

3D ReX: Causal Explanations in 3D Neuroimaging Classification

Melane Navaratnarajah¹, Sophie A. Martin², David A. Kelly¹, Nathan Blake^{1, 2}, Hana Chockler¹

¹ King’s College London, UK

² University College London, UK

melane.navaratnarajah@kcl.ac.uk, s.martin.20@ucl.ac.uk, david.a.kelly@kcl.ac.uk, nathan.blake@kcl.ac.uk, hana.chockler@kcl.ac.uk

Abstract

Explainability remains a significant problem for AI models in medical imaging, making it challenging for clinicians to trust AI-driven predictions. We introduce 3D-REX, the first causality-based post-hoc explainability tool for 3D models. 3D-REX uses the theory of actual causality to generate responsibility maps which highlight the regions most crucial to the model’s decision. We test 3D-REX on a stroke detection model, providing insight into the spatial distribution of features relevant to stroke.

1 Introduction

Machine learning, particularly deep learning algorithms, has transformed medical diagnosis by automating time-consuming and often subjective tasks, leading to more efficient and accurate results (Esteva et al. 2019). These algorithms can extract features from complex data that may not be distinguishable to the human eye, reshaping healthcare practices. In particular, 3D data provides in-depth spatial information of internal structures which is often critical to understanding complex regions of the body. Spatial insights potentially aid with earlier diagnosis, facilitate intervention planning, modeling biological processes, and enable clinicians to visualize precise anatomy with greater reliability. However, despite the increasing use of 3D models across healthcare, a significant drawback is the lack of interpretability (Sheu and Pardeshi 2022; Saeed and Omlin 2023). If clinicians cannot discern the reasoning behind a model’s final predictions, this poses a significant challenge for integrating these tools into medical practice. There is uncertainty and potential risks, as decisions based solely on opaque model outputs might have severe consequences, especially in a clinical setting.

Explainable AI (XAI) tools, including perturbation-based methods like LIME (Ribeiro, Singh, and Guestrin 2016), SHAP (Lundberg and Lee 2017), and RISE (Petsiuk, Das, and Saenko 2018), as well as backpropagation-based techniques such as Grad-CAM (Selvaraju et al. 2017), are designed to bridge the interpretability gap in AI-powered tools. These methods aim to facilitate understanding, increase trust in AI systems, and support error identification, paving the way towards safer and more transparent applications in high-risk domains such as healthcare. To this end, much attention has been given to XAI in diagnostic neuroimaging.

However, existing methods still require substantial refinement in their robustness and adaptability to diverse input types, such as 3D data, to ensure they are suitable for explaining local and global disease markers. In 3D imaging, where identifying localized regions (*e.g.* lesions) and broader patterns is critical, a notable gap exists in tools capable of delivering holistic explanations. This disparity remains largely under-explored.

Attempts to provide post-hoc explainability for 3D medical imaging models have typically focused on backpropagation techniques. For instance, Wood et al. (2022) used smooth guided backpropagation to provide slice-wise and voxel-wise saliency maps. However, these methods rely upon direct access to the internal model, which is not always available. Hence, we focus here on model-agnostic XAI tools, which require no such access.

Much of the literature pertaining to XAI for neuroimaging has focused on cancer and dementia, as these are more clinically immediate problems (Van der Velden et al. 2022). However, a common limitation of these studies is the difficulty of validating the explanations at an individual level, due to the lack of a ground truth. Stroke detection, however, provides something against which to evaluate explanations: the presence of localized lesions in the brain. Of the few studies that apply explainability to stroke detection (Gurmessa and Jimma 2023), the most common XAI methods are SHAP and LIME. Neither tool has an innate understanding of 3D data, nor of the types of occlusions required to extract maximum information from the model.

To overcome this, we introduce 3D-REX, an extension of the existing framework, REX (Chockler et al. 2024). 3D-REX is modified to accommodate 3D inputs (without access to a model’s weights) and to facilitate meaningful occlusion values. We present the algorithm for computing 3D causal explanations in Section 3 and conduct a small-scale evaluation of the method on stroke data (Section 4). We conclude with a discussion of present limitations and future work (Section 5).

