

Beyond Gate Weights: Decision-Level Interpretation of Radiomics–Image Feature Fusion for RCC Classification

Yuan Liang^{1,3}, Sourav Bhattacharjee², and Abraham Campbell³

¹ Research Ireland Centre for Research Training in Machine Learning

² School of Veterinary Medicine, University College Dublin, Dublin, Ireland

³ School of Computer Science, University College Dublin, Dublin, Ireland
yuan.liang@ucdconnect.ie

Abstract. Gate coefficients describe how a fusion layer weights intermediate features, but they do not necessarily reflect the magnitude of branch contributions at the classifier output. We investigate this distinction in a fixed 3D MedVAE–radiomics model for clear cell versus non-clear cell renal cell carcinoma classification from contrast-enhanced CT. Because the model uses a linear classification head, the binary decision logit can be decomposed exactly into a gated image term, a direct radiomics term, and a bias. We define effective radiomics dominance as the direct radiomics term’s share of the summed absolute non-bias branch contributions and compare it with structural radiomics weighting. Internally, mean structural radiomics weighting was 0.823, whereas mean effective radiomics dominance was 0.148, with a modest case-wise association ($\rho = 0.290$). The gated image term retained substantially stronger discriminative ranking information than the direct radiomics term. The same structural-to-effective discrepancy was observed in exploratory external evaluation. Branch-targeted permutation, integrated-gradient, and Shapley-style analyses further identified radiomics descriptors influencing the direct radiomics term. Because radiomics also generates the gate, the gated image term remains radiomics-dependent, so additive branch contribution should not be interpreted as total input influence. These findings show that structural gate weighting and decision-level branch contribution represent distinct aspects of gated feature fusion and should be interpreted separately.

Keywords: Feature fusion · Gated fusion · Interpretability · Radiomics · Medical imaging

1 Introduction

Renal cell carcinoma (RCC) is a major urological malignancy for which preoperative subtype characterisation is clinically relevant. Clear cell RCC (ccRCC) and non-clear cell RCC differ in prognosis and management, while their appearances can overlap on contrast-enhanced computed tomography (CT) [3]. Radiomics

describes tumour shape, intensity distribution, and texture through predefined measurements [1]. CT radiomics has been investigated for distinguishing ccRCC from non-clear cell RCC [19]. Learned image representations offer another description of tumour appearance, including transferable latent features from medical image encoders such as MedVAE [18].

Gated fusion adaptively weights intermediate feature branches [2]. Gate coefficients have also been used to interpret modality emphasis in clinical prediction. For example, Sun et al. analysed softmax self-gating weights across electronic-health-record time windows and complemented this analysis with integrated-gradient attribution [14]. Such approaches motivate a distinction between weighting at the fusion layer and contribution to the final decision. The latter additionally depends on feature magnitudes, their directions, and downstream classifier weights. When radiomics generates the gate, it influences both its own branch and the weighting of the image branch, so direct additive contribution and total input dependence are distinct quantities.

The interpretation of attention coefficients has been debated in relation to other measures of model behaviour [9,21]. Kobayashi et al. analysed transformed-vector norms alongside attention weights, obtaining a different account of information aggregation in Transformers [10]. Although attention and the gated classifier studied here have different architectures, this work motivates examining weighted features and their downstream effects. In medical imaging, specifying the model quantity being interpreted helps establish the relevance of an analysis to clinical questions [4].

We study a fixed 3D MedVAE–radiomics gated classifier through three complementary analyses: decomposition of its decision logit, comparison of structural and effective branch weighting, and feature attribution targeted to its direct radiomics term. Internal and exploratory external evaluation assess discrimination, directional agreement, and radiomics feature sensitivity. Together, these analyses connect gate behaviour to signed classifier-level contributions and provide a focused investigation of radiomics–image fusion in RCC.

