

<https://doi.org/10.1038/s41698-026-01482-2>

Robust multicentre detection and classification of colorectal liver metastases on CT: application of foundation models

Check for updates

Shruti Atul Mali¹ ✉, Zohaib Salahuddin¹, Yumeng Zhang¹, Andre Aichert², Xian Zhong³, Henry C. Woodruff^{1,4}, Maciej Bobowicz⁵, Katrine Riklund^{6,7}, Juozas Kupčinskas⁸, Lorenzo Faggioni⁹, Roberto Francischello⁹, Razvan L. Miclea¹⁰ & Philippe Lambin^{1,4} On behalf of EUCanImage working group*

Colorectal liver metastases (CRLM) remain a major cause of cancer-related mortality. Early, reliable detection on CT imaging is essential for curative treatment planning, yet the diagnostic performance of AI models often declines across scanners and institutions, limiting clinical generalizability. In this study, we developed and evaluated a multi-center foundation-model-based AI pipeline for patient-level classification and lesion-level detection of CRLM on contrast-enhanced CT. Using data from the EuCanImage consortium ($n = 2437$) and TCIA_CRLM ($n = 197$, all CRLM), we benchmarked several pretrained foundation models and identified UMedPT as the optimal encoder. The final model achieved an AUC of 0.89 and a sensitivity of 0.82 on the EuCanImage test set, with a sensitivity of 0.85 on the external TCIA cohort. Excluding the most uncertain 20% of cases improved AUC to 0.90 and balanced accuracy to 0.85. Decision-curve analysis indicated superior net benefit over “treat-all” and “treat-none” strategies for threshold probabilities between 0.35 and 0.75. The lesion-level detector identified 69.1% of lesions overall, increasing substantially with lesion size. Grad-CAM maps showed strong correspondence between attention regions and metastases in high-confidence predictions. These findings demonstrate that foundation-model-based pipelines enable robust, generalizable, and interpretable solutions for CRLM detection and classification in multi-center CT imaging.

Colorectal cancer (CRC) remains a major global health burden, ranking as the second most common type of cancer in women and third in men, with an estimated 3.2 million new cases worldwide annually, by the year 2040¹. Owing to the challenges of early detection of CRC, about 50% of the patients will ultimately progress to colorectal liver metastases (CRLM), which is one of the most aggressive types of liver tumors². Precise diagnosis and

localization of CRLM is important to inform treatment choices that can include surgical excision, systemic therapy, or locoregional treatment and directly affect patient outcomes^{3,4}. Computed tomography (CT) is frequently used in the diagnosis and monitoring of CRLM^{5,6}. Although the CT allows distinguishing between lesions and healthy liver tissue due to their contrast difference, it is challenging to differentiate between benign lesions

¹Department of Precision Medicine, GROW - Research Institute for Oncology and Reproduction, Maastricht University, Maastricht, The Netherlands. ²Digital Technology and Innovation, Siemens Healthcare GmbH, Erlangen, Germany. ³Department of Medical Ultrasonics, Institute of Diagnostic and Interventional Ultrasound, The First Affiliated Hospital of Sun Yat-sen University, Guangzhou, China. ⁴Department of Radiology and Nuclear Medicine, GROW - Research Institute for Oncology and Reproduction, Maastricht University, Medical Center+, Maastricht, The Netherlands. ⁵2nd Department of Radiology, Medical University of Gdańsk, Gdańsk, Poland. ⁶Umeå Center for Functional Brain Imaging (UFBI), Umeå University, Umeå, Sweden. ⁷Department of Diagnostics and Intervention, Umeå university, Umeå, Sweden. ⁸Department of Gastroenterology, Lithuanian University of Health Sciences, Kaunas; Institute for Digestive Research, Lithuanian University of Health Sciences, Kaunas, Lithuania. ⁹Department of Translational Research, Academic Radiology, University of Pisa, Pisa, Italy. ¹⁰Department of Radiology and Nuclear Medicine, Maastricht University Medical Center +, Maastricht, The Netherlands. *A list of authors and their affiliations appears at the end of the paper.

✉ e-mail: s.mali@maastrichtuniversity.nl

(e.g. cysts) and malignant lesions since they may share similar imaging characteristics^{7,8}. This leads to a time-consuming diagnostic procedure and a high likelihood of missing lesions. This has led to growing interest in the development of computer-aided diagnostic (CAD) tools that can aid radiologists in the classification and detection of focal liver lesions^{9,10}.

Radiomics has been given a lot of attention among other AI-based methods in medical image analysis^{11,12}. Radiomics extracts hand-crafted quantitative descriptors from medical images to characterize tissue morphology, texture, and intensity, and has been explored for prognosis, treatment response assessment, and lesion classification in CRLM^{13,14}. However, radiomic features are sensitive to scanner variability, reconstruction parameters, and acquisition protocols, requiring harmonization methods such as ComBat (corrects batch effects) for reproducibility in multi-center studies^{14,15}. ComBat is a statistical tool that corrects for scanner- or center-related batch effects in features to improve reproducibility. The analysis of medical images with the help of deep learning has evolved after the discovery of AlexNet in natural image classification¹⁶. Since then, convolutional neural networks (CNNs) and architectures such as U-Net¹⁷ have been used extensively in various applications such as classification, detection, and segmentation in medical imaging^{18,19}. Deep learning has demonstrated great potential in automated lesion detection and characterization in CRLM^{20,21}. Recent studies have investigated deep learning approaches for automatic segmentation and detection of CRLM across CT and MRI datasets, showing encouraging results for lesion delineation, treatment response assessment, and quantitative tumor burden estimation^{22–24}. However, deep learning algorithms typically need huge annotated datasets, and they do not tend to generalize to external settings and multi-center cohorts, which restricts their applicability in practice^{18,25}.

Recently, Foundation models (FMs) have become a novel paradigm in medical image analysis²⁶. FMs are trained on large-scale and heterogeneous datasets, usually employing self-supervised learning. They tend to produce generalizable feature representations that can be fine-tuned to achieve downstream tasks, especially with limited labeled data^{27,28}. FMs have shown strong performance in oncology imaging in segmentation, classification, and biomarker discovery, and provide robust and more adaptable results than traditional supervised models²⁹. Their ability to adapt to domain shifts and exploit broad prior knowledge makes them well-suited for multi-center CRLM analysis.

Beyond accuracy, explainability and reliability are also crucial for clinical translation. Black box AI systems run the risk of compromising the clinician’s trust and delaying integration into practice^{30,31}. Integration of FMs with explainable AI techniques can not only offer robust performance, but also transparency, as a result of which a clinician can interpret the prediction and incorporate it into diagnostic decision-making^{32,33}. Although

FMs have been increasingly successful in other imaging applications, their use with CRLM has never been explored in a systematic manner. Moreover, the majority of previous research has focused on accuracy measures only without considering clinical interpretability, uncertainty quantification, and the decision-making potential. To be deployed in oncologic imaging, it is necessary not only to have high diagnostic accuracy but also to know the model uncertainty and to be able to visualize the anatomical regions driving predictions.

To fill these gaps, we developed a multi-center AI pipeline based on foundation models to perform patient-level classification and lesion-level detection of CRLM on contrast-enhanced CT. By using the EuCanImage consortium dataset, we benchmarked various pretrained FMs and fine-tuned UMedPT as the best encoder for downstream tasks. Using the best FM, we designed two complementary pipelines: (i) a patient-level classification model classifying CRLM from non-CRLM and negative cases, utilizing Gradient-weighted Class Activation Mapping (Grad-CAM) for interpretable visualization of model predictions; and (ii) a lesion-level detection model to locate metastatic liver lesions across multiple imaging cohorts. In addition, we also included uncertainty quantification (UQ), decision-curve analysis (DCA), and explainability to evaluate the reliability of the model and the potential clinical benefit. This study is one of the most extensive analyses of FM-driven AI in CRLM, filling the gap between technical performance and clinical relevance in the direction of trustworthy and interpretable AI decision support in oncologic imaging.

Results

Clinical and Imaging data

The final EuCanImage cohort included 2437 patients across four centers: KAUNOS (*n* = 490), GUMED (*n* = 848), UMEA (*n* = 401), and UNIPi (*n* = 698), plus an additional 197 patients from the external TCIA_CRLM dataset (Table 1). The mean patient age ranged between 63.5 ± 13.3 and 68.2 ± 12.5 years, with a balanced gender distribution (female 35–50%; male 50–61%). The TCIA_CRLM cohort had a mean age of 59.7 ± 12.3 years and included 117 females (59.4%) and 80 males (40.6%). Among EuCanImage patients, 710 (29%) were diagnosed with colorectal liver metastases (CRLM), 1004 (41%) presented with non-CRLM lesions, and 723 (30%) were negative for healthy subjects. The TCIA_CRLM dataset consisted exclusively of CRLM-positive cases (*n* = 197, 100%). The total number of annotated CRLM lesion masks was 796 cases across both datasets: 599 from EuCanImage centers and 197 from TCIA_CRLM.

Foundation-model benchmarking

To identify the most suitable pretrained backbone for CRLM classification, we compared five candidate foundation models using logistic regression

Table 1 | Demographic and imaging overview of the EuCanImage colorectal liver metastasis (CRLM) cohort

	KAUNOS	GUMED	UMEA	UNIPi	TCIA_CRLM	<i>p</i> -value
Clinical data						
Age: Mean (SD)	63.5 ± 13.3	66.3 ± 11.8	66.6 ± 10.1	68.2 ± 12.5	59.7 ± 12.3	<0.001
Gender						<0.001
Female	242 (49.4%)	363 (42.8%)	142 (35.4%)	348 (49.9%)	117 (59.4%)	
Male	248 (50.6%)	485 (57.2%)	247 (61.6%)	350 (50.1%)	80 (40.6%)	
unknown	-	-	12 (3.0%)	-	-	
Imaging data						
Total patients	490	848	401	698	197	
CRLM cases	237 (48.36%)	353 (41.63%)	27 (6.73%)	93 (13.32%)	197 (100%)	
non-CRLM cases	149 (30.41%)	209 (24.64%)	374 (93.26%)	272 (38.96%)	0 (0.0%)	
Negative cases	104 (21.22%)	286 (33.72%)	0 (0.0%)	333 (47.70%)	0 (0.0%)	
Lesion masks (CRLM)	216 (44.08%)	308 (36.32%)	26 (6.48%)	49 (7.02%)	197 (100%)	

Values are reported as mean ± standard deviation (SD) or number (percentage). The table summarizes patient distribution, age, gender, and counts of CRLM, non-CRLM, and negative cases, along with the total number of lesion masks available per site.

Table 2 | Benchmarking of logistic regression (LR) classifiers applied to deep feature embeddings extracted from five candidate foundation models

Model	AUC (CI)	Balanced accuracy (CI)	Sensitivity (CI)	Specificity (CI)	Precision (CI)	F1-score (CI)
FMCIB + LR	0.70 (0.66–0.74)	0.63 (0.59–0.67)	0.52 (0.45–0.58)	0.74 (0.71–0.78)	0.45 (0.4–0.52)	0.48 (0.42–0.54)
UMedPT + LR	0.81 (0.77–0.84)	0.74 (0.71–0.78)	0.62 (0.55–0.68)	0.87 (0.84–0.9)	0.65 (0.58–0.72)	0.63 (0.58–0.68)
CT-FM + LR	0.74 (0.70–0.78)	0.67 (0.63–0.71)	0.58 (0.52–0.65)	0.76 (0.72–0.8)	0.5 (0.43–0.56)	0.54 (0.48–0.59)
Models Genesis + LR	0.73 (0.69–0.77)	0.64 (0.60–0.68)	0.55 (0.49–0.62)	0.73 (0.69–0.76)	0.45 (0.4–0.51)	0.5 (0.44–0.55)
Med3D + LR	0.66 (0.61–0.70)	0.62 (0.58–0.66)	0.51 (0.45–0.58)	0.72 (0.68–0.76)	0.43 (0.37–0.49)	0.47 (0.41–0.52)

Reported metrics include area under the ROC curve (AUC), balanced accuracy, sensitivity, specificity, precision, and F1-score with 95% confidence intervals. Values in bold indicate that UMedPT + LR achieved the highest overall performance across metrics.

