European Radiology (2020) 30:2365–2376
https://doi.org/10.1007/s00330-019-06553-6

IMAGING INFORMATICS AND ARTIFICIAL INTELLIGENCE

Accurate prediction of responses to transarterial chemoembolization for patients with hepatocellular carcinoma by using artificial intelligence in contrast-enhanced ultrasound

Dan Liu[1,2] • Fei Liu[2,3] • Xiaoyan Xie[1] • Liya Su[1] • Ming Liu[1] • Xiaohua Xie[1] • Ming Kuang[1,4] • Guangliang Huang[1] • Yuqi Wang[2,3] • Hui Zhou[2,3] • Kun Wang[2,3] • Manxia Lin[1] • Jie Tian[2,5]

Received: 20 June 2019 / Revised: 17 October 2019 / Accepted: 28 October 2019 / Published online: 3 January 2020

Abstract

Objectives We aimed to establish and validate an artificial intelligence–based radiomics strategy for predicting personalized responses of hepatocellular carcinoma (HCC) to first transarterial chemoembolization (TACE) session by quantitatively analyzing contrast-enhanced ultrasound (CEUS) cines.

Methods One hundred and thirty HCC patients (89 for training, 41 for validation), who received ultrasound examination (CEUS and B-mode) within 1 week before the first TACE session, were retrospectively enrolled. Ultrasonographic data was used for building and validating deep learning radiomics-based CEUS model (R-DLCEUS), machine learning radiomics-based time-intensity curve of CEUS model (R-TIC), and machine learning radiomics-based B-Mode images model (R-BMode), respectively, to predict responses (objective-response and non-response) to TACE with reference to modified response evaluation criteria in solid tumor. The performance of models was compared by areas under the receiver operating characteristic curve (AUC) and the DeLong test was used to compare different AUCs. The prediction robustness was assessed for each model.

Results AUCs of R-DLCEUS, R-TIC, and R-BMode were 0.93 (95% CI, 0.80–0.98), 0.80 (95% CI, 0.64–0.90), and 0.81 (95% CI, 0.67–0.95) in the validation cohort, respectively. AUC of R-DLCEUS shows significant difference compared with that of R-TIC ($p = 0.034$) and R-BMode ($p = 0.039$), whereas R-TIC was not significantly different from R-BMode. The performance was highly reproducible with different training and validation cohorts.

Conclusions DL-based radiomics method can effectively utilize CEUS cines to achieve accurate and personalized prediction. It is easy to operate and holds good potential for benefiting TACE candidates in clinical practice.

Dan Liu and Fei Liu contributed equally to this work.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00330-019-06553-6) contains supplementary material, which is available to authorized users.

✉ Manxia Lin
linmxia@mail.sysu.edu.cn

✉ Jie Tian
tian@ieee.org

[1] Department of Medical Ultrasonics, Institute of Diagnostic and Interventional Ultrasound, The First Affiliated Hospital of Sun Yat-sen University, Guangzhou 510080, China

[2] CAS Key Laboratory of Molecular Imaging, Institute of Automation, Chinese Academy of Sciences, 95 Zhongguancun East Road, Beijing 100190, China

[3] Department of the Artificial Intelligence Technology, University of Chinese Academy of Sciences, NO.19(A) Yuquan Road, Shijingshan District, Beijing 100049, People's Republic of China

[4] Department of Liver Surgery, The First Affiliated Hospital of Sun Yat-sen University, Guangzhou 510080, China

[5] Beijing Advanced Innovation Center for Big Data-Based Precision Medicine, Beihang University, Beijing 100191, China

Key Points

- *Deep learning (DL) radiomics-based CEUS model can accurately predict responses of HCC patients to their first TACE session by quantitatively analyzing their pre-operative CEUS cines.*
- *The visualization of the 3D CNN analysis adopted in CEUS model provided direct insight into what computers "see" on CEUS cines, which can help people understand the interpretation of CEUS data.*
- *The proposed prediction method is easy to operate and labor-saving for clinical practice, facilitating the clinical treatment decision of HCCs with very few time costs.*

Keywords Therapeutic chemoembolization · Hepatocellular carcinoma · Ultrasonography · Deep learning

Abbreviations

AFP	Alpha-fetoprotein
AI	Artificial intelligence
AUC	Area under the receiver operating characteristic curve
BCLC	Barcelona Clinical Liver Cancer
C3D	Convolutional 3D
CECT	Contrast-enhanced CT
CEMRI	Contrast-enhanced MRI
CEUS	Contrast-enhanced ultrasound
CNN	Convolutional neural network
CR	Complete response
CT	Computed tomography
DL	Deep learning
FC	Fully connected layer
GBRT	Gradient Boosted Regression Trees
HBV	Hepatitis B virus
HCC	Hepatocellular carcinoma
HCV	Hepatitis C virus
ML	Machine learning
mRECIST	Modified Response Evaluation Criteria in Solid Tumors
MRI	Magnetic resonance imaging
NPV	Negative predictive value
PD	Progression disease
PPV	Positive predictive value
PR	Partial response
RBF	Radial based function
ROI	Region of interest
SD	Stable disease
SVM	Support vector machine
T	Training cohort
TACE	Transarterial chemoembolization
TIC	Time-intensity curve
V	Validation cohort

Introduction

Hepatocellular carcinoma (HCC) is the fifth most common malignancy and the second leading cause of cancer-related death globally [1]. Although curative therapeutic modalities, such as transplantation, resection, and ablation, are recommended for early-stage HCC, a significant proportion of HCC patients are diagnosed at intermediate stage, in which transarterial chemoembolization (TACE) is the most widely used first-line therapy [2].

Conventional TACE employs the lipiodol-chemotherapeutic agent suspension and gelatin sponge particles for interventional treatment, which is currently the most common TACE modality worldwide [3]. However, local HCC responses to this therapy are highly diverse. Even at the same Barcelona Clinical Liver Cancer (BCLC) stage B, different HCC patients generally exhibit different treatment outcomes after their first TACE session [3]. Extensive randomized clinical trials have been conducted to find out the best subsequent treatment strategies after TACE regarding different response behaviors [4–6]. Despite various kinds of therapeutic strategies that were proposed, they all reached an agreement that the local HCC response to the first TACE session is significantly correlated with its following response to subsequent therapies and also the patients' overall survival [7, 8]. Therefore, the accurate and personalized prediction of local tumor responses to the first TACE session holds critical clinical impact on the overall management of HCC patients [9].

Current approaches for the prediction of HCC responses to TACE mainly based on serological biomarkers and magnetic resonance imaging (MRI) [10–12]. Besides, some score systems based on clinical, radiological, and biological data, such as HAP-score (the hepatoma arterial embolization prognostic), STATE-score (the selection for TACE treatment), and ART-score (the Assessment for Retreatment with TACE), were proposed to predict outcomes in patients with HCCs undergoing TACE and help optimize the selection for TACE [13–15]. However, these attempts only offered very limited precision of prediction, or with a limited number of enrolled patients. Different from these studies, we hypothesized that the contrast-enhanced ultrasound (CEUS) imaging may achieve better prediction efficacy, because dynamic CEUS cines can offer outstanding temporal resolution in tracing the microcirculation perfusion of tumors [16], which is considered vital in tumor responses to TACE [17].

