



OPEN ACCESS

EDITED BY

Georgios Leontidis,
University of Aberdeen, United Kingdom

REVIEWED BY

Ibrahim Kandel,
Houston Methodist Research Institute,
United States
Takashi Kuremoto,
Nippon Institute of Technology, Japan

*CORRESPONDENCE

Stefan Haufe
✉ haufe@tu-berlin.de

RECEIVED 31 October 2023

ACCEPTED 09 February 2024

PUBLISHED 26 February 2024

CITATION

Oliveira M, Wilming R, Clark B, Budding C,
Eitel F, Ritter K and Haufe S (2024)
Benchmarking the influence of pre-training
on explanation performance in MR image
classification. *Front. Artif. Intell.* 7:1330919.
doi: 10.3389/frai.2024.1330919

COPYRIGHT

© 2024 Oliveira, Wilming, Clark, Budding,
Eitel, Ritter and Haufe. This is an open-access
article distributed under the terms of the
[Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in
other forums is permitted, provided the
original author(s) and the copyright owner(s)
are credited and that the original publication
in this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

Benchmarking the influence of pre-training on explanation performance in MR image classification

Marta Oliveira¹, Rick Wilming², Benedict Clark¹, Céline Budding³, Fabian Eitel³, Kerstin Ritter³ and Stefan Haufe^{1,2,3*}

¹Division 8.44, Physikalisch-Technische Bundesanstalt, Berlin, Germany, ²Computer Science Department, Technische Universität Berlin, Berlin, Germany, ³Berlin Center for Advanced Neuroimaging (BCAN), Charité— Universitätsmedizin Berlin, Berlin, Germany

Convolutional Neural Networks (CNNs) are frequently and successfully used in medical prediction tasks. They are often used in combination with transfer learning, leading to improved performance when training data for the task are scarce. The resulting models are highly complex and typically do not provide any insight into their predictive mechanisms, motivating the field of “explainable” artificial intelligence (XAI). However, previous studies have rarely quantitatively evaluated the “explanation performance” of XAI methods against ground-truth data, and transfer learning and its influence on objective measures of explanation performance has not been investigated. Here, we propose a benchmark dataset that allows for quantifying explanation performance in a realistic magnetic resonance imaging (MRI) classification task. We employ this benchmark to understand the influence of transfer learning on the quality of explanations. Experimental results show that popular XAI methods applied to the same underlying model differ vastly in performance, even when considering only correctly classified examples. We further observe that explanation performance strongly depends on the task used for pre-training and the number of CNN layers pre-trained. These results hold after correcting for a substantial correlation between explanation and classification performance.

KEYWORDS

XAI, explainability, interpretability, pre-training, MRI, benchmark, dataset, classification

1 Introduction

Following AlexNet’s (Krizhevsky et al., 2012) victory in the ImageNet competition, CNNs developed to become the deep neural network (DNN) architecture of choice for any image-based prediction tasks. Apart from their ingenious design, the success of CNNs was made possible by ever-growing supplies of data and computational resources. However, sufficient labeled data to train complex CNNs are not widely available for every prediction task. This is especially true for medical imaging data, which is cumbersome to acquire and underlies strict data protection regulations. To address this bottleneck, Transfer Learning (TL) techniques are frequently employed (Ardalan and Subbian, 2022). In the context of DNNs, TL strategies often consist of two steps. First, a surrogate model is trained on a different prediction tasks, for which ample training data are available. This is called pre-training. And, second, the resulting model is adapted to the prediction task of

interest, where only parts of the model's parameters are updated (Pan and Yang, 2010). This is called re-training or fine-tuning and requires smaller amounts of labeled data than training a network from scratch, leading to a less computationally expensive process. TL is, therefore, frequently employed for prediction tasks in medical imaging, where it is believed that TL techniques improves the generalization by identifying common features between the two tasks (Valverde et al., 2021). Cheng and Malhi (2017) use a DNN, trained with the ImageNet dataset (Deng et al., 2009), to classify ultrasound images into eleven categories, achieving better results than human radiologists. Another example is the reconstruction of Magnetic Resonance Imaging (MRI) data with models trained on an image corpus that was augmented with ImageNet data (Dar et al., 2020). The resulting model outperformed conventional reconstruction techniques. However, it also has been argued (Shirokikh et al., 2020) that the use of pre-trained models may not be adequate for the medical field. The main argument being that structures in medical images are very different from those observed in natural images. Hence, feature representations learned during pre-training may not be useful for solving clinical tasks.

Despite the success of DNN models, their intrinsic structure makes them hard to interpret. This challenges their real-world applicability in high-stake fields such as medicine. Although many practices in medicine are still not purely evidence-based, the risk posed by faulty algorithms is exponentially higher than that of doctor-patient interactions (Topol, 2019). Thus, it has been recognized that the working principles of complex learning algorithms need to be made transparent if such algorithms are to be used on critical human data. The General Data Protection Regulation of the European Union (GDPR, Article 15), for example, states that patients have the right to receive meaningful information about how decisions are achieved based on their data, including decisions made based on artificial intelligence algorithms, such as DNNs (European Commission, 2018).

The field of "explainable artificial intelligence" (XAI) arose to address this need. A popular class of XAI methods seek to deliver so-called local *post-hoc* explanations, which are derived from a trained model's output on a test input. These methods can be either specific to a particular architecture or type of ML model or model-agnostic, where explanations can be produced for a large variety of model architectures. The outcome of such methods is often a so-called heat map, which assigns an "importance" score to each input feature. However, despite the popularity of XAI methods, their theoretical underpinnings are far from established. Most importantly, there is no agreed upon definition of what explainability means or what XAI methods are supposed to deliver (Zucco et al., 2018). Consequently, little quantitative empirical validation of XAI methods exists (Das and Rad, 2020). This limits the utility of XAI methods for quality control purposes in critical domains such as medicine.

To date, most quantitative evaluations of XAI methods focus on secondary quality aspects such as robustness or uncertainty of explanations but spare out the fundamental issue of explanation correctness. To define a notion of correctness, it is necessary to devise working definitions of what constitutes a desirable explanation for a given input datum. Such a definition would allow one to measure the explanation performance of XAI methods using

objective metrics. Synthetic data whose data-generating process is known by construction provide such a ground truth.

