Multi-metric Evaluation of the Statistical Radiomic Feature Attribution Mapping method for Explainable Radiomic Models

Oleksandr Davydko[1], Vladimir Pavlov[2] and Luca Longo[3]

[1]Technological University Dublin, Dublin, Ireland
[2]National Technical University of Ukraine Igor Sikorsky Kyiv Polytechnic Institute, Kyiv, Ukraine
[3]Artificial Intelligence and Cognitive Load Research Lab, University College Cork, Cork, Ireland

Abstract

Radiomic models can achieve strong diagnostic performance on medical images, but they remain difficult to trust when their predictions cannot be linked to clinically meaningful regions. This problem is especially important in the presence of dataset bias, where apparently accurate models may rely on spurious cues rather than pathology-relevant evidence. Existing spatial explanation approaches for radiomics, such as per-feature activation maps, provide only partial insight because they are not directly tied to model-level attributions and are often evaluated with a single metric. This research studies SRFAMap (Statistical Radiomic Feature Activation Map), a method that maps Integrated Gradients attributions from radiomic feature space back to the input pixel space, and evaluates it across three medical imaging datasets: tuberculosis chest X-rays, fatty liver ultrasound, and PneumoniaMNIST. The evaluation combines faithfulness correlation, relative input stability, and sparseness over Monte Carlo splits, with RFAM employed as a baseline, Wilcoxon signed-rank tests used for statistical comparison, and Region Of Interest (ROI) masking analysed qualitatively. Findings show that SRFAMap is consistently robust, achieving three- to four-fold better stability than RFAM and generally stronger faithfulness. More importantly, the study shows that good explanation scores do not automatically imply clinical validity: in tuberculosis, SRFAMap faithfully exposed model reliance on spurious correlations. These findings support multi-metric evaluation as a necessary requirement for trustworthy radiomic explainability and contribute to the body of knowledge by positioning SRFAMap as a useful auditing tool for medical imaging models.

Keywords

Explainable artificial intelligence, Radiomics, Medical imaging, Saliency maps, Integrated Gradients

1. Introduction

Radiomics-based models can often achieve strong predictive performance [1]. Yet predictive success alone is insufficient in medical settings. Clinicians need to know whether a model relies on pathology-relevant evidence or on accidental dataset cues. This challenge is well captured by the Clever Hans effect, where a model appears accurate while basing its decisions on the wrong evidence [2]. For radiomic predictors, the problem is even harder because the model operates in feature space, whereas clinical interpretation ultimately requires image-space evidence. Statistical Radiomic Feature Attribution Mapping (SRFAMap) [3] addresses this problem by mapping Integrated Gradients attributions from the radiomic feature space back into the input pixel domain. However, a useful explanation in medicine should satisfy more than one requirement. It should be faithful to the model, stable under small perturbations, and sufficiently focused to support interpretation without becoming misleadingly sparse. These requirements motivate the present study, which evaluates SRFAMap across three medical imaging datasets—tuberculosis chest X-rays, fatty liver ultrasound, and PneumoniaMNIST—with RFAM [4] retained as a baseline comparator. Three research questions guide the analysis: whether SRFAMap produces faithful explanations, whether those explanations remain stable under perturbation, and whether Regions Of Interest (ROI) segmentation improves their clinical plausibility. The main contribution is a compact radiomics-specific multi-metric evaluation of SRFAMap using faithfulness correlation, relative

Late-breaking work, Demos and Doctoral Consortium, colocated with the 4th World Conference on eXplainable Artificial Intelligence: July 01–03, 2026, Fortaleza, Brazil

0000-0002-6048-4639 (O. Davydko); 0000-0002-3293-5308 (V. Pavlov); 0000-0002-2718-5426 (L. Longo)

input stability, and sparseness. The remainder of the article is structured as follows. Section 2 discusses related work in explainable AI for medical imaging. Section 3 introduces the empirical design used to evaluate SRFAMap together with the selected evaluation metrics. Section 4 presents the experimental results and offers a critical discussion. Finally, Section 5 synthesises the contribution and outlines future avenues for improvement.

