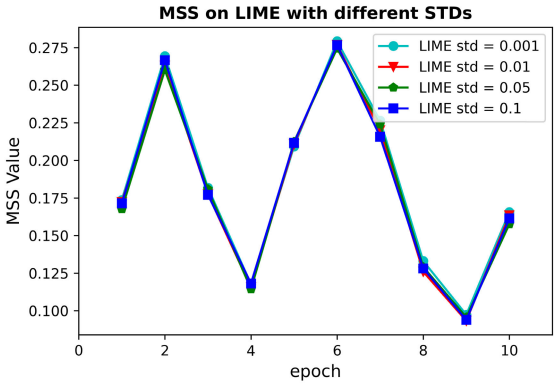
As argued in [9], IG aggregates gradients of all samples along a straight path from an input to a baseline. The focus of



(a)



(b)

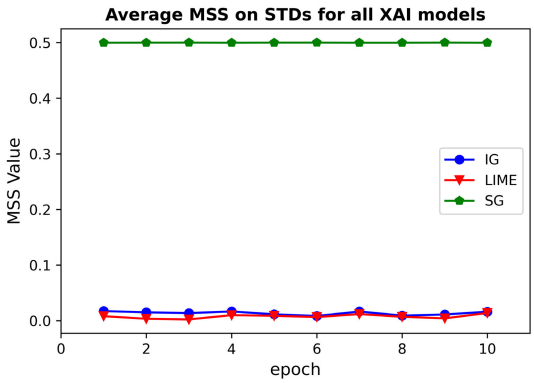


(c)

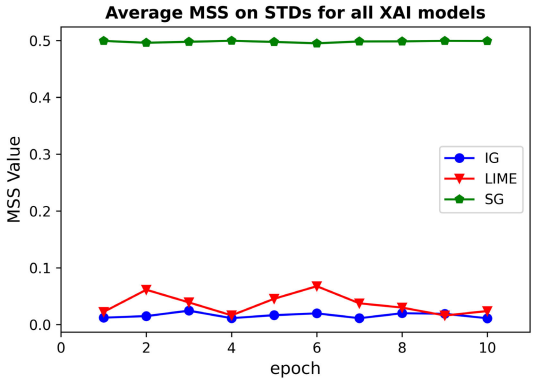
FIGURE 9. Comparison of *MSS* with different STDs of perturbation for IG (a), (b) SG and (c) LIME on MAHNOB-HCI. Similar to Figure 8, the sensitivity of IG varies under different STDs, while SG and LIME do not. In addition, the XAI models seem independent of the black-box dynamism and data properties, as their sensitivity follows an unsteady trend within epochs.

the experiments in this section is on the number of steps (*no_steps*) in which the gradients are aggregated. More precisely, we explore the impacts of the 5, 10 (the standard setting), 20 and 40 steps on the sensitivities of IG on both datasets.

According to the results shown in Table 3, the 5 and 10 number of steps implies higher sensitivities than the 20 and



(a)



(b)

FIGURE 10. Comparing *MSS* of XAI models for all the STD settings on WESAD (a) and MAHNOB-HCI (b). IG and LIME provide much lower sensitivity than SG for both datasets. Such a result could be due to the lower scale of important scores in the former models (see Figure 5).

40 steps on WESAD. When the aggregations are saturated by 20 gradients, the sensitivities remain constant in all the metrics. We also observe lower values for *ASS* and *ALS* than *MSS* and *MLS* in all settings.

With respect to MAHNOB-HCI, it could be inferred that the 5 steps of gradient provide the lowest sensitivity value in most of the metrics. However, there is no considerable difference in sensitivities between the latter and the two other settings of 20 and 40. We also observe that the values of *ASS* and *ALS* are rather close to *MSS* and *MLS*, respectively, implying a dense distribution of sensitivities.

Likewise, the results discussed in Sections V-A and V-B show fairly similar results between the *MSS* and *MLS* values in all settings. The same argument also applies between *ASS* and *ALS*.

2) SMOOTHGRAD

In the following, we discuss the impact of different noise levels on the sensitivities of SG. As mentioned before, in our standard setting, we generate noisy samples using a Gaussian kernel with a mean of 0 and an STD of 1.0. We further extend this setting by examining the STDs of 0.5 and 2.0 to generate noise.