In this work, we focus on the problem of classifying MR images of the human brain. We devise synthetic ground-truth data for this problem, thereby addressing the current lack of validation of model explanations in this context. Precisely, we overlay real MR images with artificial lesions of two different types, where the type of lesion defines class membership. Lesions are realistically designed to resemble white matter hyperintensities (WMH), which are important biomarkers of the aging brain and aging-related neurodegenerative disorders (Wharton et al., 2015; d'Arbeloff et al., 2019). As the positions of the class-discriminative lesions are fully known by construction this provide a ground-truth for model explanations. We provide an open code and data framework for generating MRI slices with different types of lesions and respective ground-truths.¹

In the second part of this work, we show the benchmark's utility by investigating both classification and explanation performance as a function of model pre-training. Concretely, we benchmark common XAI methods against each other and compare the explanation performance of models pre-trained using either within-domain data (using different MRI classification tasks) or out-of-domain data (using natural images from the ImageNet classification challenge, Deng et al., 2009) as well as models that have been retrained on the task of interest to varying degrees.

2 Related work

A number of recent works in the field of XAI have moved toward objective validation of XAI approaches using synthetic data. Kim et al. (2018) propose to validate XAI via surrogate ground-truth information employing so called concept activation vectors (TCAV), which are accessible with synthetic data. Yang and Kim (2019) propose a notion of relative feature importance to develop a metric to quantitatively assess methods such as TCAV. Known data generating processes are also increasingly being utilized to provide ground-truth information for model explanations (Ismail et al., 2019, 2020; Tjoa and Guan, 2020). An evaluation strategy for XAI methods proposed by Agarwal et al. (2022) leverages a scheme to generate synthetic data, where each class is represented by a unique spatial cluster in feature space, prompting options for quantitative evaluations. Utilizing a Visual Question-Answering (VQA) task, Arras et al. (2022) introduce a framework, based on synthetic data, to quantify an explanation's quality for a distinctive object of an image. Further, Hofmann et al. (2022) used XAI methods to find structural changes of the aging brain, which allowed the authors to identify white matter lesions associated to the aging brain and forms of dementia. They applied layerwise relevance propagation (LRP, Bach et al., 2015) and compared the resulting heat maps with white matter lesion maps. Cherti and Jitsev (2021) analyzed the effect of pre-training on model transfer in medical imaging but did not investigate aspects of explainability. Finally, our own

¹ Please find all code here: https://github.com/Marta54/Pretrain_XAI_gt, and all benchmark data here: <https://www.doi.org/10.17605/OSF.IO/XNWAJ>.

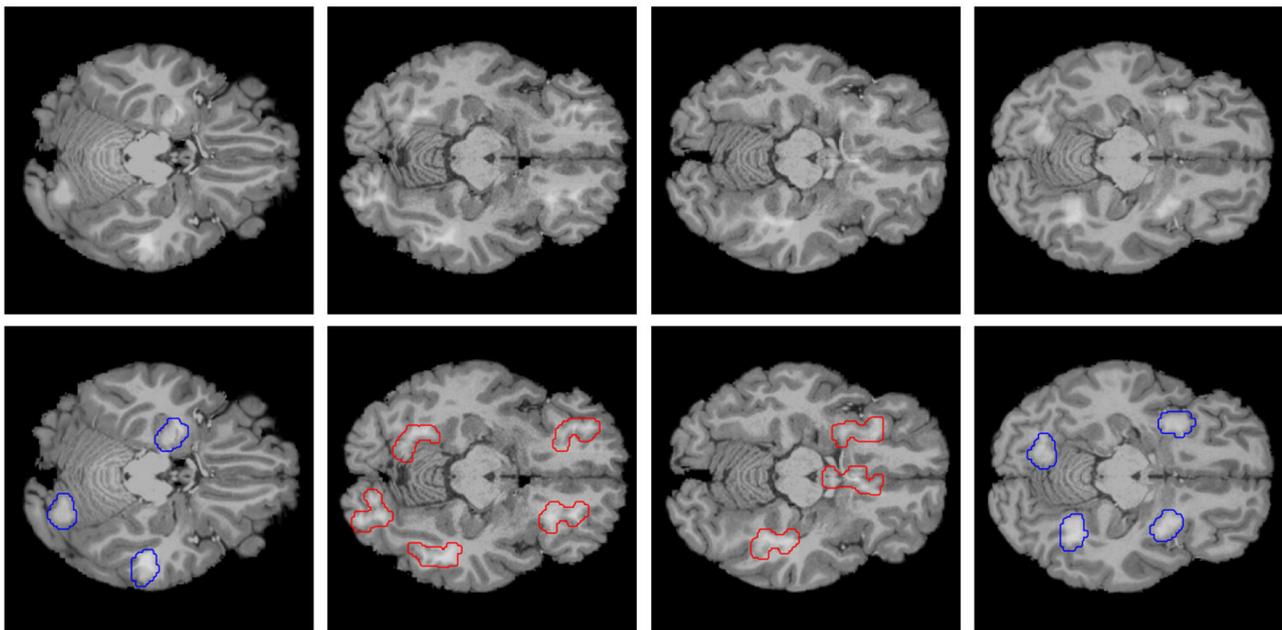


FIGURE 1

Example of images of the dataset created. The top row consists of axial MRI slices from the Human Connectome Project (HCP, Van Essen et al., 2013) healthy brain dataset, with artificial lesions added. The bottom row consists of the top row images, but with the position of the regular lesion contoured in blue and the irregular lesions contoured in red, forming the ground-truth for an explanation.

work has highlighted common misinterpretations of XAI methods. Using counterexamples and analytical derivations, Haufe et al. (2014) and Wilming et al. (2023) demonstrated that many popular XAI methods systematically attribute importance to so-called suppressor variables (Conger, 1974; Friedman and Wall, 2005), which are beneficial to the model's performance due to statistical correlations with other, informative features, but are themselves statistically unrelated to the predicted target variable. This undesirable effect was shown to be present even for linear models often assumed to be "intrinsically interpretable" (Rudin, 2019), and is incentivized by current algorithmic operationalizations of explanation correctness such as faithfulness (Bach et al., 2015). We further devised low-dimensional benchmarks to study this effect and compare different model architectures and XAI methods using a theoretically well-founded data-driven definition of explanation correctness (Wilming et al., 2022; Clark et al., 2023). Wilming et al. (2022), for example, introduce a synthetic data generation process explicitly defining discriminative and class-agnostic features such as suppressor variables. However, as these works are based on low-dimensional mathematical toy problems, the emergence of suppressor variables in realistic medical use cases such as MRI classification, and their potential influence on explanation performance in such a context, has not been studied. The present work aims to fill this gap by providing a relatively realistic clinical MR image generation scheme that induces, to a certain extent, the emergence of suppressor variables. Our work thereby provides a more challenging setting than comparable studies involving ground-truth data in medical and non-medical contexts.

