How can data augmentation improve attribution maps for disease subtype explainability?

Elina Thibeau-Sutre, Jelmer M Wolterink, Olivier Colliot, Ninon Burgos

How can data augmentation improve attribution maps for disease subtype explainability?

Elina Thibeau-Sutre[a,b], Jelmer M. Wolterink[a], Olivier Colliot[b], and Ninon Burgos[b]

[a]Department of Applied Mathematics, Technical Medical Centre, University of Twente, Enschede, The Netherlands
[b]Sorbonne Université, Institut du Cerveau - Paris Brain Institute - ICM, CNRS, Inria, Inserm, AP-HP, Hôpital de la Pitié Salpêtrière, F-75013, Paris, France

ABSTRACT

As deep learning has been widely used for computer aided-diagnosis, we wished to know whether attribution maps obtained using gradient back-propagation could correctly highlight the patterns of disease subtypes discovered by a deep learning classifier. As the correctness of attribution maps is difficult to evaluate directly on medical images, we used synthetic data mimicking the difference between brain MRI of controls and demented patients to design more reliable evaluation criteria of attribution maps. We demonstrated that attribution maps may mix the regions associated with different subtypes for small data sets while they could accurately characterize both subtypes using a large data set. We then proposed simple data augmentation techniques and showed that they could improve the coherence of the explanations for a small data set.

Keywords: Explainability, Pattern recognition and classification, Synthetic data, Subtyping, Interpretability

1. INTRODUCTION

Explaining the decisions made by deep learning models is not trivial, but could be helpful to validate their application to clinical tasks for which only few test samples are available. This is why explainability methods generating attribution maps have been developed for classifiers working on images. Many of these methods are based on gradient back-propagation.[1] Frameworks such as weakly-supervised segmentation[2] assume that the attribution map of a convolutional neural network (CNN) performing classification is coherent with prior knowledge. In this kind of framework, a deep learning network produces a label at the image-level (for example the presence or absence of a tumor), and the assumption is that the attribution map of this decision highlights which information in the image the network used to compute its output. These attribution maps are assumed to produce the segmentation of the region needed for the classification task (for example the segmentation of the tumor). However, a recent publication[3] benchmarked a series of explainability methods derived from gradient back-propagation and assessed that they had poor localization performance. This means that they were not able to correctly segment the region of interest corresponding to the classification task. However, it remains difficult to evaluate the coherence with prior knowledge of deep learning models associated with their explainability methods, especially for clinical tasks whose ground truth varies according to clinicians.[4] This is in particular true for disease subtyping as there is often no universally accepted ground-truth for subtypes. This ground truth is even more difficult to assess when there is no consensus on the definitions of the subtypes composing the disease, as is the case for example for Alzheimer's disease,[5] Parkinson's disease,[6] or frontotemporal dementia.[7]

In this paper, we show that gradient back-propagation *can* correctly highlight ground truth patterns. We first assess that a CNN can identify latent subtypes grouped under the same label by K-means clustering on its feature maps, and then show that attribution maps can accurately highlight specific patterns of these subtypes. Additionally, we propose and evaluate simple data augmentation strategies to improve the separability of these subtypes with a limited number of samples. We work with synthetic data mimicking the patterns of a neurodegenerative disease. In such a controlled setting, we know which patterns are relevant for the classification, allowing a more reliable comparison of the evaluated techniques.

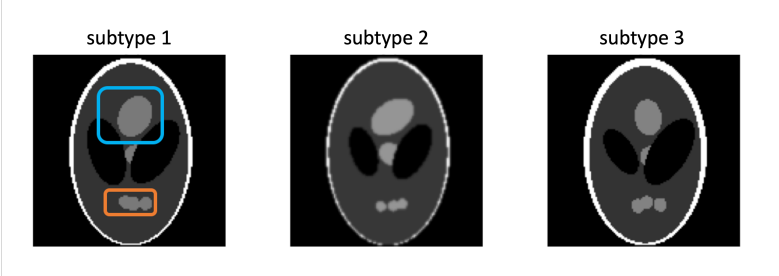

Figure 1: Display of the three subtypes. Regions *Top* and *Bottom* are framed in blue and orange, respectively. Subtype 1 has two large regions, subtype 2 has a large *Top* region and a small *Bottom* region, and subtype 3 has a small *Top* region and a large *Bottom* region.