2 Methods

2.1 Cohorts and Image Preprocessing

The internal cohort comprised 396 patients from KiTS23, a collection of contrast-enhanced abdominal CT scans with kidney and renal tumour annotations [8]. Multifocal cases were excluded to avoid ambiguity between patient histology and the analysed lesion. The binary endpoint was ccRCC ($Y = 1$) versus non-clear cell RCC ($Y = 0$). A fixed stratified patient-level split assigned 237 patients to training (168 ccRCC, 69 non-clear cell RCC), 99 to validation (70 and 29), and 60 to internal testing (42 and 18). Classifier parameters were fitted on training data, and validation data were used for checkpoint and classification-threshold selection. The held-out test partition was used only for final evaluation and was not used for model fitting.

CT images and masks were reoriented to RAS and resampled to 1.0-mm isotropic spacing, using linear interpolation for images and nearest-neighbour interpolation for masks. Intensities were clipped to $[-150, 200]$ HU. The learned image branch used tumour-centred crops with a 12-voxel contextual margin. Nonzero image voxels were z-score normalised per volume, followed by padding or centre cropping to $96 \times 96 \times 96$ voxels. Radiomics extraction used HU-scale images without this image z-score normalisation. For radiomics, multiple annotations of the same lesion were combined by STAPLE [20].

The external TCGA/AIMI cohort comprised 41 cases, including 39 ccRCC and two non-clear cell RCC cases. Cases were drawn from the TCGA-KIRC, TCGA-KIRP, and TCGA-KICH collections, using AIMI DICOM-SEG annotations [11,17]. Cohort construction retained expert-corrected masks or masks with the highest available expert quality score. External cases were not used to fit the classifier, select radiomics features, estimate training normalisation statistics, or select the checkpoint or classification threshold. The external cohort was used for exploratory evaluation of the fixed model.

2.2 Radiomics Representation

PyRadiomics was used to extract shape, first-order, and texture features [16], with reference to IBSI feature definitions [23]. Shape features were computed on the original tumour region. First-order and texture features were computed on original and Laplacian-of-Gaussian (LoG) images with $\sigma \in \{1, 2, 3\}$ mm, using a fixed discretisation bin width of 25 HU. The initial set contained 386 features.

The inherited pipeline first required inter-annotator ICC(3,1), a two-way mixed-effects consistency measure [13], of at least 0.75. Features below the 20th percentile of median absolute deviation (MAD) were then removed, followed by greedy Pearson-correlation filtering at $|r| > 0.95$. Average-linkage hierarchical clustering used distance $1 - |r|$ and a cut threshold of 0.2, retaining the feature with the largest MAD from each cluster. The classifier analysed here used the resulting fixed 73-feature representation. Feature-wise z-score statistics were estimated from the 237 training patients and applied unchanged to the validation, internal test, and external cohorts.

2.3 Gated Classifier and Model Development

The image branch used the volumetric MedVAE encoder [18]. Its latent tensor was flattened and linearly projected to $\mathbf{f}_{\text{img}} \in \mathbb{R}^{512}$. The 73-feature radiomics vector was independently projected to $\mathbf{f}_{\text{rad}} \in \mathbb{R}^{512}$. A radiomics-conditioned multilayer perceptron, with dimensions $512 \rightarrow 256 \rightarrow 512$, a hidden ReLU, and an output sigmoid, produced the feature-wise image gate:

$$\mathbf{g} = \sigma(W_2 \text{ReLU}(W_1 \mathbf{f}_{\text{rad}} + \mathbf{a}_1) + \mathbf{a}_2). \quad (1)$$

The fused representation and two-class linear head were

$$\mathbf{h} = \mathbf{g} \odot \mathbf{f}_{\text{img}} + (1 - \mathbf{g}) \odot \mathbf{f}_{\text{rad}}, \quad (z_0, z_1)^\top = W_{\text{cls}} \mathbf{h} + \mathbf{b}_{\text{cls}}, \quad (2)$$

where $\odot$ denotes element-wise multiplication. A projected dimension is a learned mixture of input features, not an individually identifiable radiomics descriptor.

