DBCE : A Saliency Method for Medical Deep Learning Through Anatomically-Consistent Free-Form Deformations

Joshua Peters† University of Queensland joshua.peters@uq.net.au

Léo Lebrat† CSIRO QUT

Rodrigo Santa Cruz CSIRO QUT

Aaron Nicolson CSIRO

Gregg Belous CSIRO

Salamata Konate CSIRO and QUT

Parnesh Raniga CSIRO

Vincent Dore CSIRO

Pierrick Bourgeat CSIRO

Jurgen Mejan-Fripp CSIRO

Clinton Fookes QUT

Olivier Salvado CSIRO and QUT

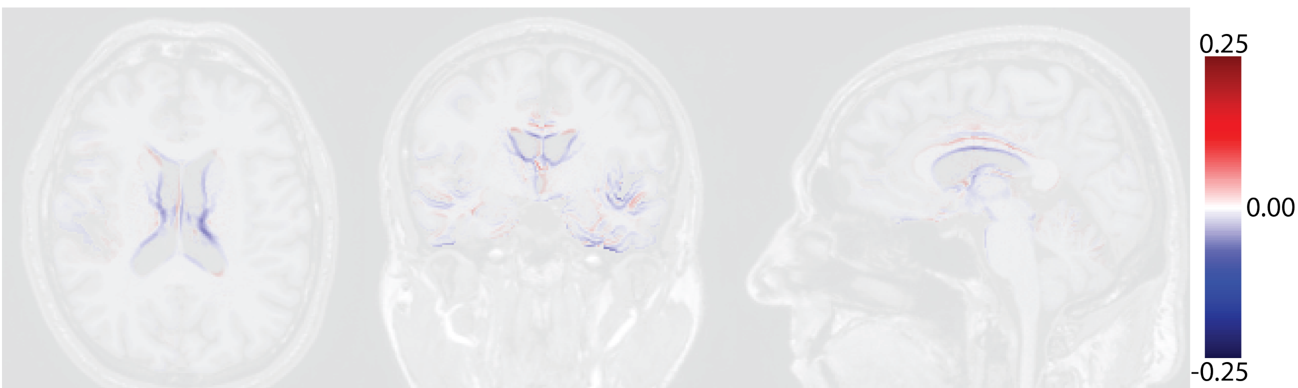

Figure 1: Difference between the initial image and the image produced by our generation-based method when applied to a neural network trained to detect Alzheimer's disease. The blue colour denotes tissue atrophy, and the red colour an increase in tissue density. DBCE outputs the minimal anatomically plausible diffeomorphism flipping one's network prediction.

Abstract

Deep learning models are powerful tools for addressing challenging medical imaging problems. However, for an ever-growing range of applications, interpreting a model's prediction remains non-trivial. Understanding decisions made by black-box algorithms is critical, and assessing their fairness and susceptibility to bias is a key step towards healthcare deployment. In this paper, we propose ***DBCE*** *(**D**eformation **B**ased **C**ounterfactual **E**xplainability). We optimise a diffeomorphic transformation that deforms a given input image to change the prediction of the model. This provides anatomically meaningful saliency maps indicating tissue atrophy and expansion, which can be easily interpreted by clinicians. In our test case,* ***DBCE*** *replicates the transition of a patient from healthy control (HC) to Alzheimer's disease (AD). We benchmark* ***DBCE*** *against three commonly used saliency methods. We show that it provides more meaningful saliency maps when applied to one subject and disease-consistent atrophy patterns when used over a larger cohort. In addition, our method fulfils a recent sanity check and is repeatable for different model initialisations in contrast to classical sensitivity-based methods.*

† Equal contribution, author ordering determined by coin flip.

1. Introduction

During the last decade, medical Deep Learning (DL) models have surpassed deterministic approaches by delivering faster runtimes and above human-expert performance [42, 60, 35]. Despite these astonishing results, DL-based methods suffer slow adoption rates in healthcare settings, with critics often referring to their lack of tractability or explainability [20, 11]. Indeed, to benefit from a black-box prediction in a critical domain, one would like to provide additional explanations that have motivated the decision to help the final user in accessing the network decision. Those explanations will first help rationalise and make the decision-making of neural networks (NNs) less opaque. In addition, it could help the end-user identify potential biases or support designing trustworthy and fair algorithms [39, 53].

To this end, numerous methods have been devised to provide the user with an understanding of the *"why"* through heatmaps that highlight the most critical parts of the input contributing to the prediction. Nevertheless, the quality and "finesse" of these saliency maps are vital for providing additive value to the diagnostic process. It has been demonstrated that for assisting in the grading of diabetic retinopathy, the prediction of a DL model along with a heatmap generated with Integrated Gradient does not provide a more significant benefit over the prediction alone [49]. In addition, recent studies suggest that diffuse saliency maps are prone to user confirmation and automation bias [48, 20], raising further questions on the actual usefulness of saliency maps. In this direction, researchers devised new sanity checks [1, 22, 4], a set of benign tests that we expect saliency methods to pass. Those tests aim to assess the utility and robustness of saliency methods and evaluate their: localisation utility, sensitivity to model weight randomisation, repeatability and reproducibility. Surprisingly, most post-hoc saliency approaches failed to pass those tests [1, 4].

Those recent findings pushed the community to rethink the way of producing saliency maps. Traditionally, sensitivity-based methods look only at what pieces of information already present in the image are used to produce the prediction. Training Data Based Explanation Methods is a new group of regimes that no longer look at an individual image but rather explore the relationships between the image and the training dataset. It provides an explanation based on comparing influential samples, concepts or prototypical parts built from the training dataset [30, 31, 8, 26].

However, relying on the training data has inherent limitations in the case of high-dimensional data and scarce datasets. Recent advances in GPU computing and the progress of generative models allow the creation of compelling counterfeit data. The ability to mimic one's dataset distribution allowed one to consider counterfactual approaches and explore the generation of explanation by producing new counterfactual examples [57, 19]. Generation-based methods provide a rich insight. It allows exploring the *"what if"* scenarios through the generation of new plausible examples close to the original image but for which the neural network predicts a different class. Indeed if the classification output is incorrect, this method allows grasping which parts of the image are needed to be changed to correct the prediction. In the case of medical imaging, this approach exhibits to a practitioner which medical images would have yielded a different diagnosis, diminishing the deep neural network's (DNN) opaqueness. This technique provides a profound insight into the decision boundary of a DNN by providing realistic data of where the decision is being flipped.

