Unboxing the black-box of deep learning based reconstruction of undersampled MRIs

Soumick Chatterjee[1,2,3,*,†], Arnab Das[1,4,†], Rupali Khatun[5,6,7] and
Andreas Nürnberger[2,3,8]

[1]*Genomics Research Centre, Human Technopole, Milan, Italy*

[2]*Faculty of Computer Science, Otto von Guericke University Magdeburg, Magdeburg, Germany*

[3]*Data and Knowledge Engineering Group, Otto von Guericke University Magdeburg, Magdeburg, Germany*

[4]*German Research Centre for Artificial Intelligence, Kaiserslautern, Germany*

[5]*Translational Radiobiology, Department of Radiation Oncology, Universitätsklinikum Erlangen, Erlangen, Germany*

[6]*Department of Radiation Oncology, Universitätsklinikum Erlangen, Friedrich-Alexnder-Universität Erlangen-Nürnberg Erlangen, Germany*

[7]*Comprehensive Cancer Centre Erlangen-EMN, Erlangen, Germany*

[8]*Centre for Behavioural Brain Sciences, Magdeburg, Germany*

Abstract
Deep learning has emerged as a very important area of research and has shown immense potential in solving different kinds of problem, including in the medical field. For tasks like undersampled MRI reconstruction - the process of speeding up MRI acquisition with the help of undersampling, deep learning has shown its dominance over the years. But one of the major problems with deep learning is trust: Complex reasonings done by these models appear black-box to the users. Therefore, to build trust and better acceptability, it is important to open up this black-box nature of these models. For classification models, several approaches have been proposed. Nevertheless, for models dealing with inverse problems, like the reconstruction of the undersampled MRIs, it is more challenging as the output of the model has the same number of output pixels as the input, making the interpretability of such models more complex. This research explores different methods to understand the working mechanism of a deep learning model for the task of undersampled MRI reconstruction.

Keywords
Deep Learning, Blackbox, Inverse Problem, Interpretability, Explainability, MRI, MR Image Reconstruction

1. Introduction

Deep learning models have been proven very successful for a wide variety of tasks. And nowadays, it is applied in various fields ranging from the study of energy, consumer, and social-

cultural linguistics to critical domains such as autonomous driving, medical image analysis, and many more. The decisions made by these models directly or indirectly affect human life. The main reason for the success of these deep learning models is the availability of digitised data and their power to find complex patterns from it - to learn to perform the trained task. For computer vision-related tasks, which are often very complex, deep models with hundreds of thousands of parameters are employed. Models can be understood as a parameterised complex function estimator that maps input domain data to decision domains of classification [1, 2], segmentation [3], regression, Image-reconstruction [4], de-noising [5], and many more. But from an external perspective, a deep learning model, making decisions after learning from the given training data, may appear as a "black box" with no direct accountability for the decisions it makes. This is often true, since these models often do not provide any reason for their predictions. Hence, for critical domains such as biomedical applications, where the slightest of mistakes may have grave effects and even can be fatal, the use of these methods becomes a widely debated topic, as it has been seen in the past that a model giving the best accuracy during the test might not be making the correct reasoning to arrive at their decision [6]. This not only increases the chances of failure while in use in production but also makes it difficult to trust such methods. So, for better acceptability and applicability, opening the black-box nature of these models is the need of the hour. This will build trust in the decisions made by deep learning models, as predictions will be better grounded and explained.

Recent years have seen an increase in different interpretability and explainability techniques to try to understand the working mechanism of these complex models. Captum [7] is one of the many available packages that enables the application of post hoc methods on deep learning models already trained to help better understand those models. The primary focus of Captum, as well as most of the existing methods, is on classification models. TorchEsegeta [8] is a unified pipeline, including Captum and several other methods, which enables developers and decision makers (e.g., doctors) to apply several post-hoc interpretability and explainability methods on already trained classification models. This pipeline also extends these methods to explain deep segmentation models. However, there has not been any significant research on reconstruction models.

Image reconstruction is another task where deep learning models have demonstrated their superiority. An example of image reconstruction in the field of medical imaging is the task of undersampled image reconstruction. Magnetic resonance imaging (MRI) is an inherently slow process - making it difficult to be used in real-time applications. Undersampling, a process of ignoring parts of the data [9], can make image acquisition faster, but can compromise image quality (e.g. loss of resolution, presence of artefacts). Deep learning models, such as UNet [4] and ReconResNet (including the NCC1701 pipeline) [9], have shown superior performance over non-deep learning-based techniques for undersampled MRI reconstruction, such as compressed sensing [10]. There are also techniques that aim to combine deep learning models with compressed sensing [11, 12]. All these models primarily aim to reduce artefacts from the given input undersampled MRIs - learning by comparing their outputs against the corresponding fully-sampled MRIs. Similarly, reconstruction models like ShuffleUNet [13] and DDoS-UNet [14] attempt to superresolve (i.e., improving the image resolution) the input of low-resolution undersampled MRIs to the resolution of the corresponding high-resolution fully sampled MRIs. Although numerous deep learning methods have been proposed, in addition to

uncertainty quantification [15, 16], not much exploration has been done from the perspective of interpretability and explainability. Hence, the objective of this research is to find ways to understand the inner working mechanism of the reconstruction models for undersampled MRI reconstruction, with the help of different analyses and visualisations, to try to interpret and explain such models.

