

Automatic Real-Time Detection and Diagnosis of Liver Tumor with Ultrasound

Chih-Horng Wu^{1,2,*}, Jin-Chuan Sheu^{3,4}, Pei-Lien Chou⁵, Jonathan Lee⁶, Hsiao-Ching Nien^{3,7,*}

¹Department of Medical Imaging, National Taiwan University Hospital, Taipei, Taiwan; ²Hepatitis Research Center, National Taiwan University Hospital, Taipei, Taiwan; ³Good Liver Foundation, Taipei, Taiwan; ⁴Department of Internal Medicine, National Taiwan University Hospital, Taipei, Taiwan; ⁵YUAN High-Tech Development Co., Ltd, Taipei, Taiwan; ⁶Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan; ⁷Department of Family Medicine, National Taiwan University Hospital, Taipei, Taiwan

*These authors contributed equally to this work

Correspondence: Hsiao-Ching Nien, Department of Family Medicine, National Taiwan University Hospital, No. 7, Zhongshan S Road, Zhongzheng District, Taipei City, Taiwan, 100, Tel +886-223123456, Email celian0916@gmail.com

Background/Aim: Ultrasonography is the most commonly used screening tool for hepatocellular carcinoma (HCC). However, the diagnostic performance of ultrasound is highly operator dependent. We aimed to develop deep learning (DL) models to automatically diagnose and detect hepatic lesions in a larger dataset, with HCC as the dominant malignancy.

Methods: We enrolled patients diagnosed with hepatic tumors using abdominal ultrasound between January 2002 and December 2020 in a retrospective cohort with a diagnosis of malignant and benign lesions. A total of 1576 patients with 4599 images and 6001 lesions were analyzed. Deep learning models included ResNet50, Xception, Inception Resnet V2, EfficientNet-B5, EfficientNetV2-S, EfficientNetV2-L, Swin-T, and Swin-B for diagnosis and YOLOR for lesion detection. We analyzed the area under the curve (AUC) to determine the diagnostic performance and choose the best model. The mean Average Precision (mAP) score was then evaluated for real-time lesion detection using the area under the precision-recall curve after the average of each category.

Results: The dataset was separated into 1061 in training, 373 in validation, and 142 testing sets. The AUC for ResNet50, Xception, Inception Resnet V2, EfficientNet-B5, EfficientNetV2-S, EfficientNetV2-L, Swin-T, and Swin-B are 0.88, 0.89, 0.88, 0.90, 0.85, 0.89, 0.89, and 0.90, respectively. The mAP scores for detecting and differentiating malignant and benign lesions for YOLOR-W6 and YOLOR-D6 in the validation and testing sets were 0.5134/0.5342 and 0.5410/0.5631.

Conclusion: Our study demonstrated that DL models can differentiate between benign and malignant lesions with high accuracy on ultrasound images. Simultaneous DL-based lesion detection and classification are also possible using real-time ultrasonography.

Keywords: deep learning, ultrasonography, artificial intelligence, diagnosis, liver neoplasms

Introduction

Hepatocellular carcinoma (HCC) accounts for approximately 90% of primary liver cancers and is closely associated with chronic hepatitis B and C infections and liver cirrhosis in Asian countries.^{1–3} Due to the implementation of effective surveillance programs, the proportion of HCC cases detected at an early stage has risen significantly—from 5–10% to 40–60%—allowing more patients to receive potentially curative treatment.⁴

Among surveillance tools, abdominal ultrasound (US) remains a first-line modality because of its higher sensitivity than serum biomarkers and greater cost-effectiveness compared to dynamic computed tomography (CT) or magnetic resonance imaging (MRI) in high-risk populations.^{5,6} In earlier meta-analyses, US demonstrated pooled sensitivity and specificity of 60% and 97%, respectively, compared to 68% and 93% for CT and 81% and 85% for MRI.⁷ However, with advancements in imaging resolution and the advent of contrast-enhanced ultrasound, updated systematic reviews and multicenter studies now report US sensitivity reaching 84–95%, comparable to that of CT and MRI.^{8,9}

Moreover, US is widely utilized for imaging-guided radiofrequency ablation (RFA) owing to its real-time feedback, lack of radiation exposure, and cost benefits. Nevertheless, its effectiveness may be hindered by anatomical challenges

and reliance on operator expertise. CT, on the other hand, excels in cross-sectional imaging clarity, enabling superior visualization of lesions and accurate needle placement—particularly for deep or inconspicuous tumors. Its limitations lie in the absence of real-time monitoring and radiation exposure. The selection between US and CT should therefore be individualized, taking into account tumor location, patient body habitus, and procedural requirements.¹⁰

Despite these advancements, the diagnostic performance of US remains heavily influenced by operator skill and the quality of the imaging system, including probe type and settings. Additional factors—such as hepatic parenchymal heterogeneity, thick abdominal walls, and ascites—can further impair image quality and diagnostic confidence.^{11,12} As a result, these limitations could lead to discrepancies in diagnostic results, as well as increase the probabilities of misdiagnosis or missed diagnosis. These challenges underscore the urgent need for automated and standardized approaches to support less experienced sonographers and enhance diagnostic accuracy.

Currently, deep learning (DL)—a subset of artificial intelligence—has shown promise as a transformative tool in medical image analysis. DL algorithms, particularly convolutional neural networks (CNNs), have demonstrated the capacity to automate complex image interpretation tasks such as lesion detection, classification, and segmentation. Deep learning (DL) has emerged as a breakthrough technology in computer vision and has the potential to perform automatic ultrasound image analysis tasks, such as lesion/nodule classification and object detection.¹³ The challenges of using DL with ultrasound images mainly lie in the fact that ultrasonography produces dynamic, cross-sectional images, whereas most DL research focuses on optically acquired images outside an object. Therefore, unlike CT or MRI scans, which have been proposed in several studies,^{14,15} more versatile training data and novel neural network architectures are required for the application of deep learning in ultrasonography than in CT or MRI.

Several studies have focused on convolutional neural network (CNN)-based DL algorithms to classify liver fibrosis/cirrhosis and steatosis by using ultrasonography.^{16,17} One study demonstrated high accuracy in detecting focal liver lesions with ultrasound in a modest dataset (367 images) and characterizing malignancy in a Western population, with only 6 (1.6%) images of HCC.¹⁸ Another study used DL to detect HCC using intraoperative ultrasonography.¹⁹ No large-scale study has used DL to simultaneously detect and diagnose hepatic tumors. Therefore, our study aimed to develop a DL model to detect lesions in a larger dataset and distinguish malignant from benign hepatic tumors in an Eastern population, with HCC as the dominant malignancy.

