ORIGINAL ARTICLE

AASLD
AMERICAN ASSOCIATION FOR
THE STUDY OF LIVER DISEASES

A flexible three-dimensional heterophase computed tomography hepatocellular carcinoma detection algorithm for generalizable and practical screening

Chi-Tung Cheng[1] | Jinzheng Cai[2] | Wei Teng[3] | Youjing Zheng[4] |
Yu-Ting Huang[5] | Yu-Chao Wang[6] | Chien-Wei Peng[3] | Youbao Tang[2] |
Wei-Chen Lee[6] | Ta-Sen Yeh[6] | Jing Xiao[2] | Le Lu[2] | Chien-Hung Liao[1,7] |
Adam P. Harrison[2]

[1]Department of Trauma and Emergency
Surgery, Chang Gung Memorial Hospital
at Linkou, Chang Gung University, Linkou,
Taiwan, Republic of China

[2]PAII Inc., Bethesda, Maryland, USA

[3]Department of Gastroenterology and
Hepatology, Chang Gung Memorial
Hospital, Linkou Medical Center, Linkou,
Taiwan, Republic of China

[4]Virginia Polytechnic Institute and State
University, Blacksburg, Virginia, USA

[5]Department of Diagnostic Radiology,
Chang Gung Memorial Hospital at
Keelung, Chang Gung University,
Keelung, Taiwan, Republic of China

[6]Department of General Surgery, Chang
Gung Memorial Hospital, Linkou, Taiwan,
Republic of China

[7]Center for Artificial Intelligence in
Medicine, Chang Gung Memorial Hospital,
Linkou, Taiwan, Taiwan, Republic of China

Correspondence
Chien-Hung Liao, Department of Trauma
and Emergency Surgery, Chang Gung
Memorial Hospital at Linkou, 5 Fu-Shin
Street, Kueishan, Taoyuan 333, Taiwan,
Republic of China.
Email: surgymet@gmail.com

Adam P. Harrison, PAII Inc., 6720b
Rockledge Dr, Bethesda, MD 20817, USA.
Email: adam.p.harrison@gmail.com

Abstract

Hepatocellular carcinoma (HCC) can be potentially discovered from abdominal computed tomography (CT) studies under varied clinical scenarios (e.g., fully dynamic contrast-enhanced [DCE] studies, noncontrast [NC] plus venous phase [VP] abdominal studies, or NC-only studies). Each scenario presents its own clinical challenges that could benefit from computer-aided detection (CADe) tools. We investigate whether a single CADe model can be made flexible enough to handle different contrast protocols and whether this flexibility imparts performance gains. We developed a flexible three-dimensional deep algorithm, called heterophase volumetric detection (HPVD), that can accept any combination of contrast-phase inputs with adjustable sensitivity depending on the clinical purpose. We trained HPVD on 771 DCE CT scans to detect HCCs and evaluated it on 164 positives and 206 controls. We compared performance against six clinical readers, including two radiologists, two hepatopancreaticobiliary surgeons, and two hepatologists. The area under the curve of the localization receiver operating characteristic for NC-only, NC plus VP, and full DCE CT yielded 0.71 (95% confidence interval [CI], 0.64–0.77), 0.81 (95% CI, 0.75–0.87), and 0.89 (95% CI, 0.84–0.93), respectively. At a high-sensitivity operating point of 80% on DCE CT, HPVD achieved 97% specificity, which is comparable to measured physician performance. We also demonstrated performance improvements over more typical and less flexible nonheterophase detectors. *Conclusion*: A single deep-learning algorithm can be effectively applied to diverse HCC detection clinical scenarios, indicating that HPVD could serve as a useful clinical aid for at-risk and opportunistic HCC surveillance.

Chi-Tung Cheng and Jinzheng Cai contributed equally to this work.

INTRODUCTION

Hepatocellular carcinoma (HCC) is the most common primary malignant tumor of the liver and is the fourth leading cause of cancer deaths worldwide.[1] The number of new cases of HCC increases annually.[1–3] HCC almost runs a fulminant course and carries a grave prognosis.[3]

Imaging plays a key role in the diagnosis of HCC. Advances in imaging technology over the past 2 decades have contributed to better characterization of hepatic lesions. Dynamic contrast-enhanced (DCE) multiphase computed tomography (CT) of the liver, which includes noncontrast (NC), arterial phase (AP), venous phase (VP), and delay phase (DP) scans, is a preferred imaging modality for surveying patients at risk of HCC.[4] The American Association for the Study of Liver Diseases guidelines recommend DCE CT as the diagnostic evaluation of HCC.[5] A recent Cochrane meta-analysis has reported that detection sensitivity is 78% and specificity is 91% for HCC lesions of any size.[6] For analysis focusing on small lesions, sensitivity has been reported to be as low as 40%.[7] Thus, HCC lesions are at danger of being missed at an early course of the disease when treatment is most promising. However, such analyses only focus on studies using pathological examination of the explanted liver or a biopsy of focal lesions as reference standards, meaning patients who are negative would typically have either very advanced chronic liver disease or a liver lesion of another type. Sensitivities and specificities from the broader clinical population may not be captured. HCC lesions are also increasingly being discovered in patients not known to be high risk for HCC, using imaging studies ordered for indications other than suspicious liver lesions.[8–10] Even though it is less informative than DCE CT, such protocols represent a potentially important opportunity for opportunistic HCC surveillance. Computer-aided solutions may help improve patient surveillance in these challenging settings.

Deep learning (DL)-based computer-aided detection (CADe) algorithms have achieved notable successes within natural imagery.[11] In medical images, researchers have applied DL CADe to find liver lesions, but apart from Kim et al.,[12] current studies either do not compare against clinical readers[13–16] or do not enroll control patients without any target lesions[16,17]; thus, gauging the CADe model screening sensitivity and specificity is difficult. Another gap is that HCC surveillance can be plausibly executed on different CT study types, each typically corresponding to its own distinct scenario. Three types are particularly prominent: DCE CT for patients at high risk for liver lesions; NC + VP CT abdominal studies acquired for general abdominal diagnostic purposes; and NC CT studies acquired for general diagnostic purposes or for patients unable to tolerate contrast agents. The effectiveness of DL CADe

for these different scenarios needs to be assessed, ideally considering the makeup of the likely imaged population for each protocol. For DCE CT and the associated higher risk population, high sensitivity is likely of greater importance. For NC + VP or NC-only and the associated more general population, high specificity is likely paramount. Other contrast-phase combinations are also possible (e.g., for centers that do not acquire DP scans). Because each contrast phase can present distinct information, all acquired scans should be exploited in detection. Correspondingly, a deployed DL CADe model should be flexible enough to handle any contrast-phase combination that it is presented with.

In this study, we develop a flexible DL HCC lesion CADe algorithm, called heterophase volumetric detection (HPVD), to deal with these challenges. Our approach integrates heteromodal learning[18] into a powerful volumetric detection framework.[13] This heteromodal (or heterophase) integration accomplishes two goals: (a) it enables HPVD to operate with any combination of CT phase inputs, providing it with maximal flexibility in deployment and (b) our single HPVD algorithm can perform noninferior to standard input-specific CADe algorithms, each trained only on one contrast-phase combination, which makes the HPVD algorithm a valuable means to survey for HCC in varied and distinct patient populations.

MATERIALS AND METHODS

Patient collection and image collection

Our patient and study selection process, which was approved by the institutional review board (201800187B0) of Chang Gung Memorial Hospital (CGMH), is illustrated in Figure 1. Our target patient population comprised patients with one or more viable HCC lesions. Our control patient population comprised patients who were high risk with no viable HCC lesions. We drew patients from the following two pools: pathology-proven studies and nonpathology-proven negative studies. All negative studies were used as controls, but pathology-proven studies were either target or control.

