Strategies to Exploit XAI to Improve Classification Systems

Andrea Apicella[1,2,3](✉), Luca Di Lorenzo[1], Francesco Isgrò[1,2,3], Andrea Pollastro[1,2,3,4], and Roberto Prevete[1,2,3]

[1] Dipartimento di Ingegneria Elettrica e delle Tecnologie dell'Informazione, Università degli Studi di Napoli Federico II, Naples, Italy
andrea.apicella@unina.it

[2] Laboratory of Augmented Reality for Health Monitoring (ARHeMLab), Naples, Italy

[3] Laboratory of Artificial Intelligence, Privacy and Applications (AIPA Lab), Naples, Italy

[4] Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
https://www.dieti.unina.it, https://arhemlab.dieti.unina.it, http://aipa.dieti.unina.it

Abstract. Explainable Artificial Intelligence (XAI) aims to provide insights into the decision-making process of AI models, allowing users to understand their results beyond their decisions. A significant goal of XAI is to improve the performance of AI models by providing explanations for their decision-making processes. However, most XAI literature focuses on how to explain an AI system, while less attention has been given to how XAI methods can be exploited to improve an AI system. In this work, a set of well-known XAI methods typically used with Machine Learning (ML) classification tasks are investigated to verify if they can be exploited, not just to provide explanations but also to improve the performance of the model itself. To this aim, two strategies to use the explanation to improve a classification system are reported and empirically evaluated on three datasets: Fashion-MNIST, CIFAR10, and STL10. Results suggest that explanations built by Integrated Gradients highlight input features that can be effectively used to improve classification performance.

Keywords: XAI · Machine Learning · DNN · Integrated Gradients · attributions

1 Introduction

Explainable Artificial Intelligence (XAI) aims to provide an understanding of how AI models work and reasons beyond the decisions they make, allowing users to understand their results. This is particularly important as AI becomes more integrated into everyday life and critical decision-making processes such as healthcare and finance. However, it is essential to note that a large part of the current XAI literature proposes methods to provide explanations to AI

L. Longo (Ed.): xAI 2023, CCIS 1901, pp. 147–159, 2023.
https://doi.org/10.1007/978-3-031-44064-9_9

systems [1,2,12,13,16], but less attention is given on how XAI can be used to improve the performance of AI models. Indeed, the goal of XAI is not only to provide explanations but also to improve the AI model performance thanks to a more profound knowledge of the AI's decision-making strategies. This is a significant shortcoming in the context of such research studies, as XAI's overall goal is also to improve the performance of AI models thanks to a more profound knowledge of the AI's decision-making strategies. In fact, by explaining their decision-making processes, XAI techniques can help AI researchers better understand the mechanisms behind AI outputs, allowing them to identify errors in their design and/or implementation. Accordingly, in this paper, we explore several well-known XAI methods typically used for Machine Learning (ML) classification tasks to see if they can be exploited both to provide explanations and to improve the model itself. The single explanation of a given ML output may not be enough to improve the ML system: the ML researcher may not be able to use the explanation directly due to the complexity of the ML system (for example, a Deep Neural Network). It would be desirable to have an automatic process that uses explanations of the ML system behaviour to improve the ML system itself automatically, or ideally, the ML system should be able to improve itself in a feedback loop fashion using the explanations provided (see Fig. 1).

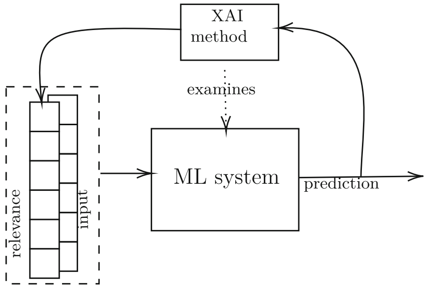

Fig. 1. General functional scheme of a Machine Learning architecture able to select/transform relevant input features relying on explanations provided by an XAI system.

The basic underlying idea is that explanations about the model outputs can help tune the ML system parameters better. In general, an explanation explains why an ML model returns a result given a specific input. However, building an explanation is particularly challenging if the model to inspect is a DNN, 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 humanly understandable. However, we start from the simpler hypothesis that the knowledge given by the explanation can be used to understand the model's strengths and weaknesses, changing the ML model to adapt itself to different inputs better. In the XAI context, explanations are built using ML system behaviour to understand its input-output relationships. Therefore, explanations of an ML system can be used to identify the critical characteristics of the input that caused a given output and thus use this knowledge to adapt/modify the ML system itself.