Materials and Methods

Study Population

This retrospective single-center cohort study enrolled patients diagnosed with hepatic tumors using abdominal ultrasound between January 2002 and December 2020 at the National Taiwan University Hospital (NTUH), a tertiary referral center in northern Taiwan. The diagnosis of malignant tumors consisting of HCC, cholangiocarcinoma, or metastasis was identified from the NTUH Cancer Registry Database and confirmed by dynamic CT, MRI, or pathologic examination according to the guidelines proposed by the American Association for the Study of Liver Disease (AASLD) and the Taiwan Liver Cancer Association (TLCA).^{2,20} In contrast, the impression of benign tumors was composed of hepatic cysts, hemangiomas, focal fatty sparing, focal nodular hyperplasia, and other benign findings diagnosed using dynamic CT, MRI, or follow-up. In addition, those who had no adequate ultrasound images for analysis or no confirmatory studies for the final diagnosis were excluded. The Institutional Review Board of our hospital approved this study (201801011RINB), and the need for informed consent was waived because of its retrospective design, which posed no additional risk to the patients.

Ultrasound Images Dataset

Ultrasound images were extracted from the imaging archive of the NTUH for review, and ultrasound examinations were performed using eight scanners (Sonolayer SSA-250A, 340A, 770A, 790A; Toshiba, Tokyo, Japan; TUS-Aplio 500, i800; Canon Medical Systems, Tokyo, Japan; Arietta 70, 850; Hitachi Healthcare, Tokyo, Japan). Only the examinations closest to the date of diagnosis were selected for patients who had undergone more than one ultrasound examination. First, 1576 patients were enrolled, with 4599 images, and 6001 lesions were analyzed (Table 1). Second,

Table 1 Patients' Characteristics

Characters	Training set (N = 1061)	Validation set (N = 373)	Testing set (N = 142)	Total (N = 1576)
Age (year)	64.9 (18.9)	66.0 (18.5)	66.0 (20.0)	65.2 (18.9)
Male	683 (64.4%)	251 (67.3%)	99 (69.7%)	1033 (65.6%)
HBV	485 (45.7%)	190 (49.8%)	85 (59.8%)	760 (48.2%)
HCV	230 (21.7%)	72 (19.3%)	32 (22.5%)	334 (21.2%)
Liver cirrhosis	436 (41.2%)	136 (36.5%)	57 (40.1%)	629 (40.0%)
Fatty liver	103 (9.7%)	20 (5.4%)	6 (4.2%)	129 (8.2%)
Platelet (K/)	175 (119)	182 (116)	169 (120)	177 (119)
AST (U/L)	39 (62)	46 (66)	54 (62)	41 (64)
ALT (U/L)	34 (45)	34 (48)	39 (69)	34 (47)
Alb (g/dL)	3.9 (1.2)	3.9 (1.0)	3.9 (1.0)	3.9 (1.2)
T-Bil (mg/dL)	0.90 (0.93)	0.90 (0.85)	0.89 (0.82)	0.90 (0.99)
Cre (mg/dL)	0.80 (0.30)	0.90 (0.30)	0.80 (0.30)	0.82 (0.30)
AFP (ng/mL)	6.31 (91.57)	8.04 (117.90)	21.1 (694.27)	7.18 (116.02)

Note: Presented as median (interquartile range) or number (percentage).

Abbreviations: HBV: seropositive for hepatitis B surface antigen; seropositive for hepatitis C antibody; AST: aspartate aminotransferase; ALT: alanine aminotransferase; Alb: albumin; T-Bil: total bilirubin; Cre: creatinine; AFP: alpha-fetoprotein.

Table 2 Number of Lesions in Each Dataset

	Lesion Type	Training Set	Validation Set	Testing Set	Total
Malignant lesions (N = 4671)	Hepatocellular carcinoma	3448	243	426	4117
	Cholangiocarcinoma	227	28	22	277
	Metastasis	239	10	28	277
Benign lesions (N = 1330)	Cyst	271	28	2	301
	Hemangioma	436	10	39	485
	Focal fatty sparing	146	31	6	183
	Focal nodular hyperplasia	281	22	33	336
	Others	21	1	3	25
Total		5069	373	559	6001

the lesions were categorized as malignant (HCC, cholangiocarcinoma, and metastasis) or benign (hepatic cyst, hemangioma, focal fatty sparing, focal nodular hyperplasia, and other benign findings [adenoma and focal fatty change]). Finally, the dataset was randomly split into training, validation, and testing sets in a ratio of 45:3:5. There were 5069 lesions (3914 malignant and 1155 benign) in the training set, 373 lesions (281 malignant and 92 benign) in the validation set, and 559 lesions (476 malignant and 83 benign) in the testing set (Table 2). Because we had a large number of cases, including 3914 benign and 1155 malignant lesions in the training set, we did not apply any data imbalance correction.

Data Preparation

Patients seropositive for hepatitis B surface antigen were considered to have hepatitis B virus (HBV) infection and those seropositive for hepatitis C antibody were considered to have hepatitis C virus (HCV) infection. The presence of liver cirrhosis and fatty liver was determined by ultrasound examination. After removing the identity of the patient, eight experienced ultrasonographers with board certification labeled the lesions with bounding boxes as our target regions. We then cropped the target regions from the original images and resized them. Finally, gamma adjustment is used to make the original images brighter or darker. Data augmentation strategies, including flip, rotation, crop, resize, and over-sampling, were employed to improve insufficient quantity and imbalanced classification.

Modeling

Distinguishing Between Malignant and Benign Hepatic Tumors

Eight different models were trained to classify the liver tumor types: ResNet50, Xception, Inception Resnet V2, EfficientNet-B5, EfficientNetV2-S, EfficientNetV2-L, Swin-T, and Swin-B. In ResNet50, He et al discovered that the result would be better if nonlinear layers were used to asymptotically approximate multiple different residual functions, instead of a single complicated nonlinear function.²¹ Inception ResNet V2 comprises the residual and inception blocks. Inception blocks can also decrease memory consumption and compute complexity.²² EfficientNet-B5 utilizes a compound scaling method that uniformly scales the depth, width, and resolution of the network, allowing it to achieve a better performance. EfficientNet-B5 has approximately 30 million parameters and achieves state-of-the-art performance in various computer vision tasks such as image classification, object detection, and segmentation.²³ It achieves high accuracy while requiring fewer computational resources than other larger models. During modeling, the training loss function was cross-entropy and the learning rate was 0.001. EfficientNetV2-L is a type of convolutional neural network that has faster training speed and better parameter efficiency than previous models. Swin-B is a new type of transformer architecture that uses sliding windows and hierarchical structures. It handles images through a hierarchical structure similar to that of a CNN, allowing the model to handle images of different scales flexibly. In addition, Swin-B uses window self-attention, which reduces the computational complexity.

Automatic Lesion Detection with Classification on Ultrasound Images

We used You Only Learn One Representation(YOLOR) for automatic detection with classification. The backbone of the initial YOLOR was CSPDarknet53, which has the same backbone as ScaledYOLOv4.²⁴ Furthermore, YOLOR can achieve higher accuracy with lower computing power, and separate into YOLOR -D6 and YOLOR -W6 in our study. The neck structure of YOLOR is based on SPP-PAN, which combines the concepts of spatial pyramid pooling and path aggregation to enhance multi-scale perception and feature representation capabilities. Therefore, YOLOR can simultaneously consider the network and hardware to optimize the model. Finally, YOLOR uses the head-layer features added and multiplied by the implicit knowledge. The implicit knowledge is a trainable vector.

