Research article

JHEP|Reports

Application of a deep learning algorithm for the diagnosis of HCC

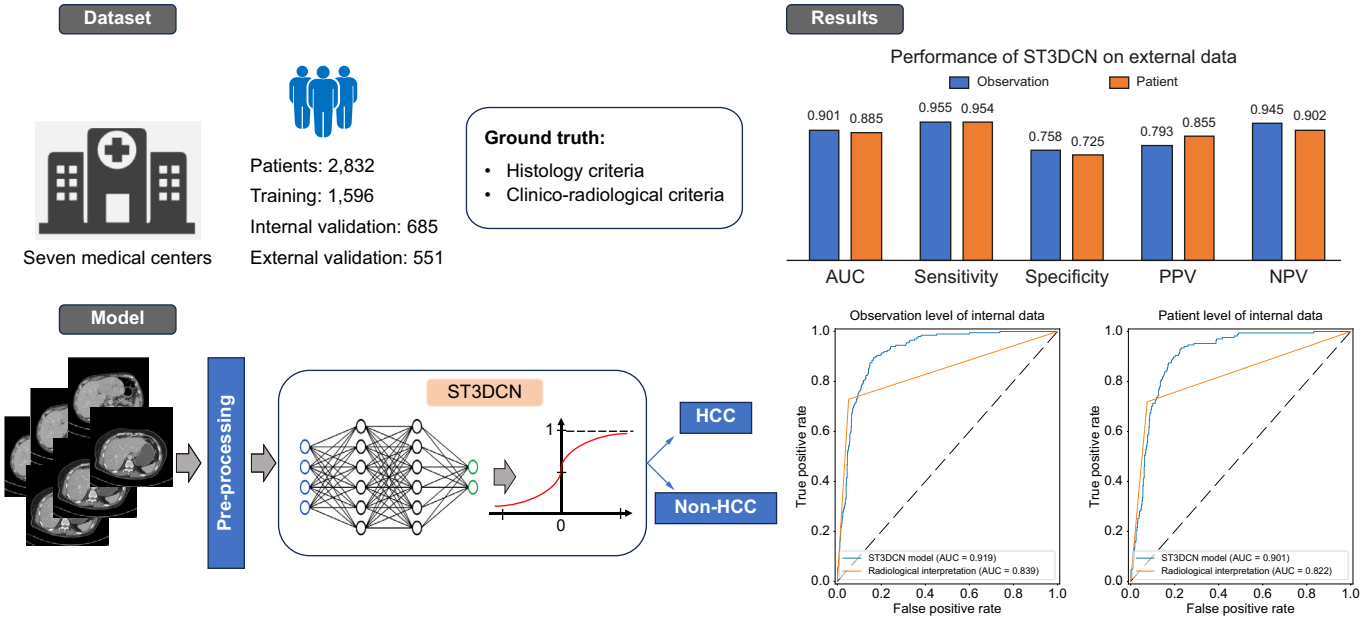
Authors

Philip Leung Ho Yu, Keith Wan-Hang Chiu, Jianliang Lu, …, Chengzhi Peng, Wai Keung Li, Man-Fung Yuen‡Wai-Kay Seto

Correspondence

mfyuen@hkucc.hku.hk (M.-F. Yuen), wkseto@hku.hk (W.-K. Seto).

Graphical abstract



Highlights:

- Deep learning can accurately diagnose HCC on CT.
- Multiple sensitivity analyses of different clinical scenarios demonstrated that deep learning remained robustly accurate.
- Deep learning explainability provided topographical transparency and quality control to the diagnostic process.
- Early diagnosis by deep learning could be used to reduce the currently high mortality rate of HCC.

Impact and implications:

The clinical applicability of deep learning in HCC diagnosis is potentially huge, especially considering the expected increase in the incidence and mortality of HCC worldwide. Early diagnosis through deep learning can lead to earlier definitive management, particularly for at-risk patients. The model can be broadly deployed for patients undergoing a triphasic contrast CT scan of the liver to reduce the currently high mortality rate of HCC.

https://doi.org/10.1016/j.jhepr.2024.101219

Research article

JHEP|Reports

Application of a deep learning algorithm for the diagnosis of HCC

Philip Leung Ho Yu[1,2,†], **Keith Wan-Hang Chiu**[3,4,5,†], **Jianliang Lu**[6,†], Gilbert C.S. Lui[2], Jian Zhou[7], Ho-Ming Cheng[5], Xianhua Mao[6], Juan Wu[8], Xin-Ping Shen[5], King Ming Kwok[9], Wai Kuen Kan[10], Y.C. Ho[11], Hung Tat Chan[5], Peng Xiao[8], Lung-Yi Mak[6,12], Vivien W.M. Tsui[6], Cynthia Hui[6], Pui Mei Lam[6], Zijie Deng[8], Jiaqi Guo[8], Li Ni[8], Jinhua Huang[13], Sarah Yu[3], Chengzhi Peng[14], Wai Keung Li[2], Man-Fung Yuen[6,8,12,‡,*], Wai-Kay Seto[6,8,12,*,‡]

JHEP Reports **2025**. vol. 7 | 1–12

Background & Aims: Hepatocellular carcinoma (HCC) is characterized by a high mortality rate. The Liver Imaging Reporting and Data System (LI-RADS) results in a considerable number of indeterminate observations, rendering an accurate diagnosis difficult.

Methods: We developed four deep learning models for diagnosing HCC on computed tomography (CT) via a training–validation–testing approach. Thin-slice triphasic CT liver images and relevant clinical information were collected and processed for deep learning. HCC was diagnosed and verified via a 12-month clinical composite reference standard. CT observations among at-risk patients were annotated using LI-RADS. Diagnostic performance was assessed by internal validation and independent external testing. We conducted sensitivity analyses of different subgroups, deep learning explainability evaluation, and misclassification analysis.

Results: From 2,832 patients and 4,305 CT observations, the best-performing model was Spatio-Temporal 3D Convolution Network (ST3DCN), achieving area under receiver-operating-characteristic curves (AUCs) of 0.919 (95% CI, 0.903–0.935) and 0.901 (95% CI, 0.879–0.924) at the observation (n = 1,077) and patient (n = 685) levels, respectively during internal validation, compared with 0.839 (95% CI, 0.814–0.864) and 0.822 (95% CI, 0.790–0.853), respectively for standard of care radiological interpretation. The negative predictive values of ST3DCN were 0.966 (95% CI, 0.954–0.979) and 0.951 (95% CI, 0.931–0.971), respectively. The observation-level AUCs among at-risk patients, 2–5-cm observations, and singular portovenous phase analysis of ST3DCN were 0.899 (95% CI, 0.874–0.924), 0.872 (95% CI, 0.838–0.909) and 0.912 (95% CI, 0.895–0.929), respectively. In external testing (551/717 patients/observations), the AUC of ST3DCN was 0.901 (95% CI, 0.877–0.924), which was non-inferior to radiological interpretation (AUC 0.900; 95% CI, 0.877–-923).

Conclusions: ST3DCN achieved strong, robust performance for accurate HCC diagnosis on CT. Thus, deep learning can expedite and improve the process of diagnosing HCC.

Introduction

Liver cancer is the seventh-most common cancer worldwide and ranks second in terms of cancer deaths.[1] Based on current global estimations, there will be a 60% increase in liver cancer-related deaths over the next two decades, reaching 1.33 million deaths in 2040.[2] The main disease burden is found in Eastern Asia, where the age-standardized mortality is 24.5 and 8.0 per 100,000 in men and women, respectively.[2] Chronic HBV infection is the primary risk factor for hepatocellular carcinoma (HCC) in Eastern Asia, with the lifetime risk of HCC among infected patients ranging from 10% to 25%.[3]

Liver cancer has a case-fatality rate of 91.4%, which is much higher than other common cancers.[2] Yet, there is a marked difference in 5-year survival rates for HCC based on disease staging, ranging from 91.5% in solitary tumors <2 cm in size to 11% in advanced disease with adjacent organ involvement,[4,5] highlighting the importance of an early and accurate diagnosis. HCC is typically diagnosed radiologically based on the highly distinctive characteristic of dynamic patterns on contrast-enhanced computed tomography (CT) or magnetic resonance (MR) imaging.[6,7] To standardize radiological lexicon and interpretation, the Liver Imaging Reporting and Data System (LI-RADS) was established to categorize liver observations based on their malignancy risk,[8] with the highest category of LR-5 being highly accurate for HCC.[9,10] Nonetheless, 49% of observations in the at-risk population fall within the indeterminate category of LR-2–LR-4;[8,11] despite having different risk profiles, with the lack of a definitive diagnosis,

* Corresponding authors. Address: Department of Medicine, The University of Hong Kong-Shenzhen Hospital, Shenzhen, China. Tel.: +86 75586913388.
E-mail addresses: mfyuen@hkucc.hku.hk (M.-F. Yuen), wkseto@hku.hk (W.-K. Seto).
† Co-first authors with equal contribution.
‡ Contributed to supervision equally.
https://doi.org/10.1016/j.jhepr.2024.101219

there is little difference in subsequent clinical evaluations, which are mainly surveillance imaging and biopsy in selected cases,[12] potentially resulting in surveillance-related harm.[13] Discrepant LI-RADS categories from serial imaging[10] and the inherent human error present in 3–5% of radiological reporting[14] further complicate clinical interpretation.

