

Radiomics–Foundation Fusion for Interpretable RCC Classification: Internal Benchmarking and Exploratory External Transfer

Yuan Liang^{1,4}, Fangyijie Wang^{1,2}, Kathleen M. Curran^{1,2}, Guénolé Silvestre^{1,4},
Sourav Bhattacharjee³, and Abraham Campbell⁴

¹ Research Ireland Centre for Research Training in Machine Learning

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

³ School of Veterinary Medicine, University College Dublin, Dublin, Ireland

⁴ School of Computer Science, University College Dublin, Dublin, Ireland
yuan.liang@ucdconnect.ie

Abstract. Accurate preoperative subtype classification of renal cell carcinoma (RCC) from contrast-enhanced CT remains clinically challenging because clear cell RCC (ccRCC) and non-clear cell RCC often show overlapping imaging appearances. This study evaluates whether foundation representations reduce reliance on handcrafted radiomics, or whether radiomics remains complementary for interpretable tumour characterisation. We compared radiomics, conventional CNN features, MedicalNet-pretrained features, MedVAE representations, and fusion variants for binary ccRCC classification on KiTS23, reporting area under the receiver operating characteristic curve (AUC) with bootstrap confidence intervals and average precision (AP) as a complementary class-imbalance-sensitive metric. We further assessed branch-removal ablation, TCGA/AIMI external transfer, and interpretability using radiomics permutation importance and gate-level analysis. Internally, 3D MedVAE gated fusion achieved the best performance, with an AUC of 82.7% and AP of 92.2%. On the external TCGA cohort, the same model achieved an AUC of 79.5% and AP of 98.9%, although specificity remains uncertain because only two external non-ccRCC cases were available. Gate analysis showed a radiomics-dominant fusion regime, suggesting that foundation representations acted as case-dependent refinement signals rather than replacements for structured tumour descriptors. These findings support radiomics as a complementary and clinically interpretable component of CT-based RCC characterisation in the foundation-model era.

Keywords: Renal cell carcinoma · Computed tomography · Radiomics · Foundation representations · External transfer · Interpretability

1 Introduction

Renal cell carcinoma (RCC) is a major urological malignancy for which accurate preoperative subtype characterisation remains clinically important. Distinguishing clear cell RCC (ccRCC) from non-clear cell RCC is clinically relevant

because histological subtype is associated with prognosis, treatment response, and management decisions. Contrast-enhanced CT is central to renal mass assessment [8,10], but visual interpretation is limited by overlapping enhancement patterns, acquisition heterogeneity, and difficulty in quantifying intratumoural heterogeneity.

Radiomics provides structured tumour descriptors by extracting predefined shape, intensity, and texture features from routine CT [2,21,14]. CT radiomics has shown promise for distinguishing ccRCC from non-clear cell subtypes [18], but remains sensitive to segmentation, acquisition protocol, and feature-selection choices. Deep learning and pretrained image representations offer complementary information, with CNN-based transfer learning explored in renal imaging [9,13,5] and foundation representations increasingly used when labelled data are limited [3,16]. This raises the practical question of whether foundation representations reduce the need for handcrafted radiomics, or whether radiomics remains useful as structured and interpretable tumour evidence. Prior RCC studies have reported promising subtype-classification performance using CT radiomics [18] and end-to-end multi-phase CNNs [13]. However, direct state-of-the-art ranking across studies is difficult because of differences in cohorts, contrast phases, subtype definitions, preprocessing, and validation protocols. We therefore position this work as a controlled within-cohort comparison of radiomics, conventional pretrained CNNs, and MedVAE representations, rather than as a claim of global state-of-the-art performance.

Interpretability and external transfer are also important for clinical adoption. Post hoc saliency methods can be visually intuitive but weakly coupled to the learned decision pathway [1,6,17,4,12]. In contrast, radiomics features correspond to defined tumour phenotype measurements and can support decision-centric interpretation within fusion models. Meanwhile, internal test performance alone is insufficient to assess robustness, motivating cautiously framed external evaluation under acquisition and segmentation shift.

