MethodsX 10 (2023) 102009

Contents lists available at ScienceDirect

MethodsX

journal homepage: www.elsevier.com/locate/mex

Generating post-hoc explanation from deep neural networks for multi-modal medical image analysis tasks

Weina Jin[a,*], Xiaoxiao Li[b], Mostafa Fatehi[c], Ghassan Hamarneh[a]

[a] School of Computing Science, Simon Fraser University, Burnaby, BC, V5A 1S6, Canada
[b] Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, BC, V6T 1Z4, Canada
[c] Division of Neurosurgery, The University of British Columbia, Vancouver, BC, V5Z 1M9, Canada

ARTICLE INFO

Method name:
Post-hoc feature attribution map explanation methods

Keywords:
Interpretable machine learning
Explainable artificial intelligence
Medical image analysis
Multi-modal medical image
Post-hoc explanation

ABSTRACT

Explaining model decisions from medical image inputs is necessary for deploying deep neural network (DNN) based models as clinical decision assistants. The acquisition of multi-modal medical images is pervasive in practice for supporting the clinical decision-making process. Multi-modal images capture different aspects of the same underlying regions of interest. Explaining DNN decisions on multi-modal medical images is thus a clinically important problem. Our methods adopt commonly-used post-hoc artificial intelligence feature attribution methods to explain DNN decisions on multi-modal medical images, including two categories of gradient- and perturbation-based methods.

- Gradient-based explanation methods – such as Guided BackProp, DeepLift – utilize the gradient signal to estimate the feature importance for model prediction.
- Perturbation-based methods – such as occlusion, LIME, kernel SHAP – utilize the input-output sampling pairs to estimate the feature importance.
- We describe the implementation details on how to make the methods work for multi-modal image input, and make the implementation code available.

Specifications table

Subject area:	Computer Science
More specific subject area:	Medical Image Analysis; Interpretable Machine Learning; Explainable Artificial Intelligence
Name of your method:	Post-hoc feature attribution map explanation methods
Name and reference of original method:	Gradient [11], deconvolution [12], Guided BackProp [13], SmoothGrad [14], input × gradient [16], integrated gradients [17], DeepLift [15], gradient SHAP [18], Grad-CAM [19], Guided GradCAM [19], occlusion [12,20], feature ablation, feature permutation [21], LIME [22], Shapley value sampling [23,24], kernel SHAP [18]
Resource availability:	The code of our implementation is available at: http://github.com/weinajin/multimodal_explanation

Method details

Background: multi-modal medical images

Multi-modal medical images are multiple medical images from different image modalities that visualize the same underlying regions of interest, such as cells, tissues, lesions, or organs. Each imaging modality provides complementary information to facilitate

DOI of original article: 10.1016/j.media.2022.102684
* Corresponding author.
E-mail address: weinaj@sfu.ca (W. Jin).
Social media: (W. Jin)

https://doi.org/10.1016/j.mex.2023.102009
Received 5 December 2022; Accepted 7 January 2023
Available online 10 January 2023

clinical reasoning. Compared to a single-modal medical image, multi-modal medical images can better visualize the regions of interest by leveraging the visualization capabilities of multiple image-capturing devices, or multiple parametric settings or perspectives from the same device.

Some examples of multi-modal medical images, which are commonly used in clinical settings, include: chest X-ray images with multiple projections, such as frontal and lateral views; magnetic resonance imaging (MRI) with multiple pulse sequences; morphological and functional MRI of the brain; CT images viewed at different levels and windows to observe different anatomical structures such as bones, lungs, and other soft tissues [1]; multi-modal endoscopy imaging [2]; photographic, dermoscopic, and hyper-spectral images of a skin lesion [3,4]; multiple stained microscopic or histopathological images [5,6]; and dual-modality imaging of positron emission tomography-computed tomography (PET-CT) [7].

Generating heatmap explanation for multi-modal medical image analysis tasks

In medical imaging-related predictive tasks, the deep neural network (DNN) models output a prediction given an input image. In clinical applications, clinical users may request an explanation from the DNN model in addition to its prediction. This is the AI interpretability or explainability problem. There are various explainable artificial intelligence (XAI) techniques to tackle this problem, and they can be categorized into two paradigms: one is to use an inherently interpretable model, and the other is to apply post-hoc explanation methods to analyze and explain the decisions of a trained or deployed black-box model. Our work focuses on post-hoc XAI algorithms.