Artificial intelligence (AI) is expected to bring about major healthcare benefits worldwide. Medical imaging is best suited for the application of AI, because complex pattern recognition of imaging data via deep learning facilitates a quantitative, rather than qualitative assessment of radiographic characteristics.[15] An AI algorithm applied in HCC diagnosis could improve diagnostic accuracy, and reduce human misinterpretation and the need for further scans and investigations, leading to decreased costs and workload in healthcare systems. With CT being the more easily accessible cross-sectional imaging in the study region,[16] we developed a deep learning algorithm for HCC diagnosis based on a multi-center CT image and clinical data set.

Materials and methods

Patient cohorts and data collection

Model training, internal validation and external testing of our deep learning algorithm was conducted following the Checklist for Artificial Intelligence in Medical Imaging (CLAIM) guideline,[17] which is modelled after the Standards for Reporting of Diagnostic Accuracy Studies (STARD). We retrospectively collected archived thin-sliced (≤1.25-mm slice thickness) multi-phasic contrast CT liver images scanned between March 2013 and August 2020 in Digital Imaging and Communications in Medicine (DICOM) format and relevant clinical information from seven medical centers in our locality, performed in Asian individuals aged ≥18 years. These included 444, 124, 882, 735, 49, and 47 patients from The University of Hong Kong 2014–2018, Queen Mary Hospital 2014–2020, The University of Hong Kong-Shenzhen Hospital 2013–2019, Pamela Youde Nethersole Eastern Hospital 2018–2020, Queen Elizabeth Hospital 2019–2020, and Kwong Wah Hospital 2020, respectively. We further recruited an independent external testing cohort, comprising 551 patients from Sun Yat-Sen University Cancer Center 2013–2018.

To improve the robustness and generalizability of the model to a real-world setting and its adaptability to heterogeneity in imaging standards, we adopted a data-driven approach, aiming to be as inclusive as possible during image and data collection.[18,19] We included all scans in which at least one untreated liver observation could be made.[8] In addition to collecting scans performed in at-risk patients,[12] for model generalizability, we also collected images from individuals without underlying chronic liver disease or increased risk of HCC, performed in individuals for the characterization of incidentally identified liver lesions. We marked patients at risk of HCC development, defined as male patients ≥40 years in age with chronic HBV or female patients ≥50 years in age with chronic HBV, or patients with any disease etiology and underlying liver cirrhosis,[6] suggested by a Fibrosis-4 (FIB-4) score of >1.45.[20]

All CT examinations typically extended from the lung base to the iliac crest, and included four phases: non-contrast, late hepatic arterial, portovenous, and delayed phases. Scans were included from different commercially available CT scanners, using different image acquisition protocols (Table S1), and observations ≥5 mm in size were analyzed.[21,22] Scans performed after locoregional therapy, including thermal ablation, transarterial chemoembolization, or radioembolization and external beam radiation therapy, were excluded because these are assessed differently via LI-RADS under the treatment response algorithm.[8] Incomplete scans, and thick-sliced-only scans (>1.25-mm slice thickness) were also excluded to ensure uniform image quality and resolution. Further image preprocessing is described in Section S1A and Fig. S1.

The present study was approved by the Institutional Review Board of the different participating institutions, in accordance with the Declaration of Helsinki.

Ground truth definitions

We opted against a purely histology-based ground truth definition,[23] given that >60% of HCCs in Asia lack histological diagnosis,[24] while histology would not be available for most non-HCCs. To ensure real-world representativeness and generalizability, a composite clinical reference standard was adopted to establish the ground truth diagnosis.[9] A diagnosis of HCC was based on either histological or clinicoradiological criteria, with histology based on surgical resection, explant or excisional biopsy. Clinicoradiological criteria were reviewed by four clinical investigators with 13, 12, 9, and 8 years of clinical experience, respectively, who in addition to the baseline index scan, reviewed the clinical and radiological progress of all observations over the subsequent 12 months. The estimated doubling time of HCC is 6 months;[10] hence, a stringent 12-month time window ensured the accuracy of our composite clinical reference standard. HCC was diagnosed if the requirement for radiological diagnosis (arterial enhancement with portal-venous washout)[6] or the abovementioned histological criteria were fulfilled within the 12-month window. An observation was considered negative for HCC if it demonstrated a negative histology, or the lack of threshold growth, spontaneous reduction, or disappearance in the absence of treatment.[9] Lesions that underwent presumptive HCC treatment without conclusive diagnosis (*e.g.* LR-4 or LR-M observations) were categorized as high-risk observations but not as definitive HCC.[25,26] Each ground truth diagnosis was reviewed and validated by two clinical investigators, with conflicting reviews handled by consensus review.

Data processing and deep learning model training

Our internal data set was randomly divided into a 7:3 training:validation set ratio. Deep learning was performed using TensorFlow 2.11, facilitated by a computational platform powered by NVIDIA Tesla V100 graphic processing units (Dell Technologies, Singapore). Given that a conventional 2D deep learning model might miss existing spatial information between slices of CT scans,[27] 3D models (see Section S1B for initial development) were trained to predict the binary outcome of HCC *vs*. non-HCC. The range of Hounsfield units (HU) was windowed to [–160, 240] to remove extraneous features and the images were normalized to [0, 1]. Augmentation strategies included horizontal and vertical flipping and 3D rotation. CT images were cropped by observation masks and the cropped volumes were resized to 70 × 70 × 70 for model training. Binary cross entropy was used as loss function and class weight was

set in the training to handle the imbalance in positive and negative samples.

The novel Spatio-Temporal 3D Convolution Network (ST3DCN) (Fig. 1; Section S1C) was developed, which uniquely considers the multiphasic nature of CT imaging, ensuring all detailed visual imaging clues from all CT phases would be used for HCC classification. For the remaining three models that implemented according to published architectures, Convolutional Three-Dimensional (C3D) processes spatiotemporal information and is used in network architectures on video classification;[28] Three-Dimensional Residual Network (3DResNet) is the 3D version of a residual network,[29] using skip connections across layers as the residual learning to facilitate identify mapping and image classification; and Three-Dimensional Squeeze-and-Excitation (3DSE) uses squeeze-and-excitation into a 3D convolutional network for CT phase recognition, utilizing fully connected layers to capture cross-channel interdependencies (Fig. S2, Section S1D–F).

Internal validation and external testing

The trained deep learning models were applied in the internal validation cohort to assess their diagnostic performance. The diagnostic performance of the models was further assessed in the independent external-testing data set, with ground-truth definitions defined similarly to the internal cohort and all HCCs being histologically confirmed.

Radiological interpretation

All liver observations were annotated and interpreted radiologically during internal validation and external testing by one and three board-certified radiologists, respectively, all with >10 years of experience in abdominal cross-sectional imaging, and blinded to the ground truth of each observation, patient clinical information, and deep learning results. Liver observations in at-risk patients were further categorized using LI-RADS version 2018,[8] with observations categorized as LR-1 to LR-5, or LR-M (probably or definitely malignant, not HCC specific).[8] For the remaining patients, radiologists graded each observation on the binary outcome of HCC *vs*. non-HCC. Any discrepancies during external testing were resolved by consensus review.

Model explainability

To understand the model explainability, Gradient-weighted Class Activation Mapping (Grad-CAM) was used to build heatmaps of analyzed images to indicate which portion of CT images contributed significantly to the classification of HCC cases[30] (further described in Section S1G). More specifically, the integrated gradients method was used as a measure of deep learning explainability to approximate Shapley values calculated spatially for displaying the contribution of each voxel to the prediction (Section S1H).[31]

Statistical analysis

Sample size calculation is described in Section S2A and Table S2. Continuous values were expressed as means (±SD) or medians (IQR) as appropriate. Statistical analysis was performed using Python 3.10.2 (Python Software Foundation), with two-sided p values <0.05 considered statistically significant. The Fleiss' Kappa statistic determined interobserver agreement during LI-RADS assessment.

To determine the performance characteristics pertaining to the binary outcome of HCC *vs*. non-HCC, area under receiver-operating-characteristic curves (AUCs) were constructed to assess overall diagnostic accuracy of the different deep learning models and LI-RADS, with Delong's test applied for the comparison of different AUC curves.[32] In addition, 95% CIs were obtained by using the Hanley and McNeil's method.[33] Given the ratio of non-HCC:HCC cases in the training data, 0.8 was selected as the threshold of the cutoff probability for HCC. Comparison of other performance metrics, including sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were calculated similarly. When non-inferiority between AUCs was to be demonstrated, the non-inferiority test for paired Receiver Operating Characteristic (ROC) curves was performed with a default margin at –0.05.[34] To determine the robustness of the classification model, different sensitivity analyses were performed during internal validation: (1) patients at risk for HCC, in which surveillance is recommended[6]; (2) assessing gold-standard HCC and excluding real-world HCC; (3)looking at indeterminate nodules as defined by LI-RADS (LR-2/LR-3/LR-4); (4) observation size, including 2–5 cm and <2 cm; and (5) assess singularly the arterial or portovenous phase. A misclassification analysis was also conducted, analyzing any false positives and false negatives that occurred.

Results

Patient selection is detailed in Fig. 2. Overall, 2,832 patients and 4,305 CT observations were included in the study. For the internal training and validation cohort, 2,630 patients were screened from six centers, with 2,281 patients (86.7%) eventually included in the analysis. Baseline characteristics are detailed in Table 1. The mean age was 58.4 (±14.3) years; 1,371 (60.1%) and 1,215 (53.3%) patients had underlying chronic liver disease or were at risk for HCC, respectively. Altogether 3,588 observations (1.57 observations per scan) were analyzed (contouring details in Table S3), with a median observation size of 21.25[13–41] mm. Following the 12-month composite clinical reference standard, 514 (22.5%) patients and 607 (16.9%) observations were categorized as HCC. The median size of HCC and non-HCC observations was 59.2 (mm 30.8–115.0) and 18.4. mm (12.0–31.5), respectively. In total, 188 (31.0%) and 419 (69.0%) of HCCs were diagnosed via histological and clinicoradiological criteria, respectively.