2. Related Work

Explainable AI for medical imaging combines attribution methods and evaluation frameworks. Widely used methods include Integrated Gradients [5], Grad-CAM [6], Layer-wise Relevance Propagation, DeepLIFT [7, 8], and SHAP [9]. In radiomics, one strand explains predictors in feature space through SHAP, permutation importance, or coefficient analysis [10], while another produces radiomic feature activation maps (RFAM) from local sliding-window responses in image space [11]. The former ranks features but does not localise evidence; the latter localises individual feature patterns but is weakly linked to model-level attribution. SRFAMap was proposed to bridge this gap by projecting feature-space attributions back to input image space [3].

Recent evaluation literature shows that explanation quality should not be judged by visual plausibility alone. Faithfulness is commonly measured with deletion, insertion, and related perturbation-based metrics [12], supplemented by infidelity measures [13] and causal tests such as Remove-and-Retrain [14]. Robustness and sensitivity are assessed through Relative Input Stability, parameter-randomisation tests, and the Random Logit Test [15, 16, 17]; localisation and compactness are assessed through overlap-, sparsity-, and entropy-based metrics [18, 19]. Empirical studies in medical imaging report strong variability across datasets, architectures, and preprocessing pipelines, including both clinically meaningful highlighting and frequent mislocalisation [20, 21]. Methodological work likewise shows that metric definitions can alter method rankings, motivating multi-metric benchmarking frameworks and toolkits [12, 22]. In synthesis, radiomics-oriented saliency methods exist, but their evaluation remains incomplete: SRFAMap and RFAM have mainly been tested through faithfulness, whereas current XAI practice supports joint assessment of faithfulness, robustness, localisation, and complexity. The contribution of the present study therefore lies in applying a focused radiomics-specific audit to SRFAMap and RFAM under matched experimental conditions.

3. Design and Methodology

The study was guided by a single comparative hypothesis: if SRFAMap and RFAM are applied to the same trained radiomic classifiers and evaluated on the same test images, then their metric distributions will differ statistically significantly, and SRFAMap will perform at least marginally better overall, namely with higher faithfulness correlation, lower relative input stability, and a more favourable sparseness profile in interpretation. This formulation supports a direct paired comparison between the two explainers while preserving the multi-metric character of the evaluation.

Three datasets were used: 800 tuberculosis chest X-rays from the Shenzhen and Montgomery collections [23], 550 fatty-liver ultrasound images [24], and 5,856 PneumoniaMNIST chest X-rays [25]. ROI masking was applied to the tuberculosis and liver datasets. Lung masks were produced with Attention U-Net for chest X-rays [26], while a fixed handcrafted mask was used for the ultrasound images. Each image was represented by 223 second-order radiomic features derived from standard texture-statistics matrices: 120 GLCM features, 16 GLRLM features, 16 GLSZM features, 56 GLDM features, and 15 NGTDM features [27, 28, 29, 30, 31]. These features were used to train a multilayer perceptron with four hidden layers (hidden size is 128) each followed by dropout layers ($p = 0.2$), and ReLU activations. Adam algorithm is used for optimisation with learning rate $1e-4$, and early stopping (patience - 40 epochs). An MLP was chosen because the object of explanation in this study is a radiomic-feature classifier; replacing the radiomic representation with a CNN would define a different model family and is left for future work. Classification quality was summarised by test

accuracy and the Matthews Correlation Coefficient (MCC) [32]. Figure 1 summarises the full study design. The pipeline begins with preprocessing and optional ROI masking, continues with extraction of second-order radiomic descriptors, and ends with explanation generation and multi-metric comparison. This compact workflow is important because the same predictive models are interrogated under both masked and unmasked conditions, allowing the observed changes in saliency to be attributed to preprocessing rather than to changes in model family.