This work investigates radiomics–foundation complementarity for ccRCC versus non-clear cell RCC classification from contrast-enhanced CT. We benchmark radiomics, CNN/MedicalNet features, MedVAE representations, and fusion variants on KiTS23. We further assess TCGA/AIMI external transfer and interpretability using radiomics permutation importance and gate-level analysis.

2 Methods

2.1 Study Cohorts and Preprocessing

Figure 1 summarises the overall study design. The internal cohort was derived from KiTS23, a public multi-institutional contrast-enhanced CT dataset with kidney, tumour, and cyst segmentations [7]. Multifocal cases were excluded to avoid lesion-level ambiguity, and the final task was formulated as binary ccRCC versus non-clear cell RCC classification. This binary formulation was selected as a clinically relevant first-stage distinction because the individual non-clear-cell subtype groups were small. Finer subtype classification is left for future work.

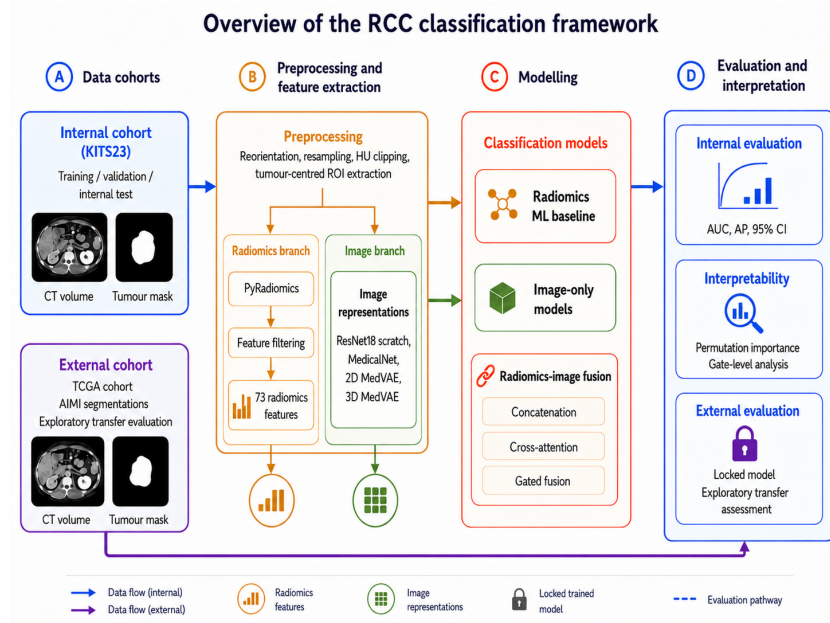


Fig. 1. Overview of the RCC framework: KiTS23 for internal evaluation, TCGA-KIRC/KIRP/KICH with AIMI segmentations for exploratory external transfer, and radiomics–image modelling with interpretability analyses.

All CT volumes and masks were reoriented to RAS, resampled to 1.0 mm isotropic spacing, and intensity-clipped to $[-150, 200]$ HU. Image volumes were resampled using linear interpolation and masks using nearest-neighbour interpolation. For each tumour, we generated a zero-margin tumour ROI for radiomics extraction and a tumour-centred ROI with a fixed 12-voxel margin for image-representation learning. When multiple tumour annotations were available, a STAPLE consensus mask was used [19].

An independent external cohort was derived from TCGA-KIRC, TCGA-KIRP, and TCGA-KICH. Tumour masks were obtained from the AIMI Annotations initiative, which provides AI-generated DICOM-SEG kidney, tumour, and cyst annotations for NCI Imaging Data Commons collections [15,11]. To reduce segmentation-quality uncertainty, we retained only expert-corrected masks or masks assigned the highest expert quality score. The external cohort was used only for evaluation, and its positive-enriched composition was interpreted as exploratory transfer evidence rather than definitive diagnostic specificity.

