Assesing the Interpretability of the Statistical Radiomic Features via Image Saliency Maps in Medical Image Classification Tasks

Oleksandr Davydko[1,*]

[1]*Technological University Dublin*

Abstract

The presented research aims to improve the interpretability of medical image classification models trained with statistical radiomic features. While showing classification results comparable with state-of-the-art convolutional neural network models, statistical radiomic features' interpretability is still understudied. Neural network models use saliency map approaches to provide a human operator with intuitive visualisation of the model's attention, but statistical radiomic-based models still have no such tools developed. This research aims to eliminate this gap and allow the saliency map generation for models trained with statistical radiomic features. Preliminary results show that the proposed approach may generate faithful saliency maps for the ResNet-50 classification model trained the first-order statistical radiomic features.

Keywords

Medical image classification, Texture analysis, Statistical radiomic features, Saliency maps

1. Context and motivation

1975 Haralik et al. introduced a way to obtain high-order image features from texture. These features describe the image texture properties using statistics. They allowed the large image to be compressed into a set of 14 features, which allowed solving the image classification tasks, as computation capabilities were limited then. Even though those feature-forming methods were originally developed for aerial photograph and satellite image classification, the most popular application of those methods is medical image classification.

Even after introducing high-performance GPUs and convolutional neural networks, some research like [1] still indicate that using statistical radiomic features can show near state-of-the-art results in medical image classification tasks. However, a lack of classification result interpretability could be a reason to choose automatically extracted features instead of statistical radiomic. In the current literature, the interpretability of statistical radiomic features is mainly addressed by attributing ad-hoc high-order statistical radiomic features, while neural network solutions utilize saliency map methods, which produce much more understandable

Late-breaking work, Demos and Doctoral Consortium, colocated with The 2nd World Conference on eXplainable Artificial Intelligence: July 17–19, 2024, Valletta, Malta

*Corresponding author.

d22125337@mytudublin.ie (O. Davydko)

visual explanations than standard numerical feature importance.

This research aims to introduce a method for generating a saliency map when a classification model is trained with statistical radiomic features, subsequently improving the explainability of statistical-radiomic-based classification models for medical images.

2. Related work

The statistical radiomic features were applied to solve many different medical image classification tasks, showing near state-of-the-art classification performance. Authors of the research [2] have attempted to fuse grey-level co-occurrence matrix (GLCM), grey-level run length matrix (GLRLM), and segmentation-based fractal texture analysis (SFTA) features to detect the COVID-19 lesions presence on the chest X-ray images. The fusion of features combined with feature selection done by principle components analysis allowed reaching 0.94 F1-score while distinguishing between healthy and COVID pneumonia lung images. Similar results were observed in the same task when using first-order statistics (FOS), GLCM, GLRLM, and grey-level size zone matrix (GLSZM) feature extraction methods in the work [1]. Here, the authors report the F1-score at a 0.98 rate. In research [3] authors report 0.975 accuracy while performing a classification of brain tumors on magnetic resonance images (MRI)

Current advances in the interpretability of radiomic-based models mostly include interpreting importances of used high-order radiomic features. Authors of the research in [4] use SHAP [5], which allowed them to identify the most influential features. In another research work [6], authors interpret radiomic feature groups importances by analysing logistic regression coefficients. A study [7] uses SHAP to reveal the most influential features to diagnose schizophrenia by brain magnetic resonance images (MRI). The authors of the work [8] use the same technique to find the connection between particular features and panic disorder signs.

At the same time, researchers utilise saliency map methods such as Integrated Gradients [9], layer-wise propagation [10], DeepLIFT [11], GradCAM [12] for interperting convolutional neural network predictions. The saliency map is much easier to understand from the point of view of human perception. For radiomic-based models, a little work discusses some analogs of saliency maps. In the research [13], authors discuss the interpretability of tumor tissue signature identification when local statistical radiomic features are used. The problem of interpretability was tackled by visualizing feature activation maps for a single high-order feature.

