

**Table S1.** Reporting Guidelines: from Image processing to features calculation steps.

Area	Topic	Description
Patient	Region of interest	Liver metastases
Acquisition	Acquisition protocol	Different acquisition protocols (2 centers)
	Scanner type	Center A: Siemens, Somaton Definition/ Sensation 64 Center B: Siemens Somaton Definition FLASH
	Imaging modality	TC
	Scan duration	Approximately 10/15 minutes
Image registration	Registration method	NONE
Data conversion	Not applicable	NONE
Post-acquisition processing	Anti-aliasing	NONE
	Non-uniformity correction	NONE
	Intensity normalization	NONE
Segmentation	Method	Manual segmentation performed on portal-phase CT scan
	Conversion to mask	NIFTI
Image Interpolation	Interpolation method	NONE
Mask Interpolation	Interpolation method	NONE
Re-segmentation	Method	Between 1 <sup>st</sup> and 99 <sup>th</sup> percentile
Discretization	Method	Fixed bin number (32 bin)
Image transformation	Image filter	NONE
Image biomarker computation	Biomarker set	Shape-based, First-order statistics, GLCM, GLRLM, GLSZM, NGTDM, GLDM
	IBSI compliance	Yes
	Software availability	Pyradiomics
Image biomarker computation - texture parameters	Texture matrix aggregation	2D averaged
	Distance weighting	No weighting
	CM symmetry	Symmetric co-occurrence matrices
	CM distance	1
List of features	First-order statistics	Energy total energy

	entropy minimum 10 <sup>th</sup> 90 <sup>th</sup> Maximum Mean Median interquartile range range mean absolute deviation robust mean absolute deviation root mean squared skewness kurtosis variance uniformity
Shape-based	Elongation Flatness Least Axis Length Major Axis Length Maximum 2D Diameter Column Maximum 2D Diameter Row Maximum 2D Diameter Slice Maximum3DDiameter Mesh Volume Minor Axis Length Sphericity Surface Area Surface Volume Ratio Voxel Volume
GLCM	Autocorrelation Cluster Prominence Cluster Shade Cluster Tendency Contrast Correlation Difference Average Difference Entropy Difference Variance Inverse Difference Normalized Inverse Difference Inverse Difference Moment Normalized Inverse Difference Moment Informational Measure of Correlation 1 Informational Measure of Correlation 2 Inverse Variance Joint Average Joint Energy Joint Entropy Maximal Correlation Coefficient Maximum Probability Sum Average Sum Entropy Sum Squares
GLRLM	Gray Level Non-Uniformity Gray Level Non-Uniformity Normalized Gray Level Variance High Gray Level Run Emphasis Long Run Emphasis