2. Methodology

In literature, interpretability and explainability methods are grouped under different rubrics, for example, local vs global, model dependent vs model agnostic, intrinsic vs post hoc, etc. [17]. This work proposes several methods for model understanding, uncertainty estimation, and interpretability in a post hoc fashion. The methods used are discussed in three subsections accordingly. As input, reconstruction models are provided with undersampled brain images, and the models are able to predict fully sampled images, and all models were trained in a supervised manner.

2.1. Uncertainty estimation

2.1.1. Model weight perturbation

Epistemic uncertainty gives rise to parameter uncertainty in trained models, which means that parameters can take several values for a region in input data space where there are no or very few data points presented during training [18]. To leverage this fact, the model looks perturbed by adding small random Gaussian noise in each run [19]. Then, this was applied several times for the same input image. Again, the pixel-wise output variance was then calculated from all the runs and produced an uncertainty heatmap from the same.

2.1.2. Monte-Carlo Simulation

This is a popular method for estimating the parameter uncertainty of predictions of a model. This is marked as epistemic uncertainty in the literature. Often the dropout layer is used in deep learning models as a Bayesian approximation of a model ensemble and as a regulariser to tackle overfitting problems [20]. But, as standard practise, dropout is only enabled in training time and disabled in test/inference time to output a deterministic and reproducible prediction [21]. The dropout was enabled at the test time and the model several times on the same input image. Then the pixel-wise variance was calculated from all the runs and produced an uncertainty heatmap of the same. Several additional dropouts were also introduced in the model at the test time and the experiments were repeated.

2.1.3. Subject Level Uncertainty score

For patch-based segmentation networks, [22] proposed a way to estimate the uncertainty at the subject level. This method estimates a multivariate Gaussian distribution over average pooled latent space activations from training patches and then calculates the Mahalanobis distance for

test patches. Then, from these distances, it calculates an uncertainty mask for the entire volume and finally provides a subject-level uncertainty score by averaging the mask over all voxels.

2.2. Model Understanding

2.2.1. Latent Space Exploration

Medical domain input data, such as MRI volume, is extremely high-dimensional, so directly estimating the prior data distribution is not a suitable task. Deep learning model architectures that incorporate an encoder-decoder structure learn a representation of the input data in its latent space, which is relatively lower dimensional. Exploring the latent space is often a go-to method for diving into the model's understanding of the data. In this work, various latent space exploration experiments [23, 24] on reconstruction models have been performed.

- The simplest technique is to directly visualise the latent space activation/feature maps.
- In the second method, 1000 images have been through the model and captured their latent space representation. Later on, t-Distributed Stochastic Neighbour Embedding (tSNE) was performed to project the high-dimensional data into a 2D map to visualise the same. The same has been repeated for training, validation, and test set images to understand or identify any distribution shift.
- Upon doing the tSNE and visualising, one could see grouping structures appearing. Therefore, the clustering was performed on the 2D representation. And identified representative input images for each cluster by backtracking the 2d representation to input data. The goal is to visualise how different latent representations correspond to the different input image.
- Latent spacewalk: Another popular approach, often performed by the deep learning community to understand whether the manifold learnt by the model in its latent space is continuous and fills the entire space or not. This can also help adapt the architecture of the model. Two different input images, in this case - MRIs, were chosen randomly from the test set and their latent representation of the model was obtained. Then, the latent vectors were linearly interpolated with uniform steps to generate the intermediate latent representations. Finally, these representations were decoded and visualised.

2.2.2. Noise Tolerance Estimation

This experiment shows the robustness of the model under test against possible noise, specifically for the decoder part of the model [25].

- In the initial approach, Gaussian noise of different magnitudes has been added to the latent representation, and then the noisy latent has been decoded. Finally, the structural similarity index (SSIM)[26] of the reconstructed input and the ground truth is calculated. After the experiment, a 2D graph of the SSIM value and the noise magnitude was plotted to see the notion of how noise tolerant the decoder is.
- In the second approach, some of the latent feature maps were randomly zeroed in different proportions and then reconstructed the resulting representation. The reconstructed image

and ground truth were compared by calculating the SSIM value. It is to be noted that the latent feature maps are suppressed randomly, and when unimportant feature maps get suppressed, no effective change occurs in the reconstructed image that is obtained by decoding the latent. However, this may also give a false notion of noise robustness. To ignore this problem, the same method has been applied several times and the median image of the reconstructed images has been selected.