To better interpret CEUS, we employed artificial intelligence (AI)–based radiomics strategies [18–20]. As an emerging technology, radiomics can provide comprehensive quantifications of large amounts of image features from medical images, which has the potential to uncover disease characteristics that fail to be appreciated by human eyes [21]. These features can be user-defined and selected by machine learning (ML) methods [22], or be computer-defined and selected by deep learning (DL) methods [19]. Comparing with the amount of radiomics studies on computed tomography (CT) or MRI, ultrasound imaging–based radiomics investigations are still few reported. In this study, we enrolled 130 HCC patients at BCLC stage B who received TACE as their first-session treatment. DL- and ML-based radiomics models were specially designed and trained by dynamic CEUS cines and static B-mode images, respectively, in order to predict individualized tumor response to TACE.

Materials and methods

Patient eligibility

For this retrospective study, ethical approval was conformed, and the requirement for informed consent was waived off. From December 2012 to December 2017, 138 consecutive patients with HCC who underwent CEUS examination before TACE were recruited. HCCs were diagnosed with reference to EASL guidelines [9]. Inclusion criteria were aged 18–80 years old, BCLC stage B, first-session TACE for target tumor without previous local-regional treatment, platelet count ≥ 50 × 10^9/L, and prothrombin time < 21 s. Eight patients were excluded due to poor imaging quality ($n = 3$) or excessive respiratory motion ($n = 5$) during CEUS. Finally, B-mode images and CEUS cines of 130 patients acquired pre-TACE were retrospectively analyzed (Fig. 1).

CEUS data acquisition

CEUS examinations were performed within the week before TACE because of a project using CEUS to assess the treatment response to TACE of HCC using Modified Response Evaluation Criteria in Solid Tumors (mRECIST) [23]. Examinations were performed by two radiologists who had over 10 years' experience in liver CEUS. There were three different ultrasound instruments used in this study (Philips iU22, Toshiba Aplio, e-saote). Before CEUS, B-mode images of lesions were acquired. One-minute minimum continuous cine was acquired after injecting 2.4 mL of the second-generation contrast agent (SonoVue, Bracco Imaging) via the elbow followed by a 5-mL saline flush. For multiple tumors, patients received additional administrations of SonoVue (range, 2–3 injections; median, 3 injections) to make sure each tumor was observed.

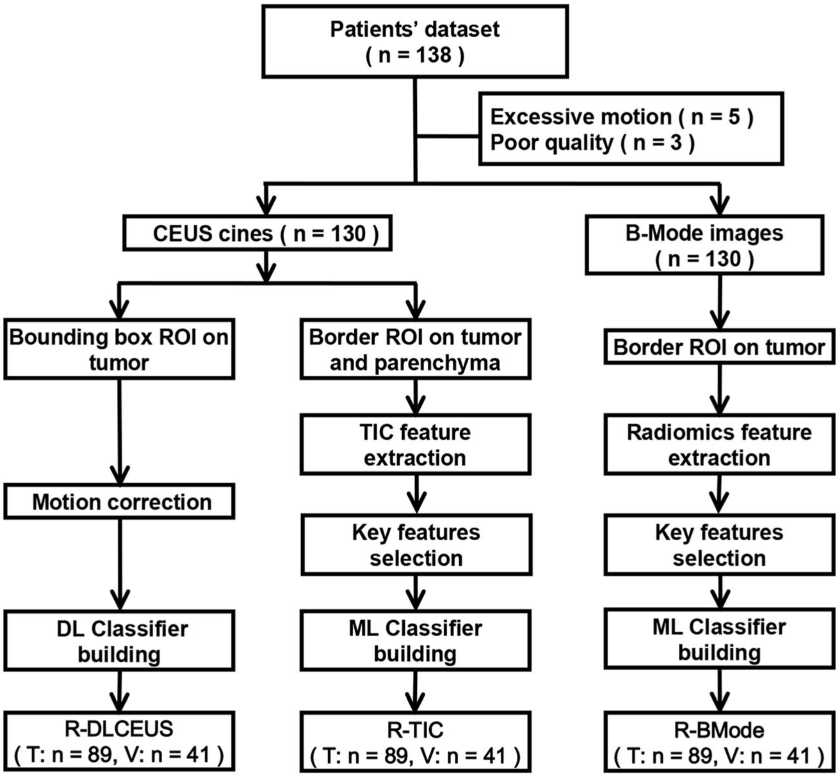

Fig. 1 Patient enrollments and radiomics model building flowchart. DL-based R-DLCEUS model was built from dynamic CEUS cines. Pre-defined TIC feature–based R-TIC model was built from dynamic CEUS cines. Pre-defined radiomics feature–based R-BMode model was built from static B-mode images. CEUS, contrast-enhanced ultrasound; DL, deep learning; ML, machine learning; ROI, region of interest; TIC, time-intensity curve; T, training cohort; V, validation cohort

TACE and tumor response assessment

TACE was performed by a radiologist with over 10 years' experience in interventional cancer therapy. Using a transfemoral approach, 5-Fr Yashiro catheters (Terumo) and microcatheters were used to superselectively catheterize the relevant arteries. Fluoroscopic guidance was used to slowly inject a mixture of liquefied lipiodol (Guerbet Laboratories) and epirubicin (Pfizer Inc.) before the embolization performed by using gelatin sponge particles (Bi-Trumed Biotech Co., Ltd.). The quantity of lipiodol (maximum volume, 20 mL), epirubicin (range, 40–80 mg), and gelatin sponge particles was selected based on patients' liver function, tumor size, and diameter of the relevant arteries. The endpoint of embolization was defined as the contrast of feeding artery had not been cleaned up after 5 heartbeats [24]. Contrast-enhanced CT (CECT) or contrast-enhanced MRI (CEMRI) performed 1 month after TACE was used to assess tumor response. The lesion with the largest diameter was selected as the target tumor. Local tumor response of this target tumor was assessed using the mRECIST guidelines [25]. Objective-response was defined as the sum of complete response (CR) and partial response (PR). Non-response was defined as the sum of stable disease (SD) and progression disease (PD) [26]. Tumor responses were interpreted in consensus by two radiologists before AI analysis, both of whom had 10 years of experience with abdominal CT/MRI interpretation.

Tumor segmentation and preprocessing

We retrospectively reviewed the TACE procedure records and CEUS documents; the largest tumor that has received TACE was selected as the target tumor. Then, the target tumor was manually segmented. Using ITK-SNAP software [27], tumor segmentation was performed by an ultrasound practitioner with work experience of 5 years, who was blinded to the result of tumor response during segmentation process. Previous studies revealed that the parameters of CEUS during wash-in were significant in assessing responses to TACE for HCC [17]; therefore, we adopted 1-min continuous CEUS for analysis [28].