2 Related Work

REX, and 3D-REX, are rooted in the framework of actual causality introduced by (Halpern and Pearl 2005). A thorough overview of the topic can be found in Halpern (2016). The REX tool is described in detail in Chockler

et al. (2024), and a full presentation of the theoretical apparatus is in Chockler and Halpern (2024). Actual causality extends simple counterfactual reasoning by considering the effect of *interventions* to the current setting. For images, this means changing the values of pixels to some masking value, or values, and observing the output (*i.e* the model classification). REX uses *causal responsibility* (Chockler and Halpern 2004) to quantify the contribution of pixels towards the final classification. We can view this quantification directly, as in Figure 1, or extract an approximately minimal set of pixels (or voxels) which are, by themselves, sufficient to get the required classification.

What constitutes a “meaningful” occlusion is open to debate and is often dependent on the dataset, task and model in question. Fong and Vedaldi (2017) suggest that using naturalistic or plausible effects leads to more meaningful occlusions and experimented with constant values, noise injection and blurring. However, Uzunova et al. (2019) and Lenis et al. (2020) find that these occlusions are not suitable for medical imaging and instead introduced occlusion values using a variational autoencoder and inpainting respectively. Blake et al. (2024) observed that REX, with an occlusion value of 0 on false RGB MRI slices of brain tumors, produces explanations which coincide well with a human provided segmentation.

3 Methodology

In this section, we present the details of REX’s inner workings, highlighting the modifications needed to accommodate 3D input data for medical imaging tasks. The algorithm functions by iteratively building a responsibility map that keeps track of each voxel’s causal responsibility. This is achieved by masking specific input regions and assessing the model output for each mask. Occlusions are sets of supervoxels (themselves groupings of contiguous voxels): Responsibility for the classification is distributed over the supervoxels which contributed to the correct classification. No responsibility is given to those supervoxels which make no contribution. Supervoxels with non-0 responsibility are further refined. This process is repeated multiple times with different random supervoxels. The final result is a causal responsibility map. A sufficient, approximately minimal, subset of the voxels can be extracted from the image using information contained in the map, which constitutes a (causal) explanation. We highlight this explanation in orange in the examples shown in Figure 1.

The process is outlined for one iteration in Algorithm 1 and begins by obtaining the model’s original prediction on an unmodified MRI scan, which serves as a *target*. All subsequent predictions on modified inputs are compared to this target to determine whether a modification “passes” or “fails” — that is, whether it maintains the original classification. The responsibility map is initialized as a zero matrix matching the dimensions of the input data. This map accumulates responsibility for each voxel in the input space based on the passing mutants, eventually forming peaks in areas responsible to the classification.

Each iteration of 3D-REX involves initializing a queue that tracks partitions of the input space. Each supervoxel is

Algorithm 1: General Algorithm for 3D-REX-*generate_resp_map()*

```

1: Input: Model  $M$ , Dataset  $D$ , Max Depth  $d_{max}$ , Search
   Limit  $l_{max}$ 
2: Output: Responsibility Map  $rm$ 
3:  $m \leftarrow M$  // Load the model with a specific checkpoint
4:  $d \leftarrow D$  // Select a valid input datapoint
5:  $rm \leftarrow \text{zeros}(\text{shape}(d))$  // Initialize Responsibility
   Map
6:  $target \leftarrow m(d)$  // Get classification for original data-
   point
7:  $queue \leftarrow [d]$  // Initialize queue with the full input
   space
8:  $passing\_mutants \leftarrow []$ 
9:  $total\_work \leftarrow 0$ 
10: while not ( $depth \geq d_{max}$  or  $steps \geq l_{max}$ ) do
11:    $head \leftarrow queue.pop()$ 
12:    $masks \leftarrow \text{generate\_masks}(head)$  // Split input
     region into subregions
13:   for all  $mask \in masks$  do
14:      $mutant \leftarrow \text{apply\_mask}(d, mask)$ 
15:      $prediction \leftarrow m(mutant)$ 
16:     if  $prediction = target$  then
17:        $queue.append(mask)$  // Store passing mu-
         nants for further splitting
18:        $update\_resp\_map(rm, mask)$ 
19:     end if
20:   end for
21:    $depth = \text{update\_depth}(depth, passing)$ 
22:    $total\_work += \text{len}(passing)$ 
23: end while
24: return  $rm$ 

```

attributed responsibility after querying the model, based on whether it was a passing partition. Supervoxels with no responsibility are discarded. This process continues until the search budget is exhausted or the supervoxels become so refined that they no longer contribute to the classification or fall below a size threshold.

