Journal of Gastroenterology and Hepatology

JGHF

doi:10.1111/jgh.16803

ORIGINAL ARTICLE - GASTROENTEROLOGY (EXPERIMENTAL)

Automatic localization and deep convolutional generative adversarial network-based classification of focal liver lesions in computed tomography images: A preliminary study

Pushpanjali Gupta,[*,†,‡,§] Yao-Chun Hsu,[¶,**] Li-Lin Liang,[†,§] Yuan-Chia Chu,[††,‡‡,§§] Chia-Sheng Chu,[¶¶,***] Jaw-Liang Wu,[†††] Jian-An Chen,[‡] Wei-Hsiu Tseng,[‡] Ya-Ching Yang,[‡] Teng-Yu Lee,[‡‡‡,§§§] Che-Lun Hung[†,‡] and Chun-Ying Wu[*,†,‡,§,¶¶,¶¶¶]

*Division of Translational Research, ††Information Management Office, ‡‡Big Data Center, Taipei Veterans General Hospital, †Health Innovation Center, ‡Institute of Biomedical Informatics, §Institute of Public Health, ¶¶Ph.D. Program of Interdisciplinary Medicine, †††School of Medicine, National Yang Ming Chiao Tung University, §§Department of Information Management, National Taipei University of Nursing and Health Sciences, ***Division of Gastroenterology and Hepatology, Taipei City Hospital Yang Ming Branch, Taipei, ¶Division of Gastroenterology and Hepatology, E-DA Hospital, **School of Medicine, I-Shou University, Kaohsiung, ‡‡‡Division of Gastroenterology and Hepatology, Taichung Veterans General Hospital, §§§School of Medicine, Chung Shan Medical University¶¶¶Department of Public Health, China Medical University, Taichung, Taiwan

Key words
artificial intelligence, early diagnostic tool, focal liver lesions, generative adversarial network, hepatocellular carcinoma.

Accepted for publication 24 October 2024.

Correspondence
Chun-Ying Wu and Che-Lun Hung, Institute of Biomedical Informatics, National Yang Ming Chiao Tung University, No. 155, Section 2, Linong Street, Taipei 11221, Taiwan.
Email: dr.wu.taiwan@gmail.com; clhung@nycu.edu.tw

Declaration of conflict of interest: The authors declare no competing interests. C.-Y. W. is an Editorial Board member of *JGH* and a co-author of this article. To minimize bias, they were excluded from all editorial decision-making related to the acceptance of this article for publication.
Author contribution: C.-Y. W. and C.-L. H. conceived and directed the study. C.-Y. W. has supervised the work regarding clinical feasibility, and C.-L. H. has supervised the work regarding technical suitability; therefore, they have equal contributions as corresponding authors. P. G. developed deep learning models, annotated and analyzed the data, and drafted the manuscript. Y.-C. H. performed data acquisition and provided clinical guidance. L.-L. L. provided strategic guidance. Y.-C. C. helped create an environment for AI analysis and troubleshooting. J.-L. W., J.-A. C., W.-H. T., and Y.-C. Y. annotated the data. Y.-C. H., C.-S. C., and T.-Y. L. checked the annotation and provided clinical guidance. All authors performed the critical revision of the manuscript and approved the final draft of the article.

Abstract

Background and Aim: Computed tomography of the abdomen exhibits subtle and complex features of liver lesions, subjectively interpreted by physicians. We developed a deep learning-based localization and classification (DLLC) system for focal liver lesions (FLLs) in computed tomography imaging that could assist physicians in more robust clinical decision-making.
Methods: We conducted a retrospective study (approval no. EMRP-109-058) on 1589 patients with 17 335 slices with 3195 FLLs using data from January 2004 to December 2020. The training set included 1272 patients (male: 776, mean age 62 ± 10.9), and the test set included 317 patients (male: 228, mean age 57 ± 11.8). The slices were annotated by annotators with different experience levels, and the DLLC system was developed using generative adversarial networks for data augmentation. A comparative analysis was performed for the DLLC system *versus* physicians using external data.
Results: Our DLLC system demonstrated mean average precision at 0.81 for localization. The system's overall accuracy for multiclass classifications was 0.97 (95% confidence interval [CI]: 0.95–0.99). Considering FLLs ≤ 3 cm, the system achieved an accuracy of 0.83 (95% CI: 0.68–0.98), and for size > 3 cm, the accuracy was 0.87 (95% CI: 0.77–0.97) for localization. Furthermore, during classification, the accuracy was 0.95 (95% CI: 0.92–0.98) for FLLs ≤ 3 cm and 0.97 (95% CI: 0.94–1.00) for FLLs > 3 cm.
Conclusion: This system can provide an accurate and non-invasive method for diagnosing liver conditions, making it a valuable tool for hepatologists and radiologists.

Ethical approval: This study gained the consent of the Institutional Review Board of E-DA Hospital, Kaohsiung (IRB no. EMRP-109-058), and was conducted ethically in accordance with the World Medical Association Declaration of Helsinki. As this study was conducted retrospectively, the requirement for written informed consent was waived by the Institutional Review Board, as per their approval.
Financial support: This study is partly financially supported by the National Science and Technology Council (NSTC) of Taiwan under grant 112-2634-F-A49-003.

Introduction

According to recent statistics, liver cancer is the sixth most commonly diagnosed cancer and the third leading cause of cancer-related deaths globally.[1] Distressingly, projections indicate that the number of liver cancer cases may increase by 55% between 2020 and 2040, with an estimated 1.4 million diagnoses and 1.3 million deaths by 2040.[2] Contrast-enhanced computed tomography (CT) is a useful dynamic imaging modality that can be used to identify focal liver lesions (FLLs) in a cost-saving and time-saving manner compared with ultrasonography and magnetic resonance imaging (MRI).[3] Nonetheless, the interpretation of CT scans is subjective because of the involvement of human perception; less experienced residents or non-academic radiologists could wrongly evaluate the images. In addition, the liver spans from 20 to 45 slices per phase in a triple-phase CT. Radiologists might have to look at approximately 120 slices per patient to check for the lesions because different lesions show better visibility in various phases of CT. The error is inevitable when handling multiple cases daily and spending only 3–4 s per slice.[4] Furthermore, different lesions exhibit different characteristics in various phases of CT. The most prevalent form of liver cancer is hepatocellular carcinoma (HCC), which has a 5-year survival rate of only 18%,[5] and hemangioma (HEM) accounts for approximately 20% of the diagnosed benign cases. Nonetheless, it was found that smaller HCCs are sometimes mistaken as HEM,[6] which causes a burden in the healthcare system where treatment of malignant cases is missed because of poor diagnosis.

Considering the remarkable achievements of deep learning (DL), we believe that computer-aided diagnosis may assist physicians in precisely identifying FLLs.[7–9] In recent years, generative adversarial networks (GANs)[10] have been widely deployed to generate synthetic images that could be used to train the models and avoid the manual method of tedious labeling. DL can identify and differentiate liver masses in contrast-enhanced CT.[11] For instance, the work in Shi *et al.*[12] used DL to differentiate HCC from other FLLs. The authors investigated the significance of different phases of CT, where a comparative study was performed considering three phases of CT—arterial, portal venous, and delayed—and all phases of CT where non-contrast phase images were also included. The model with three phases achieved an accuracy of 85.6%, and four phases of CT attained an accuracy of 83.3%. On the other hand, another study used an automatic approach for liver segmentation followed by detecting malignant lesions.[13] Although the work proposed the differentiation of malignant lesions from benign lesions, their data consisted of 93.8% of HCC; this might create bias in the model derivation.

