Research article

JHEP|Reports

Automated CT LI-RADS v2018 scoring of liver observations using machine learning: A multivendor, multicentre retrospective study

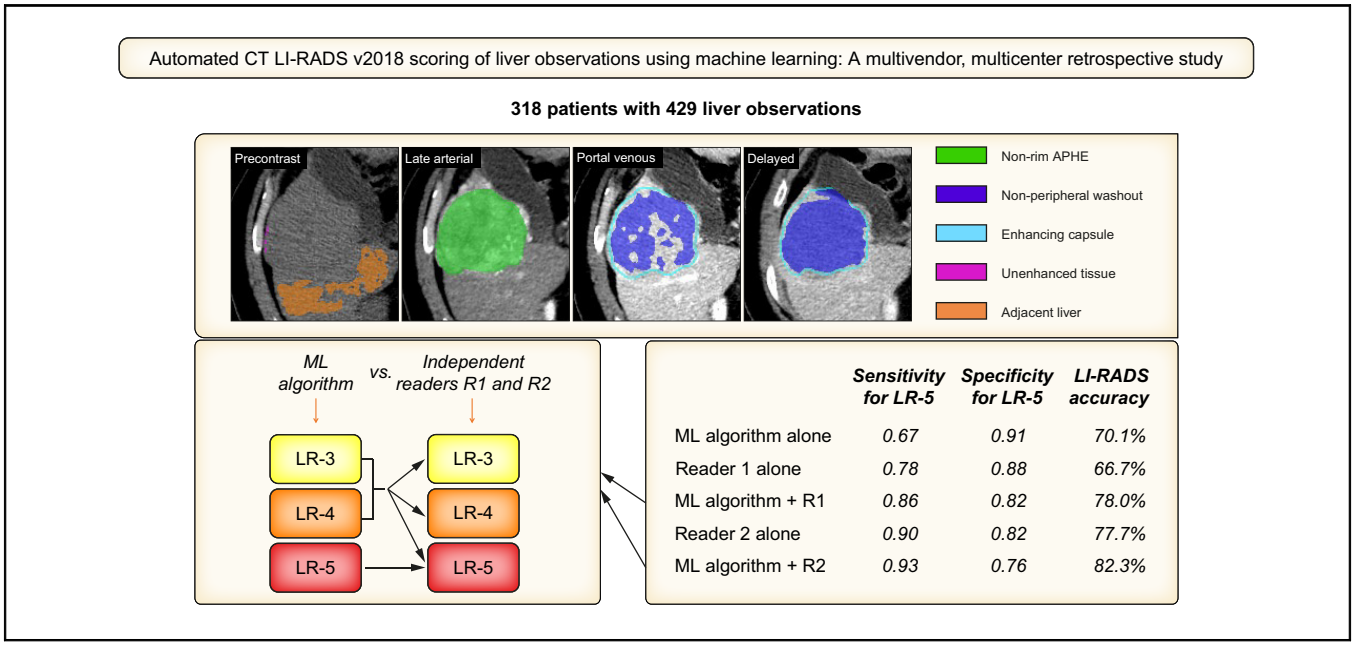
Authors

Sébastien Mulé, Maxime Ronot, Mario Ghosn, Riccardo Sartoris, Giuseppe Corrias, Edouard Reizine, Vincent Morard, Ronan Quelever, Laura Dumont, Jorge Hernandez Londono, Nicolas Coustaud, Valérie Vilgrain, Alain Luciani

Correspondence

maxime.ronot@aphp.fr (M. Ronot).

Graphical abstract



Highlights

- ML algorithm evaluated whether non-rim APHE, non-peripheral washout, and enhancing capsule were present, absent, or of uncertain presence.
- LR-5 observations were diagnosed with high specificity and good sensitivity.
- A stepwise use of the ML algorithm and the radiologist's visual analysis improved the overall performance for LR-5.

Impact and implications

Assessment of CT/MRI LI-RADS v2018 major features leads to substantial inter-reader variability and potential decrease in hepatocellular carcinoma diagnostic accuracy. Rather than replacing radiologists, our results highlight the potential benefit from the radiologist–artificial intelligence interaction in improving focal liver lesions characterisation by using the developed algorithm as a triage tool to the radiologist's visual analysis. Such an AI-enriched diagnostic pathway may help standardise and improve the quality of analysis of liver lesions in patients at high risk for HCC, especially in non-expert centres in liver imaging. It may also impact the clinical decision-making and guide the clinician in identifying the lesions to be biopsied, for instance in patients with multiple liver focal lesions.

https://doi.org/10.1016/j.jhepr.2023.100857

Research article

JHEP|Reports

Automated CT LI-RADS v2018 scoring of liver observations using machine learning: A multivendor, multicentre retrospective study

Sébastien Mulé,[1,2,3] Maxime Ronot,[4,5,*] Mario Ghosn,[1,2] Riccardo Sartoris,[4] Giuseppe Corrias,[4] Edouard Reizine,[1,2,3] Vincent Morard,[6] Ronan Quelever,[6] Laura Dumont,[6] Jorge Hernandez Londono,[6] Nicolas Coustaud,[6] Valérie Vilgrain,[4,5] Alain Luciani[1,2,3]

[1]*Service d'Imagerie Médicale, AP-HP, Hôpitaux Universitaires Henri Mondor, Créteil, France;* [2]*Faculté de Santé, Université Paris Est Créteil, Créteil, France;* [3]*INSERM IMRB, U 955, Equipe 18, Créteil, France;* [4]*Service de Radiologie, Hôpital Beaujon, AP-HP Nord, Clichy, France;* [5]*Université de Paris, CRI, INSERM U1149, Paris, France;* [6]*GE Healthcare, Buc, France*

JHEP Reports **2023.** https://doi.org/10.1016/j.jhepr.2023.100857

Background & Aims: Assessment of computed tomography (CT)/magnetic resonance imaging Liver Imaging Reporting and Data System (LI-RADS) v2018 major features leads to substantial inter-reader variability and potential decrease in hepatocellular carcinoma diagnostic accuracy. We assessed the performance and added-value of a machine learning (ML)-based algorithm in assessing CT LI-RADS major features and categorisation of liver observations compared with qualitative assessment performed by a panel of radiologists.

Methods: High-risk patients as per LI-RADS v2018 with pathologically proven liver lesions who underwent multiphase contrast-enhanced CT at diagnosis between January 2015 and March 2019 in seven centres in five countries were retrospectively included and randomly divided into a training set (n = 84 lesions) and a test set (n = 345 lesions). An ML algorithm was trained to classify non-rim arterial phase hyperenhancement, washout, and enhancing capsule as present, absent, or of uncertain presence. LI-RADS major features and categories were compared with qualitative assessment of two independent readers. The performance of a sequential use of the ML algorithm and independent readers were also evaluated in a triage and an add-on scenario in LR-3/4 lesions. The combined evaluation of three other senior readers was used as reference standard.

Results: A total of 318 patients bearing 429 lesions were included. Sensitivity and specificity for LR-5 in the test set were 0.67 (95% CI, 0.62–0.72) and 0.91 (95% CI, 0.87–0.96) respectively, with 242 (70.1%) lesions accurately categorised. Using the ML algorithm in a triage scenario improved the overall performance for LR-5. (0.86 and 0.93 sensitivity, 0.82 and 0.76 specificity, 78% and 82.3% accuracy for the two independent readers).

Conclusions: Quantitative assessment of CT LI-RADS v2018 major features is feasible and diagnoses LR-5 observations with high performance especially in combination with the radiologist's visual analysis in patients at high-risk for HCC.