TABLE 3. IG sensitivities with different number of steps for each dataset.

Dataset	Metrics	no_steps			
		5	10	20	40
WESAD	(μ_{MSS}, μ_{MLS})	(0.36, 0.36)	(0.31, 0.30)	(0.28, 0.27)	(0.28, 0.27)
	(μ_{ASS}, μ_{ALS})	(0.17, 0.17)	(0.15, 0.15)	(0.14, 0.14)	(0.14, 0.14)
MAHNOB-HCI	(μ_{MSS}, μ_{MLS})	(0.08, 0.07)	(0.11, 0.10)	(0.09, 0.08)	(0.08, 0.08)
	(μ_{ASS}, μ_{ALS})	(0.06, 0.05)	(0.08, 0.07)	(0.07, 0.06)	(0.06, 0.06)

TABLE 4. SG sensitivities with different noise STDs for each dataset.

Dataset	Metrics	Noise STD		
		0.5	1.0	2.0
WESAD	(μ_{MSS}, μ_{MLS})	(10.97, 10.91)	(11.42, 11.38)	(11.91, 12.03)
	(μ_{ASS}, μ_{ALS})	(8.41, 8.40)	(8.91, 8.89)	(9.41, 9.41)
MAHNOB-HCI	(μ_{MSS}, μ_{MLS})	(2.41, 2.44)	(2.76, 2.73)	(2.89, 2.89)
	(μ_{ASS}, μ_{ALS})	(1.75, 1.75)	(2.10, 2.10)	(2.23, 2.23)

As shown in Table 4, on WESAD, the STD of 0.5 results in lower sensitivities than the standard setting in all the metrics. In contrast, the STD of 2.0 results in higher sensitivities. On MAHNOB-HCI, the results related to the STDs of 1.0 and 2.0 are closer to each other in all the metrics. We also see similar results for *MSS* and *MLS* in all the settings. This argument also applies in the case of the *ASS* and *ALS* metrics. In other words, incorporating different standard deviations of sampling noise does not cause a remarkable change between the results of short- and long-term sensitivity metrics.

3) LIME

As discussed in [13], LIME generates several samples in the neighbourhood of to-be-explained instance by a Gaussian kernel. Later, LIME approximates a linear model to provide an explanation. In this section, we vary over the standard deviation of this kernel and investigate how this variation impacts the explanation sensitivities of LIME. To this end, we choose STDs of 0.5 and 2.0 in addition to the standard setting (1.0).

Table 5 shows that for both datasets, the highest neighborhood STD (2.0) achieves better results than the lowest STD (0.5). With respect to $STD = 1.0$ and $STD = 0.5$, we observe better results in the former setting than in the latter, with differences of approximately 0.1 for all the metrics on both datasets. We also report such a difference between the STDs of 2.0 and 1.0. Comparing *MSS-MLS* and *ASS-ALS*, we find similar sensitivities in each pair for all the settings of both datasets. However, in a comparison between the maximum and average sensitivity metrics, one could see that the average sensitivities are lower than the maximum sensitivities on both datasets.

VI. IMPLEMENTATION CHALLENGES AND TIME COMPLEXITY

Each XAI model follows a specific reasoning to explain the input of interest. The XAI models examined in this work

TABLE 5. LIME sensitivities of different neighbourhood STDs for each dataset.

Dataset	metrics	Neighborhood STD		
		0.5	1.0	2.0
WESAD	(μ_{MSS}, μ_{MLS})	(0.30, 0.30)	(0.17, 0.17)	(0.05, 0.05)
	(μ_{ASS}, μ_{ALS})	(0.22, 0.22)	(0.14, 0.14)	(0.03, 0.03)
MAHNOB-HCI	(μ_{MSS}, μ_{MLS})	(0.31, 0.31)	(0.18, 0.18)	(0.04, 0.04)
	(μ_{ASS}, μ_{ALS})	(0.25, 0.25)	(0.15, 0.15)	(0.03, 0.03)

TABLE 6. Running-time complexity of all XAI models in the scale of second on each dataset.