Diagnostic performance: overall and at risk

The overall diagnostic performance of different deep learning models compared with radiological interpretation during internal validation, at both the observation and patient levels, is depicted in Table 2A with the AUC curves shown in Fig. 3A. All four deep learning models were statistically superior to radiological interpretation. The best-performing model was ST3DCN, achieving AUCs of 0.919 (95% CI, 0.903–0.935) and 0.901 (95% CI, 0.879–0.924) at the observation and patient levels, respectively, compared with radiological interpretation, with AUCs of 0.839 (95% CI, 0.814–0.864) (p <0.001) and 0.822 (95% CI, 0.790–0.853) (p = 0.002), respectively. ST3DCN also achieved a higher NPV of 0.966 (95% CI, 0.954–0.979) and

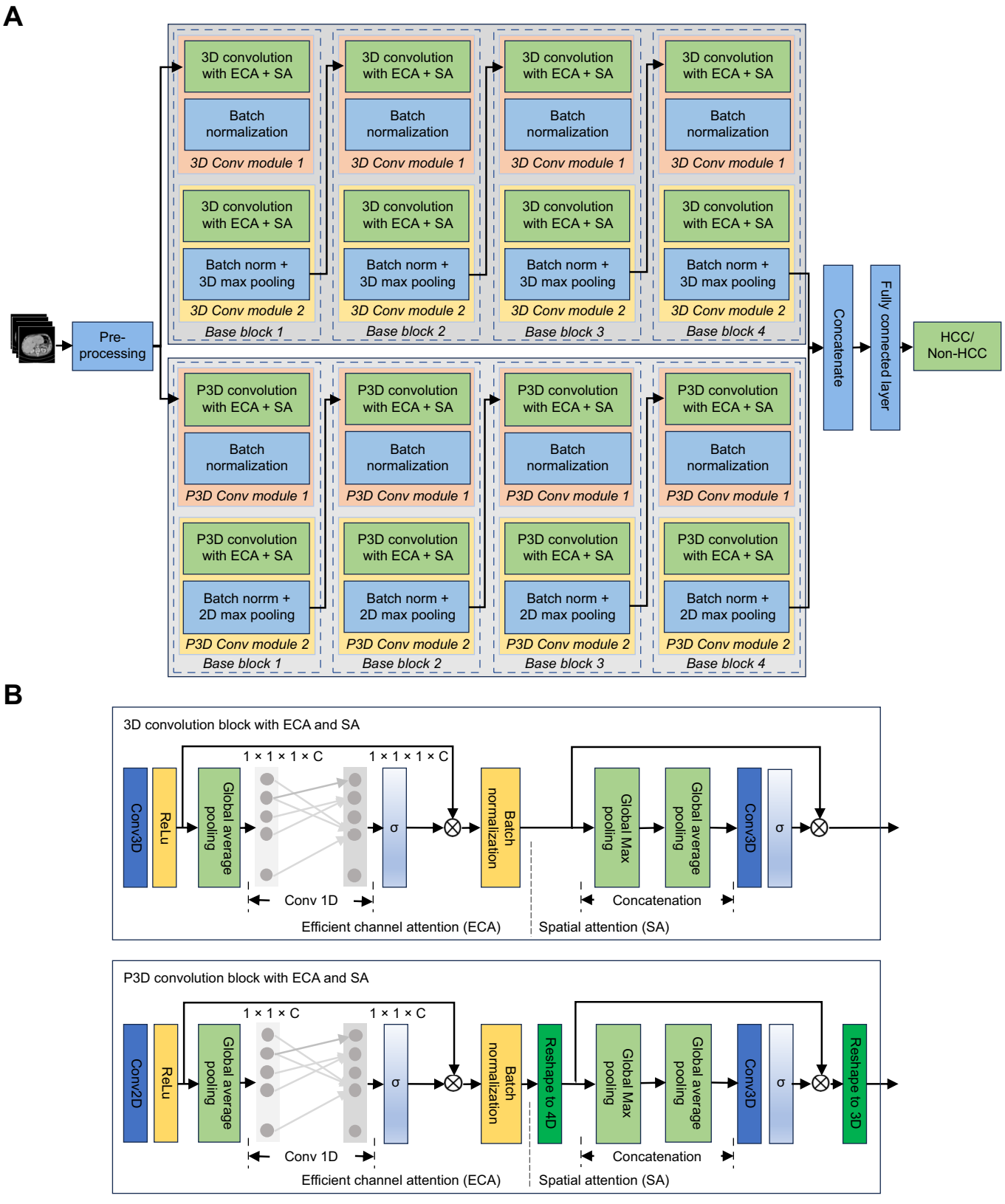


Fig. 1. Overall framework of the developed deep learning model, Spatio-Temporal 3D Convolution Network (ST3DCN), for predicting hepatocellular carcinoma (HCC)/non-HCC. (A) Complete architecture of ST3DCN. (B) Details of the efficient channel attention (ECA) and spatial attention (SA) blocks. P3D, pseudo 3D.

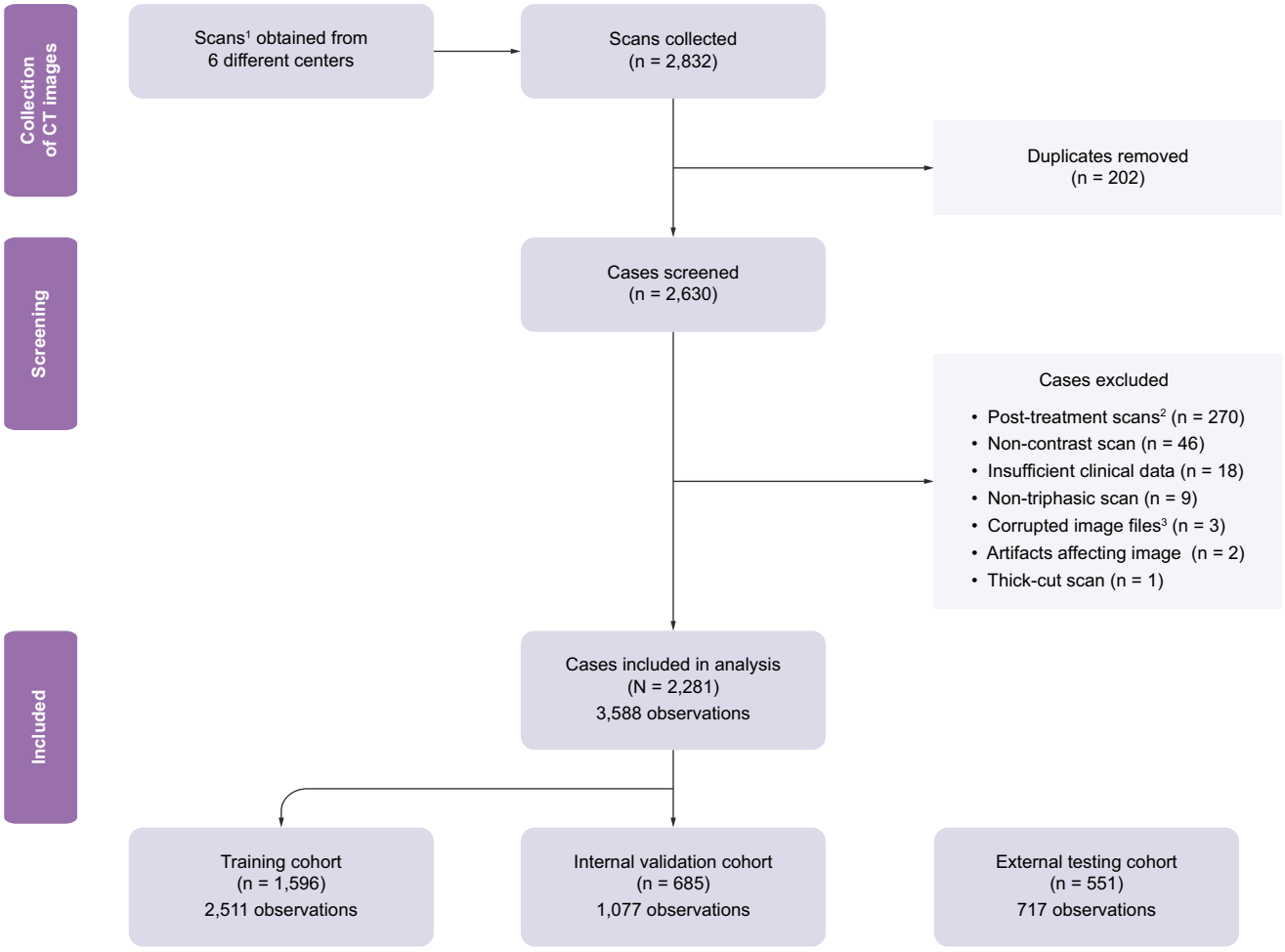


Fig. 2. Patient selection process. [1]Relevant clinical data were also collected. [2]Post-treatment scans after percutaneous, transcatheter, or radiation therapy are assessed differently via LIRADS and were excluded. [3]Two scans had prior intra-abdominal vascular coils inserted, hindering image quality. CT, computed tomography; LI-RADS, Liver Imaging Reporting and Data System.

0.951 (95% CI, 0.931–0.971) at the observation and patient levels, respectively, compared with 0.939 (95% CI, 0.923–0.955) and 0.911 (95% CI, 0.886–0.935), respectively for radiological interpretation. 3DSE, 3DResNet, and C3D achieved AUCs of 0.875–0.915, which were significantly better than radiological interpretation (p <0.001–0.047), but numerically inferior to that of ST3DCN.