In this study, three radiomics models were established (Fig. 1), which were the DL model (named R-DLCEUS) and the ML model (named R-TIC) trained by dynamic CEUS cines and another ML model (named R-BMode) trained by static B-mode images. For building R-DLCEUS model, a bounding box (Fig. 2a, the yellow box) was delineated as a region of interest (ROI) on a CEUS frame with a clear tumor margin. Following motion correction [29], the CEUS cines were cropped to keep only the bounding box ROI region for DL analysis. For the R-TIC model, two ROIs were manually delineated on all CEUS frames over the whole cine. One ROI delineated the tumor border accurately (Fig. 2a, the red contour), and the other was defined inside the normal liver parenchyma close to the tumor, warding off major blood vessels (> 3 mm diameter) and ribs (Fig. 2a, the green contour). For the R-BMode model, tumors were segmented manually around the border on three B-mode images acquired before the injection of ultrasound contrast agent (Fig. 2b, the blue contour).

Radiomics feature extraction and model building

Before data analysis, we randomly select two-thirds of the 130 enrolled patients (n = 89) as the training cohort for model training and the rest (n = 41) as the validation cohort to evaluate the predictive performance of radiomics models. This randomization was repeated several times until no significant difference in baseline characteristics (Table 1) was found between training and validation cohorts. Flow charts of the DL model design (R-DLCEUS) and ML models design (R-TIC and R-BMode) are illustrated in Fig. 2c and Supplementary Figure S2, respectively.

R-DLCEUS was built to analyze dynamic CEUS cines by a variant of convolutional 3D (C3D) neural network (CNN) method (Fig. 2c) [30], which was specially designed to analyze dynamic CEUS cines automatically. The reason why we preferred 3D CNN to 2D CNN was that C3D is simpler and more efficient in learning spatiotemporal features specific for predicting responses to TACE. Compared with C3D, 2D CNN can only analyze the spatial features from one frame of CEUS cines. We have to add more modules to modify 2D CNN in order to analyze temporal features from CEUS [31–33], which inevitably generates more parameters and hinders its overall learning performance. Our CNN model consisted of five convolution blocks for feature extraction and one perceptron for classification. The model was trained by stochastic gradient descent (SGD) solver [34] with a cross-entropy loss function [34]. For reducing overfitting, depthwise convolution [35] and global pooling [36] were used to reduce learning parameters, and video frame sampling and data augmentation [34] were conducted to improve its generalization ability. More detailed mathematical definitions and discerptions are provided in the Supplementary Method. After the automatic quantitative analysis, probabilities for each category (objective-response and non-response) were obtained for each CEUS cine.

R-TIC was established based on time-intensity curve (TIC) features extracted from dynamic CEUS cines. We defined and extracted 18 TIC features from manually segmented CEUS tumor ROIs and normal liver parenchyma ROIs of 1-min CEUS cines. Detailed definitions and calculation method were described in Supplementary method and Supplementary Table S3. The predictive contribution of each feature was measured by a preliminary model built by applying the Gradient Boosted Regression Trees (GBRT) algorithm [37]. TIC features with larger contribution were selected as key features. Based on those selected key features, a classifier for predicting TACE response was built by using the support vector machine (SVM) model with radial based function (RBF) kernel [38]. Cross-validation was applied to determine the number of key features and the basic structure parameters of GBRT and SVM [39]. The

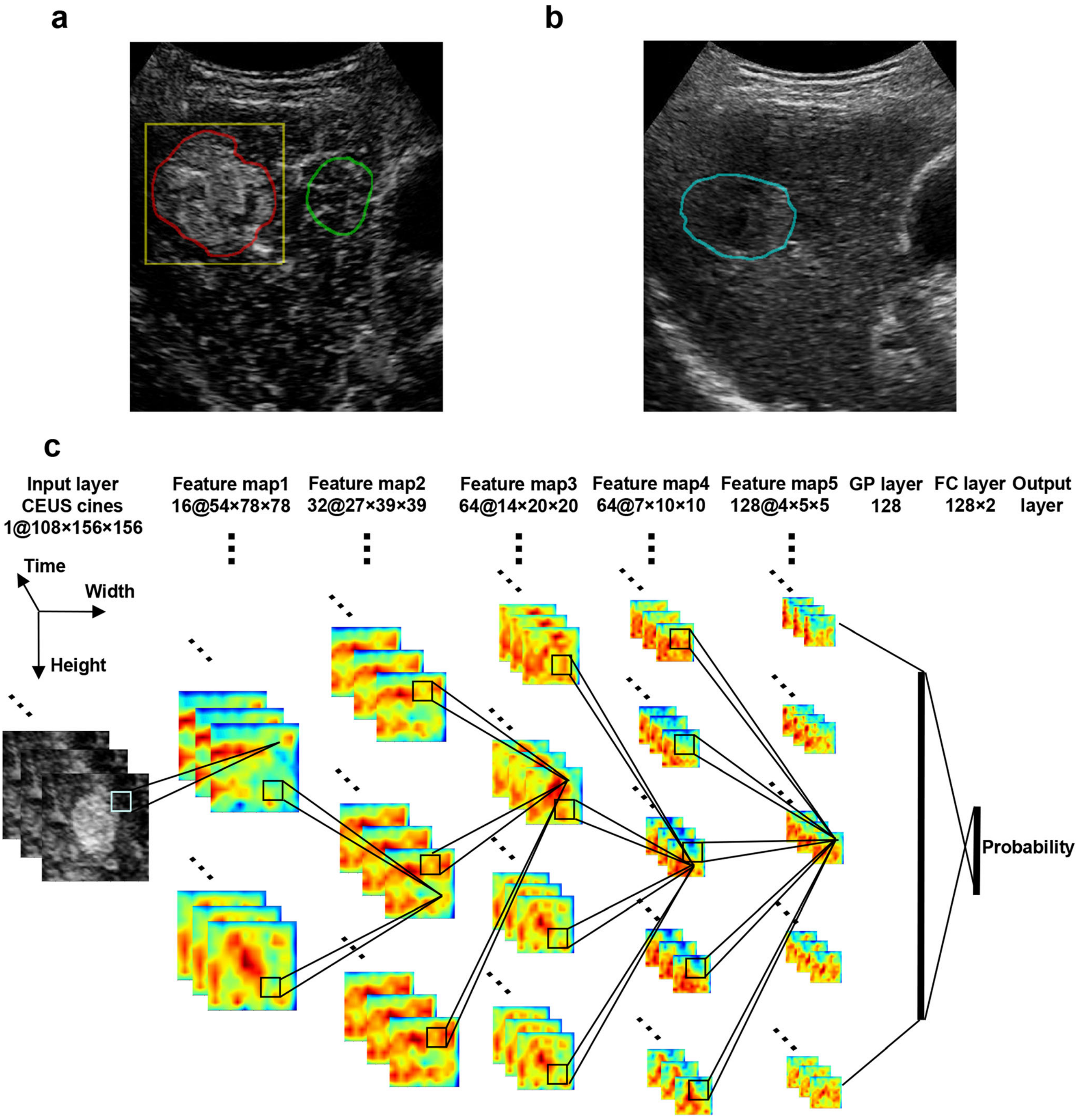

Fig. 2 The ROI segmentation and the design of R-DLCEUS. **a** An example of different ROIs in CEUS cines. Before segmentation, the cines and images were converted from color maps to gray scales for loading in ITK-SNAP. The yellow bounding box ROI contained the whole tumor, which defined the input of R-DLCEUS model. The red ROI was the manual delineation of tumor. The green ROI was a delineation of surrounding normal liver parenchyma. They were used to extract TIC features for R-TIC model. **b** An example of ROI in B-mode image. The blue ROI was the manual delineation of tumor for R-BMode model. **c** The overall topology structure of R-DLCEUS model. A CEUS cine was inputted into the CNN, and after applying five times auto-learning transformation of DL layers, the resulted feature maps are used to calculate the prediction probability. At each CNN layer, the inputted feature maps were computed to another set of more discriminating feature maps by automatically learning the way of transformation. ROI, region of interest; DL, deep learning; CEUS, contrast-enhanced ultrasound; TIC, time-intensity curve; CNN, convolutional neural network; GP layer, global average pooling layer; FC layer, fully connected layer