In this paper, we propose **DBCE** (**D**eformation **B**ased **C**ounterfactual **E**xplanability), a generation-based method which given an input image and a black-box Deep Neural Network (DNN), produces a delusive image by optimising a regular deformation on the input image. From DBCE's deformation, we derive a pixel attribution map (saliency map) derived from the norm of the deformation; we compare this method against classical post-hoc explainability methods. Further, we propose a more instructive visualisation technique based on the difference between the original input and the generated delusive image. We access the repeatability of our technique for a fixed method and architecture but different checkpoints [1, 4]. We then evaluate the intra-class (same diagnosis) repeatability of our method across multiple patients. Finally, we access the ill-posedness of DBCE for deformations supported by a single anatomical zone; we use this result to demonstrate the robustness of lightweight deep Convolutional Neural Networks (CNNs) against localised alterations of the input image. These results highlight the robustness of this architecture [23].

2. Related works

Saliency maps for medical imaging have been produced in several ways. They can be derived from the model's architecture. For example, the attention matrices of a vision Transformer highlight the sub-regions of a medical image that it deems most important [21, 12]; Generative approaches can provide the explanation of a decision's landscape through displacement in the latent space [57, 10]. However, these methods have several disadvantages: they are computationally intensive, especially for three-dimensional images, and can be challenging to train, particularly on small datasets.

Interpretation of a model's prediction can be studied through its approximation. This can be achieved locally by discretizing a complex model with a fully explainable one [45]. However, it has been found that the fidelity of this surrogate model can be brittle [3, 58].

Saliency maps can be generated by post-hoc methods, which require a trained model and input samples. There are two main categories of post-hoc methods: sensitivity-based and perturbation methods.

Sensitivity-based methods encompass both gradient-based methods [55, 68, 61, 52, 54, 59, 7, 29] and contribution propagation methods [6, 37]. Notwithstanding their computational affordability, few of these methods require access to intermediate layers [6, 52] or require architectural modifications of the CNN [68, 61]. However, their interpretation can be difficult and subjective [1, 22, 4].

Perturbation methods aim to find which localized regions contribute most to the prediction by leveraging masking or blurring of the input image [40, 18, 64, 32]. Whilst such perturbations could be appropriate for natural images (where occlusion can naturally occur), they may not be suitable for three-dimensional structural images—as such perturbations would produce unrealistic images. To amend this, three-dimensional perturbation methods often require anatomical segmentation—which is task-dependent and requires labour-intensive annotation [64].

Counterfactual methods [66, 13] are crafted to produce minimal information that will tamper with the initial prediction of a neural network. In the Computer Vision literature, those methods are also known as adversarial perturbations [38, 44] notwithstanding being robust to a physical-world context [33, 16] the insight provided by such methods remains limited. Indeed, those methods results are not sparse and interpretable by humans. Most of the time, examples provided by those methods are artificial and unrealistic to the initial training dataset. Recently, those methods have been successfully revisited using generative approaches with discriminative losses [57, 19]. However, those approaches remain very challenging for three-dimensional, limited data. Our method is situated at the interface of both counterfactual and perturbation based-methods and leverages the benefits from both.

The proposed approach does not require domain knowledge and is model agnostic. It optimises over a set of diffeomorphisms; these transformations do not annihilate the output image's verisimilitude and allow the generation of high-resolution counterfactual examples. Moreover, the resulting optimisation problem is numerically affordable in a few seconds using a GPU and precludes resorting to computationally expensive generative models whilst generating anatomically plausible images.

This paper evaluates this idea using 3D volumetric structural MR images, with the following assumptions:

- The disease continuum can be modelled by smooth and invertible mappings.
- The mapping from the healthy control to the disease group can be generated using a sufficiently refined Free-Form Deformation (FFD).

The rationale behind generating a new anatomically plausible image using smooth deformations is motivated by numerous medical imaging pipelines that utilise those diffeomorphic mappings to compare, segment and aggregate different measurements between patients [43, 17]. The particular choice towards FFD to parameterise diffeomorphisms is guided by their conciseness, allowing for easier optimisation problems over a set of diffeomorphisms.

3. Method

Given a DNN f_θ, DBCE smoothly deforms an input image $\mathbf{x}$ to produce a counterfactual example $\tilde{\mathbf{x}}$. The general flowchart diagram is presented in Figure 2. More specifically, consider a trained DNN $f_\theta : \mathbb{R}^{H\times W\times D} \mapsto \mathbb{R}$ and an input image $\mathbf{x} \in \mathbb{R}^{H\times W\times D}$ for which one wants to provide a saliency map for the decision $\mathbf{p}$. Given the input $(f_\theta, \mathbf{x})$, DBCE seeks to optimize a smooth deformation $\mathcal{T}_\Phi(\bullet)$ to produce a counterfactual prediction $f_\theta(\tilde{\mathbf{x}}) = \bar{\mathbf{p}}$ from the deformed image $\tilde{\mathbf{x}} = \mathcal{T}_{\Phi^\star}(\mathbf{x})$.

A byproduct of using this intuitive set of deformations for medical images is that one can efficiently compute the distance between two anatomies as the energy of the deformation provided by $\mathcal{T}_{\Phi^\star}$.

