SHAP-based explanations to improve classification systems

Andrea Apicella[1,2,3,*,†], Salvatore Giugliano[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

Explainable Artificial Intelligence (XAI) is a field usually dedicated to offering insights into the decision-making mechanisms of AI models. Its purpose is to enable users to comprehend the reasoning behind the results provided by these models, going beyond mere outputs. In addition, one of the main goals of XAI is to improve the performance of AI models by exploiting the explanations of their decision-making processes. However, a predominant portion of XAI research concentrates on elucidating the functioning of AI systems, with comparatively fewer studies delving into how XAI techniques can be leveraged to enhance the performance of an AI system. This underlines a potential area for further exploration and development in the field of XAI. In this paper we focus on the possibility to enhance the performance of an already trained AI model. To this aim we propose a new scheme of interaction between explanations provided by SHAP XAI method and computations of the responses of a given AI model. This new proposal was tested using the well-known CIFAR-10 dataset and EfficientNet-B2 model, showing promising results.

Keywords

XAI, Machine Learning, DNN, SHAP, 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. A large part of the past and current literature [1, 2, 3, 4, 5] focuses solely on how to explain AI systems, while less attention is paid to whether and how current XAI methods can be used to improve an AI system. This is a significant shortcoming in the context of such research studies.

"XAI.it 2023 - Italian Workshop on Explainable Artificial Intelligence", editors "Cataldo Musto, Riccardo Guidotti, Anna Monreale, Erasmo Purificato, Giovanni Semeraro"

[*] Corresponding author.

[†] These authors contributed equally.

andrea.apicella@unina.it (A. Apicella); salvatore.giugliano2@unina.it (S. Giugliano); francesco.isgro@unina.it (F. Isgrò); rprevete@unina.it (R. Prevete)

0000-0002-5391-168X (A. Apicella); 0000-0002-1791-6416 (S. Giugliano); 0000-0001-9342-5291 (F. Isgrò); 0000-0002-3804-1719 (R. Prevete)

CEUR Workshop Proceedings (CEUR-WS.org)

For example, 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.

Thus, our goal is to establish an automated process by which the explanations of the ML system's behaviour are used to improve the system's performance.

The core idea is that explanations can facilitate the discernment of pivotal input features influencing specific outputs. This acquired knowledge can then be applied to fine-tune or refine the ML system itself. In particular, we focus on the possibility to enhance an already trained ML system with slight changes to the system itself. In this work, we leverage the SHAP (SHapley Additive exPlanations, [6]) method to devise a system aimed at enhancing the performance of a classifier. SHAP, a widely recognized explainability technique, provides insights into the contribution of different features in a model's predictions. Our approach employs a surrogate of the explanations, obtained through an encoder, with the specific intention of distilling only the truly pertinent information conveyed by the explanation. This allows us to focus on the most important explanatory insights, while ensuring the effectiveness and efficiency of our framework for improving classifiers.

Two experimental setups were designed to assess the effectiveness of our approach. The first one aims to investigate whether SHAP provides explanations with informative content significant enough to enhance the model's performance. In other words, we propose a strategy to experimentally find an upper bound on the capability of the SHAP explanations to improve the system performance in a given problem. In the second experimental setup, we show how the SHAP explanations can be effectively incorporated into the processing of the ML system responses in order to improve them. In this work, we will provide the following contributions: i) We consider a model already trained, aiming to introduce minimal modifications while leveraging its existing feature extraction ability, ii) we propose a simple strategy to experimentally compute the extent to which explanations can improve the already trained model, and iii) we present a straightforward methodology for seamlessly integrating explanations with the output of a pre-trained model, offering a practical and effective framework for model improvement.

The paper is organized as follows: in Sec. 2 the current literature about XAI used to improve ML models is reported; in Sec. 3 the proposed method is described while in Sec. 4 and 5 experimental assessment and the obtained results are described and reported, respectively. Finally, in Sec. 6 we conclude the paper with final remarks.