2.2 Radiomics Feature Extraction and Classical Baseline

Radiomics features were extracted from the zero-margin tumour ROI using PyRadiomics [14] with IBSI-aligned definitions [21]. Shape features were computed from the original image, while first-order and texture features were ex-

tracted from original and Laplacian-of-Gaussian filtered images with $\sigma \in \{1, 2, 3\}$ using a fixed bin width of 25 HU, yielding 386 initial features.

Robustness filtering removed features with $\text{ICC} < 0.75$, low median absolute deviation, and high pairwise correlation ($|\rho| > 0.95$). Correlation-based hierarchical clustering was then used to retain representative features, resulting in a 73-dimensional radiomics representation for fusion models.

For the classical radiomics baseline, the 73-dimensional radiomics representation was defined once from the internal KiTS23 cohort and also used as the radiomics input for fusion models. Classifier-level feature selection, scaling, and SVC fitting were restricted to the internal training data and then applied unchanged to validation, internal test, and external cohorts.

2.3 Image Representations and Fusion Models

We compared 3D ResNet-18 trained from scratch, MedicalNet-initialised 3D ResNet-18 [5], and MedVAE-based image representations [16] under the same preprocessing and evaluation protocol. MedVAE was evaluated using both slice-based 2D and volumetric 3D branches. Image-only models passed the extracted representation directly to a classifier head. All reported MedVAE models used partial fine-tuning. In the 3D gated model, only the later encoder/output modules were updated jointly with the projection, fusion, and classification layers. The pretrained parameters used a lower learning rate than the newly added layers (10^{-5} versus 10^{-4}). The flattened MedVAE latent tensor and radiomics vector were projected into a shared 512-dimensional space. A $512 \rightarrow 256 \rightarrow 512$ MLP with ReLU activation and sigmoid output generated the feature-wise gate, and the fused representation was passed to a single linear classifier producing two logits. This lightweight classification head limited task-specific capacity and focused the comparison on the representations and fusion mechanisms.

For fusion models, the selected radiomics vector was standardised using training-set statistics and projected to the same latent dimension as the image embedding. We evaluated concatenation, cross-attention, and gated fusion as representative integration strategies. For gated fusion, given image embedding $f_{\text{img}} \in \mathbb{R}^d$ and projected radiomics embedding $h_{\text{rad}} \in \mathbb{R}^d$, the fused representation was defined as:

$$g = \sigma(\text{MLP}(h_{\text{rad}})), \quad h_{\text{fuse}} = g \odot f_{\text{img}} + (1 - g) \odot h_{\text{rad}}, \quad (1)$$

where $\odot$ denotes element-wise multiplication. Larger gate values indicate greater image-branch weighting, whereas smaller values indicate greater radiomics-branch weighting. This design allowed the model to adaptively modulate image features using structured radiomics evidence, while preserving gate values as an interpretable estimate of branch weighting.

2.4 Decision-Centric Interpretability

We performed two complementary interpretability analyses for the best-performing fusion model. First, radiomics permutation importance was used to quantify feature contribution. Each radiomics feature was randomly permuted across cases

while keeping image inputs and all other radiomics features unchanged. The procedure was repeated five times per feature, and the mean decrease in AUC was used as the feature-importance score.

Second, we analysed the learned gate values of the gated-fusion model. For each case, the gate vector was extracted during inference and summarised by its mean value. This analysis was performed on both the internal test set and the external cohort to examine how the model weighted image and radiomics branches during prediction.

3 Experiments and Results

3.1 Experimental Setup

The internal KiTS23 cohort was split at the case level into training, validation, and test sets using stratified sampling based on the binary ccRCC label, with fixed proportions of 60%, 25%, and 15%, respectively. Model selection was based on validation AUC. A centre-held-out split could not be constructed because centre identifiers were unavailable in the KiTS23 metadata used in this study. Therefore, the internal test set represents held-out case-level evaluation rather than explicit centre-wise out-of-distribution validation. Weighted cross-entropy loss mitigated class imbalance, and image-branch augmentation included random flips, 90-degree rotations, and small affine perturbations; all validation, internal test, and external inputs were processed deterministically.