Impact and implications: Assessment of CT/MRI LI-RADS v2018 major features leads to substantial inter-reader variability and potential decrease in hepatocellular carcinoma diagnostic accuracy. Rather than replacing radiologists, our results highlight the potential benefit from the radiologist–artificial intelligence interaction in improving focal liver lesions characterisation by using the developed algorithm as a triage tool to the radiologist's visual analysis. Such an AI-enriched diagnostic pathway may help standardise and improve the quality of analysis of liver lesions in patients at high risk for HCC, especially in non-expert centres in liver imaging. It may also impact the clinical decision-making and guide the clinician in identifying the lesions to be biopsied, for instance in patients with multiple liver focal lesions.

Introduction

Contrast-enhanced computed tomography (CT) and magnetic resonance imaging (MRI) are the imaging modality of choice for diagnosis and staging of liver lesions in patients at high risk for hepatocellular carcinoma (HCC). The Liver Imaging Reporting and Data System (LI-RADS) is a comprehensive system that provides standardisation of terminology and liver image acquisition and interpretation in these high-risk patients.[1–3] The CT/MRI LI-RADS algorithm weights the probability for a given observation to be an HCC, allowing non-invasive diagnosis for HCC but also identification of lesions for which biopsy may be considered. To this aim, five major imaging features are considered: non-rim arterial phase hyperenhancement (APHE), non-peripheral washout, enhancing capsule, observation size, and threshold growth.[2] The first three are qualitatively evaluated, leading to substantial inter-reader variability and potential decrease in diagnostic accuracy.[4]

Keywords: Hepatocellular carcinoma; LI-RADS; Major features; Computed tomography; Machine learning.

Received 29 November 2022; received in revised form 21 June 2023; accepted 12 July 2023; available online 22 July 2023

* Corresponding author. Address: Service de Radiologie, Hôpital Beaujon AP-HP Nord, 100 Bd du Général Leclerc, 92110 Clichy, France.

E-mail address: maxime.ronot@aphp.fr (M. Ronot).

Kang *et al.*[5] recently showed that the inter-reader reliability of CT LI-RADS major features differed significantly according to the average reader experience and to the difference in reader experience. This may be because of observations with subtle, not easily assessed APHE and/or washout. Another potential explanation is the possible heterogeneous spatial distribution of APHE and washout within lesions. Indeed, rim and non-rim APHE features – and peripheral and non-peripheral washout – have entirely different meanings. Although rim APHE – that is, APHE most pronounced in lesion periphery – may be seen in HCC, especially in the sarcomatoid subtype, and in combined tumours,[6,7] this feature is highly suggestive of non-HCC malignancy and is therefore recognised as a LR-M feature. In the same way, peripheral washout – that is, washout most pronounced in lesion periphery – is also suggestive of non-HCC malignancy and thus should lead to lesion categorisation as LR-M. Although inter-reader variability of enhancing capsule seems to be modest,[8,9] a three-dimensional (3D) assessment is necessary as this feature may be incomplete and therefore be underdiagnosed when analysed on a single section. Consequently, the LI-RADS algorithm may be less widely used in non-expert centres in liver imaging.

It has been suggested that a quantitative evaluation of both APHE and washout using two-dimensional (2D) regions of interest (ROIs) may help improve the accuracy of LI-RADS tumour categorisation and thus accuracy for the non-invasive diagnosis of HCC.[10,11] Automated major features assessment and LI-RADS tumour categorisation may be of great help for lesion characterisation, subsequent management optimisation, and help in identifying the lesions to be biopsied,[12] with different added-value in expert and non-expert centres in liver imaging. Convolutional neural network-based deep learning algorithms have shown their potential to classify focal liver lesions in a heterogeneous patient population of benign and malignant primary liver lesions and liver metastases.[13,14] However, the accuracy and reproducibility of deep learning algorithms may be compromised in the absence of large amounts of available data. Conversely, machine learning (ML)-based algorithms such as logistic regression models may require far less data for the training process, allowing the majority of data to be dedicated to the model validation process. To our knowledge, the feasibility and performance of an automated ML-based 3D analysis of LI-RADS v2018 major features and the resulting LI-RADS categorisation has been poorly investigated to date.

In this context, our study aimed to assess the performance and added-value of an ML algorithm in the assessment of LI-RADS v2018 major features and categorisation of hepatic observations at CT in patients at risk for HCC. To this aim, three scenarios were explored: the ML algorithm performances were first compared with the visual assessment of independent radiologists. The optimal integration of the ML algorithm in the clinical diagnostic pathway was then investigated by evaluating the performance of a sequential use of the ML algorithm and independent radiologists in a triage scenario, and in an add-on scenario.

Patients and methods

Study participants

This multicentric study was conducted by the HECAM (HEpatocellular Carcinoma Multi-technological) research consortium and was partly financed by 'Bpifrance' (French Banque Publique d'Investissement) together with GE Healthcare France. The institutional review boards (IRB) of the study centres approved this multicentre retrospective study, and informed consent was waived by the IRB.

Patients with cirrhosis or chronic HBV infection or prior HCC with pathologically proven liver lesions between January 2015 and March 2019 from seven centres in five countries (Henri Mondor University Hospital, Créteil, France; Beaujon University Hospital, Clichy, France; National University of Seoul, South Korea; University Hospital of Pisa, Italy; University Hospital of Angers, Angers, France; Cetir medical center, Barcelona, Spain; Zwanger-Pesiri Radiology Hospital, New York, USA) were considered for inclusion. Patients were included if they had undergone multiphase contrast-enhanced CT at diagnosis, and if liver lesions were pathologically proven. Patients were excluded if they had undergone lesion treatment before surgery/biopsy or if the late-arterial and/or portal venous phase was not acquired. A total of 326 patients (median age, 63 years; interquartile range [IQR], 56–71 years) with 440 observations were included. Recorded demographic characteristics and clinical data included age, sex, presence of cirrhosis, cause of liver disease, and serum alpha-foetoprotein (AFP) level.

CT examination

CT examinations were obtained with CT scanners from various vendors (Table S1). All patients underwent a multiphase contrast-enhanced CT scan of the abdomen, including at least late-arterial and portal-venous phases. A pre-contrast (unenhanced) phase and a delayed phase were acquired in 322/326 (98.8%) and 279/326 (85.6%) patients, respectively. All four phases were available in 270/326 (82.8%) of patients. The CT acquisition parameters varied among centres according to local routine protocols. The tube voltage was set at 120 kVp in 274/326 (84%) of cases (range, 80–140 kVp), mean effective tube current, 350 mA, median slice thickness, 1.25 mm (range, 0.625–3 mm), and median reconstruction interval, 1.25 mm (range, 0.625–3 mm).

Readers visual assessment and ground truth definition

Five independent senior readers visually assessed the 440 observations. For each observation, the presence of features suggestive of malignancy but not specific for HCC (including rim APHE) and the presence of tumour-in-vein were first evaluated, and observations with such features were respectively categorised as LR-M and LR-TIV. Observations categorised LR-1 or LR-2, LR-M, or LR-TIV were excluded from the subsequent analysis. In the remaining cases, non-rim APHE, non-peripheral washout, and enhancing capsule were visually assessed. The corresponding LI-RADS category was then proposed. As only baseline imaging was available, observation growth was not considered to define the LI-RADS category.

For each observation, the ground truth was based on the evaluation of three out of the five senior readers (SM, RS, and MG, with 6, 5, and 3 years of experience in the use of LI-RADS, respectively) and defined as the majority choice among them. For each observation, one of the three readers placed a free-hand volume of interest encompassing the entire observation.

The performances of the remaining two independent readers (with 5 and 3 years of experience in the use of LI-RADS, respectively), not implicated in the ground truth definition, were evaluated and compared with those of the ML algorithm, using the ground truth as defined above as reference standard.