Initially, the input is split into quadrants. We randomly select two axes (from x , y , and z) to partition the input space, resulting in four segments that can still adequately capture (or break) spatial dependencies in the 3D input. While dividing the input into eight partitions would provide more exhaustive coverage, it would lead to up to 256 possible combinations to assess. In a worst-case scenario where all initial mutants pass, *i.e* a truly global diagnosis, each mutant would pass, and therefore, further generate 256 additional variants, resulting in exponential growth of evaluation candidates. By reducing the partitioning to four segments as shown in Algorithm 2, 3D-REX constructs a responsibility map, highlighting causally significant regions in the 3D MRI data while reducing overwhelming computational costs.

4 Experiments

While the purpose of this paper is to demonstrate 3D-REX, we summarize here the data, model architecture and train-

Algorithm 2: The mask generation algorithm: generate_masks(h)

```

1: Input: The head of the queue,  $h$ , which is a partition of
   the input.
2: Output: It will output the input randomly partitioned
   into 4 segments as a list.
3:  $axes \leftarrow [ROW, COL, DEPTH]$  // Define enums for
   the available axes
4:  $ranges[ROW] \leftarrow [row\_start, row\_end]$ 
5:  $ranges[COL] \leftarrow [col\_start, col\_end]$ 
6:  $ranges[DEPTH] \leftarrow [depth\_start, depth\_end]$ 
7:  $axis1, axis2 \leftarrow rand(axes, 2)$  // Randomly picks
   two axes to split
8:  $coord1 \leftarrow rand(ranges[axis1])$ 
9:  $coord2 \leftarrow rand(ranges[axis2])$ 
10:  $first\_split\_boxes = split(input\_box, axis1, coord1)$ 
11: for all  $box \in first\_split\_boxes$  do
12:    $new\_box \leftarrow split(box, axis2, coord2)$ 
13:    $second\_split\_boxes.append(new\_boxes)$ 
14: end for
15: return  $second\_split\_boxes$ 

```

ing. We trained a model to classify stroke patients from healthy controls using data from the Anatomical Tracings of Lesions After Stroke (ATLAS) R2.0 (Liew et al. 2022) and IXI datasets (Imperial College London 2015). 3D isotropic T1-weighted magnetic resonance images (MRI) from a total of 1236 (655 stroke patients, 581 controls) participants were used. Since data from the ATLAS dataset was already registered to a common brain template, we registered the IXI scans to the same template provided to reduce site-specific spatial biases. All volumes were skull-stripped, scaled to an intensity range of 0 to 1 and resized to $96 \times 96 \times 96$ mm. 90% was used for training and validation, and 10% as a held-out test set. We used MONAI’s (Monai 2020) ResNet18 implementation to train a classifier using a binary cross-entropy loss with a weighted Adam optimizer. The trained model achieved an accuracy of 97% and AUROC value of 99% on the test data ($n = 124$, 57 controls and 67 stroke patients).

Settings for 3D-REX Explanations To explain the trained model, we utilized the outlined method, 3D-REX. To find the optimal occlusion value for examining the model’s decision-making process in this specific context, we compared different options, such as the mean intensity across stroke patients, patches from a healthy MRI scan and a 0 value. This value is applied to the data after it is processed for the model, so does not in general correspond to black. Rather, it is an out-of-distribution value unlikely to be associated with a specific physical part of the brain scan.

While both healthy MRI occlusion and 0-value occlusion successfully produced explanations, mean value occlusion failed to yield any interpretable results, highlighting its ineffectiveness in this context. The explanations derived from the healthy MRI and 0-value occlusion highlighted distinct regions of the brain. The 0-value occlusion produced explanations closely aligned/overlapping with lesion location, indicating that the model’s focus was on areas corresponding

to regions of interest typically associated with stroke pathology. In contrast, the healthy MRI occlusion led to explanations that pointed to a different area of the brain, not aligned with lesion location but rather with other regions that the model had focused on during its decision-making process.