2.2.3. Probabilistic Model Understanding Approximation

Image Reconstruction models with bottleneck layers can be considered similarly to auto-encoders (AE). The bottleneck layer represents the latent space representation for a given input learnt by the model. So in connection to the AE, layers up to the bottleneck layer in the model can be considered an encoder, and the rest of the layers as a decoder. Reconstruction models are capable of removing any unintended artefacts that occurred due to violation of the Nyquist-Shannon sampling theorem [27, 28], undersampling the MRI slices. So, these models can be considered analogous to energy-based generative models, and expect that after training, the model would have its understanding of data distribution. For

- Latent space representation: h
- Input data: x
- Reconstruction data: x'
- Encoder, decoder params: θ
- True data distribution : $P_{data}(x)$ or $P(x)$
- Models data distribution : $P_model(x)$ or $P_{\theta}(x)$

$P_{model}(x)$ can be estimated using a repeated Gibbs update by sampling from $P_{model}(h|x)$ and $P_{model}(x|h)$. As this is a directed model, thus the single update only means one pass-through encoder and decoder. But the problem is that, similar to AE, there is no mechanism to get the initial h to burn the chain, so two solutions have been proposed in this work, inspired by naive Markov Chain Monte Carlo (MCMC) and Contrastive Divergence (CD) [29]. In the naive MCMC, a sample of a Normal distribution with 0 mean and standard deviation of 1 has been fetched. And for the solution depicting CD, a sample from training data has been taken and passed through the encoder once to get the initial h.

2.2.4. Input Anomaly test

The goal of the Input Anomaly test is to verify the robustness of the model against any additional anomaly in the input image in real time that it has not seen at training time. This test inspired by the fact that deep learning models might be prone to adversarial noise [30, 31] or might react differently when encountered with anomalies if trained only with non-anomalous data - the idea behind unsupervised anomaly detection [32]. This work mainly deals with brain image reconstruction, so the expectation of the model is that the model should be able to perform proper reconstruction of the image if there is a tumour or a tumour-like structure present in the brain tissue. Brain images without tumours were selected, and a tumour-like circular structure was added to these images, and later on, these images were used as ground truth. The images

were then undersampled and passed through the model to visualise the final reconstructed image. This experiment has been performed for various pixel values for the circular lesion.

2.2.5. Targeted Activation Maximisation

Activation maximisation is a technique for classification models to show which types of input images activate a particular output neuron the most [33, 34]. This idea has been extended in this reconstruction model. For the classification task, there is a final output neuron (in the case of binary classification), and this is the one on which activation maximisation is applied. For the same, constantly the activation of the output neuron is taken, and the same is considered for the loss, and a gradient ascent is performed on the input to maximise the activation of this neuron as much as possible. So, the selection of this output neuron gives a hint to the discriminating network in which "concept" is to be maximised. The problem with the reconstruction network is that the selection of a particular output neuron is not possible. Because each pixel in the reconstruction output is a regressed value, selecting just one output pixel for activation maximisation does not provide a meaningful concept. So to give the network a hint about the concept that is to be maximised, a group of pixels from the reconstruction output can be chosen rather than a single one. But this raises the question of which pixels to choose? For this, one can take the support of the truth of the ground. A fully sampled ground-truth image was then used to generate two binary masks. The first masks indicate which pixels have a value greater than a threshold, and the second mask indicates which pixels have values lower than the threshold in the ground truth. These masks can be used as hints for maximisation of the activation of the network. Initially, the computation can start from random Gaussian noise, then perform gradient ascent for the pixels related to the first mask, and finally, gradient descent is performed for the pixels related to the second mask. This can be achieved by multiplying the binary mask by the reconstructed output at each step before summing. This work also showed the different maximised concepts based on the threshold value.

2.2.6. Activation/Reconstruction Comparison

The primary motivation for this experiment is to capture the model's response when presented with out-of-distribution data, which they have not seen before during the training period [35]. This will help us to understand the model's understanding of the data by analysing what sort of thing it can successfully reconstruct and what it cannot. This, in turn, helps to understand whether the models have learnt any prior knowledge about the structure that it is reconstructing or not. As input, The models have been presented with one in-distribution brain data, one noise input, and one completely out-of-distribution flower image, and then the histogram of latent space values and the reconstructed images have been compared.

2.3. Interpretability

The TorchEsegeta project [8] provides more than 40 interpretability methods from the literature and third-party libraries, catering to classification and segmentation models. As a part of the current research, some of these methods were extended for reconstruction models. This has been achieved with the help of a wrapping mechanism which converts the output of reconstruction

models to be similar to that of classification models. The wrapping mechanism was inspired by the wrapper proposed in TorchEsegeta [8]. This method performs class identification by Otsu thresholding of the reconstruction output and then sums up the pixels for each class. This task is also performed in two steps:

a. Normalisation - In this step, the output reconstructed image is normalised by the following function:

$$y_{ij_{norm}} = \frac{y_{ij} - min(y)}{max(y) - min(y)} \tag{1}$$

b. Pixel-wise binarisation - Pixel-wise binarisation is performed with the help of Otsu thresholding.

$$y_{ij} = \begin{cases} 1 & where \ \ y_{ij_{norm}} > th \\ \\ 0 & elsewhere \end{cases} \tag{2}$$

where th = otsu($y_{ij_{norm}}$)

The output of both processes is a tensor with a single value for each class. Although, the output range would not strictly be in the range [0,1]. As of now, the methods present in TorchEsegeta, belonging to the two libraries Captum and CNN Visualisation, have been extended and tested for the reconstruction models.