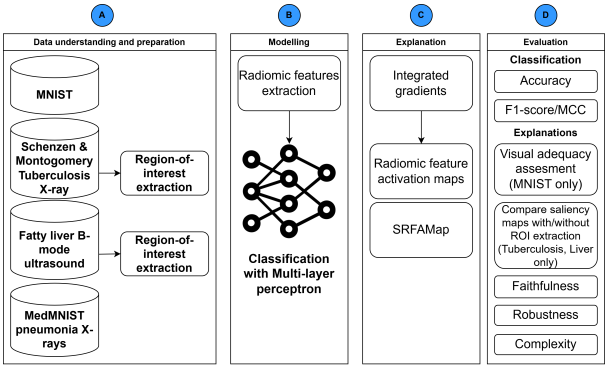


Figure 1: Overall research structure, from data preparation through radiomic feature extraction, model training, saliency generation, and evaluation.

SRFAMap was generated by mapping Integrated Gradients attributions from texture-matrix space back to image space. For a texture-statistics element x_i, the attribution was computed as:

$$\mathrm{IG}_i(x) = (x_i - x_i') \int_{\alpha=0}^{1} \frac{\partial F(R(x' + \alpha(x - x')))}{\partial x_i} \, d\alpha \tag{1}$$

where x' is a zero baseline, F is the trained classifier, and R is the radiomic feature function. Both explainers operated on the same trained classifiers and the same held-out test images; the classifier input remained the full 223-dimensional radiomic vector throughout. RFAM was used as the baseline explainer and computed with a 3×3 sliding window over the three most attributed radiomic features [4]. This choice follows the original per-feature RFAM formulation while keeping the comparison tractable: SRFAMap aggregates attribution information into a single map, whereas RFAM produces one map per selected feature, reported here as AMM1–AMM3. To keep the short-paper study compact while still covering complementary criteria, one representative metric was selected from each evaluation family: faithfulness correlation, relative input stability, and sparseness. Both methods were evaluated over 20 Monte Carlo train/validation/test splits, and Wilcoxon signed-rank tests were used to determine whether observed metric differences were statistically significant.

Figure 2 emphasises the key methodological distinction between the two explanation families. RFAM computes local feature responses directly in image space, whereas SRFAMap first attributes model importance in feature space and then projects that information back to the image domain. This difference motivates the expectation that SRFAMap should be less sensitive to local pixel perturbations and better aligned with the actual decision process of the trained classifier.

The evaluation strategy combined quantitative benchmarking with qualitative inspection. For each Monte Carlo split, both explainers were generated on the same held-out test images so that metric differences reflected the explanation method rather than changes in classifier training. Faithfulness correlation, relative input stability, and sparseness were analysed jointly because they capture complementary properties: alignment with model behaviour, robustness to perturbation, and spatial concentration. Quantitative scores were then interpreted alongside visual inspection of representative saliency maps, especially in the masked versus unmasked settings, to determine whether improved metric values also corresponded to more clinically plausible localisation. The source code is published on GitHub.[1]

[1]https://github.com/oleksandr-davydko/srfamap

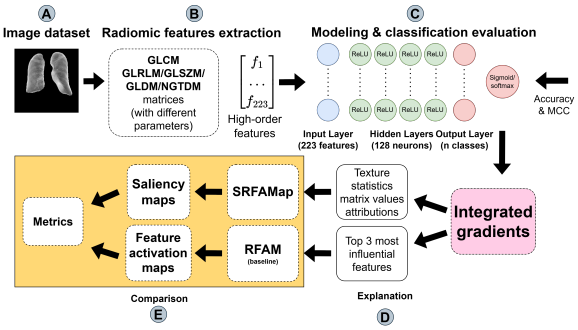


Figure 2: End-to-end experiment flow used to compare SRFAMap and RFAM on the same radiomic classifiers.

4. Results and Discussion

Tables 1 and 2 separate predictive performance from explainability findings. The radiomic MLP achieved high accuracy on PneumoniaMNIST and liver ultrasound, and lower but still usable performance on tuberculosis. From the perspective of SRFAMap itself, the most consistent result is its robustness: across all datasets, relative input stability was three to four times lower than for the RFAM baseline, and the difference was statistically significant in every comparison. Faithfulness correlation also generally favoured SRFAMap, with strong advantages on PneumoniaMNIST and liver data, weaker positive evidence on masked tuberculosis, and near-equivalent behaviour on unmasked tuberculosis. The stability margin was not marginal: in the full metric distributions, SRFAMap typically fell in the range 6.6–10.6, whereas RFAM variants were usually between 22.6 and 44.3. This scale difference matters in practice because a clinically useful explanation should not change substantially under minor acquisition noise or small perturbations. At the same time, stability alone is not treated here as proof of interpretability or clinical utility. It is interpreted jointly with faithfulness and ROI-based qualitative plausibility, which reduces but does not eliminate the possibility that part of the stability advantage reflects smoothing or aggregation effects.