Dataset	System Specification	Time Complexity (seconds)		
		IG	SG	LIME
WESAD	Core i5-7600T, 2.81 GHz 32.0 GB RAM	7666	1884	11129
MAHNOB-HCI	Core i5-7600T, 2.81 GHz 32.0 GB RAM	4949	1061	3355

are initially proposed for contexts with non-time series data. In this paper, we devoted extra efforts to making these models compatible with time series data. Moreover, the XAI models are usually employed for output explanations of traditional black-box models, e.g., support vector machines (SVM) [13], [14], as well as popular deep learning models, e.g., inception architectures [8], [9], [13]. However, the community lacks the practice of output explanation for deep learning models such as CN-Waterfall with specific structure. Such practice could entail integration challenges rather than paving a straight way to apply the XAI models. In our case, since CN-Waterfall is fed by parallel inputs, we implemented a module that maps the preprocessed data to a parallel representation (see Figure 2) to tackle the integration challenge.

We also investigated the running-time complexity of the XAI models applied on the standard setting for both datasets. As shown in Table 6, we performed all the experiments on a machine with an Intel(R) Corei5-7600T CPU, 2.81 GHz clock speed and 32 GB RAM. We noticed that LIME is computationally more expensive than IG on WESAD but less expensive on MAHNOB-HCI. Overall, the SG model is the most affordable XAI model on both datasets.

VII. CONCLUSION AND FUTURE WORKS

This paper formulated four different metrics, namely, *MSS*, *MLS*, *ASS* and *ALS*, to evaluate the sensitivities of XAI models considering temporal-based perturbations and training neighbors around the series of interest. Our hypothesis was that we would obtain similar explanations for close series with the same class labels, and thereby the XAI models would result in low sensitivities. We focused on the sensitivity evaluations of three attribution-based XAI models, namely IG, SG and LIME. These models were applied to explain the decision of CN-Waterfall [26], a highly accurate convolutional deep learning model specialized for affect computing. The experiments were conducted on a three-fold setting of

data, *metric* and *XAI hyperparameter* on the WESAD and MAHNOB-HCI datasets. We also discussed the applicability and running-time complexity of each XAI model with respect tp the sensitivity evaluations.

In summary, we found that (i) IG and LIME provide lower scales of sensitivity than SG in all the metrics and settings. We referred the result to the lower scale of important scores generated by the former models; (ii) the window size of the series plays a role in the sensitivities' variation in the XAI models. In our experiments, higher sensitivities were associated with a smaller window size; (iii) ignoring network parameters and data properties in design, the XAI models fluctuate in terms of sensitivities when the parameters and properties change; (iv) the sensitivities of XAI models vary with respect to different settings of hyperparameters.

There are several shortcomings in this research that could be further investigated in the future. *First*, in this study, the impact of a limited number of window and overlapping sizes were provided. We encourage practitioners to explore the impact of broader data settings. *Second*, we examined *equal ranges* of short- and long-term perturbations. In most of the cases, we observed similar outputs for *MSS-MLS*, and also for *ASS-ALS*. It could be interesting to focus on unequal/dynamic ranges of temporal-based perturbations and explore how the sensitivities of XAI models change under such settings. *Third*, although the evaluated XAI models are among the most prominent models in the XAI field, investigating the sensitivities of more elaborated models is recommended. Specifically, it is worth to examine the models with low running-time complexities and modular implementations. *Last*, in this paper, we focused on the time series benchmarks in affect computing. Understanding how the proposed metrics work in other domains (e.g., human activity recognition) could further acknowledge the scalability of these metrics in practice. In theory, the proposed metrics are assumed scalable to other domains of interest as there is no constraint on the context/semantic of to-be-explained series in the process of metrics' design.

REFERENCES

[1] D. H. Park, L. A. Hendricks, Z. Akata, A. Rohrbach, B. Schiele, T. Darrell, and M. Rohrbach, "Multimodal explanations: Justifying decisions and pointing to the evidence," 2018, *arXiv:1802.08129*.

[2] C. Panigutti, A. Perotti, and D. Pedreschi, "Doctor XAI: An ontology-based approach to black-box sequential data classification explanations," in *Proc. Conf. Fairness, Accountability, Transparency*, Jan. 2020, pp. 629–639.