It can be stated, that the problem of statistical radiomic features interpretability is definitely in focus of researchers but requires further investigations. The main problem to investigate is the generation of understandable image saliency maps for radiomic-based models.

3. Design and methodology

This research describes methods for saliency map generation of the first- and high-order statistical radiomics features.

3.1. Mapping first-order radiomic features' attributions

First-order statistical radiomic features are formed by computing the frequency of some texture substructure appearing in the image. Some examples of such substructures are:

1. Pair of pixels with intensities i, j
2. Run-length of pixels with the same intensity i
3. Cluster of connected pixels with same intensity i

A proposed method of saliency map generation includes the computation of features' attributions and subsequently adding those attribution values to pixels involved in forming particular feature values (figure 1). Regarding the size of the statistical radiomics features matrices (256x256 at least), it is proposed to use convolutional neural networks to build the classifiers and gradient-based methods to obtain attributions (such as Integrated Gradients).

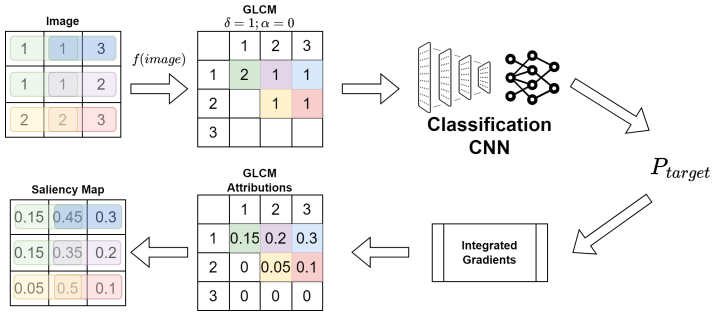

Figure 1: A visualisation of the first-order statistical features' attributions mapping to saliency map values

3.2. Mapping high-order radiomic features' attributions

A mapping of high-order statistical radiomic features could be defined as an extension of the first-order features mapping procedure. It is a noticeable fact that all high-order feature formulas are represented with differentiable functions. That allows the usage of gradient-based methods (such as Integrated Gradients, GradCAM, DeepLIFT) directly to obtain first-order feature attributions and subsequently map them to medical image pixel attributions by applying procedure from section 3.1. The graphical representation of this process is displayed in figure 2. This approach is feasible if the classification model is of a fully-connected neural network type. Known model-agnostic methods such as SHAP cannot be used to attribute the massive number (at least 65536 for GLCM and more for other methods) of the first-order features due to slowness in the calculation process, which leaves the usage of other classification models as an open question out of this research scope.

3.3. Experiment

The experiment is designed to test the faithfulness of the saliency maps generated by methods described in 3.2 and 3.2 and compare them with those generated by existing methods in

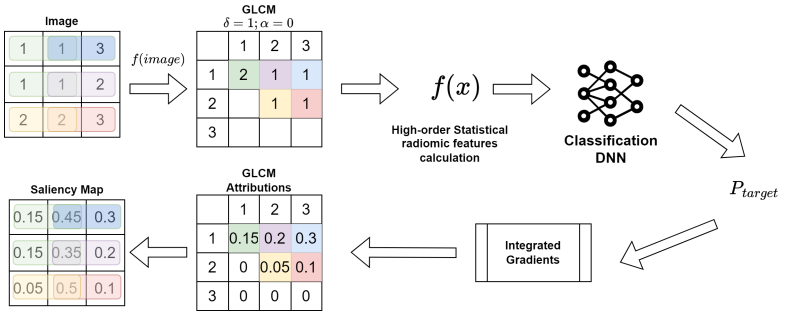

Figure 2: A visualisation of the high-order statistical features' attributions mapping to saliency map values

different classification tasks on X-ray and MRI images. The experiment's scheme is presented in the figure 3. The described experiment is run separately with different datasets to test the generalisability of the proposed approach.