To construct the pool of pathology-proven studies, we examined the pathology reports indicating liver neoplasm from November 1999 to December 2017 from CGMH Linkou, a major hospital in Taiwan. From these, we identified 1287 complete reports diagnosing the presence of HCC from either hepatic resection or liver transplantation procedures that also had associated preoperative DCE CT studies in the CGMH Picture Archiving and Communication System (PACS) repository within 3 months before the procedure. The pathological reports specify the number of lesions and their liver segment location. Using this information, a hepatopancreatobiliary (HPB) surgeon (C.T.C. with
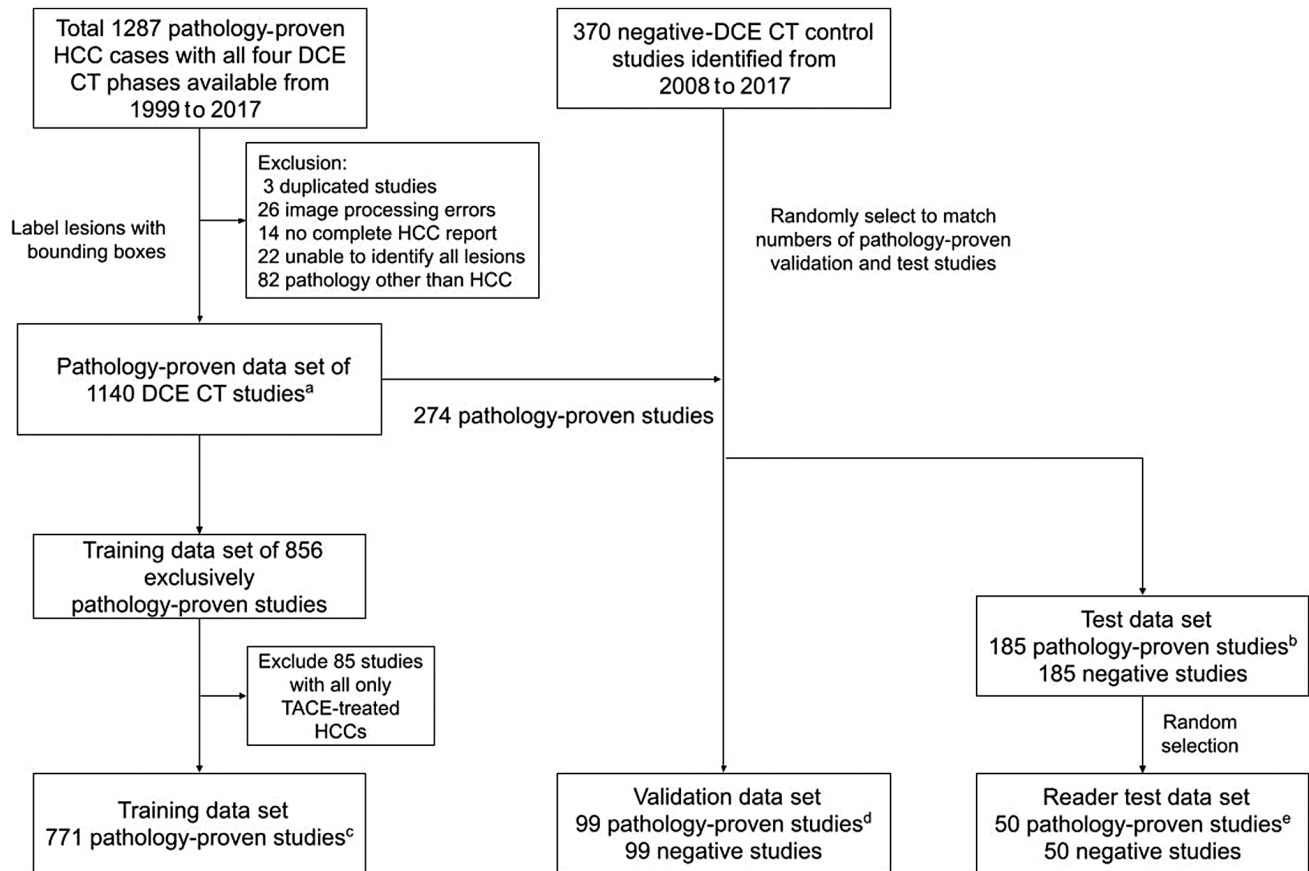
FIGURE 1 Patient selection and data set preparation. [a]Includes 1024 with untreated HCCs and 116 controls with only TACE-treated HCCs. [b]Includes 164 studies containing untreated HCCs and 21 TACE-treated control studies. [c]Includes 761 studies containing untreated HCCs, with 10 TACE-treated studies found after a final review. [d]Includes 89 studies containing untreated HCCs and 10 TACE-treated control studies. [e]Includes 45 studies containing untreated HCCs and five TACE-treated control studies. The numbers for each split are derived from splitting the total number of pathology studies into training (70%), test (20%), and validation (10%). CT, computed tomography; DCE, dynamic contrast enhanced; HCC, hepatocellular carcinoma; TACE, transcatheter arterial chemoembolization.

9 years of experience) located every lesion in the DCE CT studies, annotating each reported lesion in each CT study with a three-dimensional (3D) bounding box (bbox). If the reported lesion was not visible in the DCE CT scans, no bbox was drawn and the corresponding study was excluded. Transcatheter arterial chemoembolization (TACE)-treated HCC lesions were also annotated, but these were recorded as a separate type. After the annotation and review process, a total of 1140 DCE CT studies remained. These were split into 856, 99, and 185 studies for training, validation, and testing, respectively. Among the training and validation sets, 21 and 10 pathology-proven studies, respectively, only contained TACE-treated HCCs. These pathology-proven TACE-only studies were included in the control cohort, resulting in 164 and 89 target studies in the test and validation set, respectively. In the training data set, we excluded studies with only TACE-treated HCC lesions for algorithmic development purposes, resulting in 771 studies. However, after a final review, 10 TACE-only studies were still found among the 771 studies (details in the Supporting Material).

Patients with only TACE-treated HCC lesions from the pathology-proven group were considered part of the control cohort. To augment the control cohort further, we constructed a pool of negative studies, which were defined as nonpathology-proven DCE CT studies where the associated radiologic report indicated no lesions. To do this, we extracted the CGMH radiologic reports of DCE CT studies from 2008 to 2017, which are typically only ordered if there is a suspected liver lesion. To identify studies where no liver lesions were found, we adapted the NegBio[19] radiologic report parser to flag and exclude studies where a found liver lesion was mentioned in the radiologic report. Liver lesion types disqualifying a study included HCC, intrahepatic cholangiocarcinoma, hemangioma, secondary metastasis, focal nodular hyperplasia, adenoma, and abscess findings, where we used the MetaThesaurus to identify synonyms for each finding.[20] Because liver cysts are common and easy to distinguish from HCC, the presence of a liver cyst did not disqualify a patient from being in the control cohort. The text parser identified 2559 DCE CT studies as having no liver lesion

findings, and from these, we randomly selected 470 studies. From these, we manually double-checked the reports, excluding another 10 from the pool. We randomly selected 99 and 185 of these studies to match the numbers in the validation and test set, respectively, of the pathology-proven studies. The demographic, clinicopathological, and scanning parameters of these cohorts are outlined in Table 1 and Tables S1 and S2.

After DCE CT study collection, the phase of each scan was manually identified for validation and test sets and semiautomatically identified for the training set (details in the Supporting Material). Scans within a DCE study were cropped to the same field of view (based on world coordinates). Any nonrigid misalignments between scans were rectified by registering the NC, AP, and DP scans to the VP scan, using SimpleElastix[21] for rigid alignment followed by the dense displacement sampling (DEEDS) algorithm[22] for nonrigid alignment. Finally, for each study, a liver mask was automatically generated using the robust multiphase CHASe algorithm.[23]

Algorithmic development

The training and inference workflow of the HPVD algorithm is depicted in Figure 2. HPVD enhances and extends a state-of-the-art deep-learning 3D CADe algorithm, called volumetric universal lesion detection (VULD).[13] VULD accepts 3D CT scans as input and then outputs bboxes around any found lesion. It has been shown to outperform many popular deep-learning CADe algorithms for lesion detection.[13] However, one trained VULD model can only accept a specific input phase combination. HPVD avoids this restriction by enhancing VULD for heteromodal learning using Heteromodal Image Segmentation principles.[18] Unlike VULD, a single HPVD model accepts any combination of multiphase 3D inputs. During training, HPVD learns to predict bboxes to localize lesions from any input phase combination, which can range from single-phase to the complete four-phase DCE CT. HPVD only localizes suspicious lesion regions without discriminating their type, so a simple postprocessing step filters out any TACE-treated HCC lesions from the target untreated HCC lesions. During inference, the automatically derived liver mask is used to eliminate detected bboxes where less than 30% of its volume overlaps with the liver mask. Detailed explanations of the HPVD model construction, loss function, and training can be found in the Supporting Material.