Among at-risk individuals, the distribution of LI-RADS categorization for internal training (1,231 observations) and validation cohorts (549 observations) is depicted in Fig. S3. There was no significant difference in the distribution of LR-1 to LR-5 among the two cohorts (p = 0.356). ST3DCN maintained a similarly high diagnostic performance, achieving AUCs of 0.899 (95% CI, 0.874–0.924) and 0.862 (95% CI, 0.826–0.898), respectively at the observation and patient levels, significantly superior to LI-RADS (AUC 0.817, 95% CI, 0.783–0.852, p <0.001; and AUC 0.790, 95% CI, 0.746–0.835, p = 0.036, respectively) (Table 2B and Fig. 3B).

Heatmap visualization generated by ST3DCN was used to represent the predicted probabilities of HCC (high-risk HCC highlighted in red; Fig. 4A) and non-HCC (Fig. 4B). Deep learning interpretability was demonstrated by a visualization of Shapley values (Fig. S4), with each voxel corresponding to a spatial imaging input feature. Voxels with higher Shapley values signified a greater contribution to HCC diagnosis.

Sensitivity analysis: high-risk observations

The diagnostic performance of the four deep learning models in different patient cohorts as part of a sensitivity analysis is depicted in Table S4 and Fig. S5. When high-risk observations (n = 88) were considered together with HCCs (Table S4A and Fig. S5A), AUCs of ST3DCN at observational and patient levels were 0.953 (95% CI, 0.942–0.965) and 0.952 (95% CI, 0.938–0.967), respectively, compared with radiological interpretation at 0.855 (95% CI, 0.833–0.878) and 0.851 (95% CI, 0.823–0.879), respectively. The trends in diagnostic performances of 3DSE, 3DResNet, and C3D were similar, with AUCs ranging from 0.924 to 0.936 at the observation and patient levels, respectively.

Sensitivity analysis: indeterminate nodules

The performance of the deep learning models was further analyzed for indeterminate observations, in which follow-up imaging or investigations are required. There were 276

Table 1. Baseline characteristics of 2,281 patients and 3,588 observations undergoing model training and validation.

Characteristic	All patients (N = 2,281)	HCC* (n = 514)	Non-HCC (n = 1,767)
Age (years)	58.4 (±14.3)	61.3 (±13.2)	57.6 (±14.5)
Male patients (%)	1,391 (61.0)	416 (80.9)	975 (55.2)
Chronic liver disease (%)			
Present	1,371 (60.1)	492 (95.7)	879 (49.7)
HBV	1,016 (74.1)	392 (79.7)	624 (71.0)
HCV	84 (6.1)	30 (6.1)	52 (5.9)
MASLD	116 (8.5)	12 (2.4)	104 (11.8)
Alcoholic liver disease	48 (3.5)	12 (2.4)	40 (4.6)
≥2 liver diseases†	55 (4.0)	27 (5.5)	26 (3.0)
Others‡	52 (3.8)	19 (3.9)	33 (3.8)
No known liver disease	910 (39.9)	22 (4.3)	888 (50.3)
HCC risk§ (%)			
At risk	1,215 (53.3)	477 (92.8)	738 (41.8)
Not at risk	1,010 (44.3)	31 (6.08)	979 (55.4)
Indeterminate¶	56 (2.5)	6 (1.2)	50 (2.8)
Albumin, g/L	41 (37–44)	38 (34–43)	42 (39–45)
Bilirubin, mmol/L	13 (9–19)	15 (10–24)	12 (9–17)
ALT, U/L	27 (18–45)	37 (25–65)	25 (17–38)
AST, U/L	33 (23–60)	53 (33–103)	27 (20–44)
ALP, U/L	80 (64–112)	102 (75–168)	77 (62–99)
Platelet count, × 10^9/L	201 (146–257)	167 (113–237)	210 (162–263)
INR	1.1 (1.0–1.2)	1.1 (1.1–1.2)	1.1 (1.0–1.2)
AFP, ng/ml	4 (2–21)	40 (6–529)	3 (2–6)

*HCC ground truth was determined by a composite clinical reference standard, referencing histological, radiological, and clinical endpoints.
†Including HBV/HCV; HBV/alcoholic liver disease; and HCV/alcoholic liver disease.
‡Including cryptogenic cirrhosis, cardiac cirrhosis, autoimmune hepatitis, primary biliary cholangitis, and recurrent pyogenic cholangitis.
§At-risk individuals for HCC defined as patients with chronic HBV either male ≥40 years or female ≥50 years of age; or patients with any disease etiology with a FIB-4 score of ≥1.45.
¶Platelet count and, hence, FIB-4 not available for 57 patients. AFP, alpha-fetoprotein; ALT, alanine aminotransferase; ALP, alkaline phosphatase; AST, aspartate aminotransferase; FIB-4, Fibrosis-4; HCC, hepatocellular carcinoma; INR, international normalized ratio; MASLD, metabolic-associated steatotic liver disease.

indeterminate observations in the validation cohort, in which 48 (17.4%) had a ground truth diagnosis of HCC. ST3DCN achieved AUCs of 0.816 (95% CI, 0.762–0.871) and 0.848 (95% CI, 0.794–0.902) at the observation and patient levels, respectively (Table S4B and Fig. S5B). NPV was 0.929 at both levels. The AUCs of the other three models were 0.801–0.842 at the observation and patient levels.

Sensitivity analysis: observation size

Among 363 observations of 2–5 cm in the validation cohort, ST3DCN achieved AUCs of 0.872 (95% CI, 0.836–0.909) and 0.877 (95% CI, 0.838–0.916) at the observation and patient levels, respectively (Table S4C and Fig. S5C). NPVs remained high at 0.942 and 0.931 respectively. The AUCs of radiological interpretation were 0.748 (95% CI, 0.694–0.802) and 0.725 (95% CI, 0.663–0.787) at the observation and patient levels, respectively. Among 501 observations of <2 cm, the AUC of ST3DCN was 0.852 (95% CI, 0.783–0.920) and 0.851 (95% CI, 0.781–0.922) at the observation and patient levels, respectively, and NPVs were 0.980 and 0.968 respectively (Table S4D and Fig S5D).

Sensitivity analysis: singular CT phase

When validating the deep learning models for a singular CT phase, ST3DCN at the observation and patient levels achieved AUCs of 0.918 (95% CI, 0.901–0.934) and 0.896 (95% CI, 0.873–0.919), respectively for arterial phase (Table S4E and Fig. S5E) and 0.912 (95% CI, 0.895–0.929) and 0.890 (95% CI, 0.867–0.914), respectively for the portovenous phase (Table S4F and Fig. S5F).

External testing

Baseline characteristics of 551 patients and 717 observations from the independent external cohort are detailed in Table S5. Overall, 353 (49.2%) observations had a ground truth diagnosis of HCC, all confirmed by histology. Among the non-HCC observations, 42 (5.9%) and 58 (8.1%) were cholangiocarcinomas and liver metastasis, respectively. The at-risk cohort comprised 365 patients (66.2%) and 469 observations (65.4%); LI-RADS categorization for external testing in at-risk patients is shown in Fig. S3.

The diagnostic performances of the four deep learning models are detailed in Table 2C and Fig. 3C. ST3DCN achieved AUCs of 0.901 (95% CI, 0.877–0.924) and 0.885 (95% CI, 0.853–0.917) at the observation and patient levels, respectively, which was non-inferior to radiological interpretation (AUC 0.900, 95% CI, 0.877–0.923; AUC 0.888, 95% CI, 0.856–0.920, respectively; non inferiority margin = –0.023 and –0.034,respectively; both p <0.001). Similar to internal validation, high NPVs of 0.945 and 0.902 respectively were achieved. Diagnostic performance trends for 3DSE, 3DResNet, and C3D were comparable, with AUCs of 0.863–0.885 achieved (non-inferiority margin = –0.048 to –0.039, p = 0.010–0.036).

For at-risk patients (Table 2D and Fig. 3D), the AUCs of ST3DCN were 0.892 (95% CI, 0.861–0.922) and 0.881 (95% CI, 0.839–0.923) for the observation and patient levels, respectively, which were non-inferior to LI-RADS (AUC 0.900, 95% CI, 0.870–0.930; AUC 0.885, 95% CI, 0.844–0.926 respectively; non-inferiority margin = –0.038 and –0.043 with p = 0.011 and 0.026, respectively). If patients with cholangiocarcinoma and liver metastasis were excluded (Table 2E and Fig. 3E), the AUCs of ST3DCN increased to 0.981 (95% CI, 0.969–0.993)

Table 2. Diagnostic performance of the four deep learning models.

