2024 IEEE International Symposium on Biomedical Imaging (ISBI) | 979-8-3503-1333-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISBI56570.2024.10635506

ON THE SUITABILITY OF B-COS NETWORKS FOR THE MEDICAL DOMAIN

Isabel Rio-Torto[1,2,], Tiago Gonçalves[1,3,*], Jaime S. Cardoso[1,3] and Luís F. Teixeira[1,3]*

[*]Equal contribution [1]INESC TEC, Campus da FEUP Rua Dr. Roberto Frias s/n, 4200-465 Porto, Portugal
[2]Departamento de Ciência de Computadores, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre s/n, 4169–007 Porto, Portugal
[3]Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias s/n, 4200-465 Porto, Portugal

ABSTRACT

In fields that rely on high-stakes decisions, such as medicine, interpretability plays a key role in promoting trust and facilitating the adoption of deep learning models by the clinical communities. In the medical image analysis domain, gradient-based class activation maps are the most widely used explanation methods and the field lacks a more in depth investigation into inherently interpretable models that focus on integrating knowledge that ensures the model is learning the correct rules. A new approach, B-cos networks, for increasing the interpretability of deep neural networks by inducing weight-input alignment during training showed promising results on natural image classification. In this work, we study the suitability of these B-cos networks to the medical domain by testing them on different use cases (skin lesions, diabetic retinopathy, cervical cytology, and chest X-rays) and conducting a thorough evaluation of several explanation quality assessment metrics. We find that, just like in natural image classification, B-cos explanations yield more localised maps, but it is not clear that they are better than other methods' explanations when considering more explanation properties.

***Index Terms*—** b-cos networks, explainable artificial intelligence, interpretability, medical image analysis

1. INTRODUCTION

Explainable artificial intelligence (xAI) methods aim to clarify the inner logic of machine learning models, thus promoting transparency, trust and accountability, and guaranteeing fairness [1]. Generally, these methods can be divided into three categories: *pre-model* (e.g. exploratory data analysis and visualization), *in-model* (e.g. design of architectures with domain constraints or regularization techniques [2]), and *post-model* (e.g. gradient/saliency/perturbation/attention-based methods that quantify the relevance of the features learned by the model [3, 4]) [5]. Although some authors argue that in high-stakes decision fields only inherently interpretable models might guarantee trust and transparency [6], saliency-based methods like GradCAM [3] are still among the most popular in the general domain and, consequently, in the medical domain [7], mainly because they are model-agnostic and do not require retraining the base models. However, given that the xAI field lacks a gold-standard metric for evaluating the quality of explanations, it is unclear if other methods might be better suited, especially to medical applications.

Recently, Böhle et al. [8] proposed the B-cos transform as a replacement of all linear transforms of a deep neural network, to induce an alignment between the network's weights and task-relevant input patterns. Since a sequence (network) of such transformations can be reduced to a single linear transformation, it faithfully summarizes the full model computations, and, thus, a B-cos network constitutes an inherently interpretable model. In this work, we investigate the application of B-cos networks to the medical domain. Our contributions are the following:

1. Application of the B-cos network to several medical use cases (skin lesions, diabetic retinopathy, cervical cytology, and chest X-rays) and multiple tasks (multiclass and multilabel classification), thus expanding the proposal of the original paper (i.e., binary classification on natural images);
2. Evaluation of a more complete set of properties of the generated visual explanations using an independent framework [9];
3. Integration of the B-cos layer into a state-of-the-art custom model for the medical domain [10];
4. Investigation of the effect of different pretraining strategies on the performance of B-cos networks.

The remainder of this paper is organized as follows: section 2 details the employed methodology, section 3 presents and discusses the obtained results, and section 4 concludes the paper and suggests future work directions. The code is available in a GitHub repository[1].

This work is financed by National Funds through the FCT - Fundação para a Ciência e a Tecnologia, I.P. (Portuguese Foundation for Science and Technology) within the project CAGING, with reference 2022.10486.PTDC (DOI 10.54499/2022.10486.PTDC), and through the Ph.D. Grants "2020.06434.BD" and "2020.07034.BD".

[1]https://github.com/icrto/medical-bcos

2. METHODOLOGY

2.1. B-cos Networks

Böhle et al. [8] introduced B-cos networks: the authors proposed the replacement of all the linear transforms of a deep neural network with the B-cos transform, arguing that this new module will optimize the network weights to align with task-relevant input patterns and inherently produce explanatory saliency maps given the completely linear nature of the model.

