

Beyond Discrimination: Evaluating Probability Calibration in Radiomics–Image Fusion for Renal Cell Carcinoma Classification

Yuan Liang^{1,2*}, Morgan May^{1,3*}, Abraham Campbell², Sourav Bhattacharjee⁴,
Simon Caton², and Pierpaolo Dondio⁵

¹ Research Ireland Centre for Research Training in Machine Learning, ML-Labs,
Dublin, Ireland

² School of Computer Science, University College Dublin, Dublin, Ireland

³ School of Computer Science, Technological University Dublin, Dublin, Ireland

⁴ School of Veterinary Medicine, University College Dublin, Dublin, Ireland

⁵ School of Computing, Dublin City University, Dublin, Ireland
`yuan.liang@ucdconnect.ie`

Abstract. Reliable probability estimates are important for the clinical interpretation of machine-learning classifiers, yet radiomics–image fusion models for renal cell carcinoma (RCC) are commonly evaluated primarily by discrimination. Using 396 patients from the KiTS23 dataset, we evaluated whether adding radiomics to four image-model families improved probabilistic performance in addition to discrimination. Seventeen individual models and four probability-average ensembles were assessed using discrimination, probability calibration, and proper scoring measures. Across 12 matched image-only versus fusion comparisons, AUC and AP improved in all 12, while Brier score and negative log likelihood (NLL) improved in 11 of 12 and expected calibration error (ECE) in eight of 12. The 3D MedVAE gated-fusion model achieved the strongest overall individual-model profile. Probability averaging across concatenation, cross-attention, and gated fusion produced favourable point estimates across all five primary metrics relative to image-only models in all four families, although paired-bootstrap support was stronger for Brier score and NLL than for ECE. These findings show that radiomics–image fusion can improve probabilistic prediction quality beyond discrimination, while calibration behaviour varies across image architectures and fusion strategies. We show that our proposed probability-average ensemble models improve discrimination and calibration simultaneously across all architectures in the study.

Keywords: Probability calibration · Renal cell carcinoma · Radiomics · Feature fusion · Deep learning

1 Introduction

Renal cell carcinoma (RCC) is a heterogeneous malignancy comprising several histological subtypes with distinct biological behaviour, prognosis, and treat-

* These authors contributed equally to this work.

ment implications. Among RCC, clear cell RCC (ccRCC) is the most common subtype, and its preoperative distinction from non-clear cell RCC is clinically relevant for patient stratification and management [2]. Contrast-enhanced computed tomography (CT) plays a central role in the assessment of renal masses, but substantial overlap in tumour appearance can make subtype characterisation challenging using visual interpretation alone. Quantitative imaging approaches have therefore increasingly been investigated to extract additional information from routine CT.

Radiomics provides predefined measurements of tumour shape, intensity, and texture and has shown promise for distinguishing ccRCC from non-clear cell RCC [25]. In parallel, deep neural networks can learn higher-dimensional image representations directly from CT data, while pretrained medical imaging models offer a means of exploiting such representations when labelled datasets are limited [20,24]. These two feature extractions are conceptually complementary: radiomics provides structured descriptors of tumour phenotype, whereas learned image representations can capture complex spatial patterns that are difficult to specify manually. Consequently, integrating radiomics with learned image representations offers a potential means of improving RCC classification beyond either representation alone.

Most evaluations of such models, however, focus primarily on discrimination [3]. Here, discrimination refers to a model’s ability to distinguish or rank ccRCC cases relative to non-ccRCC cases. Metrics such as the area under the receiver operating characteristic curve (AUC) and average precision (AP) quantify this ability, but they do not determine whether the predicted probabilities are numerically reliable. A classifier may achieve strong discrimination while systematically assigning probabilities that are too extreme or otherwise inconsistent with observed outcome frequencies [22]. This distinction is particularly important in medical AI, where model outputs may be interpreted as estimates of disease probability rather than merely as ranking scores.

Probability calibration therefore provides a complementary dimension of model evaluation. A well-calibrated classifier should, for example, assign an event probability close to 0.8 to groups of patients in whom the event occurs approximately 80% of the time. Neural networks may exhibit substantial miscalibration despite high predictive accuracy [7]. Although probability calibration has been considered in multimodal imaging studies for other clinical tasks [15,27,16,11], systematic evaluation of calibration remains limited in renal cancer imaging research [21]. For renal cancer, it therefore remains unclear whether the addition of a structured radiomics modality improves only the ranking performance of deep models or also the quality of the probabilities they produce.

