Toward the application of XAI methods in EEG-based systems

Andrea Apicella[1,2,3,*,†], Francesco isgrò[1,2,3,†], Andrea Pollastro[1,2,3,†] and Roberto Prevete[1,2,3,†]

[1]*Laboratory of Augmented Reality for Health Monitoring (ARHeMLab)*
[2]*Laboratory of Artificial Intelligence, Privacy & Applications (AIPA Lab)*
[3]*Department of Electrical Engineering and Information Technology, University of Naples Federico II*

Abstract

An interesting case of the well-known Dataset Shift Problem is the classification of Electroencephalogram (EEG) signals in the context of Brain-Computer Interface (BCI). The non-stationarity of EEG signals can lead to poor generalisation performance in BCI classification systems used in different sessions, also from the same subject. In this paper, we start from the hypothesis that the Dataset Shift problem can be alleviated by exploiting suitable eXplainable Artificial Intelligence (XAI) methods to locate and transform the relevant characteristics of the input for the goal of classification. In particular, we focus on an experimental analysis of explanations produced by several XAI methods on an ML system trained on a typical EEG dataset for emotion recognition. Results show that many relevant components found by XAI methods are shared across the sessions and can be used to build a system able to generalise better. However, relevant components of the input signal also appear to be highly dependent on the input itself.

Keywords

BCI, XAI, EEG, Dataset Shift, cross-session

In this research work, we experimentally investigate the performances of several well-known eXplainable Artificial (XAI) methods proposed in the literature in the context of Brain-Computer Interface (BCI) problems using EEG input-based Machine Learning (ML) algorithms to evaluate the possibility of alleviating the *Dataset Shift problem*. This is not a trivial issue as, differently from other signals, the non-stationarity of EEG signals makes them hard to analyse. In recent years, Brain-Computer Interfaces (BCIs) have been emerging as technology allowing the human brain to communicate with external devices without the use of peripheral nerves and muscles, enhancing the interaction capability of the user with the environment. In particular, several proposals of BCI methods based on Electroencephalographic (EEG) signals are receiving growing interest by the scientific community thanks to its implication in medical purposes [1, 2, 3, 4], other than other fields such as entertainment [5], education [6], and marketing [7]. This is because measuring and monitoring the brain's electrical activity can provide important

XAI.it 2022 - Italian Workshop on Explainable Artificial Intelligence
*Corresponding author.
†These authors contributed equally.
andrea.apicella@unina.it (A. Apicella); francesco.isgro@unina.it (F. isgrò); andrea.pollastro@unina.it (A. Pollastro); rprevete@unina.it (R. Prevete)
0000-0002-5391-168X (A. Apicella); 0000-0001-9342-5291 (F. isgrò); 0000-0003-4075-0757 (A. Pollastro); 0000-0002-3804-1719 (R. Prevete)

information related to the brain's physiological, functional, and pathological status. EEG signals are particularly suitable to this aim thanks to their important qualities such as non-invasiveness and high temporal resolution [8]. Furthermore, several solutions exploiting EEG acquisition devices more comfortable and with a low number of electrodes are being proposed [9, 10, 11, 12], allowing an acquisition process less influenced by noise due to the user-device interaction. Thanks to its properties, the EEG signal is one of the most promising candidates to become one of the most used communication channels between man and machine.

Several BCI solutions adopting ML methods are proposed in the literature. Generally, EEG data acquired from persons subjected to well-known stimuli are used in the training stage. These data are labelled following some established protocol, usually dependent on the task. For example, in an Emotion Recognition (ER) task, stimuli can be images or videos considered able to elicit particular emotions. Therefore the labels can be inferred by the stimuli or declared by the subject, who will say whether or not he felt a specific emotion during the stimulus administration. If the training stage is successful, the model can generalise on new unlabelled data, such as new acquisition from another subject or the same subject in another session.

However, one of the main defects of the EEG signal is that its statistical characteristics change over time. This implies that even under the same conditions and for the same task, significantly different signals can be acquired just as time passes. It is important to highlight that this phenomenon can also occur using the same stimuli-reaction (e.g., same emotions with the same stimuli) to the same subject at different times, leading to substantially different EEG signals even for the same subject. This problem is even more present among different subjects, who, given the same stimuli and emotions, can produce very different acquisitions between them. For these reasons, EEG is considered a non-stationary signal [13]. More in detail, the following cases in an EEG-based task can arise: i) a model trained on a set of EEG data acquired from a given subject at a specific time could not work on data acquired from the same subject at different times (Cross-Session generalisation problem), or ii) a model trained on data acquired from one or more subjects should not work as expected in classifying EEG signals acquired from a different subject at different times (Cross-Subject generalisation problem).

This type of problem can be treated as an instance of the Dataset Shift problem [14]. In a nutshell, Dataset Shift arises when the distribution of the training data differs from the data distribution used outside of the training stage (that is, running or evaluation stages); therefore the standard ML assumption [14] to have the same data distribution for both training and test set does not hold. Consequently, standard ML approaches can produce ML systems which exhibit poor generalisation performances.

On another side, a sub-field of Artificial Intelligence, eXplainable Artificial Intelligence (XAI), wants to explain the behaviour of AI systems, such as ML ones. In general, an explanation gives information on why an ML model returns an output given a specific input. In particular, several XAI methods applied to Deep Neural Networks are giving promising results [15, 16, 17, 18].