Sparseness was more nuanced. RFAM often produced more concentrated maps, but higher compactness was not automatically better. Diffuse pathologies such as fatty liver can reasonably produce broader attribution patterns, whereas focal abnormalities may benefit from tighter localisation. For that reason, sparseness was interpreted jointly with faithfulness and qualitative plausibility rather than treated as an independently decisive criterion.

Wilcoxon signed-rank testing was used because the metric distributions were non-normal. The significance pattern was clear. For relative input stability, every SRFAMap versus RFAM comparison was significant in every dataset, with extremely small p-values throughout, for example 1.617×10^{-236} on liver, 7.665×10^{-225} on liver masked, and effectively zero on PneumoniaMNIST. For faithfulness, the difference was significant across liver and PneumoniaMNIST comparisons, partly significant for masked tuberculosis ($p = 2.856 \times 10^{-2}$ and 4.770×10^{-3} versus two RFAM variants), and not significant for unmasked tuberculosis. Sparseness differences were also significant in nearly all comparisons, with the main exception of one PneumoniaMNIST pairing where SRFAMap and RFAM were statistically indistinguishable ($p = 7.903 \times 10^{-1}$).

Negative faithfulness values in the liver and tuberculosis settings should not automatically be interpreted as SRFAMap failure. They may also reflect suppressor effects, non-linear interactions, or insertion/deletion protocols that generate out-of-distribution samples [12]. More broadly, the contrast between liver and tuberculosis shows that SRFAMap can either support confidence in plausible evidence or expose shortcut learning.

Remove-and-Retrain (ROAR) is therefore an important next step. If removing the most important regions causes a substantial performance drop after retraining, that would support the interpretation that SRFAMap captures genuinely influential evidence despite the negative perturbation-based correlation; a weak drop would instead suggest metric artefacts or suppressor-driven behaviour.

The metric distributions are shown in Figure 3. As in prior XAI evaluations, the distributions were non-normal, making Wilcoxon testing preferable to comparisons of means alone [12].

Table 1
Classification performance of the radiomic MLP over 20 Monte Carlo splits, reported as mean ± standard deviation.

Dataset	Test accuracy	Test MCC
PneumoniaMNIST	0.931 ± 0.010	0.826 ± 0.026
Liver	0.967 ± 0.027	0.926 ± 0.061
Liver masked	0.961 ± 0.024	0.916 ± 0.050
Tuberculosis	0.769 ± 0.045	0.545 ± 0.087
Tuberculosis masked	0.751 ± 0.044	0.509 ± 0.088

Table 2
Explainability summary for SRFAMap (SRFAMap), with RFAM used as baseline.

Dataset	Faithfulness	Stability	Main qualitative finding
PneumoniaMNIST	SRFAMap > RFAM	SRFAMap ≫ RFAM	Consistent metric advantage
Liver	SRFAMap > RFAM	SRFAMap ≫ RFAM	ROI suppresses scanner artefacts
Liver masked	SRFAMap > RFAM	SRFAMap ≫ RFAM	Attention shifts to liver parenchyma
Tuberculosis	SRFAMap ≈ RFAM	SRFAMap ≫ RFAM	Maps remain clinically implausible
Tuberculosis masked	SRFAMap mildly better	SRFAMap ≫ RFAM	Masking removes background, not bias