To generate explanations for tasks on multi-modal medical images, we chose the form of feature attribution map (heatmap) explanation, as it is the most commonly-used explanation form, and is easily understandable by physicians, as demonstrated in our pilot user study [8]. Given an input image x in the input space X, and a predictive model f that maps x from the input space to the output space Y: $f : X \rightarrow Y$, the heatmap explanations give rationales on why the model f maps x to the target output y (y could be the model's prediction, or other user-designated output values). The heatmap serves as a fuzzy mask over the input image to indicate the relative importance of each input image feature for the model prediction. A higher value on the heatmap indicates that the image region is more important. There are two major groups of post-hoc heatmap methods:

1. Gradient-based explanation algorithms: they access the internal signals of the model – such as gradients and/or neuron activations – to estimate the model decision process. Depending on the specific method, usually there are some prerequisites on the model type to apply the gradient-based explanation methods. For example, in order to compute the gradient, the model needs to be differentiable; and in order to get the neuron activations, the model needs to have the architecture of a neural network. Since gradient-based explanation utilizes backpropagation or its modification to compute the explanation, the computational time to generate an explanation is relatively fast.
2. Perturbation-based explanation algorithms: unlike gradient-based explanation, perturbation-based explanation methods do not need access to the model's internal parameters. They only utilize input-prediction pairs to gage the model decision process. Because of this property, this group of algorithms is usually model-agnostic, i.e. does not depend on the model type and can explain any black-box models. Since the computation needs to sample multiple input-output pairs, the computational time is typically longer than gradient-based algorithms. We summarize the comparison between the two groups of algorithms in Table 1.

In our implementation, since the input image is a multi-modal medical image, depending on the specific explanation needs, the generated heatmap explanation can be a single heatmap mask that indicates the joint regions of important features for all modalities (given that the multi-modal images are co-registered), or multiple heatmap masks, where each mask provides an explanation on one image modality. In our application cases, whenever possible, we generated a multi-modal heatmap explanation using multiple heatmap masks. We used the Captum interpretability library [9] for PyTorch [10] to generate explanations on multi-modal medical images. Next, we briefly review and compare each XAI algorithm, and describe our implementation details.

Gradient-based heatmap algorithms

We generated gradient-based heatmap explanations with the following XAI algorithms: gradient [11], deconvolution [12], Guided BackProp [13], SmoothGrad [14], input × gradient [16], integrated gradients [17], DeepLift [15], gradient SHAP [18], Grad-CAM [19], Guided GradCAM [19].

Table 1
Comparison of gradient- and perturbation-based feature attribution explanation algorithms.

Algorithm	Require access to model parameters	Model dependency	Computational efficiency
Gradient-based	Yes	Differentiable models	Fast
Perturbation-based	No	Model-agnostic	Slow

Review of gradient-based explanation methods

Gradient explanation computes the gradient of the target output y with respect to the input x: $\frac{\partial f_y(x)}{\partial x}$, and uses the gradient to represent how sensitive the output changes when the input changes. Gradient is the weight term in the first-order Taylor approximation of the decision function f. The vanilla gradient explanation, however, has two major problems: (1) Noisy gradient signal: the gradient signal usually generates visually noisy heatmap explanations [14]; and (2) Gradient saturation: the gradient is saturated (i.e.: zero gradient) for some input features that coincide with the flat area of the model decision manifold, i.e.: perturbations of features in x will not produce a change in the output. This may lead to underestimating the importance of features [15].

Follow-up methods on the variant of gradient were aimed at overcoming the weakness of the vanilla gradient explanation. To deal with the noisy gradient problem, deconvolution and Guided BackProp modify the backpropagation rule by discarding negative values during backpropagation calculation: deconvolution discards the negative gradient signals [12], and Guided BackProp discards the negative gradients and negative data values from the bottom layer [13]. SmoothGrad is another line of work that reduces gradient noise by averaging. It randomly samples the input x's neighborhood and computes a local average of the gradient heatmap [14].