[3] N. Safinianaini and H. Boström, "Towards interpretability of mixtures of hidden Markov models," in *Proc. AAAI Workshop Explainable Agency Artif. Intell.*, 2021, pp. 1–10.

[4] A. H. Gee, D. Garcia-Olano, J. Ghosh, and D. Paydarfar, "Explaining deep classification of time-series data with learned prototypes," 2019, *arXiv:1904.08935*.

[5] N. Fouladgar, M. Alirezaie, and K. Främling, "Exploring contextual importance and utility in explaining affect detection," in *Advances in Artificial Intelligence*, B. S. Baldoni M., Ed., vol. 12414. Cham, Switzerland: Springer, 2021, pp. 3–18.

[6] R. Assaf, I. Giurgiu, F. Bagehorn, and A. Schumann, "MTEX-CNN: Multivariate time series explanations for predictions with convolutional neural networks," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2019, pp. 952–957.

[7] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," 2013, *arXiv:1312.6034*.

[8] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "SmoothGrad: Removing noise by adding noise," 2017, *arXiv:1706.03825*.

[9] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, 2017, pp. 3319–3328.

[10] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, vol. 8689. Cham, Switzerland: Springer, 2014.

[11] T. Rojat, R. Puget, D. Filliat, J. Del Ser, R. Gelin, and N. Díaz-Rodríguez, "Explainable artificial intelligence (XAI) on timeseries data: A survey," 2021, *arXiv:2104.00950*.

[12] A. Abdelsalam Ismail, M. Gunady, H. Corrada Bravo, and S. Feizi, "Benchmarking deep learning interpretability in time series predictions," 2020, *arXiv:2010.13924*.

[13] M. Ribeiro, S. Singh, and C. Guestrin, "'Why should i trust you?' Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD, Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1135–1144.

[14] R. Guidotti, A. Monreale, S. Ruggieri, D. Pedreschi, F. Turini, and F. Giannotti, "Local rule-based explanations of black box decision systems," 2018, *arXiv:1805.10820*.

[15] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," 2017, *arXiv:1702.08608*.

[16] C.-K. Yeh, C.-Y. Hsieh, A. S. Suggala, D. I. Inouye, and P. Ravikumar, "On the (in)fidelity and sensitivity of explanations," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 32, 2019, pp. 1–12.

[17] P. Hase and M. Bansal, "Evaluating explainable AI: Which algorithmic explanations help users predict model behavior?" in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 5540–5552.

[18] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Metrics for explainable AI: Challenges and prospects," 2018, *arXiv:1812.04608*.

[19] U. Bhatt, A. Weller, and J. M. F. Moura, "Evaluating and aggregating feature-based model explanations," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, 2020, pp. 3016–3022.

[20] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31. Red Hook, NY, USA: Curran Associates, 2018.

[21] A. Ghorbani, A. Abid, and J. Zou, "Interpretation of neural networks is fragile," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 3681–3688.

[22] V. Arya, R. K. E. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. Vera Liao, R. Luss, A. Mojsilović, S. Mourad, P. Pedemonte, R. Raghavendra, J. Richards, P. Sattigeri, K. Shanmugam, M. Singh, K. R. Varshney, D. Wei, and Y. Zhang, "One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques," 2019, *arXiv:1909.03012*.

[23] A.-p. Nguyen and M. R. Martínez, "On quantitative aspects of model interpretability," 2020, *arXiv:2007.07584*.

[24] M. Singh, N. Kumari, P. Mangla, A. Sinha, V. Balasubramanian, and B. Krishnamurthy, "On the benefits of attributional robustness," 2019, *arXiv:1911.13073*.

[25] U. Schlegel, H. Arnout, M. El-Assady, D. Oelke, and D. A. Keim, "Towards a rigorous evaluation of xai methods on time series," in *Proc. CoRR*, 2019, pp. 4197–4201.

[26] N. Fouladgar, M. Alirezaie, and K. Främling, "CN-waterfall: A deep convolutional neural network for multimodal physiological affect detection," *Neural Comput. Appl.*, vol. 34, pp. 1–20, Sep. 2021.