Level	Model		AUC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)
(A) Internal validation *vs.* radiological interpretation[*] for all patients							
Observation level	Deep learning models	ST3DCN	0.919 (0.903–0.935)	0.869 (0.823–0.916)	0.852 (0.828–0.875)	0.571 (0.515–0.627)	0.966 (0.954–0.979)
		3DSE	0.900 (0.882–0.919)	0.678 (0.613–0.743)	0.911 (0.892–0.93)	0.634 (0.569–0.699)	0.926 (0.908–0.943)
		3DResNet	0.915 (0.899–0.932)	0.673 (0.608–0.739)	0.923 (0.905–0.94)	0.663 (0.598–0.729)	0.926 (0.908–0.943)
		C3D	0.913 (0.896–0.93)	0.714 (0.651–0.776)	0.905 (0.886–0.925)	0.631 (0.568–0.694)	0.933 (0.916–0.95)
	Radiological interpretation		0.839 (0.814–0.864)	0.729 (0.667–0.79)	0.949 (0.934–0.963)	0.763 (0.703–0.824)	0.939 (0.923–0.955)
Patient level	Deep learning models	ST3DCN	0.901 (0.879–0.924)	0.868 (0.817–0.92)	0.828 (0.796–0.861)	0.62 (0.557–0.682)	0.951 (0.931–0.971)
		3DSE	0.875 (0.849–0.9)	0.683 (0.612–0.753)	0.88 (0.852–0.908)	0.648 (0.577–0.718)	0.896 (0.869–0.922)
		3DResNet	0.886 (0.862–0.91)	0.677 (0.606–0.748)	0.898 (0.872–0.924)	0.681 (0.61–0.752)	0.896 (0.87–0.922)
		C3D	0.893 (0.869–0.916)	0.713 (0.644–0.781)	0.882 (0.854–0.91)	0.661 (0.592–0.73)	0.905 (0.879–0.931)
	Radiological interpretation		0.822 (0.790–0.853)	0.719 (0.65–0.787)	0.925 (0.902–0.947)	0.755 (0.688–0.822)	0.911 (0.886–0.935)
(B) Internal validation *vs.* radiological interpretation for at-risk patients[†]							
Observation level	Deep learning models	ST3DCN	0.899 (0.874–0.924)	0.862 (0.812–0.911)	0.837 (0.798–0.875)	0.733 (0.675–0.791)	0.921 (0.891–0.95)
		3DSE	0.877 (0.849–0.905)	0.660 (0.592–0.727)	0.873 (0.838–0.907)	0.729 (0.663–0.796)	0.831 (0.793–0.869)
		3DResNet	0.887 (0.86–0.914)	0.665 (0.597–0.732)	0.889 (0.857–0.922)	0.758 (0.692–0.823)	0.836 (0.799–0.873)
		C3D	0.895 (0.869–0.921)	0.707 (0.642–0.772)	0.881 (0.847–0.914)	0.756 (0.692–0.819)	0.853 (0.817–0.889)
	Radiological interpretation		0.817 (0.783–0.852)	0.729 (0.665–0.792)	0.906 (0.876–0.936)	0.801 (0.741–0.861)	0.865 (0.831–0.9)
Patient level	Deep learning models	ST3DCN	0.862 (0.826–0.898)	0.859 (0.804–0.914)	0.776 (0.721–0.831)	0.732 (0.668–0.796)	0.885 (0.84–0.93)
		3DSE	0.829 (0.788–0.869)	0.66 (0.586–0.735)	0.817 (0.766–0.869)	0.72 (0.647–0.794)	0.772 (0.718–0.826)
		3DResNet	0.843 (0.804–0.882)	0.667 (0.593–0.741)	0.845 (0.797–0.893)	0.754 (0.682–0.826)	0.781 (0.728–0.833)
		C3D	0.858 (0.821–0.895)	0.705 (0.634–0.777)	0.84 (0.792–0.889)	0.759 (0.689–0.828)	0.8 (0.748–0.852)
	Radiological interpretation		0.790 (0.746–0.835)	0.718 (0.647–0.789)	0.863 (0.817–0.909)	0.789 (0.722–0.856)	0.811 (0.761–0.861)
(C) External testing *vs.* radiological interpretation for all patients							
Observation level	Deep learning models	ST3DCN	0.901 (0.877–0.924)	0.955 (0.933–0.976)	0.758 (0.714–0.802)	0.793 (0.754–0.831)	0.945 (0.919–0.971)
		3DSE	0.885 (0.86–0.91)	0.793 (0.751–0.835)	0.841 (0.803–0.878)	0.828 (0.788–0.869)	0.807 (0.768–0.847)
		3DResNet	0.880 (0.855–0.906)	0.807 (0.766–0.849)	0.799 (0.758–0.841)	0.796 (0.754–0.838)	0.811 (0.77–0.851)
		C3D	0.878 (0.852–0.903)	0.861 (0.825–0.897)	0.794 (0.752–0.836)	0.802 (0.762–0.842)	0.855 (0.817–0.893)
	Radiological interpretation	0.900 (0.877–0.923)		0.830 (0.791–0.869)	0.970 (0.952–0.987)	0.964 (0.943–0.985)	0.855 (0.821–0.889)
Patient level	Deep learning models	ST3DCN	0.885 (0.853–0.917)	0.954 (0.932–0.976)	0.725 (0.664–0.787)	0.855 (0.82–0.89)	0.902 (0.857–0.948)
		3DSE	0.884 (0.852–0.916)	0.801 (0.759–0.843)	0.809 (0.755–0.863)	0.877 (0.841–0.913)	0.705 (0.647–0.764)
		3DResNet	0.863 (0.829–0.898)	0.81 (0.769–0.851)	0.755 (0.696–0.814)	0.849 (0.810–0.888)	0.7 (0.639–0.761)
		C3D	0.868 (0.834–0.902)	0.861 (0.825–0.898)	0.760 (0.701–0.818)	0.859 (0.823–0.896)	0.764 (0.705–0.822)
	Radiological interpretation		0.888 (0.856–0.92)	0.830 (0.79–0.869)	0.946 (0.915–0.977)	0.963 (0.942–0.985)	0.766 (0.714–0.818)
(D) External testing *vs.* radiological interpretation for at-risk patients							
Observation level	Deep learning models	ST3DCN	0.892 (0.861–0.922)	0.952 (0.926–0.979)	0.743 (0.685–0.801)	0.810 (0.765–0.855)	0.931 (0.893–0.969)
		3DSE	0.881 (0.849–0.914)	0.801 (0.751–0.85)	0.839 (0.791–0.888)	0.852(0.806–0.897)	0.785 (0.733–0.838)
		3DResNet	0.886 (0.854–0.917)	0.797 (0.747–0.847)	0.798 (0.745–0.851)	0.820 (0.771–0.868)	0.773 (0.719–0.828)
		C3D	0.881 (0.848–0.913)	0.865 (0.822–0.907)	0.789 (0.735–0.843)	0.825 (0.779–0.871)	0.835 (0.784–0.886)
	Radiological interpretation		0.900 (0.87–0.93)	0.837 (0.791–0.882)	0.963 (0.938–0.988)	0.963 (0.938–0.988)	0.837 (0.791–0.882)
Patient level	Deep learning models	ST3DCN	0.881 (0.839–0.923)	0.951 (0.924–0.978)	0.706 (0.624–0.788)	0.870 (0.83–0.91)	0.875 (0.809–0.941)
		3DSE	0.884 (0.842–0.926)	0.809 (0.76–0.858)	0.798 (0.726–0.87)	0.892 (0.852–0.933)	0.669 (0.592–0.746)
		3DResNet	0.879 (0.837–0.921)	0.801 (0.751–0.851)	0.765 (0.688–0.841)	0.876 (0.832–0.919)	0.650 (0.571–0.729)
		C3D	0.875 (0.832–0.918)	0.866 (0.823–0.908)	0.765 (0.688–0.841)	0.884 (0.843–0.924)	0.734 (0.656–0.812)
	Radiological interpretation		0.885 (0.844–0.926)	0.837 (0.791–884)	0.933 (0.888–0.978)	0.963 (0.937–0.988)	0.735 (0.665–0.805)

(*continued on next page*)

Table 2. (*continued*)

Level	Model		AUC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)
(E) External testing vs. radiological interpretation for all patients, discounting cholangiocarcinomas and liver metastases							
Observation level	Deep learning models	ST3DCN	0.981 (0.969–0.993)	0.955 (0.933–0.976)	0.931 (0.9–0.962)	0.949 (0.926–0.972)	0.938 (0.909–0.967)
		3DSE	0.969 (0.954–0.984)	0.796 (0.754–0.838)	0.973 (0.953–0.993)	0.976 (0.958–0.993)	0.778 (0.733–0.824)
		3DResNet	0.96 (0.942–0.977)	0.807 (0.766–0.849)	0.946 (0.919–0.974)	0.953 (0.929–0.977)	0.783 (0.738–0.829)
		C3D	0.966 (0.95–0.982)	0.858 (0.822–0.895)	0.962 (0.938–0.985)	0.968 (0.949–0.988)	0.833 (0.791–0.876)
	Radiological interpretation		0.932 (0.910–0.954)	0.864 (0.828–0.9)	1.0	1.0	0.844 (0.804–0.885)
Patient level	Deep learning models	ST3DCN	0.982 (0.966–0.998)	0.954 (0.932–0.976)	0.942 (0.903–0.981)	0.976 (0.96–0.993)	0.89 (0.84–0.941)
		3DSE	0.967 (0.945–0.988)	0.804 (0.762–0.846)	0.971 (0.943–0.999)	0.986 (0.972–1.0)	0.663 (0.598–0.729)
		3DResNet	0.958 (0.934–0.982)	0.81 (0.769–0.851)	0.949 (0.913–0.986)	0.976 (0.958–0.993)	0.665 (0.599–0.731)
		C3D	0.965 (0.943–0.987)	0.859 (0.822–0.895)	0.971 (0.943–0.999)	0.987 (0.974–1.0)	0.732 (0.668–0.796)
	Radiological interpretation		0.931 (0.901–0.961)	0.862 (0.825–0.898)	1.0	1.0	0.742 (0.679–0.805)

*Radiological interpretation: LR-5 required for diagnosis of HCC.
†At-risk patients defined as patient with cirrhosis of any etiology, or male ≥40 years or female ≥50 years of age with chronic HBV; HCC, hepatocellular carcinoma; ST3DCN, Spatio-Temporal 3D Convolution Network; Radiological Interpretation, Liver Imaging Reporting and Data System; NPV, negative predictive value; PPV, positive predictive value.

and 0.982 (95% CI, 0.966–0.998), respectively, superior to that of radiological interpretation (AUC 0.932, 95% CI, 0.910–0.954, p <0.001; AUC 0.931, 95% CI, 0.901–0.961, p <0.001, respectively). In addition to high NPVs (0.89–0.938), relatively higher PPVs of 0.949–0.976 were achieved. Additional sensitivity analyses are detailed in Table S6 and Fig. S6.