2.4. Experimental Setup

This research analysed the ReconResNet model [9] for the task of undersampled MRIs. Following the original article, two different publicly available brain MRI datasets were used for all experiments with the undersampled MRIs - OASIS [36] and IXI (available online:). The MRIs from the datasets were treated as fully sampled ground truth images and were artificially undersampled. The model was trained by supplying undersampled images as input (i.e., images with artefacts), and the prediction was compared with the ground-truth images.

3. Results

This section presents the results obtained using some of the methods discussed in the methodology section.

3.1. Uncertainty estimation

The rightmost heatmaps in Figure 1 are generated using the 'hot' colourmap from the Matplotlib library, which means that black represents the lowest uncertainty and bright yellow represents a high amount of uncertainty. As the pictures depict, the model is quite certain about the areas outside of the brain area, and hence no noisy undersampling artefact of the input image is transferred to the reconstructed output image. The most uncertainty arises in the skull and brain tissues. Also, it can be seen that the model is quite robust against dropout, but it produces higher uncertainty when the model weights are perturbed.
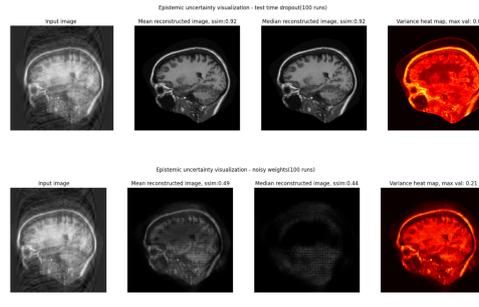
Figure 1: The top row is the outcome of the Monte-Carlo simulation, whereas the bottom row is the outcome of the model's weight perturbation.
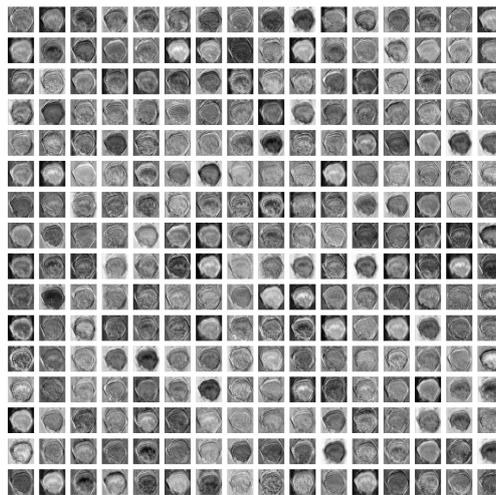

Figure 2: This image depicts all the latent space feature maps. This plot is helpful in understanding what shape or structure the different filters are looking for.

The uncertainty map in Figure 1 clearly shows an increase in uncertainty, and quantitatively the variance of the maximum uncertainty value also increases.

3.2. Model Understanding

As shown in Figure 3a, the second method of latent space exploration shows that there are indeed three clusters that exist in the latent space data with reduced dimensionality. The authors took three representative samples from each cluster and then backtracked them to the input space. The lower subplot shows the three different input images corresponding to those three samples.

Two codes have been selected from the latent code space, linearly interpolated between the codes, and reconstructed all the codes. In Figure 3b, you can see how one image is slowly interpolating into another. This experiment helps to understand whether the manifold learnt in the latent space is continuous and sufficiently covers the whole space or not. As the transition is quite smooth and intermediate images are not that blurry, one can say that the model has
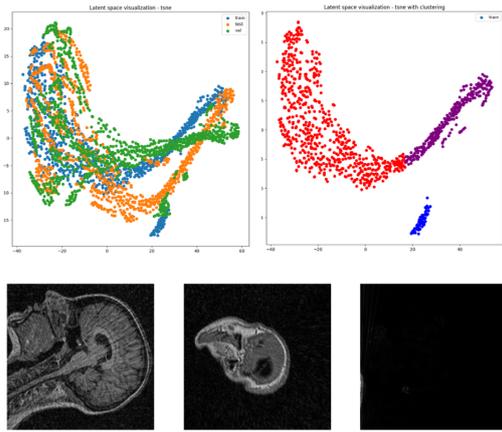
(a) Principal component analysis using the tSNE method. The top left subplot shows the distribution shift between the training, validation, and test dataset. The top right subplot showed the clustering outcome when the 2D embeddings were clustered to find any pattern.
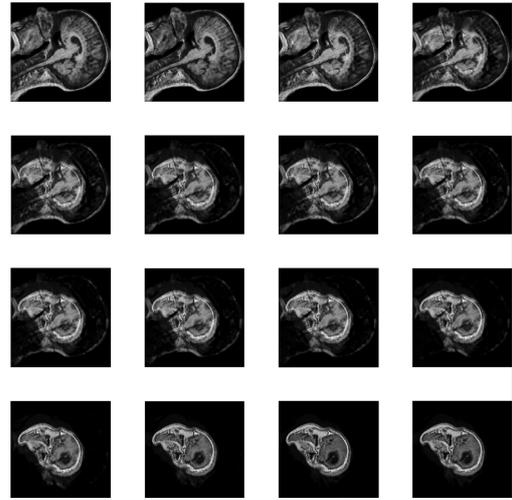
(b) Outcome of walking the latent space experiment.