Table 3 | Performance of the fine-tuned UMedPT + MLP classifier on the independent EuCanImage test set ($n = 733$), EuCanImage sites (KAUNOS, GUMED, UMEA, UNIP), and TCIA_CRLM subset ($n = 197$, all CRLM cases, portal-venous phase only)

Test set	AUC (CI)	Balanced accuracy (CI)	Sensitivity (CI)	Specificity (CI)	Precision (CI)	F1-score (CI)
Datasets						
EuCanImage ($n = 733$)	0.89 (0.86–0.91)	0.82 (0.78–0.84)	0.82 (0.77–0.88)	0.81 (0.78–0.84)	0.64 (0.58–0.69)	0.72 (0.67–0.76)
TCIA_CRLM ($n = 197$)	-	-	0.85 (0.8–0.9)	-	-	-
Centers						
KAUNOS ($n = 147$)	0.92 (0.86–0.96)	0.81 (0.74–0.86)	0.92 (0.84–0.97)	0.7 (0.59–0.8)	0.74 (0.65–0.83)	0.82 (0.75–0.88)
GUMED ($n = 255$)	0.88 (0.83–0.91)	0.81 (0.76–0.86)	0.82 (0.75–0.89)	0.81 (0.74–0.87)	0.75 (0.67–0.82)	0.78 (0.72–0.84)
UMEA ($n = 121$)	0.83 (0.65–0.96)	0.66 (0.48–0.85)	0.38 (0.00–0.75)	0.95 (0.90–0.98)	0.33 (0.00–0.67)	0.35 (0.00–0.63)
UNIP ($n = 210$)	0.81 (0.69–0.91)	0.74 (0.65–0.83)	0.71 (0.55–0.87)	0.77 (0.71–0.83)	0.32 (0.22–0.44)	0.44 (0.32–0.57)

Results are reported for portal-venous phase inputs at the optimal threshold of 0.4267. Metrics include AUC, balanced accuracy, sensitivity, specificity, precision, and F1-score with 95% confidence intervals.

applied to their extracted deep features (Table 2). In terms of AUC, values ranged from 0.66 (Med3D + LR) to 0.81 (UMedPT + LR). FMCIB + LR and ModelsGenesis + LR achieved intermediate values (0.70 and 0.73, respectively), while CT-FM + LR reached 0.74. Balanced accuracy showed the same pattern, with UMedPT + LR having the largest value (0.74) as opposed to 0.62–0.67 with other models.

In terms of sensitivity, which measures the capacity of the model to identify CRLM patients accurately, the overall performance in the LR benchmark was typically low, ranging between 0.51 and 0.62 (Med3D + LR at the lower and UMedPT + LR at the upper). The specificity was also uniformly higher among the models (at most 0.87 in UMedPT + LR), reflecting more discrimination of non-CRLM/negative cases. Precision ranged from 0.43 (Med3D + LR) to 0.65 (UMedPT + LR), and F1-scores from 0.47 to 0.63. Overall, the UMedPT model provided the best classification performance across all metrics.

These results indicate that UMedPT provided the most discriminative and robust feature space for CRLM classification across all classification metrics. In contrast, Med3D + LR and FMCIB + LR consistently underperformed, while CT-FM + LR and ModelsGenesis + LR achieved intermediate results.

To assess potential scanner-related bias, ComBat feature harmonization was applied across sites using the UMedPT features. However, no performance improvement was observed (AUC = 0.72 vs. 0.81 pre-harmonization), indicating that the foundation model already yielded site-robust representations (see Supplementary Material).

Classification performance

Building on the benchmark, we developed a dedicated classification pipeline using UMedPT as the backbone with a fine-tuned MLP head. Although training used all available contrast phases to enhance robustness, evaluation was restricted to the PV phase, which provides the most consistent lesion-to-parenchyma contrast. In our dataset, PV-phase scans were available for

55% of training cases and 94% of cases for the EuCanImage test set and all cases of TCIA-CRLM.

Using the optimal operating threshold (0.4267) derived from Youden's index on the validation folds from cross-validation, the model achieved an AUC of 0.89 (95% CI: 0.86–0.91) and a balanced accuracy of 0.82 (95% CI: 0.78–0.84) (Table 3) on the independent EuCanImage test set ($n = 733$). Sensitivity and specificity were well balanced (0.82 and 0.81, respectively), while precision (0.64) and F1-score (0.72) indicated robust discrimination between CRLM and non-CRLM/negative cases. Figure 1 illustrates the receiver operating characteristic (ROC) curves of the five cross-validation folds, and the ensemble test set, where the values of the area under the curve are consistently high in each fold and the generalization on the EuCanImage test set. The associated confusion matrix of the portal-venous phase EuCanImage test data (Fig. 1) confirms the distribution of true and false predictions, which is in accordance with the reported sensitivity and specificity values.

To further evaluate the external TCIA_CRLM cohort ($n = 197$; all-positive cases), only the recall metric was applicable, where the model achieved a sensitivity of 0.85 (95% CI: 0.8–0.9). Since the TCIA_CRLM test set had all lesions annotated, we further evaluated classification sensitivity stratified by lesion count and lesion diameter. Sensitivity increased with both lesion burden and lesion size (0.98 for patients with more than 4 lesions and 1.00 in patients with lesions more than 5 cm) (Supplementary Fig. 3). The MLP model was evaluated per site, and it gave a stable performance across each of these centers, with AUC values ranging from 0.81 to 0.92 and balanced accuracy between 0.66 and 0.81. The model achieved the highest performance at KAUNOS (AUC = 0.92, sensitivity = 0.92) and GUMED (AUC = 0.88, balanced accuracy = 0.81), indicating consistent generalization within these cohorts. The UMEA site showed a relatively lower F1-score (0.38) and high specificity (0.95), reflecting the impact of severe class imbalance and smaller positive case counts (8/121 for UMEA). Similarly, the UNIP cohort showed a lower F1-score (0.44) despite moderate AUC (0.81)

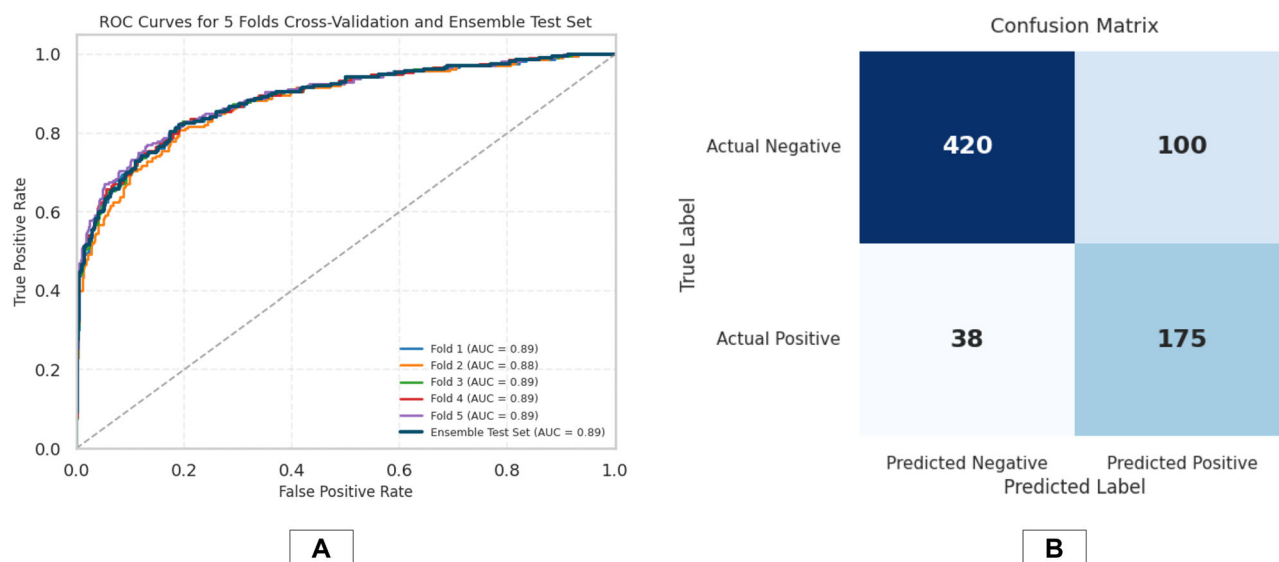


Fig. 1 | Classification performance of the UMedPT + MLP model on the EuCanImage test set (portal-venous phase). **A** ROC curves of five cross-validation folds and the ensemble model, which depicts consistently high AUC values across

folds and robust performance on the EuCanImage test set. **B** Confusion matrix, at the selected operating threshold (0.4267), which presents the distribution of correctly and incorrectly classified cases (TN = 420, FP = 100, FN = 38, TP = 175).

and balanced accuracy (0.74), reflecting the impact of severe class imbalance and relatively smaller positive cases (28/210 for UNIP). Overall, the model showed a good discrimination ability ($AUC \geq 0.81$) and balanced accuracy was in the range 0.66–0.81, giving decent sensitivity-specificity trade-offs across multiple centers.

Misclassification analysis

To better understand model behavior, we examined misclassifications across CRLM, non-CRLM, and Negative subgroups within the EuCanImage test set; for summary reporting, non-CRLM and Negative were grouped as “Other.” At the selected operating threshold, the confusion matrix showed 420 true negatives, 100 false positives, 38 false negatives, and 175 true positives.

In order to study classification failures in more detail, we examined the 29 misclassified cases in the TCIA_CRLM test set, all of which represented false-negative cases (i.e., CRLM-positive patients that were classified as Other). In all these patients, the total number of lesions annotated was 48, with a median of one lesion per case (range: 1–4). The mean lesion volume was 4.53 cm^3 (median: 1.1 cm^3 , IQR: $0.3\text{--}3.1 \text{ cm}^3$). Almost half of all lesions (47.9%) were less than 1 cm^3 , and almost half of the patients (14/29) had sub-centimeter lesions. A small subgroup (7/29) had at least one large lesion ($>10 \text{ cm}^3$). These results showed that misdiagnosed TCIA_CRLM cases were often associated with small or subtle metastases, although a few cases also had larger, possibly poorly visible lesions.

A similar pattern could be observed in the EuCanImage portal-venous subset, in which 138 test cases were misclassified, out of which only 68 had available tumor annotations. In these patients, about 99 annotated lesions were found with a median of two lesions per case (range: 1–5). The mean lesion volume was 5.26 cm^3 (median: 0.65 cm^3 , IQR: $0.3\text{--}2.2 \text{ cm}^3$), and 59.6% of the total annotated lesions were less than 1 cm^3 . Larger lesions ($\geq 10 \text{ cm}^3$) represented approximately 6% of the total. Since not all lesions were annotated, quantitative sensitivity of detection could not be accurately estimated, but the distribution observed indicates that small or less conspicuous metastases were once again common among partially annotated misclassified cases. Probability estimates of the classification (UMedPT + MLP) model were tested with calibration curves on the EuCanImage test set (refer to Supplementary Fig. 1). The model had moderate calibration of the joint test set (Brier = 0.17, slope = 3.2, intercept = -0.50), which showed a bit overconfident estimates of the probability.

Uncertainty quantification

We stratified the subjects into a group of certainty (CG) and an uncertain group (UG) at different thresholds of certainty to determine the relationship between model confidence and predictive reliability on the EuCanImage portal-venous test set (Fig. 2). The curves indicate balanced accuracy and AUC with respect to the percentage of subjects retained in the CG. Expectedly, the performance was always better with high certainty, and the CG outperformed the UG on all measures. CG was maintaining comparable accuracy and AUC values as the overall model performance, whereas UG gradually decreased and became less stable. The shaded regions represent 95% bootstrap confidence intervals. At the threshold where 80% of the subjects were retained in the CG (i.e., the most uncertain 20% were excluded), the model had a balanced accuracy value of 0.85 (95% CI: 0.81–0.89) and $AUC = 0.9$ (95% CI: 0.86–0.93). It means that the accuracy and AUC were noticeably increased by eliminating the 20% most uncertain cases. Conversely, the uncertain group (UG) of the same threshold exhibited a sharp reduction in performance (balanced accuracy = 0.51 (95% CI: 0.50–0.53), $AUC = 0.74$ (95% CI: 0.64–0.8)), which confirms that uncertainty quantification is effective to distinguish between reliable and unreliable predictions.

Decision-curve analysis

We further assessed the clinical utility of the classification model using decision-curve analysis (DCA) on the EuCanImage test set (Fig. 3). The model exhibited a consistently higher net benefit than both ‘treat-all’ and ‘treat-none’ strategies across a broad range of threshold probabilities ($\sim 0.35\text{--}0.75$). The optimum net benefit was observed around a threshold of about 0.5, which implies that the model provides the greatest clinical advantage when the default probability threshold is considered. At low thresholds (<0.35), the ‘treat-all’ and model curves are overlapping, showing that the model has no advantage in that range. However, the model maintained a higher net benefit than both baseline strategies across intermediate thresholds, which showed that it could help reduce unnecessary interventions while maintaining sensitivity for CRLM detection. The net benefit of the model became close to zero at very high thresholds (>0.8), implying that there is little value at very high thresholds unless only highly confident predictions are being used to make decisions.