The divergence in explanation for different occlusion values highlights an important insight: while the 0-value occlusion produced results that were closer to the lesion location, using the healthy MRI scan revealed that the model’s underlying reasoning might rely on a complex set of brain regions for the classification. Although interesting, further experimentation on a larger model and dataset is required to confirm and validate these results. Nonetheless, these findings draw importance to the selection of the occlusion values and how using different strategies provides valuable insight into the model’s interpretation of the stroke-related features in MRI data. From a clinician’s perspective, being able to learn a diverse set of features that do not solely focus on the obvious pathological regions provides an opportunity to discover nuanced patterns and derive new biological insights.

Examples of Results Figures 1 and 2 illustrate the responsibility map generated by 3D-REX superimposed on the MRI scans for patient A. They also outline the explanation extracted from the responsibility map and the lesion location annotation for comparison.

The two visualizations provide complementary perspectives on the model’s interpretability. Figure 1 shows a 2D overlay of the responsibility map and explanation, highlighting the areas of high responsibility in red and areas of low responsibility in blue. The visualization is presented across the axial, sagittal and coronal planes, enabling a comprehensive view of the model’s focus within the brain.

In contrast, Figure 2 showcases a 3D rendering of the explanation, offering a volumetric view of the model’s focus within the entire structure. This captures the spatial distribution of the explanation compared to the lesion location, allowing for the assessment of deeper or overlapping regions that may not be apparent in the 2D slices. In this case, the explanation, shown in Figure 2, is positioned perpendicularly in comparison to the lesion in 3D space and focuses on the right hemisphere cortical lesion. Overall, these visualizations should indicate to clinicians the necessary insight into the model, with different dimensional explanations that can aid in understanding the complex brain structures for stroke diagnosis.

5 Discussion

Limitations One significant limitation of the proposed approach is the computational complexity arising from the large search space inherent in constructing responsibility maps for 3D data. Although we mitigate this by only splitting on two axes rather than all three, which reduces the computational burden, further optimization is necessary in time-constrained tasks. Techniques such as parallel processing or batching mutants, where supported by the model, and better sampling strategies could enhance the scalability and efficiency.

We do not claim a comprehensive evaluation of 3D-REX.

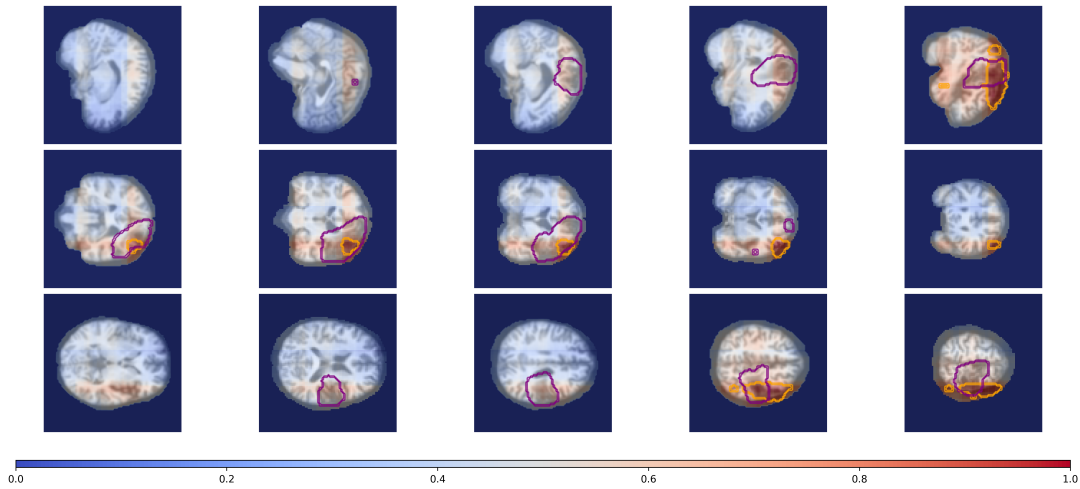


Figure 1: Visualization of the model’s explanation for patient A, predicted as stroke with a confidence of 0.9756. The **purple** outline represents the annotated lesion location from the dataset, while the **orange** outline represents the explanation generated by 3D-REX. The patient’s brain is also overlaid with the responsibility map generated by 3D-REX, and the color bar depicts the normalized intensity of feature contributions.

