XAI approach for addressing the dataset shift problem: BCI as a case study

Andrea Apicella[1,2,3,*,†], Francesco isgrò[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

In the Machine Learning (ML) literature, a well-known problem is the *Dataset Shift* problem where, differently from the ML standard hypothesis, the data in the training and test sets can follow different probability distributions leading ML systems toward poor generalisation performances. Therefore, such systems can be unreliable and risky, particularly when used in safety-critical domains. This problem is intensely felt in the Brain-Computer Interface (BCI) context, where bio-signals as Electroencephalographic (EEG) are used. In fact, EEG signals are highly non-stationary signals both over time and between different subjects. Despite several efforts in developing BCI systems to deal with different acquisition times or subjects, performance in many BCI applications remains low. Exploiting the knowledge from eXplainable Artificial Intelligence (XAI) methods can help develop EEG-based AI approaches, overcoming the performance returned by the current ones. The proposed framework will give greater robustness and reliability to BCI systems with respect to the current state of the art, alleviating the dataset shift problem and allowing a BCI system to be used by different subjects at different times without the need for further calibration/training stages.

Keywords

XAI, EEG, cross-subject, dataset shift, BCI

1. Introduction

Supervised Machine Learning (ML) models can learn from human-classified examples (labelled data) to generalise toward new unknown data (unlabelled data). In a nutshell, two different stages are needed for a supervised ML system to work properly, i) a *Training stage*, where a set of labelled examples are fed to the system, so that it can learn a good mapping between examples and the provided labels, and ii) a *Running/Production stage*, where unlabelled examples are fed to the system which returns the most probable labels using the mapping learned in the training stage. However, if labelled data not used in the training stage are available, they can be used to evaluate the trained model (*Evaluation stage*). ML classical methods start from the hypothesis that all the used data in any stage come from the same distribution probability. This assumption

1st Workshop on Bias, Ethical AI, Explainability and the role of Logic and Logic Programming, BEWARE-22, co-located with AIxIA 2022, University of Udine, Udine, Italy, 2022

*Corresponding author.
†These authors contributed equally.
andrea.apicella@unina.it (A. Apicella); francesco.isgro@unina.it (F. isgrò); rprevete@unina.it (R. Prevete)
0000-0002-5391-168X (A. Apicella); 0000-0001-9342-5291 (F. isgrò); 0000-0002-3804-1719 (R. Prevete)

can be strong in real environments since the real distributions of the data are often unknown. As a consequence, the trained model could not perform well on the production data if its distribution probability is different from the training one, failing in generalisation. Or, even worse, if the data used in the training and evaluation stage come from the same distribution, there may be an overestimation of the model performance if the production stage data come from a different distribution. In the ML literature, this is known as the Dataset Shift problem [1]. Summarising, 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 starting ML assumption does not hold. Consequently, standard ML approaches can produce ML systems which exhibit poor generalisation performances making such systems unreliable and risky, especially when used in safety-critical domains. This problem is particularly felt in Brain Computer Interface (BCI, [2]) context, where some bio-signals acquired from the brain, such as the Electroencephalographic (EEG) ones, can continually change their statistical characteristics. This implies that even under the same conditions and for the same task, significantly different signals can be acquired just as time passes, also occur using the same stimuli-reaction (e.g., same emotions with the same stimuli in an emotion detection task) on the same subject. This problem is even more evident in different subjects who, given the same stimuli and responses, can produce very different acquisitions between them. For these reasons, EEG is considered a non-stationary signal [3]. Because of this, big differences across acquisitions made at different times or across different subjects can arise leading to different data probability distributions. 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 could not work as expected in classifying EEG signals acquired from a different subject at different times (Cross-Subject generalisation problem). These conditions lead the model toward poor generalisation performance, and the construction of unreliable ML systems. Several strategies have been proposed to overcome the dataset shift, considering the differences between the possible distributions involved. Several proposals are based on Transfer Learning (TL) methods [4], a family of approaches to transfer knowledge learned from a Machine Learning system to another. TL approaches can be categorised into several subfamilies, such as Domain Adaptation (DA) [4] and Domain Generalisation (DG) [5] approaches. However, Existing DA and DG solutions are still far from being able to be adopted in real EEG-based tasks because of their low classification 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 describes why an ML model returns a given output given a specific input. In particular, several XAI methods applied to Deep Neural Networks are giving promising results, such as [6, 7]. Our idea is that explanations about the outputs of a trained ML model can help to overcome/mitigate the dataset shift problem, in general, and to generalise across subjects/sessions in case of EEG signals, in particular. In the XAI context, several explanations are built by inspecting the model's inner mechanism to understand the input-output relationships. Therefore, explanations of an ML system can be used to locate and exploit, for each given output, the main input characteristics to build a new ML system able to generalise toward different data, also coming from different probability distributions. In this research work, explanations built by an *Artificial Explainer* are used to