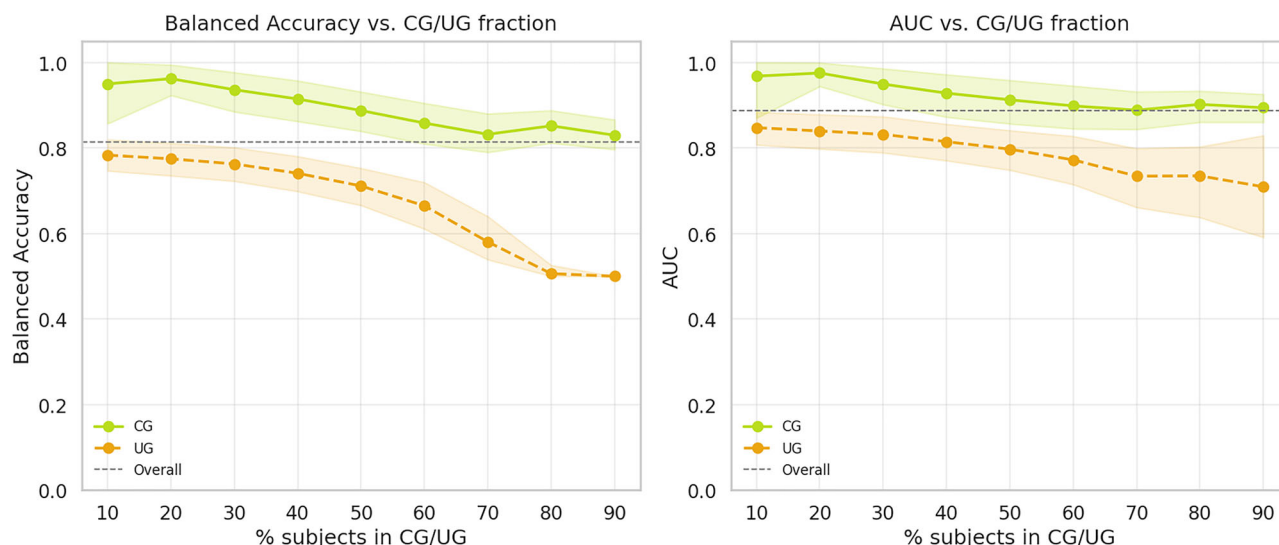


Fig. 2 | Performance analysis based on uncertainty on the EuCanImage portal-venous test cohort. Balanced accuracy (left) and Area under the curve (AUC, right) are plotted as a function of the percentage of subjects retained in the certain group

(CG) and uncertain group (UG). Solid lines indicate CG performance, dashed lines indicate UG performance, and the horizontal dashed line represents overall model performance. Shaded regions denote 95% bootstrap confidence intervals.

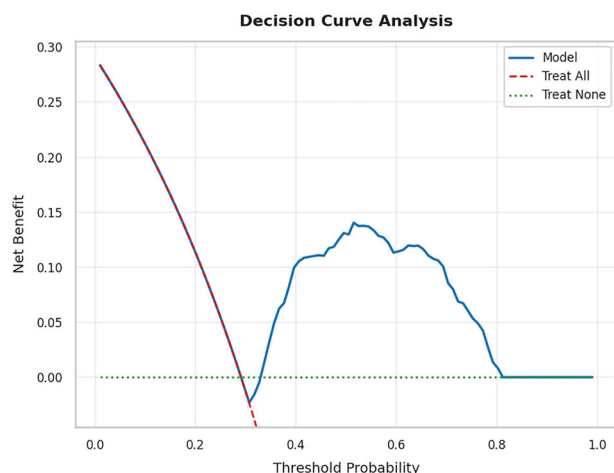


Fig. 3 | Decision-curve analysis (DCA) on the EuCanImage test set. Net benefit is plotted against threshold probability for the classification model (blue), compared with “treat-all” (red dashed) and “treat-none” (green dotted) strategies. The model shows superior net benefit across thresholds between approximately 0.35 and 0.75, peaking near 0.5.

Grad-CAM explainability analysis

To assess model explainability, Grad-CAM heatmaps were generated for scan-level classification predictions (see Fig. 4). Every panel has a corresponding CT slice at its top and the Grad-CAM visualization at its bottom, with warmer colors representing the regions that have the highest contribution to the label being predicted. The examples show model attention in cases of CRLM. The highlighted regions in panels A and B have directly overlapping visible hepatic lesions and are associated with high-confidence CRLM predictions (0.8). In comparison, panels C and D depict weaker patterns of attention. In panel C, the heatmap has a diffuse distribution throughout the abdomen, with the strongest activation observed outside the liver, even though the ground-truth label is negative, which corresponds with the lower predicted probability (0.6). In panel D, the attention is localized, although it covers the lesion; there are also some false-positive regions, which implies a lower-confidence positive prediction (0.4). These results show that the model tends to focus on clinically relevant hepatic regions in true positive cases, where

lower-confidence predictions are linked with more diffuse or off-target activation. Additional qualitative Grad-CAM visualization examples illustrating true positive, false-positive, true negative, and false-negative scenarios are provided in Supplementary Fig. 2.

Lesion-level detection performance

The quantitative analysis for lesion detection was performed on the TCIA-CRLM dataset (Fig. 5), which contains detailed lesion annotations. In the TCIA_CRLM test set, the model was able to detect 331 of 479 ground-truth lesions (69.1%) at an IoU threshold of 0.025 and with a minimum lesion size of 50 voxels ($\approx 0.125 \text{ cm}^3$). Detection performance increased consistently with lesion size, from 30.0% (36/120) in Q1 ($\leq 0.80 \text{ cm}^3$) to 98.3% (118/120) in Q4 ($> 11.49 \text{ cm}^3$). The total number of predictions made by the model was 588, with 257 false positives, a false positive per image rate of 1.3, and a false positive per lesion rate of 0.54. In the case of the EuCanImage test set, only a subset of the lesions were manually labeled; thus, there was no way to report reliable quantitative metrics on the lesion level. The results of detection on this cohort were checked qualitatively and were applied mostly to validate model generalizability to other sites.

Cross-task consistency (TCIA)

Lesion-level detection performance improved in subsets defined by classification accuracy and confidence. In correctly classified cases, sensitivity increased from 69.1% to 71.0%, and in the high-confidence correctly classified subset ($\text{CG} \cap \text{Correct}$), sensitivity reached 74.1%. False-positive rate per image dropped from 1.30 to 1.22. These results indicate that higher classification confidence corresponds to more accurate lesion localization and fewer spurious detections, demonstrating coherence between classification and detection model behavior. Refer to Table 4.

Qualitative detection analysis

Representative examples of detection outputs from all participating EuCanImage centers and the external TCIA_CRLM dataset are presented in Fig. 6. Ground-truth lesion annotations are outlined in green, and model-predicted bounding boxes are shown in red. Across cohorts, the detection model consistently localized both small and large CRLM, capturing diverse appearances and contrast levels. However, performance was less reliable for subtle, low-contrast, or peripherally located lesions, some of which were missed entirely or only partially overlapped with the reference annotation. In several cases, false-positive predictions were observed near vascular or biliary structures, reflecting residual challenges in differentiating small

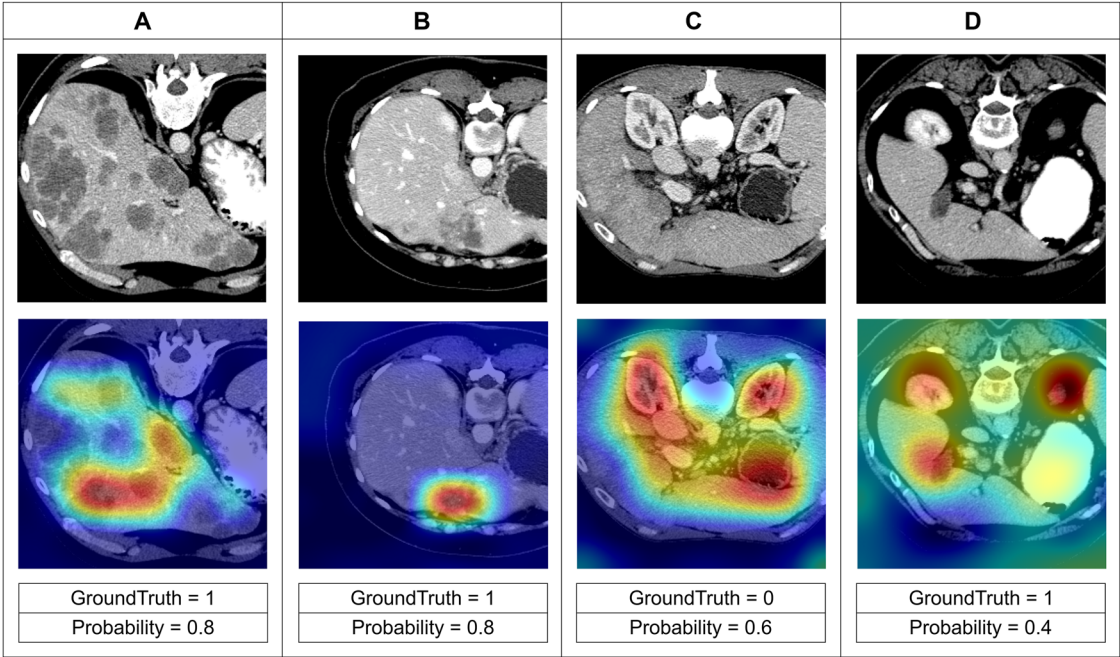


Fig. 4 | Grad-CAM visualizations for scan-level classification predictions. Each panel shows the original CT slice (top) and the corresponding Grad-CAM heatmap (bottom), with ground-truth labels and predicted probabilities displayed below. (A) and (B) demonstrate alignment between model attention and visible liver metastases, while (C) and (D) illustrate less precise activation, including off-target focus on non-hepatic structures (C) or partial lesion coverage (D).

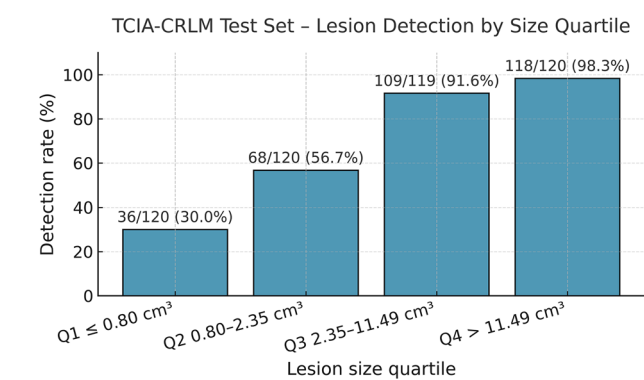


Fig. 5 | Lesion detection rates stratified by lesion size quartile on the TCIA_CRLM test set. Each bar shows the number of detected lesions relative to the total ground-truth lesions within that quartile, with corresponding percentages annotated above.

lesions from anatomical variations. These examples illustrate the model’s cross-site generalization and highlight both successful detections and typical failure modes.

RQS 2.0 assessment

The overall methodological quality and readiness of the proposed AI pipeline were evaluated using the Radiomics Quality Score (RQS) 2.0 framework³⁴. The study achieved a total of 28 out of 35 points, corresponding to a Radiomics Readiness Level (RRL) of 6, and for the deep learning track, which reflects strong methodological transparency, reproducibility, and partial clinical readiness. The scoring captured key strengths across data preparation, model development, validation, and trustworthiness. The cumulative progression of achieved versus maximum attainable scores per readiness level is shown in Fig. 7, illustrating consistent methodological completeness up to RRL-6, where requirements for calibration, interpretability, and external validation were satisfied. A detailed explanation of each criterion, with supporting evidence, is provided in Supplementary Table 2.

Table 4 Lesion-level detection performance on the TCIA_CRLM dataset across evaluation subsets			
Metric	Overall (n = 197)	Correct (n = 168)	CG (80%) ∩ Correct (n = 135)
Sensitivity (IoU > 0.025)	69.1%	71.1%	74.1%
FPR / image	1.30	1.34	1.22
FPR / lesion	0.54	0.52	0.49

Detection sensitivity and false-positive rates are reported for the full test set, correctly classified cases, and the high-confidence subset (CG ∩ Correct). Values in bold indicate that detection performance improved with classification confidence, with sensitivity increasing from 69.1% to 74.1% and decreasing false-positive rates.