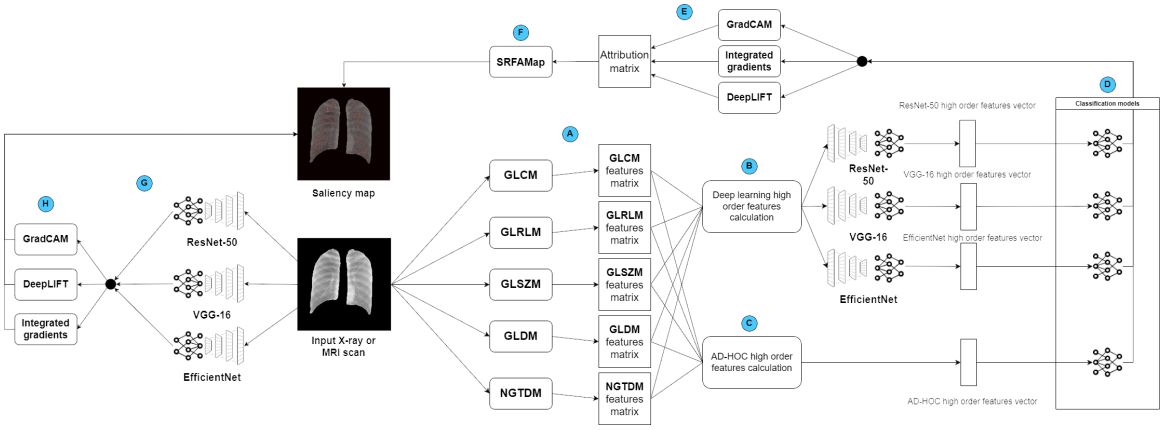

Figure 3: The experiment scheme. High-order radiomic features are computed from first-order statistical radiomic features (A) with deep learning (B) or ad-hoc formulas (C) and become an input for the classification of the multi-layer perceptron model (D). Attributions for the features are computed with GradCAM, DeepLIFT, and Integrated Gradients (E) and mapped to the image saliency map (F). The same is performed for plain image models (G and H).

3.3.1. Data Preparation

Each image in the dataset is converted into greyscale if needed, as statistical radiomic features are defined only for greyscale textures in the current literature. Images are left intact for the baseline pipeline (steps G and H in the figure 3).

3.3.2. Feature extraction

First-order statistical radiomic features are extracted with GLCM, GLRLM, GLSZM, grey-level dependency matrix (GLDM), and neighboring grey-tone difference matrix (NGTDM) methods

into matrices, as described in works [14, 15, 16, 17, 18]. While calculating statistical radiomics features, background pixels are not taken into account. The parameters for the mentioned methods are to be found with a hyperparameter search procedure. High-order statistical radiomics features are extracted out of matrices containing first-order radiomic features with formulas defined in [14, 15, 16, 17, 18] and concatenated into a single feature vector. The list of features calculated is provided in table 1

Table 1
The high-order statistical radiomic features employed in this research

Method	Features	Total	Reference
GLCM	Angular Second Moment, Contrast, Correlation, Sum of Squares, Inverse Difference Moment, Sum Average, Sum Variance, Sum Entropy, Entropy, Difference Variance, Difference Entropy, Information Measures of Correlation (2x)	13	[14]
GLRLM	Short Run Emphasis, Long Run Emphasis, Grey Level Non-Uniformity, Gray Level Non-Uniformity, Run Length Non-Uniformity, Run Length Non-Uniformity Normalized, Run Percentage, Grey Level Variance, Run Variance, Run Entropy, Low Gray Level Run Emphasis, High Gray Level Run Emphasis, Short Run Low Gray Level Emphasis, Short Run High Gray Level Emphasis, Long Run Low Gray Level Emphasis, Long Run High Gray Level Emphasis	16	[15]
GLSZM	Small Area Emphasis, Large Area Emphasis, Gray Level Non-Uniformity, Gray Level Non-Uniformity Normalized, Size-Zone Non-Uniformity, Size-Zone Non-Uniformity Normalized, Zone Percentage, Gray Level Variance, Zone Variance, Zone Entropy, Low Gray Level Zone Emphasis, High Gray Level Zone Emphasis, Small Area Low Gray Level Emphasis, Small Area High Gray Level Emphasis, Large Area Low Gray Level Emphasis, Large Area High Gray Level Emphasis	16	[16]
GLDM	Small Dependence Emphasis, Large Dependence Emphasis, Gray Level Non-Uniformity, Dependence Non-Uniformity, Dependence Non-Uniformity Normalized, Gray Level Variance, Dependence Variance, Dependence Entropy, Low Gray Level Emphasis, High Gray Level Emphasis, Small Dependence Low Gray Level Emphasis, Small Dependence High Gray Level Emphasis, Large Dependence Low Gray Level Emphasis, Large Dependence High Gray Level Emphasis	14	[17]
NGTDM	Coarseness, Contrast, Busyness, Complexity, Strength	5	[18]