Model development used partial fine-tuning, class-weighted cross-entropy, Adam, and a batch size of 2. Training augmentation comprised independent flips along two spatial axes (probability 0.5 each), random 90-degree rotations (0.2), and affine rotations/scaling (0.2; rotation range 0.1 radians and scale range 0.1 per axis). Validation and evaluation preprocessing were deterministic. The learning rate was halved after five epochs without improved validation AUC; early stopping used patience 15, and the checkpoint with the highest validation AUC was retained. All contribution analyses used this fixed checkpoint and its unscaled scores.

2.4 Decision-Level Branch Decomposition

Because the fused representation is followed by a linear classification head, the binary decision logit can be decomposed exactly into additive terms associated with the two gated branches. Let $\mathbf{w} = \mathbf{W}_{\text{cls},1} - \mathbf{W}_{\text{cls},0}$ and $b = b_{\text{cls},1} - b_{\text{cls},0}$. The ccRCC versus non-clear cell RCC logit difference is

$$z = z_1 - z_0 = \mathbf{w}^\top \mathbf{h} + b, \quad \hat{p} = \sigma(z). \quad (3)$$

Substituting the gated representation from Eq. (2) gives

$$C_{\text{img}} = \mathbf{w}^\top (\mathbf{g} \odot \mathbf{f}_{\text{img}}), \quad C_{\text{rad}} = \mathbf{w}^\top ((1 - \mathbf{g}) \odot \mathbf{f}_{\text{rad}}), \quad (4)$$

such that

$$z = C_{\text{img}} + C_{\text{rad}} + b. \quad (5)$$

We refer to C_{img} as the gated image term and C_{rad} as the direct radiomics term. Both are conditional on the radiomics-generated gate, so this is a decomposition into computational branch terms rather than an allocation of total input influence. For the saved-output analysis, C_{rad} was reconstructed from the fixed model weights and standardised radiomics inputs, and C_{img} was recovered as $z - C_{\text{rad}} - b$ from the saved class logits.

For each case, structural radiomics weighting was defined as

$$S_{\text{rad}} = 1 - \frac{1}{d} \sum_{j=1}^d g_j, \quad d = 512. \quad (6)$$

This quantity describes the average structural weighting assigned to radiomics at the fusion layer. Effective radiomics dominance was defined as

$$D_{\text{rad}} = \frac{|C_{\text{rad}}|}{|C_{\text{img}}| + |C_{\text{rad}}|}. \quad (7)$$

D_{rad} measures the relative magnitude of the direct radiomics term among the two non-bias additive branch terms. Values closer to zero indicate dominance of the gated image term, while values closer to one indicate dominance of the direct radiomics term. Because absolute values are used, contribution direction was analysed separately as described below.

2.5 Discrimination and Directional Behaviour

We evaluated AUC and average precision (AP) for z , C_{img} , and C_{rad} , with larger values oriented toward ccRCC and no post-hoc score reversal. Spearman correlation quantified the association between S_{rad} and D_{rad} .

Directional agreement was defined by $C_{\text{img}}C_{\text{rad}} > 0$, conflict by $C_{\text{img}}C_{\text{rad}} < 0$, and a neutral case by a zero product. These signs describe additions to the ccRCC logit, not separate branch classifications. For error-rate comparisons, the full prediction was classified using a threshold selected by maximising Youden’s index on the validation set [22] and then held fixed across evaluation cohorts. We summarised conflict frequency and compared full-model error rates and D_{rad} distributions between agreement and conflict groups.

2.6 Radiomics Feature Attribution

Feature attribution targeted the direct radiomics term $F(\mathbf{x}) = C_{\text{rad}}(\mathbf{x})$, where $\mathbf{x}$ is the standardised 73-feature radiomics vector. Repeated permutation importance was the primary analysis [7]. Each feature was independently shuffled across patients 200 times, separately within each evaluation cohort, while the radiomics projection and radiomics-conditioned gate were recomputed after each perturbation. Importance was quantified by the mean decrease in AUC and AP of C_{rad} ; negative values denote improved discrimination after permutation.