2. MATERIALS AND METHODS

Experiments, including data generation, network training, and attribution maps computation, were conducted with the ClinicaDL* software.[8]

2.1 Synthetic data

Synthetic images were used instead of real ones to ensure that the patterns identified by the network correspond to the true patterns specific to the underlying subtypes. These images are inspired from the Shepp-Logan phantom[9] that was used to develop and test neuroimaging reconstruction algorithms.

Our use case is Alzheimer's disease heterogeneity. We simulated a data set of T1-weighted (T1w) MR images of demented patients and cognitively normal participants. Three rules were used to generate the synthetic data sets. First, the disease is characterized by the atrophy (volume decrease) of specific regions of the brain. Second, the disease is heterogeneous, i.e. different atrophy patterns exist within a patient group. Third, some patients can hardly be distinguished from cognitively normal participants based on their T1w MRI.

The size of the images was fixed to 128 × 128 pixels. Three subtypes, displayed in Figure 1, were sampled by varying the size, position, orientation and contrast of the ellipses depicting simplified brain regions. A random smoothing was applied. Subtypes vary according to the size of the regions *Top* and *Bottom*. Subtype 1 has two large regions, subtype 2 has a large *Top* region and a small *Bottom* region, and subtype 3 has a small *Top* region and a large *Bottom* region. Two labels are associated with these subtypes: *Control* and *Atrophied*. The *Control* group comprises only images of subtype 1, whereas the *Atrophied* group comprises images of subtype 1 (errors), subtype 2 (typical) and subtype 3 (atypical).

As we do not know the proportion of each subtype in the *Atrophied* group, three training/validation data sets were sampled with different *Atrophied* distributions: *homogeneous* (500 samples per class) consists of 85% typical subtypes, 10% atypical subtypes, and 5% errors, *heterogeneous* (500 samples per class) consists of 65% typical subtypes, 25% atypical subtypes, 10% of errors, and *large* (10,000 samples per class) has the same properties as *homogeneous*. Finally, we generated a test set of 1,000 images in which the *Atrophied* group is composed of half typical and half atypical subtypes.

2.2 Baseline CNN classification

All networks (except in section 3.2.1) learn the classification task *Control* versus *Atrophied*. The subtypes are not given to the network. The chosen architecture is composed of five convolutional blocks followed by a dropout layer with probability 0.5 and a fully-connected layer. Each convolutional block is made of two convolutional layers with kernel size of 3 and padding of 1, a batch-normalization layer, a LeakyReLU activation and a max pooling layer with a kernel size and a stride of 2. Default values were used for other hyperparameters. The predicted label of the input image is the index of the output node having the maximum activation. The weights of the convolutional and fully-connected layers were initialized with the default initialization method of PyTorch.[10] During training, weights were updated based on the cross-entropy loss. The training continues to the end of

*https://github.com/aramis-lab/clinicadl

Table 1: Baseline performance for each data set. The first two rows indicate the theoretical values for a perfect or a random characterization of the patterns.

Data set	Task	CNN subtyping ability	typical subtype separability	typical subtype specificity	atypical subtype separability	atypical subtype specificity	balanced accuracy
perfect		1.00	$+\infty$	$+\infty$	$+\infty$	$+\infty$	N/A
random		0.00	1.0	1.0	1.0	1.0	N/A
homogeneous	*Control* versus *Atrophied*	0.95	12.3	5.3	1.6	4.2	0.98
heterogeneous	*Control* versus *Atrophied*	1.00	8.6	5.1	0.9	4.6	0.98
homogeneous	subtype classification	N/A	9.3	6.6	1.1	7.9	1.00
heterogeneous	subtype classification	N/A	9.8	7.3	1.1	7.8	1.00
large	*Control* versus *Atrophied*	0.99	102.5	12.3	12.9	13.7	1.00

a pre-defined number of epochs (100 or 1,000). Unless stated otherwise, only the networks trained with 1,000 epochs are presented. The final model is the one that obtained the lowest validation loss during training. The loss of the model is evaluated at the end of each epoch. By default there is no data augmentation.

Each training/validation data set was split into training and validation sets by performing a 5-fold cross-validation (CV) stratified by subtype. The 5-fold data split was performed only once, thus guaranteeing that all the experiments used the same images during CV. Models were tested on the test set of 1,000 images without data augmentation. The balanced accuracy was computed to ensure that the model correctly learnt the task.