Statistical Analysis

Data analysis was performed using Excel 2016 (Microsoft, Redmond, WA, USA) and MedCalc 20.023 (MedCalc Software Ltd., Ostend, Belgium). We performed a receiver operating characteristic (ROC) curve analysis to determine the diagnostic performance of ResNet50, Xception, Inception Resnet V2, EfficientNet-B5, EfficientNetV2-S, EfficientNetV2-L, Swin-T, and Swin-B in predicting malignancy. We assumed a 1% significance level test ($\alpha = 0.01$), 99% power ($\beta = 0.01$), diagnostic performance (area under curve [AUC]) of the deep learning algorithm = 0.90, null hypothesis value = 0.85, ratio negative/positive = 1/3.5, and the required cases in the negative/positive groups were 258/902. Therefore, we estimated the sample size to be 1200. When $\text{IoU} \geq 0.5$ was in the bounding box, we defined a positive result for lesion detection. The mean Average Precision (mAP) score was then evaluated for lesion detection using the area under the precision-recall curve after the average of each category. A p-value < 0.01 was considered at p < analyses.

Results

Patients' Characteristics

A total of 1576 patients were into 1061 in the training, 373 in validation and 142 in testing set. Of them, 1033 (65.6%) and 542 (34.4%) patients were male and female with a median age (interquartile range) of 65.2 (18.9) years. 760 (48.2%) had HBV infection, 334 (21.2%) had HCV infection, 43 (2.7%) had HBV/HCV coinfection, and 524 (33.3%) had no evidence of HBV or HCV infection. Moreover, 629 (40.0%) patients had liver cirrhosis, and 129 (8.2%) had fatty liver. There was a relative balance in the patient characteristics among the training, validation, and testing sets, as shown in Table 1.

Diagnostic Performance of DL Models

The AUC for ResNet50, Xception, Inception ResNet V2, and EfficientNet-B5 were 0.88, 0.89, 0.88, and 0.90 (Figure 1). EfficientNetV2-S, EfficientNetV2-L, Swin-T, and Swin-B were 0.85, 0.89, 0.89, and 0.90, in validation and 0.75, 0.78, 0.77, and 0.77, in testing set, respectively (Figure 2). The accuracies of EfficientNetV2-S, EfficientNetV2-L, Swin-T, and Swin-B were 0.761, 0.810, 0.796, and 0.836, respectively, in the validation set, and 0.671, 0.733, 0.710, and 0.699, respectively, in the testing set (Table 3). After the optimal threshold was selected for Swin-B, the accuracy, recall (sensitivity), precision, and F1 score for the validation and testing datasets were 0.836, 0.837, 0.626, 0.716 and 0.699, 0.699, 0.289, 0.408 (Table 3). The best accuracies for the eight classifications are Swin-B for the validation set (0.735) and Swin-T for the testing set (0.766), respectively. Among the eight types of hepatic lesions, HCC and cysts had the highest accuracies for malignant and benign lesions, respectively.

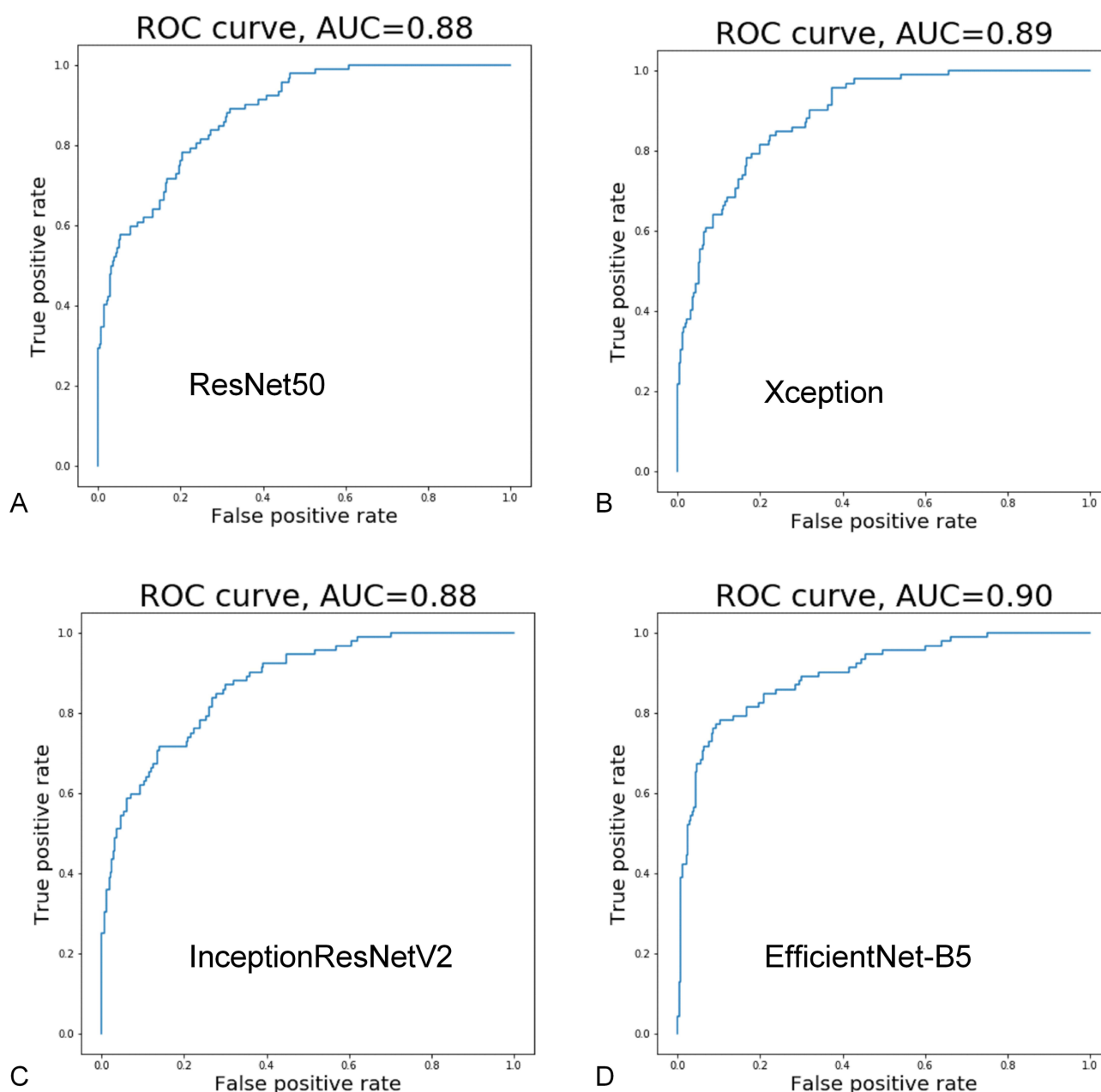


Figure 1 Diagnostic performance for four deep learning algorithms, including (A) ResNet50, (B) Xception, (C) InceptionResnetV2, and (D) EfficientNet-B5.