Figure 3: Latent space exploration

(a) Reconstructed images
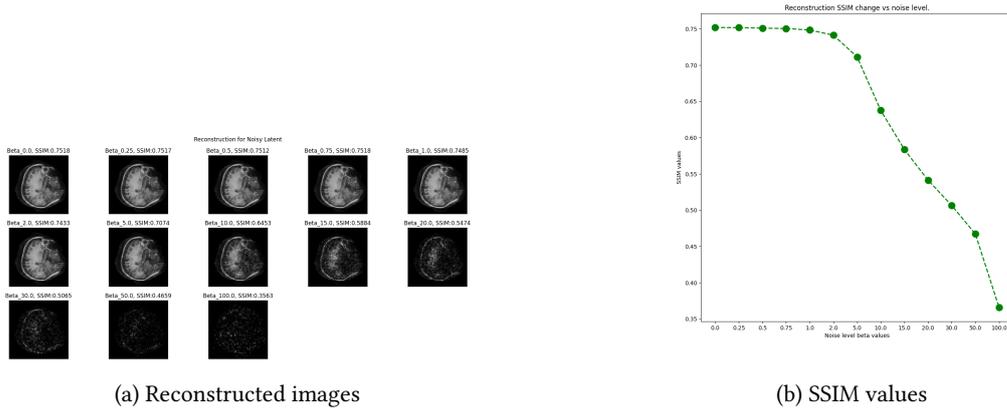

(b) SSIM values

Figure 4: Reconstruction with different amounts of noise added to the latent

learnt a manifold that can cover the latent space sufficiently and continuously.

Figure 4a and Figure 4b are from the decoder's noise tolerance experiment. The top image shows how the reconstructed image changes depending on the amount of noise added to the latent space representation. While the top image is a visual representation, the bottom image quantitatively shows the result. The beta values along the x-axis are the noise level, and the y-axis depicts the SSIM values for the reconstructed image against the ground truth. The starting of the graph is horizontal up to a certain beta value. This shows the robustness to noise of the model in the region.
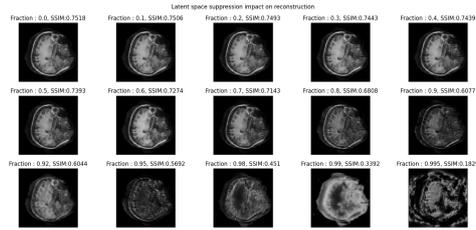
Figure 5: Visual and quantitative quality of reconstructed image with different amounts of suppressed latent feature maps
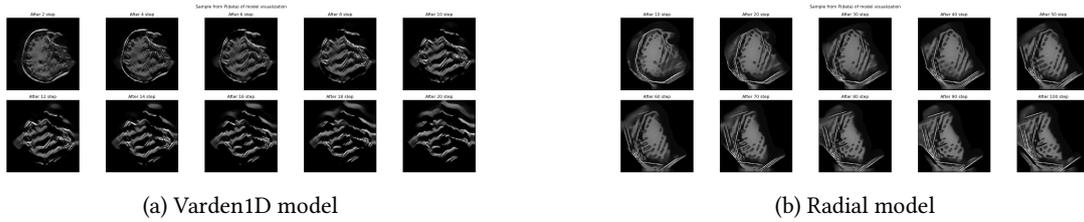

(a) Varden1D model

(b) Radial model

Figure 6: Samples from estimated $P_{model}(x)$

Figure 7: Activation maximisation for a different group of pixels based on different threshold values.

Figure 5 shows the quality (visual and quantitative) of the reconstructed image when randomly suppressing a certain fraction of the latent feature maps.

The result of the probabilistic model understanding approximation experiment is depicted in Figure 6a. The outcome shows different $P_{model}(x|h)$ output after running different numbers of CD steps. This is the result of the Varden1D reconstruction model. The same experiment was also performed for the radial model, Figure 6b shows the same.