2.2. Explanation Quality Assessment Metrics

Recognizing the importance of a proper methodology to assess the quality of xAI methods, we decided to evaluate the approach proposed by Böhle et al. [8] with the framework developed by Hedström et al. [9]. It consists of a battery of quantitative measurements of different properties that a good explanation should have. We consider the following properties and metrics [9]:

- Faithfulness (do features deemed relevant affect model decisions more strongly?): Selectivity (drop in performance when features deemed relevant by a xAI method are iteratively removed) and Faithfulness Correlation (correlation between the sum of the subset of randomly replaced attributions and the difference in function output)
- Robustness (are explanations stable when subject to slight perturbations in the input?): Maximum Sensitivity (maximum sensitivity of an explanation using Monte Carlo sampling) and Relative Representation Stability (distance between the internal model representation of an explanation and its perturbed version)
- Complexity (are explanations concise?): Sparsity (Gini Index to measure if only highly attributed features are predictive of the model output) and Complexity (entropy of the contribution of all features to the total magnitude of the attribution)
- Localisation (is the explanatory evidence centered around a given region of interest?): Focus (quantifies the precision of the explanation by creating mosaics of images from different classes) and Attribution Localisation (ratio of positive attributions within a targeted object and the total positive attributions)

2.3. Pretraining Strategies

In the work by Böhle et al. [8], the authors compared B-cos architectures with their original counterparts, but only on ImageNet [11]. Thus, their experiments involved training both types of networks from scratch. In the medical imaging domain, it is common practice to employ transfer learning [12]. Therefore, we test different pretraining strategies for both B-cos and original networks: no pretraining, pretraining on ImageNet, and initializing B-cos networks from differently trained baselines.

2.4. B-cos Version of a Custom Medical Imaging Model

To understand the feasibility of integrating the B-cos transform in medical deep learning models, we adopted the work of Liu et al. [10] as a use case. Their Clinical-Inspired Network (CI-Net) is an attention-based model that aims to mimic the diagnostic process of clinicians, having achieved state-of-the-art performance on 6 benchmark databases for skin lesion analysis. Contrary to commonly used networks like the DenseNet [13], the CI-Net contains two branches and a distinguish module, which learns to discriminate images from different classes, making it more difficult to integrate the B-cos transform into this architecture.

2.5. Datasets

Given the diversity of use cases and tasks in the medical domain, we test B-cos networks on different tasks (multiclass and multilabel classification), and applications (skin lesions, diabetic retinopathy, cervical cytology, and chest X-rays). For skin lesion analysis, we used the ISIC2018 dataset [14, 15], containing 7 classes from melanoma to vascular lesions. For diabetic retinopathy (DR) classification, we used the APTOS2019 dataset [16], which includes images with 5 different severity scores, ranging from no DR to proliferative DR. For cervical cytology we used the DTU/Herlev Pap Smear Database [17], consisting of images of individual cervical cells in different stages of cervical cancer. For chest X-ray diagnosis (and multilabel classification), we resorted to the VinDrCXR [18], which contains 22 labels and corresponding bounding boxes for pathologies like pneumonia or cardiomegaly.

2.6. Implementation Details

All the models followed the training pipelines described by Böhle et al. [8], except for CI-Net-based models, which followed the training pipeline described by its authors [10]. Regarding data processing and augmentation, we resized all the images to 448×448, applied random affine transforms (i.e., rotation, translation, scale, and shear) according to specific literature on each use case (more details in the code), and normalized the inputs accordingly. All datasets were divided into stratified train-val-test splits with 20% for test for datasets without test set, and 10% of the remaining data for validation. For ISIC2018, we used the official split. We saved the best models according to the best F1-score on the validation set.

For the explanation quality metrics, we ensured that all models were being evaluated under the same conditions: models were evaluated only on the set of correctly predicted images by all models under test. For the multilabel classification scenario (VinDr-CXR), an image is considered correctly

classified only when all labels are correctly predicted, and we explain only the label with the lowest index. For datasets where more than 100 images are correctly predicted by all models, we evaluate only 100 (for Herlev we evaluated 40 images). More details can be found in the available code.

3. RESULTS AND DISCUSSION

We present our results on the 4 datasets considered in Table 1.