The primary metric was AUC, with AP, sensitivity, specificity, and other threshold-dependent metrics also reported. Threshold-dependent metrics used a Youden-index threshold selected on the internal validation set and locked for internal test and external TCGA evaluation [20]. Bootstrap 95% confidence intervals were estimated for AUC, and additionally for AP on the external cohort. The TCGA/AIMI cohort was used only for locked-model transfer assessment.

3.2 Internal Performance on KiTS23

Table 1 summarises the internal held-out test performance. The radiomics SVC baseline achieved an AUC of 74.4%, outperforming both image-only MedVAE branches. Fusion improved performance over image-only models, with 3D MedVAE concatenation, cross-attention, and gated fusion achieving AUCs of 73.8%, 79.5%, and 82.7%, respectively. The best internal result was obtained by 3D MedVAE gated fusion (AUC 82.7%, AP 92.2%). MedicalNet and MedVAE differ in architecture, pretraining objective, and pretraining data. Their comparison should therefore be interpreted as a comparison of complete pretrained representation pipelines rather than as an isolated assessment of pretraining strategy. Nevertheless, the 3D MedVAE gated-fusion model achieved the best internal performance among all evaluated models. Among the evaluated 3D MedVAE fusion variants, gated fusion achieved a higher AUC than concatenation and cross-attention, supporting constrained channel-wise modality weighting in this

Table 1. Internal held-out test performance for ccRCC versus non-ccRCC classification. AUC includes bootstrap 95% confidence intervals. All values are percentages. Conventional CNNs are reported as gated-fusion reference baselines. The full fusion-strategy comparison is reported for 3D MedVAE.

Category	Model	AUC (95% CI)	AP	Accuracy	F1	Sensitivity	Specificity
Radiomics	SVC baseline	74.4 [61.2, 87.3]	84.8	62.0	65.6	52.6	83.8
CNN fusion	ResNet18 gated fusion	78.4 [63.5, 90.9]	86.5	75.0	81.9	81.0	61.1
CNN fusion	MedicalNet gated fusion	78.6 [63.9, 91.3]	86.4	70.0	75.0	64.3	83.3
Foundation only	2D MedVAE	59.5 [42.9, 75.4]	78.8	63.3	75.0	78.6	27.8
Foundation only	3D MedVAE	63.2 [46.0, 79.9]	77.2	71.7	82.5	95.2	16.7
Foundation fusion	2D MedVAE gated fusion	79.6 [67.0, 90.4]	90.1	76.7	84.4	90.5	44.4
Foundation fusion	3D MedVAE concatenation	73.8 [60.5, 87.3]	86.3	61.7	67.6	57.1	72.2
Foundation fusion	3D MedVAE cross-attention	79.5 [67.6, 90.5]	88.9	71.7	79.5	78.6	55.6
Foundation fusion	3D MedVAE gated fusion	82.7 [70.7, 92.2]	92.2	71.7	77.9	71.4	72.2

limited-data setting. Overall, radiomics remained competitive as a standalone descriptor family and provided complementary information when fused with MedVAE representations.

3.3 Branch-Removal Ablation of the Best Fusion Model

We further performed branch-removal ablation on the best-performing 3D MedVAE gated-fusion model by routing either the image representation or the projected radiomics representation alone through the trained gated-fusion classifier, while keeping all model parameters fixed. As shown in Table 2, the full fusion model achieved an AUC of 82.7%, whereas image-branch-only and radiomics-branch-only inference reached 60.4% and 52.9%, respectively. This supports radiomics-image complementarity within the trained fusion pathway. These single-branch variants are not independently optimised classifiers, and therefore assess pathway dependence rather than standalone modality performance. The large gap between full fusion and either branch alone indicates that the classifier did not simply rely on a single dominant input, but required joint access to both representation spaces.