Lesion-flagging criterion

Evaluating CADe models requires a criterion for whether a bbox represents a good localization. No option is

TABLE 1 Demographic distribution of data sets

Feature	Target (n = 164)	Control (n = 206)
Entire test data set (pathology proven + negatives)		
TACE present, n (%)	7 (4.3)	21 (10.2)
Age, mean (SD)	59.3 (11.1)	53.9 (13.6)
Sex, n (%)		
Male	39 (23.8)	75 (36.4)
Female	125 (76.2)	131 (63.6)
Pathology-proven cohorts (target + TACE-only controls)		
Feature	Training (n = 771)	Test (n = 185)
Total HCC lesion numbers	851	175
Total TACE lesion numbers	73	50
HCC makeup, n (%)		
None (TACE-only)	10 (1.3)	21 (11.4)
Solitary	688 (89.2)	155 (83.8)
Multiple	73 (9.5)	9 (4.9)
Procedure, n (%)		
Resection	692 (89.8)	158 (85.4)
Transplantation	79 (10.2)	27 (14.6)
Hepatitis, n (%)		
Hepatitis B	417 (54.1)	86 (46.5)
Hepatitis B and C	36 (4.7)	9 (4.9)
Hepatitis C	147 (19.1)	46 (24.9)
Not hepatitis B or C	56 (7.3)	17 (9.2)
Unknown	115 (14.9)	27 (14.6)
Cirrhosis, n (%)	393 (51.0)	104 (56.2)
Fatty liver, n (%)		
No	539 (70.2)	133 (71.9)
Mild	167 (21.7)	39 (21.1)
Moderate	61 (7.9)	12 (6.5)
Severe	1 (0.1)	1 (0.5)

Note: We use HCC to denote untreated HCC lesions, whereas TACE denotes TACE-treated HCC lesions. Pathology-proven and negative-DCE CT studies (185 each) were enrolled as part of the test set. Of the pathology-proven studies, 21 were TACE-only and are considered part of the control cohort, with the remainder considered as part of the target cohort. We list characteristics of the entire test data set first, followed by a listing of characteristics only available in the pathology-proven subset of the test set. We also list characteristics of the training set, which is entirely pathology proven.
Abbreviations: CT, computed tomography; DCE, dynamic contrast enhanced; HCC, hepatocellular carcinoma; TACE, transcatheter arterial chemoembolization.

perfect,[24] but it must reflect the end clinical goals and must be chosen *a priori*. Note the chosen criterion is not affected by nor does it affect algorithmic training, it is only used in evaluation to separate true-positive bboxes from false-positive (FP) ones. The main goal of HPVD is to flag lesions and bring them to the attention
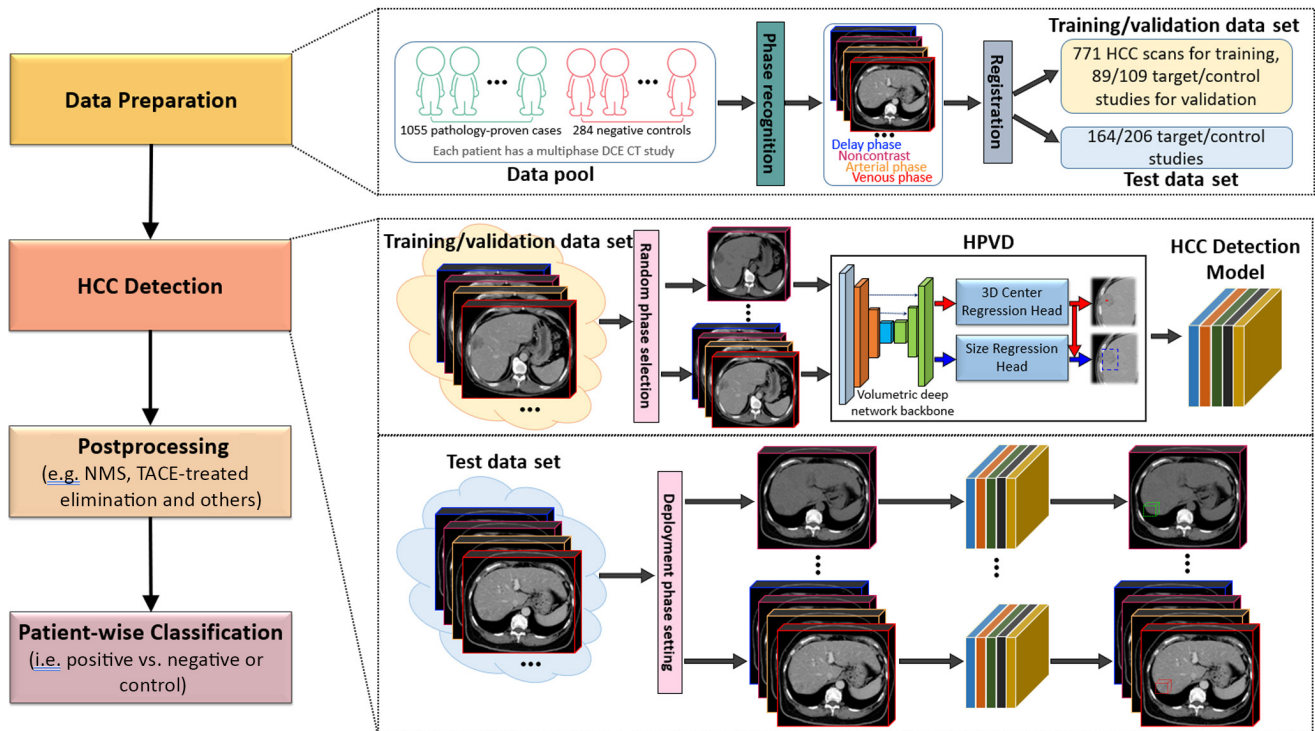
FIGURE 2 Complete workflow of the image preprocessing and detection algorithm. 3D, three dimensional; CT, computed tomography; DCE, dynamic contrast enhanced; HCC, hepatocellular carcinoma; HPVD, heterophase volumetric detection; NMS, non-maximum suppression, TACE, transcatheter arterial chemoembolization.

of the physician to avoid missing observable HCCs. A common evaluation criterion is the intersection over union (IoU) with the ground-truth bboxes.[24] However, a good or bad IoU does not necessarily capture whether a lesion is successfully flagged. We use a different criterion, which we call the lesion-flagging criterion. First, the center of a predicted bbox must lie inside the ground-truth bbox, which is sometimes called the pointing game in machine-learning literature.[25] Second, a predicted bbox must also have a high enough intersection over the detected bbox area (IoBB),[26] which does not penalize predicted bboxes smaller in area than the ground-truth bbox. We used a 3D IoBB threshold ≥0.3. The lesion-flagging threshold and the difference between IoU and IoBB are pictorially demonstrated in Figure 3.

Reader study

We recruited six board-certified physicians as our clinical readers. We randomly selected 50 pathology-proven and 50 negative-CT studies from the larger test set. Among the 50 pathology-proven studies, 45 were target studies containing one or more untreated HCC lesions and five were control studies with only TACE-treated HCC lesions. The demographic data of the reader study subset are shown in Table S3. We

developed a plugin for the Medical Imaging Interaction Toolkit (MITK) viewer[27] that allowed readers to quickly navigate through a set of multiphase CT studies and mark suspicious lesions with a response evaluation criteria in solid tumors (RECIST) marks.[28] A 2D box can be generated from the resulting RECIST marks. Each reader was given the MITK software and the custom plugin, and we asked the readers to decide if any non-treated HCC lesions were present and if so, to mark the maximally suspicious finding (details in the Supporting Material).