Training. Except for the CI-Net, we focus on DenseNet-121 [13], since it is a popular architecture among the medical imaging community and Böhle et al. [8] already showed that their conclusions hold for different architectures. One important aspect to highlight is that B-cos networks take twice as long to train and require about twice as much memory.

Pretraining strategies. For ISIC and APTOS, we study the effect of different pretraining strategies. In both cases, the performance improves on baseline and B-cos models when ImageNet pretraining is used (e.g. model 3 vs 6 or 4 vs 7). Interestingly, on APTOS2019, transferring knowledge from the baseline model already adapted to the final domain worsens the results compared to the B-cos network trained from scratch (models 13 and 16 vs 12), and, on both datasets, it falls short of the performance obtained when transferring knowledge from a B-cos network trained on another domain (models 5 and 8 vs 7, 13 and 16 vs 15). This shows that it is preferable to leverage knowledge from another B-cos network that already learned the weight-input alignment pressure, even though on another domain, than to learn from a non B-cos network (i.e. non-aligned) on the same domain.

Adapting a custom model. Transforming a custom model into a B-cos network is a straightforward process given the code made available by Böhle et al. [8]. As expected, this custom architecture outperforms DenseNet-121 on both baseline and B-cos versions in terms of Localisation. Interestingly, the baseline CI-Net is not better in terms of classification than the DenseNet-121, but the B-cos version achieves the best results (see models 9 and 10).

Explanation quality. We started by applying our evaluation protocol to ImageNet-trained networks made available by the original authors to verify if the results they reported held for different properties of explanations (in their work [8], only the localisation property is evaluated with a metric similar to Focus). In 4 out of 7 metrics, their conclusions hold, i.e. post-hoc methods applied on B-cos networks yield better explanations and, in turn, the inherent B-cos explanations surpass other methods both on B-cos and non B-cos networks. On the medical use cases tested, the results vary. For ISIC and APTOS, the trend in terms of Selectivity is reversed compared to the trend on ImageNet. For Herlev and VinDr-CXR, explanations obtained from B-cos networks are better than on the original networks. Still, the inherent B-cos explanations do not surpass other xAI methods (e.g. 4.56 for GradCAM vs 5.04 on VinDr-CXR). In terms of Faithfulness Correlation, B-cos explanations surpass other methods except on VinDr-CXR. Regarding robustness, B-cos networks and explanations are, in general, better for RRS but worse for Maximum Sensitivity. The same applies to Sparseness and Complexity. Finally, on Localisation metrics, B-cos explanations surpass the best explanation methods on B-cos and non B-cos models for 3 out of 4 datasets, but only for ImageNet pretrained models (c.f. models 6/7, 9/10, 14/15, 19/20). Further, on VinDr-CXR, where local bounding box annotations are available, the Attribution Localisation metric shows the trend the original authors found. These results indicate that B-cos explanations are, indeed, generally more focused than other explanations, which is also corroborated by the qualitative results of Fig. 1. Visually, Input×Gradient produces more sparse maps, while GradCAM and B-cos explanations