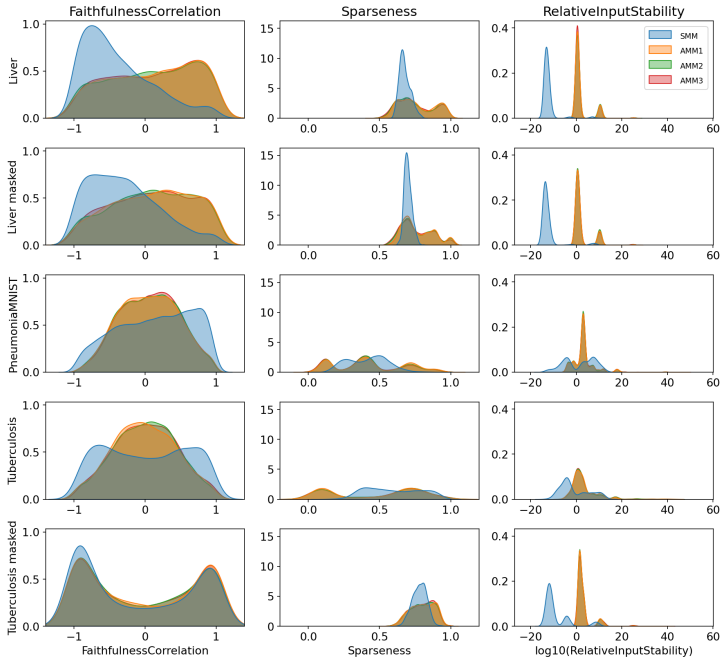
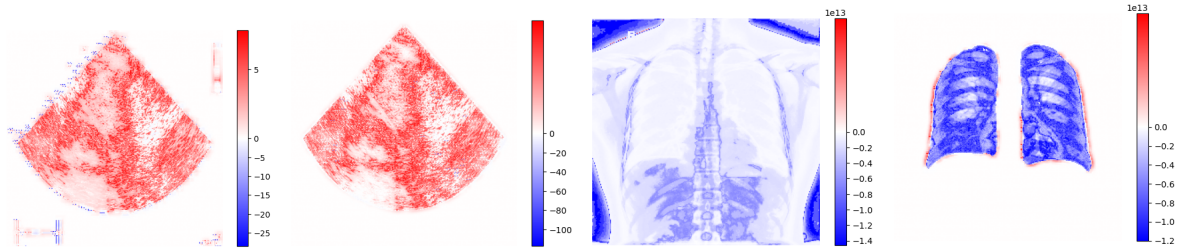


Figure 3: Kernel density estimates of faithfulness correlation, relative input stability, and sparseness for SRFAMap (SMM) and the three top-ranked RFAM activation maps (AMM1–AMM3) across all datasets. Higher faithfulness and sparseness indicate stronger alignment and concentration, while lower stability more robust explanations.

The qualitative ROI analysis in Figure 4 clarifies these metrics. For liver ultrasound, masking redirected attention from scanner markings and peripheral artefacts to the liver parenchyma, suggesting that SRFAMap can localise clinically plausible texture once distracting content is removed. For tuberculosis, masking removed background content but did not recover pathology-centred explanations, indicating that SRFAMap may faithfully reflect clinically flawed model behaviour.

This divergence is consistent with the underlying pathology. Fatty liver disease produces comparatively diffuse textural change, making global radiomic descriptors and ROI restriction a natural combination. Tuberculosis lesions are more heterogeneous and spatially irregular, so global second-order descriptors may remain predictive while still failing to capture lesion-centred evidence. SRFAMap is therefore informative in both settings, but for different reasons: in liver ultrasound it highlights plausible evidence, while in tuberculosis it exposes a mismatch between predictive success and clinical plausibility. Overall, SRFAMap appears robust and practically useful for radiomic models, with stability as its clearest strength across datasets and preprocessing conditions. At the same time, the tuberculosis case shows that explanation quality must still be interpreted with domain knowledge: a stable and faithful map can reveal a clinically flawed model. Further investigation with ROAR is needed before drawing a final conclusion.



(a) Liver, no ROI (b) Liver, ROI (c) Tuberculosis, no ROI (d) Tuberculosis, ROI

Figure 4: Qualitative effect of ROI segmentation on SRFAMap. For liver ultrasound, ROI masking suppresses spurious peripheral cues and shifts attribution to the liver tissue. For tuberculosis chest X-rays, masking removes background content but does not recover pathology-centred explanations.