Furthermore, when examining the use of MRI in FLL diagnosis, previous research[14] has suggested a binary classification model to differentiate HCC from other FLLs. This model achieved an overall accuracy of 87.3%. On the other hand, extending binary classification, the multiclass classification model was derived in Wang *et al.*[15] considering seven types of FLLs; the model has areas under the curve (AUCs) of 0.969 and 0.974 for binary and multiclass models, respectively. It is essential to remember that some tumors may be difficult to spot at first glance. This could lead to variability in the input data. To fully automate the diagnosis process, automatic detection and classification of FLLs were conducted in Zhou *et al.*,[16] but only 616 nodules were considered in the study. While the detection model achieved an F1-score of 87.8%, the binary and multiclass classification models showed lower accuracy results of 82.5% and 73.4%, respectively. The model's performance could have been improved when more samples were used as training data. Considering the increasing incidence of cancers, it is essential to have early diagnosis and correct identification of the different types of lesions for planning effective treatment strategies. Therefore, this study aimed to develop a DL-based system for the automatic localization and classification of FLLs in CT imaging that could assist physicians in faster and more robust clinical decision-making. The overall flow of our study is shown in Figure 1.

Methods

This retrospective study was performed according to the Declaration of Helsinki and approved under institutional review board approval no. EMRP-109-058. Written informed consent from patients was waived. The authors had control of the data and information submitted for publication.

Patient eligibility. Our study included data collected from January 2004 to December 2020, screening 69 004 patients. For

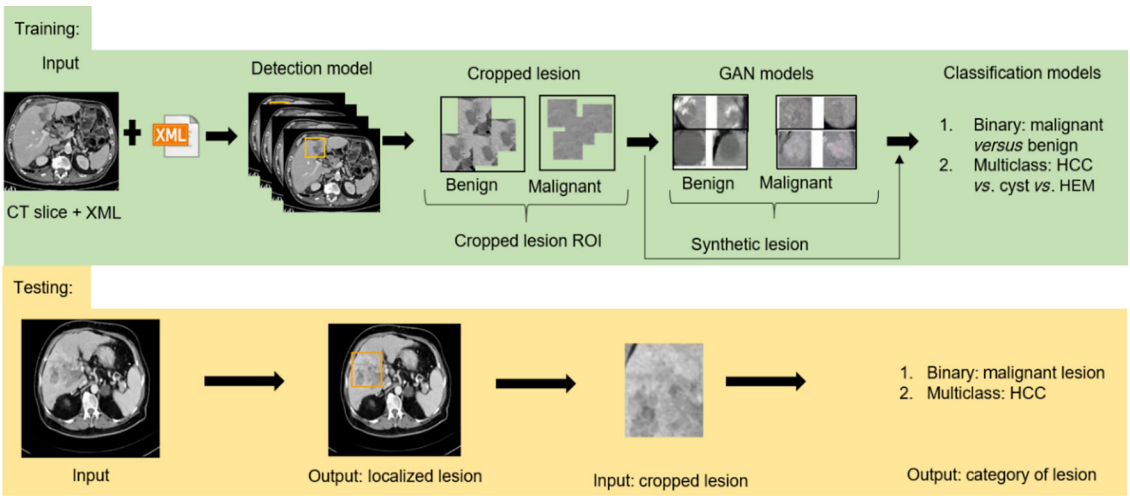

Figure 1 Outline showing end-to-end workflow for localizing and classifying focal liver lesions in liver computed tomography (CT) image. GAN, generative adversarial network; HCC, hepatocellular carcinoma; HEM, hemangioma; ROI, region of interest.

contrast-enhanced CT examinations, we utilized advanced imaging technology such as Sensation 16 (Siemens, Germany), LightSpeed VCT (GE Medical Systems, United States), BrightSpeed (GE Medical Systems), SOMATOM Definition AS (Siemens), Emotion 16 (Siemens), and Optima CT660 (GE Medical Systems). To ensure the accuracy of our findings, we excluded patients who did not have a suitable CT before treatment, had inappropriate artifacts on preoperative CT, showed massive intra-abdominal bleeding due to rupture, had an unclear tumor on preoperative CT, did not undergo contrast CT examination, or were under 20 years of age. For HCC cases, we preferred patients who had quadruple phases (precontrast, arterial, venous, and delay phases). However, obtaining a quadruple phase for benign lesions was challenging. We considered the precontrast, contrast (slices were included in the venous phase during analysis), and delay phases for benign cases. Ultimately, our study included 1589 patients, with 736 patients having naïve HCC primarily, as shown in Figure 2. Of the 853 patients with benign lesions, 593 had cysts, and 454 had HEM. In addition, demographic data such as age and gender were also collected for all 1589 patients in our study. After eligibility assessment, a total of 1589 patients (mean age, 60 years ± 11.26 [SD]; 1004 male [63%]) with the age of patients ranging from 20 to 95 years were included in the study (Table 1).

Data preparation and annotations. In this retrospective study, all preoperative CT images were obtained from the picture archiving and communication system. The patients' data collected in digital imaging and communications in medicine (DICOM) format were converted to JPEG format, widely accepted for medical image analysis, using MicroDicom, a free DICOM viewer (https://www.microdicom.com/). All the slices were viewed in the abdomen window, and the slices in axial planes were exported to a fixed dimension of 512 × 512 with optimal lossless quality maintained and patients' information anonymized. Moreover, the converted slices were verified by experienced radiologists and hepatologists to ensure no eye-catching distortion of the liver tissue. A patient with an abdomen CT contains approximately 25–40 slices per phase; however, the FLL may not appear on all slices. As a result, the experts separated the slices containing FLL manually. Consequently, 17 335 slices were extracted and used for the annotation process. To use a supervised learning-based detection approach for the localization of FLLs in the CT slices, the labeled samples were created using the extracted slices and adopting the LabelImg tool (https://pypi.org/project/labelImg/1.4.0/), where boxes were drawn around the FLL region and the coordinates of the bounding box for FLL in the form of region of interest (ROI) were stored in Extensible Markup Language (XML) format and text format. The annotation was initially made in slices obtained from phase A of CT because HCC is hyper-enhanced and conveniently visible in phase A. This was followed by annotating the slices obtained from phase V of CT. The radiology report was used to categorize all FLLs; annotator 1 (a physician with 4 years of experience) initially made the bounding box. The bounding box made by annotator 1 was verified by annotator 2 (a hepatologist with more than 8 years of experience). Another independent annotation was made by annotator 3 (a hepatologist with more than 15 years of experience). Annotator 2 and annotator 3 were unaware of each other's annotations. For qualitative ground truth generation, the inter-reader agreement was obtained between annotator 2 and annotator 3. There was variability of FLL sizes in patients, as shown in Figure S1, where the first row shows the images and the second row shows bounding boxes for the labeled ROI. In total, 17 335 slices were labeled out of 45 049 from 1589 patients, resulting in 1535 HCC, 866 cysts, and 794 HEMs, altogether 3195 lesions.