In this study, we investigate this question through a controlled comparison of image-only and radiomics–image fusion models for binary ccRCC classification across various combinations of neural network architectures and radiomics fusion methods.

2 Methods

To systematically evaluate the effect of radiomics–image fusion, we used a structured 4×5 model evaluation grid comprising four image model families and five prediction configurations within each family: an image-only baseline, three radiomics-fusion variants, and a probability-average ensemble of the three fusion models. The specific image backbones and fusion mechanisms are described in Section 2.3. A radiomics-only SVC was evaluated separately as an additional baseline.

2.1 Dataset and Preprocessing

The study used the Kidney Tumour Segmentation 2023 (KiTS23) dataset [10], a multi-institutional collection of contrast-enhanced abdominal CT scans with kidney and renal-tumour annotations. The classification task was formulated as clear-cell renal cell carcinoma (ccRCC) versus non-clear-cell RCC. Multifocal cases were excluded to avoid ambiguity between patient- and lesion-level labels. A total of 396 patients were included. A fixed patient-level split was used consistently across all model families, with approximately 60% of cases assigned to training, 25% to validation, and 15% to held-out testing. The validation and test partitions had the same class distributions, with ccRCC accounting for approximately 70% of cases in both partitions. The validation partition was used for model selection, whereas the held-out test partition was used only for final evaluation and was not used for model fitting or checkpoint selection.

All CT volumes and tumour masks were reoriented to RAS orientation and re-sampled to 1.0-mm isotropic spacing. CT intensities were clipped to $[-150, 200]$ HU. Linear interpolation was used for images and nearest-neighbour interpolation for segmentation masks [28]. A zero-margin tumour region of interest was used for radiomics extraction, whereas image-based models used a tumour-centred context region with a 12-voxel margin. When multiple tumour annotations were available, a STAPLE consensus mask was generated [26]. The underlying patients, labels, and split membership were kept identical across model comparisons.

2.2 Radiomics Representation

Radiomics features were extracted from the tumour region using PyRadiomics [23] with IBSI-aligned definitions [28]. Shape features were calculated from the original image, while first-order and texture features were extracted from both the original and Laplacian-of-Gaussian-filtered images with $\sigma \in \{1, 2, 3\}$ using a fixed bin width of 25 HU.

The initial 386 features were reduced by removing features with $\text{ICC} < 0.75$, the lowest 20% by median absolute deviation, and highly correlated features with $|\rho| > 0.95$, followed by hierarchical clustering with a distance threshold of 0.2. This resulted in a 73-dimensional radiomics representation. For fusion models, radiomics features were standardised using statistics estimated from the training

partition and applied unchanged to the validation and test partitions. A support vector classifier (SVC) using the selected radiomics features was included as a radiomics-only baseline.

2.3 Image Representations and Classification Models

We evaluated conventional convolutional and pretrained image representations alongside MedVAE representations [24]. The conventional image models used a 3D ResNet-18 architecture [9] either trained from random initialisation or initialised from MedicalNet weights [5]. For each backbone, an image-only classifier and three corresponding radiomics-fusion variants, concatenation [12], cross-attention [19], and gated fusion [1], were evaluated under the same patient-level split and training protocol. For fusion models, the image and radiomics representations were projected into compatible latent spaces before fusion.

MedVAE [24] was evaluated in both 2D and 3D configurations. For the 2D models, tumour-containing axial slices were selected according to tumour area and the top 11 slices were encoded individually. Their latent representations were projected and mean-pooled to obtain a case-level image representation. For the 3D models, the tumour-centred CT volume was encoded directly using the volumetric MedVAE encoder, and the resulting latent representation was flattened and projected to a case-level embedding. Image-only 2D and 3D MedVAE classifiers were evaluated alongside their radiomics-fusion counterparts.

Concatenation. The projected image and radiomics embeddings were directly concatenated [12],

$$\mathbf{h}_{\text{concat}} = [\mathbf{f}_{\text{img}}; \mathbf{h}_{\text{rad}}], \quad (1)$$

and the joint representation was passed to a classification head producing two logits. This provides a simple fusion baseline in which the classifier learns the joint contribution of both modalities without an explicit cross-modal interaction mechanism.