Statistical analysis

We performed two primary types of analysis. The first was free-response receiver operating characteristic (FROC) analysis,[29] which is common in CADe evaluations and provides a measure of how well the CADe models can localize all lesion findings in the CT studies. The sensitivity and FPs per study were measured across different confidence thresholds to produce the FROC curve. We compared a single HPVD model against three VULD models, each trained specifically for one of the three input phase combinations. Other main figures of merit were derived from localization ROC (LROC) analysis.[30] The LROC curve can be interpreted similarly as ROC curves, except that the
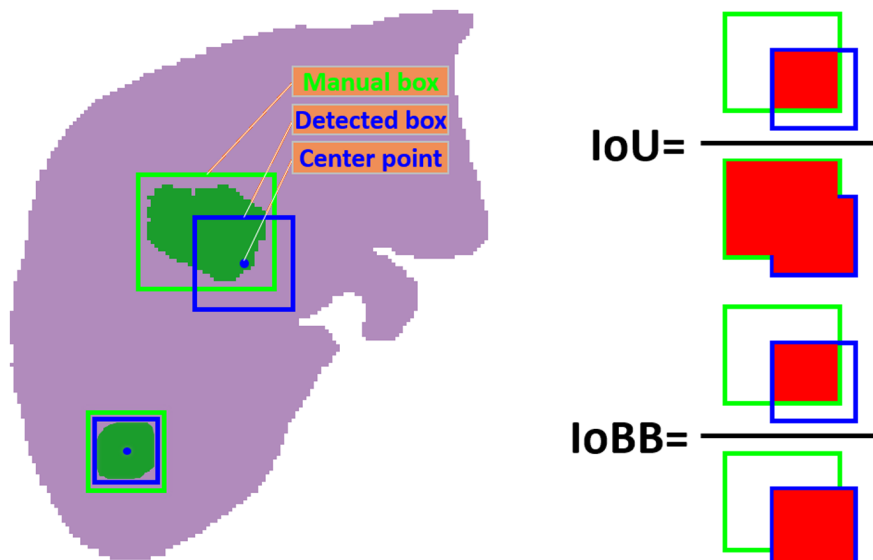
FIGURE 3 Bounding box detection. A bounding box detection is considered as a true positive if (A) its center falls inside the ground-truth lesion bounding box and (B) if the three-dimensional IoBB is >0.3. Here for clarity, we use two-dimensional illustrations. On the right, the overlap area used in the numerator and denominator for the IoBB metric and the more common IoU metric are pictorially illustrated. IOBB, intersection over the bounding box; IoU, intersection over union.
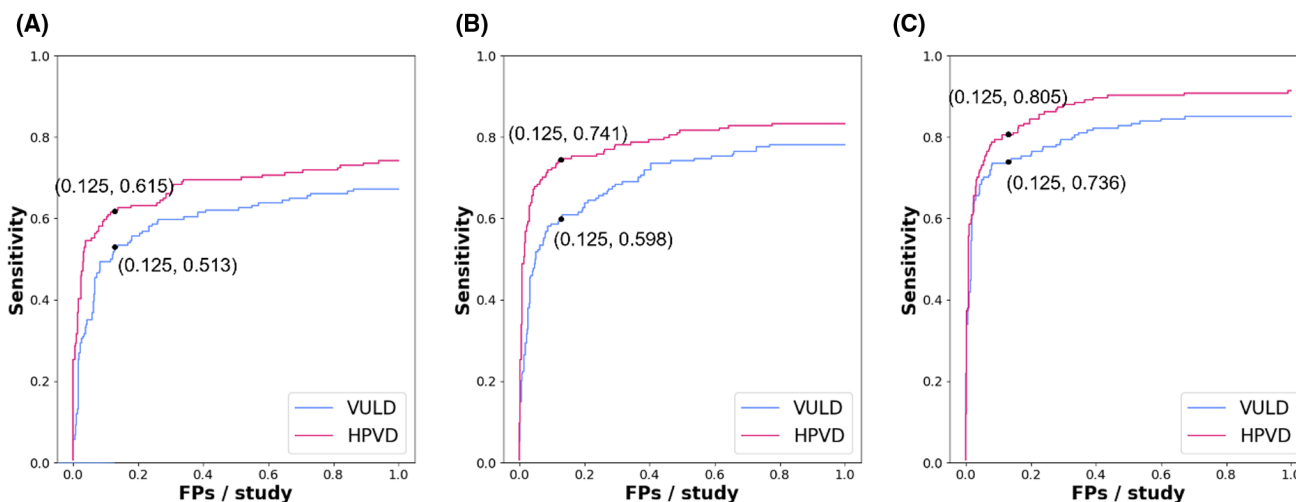

FIGURE 4 FROC curves of detection performance of the VULD and HPVD models. FROC curves depict performance using the complete test set (n = 370), where detection performance is depicted on (A) NC-only, (B) NC + VP, and (C) full DCE CT studies. CT, computed tomography; DCE, dynamic contrast enhanced; FP, false positive; FROC, free-response receiver operating characteristic; HPVD, heterophase volumetric detection; NC, noncontrast; VP, venous phase; VULD, volumetric universal lesion detection.

ordinate of the curve is the true-positive localization rate (TPLR), meaning the CADe model must both (1) detect the presence of HCC lesion(s) when one or more are present and (2) localize one of them successfully. The confidence threshold can be varied, and the corresponding TPLR and specificities can be calculated. We calculated the area under the curve (AUC) of the LROC with 95% confidence intervals and statistical significance tests on the improvements of the HPVD AUC over VULD, calculated using the nonparametric procedure of Wunderlich and Noo.[30] The Bonferroni correction was applied to correct for multiple comparisons

of the latter.[31] Please see our Supporting Material for more explanation on these designs.

RESULTS

High lesion-wise detection performance

FROC curves for the complete test set (n = 370) are rendered in Figure 4A–C. Focusing on the 0.125-FPs/study operating point for the NC-only, NC + VP, and DCE CT settings, HPVD achieved 61.5%, 74.1%, and 80.5%

sensitivity, respectively, compared to the VULD lower sensitivities of 51.3%, 59.8%, and 73.6%, respectively. As expected, performance was best under the full DCE CT setting and sensitivities decreased as fewer phases were available. The increased performance of one HPVD model over three different input-specific VULD models suggests that heterophase learning can impart performance improvements over and above any benefits from increased flexibility. In particular, the HPVD high NC+VP performance suggested it still provided good lesion-wise detection performance despite the lack of arterial and delayed contrast phases.

High study-wise detection performance

Study-wise LROC curves are rendered in Figure 5. When using the complete test set (n = 370), the single HPVD model can yield a higher AUC than each of the three VULD models, no matter the input settings (Figure 5A–C). However, statistically significant improvements were only achieved for the NC+VP setting. Unsurprisingly, performance was best when full DCE CT was available, but HPVD still managed to post an AUC of 0.81 when only given NC+VP CTs. The NC-only input challenged HPVD the most, but the TPLR may be high enough for opportunistic screening. The performance at high-specificity and high-TPLR operating points, which may be good settings for low-risk opportunistic screening and high-risk screening, respectively, is highlighted in Figure 5A–C. When using full DCE CT (e.g., for high-risk screening), HPVD can reach a specificity of 97% at the high-TPLR operating point of 80%. On the other hand, for potential low-risk screening applications, HPVD can achieve a TPLR of 59% and 74% for the NC-only and NC+VP settings, respectively, at a high-specificity operating point of 95%.

In terms of the reader-study subset (n = 100), Figure 5D–F illustrates HPVD and reader performance. As can be seen, all readers performed well when given the four-phase DCE CT, but HPVD managed to outperform or match them. As expected, reader performance dropped substantially with the NC+VP and NC-only studies, with the two radiologists tending to perform better than the other specialties. In the NC+VP and NC-only studies, HPVD still performed well but its performance dropped like the readers. These results suggest that HPVD can match the performance trends of a diverse strata of clinical readers.

Because small lesions can challenge discovery, we also evaluated the effect of lesion size on performance. To do this, we stratified the pathology-proven cohort by studies that had the largest lesion ≤3 cm (n = 92, 49.7%) or >3 cm (n = 93, 50.3%), which is a size threshold that can be used to discriminate between early and intermediate-stage HCC.[32] We randomly selected negative-control cases to match the number of target

cases in each size stratification. Smaller sized lesions indeed challenged the HPVD model (Figure 6A,B), with HPVD reporting AUCs of 0.95 and 0.81 for large and small lesions, respectively, under the DCE CT setting. Results from the reader study subset (Figure 6C,D) demonstrated that such lesions also seriously challenged physician readers. Importantly, the performance of HPVD was better than clinical readers, suggesting that the CADe model could be beneficial in discovering HCC lesions early in their course. The LROC curves of NC+VP and NC-only are shown in Figures S1 and S2, respectively.

Finally, some qualitative examples are highlighted in Figure 7. Cases that were localized by HPVD but missed by some of the readers, exemplifying the interobserver variation among human readers, are highlighted in Figure 7A,B. A small HCC, sized at 1.8 cm, was missed by all readers for all three input phase settings (Figure 7C). However, HPVD correctly localized the lesion if given either an abdominal CT (NC+VP) or a full DCE CT.

DISCUSSION

This work highlights several important results. Foremost, a single HPVD CADe model can accurately localize HCC lesions in various input phase settings and can outperform multiple input-specific competitors. On our data set, HPVD performs comparably or better than two radiologists, two hepatologists, and two HPB surgeons; for small lesions, HPVD performs better than all tested readers. In medical centers, liver imaging interpretations are performed by a specialized radiologist with a multidisciplinary team. However, most patients are diagnosed initially in regional hospitals lacking specialized radiologists where our CADe model could play an important role.

