This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Supplementary Material
Knowledge-injection for small-cohort medical image learning

Marlene Mertens[1,2], Felix Pfister[3], and Stefan Höhme[1,2]

[1] Interdisciplinary Center for Bioinformatics (IZBI), Leipzig, Germany
[2] Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI), Leipzig, Germany
mertens@informatik.uni-leipzig.de
hoehme@uni-leipzig.de
[3] Department of Diagnostic and Interventional Radiology, University Hospital Leipzig, Leipzig, Germany
Felix.Pfister@medizin.uni-leipzig.de

S1 Feature extraction and definitions

Table S1. List of our 14 knowledge-based features for cirrhosis classification in NAFLD and their source sequences.

Feature	Source sequence
Fat fraction, mean	Q-Dixon PDFF
First-order entropy	T1-weighted hepatobiliary phase
NGTDM busyness	T1-weighted hepatobiliary phase
Liver iron content	Q-Dixon Report of R2*
Liver relative enhancement (LRE)	Native and hepatobiliary phase T1
Liver-to-spleen ADC ratio	DWI ADC (b = 50, 1000 s/mm^2)
Spleen volume	T1-weighted hepatobiliary phase
Portal vein diameter (PV)	T1-weighted hepatobiliary phase
Caudate-to-right-lobe ratio (C/RL)	T1-weighted hepatobiliary phase
Liver-spleen volume ratio (LSVR)	T1-weighted hepatobiliary phase
Age	Clinical record
BMI	Clinical record
Type 2 diabetes (T2DM)	Clinical record
Arterial hypertension	Clinical record

Sequences. We derive the imaging features from five MRI sequences. The Q-Dixon acquisition yields a proton density fat fraction map (PDFF) in which every voxel encodes the local hepatic fat percentage, and through its report the liver iron value is directly computed from the R2* echo. The diffusion weighted imaging provides an apparent diffusion coefficient map, with b-values of 50 and

1000. A native T1-weighted sequence is acquired before contrast, and a T1-weighted hepatobiliary phase sequence is acquired 15 minutes after gadoxetate administration (Primovist).

Segmentations. The liver on the T1-weighted hepatobiliary phase was segmented manually by a radiologist in 3D Slicer, and this mask defines the region for the two hepatobiliary texture features, the first-order entropy [1] and the NGTDM busyness [1]. The focal liver lesions and intrahepatic vessels were also delineated manually on the same sequence and serve as exclusion masks. The liver on the Q-Dixon PDFF sequence was also segmented by the same radiologist in 3D Slicer. All remaining structures were segmented automatically with TotalSegmentator, namely the spleen, the portal and splenic vein, the eight Couinaud liver segments, and the liver on the native T1-weighted, and the DWI ADC.

Preprocessing. All T1-weighted volumes underwent N4 bias field correction [2] restricted to a liver or body mask, which removes the low frequency intensity drift that would otherwise bias the liver versus spleen comparison. The quantitative parameter maps, meaning PDFF and DWI ADC, were not bias corrected because their voxel values are physically meaningful. Volumes were resampled to a defined voxel grid where texture extraction required it. Features that combine two sequences were computed after a mask-guided rigid registration using the liver masks as the registration region. For the intensity features a combined vessel and lesion mask was dilated and removed from the liver so that bright vessels and focal lesions do not bias the parenchymal signal.

Feature descriptions.

- **Fat fraction, mean.** Mean of the PDFF voxels inside the liver after excluding voxels above 80 percent, which mainly contain residual perihepatic fat and vessels.
- **First-order entropy.** PyRadiomics first-order entropy [1] of the hepatobiliary phase T1 inside the liver, a measure of signal heterogeneity.
- **NGTDM busyness.** PyRadiomics NGTDM busyness [1] of the hepatobiliary phase T1 inside the liver, which captures parenchymal nodularity.
- **Liver iron content.** Read directly from the vendor Q-Dixon report produced by a Siemens scanner.
- **Liver relative enhancement (LRE).** Ratio of the liver to spleen signal in the hepatobiliary phase divided by the same ratio in the native phase, which isolates hepatocyte gadoxetate uptake and cancels scanner gain.
- **Liver-to-spleen ADC ratio.** Mean liver ADC divided by mean spleen ADC over the cleaned masks.
- **Spleen volume.** Voxel count of the spleen mask multiplied by the voxel volume, in millilitres.
- **Portal vein diameter (PV).** Twice the maximum of a Euclidean distance transform of the portal and splenic vein mask, giving the inscribed diameter in millimetres.

- **Caudate-to-right-lobe ratio (C/RL).** Volume of Couinaud segment 1 divided by the summed volume of segments 5 to 8.
- **Liver-spleen volume ratio (LSVR).** Liver volume divided by spleen volume.
- **Age.** Patient age from the clinical record.
- **BMI.** Body mass index from the clinical record.
- **Type 2 diabetes.** Binary diabetes status from the clinical record.
- **Arterial hypertension.** Binary hypertension status from the clinical record.