In the XAI context, several explanations are built by inspecting the model's inner mechanism to understand the input-output relationships, assigning a relevance score to each input component. However, building an explanation is particularly challenging if the model to inspect is a DNN; this is mainly for two reasons: i) DNNs offer excellent performances in several tasks, but at the price of high inner complexity of the models, leading toward low interpretability, ii) to help the ML user to understand the system behaviours, typical explanations have to be

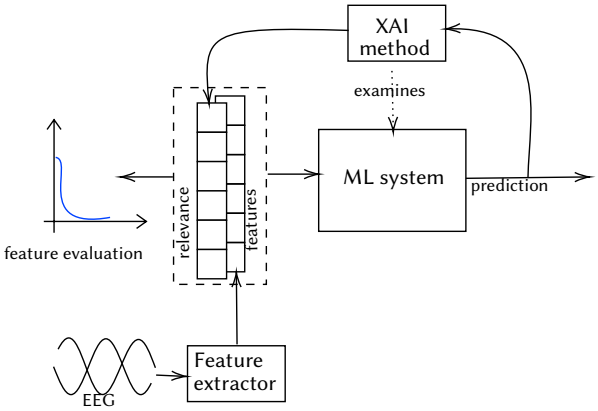

Figure 1: A general functional scheme of a Machine Learning (ML) architecture based on XAI methods to select and transform relevant input features with the aim of improving the performance of ML systems in the context of the dataset-shift problem.

humanly understandable.

The general idea of this work is that outputs' explanations of a trained ML model on given inputs can help the setup of new models able to overcome/mitigate the dataset shift problem, in general, and to generalise across subjects/sessions in case of EEG signals, in particular.

More specifically, in this work, we focus on how several well-known XAI methods proposed in literature behave in explaining decisions made by an ML system based on EEG input features (Fig. 1). Notice that several current XAI methods are usually tested on datasets, such as image and text recognition datasets [17, 19], where the domain shift problem is slight or not present. Therefore, this work is a first step toward a long term goal consisting in exploiting explanations made by XAI methods to locate and transform the main characteristics of the input for each given output, and to build ML systems able to generalise toward different data coming from different probability distributions (in this context, sessions and subjects). To this end, in this paper, we evaluate and analyse the explanations produced by a set of well-known XAI methods on an ML system trained on data taken from SEED [20], a public EEG dataset for an emotion classification task. The results obtained show, on oneside, that only some well-known XAI methods produce reliable explanations in the EEG domain in the analysed task. On another side, it is shown that the relevant components found in the training data can only be partially used on data acquired outside of the training stage. Notably, many relevant components found in the training data are still relevant across the sessions.

The paper is organised as follows: In Section 1, a brief description of the related works is reported. In Section 2 the proposed evaluation framework is presented. In Section 3 the obtained results are discussed. Finally, in Section 4 is devoted to final remarks and future developments.

1. Related works

In general, Modern ML approaches, as Deep learning, are characterised by a lack of transparency of their internal mechanisms, making it not easy for the AI scientist to understand the real reasons behind the inner behaviours. In this case, the relationships of the classified emotion with the EEG input are often challenging to understand. In the EEG-based applications, works based on simple features selection strategies to choose the best EEG features are widely proposed in the literature, such as [21, 22]. These studies, however, are based on standard feature selection methods, without exploiting information given by XAI methods. XAI is a branch of AI interested to "explain" ML behaviours. This is done providing methods for generating possible explanations of the model's outputs. XAI methods are gaining prominence in explaining several classification systems based on several inputs, such as images [17, 23], natural language processing [24], clinical decision support systems [25], and so on. To the best of our knowledge, however, the number of research works which attempt to improve the performance of ML models on the relying on XAI's methods is enough limited, especially in the context of bio-signal classification problems. For example, in [26, 27] feature selection procedures are performend on biomedical data by exploiting Correlation-based Feature Selection and Chaotic Spider Monkey Optimization methods. In [28] the authors propose to use an occlusion sensitivity analysis strategy [29] to locate the most relevant cortical areas in a motor imagery task. In [30] the use of XAI methods to interpret the answer of Epilepsy Detection systems is discussed.

2. Methods

Bearing in mind that we want to use the XAI method to alleviate the dataset shift problem in the BCI context, we conducted a series of experiments having the following goals: 1) testing the capability of the selected XAI methods to find relevant components for this specific signal; 2) verifying how much relevant components are dependent on the single sample of the dataset where the relevance are computed; 3) how much relevant components can be considered shared among samples of the same session, and finally 4) how much relevant components can be considered shared between samples of two different sessions, where the data shift problem is typically present.

In the remaining of this section, a brief description of the tested XAI methods is reported, followed by the used data and model descriptions. Finally experimental assessment and the evaluation strategy adopted are reported.

2.1. Investigated XAI Methods

In this work, we analyse XAI methods proposing explanations in terms of relevance of the input components on the output returned by a given classifier. More in detail, the following XAI methods are investigated: Saliency [31], Guided Backpropagation [32], Layer-wise Relevance Propagation (LRP) [33], Integrated Gradients [34], and DeepLIFT [35].