Integrated gradients [15] and a Monte Carlo Shapley-style analysis [12,5] were complementary methods, computed separately within each cohort. Both used the zero standardised-feature vector, corresponding to training-set feature means, as the reference. Integrated gradients were approximated using 64 equally spaced right-endpoint evaluations along the straight-line path to each patient’s feature vector; absolute attributions were then averaged across patients. The Shapley-style analysis sampled 64 random orders of all 73 features. Starting from the reference, each feature was replaced by its patient-specific value, and its marginal change in C_{rad} was recorded. Absolute marginal changes were averaged over orders and patients. This magnitude summary differs from taking the absolute value of each patient’s order-averaged Shapley estimate.

2.7 Statistical Analysis

AUC/AP 95% confidence intervals (CIs) were the 2.5th and 97.5th percentiles from 1,000 class-stratified patient bootstrap replicates [6]. Paired differences used identical resampled patients for the compared scores. These intervals condition on the observed class counts and fixed model. Spearman CIs used 1,000 ordinary patient bootstrap replicates. Error-rate comparisons used Fisher’s exact test, and D_{rad} distributions were compared using the two-sided Mann–Whitney test. Comparisons were unadjusted for class composition or other covariates. All tests were exploratory and were not corrected for multiple comparisons.

For feature permutation, the 2.5th–97.5th percentile range of the 200 score decreases summarised variability across random shuffles of the fixed cohort.

These ranges quantify permutation variability, whereas patient bootstrap CIs quantify patient-sampling uncertainty conditional on the specified resampling design. Neither incorporates uncertainty from training or feature selection.

3 Results and Discussion

3.1 Discrimination of the Decomposed Scores

The gated image term retained discrimination close to the full classifier, whereas the radiomics branch term showed substantially weaker ranking performance (Table 1). Internally, paired bootstrap analysis gave a full-minus-image AUC difference of -0.017 (95% CI -0.042 to 0.000), compared with 0.308 (95% CI 0.165 – 0.468) for the full-minus-radiomics difference. These results indicate that substantially more of the observed discriminative ranking information was carried by the gated image term than by the radiomics branch term, without establishing superiority or equivalence between the gated image term and the complete classifier.

Table 1. Discrimination of the full score and the gated branch terms. Parentheses contain class-stratified patient bootstrap 95% CIs, conditional on the fixed model. ccRCC prevalence was 0.700 internally and 0.951 externally.

Cohort	Score	AUC (95% CI)	AP (95% CI)
Internal	Full score	0.827 (0.710–0.923)	0.922 (0.867–0.967)
	C_{img}	0.844 (0.738–0.938)	0.928 (0.877–0.972)
	C_{rad}	0.519 (0.352–0.689)	0.720 (0.629–0.833)
External	Full score	0.782 (0.641–0.897)	0.988 (0.979–0.995)
	C_{img}	0.756 (0.615–0.872)	0.987 (0.978–0.993)
	C_{rad}	0.462 (0.000–0.974)	0.926 (0.862–0.999)

The same qualitative pattern was observed externally, although the estimates were exploratory given the positive-enriched cohort (Table 1).

The discrimination of C_{img} should not be interpreted as the performance of an independently trained image-only model. The image representation was learned within the fusion model, and its weighting remains conditioned on radiomics through the gate. The branch-score analysis therefore characterises additive terms within the intact classifier rather than the effect of removing either feature branch.

3.2 Structural Weighting versus Effective Contribution

Mean structural radiomics weighting S_{rad} was 0.823 internally, whereas mean effective radiomics dominance D_{rad} was 0.148. The corresponding external means were 0.815 and 0.078 (Fig. 1A). Thus, strong structural weighting toward radiomics coexisted with a substantially smaller radiomics share of the absolute net additive branch terms.