Cross-attention. Cross-attention was used to model conditional interactions between the image and radiomics representations [19]. Image token(s) were used as queries, whereas the projected radiomics representation was transformed into a small set of radiomics tokens serving as keys and values. For the 2D model, let $\mathbf{U} \in \mathbb{R}^{k \times d}$ denote the k slice-level image embeddings and let $\mathbf{h}_{\text{rad}} \in \mathbb{R}^d$ denote the projected radiomics representation. The image queries and radiomics tokens were constructed as

$$\mathbf{Q} = \mathbf{U}\mathbf{W}_Q, \quad \mathbf{R} = \text{reshape}(\mathbf{h}_{\text{rad}}\mathbf{W}_R), \quad (2)$$

where $\mathbf{Q} \in \mathbb{R}^{k \times d_c}$ and $\mathbf{R} \in \mathbb{R}^{T \times d_c}$. Cross-modal attention was then computed as

$$\text{Attn}(\mathbf{Q}, \mathbf{R}, \mathbf{R}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{R}^T}{\sqrt{d_c}}\right)\mathbf{R}. \quad (3)$$

The attended image tokens were further processed using residual connections, layer normalisation, and a feed-forward block, then mean-pooled and combined

with the pooled image representation for classification. For ResNet-18, MedicalNet, and 3D MedVAE, the projected case-level image representation served as a single query under the same formulation. Thus, the radiomics tokens condition the image representation before the final classification stage rather than being combined only by late feature concatenation.

Gated fusion. For gated fusion, the projected image representation $\mathbf{f}_{\text{img}} \in \mathbb{R}^d$ and radiomics representation $\mathbf{h}_{\text{rad}} \in \mathbb{R}^d$ were combined using a learned feature-wise gate [1]:

$$\mathbf{g} = \sigma(\text{MLP}(\mathbf{h}_{\text{rad}})), \quad (4)$$

$$\mathbf{h}_{\text{fuse}} = \mathbf{g} \odot \mathbf{f}_{\text{img}} + (1 - \mathbf{g}) \odot \mathbf{h}_{\text{rad}}, \quad (5)$$

where $\odot$ denotes element-wise multiplication. The fused representation was subsequently passed to a linear classifier producing two class logits. For the 3D MedVAE gated model, both modalities were projected into a 512-dimensional shared space and the gate was generated by a $512 \rightarrow 256 \rightarrow 512$ multilayer perceptron with ReLU and sigmoid activations.

Probability-average ensembles. As a secondary robustness analysis, equal-weight probability-average ensembles were constructed separately for the ResNet-18, MedicalNet, 2D MedVAE, and 3D MedVAE families using the model-native predicted probabilities of ccRCC from concatenation, cross-attention, and gated fusion:

$$p_{i,\text{ensemble}} = \frac{p_{i,\text{concat}} + p_{i,\text{cross}} + p_{i,\text{gated}}}{3}, \quad i = 1, \dots, n, \quad (6)$$

where $p_{i,m} = P_m(Y_i = 1 \mid \mathbf{x}_i)$ denotes the model-native predicted probability of ccRCC for the i -th sample produced by model $m \in \{\text{concat}, \text{cross}, \text{gated}\}$, and $p_{i,\text{ensemble}}$ denotes the corresponding ensemble probability. The ensemble weights were fixed a priori at $1/3$; no weights were estimated from the data, and test set outcomes were not used in ensemble construction.

Training Protocol. All neural network models used the same fixed patient-level split and were trained with class-weighted cross-entropy and Adam optimisation. CT inputs were standardised over non-zero voxels, centre-cropped or padded to the architecture-specific size, and augmented using random flips, 90° rotations, and affine transformations. Validation AUC guided learning rate scheduling, early stopping, and checkpoint selection. ResNet-18 and MedicalNet used 128^3 inputs, while 2D MedVAE used the 11 largest tumour-containing axial slices from a 128^3 context volume and 3D MedVAE used 96^3 inputs. MedVAE encoders were partially fine-tuned, with pretrained parameters assigned a lower learning rate than newly initialised downstream layers (10^{-5} versus 10^{-4}). All fusion models used the same 73-dimensional radiomics representation standardised using training-set statistics. In total, 17 individual models and four probability-average ensembles were evaluated. Validation-selected checkpoints were frozen before final test set evaluation.