2. Related works

The internal mechanisms of modern ML approaches, such as Deep Learning, are typically opaque, posing challenges for AI scientists attempting to comprehend the underlying processes governing their behaviors. Consequently, establishing a clear understanding of the relationships between inputs and outputs can prove to be challenging. As a result, the utilization of eXplainable AI (XAI) methods span various domains, including but not limited to images [1, 2, 5, 7, 8], natural language processing [9, 10], clinical decision support systems [11], and more. The integration of explanations into the machine learning pipeline has garnered significant attention in very recent years within the academic literature in several fields. In a related vein, [12] introduces a dataset for hate speech detection that incorporates rationales explaining the labels assigned. Another notable contribution by [13] presents a text augmentation technique tailored for Natural Language Processing (NLP) tasks. This innovative method leverages XAI techniques to discern the significance of individual words. For example, In [14, 15] several XAI methods are empirically evaluated on an ML system trained on EEG data for BCI applications, reporting that many components considered relevant by XAI methods can be potentially used to improve a ML system. More in general, Weber et al. conduct a comprehensive survey in [16] that explores various instances of eXplainable AI (XAI) methodologies employed to enhance classification systems. The work of [17] introduces a framework wherein both data and explanations are utilized jointly to train a machine learning model, equipping it to provide both interpretable explanations and a predictive model. [18] employed Deep Taylor Decomposition (DTD) relevance [4] to construct a robust classifier for identifying the presence of orca whales in hydrophone recordings. The DTD relevance serves as a binary mask, enabling the selection of the most crucial input features. Similarly, in [19] and [20], LRP (Layer-wise Relevance Propagation) explanations [5] guided the training process of a classification task, ensuring emphasis on the salient features. LRP is also harnessed in [21], where the explanations provided are manually scrutinized to eliminate extraneous information from the dataset, enhancing the overall model performance. [22] embedded Contextual Decomposition (CD) explanations [23] directly into the training loss function of a DNN model. [24] proposes Attention Branch Network (ABN), a model designed to leverage on CAM-based explanations to weight the input feature maps introducing a branch structure. In [25] a set of well-known XAI methods are investigated to verify if they can be exploited to improve the performance of a classification system. In [26, 27, 28] different feature priors schemes are integrated as penalty term into the training loss functions. It is important to highlight that these methods often require a human expert with prior knowledge to build a domain-specific prior knowledge. Similarly, [29] and [30] implement a classifier that focuses exclusively on a predefined set of features, thereby constraining the training loss function.

In [31] a retraining strategy to improve the model predictions leveraging on SHapley Additive exPlanations (SHAP) values to give specific training weights to misclassified data samples is proposed. The effectiveness of SHAP values is also investigated in [32] where SHAP explanations are used to implement a feature selection strategy. SHAP was also used in [33] to improve the performance of an autoencoder for a network anomaly detection problem. In [34] a feature augmentation method (Shapley Feature Augmentation, SFA) based on Shapley values is proposed. Differently from the proposed work, SFA augments features leveraging on both the model outputs and the Shapley explanations. Furthermore, the augmented feature vectors is composed of both the original features and all their Shapley values, doubling the feature dimensions.

3. Method

3.1. Notation

We assume that a dataset $D = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{n}$ of n labeled data is available, where $\mathbf{x}^{(i)}$ is a sample point in a given feature space X and $y^{(i)}$ is the corresponding class in a class label space $\{1, 2, \ldots, C\}$.

In this work, we adopt SHAP explanations to improve the performance of a given classifier

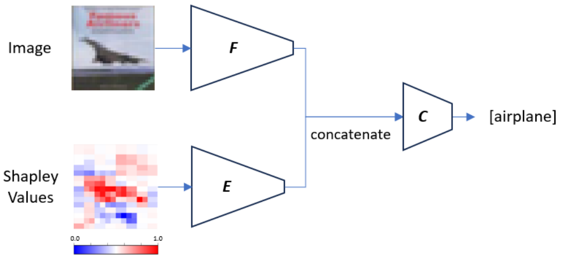

Figure 1: The architecture that takes two inputs: the image and its corresponding explanation. The first input is directed to the pre-trained feature extractor F, while the second input is processed by an explanation Encoder E designed to handle explanations based on Shapley values.

M on an input $\mathbf{x}^{(i)}$. The model M can be formalized as a function $M : X \rightarrow \mathbb{R}^C$. The model output can be viewed as a score function for each possible class, and the predicted class can be computed as $\hat{y}^{(i)} = \arg\max\{M(\mathbf{x}^{(i)})\}$. For each input $\mathbf{x}^{(i)}$ and output $\hat{y}^{(i)}$ by M, SHAP provides explanations about all the possible classes. In this work, we will refer to the explanation provided by SHAP on an input $\mathbf{x}^{(i)}$ respect to a generic class $c \in \{1, 2, \ldots, C\}$ as $e_{\mathbf{x}^{(i)},c}$, instead the SHAP explanation respect to the *inferred* class $\hat{y}^{(i)}$ provided by M will be indicated by $e_{\mathbf{x}^{(i)}}$, and finally we will refer to the SHAP explanation respect to the *real* target class $y^{(i)}$ as $e^*_{\mathbf{x}^{(i)}}$.