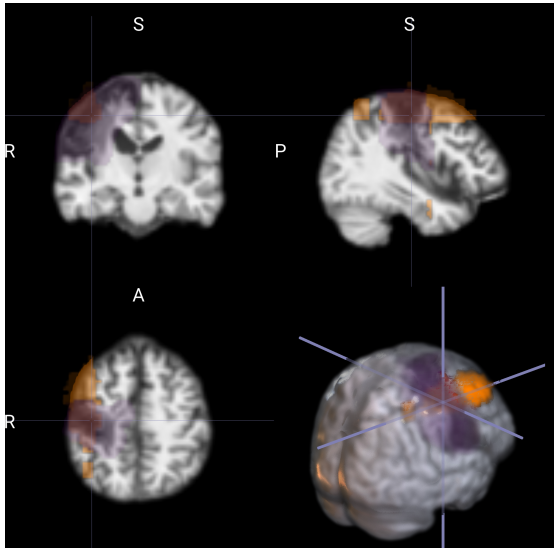


Figure 2: 3D rendering of patient A’s results showing the patient’s brain with overlaid annotated lesion location from the dataset (**purple**) and 3D-ReX’s explanation (**orange**). This visualization provides a spatial perspective of the explanation.

Numerical evaluation of a large-scale dataset remains to be done. While individual examples and visual comparisons provide insight into our technique, the absence of a systematic, standardized assessment framework over a broader dataset limits generalizability. This issue is further emphasized in a broader call for standardized criteria in clinical explainability evaluation, outlined in Jin et al. (2023). This paper highlights the need for explainability tools to be validated across diverse and clinically relevant datasets and user

studies to ensure their utility and reliability in real-world applications. As a next step, we plan to conduct a large-scale evaluation to address this limitation and enhance the robustness of our approach.

Future Work We intend to extend 3D-REX’s capabilities across different medical imaging modalities. We will validate this process with different applications that involve more heterogeneous data and complex pathological features. Another avenue for future work is to further investigate the plausibility of finding a range of different explanations using different occlusion values, such as leveraging a healthy brain MRI to create “real brain” occlusions, in a diverse set of contexts and larger models, potentially uncovering novel insight into the model’s behavior.

In cases where the model identifies unexpected features or patterns in the data, we will investigate the use of explanations to disentangle whether the model is learning incorrect features or the model has learned a feature that has biological relevance. This aligns with the broader goal of enabling XAI tools to find explanations that are clinically meaningful but also capable of revealing new scientific knowledge.

Future work concerning REX’s integration with other modes of data input type should include multi-modal models, combining both 3D structural imaging such as CT, MRI *etc.* with genomic, patient history and other details. Multi-modal integration will provide a comprehensive understanding and accurate classification, and being able to produce coherent explanations will help mitigate potential bias and validate model decisions in multimodal contexts.

References

Blake, N.; Kelly, D.; Peña, S.; Chanchal, A.; and Chockler, H. 2024. Explainable AI for the classification of brain MRIs. *Nature Portfolio*.