To deal with the gradient saturation problem, instead of computing the local gradient at a given point x, follow-up works introduce a reference input x_0 (which is usually a non-informative input), and the heatmap explains the decision difference between $y_0 = f(x_0)$ and $y = f(x)$. As a baseline method, input × gradient sets the reference input $x_0 = 0$, and explains $f(x) - f(0)$ via the first-order Taylor approximation of f, which equals to multiplying the gradient by the input x [16]. Since input × gradient uses the first-order Taylor approximation, which is a linear function and may not truthfully approximate the original function f, integrated gradients improves the baseline method input × gradient by approximating $f(x) - f(x_0)$ via the path integral of the gradients along the line path from x_0 to x [17], and DeepLift improves the baseline input × gradient by modifying the backpropagation rule. It introduces the difference-from-reference operation to the backpropagation, so that to attribute $f(x) - f(x_0)$ to each input feature via the backward pass [15].

There are other gradient-based explanation methods: gradient SHAP approximates the Shapley value (which will be introduced in Section "Perturbation-based heatmap algorithms") of the feature by computing the expectation of gradients [18]. Grad-CAM combines the gradient signal with the neuron activation signal from a latent representation layer of DNN [19]. Guided GradCAM further combines the signal from Grad-CAM with the signal from Guided BackProp [19].

Implementation detail

In our implementation to generate gradient-based explanations for tasks on multi-modal medical images, using the Captum interpretability library [9], we specified the input and the target output value, and calculated heatmap explanations from gradient, deconvolution, Guided BackProp, SmoothGrad, input × gradient, Grad-CAM, and Guided GradCAM methods. Note that Grad-CAM and Guided GradCAM also need to specify the layer name to read the neuron activation signals, which we set as the topmost convolutional layer right before the multilayer perceptron decision layer. In addition, as described above, a reference input x_0 is required for integrated gradients, DeepLift, and gradient SHAP, and we set x_0 to be an all-zero vector of the same size as the input x.

Perturbation-based heatmap algorithms

We generated perturbation-based heatmap explanations with the following XAI algorithms: occlusion [12,20], feature ablation, feature permutation [21], LIME [22], Shapley value sampling [23,24], and kernel SHAP [18].

Review of perturbation-based explanation methods

Perturbation-based explanation methods compute the importance of a feature by probing the model f and getting samples of input-output pairs from the perturbed inputs. Occlusion [12,20], feature ablation, and feature permutation [21] methods directly compute the marginal contribution of a feature using the prediction difference before and after a feature removal. To perturb the input and remove input features, occlusion method uses a sliding window approach to remove image features that are covered by the window region; feature ablation method uses a feature mask; and feature permutation method shuffles the feature values within a batch of inputs.

Other perturbation-based methods utilize structured frameworks to quantify the feature importance during input perturbation: LIME (Local Interpretable Model-agnostic Explanations) [22] is an explanation framework that constructs a local interpretable model as a proxy of f to explain the feature importance. The local proxy model is constructed by learning from input-output data pairs that are sampled around the neighborhood region of the input. Shapley value sampling [23,24] and kernel SHAP [18] are both based on the framework of Shapley value [25]. Shapley value computes the weighted average of the marginal contribution of a feature among all possible feature combinations. It is the unique solution to the task of fairly attributing the prediction to each feature. Shapley value sampling method utilizes sampling to estimate the Shapley value, and reduces the computational complexity from exponential to polynomial time. Kernel SHAP also aims to improve the sampling efficiency by modifying the LIME framework (including the loss function, weighting kernel, and regularization term) to recover the Shapley value.

Implementation detail

In our implementation, we fed all perturbation-based explanation methods (except occlusion) with a superpixel mask of the input image, so that methods treat each superpixel — rather than each image voxel/pixel — as a feature during the feature removal or sampling process. This greatly reduced the number of features, and kept the computational time within a tolerable duration. The superpixel mask groups the input image pixel/voxel to their semantic similar neighbors. We used the simple linear iterative clustering