3.2. Method description

In our proposal, we consider a neural network model M as composed of a feature extractor F followed by a classification layer L, i.e. $M(\mathbf{x}^{(i)}) = L\big(F(\mathbf{x}^{(i)})\big)$. We want to exploit both the knowledge given by explanations $e_{\mathbf{x}^{(i)}}$ on the model encoding $M(\mathbf{x}^{(i)})$ together with features extracted by $F(\mathbf{x}^{(i)})$, to obtain an improved ML model $C\big(M(\mathbf{x}^{(i)}), F(\mathbf{x}^{(i)})\big)$ able to provide improved performance respect to M. The overall model is shown in Fig. 1. Summarizing, the proposed framework is composed of three main components: i) a feature extractor F, taken from a model M with a given performance that we want to improve, ii) an explanation Encoder E which extract the relevant information from the provided explanations, and iii) a final classifier C. The input $\mathbf{x}^{(i)}$ and explanations $e_{\mathbf{x}^{(i)}}$ are fed to F and E respectively. The resulting outputs $F(\mathbf{x}^{(i)})$ and $E(e_{\mathbf{x}^{(i)}})$ are then concatenated together and fed to C, which will provide the final classification. In other words, C will replace L in the classification step, taking into account information provided by SHAP explanations. Differently from similar work such as [24], in this work we adopt an encoded version of the SHAP explanations (which we will refer as SHAP *surrogate*) computed by an encoder E. Indeed, we assume that an encoder can easily extract only the useful information from a given explanation.

4. Experimental assessment

The main goal of this work is to improve the classification performance of a model exploiting SHAP explanations. To this aim, we want first to investigate if SHAP provides explanations containing enough and useful information to improve the classification performance of M.

Therefore, initially we investigate if and how much SHAP surrogates can help the classification of a given model. Once it has been established that SHAP surrogates are useful to improve classification performance, they can be used for the effective classification system. Therefore, the experimental assessment is composed of two sets of experiments, the former to investigate if SHAP explanations could be used as additional input to enhance effectiveness of the classification model, the latter to adopt them in a real scenario. Since the first experimental scenario is purely exploratory, we assume that the explanations $e^*_{\mathbf{x}^{(i)}}$ about the real class for each input $\mathbf{x}^{(i)}$ are available, while in the second experimental scenario we adopt the effective explanation $e^*_{\mathbf{x}^{(i)}}$ provided by the SHAP method, therefore assuming the real case where information about the effective input class may not be available.

4.1. CIFAR-10 dataset

CIFAR-10 [35] was used as benchmark dataset. 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.

Adopted model and training

As base model M we employed EfficientNet-B2 model [36], pre-trained on the ImageNet dataset [37]. The model M is then fine-tuned on CIFAR-10 training data, resulting in an accuracy of 97.15% on the test set. Since EfficientNet is trained on ImageNet with images having dimensions of 224×224, bicubic interpolation was employed on CIFAR-10 to scale images from the original resolution of 32×32 to 224×224. The best model hyperparameters have been found with a grid-search strategy, where hyperparameters and variation ranges are reported in Tab. 1.

The resulting baseline model was then used to generate SHAP explanations for a given input. SHAP explanations were generated using the method described in [6].

As explained in Sec. 3, the proposed framework (Fig. 1) is a double-branch neural network architecture, where the first branch consists in the pre-trained feature extractor F fed with an input $\mathbf{x}^{(i)}$, while the second branch consists in an encoder E, whose architecture is shown in Fig. 2. The encoder E processes a SHAP value-based explanation with the aim of capturing meaningful information and relations about both the visual content of the input image and the associated explanations. The outputs of F and E, that are the features extracted by $\mathbf{x}^{(i)}$ and the surrogate explanation respectively, are then concatenated together and fed to the final classifier C. In this work, we adopt as classifier C a simple shallow fully-connected neural network with a number of neurons equals to the number of classes. Keeping the inner parameters of F fixed, parameters of E and C are learned during the training stage. The best hyperparameters have been found adopting a grid-search strategy with ranges listed in Tab. 1.