2.3 Data augmentation

Data augmentation was only applied to the training set during the training procedure. Once an image is loaded, it may be randomly altered to produce a slightly different version. The possible transformations are the following:

- **CropPad** samples between 0 and 10 pixels to crop the boundaries on one side and pad the opposite side;
- **Erasing**[11] draws a black rectangle in the image with a probability of 0.5. The default PyTorch values were used to choose the size and location of the rectangle;
- **Noise** adds Gaussian noise to the original image with a standard deviation sampled between 0 and 0.1;
- **Smoothing** smooths the image with a Gaussian kernel of standard deviation sampled between 0 and 1.

Note that we limited our search to transformations that could easily be applied to 3D volumes.

2.4 Subtype identification using attribution maps

Though many explainability methods exist, we restricted this study to attribution maps produced by gradient back-propagation,[1] as it is widely used and conceptually simple. An individual attribution map corresponds to the gradients of an output node with respect to an image. In our case, the output node is the one corresponding to the *Control* group. The intensities of the attribution map of an image correspond to the changes needed to transform this image into a sample of the *Control* group. A group attribution map is the average of the individual attribution maps of the correctly classified images of this group.

The coherence of the CNN coupled with attribution maps in identifying *Atrophied* subtypes and their characteristics was measured with three criteria. In the following we refer to the regions *Top*, *Bottom* and *Background*, where *Background* includes all pixels of the image which are neither in *Top* nor in *Bottom*. S_X is the surface of region X and $I(p)$ is the intensity of the saliency map at position p.

CNN subtyping ability. The ability to identify the subtypes was evaluated by clustering ten times the feature maps before the dropout layer and evaluating the mean adjusted rand index[†] between the true subtypes and the clusters found on the evaluation set.

[†]the mean adjusted rand index measures the similarity of two clusterings, corrected for chance

Attribution maps separability. The ability to distinguish patterns of typical vs atypical subtypes was evaluated by comparing the sum of the absolute values of the normalized intensities in the regions *Top* and *Bottom*.

$$separability_{typical} = \frac{\sum_{p \in Top} |I(p)|}{\sum_{p \in Bottom} |I(p)|} * \frac{S_{Bottom}}{S_{Top}} \qquad separability_{atypical} = \frac{\sum_{p \in Bottom} |I(p)|}{\sum_{p \in Top} |I(p)|} * \frac{S_{Top}}{S_{Bottom}} \tag{1}$$

Attribution maps specificity. The specificity was evaluated by computing the ratio between the sum of the absolute values of the normalized intensities in the regions *Top* and *Bottom* and in the *Background*.

$$specificity_{typical} = specificity_{atypical} = \frac{\sum_{p \in Top \cup Bottom} |I(p)|}{\sum_{p \in Background} |I(p)|} * \frac{S_{Background}}{S_{Top \cup Bottom}} \tag{2}$$

Theoretical values of each criterion for a perfect or random characterization of the patterns are given in Table 1. Intensities in a region are normalized, i.e. divided by the surface of the whole region. As discussed in Section 4, one should not compare the values of the criteria between the two subtypes.

3. RESULTS

3.1 Baseline results

The three evaluation criteria were applied to networks trained on the *homogeneous* and *heterogeneous* data sets (Table 1). Group attribution maps of the first fold of each network series are displayed in Figure 2A. Balanced accuracy is very high (>= 0.98) for all the networks. The subtyping ability of the CNN is maximal for all the networks, the difference on *homogeneous* being caused by results on one fold being lower than the others. The separability and specificity of attribution maps are equivalent for both data sets. On both data sets the separability of the atypical saliency map is equivalent to a random map, even though these maps are specific.

3.2 Ideal cases

To understand how these baseline results could be improved, saliency maps obtained in two ideal cases were evaluated. In the first case, the same amount of data is used but the subtypes are actually known and used to train the network. In the second case, the labels are unknown but the data set is larger.

3.2.1 Known subtypes

In this section, the networks are not trained on the binary class *Control* vs *Atrophied* as in the other sections. Instead, a three-class classification is learnt based on the subtype. Results with known labels are displayed in Figure 2B and in Table 1. In this case the CNN subtyping ability was not evaluated, as the networks are actually learning to differentiate the subtypes, so it actually corresponds to the balanced accuracy evaluated on the test set. We used 100 epochs as convergence was reached sooner.