2.4 Probability Calibration

Calibration was evaluated using the model-native probability of ccRCC, $p_i = \hat{P}(Y_i = 1 \mid x_i)$, where $Y_i = 1$ denotes ccRCC. For neural networks, p_i was obtained from the ccRCC component of the two-class softmax output. For the radiomics SVC, rather than softmax outputs, p_i was obtained from the classifier’s fitted class probability estimates, which use an internal probability mapping procedure.

Probability calibration was quantified using expected calibration error (ECE) [7,17], calibration intercept and slope [18], and reliability diagrams, while the Brier score and NLL were used as complementary proper scoring rules for overall probabilistic performance. Proper scoring rules are instance-wise measures for evaluating the quality of probability estimates that avoid the need for score binning [17].

To limit sparse bins in the relatively small test cohort, ECE was calculated using an equal-frequency binning scheme [17] as a scalar summary of calibration error:

$$\text{ECE} = \sum_{m=1}^M \frac{|B_m|}{n} |\bar{p}_m - \bar{y}_m|, \quad (7)$$

where $\bar{p}_m$ denotes the mean predicted probability and $\bar{y}_m$ the observed ccRCC proportion in bin B_m , M denotes the number of the bins, and n is the total number of samples. ECE was calculated using five equal-frequency bins ($M = 5$).

The Brier score and NLL are defined as

$$\text{Brier} = \frac{1}{n} \sum_{i=1}^n (p_i - y_i)^2, \quad (8)$$

and

$$\text{NLL} = -\frac{1}{n} \sum_{i=1}^n [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)], \quad (9)$$

where p_i and y_i denote the predicted probability and the true label (according to the ground truth) of the i -th sample, respectively. Lower values for Brier score and NLL indicate better probabilistic prediction performance. ECE was interpreted together with the Brier score and NLL rather than as a standalone measure [4,22].

Calibration intercept and slope were estimated using logistic calibration regression [22,13,8,6]:

$$\text{logit}\{P(Y_i = 1)\} = \alpha + \beta \text{logit}(p_i), \quad (10)$$

where α is the calibration intercept and β is the calibration slope. Perfect calibration corresponds to $\alpha = 0$ and $\beta = 1$. The reported intercept was the joint intercept from the free-intercept/free-slope calibration model rather than calibration-in-the-large.

Reliability diagrams [7,14] were used to plot the observed ccRCC proportion against the mean predicted probability in each of the five equal-frequency bins. Wilson 95% confidence intervals were calculated for the observed event proportion. Points above the identity line indicate underconfidence, whereas points below the line indicate overconfidence.

The primary analysis comprised 12 matched image-only versus radiomics-fusion comparisons, with each fusion model compared against the corresponding image-only model using the same backbone, patient split, and evaluation protocol. Ninety-five percent confidence intervals were estimated using 1,000 ordinary patient-level bootstrap samples with percentile limits. The four probability-average ensembles were evaluated separately against their corresponding image-only models, and ensemble-minus-image-only differences in ECE, Brier score, and NLL were assessed using 1,000 paired patient-level bootstrap samples with identical resampled patient indices.

3 Results and Discussion

3.1 Discrimination, Calibration, and Probabilistic Performance

Table 1 summarises performance across the 17 individual models and four probability-average ensembles. Across the 12 matched image-only versus radiomics-fusion comparisons, AUC and AP were higher for every fusion model. The Brier score and NLL were lower in 11 of 12 comparisons each, whereas ECE was lower in eight of 12. Thus, the discrimination gain associated with radiomics fusion was highly consistent across architectures, while improvements in the proper scoring rules were also frequent and reductions in calibration error were more fusion-dependent.

Architecture-specific results showed that the fusion strategy with the strongest discrimination was not necessarily the one with the lowest calibration error. For ResNet-18, cross-attention achieved the highest discrimination, whereas concatenation yielded the lowest ECE and gated fusion the lowest Brier score and NLL. A similar divergence was observed for MedicalNet and 2D MedVAE. In contrast, 3D MedVAE gated fusion showed the strongest overall individual model profile, increasing AUC from 0.632 to 0.827 and AP from 0.772 to 0.922 while reducing ECE from 0.178 to 0.106, Brier score from 0.230 to 0.150, and NLL from 1.359 to 0.472. The radiomics-only SVC nevertheless achieved the lowest ECE (0.063).