3 Methods

3.1 Data generation

To generate the data used for this analysis, we use 2-dimensional T1-weighted axial MRI slices from 1007 healthy adults aged between 22 and 37 years, sourced from the Human Connectome Project (HCP, Van Essen et al., 2013 see also Supplementary Section 1). The MRI data consists of 3D MRI slices pre-processed with the FSL (Jenkinson et al., 2012) and FreeSurfer (Fischl, 2012) tools as described in Jenkinson et al. (2002) and Glasser et al. (2013) and defaced as reported in Milchenko and Marcus (2013). These slices provide the background on which a random number of artificial "lesions" are overlaid. Regular and irregular lesions are generated and added to the slices. Each slice contains only one type of lesions. This defines a binary classification problem, which we solve using CNNs. The lesions are added so that the dataset created is balanced.

For this study, we keep only slices with <55% black pixels. These images are 260×311 pixels in size. To obtain square images, they are padded vertically with zeros and cropped horizontally. The final size of the slices result in background images $B \in \mathbb{R}^{270 \times 270}$, where we keep the intensity values $B_{ij} \in [0, 0.7]$ for $i, j = 1, \dots, 270$.

Artificial lesions are created from a 256×256 pixel noise image, to which a Gaussian filter with a radius of 2 pixels is applied. The Otsu method (Otsu, 1979) is used to binarize the smoothed image. After the application of the morphological operations erosion and opening, a second erosion is applied to create more irregular shapes (see Supplementary Section 3). Since these shapes occur less

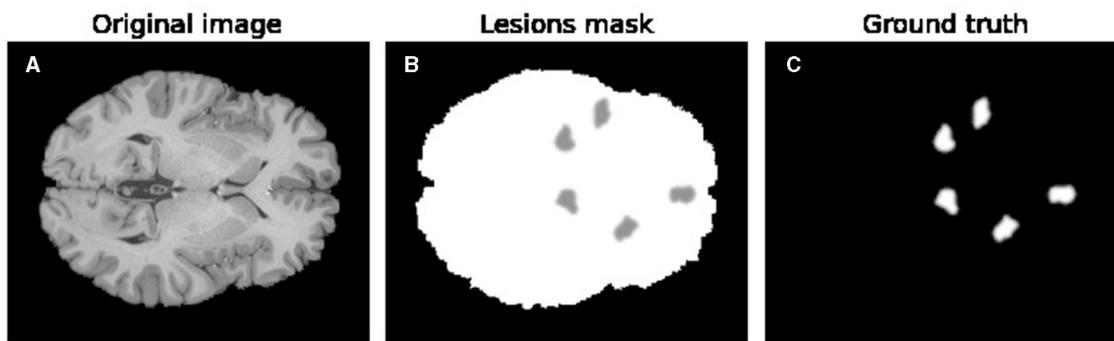


FIGURE 2

Example of the lesion creation process. (A) Depicts an original axial MRI slice from the Human Connectome Project and denoted as background image B . (B) Showscases an example lesion mask L composed of several lesions. (C) Represents ground-truth explanations used for the XAI performance study.

frequently than regular shapes, these determine the number of different noise images necessary to create a given number of lesions.

From the images obtained after the application of the morphological operations, the connected components (contiguous groups of non-zero intensity pixels fully surrounded by zero intensity pixels) are identified, which serve as lesion candidates. Further, lesions are selected based on the compactness of their shape. Here, it is sufficient to consider the isoperimetric inequality on a plane $A \leq p^2/4\pi$, where $A \in \mathbb{R}$ is the area of a particular lesion shape and $p \in \mathbb{R}$ its perimeter. The compactness is obtained by comparing the shape of the lesion candidate to a circle with the same perimeter. The larger the compactness, the rounder the shape. Here, regular lesions are required to have a compactness above 0.8 and irregular lesions have a compactness below 0.4. After selecting the lesions, they are padded with a 2-pixel margin, and a Gaussian filter with a radius of 0.75 pixels is applied to smooth the lesion boundaries. Examples of obtained lesions are displayed in Figure 1.

Three to five lesions of the same type (regular or irregular) are composed in one image $L \in \mathbb{R}^{270 \times 270}$ in random locations within the brain, without overlapping and pixel-wise multiplied with the background MRI B (see Figure 2). For the lesions we consider the intensity values $L_{ij} \in [0, w]$, where i, j correspond to pixels representing lesions. The parameter w is a constant that controls the SNR. Higher w values lead to whiter lesions and higher SNR, leading to easier classification and explanation tasks. In this study, we set $w = 0.5$. Note also, that this setup may lead to the emergence of so-called suppressor variables. These would be pixels of the background outside any lesion, which could still provide a model with information on how to remove background content from lesion areas in order to improve the model's predictions. Suppressor variables have been shown to be often misinterpreted for important class-dependent features by XAI methods (Haufe et al., 2014; Wilming et al., 2022, 2023).

In parallel to the generation of the actual synthetic MR images, the same lesions are added to a black image to create ground-truth masks. We summarize the ground-truth explanations via the set

$$\mathcal{F}_{lesions} := \{i, j \in [270] \mid L_{ij} \neq 1 \text{ or } L_{ij} \neq 0\}, \quad (1)$$

where $[270] := \{1, \dots, 270\}$. The ground-truth explanation Equation 1 is different for each image and an example of $\mathcal{F}_{lesions}$ represented as an image can be seen in Figure 2C.

Out of the 1,006 subjects in the HCP dataset, 60% were used to create the training dataset, 20% to create the validation dataset, and another 20% to create the holdout dataset, corresponding to 24,924, 8,319, and 8,319 slices, respectively.