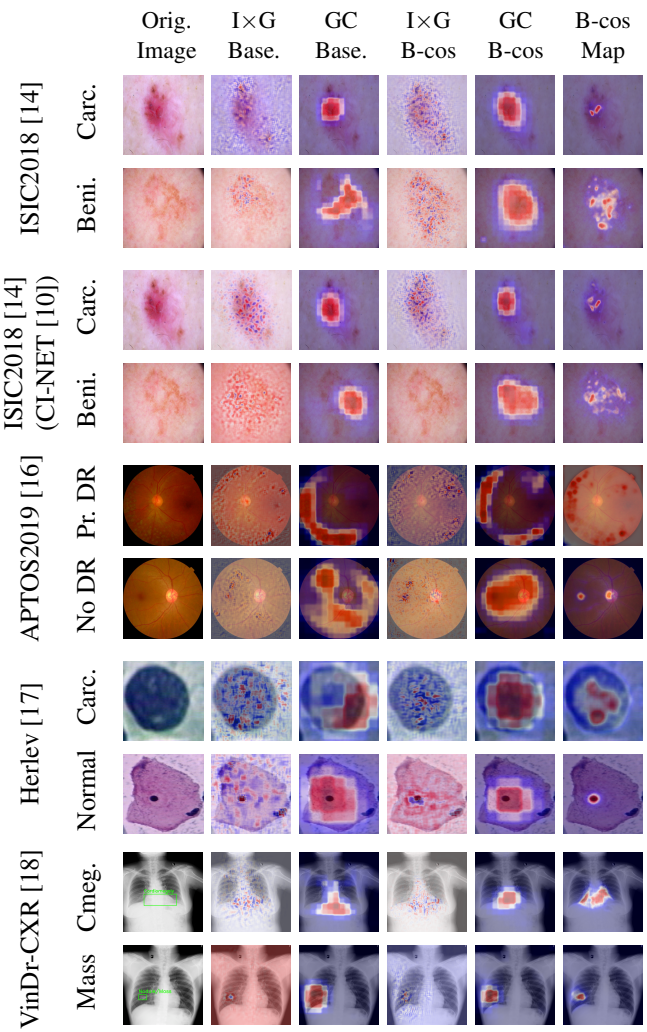

Fig. 1: Comparing explanations from Input×Gradient [19] (I×G), GradCAM [3] (GC) and B-cos on original (Base.) and B-cos networks. The colourmap ranges from blue to red, blue(red) meaning less(more) relevant pixels.

Table 1: Results on model performance and xAI quality assessment obtained for each model across several datasets. For all models we show the results for the best xAI method. For B-cos networks we also show results for the intrinsic B-cos explanations. We compare against the following xAI methods: G - Grad, GC - GradCAM [3], I×G - Input×Gradient [19], IG - Integrated Gradients [20]. For each metric ↑ (↓) means higher (lower) is better. Underlined numbers represent the best result for non B-cos models on a given dataset, while bold numbers highlight the best B-cos results. DN121 stands for DenseNet-121 and RN18 IN corresponds to an ImageNet trained ResNet-18.