2.1.1. Saliency

Saliency method is one the of the simplest and more intuitive method to build an explanation of a ML system. Proposed in [31], Saliency method is based on the gradient of the output function of the ML system respect to its input. In a nutshell, an explanation of the output $C(\mathbf{x})$ of a ML system fed with an input $\mathbf{x} \in \mathbb{R}^d$ is built generating a saliency map leveraging on the gradient $\frac{\partial C}{\partial \mathbf{x}}$ of C with respect to its input computed through backpropagation. The magnitude of the gradient indicates how much the features need to be changed to affect the class score.

2.1.2. Guided BackPropagation

Guided BackPropagation (Guided BP) [32] can be seen as a slight variation of Saliency method proposed in [31]. The main difference is in the value used as gradient in case of rectified activation functions (ReLU): in Saliency method, the real gradient is used in computing the features relevance. Instead, Guided BP starts from the hypothesis that the user is not interested if a feature "decreases" (i.e., negative value) a neuron activation, but only in the most relevant ones. Therefore, instead of the true gradient, in guided BP a gradient transformation is used to prevent backward flow of negative values, avoiding to decrease the neuron activations and highlighting the most relevant features. Obviously, Guided BP can fail to highlight inputs that contribute negatively to the output due to "zero-ing" the negative values.

2.1.3. Layer-wise Relevance Propagation

Layer-wise Relevance Propagation (LRP) associates a relevance value to each input element (pixels in case of images) to build explanations for the ML model answer. In a nutshell, the output $C(\mathbf{x})$ of a ML system on an input $\mathbf{x} \in \mathbb{R}^d$ is decomposed as a sum of relevances on the single features composing $\mathbf{x}$, i.e. $C(\mathbf{x}) \simeq \sum_{i=1}^{d} R_i$ where R_i is a score of the local contribution of the i-th feature on the produced output. In particular, positive values denote positive contributions, while negative values negative contributions. Applied to ANN, this principle can be generalised across each pair of consecutive layers l and $l+1$ of a network composed of L layers such that $\sum_{i=1}^{q} R_i^{(l+1)} = \sum_{i=1}^{q'} R_i^{(l)}$ where q and q' are the features of the layers $l+1$ and l respectively. Since the final network output $C(\mathbf{x})$ of an ANN is the output of the L-th layer, it results that $C(\mathbf{x}) = \cdots = \sum_{i=1}^{q} R_i^{(l+1)} = \sum_{i=1}^{q'} R_i^{(l)} = \cdots = \sum_{i=1}^{d} R_i$. This rule can be interpreted as a conservation rule, and leveraging on that different methods to compute the relevance have been proposed, depending on the type of features involved. In case of densely connected layers, the most known rule is the $z - rule$ [33], which takes care of the neuron activations of each layer to compute the final relevance of each layer.

2.1.4. Integrated Gradients

One of the main drawbacks of simple gradient-based method is that the gradient respect to the input should be small in the neighbourhood of the input features also for relevant ones.

Instead of using only the gradient respect to the original input, [34] proposed to average all the gradients between the original input $\mathbf{x}$ and a baseline input $\mathbf{x}^{ref}$ (that is, an input s.t. $C(\mathbf{x}^{ref})$ results in a neutral prediction). In this way, if features of inputs closer to the baseline have higher gradient magnitudes, they are taken into account thanks to the average operator. More formally, the importance of each feature x_i computed by Integrated Gradient (IG) is defined as $IG(x_i) = (x_i - x_i^{ref}) \int_{\alpha=0}^{1} \frac{\partial C\left(x_i^{ref} + \alpha(x_i - x_i^{ref})\right)}{\partial x_i} d\alpha$. In other words, IG aggregates the gradients along the intermediate inputs on the straight-line between the baseline and the input, selected as $\alpha \in [0, 1]$ changes.

2.1.5. DeepLIFT

In [35] a method consisting in assigning feature relevance scores according to the difference between the neurons activation and a reference activation (such as the baseline for Integrated Gradient method) is proposed. The authors proposed to compute for each feature a multiplier entity similar to a partial derivative, but leveraging over finite differences instead of infinitesimal ones. Each multiplier can be defined as $m_{\Delta x \Delta t} = \frac{R_{\Delta x \Delta t}}{\Delta x}$ and represents the ratio between i) the contribution $R_{\Delta x \Delta t}$ of the difference $\Delta x = x - x^{ref}$ from the reference x^{ref} of each feature x to the difference $\Delta t = t - t^{ref}$ between the output t and the reference output t^{ref}, and ii) the difference Δx. Therefore, the authors proposed a set of rules to compute the features relevance based on the proposed multipliers exploiting a Back Propagation-based approach.

2.2. Data

The SEED dataset consists of EEG signals recorded from 15 subjects stimulated by 15 film clips carefully chosen to induce negative, neutral and positive emotions. Each film clip has a duration of approximately 4 minutes. Three sessions of 15 trials were collected for each subject. EEG signals were recorded in 62 channels using the ESI Neuroscan System[1]. During our experiments, we considered the pre-computed differential entropy (DE) features smoothed by linear dynamic systems (LDS) for each second, in each channel, over the following five bands: delta (1–3 Hz); theta (4–7 Hz); alpha (8–13 Hz); beta (14–30 Hz); gamma (31–50 Hz).