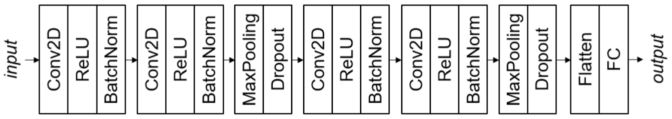

Figure 2: Explanation Encoder E which extracts the relevant information from the provided explanations.

Explanations

Two distinct types of experiments were conducted, the former to investigate if SHAP explanations could be used as additional input to enhance effectiveness of the classification model, therefore exploiting $e^*_{\mathbf{x}^{(i)}}$ about the true classes $y^{(i)}$, while the latter to adopt them in a real scenario, therefore no knowledge of the true class of the input was considered available. In the second case the following two strategies to build explanations have been adopted:

A. **Most probable class explanation** $e_{\mathbf{x}^{(i)}}$: the explanation $e_{\mathbf{x}^{(i)}}$ of the most probable $\hat{y}^{(i)}$ class according to M is generated.

B. **Weighted average explanation** $\overline{e}^{(i)}_{\mathbf{x}}$: An averaged explanation $\overline{e}^{(i)}_{\mathbf{x}}$ is computed from the set of C explanations generated by SHAP on the input $\mathbf{x}^{(i)}$ for the model M, with each explanation's contribution weighted by the classification scores provided by model M, as shown in Equation 1. $\mathbf{x}$ is the input image, C represents the total number of classes, $M(\mathbf{x}^{(i)})_c$ is the score assigned to class c by M, and $e_{\mathbf{x}^{(i)},c}$ is the SHAP values attributed to the class c on the input $\mathbf{x}^{(i)}$.

$$\overline{e}_{\mathbf{x}^{(i)}} = \sum_{c=1}^{C} M(\mathbf{x}^{(i)})_c \cdot e_{\mathbf{x}^{(i)},c} \tag{1}$$

Table 1
Variation ranges for the grid search optimization strategy for the models M, E and C.

Hyperparameter	Range
Batch Size	$\{32, 64, 128\}$
Learning Rate	$[0.0001, 0.01]$ with step of 0.0005
Validation Fraction	$\{0.05, 0.1, 0.2\}$

To ensure a sufficient level of granularity for superpixels, SHAP explanations were built with 2000 evaluation.

5. Results

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

In Table 2 performance accuracy on the test set of the CIFAR-10 dataset is shown. Specifically, the baseline model M, without additional information provided by SHAP, achieved an accuracy of 97.15%. Instead, using the adopted framework and explanations $e^*_{\mathbf{x}}$ (that are explanation on the correct classes), an higher accuracy of 98.64% was achieved. Differently, relying on the most probable class explanation $e_{\mathbf{x}}$ or the weighted average explanations $\overline{e}_{\mathbf{x}}$, accuracy drops to 97.06% and 97.21%, respectively. First results suggest that SHAP explanations on the model performance, in general, can have a significant impact, leading to performance improvements. Despite this, last experiments showed that the proposed approaches adopting effective explanations $e_{\mathbf{x}}$ lead to performance comparable with the baseline.

Figure 3 provides predictions for some of the images in the test data. It is relevant to note that the use of explanations corresponding to the real classes e^* leads to correct predictions in

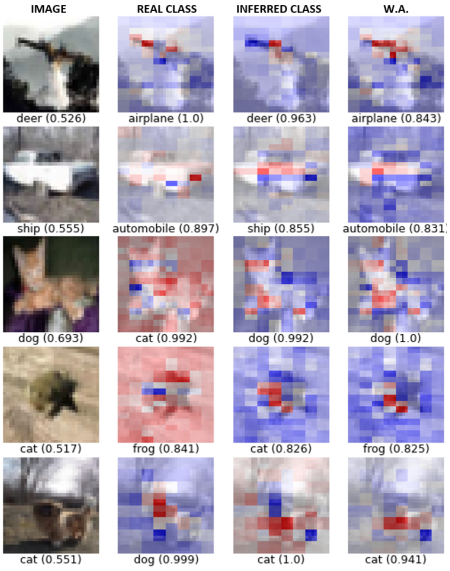

Figure 3: Images from the CIFAR-10 test set accompanied by predictions and scores from model M (first column), along with SHAP explanations displaying predicted classes and scores provided by C (second, third, and fourth columns). In greater detail, in the second column the explanations e^* on the model M together with the predicted class and score provided by C is reported. Similarly, explanations e considering the inferred class and explanations $\overline{e}$ considering the Weighted Average technique according to the M model are reported together with classes and scores given by C in third and fourth column, respectively.

without explanations (baseline)	with real class explanations e^*	with inferred class explanations e	with weighted average explanations $\overline{e}$
97.15	98.64	97.06	97.21

Table 2
Accuracy (%) scores on CIFAR-10 test set.

several cases, revealing a considerable information value in such explanations. In some cases, the weighted average technique is equally effective, also contributing to correct predictions.