Thus, this paper aims to investigate whether the relevant features highlighted by an XAI method can be used with the input data to improve the classification performance of an ML system. However, we also experimentally evaluate which XAI methods can effectively highlight the most relevant features for our goal, as the performance of existing XAI methods may depend on the specific problem.

2 Related Works

The internal workings of Modern ML approaches, such as Deep learning, are typically opaque, letting to the AI scientist's ability to grasp the underlying processes that drive their behaviours. Subsequently, comprehending the connections between outputs and inputs can be extremely difficult. The use of XAI methods is becoming increasingly prominent for explaining various classification systems that rely on multiple inputs, including, but not limited to, images [1,12,16], natural language processing [10,14], clinical decision support systems [20], and others. However, using XAI methods to improve the performance of ML models in classification problems is relatively scarce in current research. In [27], a survey of the works leveraging XAI methods to improve classification systems is reported. The authors of [7] provide a general framework to train a model both with data and explanations with the aim of not only to get to the correct answer, but also to provide a correct explanation. In general, the importance of involving the explanations in the ML pipeline has gaining attention in literature. For instance, in [11] a dataset for a hate Speech detection including rationales about their labels is described. In [19], a first study is proposed to use the relevance produced by Deep Taylor Decomposition [13] to build a reliable classifier to build a system able to detect the presence of orca whales in hydrophone recordings. The relevance is used as a binary mask to select the most relevant input features. Differently from [19], our study focuses on image classification tasks on publicly available datasets, selecting an XAI strategy to build the relevance mask by a preliminary study on a family of XAI methods available in the literature. In [17], the training loss function is constrained to lead the classifier to focus only on a prior-defined set of features. Similarly, in [25], LRP explanations [12] are exploited to lead the training stage of an ML model to emphasize the important features of a classification task. In [21], eXplanatory interactive Learning (XiL) is proposed. XiL is a mechanism consisting of interactively querying the user (or some other information source) during the training stage to obtain the desired outputs of the data points. In particular, the model considers an input and predicts a label together with an explanation of its prediction. If necessary, the user responds by correcting the learner and providing improved (but not necessarily optimal) feedback to the model during its training.

In the biomedical field, [9,22] attempted to enhance the models' abilities to select features by utilizing Correlation-based Feature Selection and Chaotic Spider Monkey Optimization methods on biomedical data. Additionally, an occlusion sensitivity analysis technique [29] is suggested in [8] to identify the most pertinent cortical regions for a motor imagery task. The usage of XAI methods to interpret the outputs of Epilepsy Detection systems is also explored in [15]. In [3,4] an experimental analysis of several well-known XAI methods applied on an ML system trained on EEG data was carried out, showing that many components considered relevant by XAI methods are shared across the signal and can be potentially used to build a system able to generalize better. Instead, the main goal of the current study is to analyze the effectiveness of a set of selected XAI methods in improving the performance of a machine learning system for an image classification task. Additionally, this study explores various approaches to combining input and explanation to optimize the ML system's performance.

3 Method

We conducted a series of experiments with the following goals: i) testing the capability of a set of well-known selected XAI methods to provide information able to effectively improve the ML system performance in an image classification task on different datasets; ii) testing several strategies to combine input and explanation for improving the ML system performance.

i) Evaluating explanation methods

The following XAI methods have been tested and evaluated to detect an explanation method able to enhance the model performance positively: Saliency [23], Guided BackPropagation [24], and Integrated Gradients [26]. The explanations provided by these methods are evaluated by computing MoRF (Most Relevant First) curve and LeRF (Least Relevant First) curves, proposed in [5,18]. The MoRF curve is computed as follows: given a classifier, an input $\mathbf{x}$ and the respective classification output $C(\mathbf{x})$, the input features are iteratively replaced by zeros, following the descending order with respect to the relevance values returned by the explanation method. Therefore, the expected MoRF curve is such that the more relevant the identified components are for the classification output, the steepest the curve. Conversely, LeRF curves are built iteratively, removing the input features following the ascending order with respect to the relevance values returned by the explanation method. Consequently, we expect the classification output to be very close to the original value in the first iterations (corresponding to less relevant features removed), dropping quickly to zero as the process goes toward. While the MoRFs report how much the classifier output is altered by removing highly relevant components, LeRFs report how much the least relevant components leave the output intact. In the following of this subsection, the investigated XAI methods are described briefly.