improve the generalisation performance by analysing a set of ML models and understanding their inner input/output relations. In particular, we plan to develop solutions in the context of EEG signal classification problems which can lead to Subject-Independent models.

2. Related works

In recent years, BCI systems based on Electroencephalographic (EEG) input signals are receiving a strong interest by the scientific community thanks to the opportunity to exploit ML together with the EEG qualities, such as non-invasiveness and high temporal resolution [8], in several BCI applications such as healthcare [9, 10, 11, 12] and education [13]. However, 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 between the classifier's output and the EEG input are often challenging to understand. In the EEG-based applications, works based on features selection to choose the best EEG features are widely proposed in the literature, such as [14]. However, the greatest part of these studies does not take into account the inner state of the model, relying only on its input-output functional relations. XAI is a branch of AI concerned to "explain" ML behaviours. This is made providing methods for generating possible explanations of the model's outputs. To the best of our knowledge, the number of research works which attempt to improve the performance of ML models on the basis of XAI's methods is enough limited, especially in the context of bio-signal classification problems. For example, in [15, 16] feature selection procedures are carried out on biomedical data leveraging on feature selection and swarm intelligence methods. In [17] an occlusion sensitivity analysis strategy [18] to locate the most relevant cortical areas in a motor imagery task is used. In [19] the use of XAI methods to interpret the answer of Epilepsy Detection systems is discussed.

3. Methods

We start from the hypothesis that possible hidden relationships between the input EEG signal and the ML outputs can be identified by XAI methods, allowing the scientist to focus on the best features for the used model and, on the other side, if there are features leading to poor results as they are, for example, subject-specific features. In this work, we want to exploit XAI methods to select the best features, or the best feature functional transformations, involved in an EEG classification task. Our hypothesis relied on the fact that XAI methods can be used to locate which part of the input are more relevant for the model classification. Indeed, leveraging on the model inner parameters, several XAI methods are able to trace which input features are more involved for a given output, that can be interpreted as explanations about the produced output. Consequently, we propose to take advantage of XAI methods to select/transform suitable features to enhance the performance of a ML system in the context of domain-shift problems. A key step of this research work is to investigate the ability of the current main XAI methods to select the best input features for our aims. Therefore, a first step is to select a proper EEG feature transformation among those proposed in literature [20] for the task in exam. Next, existing methods able to build explanations suitable for the EEG domain in the selected feature

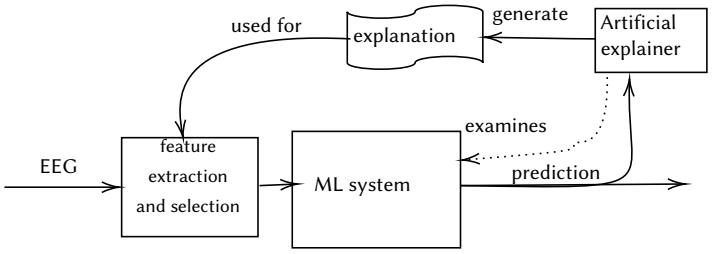

Figure 1: The proposed framework. Given EEG data as input for a given task, the ML system builds a prediction, interpreted by an Artificial explainer which generates an explanation to the given output. The explanation is then used by a feature extraction/selection method to leverage on the more effective set of feature for the next inputs.