In Figure 7, the result of the targeted activation maximisation experiment is shown here. The result shows different input images, which maximises the output of the model against different selected threshold values. It is interesting to see, for different threshold values, how the activation of the is maximised for different parts of the brain when the input is mere noise. The images are from threshold value 0.0 to 0.6 as you go along from left to right and top to bottom.

Figure 8a presents the result of the input anomaly test experiment, in which the authors check the robustness or ability of the network to reconstruct the lesion in brain cells, which the model has not seen during training. When presenting lesion images with various pixel values as output, the model generates plausible results when the pixel value is in a higher range. When the pixel value of the lesion is similar to the neighbouring brain cells, the reconstructed pixel values are underpredicted, and the reconstruction is not that prominent.

In the activation/reconstruction comparison experiment, it was found that the reconstruction model failed to reconstruct the out-of-distribution input images. Figure 8b shows the experiment
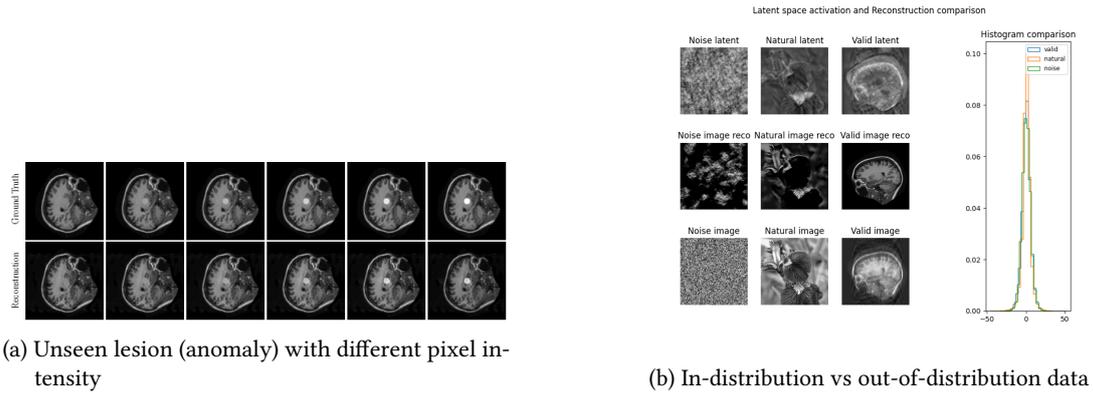
(a) Unseen lesion (anomaly) with different pixel intensity

(b) In-distribution vs out-of-distribution data

Figure 8: Reconstruction comparisons for extreme tests
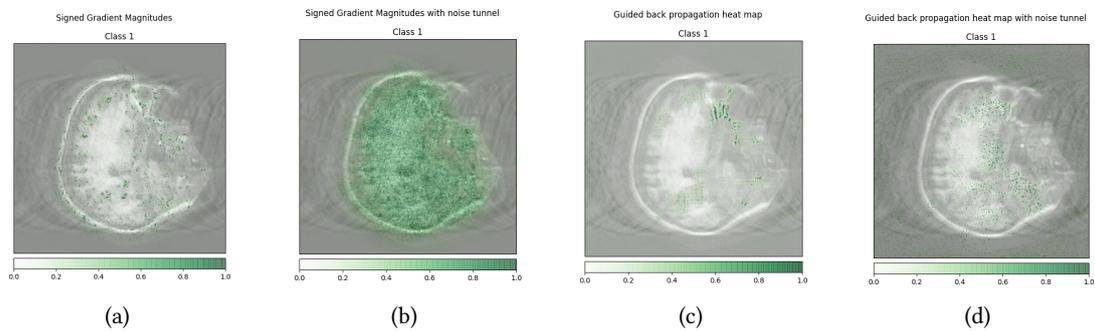

(a) (b) (c) (d)

Figure 9: (a-b) Saliency and (c-d) Guided back-propagation

output. For the flower image in the middle, the model reconstructed the parts that are similar to the in-distribution brain image. But it made most of the reconstructed pixel zero, following the distribution of brain images. Furthermore, the histogram shows that for the flower image, the latent space activation is zero for more neurons compared to the in-distribution data. That means that most neurons are not activated when presented with the flower image.

3.3. Interpretability

The attribution results of the reconstructed model generated by TorchEsegeta [8] are shown in Figures 9 and 10. As elaborated in the Methods section, an Otsu-based wrapper has been used for generating these attributions. For all figures, the positive attribution of the corresponding methods is overlaid on top of the input images.