[27] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, "Introducing wesad, a multimodal dataset for wearable stress and affect detection," in *Proc. 20th ACM Int. Conf. Multimodal Interact.*, 2018, pp. 400–408.

[28] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 42–55, Aug. 2012.

[29] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, "Towards better understanding of gradient-based attribution methods for deep neural networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018, pp. 1–16.

[30] R. Guidotti and S. Ruggieri, "On the stability of interpretable models," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2019, pp. 1–8.

[31] S. Siddiqui, D. Mercier, M. Munir, A. Dengel, and S. Ahmed, "Tsviz: Demystification of deep learning models for time-series analysis," *IEEE Access*, vol. 7, pp. 67027–67040, 2019.
[32] M. Munir, S. A. Siddiqui, F. Küsters, D. Mercier, A. Dengel, and S. Ahmed, "TSXplain: Demystification of dnn decisions for time-series using natural language and statistical features," in *Proc. Int. Conf. Artif. Neural Netw. (ICANN)*, 2019, pp. 426–439.
[33] S. Tonekaboni, S. Joshi, D. Duvenaud, and A. Goldenberg. (2020). *Explaining Time Series by Counterfactuals*. [Online]. Available: https://openreview.net/forum?id=HygDF1rYDB
[34] E. Delaney, D. Greene, and M. T. Keane, "Instance-based counterfactual explanations for time series classification," 2020, *arXiv:2009.13211*.
[35] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," in *Proc. Int. Conf. Manage. Data (SIGMOD)*, vol. 29, 2000, pp. 93–104.
[36] R. Guidotti, A. Monreale, F. Spinnato, D. Pedreschi, and F. Giannotti, "Explaining any time series classifier," in *Proc. IEEE 2nd Int. Conf. Cognit. Mach. Intell. (CogMI)*, Oct. 2020, pp. 167–176.
[37] N. Fouladgar, M. Alirezaie, and K. Främling, "Decision explanation: Applying contextual importance and contextual utility in affect detection," in *Proc. Italian Workshop Explainable Artif. Intell.*, vol. 2742, 2020, pp. 1–13.

NAZANIN FOULADGAR is currently pursuing the Ph.D. degree with the Department of Computing Science, Umeå University, Sweden.

She is a member of the Explainable AI (XAI) Research Group and supervised by Prof. Kary Främling, Assistant Prof. Marjan Alirezaie, and Prof. Frank Drewes. She is the author of several journals and conference papers. Her current research interests include explainable artificial intelligence, machine learning, and time series analysis. Specifically, she is interested in the development of deep learning models and their applications in health care and sensor-based systems. Previously, she was awarded as a Research Fellowship by Young Researchers and Elite Club, Esfahan, Iran, in 2016, and computational resources by the Swedish National Infrastructure for Computing (SNIC), in 2020.

MARJAN ALIREZAIE received the M.Sc. degree in computer science from Linköping University, Sweden, in 2010, and the Ph.D. degree from Örebro University, Sweden, in 2015, for a dissertation entitled "Bridging the Semantic Gap Between Sensor Data and Ontological Knowledge."

She is currently an Assistant Professor with the Center for Applied Autonomous Sensor Systems (AASS), Örebro University. From 2015 to 2020, she was a Postdoctoral Fellow and a Researcher at Örebro University, where she continued her research on semantic perception. Her research interests include knowledge representation and reasoning, machine learning, and the integration of both. In particular, she is interested in the development of neuro-symbolic (NeSy) models and the extension of their applications in robotics and sensory systems.

KARY FRÄMLING (Member, IEEE) is currently a Professor in data science with Umeå University and an Adjunct Professor in computer science with Aalto University. He is also the founder of several companies and the Former Chair of the The Open Group IoT Work Group. He is the author of over 150 papers published in scientific journals and conferences, including the (presumably) first article that mentions the IoT and describes an operational implementation, in 2002. Since then, he has been the main Architect of IoT systems in various domains, such as buildings, HVAC equipment, vehicles, supply chain management, and smart cities, under the umbrella concept of intelligent products. His current core interest is developing new methods for explainable artificial intelligence and the related machine learning technologies. His Contextual Importance and Utility Method published, in 1995, where he is presumably the first that addresses post-hoc explainability.