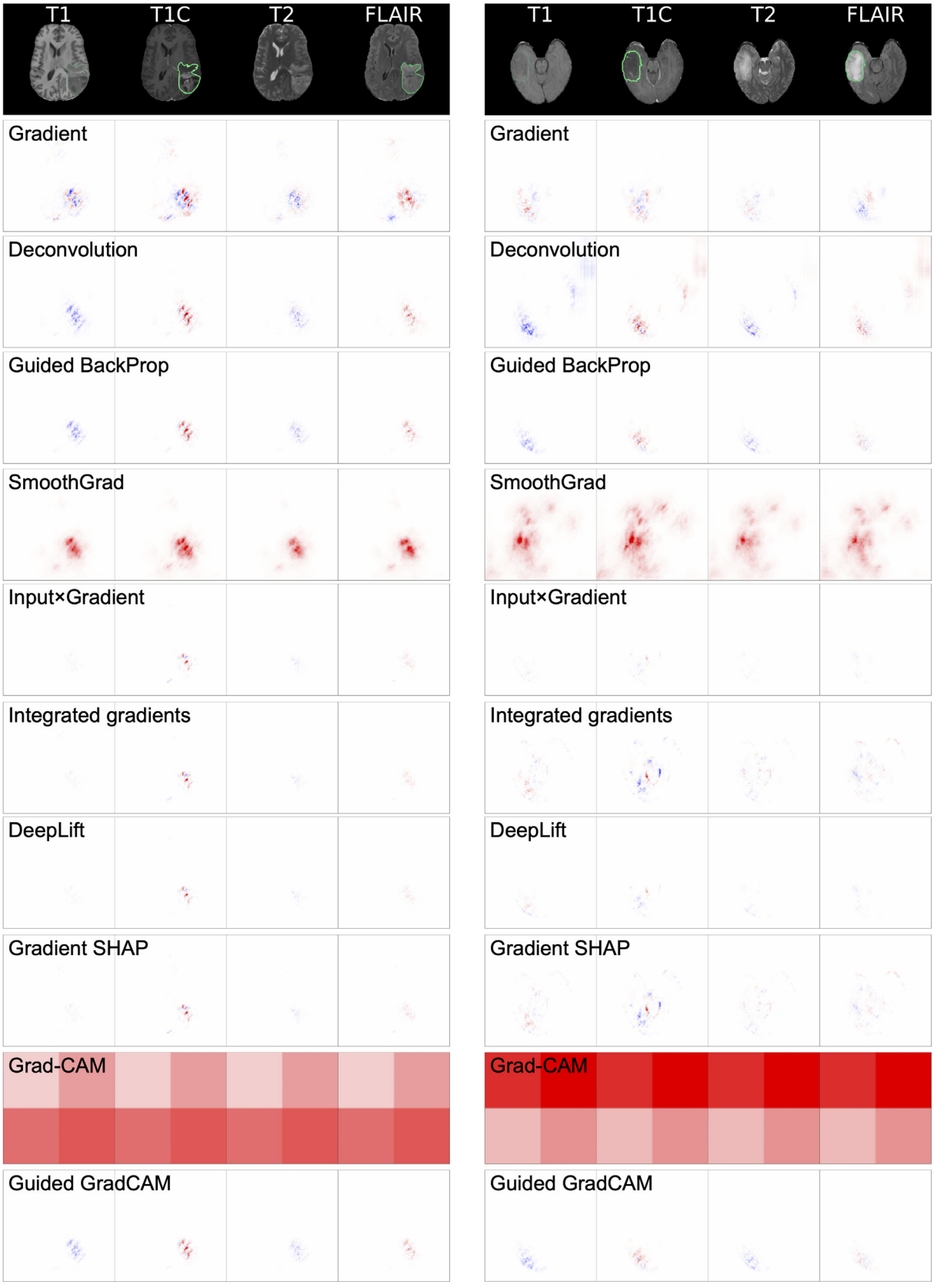

Fig. 1. Heatmap explanations generated from the gradient-based explanation methods.