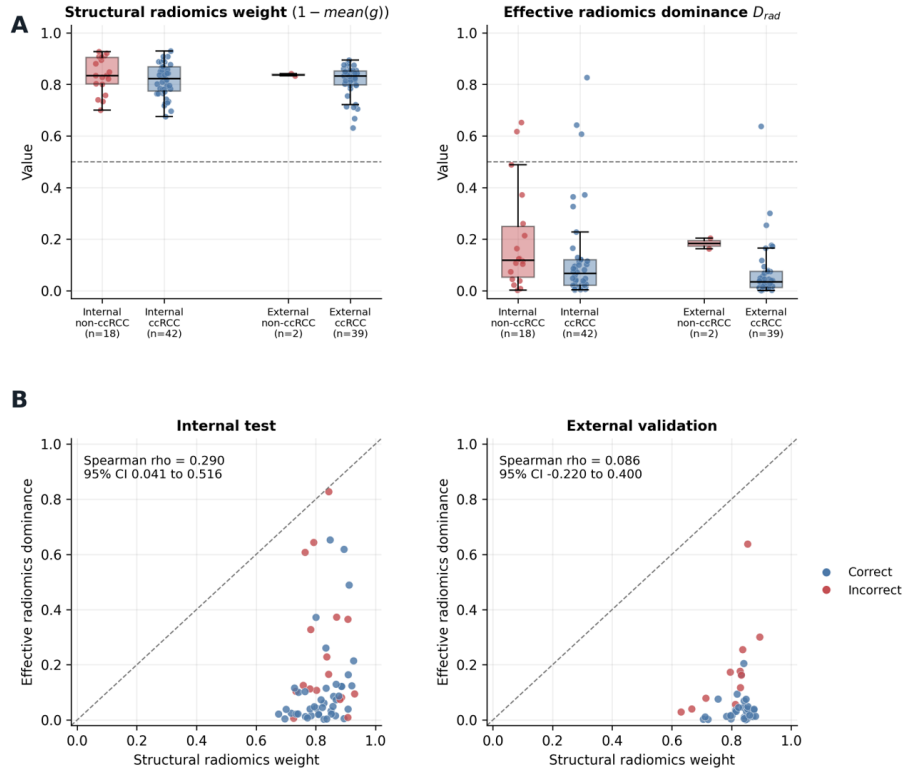


Fig. 1. Structural and effective radiomics weighting. (A) Distributions by cohort and true class. (B) Case-wise association, coloured by full-model correctness at the validation-selected threshold. Spearman 95% CIs use ordinary patient bootstrap resampling. The diagonal denotes numerical equality. Both branch terms retain radiomics-conditioned gating.

Structural radiomics weighting showed a modest rank association with effective radiomics dominance internally ($\rho = 0.290$, 95% CI 0.041–0.516). The external correlation estimate was smaller and remained uncertain ($\rho = 0.086$, 95% CI –0.220 to 0.400; Fig. 1B).

Structural weighting and effective contribution describe different stages of the model computation. Gate coefficients scale intermediate representations, whereas the branch terms additionally incorporate feature values, classifier weights, and cancellation across dimensions. Gate weighting alone is therefore insufficient to infer output-level branch contribution. At the same time, a small radiomics branch term does not imply that the classifier is independent of radiomics because radiomics also determines the gate applied to the image branch.

The distinction observed here is consistent with previous work showing that learned weighting coefficients alone may incompletely characterise downstream model outputs. Kobayashi et al. demonstrated that incorporating the norms

of transformed representations can change the interpretation obtained from attention weights alone [10]. Although attention and gated feature fusion are different mechanisms, both motivate examining weighted representations rather than coefficients in isolation. Here, the analysis additionally incorporates the final classifier weights to quantify signed contributions to the binary decision logit. These branch terms depend on projected feature values, gate coefficients, classifier weights, and cancellation across dimensions. Moreover, radiomics influences the gated image term through $\mathbf{g}$, so a small direct radiomics contribution does not imply limited total radiomics influence. Gate statistics and classifier-level branch terms therefore provide complementary, rather than interchangeable, descriptions of fusion behaviour.

3.3 Directional Agreement and Conflict

Conflict occurred in 28.3% of internal cases and 34.1% of external cases (Table 2). Internally, conflict cases showed a numerically higher error rate than agreement cases, but neither error rates nor the distribution of D_{rad} differed significantly between groups (Fisher’s exact $p = 0.530$; Mann–Whitney $p = 0.682$). External comparisons likewise showed no clear differences (Fisher’s exact $p = 1.000$; Mann–Whitney $p = 0.967$).