Data augmentation. The main issue of supervised learning-based training of models is the requirement of a large, labeled training dataset. To increase the training samples and improve the model performance in classifying the FLL into different categories, we used two approaches for augmenting the dataset: (i) conventional augmentation that uses image transformation techniques and (ii) generation of synthetic new samples after

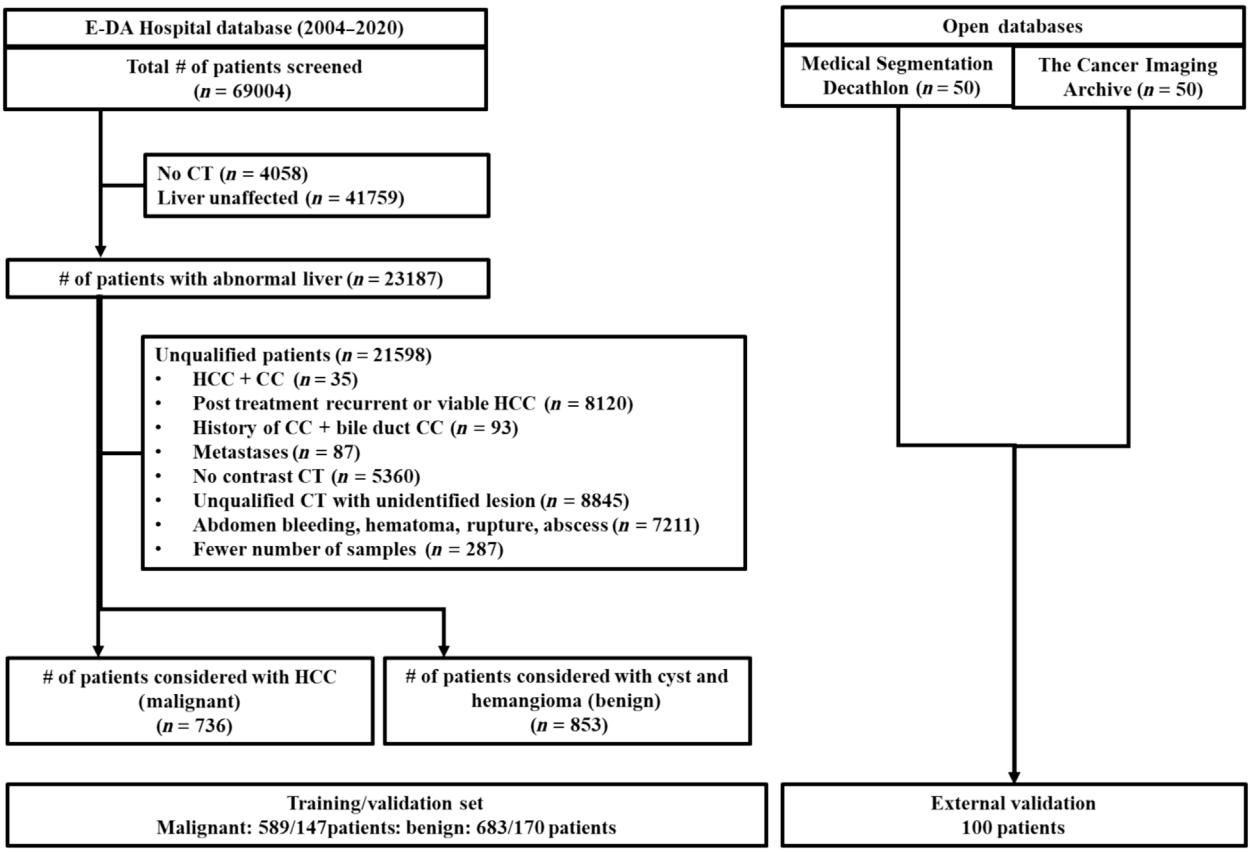

Figure 2 The flowchart of patients is considered in this study. Focal liver lesions with fewer available samples were excluded. Patients may be excluded because of more than one criterion. Therefore, the total number of excluded patients in each step may outnumber the sum of patients excluded by individual criteria. CC, cholangiocarcinoma; CT, computed tomography; HCC, hepatocellular carcinoma.

learning from given real images using generative models. Both approaches are described in the following.

Conventional augmentation. There is a huge possibility of overfitting when training deep convolutional neural networks (NNs) with limited data. The most popular and basic solution for enlarging the dataset is to use conventional augmentation techniques such as rotation, scaling, translation, flipping, and shearing. In this study, after the detection model detected the FLLs, the FLLs were cropped. Each FLL ROI was rotated at different angles = {0°, 30°, 45°, 60°, 90°, 120°, 180°}. This was followed by horizontal and vertical flipping of each rotated sample. We used translational and scaling samples with a factor of 0.2 and 0.04, respectively. In this work, shearing was not used to prevent the shape deformation of FLLs. Finally, all samples were resized to 128 × 128 pixels.

Table 1 Demographic data of patients considered in this study

Characteristic	Training set (*n* = 1272)	Testing set (*n* = 317)	*P*-value
Age (years), mean ± SD	61.75 ± 10.97	56.94 ± 11.84	< 0.001
Gender			< 0.001
Female	496 (39)	89 (28)	
Male	776 (61)	228 (72)	
No. of patients with malignancy diagnosis	589 (46)	147 (46)	< 0.001

Unless otherwise stated, data are numbers with percentages in parentheses.

Generative adversarial network-based augmentation. The GAN is a specific framework where a DL model learns to capture the training data's distribution to generate new data from the same distribution. GANs typically learn in an unsupervised manner using a zero-sum game theory approach, where one NN's gain is equal to another NN's loss. We used the deep convolutional GAN (DCGAN) structure mentioned in Radford *et al*.,[17] where the network used transpose convolution and convolutional layers in the generator and discriminator networks, respectively. The overall architectural outline of the DCGAN (Default Architecture of DCGAN section of the supporting information) is depicted in Figure S2a. The generator (Fig. S2b) maps the input from the latent space to the vector space to generate an image of the same dimensions as the real image (128 × 128 × 3). Conversely, the discriminator (Fig. S2c) processes the input image of dimension 128 × 128 × 3, extracting its features and ascertaining whether the image is real or fake. The input image size was set to 128

pixels, with noise size of 100, learning rate for both discriminator and generator as 0.00002, batch size of 64, number of epochs of 500, and epsilon set to 0.00005. Both the generator and discriminator use binary cross-entropy loss. We used three independent DCGANs to generate synthetic images for each type of lesion (HCC, HEM, and cyst) separately. Therefore, each DCGAN generated synthetic lesion for one class only. In addition, the DCGANs were primarily used to generate synthetic images that could improve the classification model's performance. Consequently, we did not use the DCGAN model during the inference time.

Deep learning models

Detection network. We experimented with different object detection models such as single shot detector with backbone MobileNetv2,[18] EfficientDet,[19] RetinaNet,[20] and You Only Look Once (YOLO).[21,22] The application of RetinaNet is popular in the medical image domain, especially when considering detection using CT images.[23] Another influential object detection model, YOLOv8,[24] with different variants useful to detect objects of different scales, was considered suitable because the FLLs were of variable sizes. We adopted the structure of the YOLOv8 model (Default Architecture of YOLOv8 section of the supporting information), briefly represented in Figure S3a, provided by Ultralytics.[25]

Classification network. During classification model selection, both real and synthetic images were used during the training phase, and only real images were used during the testing phase. To select the best model for the classification of FLLs, we compared the performance of several popular models such as VGG16, ResNet50, DenseNet121, Inceptionv3, Inception-ResNetv2 (IRv2),[26] and EfficientNet.[27] After comparison, the best-performing model was EfficientNetB5 (Default Architecture of EfficientNetB5 section of the supporting information), represented in Figure S3b. The two steps of classification were performed to differentiate among the FLLs. Initially, we performed binary classification to distinguish benign *versus* malignant lesions and then differentiated lesions into respective kinds. The best-performing model's hyperparameters were a learning rate of 0.0008, Adam optimizer, batch size of 128, 80:20 ratio for training : testing, and number of epochs of 30.