violin plot was drawn to show the distribution of selected key features in different categories. In a violin plot, the more obvious distribution difference of two categories reveals the more significant classification ability of a particular feature.

Table 1 Baseline characteristics of patients

Characteristic	All patients	Training cohort	Validation cohort	*p* value
Patient number (%)	130	89 (68.5%)	41 (31.5%)	-
Age, years	55.0 ± 12.3 (18–80)	55.6 ± 12.2 (18–80)	53.9 ± 12.7 (18–72)	0.467
Sex				0.812
Male	106 (81.5%)	73 (82.0%)	33 (80.5%)	
Female	24 (18.5%)	16 (18.0%)	8 (19.5%)	
Etiology				0.301
HBV	109 (83.8%)	74 (83.1%)	35 (85.4%)	
HCV	6 (4.6%)	4 (4.5%)	2 (4.9%)	
Unknown/other	15 (11.6%)	11 (12.4%)	4 (9.7%)	
Performance status				0.546
0	117 (90.0%)	81 (91.0%)	36 (87.8%)	
1	13 (10%)	8 (9.0%)	5 (12.2%)	
Child-Pugh class				0.677
A	123 (94.6%)	85 (95.5%)	38 (92.7%)	
B	7 (5.4%)	4 (4.5%)	3 (7.3%)	
AFP, ng/mL				0.109
< 20	37 (28.5%)	25 (28.1%)	12 (29.2%)	
20–200	37 (28.5%)	23 (25.8%)	14 (34.1%)	
≥ 200	56 (43.0%)	41 (46.1%)	15 (36.7%)	
Tumor number				0.648
Unifocal	28 (21.5%)	18 (20.2%)	10 (24.4%)	
Multifocal	102 (78.5%)	71 (79.8%)	31 (75.6%)	
Tumor size, cm	4.2 ± 3.2 (1.0–15.8)	4.3 ± 3.2 (1.0–15.8)	4.1 ± 3.3 (1.0–14.2)	0.496
Tumor location				0.321
Right lobe	90 (69.2%)	65 (73.0%)	25 (60.9%)	
Left lobe	32 (24.6%)	20 (22.5%)	12 (29.2%)	
Bilobar	8 (6.2%)	4 (4.5%)	4 (9.9%)	
Tumor response				0.846
Objective-response	46 (35.4%)	31 (34.8%)	15 (36.6%)	
Non-response	84 (64.6%)	58 (65.2%)	26 (63.4%)	

Data are shown as means ± standard deviation (range) or number (%). Student's *t* test was used to compare continuous variables. The χ^2 test was used to compare categorical variables. *HBV*, hepatitis B virus; *HCV*, hepatitis C virus; *AFP*, alpha-fetoprotein

R-BMode was constructed based on radiomics features extracted from static B-mode images (Fig. 1). We averaged B-mode images and blurred in order to reduce speckle [40]. In total, we extracted 934 user-defined radiomics features, i.e., 181 statistics features, 13 tumor shape features, and 740 texture features. The key feature selection and model building method were the same as R-TIC. Due to the limited numbers of B-mode images and their lack of hemodynamic perfusion information compared with CEUS cines, it was unsuitable to analyze B-mode images by using the DL method in this study. Therefore, we did not include DL analysis of the B-mode image.

Radiomics model accuracy and robustness assessment

Using the dataset split strategy as described above, the enrolled 130 patients were randomly split into three different training cohorts and validation cohorts (89 vs. 41, 85 vs. 45, and 86 vs. 44) [19]. Three validation cohorts were not overlapped. Each of the three radiomics models was trained three times, and the corresponding prediction accuracy was assessed separately in three validation cohorts. The variation of prediction accuracy indicated the robustness of each model.

HAP-score assessment and comparison with AI models

The STATE-score could not be calculated since C-reactive protein was not a routine test in our center. ART-score was calculated prior to the second TACE session [15], which was inconsistent with the design of this study. Therefore, we assessed the prediction accuracy of the HAP-scores based on albumin, bilirubin, α-fetoprotein, and the size of dominant tumor [13]. AUCs of HAP-score were compared with those of our proposed models in the training and validation cohorts, respectively.

DL model visualization and its acceptability of human readers

For a better understanding of how the DL model predicts different treatment responses, we converted DL feature maps into pseudo-colored maps using Selvaraju R.'s method [41]. To validate the value of DL feature maps in clinical practice, firstly we concluded the general criteria of the corresponding relation between the characteristics of DL feature maps and responses to TACE. Then, two doctors (a PhD student and a radiologist with more than 10 years of experience) were told the criteria and asked to predict the responses to TACE based on the criteria. Finally, we computed two doctors' prediction accuracy and their statistical significance using NRI test [42]. We also compared the prediction accuracy of two doctors with the three AI models in area under the receiver operating characteristic curves (AUC).

Statistical analysis

Student's *t* test or the Mann-Whitney test, as appropriate, was used to compare continuous variables. The χ^2 test was used to compare categorical variables. AUC was used to estimate the probability of correct prediction of objective-response and non-response to TACE. Differences between AUCs were calculated using the DeLong test. Data analysis was performed using GraphPad Prism (version 5.0; GraphPad Software). Model building and evaluation were conducted using python 2.7, pytorch (version 0.4.0), pyradiomics (version 2.0.0) [43], scipy, lmfit, and scikit-sklearn package. All statistical tests were two-sided. Differences were considered significant at $p < 0.05$.

Results

Baseline characteristics

Three different ultrasound instruments were used to examine the patients: Philips iU22 ($n = 37$), Toshiba Aplio ($n = 70$) and e-saote ($n = 23$). Tumor response was assessed using CECT in 80 (61.5%) patients and CEMRI in 50 (38.5%) patients. The numbers of objective-response and non-response HCC patients to TACE were 46 (35.4%) and 84 (64.6%), respectively. The objective-response rate to TACE in our study was in accordance with previous report [44]. There were no significant differences in the variables between the training and validation cohorts (Table 1).