Machine learning algorithm

Observations were divided into a training set (n = 84 observations) and a test set (n = 345 observations) using a stratified approach based on LI-RADS features and common characteristics of the observations to warrant balanced distribution of major features in the training set. A subset of the training set was defined as the validation set and allowed to monitor the model's performance during the training process.

The image processing and ML pipeline for LI-RADS major features analysis was divided into different steps (Fig. 1): automatic phases identification, non-rigid registration to enable a voxel-wise temporal profile analysis leading to lesion subsegmentation and surrounding parenchyma delineation (Fig. 2). The algorithm design was driven according to the two following principles: accuracy and model explainability. In combination with an extensive data augmentation strategy, this approach allowed to train the model using few observations (see Supplementary material).

The model outcome was a probability of major feature presence, and a threshold was set for the final decision. This threshold was chosen as the value providing the highest specificity and at least 0.8 sensitivity. A scrutiny zone on both sides of the threshold value was introduced, and the presence of a feature was stated as uncertain in case of a probability close to the threshold value (see Supplementary material).

Statistical analysis

Continuous variables are provided as median and interquartile range. Categorical variables as count and percentages. Chi-square and Fisher exact tests were used to compare frequencies of categoric variables between the training and test sets, as appropriate. The paired sample *t* test and Wilcoxon rank-sum test were performed for continuous variables, as appropriate.

The sensitivity and specificity of the ML algorithm and of the two independent readers for the assessment of major features compared with the ground truth were estimated. Bootstrap resampling was performed to estimate the precision and confidence of the results by obtaining bootstrap 95% CIs. The accuracy of the resulting observation categorisation and the sensitivity and specificity for LR-5 were also calculated.

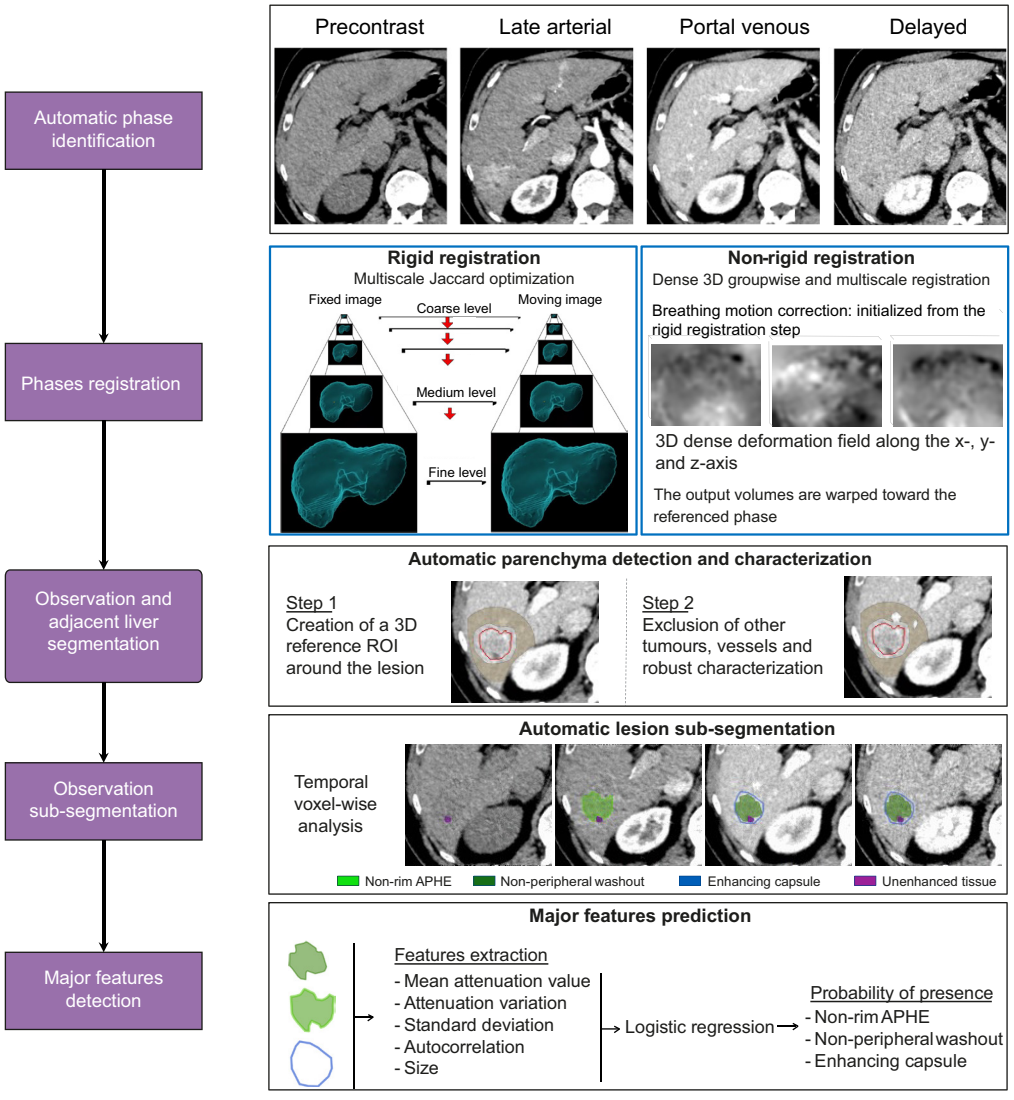


Fig. 1. Overview and major steps of the machine learning-based algorithm.

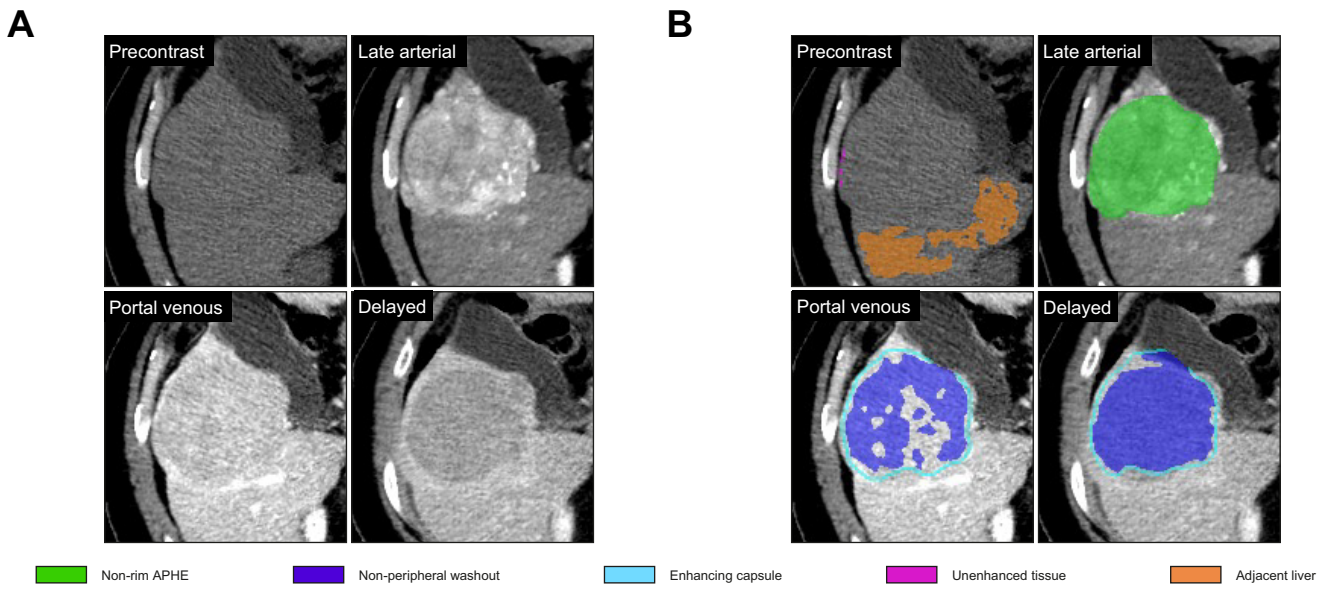


Fig. 2. Axial multiphasic CT images in a 70-year-old man with 76-mm hepatocellular carcinoma in segment 5 of the liver. Multiphasic images are presented (A) without and (B) with superimposition of areas in which non-rim arterial phase hyperenhancement (APHE), non-peripheral washout and enhancing capsule were detected by the machine learning algorithm. The adjacent liver parenchyma automatically segmented for the analysis of the observation enhancement patterns is also shown. Non-rim APHE, non-peripheral washout, and enhancing capsule were present and accurately detected by both the machine learning algorithm and the two independent readers in this LR-5 observation.