Parameterisation of the deformation. To produce smooth deformations that are invertible and anatomically plausible, we make use of a FFD parameterised by a grid of points $\Phi \in \mathbb{R}^{N\times N\times N\times 3}$ [47]. More generally, $\mathcal{T}_\Phi(\bullet)$ is the re-sampled digital image determined by the deformation vector field $v_\Phi : \mathbb{R}^3 \mapsto \mathbb{R}^3$ which is defined as,

$$v_\Phi(x_1, x_2, x_3) = \mathbf{Id}_{\mathbb{R}^3}(x_1, x_2, x_3) + u_\Phi(x_1, x_2, x_3), \quad (1)$$

with $\mathbf{Id}$ the identity map, and with u_Φ defined as,

$$u_\Phi(x_1, x_2, x_3) = \sum_{l,m,n=0}^{3} \beta_l(u)\beta_m(v)\beta_n(w)\phi_{i+l,j+m,k+n}, \quad (2)$$

where (i, j, k) are the local indexes within the FFD grid and (u, v, w) their relative position. For instance along the first dimension $i = \lfloor \frac{x_1}{N-1} \rfloor$ and $u = \frac{x_1}{N-1} - (i + 1)$. β_i denotes the polynomial decomposition of the third order B-spline function [34] with,

$$\beta_0(t) = \frac{1}{6}(1-t)^3$$
$$\beta_1(t) = \frac{1}{6}\left(3t^3 - 6t^2 + 4\right)$$
$$\beta_2(t) = \frac{1}{6}\left(-3t^3 + 3u_i^2 + 3t + 1\right)$$
$$\beta_3(t) = \frac{1}{6}t^3.$$

DBCE's optimisation algorithm. Given the aforementioned FFDs, we can now describe the optimisation problem that DBCE seeks to solve,

$$\Phi^\star = \underset{\substack{\Phi \\ f_\theta(\mathbf{x})f_\theta(\mathcal{T}_\Phi(\mathbf{x}))\geq 0}}{\arg\min} \quad \operatorname{sign}\left(f_\theta(\mathbf{x})\right) \quad f_\theta\left(\mathcal{T}_\Phi(\mathbf{x})\right) + \sum_{q=1}^{3} \lambda_q \mathcal{R}_q(\Phi). \quad \text{(Opt-DBCE)}$$

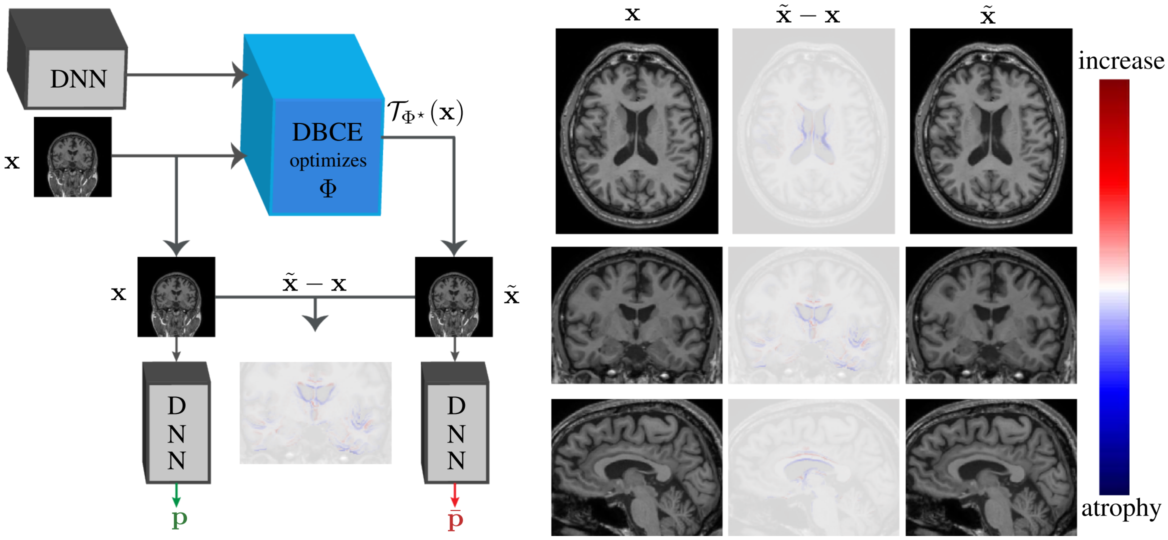

Figure 2: Left: DBCE method: given a Deep Neural Network (DNN), an input image $\mathbf{x}$ and its prediction $\mathbf{p}$, DBCE seeks to optimize the FFD grid-points Φ such that when the warped image $\tilde{x}$ is fed to the network the prediction is swapped. Right: Difference between the original image and the image perturbed by DBCE.

Suppose that $f_\theta : \mathbb{R}^{H \times W \times D} \mapsto [-1, 1]$ is a binary classifier occulting from the regularisation terms $\mathcal{R}$, the solution is met when $f_\theta(\mathcal{T}_\Phi(\mathbf{x})) = 0$ for a modified image $\mathcal{T}_\Phi(\mathbf{x})$ on the decision boundary of f_θ. We compute (Opt-DBCE) using a gradient descent approach, as described by Algorithm 1.

Algorithm 1 DBCE Algorithm

Input Image $\mathbf{x}$, model f_θ

Output Image $\tilde{\mathbf{x}}$

Initalize $i \leftarrow 0,\ \mathbf{x}_0 \leftarrow \mathbf{x},\ \Phi_0 \leftarrow \mathbf{0}_{\mathbb{R}^{N \times N \times N \times 3}}$

$\bullet_i$ denotes the variable $\bullet$ at the i-th iteration.

while $f_\theta(\mathbf{x}) f_\theta(\mathbf{x}_i) > 0$ **do**

  $\mathcal{L}_i \leftarrow \text{sign}(f_\theta(\mathbf{x})) f_\theta(\mathbf{x}_i) + \sum_{q=1}^{r} \lambda_q \mathcal{R}_q(\Phi_i)$

  $s_i \leftarrow \nabla_{\Phi_i} \mathcal{L}_i$

  $\Phi_{i+1} \leftarrow \Phi_i - \tau s_i$ ▷ Gradient step

  $\mathbf{x}_{i+1} \leftarrow \mathcal{T}_{\Phi_{i+1}}(\mathbf{x}_i)$ ▷ Image update

  $i \leftarrow i + 1$

end while

return $\tilde{\mathbf{x}} = \mathbf{x}_i$

Penalties. To meet the assumptions presented at the end of Section 2, one has to enforce constraints on the deformation $\mathcal{T}_\Phi$. This is implemented by adding penalty terms to the loss function.