The specificity is increased when knowing the labels, i.e. regions that are not correlated with any subtype are less highlighted by the saliency maps compared to the ones correlated with the patterns. However, the separability remains low, which means that both regions of interest are highlighted for both subtypes, so it is not possible to differentiate subtypes based on their heatmaps.

3.2.2 Large data set

We studied the impact of the number of samples by comparing the results obtained with the *large* and *homogeneous* data sets (Table 1 and Figure 2C). For the *large* data set, we used 100 epochs because training was more computationally costly and convergence was reached sooner.

For the *large* data set, the separability and the specificity of the attribution maps are much higher than for the *homogeneous* data set, even when the task is to differentiate the subtypes. Moreover, the areas highlighted by the attribution maps appear more separated at visual inspection. This indicates that attribution maps may better reflect the patterns of each sample without being contaminated by the global pattern of the disease when CNNs are trained with more data. However, data sets in medical imaging tend to be small so we need to find solutions to improve the separability and specificity of our attribution maps on small data sets.

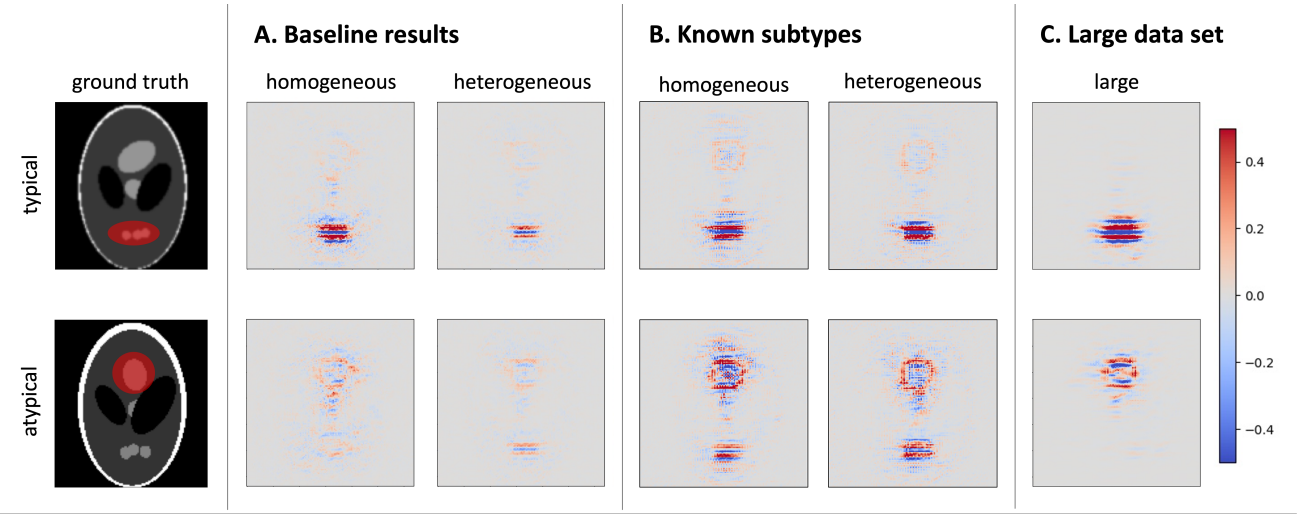

Figure 2: Group attribution maps obtained for the first fold of the 5-fold CV without data augmentation (A) with unknown subtypes, (B) with known subtypes and (C) with a large data set of unknown subtypes.

3.3 Benchmark of data augmentation strategies

We proposed to use data augmentation (see section 2.3) to improve the separability and specificity of attribution maps. Notable improvement was only observed for the *heterogeneous* data set. Only **CropPad** led to a statistical improvement according to a T-test with Bonferroni correction on the five folds of each experiment, and on atypical maps only. The two best-performing techniques were selected to be applied in the following order: **CropPad** and **Noise**. This combination led to a statistically significant improvement compared with the baseline values on the separability and the specificity of the typical subtype, and the specificity of the atypical subtype. Criterion values are summarized in Table 2 and group attribution maps obtained with the *homogeneous* data set are displayed in Figure 3.