Effective use of CT imaging for HCC detection is critical for increasing patient survival, but it is complicated and difficult. For one, patient distributions are challenging. Unlike other liver lesions, up to 76%–90% of patients with HCC in Western countries usually have cirrhosis,[33,34] causing morphologic changes that can obscure lesions. Additionally, cirrhosis is associated with hepatic nodular lesions that can be easily mistaken as HCC, even by expert clinicians.[35] For patients with cirrhosis, reported diagnostic sensitivity ranges from 44% to 87%.[36,37] Furthermore, surgical or TACE treatment of HCC can cause their own obscuring morphologic changes as tissue regenerates, making it difficult to identify viable or new tumors. Our data collection protocol mirrored many of these challenges. All patients with a pathological HCC diagnosis from a resection or transplantation between 2000 and 2017 and who had an accompanying preoperative DCE CT study were enrolled from the CGMH, which is a major hospital and a
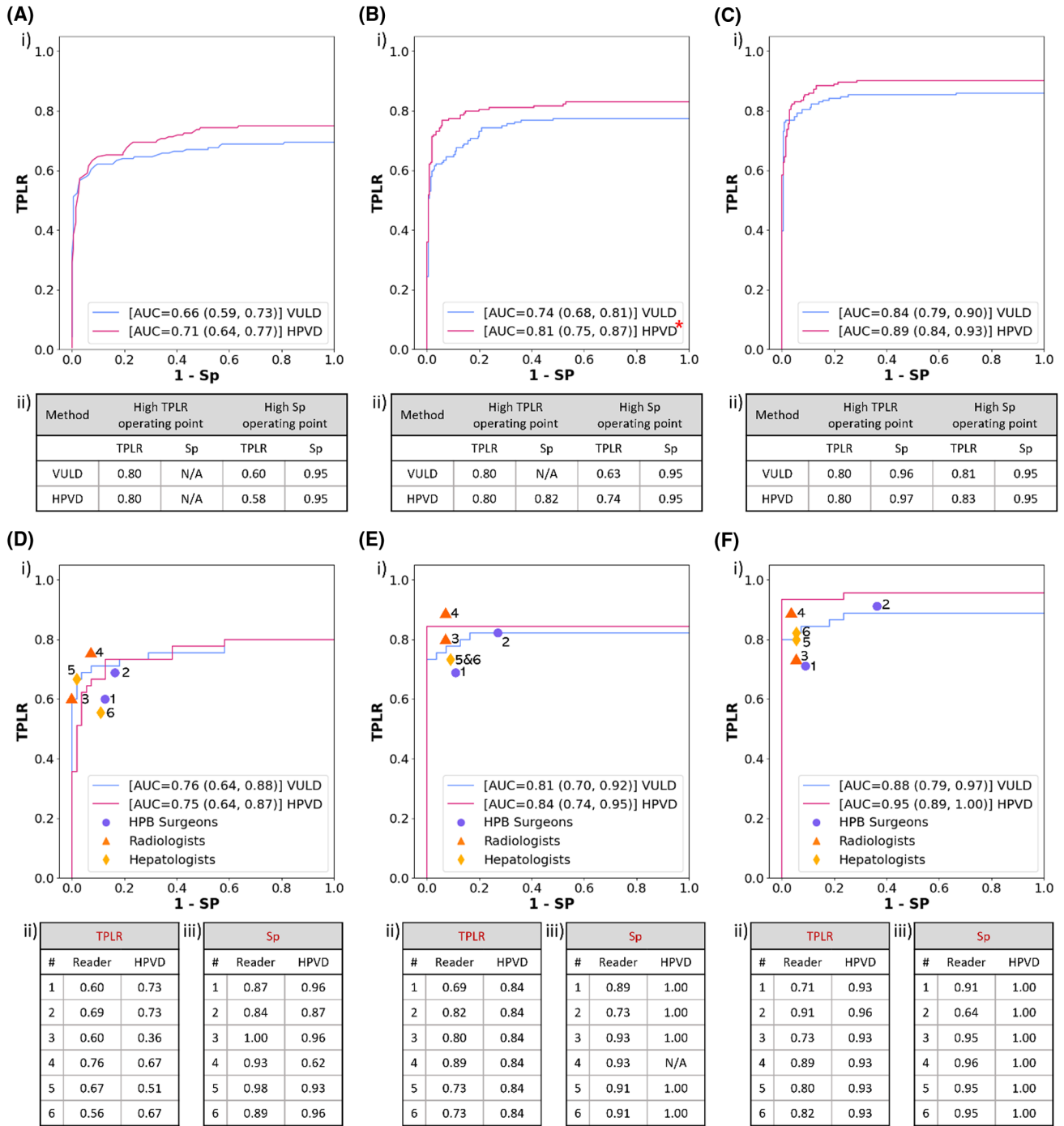
(A)

Method	High TPLR operating point		High Sp operating point	
	TPLR	Sp	TPLR	Sp
VULD	0.80	N/A	0.60	0.95
HPVD	0.80	N/A	0.58	0.95

(B)

Method	High TPLR operating point		High Sp operating point	
	TPLR	Sp	TPLR	Sp
VULD	0.80	N/A	0.63	0.95
HPVD	0.80	0.82	0.74	0.95

(C)

Method	High TPLR operating point		High Sp operating point	
	TPLR	Sp	TPLR	Sp
VULD	0.80	0.96	0.81	0.95
HPVD	0.80	0.97	0.83	0.95

(D)

	TPLR				Sp	
#	Reader	HPVD		#	Reader	HPVD
1	0.60	0.73		1	0.87	0.96
2	0.69	0.73		2	0.84	0.87
3	0.60	0.36		3	1.00	0.96
4	0.76	0.67		4	0.93	0.62
5	0.67	0.51		5	0.98	0.93
6	0.56	0.67		6	0.89	0.96

(E)

	TPLR				Sp	
#	Reader	HPVD		#	Reader	HPVD
1	0.69	0.84		1	0.89	1.00
2	0.82	0.84		2	0.73	1.00
3	0.80	0.84		3	0.93	1.00
4	0.89	0.84		4	0.93	N/A
5	0.73	0.84		5	0.91	1.00
6	0.73	0.84		6	0.91	1.00

(F)

	TPLR				Sp	
#	Reader	HPVD		#	Reader	HPVD
1	0.71	0.93		1	0.91	1.00
2	0.91	0.96		2	0.64	1.00
3	0.73	0.93		3	0.95	1.00
4	0.89	0.93		4	0.96	1.00
5	0.80	0.93		5	0.95	1.00
6	0.82	0.93		6	0.95	1.00

FIGURE 5 Patient-wise HCC detection LROC analysis. Detection performance is depicted for (A) NC-only, (B) NC + VP, and (C) DCE CT studies. The complete test set results (n = 370) are presented, where performance of VULD and HPVD are illustrated. (A–C, i) LROC curves using the top-1 detected bbox in each patient. AUCs are reported with 95% confidence intervals. The * in (B) indicates the difference between the VULD and HPVD AUC is statistically significant. (A–C, ii) Patient-wise LROC analysis at two possible operating points, one that could be appropriate for surveilling patients at high risk (high TPLR of 0.80) and the other for opportunistic screening of patients at low risk (high Sp of 0.95). Some models could not achieve the high TPLR so their Sp is set as N/A. (D-F) Performance is presented for the reader study subset (n = 100), where reader performance is also marked. (D–F, i) Each data point represents a single reader, and each LROC curve represents the performance of the deep-learning model using the top-1-detected bbox in each patient. (D–F, ii) TPLR of each reader and the TPLR of the proposed model at a Sp chosen to match each reader's Sp. (D–F, iii) Sp of each reader and the Sp of the proposed model at a TPLR chosen to match each reader's TPLR. Note, VULD requires a separate model for each input contrast-phase setting. AUC, area under the curve; bbox, bounding box; CT, computed tomography; DCE, dynamic contrast enhanced; HCC, hepatocellular carcinoma; HPB, hepatopancreaticobiliary; HPVD, heterophase volumetric detection; LROC, localization receiver operating characteristic; N/A, not applicable; NC, noncontrast; SP/Sp, specificity; TPLR, true-positive localization rate; VP, venous phase; VULD, volumetric universal lesion detection.
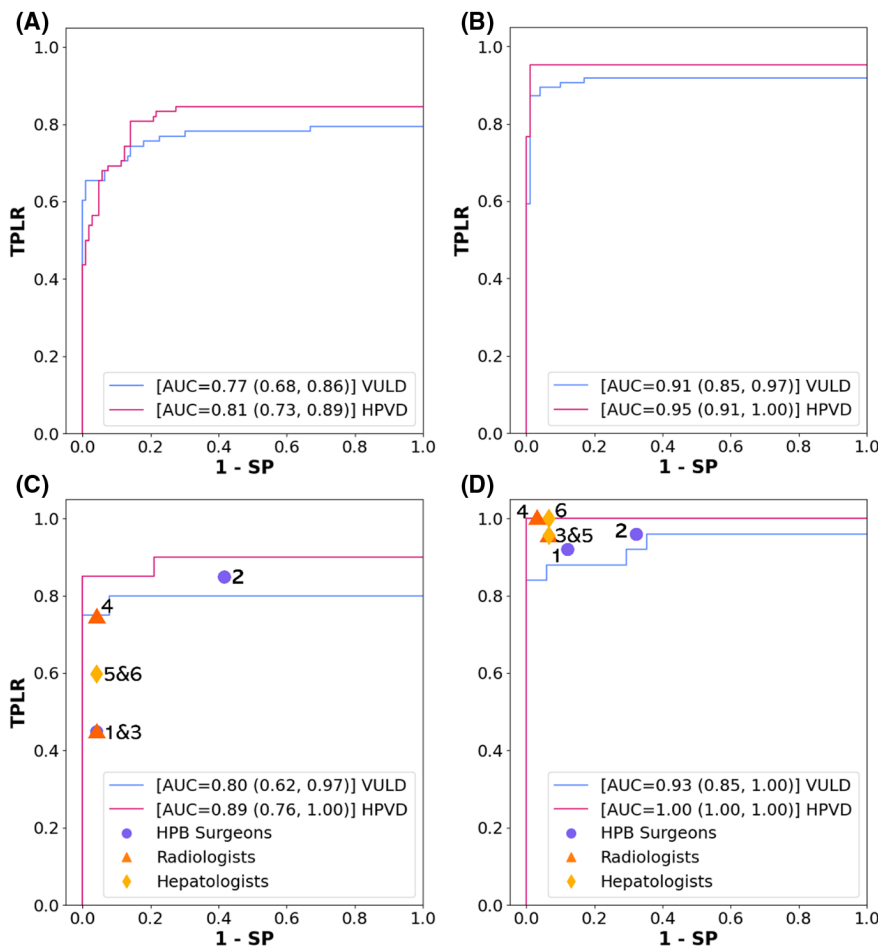
FIGURE 6 LROC curves using full DCE CT studies stratified by the size of the largest lesion. (A,B) Results on the full test set for studies with small (n = 184) and large (n = 186) lesions, respectively. (C,D) Results on the reader study subset for studies with small (n = 44) and large (n = 56) lesions, respectively, with reader performance also indicated. AUC, area under the curve; CT, computed tomography; DCE, dynamic contrast enhanced; HPB, hepatopancreaticobiliary; HPVD, heterophase volumetric detection; LROC, localization receiver operating characteristic; SP, specificity; TPLR, true-positive localization rate; VULD, volumetric universal lesion detection.