These exploratory comparisons show that the two signed branch terms can reinforce or oppose one another within individual predictions. However, the results provide limited evidence that contribution conflict identifies difficult cases or that radiomics plays a corrective role. Establishing such a role would require a direct intervention on the radiomics-mediated pathway rather than comparison of signed branch terms alone.

Table 2. Directional agreement and conflict. Errors refer to the full classifier at the fixed validation-selected threshold. Mean and median effective radiomics dominance are reported for each group.

Cohort	Relation	n	Errors/ n (%)	Mean D_{rad}	Median D_{rad}
Internal	Agreement	43	11/43 (25.6)	0.128	0.086
	Conflict	17	6/17 (35.3)	0.198	0.043
External	Agreement	27	7/27 (25.9)	0.060	0.035
	Conflict	14	4/14 (28.6)	0.112	0.039

3.4 Radiomics Feature Attribution

The largest internal mean permutation-AUC decrease for C_{rad} was associated with LoG ($\sigma = 3$ mm) GLSZM GrayLevelNonUniformityNormalized (ΔAUC 0.043, ΔAP 0.041). The same feature ranked first under both integrated-gradient and Shapley-style attribution, showing consistent prominence across the three complementary attribution analyses (Fig. 2).

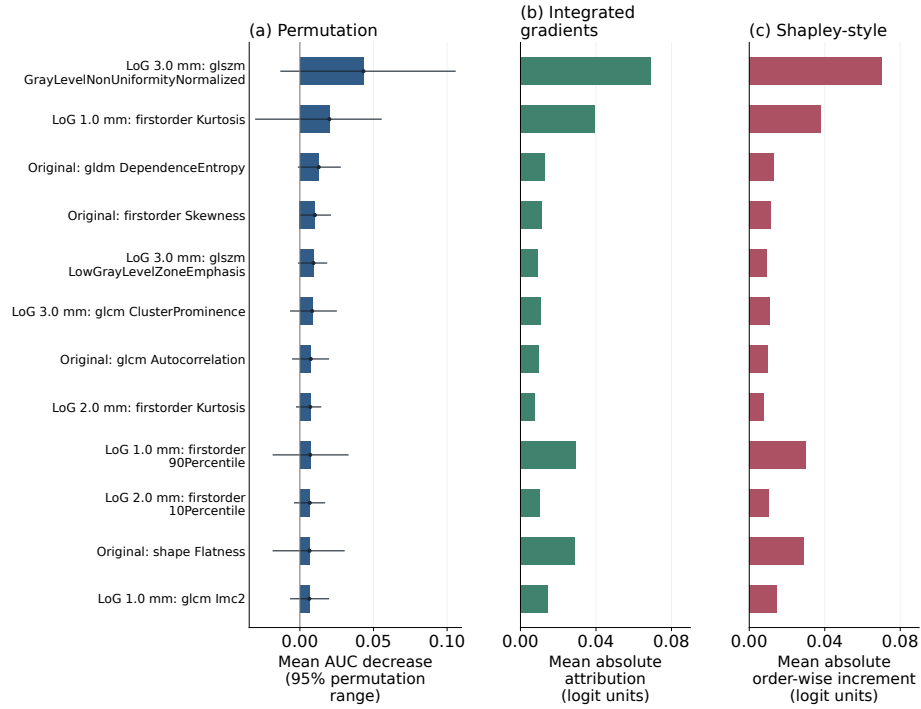


Fig. 2. Internal feature attribution to C_{rad} . The 12 leading features by mean permutation-AUC decrease appear in a common row order. (a) Mean AUC decreases and central 95% ranges across 200 within-cohort permutations; ranges describe shuffle variability rather than patient-sampling CIs. (b) Mean absolute integrated gradients. (c) Mean absolute order-wise marginal increments, averaged over 64 feature-addition orders and patients. Panels (b) and (c) are in logit units. LoG scales are in mm.