5. Conclusions and Future Work

This research shows that SRFAMap is best understood through a multi-metric evaluation lens. Across three datasets, the method remained consistently stable and usually faithful, with Wilcoxon signed-rank tests supporting that these differences were systematic rather than incidental. The comparison with RFAM confirms such an advantage, but the broader contribution is conceptual: SRFAMap is most valuable when its outputs are interpreted as evidence about model behaviour rather than as automatic proof of clinical correctness. Findings also offer an important remark for the wider explainability literature. High-quality explanation scores do not guarantee that a model has learned clinically valid evidence. In the liver setting, SRFAMap supported confidence by shifting attention toward plausible tissue regions after ROI masking. In the tuberculosis setting, it played an equally important diagnostic role by exposing shortcut learning and spurious reliance. This dual role makes SRFAMap useful not only for validating models, but also for challenging them. Accordingly, the hypothesis was partially accepted. The expected statistically significant separation was strongly supported for relative input stability and largely supported for sparseness, but not uniformly for faithfulness across all dataset conditions, especially in unmasked tuberculosis where the two methods were not distinguishable by Wilcoxon testing. The main practical implication is that radiomic explainability should be assessed with complementary quantitative metrics together with qualitative inspection of anatomical plausibility. Faithfulness alone would have painted an incomplete picture, whereas stability and ROI-based analysis made the strengths and limits of SRFAMap much clearer. Future work will prioritise Remove-and-Retrain because it provides an especially direct test of whether the regions highlighted by SRFAMap are causally important for prediction, and will combine it with lesion-overlap metrics, cross-centre validation, expert review, feature-family ablations, and more spatially aware radiomic or CNN-radiomics hybrid representations to reduce shortcut-learning behaviour in chest X-rays.

Acknowledgments

We thank IBM for granting access to IBM Power systems on which the models were trained.

Disclosure of Interests

The authors have no competing interests to declare relevant to this article's content.

Declaration on Generative AI

During the preparation of this work, the author(s) used GitHub Copilot and GPT-5.4 model to assist with language refinement. After using this tool, the author(s) reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] J. E. Van Timmeren, D. Cester, S. Tanadini-Lang, H. Alkadhi, B. Baessler, Radiomics in medical imaging—"how-to" guide and critical reflection, Insights into Imaging 11 (2020) 91.

[2] D. Wallis, I. Buvat, Clever hans effect found in a widely used brain tumour mri dataset, Medical Image Analysis 77 (2022) 102368.

[3] 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, in: Explainable Artificial Intelligence, Springer Nature Switzerland, 2024, pp. 3–23.

[4] 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.

[5] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: Proceedings of the 34th International Conference on Machine Learning (ICML), volume 70 of *Proceedings of Machine Learning Research*, PMLR, 2017, pp. 3319–3328.

[6] 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.

[7] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, K.-R. Muller, Layer-wise relevance propagation: An overview, in: W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, K.-R. Muller (Eds.), Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, volume 11700, Springer International Publishing, Cham, 2019, pp. 193–209.

[8] A. Shrikumar, P. Greenside, A. Kundaje, Learning important features through propagating activation differences, in: Proceedings of the 34th International Conference on Machine Learning (ICML), volume 70 of *Proceedings of Machine Learning Research*, PMLR, 2017, pp. 3145–3153.

[9] 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.

[10] E. Lavrova, E. Lommers, H. Woodruff, A. Chatterjee, P. Maquet, E. Salmon, P. Lambin, C. Phillips, Exploratory radiomic analysis of conventional vs. quantitative brain mri: Toward automatic diagnosis of early multiple sclerosis, Frontiers in Neuroscience 15 (2021) 679941.

[11] 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.

[12] T. Gomez, T. Freour, H. Mouchere, Metrics for saliency map evaluation of deep learning explanation methods, in: M. El Yacoubi, E. Granger, P. C. Yuen, U. Pal, N. Vincent (Eds.), Pattern Recognition and Artificial Intelligence, volume 13363, Springer International Publishing, Cham, 2022, pp. 84–95.