leading center for HCC surveillance and treatment.[38] Hence, its patient population should be a good representation of the target population in East Asia, where HCC prevalence is particularly acute.[32] Our negative-control population was also carefully selected as they were drawn from patients imaged with DCE CT, which is only typically ordered for patients at high risk who match many of the profiles of the target population.

Because the liver tissue background and other anatomic structures are frequently hard to distinguish from HCC lesions, multiple contrast phases are needed to confidently detect HCC. Consequently, DCE CT is the standard diagnostic tool,[5] but it adds its own difficulties. Imaging is conducted at several time intervals after contrast injection, and not all images can be acquired at the right timing, which makes contrast enhancement nonstandardized across patients. The amount and rate of administration of contrast and the imaging speed also impact DCE CT quality.[39,40] Finally, patient movement and breathing can cause nonrigid misalignments between phases. Typical identifying DCE CT features

may only be present in 26%–62% of HCC lesions, and a non-negligible number of HCCs can be hypovascular (with poor contrast enhancement), which can lead to a false-negative diagnosis.[36,37]

Despite these challenges, the current HPVD algorithm can achieve a specificity of 97% at a high TPLR operating point of 80%, which compares well to reported clinician performance.[36,37] HPVD can achieve a higher TPLR than all six tested readers when evaluating at their measured specificities, as Figure 5F demonstrates. In terms of small lesions, reader and HPVD performance drop, but the relative difference between HPVD and the readers (Figure 6C) is greater than for large lesions. When applied to DCE CT images typically used for high-risk populations, the HPVD model may help avoid missed lesions and may help avoid additional imaging or invasive biopsy procedures.

The results with DCE CT are highly promising, but DCE CT is not a frontline tool for opportunistic detection as it takes more time to conduct and is not readily obtainable in remote centers. Thus, HCC detection
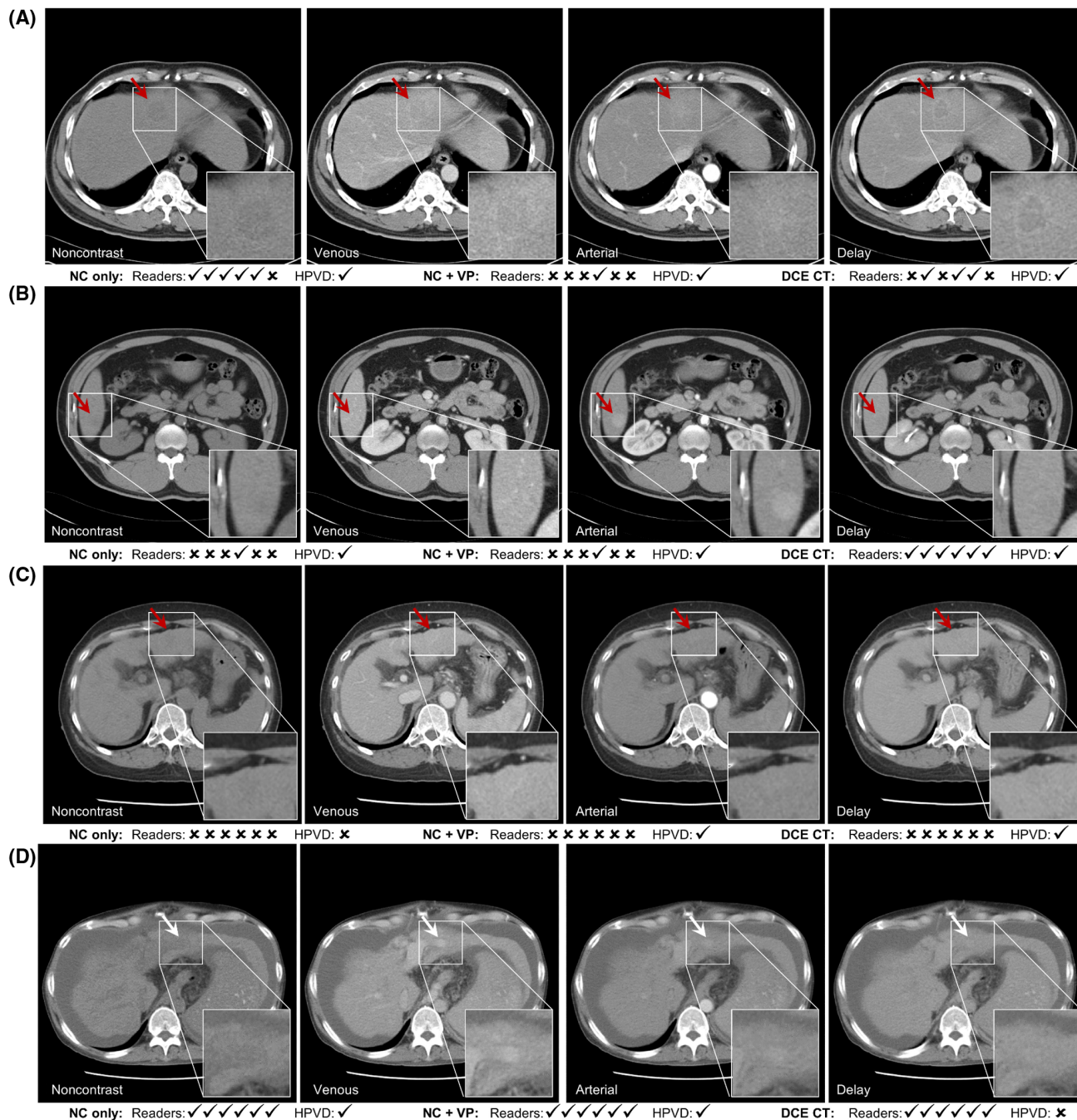
FIGURE 7 Examples of HCC detection. Each row presents DCE CT scans from one study. (A–C) Studies with pathology-proven HCC. (D) Negative study. In (A–C), the most representative axial slice of the lesion from the NC, VP, AP, and DP scans are shown from left to right. Instead of a lesion, (D) shows a false-positive detection produced by HPVD when it is given DCE CT input. Reader and HPVD performance are indicated by the check marks or cross marks below each study, where the two hepatopancreaticobiliary surgeons, two radiologists, and two hepatologists are arranged from left to right. A check mark is used to indicate that the reader or model has correctly classified the corresponding case. Otherwise, cross marks are used to indicate incorrect classifications. Red arrows indicate HCC lesions, and white arrows indicate the false-positive detection in (D). CT images were rendered using a [40, 400] Hounsfield-unit window. AP, arterial phase; CT, computed tomography; DCE, dynamic contrast enhanced; DP, delay phase; HCC, hepatocellular carcinoma; HPVD, heterophase volumetric detection; NC, noncontrast; VP, venous phase.