Discussion

In this multi-center study, we developed and evaluated an AI pipeline for CRLM using foundation-model features for patient-level classification and lesion-level detection across heterogeneous contrast-enhanced CT. A systematic benchmark of pretrained backbones showed that UMedPT + MLP gave the most balanced performance with an AUC of 0.89 and balanced accuracy of 0.82 on the EuCanImage test cohort at an objectively selected operating point (Youden’s index). The sensitivity of our classification model was also high at about 85% when applied to the external TCIA_CRLM data and 82% when applied to the EuCanImage hold-out test set. This highlights the clinical importance of sensitivity in the detection of CRLM because false-negative detection of true metastases may result in under-treatment or delayed treatment³⁵. At the lesion level for the TCIA_CRLM dataset, the detection model achieved an overall sensitivity of 69.1% and exceeded 98% for the largest quartile. It was more challenging for small ($<1 \text{ cm}^3$) lesions. Grad-CAM examples illustrated this pattern, with high-confidence cases showing focused hepatic attention and lower-confidence cases exhibiting diffuse or off-target activations.

Relative to previous AI research in CRLM that has largely focused on radiomics or segmentation tasks rather than integrated classification and detection, our pipeline combines foundation-model transfer, uncertainty

Fig. 6 | Qualitative detection results across EuCanImage centers and the external TCIA_CRLM dataset. Examples from GUMED, KAUNOS, UMEA, UNIPi, and TCIA cohorts are shown. Ground-truth annotations are displayed in green, and model predictions in red. The figure demonstrates accurate detections, partial overlaps, and typical false-positive or missed cases.

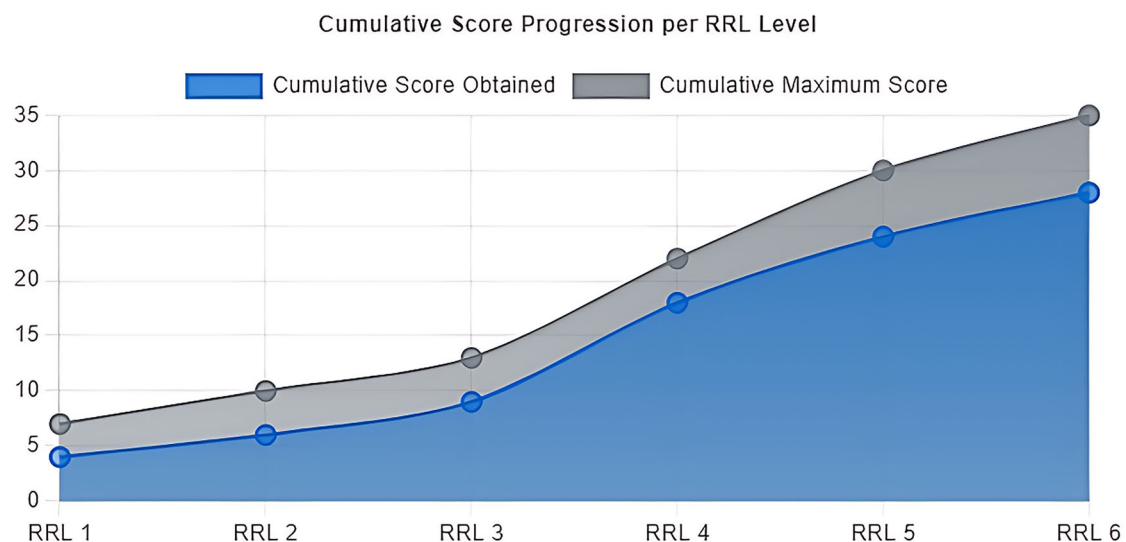
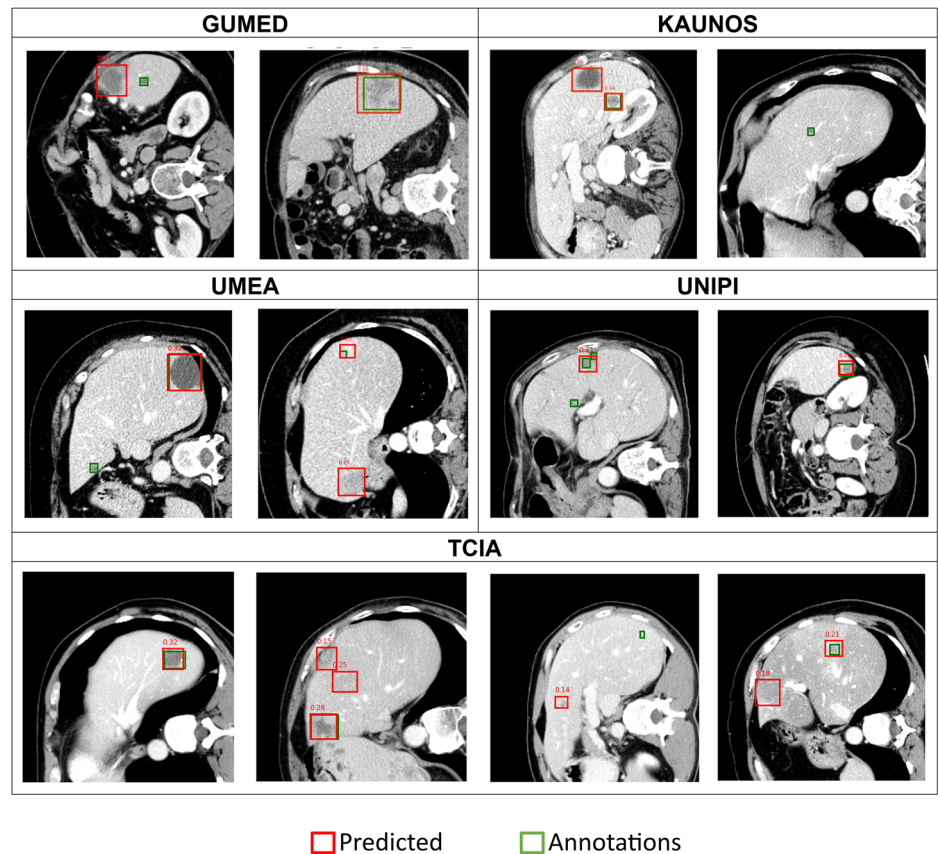


Fig. 7 | Cumulative Radiomics Quality Score (RQS 2.0) progression across Radiomics Readiness Levels (RRL 1–6). The study achieved an RQS 2.0 score of 28/35, corresponding to RRL-6, indicating high methodological rigor and partial clinical readiness according to the RQS 2.0 framework.

quantification, decision-curve analysis, and explainability in a single framework. For example, Tharmaseelan et al.³⁶ reported that radiomics-based classifiers could discriminate colorectal vs pancreatic liver metastases with AUC ~ 0.87, outperforming DenseNet-121 in their task, though their focus was on primary tumor origin rather than general metastasis detection. Luo et al.³⁷ built radiomics signatures for disease-free survival in CRLM and demonstrated the prognostic value of radiomic features in contrast-enhanced CT, although external validation performance was modest. Montagnon et al.³⁸ similarly emphasized CT texture analysis in predicting

chemotherapy response among CRLM lesions, finding that texture features add beyond size/volume but that performance is challenged when lesions are small. Wesdorp et al.²¹ developed deep learning autosegmentation models for liver and CRLM, achieving tumor DSC ~ 0.86 and supporting the feasibility of automated burden assessment. Several studies have further investigated deep learning methodologies for automated segmentation of CRLM^{22–24,39}. For example, convolutional neural networks have been used to detect and segment liver metastases on CT^{22,40}. While these models have shown satisfactory performance for larger lesions, their sensitivity decreased

when dealing with smaller lesions²². Some studies have also utilized 3D-U-Net architectures and generative adversarial networks for the segmentation of hepatic metastases in CT and MRI for applications such as treatment response assessment and tumor burden quantification^{23,24,39}. Compared with these approaches, our pipeline achieved comparable or superior classification performance (AUC 0.89) for a patient-level CRLM classification in a more heterogeneous multi-center cohort, while also explicitly reporting sensitivity (>80%), which is a clinically decisive metric for CRLM management. Detecting CRLM remains a challenge because lesions can be small, heterogeneous (CRLM vs non-CRLM), and sometimes visually similar to benign hepatic findings, with sensitivity decreasing particularly for sub-centimeter lesions, even among expert readers^{22,41}. To address these challenges, our study goes beyond reporting a single classification metric by systematically benchmarking several pretrained medical foundation models and identifying UMedPT as the most transferable backbone for this task. The resulting framework integrates patient-level classification with lesion-level detection while incorporating uncertainty quantification, decision-curve analysis, and Grad-CAM-based visualization to provide interpretable and reliable model outputs. These findings are further supported by the supplementary analysis (Supplementary Fig. 3) stratified by lesion number and diameter that showed an increase in sensitivity with lesion count/diameter, highlighting the challenge of detecting small CRLM lesions on CT. Together, these elements provide additional insights into model performance and clinical usability.

Another key strength of this study is that it combines classification, detection, quantification of uncertainty, and explainability within a single foundation-model framework. Despite being pretrained on heterogeneous medical imaging tasks, the UMedPT backbone demonstrated strong transferability to hepatic metastasis detection, achieving robust discrimination and clinically meaningful sensitivity. In addition, the models were trained on heterogeneous multi-phase CT data across multiple institutions, exposing the network to variations in contrast enhancement patterns and acquisition protocols and thereby improving robustness to real-world imaging variability. Notably, the model showed consistent performance in the various test sets, which indicated that it was resilient to data heterogeneity, which is a typical problem in multi-institutional imaging. It was a large multi-center cohort of 2437 patients across four European centers, across a wide range of scanners and protocols, and thus it is one of the most diverse datasets to have been studied in CRLM.

There were two design choices that were emphasized in our study. First, despite using all the contrast phases available to train the model to make the model more diverse and robust, the model was assessed only on images in portal-venous (PV) phase images. The PV phase is widely recognized as the most informative and consistent phase for identifying liver metastases, as it provides stable parenchymal enhancement and clearer lesion delineation across scanners and institutions^{42,43}. The use of all phases in the training process served as a form of data-augmentation technique, exposing the model to variations in contrast appearance and consequently improving its generalization capabilities, and testing the model on the PV phase allowed maintaining the comparability to standard clinical protocols. Second, the uncertainty-aware analysis provided further insights into model reliability. Following the framework of Alves et al.⁴⁴, we observed that cases classified as “certain” consistently yielded higher AUC and balanced accuracy compared with “uncertain” cases, validating the utility of uncertainty estimates as a triage tool. In clinical workflows, such stratification could support selective human-AI collaboration, where high-confidence cases may be safely automated while uncertain cases prompt expert review. Decision-curve analysis reinforced this interpretation by showing a positive net benefit across a broad range of threshold probabilities (~0.35–0.75), with the maximum benefit observed around a threshold of 0.5. This indicates that incorporating AI-based predictions may be useful in preventing unnecessary procedures without being less sensitive to clinically relevant detection of CRLM through contrast-enhanced CT images. Together with moderate calibration (Brier 0.17; slope 3.2; intercept –0.50), these results highlight the potential of pairing model predictions with uncertainty estimates and

threshold-based decision policies to enable safe and interpretable clinical decision support. Such decision-support tools can help clinicians to prioritize cases with a high likelihood of CRLM for further diagnostic evaluation or treatment planning.

The findings of misclassifications (failure mode analysis) show that lesion size and visual subtlety remain key factors contributing to classification errors, particularly in multi-center datasets with heterogeneous acquisition protocols. These results show that UMedPT provides the most transferable and discriminative representations among the tested FMs, and that fine-tuning with an MLP head yields a practically deployable classifier with controllable operating points, high specificity on mixed cohorts, and stable behavior across centers, CT phases, and datasets, including an all-positive external subset where recall can be emphasized by threshold selection.

At the lesion level, the detection model achieved a sensitivity of around 69.1% across the TCIA_CRLM set. Detection performance was strongly dependent on lesion size, exceeding more than 90% for the largest quartile of lesions (>2–11 cm³) but dropping to ~30% for the smallest quartile (<1 cm³). This pattern aligns with known clinical challenges in identifying small or low-contrast metastases on CT, where even expert readers exhibit reduced sensitivity^{22,41}. Quantitative false-positive rates were assessed only on the TCIA-CRLM dataset, which includes complete lesion annotations, yielding an average of 1.3 false positives per image and 0.54 per lesion. For the EuCanImage cohort, lesion annotations were incomplete, so false positives could not be reliably estimated; however, qualitative inspection indicated similar trends, with most spurious detections occurring near vascular or biliary structures rather than within healthy parenchyma.