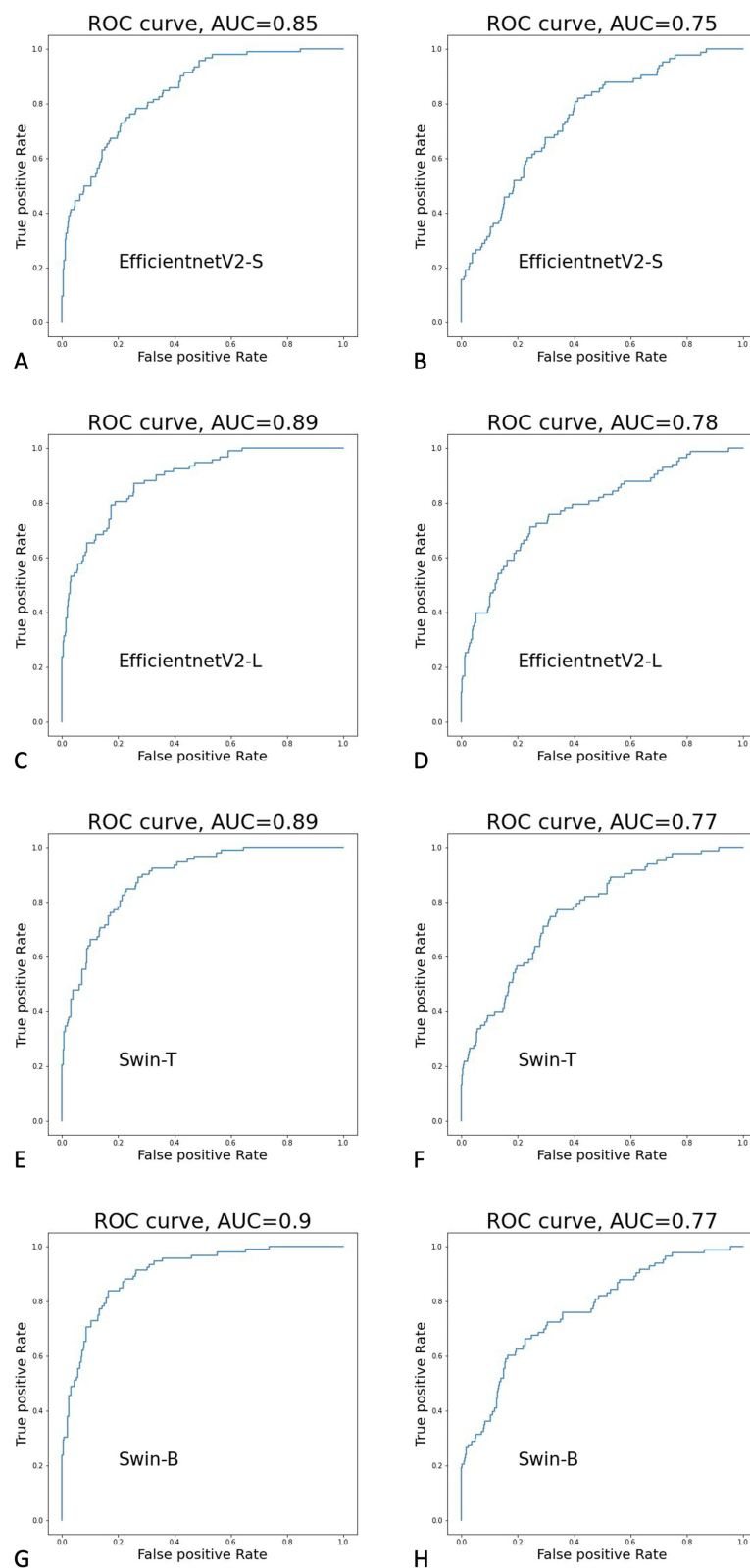


Figure 2 Diagnostic performance in validation and testing for EfficientNetV2-S (A and B), EfficientNetV2-L (C and D), Swin-T (E and F), and Swin-B (G and H).

Table 3 Diagnostic Performance of Differentiation Between Malignant and Benign Lesions and Classifying 8 Kinds of Hepatic Lesions

Differentiation between benign and malignant lesions										
	Model type	Threshold	Accuracy (95% CI)		Recall (95% CI)		Precision (95% CI)		F1 score (95% CI)	
Validation set	EfficientNetV2-S	0.312	0.761 (0.697–0.804)		0.761 (0.696–0.81)		0.511 (0.4–0.587)		0.611 (0.515–0.673)	
	EfficientNetV2-L	0.210	0.810 (0.745–0.842)		0.804 (0.744–0.843)		0.583 (0.459–0.657)		0.676 (0.571–0.737)	
	Swin-T	0.219	0.796 (0.748–0.836)		0.804 (0.747–0.84)		0.561 (0.473–0.657)		0.661 (0.583–0.73)	
	Swin-B	0.195	0.836 (0.777–0.871)		0.837 (0.776–0.873)		0.626 (0.496–0.705)		0.716 (0.608–0.782)	
Testing set	EfficientNetV2-S	0.237	0.671 (0.624–0.742)		0.675 (0.627–0.739)		0.263 (0.21–0.346)		0.378 (0.315–0.465)	
	EfficientNetV2-L	0.104	0.733 (0.657–0.775)		0.723 (0.654–0.78)		0.323 (0.234–0.4)		0.446 (0.345–0.527)	
	Swin-T	0.103	0.710 (0.648–0.753)		0.711 (0.656–0.761)		0.299 (0.222–0.374)		0.421 (0.333–0.494)	
	Swin-B	0.077	0.699 (0.642–0.76)		0.699 (0.643–0.763)		0.289 (0.225–0.382)		0.408 (0.335–0.507)	
Classifying 8 kinds of hepatic lesions										
	Model type	Total Acc.	HCC Acc.	Cho Acc.	Met Acc.	Cys Acc.	Hem Acc.	FFS Acc.	FNH Acc.	Oth Acc.
Validation set	EfficientNetV2-S	0.694 (0.654–0.737)	0.922 (0.885–0.953)	0.036 (0.0–0.115)	0.000 (0.0–0.0)	0.964 (0.885–1.0)	0.300 (0.0–0.6)	0.032 (0.0–0.1)	0.136 (0.0–0.286)	0.000 (0.0–0.0)
	EfficientNetV2-L	0.700 (0.657–0.737)	0.909 (0.875–0.937)	0.000 (0.0–0.0)	0.000 (0.0–0.0)	0.929 (0.833–1.0)	0.300 (0.0–0.6)	0.290 (0.136–0.5)	0.091 (0.0–0.214)	0.000 (0.0–0.0)
	Swin-T	0.727 (0.684–0.767)	0.914 (0.876–0.944)	0.071 (0.0–0.167)	0.000 (0.0–0.0)	0.964 (0.885–1.0)	0.300 (0.0–0.6)	0.419 (0.25–0.609)	0.182 (0.045–0.35)	0.000 (0.0–0.0)
	Swin-B	0.735 (0.697–0.772)	0.881 (0.843–0.918)	0.286 (0.12–0.444)	0.000 (0.0–0.0)	1.000 (1.0–1.0)	0.200 (0.0–0.5)	0.645 (0.476–0.829)	0.091 (0.0–0.227)	0.000 (0.0–0.0)
Testing set	EfficientNetV2-S	0.755 (0.717–0.787)	0.944 (0.918–0.966)	0.000 (0.0–0.0)	0.071 (0.0–0.168)	1.000 (0.0–1.0)	0.359 (0.222–0.488)	0.000 (0.0–0.0)	0.061 (0.0–0.143)	0.000 (0.0–0.0)
	EfficientNetV2-L	0.757 (0.719–0.794)	0.930 (0.907–0.95)	0.045 (0.0–0.167)	0.107 (0.0–0.223)	1.000 (0.0–1.0)	0.462 (0.313–0.606)	0.333 (0.0–0.9)	0.030 (0.0–0.1)	0.000 (0.0–0.0)
	Swin-T	0.766 (0.733–0.8)	0.944 (0.918–0.967)	0.136 (0.0–0.319)	0.143 (0.034–0.286)	1.000 (0.0–1.0)	0.385 (0.267–0.524)	0.167 (0.0–0.626)	0.030 (0.0–0.1)	0.000 (0.0–0.0)
	Swin-B	0.744 (0.708–0.782)	0.920 (0.896–0.95)	0.045 (0.0–0.158)	0.107 (0.0–0.214)	1.000 (0.0–1.0)	0.385 (0.225–0.535)	0.333 (0.0–0.903)	0.030 (0.0–0.1)	0.000 (0.0–0.0)