Saliency: The saliency method [23] is a straightforward and intuitive way to explain a machine learning (ML) system. Originally presented in [23], Essentially, an explanation of the ML system's output $C(\mathbf{x})$ for an input $\mathbf{x} \in \mathbb{R}^d$ is created by generating a saliency map using the gradient $\frac{\partial C}{\partial \mathbf{x}}$. The gradient's magnitude indicates how much the features must be adjusted to impact the class score.

Guided BackPropagation: Guided BackPropagation (Guided BP) [24] is a method similar to the Saliency one, with the main difference being that in Guided BP, a gradient transformation is used preventing the backward flow of negative values, rather than using the real gradient. This method starts from the assumption that negative values may decrease the neuron activations and are not considered as important by the user. The main drawback is that it can failure to highlight inputs that negatively contribute to the output.

Integrated Gradients: [26] proposed an approach involving the average of all gradients between the original input $\mathbf{x}$ and a baseline input $\mathbf{x}^{ref}$, where $C(\mathbf{x}^{ref})$ results in a neutral prediction. This method, known as Integrated Gradient (IG), takes into account the magnitude of gradients of features of inputs closer to the baseline. The importance of each feature x_i is computed aggregating the gradients along the intermediate inputs on the straight-line between the baseline and the input by changing α over the range $[0, 1]$.

ii) Merging schemes

This work aims to propose a valid method to exploit an XAI explanation to improve the results of a classifier. However, it is important to highlight that we start from the assumption that, for a given input, an explanation of the model's output with respect to the correct target class is available. In real scenarios, where the correct class is not available for new input, this assumption is unrealistic. Despite this, this assumption is adopted to effectively explore the the improvement of classification performance exploiting the explanations. In other words, we try to answer the question, "If the explanation on how an ideal model should behave when fed with a given input, could it help the actual classifier?".

We propose two possible strategies to merge the explanations into the classification process: *binary mask* and *soft-masking* schema. These strategies are described in the following two sub-paragraphs.

Binary Mask Strategy: Similarly to [19], the first strategy starts from the assumption that the explanation's scores can be considered as a measurement of the "attention" that the model has to give to each feature to produce a certain output. In particular, given an input $\mathbf{x} \in \mathbf{R}^d$, and an explanation in terms of input relevance map $A(\mathbf{x}, C) \in \mathbf{R}^d$ for the output $C(\mathbf{x})$ of a classifier C, we use the following simple rule to construct a mask M:

$$M_i = \begin{cases} 1 & \text{if } A(\mathbf{x}, C)_i > 0, \\ 0 & \text{otherwise.} \end{cases}$$

In other words, the goal is fed the model C with only the features which contribute positively to the output. The aim is to understand if the feature highlighted by an explanation can actually lead the model toward the correct classification. Therefore, a masked version of the input $M * \mathbf{x}$, where $*$ is the dot-wise product, can be fed to the model C. Differently to [19], our study focuses on image classification tasks on well-known dataset, selecting as XAI strategy to build the relevance mask by a preliminary study on a family of XAI methods available in literature.

Soft-Masking Strategy: In the previous schema, features having negative relevance scores are removed from the input of the classifier. However, negative scores can be a source of information which could lead the classifier toward the right response, as well as the positive ones. The problem is how to integrate this kind of information into the input. Instead of defining a prior given merging rule, we consider to delegate a ML model to find the best one. In other word, we delegate the model to merge together relevance $A(\mathbf{x}, C)$ and the input $\mathbf{x}$. To this aim, a supplementary mixer network to merge together $\mathbf{x}$ and $A(\mathbf{x}, C)$ is adopted, connected to the classifier C as shown in Fig. 2. From now on, this network is called *Mixer*. We adopt two further networks, $E_{\mathbf{x}}$ and E_A to reduce the dimensionality of $E_{\mathbf{x}}$ and $A(\mathbf{x}, C)$ respectively. The $E_{\mathbf{x}}$ and E_A are then concatenated and fed to the Mixer. The resulting Mixer output can be considered as an input mask M and used for weighting the C input $\mathbf{x}$. Mixer, $E_{\mathbf{x}}$, and E_A can be learned freezing the C parameters and using classical training procedure on the remaining ones, corresponding to search for the best Mixer, $E_{\mathbf{x}}$, and E_A parameters able to reduce and join together $A(\mathbf{x}, C)$ and $\mathbf{x}$, for a given classifier C.