In this work, the relevant components of an EEG signal can be considered taking into account three different aspects of the signal: i) considering each single feature composing the input, ii) considering each single band composing the EEG signal, that are alpha, beta, theta, and delta, and iii) considering each single channel/electrode from which the input EEG signal was acquired. Cases ii) and iii) can be viewed as different aggregations of fixed features of the EEG signals. In the following of this work, we refer generically with the term "components" where it is not necessary to specify if we are talking about features, bands or channels.

2.3. Experimental assessment

To achieve the goals defined at the beginning of this section, the following experiments are made: firstly, to evaluate the capability of the selected XAI methods to find relevant components, we analysed the explanations of model responses on data coming from the same session where the

[1]https://compumedicsneuroscan.com

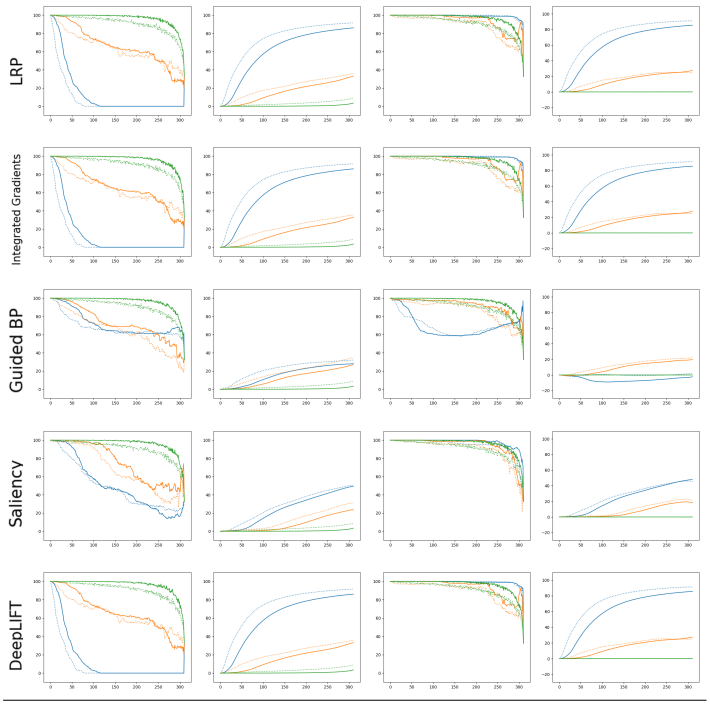

Figure 2: MoRF (first column), AOPC (second column), LeRF (third column), and ABPC (fourth column) curves using the tested XAI methods for both intra-session (solid line) and inter-session (dotted lines) considering features as signal components. Results scoring the input components using effective relevance (blue lines) and averaged relevance computed on training data (orange lines) are reported for each case and compared with a random component scoring (green lines). On the $x$and y axes are reported the iteration step in the curve generation and the accuracy, respectively.

training data was extracted; then, to evaluate how much relevant components can be considered shared among samples of the same session, we analysed the explanations of the model responses on data belonging to a session different from the training one. Finally, to evaluate if relevant components can be considered shared between samples of two different sessions and how much relevant components are dependent on the single data sample where the relevance are computed, the components' average relevance of data coming from the training session are used as sorting score and select the components belonging to another session.

Summarising, the following cases are considered: i) intra-session case: given a model C trained on data coming from a session s_{tr}, explanations of the responses on input data belonging to the same session s_{tr} are built. ii) inter-session case: given a model C trained on data coming from a session s_{tr}, explanation of responses on inputs belonging to a sessions s_{te} different from s_{tr} are built. Each of these cases can be in turn evaluated considering two different relevance: a) real relevance: we assume that it is possible to compute the relevance of the input, since the classification output is known; b) presumed relevance: we assume that the relevance of the input is not available, since we are outside the training stage. In this case, we use the average of

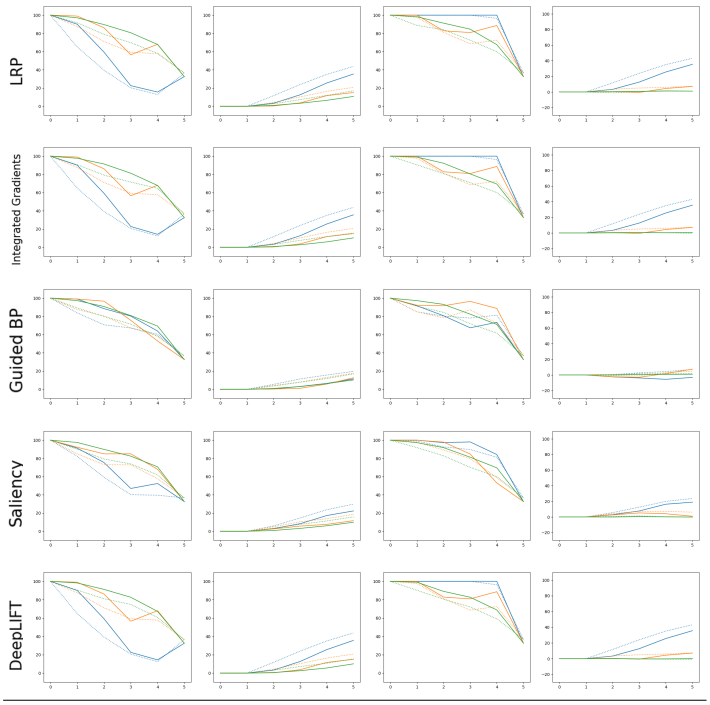

Figure 3: MoRF (first column), AOPC (second column), LeRF (third column), and ABPC (fourth column) curves using the tested XAI methods for both intra-session (solid line) and inter-session (dotted lines) considering delta, theta, alpha, beta, gamma EEG bands as signal components. Results scoring the input components using effective relevance (blue lines) and averaged relevance computed on training data (orange lines) are reported for each case and compared with a random component scoring (green lines). On the x and y axes are reported the iteration step in the curve generation and the accuracy, respectively.

the same component relevance obtained on training data as component relevance.