The external permutation ranking differed, with LoG ($\sigma = 1$ mm) first-order Skewness showing the largest mean AUC decrease. Given the small number of external non-clear cell RCC cases, this difference should be interpreted descriptively rather than as evidence of a reproducible shift in feature importance.

The three attribution methods characterise different properties of C_{rad} . Permutation importance measures changes in branch-score discrimination, whereas integrated gradients and the Shapley-style analysis quantify reference-relative changes in the branch logit. Since perturbing radiomics also changes the gate applied to the image representation, these rankings should be interpreted as sensitivity of the radiomics branch term rather than importance for the complete classifier.

3.5 Scope and Limitations

This study characterises one fixed classifier and one internal split. The decomposition is exact for its additive fusion and linear head, while its interpretation

is conditional on the chosen branch representation. The dominance ratio concerns net non-bias terms: cancellation within each branch occurs before taking absolute values, and direction is assessed separately. It is therefore a branch-level description rather than a causal allocation of information sources.

The small, positive-enriched external cohort limits generalisation. Resampling only two observed non-clear cell RCC cases cannot represent population heterogeneity, and the external association estimate remains uncertain. High external AP must also be interpreted relative to the high ccRCC prevalence. Differences in case mix and segmentation could affect transportability, but their individual effects were not isolated. Group comparisons were unadjusted, and attribution estimates depend on the chosen perturbations and reference. Unconditional shuffling and reference-based feature combinations may depart from the joint distribution of correlated radiomics measurements. These analyses support computational interpretability; biological interpretation and clinical utility would require additional validation.

4 Conclusion

This study shows that gate inspection and decision-level decomposition answer different interpretability questions in gated radiomics–image feature fusion. In the analysed RCC classifier, strong structural weighting toward radiomics coexisted with a substantially smaller radiomics branch contribution at the classifier output, while the gated image term retained most of the observed discriminative ranking information. Because the gated image term remains conditioned on radiomics, these findings should not be interpreted as evidence that the radiomics input is dispensable. Instead, they distinguish structural weighting, additive branch contribution, and total input dependence as related but non-equivalent descriptions of model behaviour. Reporting gate statistics together with classifier-weighted branch terms therefore provides a more precise account of gated-fusion behaviour. Future work should examine whether this structural-to-effective discrepancy persists across other fusion architectures and use intervention-based analyses to quantify total input dependence in larger and more balanced external cohorts.

References

1. Aerts, H.J., Velazquez, E.R., Leijenaar, R.T., Parmar, C., Grossmann, P., Carvalho, S., Bussink, J., Monshouwer, R., Haibe-Kains, B., Rietveld, D., et al.: Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature communications* **5**(1), 4006 (2014)
2. Arevalo, J., Solorio, T., Montes-y Gómez, M., González, F.A.: Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992* (2017)
3. Bex, A., Abu Ghanem, Y., Albiges, L., et al.: European association of urology guidelines on renal cell carcinoma: The 2025 update (2025). <https://doi.org/10.1016/j.eururo.2025.02.020>