Abbreviations: Acc., Accuracy; HCC, Hepatocellular carcinoma; Cho, Cholangiocarcinoma; Met, Metastasis; Cys, Cyst; Hem, Hemangioma; FFS, Focal fatty sparing; FNH, Focal nodular hyperplasia; Oth, Others.

Automatic Lesion Detection and Diagnosis on Ultrasound Images

The mAP scores for differentiating malignant and benign lesions for YOLOR -W6/ YOLOR -D6 in the validation and testing sets were 0.5134/0.5342 and 0.5410/0.5631. The mAP scores for the classification of the eight types of lesions for YOLOR -W6/ YOLOR -D6 in the validation and testing sets were 0.2958/0.2938 and 0.3029/0.3135 (Figure 3). A visual presentation of the ultrasound images with a predictive bounding box is shown in Figure 4.

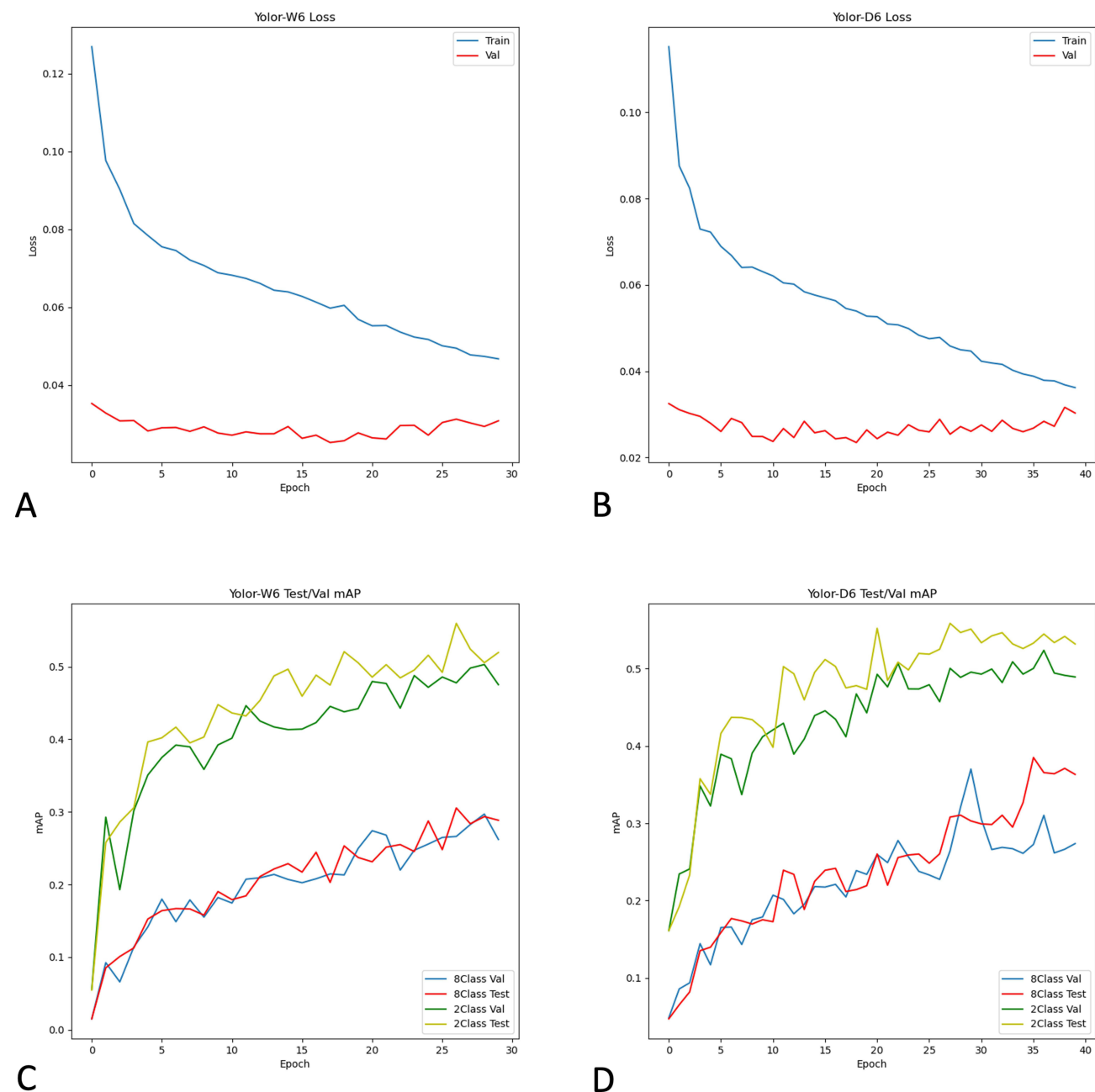


Figure 3 The loss to differentiate between malignant and benign lesions in training and validation for YOLOR-W6 (A) and YOLOR-D6 (B). The mean Average Precision score to differentiate between malignant and benign lesions and classify 8 kinds of hepatic lesions in validation and testing for YOLOR-W6 (C) and YOLOR-D6 (D).