Calibration regression provided a complementary view of probability calibration (Table 1). Radiomics fusion yielded slope point estimates closer to the ideal value of 1 for both CNN backbones and for 3D MedVAE, although confidence intervals were wide and the magnitude of improvement varied by fusion mechanism. The 2D MedVAE image-only model already had a slope near 1, and fusion produced mixed changes. These findings further indicate that calibration behaviour varied across image representations and fusion strategies.

Table 1. Internal discrimination, probability calibration, probabilistic performance, and calibration-regression parameters on the held-out KiTS23 test cohort ($n = 60$). Values in brackets denote 95% ordinary patient-bootstrap confidence intervals. Ideal calibration corresponds to slope = 1 and intercept = 0.

Image Model	Radiomics	AUC	AP	ECE	Brier	NLL	Calibration slope		Calibration intercept	
-	radiomics only	0.766	0.894	0.063	0.186	0.550	0.91	[0.26, 2.25]	-0.02	[-1.47, 0.98]
ResNet18	image only	0.532	0.738	0.153	0.216	0.646	0.20	[-0.59, 1.21]	0.62	[-0.35, 1.73]
	concatenation	0.663	0.799	0.096	0.207	0.607	0.61	[-0.03, 1.52]	0.56	[-0.04, 1.28]
	cross-attention	0.790	0.897	0.136	0.190	0.556	0.79	[0.38, 1.51]	0.78	[0.15, 1.65]
	gated fusion	0.784	0.865	0.124	0.176	0.533	1.22	[0.53, 2.32]	0.61	[0.02, 1.36]
	ensemble	0.780	0.879	0.119	0.178	0.537	1.18	[0.56, 2.31]	0.56	[-0.11, 1.30]
MedicalNet	image only	0.639	0.852	0.206	0.259	0.741	0.20	[-0.56, 0.59]	0.86	[0.34, 1.53]
	concatenation	0.746	0.866	0.129	0.188	0.559	1.70	[0.75, 3.51]	0.41	[-0.20, 1.11]
	cross-attention	0.776	0.902	0.219	0.221	0.652	0.80	[0.41, 1.60]	1.15	[0.59, 2.01]
	gated fusion	0.786	0.864	0.115	0.169	0.518	1.14	[0.51, 2.33]	0.57	[-0.07, 1.35]
	ensemble	0.815	0.905	0.156	0.178	0.537	1.64	[0.90, 3.53]	0.77	[0.12, 1.62]
2D MedVAE	image only	0.595	0.788	0.134	0.221	0.650	1.11	[-0.47, 2.81]	-0.97	[-3.78, 1.62]
	concatenation	0.747	0.880	0.149	0.220	0.669	0.42	[0.17, 0.89]	0.72	[0.14, 1.48]
	cross-attention	0.761	0.875	0.065	0.171	0.532	0.62	[0.34, 1.14]	0.06	[-0.70, 0.79]
	gated fusion	0.796	0.901	0.174	0.200	0.604	1.02	[0.56, 1.94]	-1.35	[-3.13, -0.15]
	ensemble	0.792	0.899	0.071	0.162	0.492	0.94	[0.48, 1.80]	-0.16	[-0.99, 0.63]
3D MedVAE	image only	0.632	0.772	0.178	0.230	1.359	0.07	[-0.05, 0.35]	0.50	[-0.39, 1.35]
	concatenation	0.738	0.863	0.121	0.194	0.624	0.41	[0.17, 0.80]	0.31	[-0.47, 1.02]
	cross-attention	0.795	0.889	0.201	0.241	1.204	0.16	[0.08, 0.30]	0.37	[-0.31, 1.02]
	gated fusion	0.827	0.922	0.106	0.150	0.472	0.52	[0.34, 0.93]	0.02	[-0.89, 0.72]
	ensemble	0.807	0.911	0.140	0.165	0.494	0.59	[0.36, 0.96]	0.14	[-0.59, 0.87]

3.2 Calibration and Probabilistic Performance Across Fusion Strategies

The expanded comparison across four image backbones revealed consistent but non-uniform effects of the three fusion strategies. Gated fusion reduced the Brier score and NLL relative to the corresponding image-only model in all four backbones and reduced ECE in three of four. Concatenation reduced the Brier score in all four backbones and reduced both ECE and NLL in three of four. Cross-attention reduced NLL in all four backbones and Brier score in three of four, but reduced ECE in only two of four. Thus, improvements in the proper scoring rules were more reproducible across fusion mechanisms than reductions in calibration error.