2.4. Evaluation

For each case, we investigated the explanations returned by XAI method in order to analyse if the explanations built can correctly identify the impact that i) each input feature, ii) each electrode, and iii) each frequency band has on the classification performances. To this aim, we consider as relevance for each feature the relevance score returned by the XAI method, for each electrode the mean relevance score of all the feature belonging to the electrode, and for each frequency bands the mean average score of all the features belonging to the frequency band. Therefore, the following evaluation strategies are then adopted and repeated considering features, electrodes, and frequency bands as EEG components in turn: a) analysis of the MoRF (Most Relevant First) curve, proposed in [33, 36]. In case of evaluating the components relevance returned by the explanation method, the MoRF curve can be computed as follows: given a classifier, an input EEG signal $\mathbf{x}$ and the respective classification output $C(\mathbf{x})$, the EEG components are

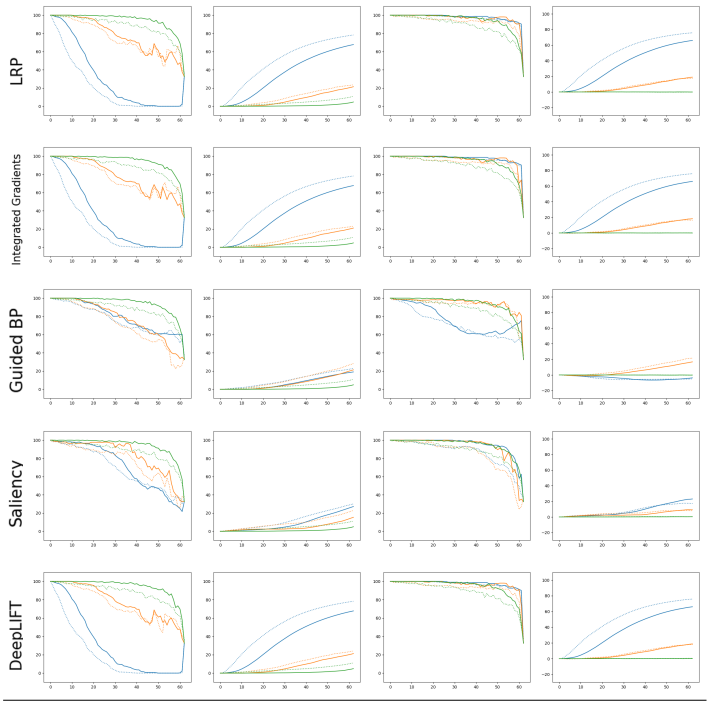

Figure 4: MoRF (first column), AOPC (second column), LeRF (third column), and ABPC (fourth column) curves using the tested XAI methods for both intra-session (solid line) and inter-session (dotted lines) considering the acquisition electrodes as signal components. Results scoring the input components using effective relevance (blue lines) and averaged relevance computed on training data (orange lines) are reported for each case and compared with a random component scoring (green lines). On the x and y axes are reported the iteration step in the curve generation and the accuracy level reached, respectively.

iteratively replaced by zeros, following the descending order with respect to the relevance values returned by the explanation method. In other words, performances were analysed by removing (i.e. setting to zero) components in a decreasing order of impact on the predictions supplied by the explanation. In this way, the expected curve is such that more relevant the identified components are for the classification output, steepest is the curve. Furthermore, the change in the AOPC (Area Over Perturbation Curve) value is reported for each MoRF iteration. AOPC is computed as $AOPC = \frac{1}{K+1}\langle \sum_{k=0}^{K} C(\mathbf{x}^{(0)}) - C(\mathbf{x}^{(k)}))\rangle$ where K is the total number of iterations, $\mathbf{x}^{(0)}$ is the original input, $\mathbf{x}^{(k)}$ is the input at the iteration k, and $\langle\cdot\rangle$ is the average operator over a set of inputs. MoRFs and AOPCs are reported also considering channels and bands as characteristics to analyse.

b) the analysis of the LeRF (Least Relevant First) curve, proposed in [36]. Differently from the MoRF curve, in this case the EEG components are iteratively removed following the ascending order with respect to the relevance values returned by the explanation method. In the resulting