Prediction accuracy and robustness

In the training cohort, AUCs of R-DLCEUS, R-TIC, and R-BMode reached 0.98, 0.84, and 0.82, respectively. DL-based radiomics model using dynamic CEUS cines (R-DLCEUS) achieved the best prediction performance compared with two ML models (R-TIC and R-BMode) (Fig. 3a, b). The same results were confirmed in the validation cohort; AUCs of R-DLCEUS, R-TIC and R-BMode were 0.93, 0.80, and 0.81, respectively (Fig. 3c, d). Table 2 summarizes the quantitative analysis of all three models. R-DLCEUS offered significantly higher AUCs in comparison with R-TIC and R-BMode in both the training ($p = 0.002$, $p = 0.001$) and validation cohorts ($p = 0.034$, $p = 0.039$). However, there is no significant difference of AUCs between R-TIC and R-BMode in either the training or validation cohort ($p = 0.461$, $p = 0.592$). (Please find detailed results and analysis about key features selection for R-TIC and R-BMode in supplementary results.)

The robustness experiment was performed three times with randomized patient compositions in training and validation cohorts. The predictive accuracy was highly reproducible in all three kinds of radiomics models (Supplementary Figure S3 and Supplementary Table S4). No significant variation was found in any cases (Supplementary Table S5), which indicated the remarkable robustness of both DL and ML radiomics strategies.

Comparison of HAP-score with AI models.

AUCs of HAP-score were 0.623 and 0.617 in training and validation cohorts (Supplementary Figure S6), which were consistent with the previous study [13]. There were significant differences between AUCs of HAP-score and three AI models in both cohorts (Supplementary Table S6).

DL model visualization and prediction accuracy of human readers

Because R-DLCEUS demonstrated the outstanding ability in predicting responses to TACE, we further investigated how it worked. After converting the CNN-based DL feature maps into pseudo-colored maps [41], the CEUS cines were interpreted and re-visualized in the eyes of DL. By reading feature map images, we preliminarily concluded two general criteria for human readers to predict responses to TACE based on the characteristics of DL feature maps. Firstly, if there were feeding arteries marked by red color, the corresponding case belonged to objective-response with high probability (Figure 4a). Secondly, if the red/warm color regions occurred in arterial phase and last for a long time, the probability of objective-response was high (Fig. 4b). On the contrary, if the red/warm color regions occurred late until the portal phase, the probability of non-response was high (Fig. 4c). Finally, if there was contradiction between the first and second criteria, we made a prediction based on the first criteria.

Based on the concluded criteria, two doctors conducted prediction-based CEUS cines and corresponding DL feature map images. The prediction accuracy of two doctors was summarized in Supplementary Table S7. There was no significant difference between two doctors' predictions ($p = 0.474$, 0.444 in training and validation cohorts, respectively). The ROC curves showed that the prediction performances of doctors based on DL feature map were better than on R-TIC and R-BMode, but still inferior to R-DLCEUS (Supplementary Figure S7).

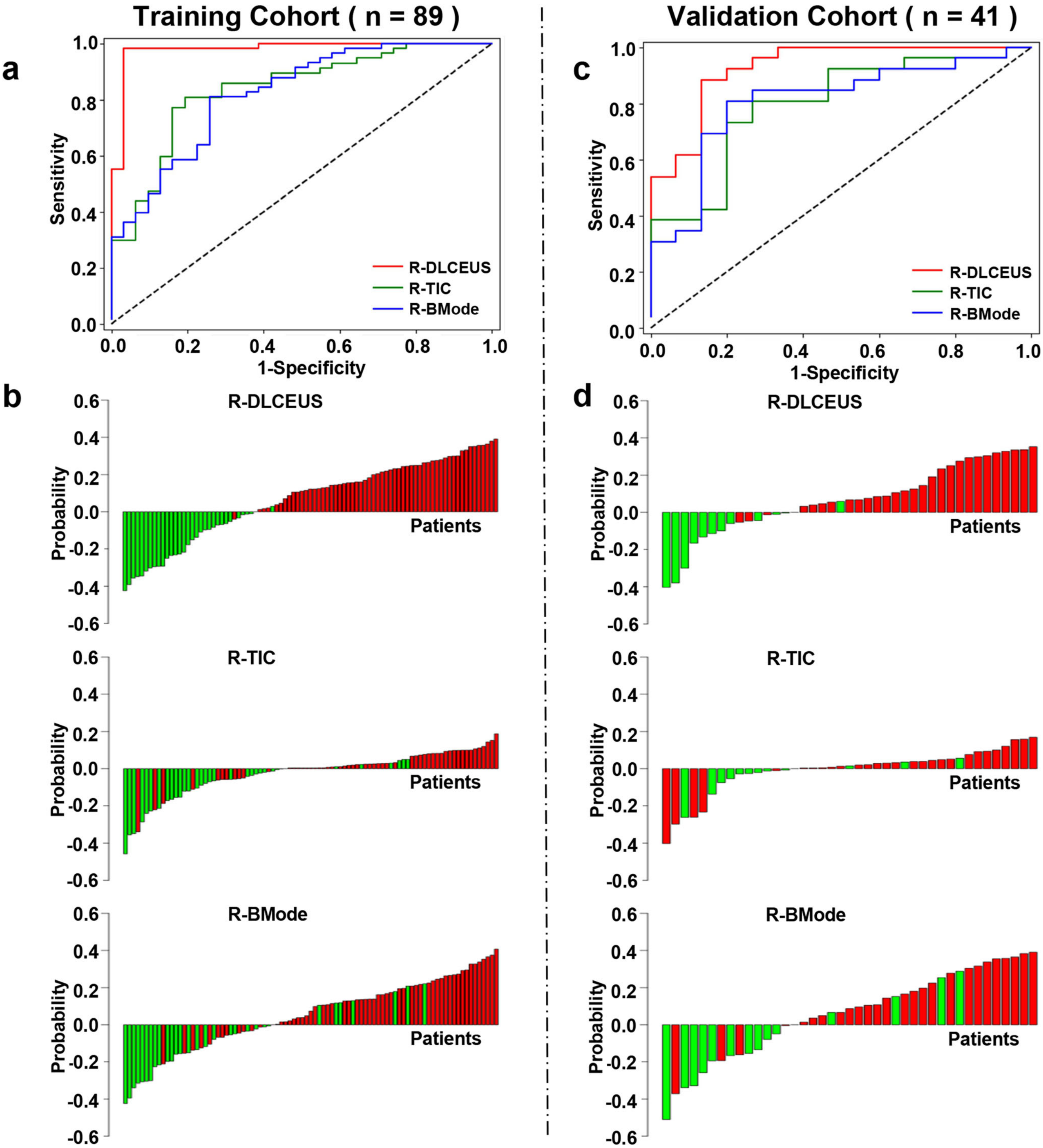

Fig. 3 The predictive performance of R-DLCEUS, R-TIC, and R-BMode in training and validation cohorts. **a** The ROC curves of three radiomics models in the training cohort. **b** The predictive probability for each HCC patient in the training cohort based on three models, respectively. Positive and negative probability indicate non- and objective-response to TACE considered by models. Red and green represent the ground truth of non- and objective-response obtained after 1-month follow-up. **c**, **d** The corresponding ROC curves and individualized predictive probability of three radiomics models in the validation cohort. In (**b**, **d**), the probabilities were the output of models subtracted by cutoff values. ROC curve, receiver operating characteristic curve; HCC, hepatocellular carcinoma; TACE, transarterial chemoembolization

