Preoperative diagnosis of hepatocellular carcinoma patients with bile duct tumor thrombus using deep learning method

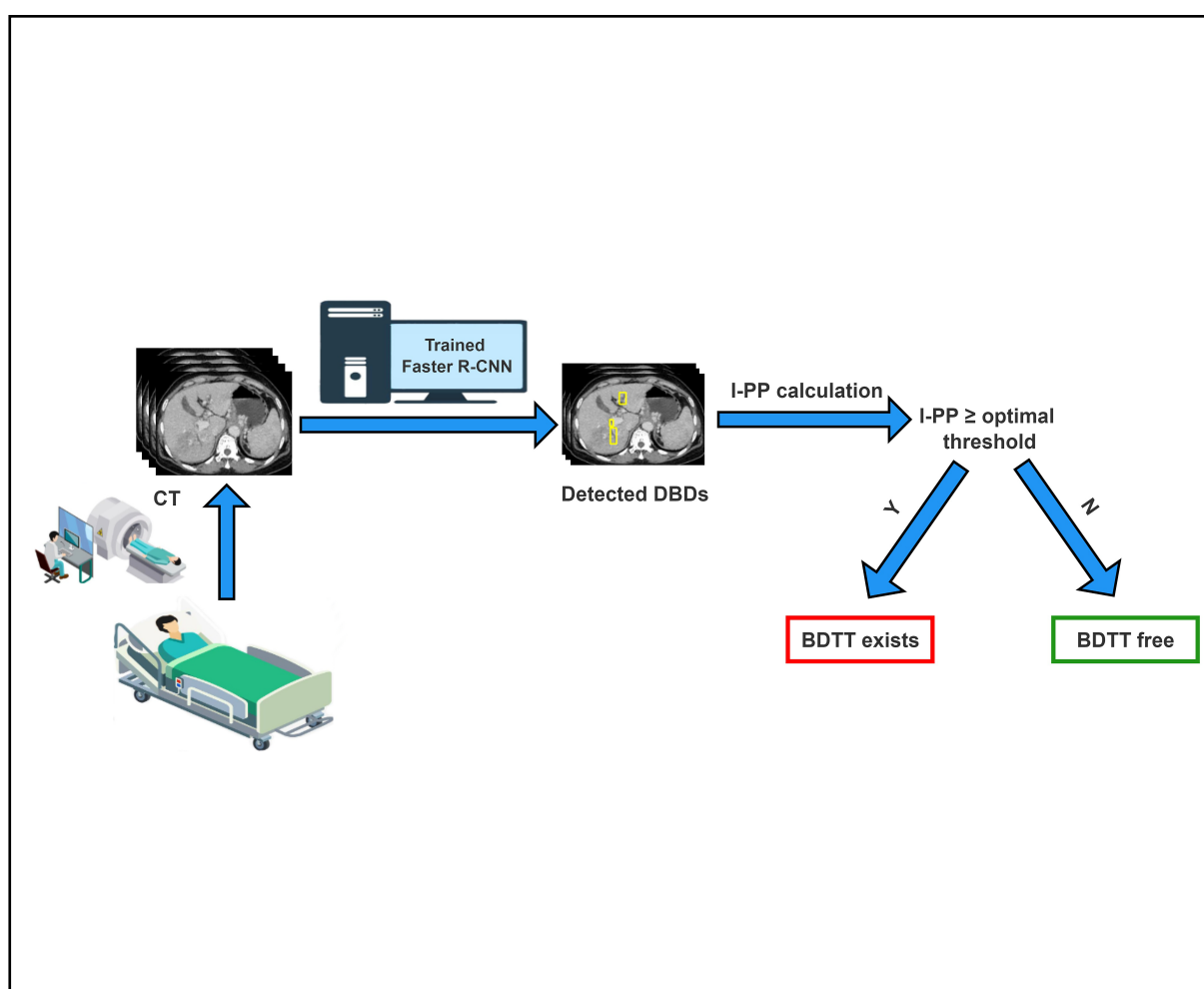Jinming Liu[1], Jiayi Wu[2], Anran Liu[1], Yannan Bai[2], Hong Zhang[1] ✉, and Maolin Yan[2]

[1]School of Management, University of Science and Technology of China, Hefei 230026, China;
[2]Department of Hepatobiliary Pancreatic Surgery, Fujian Provincial Hospital, Fuzhou 350001, China

✉Correspondence: Hong Zhang, E-mail: zhangh@ustc.edu.cn

Graphical abstract

Overview of the proposed computer-aided diagnosis process for BDTT.