The fusion strategy that maximised discrimination was not necessarily the one with the lowest calibration error. For ResNet-18, cross-attention achieved the highest AUC and AP, whereas concatenation produced the lowest ECE and gated fusion the lowest Brier score and NLL. MedicalNet cross-attention similarly improved discrimination while slightly increasing ECE relative to the image-only model. In contrast, 3D MedVAE gated fusion provided the most favourable joint profile across discrimination, calibration error, and proper scoring rules. These findings indicate that the probabilistic consequences of radiomics fusion cannot be inferred from discrimination alone and depend on how the two modalities are integrated. This variation across individual fusion mechanisms motivated the

Table 2. Paired patient-bootstrap differences between probability-average ensembles and their corresponding image-only models. Differences are defined as ensemble minus image-only; negative values therefore indicate lower probabilistic error for the ensemble. Values in brackets denote 95% paired-bootstrap confidence intervals. Bold values indicate intervals excluding zero.

Model family	Δ ECE (95% CI)	Δ Brier (95% CI)	Δ NLL (95% CI)
ResNet18	-0.034 [-0.122, 0.124]	-0.037 [-0.105, 0.028]	-0.109 [-0.286, 0.060]
MedicalNet	-0.050 [-0.175, 0.040]	-0.080 [-0.136, -0.022]	-0.204 [-0.354, -0.061]
2D MedVAE	-0.063 [-0.133, 0.043]	-0.059 [-0.115, -0.008]	-0.157 [-0.293, -0.031]
3D MedVAE	-0.038 [-0.163, 0.070]	-0.066 [-0.151, 0.012]	-0.864 [-1.783, -0.163]

ensemble analysis, which tested whether averaging their complementary predictions could yield a more consistent performance profile.

3.3 Probability-Average Ensemble Robustness Analysis

To examine whether the broader probabilistic-performance pattern depended on selecting a single fusion mechanism, equal-weight probability-average ensembles of concatenation, cross-attention, and gated fusion were compared with the corresponding image-only models. Importantly, this simple averaging strategy improved all five primary performance metrics at the point-estimate level across all four model families: AUC and AP increased, while ECE, Brier score, and NLL decreased (Table 1).

Paired patient-bootstrap differences for the probability-average ensembles are summarised in Table 2. Negative differences indicate lower probabilistic error for the ensemble. None of the ECE confidence intervals excluded zero, although the point estimates favoured the ensemble across all four model families. Notably, for 3D MedVAE, probability averaging reduced ECE from 0.178 to 0.140, Brier score from 0.230 to 0.165, and NLL from 1.359 to 0.494, corresponding to relative reductions of 21.3%, 28.3%, and 63.6%, respectively. Thus, averaging the three complementary fusion strategies produced consistently favourable point estimates across discrimination, calibration error, and overall probabilistic performance, while paired-bootstrap evidence was stronger for the proper scoring rules than for ECE.

3.4 Reliability of the 3D MedVAE Gated-Fusion Model

The predicted-risk distribution and reliability diagram for 3D MedVAE gated fusion are shown in Figure 1. Predictions were concentrated toward high ccRCC probabilities and showed heterogeneous deviations from the identity line across probability bins. The model achieved ECE 0.106, Brier score 0.150, and NLL 0.472. Calibration regression yielded an intercept of 0.021 [95% CI: -0.890, 0.722] and a slope of 0.517 [0.340, 0.930]. The near-zero joint intercept suggested limited overall probability offset after accounting for the slope, whereas the slope below 1 indicated that the predicted probabilities remained overly extreme across the probability range.