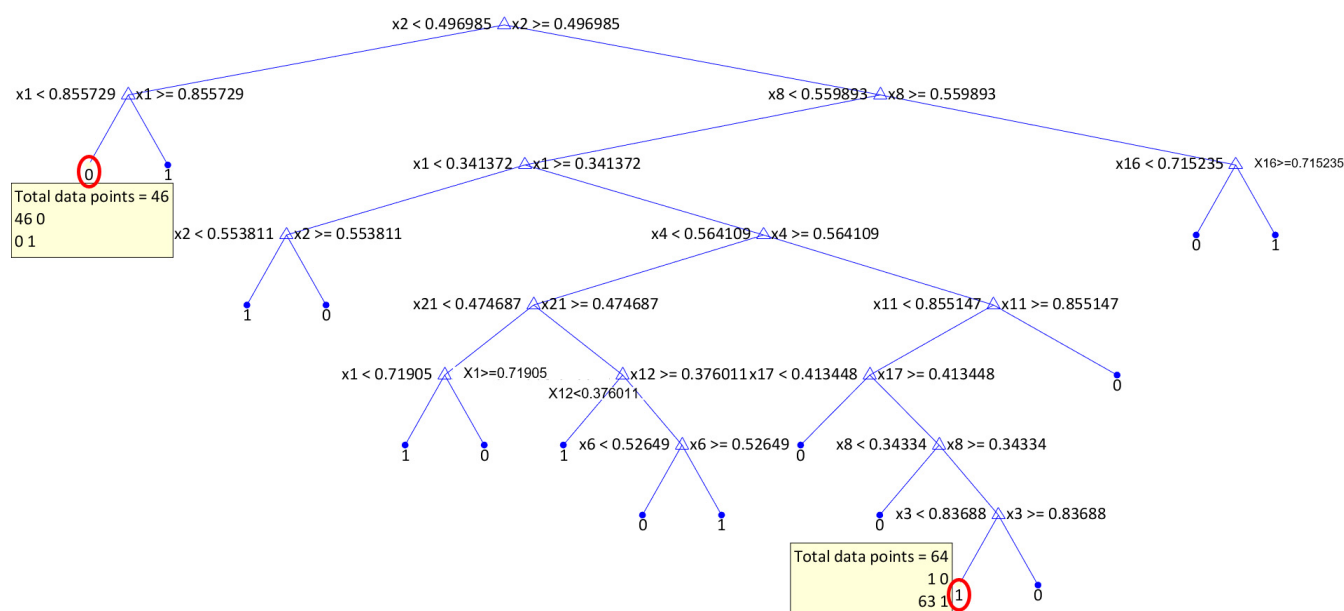
	Long Run High Gray Level Emphasis
	Long Run Low Gray Level Emphasis
	Low Gray Level Run Emphasis
	Run Entropy
	Run Length Non-Uniformity
	Run Length Non-Uniformity Normalized
	Run Percentage
	Run Variance
	Short Run Emphasis
GLSZM	Gray Level Non-Uniformity
	Gray Level Non-Uniformity Normalized
	Gray Level Variance
	High Gray Level Zone Emphasis
	Large Area Emphasis
	Large Area High Gray Level Emphasis
	Large Area Low Gray Level Emphasis
	Low Gray Level Zone Emphasis
	Size Zone Non-Uniformity
	Size Zone Non-Uniformity Normalized
	Small Area Emphasis
	Small Area High Gray Level Emphasis
	Small Area Low Gray Level Emphasis
	Zone Entropy
NGTDM	Busyness
	Coarseness
	Complexity
	Contrast
	Strength
GLDM	Dependence Entropy
	Dependence Non-Uniformity
	Dependence Non-Uniformity Normalized
	Dependence Variance
	Gray Level Non-Uniformity
	Gray Level Variance
	High Gray Level Emphasis
	Large Dependence Emphasis
	Large Dependence High Gray Level Emphasis
	Large Dependence Low Gray Level Emphasis
	Low Gray Level Emphasis
	Small Dependence Emphasis
	Small Dependence High Gray Level Emphasis
	Small Dependence Low Gray Level Emphasis

Table S2. Performances of other ML techniques on both training and validation sets.

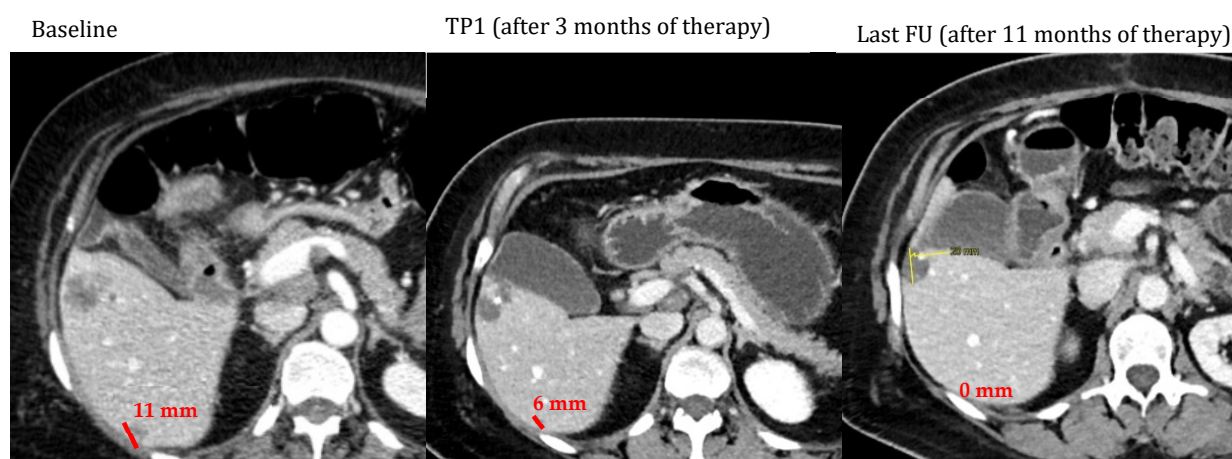
	Train					Validation				
	ACC % (95% CI)	SE % (95% CI)	SP % (95% CI)	PPV % (95% CI)	NPV % (95% CI)	ACC % (95% CI)	SE % (95% CI)	SP % (95% CI)	PPV % (95% CI)	NPV % (95% CI)
Decision Tree	97 (89-100)	99 (94-99)	94 (85-98)	95 (89-98)	99 (91-100)	86 (81-92)	85 (68-95)	92 (78-98)	90 (76-96)	87 (75-94)
LR Step-wise binomial	81.5 (77-84)	79 (73-81)	84 (73-85)	86 (72-90)	76 (73-80)	63 (55-69)	61 (50-65)	65 (51-70)	61 (50-65)	65 (50-65)
LR Step-wise poisson	78.5 (73-81)	79 (73-81)	78 (72-82)	81 (74-83)	74 (71-80)	77.5 (69-84)	82 (70-85)	73 (68-79)	73 (69-79)	82 (69-85)
SVM Linear	79	84	73.5	80	61	67.5	67.5	67.5	62.5	67.5