The interobserver agreement of LI-RADS interpretation among the three reporting radiologists was excellent, with a Fleiss' κ value of 0.972 (95% CI, 0.960–0.984) and 0.966 (95% CI, 0.951–0.981) at the patient and observation levels, respectively. Discrepancies that required consensus reading were found in 17 (3.09%) patients and 22 (2.82%) observations.

Misclassification analysis

Details of misclassification analyses are provided in Table S7, with examples of misclassified patients depicted in Fig. S7. False-positive observations were present in 12.1% (n = 130) and 14.5% (n = 104) of internal validation and external testing observations, respectively. When considered in combination, 44.4% (n = 104) of false positives were cholangiocarcinoma or liver metastasis. False-negative observations were less frequent, present in 2.4% (n = 26) and 2.2% (n = 16) in internal validation and external testing observations, respectively, with LR-5 observations comprising 45.2% (n = 19).

Discussion

The current study illustrated the high accuracy of 3D deep learning models for diagnosing HCC on CT, with ST3DCN achieving observation-level AUCs of 0.901–0.919 during internal validation and external testing, compared with radiological interpretation and the other three models, with AUCs of 0.839–0.900 and 0.878–0.915, respectively. Corresponding NPVs in ST3DCN were similarly high at 0.945–0.966, compared with radiological interpretation at 0.855–0.939 and the other models at 0.807–0.933. The clinical applicability of ST3DCN is potentially huge, especially when, over the next two decades, the incidence and mortality of liver cancer is expected to increase by 60% worldwide.[2]

The foundation of our deep learning models was epitomized by several important features. First, the quality and representativeness of collected data are key. A crucial component of our deep learning model is foregoing a purely histology-based ground truth, which can result in considerable ascertainment bias by neglecting substantial cohorts of patients without histology.[23] Instead, we adopted a clinical composite reference standard,[9] via integration of clinical and radiological characteristics, to ensure that our model was representative and reflective of real-world practice. The current study differed from other published research, which specifically focused on differentiating between specific types of liver tumor, such as between HCC and a combination of intrahepatic cholangiocarcinoma, metastasis, and hemangioma.[35–37] Other key components were our data-driven approach and adequate sample size, which were important in anticipation of heterogeneous standards and large variations in medical imaging equipment and scanner settings.[19] A data-driven approach in deep learning avoids overfitting and improves versatility in clinical application.[18]

Second, the performance of our deep learning models was robust, performing well in different clinical scenarios. Most

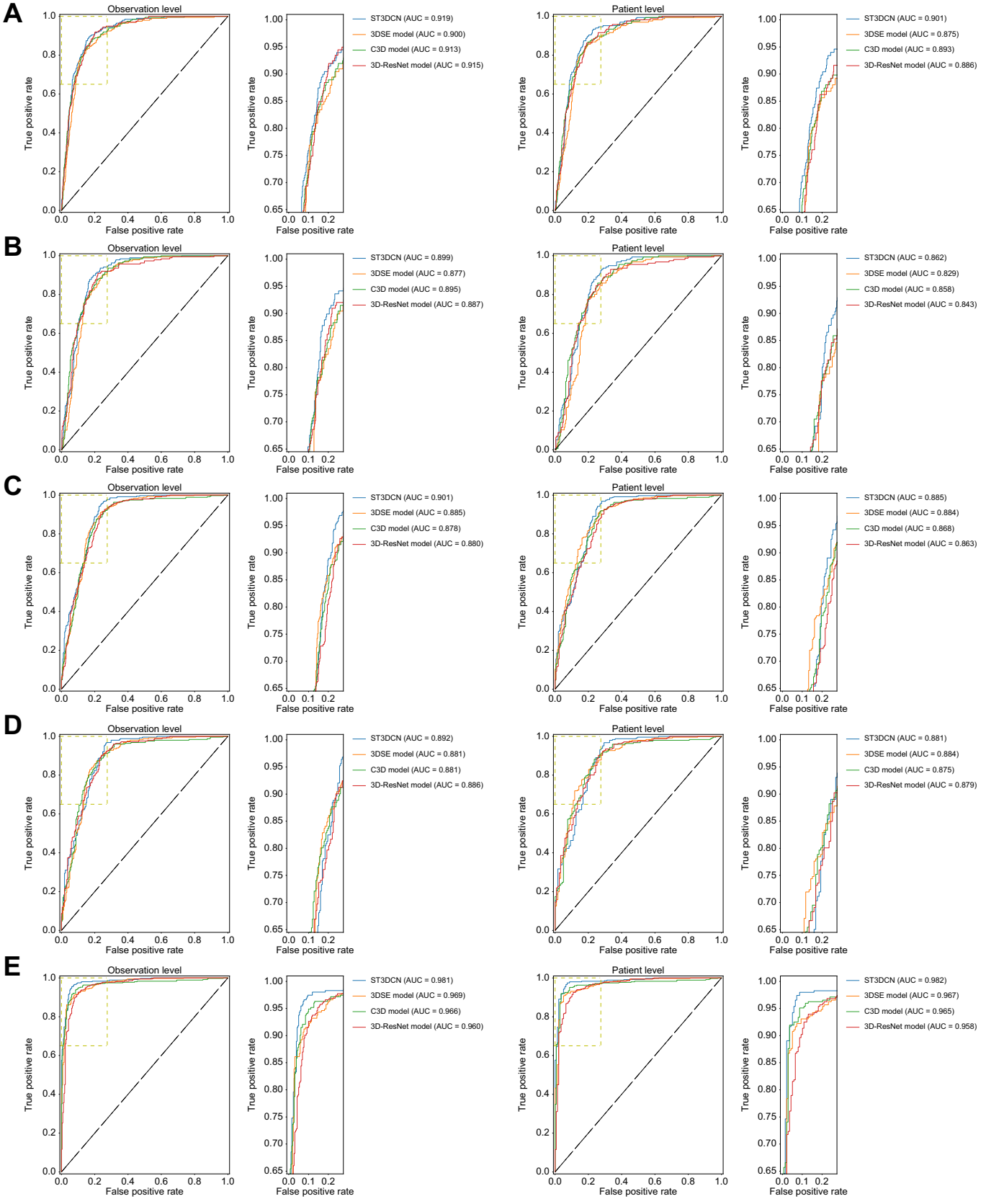


Fig. 3. AUC curves of different 3D deep learning models. (A) Internal validation for overall patients; (B) internal validation for at-risk patients; (C) external testing for overall patients; (D) external testing for at-risk patients. Among all models, Spatio-Temporal 3D Convolution Network (ST3DCN; blue line) had the highest diagnostic performance, achieving AUCs of 0.862 and 0.919 at the observation and patient level, respectively during internal validation and 0.881 and 0.901, respectively during external testing. (E) When discounting cholangiocarcinomas and liver metastases for overall patients in external testing, AUCs of 0.981 and 0.982, respectively were achieved. 3DResNet, Three-Dimensional Residual Network; 3DSE, Three-Dimensional Squeeze-and-Excitation; C3D, Convolutional Three-Dimensional.