Dataset	ID	Model	Pretraining	Task Performance			Faithfulness		Robustness		Complexity		Localisation	
				(B)ACC ↑	F1 ↑	AUC ↑	Select. ↓	Correl. ↑	Sens. ↓	RRS ↓	Sparse. ↑	Complex. ↑	Focus ↑	Att. Loc. ↑
ImageNet [11]	1	DN121	-	72.6	72.3	99.8	7.63 (IG)	0.073 (I×G)	0.545 (GC)	1.62 (G)	0.610 (I×G)	10.5 (GC)	0.722 (GC)	-
	2	B-cos DN121	-	**73.6**	**73.2**	99.2	5.92 (G) **4.85**	**0.138** (IG) 0.043	0.994 (GC) 1.54	1.22e-4 (G) **6.92e-5**	0.604 (IG) **0.624**	10.3 (GC) 10.1	0.748 (GC) **0.851**	- -
ISIC2018 [14]	3	DN121	-	76.7	62.5	92.4	25.9 (IG)	0.063 (GC)	1.80 (IG)	0.012 (IG)	0.583 (G)	11.6 (IG)	0.707 (GC)	-
	4	B-cos DN121	-	74.9	57.8	89.1	48.1 (GC) 91.4	**0.487** (I×G) -0.020	**0.964** (GC) 1.13	4.33e-6 (I×G) 8.87e-8	0.612 (I×G) 0.443	**11.9** (GC) 11.8	0.434 (GC) 0.585	- -
	5	B-cos DN121	3	75.5	59.6	90.4	35.7 (IG) 70.7	0.297 (I×G) 0.057	2.19 (GC) 1.94	2.84e-6 (IG) **4.81e-8**	0.657 (IG) 0.550	11.4 (G) 11.6	0.539 (GC) 0.614	- -
	6	DN121	1	85.6	78.5	97.6	100.3 (IG)	0.288 (GC)	1.32 (G)	0.507 (IG)	0.606 (GC)	11.7 (IG)	0.772 (GC)	-
	7	B-cos DN121	2	82.3	71.2	93.7	65.6 (IG) 46.6	0.114 (I×G) 0.024	2.44 (G) 6.62	9.31e-7 1.70e-5	0.707(GC) **0.767**	11.5 (I×G) 10.9	0.619 (I×G) 0.796	- -
	8	B-cos DN121	6	77.1	62.0	90.9	**26.8** (GC) 48.4	0.170 (I×G) 0.027	2.68 (GC) 1.62	2.58e-7 (G) 1.33e-7	0.631 (I×G) 0.616	11.5 (G) 11.5	0.668 (GC) 0.656	- -
	9	Baseline CI-Net [10]	RN18 IN	84.2	74.2	96.3	30.3 (I×G)	0.173 (GC)	2.00 (IG)	0.250 (G)	0.716 (GC)	11.4 (IG)	0.887 (GC)	-
	10	B-cos CI-Net [10]	RN18 IN	**82.9**	**71.4**	**94.4**	52.8 (I×G) 77.3	0.144 (IG) 0.061	1.40 (G) 3.51	1.55e-5 (G) 1.10e-5	0.744 (GC) 0.709	11.5 (I×G) 11.2	0.591 (IG) **0.889**	- -
APTOS2019 [16]	11	DN121	-	79.3	57.6	93.0	32.3 (I×G)	0.049 (GC)	1.49 (G)	0.038 (IG)	0.791 (GC)	11.4 (IG)	0.695 (GC)	-
	12	B-cos DN121	-	79.3	58.6	90.9	88.1 (GC) 82.3	**0.365** (IG) 0.019	29.1 (G) 18.4	1.02e-5 (I×G) 1.91e-5	0.723 (IG) 0.735	**11.4** (GC) 11.0	0.413 (GC) 0.585	- -
	13	B-cos DN121	11	75.7	46.2	88.1	**41.5** (I×G) 51.7	0.104 (I×G) 0.027	1.66 (GC) **1.46**	1.08e-6 (IG) **7.79e-7**	0.721 (GC) 0.723	11.1 (G) 11.0	0.549 (GC) 0.642	- -
	14	DN121	1	82.9	66.2	94.7	37.1 (IG)	0.180 (GC)	1.41 (G)	0.156 (G)	0.802 (GC)	11.2 (IG)	0.845 (GC)	-
	15	B-cos DN121	2	**81.3**	**61.9**	**92.6**	56.1 (I×G) 69.3	0.061 (GC) 0.034	2.61 (IG) 5.10	9.64e-7 (G) 1.35e-6	0.784 (GC) **0.826**	11.2 (I×G) 10.5	0.729 (GC) **0.870**	- -
	16	B-cos DN121	14	78.9	54.8	91.4	51.0 (IG) 53.6	0.0578 (GC) 0.046	1.61 (I×G) 4.99	1.05e-6 (G) 1.67e-5	0.792 (GC) 0.808	10.8 (I×G) 10.6	0.791 (GC) 0.818	- -
Herlev [17]	17	DN121	1	66.2	71.0	94.3	10.1 (I×G)	0.143 (GC)	0.912 (GC)	0.0530 (Grad)	0.553 (I×G)	10.6 (GC)	0.732 (GC)	-
	18	B-cos DN121	2	63.5	67.1	91.4	**8.67** (I×G) 9.65	0.124 (GC) **0.144**	2.03 (IG) 2.68	**1.29e-6** 2.24e-6	0.557 (IG) **0.572**	10.3 (GC) 10.2	0.606 (GC) 0.666	- -
VinDr-CXR [18]	19	DN121	1	96.2	28.4	74.8	9.13 (I×G)	0.119 (GC)	1.00 (I×G)	2.97e-4 (I×G)	0.740 (GC)	11.6 (IG)	0.277 (I×G)	0.137 (GC)
	20	B-cos DN121	2	95.6	22.7	**77.8**	**4.56** (GC) 5.04	0.093 (GC) 0.066	2.58 (I×G) 13.2	**1.18e-6** (IG) 1.60e-5	**0.763** (GC) 0.726	11.5 (G) 11.0	0.297 (I×G) **0.301**	0.237 (GC) **0.287**

are more concentrated around a single area. Furthermore, B-cos maps tend to have less highly activated pixels than Grad-CAM's.

4. CONCLUSION AND FUTURE WORK

This work presents the first study on the applicability and suitability of B-cos networks, a weight-input aligned architecture, for the medical domain. Although achieving better explanation quality than other xAI methods on ImageNet trained models, it is not clear that the same happens for the medical use cases considered, especially when taking into account all explanation quality assessment metrics. However, B-cos explanations yield more localised saliency maps, a finding that corroborates the original authors' conclusions. Given these results and the slight drop in classification performance, it is not clear that B-cos networks produce better explanations in the medical domain. Further work should be devoted to integrating B-cos networks in different tasks (e.g., image segmentation), studying of different xAI quality assessment metrics with more post-hoc methods, and a deep focus on model optimization or regularization to promote the integration of clinical-domain knowledge into the models (e.g., integrating clinical metadata through data fusion techniques, using class weights in the loss function).