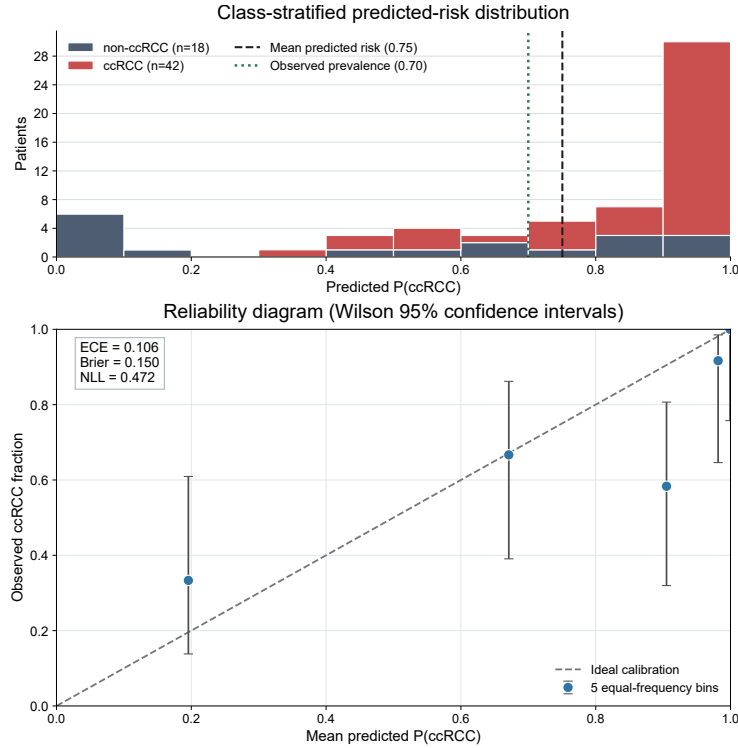


Fig. 1. Predicted-risk distribution and reliability diagram for the 3D MedVAE gated-fusion model on the held-out KiTS23 test cohort. Reliability was evaluated using five equal-frequency bins; error bars denote Wilson 95% confidence intervals and the dashed line represents perfect calibration.

4 Conclusion

Radiomics–image fusion improved discrimination consistently across four image-model families, while improvements in overall probabilistic performance were more reproducible than reductions in calibration error. Across 12 matched comparisons, AUC and AP were higher in all 12 comparisons, Brier score and NLL were lower in 11 of 12, and ECE was lower in eight of 12. Calibration behaviour nevertheless varied across image architectures and fusion mechanisms. Among individual models, 3D MedVAE gated fusion achieved the strongest joint discrimination–probabilistic performance profile. Probability averaging across the three fusion strategies also produced favourable point estimates relative to image-only models across all four families, with stronger paired-bootstrap support for Brier score and NLL than for ECE. These results support evaluating multimodal medical-image classifiers using complementary discrimination, calibration, and proper-scoring measures rather than discrimination alone. External multi-centre validation is required to determine the generalisability of these findings.

References

1. Arevalo, J., Solorio, T., Montes-y Gómez, M., González, F.A.: Gated multimodal units for information fusion. In: International Conference on Learning Representations Workshop (2017)
2. Bex, A., Abu Ghanem, Y., Albiges, L., Bonn, S., Campi, R., Capitanio, U., Dabestani, S., Hora, M., Klatte, T., Kuusk, T., et al.: European association of urology guidelines on renal cell carcinoma: The 2025 update. *European Urology* **87**(6), 683–696 (2025). <https://doi.org/10.1016/j.eururo.2025.02.020>
3. Bhandari, A., Ibrahim, M., Sharma, C., Liong, R., Gustafson, S., Prior, M.: CT-based radiomics for differentiating renal tumours: A systematic review. *Abdominal Radiology* **46**(5), 2052–2063 (2021)
4. Brier, G.W.: Verification of forecasts expressed in terms of probability. *Monthly Weather Review* **78**(1), 1–3 (1950)
5. Chen, S., Ma, K., Zheng, Y.: Med3D: Transfer learning for 3d medical image analysis. arXiv preprint arXiv:1904.00625 (2019)
6. Christiansen, F., Konuk, E., Ganeshan, A.R., Welch, R., Pales Huix, J., Czekierdowski, A., Leone, F.P.G., Haak, L.A., Fruscio, R., Gaurilcikas, A., et al.: International multicenter validation of ai-driven ultrasound detection of ovarian cancer. *Nature medicine* **31**(1), 189–196 (2025)
7. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 70, pp. 1321–1330. PMLR (2017)
8. Hatt, M., Krizsan, A., Rahmim, A., Bradshaw, T., Costa, P.F., Forgács, A., Seifert, R., Zwanenburg, A., El Naqa, I., Kinahan, P., et al.: Joint eanm/snmml guideline on radiomics in nuclear medicine: Jointly supported by the eanm physics committee and the snmml physics, instrumentation and data sciences council. *European Journal of Nuclear Medicine and Molecular Imaging* **50**(2), 352–375 (2023)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
10. Heller, N., Isensee, F., Tejpaul, R., Wood, A., Papanikolopoulos, N., Weight, C.: 2023 kidney and kidney tumor segmentation challenge (2023). <https://doi.org/10.5281/zenodo.7840134>
11. Hu, S., Wang, X.: Robust liver fibrosis staging under scanner shift with deep and radiomic streams (2026), <https://openreview.net/forum?id=i3ainP2BVH>
12. Huang, S.C., Pareek, A., Seyyedi, S., Banerjee, I., Lungren, M.P.: Fusion of medical imaging and electronic health records using deep learning: A systematic review and implementation guidelines. *npj Digital Medicine* **3**(1), 136 (2020). <https://doi.org/10.1038/s41746-020-00341-z>
13. Lambin, P., Leijenaar, R.T., Deist, T.M., Peerlings, J., De Jong, E.E., Van Timmeren, J., Sanduleanu, S., Larue, R.T., Even, A.J., Jochems, A., et al.: Radiomics: the bridge between medical imaging and personalized medicine. *Nature reviews Clinical oncology* **14**(12), 749–762 (2017)
14. Murphy, A.H., Winkler, R.L.: Reliability of subjective probability forecasts of precipitation and temperature. *Journal of the Royal Statistical Society Series C: Applied Statistics* **26**(1), 41–47 (1977)
15. Nelson, M.A., Tasci, E.S., Wang, Z., Zhou, Z., Aktas, H.E., Bejar, A.M., Keles, E., Hong, Z., Taflan, S.S., Tasci, M.E., et al.: Gaussian meta-space augmenta-