S2 Radiomics pipeline

As a data-driven counterpart to the knowledge-based feature set, we assemble a broad radiomics pool and select from it statistically. Using PyRadiomics [1] on the available segmentation, we compute the full first-order and texture panels (GLCM, GLRLM, GLSZM, GLDM, NGTDM) over the liver region in the PDFF, T1-weighted hepatobiliary phase, and DWI ADC sequences, and add 3D shape descriptors for the liver, spleen, Couinaud segments, and portal vein together with the volumetric features. This yields 456 candidate features for the 19 patients.

The baseline follows the same protocol as all other methods, namely stratified k-fold cross-validation for $k = 5$ and $k = 3$, each repeated five times with per-repeat seeds derived from a fixed base seed. Preprocessing and feature selection are performed on the training fold only. Within each fold we median-impute, drop near-constant features, standardize, and select features with an elastic-net logistic regression whose penalty is tuned by an inner cross-validation. The number of selected features is not fixed but emerges from the penalty, varying across folds: for the 3-fold setting a mean of 81.3 features in range 9 to 398 features and for the 5-fold setting a mean of 106.1 features in the range 3 to 373 features. The selected features are classified with a Random Forest (500 trees, balanced class weights). The same configuration is used for the knowledge-based baseline.

The Radiomics baseline classification and the knowledge-based Random Forest approach differ only in how the features are selected, everything else is fixed. This isolates the feature-selection component, so any difference between the two is attributable to selecting features from clinical knowledge rather than estimating them from the data.

S3 Cohort-specific threshold computation

The cohort-specific markers for LSVR and NGTDM busyness are recomputed within each cross-validation fold, using only the non-cirrhotic training samples of that fold, so that no held-out sample influences threshold estimation. Tables S2-S3 report the resulting threshold values for every fold across all repeats, in both the 3-fold and 5-fold settings. The variation across folds reflects the small per-fold training size and is consistent with the fold-to-fold instability reported in the main results.

Table S2. Per-fold LSVR marker (25th percentile of the non-cirrhotic training class) for the 5×3-fold and 5×5-fold settings. Each value is computed from the training partition of the respective fold only.

LSVR 25%	**Repeat 1**	**Repeat 2**	**Repeat 3**	**Repeat 4**	**Repeat 5**
3-Fold: Fold 1	7.84	7.64	7.64	6.86	6.86
3-Fold: Fold 2	7.64	8.41	8.12	8.61	9.20
3-Fold: Fold 3	7.89	7.89	7.75	7.89	7.00
5-Fold: Fold 1	7.75	7.89	7.89	7.00	7.00
5-Fold: Fold 2	7.14	7.14	7.14	7.14	8.14
5-Fold: Fold 3	8.14	8.14	8.14	8.14	8.14
5-Fold: Fold 4	7.14	9.09	8.14	9.09	8.14
5-Fold: Fold 5	8.14	7.14	7.14	7.14	7.14

Table S3. Per-fold NGTDM busyness marker (75th percentile of the non-cirrhotic training class) for the 5×3-fold and 5×5-fold settings. Each value is computed from the training partition of the respective fold only.

Busyness 75%	**Repeat 1**	**Repeat 2**	**Repeat 3**	**Repeat 4**	**Repeat 5**
3-Fold: Fold 1	74.12	63.20	52.71	82.82	54.50
3-Fold: Fold 2	63.20	74.12	63.70	50.32	82.82
3-Fold: Fold 3	59.35	59.35	77.61	64.56	64.56
5-Fold: Fold 1	64.56	58.60	63.81	77.61	59.35
5-Fold: Fold 2	72.39	72.39	55.00	55.00	55.00
5-Fold: Fold 3	54.00	55.00	55.00	54.00	72.39
5-Fold: Fold 4	55.00	72.39	72.39	72.39	55.00
5-Fold: Fold 5	72.39	55.00	72.39	55.00	72.39

Table S4. Model architecture and training hyperparameters. The Multi-layer perceptron and logistic regression share the input, task loss and training procedure, and differ only in capacity.

Hyperparameter	Value
Architecture	
Hidden layers (MLP)	16, 8
Normalization	Layer normalization
Activation	ReLU
Dropout (MLP)	0.4
Training	
Task loss	BCE with logits
Optimizer	AdamW
Learning rate	1e−3
Weight decay	1e−2
Batch size	1
Epochs	100
Gradient clipping	max-norm 1.0
Feature standardization	training-fold statistics only

S4 Model architecture and training hyperparameters

The shallow network is a Multi-Layer Perceptron (MLP) with two hidden layers of 16 and 8 units, each followed by layer normalization, a ReLU activation and dropout, before a final linear layer produces a single logit. In total the network has roughly 430 trainable parameters, keeping the parameter count low relative to the 19 available cases. The full set of architecture and training hyperparameters is given in Table S4. Three choices target the small-data regime in particular: a high dropout rate of 0.4 to limit overfitting, together with a batch size of 1 and a deliberately small learning rate of 1e-3, which produce many small gradient updates and keep the optimization stable when only a handful of samples are available.

The logistic-regression baseline shares the same learning rate and training procedure, replacing the hidden layers with a single linear layer and using L2 weight decay as regularization in its plain (BCE-only) variant.