Statistical analysis. The Cohen κ test[28] was used to evaluate the level of agreement between annotator 2 and annotator 3 in their annotation of images. The statistical significance of normally distributed variables was tested using a *t*-test, and differences in count variables were tested using the chi-squared test. The *P*-value less than 0.05 was considered statistically significant. Identifying FLL as malignant was considered positive, and benign was considered negative in both detection and classification. Similarly, for multiclass, the one-*versus*-all scheme was applied while viewing each class as positive at each instance, and finally, the average value was obtained. Consequently, true positives (TPs), true negatives (TNs), false positives (FPs), and false negatives (FNs) were obtained to calculate accuracy = (TP + TN)(TP + FP + TN + FN), sensitivity = TP(TP + FN), specificity = TN(TN + FP), precision = TP(TP + FP), and F1-score = (2 × precision × sensitivity)(precision + sensitivity) to evaluate the performance of different models. In addition, the mean Average Precision (mAP) of the model showing the tradeoff between precision and sensitivity was also determined during detection. In addition, receiver operating characteristic (ROC) curves were generated, and the AUC was also determined to assess the performance of models. All experiments were conducted on a Tesla V100 PCIe GPU with 32 GB of system RAM. Torch 2.0.0 +cu117 and TensorFlow 2.12[29] were implemented in the Python 3.10 environment. Scikit-learn was primarily used to perform statistical analyses.

Results

Patient demographics and imaging data. Out of 1589 patients, 1272 (mean age, 61.75 ± 10.97 [SD]; male, 776 [61%]) were considered for training, and 317 (mean age, 56.94 ± 11.84 [SD]; male, 228 [72%]) were considered for testing. The FLL ranged from 0.02 to 25 cm (median: 2 cm) (Fig. S4). During patient separation, care was taken to have lesion separation without any bias in size. Therefore, out of 3195 FLLs, we used 1545 (49%) (FLL $\leq$ 3 cm) and 928 (29%) (FLL $>$ 3 cm) for training and 395 (12%) (FLL $\leq$ 3 cm) and 321 (10%) (FLL $>$ 3 cm) for testing.

Inter-reader agreement for qualitative computed tomography images. Table S1 shows the inter-reader agreement between the readers and κ statistics for each imaging. The agreement was considered slight (κ = 0.01–0.20), fair (κ = 0.21–0.40), moderate (κ = 0.41–0.60), substantial (κ = 0.61–0.80), almost perfect (κ = 0.81–0.99), and perfect (κ = 1). Annotator 2 and annotator 3 agreed with κ between 0.81 and 0.99 for all three FLLs, showing almost perfect inter-reader agreement. It was observed that for even HCC $\leq$ 3 cm, the inter-agreement was almost perfect κ = 0.95 (0.90–1), and for cyst $\leq$ 3 cm, the κ value was 0.91 (0.88–0.93). Further, it was observed that the annotation of HEM had better agreement in comparison with HCC and cyst with κ = 0.97 (0.95–1) for size $\leq$ 3 cm and κ = 1 (0.99–1) for size $>$ 3 cm.

Selection of detection model. In the optimization phase of the detection model, various models were evaluated based on their performance with the hyperparameter configurations specified in Table S2. It was observed that YOLOv8 had the best performance with mAP 0.81 and an FN rate of 0.27 (instance-wise) for the validation set. The different performance metrics obtained for each model were recorded (Table S3). Nonetheless, during patient-wise analysis, if a lesion could be identified in at least 50% (threshold) of the total number of labeled slices, the model accuracy was 100%. Such a threshold was efficient when considering lesions of variable sizes ranging from 0.02 to 25 cm. For instance, it was found that a small FLL was primarily visible in two slices of patients; failing to localize FLL in one slice made the accuracy 50%; on the other hand, localizing 10 out of 16 slices for an FLL of size 8 cm resulted in 100% accuracy. With the significant

potential of recognizing smaller ROIs, YOLOv8 achieved an average accuracy of 0.92 (95% confidence interval [CI]: 0.86–0.96) during patient-wise testing. Consequently, few samples for the slice-wise performances of YOLOv8 are shown in Figure S5. The ground truths provided by experts are shown in Figure S5a, and the corresponding outputs provided by the artificial intelligence model—YOLOv8—are shown in Figure S5b. It can be observed that the model performed well in localizing the FLLs, even of smaller size (≤ 3 cm). In addition, patient-wise performance of YOLOv8 is demonstrated (Fig. S6). Figure S6a,b contains slices with the FLL localized in slices for phases A and V, respectively. In addition, Figure S6c,d includes FLLs localized in slices for the non-contrast phase and delayed phase, respectively. It can be observed that although our model was not trained with images of non-contrast and delay phases, the model could localize the FLLs with 0.70 (95% CI: 0.619–0.781) and 0.84 (95% CI: 0.789–0.901) accuracies, which suggest better usability of the model for all phases of CT.

Generation of generative adversarial network-based augmented images. The ROIs for HCC, HEM, and cyst were input to the DCGAN model separately, and the model generated the augmented samples to obtain many training examples. Approximately 12 000 samples for phase A and 23 000 for phase V for each type of lesion were used. In Figure S7a–c, the top row represents the input samples for the model showing HCC, HEM, and cyst, respectively. Correspondingly, the bottom row of Figure S7a–c shows the output produced by the model for the three types of FLLs. Both the real images and synthetic samples were used to train the classification model in the third phase.

Selection of classification model. The real image and synthetic images were used as input to select the best-performing classification models. During model selection, a fivefold cross-validation study was performed, and average values of training, validation, and testing accuracies were recorded (Table S4). Based on the observation, hyper-tuned EfficientNetB5 was considered our classification model.

Diagnostic accuracy of classification model. After the model selection, the testing performances of different models with EfficientNetB5 as the base, considering phase A and phase V of CT individually and together, were recorded in Table 2. It can be observed that the model derived with only phase A images performed better with the addition of GAN-based images, where the AUC increased from 0.97 (95% CI: 0.94–1.00) to 0.98 (95% CI: 0.98–1.00) (Fig. S8a,c) when classifying the lesion into benign or malignant categories. Moreover, during multiclass differentiation, the model performance increased by 5% (Table 2). This can be again justified from Figure S8b,d, where the area under the ROC curve has increased from 0.95 (95% CI: 0.90–1.00) to 0.98 (95% CI: 0.98–1.00) (Table 2).