6. Conclusions

Explainable Artificial Intelligence (XAI) plays a crucial role in unraveling the decision-making processes of AI models, shedding light on their inner workings and enhancing transparency. While the literature has predominantly focused on explaining AI systems, this work takes a step further to investigate how these explanations can be harnessed to improve AI system performance. In an ideal scenario, we envision an automated process where AI system explanations are leveraged to autonomously enhance system performance and understanding. Our assumption is that insights gained from explanations can enhance the adaptability of AI models to a wide range of inputs. In this study, we employ the SHAP (SHapley Additive exPlanations) method to develop a system aimed at improving classifier performance. SHAP, a widely recognized explainability technique, provides insights into the contribution of different features in a model's predictions. Our approach involves using a surrogate of SHAP explanations obtained through an encoder, distilling only the pertinent information from the explanation. This ensures a focus on crucial insights while maintaining the efficiency of our classifier improvement framework. We conduct two sets of experimental setups to assess the effectiveness of our approach. The first set aims to investigate whether SHAP explanations contain sufficient and useful information to enhance model performance. This setup primarily evaluates SHAP's explanatory power in isolation. In the second setup, we actively incorporate these explanations into our framework to demonstrate their real-world impact and effectiveness in enhancing classifier performance. This setup represents a practical application of SHAP insights. The results are encouraging. In the first experimental setup, when utilizing explanations corresponding to the true classes, we achieved the highest accuracy of 98.64%. This highlights the significant impact of explanations on model performance, especially when true class information is available. In real-world scenarios where this information is not available, we have shown that solutions can be found where the model performance is at least comparable to the baseline. This highlights the potential of explanations to improve the performance of AI systems. Further investigations to improve the proposed method will be made and proposed in future works. In conclusion, this work provides valuable insights into the potential of XAI methods, such as SHAP, to not only explain AI systems but also to enhance their functionality and adaptability. As AI continues to play a pivotal role in various domains, understanding and harnessing the power of explanations for improvement becomes increasingly important.

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 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, by the Ministry of Economic Development, “INtegrated Technologies and ENhanced SEnsing for cognition and rehabilitation” (INTENSE) project, and by Centro Nazionale HPC, Big Data e Quantum Computing (PNRR CN1 spoke 9 Digital Society & Smart Cities, CUP: E63C22000980007). Furthermore, we acknowledge financial support from the PNRR MUR project PE0000013-FAIR (CUP: E63C22002150007).

References

[1] 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.

[2] 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.

[3] 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.

[4] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, K.-R. Müller, Explaining nonlinear classification decisions with deep taylor decomposition, Pattern recognition 65 (2017) 211–222.

[5] 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.

[6] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30, Curran Associates, Inc., 2017, pp. 4765–4774. URL: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

[7] A. Apicella, S. Giugliano, F. Isgrò, R. Prevete, A general approach to compute the relevance of middle-level input features, in: Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part III, Springer, 2021, pp. 189–203.

[8] A. Apicella, S. Giugliano, F. Isgro, R. Prevete, et al., Explanations in terms of hierarchically organised middle level features, in: CEUR WORKSHOP PROCEEDINGS, volume 3014, CEUR-WS, 2021, pp. 44–57.

[9] 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.

[10] T. Lei, R. Barzilay, T. Jaakkola, Rationalizing neural predictions, arXiv preprint arXiv:1606.04155 (2016).

[11] 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.

[12] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, A. Mukherjee, Hatexplain: A benchmark dataset for explainable hate speech detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, 2021, pp. 14867–14875.

[13] S. Kwon, Y. Lee, Explainability-based mix-up approach for text data augmentation, ACM Transactions on Knowledge Discovery from Data 17 (2023) 1–14.

[14] A. Apicella, F. Isgrò, A. Pollastro, R. Prevete, 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, November 28 - December 3, 2022, volume 3277 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 1–15.