4. Conclusion and Future Work

This research presented several methods for understanding ReconResNet, a deep learning-based undersampled MRI reconstruction model. This paper serves as a starting point for exploring these and other methods for the explainability and interpretability of such models. Here, some
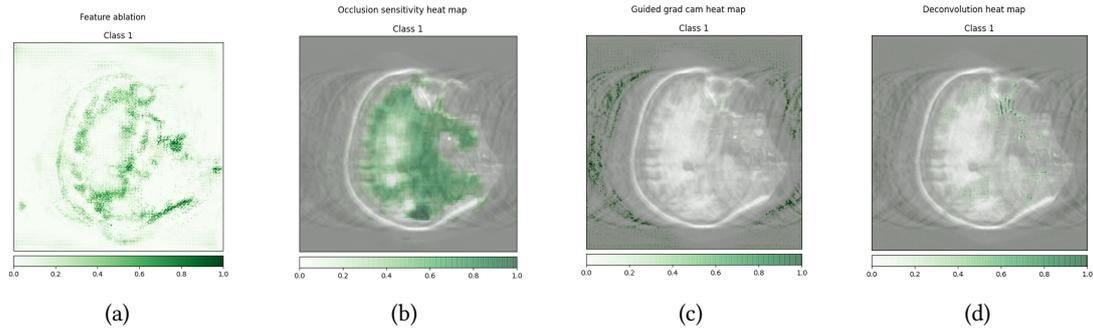
Figure 10: (a) Feature Ablation, (b) Occlusion sensitivity, (c) Guided grad-cam, and (d) Deconvolution

of the proposed methods were applied at a limited scale. In the future, all these methods will be evaluated in more detail and a user study will be conducted involving medical professionals to evaluate the advantage of these methods in terms of trust-building in clinical practise. In addition, more undersampling techniques for reconstruction models, different datasets, and comparisons between different models will also be performed in the near future. The remaining interpretability methods in the TorchEsegeta pipeline would also be extended and evaluated for reconstruction models.

References

[1] L. Mou, P. Ghamisi, X. X. Zhu, Unsupervised spectral–spatial feature learning via deep residual conv–deconv network for hyperspectral image classification, IEEE Transactions on Geoscience and Remote Sensing 56 (2017) 391–406.

[2] J. Zhang, Y. Xie, Y. Xia, C. Shen, Attention residual learning for skin lesion classification, IEEE transactions on medical imaging 38 (2019) 2092–2103.

[3] D. Pakhomov, V. Premachandran, M. Allan, M. Azizian, N. Navab, Deep residual learning for instrument segmentation in robotic surgery, in: International Workshop on Machine Learning in Medical Imaging, Springer, 2019, pp. 566–573.

[4] C. M. Hyun, H. P. Kim, S. M. Lee, S. Lee, J. K. Seo, Deep learning for undersampled mri reconstruction, Physics in Medicine & Biology 63 (2018) 135007.

[5] W. Jifara, F. Jiang, S. Rho, M. Cheng, S. Liu, Medical image denoising using convolutional neural network: a residual learning approach, The Journal of Supercomputing 75 (2019) 704–718.

[6] S. Chatterjee, F. Saad, C. Sarasaen, S. Ghosh, R. Khatun, P. Radeva, G. Rose, S. Stober, O. Speck, A. Nürnberger, Exploration of interpretability techniques for deep covid-19 classification using chest x-ray images, arXiv preprint arXiv:2006.02570 (2020).

[7] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, et al., Captum: A unified and generic model interpretability library for pytorch, arXiv preprint arXiv:2009.07896 (2020).

[8] S. Chatterjee, A. Das, C. Mandal, B. Mukhopadhyay, M. Vipinraj, A. Shukla, R. Nagaraja Rao, C. Sarasaen, O. Speck, A. Nürnberger, Torchesegeta: Framework for interpretability and

explainability of image-based deep learning models, Applied Sciences 12 (2022) 1834.

[9] S. Chatterjee, M. Breitkopf, C. Sarasaen, H. Yassin, G. Rose, A. Nürnberger, O. Speck, Reconresnet: Regularised residual learning for mr image reconstruction of undersampled cartesian and radial data, Computers in Biology and Medicine (2022) 105321.

[10] M. Lustig, D. Donoho, J. M. Pauly, Sparse mri: The application of compressed sensing for rapid mr imaging, Magnetic resonance in medicine 58 (2007) 1182–1195. doi:10.1002/mrm.21391.

[11] K. Hammernik, T. Klatzer, E. Kobler, M. P. Recht, D. K. Sodickson, T. Pock, F. Knoll, Learning a variational network for reconstruction of accelerated mri data, Magnetic resonance in medicine 79 (2018) 3055–3071.

[12] A. Sriram, J. Zbontar, T. Murrell, A. Defazio, C. L. Zitnick, N. Yakubova, F. Knoll, P. Johnson, End-to-end variational networks for accelerated mri reconstruction, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2020, pp. 64–73.

[13] S. Chatterjee, A. Sciarra, M. Dünnwald, R. V. Mushunuri, R. Podishetti, R. N. Rao, G. D. Gopinath, S. Oeltze-Jafra, O. Speck, A. Nürnberger, Shuffleunet: Super resolution of diffusion-weighted mris using deep learning, in: 2021 29th European Signal Processing Conference (EUSIPCO), IEEE, 2021, pp. 940–944.