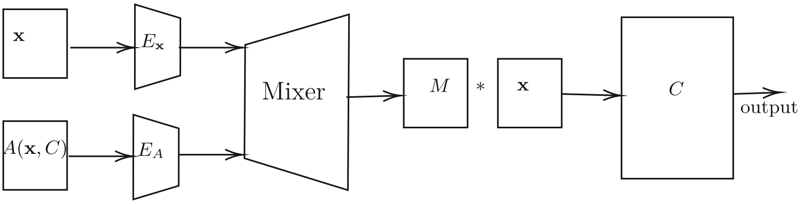

Fig. 2. Architecture of the soft-masking schema

4 Experimental Assessment

Fashion-MNIST, CIFAR10 and STL10 datasets were used as benchmark dataset. The Fashion-MNIST dataset contains images depicting various fashion articles [28]. It shares the same image size and training/testing splits of MNIST dataset. The dataset contains 60,000 training images and 10,000 test images, each of size 28×28 and grayscale in nature. There are ten classes in the dataset, including T-shirt/top, Trouser, Pullover, Dress, Coat, Sandal, Shirt, Sneaker, Bag, and Ankle boot.

CIFAR-10 is a collection of 60,000 color images grouped into ten categories, that are airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck. The dataset offers 50,000 training images and 10,000 test images, all of size 32×32.

The STL10 dataset consists of images belonging to ten different classes, that are airplane, bird, car, cat, deer, dog, horse, monkey, ship, and truck. Each image has a size of 96 × 96 pixels.

As classifier C, we adopt a ResNet18 [6] pre-trained on ImageNet dataset for the CIFAR10 and STL10 dataset, and a custom model composed of two fully-connected layers with ReLU activation function for Fashion-MNIST dataset.

A first experiment to compute the baseline consisting in the fine tuning of C using the training set provided in each investigated dataset was carried out. Baseline models are then used to build the explanations of a model prediction for each input. Therefore, the produced explanations were evaluated computing the MoRF and the LeRF curves, as described in Sect. 3. The best explanation method found was used to evaluate the proposed merging schemes.

For the experiments involving binary masking scheme, the following two learning strategies have been used:

A. **fine tuning on masked data**: a fine tuning of C was made using only masked training data, as discussed in Sect. 3;
B. **fine tuning on both original and masked data**: a two-step fine tuning procedure was adopted, the former on the unaltered training data provided by the evaluated datasets, the latter on the masked training data.

Instead, in the soft-masking case, we adopt as $E_{\mathbf{x}}$ and E_A as two networks composed of 5 fully-connected layers equipped with ReLU activation function interspersed by Batch Normalization for experiments involving CIFAR10 and STL10, and of 2 fully-connected layers equipped with ReLU activation function interspersed by Batch Normalization for experiments involving Fashion-MNIST. Further details about the modules are reported in Table 1. The training consisted in two steps: firstly, a fine tuning of C was made using training data without any change on it. Secondly, the Mixer network, $E_{\mathbf{X}}$, and E_A are trained on the same data freezing the C parameters. The training was made with the Adam optimization algorithm. Best batch size and learning rate were found with a grid-search approach, testing batch sizes $\{64, 128, 256\}$ and learning rates in range $[0.001, 0.01]$ with step of 0.02. A validation set of 30% of the training data was adopted to stop the iterative learning process, with a maximum number of 100 iterations.

5 Results

In this section, results of the experimental assessments are reported.

5.1 Explanation Methods

In Fig. 3 the average MoRF and LeRF curves computed on the explanations obtained on Fashion-MNIST, CIFAR10, and STL10 test sets using the network models trained with the respective training sets are shown. Regarding the

Table 1. Architectures of the modules used. The numbers indicate how many neurons are employed in each fully-connected layer. The C module adopted for CIFAR10 and STL10 was a ResNet18 pretrained on ImageNet.