Cross-task analysis further revealed that detection precision improved in high-confidence, correctly classified cases. Within this subset, lesion sensitivity increased from 69.1% to 74.1%, while the false-positive rate decreased from 1.30 to 1.22 per image, demonstrating that higher classification confidence corresponded to more accurate lesion localization and fewer spurious detections. This coherence between classification confidence and detection accuracy underscores that uncertainty information is not only a proxy for case-level reliability but also reflects underlying feature consistency across tasks.

Grad-CAM explainability also improved interpretability, showing a high correlation between the heatmaps and annotated lesions in high-confidence cases. In contrast, diffuse or off-target activations were often associated with uncertain predictions, reinforcing the complementarity of uncertainty estimation and visual explainability. These results are in line with a clinically oriented paradigm where foundation models not only give good diagnostic performance but also give calibrated confidence and interpretable outputs that are relied on in decision support.

However, there are a few limitations to this study. Sensitivity for small lesions (first quartile) remains modest. Small lesions were a principal failure mode in detection, reflecting known CT constraints for sub-centimeter disease; false positives clustered around vascular/biliary structures. Second, EuCanImage lesion annotations are partial, precluding reliable lesion-wise sensitivity on that cohort; TCIA_CRLM therefore served as the quantitative reference for detection. While this does not diminish the model’s high specificity at the patient level, it highlights the need for refined post-processing or uncertainty-guided filtering to suppress spurious predictions. Third, in Grad-CAM visualization, off-target attention in lower-confidence cases suggests that the model sometimes leverages non-hepatic cues or background context, which could mislead in certain cases. Finally, although our data encompassed multiple European centers, broader external validation beyond the EuCanImage and TCIA_CRLM cohorts, ideally through prospective evaluation, will be critical to confirm generalizability and clinical applicability before deployment.

Future directions should focus on improving the detection of small lesions, which remain the principal challenge in both clinical and automated CRLM assessment. Potential strategies include higher-resolution or multi-phase contrast CT inputs, specialized small-lesion-aware detection heads, and post-processing guided by uncertainty estimates to suppress spurious

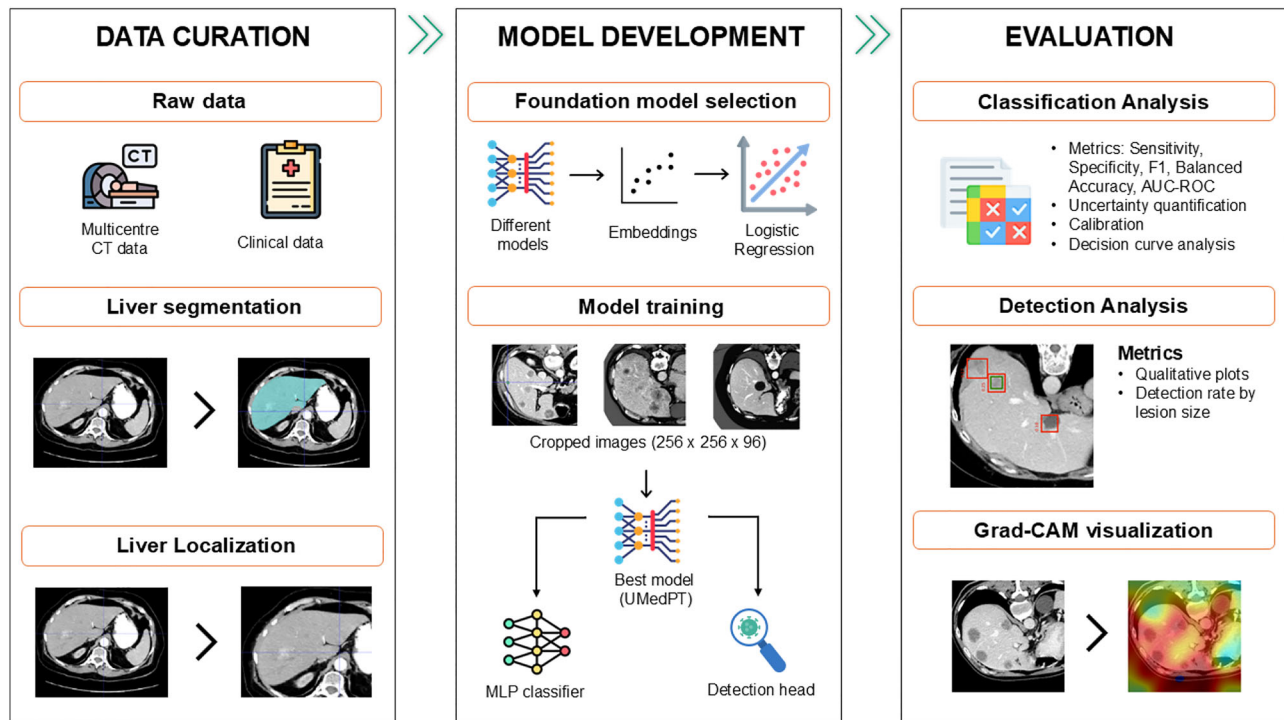


Fig. 8 | Schematic overview of the proposed pipeline. The pipeline consisted of three stages: data curation (multicentre CT and clinical data), model development (foundation-model selection and training classification and detection models), and evaluation (classification metrics, uncertainty quantification, decision-curve

analysis, lesion detection metrics, and Grad-CAM explainability). Preprocessing of images included resampling to $1 \times 1 \times 2.5$ mm, cropping to $256 \times 256 \times 96$ voxels, intensity windowing (-37 to 171 HU), and normalization per scan.

predictions. Additionally, cases could be explored where patients with cardiac dysfunction may show heterogeneous enhancement of the liver during the PV-phase acquisition, potentially increasing the likelihood of false-negative predictions (especially in failure mode analysis). This effect has been indirectly observed in previous studies^{45,46} where early or delayed PVP timing and perfusion variability can significantly reduce lesion detection in CRLM imaging. Integration of clinical and pathological metadata (e.g., tumor markers, genetic profiles, treatment history) with imaging features could further enhance patient-level classification and prognostic modeling. For interpretability, extending beyond Grad-CAM to attention-based or perturbation-driven or spatial-variance-driven explainers may provide higher-fidelity insight into model reasoning. Finally, prospective, multi-center validation and clinical trials will be essential to confirm performance in real-world diagnostic workflows and to quantify the clinical utility of uncertainty-informed AI support. Overall, these directions align with a broader goal of building trustworthy, interpretable, and clinically actionable foundation-model pipelines for CRLM detection and staging on CT.

In conclusion, this study demonstrates that foundation-model-based pipelines, when systematically benchmarked and fine-tuned with appropriate preprocessing and interpretability strategies, can deliver robust and generalizable performance for both patient-level classification and lesion-level detection of colorectal liver metastases in multi-center CT data. The integration of classification, detection, calibration, uncertainty, and explainability within a single unified framework represents a significant advance toward trustworthy AI in oncology imaging. Despite persistent challenges, most notably the limited sensitivity for small or subtle lesions and the occurrence of false positives near vascular structures, the proposed approach achieves clinically relevant recall and interpretability across heterogeneous multicentre cohorts. The inclusion of decision-curve analysis and uncertainty-aware evaluation provides a transparent assessment of clinical benefit and model

reliability, addressing key aspects of translational readiness. Collectively, these results highlight the feasibility of deploying foundation-model-driven methods as a critical step toward clinically applicable, transparent, and reliable AI tools for CRLM detection, staging, and treatment planning in contrast-enhanced CT.

Methods

Figure 8 shows the schematic overview of the pipeline. The workflow consists of three main stages: data curation, model development, and evaluation. Multicentre CT data and associated clinical information were curated within the EuCanImage infrastructure. TotalSegmentator was used to automatically segment the liver region, which was followed by localizing the liver to obtain liver patches. The images and masks were then resampled to the median voxel spacing of $1 \times 1 \times 2.5$ mm, cropped to a fixed bounding box of $256 \times 256 \times 96$ voxels based on the 95th percentile liver dimensions, and windowed between -37 and 171 HU to enhance liver contrast. Then, intensity normalization was applied on a per-scan basis to equalize the intensity values to a standard range.

The foundation-model benchmarking, classification, and detection pipelines, and uncertainty quantification and explainability evaluation were part of model development. Both classification and detection were evaluated quantitatively and qualitatively, and additional evaluations were done to assess uncertainty-aware performance, decision-curve analysis (DCA), and model interpretability using Grad-CAM.

Dataset

This is a retrospective multi-site study that used the EuCanImage project data and publicly available Colorectal Liver Metastases (CRLM) data in the public database on The Cancer Imaging Archive (TCIA)⁴⁷. The retrospective EuCanImage multi-center study was approved by the ethics committees of Comitato Etico Regionale per la Sperimentazione Clinica della Regione Toscana Sezione: AREA VASTA NORD OVEST

(No. EUCANIMAGE Prot n 19292), Ethics Review Authority (No. 2021-03391), KAUNAS REGIONAL ETHICS COMMITTEE FOR BIOMEDICAL RESEARCH (No. 2021-02-04 Nr. BE-2-33) and Independent Bioethics Committee for Scientific Research at The Medical University of Gdansk (No. NKBBN/355/2020) for UNIPI, UMEA, KAUNOS and GUMED respectively. Due to the retrospective nature of this study, informed consent was obtained for the UNIPI cohort, while it was waived by the respective ethics committees for UMEA, KAUNOS, and GUMED. This study was conducted in accordance with the Declaration of Helsinki. EuCanImage is a European initiative that is focused on creating a cancer imaging platform to design and train AI algorithms on a variety of different tumors, such as colorectal cancer, lung cancer, and breast cancer. Imaging and clinical data were curated in the EuCanImage infrastructure that offers standard pipelines to process secure data and, in the meantime, adheres to the European regulations of data protection and privacy.

For the present work, we focused on the colorectal liver metastases (CRLM) use case, addressing the research question: “Can AI identify liver metastases in colorectal cancer from pre- and post-operative contrast-enhanced CT scans?” Patient inclusion criteria were: (i) Availability of contrast-enhanced CT imaging with liver coverage, (ii) Clinical and histopathological confirmation of diagnosis, and (iii) Diagnosis falling into one of three categories: CRLM, non-CRLM liver lesions (e.g., benign cysts, haemangiomas, or non-colorectal malignancies), or normal liver scans (negative). No specific exclusion criteria were applied. The final EuCanImage cohort consisted of 2437 patients. A total of 2437 patients with available CT imaging were included from four European hospitals (Kaunos, GUMED, Umeå, and Unipi). Each patient could contribute one or more CT scans, acquired either pre- or post-operatively as part of routine clinical workflows. Patients were stratified into three categories based on histopathological confirmation and radiology reports: (i) CRLM: colorectal liver metastases, (ii) non-CRLM: other liver lesions (e.g., benign cysts, non-colorectal malignancies), and (iii) negative: no liver lesions. For CRLM cases, up to three liver lesions per patient were annotated by expert radiologists. Since not all lesions were annotated, lesion masks were used only for lesion-level detection tasks, while patient-level binary labels (CRLM vs. non-CRLM/negative) were used for classification.

To improve robustness and evaluate generalizability, we additionally included the TCIA_CRLM collection ($n \approx 197$ patients, portal-venous phase CT) as an independent external dataset. This dataset consists only of CRLM-positive cases with full annotated lesions segmented semi-automatically. Therefore, it makes the data suitable for only lesion-level detection but not for balanced classification analyses.

The dataset was sampled on a per-patient, per-center basis to prevent scan-level leakage, and stratified sampling was done to maintain class balance. The final dataset splits were:

- Training/validation cohort (EuCanImage): 70% ($n = 1704$ patients, 3900 scans)
- Independent hold-out test cohort (EuCanImage): 30% ($n = 733$ patients, 1692 scans)
- External test cohort (TCIA_CRLM): 197 patients (portal-venous phase CT, all CRLM-positive).

The EuCanImage test set was used for patient-level classification (CRLM vs. non-CRLM/negative). The TCIA_CRLM cohort, which contains only the positive-CRLM cases, was used to determine the classification model’s sensitivity to metastatic cases and to support lesion-level detection tasks in a multi-institutional setting. Figure 8 shows the schematic of the study pipeline, including rigorous benchmarking of the FM-based AI methods and classification and detection tasks for CRLM.

Statistical comparisons of patient characteristics across datasets were performed using a χ^2 test for categorical variables (sex) and the Kruskal–Wallis test for continuous variables (age) due to non-normal distribution. Corresponding p -values are reported in Table 1.