S5 Clinical loss and adversarial hyperparameters

Table S5 lists the hyperparameters of the two knowledge-based loss terms. Because the cohort is too small to set aside a separate validation split, these hyperparameters were not tuned on held-out data but fixed after inspecting the training-loss curves over the epochs. We chose the weights λ_{logic} and λ_{adv} so that the clinical and adversarial terms are of a comparable magnitude to the task loss without dominating it, and set the thresholds $\tau_{\text{high}} = 0.6$ and $\tau_{\text{low}} = 0.2$ to leave a margin around the decision boundary. Since all features are standardized

Table S5. Clinical loss and adversarial training hyperparameters.

Hyperparameter	Value
Clinical loss	
Weight λ_{logic}	1.0
τ_{high} / τ_{low}	0.6 / 0.2
Adversarial training (PGD)	
Weight λ_{adv}	1.0
Ball radius ϵ	$0.5\,\sigma$ (L_∞)
Steps k	7
Step size	$\alpha \cdot \epsilon / k, \quad \alpha = 1.5$
Random start	yes
Frozen features	T2DM, Arterial Hypertension

Table S6. Fisher's exact test against chance (one-sided) on the pooled out-of-fold predictions. Values are p-values.

Method	5-fold	3-fold
Radiomics-RF	0.144	0.506
KB-RF	0.005	0.005
Clinical rule	0.022	0.022
LR + L2	0.022	0.055
LR + clin	0.054	0.022
LR + clin + adv	0.055	0.022
MLP	0.071	0.024
MLP + clin	0.005	0.005
MLP + clin + adv	0.022	0.005

to zero mean and unit variance beforehand, the perturbation radius $\epsilon = 0.5\,\sigma$ is expressed in standard-deviation units and therefore applies the same relative search range to every feature, regardless of its original scale. Together with $k = 7$ PGD steps, this restricts the search to a small region around each sample, so the perturbed points remain close to the observed data. The two binary clinical features are held fixed during the search, since they are categorical and a continuous perturbation would move them to values with no valid interpretation.

S6 Statistical tests

All tests operate on out-of-fold predictions, computed identically for all models with the same 5×repeated stratified cross-validation used in the main experiments. Within each repeat the folds partition the cohort, so every patient is held out exactly once and receives one prediction per repeat. For the probabilistic models we average the five predicted probabilities and threshold at 0.5, while the clinical rule is deterministic. This yields one prediction per patient (n = 19) for every model, forming the 2×2 prediction versus label table used by the tests.

Table S7. McNemar's exact test on paired per-patient out-of-fold predictions. For each comparison (A vs. B), b and c count the discordant patients classified correctly by *only* A or *only* B, respectively; the test uses these discordant pairs alone. Identical marks pairs that agreed on every patient, for which no test is defined (p = n/a). Only KB-RF vs. Radiomics-RF in the five-fold setting reaches significance.

Comparison (A vs. B)	Setting	b (A only)	c (B only)	p
MLP vs. MLP+clin	5-fold	0	2	0.500
	3-fold	0	1	1.000
MLP vs. MLP+clin+adv	5-fold	2	3	1.000
	3-fold	0	1	1.000
MLP+clin vs. MLP+clin+adv	5-fold	2	1	1.000
	3-fold	identical		n/a
KB-RF vs. Radiomics-RF	5-fold	6	0	**0.031**
	3-fold	6	1	0.125
KB-RF vs. MLP+clin+adv	5-fold	2	0	0.500
	3-fold	identical		n/a

These pooled predictions differ from the fold-averaged balanced accuracy in Table 2, so two models can share identical pooled predictions while still differing in Table 2.

For each model we apply a one-sided Fisher's exact test to the 2×2 table of predicted versus true label, testing whether the predictions are associated with the labels beyond chance. The exact test is used because the asymptotic approximation is unreliable at this sample size. The p-values are given in Table S6. Most models exceed chance at the 0.05 level. The exceptions are the plain MLP in the 5-fold setting, the plain LR in the 3-fold setting, LR+clin and LR+clin+adv in the 5-fold setting, and the radiomics baseline in both settings.

We compare two models with McNemar's exact test on their paired per-patient predictions, using only the discordant patients. The knowledge-based features differ significantly from elastic-net selection in the 5-fold setting ($p < 0.05$), but not in the 3-fold setting. No other comparison reaches significance in either setting, including the stepwise MLP comparisons and the knowledge-based random forest against MLP+clin+adv (Table S7). With few discordant patients these tests are underpowered, so a non-significant result means a difference cannot be resolved at this cohort size rather than that the models are equivalent.

References

1. van Griethuysen, J.J.M., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R.G.H., Fillion-Robin, J.C., Pieper, S., Aerts, H.J.W.L.: Computational radiomics system to decode the radiographic phenotype. Cancer Research **77**(21), e104–e107 (2017). https://doi.org/10.1158/0008-5472.CAN-17-0339
2. Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C.: N4ITK: Improved N3 bias correction. IEEE Transactions on Medical Imaging **29**(6), 1310–1320 (2010). https://doi.org/10.1109/TMI.2010.2046908