The first restriction considered is the local invertibility of the mapping provided by the FFD. We borrow the quadratic penalisation proposed by [9],

$$\begin{aligned}
&\mathcal{R}_1(\Phi) = \\
&\frac{1}{N^3} \sum_{d=1}^{3} \sum_{i,j,k=1}^{N} \Big(p(\phi_{i+1,j,k,d} - \phi_{i,j,k,d}; \chi_1^{d,x_1}, \chi_2^{d,x_1}) \\
&\qquad + p(\phi_{i,j+1,k,d} - \phi_{i,j,k,d}; \chi_1^{d,x_2}, \chi_2^{d,x_2}) \\
&\qquad + p(\phi_{i,j,k+1,d} - \phi_{i,j,k,d}; \chi_1^{d,x_3}, \chi_2^{d,x_3}) \Big),
\end{aligned}$$

with p a quadratic penalty function defined as,

$$p(t; \chi_1, \chi_2) = \begin{cases} \frac{1}{2}(t - \chi_1)^2 & \text{if } t < \chi_1 \\ \frac{1}{2}(t - \chi_2)^2 & \text{if } t > \chi_2 \\ 0 & \text{otherwise.} \end{cases}$$

This ensures that the deformation v_ϕ is invertible and averts the creation of foldings and singularity points. It is here where anatomical information is lost and artifacts can be created.

The second constraint is an elementary L^1-regularisation. For height H, width W and depth D of the image, we enforce the sparsity of the transformation via,

$$\mathcal{R}_2(\Phi) = \frac{1}{HWD} \|u_\Phi\|_1. \tag{3}$$

Finally, we enforce the support of u to be restrained to $\Omega \subset \mathbb{R}^3$,

$$\mathcal{R}_3(\Phi) = \frac{1}{N^3} \|(1 - \mathcal{M}_\Omega) \odot \Phi\|_1, \tag{4}$$

where $\odot$ denotes the Hadamard product and $\mathcal{M}_\Omega$ is a discrete binary mask down-sampled to the grid points' resolution. This term allows us to localise the transformation

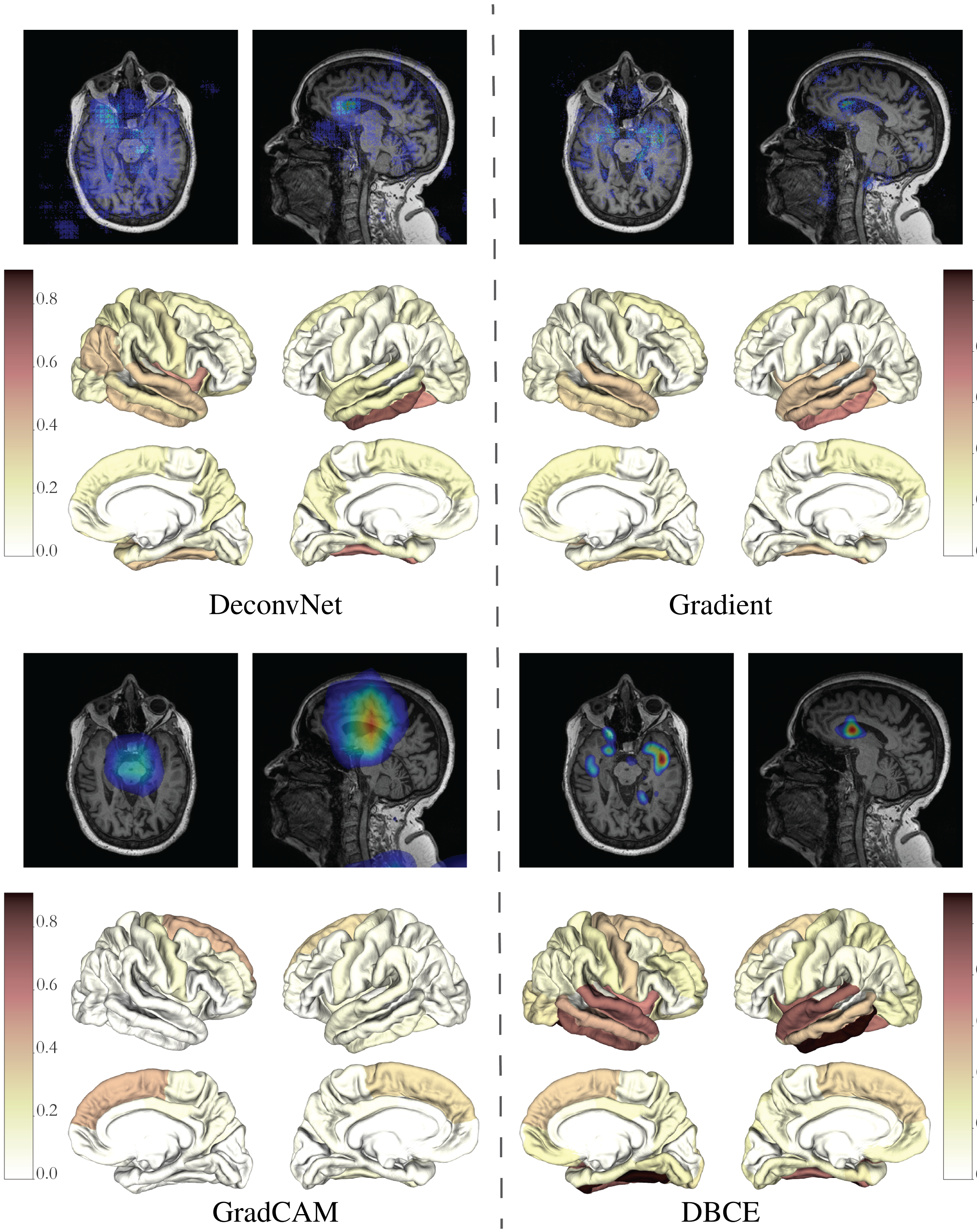

Figure 3: Saliency maps for each method of an axial and a sagittal view for a single prediction. Percentage of total saliency map averaged for all of the scans present in the experimental split and correctly classified as HC.

to specific anatomic regions of the input image. Throughout Section 4 and 5 (except for Section 5.2) we restrict Ω to the patient's brain. This ensures that deformations occur in parts of the patient's anatomy that are susceptible in changing the diagnosis.