applied to other CT protocols is also important to develop. For instance, standard general-purpose abdominal CT protocols consist of just the NC+VP scans. Because early stage HCC mostly manifests without symptoms,[3,41] opportunistically discovering incidental

HCC findings from such protocols may be a fruitful way to detect early HCC; such "incidentalomas" are increasingly being detected as the prevalence and use of imaging has increased worldwide.[8–10] Solely using human clinical readers to detect incidental findings is difficult or

infeasible as the examination of the liver may not be as thorough as for CT studies ordered specifically for HCC surveillance. In such scenarios, high-specificity operating points are likely critical to reduce FPs because the HCC prevalence will be lower than with high-risk populations. The HPVD TPLR is 74% at a specificity of 95% when using the NC + VP input (Figure 5B). Apart from one radiologist, HPVD can outperform clinical readers for this setting (Figure 5E). HPVD could be applied to any abdominal CT study to automatically flag suspicious regions requiring more investigation. The NC-only input setting is also clinically important because cirrhotic hepatic disease is associated with renal insufficiency,[42] which can prohibit the use of contrast media to surveil HCC. Additionally, a high proportion of general patients who only undergo medical examinations have NC CT acquired, and incidental identification of liver tumors is not uncommon.[8–10] In patients with hepatic steatosis, the fatty sparing infiltration around the tumor in NC-only CT can be a highly suspicious presentation of HCC; however, other liver tumors also showed similar presentation.[43–45] Although TPLRs suffer compared to an NC + VP protocol (59% compared to 74%), HPVD can still localize HCCs at high specificities of 95%.

Several published works have proposed liver lesion CADe solutions, including some examples that have only been validated on small data sets[16] or did not present comparative clinical reader performance.[13–16] Zhou et al.[17] investigated reader performance against their DCE CT CADe model, but only for classification and not for localization. They also only reported CADe performance at one operating point and did not specify their criteria for judging a detection as a true positive or not, which is a crucial detail for CADe evaluation.[24] Furthermore, they excluded patients with treated lesions, including TACE, and those who had surgery. Given the critical importance of detecting recurrent HCC lesions, this eliminates a major target population. Kim et al.[12] reported a DCE CT CADe solution for secondary liver metastases from colorectal cancer and demonstrated good performance against clinical readers, but again no criteria were given for what constitutes a good localization. Moreover, they developed a 2D CADe model instead of a true 3D one, like ours. Because they do not have the requisite 3D context, 2D CADe models are more susceptible to FPs, which may explain why they reached a sensitivity of 81.3% at the cost of 1.3 FPs/study. By contrast, HPVD reaches 80.5% sensitivity at only 0.125 FPs/study (Figure 4C).

Apart from any distinctions in evaluation and performance, the proposed HPVD offers a notable technological distinction from all prior work—its heterophase capabilities. HPVD is flexible enough to handle any input phase combination setting, and we evaluated performance for NC-only, NC + VP, and DCE CT inputs. Although only three prominent input phase settings were evaluated, HPVD can, in principle, operate with any of the 15 possible phase combinations. Therefore, a single HPVD model can readily adapt to centers that do not typically acquire the DP or to studies where one or more phases are missing. Standard models, like VULD,[13] require a trained model for each input phase setting. Ideally, CADe models can accept whatever phases are available, with stronger performance as more inputs are given. Indeed, the single HPVD model outperformed VULD in each phase setting (Figure 5A–C), although statistical significance was only achieved for the NC + VP setting. It should be emphasized that HPVD is adapted from VULD,[13] and the distinction of note is the use of heterophase learning for the former. We are the first to demonstrate that a single CADe model can exhibit better performance than several input-specific models together while also being much more flexible.

There are several limitations of this study. To focus data collection and analyses on what is currently the most prevalent malignant liver lesion, our study focused specifically on detecting HCC and not all liver lesion types. Further algorithmic development on universal hepatic lesion CADe remains an important future goal. Even if the focus remains on HCC detection, evaluating performance in the presence of other types should also be performed and would better represent clinical populations. Nevertheless, for this preliminary study, we did not include patients with other focal lesion types because each additional type is a potential confounding factor, requiring a commensurate increase in sample size for each factor to statistically measure performance. Thus, we leave it for future work to characterize performance under more varied scenarios. Our target cohort also is susceptible to selection bias as it was collected from patients who had undergone a liver resection or transplantation. Evaluation on a larger, nonpathology-proven, target-patient cohort is key, although obtaining reliable gold-standard labels would become a challenge. Additionally, subgroup analysis on different etiologies of chronic liver disease would help further illuminate any strengths or weaknesses of the HPVD model. Our data are currently single center and retrospective; multicenter and prospective evaluation are other important future aims. In terms of experimental settings, our NC + VP input phase setting is meant to simulate the routine abdominal CT. In reality, a routine abdominal CT can use different contrast-injection protocols than DCE CT, and these may enhance tissues differently. Regarding the reader study, the custom CT viewing (and annotation software) used is not the same clinical Digital Imaging and Communications in Medicine viewer that readers would have been familiar with. In addition, the scans were preprocessed with a registration algorithm that could decrease the resolution of the images. Moreover, some studies had to be re-annotated because of data transfer errors. The above factors might affect the performance of the readers and introduce a bias that differs from real-world

clinical settings. Finally, the readers included four non-radiologists and one general radiologist who may not be representative of a clinical care team at a transplant or other liver center. Nonetheless, the application of the HPVD algorithm is not limited to such centers as its greatest benefit could be for non-liver centers or regional hospitals that lack the requisite expertise.

In conclusion, we developed a flexible and powerful 3D HPVD algorithm that can operate with any combination of DCE CT input phases. The HPVD model demonstrated its ability by its noninferior performance compared to clinical specialists on a balanced reader test across three different input phase settings. Further prospective study is warranted to demonstrate the benefit of using HPVD for HCC surveillance in the clinical setting.

AUTHOR CONTRIBUTIONS

Chi-Tung Cheng, Adam P. Harrison, Jinzheng Cai, Le Lu, and Chien-Hung Liao designed the experiments. Wei Teng, Yu-Ting Huang, Chi-Tung Cheng, and Chien-Hung Liao acquired radiographics for use in the study and provided strategic support. Jinzheng Cai, Youbao Tang, and Adam P. Harrison wrote code to achieve different tasks and carried out all experiments. Jinzheng Cai and Adam P. Harrison implemented the annotation tools for data annotation. Chi-Tung Cheng, Yu-Chao Wang, Chien-Wei Peng, and Chien-Hung Liao provided labels for use in measuring algorithm performance. Chi-Tung Cheng, Jinzheng Cai, Le Lu, Adam P. Harrison, and Chien-Hung Liao drafted the manuscript. Ta-Sen Yeh helped extensively with writing the manuscript. Jing Xiao, Wei-Chen Lee, and Ta-Sen Yeh provided clinical supervision. Jing Xiao, and Youbao Tang, revised the manuscript. Chi-Tung Cheng, Adam P. Harrison, and Chien-Hung Liao supervised the project. All authors have read and agreed to the published version of the manuscript.

ACKNOWLEDGMENT

FUNDING INFORMATION

CONFLICT OF INTEREST

The authors declare that there are no competing interests.