Subgroup analyses according to the size of observations (<20 mm and ≥20 mm observations) were also performed.

The above-mentioned analyses were performed in two different situations. First, the major features with uncertain presence according to the ML algorithm were categorised as absent, in agreement with LI-RADS recommendations.[13] All observations in the test set were included. Second, the major features with uncertain presence according to the ML algorithm (*i.e.* within the scrutiny zone) were excluded from the analyses. Hence, only features for which the algorithm was able to decide whether they were present or not were included.

To investigate how the developed ML algorithm may be optimally integrated in the clinical diagnostic pathway, we evaluated the sequential performances of a stepwise use of the ML algorithm and the independent readers (Fig. 3). On one hand, the ML algorithm was used to categorise indeterminate observations (LR-3 and LR-4 observations) after visual analysis of the independent readers (add-on role). On the other hand, the ML algorithm was applied to all hepatic observations (triage role), and (i) only LR-3 and LR-4 observations per the ML algorithm were further visually analysed by the readers (triage #1); (ii) only observations with at least one major feature of uncertain presence were visually analysed by the readers (triage #2).

The agreement between pairs of readers was evaluated using Cohen's kappa statistics, and between all readers using Fleiss kappa. In addition, the overall inter-reader agreement was introduced. For each observation, the overall inter-reader agreement for each major feature was arbitrarily considered high if at least four of the five readers agreed with each other. Otherwise, the overall inter-reader agreement was considered low.

A p value <0.05 was considered statistically significant. Statistical analyses were conducted in Python using the scikit-learn library.

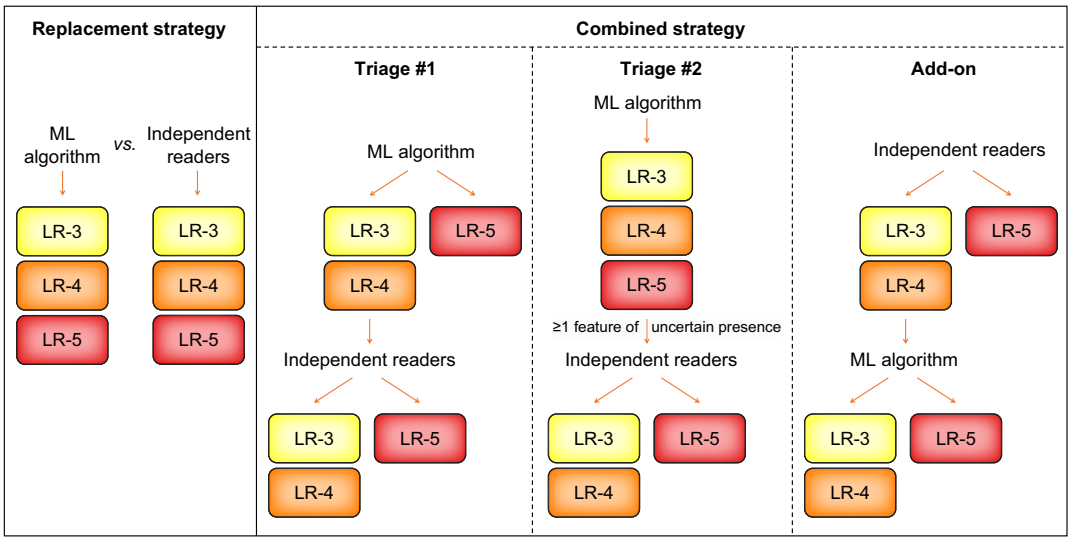
Results

Participant characteristics

According to the ground truth, four observations were categorised LR-1 or LR-2, two observations were categorised LR-M, and tumour-in-vein was present for five observations. After excluding these 11 observations, a total of 318 patients bearing 429 observations were included, with a large majority of HCC lesions (Table 1). The median tumour size was 28 mm (IQR, 18–47 mm), 129/429 observations (30.1%) were smaller than 20 mm, whereas 98/429 (22.8%) were larger than 50 mm. The vast majority of observations had a nodular growth pattern (352/429, 82.1%), the remaining 77/429 (17.9%) being infiltrative. Two hundred and eighty-three (283/318, 89.0%) patients had cirrhosis, including six with a history of HCC. The remaining 35 patients had either prior HCC (n = 4) and/or chronic HBV infection (n = 31). The median AFP serum level was 9 ng/ml (IQR, 4.9–49.8 ng/ml).

Ground truth

Non-rim APHE, non-peripheral washout, and enhancing capsule were present in 360/429 (83.9%), 336/429 (76.1%), and 110/429 (25.6%) observations, respectively. Overall, 82/429 (19.1%), 60/429 (14.0%), and 287/429 (66.9%) observations were categorised LR-3, LR-4 and LR-5, respectively. Sensitivity and specificity of LR-5 for HCC were 0.55 (95% CI, 0.48–0.61) and 0.88 (95% CI, 0.62–0.98), respectively. Those 429 observations were divided



Sensitivity for LR-5	0.67	0.78/0.90	0.86/0.93	0.77/0.80	0.84/0.88
Specificity for LR-5	0.91	0.88/0.82	0.82/0.76	0.88/0.87	0.85/0.77
LI-RADS accuracy	70.1%	66.7%/77.7%	78.0%/82.3%	79.1%/75.9%	74.2%/78.6%

Fig. 3. Assessment of the performance and added-value of the developed ML algorithm in the assessment of LI-RADS v2018 major features and categorisation of hepatic observations at CT in patients at risk for HCC. The ML algorithm performance was first compared with the visual assessment of independent radiologists (replacement scenario). The performance of a sequential use of the ML algorithm and independent radiologists was then evaluated in triage scenarios, and in an add-on scenario. CT, computed tomography; HCC, hepatocellular carcinoma; LI-RADS, Liver Imaging Reporting and Data System; LR, LI-RADS; ML, machine learning.

Table 1. Characteristics of the study population.

Characteristics	Available data (n)	Training set (n = 39)	Test set (n = 279)	*p* value
Sex, n (%)	286			0.50
Men		28 (85)	196 (77)	
Women		5 (15)	57 (23)	
Age (years)	296	62 (54–71)	63 (56–71)	0.68
Cause of liver disease, n (%)	275			0.26
Alcohol abuse		7 (22.6)	61 (23.4)	
Chronic hepatitis B		4 (12.9)	76 (29.1)	
Chronic hepatitis C		10 (32.3)	70 (26.8)	
NASH		10 (32.3)	48 (18.4)	
Other		0 (0.0)	6 (2.3)	
Cirrhosis, n (%)	318	35 (89.7)	241 (86.4)	0.80
AFP serum level (ng/ml)	122	33 (6–326)	9 (4–38)	0.11
Tumour size (mm)*	429	22 (16–39)	29 (20–49)	0.10
Type of tumour*, n (%)	429			0.45
HCC		83 (98.8)	332 (96.2)	
iCCA		0	2 (0.6)	
cHCC-CCA		0	6 (1.7)	
Liver metastasis		1 (1.2)	0 (0.0)	
FNH		0	1 (0.3)	
Infection		0	2 (0.6)	
Focal steatosis		0	1 (0.3)	
Low-grade dysplastic nodule		0	1 (0.3)	
LI-RADS classification	429			0.014
LR-3		25	57	
LR-4		14	52	
LR-5		45	236	

Categorical data are represented using number (percentages) and continuous data using median (interquartile range). Level of significance: $p = 0.05$ (Chi-square and Fisher exact tests for categoric variables; paired sample *t* test and Wilcoxon rank-sum test for continuous variables).
AFP, alpha-foetoprotein; cHCC-CCA, combined hepatocellular-cholangiocarcinoma; FNH, focal nodular hyperplasia; HCC, hepatocellular carcinoma; iCCA, intrahepatic cholangiocarcinoma; NASH, non-alcoholic steatohepatitis.
* There were 429 observations in 318 patients.

into a training set (84 observations) and a test set (345 observations). Both sets included 25 (30%)/14 (17%)/45 (54%) and 57 (17%)/52 (15%)/236 (68%) observations categorised as LR-3/LR-4/LR-5, respectively.