[13] C.-K. Yeh, C.-Y. Hsieh, A. Suggala, D. I. Inouye, P. K. Ravikumar, On the (in)fidelity and sensitivity of explanations, in: Advances in Neural Information Processing Systems 32 (NeurIPS), 2019.
[14] S. Hooker, D. Erhan, P.-J. Kindermans, B. Kim, A benchmark for interpretability methods in deep neural networks, in: Advances in Neural Information Processing Systems 32 (NeurIPS), 2019.
[15] C. Agarwal, N. Johnson, M. Pawelczyk, S. Krishna, E. Saxena, M. Zitnik, H. Lakkaraju, Rethinking stability for attribution-based explanations, 2022. ArXiv:2203.06877.
[16] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, B. Kim, Sanity checks for saliency maps, in: Advances in Neural Information Processing Systems 31 (NeurIPS), 2018.
[17] L. Sixt, M. Granz, T. Landgraf, When explanations lie: Why many modified bp attributions fail, in: Proceedings of the 37th International Conference on Machine Learning, ICML '20, JMLR.org, 2020.
[18] P. Chalasani, J. Chen, A. R. Chowdhury, S. Jha, X. Wu, Concise explanations of neural networks using adversarial training, in: Proceedings of the 37th International Conference on Machine Learning, JMLR.org, 2020.
[19] U. Bhatt, A. Weller, J. M. F. Moura, Evaluating and aggregating feature-based model explanations, in: International Joint Conference on Artificial Intelligence, 2020.
[20] F. Fumero, J. Sigut, J. Estevez, T. Diaz-Aleman, Systematic application of saliency maps to explain the decisions of convolutional neural networks for glaucoma diagnosis based on disc and cup geometry, Biomedical Physics & Engineering Express 11 (2025) 025008.
[21] M. Byra, K. Dobruch-Sobczak, H. Piotrzkowska-Wroblewska, Z. Klimonda, J. Litniewski, Explaining a deep learning based breast ultrasound image classifier with saliency maps, Journal of Ultrasonography 22 (2022) 70–75.
[22] A. Hedstrom, L. Weber, D. Krakowczyk, D. Bareeva, F. Motzkus, W. Samek, S. Lapuschkin, M. M. M.-C. Hohne, Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond, Journal of Machine Learning Research 24 (2023) 1–11.
[23] S. Jaeger, S. Candemir, S. Antani, Y.-X. J. Wang, P.-X. Lu, G. Thoma, Two public chest x-ray datasets for computer-aided screening of pulmonary diseases, Quantitative Imaging in Medicine and Surgery 4 (2014) 475–477.
[24] M. Byra, et al., Transfer learning with deep convolutional neural network for liver steatosis assessment in ultrasound images, International Journal of Computer Assisted Radiology and Surgery 13 (2018) 1895–1903.
[25] J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke, H. Pfister, B. Ni, Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis, in: IEEE 18th International Symposium on Biomedical Imaging (ISBI), 2021, pp. 191–195.
[26] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, D. Rueckert, Attention u-net: Learning where to look for the pancreas, 2018. ArXiv:1804.03999.
[27] R. M. Haralick, K. Shanmugam, I. Dinstein, Textural features for image classification, IEEE Transactions on Systems, Man, and Cybernetics SMC-3 (1973) 610–621.
[28] M. M. Galloway, Texture analysis using gray level run lengths, Computer Graphics and Image Processing 4 (1975) 172–179.
[29] G. Thibault, B. Fertil, C. Navarro, S. Pereira, N. Levy, J. Sequeira, J.-L. Mari, Texture indexes and gray level size zone matrix application to cell nuclei classification, in: International Conference on Pattern Recognition and Information Processing, 2009.
[30] C. Sun, W. G. Wee, Neighboring gray level dependence matrix for texture classification, Computer Vision, Graphics, and Image Processing 23 (1983) 341–352.
[31] M. Amadasun, R. King, Textural features corresponding to textural properties, IEEE Transactions on Systems, Man, and Cybernetics 19 (1989) 1264–1274.
[32] D. Chicco, G. Jurman, The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation, BMC Genomics 21 (2020) 6.