Table 2. Branch-removal ablation within the trained 3D MedVAE gated-fusion model. AUC is reported on the internal test set.

Variant	Image branch	Radiomics branch	AUC (%)
Radiomics branch only	✗	✓	52.9
Image branch only	✓	✗	60.4
Full fusion	✓	✓	82.7

3.4 Exploratory External Transfer on TCGA with AIMI Segmentations

Table 3 reports locked-model external evaluation on the TCGA renal cancer cohort with AIMI segmentations. This analysis was intended as exploratory transfer assessment rather than definitive external validation, because the cohort was strongly positive-enriched and contained only two non-ccRCC cases. Therefore, specificity and AUC should be interpreted cautiously, and confusion-matrix counts are reported explicitly.

Table 3. Exploratory external transfer performance on the TCGA cohort with AIMI segmentations. Models were evaluated using locked internal-validation thresholds without external refitting. AUC and AP are reported with bootstrap 95% confidence intervals. All metric values are percentages.

Category	Model	<i>n</i>	AUC (95% CI)	AP (95% CI)	F1	Sensitivity	Specificity	TN/FP/FN/TP
Radiomics	SVC baseline	41	83.3 [59.0, 100.0]	99.0 [97.5, 100.0]	97.4	97.4	50.0	1/1/1/38
CNN fusion	ResNet18 gated fusion	41	93.6 [79.5, 100.0]	99.7 [98.9, 100.0]	93.3	89.7	50.0	1/1/4/35
CNN fusion	MedicalNet gated fusion	41	96.2 [87.2, 100.0]	99.8 [99.3, 100.0]	90.1	82.1	100.0	2/0/7/32
Foundation only	2D MedVAE	41	70.5 [30.8, 100.0]	97.9 [94.7, 100.0]	87.3	79.5	50.0	1/1/8/31
Foundation only	3D MedVAE	41	64.1 [33.3, 92.3]	97.6 [95.1, 99.6]	92.1	89.7	0.0	0/2/4/35
Foundation fusion	2D MedVAE gated fusion	41	83.3 [65.4, 97.4]	99.1 [98.0, 99.9]	93.3	89.7	50.0	1/1/4/35
Foundation fusion	3D MedVAE gated fusion	41	79.5 [66.6, 92.3]	98.9 [98.1, 99.6]	87.0	76.9	100.0	2/0/9/30

Several internally trained models retained high positive-class ranking under TCGA/AIMI domain and segmentation shift. The 3D MedVAE gated-fusion model achieved an external AUC of 79.5% and AP of 98.9%, while MedicalNet gated fusion achieved the highest external AUC of 96.2%. This external ranking should be interpreted cautiously, as the apparent MedicalNet advantage may reflect model robustness under TCGA/AIMI shift but was estimated from only two negative cases. This may indicate that different pretrained representations respond differently to acquisition and segmentation shift, rather than a stable superiority of one model family. High AP values mainly reflect positive-class ranking in a cohort with 95.1% ccRCC prevalence, and the very small number of negative cases limits conclusions about balanced diagnostic performance. These results therefore support exploratory transfer feasibility, but not definitive population-level generalisation.

3.5 Radiomics-Based Interpretability

To examine which structured tumour descriptors contributed most strongly to the best-performing fusion model, we performed repeated permutation importance analysis on the radiomics branch. Figure 2 shows the top-ranked radiomics features according to the mean decrease in AUC after feature-wise permutation. The most influential descriptors were dominated by texture and heterogeneity-related feature families, including GLDM, GLCM, GLSZM, GLRLM, first-order, and Laplacian-of-Gaussian-derived features.