Compliance with Ethical Standards This research study was conducted retrospectively using human subject data made available in open access. Ethical approval was *not* required, as confirmed by the license attached with the open access data.

5. REFERENCES

[1] Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso, "Machine Learning Interpretability: A Survey on Methods and Metrics," *Electronics*, vol. 8, 2019.

[2] Isabel Rio-Torto, Kelwin Fernandes, and Luís F. Teixeira, "Understanding the decisions of CNNs: An in-model approach," *Pattern Recognition Letters*, vol. 133, pp. 373–380, 2020.

[3] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, "Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization," in *ICCV*, 2017, pp. 618–626.

[4] Tiago Gonçalves, Isabel Rio-Torto, Luís F. Teixeira, and Jaime S. Cardoso, "A Survey on Attention Mechanisms for Medical Applications: are we Moving Toward Better Algorithms?," *IEEE Access*, vol. 10, pp. 98909–98935, 2022.

[5] Jianlong Zhou, Fang Chen, and Andreas Holzinger, "Towards Explainability for AI Fairness," in *XxAI - Beyond Explainable AI*, 2020, p. 375–386.

[6] Cynthia Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[7] Cristiano Patrício, João C. Neves, and Luís F. Teixeira, "Explainable Deep Learning Methods in Medical Image Classification: A Survey," *ACM Computing Surveys*, vol. 56, no. 4, oct 2023.

[8] Moritz Böhle, Navdeeppal Singh, Mario Fritz, and Bernt Schiele, "B-cos Alignment for Inherently Interpretable CNNs and Vision Transformers," *arXiv:2306.10898*, 2023.

[9] Anna Hedström, Leander Weber, Daniel Krakowczyk, Dilyara Bareeva, Franz Motzkus, Wojciech Samek, Sebastian Lapuschkin and Marina Marina M.-C. Höhne, "Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations and Beyond," *Journal of Machine Learning Research*, vol. 24, no. 34, pp. 1–11, 2023.

[10] Zihao Liu, Ruiqin Xiong, and Tingting Jiang, "CI-Net: Clinical-Inspired Network for Automated Skin Lesion Recognition," *IEEE Transactions on Medical Imaging*, vol. 42, no. 3, pp. 619–632, 2023.

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR*, 2009, pp. 248–255.

[12] Christos Matsoukas, Johan Fredin Haslum, Moein Sorkhei, Magnus Söderberg, and Kevin Smith, "What Makes Transfer Learning Work for Medical Images: Feature Reuse & Other Factors," in *CVPR*, 2022, pp. 9225–9234.

[13] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger, "Densely Connected Convolutional Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4700–4708.

[14] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler, "Data descriptor: the HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific Data*, vol. 5, no. 1, 2018.

[15] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al., "Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC)," *arXiv:1902.03368*, 2019.

[16] "APTOS 2019 Blindness Detection: Detect diabetic retinopathy to stop blindness before it's too late," `https://www.kaggle.com/c/aptos2019-blindness-detection`, Accessed: 2023-11-07.

[17] Jan Jantzen, Jonas Norup, Georgios Dounias, and Beth Bjerregaard, "Pap-smear Benchmark Data For Pattern Classification," *Nature inspired Smart Information Systems (NiSIS)*, pp. 1–9, 2005.

[18] Ha Q. Nguyen, Khanh Lam, Linh T. Le, Hieu H. Pham, Dat Q. Tran, Dung B. Nguyen, Dung D. Le, Chi M. Pham, Hang T. T. Tong, Diep H. Dinh, Cuong D. Do, Luu T. Doan, Cuong N. Nguyen, Binh T. Nguyen, Que V. Nguyen, Au D. Hoang, Hien N. Phan, Anh T. Nguyen, Phuong H. Ho, Dat T. Ngo, Nghia T. Nguyen, Nhan T. Nguyen, Minh Dao, and Van Vu, "VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations," *Scientific Data*, vol. 9, no. 1, pp. 429, 2022.

[19] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje, "Learning Important Features Through Propagating Activation Differences," in *Proceedings of the 34th International Conference on Machine Learning*, 2017, vol. 70 of *Proceedings of Machine Learning Research*, pp. 3145–3153.

[20] Mukund Sundararajan, Ankur Taly, and Qiqi Yan, "Axiomatic Attribution for Deep Networks," in *ICML*, 2017, vol. 70, p. 3319–3328.