3.2 Pre-training

We apply the XAI methods to the VGG-16 (Simonyan and Zisserman, 2014) architecture, included in the Torchvision package, version 0.12.0+cu102. Two models are pre-trained using two different corpora, and serve as starting points for our study. The first model is pre-trained using the ImageNet dataset (Deng et al., 2009) (out-of-domain pre-training). The weights used are included in the same version of Torchvision. The second model is pre-trained using MRI slices extracted from the HCP as described before but without artificial lesions (within-domain pre-training). Here, the task is to classify slices according to whether they were acquired from female or male subjects. To train the latter model, 24,924 slices are used, 46% of which belong to male subjects and 54% to female subjects. These slices are arranged into batches of 32 data points. The model is trained using stochastic gradient descent (SGD) with a learning rate (LR) of 0.02 and momentum of 0.5. The learning rate is reduced by 10% every 5 epochs. Cross-entropy is used as the loss function.

3.3 Fine-tuning

After pre-training, the models are fine-tuned layer-wise on the lesion-classification problem, with images chosen from the holdout dataset, which we split into train/validation/test again (see Supplementary Section 4). Each degree of fine-tuning includes the convolutional layers between two consecutive max-pooling layers. Thus, the five degrees of fine-tuning are: *1conv* (fine-tuning up to the first max-pooling layer), *2conv* (fine-tuning up to the second

max-pooling layer), and so on, up to *all* (fine-tuning of all VGG-16 layers). Weights in layers that are not to be fine-tuned are frozen. SGD and Cross-entropy loss with the same parameters as used for the pre-training are employed in this phase. However, several different LRs are used.

3.4 XAI methods

We apply XAI methods from the Captum library (version 0.5.0). These methods have been proposed to provide “explanations” of the models’ output in the form of a heat map $\mathbf{s} \in \mathbb{R}^{270 \times 270}$, assigning an “importance” score to each input feature of an example. We use the default settings from Captum for all XAI methods. Wherever a baseline—a reference point to begin the computation of the explanation—is needed, an all-zeros image is used. This is done for Integrated Gradients, DeepLift, and GradientSHAP. The absolute value of the obtained importance score or heat map constitutes the basis for our visualizations and quantitative explanation analyses. For visualization purposes, we further transformed the intensity of the importance scores by $-\log(1 - s_{ij}(1 - 1/b))/\log(b)$, where \log is the natural logarithm and $b = 0.5$. The XAI methods used were Integrated Gradients (Sundararajan et al., 2017), Gradient SHAP (Lundberg and Lee, 2017), LRP (Bach et al., 2015)s, DeepLIFT (Shrikumar et al., 2017), Saliency (Simonyan et al., 2013), Deconvolution (Zeiler and Fergus, 2014), and Guided Backpropagation (Springenberg et al., 2015).

3.5 Explanation performance

Our definition of quantitative explanation performance is the precision to which the generated importance or heat maps resemble the ground-truth, i.e., the location of the lesions (cf. Figure 2). It would be expected that the best explanation would only highlight the pixels of the ground-truth, since those are the ones that are relevant to the classification task at hand. We determine the explanation performance by finding the n most intense pixels $\text{Top}_{n(\mathbf{s})}$ of the heat map \mathbf{s} , where $n(\mathbf{s}) := |\mathcal{F}_{\text{lesions}}|$ is the number of pixels in the ground-truth of each image. Then we calculate the number of these pixels that were in the ground-truth (true positives). The precision or explanation performance **EP** is obtained by calculating the ratio between the true positives and all positives (the number of pixels in the ground-truth)

$$\text{EP} := \frac{|\text{Top}_{n(\mathbf{s})} \cap \mathcal{F}_{\text{lesions}}|}{|\mathcal{F}_{\text{lesions}}|}. \quad (2)$$

3.6 Baselines

The performance of each explanation is then compared to several baseline methods, which act as “null models” for explanation performance. These baselines are models that are initialized randomly and not trained (random model) and two edge detection methods, the Laplace and Sobel filters.

4 Experiments

Showcasing the proposed dataset’s utility, we fine-tune two VGG-16 models that have been previously pre-trained with the two corpora (ImageNet and MRI), to five different degrees. For each degree of fine-tuning, we fine-tuned 15 models with different seeds. Then we select the three best-performing models, where performance is measured on test data in terms of accuracy. We further analyze the model explanation performance of common XAI methods with respect to the ground-truth explanations in the form of lesion maps provided by our dataset. A reference to the Python code to reproduce our experiments is provided in the [Supplementary Section 2](#).