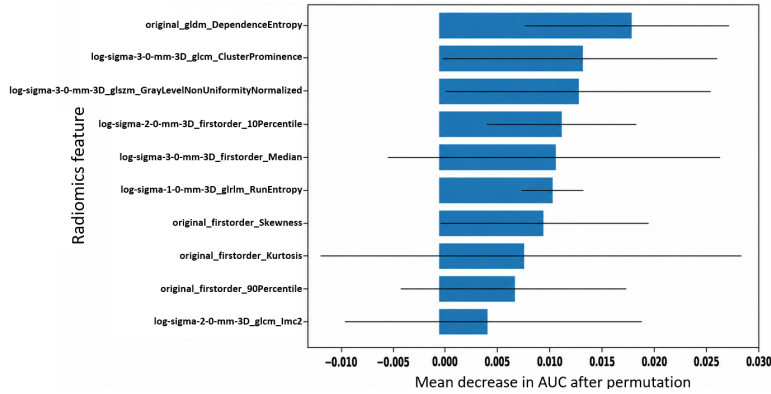


Fig. 2. Top radiomics features ranked by mean AUC decrease after repeated feature-wise permutation. Error bars show variation across repetitions.

These results indicate that the fusion model relied on interpretable descriptors of intratumoural heterogeneity and intensity distribution, rather than only on abstract image embeddings. Because these features are defined radiomics measurements, they provide an imaging-level interpretation of model behaviour. However, they should be interpreted as tumour phenotype descriptors rather than direct histopathological surrogates.

3.6 Gate-Level Interpretability

We analysed gate values from the 3D MedVAE gated-fusion model. Each case produced a 512-dimensional gate vector, where larger values indicate greater image-branch weighting and smaller values indicate greater radiomics-branch weighting. The analysis was performed on the internal test set and on the 41-case external TCGA cohort with AIMI segmentations, which contained 39 ccRCC and 2 non-ccRCC cases.

As shown in Fig. 3, mean gate values were consistently below 0.5 in both cohorts. The internal test set showed a mean gate value of 0.1227 ± 0.0675 , while the external cohort showed a similar mean value of 0.1328 ± 0.0619 . This indicates that the best-performing fusion model operated in a radiomics-dominant regime rather than using balanced multimodal weighting.

Because the external cohort contained only two non-ccRCC cases, class-stratified gate comparisons on the external cohort should be interpreted cautiously. Nevertheless, the overall gate distribution suggests that the model primarily relied on structured radiomics descriptors, with image representations acting as case-dependent modulation signals.

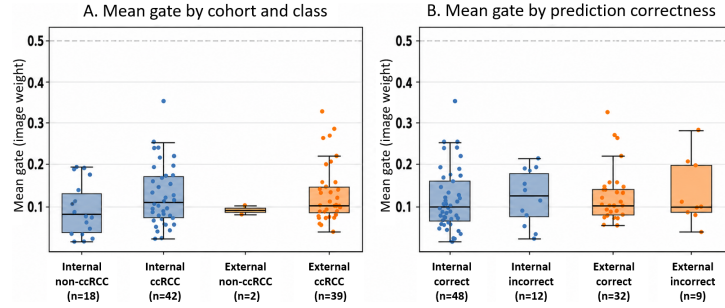


Fig. 3. Gate-level analysis of the 3D MedVAE gated-fusion model. Each point shows the case-level mean of the 512-dimensional gate vector; higher values indicate greater image weighting and lower values greater radiomics weighting. External class-stratified patterns should be interpreted cautiously due to only two non-ccRCC cases.

4 Conclusion

This study shows that handcrafted radiomics remains clinically relevant for CT-based RCC subtype classification in the foundation-model era. Radiomics was competitive as a standalone descriptor family, while branch-removal ablation and internal performance indicated that the best 3D MedVAE fusion pathway depended on both radiomics and image representations, supporting complementary rather than substitutive value. Exploratory TCGA/AIMI evaluation suggested transfer feasibility under acquisition and segmentation shift, but the positive-enriched external cohort and limited non-ccRCC cases preclude definitive claims about diagnostic specificity. Gate-level and permutation analyses further showed a radiomics-dominant decision pathway, supporting structured tumour descriptors as clinically communicable model evidence. Future work should evaluate larger balanced external cohorts, finer RCC subtype classification, and alternative volumetric encoders, including vision transformers and newer CT foundation models, while examining whether radiomics-guided gating transfers across encoder architectures.