	(71-83)	(72-85)	(71-81)	(71-84)	(56-67)	(58-75)	(58-71)	(61-75)	(55-68)	(61-75)
Random	93	94	92	94	92	75	64	86	80	72
Forest	(88-96)	(84-96)	(83-94)	(86-97)	(81-99)	(67-81)	(55-71)	(71-89)	(71-89)	(65-83)

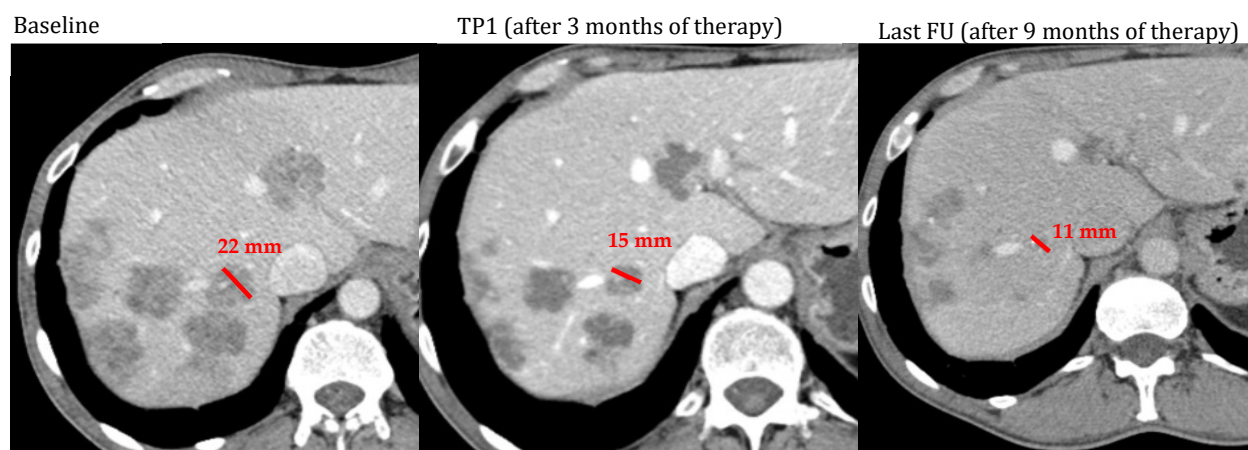
LR: logistic regression, ACC: accuracy, SE: sensitivity, SP: specificity, PPV: positive predictive value, NPV: negative predictive value.