Table 2 Comparison of the predictive performance of R-DLCEUS, R-TIC, and R-BMode in training and validation cohorts

	Patient number	AUC	Sensitivity, %	Specificity, %	Accuracy	PPV, %	NPV, %
R-DLCEUS							
T	89 (68.5%)	0.98 (0.92–0.99)	98.2 (90.8–100.0)	96.7 (83.1–99.9)	0.98 (0.92–0.99)	98.3 (90.9–100.0)	96.8 (83.3–99.9)
V	41 (31.5%)	0.93 (0.80–0.98)	89.3 (70.0–97.6)	92.3 (68.0–98.3)	0.90 (0.77–0.97)	92.0 (74.0–99.0)	81.2 (54.4–96.0)
R-TIC							
T	89 (68.5%)	0.84** (0.74–0.90)	78.6 (66.1–88.6)	81.8 (62.5–92.5)	0.80 (0.69–0.87)	88.2 (76.1–95.6)	67.6 (50.2–82.0)
V	41 (31.5%)	0.80* (0.64–0.90)	82.1 (60.6–93.4)	76.9 (55.0–92.2)	0.81 (0.65–0.91)	84.0 (63.9–95.5)	68.7 (41.3–89.0)
R-BMode							
T	89 (68.5%)	0.82** (0.73–0.91)	79.3 (66.6–88.8)	74.2 (55.4–88.1)	0.78 (0.67–0.86)	85.2 (75.7–91.4)	65.7 (52.6–76.8)
V	41 (31.5%)	0.81* (0.67–0.95)	84.6 (65.1–95.6)	73.3 (44.9–92.2)	0.80 (0.62–0.89)	84.6 (70.0–92.8)	73.3 (51.5–87.7)

95% confidence intervals for quantitative data are shown in brackets, when applicable

AUC, area under the receiver operating characteristic curve; *PPV*, positive predictive value; *NPV*, negative predictive value; *T*, training cohort; *V*, validation cohort

AUCs of three models were statistically compared with each other in training and validation cohorts. Statistical significance was assessed with the DeLong test. Differences were significant when AUCs (both training and validation) of R-DLCEUS were compared to R-TIC or R-BMode (*$p < 0.05$, **$p < 0.01$) Differences between R-TIC and R-BMode were not significant ($p > 0.05$)

Discussion

In this study, the value of CEUS cines and B-mode ultrasound images in predicting HCC responses to TACE was investigated using AI radiomics strategies. Our results revealed that the local response to TACE is a highly predictable preoperation for HCC patients by R-DLCEUS, R-TIC, and R-BMode. R-DLCEUS offered significantly higher AUCs in comparison with R-TIC and R-BMode in both the training ($p = 0.002$, $p = 0.001$) and validation cohorts ($p = 0.034$, $p = 0.039$), which demonstrated CEUS plus DL analysis was capable of offering accurate prediction of HCC responses to TACE. To the best of our knowledge, there were no non-invasive imaging radiomics approaches or any other methods reported before offering comparable performance for personalized local response prediction to TACE. AUCs of previous studies typically ranged from 0.70 to 0.81 [10–12]. Besides, the number of enrolled patients in previous studies was much smaller compared with our study. The detailed comparison between these studies and ours was summarized in Supplementary Table S8. Moreover, our results showed that the AI analysis tool was much more efficient than the existing HAP-score in predicting prognosis of TACE. HAP-score was established based on only three indexes regarding liver function and tumor load, which might limit its prediction performance [13]. By utilizing DL technology, sophisticated radiomics model can be established for quantitative analysis of CEUS cines and providing personalized prediction of TACE efficacy with high accuracy. This holds great potential for the improved selection of BCLC stage B candidates to receive TACE and for better optimization of the treatment planning and follow-up monitoring in the HCC management process [9].

In the robustness experiment, all three models showed excellent robustness, with no significant variation of the prediction accuracy in three training-validation cohort sets. It is worth noting that all ultrasound data were acquired across three different scanning systems, indicating the superior universal application ability of our radiomics models over various manufacturers.

Compared with the ML analysis method, the DL method has a great advantage in that numerous discriminating features are automatically learned from CEUS cines directly, instead of using human-defined features limited by human experiences [45]. In addition, it is time- and labor-consuming for clinical doctors in daily work to precisely delineate the border of tumors in numerous frames for R-TIC and R-BMode methods. In contrast, for R-DLCEUS, manually delineating a bounding box in only one CEUS frame to contain the whole tumor is an easy-to-operate and practical job. Besides, R-DLCEUS made full use of all the spatiotemporal features in the whole CEUS cines, thus, offered a more thorough and comprehensive assessment. Therefore, we strongly believe that R-DLCEUS based on DL analysis method and dynamic CEUS cines is the best choice in clinical practice for its effectiveness and convenience in quantitatively predicting responses to TACE for HCCs. For the similarity of spatiotemporal structure and CT/MRI's 3D space structure, our DL model is not limited to CEUS data and can potentially be applied to analyze CT and MRI with a necessary modification in parameters.

The visualization of the CNN analysis [41] adopted in R-DLCEUS provided direct insight into what computers "see" on CEUS cines. After training, the DL model intelligently recognized feeding arteries at early arterial phase of CEUS and considered them to be positively related to objective-response HCCs. These findings are essentially consistent with