Acknowledgments. Yuan Liang acknowledges doctoral scholarship support from the China Scholarship Council (CSC). This work was funded by Taighde Éireann – Research Ireland through the Research Ireland Centre for Research Training in Machine Learning (18/CRT/6183).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I.J., Hardt, M., Kim, B.: Sanity checks for saliency maps. In: *Advances in Neural Information Processing Systems (NeurIPS)*. pp. 9525–9536 (2018), <https://papers.nips.cc/paper/8160-sanity-checks-for-saliency-maps>
2. 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)
3. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., et al.: On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021)
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). <https://doi.org/10.1016/j.ejrad.2023.110786>
5. Chen, S., Ma, K., Zheng, Y.: Med3d: Transfer learning for 3d medical image analysis. In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. pp. 149–159. Springer (2019)
6. Ghassemi, M., Oakden-Rayner, L., Beam, A.L.: The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health* **3**(11), e745–e750 (2021). [https://doi.org/10.1016/S2589-7500\(21\)00208-9](https://doi.org/10.1016/S2589-7500(21)00208-9)
7. Heller, N., et al.: Kits23 dataset (2023). <https://doi.org/10.5281/zenodo.8253742>
8. International Agency for Research on Cancer: Kidney fact sheet. <https://gco.iarc.who.int/media/globocan/factsheets/cancers/29-kidney-fact-sheet.pdf> (2024), global Cancer Observatory, GLOBOCAN 2022, version 1.1
9. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. *Medical image analysis* **42**, 60–88 (2017)
10. Ljungberg, B., Albiges, L., Bedke, J., Bex, A., Capitanio, U., Giles, R., Hora, M., Klatte, T., Marconi, L., Powles, T., et al.: Eau guidelines on renal cell carcinoma (2023)
11. Murugesan, G.K., McCrumb, D., Aboian, M., Verma, T., Soni, R., Memon, F., Farahani, K., Pei, L., Wagner, U., Fedorov, A.Y., Clunie, D., Moore, S., Van Oss, J.: Ai-generated annotations dataset for diverse cancer radiology collections in nci image data commons. *Scientific Data* **11**(1), 1165 (2024). <https://doi.org/10.1038/s41597-024-03977-8>
12. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* **1**, 206–215 (2019). <https://doi.org/10.1038/s42256-019-0048-x>
13. Uhm, K.H., Jung, S.W., Choi, M.H., Shin, H.K., Yoo, J.I., Oh, S.W., Kim, J.Y., Kim, H.G., Lee, Y.J., Youn, S.Y., et al.: Deep learning for end-to-end kidney cancer diagnosis on multi-phase abdominal computed tomography. *NPJ precision oncology* **5**(1), 54 (2021)
14. 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–e107 (2017)

15. Van Oss, J., Murugesan, G.K., McCrumb, D., Soni, R.: Image segmentations produced by the aimi annotations initiative (2023). <https://doi.org/10.5281/zenodo.8436591>
16. Varma, M., Kumar, A., Sluijs, R., Ostmeier, S., Blankemeier, L., Chambon, P., Blüthgen, C., Prince, J., Langlotz, C., Chaudhari, A.: Medvae: Efficient automated interpretation of medical images with large-scale generalizable autoencoders (02 2025). <https://doi.org/10.48550/arXiv.2502.14753>
17. van der Velden, B.H.M., Kuijf, H.J., Gilhuijs, K.G.A., Viergever, M.A.: Explainable artificial intelligence (xai) in deep learning-based medical image analysis. *Medical Image Analysis* **79**, 102470 (2022). <https://doi.org/10.1016/j.media.2022.102470>
18. 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)
19. 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)
20. 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)
21. 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)