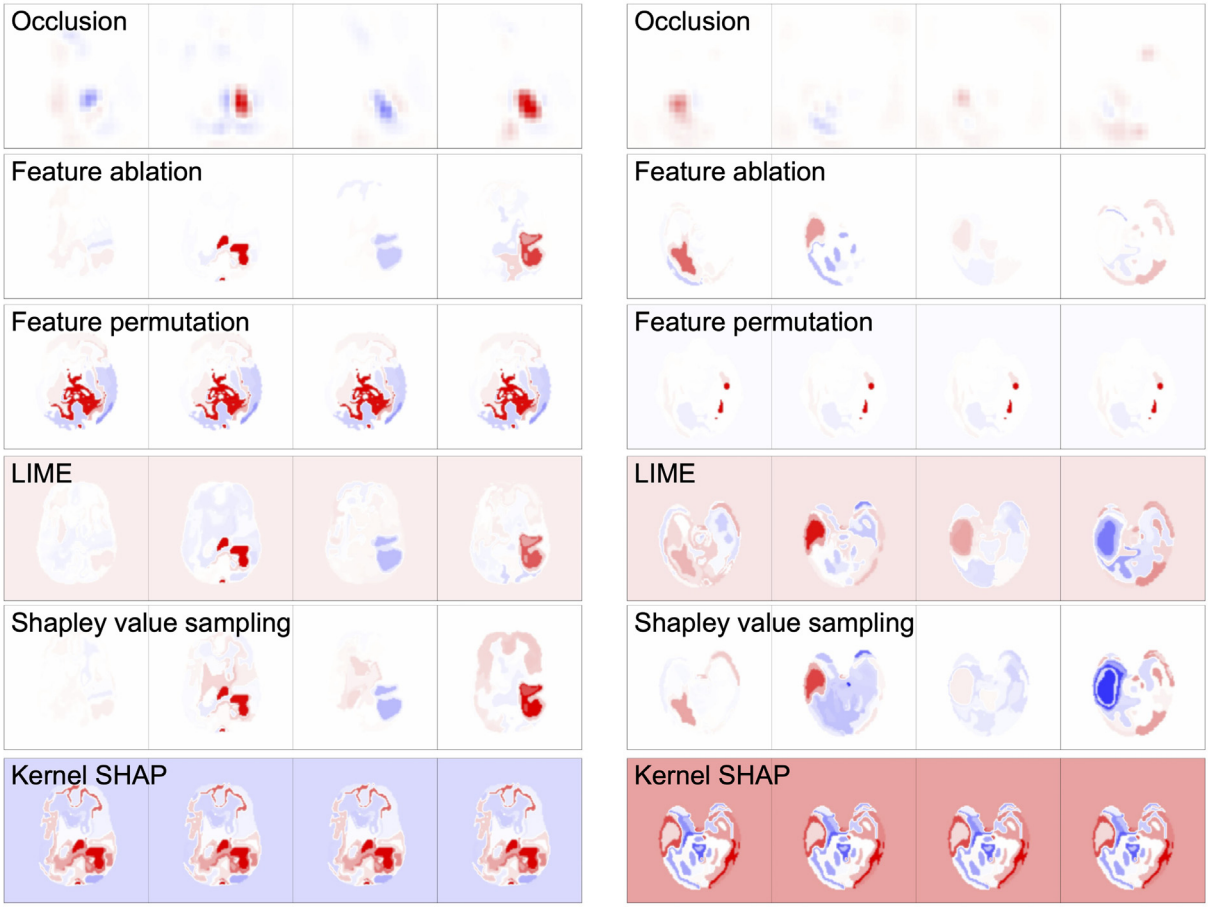

Fig. 2. Heatmap explanations generated from the perturbation-based explanation methods. The input MRIs are the same ones as those in the first row of Fig. 1.

(SLIC) [26] algorithm to segment the input images into superpixel masks, as SLIC is a commonly-used method and can also segment 3D images in the scikit-image library [27] implementation. SLIC utilizes *k*-means clustering to generate superpixel masks. To avoid segmenting superpixels for the non-informative blank background in our application, we fed a binary background mask to the SLIC algorithm to only compute superpixels for the foreground. The generated heatmap explanation assigns a feature importance score to an image region covered by the superpixel mask with the same label. To get heatmap explanations that are specific for each imaging modality, we generated a unique superpixel mask for each modality; if the image modalities were co-registered, the superpixel mask for each modality had the same shape of segmented mask, but each mask was assigned a unique label.

We filled the removed feature regions with uninformative values (all 0 s or random samples from a Gaussian distribution with the mean and standard deviation of the imaging modality). We set the batch size of input data to be four for the feature permutation method, and batch size to be one for the rest gradient- and perturbation-based methods. To balance explanation granularity and computational efficiency, we experimented with different parameter settings of the methods, such as the number of superpixels, the window size and stride step for occlusion method, the number of data pair samples for LIME and kernel SHAP methods. In our experiment, the time to generate a heatmap explanation for multi-modal medical images ranges from a few seconds to a maximum of 30–40 min (for Shapley value sampling method).