Each of the described features has its unique formula for calculation. For example, a formula for the Contrast feature of the GLCM matrix is defined as:

$$Contrast = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (i-j)^2 GLCM(i,j)$$

Here Ng - number of grey shades in the image (usually equals 256), $GLCM$ - matrix containing grey-level co-occurrence first-order radiomic features.

3.3.3. Classification models and their training

Three types of classification models are trained. A first type receives first-order statistical radiomic feature matrices. Models of the first type are represented with convolutional neural

networks. In this particular experiment, ResNet-50, VGG-16, and EfficientNet architectures are used. Models of the second type are represented with special architecture, which implements high-order statistical formulas from [14, 15, 16, 17, 18]. A multi-layer perceptron follows the implementation of the block with formulas along with a sigmoid or softmax layer. The baseline models receive plain images as input and are represented by ResNet-50, VGG-16, and EfficientNet architectures. Each classification model of every type is trained from scratch. The training is held until the F1-score on the development set stops to improve for ten subsequent epochs. Adam algorithm is used as an optimizer with a $1e-4$ learning rate.

3.3.4. Statistical radiomic feature attribution and saliency maps

As all classification models described in 3.3.3 are represented by neural networks, is is possible to use gradient-based methods for feature attributing. In this research, it is proposed to use Integrated Gradients, GradCAM, and DeepLIFT without any modifications. Subsequently, for radiomic-based models, an additional mapping, which is described in section 3.1, is conducted to obtain the saliency map. For plain images, attributions may be used as ready saliency map without additional transformations.

3.3.5. Evaluation

The classification models' performance are measured with accuracy and F1-score metrics. The faithfulness of the saliency maps is measured numerically by Increase-In-Confidence and Average Drop metrics [19] and compared to the same metrics for plain statistical radiomic features attributions. Additionally, the same evaluation is conducted with Insertion Correlation (IC) and Deletion Correlation (DC) metrics [20] as they also taking into account magnitudes of saliency values. However, for IC and DC some iterations in the computation procedure will be merged to drastically reduce the number of iterations, as the 256x256 saliency map requires more than 3000 predictions to compute these metrics.

3.3.6. Datasets for experiment

Schenzen tuberculosis open-access dataset [21] contains 662 x-ray scans. The dataset is balanced; there are 326 images of the healthy lungs and 336 images of the lung with the signs of tuberculosis. No additional transformations are applied to this dataset, except those described in 3.3.1. During the experiment, the task of distinguishing between healthy and tuberculosis lungs was assessed with this dataset.COVIDx CXR-4 dataset [22] contains 84,818 chest X-ray scans. There are 65,681 scans containing COVID-19 lesions, and 19,137 are healthy lungs. Test and validation sets for this dataset were formed balanced while leaving the train set unbalanced to ensure faithful classification metrics. No additional transformations are applied to this dataset, except those described in 3.3.1. During the experiment, the task of distinguishing between healthy and COVID-19-damaged lungs was solved with this dataset. Cancer-Net BCa contains 253 breast MRI scans with evidence of breast cancer. During the experiment, the task of full remission prediction is considered.