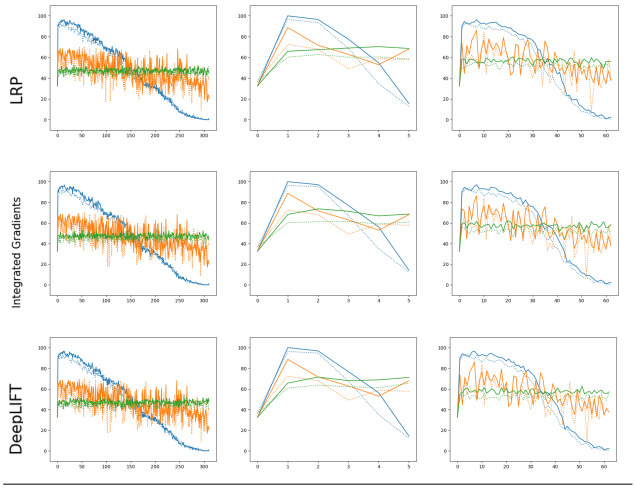

Figure 5: A first analysis of the discriminative power of the components alone. Signals composed of only one component following the XAI relevance order are fed to the ML system in an iterative manner. Results are reported for both intra-session (solid line) and inter-session (dotted lines) considering features (1st column), bands (2nd column), and electrodes (3rd column) as signal components. Results scoring the input components using effective relevance (blue lines) and averaged relevance computed on training data (orange lines) are compared with a random component scoring (green lines).

curve, we expect that the classification output should be very close to the original value when the less relevant components are removed (corresponding to the first iterations), dropping quickly to zero as the process goes toward the remotion of relevant elements. While the MoRFs report how much the classifier output is destroyed removing highly relevant components, LeRFs report how much the least relevant components leave the output intact. These indications can be combined in the ABPC (Area Between Perturbation Curves, [36]) quantity, defined as $ABPC = \frac{1}{K+1}\langle \sum_{k=0}^{K} C(\mathbf{x}_{MoRF}^{(k)}) - C(\mathbf{x}_{LeRF}^{(k)}))\rangle$ where $\mathbf{x}_{MoRF}^{(k)}$, $\mathbf{x}_{LeRF}^{(k)}$ are the values of the MoRF and LeRF values obtained at the k-th iteration step. ABPC is an indicator of how good the XAI method is. The larger the ABPC value, the better the XAI method. LeRFs and ABPCs are reported also for channels and bands analysis.

c) an analysis of the discriminative power of each component alone is made. Signals composed of only one component following the relevance order given by the XAI method are fed to the ML system in an iterative manner, and the relative performance curves are plotted.

All the experiments were carried out only on correctly classified samples.

2.5. Classification model

The XAI methods are evaluated on a feed-forward fully connected multi layered neural networks. Hyperparameters were tuned through bayesian optimisation [37]: the number of layers was constrained to a maximum of 3; for each layer, the number of nodes was searched in the space $\{2^n | n \in \{4, 5, ..., 10\}\}$ having the ReLU as activation function. Each experiment was

run having early stopping as convergence criterion with 20 epochs of patience. The 10 % of the training set was extracted using stratified sampling [38] on class labels and considered as validation set. Network optimisation was performed using Adam optimiser [39], whose learning rate that was searched in the space $\{0.1, 0.01, ..., 0.0001\}$.

As a result from the model selection stage, the best setting consisted in ANN having 3 layers with 128, 256 and 128 neurons respectively. The learning rate was set to 0.01, and reduced to its 10 % whenever the loss on validation set plateaus for 10 consecutive epochs.

3. Results & discussions

Since the behaviour of the explored XAI methods resulted in being similar across all the subjects, we report only the results obtained on just one subject. In Fig. 2, 3, and 4 MoRF and LeRF curves using the tested XAI methods are reported for both intra-session and inter-session cases, considering as components to remove at each step features (Fig. 2), bands (Fig. 3), and channels (Fig. 4), respectively. Results related to the intra-session cases are reported with solid lines, while those regarding the inter-session case are marked with dotted lines. On the x axis and y axis are reported the iteration step in the curve generation and the accuracy level reached, respectively. With blue lines, results scoring the input components using effective relevance are reported; with orange lines, results scoring the components using averaged relevance computed on training data are reported; with green lines, results related to random choice.

All the curves were compared with the random curve obtained by removing the components in random order. Several interesting points can be highlighted:

1) In all the cases, LRP, IG and Deep LIFT resulted in being more reliable XAI methods with respect to Saliency and Guided BP. Indeed, MoRF curves of LRP, IG and Deep LIFT have high slopes, however similar to each other, differently from Saliency and Guided BP. In particular, the latter is the only method among those tested whose explanations do not always seem to capture the relevant components, especially in the case of intra-session. These considerations seem consistent with what is reported in LeRF, AOPC, and ABPC.

2) counterintuitively, in almost all the cases, explanations built in inter-session cases seem to be more reliable (i.e., highlighting more relevant features) with respect to intra-session cases. This behaviour can be explained by a more significant "robustness" to input changes of the trained classifier toward data from the same training session (intra-session case), where with "robustness" we mean the ability of the classifier to give a much higher score to the chosen class respect to the other ones. Instead, data coming from different sessions (inter-session case) leads the classifier toward more borderline class scores, therefore minimum perturbations of the input data can lead to different classes, influencing the final performance and the resulting MoRF curve during the features' removal.