We visualized the generated heatmap explanations for the multi-modal brain tumor MRI in Figs. 1 and 2. The task is to classify the glioma (a type of primary brain tumor) as either a lower-grade or a high-grade glioma, based on the multi-modal brain MRI. We used the publicly-available BraTS (Multimodal Brain Tumor Segmentation) 2020 dataset [28–32]. All heatmaps were generated from the same 3D VGG-like [33] DNN model [8] and the same test images, with the input image and the tumor location outlined in the first row of Fig. 1. The greenness of the tumor contours indicates the relative importance of the imaging modality for model prediction. We showed two cases of a high-grade glioma (left column of Figs. 1 and 2) and a lower-grade glioma (right column of Figs. 1 and 2), and both cases were correctly classified by the model. Before visualization, we post-processed the heatmap by scaling the heatmap values to [−1, 1], and applied a Gaussian kernel to visually smoothen the heatmaps. The scaling operation kept the signs of positive, zero, and negative values of feature importance. We used the red-white-blue color map to represent positive, zero,

and negative values of feature importance, respectively, with the degree of redness representing the degree of importance, and the degree of blueness representing the degree of unimportance. The original MRI and heatmap explanation are in 3D format, and we only visualized one axial slice that has the maximum area of the tumor.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Weina Jin: Conceptualization, Methodology, Software, Investigation, Visualization, Writing – original draft. **Xiaoxiao Li:** Conceptualization, Methodology, Supervision, Writing – review & editing. **Mostafa Fatehi:** Conceptualization, Writing – review & editing. **Ghassan Hamarneh:** Conceptualization, Methodology, Writing – review & editing, Supervision, Funding acquisition.

Data Availability

We used the publicly-available BraTS (Multimodal Brain Tumor Segmentation) 2020 dataset (Menze, 2015; Bakas, et al., 2017, 2017a, 2017b, 2018).

Acknowledgments

This research was supported by the BC Cancer Foundation-BrainCare Fund, Borealis AI through the Borealis AI Global Fellowship Award, and Simon Fraser University Open Access Fund. This research was also enabled in part by the computational resources provided by NVIDIA and the Digital Research Alliance of Canada (alliancecan.ca).