4. Borys, K., Schmitt, Y.A., Nauta, M., Seifert, C., Krämer, N., Friedrich, C.M., Nensa, F.: Explainable ai in medical imaging: An overview for clinical practitioners–beyond saliency-based xai approaches. *European journal of radiology* **162**, 110786 (2023)
5. Castro, J., Gómez, D., Tejada, J.: Polynomial calculation of the Shapley value based on sampling. *Computers & Operations Research* **36**(5), 1726–1730 (2009). <https://doi.org/10.1016/j.cor.2008.04.004>
6. Efron, B.: Bootstrap methods: Another look at the jackknife. *The Annals of Statistics* **7**(1), 1–26 (1979). <https://doi.org/10.1214/aos/1176344552>
7. Fisher, A., Rudin, C., Dominici, F.: All models are wrong, but many are useful: Learning a variable’s importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research* **20**(177), 1–81 (2019), <https://jmlr.org/papers/v20/18-760.html>
8. Heller, N., Isensee, F., Tejpaul, R., Wood, A.W., Papanikolopoulos, N.P., Weight, C.: 2023 kidney and kidney tumor segmentation challenge (2023). <https://doi.org/10.5281/zenodo.7840134>
9. Jain, S., Wallace, B.C.: Attention is not explanation. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. pp. 3543–3556 (2019)
10. Kobayashi, G., Kuribayashi, T., Yokoi, S., Inui, K.: Attention is not only a weight: Analyzing transformers with vector norms. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. pp. 7057–7075. Association for Computational Linguistics (2020). <https://doi.org/10.18653/v1/2020.emnlp-main.574>
11. Murugesan, G.K., McCrumb, D., Aboian, M., Verma, T., Soni, R., Memon, F., Farahani, K., Pei, L., Wagner, U., Fedorov, A.Y., et al.: Ai-generated annotations dataset for diverse cancer radiology collections in nci image data commons. *Scientific data* **11**(1), 1165 (2024)
12. Shapley, L.S.: A value for n-person games. In: Kuhn, H.W., Tucker, A.W. (eds.) *Contributions to the Theory of Games II*, pp. 307–317. Princeton University Press, Princeton (1953). <https://doi.org/10.1515/9781400881970-018>
13. Shrout, P.E., Fleiss, J.L.: Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin* **86**(2), 420–428 (1979). <https://doi.org/10.1037/0033-2909.86.2.420>
14. Sun, M., Engelhard, M.M., Goldstein, B.: Borrowing from the future: Enhancing early risk assessment through contrastive learning. In: *Proceedings of the 10th Machine Learning for Healthcare Conference. Proceedings of Machine Learning Research*, vol. 298. PMLR (2025), <https://proceedings.mlr.press/v298/sun25a.html>
15. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: *International conference on machine learning*. pp. 3319–3328. PMLR (2017)
16. Van Griethuysen, J.J., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R.G., Fillion-Robin, J.C., Pieper, S., Aerts, H.J.: Computational radiomics system to decode the radiographic phenotype. *Cancer research* **77**(21), e104 (2017)
17. Van Oss, J., Murugesan, G.K., McCrumb, D., Soni, R.: Image segmentations produced by bamf under the aimi annotations initiative (2023). <https://doi.org/10.5281/zenodo.10081112>
18. Varma, M., Kumar, A., van der Sluijs, R., Ostmeier, S., Blankemeier, L., Chambon, P.J.M., Bluethgen, C., Prince, J., Langlotz, C., Chaudhari, A.S.: Medvae: Efficient

- automated interpretation of medical images with large-scale generalizable autoencoders. In: Tasdizen, T., Elhabian, S., Summers, R., Chen, C., Koch, L., Zhuang, Y. (eds.) Proceedings of The 8th International Conference on Medical Imaging with Deep Learning. Proceedings of Machine Learning Research, vol. 301, pp. 1597–1626. PMLR (09–11 Jul 2026), <https://proceedings.mlr.press/v301/varma26a.html>
19. Wang, P., Pei, X., Yin, X.P., Ren, J.L., Wang, Y., Ma, L.Y., Du, X.G., Gao, B.L.: Radiomics models based on enhanced computed tomography to distinguish clear cell from non-clear cell renal cell carcinomas. *Scientific Reports* **11**(1), 13729 (2021)
 20. Warfield, S.K., Zou, K.H., Wells, W.M.: Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. *IEEE transactions on medical imaging* **23**(7), 903–921 (2004)
 21. Wiegrefe, S., Pinter, Y.: Attention is not not explanation. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP). pp. 11–20 (2019)
 22. Youden, W.J.: Index for rating diagnostic tests. *Cancer* **3**(1), 32–35 (1950). [https://doi.org/10.1002/1097-0142\(1950\)3:1<32::AID-CNCR2820030106>3.0.CO;2-3](https://doi.org/10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3)
 23. Zwanenburg, A., Vallières, M., Abdalah, M.A., Aerts, H.J., Andrearczyk, V., Apte, A., Ashrafinia, S., Bakas, S., Beukinga, R.J., Boellaard, R., et al.: The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology* **295**(2), 328–338 (2020)