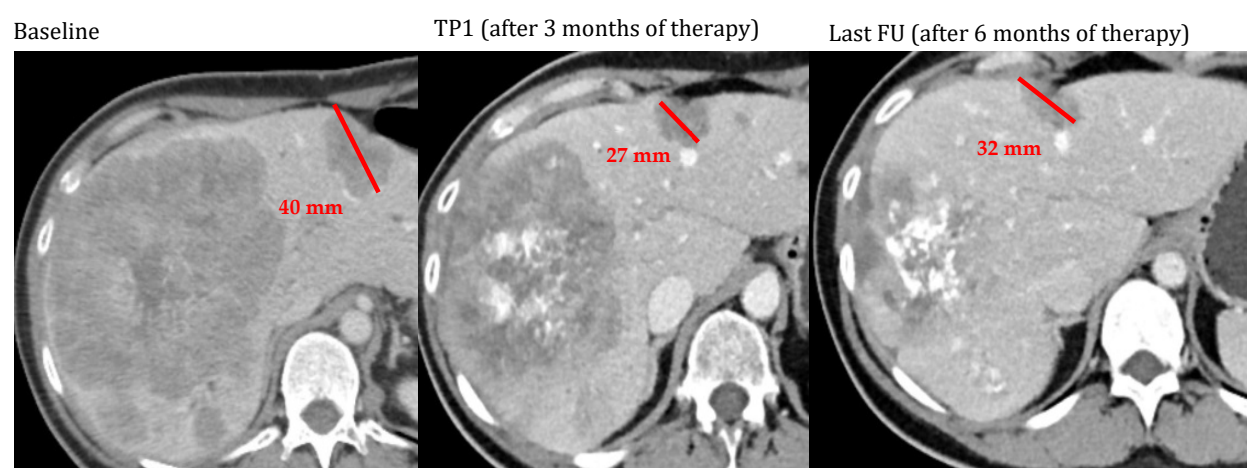
**Figure S1.** Trained Decision Tree containing the 11 selected features: shape Sphericity (named x1 in Figure 2), shape Surface Volume Ratio (x2), GLCM Contrast (x3), GLCM Difference Average (x4), GLCM Difference Variance (x6), GLCM Maximum Probability (x8), GLDM Small Dependence Low Gray Level Emphasis (x11), GLRLM Run Length Non Uniformity Normalized (x12), GLSZM Size Zone Non Uniformity Normalized (x16), GLSZM Small Area Emphasis (x17), and NGTDM Complexity (x21). 46 non responder lesions in the training set (27%) were correctly classified using only 2 variables (shape sphericity and shape surface volume ratio). In particular, if the value of both features was lower than the threshold computed by the model, these lmCRC were classified as non-responder. Conversely, to classify a lesion as R+ more features were needed, as it is visible from the last leaf of the DT, that classifies 63 lesions as R+.



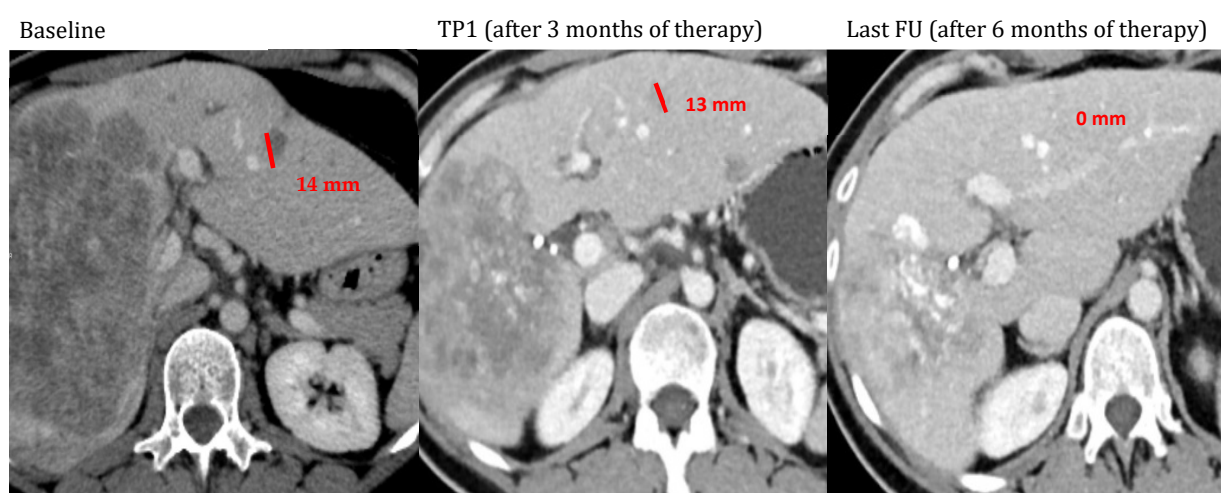
**Figure S2.** Lesion 1 in patient 1010 at baseline, TP1 and last FU during first-line chemotherapy. Lesion 1 was a good responder misclassified as R- by the algorithm (CR at 11 months). Probably, small size could have hindered a correct segmentation in this case.



**Figure S3.** Lesion 4 in patient 1016 at baseline, TP1 and last FU during first-line chemotherapy. This metastasis was a good responder misclassified as R- by the algorithm (PR at 9 months).