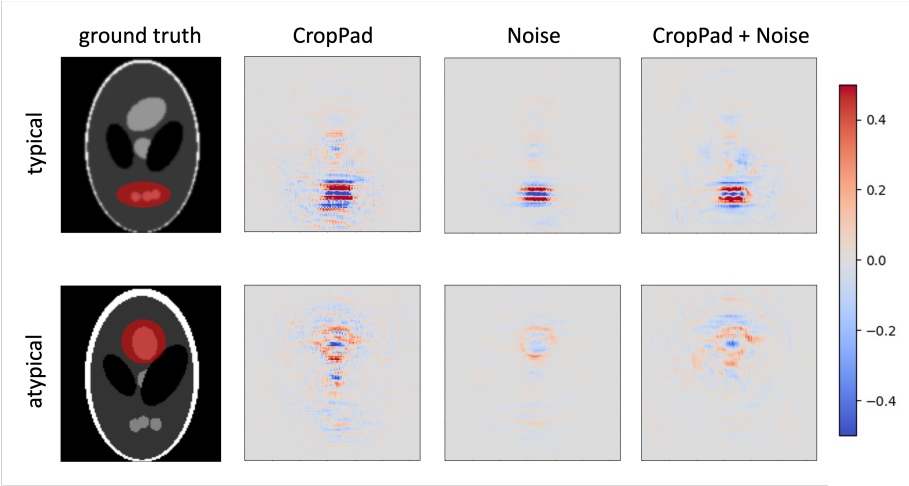

Figure 3: Group attribution maps obtained for the first fold of the 5-fold CV with the *homogeneous* data set and different data augmentation strategies (**CropPad**, **Noise** and their combination), for both the typical and atypical subtypes.

4. CONCLUSION

We studied the ability of CNNs and attribution maps to identify subtypes composing a heterogeneous disease. Our synthetic data was made to simulate Alzheimer's disease heterogeneity, but the whole framework can be

Table 2: Benchmark of data augmentation techniques on the *homogeneous* and *heterogeneous* data sets. * denotes a statistical difference with the baseline values assessed using a T-test ($p<0.05$) with Bonferroni correction.

Experiment	CNN subtyping ability	typical subtype separability	typical subtype specificity	atypical subtype separability	atypical subtype specificity	balanced accuracy
CropPad	1.00	13.2	4.7	2.0	5.1	0.99
Erasing	1.00	19.6	5.6	1.8	4.0	0.99
Noise	1.00	18.4	9.1	2.3	5.5	0.99
Smoothing	1.00	10.4	2.9	1.1	4.3	0.99
CropPad + Noise	1.00	19.1	6.4	4.2	7.1	1.00
None (baseline)	0.95	12.3	5.3	1.6	4.2	0.98

(a) *Homogeneous* data set

Experiment	CNN subtyping ability	typical subtype separability	typical subtype specificity	atypical subtype separability	atypical subtype specificity	balanced accuracy
CropPad	1.00	20.6	5.8	2.6*	7.3*	1.00
Erasing	1.00	13.7	6.3	2.0	5.2	0.99
Noise	1.00	17.3	8.1	2.3	5.4	0.99
Smoothing	1.00	12.1	5.1	1.3	3.9	0.99
CropPad + Noise	1.00	41.8*	7.6*	3.7	8.2*	1.00
None (baseline)	1.00	8.6	5.1	0.9	4.6	0.99

(b) *Heterogeneous* data set

adapted to another clinical problem. We observe that even when models obtain a very high accuracy and disease subtypes are correctly identified, the attribution maps fail to represent the underlying ground truth at the subtype level on small data sets. We then studied two ideal cases: first the case in which the subtypes are actually known and learnt by the network and then a case with a larger amount of data. We observed that the saliency maps are only slightly improved by the ground truth subtypes knowledge, and that a better coherence with the ground truth patterns is obtained with a larger data set.

As medical imaging data sets are often small, we introduced different data augmentation techniques to better identify patterns with attribution maps. This improved the separability and specificity of group attribution maps on a small data set. We showed that this improvement depends on the subtype distribution. For the *homogeneous* data set, the increase in performance was generally smaller than for the *heterogeneous* data set. The reduced number of techniques reaching a statistically significant improvement compared with the baseline could also be due to the limited number of models run per technique (five, one for each fold). This analysis could be improved by running more models for each experiment.

This study is limited to the use of gradient back-propagation to generate attribution maps, but other methods exist. Comparing them using the proposed criteria would also allow assessing the reliability of these criteria, by comparing the inter-rater reliability on individual assessments of each explainability method.[12] Another current limitation is the restriction to synthetic data.