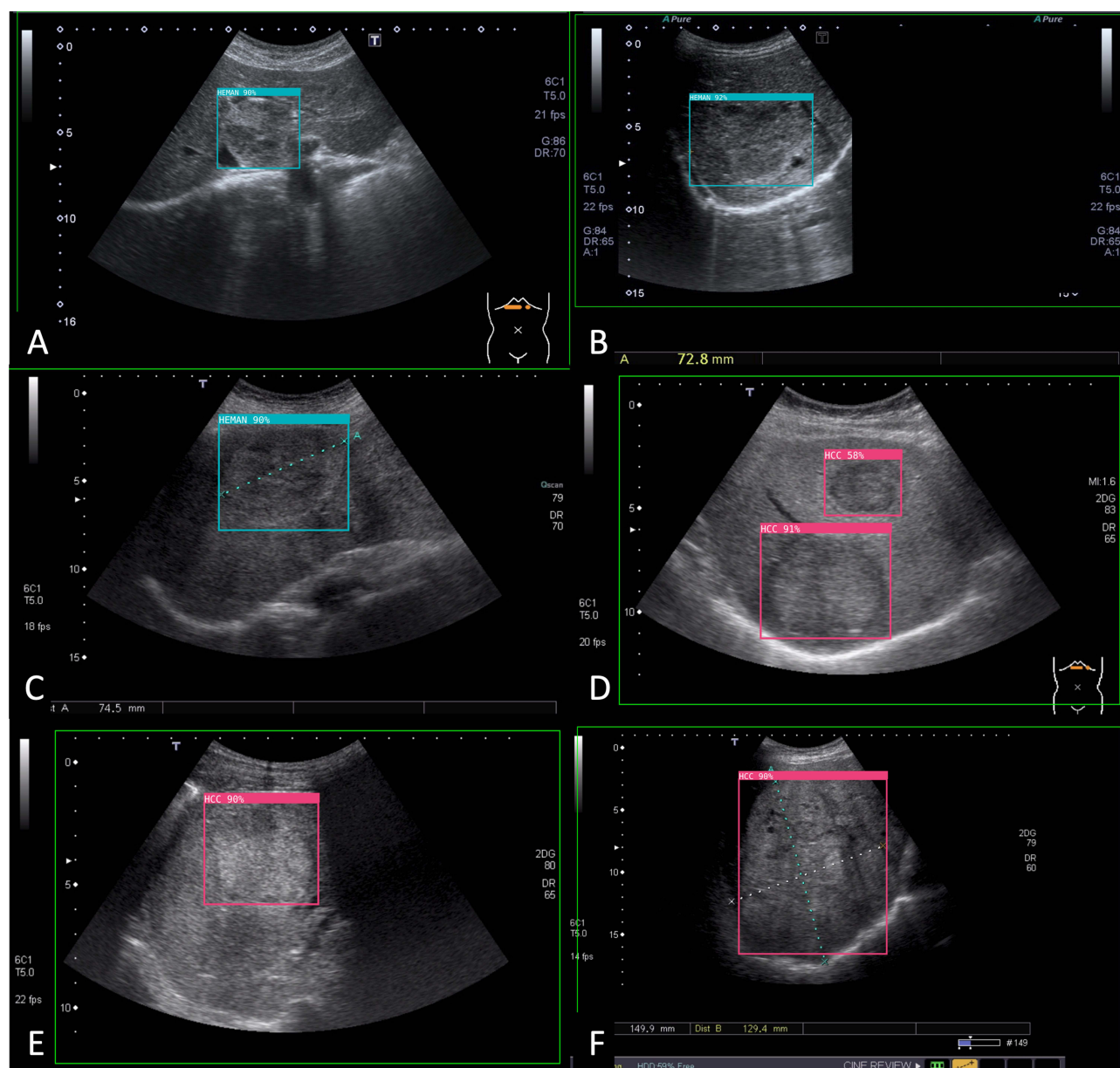


Figure 4 Visual presentation ultrasound images with predictive bounding box. The blue/red bounding boxes represented benign (A–C) and malignant (D–F) lesions. The lesions can be detected at left lobe (A), split view (B), right lobe (C), two lesions in one view (D) and sine images (F).

Subgroup Analysis

The mAP scores for lesions > 5 cm to differentiate malignant and benign lesions for YOLOR -W6/ YOLOR -D6 in the validation and testing sets were 0.561, 0.628, and 0.390/0.397, respectively. In contrast, the mAP score for lesions less than 5 cm to differentiate malignant and benign lesions for YOLOR -W6/ YOLOR -D6 in validation and testing sets were 0.435/0.491 and 0.473/0.599. Furthermore, the mAP scores for diagnosing the eight types of lesions > 5 cm using YOLOR -W6/ YOLOR -D6 in the validation and testing sets were 0.358/0.402 and 0.203/0.187, respectively. Furthermore, the mAP scores for diagnosing lesions less than 5 cm using YOLOR -W6/ YOLOR -D6 in the validation and testing sets were 0.291/0.284 and 0.223/0.330 (Table 4).

Table 4 Automatic Lesion Detection and Diagnostic Performance of Differentiation Between Malignant and Benign Lesions and Classifying 8 Kinds of Hepatic Lesions

Differentiation between benign and malignant lesions								
	Validation				Testing			
	≥ 5 cm	< 5 cm	≥ 3 cm	< 3 cm	≥ 5 cm	< 5 cm	≥ 3 cm	< 3 cm
YOLOR-W6	0.561 (0.448–0.695)	0.435 (0.377–0.517)	0.438 (0.375–0.532)	0.460 (0.361–0.562)	0.390 (0.345–0.501)	0.473 (0.434–0.534)	0.454 (0.409–0.536)	0.449 (0.403–0.525)
YOLOR-D6	0.628 (0.517–0.76)	0.491 (0.42–0.585)	0.526 (0.445–0.617)	0.521 (0.431–0.636)	0.397 (0.359–0.438)	0.599 (0.547–0.672)	0.577 (0.499–0.626)	0.540 (0.477–0.635)
Classifying 8 kinds of hepatic lesions								
	≥ 5 cm	< 5 cm	≥ 3 cm	< 3 cm	≥ 5 cm	< 5 cm	≥ 3 cm	< 3 cm
YOLOR-W6	0.358 (0.319–0.475)	0.291 (0.237–0.417)	0.280 (0.228–0.413)	0.252 (0.206–0.353)	0.203 (0.166–0.315)	0.223 (0.179–0.327)	0.268 (0.203–0.386)	0.161 (0.146–0.261)
YOLOR-D6	0.402 (0.336–0.524)	0.284 (0.242–0.429)	0.291 (0.256–0.397)	0.321 (0.251–0.47)	0.187 (0.161–0.284)	0.330 (0.25–0.459)	0.309 (0.248–0.436)	0.269 (0.209–0.434)

Discussion

In this study, DL models were developed using a single-center cohort to detect and diagnose hepatic tumors using abdominal ultrasound. The study revealed that EfficientNetV2-S, EfficientNetV2-L, Swin-T, and Swin-B achieved high diagnostic accuracies in distinguishing between malignant and benign hepatic tumors (AUC of 0.85, 0.89, 0.89, and 0.90, respectively). The performance of these models surpassed that of the conventional B-mode and contrast-enhanced ultrasonography.²⁵ Therefore, the diagnostic accuracy of B-mode ultrasound can be significantly enhanced with the assistance of DL. This improvement is crucial for HCC screening and ultimately improving patient outcomes.