Table 5 | Summary table of foundation models

Model	Model type	Architecture	Pretraining strategy	Pretraining dataset
CT-FM	Foundation	SegResEncoder (3D convolutional residual encoder) (~77 M parameters)	Self-supervised (modified SimCLR, intra-sample contrastive)	$\approx 148,000$ CT scans
FMCIB	Foundation	3D ResNet-50 backbone (~184 M parameters)	Self-supervised (contrastive lesion-centered SSL)	$\approx 11,467$ CT lesions from ~2312 patients
Med3D	Supervised	3D ResNet family (ResNet-10/18/34) (~46 M parameters)	Supervised segmentation	3DSeg-8 (8 public segmentation datasets, multi-organ CT/MRI)
Models Genesis	Self-supervised	3D U-Net-like encoder-decoder (~19 M parameters)	Self-supervised (image restoration, inpainting, outpainting, context prediction)	Large-scale unlabeled CT and MRI volumes (incl. 623 chest CTs)
UMedPT	Foundation	Swin Transformer encoder + task-specific heads (~87 M parameters)	Multi-task supervised pretraining across classification, segmentation, detection, and survival prediction	≈ 4 million images across 15 biomedical datasets (CT, MRI, PET, X-ray, WSI)

Data preprocessing

Standardized preprocessing was applied to all CT volumes to provide cross-site consistency. First, liver segmentation masks were automatically obtained with the help of TotalSegmentator⁴⁸, which provided liver volume delineations for all scans. All CT images and masks were resampled to the median voxel spacing of $1 \times 1 \times 2.5$ mm. After resampling, each volume was cropped to a fixed bounding-box set by the 95th percentile liver dimensions in the cohort, resulting in a standardized input size of $256 \times 256 \times 96$ voxels. This cropping strategy made sure that the entire liver volume was always captured and removed irrelevant background. The intensities of voxels were windowed between -37 and 171 HU, which corresponds to the 0.5th–99.5th percentile in the EuCanImage contrast-enhanced CT training set to make liver parenchyma and lesions visible. This percentile-based windowing was estimated from all liver voxels across sites and contrast phases to minimize the effects of outliers and even out the intensity distributions for the liver ROI. The model was trained on all available contrast-enhanced CT phases (arterial, portal venous, and delayed) to provide complementary information for lesion characterization. Intensity normalization was applied on a per-scan basis to harmonize voxel intensity distributions across the dataset.

Foundation-model benchmarking

To identify the most suitable pretrained representation for CRLM classification, we systematically benchmarked five models that differed in pretraining strategy, dataset scale, and architecture. The evaluated models included CT-FM, FMCIB, Med3D, ModelsGenesis, and UMedPT (Table 5).

CT-FM⁴⁹ was trained using a modified SimCLR framework for self-supervised contrastive learning on approximately 148,000 CT scans, where intra-sample contrast was enforced by sampling patches within the same scan to capture spatial-semantic relationships; augmentations included random cropping, histogram shifting, and intensity scaling. Pretraining was performed for 500 epochs (best checkpoint obtained at epoch 449). The model uses a SegResEncoder backbone, a 3D convolutional residual encoder with roughly 77 million parameters, and is intended as a general-purpose CT foundation model for segmentation, classification, and retrieval. Segmentation tasks were optimized with Dice and cross-entropy losses, whereas classification used binary cross-entropy with windowing tailored to CT-specific contrasts (e.g., blood, bone).

FMCIB (Foundation Model for Cancer Imaging Biomarkers)²⁶ was trained on $\approx 11,467$ radiographic lesions from ~ 2312 patients, spanning multiple cancer types. It employed a self-supervised contrastive learning objective, maximizing similarity between different augmentations of the same lesion and minimizing similarity across lesions. Its backbone is a 3D ResNet-50 adapted for volumetric lesion patches. The resulting embeddings are lesion-centric and have been shown to outperform prior baselines (e.g., Med3D, ModelsGenesis) for lesion classification and prognostic prediction, particularly in limited-data settings.

Med3D⁵⁰ was pretrained on the 3DSeg-8 dataset, comprising eight publicly available 3D segmentation datasets covering multiple organs and pathologies. It follows a supervised pretraining strategy. The family of encoders (3D ResNet-10/18/34) was trained for volumetric inputs and provides a strong initialization for downstream tasks, typically accelerating convergence and improving performance in both segmentation and classification (e.g., LiTS liver segmentation, lung nodule detection).

ModelsGenesis⁵¹ introduced self-supervised learning on 3D medical volumes, through image restoration and transformation tasks (e.g., recovering masked patches, outpainting, inpainting, and context restoration). Pretraining was performed on large-scale, unlabeled CT and MRI volumes. The network adopts a 3D U-Net-like encoder-decoder (with residual connections in some implementations), yielding a robust initialization that consistently outperforms training from scratch for segmentation and classification when annotations are limited.

UMedPT⁵² represents a general-purpose medical foundation model trained in a multi-task supervised framework across 15 large-scale datasets,

covering classification, segmentation, detection, and survival prediction and spanning multiple modalities (CT, MRI, PET, X-ray). It employs a hierarchical Swin Transformer encoder with shifted windows coupled to task-specific heads (e.g., segmentation decoders, classification layers). Pretraining at this scale (≈ 4 million images across the constituent datasets) supports broad transferability for both 2D and 3D medical imaging tasks.

All FMs were used as fixed (frozen) feature encoders without additional fine-tuning on the CRLM dataset. For each model, deep feature embeddings were extracted from the pretrained backbone and used as input to a downstream machine learning classifier. This design ensured a fair comparison between models by keeping the downstream classification pipelines identical across all feature representations. The pretrained weights released by the original authors were used for all models. To address class imbalance, we employed the Synthetic Minority Oversampling Technique (SMOTE)⁵³ to generate balanced training folds. Before classification, features were reduced in dimensionality using principal component analysis (PCA), with the number of retained components optimized during model selection. Then, a logistic regression classifier was built to carry out a binary classification task (CRLM vs. non-CRLM/negative).

Hyperparameter optimization was done using Optuna⁵⁴ with a search space that consisted of the number of components for PCA, regularization parameter (C), type of penalty (L1, L2, elastic net), solver, class weights (balanced), maximum iterations (varied to ensure solver convergence), and mixing ratio of elastic net. Each candidate configuration was evaluated using 10-fold stratified cross-validation on the training set, and the mean validation AUC served as the optimization objective.

The best-performing configuration was retrained on the entire training dataset and evaluated on the EuCanImage test set. The foundation model achieving the highest AUC was selected for the downstream development of the classification and detection models

Classification (patient-level)

For the classification task, the selected FM backbone (UMedPT) was used as the feature encoder. The input 3D CT volume was decomposed into a sequence of 2D slices, which were processed independently through the UMedPT encoder-squeezer module. The encoder is based on a Swin Transformer architecture and extracts hierarchical visual representations from each slice. The backbone contains four hierarchical stages with multiple Swin Transformer blocks, producing multi-scale feature maps with channel dimensions increasing from 128 to 1024. These features are compressed by the squeezer module using a 1×1 convolution and adaptive pooling to produce 512-dimensional slice-level embeddings. The resulting embeddings are aggregated by the grouper module using a learnable weighted average pooling mechanism that generates a single 512-dimensional patient-level representation for the entire 3D CT volume. Although slices are encoded individually, the grouper integrates information across all slices to capture contextual information from the volume. The architecture of the UMedPT encoder is described in detail in the original publication⁵². This slice-wise processing strategy follows the UMedPT framework and enables efficient handling of volumetric CT scans with varying depths while being compatible with the pretrained encoder. The network was fine-tuned by unfreezing a few encoder blocks, and an MLP classification head was attached to the extracted features. The MLP accepted a 512-dimensional embedding produced by the UMedPT squeezer-grouper module and consisted of two fully connected layers ($512 \rightarrow 256$ and $256 \rightarrow 128$), each followed by instance normalization, a LeakyReLU activation (negative slope = 0.01), and dropout ($p = 0.2$) for regularization. The final fully connected layer mapped the 128-dimensional representation to a single output logit, corresponding to the binary classification task (CRLM vs. non-CRLM/negative). The model was trained using 5-fold cross-validation on the EuCanImage training set, with patient-level stratification to preserve class balance. Predictions for the EuCanImage test set were obtained by averaging probabilities across the 5 trained folds (ensemble prediction). The inference was done at the patient level. This design

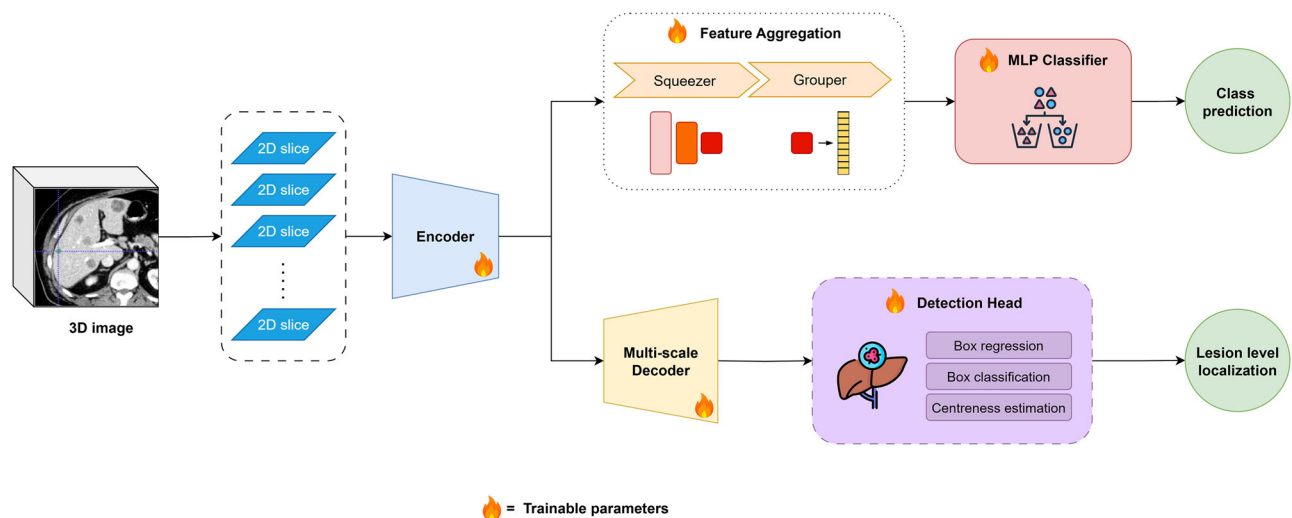


Fig. 9 | Overview of the proposed UMedPT-based CRLM classification and detection framework. The model integrates two complementary pipelines: (1) a classification branch that aggregates deep features extracted from the pretrained UMedPT encoder and squeezer-grouper module and predicts the patient-level

probability of CRLM via an MLP classifier, and (2) a lesion-level detection branch using a multi-scale decoder and FCOS-style detection head for bounding-box regression, classification, and centerness estimation.

provided sufficient non-linear capacity to discriminate between patient classes while incorporating regularization to reduce overfitting. The classification head was trained jointly with the UMedPT encoder (Fig. 9).

Uncertainty quantification and calibration

Uncertainty quantification (UQ) was performed to evaluate the reliability of model predictions, following the approach of Alves et al.⁴⁴. Prediction uncertainty was estimated by calculating the variance across five independently trained UMedPT+MLP folds. The ensemble-averaged probability was considered as the final prediction, and the variance across all 5 folds was used as an uncertainty score per patient (lower variance indicates higher confidence). This uncertainty quantification was used to rank the patients and stratify cohorts into a certain group (CG) and an uncertain group (UG) at different percentile thresholds (10–90%). At each threshold, performance metrics such as area under the ROC curve (AUC) and balanced accuracy were recalculated separately within the CG and UG subsets. Ninety-five percent confidence intervals were estimated using 1000 bootstrap resamples at the patient level.

Calibration was assessed on the EuCanImage test set using calibration curves and metrics, including Brier score, calibration slope, and intercept. Perfect calibration corresponds to slope = 1 and intercept = 0; deviations indicate over- or under-confidence.

Decision-curve analysis

To assess potential clinical utility, decision-curve analysis (DCA) was performed using the ensemble-averaged predictions from the 5-fold UMedPT+MLP model. Net benefit was calculated as:

$$\text{Net benefit} = \frac{TP}{n} - \frac{FP}{n} \times \frac{p}{1-p} \quad (1)$$

Where TP and FP are the number of true and false positives, n is the total number of patients, and p is the threshold probability. Curves were generated across the full range of threshold probabilities and compared against “treat-all” and “treat-none” strategies.