Contrary to [3], we assessed that an explanation method in association with the deep learning network it explains could be coherent with prior knowledge. As we found that the coherence of our setup could be improved using the same deterministic explanation method by changing the training of the CNN, we conclude that the poor localization utility that they observed in their study is not due to the explanation methods as they hypothesize, but to the CNNs which were not able to learn the clinical relevant features and relied instead on confounding variables. These findings support our previous study on the lack of robustness of CNNs trained on neuroimaging data.[13] This analysis also shows a limitation of semi-supervised segmentation frameworks described in the introduction, in which the number of samples may not be sufficient to highlight the diversity of regions necessary to characterize a label learnt at the level of the image, even if the labels are accurate and homogeneous.

ACKNOWLEDGMENTS

The research leading to these results has received funding from the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute) and reference ANR-10-IAIHU-06 (Agence Nationale de la Recherche-10-IA Institut Hospitalo-Universitaire-6). We also acknowledge funding from the 4TU Precision Medicine program supported by High Tech for a Sustainable Future, a framework commissioned by the four Universities of Technology of the Netherlands. The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Simonyan, K., Vedaldi, A., and Zisserman, A., "Deep inside convolutional networks: Visualising image classification models and saliency maps," in [*In Workshop at International Conference on Learning Representations*], (2014).

[2] Patel, G. and Dolz, J., "Weakly supervised segmentation with cross-modality equivariant constraints," *Medical Image Analysis* **77**, 102374 (2022).

[3] Arun, N., Gaw, N., Singh, P., Chang, K., Aggarwal, M., Chen, B., Hoebel, K., Gupta, S., Patel, J., Gidwani, M., Adebayo, J., Li, M. D., and Kalpathy-Cramer, J., "Assessing the Trustworthiness of Saliency Maps for Localizing Abnormalities in Medical Imaging," *Radiology: Artificial Intelligence* **3**(6), e200267 (2021).

[4] Saporta, A., Gui, X., Agrawal, A., Pareek, A., Truong, S. Q., Nguyen, C. D., Ngo, V.-D., Seekins, J., Blankenberg, F. G., Ng, A. Y., Lungren, M. P., and Rajpurkar, P., "Benchmarking saliency methods for chest X-ray interpretation," (2021).

[5] Ferreira, D., Nordberg, A., and Westman, E., "Biological subtypes of Alzheimer disease: A systematic review and meta-analysis," *Neurology* **94**(10), 436–448 (2020).

[6] Thenganatt, M. A. and Jankovic, J., "Parkinson Disease Subtypes," *JAMA Neurology* **71**(4), 499–504 (2014).

[7] Seelaar, H., Rohrer, J. D., Pijnenburg, Y. A. L., Fox, N. C., and van Swieten, J. C., "Clinical, genetic and pathological heterogeneity of frontotemporal dementia: A review," *Journal of Neurology, Neurosurgery & Psychiatry* **82**(5), 476–486 (2011).

[8] Thibeau-Sutre, E. et al., "ClinicaDL: An open-source deep learning software for reproducible neuroimaging processing," *Computer Methods and Programs in Biomedicine* (2022).

[9] Shepp, L. A. and Logan, B. F., "The Fourier reconstruction of a head section," *IEEE Transactions on Nuclear Science* **21**(3), 21–43 (1974).

[10] He, K., Zhang, X., Ren, S., and Sun, J., "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," in [*2015 IEEE International Conference on Computer Vision (ICCV)*], 1026–1034, IEEE (2015).

[11] Zhong, Z., Zheng, L., Kang, G., Li, S., and Yang, Y., "Random Erasing Data Augmentation," *Proceedings of the AAAI Conference on Artificial Intelligence* **34**(07), 13001–13008 (2020).

[12] Tomsett, R., Harborne, D., Chakraborty, S., Gurram, P., and Preece, A., "Sanity Checks for Saliency Metrics," *Proceedings of the AAAI Conference on Artificial Intelligence* **34**(04), 6021–6029 (2020).

[13] Thibeau-Sutre, E., Colliot, O., Dormont, D., and Burgos, N., "Visualization approach to assess the robustness of neural networks for medical image classification," in [*Medical Imaging 2020: Image Processing*], **11313**, 113131J (2020).