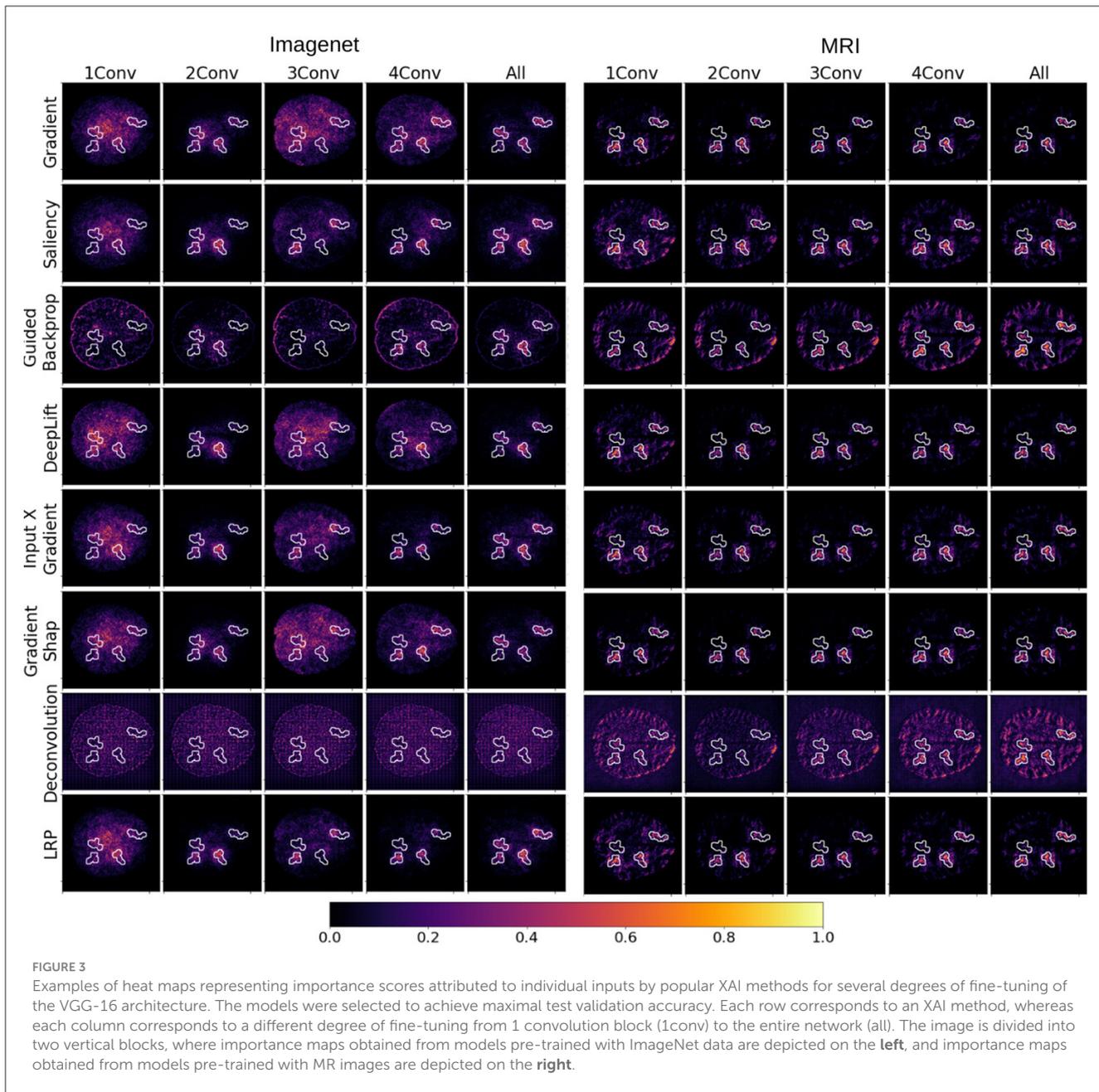
5 Results

All models, except the least fine-tuned ones (1conv), reached accuracies above 90%. The models pre-trained with ImageNet achieved higher accuracy than the ones pre-trained with MR images.

5.1 Qualitative analysis of explanations

Figure 3 displays importance heat maps for a test sample with four irregular lesions. These explanations are obtained by eight XAI methods for five degrees of fine-tuning. Plots are divided into two sections reflecting the two corpora used for pre-training (ImageNet and MRI female vs. male). The white contours in each heat map represent the ground-truth of the explanation. A good explanation should give high attribution to regions inside the white contour and low everywhere else. In this respect, most of the explanations appear to perform well, identifying most of the lesions, especially for high degrees of fine-tuning. However, the explanations generally do not highlight all of the lesions in the ground-truth. This image also shows that, for some XAI methods, the explanation may deteriorate for an intermediate degree of fine-tuning, and then improve again. This can be seen especially in the results of the model pre-trained with ImageNet data. Heat maps of the untrained baseline model are shown in [Supplementary Section 6](#).

When comparing the “explanations” obtained from models pre-trained on ImageNet data with the ones from models pre-trained on MRI data, the latter seems to contain less contamination from the structural features of the MRI background, especially for Deconvolution and Guided Backpropagation. We can further argue that some models seem to do a better job identifying the lesions than others. Particularly noisy explanations are obtained with Deconvolution, especially for models pre-trained with ImageNet data. In this case, pixels with higher importance attribution seem to form a regular grid, roughly covering the shape of the brain of the underlying MRI slice. For models pre-trained on the MRI corpus, Deconvolution is able to place higher importance within the lesions for higher degrees of fine-tuning.



5.2 Quantitative analysis of explanation performance

Figure 4 shows quantitative explanation performance Equation 2. Here, each boxplot was derived from the intersection of test images that were correctly classified by all models ($N = 2,371$). The results obtained for the edge filter baseline as well as the random baseline model are derived from the same 2,371 images. Note that the edge detection filters only depend on the given image and are independent of models and XAI methods. Thus, identical results are presented for edge filters in each subfigure. The lines in the background correspond to the average classification performance (test accuracy) of the five models for each degree of fine-tuning. The random baseline model is only one and has a test accuracy

of 50%. Interestingly, models pre-trained with ImageNet data consistently achieved higher classification performance than models pre-trained with MR images. The classification performance of the models pre-trained with MR images peaks at an intermediate degree of fine-tuning (3conv), while the models pre-trained with ImageNet improve with higher fine-tuning degrees.

In some settings, ImageNet pre-training leads to considerably worse explanation performance. This is the case for specific methods such as Deconvolution and, to some extent, Guided Backpropagation. Moreover, ImageNet pre-training leads to worse explanations across all XAI methods for lower degrees of fine-tuning (1conv and 2conv), where large parts of the models are prohibited to depart from the internal representations learned on the ImageNet data.