Detection (lesion-level)

For CRLM lesion detection, the UMedPT encoder was extended with a fully convolutional one-stage (FCOS) detection head adapted for volumetric CT data, but in a 2D slice-wise manner (Refer to Fig. 9). A feature pyramid

network (FCOS decoder) generated multi-scale representations from the encoder outputs, and these were passed to the detection head (MTLFCOSHead). The detection head consisted of three parallel convolutional branches for classification, bounding-box regression, and centerness prediction. Each branch was composed of two stacked 3×3 convolutional layers with normalization and non-linear activations. The classification branch predicted lesion probability maps, the regression branch estimated distances to lesion boundaries for bounding-box construction, and the centerness branch penalized off-center predictions to improve localization quality. The model was optimized with a multi-task objective comprising focal loss for classification ($\gamma = 2.0$, $\alpha = 0.25$, weighted by $1/\log(2)$), IoU loss for bounding-box regression (weight = 1.0), and binary cross-entropy for centerness prediction (weight = 0.2). This anchor-free method was able to detect lesions of different sizes effectively. The model was weakly supervised using partially annotated lesions (up to three lesions per CRLM patient). The model was assessed at a lesion level. A lesion was counted as detected when the IoU was more than 0.025. Bounding boxes predicted with an $\text{IoU} \leq 0.025$, relative to all ground-truth lesions, were considered false positives, and false-positive rates (FPR) were reported per image and per ground-truth lesion. In order to explore the impact of the lesion size, lesions were classified into quartiles based on their volume. The volumes were calculated by multiplying the number of voxels and the size of the voxels ($1 \times 1 \times 2.5 \text{ mm}$, $\approx 0.0025 \text{ cm}^3$ per voxel). The distribution of the lesion volumes in the test set was used to define the boundaries of the quartiles, and the rate of detection by each of the quartiles (Q1–Q4) was reported separately. This stratification enabled the determination of the sensitivity of lesion size in the TCIA-CRLM cohort, since it was a fully annotated dataset.

Explainability

In order to achieve interpretability, we utilized Gradient-weighted Class Activation Mapping (Grad-CAM) for the classification pipeline. Grad-CAM is an explainability method that generates visual explanations of convolutional neural networks, based on the gradient of target class scores relative to the final convolutional feature maps of the network. This results in heatmaps that highlight the regions of the input most influential in the model’s decision-making process⁵⁵. In our implementation, Grad-CAM was applied to the UMedPT encoder during inference for classification. Heatmaps were generated from the ‘norm1’ layer of the first Swin-TransformerBlockV2 in the seventh encoder block. To provide robust and

stable interpretability, we generated Grad-CAM maps for each of the five cross-validation folds and then computed a consensus heatmap by averaging the fold-level heatmaps. These consensus maps were overlaid onto the original CT slices, enabling visualization of the anatomical regions that contributed most strongly to predictions of CRLM versus non-CRLM/negative.

Training

All models were implemented in PyTorch with MONAI for data loading and augmentation and trained on NVIDIA H100 GPUs using mixed precision (AMP) to accelerate training and reduce memory consumption.

Training workflows for both the classification and detection pipelines are summarized in Fig. 9. For the classification model, training followed a 5-fold cross-validation design with patient-level stratification. Each fold was optimized with focal binary cross-entropy loss ($\gamma = 3.0$, $\alpha = 0.7$) to address class imbalance. The optimizer was AdamW with a learning rate of $5e-4$. Training was run for a maximum of 1000 epochs, with early stopping if no improvement in validation loss was observed for 10 epochs. The batch size was set to 12. During training, performance was monitored using accuracy, AUC, precision, recall, and F1-score, and the best checkpoint was saved based on validation loss. Final test set predictions were obtained by averaging probabilities across the 5 folds (ensemble prediction).

For the detection model, training was based on the FCOS anchor-free detection framework. The detection head was optimized with a multi-task loss comprising focal loss for classification ($\gamma = 2.0$, $\alpha = 0.25$, weighted by $1/\log(2)$), IoU loss for bounding-box regression (weight = 1.0), and binary cross-entropy for centeredness prediction (weight = 0.2). The optimizer was AdamW with a learning rate of $1e-4$, trained for up to 1000 epochs with early stopping (patience = 10). The batch size was 12, reflecting the higher memory demand for the detection module.

Loss functions

For classification, the primary objective was focal loss⁵⁶ to address class imbalance. Unlike BCE, focal loss introduces a modulating factor that reduces the relative loss for well-classified examples and focuses training on hard or misclassified cases. The loss is defined as:

$$L_{\text{Focal}} = -\alpha_t(1 - p_t)^\gamma \log(p_t) \quad (2)$$

where $p = \sigma(\hat{y})$. Here, $\alpha \in [0, 1]$ balances positive and negative examples while $\gamma \geq 0$ adjusts the focus on misclassified cases.

For detection, the training objective followed the anchor-free FCOS framework, combining three components:

- Classification loss: focal loss, encouraging correct detection of lesion voxels while handling severe class imbalance between lesion and background.
- Bounding-box regression loss: IoU loss⁵⁷, which directly optimizes the overlap between predicted and ground-truth bounding boxes. It is defined as:

$$L_{\text{IoU}} = 1 - \frac{|B_p \cap B_{gt}|}{|B_p \cup B_{gt}|} \quad (3)$$

Where B_p and B_{gt} are predicted and ground-truth bounding boxes.

Centreness loss: Binary cross-entropy loss, encouraging predictions close to lesion centers.

$$L_{\text{BCE}} = -[y_i \log \sigma(\hat{y}_i) + (1 - y_i) \log(1 - \sigma(\hat{y}_i))] \quad (4)$$

Where $y_i \in \{0, 1\}$ is the ground-truth label, $\hat{y}_i$ the predicted logit, and σ the sigmoid activation.

Evaluation metrics

For classification, metrics like balanced accuracy, area under the receiver operating characteristic curve (AUC), sensitivity, specificity, precision, and F1-score were reported.

- Sensitivity (Recall): the proportion of CRLM patients correctly identified,

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (5)$$

Where TP and FN are true positives and false negatives.

- Specificity: the proportion of non-CRLM/negative patients correctly identified,

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (6)$$

Where TN and FP are true negatives and false positives.

- Balanced accuracy: the average of sensitivity and specificity, which accounts for class imbalance,

$$\text{Balanced Accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2} \quad (7)$$

- Area Under the Receiver Operating Characteristic Curve (ROC AUC): The ROC AUC measures the model's ability to discriminate between CRLM and non-CRLM/negative patients and across all possible classification thresholds. It is calculated as the area under the ROC curve by plotting the true positive rate (sensitivity) against the false-positive rate ($1 - \text{specificity}$).
- Precision (Positive Predictive Value): the proportion of patients with predicted CRLM that were correctly classified,

$$\text{Precision} = \frac{TP}{TP + FP} \quad (8)$$

- F1-score: the harmonic mean of precision and recall,

$$F1 = 2 * \frac{\text{Precision} * \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}} \quad (9)$$

For lesion detection, performance was assessed using appropriate metrics for bounding-box regression:

- Detection rate: the rate that defines the fraction of annotated lesions that were detected with IoU above a given threshold,

$$\text{Detection Rate} = \frac{\# \text{ Lesions detected at } \text{IoU} \geq \text{threshold}}{\# \text{ Total ground truth lesions}} \quad (10)$$

- False-positive rate: The number of predicted bounding boxes that do not overlap with any ground-truth lesions ($\text{IoU} \leq 0.025$), expressed either per image or per lesion,

$$\text{FPR}_{\text{image}} = \frac{\# \text{ Total predicted lesions} - \# \text{ Total ground truth lesions}}{\# \text{ Total images}} \quad (11)$$

$$\text{FPR}_{\text{lesion}} = \frac{\# \text{ Total predicted lesions} - \# \text{ Total ground truth lesions}}{\# \text{ Total ground truth lesions}} \quad (12)$$

- Size-stratified detection performance: lesions were grouped by quartiles of lesion volume (in voxels), and detection rates were computed within each quartile (Q1, Q2, Q3, Q4) to evaluate sensitivity to lesion size.

RQS 2.0 assessment

The methodological quality of the proposed pipeline was assessed with the help of the Radiomics Quality Score (RQS 2.0) framework³⁴. All of the RQS requirements, including data preparation, model development, validation, and trustworthiness, were evaluated using the official scoring requirements. The cumulative score was mapped to the corresponding Radiomics Readiness Level (RRL) to quantify methodological maturity. Detailed scoring criteria and evidence mapping are provided in Supplementary Table 2.

Data availability

The EuCanImage dataset used in this study is part of an ongoing project and is not yet publicly available. Upon project completion, the data will be made available under controlled access via the European Genome-Phenome Archive (clinical data) and EuroBioImaging (imaging data). The TCIA_CRLM dataset is publicly available from The Cancer Imaging Archive (TCIA). <https://www.cancerimagingarchive.net/collection/colorectal-liver-metastases/>.

Code availability

The code and data analysis scripts used in this study are publicly available at the following GitHub repository: <https://github.com/shruti26mali/Multicentre-CRLM-Classification>.

Received: 12 January 2026; Accepted: 5 May 2026;

Published online: 27 May 2026

References

- Colorectal cancer. <https://www.who.int/news-room/fact-sheets/detail/colorectal-cancer>. Accessed 11 Sep 2025.
- Engstrand, J., Nilsson, H., Strömberg, C., Jonas, E. & Freedman, J. Colorectal cancer liver metastases - a population-based study on incidence, management and survival. *BMC Cancer* <https://doi.org/10.1186/s12885-017-3925-x> (2018).
- Martin, J. et al. Colorectal liver metastases: current management and future perspectives. *World J. Clin. Oncol.* **11**, 761 (2020).
- Aykut, B. & Lidsky, M. E. Colorectal cancer liver metastases: multimodal therapy. *Surg. Oncol. Clin. North Am.* <https://doi.org/10.1016/j.soc.2022.07.009> (2023).
- Nadarevic, T. et al. Computed tomography for the diagnosis of hepatocellular carcinoma in adults with chronic liver disease. *Cochrane Database Syst. Rev.* <https://doi.org/10.1002/14651858.CD013362.pub2> (2021).
- Mahboobi, H., Hemmasi, G. R., Aghajani, A., Sedighi, E. & Babaei, M. R. Diagnostic value of CT scan compared with ultrasound-guided needle biopsy in the diagnosis of liver masses; a retrospective diagnostic study. *Immunopathol. Persa* **11**, 42733–42733 (2025).
- Matteini, F. et al. Benign and malignant focal liver lesions displaying rim arterial phase hyperenhancement on CT and MRI. *Insights Imaging* **15**, 1–17 (2024).
- Heiken, J. P. Distinguishing benign from malignant liver tumours. *Cancer Imaging* **7**, S1 (2007).
- Lakshmipriya, B., Pottakkat, B. & Ramkumar, G. Deep learning techniques in liver tumour diagnosis using CT and MR imaging - a systematic review. *Artif. Intell. Med.* **141**, 102557 (2023).
- Nayantara, P., Kamath, S., Kn, M. & Kadavigere, R. Computer-aided diagnosis system for classifying the liver lesions from multiphase CT images. *Sens. Imaging* **26**, 1–25 (2025).
- Lambin, P. et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nat. Rev. Clin. Oncol.* <https://doi.org/10.1038/nrclinonc.2017.141> (2017).
- Gillies, R. J., Kinahan, P. E. & Hricak, H. Radiomics: images are more than pictures, they are data. *Radiology* <https://doi.org/10.1148/radiol.2015151169> (2015).
- Fiz, F. et al. Radiomics of liver metastases: a systematic review. *Cancers* **12**, 2881 (2020).
- Karagkounis, G. et al. Computed tomography-based radiomics with machine learning outperforms radiologist assessment in estimating colorectal liver metastases pathologic response after chemotherapy. *Ann. Surg. Oncol.* <https://doi.org/10.1245/s10434-024-15373-y> (2024).
- Da-ano, R. et al. Performance comparison of modified ComBat for harmonization of radiomic features for multicenter studies. *Sci. Rep.* **10**, 1–12 (2020).
- Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* <https://doi.org/10.1145/3065386> (2017).
- Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015 234–241 (2015).
- Litjens, G. et al. A survey on deep learning in medical image analysis. *Med. Image Anal.* <https://doi.org/10.1016/j.media.2017.07.005> (2017).
- Shen, D., Wu, G. & Suk, H.-I. Deep learning in medical image analysis. *Annu. Rev. Biomed. Eng.* **19**, 221 (2017).
- Yan, X. et al. A multimodal MRI-based model for colorectal liver metastasis prediction: integrating radiomics, deep learning, and clinical features with SHAP interpretation. *Curr. Oncol.* <https://doi.org/10.3390/curroncol32080431> (2025).
- Wesdorp, N. J. et al. Deep learning models for automatic tumor segmentation and total tumor volume assessment in patients with colorectal liver metastases. *Eur. Radiol. Exp.* <https://doi.org/10.1186/s41747-023-00383-4> (2023).
- Vorontsov, E. et al. Deep learning for automated segmentation of liver lesions at CT in patients with colorectal cancer liver metastases. *Radiol. Artif. Intell.* <https://doi.org/10.1148/ryai.2019180014> (2019).
- Liu, X. et al. Automatic segmentation of hepatic metastases on DWI images based on a deep learning method: assessment of tumor treatment response according to the RECIST 1.1 criteria. *BMC Cancer* <https://doi.org/10.1186/s12885-022-10366-0> (2022).
- Bereska, J. I. et al. Development and external evaluation of a self-learning auto-segmentation model for Colorectal Cancer Liver Metastases Assessment (COALA). *Insights Imaging* **15**, 279 (2024).
- Rockenschaub, P. et al. From single-hospital to multi-centre applications: enhancing the generalisability of deep learning models for adverse event prediction in the ICU. Preprint at <http://arxiv.org/abs/2303.15354> (2023).
- Pai, S. et al. Foundation model for cancer imaging biomarkers. *Nat. Mach. Intell.* **6**, 354–367 (2024).
- Kirillov, A. et al. Segment anything. Preprint at <http://arxiv.org/abs/2304.02643> (2023).
- Wei, M., Chen, S., Wu, S. & Xu, D. Rep-MedSAM: towards real-time and universal medical image segmentation. https://doi.org/10.1007/978-3-031-81854-7_4. Accessed 11 Sep 2025.
- D'Antonoli, T. A. et al. Foundation models for radiology: fundamentals, applications, opportunities, challenges, risks, and prospects. *Diagn. Interv. Radiol.* <https://doi.org/10.4274/dir.2025.253445> (2026).
- Wadden, J. J. Defining the undefinable: the black box problem in healthcare artificial intelligence. *J. Med. Ethics* **48**, 764–768 (2022).
- Madan Y. The pitfalls of black box AI in medicine. *McMaster Brain Res. Soc.* <https://journals.mcmaster.ca/macbrs/article/view/3498> (2023).
- Salahuddin, Z., Woodruff, H. C., Chatterjee, A. & Lambin, P. Transparency of deep neural networks for medical image analysis: a review of interpretability methods. *Comput. Biol. Med.* **140**, 105111 (2022).