STL10 $E_{\mathbf{x}}, E_A$	Mixer	CIFAR10 $E_{\mathbf{x}}, E_A$	Mixer	F-Mnist $E_{\mathbf{x}}, E_A$	Mixer	C
4096	512	2048	512	512	512	128
batch norm.	batch norm.	batch norm.	batch norm.	batch norm.	batch norm.	
ReLU	ReLU	ReLU	ReLU	ReLU	ReLU	ReLU
2048	1024	1024	1024	256	784	64
batch norm.	batch norm.	batch norm.		batch norm.		
ReLU	ReLU	ReLU		ReLU		ReLU
1024	4096	512		128		10
batch norm.	batch norm.	batch norm.				
ReLU	ReLU	ReLU				
512	9216	256				
batch norm.		batch norm.				
ReLU		ReLU				
256		128				
batch norm.						
ReLU						
128						

Fashion-MNIST dataset, all the investigating methods produced good explanations respect to the MoRF and LeRF curves. This can be due to the simplicity of Fashion-MNIST dataset, leading the explanation methods to extract the actual real features in an easy way. Instead, for CIFAR10 dataset, it is easy to see that both MoRF and LeRF curves have the expected behavior only with Integrated Gradient. Indeed, it is evident that the Integrated Gradient MoRF curve quickly decreases toward zero, indicating that removing features reported as relevant by Integrated Gradient leads to a decrease in accuracy. On the other side, Integrated Gradient LeRF curve slowly tends to zero, indicating that removing features reported as not many relevant by the XAI method does not change the performance so much. Instead, in the Guided BackPropagation and Saliency method this behavior is not present, both for MoRF and LeRF curves. For STL10, Guided BackPropagation method produces poor explanations, accordingly with STL10 and CIFAR10 case. Integrated Gradient and Saliency produce similar results, but also in this case MoRF and LeRF curves are better in the former case respect to the latter. In conclusion, among the analyzed XAI methods, Integrated Gradient results the method providing the most reliable explanations among the analyzed datasets. Therefore, in the experiments dedicated to test how to merge input and explanations we adopted the Integrated Gradients method to build the explanations.

5.2 Merging Schema: Binary Masking

In Table 2 the results adopting the binary masks are reported for all the investigated datasets. Performance on the original-data fine-tuned model (baseline in the table) are reported on the original test set, differently from the masked-data

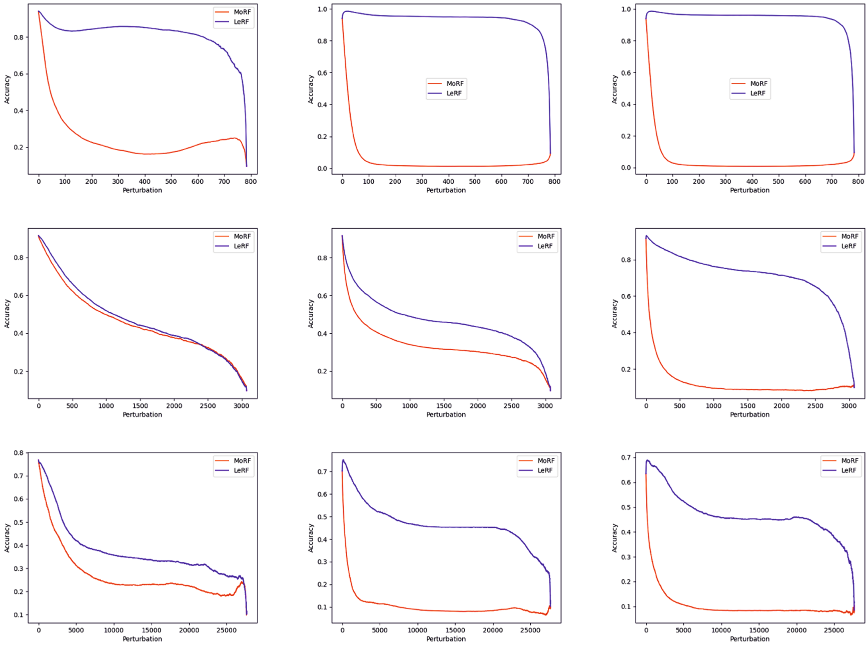

Fig. 3. Quantitative evaluation of the models' explanations on Fashion-MNIST (first row), CIFAR10 (second row) and STL10 (third row) test sets using Guided Back-Propagation (first column), Saliency (second column), and Integrated Gradient (third column). In each plot, MoRF (red) and LeRF (blue) curves are shown. (Color figure online)

fine-tuned models (cases A and B, described in Sect. 3). It is easy to see that the adoption of the relevance as masks can positively affect the model performance, leading the model toward an increment up to 10% points of accuracy in the case of CIFAR10 respect to the baseline and Fashion-MNIST, and 36% in case of STL10. In addition, it is interesting to notice that the data used during the fine-tuning stage, also if different from the data involved in the test stage, can affect the results. Interestingly, it seems that fine-tuning stage on the original