References

[1] K. Harris, H. Adams, D. Lloyd, D. Harvey, The effect on apparent size of simulated pulmonary nodules of using three standard CT window settings, Clin. Radiol. 47 (1993) 241–244. URL: https://www.sciencedirect.com/science/article/pii/S0009926005811304, doi:10.1016/S0009-9260(05)81130-4.
[2] A. Hoffman, H. Manner, J.W. Rey, R. Kiesslich, A guide to multimodal endoscopy imaging for gastrointestinal malignancy - an early indicator, Nat. Rev. Gastroenterol Hepatol 14 (7) (2017) 421–434. Epub 2017 Jun 14. PMID: 28611477, doi:10.1038/nrgastro.2017.46.
[3] J. Kawahara, S. Daneshvar, G. Argenziano, G. Hamarneh, Seven-point checklist and skin lesion classification using multitask multimodal neural nets, IEEE J. Biomed. Health Inform. 23 (2019) 538–546, doi:10.1109/JBHI.2018.2824327.
[4] E. Zherebtsov, V. Dremin, A. Popov, A. Doronin, D. Kurakina, M. Kirillin, I. Meglinski, A. Bykov, Hyperspectral imaging of human skin aided by artificial neural networks, Biomed. Optics Express 10 (2019) 3545, doi:10.1364/boe.10.003545.
[5] R.K.M. Long, K.P. Moriarty, B. Cardoen, G. Gao, A.W. Vogl, F. Jean, G. Hamarneh, I.R. Nabi, Super resolution microscopy and deep learning identify Zika virus reorganization of the endoplasmic reticulum, Sci. Rep. 10 (2020), doi:10.1038/s41598-020-77170-3.
[6] Y. Song, D. Treanor, A. Bulpitt, D. Magee, 3D reconstruction of multiple stained histology images, J. Pathol. Inform. 4 (2013) 7, doi:10.4103/2153-3539.109864.
[7] T. Beyer, D.W. Townsend, T.M. Blodgett, Dual-modality PET/CT tomography for clinical oncology, Q. J. Nucl. Med. 46 (2002) 24–34.
[8] W. Jin, X. Li, M. Fatehi, G. Hamarneh, Guidelines and evaluation of clinical explainable AI in medical image analysis, Med. Image Anal. (2023), doi:10.1016/j.media.2022.102684.
[9] Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Reblitz-Richardson, O. (2020). Captum: a unified and generic model interpretability library for PyTorch. doi:10.48550/ARXIV.2009.07896.
[10] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, S. Chintala, PyTorch: an imperative style, high-performance deep learning library, in: Proceedings of the 33rd International Conference on Neural Information Processing Systems, Red Hook, NY, USA, Curran Associates Inc, 2019.
[11] Simonyan, K., Vedaldi, A., Zisserman, A., 2014. Deep inside convolutional networks: visualising image classification models and saliency maps. arXiv:1312.6034.
[12] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (Eds.), Computer Vision – ECCV 2014, Cham, Springer International Publishing, 2014, pp. 818–833.
[13] Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M., 2015. Striving for simplicity: the all convolutional net. arXiv:1412.6806.
[14] Smilkov, D., Thorat, N., Kim, B., Vie´gas, F., Wattenberg, M., 2017. Smooth-grad: removing noise by adding noise. arXiv:1706.03825.
[15] A. Shrikumar, P. Greenside, A. Kundaje, Learning important features through propagating activation differences, in: Proceedings of the 34th International Conference on Machine Learning - Volume 70, JMLR.org, 2017, pp. 3145–3153.
[16] Shrikumar, A., Greenside, P., Shcherbina, A., Kundaje, A., 2017. Not just a black box: learning important features through propagating activation differences. arXiv:1605.01713.
[17] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: Proceedings of the 34th International Conference on Machine Learning - Volume 70, JMLR.org, 2017, pp. 3319–3328.
[18] S.M. Lundberg, S.I. Lee, A Unified Approach to Interpreting Model Predictions, NeurIPS, 2017 URL: https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf .
[19] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: visual explanations from deep networks via gradient- based localization, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618–626, doi:10.1109/ICCV.2017.74.
[20] L.M. Zintgraf, T.S. Cohen, T. Adel, M. Welling, Visualizing deep neural network decisions: prediction difference analysis, in: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, 2017 April 24-26, 2017, Conference Track Proceedings, OpenReview.net. URL:.
[21] A. Fisher, C. Rudin, F. Dominici, All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously, J. Mach. Learn. Res. 20 (2019) 1–81 URL: http://jmlr.org/papers/v20/18-760.html .
[22] M.T. Ribeiro, S. Singh, C. Guestrin, Why should i trust you?": explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16, New York, USA, ACM Press, 2016, pp. 1135–1144, doi:10.1145/2939672.2939778.
[23] J. Castro, D. Go´mez, J. Tejada, Polynomial calculation of the Shapley value based on sampling, Comput. Oper. Res. 36 (2009) 1726–1730 URL: https://www.sciencedirect.com/science/article/pii/S0305054808000804 , doi:10.1016/j.cor.2008.04.004. selected papers presented at the Tenth International Symposium on Locational Decisions (ISOLDE X).
[24] E. Štrumbelj, I. Kononenko, An efficient explanation of individual classifications using game theory, J. Mach. Learn. Res. 11 (2010) 1–18.
[25] L.S. Shapley, Notes on the n-Person Game—II: The Value of an n-Person, Game (1951).

[26] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, S. Süsstrunk, SLIC superpixels compared to state-of-the-art superpixel methods, IEEE Trans. Pattern Anal. Mach. Intell. 34 (11) (2012) 2274–2282, doi:10.1109/TPAMI.2012.120.
[27] S. van der Walt, J.L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J.D. Warner, N. Yager, E. Gouillart, T. Yu, scikit-image: image processing in Python, PeerJ. 2 (2014) e453, doi:10.7717/peerj.453.
[28] B.H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, et al., The multimodal brain tumor image segmentation benchmark (BRATS), IEEE Trans. Med. Imaging 34 (10) (2015) 1993–2024, doi:10.1109/TMI.2014.2377694.
[29] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J.S. Kirby, et al., Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features, Nat. Sci. Data 4 (2017) 170117, doi:10.1038/sdata.2017.117.
[30] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, et al., "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge", arXiv preprint arXiv:1811.02629. 2018.
[31] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. Kirby, et al., Segmentation labels and radiomic features for the pre-operative scans of the TCGA-GBM collection, Cancer Imaging Arch. (2017), doi:10.7937/K9/TCIA.2017.KLXWJJ1Q.
[32] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. Kirby, et al., Segmentation labels and radiomic features for the pre-operative scans of the TCGA-LGG collection, Cancer Imaging Arch. (2017), doi:10.7937/K9/TCIA.2017.GJQ7R0EF.
[33] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 2015 May 7-9, 2015, Conference Track Proceedings. http://arxiv.org/abs/1409.1556.