The data collected in this study consisted of approximately 3195 lesions (61% FLL ≤ 3 cm), where 1452/1545 (accuracy: 0.94 [95% CI: 0.85–1.00]) lesions were correctly localized during training and 327/395 (accuracy: 0.83 [95% CI: 0.68–0.98]) could be correctly localized during testing. Furthermore, when classifying the FLLs and GAN-based images for FLL ≤ 3 cm, the accuracy for binary and multiclass classification was 0.95 (95% CI: 0.92–0.98) and 0.94 (95% CI: 0.89–0.99) during testing, respectively. On the other hand, for FLLs > 3 cm, the models achieved accuracy of 0.97 (95% CI: 0.94–1.00) and 0.95 (95% CI: 0.90–1.00) for binary and multiclass classification, respectively. Furthermore, the performance of the model was observed to vary with the size of FLLs; during training, for size ≤ 3 cm, the accuracy for HCC was 0.97 (95% CI: 0.93–1.00), HEM was 0.96 (95% CI: 0.94–0.98), and the cyst was 0.97 (95% CI: 0.94–1.00). On the other hand, for size > 3 cm, the accuracy was 0.98 (95% CI: 0.96–1.00) for both HCC and HEM and 0.97 (95% CI: 0.96–0.98) for the cyst. Additionally, during testing, for size ≤ 3 cm, the accuracy for HCC was 0.94 (95% CI: 0.88–1.00), HEM was 0.94 (95% CI: 0.91–0.97), and the cyst was 0.94 (95% CI: 0.89–0.99). On the other hand, for size > 3 cm, the accuracy was 0.95 (95% CI: 0.91–0.99) for

Table 2 Testing performance metrics for the classification models' binary and multiclass classification

Phase	Class	Use of GAN	PPV	Sensitivity	Specificity	F1-score	Accuracy	AUC
A	Binary	No	0.88 (0.78–0.98)	0.88 (0.81–0.95)	0.88 (0.76–1.00)	0.88 (0.80–0.96)	0.88 (0.81–0.95)	0.91 (0.85–0.97)
		Yes	0.97 (0.95–0.99)	0.97 (0.96–0.98)	0.97 (0.95–0.99)	0.97 (0.95–0.99)	0.97 (0.96–0.98)	0.97 (0.96–0.98)
	Multiclass	No	0.87 (0.81–0.93)	0.89 (0.81–0.97)	0.95 (0.91–0.99)	0.88 (0.81–0.95)	0.92 (0.85–0.99)	0.94 (0.89–0.99)
		Yes	0.95 (0.91–0.99)	0.95 (0.92–0.98)	0.98 (0.97–0.99)	0.95 (0.91–0.99)	0.97 (0.95–0.99)	0.97 (0.95–0.99)
V	Binary	No	0.99 (0.98–1.00)	0.99 (0.98–1.00)	0.99 (0.98–1.00)	0.99 (0.98–1.00)	0.99 (0.98–1.00)	0.99 (0.98–1.00)
		Yes	0.99 (0.98–1.00)	0.99 (0.98–1.00)	0.99 (0.98–1.00)	0.99 (0.98–1.00)	0.99 (0.98–1.00)	1.00 (0.99–1.00)
	Multiclass	No	0.94 (0.89–0.99)	0.94 (0.88–1.00)	0.97 (0.94–1.00)	0.94 (0.89–0.99)	0.96 (0.93–0.99)	0.98 (0.96–1.00)
		Yes	0.97 (0.94–1.00)	0.97 (0.95–0.99)	0.99 (0.98–1.00)	0.97 (0.94–1.00)	0.98 (0.98–1.00)	0.99 (0.98–1.00)
AV	Binary	No	0.96 (0.93–0.99)	0.96 (0.93–0.99)	0.96 (0.92–1.00)	0.96 (0.93–0.99)	0.96 (0.92–1.00)	0.97 (0.94–1.00)
		Yes	0.99 (0.98–1.00)	0.98 (0.98–1.00)	0.98 (0.98–1.00)	0.98 (0.98–1.00)	0.98 (0.98–1.00)	0.98 (0.98–1.00)
	Multiclass	No	0.92 (0.85–0.99)	0.92 (0.86–1.00)	0.96 (0.93–0.99)	0.92 (0.85–0.99)	0.92 (0.86–1.00)	0.95 (0.90–1.00)
		Yes	0.96 (0.92–1.00)	0.96 (0.93–0.99)	0.98 (0.98–1.00)	0.96 (0.93–0.99)	0.97 (0.95–0.99)	0.98 (0.98–1.00)

Data show values with 95% CIs in parentheses. The performances for both cases, use of GAN and no use of GAN (using only phase A slices [A], using only phase V slices [V], and using both phase A and phase V slices together [AV]).
AUC, area under the curve; CIs, confidence intervals; GAN, generative adversarial network; PPV, positive predictive value.

HCC, 0.95 (95% CI: 0.90–1.00) for HEM, and 0.96 (95% CI: 0.93–0.99) for the cyst. The summaries of model performances during training and testing with demographics (gender and age) and size of patients are given in Table S5.

Although the interpretation made by DL models cannot be explained entirely because of the black-box nature of the DL models,[30] we have attempted to produce a visual explanation in Figure 3, where the heatmaps were generated using Grad-CAM.[31] The whole slices in Figure 3a show the features focused on by YOLOv8. It can be observed that the model captured the region with FLLs with more attention and, therefore, localized the FLLs correctly. Similarly, when performing FLL classification, EfficientNetB5 could focus on the fast washout pattern of HCC (Fig. 3b, row 1), peripheral enhancement in the case of HEM (Fig. 3b, row 2), and hypoattenuation pattern of the cyst (Fig. 3b, row 3) to differentiate the lesions into respective categories.

Comparison of physicians versus artificial intelligence model considering open data source. To verify the robustness of the derived DL-based localization and classification (DLLC) system, we performed a comparative analysis with two physicians: physician 1, who has 4 years of experience, and physician 2, who has more than 8 years of experience in liver CT imaging. The physicians were unaware of FLL location and category. Each physician localized and labeled the FLLs in the LabelImg tool, with the liberty to view the slices multiple times. We randomly used 50 patients each from the Task03_Liver dataset, Medical Segmentation Decathlon (MSD),[34] and the hcc_tace_seg dataset, The Cancer Imaging Archive (TCIA).[32] During localization, the accuracy achieved was 0.858 (95% CI: 0.8–0.918) for MSD and 0.95 (95% CI: 0.89–1.00) for TCIA (Table 3). Figure 4 shows the performance of the DLLC system in comparison with that of both physicians. The performance of our derived system was better than that of physician 1 and similar to that of physician 2.

Comparison of two-stage approach versus one-stage end-to-end approach. When considering FLL diagnosis, especially when focusing on HCC, HEM, and cyst, these lesions exhibit some similarities, which can be observed from the output obtained from the *t*-distributed stochastic neighbor embedding (t-SNE) algorithm (Fig. S9a). Because of some similar features, we also noticed that although some FLLs were correctly localized, the model could not determine the type of lesion as benign or malignant (Fig. S9b), where the blue box represents the ground truth. In contrast, the red and yellow boxes represent benign and malignant lesions. It was observed that the model considers the lesions as both benign and malignant with similar confidence. Consequently, we localized lesions as a single class (FLL) and generated synthetic images for each FLL separately using GAN. Finally, we used classification models to differentiate the lesions. Incorporating GAN-based images remarkably differentiated the FLLs, shown using t-SNE (Fig. S9c), ultimately improving our two-stage approach performance by 5%.

Discussion

Demographic data and imaging features. During model derivation, the data were separated in such a way that the age of patients ranged from 20 to 95 years (mean ± SD: 61.75 ± 10.97) during training and from 22 to 93 years (mean ± SD: 56.94 ± 11.84) during testing, with a P-value < 0.001. A similar observation was made with the ratio of male to female (61:39) during training and (72:28), with a P-value < 0.001. The proper data distribution shows no gender or age bias in the data included for model derivation.