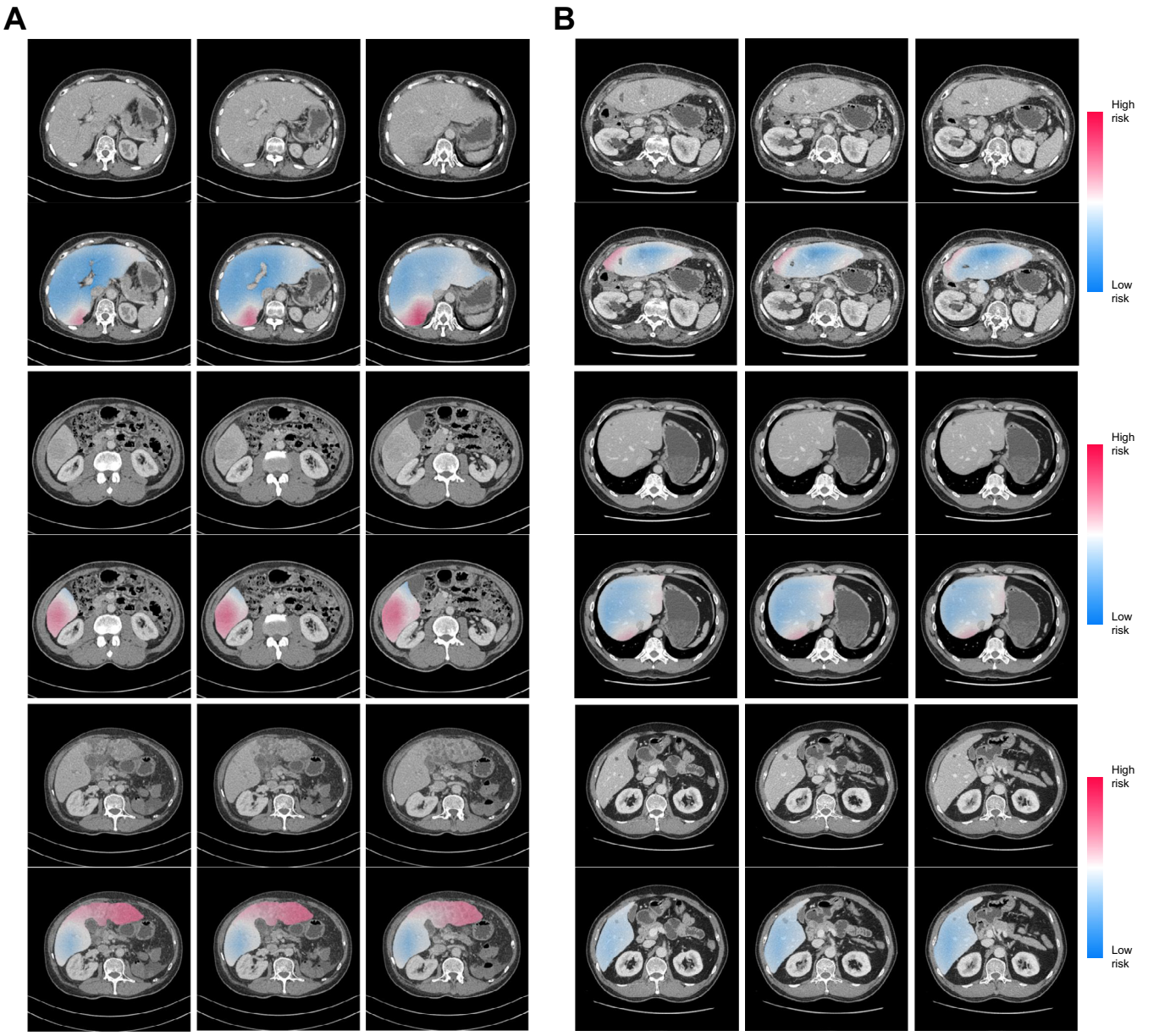


Fig. 4. Visualization of deep learning classification results. Heatmap plots for three different slices from the same computed tomography (CT) scan of three patients (A) with hepatocellular carcinoma (HCC) and (B) non-HCC. The odd rows depict the CT scan in the portovenous phase. For heatmaps in the even rows, the red color indicates the most risky area for HCC, while the blue area indicates the least risky area. Grad-CAM, gradient-weighted Class Activation Mapping.

published research lacks our more comprehensive approach because of difficulties in obtaining large, well-curated cohorts,[38] and even fewer include the active comparator of radiological interpretation or LI-RADS.[39] The AUCs of ST3DCN remained high in specifically at-risk patients and small observation sizes, as well as during external testing, an indication of the generalizability of the model. The similarly high AUC during the singular analysis of either the arterial or portovenous phase is an intriguing finding, and could suggest a potential role for opportunistic screening in general abdominal CT scans, which might only comprise a portovenous phase.[40] The consistently high NPVs of ST3DCN is important and implies that the model could be deployed as an initial screening tool for excluding HCC, facilitating the reporting prioritization of radiologists.

Third, our deep learning models were supported by strong technological foundation. We adopted a 3D approach in the architecture of our models, with ST3DCN uniquely considering the multiphasic nature of CT. This distinctive approach enabled the acquisition of different visual spatiotemporal information, which we believe was crucial for achieving a high diagnostic performance. The risk of misclassification was inevitably present, but small (Table S7), with the main issue being the number of false positives that were cholangiocarcinoma and liver metastases. Given that most metastases are hypovascular, the merits of routine acquisition of arterial dominant-phase images

during CT is disputable,[41] and with our inclusion criteria requiring an arterial phase triphasic CT scan, the subsequently low number of cholangiocarcinoma and metastases in our internal training cohort (n = 52) could had contributed to the eventual diagnostic performance. Nevertheless, when discounting these two conditions, the ST3DCN achieved a very high AUC of >0.95 during external testing. Future dedicated training of cholangiocarcinoma and metastases triphasic images could further improve its performance.

There is widespread consensus among the radiology field that AI is a tool that can assist in optimizing the decision-making process, rather than being a replacement for radiologists and clinicians.[42,43] Our present study emphasized algorithm interpretability, data quality, and generalizability, which are all important factors for a successful deep learning healthcare model.[44] However, our present model is not applicable to MR; given that MR technology differs fundamentally from that of CT, a MR-based model will require future research of comparable magnitude. Currently, both imaging modalities are equally recommended by international guidelines,[12] although CT has greater accessibility because of its rapid acquisition time, lower costs, and wider availability, especially in Asia.[16] Our study was also limited by the relatively low number of histologically confirmed HCCs and small HCCs of <2 cm, inter-reader variability during radiological interpretation, as well as only including data from Asian patients; nevertheless, >70% of liver cancers worldwide are diagnosed in Asia,[45] where our developed model is likely to have the highest clinical appeal.

In conclusion, our developed deep learning models, especially ST3DCN, were highly accurate for the diagnosis of HCC on CT, performing significantly better than radiological interpretation during internal validation and being non-inferior to radiological interpretation during external testing. The strong performance of ST3DCN was similarly retained in different sensitivity analysis, demonstrating its robustness. Thus, if widely applied, deep learning could facilitate a timely and precise diagnosis of HCC, expediting outcomes for patients, and could be a valuable tool for reducing the high mortality rate of liver cancer.

Affiliations

[1]*Department of Computer Science, The University of Hong Kong, Hong Kong, China;* [2]*Department of Mathematics and Information Technology, The Education University of Hong Kong, Hong Kong, China;* [3]*Department of Diagnostic Radiology, School of Clinical Medicine, The University of Hong Kong, Hong Kong, China;* [4]*Department of Radiology and Imaging, Queen Elizabeth Hospital, Hong Kong, China;* [5]*Department of Medical Imaging, The University of Hong Kong-Shenzhen Hospital, Shenzhen, China;* [6]*Department of Medicine, School of Clinical Medicine, The University of Hong Kong, Hong Kong, China;* [7]*Department of Medical Imaging, Sun Yat-sen University Cancer Center, State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Guangzhou, China;* [8]*Department of Medicine, The University of Hong Kong-Shenzhen Hospital, Shenzhen, China;* [9]*Department of Diagnostic and Interventional Radiology, Kwong Wah Hospital, Hong Kong, China;* [10]*Department of Radiology, Pamela Youde Nethersole Eastern Hospital, Hong Kong, China;* [11]*Department of Radiology, Queen Mary Hospital, Hong Kong, China;* [12]*State Key Laboratory of Liver Research, The University of Hong Kong, Hong Kong, China;* [13]*Department of Minimal Invasive Interventional Therapy, Sun Yat-sen University Cancer Center, State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Guangzhou, China;* [14]*Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, China*

Abbreviations

3DResNet, Three-Dimensional Residual Network; 3DSE, Three-Dimensional Squeeze-and-Excitation; AFP, alpha-fetoprotein; AI, artificial intelligence; ALP, alkaline phosphatase; ALT, alanine aminotransferase; AST, aspartate aminotransferase; C3D, Convolutional Three-Dimensional; CLAIM, Checklist for Artificial Intelligence in Medical Imaging; CT, computed tomography; DICOM, Digital Imaging and Communications in Medicine; FIB-4, Fibrosis-4; Grad-CAM, Gradient-weighted Class Activation Mapping; HCC, Hepatocellular carcinoma; LI-RADS, Liver Imaging Reporting and Data System; MASLD, metabolic associated steatotic liver disease; MR, magnetic resonance; NPV, negative predictive value; PPV, positive predictive value; ROC, Receiver Operating Characteristic; ST3DCN, Spatio-Temporal 3D Convolution Network; STARD, Standards for Reporting of Diagnostic Accuracy Studies.

Financial support

This study was supported by the Innovation and Technology Fund, The Government of the Hong Kong SAR (ref no: ITS/122/18FP); Health and Medical Research Fund (ref no: 08192936), and United Ally Research Limited, a subsidiary of Hong Kong Sanatorium and Hospital Limited.

Conflicts of interest

W-KS received speaker's fees from AstraZeneca, is an advisory board member and received speaker's fees from Abbott, received research funding from Pfizer, Alexion Pharmaceuticals, Ribo Life Sciences and Boehringer Ingelheim, and is an advisory board member, received speaker's fees and researching funding from Gilead Sciences. M-FY is an advisory board member and/or received research funding from AbbVie, Arbutus Biopharma, Assembly Biosciences, Bristol Myer Squibb, Dicerna Pharmaceuticals, GlaxoSmithKline, Gilead Sciences, Janssen, Merck Sharp and Dohme, Clear B Therapeutics, and Springbank Pharmaceuticals; and received research funding from Arrowhead Pharmaceuticals, Fujirebio Incorporation, and Sysmex Corporation. The remaining authors have no conflict of interests.

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

Authors' contributions

Study conception and design: PLHY, KW-HC, JL, WKL, M-FY, W-KS. Securing research funding: W-KS. Acquisition of data: JZ, JW, X–PS, WKK, YCH, HTC, XP, L-YM, VWMT, CH, PML, ZD, JG, NL, SY, W-KS. Analysis and interpretation of data: PLHY, KW-HC, JL, GCSL, H-MC, JZ, KMK, JH, XM, CP, W-KS. Drafting of manuscript: PLHY, KW-HC, JL, GCSL, H-MC, W-KS. Critical revision of manuscript: WKL, M-FY.

Data availability statement

Codes are available from the public Github repository https://github.com/HKUMedicineLiverAI/ST3DCN. Patient data from this study are confidential and cannot be shared in a publicly available database. Please contact the corresponding authors for limited access to the data or any further information.