space will be investigated, leveraging on classical XAI evaluation metrics (e.g., MoRF curves [21, 22]). Thereafter, an Artificial Explainer built upon a selected XAI method will provide an explanation to the model outputs. This explanation will be used to select/extract the proper feature space for a ML system, suppressing the features that can lead toward bad classification. In Fig. 1 a functional schema of the proposed work is reported. The proposed framework can be validated both for cross-session and cross-subject generalisation, comparing the performances obtained with the current state-of-art strategies.

4. Conclusions

This research project's main value is exploring XAI's use to overcome/mitigate the dataset shift problem. Overall this could lead to safer, more reliable and more transparent AI systems. In particular, we take advantage of this new approach in the context of BCI classification problems, where the dataset shift is especially relevant. In particular, we have the following main advantages: i) Improved BCI performances: selecting effective features exploiting methods such XAI-based can lead Subject-Independent BCI systems toward performances comparable to Subject-dependent ones, but without the main disadvantages of Subject-Dependent systems. ii) less expensive classification models: XAI method will be able to guide the model to select the best features for better performances, avoiding the problems related to the non-stationarity of the EEG signal in an automatic way, without any further operator interaction; iii) more comfortable systems: the lack of need for subject-specific training data leads to less time required for the users, resulting in greater comfort and less stress for the subjects; iv) development of more specific acquisition devices: a better understanding of the relationships between the system inputs and outputs provided by XAI explanations can lead toward 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] J. Quinonero-Candela, M. Sugiyama, A. Schwaighofer, N. D. Lawrence, Dataset shift in machine learning, Mit Press, 2008.

[2] T. M. Vaughan, W. J. Heetderks, L. J. Trejo, W. Z. Rymer, M. Weinrich, M. M. Moore, A. Kübler, B. H. Dobkin, N. Birbaumer, E. Donchin, et al., Brain-computer interface technology: a review of the second international meeting., IEEE transactions on neural systems and rehabilitation engineering 11 (2003) 94–109.

[3] 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.

[4] S. J. Pan, Q. Yang, A survey on transfer learning, IEEE Transactions on knowledge and data engineering 22 (2009) 1345–1359.

[5] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, C. C. Loy, Domain generalization: A survey, IEEE Transactions on Pattern Analysis and Machine Intelligence (2022).

[6] 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.

[7] 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.

[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] A. A., A. P., M. G., M. N., Eeg-based detection of emotional valence towards a reproducible measurement of emotions, Scientific Reports 11 (2021). doi:10.1038/s41598-021-00812-7.

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

[11] 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.

[12] 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.

[13] 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.

[14] A. Wosiak, A. Dura, Hybrid method of automated eeg signals' selection using reversed correlation algorithm for improved classification of emotions, Sensors 20 (2020) 7083.
[15] 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.
[16] 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.
[17] 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.
[18] M. D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: European conference on computer vision, Springer, 2014, pp. 818–833.
[19] 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.
[20] X. Li, D. Song, P. Zhang, Y. Zhang, Y. Hou, B. Hu, Exploring eeg features in cross-subject emotion recognition, Frontiers in neuroscience 12 (2018) 162.
[21] 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.
[22] A. Apicella, S. Giugliano, F. Isgró, R. Prevete, Explanations in terms of hierarchically organised middle level features, volume 3014, CEUR-WS, 2021, p. 44 – 57. Conference name: 2nd Italian Workshop on Explainable Artificial Intelligence, XAI.it 2021.