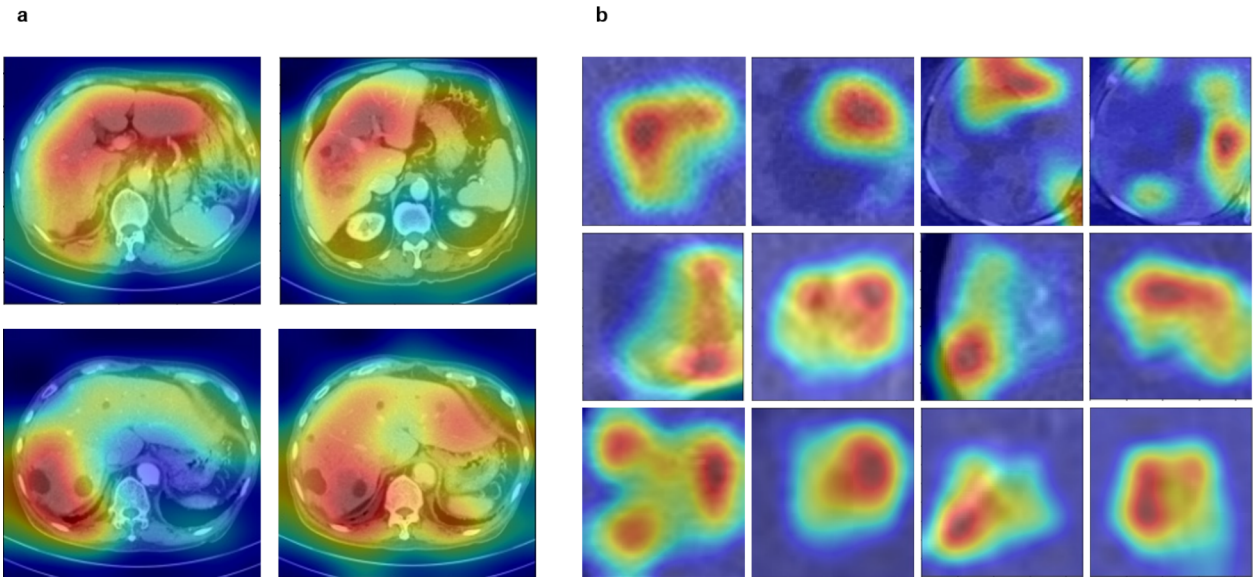

Figure 3 A heat map using Grad-CAM shows the visualization of features focused on by the models. (a) Detection in whole slices. (b) Classification of focal liver lesion using cropped region.

Table 3 Diagnostic comparisons of developed DLLC system with two physicians during external validation using Task03_Liver dataset from (MSD) and hcc-tace-seg dataset from TCIA

Dataset, task	Reader	PPV	Sensitivity	Specificity	F1-score	Accuracy
Task03_Liver (MSD), localization	Physician 1	0.576 (0.518–0.635)	0.516 (0.458–0.576)	0.804 (0.746–0.864)	0.541 (0.482–0.6)	0.736 (0.678–0.796)
	Physician 2	0.66 (0.602–0.72)	0.781 (0.722–0.84)	0.908 (0.850–0.968)	0.711 (0.653–0.77)	0.892 (0.833–0.951)
	DLLC system	0.604 (0.545–0.663)	0.748 (0.69–0.80)	0.875 (0.816–0.934)	0.665 (0.607–0.724)	0.858 (0.8–0.918)
hcc-tace-seg (TCIA), localization	Physician 1	0.502 (0.451–0.554)	0.616 (0.565–0.668)	0.911 (0.86–0.962)	0.5498 (0.498–0.601)	0.879 (0.828–0.931)
	Physician 2	0.814 (0.763–0.866)	0.793 (0.742–0.845)	0.98 (0.96–1)	0.792 (0.746–0.849)	0.96 (0.92–1)
	DLLC system	0.801 (0.75–0.853)	0.696 (0.645–0.748)	0.98 (0.96–1)	0.735 (0.684–0.787)	0.95 (0.89–1)
hcc-tace-seg (TCIA), classification	Physician 1	0.595 (0.544–0.647)	0.516 (0.465–0.568)	0.948 (0.897–1)	0.548 (0.497–0.599)	0.897 (0.846–0.949)
	Physician 2	0.883 (0.832–0.934)	0.921 (0.870–0.973)	0.99 (0.99–1)	0.90 (0.849–0.951)	0.983 (0.966–1)
	DLLC system	0.856 (0.805–0.907)	0.922 (0.971–0.973)	0.98 (0.96–1)	0.875 (0.824–0.926)	0.982 (0.964–1)

Physician 1 had 4 years of experience, and physician 2 had more than 8 years of experience in liver CT imaging.
CT, computed tomography; DLLC, deep learning-based localization and classification; MSD, Medical Segmentation Decathlon; PPV, positive predictive value; TCIA, The Cancer Imaging Archive.

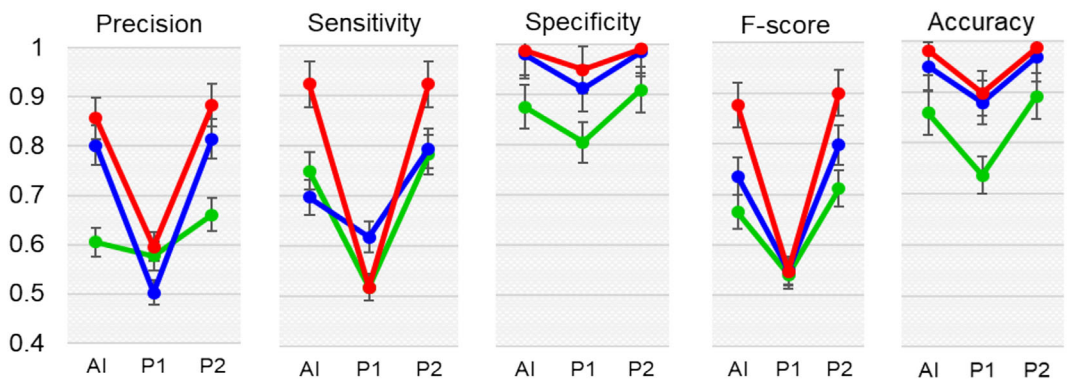

Figure 4 Diagnostic comparisons of our proposed deep learning-based localization and classification system with two physicians during external validation. —●—, Medical Segmentation Decathlon (MSD) data_Detection; —●—, The Cancer Imaging Archive (TCIA) data_Detection; —●—, TCIA data_Classification.

Size-wise analysis for lesion detection and classification. In addition, during data annotation, it was observed that 16% of HCC patients (117/736) had infiltrative lesions, where the bounding box was generated with utmost mindfulness. Consequently, our model had a lower mAP @ IOU 0.5 when localizing the FLLs. Our objective was to localize the FLLs with IOU @ 0.5; therefore, we achieved better accuracy of 0.83 (95% CI: 0.68–0.98) and 0.87 (95% CI: 0.77–0.97) for FLL $\leq$ 3 and $>$ 3 cm, respectively. It is to be noted that, although our work aims to develop a fully automatic system for the localization and classification of FLLs, because of some FNs in the current derived localization model, we used cropped FLLs for those cases missed during localization and performed classification. Another observation made during annotation was that approximately 50% (1565 out of 3195) of FLLs had a size less than 2 cm. In most cases, diagnosis of smaller HCC as HEM could be possible. In addition, because the HCC data considered were of different types, such as confluent with nodular growth of 27.8% (205/736), it was challenging to annotate the FLL, which also led to misclassification of HCC as cyst during model testing.