3) Although the best XAI methods can locate relevant features/channels/bands for each input data sample, they don't seem able to locate a set of relevant components for all the samples. In other words, the examined XAI methods fail to "generalise" to a set of general features/channels/bands relevant to the most significant part of the possible inputs. Indeed, removing the components following the average relevance (obtained in the training stage) in reverse order (MoRF orange curves) does not lead to a steep drop in performance, as in the

other case (MORF blue curves). Even in some cases, such as using bands as a component to assign the relevance (Fig. 3), the obtained curves overlap with the random ones, highlighting that removing bands in random order is almost the same that following the relevance assigned by the XAI method. This is confirmed by the other evaluation metrics adopted, i.e. MeRF, AOPC and ABPC curves.

In Fig. 5 a first analysis of the discriminative power of the components alone is made. Signals composed of only one component following the relevance order given by the XAI method are iteratively fed to the ML system. We limit the analysis only to the best XAI methods identified in the previous step: DeepLIFT, IG and LRP. From the obtained results, it is interesting to notice that the components considered most relevant for each sample fed to the classifier are enough to reach high performances. However, considering the average relevance detected during the training stage, the best components do not seem to lead toward similar performance, although they are still better than a random choice.

4. Conclusions

In this work, the performances of several XAI methods proposed in the literature in the context of Brain-Computer Interface (BCI) problems using EEG input-based Machine Learning (ML) algorithms are experimentally evaluated. The focus was on how much the relevant components selected by XAI methods be shared between different samples of the same dataset (in this case, same session) or samples of different datasets (in this case, different sessions). The final results show that the components considered most relevant for each sample fed to the classifier are enough to achieve high performances. However, the components detected considering the best average relevance during the training stage do not seem to lead toward performance returned by components scored according to their effective relevance returned by the XAI method.

This work is the first step toward developing a BCI system able to exploit XAI methods to alleviate the dataset shift problem. However, in this work, only data belonging to different sessions but acquired from the same subjects are taken into account. In future work, we plan to analyse the behaviour of XAI methods with inter-subject classifiers. Several benefits can be obtained in the EEG-based BCI applications by the proposed project. For example, a BCI system can work across different subjects without retraining the model on each new unseen subject (subject-independent model). Furthermore, a better understanding of the relationships between the system inputs and outputs provided by XAI explanations can lead to the developing and producing more effective EEG acquisition devices.

Acknowledgments

This work is supported by the European Union - FSE-REACT-EU, PON Research and Innovation 2014-2020 DM1062/2021 contract number 18-I-15350-2 and by the Ministry of University and Research, PRIN research project "BRIO – BIAS, RISK, OPACITY in AI: design, verification and development of Trustworthy AI.", Project no. 2020SSKZ7R .

References

[1] A. Apicella, P. Arpaia, G. Mastrati, N. Moccaldi, Eeg-based detection of emotional valence towards a reproducible measurement of emotions, Scientific Reports 11 (2021) 1–16.

[2] A. Apicella, P. Arpaia, M. Frosolone, N. Moccaldi, High-wearable eeg-based distraction detection in motor rehabilitation, Scientific Reports 11 (2021) 1–9.

[3] P. Arpaia, S. Criscuolo, E. De Benedetto, N. Donato, L. Duraccio, A wearable ar-based bci for robot control in adhd treatment: Preliminary evaluation of adherence to therapy, in: 2021 15th International Conference on Advanced Technologies, Systems and Services in Telecommunications (TELSIKS), IEEE, 2021, pp. 321–324.

[4] P. Arpaia, A. Esposito, A. Natalizio, M. Parvis, How to successfully classify eeg in motor imagery bci: a metrological analysis of the state of the art, Journal of Neural Engineering (2022).

[5] D. Marshall, D. Coyle, S. Wilson, M. Callaghan, Games, gameplay, and bci: the state of the art, IEEE Transactions on Computational Intelligence and AI in Games 5 (2013) 82–99.

[6] A. Apicella, P. Arpaia, M. Frosolone, G. Improta, N. Moccaldi, A. Pollastro, Eeg-based measurement system for monitoring student engagement in learning 4.0, Scientific Reports 12 (2022) 1–13.

[7] F. R. Mashrur, K. M. Rahman, M. T. I. Miya, R. Vaidyanathan, S. F. Anwar, F. Sarker, K. A. Mamun, Bci-based consumers' choice prediction from eeg signals: An intelligent neuromarketing framework, Frontiers in Human Neuroscience 16 (2022).

[8] D. P. Subha, P. K. Joseph, R. Acharya U, C. M. Lim, et al., Eeg signal analysis: a survey, Journal of medical systems 34 (2010) 195–212.

[9] P. Arpaia, L. Callegaro, A. Cultrera, A. Esposito, M. Ortolano, Metrological characterization of a low-cost electroencephalograph for wearable neural interfaces in industry 4.0 applications, in: 2021 IEEE International Workshop on Metrology for Industry 4.0 & IoT (MetroInd4. 0&IoT), IEEE, 2021, pp. 1–5.

[10] A. J. Casson, D. C. Yates, S. J. Smith, J. S. Duncan, E. Rodriguez-Villegas, Wearable electroencephalography, IEEE engineering in medicine and biology magazine 29 (2010) 44–56.

[11] A. Apicella, P. Arpaia, E. De Benedetto, N. Donato, L. Duraccio, S. Giugliano, R. Prevete, Enhancement of ssveps classification in bci-based wearable instrumentation through machine learning techniques, IEEE Sensors Journal 22 (2022) 9087–9094.