[15] A. Apicella, F. Isgrò, R. Prevete, 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, December 2, 2022, volume 3319 of *CEUR Workshop Proceedings*,

2022, pp. 83–88.

[16] L. Weber, S. Lapuschkin, A. Binder, W. Samek, Beyond explaining: Opportunities and challenges of xai-based model improvement, Information Fusion (2022).

[17] M. Hind, D. Wei, M. Campbell, N. C. Codella, A. Dhurandhar, A. Mojsilović, K. Natesan Ramamurthy, K. R. Varshney, Ted: Teaching ai to explain its decisions, in: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 2019, pp. 123–129.

[18] D. Schiller, T. Huber, F. Lingenfelser, M. Dietz, A. Seiderer, E. André, Relevance-based feature masking: Improving neural network based whale classification through explainable artificial intelligence (2019).

[19] J. ha Lee, I. hee Shin, S. gu Jeong, S.-I. Lee, M. Z. Zaheer, B.-S. Seo, Improvement in deep networks for optimization using explainable artificial intelligence, in: 2019 International Conference on Information and Communication Technology Convergence (ICTC), IEEE, 2019, pp. 525–530.

[20] J. Sun, S. Lapuschkin, W. Samek, Y. Zhao, N.-M. Cheung, A. Binder, Explanation-guided training for cross-domain few-shot classification, in: 2020 25th International Conference on Pattern Recognition (ICPR), IEEE, 2021, pp. 7609–7616.

[21] V. Bento, M. Kohler, P. Diaz, L. Mendoza, M. A. Pacheco, Improving deep learning performance by using explainable artificial intelligence (xai) approaches, Discover Artificial Intelligence 1 (2021) 1–11.

[22] L. Rieger, C. Singh, W. Murdoch, B. Yu, Interpretations are useful: penalizing explanations to align neural networks with prior knowledge, in: International conference on machine learning, PMLR, 2020, pp. 8116–8126.

[23] W. J. Murdoch, P. J. Liu, B. Yu, Beyond word importance: Contextual decomposition to extract interactions from lstms (2018).

[24] H. Fukui, T. Hirakawa, T. Yamashita, H. Fujiyoshi, Attention branch network: Learning of attention mechanism for visual explanation, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 10705–10714.

[25] A. Apicella, L. D. Lorenzo, F. Isgrò, A. Pollastro, R. Prevete, Strategies to exploit xai to improve classification systems, 2023. arXiv:2306.05801.

[26] G. Erion, J. D. Janizek, P. Sturmfels, S. M. Lundberg, S.-I. Lee, Learning explainable models using attribution priors (2019).

[27] F. Liu, B. Avci, Incorporating priors with feature attribution on text classification, arXiv preprint arXiv:1906.08286 (2019).

[28] E. Weinberger, J. Janizek, S.-I. Lee, Learning deep attribution priors based on prior knowledge, Advances in Neural Information Processing Systems 33 (2020) 14034–14045.

[29] A. S. Ross, M. C. Hughes, F. Doshi-Velez, Right for the right reasons: Training differentiable models by constraining their explanations, arXiv preprint arXiv:1703.03717 (2017).

[30] X. Shao, A. Skryagin, W. Stammer, P. Schramowski, K. Kersting, Right for better reasons: Training differentiable models by constraining their influence functions, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, 2021, pp. 9533–9540.

[31] H. Sun, L. Servadei, H. Feng, M. Stephan, A. Santra, R. Wille, Utilizing explainable ai for improving the performance of neural networks, in: 2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA), IEEE, 2022, pp. 1775–1782.

[32] W. E. Marcílio, D. M. Eler, From explanations to feature selection: assessing shap values as

feature selection mechanism, in: 2020 33rd SIBGRAPI conference on Graphics, Patterns and Images (SIBGRAPI), Ieee, 2020, pp. 340–347.

[33] K. Roshan, A. Zafar, Utilizing xai technique to improve autoencoder based model for computer network anomaly detection with shapley additive explanation (shap), arXiv preprint arXiv:2112.08442 (2021).

[34] L. Antwarg, C. Galed, N. Shimoni, L. Rokach, B. Shapira, Shapley-based feature augmentation, Information Fusion 96 (2023) 92–102.

[35] A. Krizhevsky, G. Hinton, et al., Learning multiple layers of features from tiny images (2009).

[36] M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: International conference on machine learning, PMLR, 2019, pp. 6105–6114.

[37] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee, 2009, pp. 248–255.