This study focused on developing a DL-based diagnostic system for the automatic localization and categorization of FLLs. Existing works in this field include a study[12] that uses manual cropping of lesions to differentiate HCC from non-HCC lesions. With 342 patients (449 lesions), the model achieved 85.6% accuracy with no external validation performed. Another similar work[13] focused on detecting primary hepatic malignancies in patients with high risk for HCC, with 1350 multiphase CT scans used for model development. The model achieved 84.8% sensitivity with no external validation performed. Unlike previous studies, multiclass classification of FLLs was performed in Wang *et al.*,[15] considering 445 patients (557 lesions) with lesions cropped from multisequence MRI. The model achieved 79.6% accuracy during seven-class classification. When considering the earlier-derived models for clinical implementation, it is necessary to identify and crop the lesions from CT/MRI slices. The human intervention might affect the efficiency of DL models because the input to the model depends on the physicians' experience. The lesions must be manually cropped before evaluation. On the other hand, a fully automatic method for detecting and classifying lesions is proposed in Zhou *et al.*,[16] using multiphase CT of 435 patients (616 nodules) to avoid human intervention. Although the model achieved 82.8% precision during detection and an average of 82.5% and 73.4% accuracy during binary and six-class classification, the

model might suffer from overfitting due to limited data and a single-center study with no external validation. In summary, although few works have proposed detecting and classifying lesions into different categories in either CT or MRI, our study stands out from similar research in multiple ways. Firstly, it attempts to use fully automatic data input in the form of abdomen slices, which sets it apart from other studies that rely on cropped regions. Secondly, adopting GAN-based augmentation significantly improved the models' performance by 5%. Moreover, this study validates the robustness of the developed system for FLL localization and categorization externally with two different datasets (Task03_Liver and hcc_tace_seg) obtained from well-known sources (MSD and TCIA).

We used external data to verify the robustness of our derived model. The Task03_Liver dataset from MSD had 131 training samples with mask annotation of the tumor and liver. Another dataset, hcc_tace_seg, contained 105 confirmed HCC patients who underwent CT prior to and after transarterial chemoembolization procedure. We used pretreatment CT of randomly 50 patients each for both datasets. All the images were converted from DICOM to JPG using similar procedure as during model derivation. These masks helped generate the bounding box (by annotator 3, a hepatologist with more than 15 years of experience). When observing the datasets, it was found that hcc_tace_seg (TCIA) had larger lesions (HCC) than Task03_Liver (MSD); this could be one potential reason for better localization. The second phase of our DLLC system is the classification of FLLs; however, the type of lesion was unknown in the MSD dataset, which could affect our evaluation. Besides, the TCIA dataset contained HCC only. As a result, we only used the TCIA dataset to classify lesions into benign or malignant. It was observed that physician 1 (annotator 1) misclassified 11% of FLL as benign; however, only 2% were misclassified as benign lesions by the DLLC system, which is similar to the performance of physician 2 (annotator 2). Potential reasons that physician 2 had better diagnoses are due to experience in the field and being able to view slices continuously and multiple times.

This study has several limitations. First, because of the availability of a smaller number of samples, we could only include some types of FLLs in this study. The data containing other FLLs, such as dysplastic and regenerative nodules, might bias the model performance. In addition, CT images of patients with prior treatment, such as ablation and resection, might also result in misdetection and misclassification of the treated region as FLL. Second, our study was a single-center study where using a single kind of contrast agent might affect the imaging features learned, thus limiting the model's applicability. Third, in this study, we used phase A and phase V of CT for HCC; most benign cases did not have quadruple CT, and fewer images in phase A affected the models' performances. Our preliminary study suggested that if the model is trained well with arterial and portal venous phase images, the FLLs can be well localized in precontrast and delayed phase images. However, extensive study is suggested to support our findings. Therefore, in our future work, we will perform a nationwide study to collect images from different hospitals and include several categories of FLLs, including cholangiocarcinoma and dysplastic nodules, with preferably more samples for each phase of CT, comparing both 2D and 3D detection approaches. Furthermore, we will perform model validation by utilizing newly obtained data collected prospectively. This approach aims to incorporate the validated model into ongoing clinical trials, allowing for real-time validation within clinical settings. The process will involve comprehensive validation to ensure the model's accuracy and efficacy in diverse clinical scenarios, thereby establishing its robustness and reliability for real-world implementation.

Conclusions

This study proposed a DL tool for automatic localization and categorizing FLLs. The model takes the whole slice as input to localize the FLL and further categorize the lesion into benign or malignant categories or specify the lesion type as HCC, HEM, or cyst, as per requirement. Besides, using two phases of CT could achieve an accuracy of 0.97 (95% CI: 0.95–0.99), suggesting the possibility of reducing exposure to radiation and reducing the time required by physicians to study the CT images of a patient. Furthermore, GAN can be used as an efficient substitute to reduce the time-consuming job of data collection and annotation. The synthetic augmentation of data improved the diagnostic performance of the classification model, primarily when multiclass classification was performed. In conclusion, the developed DLLC system can assist inexperienced radiologists and hepatologists in faster and more robust identification of FLLs.

Data availability statement. The data generated or analyzed during the study are available from the corresponding author by request.

References

1 Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA-Cancer J. Clin.* 2021; **71**: 209–49. https://doi.org/10.3322/caac.21660 PubMed PMID: WOS:000614520100001.

2 Rumgay H, Arnold M, Ferlay J *et al*. Global burden of primary liver cancer in 2020 and predictions to 2040. *J. Hepatol.* 2022; **77**: 1598–606. https://doi.org/10.1016/j.jhep.2022.08.021 PubMed PMID: WOS:000928043800015.

3 Liver EAS. EASL Clinical Practice Guidelines: management of hepatocellular carcinoma. *J. Hepatol.* 2018; **69**: 182–236 PubMed PMID: WOS:000436584100020.

4 McDonald RJ, Schwartz KM, Eckel LJ *et al*. The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload. *Acad. Radiol.* 2015; **22**: 1191–8. https://doi.org/10.1016/j.acra.2015.05.007 PubMed PMID: WOS:000359876600015.

5 Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2022. *CA-Cancer J. Clin.* 2022; **72**: 7–33. https://doi.org/10.3322/caac.21708 PubMed PMID: WOS:000741534100001.

6 Yildirim MB, Sahiner IT, Poyanli A *et al*. Malignant tumors misdiagnosed as liver hemangiomas. *Front. Surg.* 2021: 715429. https://doi.org/10.3389/fsurg.2021.715429 PubMed PMID: WOS:000687883400001.

7 Park HJ, Park B, Lee SS. Radiomics and deep learning: hepatic applications. *Korean J. Radiol.* 2020; **21**: 387–401. https://doi.org/10.3348/kjr.2019.0752 PubMed PMID: WOS:000523560800001.