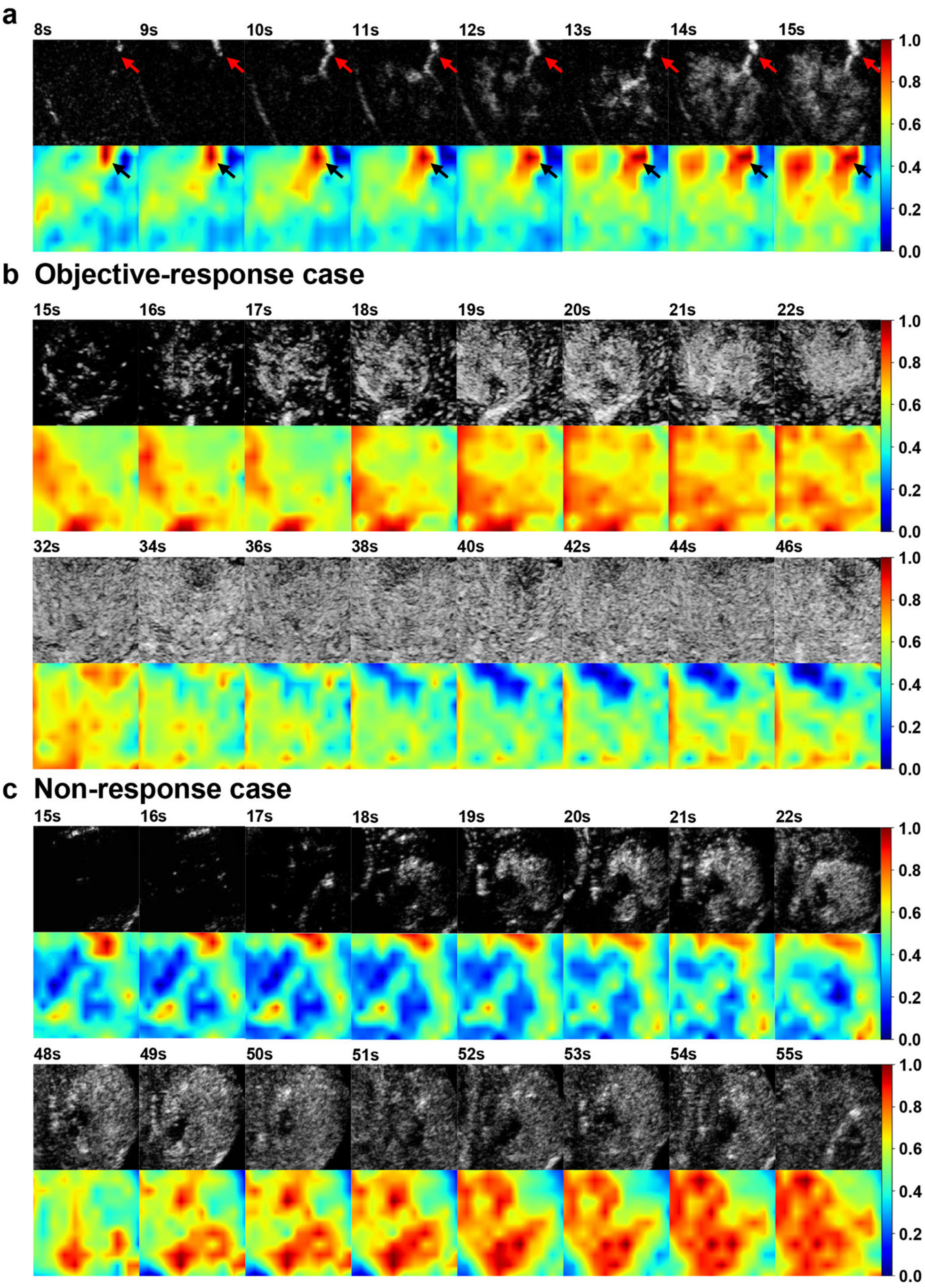

Fig. 4 Visualization of the DL analysis on CEUS cines. Monochrome CEUS cines were converted into pseudo-colored maps frame-by-frame based on the DL model. Red and warm color represent stronger contribution to the predictive classification, whereas blue and cold color represent weaker contribution. **a** Feeding artery (red arrow) was coded with red color (black arrow) for HCC with objective-response to TACE. **b**, **c** Red and blue areas dominated the ROI in different time intervals and patterns for HCCs with objective- and non-responses to TACE. DL, deep learning; CEUS, contrast-enhanced ultrasound; HCC, hepatocellular carcinoma; TACE, transarterial chemoembolization; ROI, region of interest

previous studies [46, 47] indicating that the imaging-based presence of feeding arteries is favorable for identifying the correct tumor feeding vessels and is associated with good response to TACE. Furthermore, the DL model highlighted large portion of positive contribution pixels at the arterial phase in objective-response HCCs, whereas such dominant highlighting appeared at the portal venous phase in non-response HCCs. The prediction results of two doctors based on the above findings showed that it was possible for human readers to predict the responses to TACE accurately based on the DL feature map images. Moreover, such a prediction was reproducible for readers with different clinical experiences. Therefore, this re-visualization strategy converted conventional CEUS cines into distinctive pseudo-color videos perceptible to human eyes for TACE response classification. The entire phenomenon demonstrated by this re-visualization strategy also suggested the dynamic enhancing process during CEUS examination is extraordinarily valuable for achieving the personalized prediction.

There were several major limitations of this study, including the limited population size, the disadvantages of the single-center retrospective data, and the lack of patients from different ethnic populations. Therefore, the results in the current stage are initial, and multi-center prospective studies with a larger HCC patient population are necessary to further verify the performance of proposed radiomics models. Their accuracy across different ethnics is also worth of investigation. Moreover, long-term follow-up studies are desired to truly reveal its clinical impact for benefiting the overall survival of HCC patients. Prediction system incorporating high-level medical knowledge is also a potential direction to improve our models in the future [45].

In conclusion, pre-operative CEUS examinations are of great value in predicting HCC responses to TACE. Artificial intelligence–based radiomics method can effectively utilize CEUS cines to achieve accurate and personalized prediction. The radiomics models exhibited excellent robustness with different training cohorts, even though the data were acquired by various ultrasound systems. All these findings suggest that the proposed DL-based radiomics model combining CEUS examination holds good potential for benefiting TACE candidates in the clinical practice.

Funding information This work was supported by grants from the Ministry of Science and Technology of China under Grant (No. 2017YFA0205200), National Natural Science Foundation of China under Grant (Nos. K0109003, 61671449, 81227901, and 81527805), and Chinese Academy of Sciences under Grant (Nos. GJJSTD20170004, XDBS01030200, and QYZDJ-SSW-JSC005).

Compliance with ethical standards:

Guarantor The scientific guarantor of this publication is Jie Tian.

Conflict of interest The authors of this manuscript declare no relationships with any companies whose products or services may be related to the subject matter of the article.

Statistics and biometry No complex statistical methods were necessary for this paper.

Informed consent Written informed consent was waived by the Institutional Review Board.

Ethical approval Institutional Review Board approval was obtained.

Methodology
- Retrospective
- Diagnostic or prognostic study
- Performed at one institution