Implementation details. DBCE is model agnostic and is not restricted to only convolutional architectures [68, 52, 32] but is available for any differentiable architecture. We efficiently implement the FFD by leveraging strided transposed convolutions resulting in a small memory footprint of 4.92GiB and a fast runtime of 0.256 ± 0.008s per gradient step. Algorithm 1 converges in 32.9 ± 13.4 iterations (≈ 8.45s) for a $256 \times 256 \times 256$ image.

4. Experiments

Early diagnosis of Alzheimer's Disease (AD) is a crucial task in neuroscience. The sooner the disease is detected, the more time is given for medical intervention and improving the patient's quality of life [2]. Promising deep-learning-based approaches have been devised for the early detection of AD [27, 63]. However, the lack of interpretation for their decisions slows down their implementation in clinical practice, and it is difficult to derive new knowledge from those models. In this section, we evaluate DBCE for a DNN, which, given a three-dimensional T1w structural image, predicts if the patient is healthy or has Alzheimer's disease. The architecture is inspired by a lightweight CNN [23]. We train this model using early stopping on an augmented version of the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset with a train/validation/test split of 776/342/520 images. We follow the procedure described in [67] to prevent any data leakage. More details on the architecture and dataset splits are provided in the supplementary materials.

4.1. A new visualisation technique provided by DBCE

The counterfactual images $\tilde{\mathbf{x}}$ provided by Algorithm 1 are very close to the initial images $\mathbf{x}$. Indeed, the global deformation of the FFD defined in Equation (1) is set to be the identity. Moreover, the regularisation term in Equation (3) ensures the deformation is both sparse and has small energy. As a result, the difference of $\mathbf{x} - \tilde{\mathbf{x}}$ for a given patient can be used to display voxel intensity variations between the original and the delusive image accurately. In addition, the brain structures of adversarial examples are aligned to the original patient, which accord with the use of Voxel-Based Morphometry approaches [5]. This method is also suitable for *"gifsplanation"* [10]. We attached such videos in our supplementary material. As depicted in Figure 6, DBCE appears to modify the size of the ventricles and the thickness of the temporal cortices and the hippocampal formation.

4.2. Qualitative analysis of DBCE's saliency map and comparison to pre-existing methods

In this section, we evaluate the saliency maps derived from the local norm of the transformation produced by DBCE against the Gradient method [56], DeconvNet [68], and GradCam [51]—three standard techniques for post-hoc interpretability.[1]

In Figure 3, we display saliency maps for each of the methods when considering images that are correctly classified as a Healthy Control (HC). Saliency maps for DeconvNet and Gradient are very scattered and diffused, thus challenging to interpret. On the contrary, maps issued from GradCAM are very evasive, highlighting a large zone that leaves room for interpretation bias. Finally, we propose to compare the methods by computing the averaged percentage of energy present in the patient's cortical sub-structures using FreeSurfer's atlas [14]. The saliency map derived from DBCE indicates that most of the deformations are taking place in the temporal lobes, which are discriminating regions for diagnosing AD [28].

5. Accessing the trustworthiness of the proposed method

The validation of saliency maps is a non-trivial topic as no ground truth exists, and the results could depend on the neural network architecture used. In this section, we will present different experiments that assess the reliability and usefulness of the saliency maps produced by DBCE.

5.1. Saliency maps repeatability

One desirable characteristic of a saliency method is its repeatability. When training the same architecture with different initialisations, one hopes to converge towards models that have learned the same patterns in the data [4]. Different saliency maps produced for different models should be comparable for the same input image. To access this behaviour, we propose to compare the resulting saliency maps for four different initialisations (repetitions) of the investigated network [23], and for saliency maps produced by different saliency methods [56, 68, 51, 59, 7, 29].

[1] Implementation recovered from TorchRay `https://facebookresearch.github.io/TorchRay/`

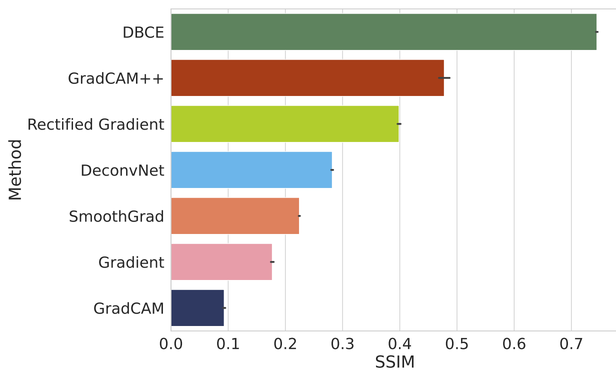

Figure 4: SSIM computed between different repetitions.

Following [1], we use the structural similarity index (SSIM) and the Pearson's correlation of the histogram of gradients (HOGs) to compute the repeatability of two saliency heatmaps for two different models on the same image. We used the implementations provided by scikit-image toolbox [65] for the computation of the HOGs. We computed the gradient deep-wise with $(16, 16)$ pixels per cell and concatenated all of the resulting histograms before computing the Pearson's correlation. The quantitative comparison of the method is reported in Figures 4 and 5.

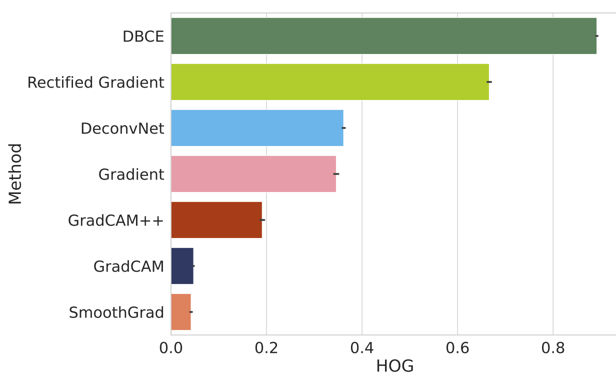

Figure 5: Pearson's correlation of HOG features.

According to [4], intra-architecture repeatability is achieved for DBCE with a structural similarity index greater than 0.5 (0.745 ± 0.039).