33. Muhammad, D. & Bendechache, M. Unveiling the black box: a systematic review of explainable artificial intelligence in medical image analysis. *Comput. Struct. Biotechnol. J.* **24**, 542–560 (2024).
34. Lambin, P. et al. Radiomics Quality Score 2.0: towards radiomics readiness levels and clinical translation for personalized medicine. *Nat. Rev. Clin. Oncol.* **22**, 831–846 (2025).
35. Rompianesi, G., Pegoraro, F., Ceresa, C. D., Montalti, R. & Troisi, R. I. Artificial intelligence in the diagnosis and management of colorectal cancer liver metastases. *World J. Gastroenterol.* **28**, 108–122 (2022).
36. Tharmaseelan, H. et al. Tumor classification of gastrointestinal liver metastases using CT-based radiomics and deep learning. *Cancer Imaging* **23**, 95 (2023).
37. Luo, X. et al. Prognostication of colorectal cancer liver metastasis by CE-based radiomics and machine learning. *Transl. Oncol.* **47**, 101997 (2024).
38. Montagnon, E. et al. Radiomics analysis of baseline computed tomography to predict oncological outcomes in patients treated for resectable colorectal cancer liver metastasis. *PLoS ONE* **19**, e0307815 (2024).
39. Xia, S.-J. et al. CRLM-GAN: a feature-constrained GAN-based deep learning framework for multi-parametric MRI-based segmentation of colorectal liver metastases before and after chemotherapy. *Cancer Imaging* **26**, 10 (2025).
40. Anderson, B. M. et al. Automated segmentation of colorectal liver metastasis and liver ablation on contrast-enhanced CT images. *Front. Oncol.* **12**, 886517 (2022).
41. van Kessel, C. S. et al. Preoperative imaging of colorectal liver metastases after neoadjuvant chemotherapy: a meta-analysis. *Ann. Surg. Oncol.* **19**, 2805–2813 (2012).
42. Paulatto, L. et al. Colorectal liver metastases: radiopathological correlation. *Insights Imaging* **11**, 1–19 (2020).
43. Scott, D. J. et al. Dual phase helical CT versus portal venous phase CT for the detection of colorectal liver metastases: correlation with intra-operative sonography, surgical and pathological findings. *Clin. Radiol.* <https://doi.org/10.1053/crad.2000.0668> (2001).
44. Alves, N. et al. Prediction variability to identify reduced AI performance in cancer diagnosis at MRI and CT. *Radiology* **308**, e230275 (2023).
45. Dercle, L. et al. Impact of variability in portal venous phase acquisition timing in tumor density measurement and treatment response assessment: metastatic colorectal cancer as a paradigm. *JCO Clin. Cancer Informat.* <https://doi.org/10.1200/CCI.17.00108> (2017).
46. Detsky, J. S. et al. Perfusion imaging of colorectal liver metastases treated with bevacizumab and stereotactic body radiotherapy. *Phys. Imaging Radiat. Oncol.* **5**, 9 (2018).
47. Simpson, A. L. et al. Preoperative CT and survival data for patients undergoing resection of Colorectal Liver Metastases (Colorectal-Liver-Metastases). *Cancer Imaging Arch.* <https://doi.org/10.7937/QXK2-QG03> (2023).
48. Wasserthal, J. et al. TotalSegmentator: robust segmentation of 104 anatomical structures in CT images. *Radiol. Artif. Intell.* <https://doi.org/10.1148/ryai.230024> (2022).
49. Pai, S. et al. Vision foundation models for computed tomography. Preprint at <http://arxiv.org/abs/2501.09001> (2025).
50. Chen, S., Ma, K. & Zheng, Y. Med3D: transfer learning for 3D medical image analysis. Preprint at <http://arxiv.org/abs/1904.00625> (2019).
51. Zhou, Z. et al. Models genesis: generic autodidactic models for 3D medical image analysis. *Med. Image Comput. Comput. Assist. Inter.* **11767**, 384–393 (2019).
52. Schäfer, R. et al. Overcoming data scarcity in biomedical imaging with a foundational multi-task model. *Nat. Comput. Sci.* **4**, 495–509 (2024).
53. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002).
54. Optuna. <https://doi.org/10.1145/3292500.3330701>. Accessed 12 Sep 2025.
55. Selvaraju, R. R. et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)* <https://doi.org/10.1109/iccv.2017.74> (IEEE, 2017).
56. Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal loss for dense object detection. Preprint at <http://arxiv.org/abs/1708.02002> (2017).
57. UnitBox. <https://doi.org/10.1145/2964284.2967274>. Accessed 11 Sep 2025.

Acknowledgements

Authors acknowledge financial support from ERC advanced grant (ERC-ADG-2015 n° 694812 - Hypoximmuno), ERC-2020-PoC: 957565-AUTO.DISTINCT. Authors also acknowledge financial support from the European Union's Horizon research and innovation programme under grant agreement: EuCanImage n° 952103 (main contributor), ImmunoSABR n° 733008, CHAIMELEON n° 952172, TRANSCAN Joint Transnational Call 2016 (JTC2016 CLEARLY n° UM 2017-8295), IMI-OPTIMA n° 101034347, AIDAVA (HORIZON-HLTH-2021-TOOL-06) n° 101057062, REALM (HORIZON-HLTH-2022-TOOL-11) n° 101095435, RADIOVAL (HORIZON-HLTH-2021-DISEASE-04-04) n° 101057699 and EUCAIM (DIGITAL-2022-CLOUD-AI-02) n° 101100633.

Author contributions

Conceptualization: S.A.M. and Z.S.; methodology: S.A.M. and Z.S.; software: S.A.M. and Z.S.; validation: S.A.M. and Z.S.; formal analysis: S.A.M. and Z.S.; investigation: S.A.M. and Z.S.; writing - original draft preparation: S.A.M. and Z.S.; writing - review and editing: S.A.M., Z.S., Y.Z., A.A., X.Z., H.C.W., M.B., K.R., J.K., L.F., R.F., R.L.M. and P.L.; supervision: Z.S., H.C.W. and P.L.; funding acquisition: P.L. All authors have read and agreed to the published version of the manuscript.

Competing interests

P.L.: none related to the current manuscript; outside of current manuscript: grants/sponsored research agreements from Radiomics SA, Convert Pharmaceuticals, and LivingMed Biotech. He received a presenter fee (in cash or in kind) and/or reimbursement of travel costs/ consultancy fee (in cash or in kind) from Radiomics SA, BHV & Roche. P.L. has shares in the companies Radiomics SA, Convert Pharmaceuticals, Comunicare, LivingMed Biotech, BHV, and Bactam. P.L. is co-inventor of two issued patents with royalties on radiomics (PCT/NL2014/050248 and PCT/NL2014/050728), licensed to Radiomics SA; one issued patent on mtDNA (PCT/EP2014/059089), licensed to ptTheragnostic/DNAmito; one non-issued patent on LSRT (PCT/ P126537PC00, US: 17802766), licensed to Varian; three non-patented inventions (softwares) licensed to ptTheragnostic/DNAmito, Radiomics SA and Health Innovation Ventures and two non-issued, non-licensed patents on Deep Learning-Radiomics (N2024482, N2024889). He confirms that none of the above entities was involved in the preparation of this paper. The other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41698-026-01482-2>.

Correspondence and requests for materials should be addressed to Shruti Atul Mali.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2026

On behalf of EUCanImage working group

Karim Lekadir¹¹, Lidia Garrucho¹¹, Philippe Lambin^{1,4}, Henry Woodruff^{1,4}, Martijn Starmans^{12,13}, Josep Lluís Gelpi¹⁴, Salvador Capella-Gutierrez¹⁴, Jordi Rambla¹⁵, Fred Prior¹⁶, Melanie Goisauf¹⁷, Pilar Nicolas¹⁸, Davide Zaccagnini¹⁹, Pär Kragsterman²⁰, Samuele Ghezzi²¹, Jeroen Cant²¹, Tobias Heimann²², Peter Gordebeke²³, Emanuele Neri²⁴, Jane Smith²⁵, Lorenzo Faggioni²⁶, Jordi Rimola²⁷, Xavier Bargalló²⁸, Katrine Riklund^{6,7}, Maciej Bobowicz⁵, Juozas Kupčinskas⁸ & Javier del Riego²⁹

¹¹Artificial Intelligence in Medicine Lab (BCN-AIM), Department of Mathematics and Computer Science, University of Barcelona, Barcelona, Spain. ¹²Department of Radiology & Nuclear Medicine, Erasmus University Medical Center, Rotterdam, the Netherlands. ¹³Department of Pathology, Erasmus University Medical Center, Rotterdam, the Netherlands. ¹⁴Barcelona Supercomputing Center (BSC), Barcelona, Spain. ¹⁵Computational Biology and Health Genomics Programme, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain. ¹⁶Department of Biomedical Informatics, University of Arkansas for Medical Sciences, Little Rock, AR, USA. ¹⁷Department of ELSI Services and Research, BBMRI-ERIC, Graz, Austria. ¹⁸Faculty of Law, University of the Basque Country, Leioa, Spain. ¹⁹Lynkeus (LYN), Rome, Italy. ²⁰Collective Minds Radiology AB (CMRAD), Täby, Sweden. ²¹Research&Development, Radiomics.bio, Liège, Belgium. ²²Siemens Healthcare GmbH (SIE), Erlangen, Germany. ²³EIBIR Gemeinnützige GmbH zur Förderung der Erforschung der Biomedizinischen Bildgebung (EIBIR), Wien, Austria. ²⁴European Society of Oncologic Imaging—ESOI Europäische Gesellschaft für Onkologische Bildgebung (ESOI), Vienna, Austria. ²⁵European Association for Cancer Research (EACR), Nottingham, UK. ²⁶Academic Radiology, Department of Translational Research, University of Pisa, Pisa, Italy. ²⁷Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain. ²⁸Department of Radiology, Hospital Clínic of Barcelona, Barcelona, Spain. ²⁹Institut d'Investigació i Innovació Parc Taulí I3PT, Universitat Autònoma de Barcelona, Barcelona, Spain.