Performance of ML algorithm compared with independent readers

Performance in the training set is provided in the Supplementary material. In the test set, by categorising features of uncertain presence as being absent, sensitivity and specificity for non-rim APHE were the highest (0.85 and 0.96, respectively) compared with the ground truth, whereas sensitivity for enhancing capsule was the lowest (0.69) (Fig. 4). Two hundred and forty-two (70.1%) of the 345 observations were accurately categorised by the ML algorithm. Sensitivity and specificity for LR-5 were, respectively, 0.67 (95% CI, 0.62–0.72) and 0.91 (95% CI, 0.87–0.96) (Table 2 and Fig. 4), whereas sensitivity and specificity of LR-5 for HCC were 0.51 (95% CI, 0.46–0.57) and 0.78 (95% CI, 0.52–0.94), respectively. Subgroup analyses according to the observation size (<20 mm or ≥20 mm) are presented in Tables S2 and S3. Sensitivity and specificity for LR-5 were, respectively, significantly higher and lower in ≥20 mm observations compared with <20 mm ones (sensitivity, 0.21 [0.10–0.32] *vs.* 0.76 [0.71–0.81], p = 0.012; specificity, 1.00 [0.93–1.00] *vs.* 0.83 [0.74–0.91], p = 0.039).

A majority of under-classified observations (56/92, 60.9%) had at least one major feature whose presence was considered uncertain according to the ML algorithm (Fig. 5), whereas three of 11 (27.3%) over-classified observations had at least one major feature of uncertain presence. The performance of the ML algorithm in the subgroup of observations without any uncertain major feature (n = 222) are detailed in the Supplementary material.

Compared with the ground truth, the sensitivity of the two readers was greater than 0.90 for non-rim APHE, significantly higher than that of the ML algorithm, whereas specificity was lower than that of the ML algorithm, with a significant difference for reader 1. Overall, 230 (66.7%) and 268 (77.7%) of the 345 observations were accurately categorised by reader 1 and reader

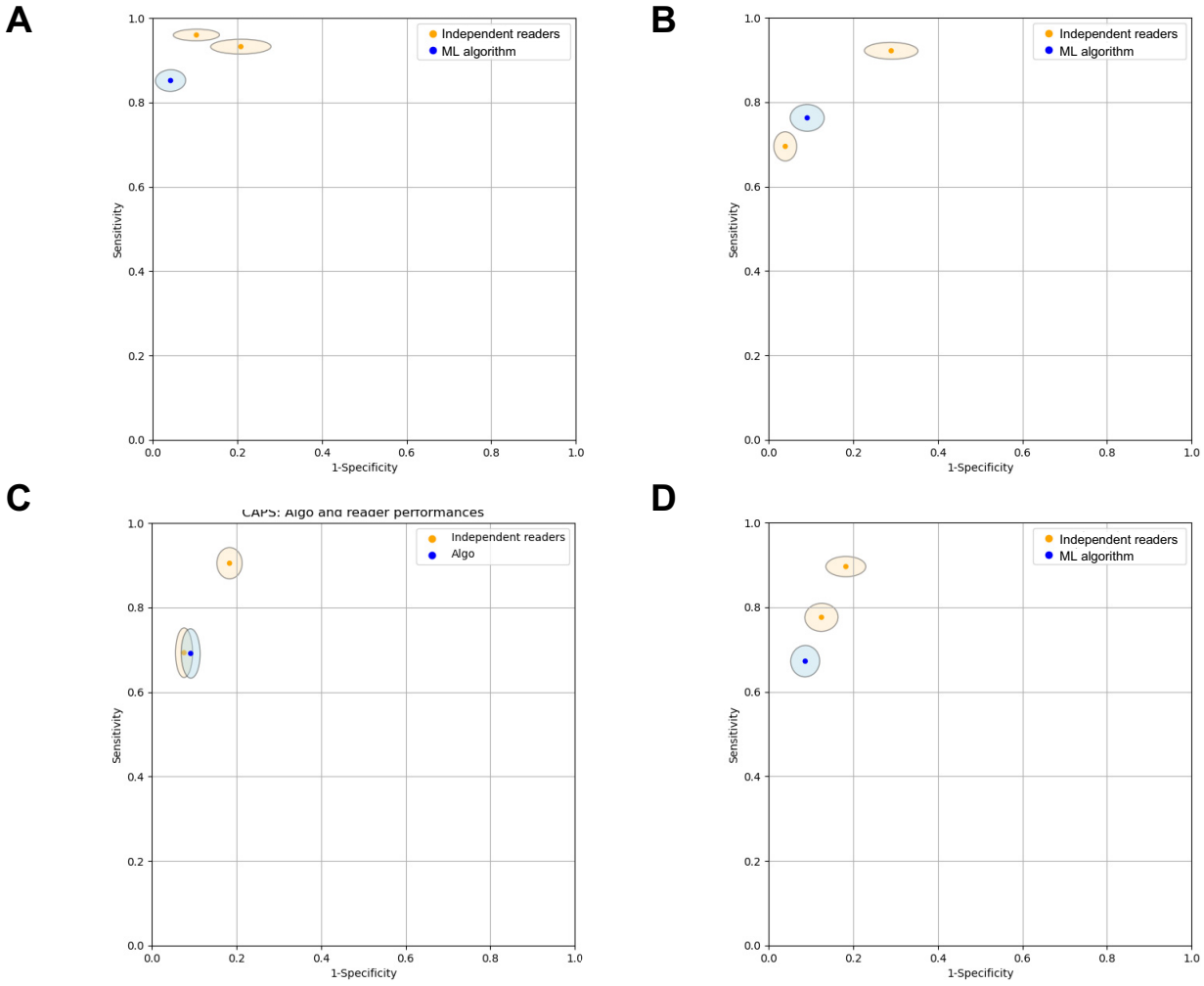


Fig. 4. Sensitivity and specificity. (A) non-rim arterial phase hyperenhancement, (B) non-peripheral washout, (C) enhancing capsule, and (D) LR-5 observation categorisation of ML algorithm and independent readers in the test set assessed using bootstrap resampling. Ellipses represent 95% confidence intervals. LR, LI-RADS; ML, machine learning.

Table 2. Performance of machine learning (ML) algorithm and independent readers in the test set.