REFERENCES

1. Younossi Z, Anstee QM, Marietti M, Hardy T, Henry L, Eslam M, et al. Global burden of NAFLD and NASH: trends, predictions, risk factors and prevention. Nat Rev Gastroenterol Hepatol. 2018;15:11–20.
2. Mittal S, El-Serag HB. Epidemiology of hepatocellular carcinoma: consider the population. J Clin Gastroenterol. 2013;47(Suppl):S2–6.
3. Njei B, Rotman Y, Ditah I, Lim JK. Emerging trends in hepatocellular carcinoma incidence and mortality. Hepatology. 2015;61:191–9.
4. Oliva MR, Saini S. Liver cancer imaging: role of CT, MRI, US and PET. Cancer Imaging. 2004;4:S42–6.
5. Heimbach JK, Kulik LM, Finn RS, Sirlin CB, Abecassis MM, Roberts LR, et al. AASLD guidelines for the treatment of hepatocellular carcinoma. Hepatology. 2018;67:358–80.
6. Nadarevic T, Giljaca V, Colli A, Fraquelli M, Casazza G, Miletic D, et al. Computed tomography for the diagnosis of hepatocellular carcinoma in adults with chronic liver disease. Cochrane Database Syst Rev. 2021;10:CD013362.
7. Yu NC, Chaudhari V, Raman SS, Lassman C, Tong MJ, Busuttil RW, et al. CT and MRI improve detection of hepatocellular carcinoma, compared with ultrasound alone, in patients with cirrhosis. Clin Gastroenterol Hepatol. 2011;9:161–7.
8. Semaan A, Branchi V, Marowsky A-L, Von Websky M, Kupczyk P, Enkirch SJ, et al. Incidentally detected focal liver lesions - a common clinical management dilemma revisited. Anticancer Res. 2016;36:2923–32.
9. Ehrl D, Rothaug K, Herzog P, Hofer B, Rau H-G. "Incidentaloma" of the liver: management of a diagnostic and therapeutic dilemma. HPB Surg. 2012;2012:891787.
10. Marrero JA, Ahn J, Reddy RK, American College of Gastroenterology. ACG clinical guideline: the diagnosis and management of focal liver lesions. Am J Gastroenterol. 2014;109:1328–47.
11. Zhao Z-Q, Zheng P, Xu S-T, Wu X. Object detection with deep learning: a review. IEEE Trans Neural Netw Learn Syst. 2019;30:3212–32.
12. Kim K, Kim S, Han K, Bae H, Shin J, Lim JS. Diagnostic performance of deep learning-based lesion detection algorithm in CT for detecting hepatic metastasis from colorectal cancer. Korean J Radiol. 2021;22:912–21.
13. Cai J, Yan K, Cheng C-T, Xiao J, Liao C-H, Lu L, et al. Deep volumetric universal lesion detection using light-weight pseudo 3D convolution and surface point regression. In: Martel AL, Abolmaesumi P, Stoyanov D, Mateus D, Zuluaga MA, Zhou SK, et al., editors. Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. Cham, Switzerland: Springer International Publishing; 2020. p. 3–13.
14. Kim DW, Lee G, Kim SY, Ahn G, Lee J-G, Lee SS, et al. Deep learning-based algorithm to detect primary hepatic malignancy in multiphase CT of patients at high risk for HCC. Eur Radiol. 2021;31:7047–57.
15. Romero FP, Diler A, Bisson-Gregoire G, Turcotte S, Lapointe R, Vandenbroucke-Menu F, et al. End-to-end discriminative deep network for liver lesion classification. 2019. Available from: https://arxiv.org/abs/1901.09483
16. Ben-Cohen A, Klang E, Kerpel A, Konen E, Amitai MM, Greenspan H. Fully convolutional network and sparsity-based dictionary learning for liver lesion detection in CT examinations. Neurocomputing. 2018;275:1585–94.
17. Zhou J, Wang W, Lei B, Ge W, Huang Y, Zhang L, et al. Automatic detection and classification of focal liver lesions based on deep convolutional neural networks: a preliminary study. Front Oncol [Internet]. 2020 [cited 2021 Apr 15];10. Available from: https://www.frontiersin.org/articles/10.3389/fonc.2020.581210/full
18. Havaei M, Guizard N, Chapados N, Bengio Y. HeMIS: hetero-modal image segmentation. In: Ourselin S, Joskowicz L, Sabuncu MR, Unal G, Wells W, editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI

2016. Cham, Switzerland: Springer International Publishing; 2016:469–77. Available from: https://arxiv.org/abs/1607.05194

19. Peng Y, Wang X, Lu L, Bagheri M, Summers R, Lu Z. NegBio: a high-performance tool for negation and uncertainty detection in radiology reports. AMIA Joint Summits on Translational Science proceedings. 2018. Available from: https://arxiv.org/abs/1712.05898

20. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004;32:D267–70.

21. Marstal K, Berendsen F, Staring M, Klein S. SimpleElastix: a user-friendly, multi-lingual library for medical image registration. 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2016. p. 574–82. https://doi.org/10.1109/CVPRW.2016.78

22. Heinrich MP, Jenkinson M, Brady SM, Schnabel JA. Globally optimal deformable registration on a minimum spanning tree using dense displacement sampling. Med Image Comput Comput Assist Interv. 2012;15:115–22.

23. Raju A, Cheng C-T, Huo Y, Cai J, Huang J, Xiao J, et al. Co-heterogeneous and adaptive segmentation from multi-source and multi-phase CT imaging data: a study on pathological liver and lesion segmentation. 2021. Available from: https://arxiv.org/abs/2005.13201

24. Petrick N, Sahiner B, Armato SG, Bert A, Correale L, Delsanto S, et al. Evaluation of computer-aided detection and diagnosis systems. Med Phys. 2013;40:087001.

25. Zhang J, Bargal SA, Lin Z, Brandt J, Shen X, Sclaroff S. Top-down neural attention by excitation backprop. Int J Comput Vis. 2018;126:1084–102.

26. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-Ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. 2017. Available from: https://arxiv.org/abs/1705.02315

27. Wolf I, Vetter M, Wegner I, Nolden M, Bottger T, Hastenteufel M, et al. The medical imaging interaction toolkit (MITK): a toolkit facilitating the creation of interactive software by extending VTK and ITK. Proc. SPIE - The International Society for Optical Engineering; 2004. https://doi.org/10.1117/12.535112

28. Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). Eur J Cancer. 2009;45:228–47.

29. Bunch PC, Hamilton JF, Sanderson GK, Simmons AH. A free response approach to the measurement and characterization of radiographic observer performance. Proc. SPIE - The International Society for Optical Engineering; 1977. p. 124–35. https://doi.org/10.1117/12.955926

30. Wunderlich A, Noo F. A nonparametric procedure for comparing the areas under correlated LROC curves. IEEE Trans Med Imaging. 2012;31:2050–61.

31. Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. BMJ. 1995;310:170.

32. Villanueva A. Hepatocellular carcinoma. N Engl J Med. 2019;380:1450–62.

33. Yang JD, Ahmed Mohammed H, Harmsen WS, Enders F, Gores GJ, Roberts LR. Recent trends in the epidemiology of hepatocellular carcinoma in Olmsted County, Minnesota: a US population-based study. J Clin Gastroenterol. 2017;51:742–8.

34. Kanwal F, Singal AG. Surveillance for hepatocellular carcinoma: current best practice and future direction. Gastroenterology. 2019;157:54–64.

35. Sangster GP, Previgliano CH, Nader M, Chwoschtschinsky E, Heldmann MG. MDCT imaging findings of liver cirrhosis: spectrum of hepatic and extrahepatic abdominal complications. HPB Surg. 2013;2013:129396.

36. 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.

37. Roberts LR, Sirlin CB, Zaiem F, Almasri J, Prokop LJ, Heimbach JK, et al. Imaging for the diagnosis of hepatocellular carcinoma: a systematic review and meta-analysis. Hepatology. 2018;67:401–21.

38. Lee C-W, Yu M-C, Wang C-C, Lee W-C, Tsai H-I, Kuan F-C, et al. Liver resection for hepatocellular carcinoma larger than 10 cm: a multi-institution long-term observational study. World J Gastrointest Surg. 2021;13:476–92.

39. Bialecki ES, Di Bisceglie AM. Diagnosis of hepatocellular carcinoma. HPB (Oxford). 2005;7:26–34.

40. Brancatelli G, Baron RL, Peterson MS, Marsh W. Helical CT screening for hepatocellular carcinoma in patients with cirrhosis: frequency and causes of false-positive interpretation. AJR Am J Roentgenol. 2003;180:1007–14.

41. Kansagara D, Papak J, Pasha AS, O'Neil M, Freeman M, Relevo R, et al. Screening for hepatocellular carcinoma in chronic liver disease: a systematic review. Ann Intern Med. 2014;161:261–9.

42. Pipili C, Cholongitas E. Renal dysfunction in patients with cirrhosis: where do we stand? World J Gastrointest Pharmacol Ther. 2014;5:156–68.

43. Kim KW, Kim MJ, Lee SS, Kim HJ, Shin YM, Kim PN, et al. Sparing of fatty infiltration around focal hepatic lesions in patients with hepatic steatosis: sonographic appearance with CT and MRI correlation. AJR Am J Roentgenol. 2008;190:1018–27.

44. Gabata T, Kadoya M, Matsui O, Ueda K, Kawamori Y, Terayama N, et al. Peritumoral spared area in fatty liver: correlation between opposed-phase gradient-echo MR imaging and CT arteriography. Abdom Imaging. 2001;26:384–9.

45. Hamer OW, Aguirre DA, Casola G, Lavine JE, Woenckhaus M, Sirlin CB. Fatty liver: imaging patterns and pitfalls. Radiographics. 2006;26:1637–53.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.