Acknowledgements

We sincerely thank the following for research support: Carmen Chan and Carol Chu from the Department of Medicine, The University of Hong Kong; Siu-Tong Tse from the Department of Diagnostic Radiology, The University of Hong Kong; Min Wang, Pei Mu, Lishi Zhou, Chuan Chen, Minfeng Xu, Chun Ning, and Xiaoyu Liu from the Department of Medicine, The University of Hong Kong-Shenzhen Hospital for the. We also express our gratitude to Esther MF Wong, formerly of Pamela Youde Nethersole Eastern Hospital, and Kevin Wong, formerly of Queen Elizabeth Hospital, for their contributions to the initial phase of the study. In addition, we thank W.C. Wong of Queen Mary Hospital, Danny Cho and Brian Lee of Kwong Wah Hospital, and King Kwong Chan of Queen Elizabeth Hospital for the retrieval of radiological images. We also thank the following MBBS medical students from the University of Hong Kong: Kevin Chuek-Kin Cheng, Shing-Chung Chow, Hoi-Fan Li, Cheuk-Yin Wong, Tiffany Hoi-Ching Chan, Jonathan Tin-Kit Chung, Darren Li-Liang Wong, Matthew Shing Him Lee, and Ernest Sing-Hong Chui, for their participation in this study.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jhepr.2024.101219.

References

Author names in bold designate shared co-first authorship.

[1] McGlynn KA, Petrick JL, El-Serag HB. Epidemiology of hepatocellular carcinoma. Hepatology 2021;73(Suppl 1):4–13.
[2] Ferlay J, Erivik M, Lam F, et al. Global cancer observatory: cancer today. Lyon, France: International Agency for Research on Cancer; 2022. https://gco.iarc.fr/today. [Accessed 24 September 2024].
[3] McGlynn KA, Petrick JL, London WT. Global epidemiology of hepatocellular carcinoma: an emphasis on demographic and regional variability. Clin Liver Dis 2015;19:223–238.
[4] **Wang JH**, **Wang CC**, Hung CH, et al. Survival comparison between surgical resection and radiofrequency ablation for patients in BCLC very early/early stage hepatocellular carcinoma. J Hepatol 2012;56:412–418.
[5] Shindoh J, Andreou A, Aloia TA, et al. Microvascular invasion does not predict long-term survival in hepatocellular carcinoma up to 2 cm: reappraisal of the staging system for solitary tumors. Ann Surg Oncol 2013;20:1223–1229.
[6] Marrero JA, Kulik LM, Sirlin CB, et al. Diagnosis, staging, and management of hepatocellular carcinoma: 2018 practice guidance by the American association for the study of liver diseases. Hepatology 2018;68:723–750.
[7] Omata M, Cheng AL, Kokudo N, et al. Asia-Pacific clinical practice guidelines on the management of hepatocellular carcinoma: a 2017 update. Hepatol Int 2017;11:317–370.
[8] Chernyak V, Fowler KJ, Kamaya A, et al. Liver imaging reporting and data system (LI-RADS) version 2018: imaging of hepatocellular carcinoma in at-risk patients. Radiology 2018;289:816–830.
[9] **van der Pol CB**, **Lim CS**, Sirlin CB, et al. Accuracy of the Liver Imaging Reporting and Data System in computed tomography and magnetic resonance image analysis of hepatocellular carcinoma or overall malignancy-a systematic review. Gastroenterology 2019;156:976–986.
[10] Ronot M, Fouque O, Esvan M, et al. Comparison of the accuracy of AASLD and LI-RADS criteria for the non-invasive diagnosis of HCC smaller than 3cm. J Hepatol 2018;68:715–723.
[11] Lee S, Kim YY, Shin J, et al. Percentages of hepatocellular carcinoma in LI-RADS categories with CT and MRI: a systematic review and meta-analysis. Radiology 2023;307:e220646.
[12] Heimbach JK, Kulik LM, Finn RS, et al. AASLD guidelines for the treatment of hepatocellular carcinoma. Hepatology 2018;67:358–380.
[13] Atiq O, Tiro J, Yopp AC, et al. An assessment of benefits and harms of hepatocellular carcinoma surveillance in patients with cirrhosis. Hepatology 2017;65:1196–1205.
[14] Bruno MA, Walker EA, Abujudeh HH. Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction. Radiographics 2015;35:1668–1676.
[15] Hosny A, Parmar C, Quackenbush J, et al. Artificial intelligence in radiology. Nat Rev Cancer 2018;18:500–510.
[16] He L, Yu H, Shi L, et al. Equity assessment of the distribution of CT and MRI scanners in China: a panel data analysis. Int J Equity Health 2018;17:157.
[17] Mongan J, Moy L, Kahn Jr CE. Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. Radiol Artif Intell 2020;2:e200029.
[18] Rueckert D, Schnabel JA. Model-based and data-driven strategies in medical image computing. Proc IEEE 2019;108:110–124.
[19] Zhou SK, Greenspan H, Davatzikos C, et al. A review of deep learning in medical imaging: imaging traits, technology trends, case studies with progress highlights, and future promises. Proc IEEE 2021;109:820–838.
[20] EASL-ALEH Clinical Practice Guidelines: non-invasive tests for evaluation of liver disease severity and prognosis. J Hepatol 2015;63:237–264.
[21] Bhattacharya S, Dhillon AP, Rees J, et al. Small hepatocellular carcinomas in cirrhotic explant livers: identification by macroscopic examination and lipiodol localization. Hepatology 1997;25:613–618.
[22] Choi MH, Choi JI, Lee YJ, et al. MRI of small hepatocellular carcinoma: typical features are less frequent below a size cutoff of 1.5 cm. Am J Roentgenol 2017;208:544–551.
[23] van der Pol CB, McInnes MDF, Salameh JP, et al. Impact of reference standard on CT, MRI, and contrast-enhanced US LI-RADS diagnosis of hepatocellular carcinoma: a meta-analysis. Radiology 2022;303:544–545.
[24] Hsu C, Chen BB, Chen CH, et al. Consensus development from the 5th Asia-Pacific primary liver cancer expert meeting (APPLE 2014). Liver Cancer 2015;4:96–105.
[25] Kim T-H, Kim SY, Tang A, et al. Comparison of international guidelines for noninvasive diagnosis of hepatocellular carcinoma: 2018 update. Clin Mol Hepatol 2019;25:245–263.
[26] **Piñero F**, **Thompson MA**, **Telli FD**, et al. LI-RADS 4 or 5 categorization may not be clinically relevant for decision-making processes: a prospective cohort study. Ann Hepatol 2020;19:662–667.
[27] Singh SP, Wang L, Gupta S, et al. 3D deep learning on medical images: a review. Sensors 2020;20:5097.
[28] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3D convolutional networks. In: 2015 IEEE International Conference on Computer Vision. Piscataway: IEEE; 2015. p. 4489–4497.
[29] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE; 2016. p. 770–778.
[30] Selvaraju RR, Cogswell M, Das A, et al. Grad-cam: visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International Conference on Computer Vision. Piscataway: IEEE; 2017. p. 618–626.
[31] Sundararajan M, Najmi A. The many Shapley values for model explanation. PMLR 2020;119:9269–9278.
[32] Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 2011;12:1–8.
[33] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29–36.
[34] Liu JP, Ma MC, Wu CY, et al. Tests of equivalence and non-inferiority for diagnostic accuracy based on the paired areas under ROC curves. Stat Med 2006;25:1219–1238.
[35] **Huang JL**, **Sun Y**, Wu ZH, et al. Differential diagnosis of hepatocellular carcinoma and intrahepatic cholangiocarcinoma based on spatial and channel attention mechanisms. J Cancer Res Clin Oncol 2023;149:10161–10168.
[36] Wang X, Li N, Yin X, et al. Classification of metastatic hepatic carcinoma and hepatocellular carcinoma lesions using contrast-enhanced CT based on EI-CNNet. Med Phys 2023;50:5630–5642.
[37] **Gao R**, **Zhao S**, **Aishanjiang K**, et al. Deep learning for differential diagnosis of malignant hepatic tumors based on multi-phase contrast-enhanced CT and clinical data. J Hematol Oncol 2021;14:154.
[38] **Shi W**, **Kuang S**, Cao S, et al. Deep learning assisted differentiation of hepatocellular carcinoma from focal liver lesions: choice of four-phase and three-phase CT imaging protocol. Abdom Radiol (NY) 2020;45:2688–2697.
[39] **Wang M**, **Fu F**, **Zheng B**, et al. Development of an AI system for accurately diagnose hepatocellular carcinoma from computed tomography imaging data. Br J Cancer 2021;125:1111–1121.
[40] Pickhardt PJ. Value-added opportunistic CT screening: state of the art. Radiology 2022;303:241–254.
[41] Kanematsu M, Kondo H, Goshima S, et al. Imaging liver metastases: review and update. Eur J Radiol 2006;58:217–228.
[42] What the radiologist should know about artificial intelligence - an ESR white paper. Insights Imaging 2019;10:44.
[43] Summary of the proceedings of the International Forum 2021. A more visible radiologist can never be replaced by AI. Insights Imaging 2022;13:43.
[44] Wang F, Casalino LP, Khullar D. Deep learning in medicine-promise, progress, and challenges. JAMA Intern Med 2019;179:293–294.
[45] Petrick JL, Florio AA, Znaor A, et al. International trends in hepatocellular carcinoma incidence, 1978-2012. Int J Cancer 2020;147:317–330.

Keywords: HCC; AI; Liver cancer; CT; LIRADS; Imaging.
Received 8 April 2024; received in revised form 10 September 2024; accepted 10 September 2024; Available online 16 September 2024