References

1. Heimbach JK, Kulik LM, Finn RS et al (2018) AASLD guidelines for the treatment of hepatocellular carcinoma. Hepatology 67:358–380
2. Park J, Chen M, Colombo M et al (2015) Global patterns of hepatocellular carcinoma management from diagnosis to death: the BRIDGE study. Liver Int 35:2155–2166
3. Lencioni R, de Baere T, Soulen MC, Rilling WS, Geschwind JF (2016) Lipiodol transarterial chemoembolization for hepatocellular carcinoma: a systematic review of efficacy and safety data. Hepatology 64:106–116
4. Llovet JM, Real MI, Montana X et al (2002) Arterial embolisation or chemoembolisation vs. symptomatic treatment in patients with unresectable hepatocellular carcinoma: a randomized controlled trial. Lancet 359:1734–1739
5. Peng ZW, Zhang YJ, Chen MS et al (2013) Radiofrequency ablation with or without transcatheter arterial chemoembolization in the treatment of hepatocellular carcinoma: a prospective randomized trial. J Clin Oncol 31:426–432
6. Meyer T, Fox R, Ma YT et al (2017) Sorafenib in combination with transarterial chemoembolisation in patients with unresectable hepatocellular carcinoma (TACE 2): a randomised placebo-controlled, double-blind, phase 3 trial. Lancet Gastroenterol Hepatol 2:565–575
7. Kim BK, Kim SU, Kim KA et al (2015) Complete response at first chemoembolization is still the most robust predictor for favorable outcome in hepatocellular carcinoma. J Hepatol 62:1304–1310
8. Kim BK, Kim KA, Park JY et al (2013) Prospective comparison of prognostic values of modified response evaluation criteria in solid tumours with European Association for the Study of the liver criteria in hepatocellular carcinoma following chemoembolisation. Eur J Cancer 49:826–834
9. European Association for the Study of the Liver (2018) EASL clinical practice guidelines: management of hepatocellular carcinoma. J Hepatol 69:182–236
10. Loosen SH, Schulze-Hagen M, Leyh C et al (2018) IL-6 and IL-8 serum levels predict tumor response and overall survival after TACE for primary and secondary hepatic malignancies. Int J Mol Sci 19:E1766
11. Abajian A, Murali N, Savic LJ et al (2018) Predicting treatment response to intra-arterial therapies for hepatocellular carcinoma with the use of supervised machine learning-an artificial intelligence concept. J Vasc Interv Radiol 29:850–857
12. Lahrsow M, Albrecht MH, Bickford MW, Vogl TJ (2017) Predicting treatment response of colorectal cancer liver metastases to conventional lipiodol-based transarterial chemoembolization using diffusion-weighted MR imaging: value of pretreatment apparent diffusion coefficients (ADC) and ADC changes under therapy. Cardiovasc Intervent Radiol 40:852–859
13. Kadalayil L, Benini R, Pallan L et al (2013) A simple prognostic scoring system for patients receiving transarterial embolisation for hepatocellular cancer. Ann Oncol 24:2565–2570
14. Hucke F, Pinter M, Graziadei I et al (2014) How to STATE suitability and START transarterial chemoembolization in patients with intermediate stage hepatocellular carcinoma. J Hepatol 61:1287–1296
15. Sieghart W, Hucke F, Pinter M et al (2013) The ART of decision making: retreatment with transarterial chemoembolization in patients with hepatocellular carcinoma. Hepatology 57:2261–2273
16. Alzaraa A, Gravante G, Chung WY et al (2013) Contrast-enhanced ultrasound in the preoperative, intraoperative and postoperative assessment of liver lesions. Hepatol Res 43:809–819
17. Moschouris H, Malagari K, Marinis A et al (2012) Hepatocellular carcinoma treated with transarterial chemoembolization: evaluation with parametric contrast-enhanced ultrasonography. World J Radiol 4:379–386
18. Song J, Shi J, Dong D et al (2018) A new approach to predict progression-free survival in stage IV EGFR-mutant NSCLC patients with EGFR-TKI therapy. Clin Cancer Res 24:3583–3592

19. Wang K, Lu X, Zhou H et al (2019) Deep learning Radiomics of shear wave elastography significantly improved diagnostic performance for assessing liver fibrosis in chronic hepatitis B: a prospective multicentre study. Gut 68:729–741
20. Liu Z, Zhang XY, Shi YJ et al (2017) Radiomics analysis for evaluation of pathological complete response to neoadjuvant chemoradiotherapy in locally advanced rectal cancer. Clin Cancer Res 23: 1038–2017
21. Kermany DS, Goldbaum M, Cai W et al (2018) Identifyingmedical diagnoses and treatable diseases by image-based deep learning. Cell 172:1122–1131
22. Huang YQ, Liang CH, He L et al (2016) Development and validation of a radiomics nomogram for preoperative prediction of lymph node metastasis in colorectal cancer. J Clin Oncol 34:2157–2164
23. Liu M, Lin M, Lu M et al (2015) Comparison of contrast-enhanced ultrasound and contrast-enhanced computed tomography in evaluating the treatment response to transcatheter arterial chemoembolization of hepatocellular carcinoma using modified RECIST. Eur Radiol 25:2502–2511
24. Georgiades C, Geschwind JF, Harrison N et al (2012) Lack of response after initial chemoembolization for hepatocellular carcinoma: does it predict failure of subsequent treatment? Radiology 265: 115–123
25. Lencioni R, Llovet JM (2010) Modified RECIST (mRECIST) assessment for hepatocellular carcinoma. Semin Liver Dis 30:52–60
26. Edeline J, Boucher E, Rolland Y et al (2012) Comparison of tumor response by response evaluation criteria in solid Tumors (RECIST) and modified RECIST in patients treated with sorafenib for hepatocellular carcinoma. Cancer 118:147–156
27. Yushkevich PA, Piven J, Hazlett HC et al (2006) User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. Neuroimage 31:1116–1128
28. Kono Y, Lyshchik A, Cosgrove D et al (2017) Contrast enhanced ultrasound (CEUS) liver imaging reporting and data system (LIRADS®): the official version by the American College of Radiology (ACR). Ultraschall Med 38:85–86
29. Ta CN, Eghtedari M, Mattrey RF et al (2014) 2-tier in-plane motion correction and out-of-plane motion filtering for contrast-enhanced ultrasound. Investig Radiol 49:707–719
30. Tran D, Bourdev L, Fergus R et al (2015) Learning spatiotemporal features with 3D convolutional networks. arXiv:1412.0767
31. Feichtenhofer C, Pinz A, Zisserman A (2016) Convolutional twostream network fusion for video action recognition. arXiv: 1604.06573
32. Wang L, Xiong Y, Wang Z et al (2016) Temporal segment networks: towards good practices for deep action recognition. arXiv: 1608.00859
33. Srivastava N, Mansimov E, Salakhutdinov R (2016) Unsupervised learning of video representations using LSTMs. arXiv:1502.04681
34. Goodfellow I, Bengio Y, Courville A (2016) Deep Learning, MIT Press
35. Francois C (2016) Xception: deep learning with depthwise separable convolutions. arXiv:1610.02357
36. Zhang B, Zhao Q, Feng W et al (2018) AlphaMEX: a smarter global pooling method for convolutional neural networks. Neurocomputing:36–48
37. Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29:1189–1232
38. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20:273–297
39. Varma S, Simon R (2006) Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics 7:91
40. Ta CN, Kono Y, Eghtedari M et al (2017) Focal liver lesions: computer-aided diagnosis by using contrast-enhanced US cine recordings. Radiology 286:1062–1071
41. Selvaraju RR, Cogswell M, Das A (2017) Grad-CAM: visual explanations from deep networks via gradient-based localization. arXiv:1610.02391
42. MJG L, Vedder MM, JCM W et al (2014) Net reclassification improvement: computation, interpretation, and controversies: a literature review and clinician's guide. Ann Intern Med 160:122–131
43. van Griethuysen JJM, Fedorov A, Parmar C et al (2017) Computational radiomics system to decode the radiographic phenotype. Cancer Res 77:104–107
44. Bruix J, Sala M, Llovet JM (2004) Chemoembolization for hepatocellular carcinoma. Gastroenterology 127:S179–S188
45. Zhou LQ, Wang JY, Yu SY et al (2019) Artificial intelligence in medical imaging of the liver. World J Gastroenterol 25:672–682
46. Iwamoto S, Sanefuji H, Okuda K (2003) Angiographic subsegmentectomy for the treatment of patients with small hepatocellular carcinoma. Cancer 97:1051–1056
47. Matsui O, Kadoya M, Yoshikawa J et al (1993) Small hepatocellular carcinoma: treatment with subsegmental transcatheter arterial embolization. Radiology 188:79–83