LI-RADS features	ML algorithm (n = 345)	Independent reader 1 (n = 345)	Independent reader 2 (n = 345)
Non-rim APHE			
Present/absent/uncertain*	255/64/26	287/58	289/56
Sensitivity	0.85 (0.82–0.88)	0.93 (0.91–0.96)	0.96 (0.94–0.98)
Specificity	0.96 (0.91–1.00)	0.79 (0.69–0.89)	0.90 (0.82–0.97)
Non-peripheral washout			
Present/absent/uncertain*	212/83/50	190/155	269/76
Sensitivity	0.76 (0.72–0.81)	0.70 (0.66–0.74)	0.92 (0.89–0.95)
Specificity	0.91 (0.85–0.96)	0.96 (0.93–0.99)	0.71 (0.62–0.80)
Enhancing capsule			
Present/absent/uncertain*	88/192/65	131/214	84/261
Sensitivity	0.69 (0.61–0.77)	0.91 (0.85–0.95)	0.69 (0.61–0.77)
Specificity	0.91 (0.88–0.94)	0.82 (0.78–0.86)	0.92 (0.90–0.95)
LI-RADS classification, n (%)			
LR-3	97 (28.1)	40 (11.6)	41 (11.9)
LR-4	77 (22.3)	50 (14.5)	53 (15.4)
LR-5	171 (49.6)	197 (57.1)	221 (64.1)
LR-1 or LR-2	–	56 (16.2)	0 (0.0)
LR-M	–	1 (0.3)	24 (7.0)
LR-TIV	–	1 (0.3)	6 (1.7)
LR-5 performance			
Sensitivity	0.67 (0.62–0.72)	0.78 (0.74–0.82)	0.90 (0.87–0.93)
Specificity	0.91 (0.87–0.96)	0.88 (0.83–0.93)	0.82 (0.77–0.86)

Numbers in parentheses are the 95% confidence interval.
APHE, arterial phase hyperenhancement; LI-RADS, Liver Imaging Reporting and Data System; LR, LI-RADS; TIV, tumour-in-vein.
* The ML algorithm categorised each major feature as present, absent or of uncertain presence. The 'uncertain' category is not applicable to readers.

2. The complete description of the visual assessment of the independent readers in the test set is reported in Table 2.

Place of the ML algorithm in the diagnostic pathway of hepatic observations

Triage scenarios

The visual assessment by independent readers 1 and 2 to categorise LR-3 and LR-4 observations according to the ML algorithm led to a sensitivity and specificity for LR-5 of 0.86 (95% CI, 0.82–0.90) and 0.82 (95% CI, 0.76–0.88) (reader 1), and of 0.93 (95% CI, 0.88–0.98) and 0.76 (95% CI, 0.70–0.82) (reader 2). Two hundred and sixty-nine (78.0%) (reader 1) and 284 (82.3%) (reader 2) of the 345 observations were accurately categorised.

The visual assessment by independent readers to categorise only observations with at least one major feature of uncertain presence according to the ML algorithm leaded to a sensitivity

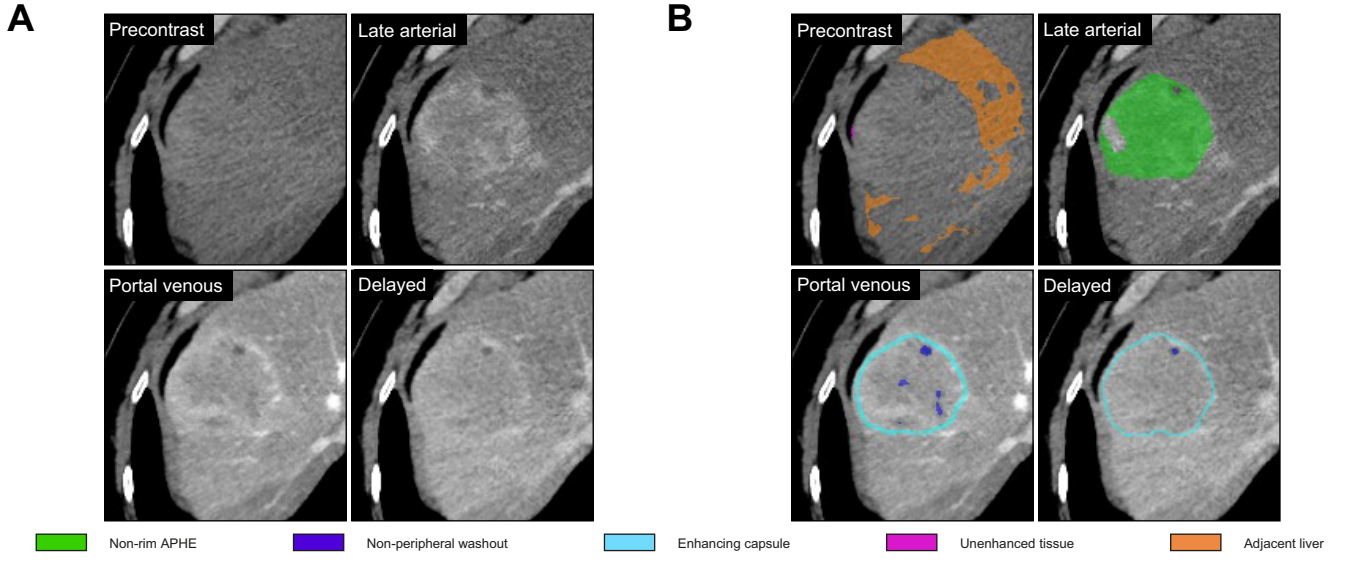


Fig. 5. Axial multiphasic CT images in a 74-year-old man with NASH-related cirrhosis and 63-mm hepatocellular carcinoma in segment 8 of the liver. Multiphasic images are presented (A) without and (B) with superimposition of areas in which non-rim arterial phase hyperenhancement (APHE), non-peripheral washout, and enhancing capsule were detected by the machine learning algorithm. The adjacent liver parenchyma automatically segmented for the analysis of the observation enhancement patterns is also shown. The three major features were considered present according to the ground truth (LR-5 observation) and by the two independent readers except enhancing capsule considered absent by one independent reader (LR-5 observation for both readers). The ML algorithm considered non-rim APHE present but washout and enhancing capsule of uncertain presence, leading to LR-4 categorisation. CT, computed tomography; LR, LI-RADS; ML, machine learning; NASH, non-alcoholic steatohepatitis.

and specificity for LR-5 of 0.77 (95% CI, 0.73–0.81) and 0.88 (95% CI, 0.83–0.93) (reader 1), and of 0.80 (95% CI, 0.75–0.85) and 0.87 (95% CI, 0.83–0.91) (reader 2). Two hundred and seventy-three (79.1%) (reader 1) and 262 (75.9%) (reader 2) of the 345 observations were accurately categorised.

Add-on scenario

Applying the ML algorithm to further categorise LR-3 and LR-4 observations according to the visual analysis of the independent readers led to a sensitivity and specificity for LR-5 of 0.84 (95% CI, 0.79–0.89) and 0.85 (95% CI, 0.79–0.91) (reader 1) and of 0.88 (95% CI, 0.84–0.92) and 0.77 (95% CI, 0.71–0.83) (reader 2). Two hundred and sixty-three (74.2%) (reader 1) and two hundred and seventy-one (78.6%) (reader 2) of the 345 observations were accurately categorised.

Inter-reader agreement

The values of inter-reader agreement between pairs of readers are summarised in Fig. 6A. Inter-reader agreement for all readers was 0.70 for non-rim APHE, 0.50 for non-peripheral washout, and 0.55 for enhancing capsule. The overall inter-reader agreement was high in 95%, 85%, and 86% of observations for non-rim APHE, non-peripheral washout, and enhancing capsule, respectively. The overall inter-reader agreement was more frequently low for major features classified as uncertain by the ML algorithm compared with those classified as present or absent, whatever the major feature (Fig. 6B).