Table 2. Accuracy scores on test set using relevance as binary masks on CIFAR10, STL10 and Fashion MNIST test sets

Model	CIFAR10	STL10	F-MNIST
baseline	85.7%	66.3%	87.3%
binary masking (A)	90.9%	97.6%	97.4%
binary masking (B)	95.8%	98.2%	97.8%

data before of the masked one (case B) leads toward an improvement in the results respect to fine-tuned model using only the masked data (case A).

5.3 Merging Schema: Soft Masking

In Table 3 the results adopting the soft-masking schema are reported. Also in this case, the proposed strategies lead to an improvement in accuracy in all the three datasets. However, except that in Fashion-MNIST case, the improvement is lower respect to the binary-masking case. This can be due to several factors, such as a possible information loss due to the dimensionality reduction applied by $E_{\mathbf{x}}$ and E_A networks, or by a non optimal architecture of the Mixer network. However, the obtained results suggest that there is room for improvements adopting more appropriate Mixer and $E_{\mathbf{x}}, E_{\mathbf{A}}$ architectures.

Table 3. Accuracy scores on test set using relevance as soft masks on CIFAR10, STL10 and Fashion MNIST test sets

Model	CIFAR10	STL10	F-MNIST
baseline	85.7%	66.3%	87.3%
soft-masking	87.6%	68.6%	99.9%

6 Conclusions

This work reports an empirical analysis of three XAI techniques on the effectiveness of explanations for an image classification problem on three well-known datasets. Next, two strategies to merge explanations and input data to enhance the model's classification performance are provided. The former strategy consists of binary masking criteria to select the input features; the latter consists of letting the model find the better mixing strategy through a learning strategy. The results are promising in both cases, especially in the binary mask case. However, all the results are obtained under the hypothesis that the explanations on the right classes are available for the test data. This is an unrealistic hypothesis, since the right class of the testing data is unknown. Thus, the results of this research work can be exploited to improve the performance of a classifier by building a system capable of giving a good approximation of the explanations even in the test phase. We plan to pursue this line of research in our future works.

Acknowledgements. 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 was partially supported 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, and by the Ministry of Economic Development, "INtegrated Technologies and ENhanced SEnsing for cognition and rehabilitation" (INTENSE) project. Furthermore, we acknowledge financial support from the PNRR MUR project PE0000013-FAIR (CUP: E63C22002150007).

References

1. Apicella, A., Isgrò, F., Prevete, R., Sorrentino, A., Tamburrini, G.: Explaining classification systems using sparse dictionaries. In: ESANN 2019 - Proceedings, 27th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, pp. 495–500 (2019)
2. Apicella, A., Giugliano, S., Isgrò, F., Prevete, R.: Exploiting auto-encoders and segmentation methods for middle-level explanations of image classification systems. Knowl.-Based Syst. **255**, 109725 (2022)
3. Apicella, A., Isgrò, F., Pollastro, A., Prevete, R.: Toward the application of XAI methods in EEG-based systems. In: Proceedings of the 3rd Italian Workshop on Explainable Artificial Intelligence co-located with 21th International Conference of the Italian Association for Artificial Intelligence(AIxIA 2022), Udine, Italy, 28 November–3 December 2022. CEUR Workshop Proceedings, vol. 3277, pp. 1–15. CEUR-WS.org (2022)
4. Apicella, A., Isgrò, F., Prevete, R.: XAI approach for addressing the dataset shift problem: BCI as a case study (short paper). In: Proceedings of 1st Workshop on Bias, Ethical AI, Explainability and the Role of Logic and Logic Programming (BEWARE 2022) co-located with the 21th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2022), Udine, Italy, 2 December 2022. CEUR Workshop Proceedings, vol. 3319, pp. 83–88 (2022)
5. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE **10**(7), e0130140 (2015)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
7. Hind, M., et al.: TED: teaching AI to explain its decisions. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pp. 123–129 (2019)
8. Ieracitano, C., Mammone, N., Hussain, A., Morabito, F.C.: A novel explainable machine learning approach for EEG-based brain-computer interface systems. Neural Comput. Appl. **34**(14), 11347–11360 (2022)
9. Laxmi Lydia, E., Anupama, C.S.S., Sharmili, N.: Modeling of explainable artificial intelligence with correlation-based feature selection approach for biomedical data analysis. In: Khamparia, A., Gupta, D., Khanna, A., Balas, V.E. (eds.) Biomedical Data Analysis and Processing Using Explainable (XAI) and Responsive Artificial Intelligence (RAI). ISRL, vol. 222, pp. 17–32. Springer, Singapore (2022). https://doi.org/10.1007/978-981-19-1476-8_2
10. Lei, T., Barzilay, R., Jaakkola, T.: Rationalizing neural predictions. arXiv preprint arXiv:1606.04155 (2016)
11. Mathew, B., Saha, P., Yimam, S.M., Biemann, C., Goyal, P., Mukherjee, A.: HateXplain: a benchmark dataset for explainable hate speech detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 14867–14875 (2021)