8 Ahn JC, Connell A, Simonetto DA, Hughes C, Shah VH. Application of artificial intelligence for the diagnosis and treatment of liver diseases. *Hepatology* 2021; **73**: 2546–63.
9 Survarachakan S, Prasad PJR, Naseem R *et al*. Deep learning for image-based liver analysis—a comprehensive review focusing on malignant lesions. *Artif. Intell. Med.* 2022: 102331. https://doi.org/10.1016/j.artmed.2022.102331 PubMed PMID: WOS:000823230600001.
10 Goodfellow I, Pouget-Abadie J, Mirza M *et al*. Generative adversarial networks. *Commun. ACM.* 2020; **63**: 139–44.
11 Yasaka K, Akai H, Abe O, Kiryu S. Deep learning with convolutional neural network for differentiation of liver masses at dynamic contrast-enhanced CT: a preliminary study. *Radiology* 2018; **286**: 899–908. https://doi.org/10.1148/radiol.2017170706 PubMed PMID: WOS:000425583200019.
12 Shi WQ, Kuang SC, 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.* 2020; **45**: 2688–97. https://doi.org/10.1007/s00261-020-02485-8 PubMed PMID: WOS:000522574300001.
13 Kim DW, Lee G, Kim SY *et al*. Deep learning-based algorithm to detect primary hepatic malignancy in multiphase CT of patients at high risk for HCC. *Eur. Radiol.* 2021; **31**: 7047–57. https://doi.org/10.1007/s00330-021-07803-2 PubMed PMID: WOS:000630283200002.
14 Oestmann PM, Wang CJ, Savic LJ *et al*. Deep learning-assisted differentiation of pathologically proven atypical and typical hepatocellular carcinoma (HCC) versus non-HCC on contrast-enhanced MRI of the liver. *Eur. Radiol.* 2021; **31**: 4981–90. https://doi.org/10.1007/s00330-020-07559-1 PubMed PMID: WOS:000605578600009.
15 Wang SH, Han XJ, Du J *et al*. Saliency-based 3D convolutional neural network for categorising common focal liver lesions on multisequence MRI. *Insights Imaging* 2021; **12**: 173. https://doi.org/10.1186/s13244-021-01117-z PubMed PMID: WOS:000722208600005.
16 Zhou JR, Wang WZ, Lei BW *et al*. Automatic detection and classification of focal liver lesions based on deep convolutional neural networks: a preliminary study. *Front. Oncol.* 2021; **10**: 581210. https://doi.org/10.3389/fonc.2020.581210 PubMed PMID: WOS:000617148400001.
17 Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:151106434. 2015.
18 Wang RJ, Li X, Ling CX. Pelee: a real-time object detection system on mobile devices. *Adv. Neur. In.* 2018: 31 PubMed PMID: WOS:000461823301091.
19 Tan M, Pang R, Le QV, eds. EfficientDet: scalable and efficient object detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.
20 Lin TY, Goyal P, Girshick R, He KM, Dollar P. Focal loss for dense object detection. *IEEE T. Pattern Anal.* 2020; **42**: 318–27. https://doi.org/10.1109/Tpami.2018.2858826 PubMed PMID: WOS:000508386100006.
21 Redmon J, Farhadi A. Yolov3: an incremental improvement. arXiv preprint arXiv:180402767. 2018.
22 Wang C-Y, Bochkovskiy A, Liao H-YM, eds. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.
23 Zlocha M, Dou Q, Glocker B. Improving RetinaNet for CT lesion detection with dense masks from weak RECIST labels. *Lect. Notes Comput. Sci.* 2019; 11769: 402–10. https://doi.org/10.1007/978-3-030-32226-7_45 PubMed PMID: WOS:000548737100045.
24 Jocher G, Chaurasia A, Qiu J. YOLO by Ultralytics. Available from URL: https://github.com/ultralytics/. 2023.
25 Jacob S, Francesco. What is YOLOv8? A complete guide. https://blog.roboflow.com/whats-new-in-yolov8/. 2023.
26 Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-ResNet and the impact of residual connections on learning. arXiv preprint arXiv:160207261. 2016.
27 Tan MX, Le QV. EfficientNet: rethinking model scaling for convolutional neural networks. *Proc. Mach. Learn. Res.* 2019: 97 PubMed PMID: WOS:000684034306026.
28 Vieira SM, Kaymak U, Sousa JM, eds. Cohen's kappa coefficient as a performance measure for feature selection. In: *International Conference on Fuzzy Systems*. IEEE, 2010.
29 Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, *et al*. Tensorflow: large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:160304467. 2016.
30 Liu XX, Faes L, Kale AU *et al*. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit Health.* 2019; **1**: E271–97. https://doi.org/10.1016/S2589-7500(19)30123-2 PubMed PMID: WOS:000525871300011.
31 Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. *IEEE Int. Conf. Comput. Vis.* 2017: 618–26. https://doi.org/10.1109/Iccv.2017.74 PubMed PMID: WOS:000425498400065.
32 Clark K, Vendt B, Smith K *et al*. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J. Digit. Imaging* 2013; **26**: 1045–57. https://doi.org/10.1007/s10278-013-9622-7 PubMed PMID: WOS:000326698800008.
33 Antonelli M, Reinke A, Bakas S, Farahani K, Kopp-Schneider A, Landman BA, et al. The medical segmentation decathlon. *Nat Commun*. 2022; **13**(1). https://doi.org/10.1038/s41467-022-30695-9. PubMed PMID: WOS:000826101400020.

Supporting information

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

Table S1. Interreader agreement with κ value.
Table S2. Hyperparameters used for the localization models.
Table S3. Performance metrics obtained during model validation.
Table S4. Accuracies obtained during classification model selection.
Table S5. Overall size-wise FLL localization and classification accuracy achieved during training and testing.
Figure S1. Samples of CT slices used in this study. (a): Different types and sizes of focal liver lesions. (b): Bounding box annotations for the samples.
Figure S2. Structure of deep convolutional generative adversarial network used in our study. (a) Overall structure of deep convolutional generative adversarial network (DCGAN). (b) Generator. (c) Discriminator.
Figure S3. Brief structure of adopted deep learning models for detection and classification tasks. (a): Detection model: YOLOv8. (b) Classification model: EfficientNetB5.
Figure S4. Histogram for the size of FLLs included in the study.
Figure S5. Slice-wise testing performance obtained from the selected hyper-tuned detection model: YOLOv8. (a): Ground truth. (b): Localization by AI model.

Figure S6. The quadruple CT of an 86-year-old female with confluent hepatocellular carcinomas with multiple satellite nodules in the right hepatic lobe. The patient-wise performance for YOLOv8 shows the FLL localized in different CT phases where all slices for the patient were provided as input. (a): Phase A slices. (b): Phase V slices. (c): Non-contrast phase slices. (d) Delayed phase slices.

Figure S7. Input samples were provided to DCGAN, and the model generated output samples. (a) Samples for HCC. (b) Samples for HEM. (c) Samples for cyst.

Figure S8. Receiver operating characteristic curves. (a, c), and (b, d) show binary and multiclass classification for Phase AV, respectively.

Figure S9. Comparison of two-stage approach versus one-stage end-to-end approach. (a, b) Visualization of Hepatocellular Carcinoma (HCC), Hemangioma (HEM), and cysts through the t-SNE algorithm, before and after incorporating GAN-derived image, respectively. (c) A few samples show the detection of FLLs using a one-step approach, where the blue, red, and yellow boxes represent ground truth and benign and malignant lesions, respectively. The lesions were considered both benign and malignant, with similar confidence.