- tion for stacking ensembles in multimodal ipmn risk stratification. arXiv preprint arXiv:2608.11472 (2026)
16. Raposo, H.: Is imaging redundant? leakage-controlled multi-representation assessment of ct-clinical fusion for two-year survival prediction in non-small-cell lung cancer. (June 29, 2026). Available at SSRN: <https://ssrn.com/abstract=7023198>
 17. Silva Filho, T., Song, H., Perello-Nieto, M., Santos-Rodriguez, R., Kull, M., Flach, P.: Classifier calibration: a survey on how to assess and improve predicted class probabilities. *Machine Learning* **112**(9), 3211–3260 (2023)
 18. Stevens, R.J., Poppe, K.K.: Validation of clinical prediction models: what does the “calibration slope” really measure? *Journal of clinical epidemiology* **118**, 93–99 (2020)
 19. Tsai, Y.H.H., Bai, S., Liang, P.P., Kolter, J.Z., Morency, L.P., Salakhutdinov, R.: Multimodal transformer for unaligned multimodal language sequences. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. pp. 6558–6569 (2019). <https://doi.org/10.18653/v1/P19-1656>
 20. 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)
 21. Ursprung, S., Beer, L., Bruining, A., Woitek, R., Stewart, G.D., Gallagher, F.A., Sala, E.: Radiomics of computed tomography and magnetic resonance imaging in renal cell carcinoma—a systematic review and meta-analysis. *European Radiology* **30**(6), 3558–3566 (2020). <https://doi.org/10.1007/s00330-020-06666-3>
 22. Van Calster, B., McLernon, D.J., van Smeden, M., Wynants, L., Steyerberg, E.W., et al.: Calibration: The achilles heel of predictive analytics. *BMC Medicine* **17**(1), 230 (2019). <https://doi.org/10.1186/s12916-019-1466-7>
 23. Van Griethuysen, J.J.M., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R.G.H., Fillion-Robin, J.C., Pieper, S., Aerts, H.J.W.L.: Computational radiomics system to decode the radiographic phenotype. *Cancer Research* **77**(21), e104–e107 (2017). <https://doi.org/10.1158/0008-5472.CAN-17-0339>
 24. 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: *Proceedings of the 8th International Conference on Medical Imaging with Deep Learning*. vol. 301, pp. 1597–1626. PMLR (2026)
 25. 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). <https://doi.org/10.1038/s41598-021-93069-z>
 26. 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)
 27. Xi, I.L., Zhao, Y., Wang, R., Chang, M., Purkayastha, S., Chang, K., Huang, R.Y., Silva, A.C., Vallieres, M., Habibollahi, P., et al.: Deep learning to distinguish benign from malignant renal lesions based on routine MR imaging. *Clinical Cancer Research* **26**(8), 1944–1952 (2020)
 28. 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)