Based on our findings, our model achieved a level of accuracy similar to that of CT or MRI scans.²⁶ It is important to consider that in the case of HCC, confirmation through pathology may not always be necessary for diagnosis. In many cases, dynamic CT or MRI can provide sufficient information for diagnosis.^{2,27} However, it should be noted that contrast injection is required for dynamic scans and may pose a risk for patients with renal insufficiency.²⁸ In contrast, ultrasound imaging with the assistance of DL has an advantage over CT or MRI, as it poses no risk of causing acute kidney injury. Furthermore, the real-time US with DL-based diagnosis also helps to guide the intervention.¹⁰

However, one of the key difficulties in developing this DL model is the unique nature of liver tumors in ultrasound images with fixed anatomic landmarks. Ultrasound interpretation relies heavily on the operator’s experience. Consequently, during the training of the DL model, it is crucial to have a large dataset of US images annotated by specialized physicians, as in our study. This step is of utmost importance for ensuring an accurate interpretation of the images. When differentiating between benign and malignant lesions in the testing set, EfficientNetV2-L achieved the highest accuracy, indicating that a larger number of parameters may contribute to improved performance. When classifying eight types of hepatic lesions, Swin-T achieved the highest overall accuracy, particularly in the diagnosis of HCC and hepatic cysts. These results suggest that a lightweight transformer model may be sufficient for effective multiclass classification in this setting.

Moreover, this study achieved an impressive detection rate, with mAP scores of 0.5134/0.5342 and 0.5410/0.5631 for YOLOR -W6/ YOLOR -D6, respectively, in both validation and testing sets. These results demonstrated the high accuracy of the DL model for detection and diagnosis. These results allow the integration of automatic detection and diagnosis, providing physicians with a quicker and more reliable screening reference. This improved the efficiency and effectiveness of the diagnostic process, particularly in the absence of an abdominal ultrasound specialist. However, the mAP score for classifying eight types of lesions for YOLOR -W6/ YOLOR -D6 in the validation and testing sets was 0.2958/0.2938 and 0.3029/0.3135, respectively, much lower than that of the benign/malignant model. Therefore, the classification of the eight types of lesions remains difficult. Given the growing interest in radiomics for extracting quantitative imaging features, integrating radiomic analysis with deep learning may offer complementary insights, particularly in capturing tumor heterogeneity and subtle imaging patterns not readily recognized by convolutional networks. Future studies could explore this hybrid approach to further enhance diagnostic accuracy and model

interpretability.²⁹ Overall, this study represents a significant step forward in medical imaging and has the potential to lead to further advancements in the detection and diagnosis of hepatic tumors.

Our study population was comprised of diverse groups with varying liver conditions. Almost half of the patients (48.2%) were HBV carriers, whereas a significant proportion (21.2%) had contracted HCV. Most patients were diagnosed with liver cirrhosis (40.0%) or fatty liver disease (8.2%). These findings have important implications for the clinical utility of our model, as they suggest that our approach is well suited for patients with parenchymal liver disease. Furthermore, we performed a subgroup analysis of lesions measuring > three centimeters and lesions measuring > five centimeters. The aim of this study was to assess whether lesion size affected the accuracy of the AI model. As shown in Table 4, accuracy was higher for lesions measuring > five centimeters. This observation suggests that size still significantly affects the overall precision of the AI model.

However, our study had several limitations. First, this was a single-center retrospective study; therefore, possible selection bias existed and external validation was still unavailable. Further prospective multicenter studies should be conducted. However, we included images from eight scanners and three manufacturers. Therefore, the variabilities in image acquisition among the different machines were trained in DL modeling. Second, we used dynamic CT, MRI, or pathologic examination to diagnose hepatic lesions. However, according to the current guidelines, only HCC can be diagnosed using images. Other malignancies such as cholangiocarcinoma should only be diagnosed via pathological examination. However, HCC was still the leading cause of malignancy in our cohort (4117/6001: 68.6%); therefore, the results may still apply to clinical practice. Third, approximately 50%, 20%, 40%, and 10% of patients had HBV, HCV infection, liver cirrhosis, and fatty liver, respectively. The factors affecting differences in underlying liver diseases were not adjusted. However, we used a bounding box to minimize the effect of the liver parenchyma with relatively high accuracy. Further studies with stratification of different etiologies of hepatic tumors should be performed. Finally, the AUCs of several models, such as EfficientNetV2-S, decreased from 0.85 in the validation set to 0.75 in the test set, suggesting potential overfitting. This discrepancy highlights the need for cautious model evaluation, as strong performance in internal validation does not necessarily guarantee generalizability to unseen data.

Moreover, the diagnostic accuracy for rare classes such as cholangiocarcinoma, metastasis, and others was substantially lower (often <0.1) compared to more prevalent classes like HCC and cysts. This limitation is likely due to the severe class imbalance and the relatively small number of training samples in these categories. To address these challenges, future studies may consider applying cross-validation strategies, data augmentation, class-weighted loss functions, or transfer learning to mitigate overfitting and improve performance in underrepresented classes. Our current work only focused on deep learning-based classification using US images.

In conclusion, our study demonstrated the potential of DL models for the real-time detection and diagnosis of hepatic lesions in an Asian population with high accuracy. HCCs and cysts had the highest accuracy in classifying the eight types of hepatic lesions, and lesions more than five centimeters had the highest detection rate. However, further improvements and broader application of DL models in the diagnosis of liver lesions are still needed.

Abbreviations

HCC, hepatocellular carcinoma; CT, computed tomography; MRI, magnetic resonance imaging; CNN, convolutional neural network; DL, deep learning; AUC, area under curve; mAP, mean average precision; HBV, hepatitis B virus; HCV, hepatitis C virus.

Code/Data Availability

Related data are accessible after requesting the corresponding author.

Ethics Approval

No study subjects or cohorts have been reported previously. Written informed consent was waived and approval was obtained from the Institutional Review Board of National Taiwan University Hospital (No. 201801011RINB). This study followed the Health Insurance Portability and Accountability Act guidelines and the 1975 Declaration of Helsinki revised in 2008.

Author Contributions

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Funding

This study was partially supported by a research grant from the Good Liver Foundation of Taipei, Taiwan.

Disclosure

Dr Chih-Horng Wu, Jin-Chuan Sheu, Pei-Lien Chou, and Hsiao-Ching Nien report a patent TWI810498 issued.

The authors have no conflict of interest to declare.