[12] A. Apicella, P. Arpaia, F. Isgrò, G. Mastrati, N. Moccaldi, A survey on eeg-based solutions for emotion recognition with a low number of channels, IEEE Access (2022) 1–1. doi:10.1109/ACCESS.2022.3219844.

[13] A. Y. Kaplan, A. A. Fingelkurts, A. A. Fingelkurts, S. V. Borisov, B. S. Darkhovsky, Nonstationary nature of the brain activity as revealed by eeg/meg: methodological, practical and conceptual challenges, Signal processing 85 (2005) 2190–2212.

[14] J. Quinonero-Candela, M. Sugiyama, A. Schwaighofer, N. D. Lawrence, Dataset shift in machine learning, Mit Press, 2008.

[15] A. A., I. F., P. R., T. G., Contrastive explanations to classification systems using sparse dictionaries, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 11751 LNCS (2019) 207 – 218.

[16] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, K.-R. Müller, Layer-wise relevance propagation: an overview, Explainable AI: interpreting, explaining and visualizing deep learning (2019) 193–209.
[17] M. T. Ribeiro, S. Singh, C. Guestrin, " why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.
[18] A. Apicella, F. Isgro, R. Prevete, G. Tamburrini, A. Vietri, Sparse dictionaries for the explanation of classification systems, in: PIE, 2015, p. 009.
[19] A. Apicella, S. Giugliano, F. Isgrò, R. Prevete, Exploiting auto-encoders and segmentation methods for middle-level explanations of image classification systems, Knowledge-Based Systems 255 (2022) 109725.
[20] W.-L. Zheng, B.-L. Lu, Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks, IEEE Transactions on Autonomous Mental Development 7 (2015) 162–175. doi:10.1109/TAMD.2015.2431497.
[21] A. Wosiak, A. Dura, Hybrid method of automated eeg signals' selection using reversed correlation algorithm for improved classification of emotions, Sensors 20 (2020) 7083.
[22] X. Zheng, X. Liu, Y. Zhang, L. Cui, X. Yu, A portable hci system-oriented eeg feature extraction and channel selection for emotion recognition, International Journal of Intelligent Systems 36 (2021) 152–176.
[23] A. Apicella, F. Isgrò, R. Prevete, A. Sorrentino, G. Tamburrini, Explaining classification systems using sparse dictionaries, ESANN 2019 - Proceedings, 27th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (2019) 495 – 500.
[24] K. Qian, M. Danilevsky, Y. Katsis, B. Kawas, E. Oduor, L. Popa, Y. Li, Xnlp: A living survey for xai research in natural language processing, in: 26th International Conference on Intelligent User Interfaces-Companion, 2021, pp. 78–80.
[25] T. A. Schoonderwoerd, W. Jorritsma, M. A. Neerincx, K. Van Den Bosch, Human-centered xai: Developing design patterns for explanations of clinical decision support systems, International Journal of Human-Computer Studies 154 (2021) 102684.
[26] E. Laxmi Lydia, C. Anupama, N. Sharmili, Modeling of explainable artificial intelligence with correlation-based feature selection approach for biomedical data analysis, in: Biomedical Data Analysis and Processing Using Explainable (XAI) and Responsive Artificial Intelligence (RAI), Springer, 2022, pp. 17–32.
[27] R. P. Selvam, A. S. Oliver, V. Mohan, N. Prakash, T. Jayasankar, Explainable artificial intelligence with metaheuristic feature selection technique for biomedical data classification, in: Biomedical Data Analysis and Processing Using Explainable (XAI) and Responsive Artificial Intelligence (RAI), Springer, 2022, pp. 43–57.
[28] C. Ieracitano, N. Mammone, A. Hussain, F. C. Morabito, A novel explainable machine learning approach for eeg-based brain-computer interface systems, Neural Computing and Applications 34 (2022) 11347–11360.
[29] M. D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: European conference on computer vision, Springer, 2014, pp. 818–833.
[30] P. Rathod, S. Naik, Review on epilepsy detection with explainable artificial intelligence, in: 2022 10th International Conference on Emerging Trends in Engineering and Technology-

Signal and Information Processing (ICETET-SIP-22), IEEE, 2022, pp. 1–6.
[31] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, arXiv preprint arXiv:1312.6034 (2013).
[32] J. T. Springenberg, A. Dosovitskiy, T. Brox, M. Riedmiller, Striving for simplicity: The all convolutional net, arXiv preprint arXiv:1412.6806 (2014).
[33] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PloS one 10 (2015) e0130140.
[34] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: International conference on machine learning, PMLR, 2017, pp. 3319–3328.
[35] A. Shrikumar, P. Greenside, A. Kundaje, Learning important features through propagating activation differences, in: International conference on machine learning, PMLR, 2017, pp. 3145–3153.
[36] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, K.-R. Müller, Evaluating the visualization of what a deep neural network has learned, IEEE transactions on neural networks and learning systems 28 (2016) 2660–2673.
[37] J. Snoek, H. Larochelle, R. P. Adams, Practical bayesian optimization of machine learning algorithms, Advances in neural information processing systems 25 (2012).
[38] J. Neyman, On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection, in: Breakthroughs in statistics, Springer, 1992, pp. 123–150.
[39] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).