Discussion

Our study demonstrates the ability of an ML algorithm to assess LI-RADS v2018 major features and categorise liver observations at CT in high-risk patients. The ML algorithm evaluated whether non-rim APHE, non-peripheral washout and enhancing capsule were present, absent or of uncertain presence. Sensitivity and specificity for non-rim APHE were the highest, while sensitivity for enhancing capsule was the lowest. LR-5 observations were diagnosed with 0.91 specificity and 0.67 sensitivity, reaching respectively 0.92 and 0.79 when features of uncertain presence were not considered. When used as a triage tool before visual analysis by radiologists or used as add-on in the subgroup of patients with LR-3 or LR-4 lesions according to the radiologists, the sensitivity for LR-5 significantly increased (0.84–0.93), associated with a mild decrease in specificity (range 0.76–0.85), but a final higher percentage of accurately categorised lesions (range 74.2–82.3%).

The LI-RADS diagnostic algorithm aims at standardising the imaging diagnosis of HCC, based on both major and ancillary imaging features. Hence, accurate assessment of those features is mandatory to warrant satisfying HCC diagnostic accuracy in patients at high risk for HCC. Methods to quantitatively assess APHE and washout appearance at CT or MRI have been previously proposed.[10,11,15,16] However, the quantitative analysis of major features was limited to 2D circular ROIs, preventing an accurate analysis of the observation heterogeneity and potentially failing to detect presence of the major features, as APHE and washout appearance do not need to coincide in the same observation part.[17] Deep learning-based approaches were also proposed for automated classification of liver observations, requiring large amounts of available data.[18] In the present study, an ML algorithm was developed, requiring far less data for the training process and thus allowing the majority of data to be dedicated to the model validation process. Moreover, bootstrap resampling was implemented to assess the precision and confidence of the results obtained.

The ML algorithm presented in this study detects the presence of non-rim APHE, non-peripheral washout, and enhancing capsule by estimating a probability of presence associated with each one of these features. An original and major strength of this study is the definition of an uncertainty area in which probability values were considered too close to the cut-off (i.e., in the scrutiny zone) to confidently affirm or reject the presence of the features. In that case, the presence of the feature was stated as

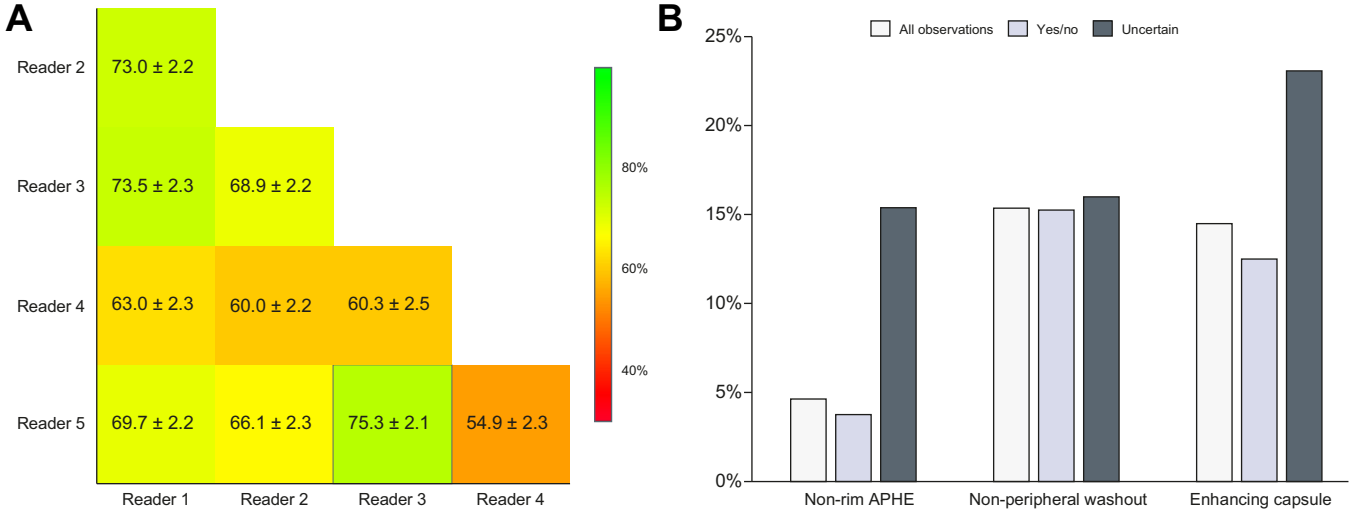


Fig. 6. Inter-reader agreement. (A) Inter-reader agreement between pairs of readers in the test set (Cohen's kappa). (B) Percentage of observations of the test set with low overall inter-reader agreement, according to their categorisation by the machine learning algorithm (present, absent, or uncertain).

uncertain. Such an approach allows first to identify the most challenging cases and to warn readers about it. Noteworthy, the agreement between independent readers was significantly lower in features of uncertain presence compared with those unequivocally categorised as present or absent by the ML algorithm, suggesting that features of uncertain presence were more challenging for readers to analyse. Hence, in a potential triage application of automatic analysis of observation enhancement features, our approach would allow to accurately separate observations with typical enhancement profiles from those that would require further expert reader visual analysis. Second, the introduction of an uncertainty area may also prevent from overdiagnosis of HCC. This point is of paramount importance, as one of the aims of the LI-RADS v2018 algorithm is to maximise specificity for HCC by applying major features only when their presence is unequivocal.[17]

Performances of the ML algorithm for LR-5 were similar to those of the independent readers when features of uncertain presence were excluded, both in terms of sensitivity and specificity. When features of uncertain presence were considered absent, the specificity for LR-5 remained as high as that of independent readers, while an expected decrease in sensitivity – from 0.79 to 0.67 – was seen. However, the major point in identifying features of uncertain presence was to highlight observations that would require further expert reader visual analysis, leading to a combined analysis of the observation by both the ML algorithm and the reader. To achieve these goals, the optimal integration of such artificial intelligence applications in the clinical pathway should be defined.[19] Rather than replacing radiologists, our results highlight the possible value of using the developed artificial intelligence (AI) algorithm as a triage tool or as an add-on to the radiologist's visual analysis, highlighting the potential benefit from the radiologist–AI interaction in improving focal liver lesions characterisation in patients at risk for HCC. We believe that such an AI-enriched diagnostic pathway may help standardise and improve the quality of analysis of liver lesions in patients at high risk for HCC, especially in non-expert centres in liver imaging. Because our algorithm provides a probability of HCC for each analysed lesion, it may also impact the clinical decision-making and guide the clinician in identifying the lesions to be biopsied, for instance in patients with multiple liver focal lesions.

The dataset split was performed according to developed model. Indeed, our ML model is a logistic regression model with only few physical features as input (<10 parameters). As the performance of traditional ML techniques grows according to a power law and then reaches a plateau after a certain quantity of input training samples, our model was trained on 82 representative patients using stratified sampling. With our data augmentation technique used, the model was trained on 4,100 samples which was enough to fit the hyperplane of $R^{N<10}$. This splitting strategy allowed us to keep the large majority of the data in the test set to reliably validate the model and to reduce the 95% CIs, which in practice increases the generalisability of the model.

Some limitations should be noted. First, this is a retrospective study, which may cause selection bias. However, multiple centres from different Western and Asian countries were involved, which may reinforce the external validity of our results in different populations. Second, only pathologically proven observations were included with a large majority of HCC, which may have an impact on the evaluated diagnostic performance of both the developed ML algorithm and the independent readers. However, considering pathology-proven lesions is of paramount importance when evaluating the diagnostic performance of a new algorithm.

Further improvements of this work may include a better characterisation of APHE and washout spatial distribution to distinguish between rim and non-rim APHE, and between peripheral and non-peripheral washout, as they contain diagnostic and prognostic information.[20] The ability of the developed ML algorithm to further improve readers performances has also to be investigated. Last, as for all AI applications, the ML algorithm is bound to be evolutive, to match potential evolutions of LI-RADS criteria and also to potentially include MR data.