References

1. Akinyemiju T, Abera S, Ahmed M; Global Burden of Disease Liver Cancer Collaboration. The burden of primary liver cancer and underlying etiologies from 1990 to 2015 at the global, regional, and national level: results from the global burden of disease study 2015. *JAMA Oncol.* **2017**;3(12):1683–1691. doi:10.1001/jamaoncol.2017.3055
2. Shao -Y-Y, Wang S-Y, Lin S-M, et al. Management consensus guideline for hepatocellular carcinoma: 2020 update on surveillance, diagnosis, and systemic treatment by the Taiwan Liver Cancer Association and the Gastroenterological Society of Taiwan. *J Formos Med Assoc.* **2021**;120(4):1051–1060. doi:10.1016/j.jfma.2020.10.031
3. Choo SP, Tan WL, Goh BKP, Tai WM, Zhu AX. Comparison of hepatocellular carcinoma in Eastern versus Western populations. *Cancer.* **2016**;122(22):3430–3446. doi:10.1002/cncr.30237
4. Llovet JM, Bruix J. Novel advancements in the management of hepatocellular carcinoma in 2008. *J Hepatol.* **2008**;48 Suppl 1:S20–37. doi:10.1016/j.jhep.2008.01.022
5. Sheu JC, Sung JL, Chen DS, et al. Early detection of hepatocellular carcinoma by real-time ultrasonography. A prospective study. *Cancer.* **1985**;56(3):660–666. doi:10.1002/1097-0142(19850801)56:3<660::aid-cncr2820560338>3.0.co;2-f
6. Andersson KL, Salomon JA, Goldie SJ, Chung RT. Cost Effectiveness of Alternative Surveillance Strategies for Hepatocellular Carcinoma in Patients With Cirrhosis. *Clin Gastroenterol Hepatol.* **2008**;6(12):1418–1424. doi:10.1016/j.cgh.2008.08.005
7. Colli A, Fraquelli M, Casazza G, et al. Accuracy of ultrasonography, spiral CT, magnetic resonance, and alpha-fetoprotein in diagnosing hepatocellular carcinoma: a systematic review. *Am J Gastroenterol.* **2006**;101(3):513–523. doi:10.1111/j.1572-0241.2006.00467.x
8. Hanna RF, Milousev VZ, Tang A, et al. Comparative 13-year meta-analysis of the sensitivity and positive predictive value of ultrasound, CT, and MRI for detecting hepatocellular carcinoma. *Abdom Radiol.* **2016**;41(1):71–90. doi:10.1007/s00261-015-0592-8
9. Burrowes DP, Medellin A, Harris AC, Milot L, Lethebe BC, Wilson SR. Characterization of focal liver masses: a multicenter comparison of contrast-enhanced ultrasound, computed tomography, and magnetic resonance imaging. *J Ultrasound Med.* **2021**;40(12):2581–2593. doi:10.1002/jum.15644
10. Wu C-H, Liang P-C, Su T-H, et al. Iodized oil computed tomography versus ultrasound-guided radiofrequency ablation for early hepatocellular carcinoma. *Hepatol Int.* **2021**;15(5):1247–1257. doi:10.1007/s12072-021-10236-0
11. Galle PR, Forner A, Llovet JM, et al. EASL clinical practice guidelines: management of hepatocellular carcinoma. *J Hepatol.* **2018**;69(1):182–236. doi:10.1016/j.jhep.2018.03.019
12. Pinto A, Pinto F, Faggian A, et al. Sources of error in emergency ultrasonography. *Crit Ultrasound J.* **2013**;5(Suppl 1):S1. doi:10.1186/2036-7902-5-S1-S1
13. Liu S, Wang Y, Yang X, et al. Deep learning in medical ultrasound analysis: a review. *Engineering.* **2019**;5(2):261–275. doi:10.1016/j.eng.2018.11.020
14. Rahman H, Bukht TFN, Imran A, Tariq J, Tu S, Alzahrani A. A deep learning approach for liver and tumor segmentation in CT images using ResUNet. *Bioengineering.* **2022**;9(8):368. doi:10.3390/bioengineering9080368
15. Zhen S, Cheng M, Tao Y, et al. Deep learning for accurate diagnosis of liver tumor based on magnetic resonance imaging and clinical data. *Front Oncol.* **2020**;10:680.
16. Meng D, Zhang L, Cao G, Cao W, Zhang G, Hu B. Liver fibrosis classification based on transfer learning and FCNet for ultrasound images. *IEEE Access.* **2017**;5:5804–5810. doi:10.1109/ACCESS.2017.2689058
17. Byra M, Styczynski G, Szmigielski C, et al. Transfer learning with deep convolutional neural network for liver steatosis assessment in ultrasound images. *Int J Comput Assist Radiol Surg.* **2018**;13(12):1895–1903. doi:10.1007/s11548-018-1843-2
18. Schmauch B, Herent P, Jehanno P, et al. Diagnosis of focal liver lesions from ultrasound using deep learning. *Diagn Interv Imaging.* **2019**;100(4):227–233. doi:10.1016/j.diii.2019.02.009
19. Karako K, Mihara Y, Arita J, et al. Automated liver tumor detection in abdominal ultrasonography with a modified faster region-based convolutional neural networks (Faster R-CNN) architecture. *Hepatobiliary Surg Nutr.* **2022**;11(5):675–683. doi:10.21037/hbsn-21-43
20. 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: Marrero et al. *Hepatology.* **2018**;68(2):723–750. doi:10.1002/hep.29913
21. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE; **2016**. p. 770–778. doi: 10.1109/CVPR.2016.90.

22. Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. arXiv:160207261 [cs]. 2016; Available from: <http://arxiv.org/abs/1602.07261>. Accessed April 14, 2022.
23. Tan M, Le QV. EfficientNet: rethinking Model Scaling for Convolutional Neural Networks. *arXiv*. 2020. doi:10.48550/arXiv.1905.11946
24. Wang C-Y, Yeh I-H, Liao H-YM. You only learn one representation: unified network for multiple tasks. *arXiv*. 2021. doi:10.48550/arXiv.2105.04206
25. Jiang Y, Zhang M, Zhu Y, Zhu D. Diagnostic role of contrast-enhanced ultrasonography versus conventional B-mode ultrasonography in cirrhotic patients with early hepatocellular carcinoma: a retrospective study. *J Gastrointest Oncol*. 2021;12(5):2403–2411. doi:10.21037/jgo-21-611
26. Chen X, Li M, Guo R, et al. The diagnostic performance of contrast-enhanced CT versus extracellular contrast agent-enhanced MRI in detecting hepatocellular carcinoma: direct comparison and a meta-analysis. *Abdom Radiol*. 2022;47(6):2057–2070. doi:10.1007/s00261-022-03484-7
27. Ayuso C, Rimola J, Vilana R, et al. Diagnosis and staging of hepatocellular carcinoma (HCC): current guidelines. *Eur J Radiol*. 2018;101:72–81. doi:10.1016/j.ejrad.2018.01.025
28. Safi W, Rauscher I, Umgelter A. Contrast-induced acute kidney injury in cirrhotic patients. A retrospective analysis. *Ann Hepatol*. 2015;14(6):895–901. doi:10.5604/16652681.1171779
29. Park HJ, Park B, Lee SS. Radiomics and deep learning: hepatic applications. *Korean J Radiol*. 2020;21(4):387. doi:10.3348/kjr.2019.0752

Journal of Hepatocellular Carcinoma

Publish your work in this journal

The Journal of Hepatocellular Carcinoma is an international, peer-reviewed, open access journal that offers a platform for the dissemination and study of clinical, translational and basic research findings in this rapidly developing field. Development in areas including, but not limited to, epidemiology, vaccination, hepatitis therapy, pathology and molecular tumor classification and prognostication are all considered for publication. The manuscript management system is completely online and includes a very quick and fair peer-review system, which is all easy to use. Visit <http://www.dovepress.com/testimonials.php> to read real quotes from published authors.

Submit your manuscript here: <https://www.dovepress.com/journal-of-hepatocellular-carcinoma-journal>

Dovepress
Taylor & Francis Group