5.2. Robustness to localised deformations

We propose to evaluate the robustness of a prediction to deformations carried out on individual anatomical substructures. Using the segmentation masks of FreeSurfer [14], we modify Algorithm 1 to restrict the deformation sequentially to every parcellation from this atlas. In Figure 6, we report the success rate in producing such a deformation that changes the DNN's prediction from Healthy Control (HC) to Alzheimer's Disease (AD).

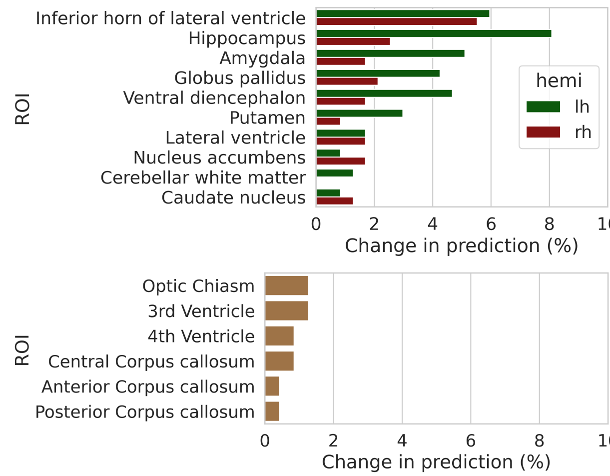

Figure 6: Top 15 Anatomical zones that when modified by DBCE can result in a change of prediction of the DNN.

Surprisingly, few of these deformations manage to change the prediction initially made by the neural network. This experiment suggests that the decision taken by a lightweight CNN [23] stems from a combination of different image features that are spatially disjoint. Similarly, this result indicates that smoothly deforming local features in isolation (a scenario that is not anatomically plausible and not described in the dataset) is not likely to tamper with the prediction of a CNN. This result could seem confusing in comparison to ultra-localized adversarial attacks where one-pixel change affects the whole prediction [62]. Nevertheless, one has to recall that by construction, FFD will continuously change the value of the input image, whereas flipping a well-chosen pixel value can create significant discontinuity that can brutally affect the value of the downstream feature maps. In addition, DBCE highlights an asymmetry in the decision made by the DNN, which might reflect an asymmetry in atrophy due to AD or a bias in the model [46, 64].

5.3. Intra-class reproducibility

The display of individual examples does not allow for a general grasp of one's network behaviour. A pleasant characteristic of numerous medical conditions is that for a given diagnosis, the causes of the disease should be similar across the whole cohort. For the particular case of AD, the disease's development is associated with atrophies of the temporal cortices and the hippocampal areas [15, 41]. In order to evaluate across all of the images with a similar prognosis, we propose to visualise the averaged absolute difference between the initial image and the deformed image produced

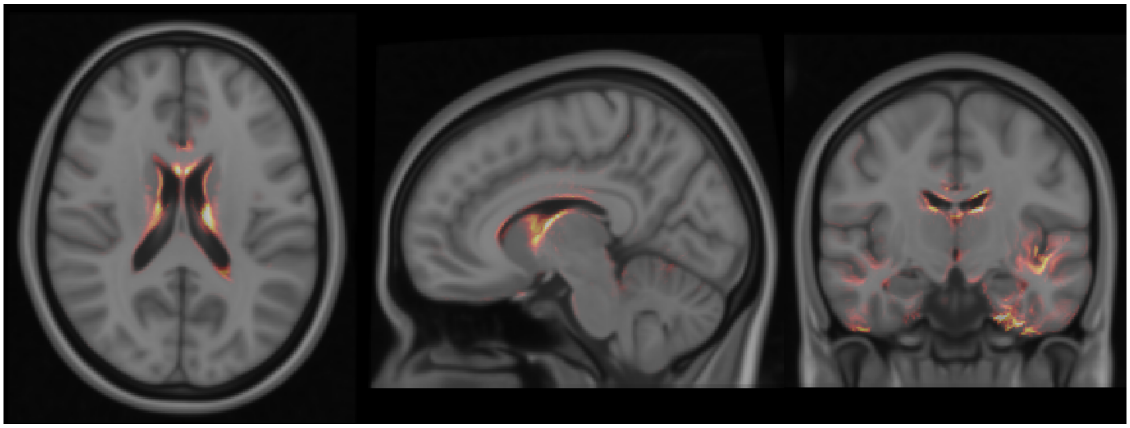

Figure 7: Absolute difference registered towards a brain atlas, showing deformations that are localised on specific anatomical regions.

by DBCE. To unify this result across a broad range of patients, we perform a non-rigid registration [36] towards an averaged brain-atlas [24].

As depicted in Figure 7, the absolute difference averaged across 235 patients tends to be localised in the temporal and the ventricle regions. The frontal, parietal and occipital areas are not highlighted by DBCE, which appear to be coherent with the current understanding of AD [25, 50, 28].

In addition, this experiment illustrates the robustness of the presented method. Indeed, for different anatomies and MR images, the presented method tends to consistently highlight disease-specific anatomical areas. In our supplementary material, we provide additional experiments on the robustness of our approach to additive Gaussian noises.

6. Conclusion

The adoption and advancement of deep-learning methods applied to healthcare applications will hinge on researchers' efforts to provide robust analysis and explanation methods that reduce the opaqueness of DNNs. This paper introduces DBCE, a generation-based interpretability method based on smooth deformations of the input image, available in 3D, and which does not require additional training data or the use of generative models. In contrast, the generation of anatomically plausible images relies on the resolution of an optimisation problem. To ensure deformations are anatomically plausible, we derive three penalty functions that allow one to tune the sparsity, localisation and invertibility of those deformations. Adjacently, we introduce two visualisation techniques for this method that monitors the voxel changes in intensity between the original and counterfactual image, or the local energy of the deformation. Finally, qualitative and quantitative tests are described to evaluate the usefulness and trustworthiness of our approach.

7. Compliance with Ethical Standards

This research was approved by CSIRO ethics 2021 068 LR

8. Acknowledgements

This work was funcded in part through an Australian Department of Industry, Energy and Resources CRC-P project between CSIRO, Maxwell Plus and I-Med Radiology Network.

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018.

[2] Ane Alberdi, Asier Aztiria, and Adrian Basarab. On the early diagnosis of alzheimer's disease from multimodal signals: A survey. *Artificial intelligence in medicine*, 71:1–29, 2016.