[14] S. Chatterjee, C. Sarasaen, G. Rose, A. Nürnberger, O. Speck, Ddos-unet: Incorporating temporal information using dynamic dual-channel unet for enhancing super-resolution of dynamic mri, arXiv preprint arXiv:2202.05355 (2022).

[15] V. Edupuganti, M. Mardani, S. Vasanawala, J. Pauly, Uncertainty quantification in deep mri reconstruction, IEEE Transactions on Medical Imaging 40 (2020) 239–250.

[16] S. Chatterjee, A. Sciarra, M. Dünnwald, A. B. Talagini Ashoka, S. Oeltze-Jafra, O. Speck, A. Nürnberger, Uncertainty quantification for ground-truth free evaluation of deep learning reconstructions, in: Joint Annual Meeting ISMRM-ESMRMB 2022, 2022, p. 5631.

[17] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, Information Fusion 58 (2020) 82–115. URL: https://www.sciencedirect.com/science/article/pii/S1566253519308103. doi:https://doi.org/10.1016/j.inffus.2019.12.012.

[18] S. MacDonald, H. Foley, M. Yap, R. L. Johnston, K. Steven, L. T. Koufariotis, S. Sharma, S. Wood, V. Addala, J. V. Pearson, et al., Generalising uncertainty improves accuracy and safety of deep learning analytics applied to oncology, Scientific Reports 13 (2023) 7395.

[19] Y.-L. Tsai, C.-Y. Hsu, C.-M. Yu, P.-Y. Chen, Formalizing generalization and adversarial robustness of neural networks to weight perturbations, Advances in Neural Information Processing Systems 34 (2021) 19692–19704.

[20] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, R. R. Salakhutdinov, Improving neural networks by preventing co-adaptation of feature detectors, arXiv preprint arXiv:1207.0580 (2012).

[21] Y. Gal, Z. Ghahramani, Dropout as a bayesian approximation: Representing model uncertainty in deep learning, in: international conference on machine learning, PMLR, 2016, pp. 1050–1059.

[22] C. Gonzalez, K. Gotkowski, A. Bucher, R. Fischbach, I. Kaltenborn, A. Mukhopadhyay, Detecting when pre-trained nnu-net models fail silently for covid-19 lung lesion segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2021, pp. 304–314.

[23] L. Fetty, M. Bylund, P. Kuess, G. Heilemann, T. Nyholm, D. Georg, T. Löfstedt, Latent space manipulation for high-resolution medical image synthesis via the stylegan, Zeitschrift für Medizinische Physik 30 (2020) 305–314.

[24] C. Qin, S. Wang, C. Chen, W. Bai, D. Rueckert, Generative myocardial motion tracking via latent space exploration with biomechanics-informed prior, Medical Image Analysis 83 (2023) 102682.

[25] Into the latent space, Nature Machine Intelligence 2 (2020) 151–151. URL: https://doi.org/10.1038/s42256-020-0164-7. doi:10.1038/s42256-020-0164-7.

[26] G. P. Renieblas, A. T. Nogués, A. M. González, N. G. León, E. G. Del Castillo, Structural similarity index family for image quality assessment in radiological images, Journal of medical imaging 4 (2017) 035501.

[27] H. Nyquist, Certain topics in telegraph transmission theory, Transactions of the American Institute of Electrical Engineers 47 (1928) 617–644.

[28] C. E. Shannon, Communication in the presence of noise, Proceedings of the IRE 37 (1949) 10–21.

[29] I. J. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, Cambridge, MA, USA, 2016. http://www.deeplearningbook.org.

[30] A. Fawzi, S.-M. Moosavi-Dezfooli, P. Frossard, Robustness of classifiers: from adversarial to random noise, Advances in neural information processing systems 29 (2016).

[31] T. Y. Liu, Y. Yang, B. Mirzasoleiman, Friendly noise against adversarial noise: a powerful defense against data poisoning attack, Advances in Neural Information Processing Systems 35 (2022) 11947–11959.

[32] S. Chatterjee, A. Sciarra, M. Dünnwald, P. Tummala, S. K. Agrawal, A. Jauhari, A. Kalra, S. Oeltze-Jafra, O. Speck, A. Nürnberger, Strega: Unsupervised anomaly detection in brain mris using a compact context-encoding variational autoencoder, Computers in Biology and Medicine 149 (2022) 106093.

[33] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014. arXiv:1312.6034.

[34] D. Erhan, Y. Bengio, A. Courville, P. Vincent, Visualizing higher-layer features of a deep network, Technical Report, Univeristé de Montréal (2009).

[35] J. Linmans, S. Elfwing, J. van der Laak, G. Litjens, Predictive uncertainty estimation for out-of-distribution detection in digital pathology, Medical Image Analysis 83 (2023) 102655.

[36] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, R. L. Buckner, Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults, Journal of cognitive neuroscience 19 (2007) 1498–1507.