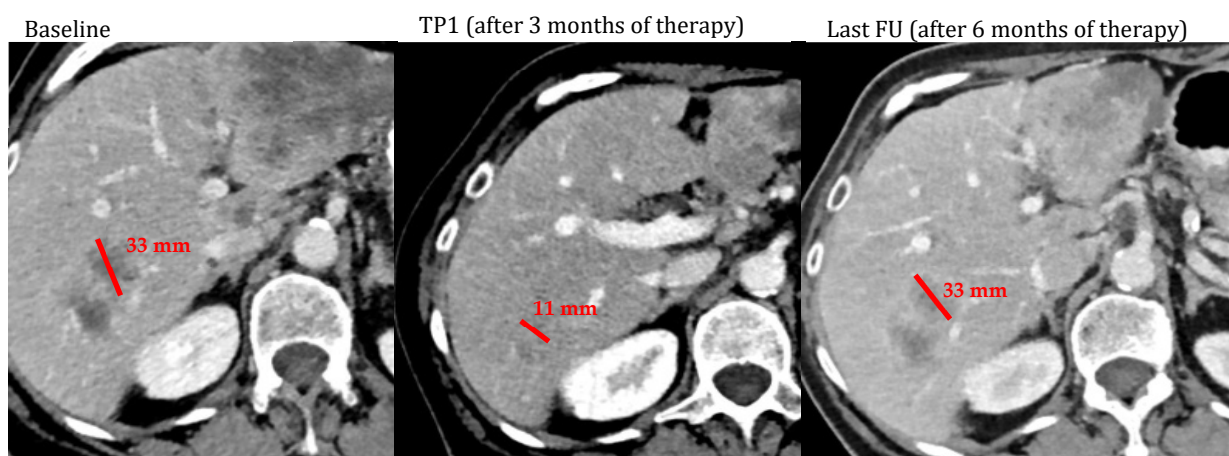
(a)



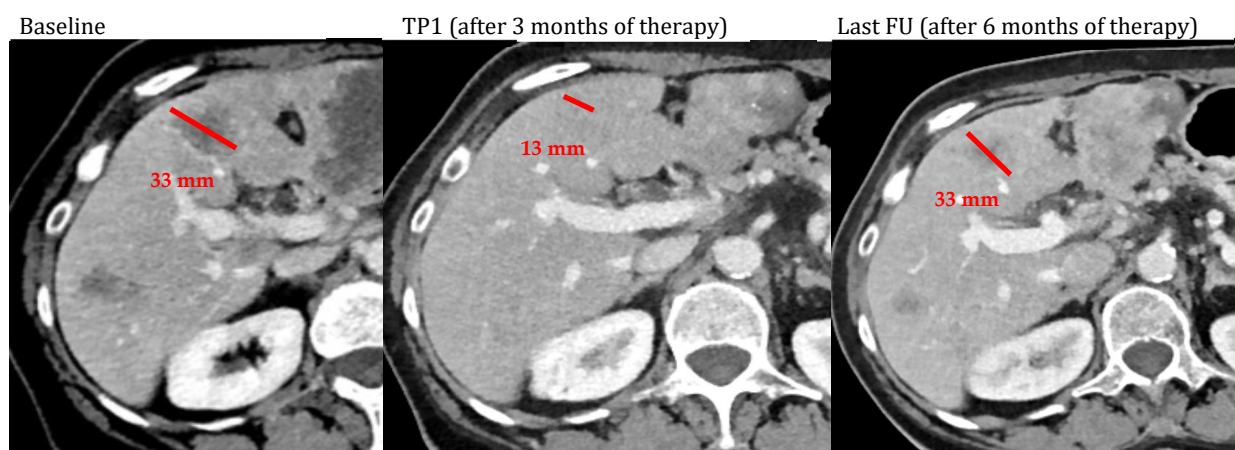
(b)

**Figure S4.** (a) Lesion 2 in patient 1017 at baseline, TP1 and last FU during first-line chemotherapy. Lesion 2 was a bad responder misclassified as R+ by the algorithm (PD at 6 months). (b) Lesion 4 in patient 1017 at baseline, TP1 and last FU during first-line chemotherapy. Lesion 7 was a good responder misclassified as R- by the algorithm (CR at 6 months).





(a)



(b)

**Figure S5.** (a) Lesion 4 in patient 1010 at baseline, TP1 and last FU during first-line chemotherapy. Lesion 4 was a bad responder misclassified as R+ by the algorithm (PD at 6 months). To note that a PR was recorded at TP1. (b) Lesion 7 in patient 1010 at baseline, TP1 and last FU during first-line chemotherapy. Lesion 4 was a bad responder misclassified as R+ by the algorithm (PD at 6 months). To note that a PR was recorded at TP1. The lesion was almost indistinguishable from normal liver parenchyma at TP1, being a possible source for incorrect segmentation.