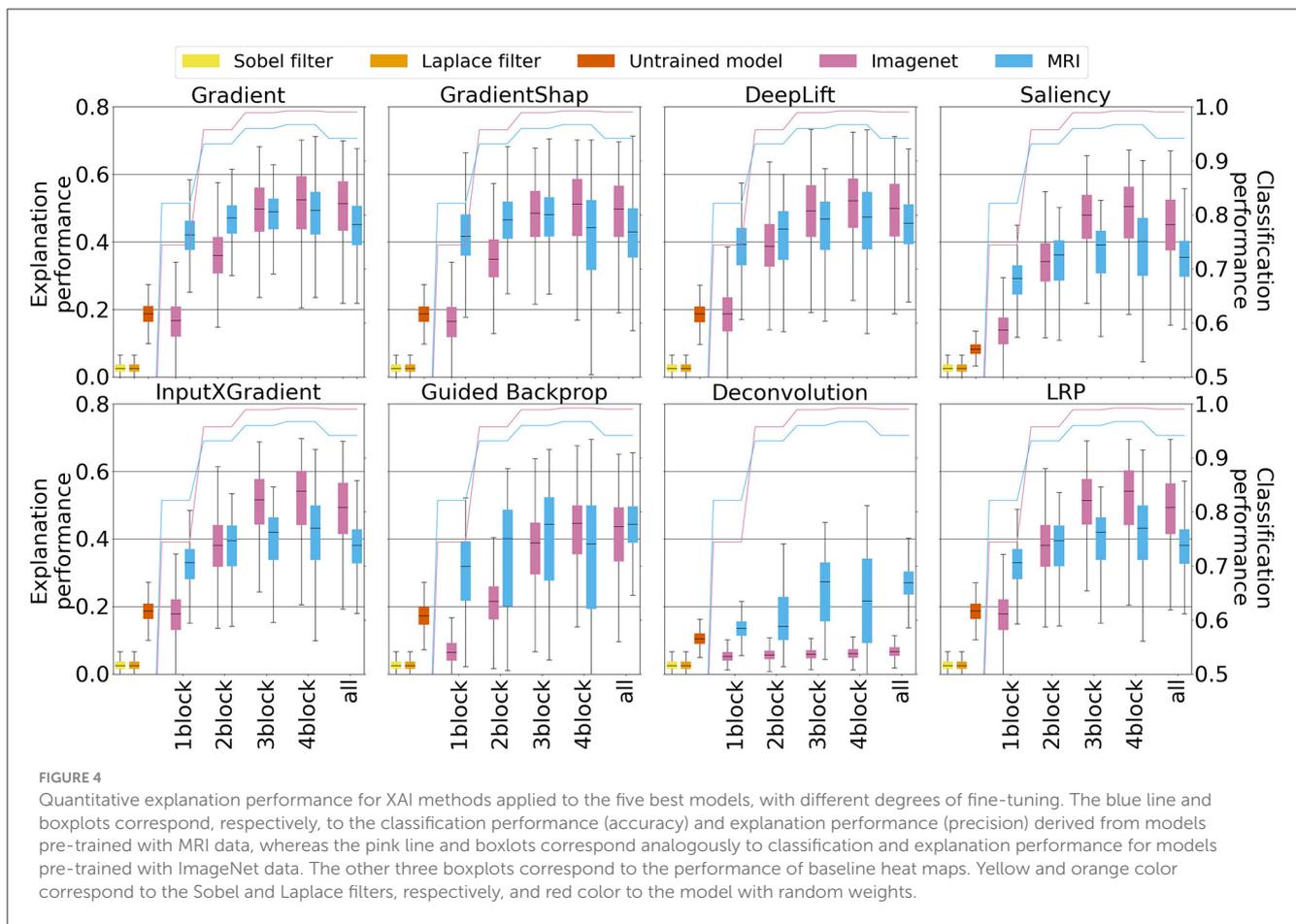


FIGURE 4
Quantitative explanation performance for XAI methods applied to the five best models, with different degrees of fine-tuning. The blue line and boxplots correspond, respectively, to the classification performance (accuracy) and explanation performance (precision) derived from models pre-trained with MRI data, whereas the pink line and boxlots correspond analogously to classification and explanation performance for models pre-trained with ImageNet data. The other three boxplots correspond to the performance of baseline heat maps. Yellow and orange color correspond to the Sobel and Laplace filters, respectively, and red color to the model with random weights.

As a function of the amount of fine-tuning, explanation performance generally increases with higher degrees of fine-tuning. However, depending on the XAI method used, and the corpus used for pre-training, this trend plateaus or even slightly reverses at a high degree of fine-tuning (4conv).

Importantly, explanation performance appears to strongly correlate with the classification performance of the underlying model. As classification accuracy could represent a potential confound to our analysis, we repeated our quantitative analysis of explanation performance based on five models with similar classification performance per pre-training corpus and degree of fine-tuning. Here, it is apparent that, when controlling for classification performance, models pre-trained on MRI data consistently outperform equally well-predicting models that were pre-trained on ImageNet data in terms of explanation performance. These results are presented in [Supplementary Section 7](#).

6 Discussion

The field of XAI has produced a plethora of methods whose goal it is to “explain” predictions performed by deep learning and other complex models, including CNNs. However, quantitative evaluations of these methods based on ground-truth data are scarce. Even if these methods are based on seemingly intuitive principles, XAI can only serve its purpose if it is itself properly

validated, which is so far not often done. The present study was designed to create a benchmark within which explanation quality can be objectively quantified. To this end, we designed a well-defined ground-truth dataset for model explanations, where we modeled artificial data to resemble the important clinical use case of structural MR image classification for the diagnosis of brain lesions. With this benchmark dataset, we propose a framework to evaluate the influence of pre-training on explanation performance.

We observed a correlation between classification accuracy and explanation performance, which could be expected since a more accurate model is likely to more successfully focus on relevant input features. Networks trained on ImageNet data may have learned representations for objects occurring only outside the domain of brain images (e.g., cats and dogs). The existence of such representations in the network seems to negatively affect XAI methods, whose importance maps are in parts derived by propagating network activations backwards through the network. Consistent with this remark is the observation that for lower degrees of fine-tuning (1conv and 2conv), the explanation quality of models pre-trained with ImageNet data is worse compared to models pre-trained with MR images. These findings challenge the popular view that the low-level information captured by the first layers of a CNN can be shared across domains.

Our quantitative analysis suggests a large dispersion of explanation performance for all XAI methods, which may be unexpected given the controlled setting in which these methods

have been applied here. Individual explanations can range from very good to very poor even for high overall classification accuracy, indicating a high risk of misinterpretation for a considerable fraction of inputs.

6.1 Limitations

Note, our analysis of XAI methods is limited to one DNN architecture, VGG-16, mainly showcasing the utility of our devised ground-truth dataset for model explanations. We stress, that, rather than conducting an exhaustive study of the behavior of popular XAI methods in relation to specific model architectures, with our work, we aim to predominantly contribute to the evaluation of XAI methods by providing a controlled ground-truth dataset, with known explanations, class-related features, enabling future research to benchmark new XAI methods.

We emphasize that we purposely refrain from expert annotated data as it does not constitute a stable ground-truth, and that full knowledge about the underlying ground-truth is needed to validate methods, a purpose that is only served by synthetically crafted data. In the medical domain, ML methods are often used with the expectation that they will uncover statistical relations that are either unknown or too complex (e.g., involving non-linear interactions of features) for human experts to discern. When experts annotate data, they may inadvertently overlook these features, potentially leading to false-positive detections if an XAI method indeed succeeds in highlighting them. Conversely, human experts may provide annotations that are simply incorrect. They can be influenced by pseudo-correlations in the data resulting from limited sample size in prior studies, or mistakenly base their judgement on confounders or even suppressors. In such instances, a correctly functioning XAI method may be mistakenly accused of delivering false-negative detections. Note in this context that clinical doctrines are highly fluctuating as new evidence is constantly being produced. For example, the assumed causal role of beta-amyloid and tau protein plaques in the brain for various types of dementia is currently being challenged. To address these challenges and strive toward real-world validity, experiments involving annotated real data are valuable and complementary next steps. However, they cannot entirely replace ground-truth experiments involving synthetic or manipulated real data due to their intrinsic biases. And generating realistic artificial and controllable image data for the MRI domain is, in itself, a very hard problem.