12. Montavon, G., Binder, A., Lapuschkin, S., Samek, W., Müller, K.-R.: Layer-wise relevance propagation: an overview. In: Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., Müller, K.-R. (eds.) Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. LNCS (LNAI), vol. 11700, pp. 193–209. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-28954-6_10
13. Montavon, G., Lapuschkin, S., Binder, A., Samek, W., Müller, K.R.: Explaining nonlinear classification decisions with deep Taylor decomposition. Pattern Recogn. **65**, 211–222 (2017)
14. Qian, K., et al.: XNLP: A living survey for XAI research in natural language processing. In: 26th International Conference on Intelligent User Interfaces-Companion, pp. 78–80 (2021)
15. Rathod, P., Naik, S.: 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), pp. 1–6. IEEE (2022)
16. Ribeiro, M.T., Singh, S., Guestrin, C.: "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, pp. 1135–1144 (2016)
17. Ross, A.S., Hughes, M.C., Doshi-Velez, F.: Right for the right reasons: training differentiable models by constraining their explanations. arXiv preprint arXiv:1703.03717 (2017)
18. Samek, W., Binder, A., Montavon, G., Lapuschkin, S., Müller, K.R.: Evaluating the visualization of what a deep neural network has learned. IEEE Trans. Neural Netw. Learn. Syst. **28**(11), 2660–2673 (2016)
19. Schiller, D., Huber, T., Lingenfelser, F., Dietz, M., Seiderer, A., André, E.: Relevance-based feature masking: improving neural network based whale classification through explainable artificial intelligence (2019)
20. Schoonderwoerd, T.A., Jorritsma, W., Neerincx, M.A., Van Den Bosch, K.: Human-centered XAI: developing design patterns for explanations of clinical decision support systems. Int. J. Hum Comput Stud. **154**, 102684 (2021)
21. Schramowski, P., et al.: Making deep neural networks right for the right scientific reasons by interacting with their explanations. Nat. Mach. Intell. **2**(8), 476–486 (2020)
22. Selvam, R.P., Oliver, A.S., Mohan, V., Prakash, N.B., Jayasankar, T.: Explainable artificial intelligence with metaheuristic feature selection technique for biomedical data classification. In: Khamparia, A., Gupta, D., Khanna, A., Balas, V.E. (eds.) Biomedical Data Analysis and Processing Using Explainable (XAI) and Responsive Artificial Intelligence (RAI). ISRL, vol. 222, pp. 43–57. Springer, Singapore (2022). https://doi.org/10.1007/978-981-19-1476-8_4
23. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)
24. Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for simplicity: the all convolutional net. arXiv preprint arXiv:1412.6806 (2014)
25. Sun, J., Lapuschkin, S., Samek, W., Zhao, Y., Cheung, N.M., Binder, A.: Explanation-guided training for cross-domain few-shot classification. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 7609–7616. IEEE (2021)
26. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: International Conference on Machine Learning, pp. 3319–3328. PMLR (2017)

27. Weber, L., Lapuschkin, S., Binder, A., Samek, W.: Beyond explaining: opportunities and challenges of XAI-based model improvement. Inf. Fusion (2022)
28. Xiao, H., Rasul, K., Vollgraf, R.: Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747 (2017)
29. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8689, pp. 818–833. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10590-1_53