[3] David Alvarez-Melis and Tommi S Jaakkola. On the robustness of interpretability methods. *arXiv:1806.08049*, 2018.

[4] Nishanth Arun, Nathan Gaw, Praveer Singh, Ken Chang, Mehak Aggarwal, Bryan Chen, Katharina Hoebel, Sharut Gupta, Jay Patel, Mishka Gidwani, et al. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiology: Artificial Intelligence*, 3(6):e200267, 2021.

[5] John Ashburner and Karl J Friston. Voxel-based morphometry—the methods. *Neuroimage*, 11(6):805–821, 2000.

[6] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

[7] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolu-

tional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE, 2018.

[8] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019.

[9] Se Young Chun and Jeffrey A Fessler. A simple regularizer for b-spline nonrigid image registration that encourages local invertibility. *IEEE journal of selected topics in signal processing*, 3(1):159–169, 2009.

[10] Joseph Paul Cohen, Rupert Brooks, Sovann En, Evan Zucker, Anuj Pareek, Matthew P Lungren, and Akshay Chaudhari. Gifsplanation via latent shift: a simple autoencoder approach to counterfactual generation for chest x-rays. In *(MIDL)*, pages 74–104. PMLR, 2021.

[11] Joseph Paul Cohen, Tianshi Cao, Joseph D Viviano, Chin-Wei Huang, Michael Fralick, Marzyeh Ghassemi, Muhammad Mamdani, Russell Greiner, and Yoshua Bengio. Problems in the deployment of machine-learned models in health care. *CMAJ*, 193(35):E1391–E1394, 2021.

[12] Yin Dai, Yifan Gao, and Fayu Liu. Transmed: Transformers advance multi-modal medical image classification. *Diagnostics*, 11(8):1384, 2021.

[13] Susanne Dandl, Christoph Molnar, Martin Binder, and Bernd Bischl. Multi-objective counterfactual explanations. In *International Conference on Parallel Problem Solving from Nature*, pages 448–469. Springer, 2020.

[14] Rahul S Desikan, Florent Ségonne, Bruce Fischl, Brian T Quinn, Bradford C Dickerson, Deborah Blacker, Randy L Buckner, Anders M Dale, R Paul Maguire, Bradley T Hyman, et al. An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *Neuroimage*, 31(3):968–980, 2006.

[15] Charles Duyckaerts, Benoît Delatour, and Marie-Claude Potier. Classification and basic pathology of alzheimer disease. *Acta neuropathologica*, 118(1):5–36, 2009.

[16] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1625–1634, 2018.

[17] Bruce Fischl. Freesurfer. *Neuroimage*, 62(2):774–781, 2012.

[18] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *(ICCV)*, pages 2950–2958, 2019.

[19] Asma Ghandeharioun, Been Kim, Chun-Liang Li, Brendan Jou, Brian Eoff, and Rosalind W Picard. Dissect: Disentangled simultaneous explanations via concept traversals. *arXiv preprint arXiv:2105.15164*, 2021.

[20] Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L Beam. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3(11):e745–e750, 2021.

[21] Behnaz Gheflati and Hassan Rivaz. Vision transformer for classification of breast ultrasound images. *arXiv:2110.14731*, 2021.

[22] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *Proceedings of the AAAI*, volume 33, pages 3681–3688, 2019.

[23] Weikang Gong, Christian F Beckmann, Andrea Vedaldi, Stephen M Smith, and Han Peng. Optimising a simple fully convolutional network for accurate brain age prediction in the pac 2019 challenge. *Frontiers in Psychiatry*, 12, 2021.

[24] GŘnther Grabner, Andrew L Janke, Marc M Budge, David Smith, Jens Pruessner, and D Louis Collins. Symmetric atlasing and model based segmentation: an application to the hippocampus in older adults. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 58–66. Springer, 2006.

[25] Clifford R Jack, Ronald C Petersen, Peter C O'brien, and Eric G Tangalos. Mr-based hippocampal volumetry in the diagnosis of alzheimer's disease. *Neurology*, 42(1):183–183, 1992.

[26] Jeya Vikranth Jeyakumar, Joseph Noor, Yu-Hsi Cheng, Luis Garcia, and Mani Srivastava. How can i explain this to you? an empirical study of deep neural network explanation methods. *Advances in Neural Information Processing Systems*, 33:4211–4222, 2020.

[27] Taeho Jo, Kwangsik Nho, and Andrew J Saykin. Deep learning in alzheimer's disease: diagnostic classification and prognostic prediction using neuroimaging data. *Frontiers in aging neuroscience*, 11:220, 2019.

[28] Ronald J Killiany, Mark B Moss, Marilyn S Albert, Tamas Sandor, James Tieman, and Ferenc Jolesz. Temporal lobe regions on magnetic resonance imaging identify patients with early alzheimer's disease. *Archives of neurology*, 50(9):949–954, 1993.

[29] Beomsu Kim, Junghoon Seo, Seunghyeon Jeon, Jamyoung Koo, Jeongyeol Choe, and Taegyun Jeon. Why are saliency maps noisy? cause of and solution to noisy saliency maps. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 4149–4157. IEEE, 2019.

[30] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018.

[31] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning*, pages 5338–5348. PMLR, 2020.

[32] Salamata Konate, Léo Lebrat, Rodrigo Santa Cruz, Pierrick Bourgeat, Vincent Dorê, Jurgen Fripp, Andrew Bradley, Clinton Fookes, and Olivier Salvado. Smocam: Smooth conditional attention mask for 3d-regression models. In *2021 IEEE (ISBI)*, pages 362–366. IEEE, 2021.

[33] Alexey Kurakin, Ian Goodfellow, Samy Bengio, et al. Adversarial examples in the physical world, 2016.

[34] Seungyong Lee, G. Wolberg, Kyung-Yong Chwa, and Sung Yong Shin. Image metamorphosis with scattered feature constraints. 2(4):337–354, 1996.

[35] Yuan Liu, Ayush Jain, Clara Eng, David H Way, Kang Lee, Peggy Bui, Kimberly Kanada, Guilherme de Oliveira Marinho, Jessica Gallegos, Sara Gabriele, et al. A deep learning

system for differential diagnosis of skin diseases. *Nature medicine*, 26(6):900–908, 2020.