4. Research question and hypothesis

This research aims to answer the next question: how a medical image saliency map could be generated to explain a classification result when a classification model is trained with the statistical radiomic features? According to this, the research hypothesis may be defined as follows:

Research hypothesis IF neural network trained with first- or high-order statistical radiomic features to perform the medical image classification AND aforementioned features attributed with Integrated Gradients, GradCAM, DeepLIFT AND image saliency map generated with proposed mapping method THEN Increase-In-Confidence, Average Drop, Insertion Correlation, and Deletion Correlation metrics for generated image saliency maps will be at least the same or statistically significantly higher as for direct feature attributions.

5. Preliminary results

The preliminary results, described in the author's previous work [23], indicate that the saliency maps generated method described in 3.1 could be considered faithful in terms of a numerical evaluation, maintaining Increase-In-Confidence metric at $50\% - 80\%$ level and Average Drop at $10\% - 38\%$. Also, results indicate that the ResNet-50 classification model trained with only first-order statistical radiomic features yields the same classification quality as the ResNet-50 model with raw image input, indicating that results are eligible for practical usage.

6. Expected final contribution to knowledge

The final contribution of the described research is expected to be a method for visually explaining a classification result via saliency maps when the classification model is trained with first- or high-order statistical radiomic features. The newly proposed method should allow for the explanation and validation of the results of the previous work which uses statistical radiomic features to solve medical image classification problems.

References

[1] H. Koyuncu, M. Barstuğan, Covid-19 discrimination framework for x-ray images by considering radiomics, selective information, feature ranking, and a novel hybrid classifier, Signal Processing: Image Communication 97 (2021) 116359. URL: https://www.sciencedirect.com/science/article/pii/S092359652100165X. doi:https://doi.org/10.1016/j.image.2021.116359.

[2] Ş. Öztürk, U. Özkaya, M. Barstuğan, Classification of coronavirus (covid-19) from x-ray and ct images using shrunken features, International Journal of Imaging Systems and Technology 31 (2020) 5–15. doi:10.1002/ima.22469.

[3] N. Zulpe, V. Pawar, Glcm textural features for brain tumor classification, IJ CSI 9 (2012) 354–359.

[4] J.-Y. Ye, P. Fang, Z.-P. Peng, X.-T. Huang, J.-Z. Xie, X.-Y. Yin, A radiomics-based interpretable model to predict the pathological grade of pancreatic neuroendocrine tumors, European Radiology 34 (2023) 1994–2005. doi:10.1007/s00330-023-10186-1.

[5] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30, Curran Associates, Inc., 2017, pp. 4765–4774.

[6] M. R. Orton, E. Hann, S. J. Doran, S. T. C. Shepherd, D. Ap Dafydd, C. E. Spencer, J. I. López, V. Albarrán-Artahona, F. Comito, H. Warren, J. Shur, C. Messiou, J. Larkin, S. Turajlic, D.-M. Koh, Interpretability of radiomics models is improved when using feature group selection strategies for predicting molecular and clinical targets in clear-cell renal cell carcinoma: insights from the tracerx renal study, Cancer Imaging 23 (2023). doi:10.1186/s40644-023-00594-3.

[7] M. Bang, J. Eom, C. An, S. Kim, Y. W. Park, S. S. Ahn, J. Kim, S.-K. Lee, S.-H. Lee, An interpretable multiparametric radiomics model for the diagnosis of schizophrenia using magnetic resonance imaging of the corpus callosum, Translational Psychiatry 11 (2021). doi:10.1038/s41398-021-01586-2.

[8] M. Bang, Y. W. Park, J. Eom, S. S. Ahn, J. Kim, S.-K. Lee, S.-H. Lee, An interpretable radiomics model for the diagnosis of panic disorder with or without agoraphobia using magnetic resonance imaging, Journal of Affective Disorders 305 (2022) 47–54. doi:10.1016/j.jad.2022.02.072.