- Chockler, H.; and Halpern, J. Y. 2004. Responsibility and Blame: A Structural-Model Approach. *J. Artif. Intell. Res.*, 22: 93–115.
- Chockler, H.; and Halpern, J. Y. 2024. Explaining Image Classifiers. *arXiv preprint arXiv:2401.13752*.
- Chockler, H.; Kelly, D. A.; Kroening, D.; and Sun, Y. 2024. Causal Explanations for Image Classifiers. *arXiv preprint arXiv:2411.08875*.
- Esteva, A.; Robicquet, A.; Ramsundar, B.; Kuleshov, V.; DePristo, M.; Chou, K.; Cui, C.; Corrado, G.; Thrun, S.; and Dean, J. 2019. A guide to deep learning in healthcare. *Nature medicine*, 25(1): 24–29.
- Fong, R. C.; and Vedaldi, A. 2017. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE international conference on computer vision*, 3429–3437.
- Gurmessa, D. K.; and Jimma, W. 2023. A comprehensive evaluation of explainable Artificial Intelligence techniques in stroke diagnosis: A systematic review. *Cogent Engineering*, 10(2): 2273088.
- Halpern, J. Y. 2016. *Actual Causality*. The MIT Press. ISBN 9780262336611.
- Halpern, J. Y.; and Pearl, J. 2005. Causes and Explanations: A Structural-model Approach. Part II: Explanations. *British Journal for the Philosophy of Science*, 56(4).
- Imperial College London. 2015. IXI Dataset – Brain Development.
- Jin, W.; Li, X.; Fatehi, M.; and Hamarneh, G. 2023. Guidelines and evaluation of clinical explainable AI in medical image analysis. *Medical Image Analysis*, 84: 102684.
- Lenis, D.; Major, D.; Wimmer, M.; Berg, A.; Sluiter, G.; and Bühler, K. 2020. Domain aware medical image classifier interpretation by counterfactual impact analysis. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*, 315–325. Springer.
- Liew, S.-L.; Lo, B. P.; Donnelly, M. R.; Zavaliangos-Petropulu, A.; Jeong, J. N.; Barisano, G.; Hutton, A.; Simon, J. P.; Juliano, J. M.; Suri, A.; Wang, Z.; Abdullah, A.; Kim, J.; Ard, T.; Banaj, N.; Borich, M. R.; Boyd, L. A.; Brodtmann, A.; Buetefisch, C. M.; Cao, L.; Cassidy, J. M.; Ciullo, V.; Conforto, A. B.; Cramer, S. C.; Dacosta-Aguayo, R.; de la Rosa, E.; Domin, M.; Dula, A. N.; Feng, W.; Franco, A. R.; Geranmayeh, F.; Gramfort, A.; Gregory, C. M.; Hanlon, C. A.; Hordacre, B. G.; Kautz, S. A.; Khlif, M. S.; Kim, H.; Kirschke, J. S.; Liu, J.; Lotze, M.; MacIntosh, B. J.; Mataró, M.; Mohamed, F. B.; Nordvik, J. E.; Park, G.; Pienta, A.; Piras, F.; Redman, S. M.; Revill, K. P.; Reyes, M.; Robertson, A. D.; Seo, N. J.; Soekadar, S. R.; Spalletta, G.; Sweet, A.; Telenczuk, M.; Thielman, G.; Westlye, L. T.; Winstein, C. J.; Wittenberg, G. F.; Wong, K. A.; and Yu, C. 2022. A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms. *Scientific Data*, 9(1): 320. Publisher: Nature Publishing Group.
- Lundberg, S. M.; and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. In Guyon, I.; Luxburg,
- U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *NIPS*, 4765–4774. Curran Associates, Inc.
- Monai, P. 2020. The MONAI Consortium. Type: Computer Program.
- Petsiuk, V.; Das, A.; and Saenko, K. 2018. RISE: Randomized Input Sampling for Explanation of Black-box Models. In *British Machine Vision Conference (BMVC)*. BMVA Press.
- Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "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, KDD '16*, 1135–1144. New York, NY, USA: Association for Computing Machinery. ISBN 9781450342322.
- Saeed, W.; and Omlin, C. 2023. Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems*, 263: 110273.
- Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *International Conference on Computer Vision (ICCV)*, 618–626. IEEE.
- Sheu, R.-K.; and Pardeshi, M. S. 2022. A Survey on Medical Explainable AI (XAI): Recent Progress, Explainability Approach, Human Interaction and Scoring System. *Sensors*, 22(20).
- Uzunova, H.; Ehrhardt, J.; Kepp, T.; and Handels, H. 2019. Interpretable explanations of black box classifiers applied on medical images by meaningful perturbations using variational autoencoders. In *Medical Imaging 2019: Image Processing*, volume 10949, 264–271. SPIE.
- Van der Velden, B. H.; Kuijf, H. J.; Gilhuijs, K. G.; and Viergever, M. A. 2022. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Medical Image Analysis*, 79: 102470.
- Wood, D. A.; Kafiabadi, S.; Al Busaidi, A.; Guilhem, E.; Montvila, A.; Lynch, J.; Townend, M.; Agarwal, S.; Mazumder, A.; Barker, G. J.; et al. 2022. Deep learning models for triaging hospital head MRI examinations. *Medical Image Analysis*, 78: 102391.