In conclusion, our study demonstrates that the proposed ML algorithm can assess LI-RADS v2018 non-rim APHE, non-peripheral washout, and enhancing capsule and categorise liver observations at CT which could help radiologists to standardise their reports according to the latest recommendations. Rather than replacing radiologists, our results highlight the potential benefit from the radiologist–AI interaction in improving focal liver lesions characterisation in patients at risk for HCC by using the developed ML algorithm as a triage tool or as an add-on to the radiologist's visual analysis. The proposed ML algorithm is a step towards a more robust and automated analysis of focal liver lesions in patients at risk of HCC through a wider use of the LI-RADS algorithm even in non-expert centres. Our algorithm may also impact the clinical decision and guide the clinician in identifying the lesions to be biopsied.

Abbreviations

2D, two-dimensional; 3D, three-dimensional; AFP, alpha-foetoprotein; APHE, arterial phase hyperenhancement; cHCC-CCA, combined hepatocellular-cholangiocarcinoma; CT, computed tomography; FNH, focal nodular hyperplasia; HCC, hepatocellular carcinoma; iCCA, intrahepatic cholangiocarcinoma; LI-RADS, Liver Imaging Reporting and Data System; ML, machine learning; MRI, magnetic resonance imaging; NASH, non-alcoholic steatohepatitis; ROIs, regions of interest.

Financial support

This work was supported by a national grant by Bpifrance with financial support by GE Healthcare France. Authors acknowledge Medicen Pôle de compétitivité Paris - Ile de France, domaine d'action stratégique imagerie médicale and Bpifrance for their support in promoting the HECAM project.

Conflicts of interest

Five authors (VM, RQ, LD, JHL and NC) are GE Healthcare France employees. The remaining authors, who are not employees of or consultants for GE Healthcare, had control of inclusion of all data and information that might present a conflict of interest for authors who are employees for GE Healthcare.

Please refer to the accompanying ICMJE disclosure forms for further details.

Authors' contributions

Conceptualisation: SM, MR, JHL, NC, VV, AL. Data curation: SM, MR, MG, RS, GC, ER, VM, RQ, LD. Formal analysis: MG, RS, GC, VM. Methodology: SM, MR, JHL, NC, VV, AL. Writing – original draft: SM, MR. Writing – review and editing: all authors.

Data availability statement

The data shown in this article are available from the corresponding authors upon a reasonable request.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jhepr.2023.100857.

References

Author names in bold designate shared co-first authorship

[1] Marrero JA, Kulik LM, Sirlin CB, Zhu AX, Finn RS, Abecassis MM, et al. Diagnosis, staging, and management of hepatocellular carcinoma: 2018 practice guidance by the American Association for the Study of Liver Diseases. Hepatology 2018;68:723–750.

[2] Tang A, Bashir MR, Corwin MT, Cruite I, Dietrich CF, Do RKG, et al. Evidence supporting LI-RADS major features for CT- and MR imaging-based diagnosis of hepatocellular carcinoma: a systematic review. Radiology 2018;286:29–48.

[3] Chernyak V, Flusberg M, Law A, Kobi M, Paroder V, Rozenblit AM. Liver imaging reporting and data system: discordance between computed tomography and gadoxetate-enhanced magnetic resonance imaging for detection of hepatocellular carcinoma major features. J Comput Assist Tomogr 2018;42:155–161.

[4] Haimerl M, Wächtler M, Zeman F, Verloh N, Platzek I, Schreyer AG, et al. Quantitative evaluation of enhancement patterns in focal solid liver lesions with Gd-EOB-DTPA-enhanced MRI. PLoS One 2014;9:e100315.

[5] Kang JH, Choi SH, Lee JS, Kin KW, Kim Sy, Lee SS, et al. Inter-reader reliability of CT Liver Imaging Reporting and Data System according to imaging analysis methodology: a systematic review and meta-analysis. Eur Radiol 2021;31:6856–6867.

[6] Rhee H, An C, Kim H-Y, Yoo JE, Park YN, Kim MJ. Hepatocellular carcinoma with irregular rim-like arterial phase hyperenhancement: more aggressive pathologic features. Liver Cancer 2019;8:24–40.

[7] Seo N, Kim M-J, Rhee H. Hepatic sarcomatoid carcinoma: magnetic resonance imaging evaluation by using the liver imaging reporting and data system. Eur Radiol 2019;29:3761–3771.

[8] Ehman EC, Behr SC, Umetsu SE, Fidelman N, Yeh BM, Ferrell LD, et al. Rate of observation and inter-observer agreement for LI-RADS major features at CT and MRI in 184 pathology proven hepatocellular carcinomas. Abdom Radiol (NY) 2016;41:963–969.

[9] Cannella R, Ronot M, Sartoris R, Cauchy F, Hobeika C, Beaufrere A, et al. Enhancing capsule in hepatocellular carcinoma: intra-individual comparison between CT and MRI with extracellular contrast agent. Diagn Interv Imaging 2021;102:735–742.

[10] Stocker D, Becker AS, Barth BK, Skawran S, Kaniewska M, Fischer MA, et al. Does quantitative assessment of arterial phase hyperenhancement and washout improve LI-RADS v2018-based classification of liver lesions? Eur Radiol 2020;30:2922–2933.

[11] Liu YI, Shin LK, Jeffrey RB, Kamaya A. Quantitatively defining washout in hepatocellular carcinoma. Am J Roentgenol 2013;200:84–89.

[12] Allaire M, Bruix J, Korenjak M, Manes S, Maravic Z, Reeves H, et al. What to do about hepatocellular carcinoma: recommendations for health authorities from the International Liver Cancer Association. JHEP Rep 2022;4:100578.

[13] Hamm CA, Wang CJ, Savic LJ, Ferrante M, Schobert I, Schlachter T, et al. Deep learning for liver tumor diagnosis part I: development of a convolutional neural network classifier for multi-phasic MRI. Eur Radiol 2019;29:3338–3347.

[14] Nam D, Chapiro J, Paradis V, Seraphin TP, Kather JN. Artificial intelligence in liver diseases: improving diagnostics, prognostics and response prediction. JHEP Rep 2022;4:100443.

[15] Fronda M, Doriguzzi Breatta A, Gatti M, Calandri M, Maglia C, Bergamasco L, et al. Quantitative assessment of HCC wash-out on CT is a predictor of early complete response to TACE. Eur Radiol 2021;31:6578–6588.

[16] Kloeckner R, Pinto Dos Santos D, Kreitner K-F, Leicher-Düber A, Weinmann A, Mittler J, et al. Quantitative assessment of washout in hepatocellular carcinoma using MRI. BMC Cancer 2016;16:758.

[17] Chernyak V, Fowler KJ, Kamaya A, Kielar AZ, Elsayes KM, Bashir MR, et al. Liver imaging reporting and data system (LI-RADS) version 2018: imaging of hepatocellular carcinoma in at-risk patients. Radiology 2018;289:816–830.

[18] Wang CJ, Hamm CA, Savic LJ, Ferrante M, Schobert I, Schalchter T, et al. Deep learning for liver tumor diagnosis part II: convolutional neural network interpretation using radiologic imaging features. Eur Radiol 2019;29:3348–3357.

[19] Tang A, Tam R, Cadrin-Chênevert A, Guest W, Chong J, Barfett J, et al. Canadian Association of Radiologists white paper on artificial intelligence in radiology. Can Assoc Radiol J 2018;69:120–135.

[20] Petukhova-Greenstein A, Zeevi T, Yang J, Chai N, DiDomenico P, Deng Y, et al. MR imaging biomarkers for the prediction of outcome after radiofrequency ablation of hepatocellular carcinoma: qualitative and quantitative assessments of the liver imaging reporting and data system and radiomic features. J Vasc Interv Radiol 2022;33:814–824.e3.