Furthermore, the proposed lesion generation process resembles the idea of white matter hyperintensities where we aim to approximate specific neurodegenerative disorders from a “model perspective”, where a natural prediction task would be “healthy” vs. “lesioned brain”. But it would be difficult to define a ground-truth for the class “healthy”. Hence, we chose to create a classification problem based on two different shapes of lesions: round vs. elongated. Admittedly, this distinction has no immediate physiological basis and serves purely the purpose of this benchmark, i.e., we can solve a classification task well enough by using a model architecture considered popular in this field. However, we provide a classification scenario where

the background, real brain slice images, provides features that are partially leveraged by ML models, which put XAI methods in the position to differentiate between class-related features, artificial lesions, and realistic brain-related features. Where we think that this distinction constitutes a realistic environment for XAI methods. In this light, our dataset can be seen as a first instance of contributing to the performance quantification of explanations produced by XAI methods for the MRI domain.

We argue that the quantitative validation of the *correctness* of XAI methods is still a greatly under-investigated topic given how popular some of the methods have become. Major efforts both on the theoretical and empirical side are needed to create a framework within which evidence for the correctness of such methods can be provided. As a first step toward such a goal, meaningful definitions of what actually constitutes a correct explanation need to be devised. While in our study, ground-truth explanations were defined through a data generation process, other definitions, depending on the intended use of the XAI, are conceivable. The existence of such definitions would then pave the way for a theoretical analysis of XAI methods as well as for use-case-dependent empirical validations.

7 Conclusion

In this work we created a versatile synthetic image dataset that allows us to quantitatively study the classification and explanation performances of CNN and similar complex ML methods in a highly controlled yet realistic setting, resembling a clinical diagnosis/anomaly detection task based on medical imaging data. Concretely, we overlaid structural brain MRI data with synthetic lesions representing clinically relevant white matter hyperintensities. We propose this dataset, to evaluate the explanations obtained from pre-trained models. Our study is set apart from the majority of work on XAI in that it uses a well-defined ground-truth for explanations, which allows us to quantitatively evaluate the “explanation” performance of several XAI methods.

Our study revealed a strong correlation between the classification performance of the model and the explanation performance of the XAI methods. Despite this correlation, models fine-tuned to a greater extent were shown to lead to better explanations. Controlling for classification performance, models pre-trained on MRI data lead to better explanations for every XAI method. The explanation performance of models pre-trained on within-domain images seem to have more stable explanation performance for a bigger range of classification accuracies. On the other hand, the explanation performance of models pre-trained with more general images quickly degrades with lower classification performance.

The quantitative analysis of the explanations also shows a concerning variability of explanation performance values, suggesting that, when these methods are used to explain an individual prediction, a large uncertainty is associated with the correctness of the resulting importance map. This is a critical issue when using XAI methods to “explain” predictions in high-stake fields such as medicine.

Data availability statement

The original contributions presented in the study are included in the article/[Supplementary material](#), further inquiries can be directed to the corresponding author.

Ethics statement

Ethical approval was not required for the study involving humans in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and the institutional requirements. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

MO: Writing – original draft, Data curation, Investigation, Visualization. RW: Writing – review & editing, Supervision, Validation. BC: Writing – review & editing, Supervision, Validation. CB: Writing – review & editing, Investigation. FE: Writing – review & editing, Supervision. KR: Writing – review & editing, Supervision. SH: Writing – review & editing, Conceptualization, Funding acquisition, Methodology, Supervision, Writing – original draft.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This result

References

- Agarwal, C., Saxena, E., Krishna, S., Pawelczyk, M., Johnson, N., Puri, I., et al. (2022). Openxai: towards a transparent evaluation of model explanations. *arXiv [preprint]*. doi: 10.48550/arXiv.2206.11104
- Ardalan, Z., and Subbian, V. (2022). Transfer learning approaches for neuroimaging analysis: a scoping review. *Front. Artif. Intell.* 5:780405. doi: 10.3389/frai.2022.780405
- Arras, L., Osman, A., and Samek, W. (2022). Clevr-xai: a benchmark dataset for the ground truth evaluation of neural network explanations. *Inf. Fus.* 81, 14–40. doi: 10.1016/j.inffus.2021.11.008
- Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE* 10, 1–46. doi: 10.1371/journal.pone.0130140
- Cheng, P. M., and Malhi, H. S. (2017). Transfer learning with convolutional neural networks for classification of abdominal ultrasound images. *J. Digit. Imaging* 30, 234–243. doi: 10.1007/s10278-016-9929-2
- Cherti, M., and Jitsev, J. (2021). *Effect of Pre-Training Scale on Intra- and Inter-Domain Full and Few-Shot Transfer Learning for Natural and Medical X-Ray Chest Images*. Sydney, NSW: Medical Imaging Meets NeurIPS (MedNeurIPS), 1–6
- Clark, B., Wilming, R., and Haufe, S. (2023). Xai-tris: non-linear benchmarks to quantify ml explanation performance. *arXiv [preprint]*. doi: 10.48550/arXiv.2306.12816
- Conger, A. J. (1974). A revised definition for suppressor variables: a guide to their identification and interpretation. *Educ. Psychol. Meas.* 34, 35–46. doi: 10.1177/001316447403400105
- Dar, S. U. H., Özbey, M., Çatlı, A. B., and Çukur, T. (2020). A transfer-learning approach for accelerated MRI using deep neural networks. *Magn. Reson. Med.* 84, 663–685. doi: 10.1002/mrm.28148
- d'Arbeloff, T., Elliott, M. L., Knodt, A. R., Melzer, T. R., Keenan, R., Ireland, D., et al. (2019). White matter hyperintensities are common in midlife and already associated with cognitive decline. *Brain Commun.* 1:fcz041. doi: 10.1093/braincomms/fcz041
- Das, A., and Rad, P. (2020). Opportunities and challenges in explainable artificial intelligence (XAI): A survey. *arXiv [Preprint]*.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). "Imagenet: a large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition (IEEE)*, 248–255.
- European Commission (2018). *2018 Reform of EU Data Protection Rules*.
- Fischl, B. (2012). FreeSurfer. *Neuroimage* 62, 774–781. doi: 10.1016/j.neuroimage.2012.01.021
- Friedman, L., and Wall, M. (2005). Graphical views of suppression and multicollinearity in multiple linear regression. *Am. Stat.* 59, 127–136. doi: 10.1198/000313005X41337

is part of a project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant Agreement No. 758985), the German Federal Ministry for Economy and Climate Action (BMWK) in the frame of the QI-Digital Initiative, and the Heidenhain Foundation.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/frai.2024.1330919/full#supplementary-material>

- Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., et al. (2013). The minimal preprocessing pipelines for the human connectome project. *Neuroimage* 80, 105–124. doi: 10.1016/j.neuroimage.2013.04.127
- Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., et al. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *Neuroimage* 87, 96–110. doi: 10.1016/j.neuroimage.2013.10.067
- Hofmann, S. M., Beyer, F., Lapuschkin, S., Goltermann, O., Loeffler, M., Müller, K.-R., et al. (2022). Towards the interpretability of deep learning models for multi-modal neuroimaging: finding structural changes of the ageing brain. *Neuroimage* 261:119504. doi: 10.1016/j.neuroimage.2022.119504
- Ismail, A. A., Gunady, M., Bravo, H. C., and Feizi, S. (2020). Benchmarking deep learning interpretability in time series predictions. *Adv. Neural. Inf. Process. Syst.* 33, 6441–6452.
- Ismail, A. A., Gunady, M., Pessoa, L., Bravo, H. C., and Feizi, S. (2019). *Input-Cell Attention Reduces Vanishing Saliency of Recurrent Neural Networks*, 10814–10824.
- Jenkinson, M., Bannister, P., Brady, M., and Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage* 17, 825–841. doi: 10.1006/nimg.2002.1132
- Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., and Smith, S. M. (2012). FSL. *Neuroimage* 62, 782–790. doi: 10.1016/j.neuroimage.2011.09.015
- Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al. (2018). “Interpretability beyond feature attribution: quantitative testing with concept activation vectors (TCAV),” in *International Conference on Machine Learning (PMLR)*, 2668–2677.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). “ImageNet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems 25*, eds F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Curran Associates, Inc.), 1097–1105.
- Lundberg, S. M., and Lee, S.-I. (2017). “A unified approach to interpreting model predictions,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, CA), 4768–4777.
- Milchenko, M., and Marcus, D. (2013). Obscuring surface anatomy in volumetric imaging data. *Neuroinformatics* 11, 65–75. doi: 10.1007/s12021-012-9160-3
- Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man Cybern.* 9, 62–66. doi: 10.1109/TSMC.1979.4310076
- Pan, S. J., and Yang, Q. (2010). A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* 22, 1345–1359. doi: 10.1109/TKDE.2009.191
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* 1, 206–215. doi: 10.1038/s42256-019-0048-x
- Shirokikh, B., Zakazov, I., Chernyavskiy, A., Fedulova, I., and Belyaev, M. (2020). “First U-Net layers contain more domain specific information than the last ones,” in *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning*, eds S. Albarqouni, S. Bakas, K. Kamnitsas, M. J. Cardoso, B. Landman, W. Li, et al. (Cham: Springer International Publishing), 117–126.
- Shrikumar, A., Greenside, P., and Kundaje, A. (2017). “Learning important features through propagating activation differences,” in *International Conference on Machine Learning (PMLR)*, 3145–3153.
- Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep inside convolutional networks: visualising image classification models and saliency maps. *CoRR* abs/1312.6034.
- Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale IMAGE recognition. *arXiv [preprint]*. doi: 10.48550/arXiv.1409.1556
- Springenberg, J., Dosovitskiy, A., Brox, T., and Riedmiller, M. (2015). “Striving for simplicity: the all convolutional net,” in *ICLR (Workshop Track)*.
- Sundararajan, M., Taly, A., and Yan, Q. (2017). “Axiomatic attribution for deep networks,” in *Proceedings of the 34th International Conference on Machine Learning, Vol. 70*, eds D. Precup, and Y. W. Teh (PMLR), 3319–3328.
- Tjoa, E., and Guan, C. (2020). Quantifying explainability of saliency methods in deep neural networks with a synthetic dataset. *IEEE Trans. Artif. Intell.* 4, 858–870.
- Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* 25, 44–56. doi: 10.1038/s41591-018-0300-7
- Valverde, J. M., Imani, V., Abdollahzadeh, A., De Feo, R., Prakash, M., Ciszek, R., et al. (2021). Transfer learning in magnetic resonance brain imaging: a systematic review. *J. Imaging* 7:66. doi: 10.3390/jimaging7040066
- Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., and Ugurbil, K. (2013). The WU-Minn human connectome project: an overview. *Neuroimage* 80, 62–79. doi: 10.1016/j.neuroimage.2013.05.041
- Wharton, S. B., Simpson, J. E., Brayne, C., and Ince, P. G. (2015). Age-associated white matter lesions: the MRC cognitive function and ageing study. *Brain Pathol.* 25, 35–43. doi: 10.1111/bpa.12219
- Wilming, R., Budding, C., Müller, K.-R., and Haufe, S. (2022). Scrutinizing XAI using linear ground-truth data with suppressor variables. *Mach. Learn.* 111, 1903–1923. doi: 10.1007/s10994-022-06167-y
- Wilming, R., Kieslich, L., Clark, B., and Haufe, S. (2023). “Theoretical behavior of xai methods in the presence of suppressor variables,” in *Proceedings of the 40th International Conference on Machine Learning (ICML), Vol. 202* (PMLR), 37091–37107.
- Yang, M., and Kim, B. (2019). Benchmarking attribution methods with relative feature importance. *arXiv [Preprint]*. arXiv:1907.09701.
- Zeiler, M. D., and Fergus, R. (2014). “Visualizing and understanding convolutional networks,” in *Computer Vision-ECCV 2014*, eds D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars (Cham: Springer International Publishing), 818–833.
- Zucco, C., Liang, H., Fatta, G. D., and Cannataro, M. (2018). “Explainable sentiment analysis with applications in medicine,” in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 1740–1747.