[9] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17, JMLR.org, 2017, p. 3319–3328.

[10] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, K.-R. Müller, Layer-Wise Relevance Propagation: An Overview, in: W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, K.-R. Müller (Eds.), Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, volume 11700, Springer International Publishing, Cham, 2019, pp. 193–209. URL: http://link.springer.com/10.1007/978-3-030-28954-6_10. doi:10.1007/978-3-030-28954-6_10, series Title: Lecture Notes in Computer Science.

[11] A. Shrikumar, P. Greenside, A. Kundaje, Learning Important Features Through Propagating Activation Differences, 2019. URL: http://arxiv.org/abs/1704.02685, arXiv:1704.02685 [cs].

[12] 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.

[13] D. Vuong, S. Tanadini-Lang, Z. Wu, R. Marks, J. Unkelbach, S. Hillinger, E. I. Eboulet, S. Thierstein, S. Peters, M. Pless, M. Guckenberger, M. Bogowicz, Radiomics Feature Activation Maps as a New Tool for Signature Interpretability, Frontiers in Oncology 10 (2020) 578895. URL: https://www.frontiersin.org/articles/10.3389/fonc.2020.578895/full. doi:10.3389/fonc.2020.578895.

[14] R. M. Haralick, K. Shanmugam, I. Dinstein, Textural features for image classification, IEEE Transactions on Systems, Man, and Cybernetics SMC-3 (1973) 610–621. doi:10.1109/TSMC.1973.4309314.

[15] M. M. Galloway, Texture analysis using gray level run lengths, Computer Graphics and Image Processing 4 (1975) 172–179. doi:10.1016/S0146-664X(75)80008-6.

[16] G. Thibault, B. FERTIL, C. Navarro, S. Pereira, N. Lévy, J. Sequeira, J.-L. Mari, Texture indexes and gray level size zone matrix application to cell nuclei classification, 2009.

[17] C. Sun, W. G. Wee, Neighboring gray level dependence matrix for texture classification, Computer Vision, Graphics, and Image Processing 23 (1983) 341–352. doi:10.1016/0734-189X(83)90032-4.

[18] M. Amadasun, R. King, Textural features corresponding to textural properties, IEEE Transactions on Systems, Man, and Cybernetics 19 (1989) 1264–1274. doi:10.1109/21.44046.

[19] A. Chattopadhay, A. Sarkar, P. Howlader, V. N. Balasubramanian, Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks, in: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), 2018, pp. 839–847. doi:10.1109/WACV.2018.00097.

[20] T. Gomez, T. Fréour, H. Mouchère, Metrics for Saliency Map Evaluation of Deep Learning Explanation Methods, Springer International Publishing, 2022, p. 84–95. doi:10.1007/978-3-031-09037-0_8.

[21] S. Jaeger, S. Candemir, S. Antani, Y.-X. J. Wáng, P.-X. Lu, G. Thoma, Two public chest x-ray datasets for computer-aided screening of pulmonary diseases, Quant. Imaging Med. Surg. 4 (2014) 475–477.

[22] Y. Wu, H. Gunraj, C.-e. A. Tai, A. Wong, COVIDx CXR-4: An Expanded Multi-Institutional Open-Source Benchmark Dataset for Chest X-ray Image-Based Computer-Aided COVID-19 Diagnostics, 2023. URL: http://arxiv.org/abs/2311.17677, arXiv:2311.17677 [cs, eess].

[23] O. Davydko, V. Pavlov, P. Biecek, L. Longo, SRFAMap: A Method for Mapping Integrated Gradients of a CNN Trained with Statistical Radiomic Features to Medical Image Saliency Maps, Springer Nature Switzerland, 2024, p. 3–23. URL: http://dx.doi.org/10.1007/978-3-031-63803-9_1. doi:10.1007/978-3-031-63803-9_1.