[36] Marc Modat, Gerard R Ridgway, Zeike A Taylor, Manja Lehmann, Josephine Barnes, David J Hawkes, Nick C Fox, and Sébastien Ourselin. Fast free-form deformation using graphics processing units. *Computer methods and programs in biomedicine*, 98(3):278–284, 2010.

[37] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern recognition*, 65:211–222, 2017.

[38] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582, 2016.

[39] Cristian Navarrete-Dechent, Stephen W Dusza, Konstantinos Liopyris, Ashfaq A Marghoob, Allan C Halpern, and Michael A Marchetti. Automated dermatological diagnosis: hype or reality? *The Journal of investigative dermatology*, 138(10):2277, 2018.

[40] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *arXiv:1806.07421*, 2018.

[41] Vincent Planche, José V Manjon, Boris Mansencal, Enrique Lanuza, Thomas Tourdias, Gwenaëlle Catheline, and Pierrick Coupé. Structural progression of alzheimer's disease over decades: the mri staging scheme. *Brain Communications*, 4(3):fcac109, 2022.

[42] Ryan Poplin, Avinash V Varadarajan, Katy Blumer, Yun Liu, Michael V McConnell, Greg S Corrado, Lily Peng, and Dale R Webster. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical Engineering*, 2(3):158–164, 2018.

[43] Gheorghe Postelnicu, Lilla Zollei, and Bruce Fischl. Combined volumetric and surface registration. *IEEE transactions on medical imaging*, 28(4):508–522, 2008.

[44] Omid Poursaeed, Isay Katsman, Bicheng Gao, and Serge Belongie. Generative adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4422–4431, 2018.

[45] Marco Tulio Ribeiro, Sameer Singh, and Carlos 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*, pages 1135–1144, 2016.

[46] James M Roe, Didac Vidal-Piñeiro, Øystein Sørensen, Andreas M Brandmaier, Sandra Düzel, Hector A Gonzalez, Rogier A Kievit, Ethan Knights, Simone Kühn, Ulman Lindenberger, et al. Asymmetric thinning of the cerebral cortex across the adult lifespan is accelerated in alzheimer's disease. *Nature communications*, 12(1):1–11, 2021.

[47] Torsten Rohlfing, Calvin R Maurer, David A Bluemke, and Michael A Jacobs. Volume-preserving nonrigid registration of mr breast images using free-form deformation with an incompressibility constraint. *IEEE trans. on medical imaging*, 22(6):730–741, 2003.

[48] Adriel Saporta, Xiaotong Gui, Ashwin Agrawal, Anuj Pareek, SQ Truong, CD Nguyen, Van-Doan Ngo, Jayne Seekins, Francis G Blankenberg, Andrew Ng, et al. Deep learning saliency maps do not accurately highlight diagnostically relevant regions for medical image interpretation. *MedRxiv*, 2021.

[49] Rory Sayres, Ankur Taly, Ehsan Rahimy, Katy Blumer, David Coz, Naama Hammel, Jonathan Krause, Arunachalam Narayanaswamy, Zahra Rastegar, Derek Wu, et al. Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy. *Ophthalmology*, 126(4):552–564, 2019.

[50] Philip Scheltens, D Leys, F Barkhof, D Huglo, HC Weinstein, P Vermersch, M Kuiper, M Steinling, E Ch Wolters, and J Valk. Atrophy of medial temporal lobes on mri in" probable" alzheimer's disease and normal ageing: diagnostic value and neuropsychological correlates. *Journal of Neurology, Neurosurgery & Psychiatry*, 55(10):967–972, 1992.

[51] 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 *Proceedings of the IEEE (ICCV)*, pages 618–626, 2017.

[52] Ramprasaath R Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? *arXiv:1611.07450*, 2016.

[53] Laleh Seyyed-Kalantari, Haoran Zhang, Matthew McDermott, Irene Y Chen, and Marzyeh Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature medicine*, 27(12):2176–2182, 2021.

[54] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMLR, 2017.

[55] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv:1312.6034*, 2013.

[56] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *In Workshop at International Conference on Learning Representations*. Citeseer, 2014.

[57] Sumedha Singla, Brian Pollack, Junxiang Chen, and Kayhan Batmanghelich. Explanation by progressive exaggeration. *arXiv preprint arXiv:1911.00483*, 2019.

[58] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 180–186, 2020.

[59] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

[60] Simeon Spasov, Luca Passamonti, Andrea Duggento, Pietro Lio, Nicola Toschi, Alzheimer's Disease Neuroimaging Initiative, et al. A parameter-efficient deep learning approach

to predict conversion from mild cognitive impairment to alzheimer's disease. *Neuroimage*, 189:276–287, 2019.

[61] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv:1412.6806*, 2014.

[62] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, 2019.

[63] Ali Haidar Syaifullah, Akihiko Shiino, Hitoshi Kitahara, Ryuta Ito, Manabu Ishida, and Kenji Tanigaki. Machine learning for diagnosis of ad and prediction of mci progression from brain mri using brain anatomical analysis using diffeomorphic deformation. *Frontiers in Neurology*, 11:576029, 2021.

[64] Elina Thibeau-Sutre, Olivier Colliot, Didier Dormont, and Ninon Burgos. Visualization approach to assess the robustness of neural networks for medical image classification. In *Medical Imaging 2020: Image Processing*, volume 11313, page 113131J. International Society for Optics and Photonics, 2020.

[65] Stefan Van der Walt, Johannes L Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D Warner, Neil Yager, Emmanuelle Gouillart, and Tony Yu. scikit-image: image processing in python. *PeerJ*, 2:e453, 2014.

[66] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

[67] Junhao Wen, Elina Thibeau-Sutre, Mauricio Diaz-Melo, Jorge Samper-González, Alexandre Routier, Simona Bottani, Didier Dormont, Stanley Durrleman, Ninon Burgos, Olivier Colliot, et al. Convolutional neural networks for classification of alzheimer's disease: Overview and reproducible evaluation. *Medical image analysis*, 63:101694, 2020.

[68] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *(ECCV)*, pages 818–833. Springer, 2014.