Public summary

■ We proposed the first AI pipeline for the diagnosis and evaluation of BDTT through indentifying DBDs on CT images.

■ Our AI pipeline achieved a high AUC of 0.94 (95% CI: 0.87, 1.00) in the application to a real dataset.

Preoperative diagnosis of hepatocellular carcinoma patients with bile duct tumor thrombus using deep learning method

Jinming Liu[1], Jiayi Wu[2], Anran Liu[1], Yannan Bai[2], Hong Zhang[1] ✉, and Maolin Yan[2]

1School of Management, University of Science and Technology of China, Hefei 230026, China;
2Department of Hepatobiliary Pancreatic Surgery, Fujian Provincial Hospital, Fuzhou 350001, China

✉Correspondence: Hong Zhang, E-mail: zhangh@ustc.edu.cn

Abstract: Preoperative diagnosis of bile duct tumor thrombus (BDTT) is clinically important as the surgical prognosis of hepatocellular carcinoma (HCC) patients with BDTT is significantly different from that of patients without BDTT. Although dilated bile ducts (DBDs) can act as biomarkers for diagnosing BDTT, it is easy for doctors to ignore DBDs when reporting the imaging scan result, leading to a high missed diagnosis rate in practice. This study aims to develop an artificial intelligence (AI) pipeline for automatically diagnosing HCC patients with BDTT using medical images. The proposed AI pipeline includes two stages. First, the object detection neural network Faster R-CNN was adopted to identify DBDs; then, an HCC patient was diagnosed with BDTT if the proportion of images with at least one identified DBD exceeded some threshold value. Based on 2354 CT images collected from 32 HCC patients (16 with BDTT and 16 without BDTT, 1∶1 matched), the proposed AI pipeline achieves an average true positive rate of 0.92 for identifying DBDs per patient and a patient-level true positive rate of 0.81 for diagnosing BDTT. The AUC value of the patient-level diagnosis of BDTT is 0.94 (95% CI: 0.87, 1.00), compared with 0.71 (95% CI: 0.51, 0.90) achieved by random forest based on preoperative clinical variables. The high accuracies demonstrate that the proposed AI pipeline is successful in the diagnosis and localization of BDTT using CT images.

Keywords: hepatocellular carcinoma; bile duct tumor thrombus; object detection; artificial intelligence; deep learning

**CLC number:** TP391.41   **Document code:** A

1 Introduction

In 1947, Mallory et al.[1] reported the first hepatocellular carcinoma (HCC) patient with bile duct tumor thrombus (BDTT) who had a typical symptom of obstructive jaundice. BDTT is uncommon in clinic, with an incidence of 0.5%−12.9% among HCC patients[1−4]. HCC patients with BDTT were shown to have worse prognoses than HCC patients without BDTT after liver resection or liver transplantation[1−8]. Specifically, HCC patients with BDTT were shown to have a higher recurrence rate[4−6] (e.g., one-year recurrence rates were estimated to be 70.3% and 34.8% for HCC patients with and without BDTT, respectively[6]) and a lower overall survival rate[1−8] (refer to Table S1 in the supporting information). Early diagnosis is important in improving the prognosis of such HCC patients, as it can help doctors make better preoperative decisions in determining surgical areas[9]. Currently, the diagnosis and evaluation of HCC with BDTT before surgery mainly depend on clinical symptom judgment and imaging examination[2, 5]. However, most HCC patients with BDTT have no specific clinical symptoms at early stages[10]. Furthermore, obstructive jaundice caused by BDTT's invasion into the common hepatic duct or common bile duct can be easily misdiagnosed as a symptom of biliary stones or cholangiocarcinoma[1, 10−12].

With the development of medical imaging technology, CT and MRI scans have been widely accepted as safe and valuable methods for diagnosing BDTT. Although the imaging features of BDTT and their correlations with the corresponding histopathologic manifestations have been reported in the literature[10−13], it is still difficult for doctors to identify BDTT on medical images by the naked eye. Many clinical studies have shown that there are three general clinical features of HCC with BDTT on CT/MRI images: (ⅰ) the presence of intraluminal soft-tissue masses in the bile duct, with biliary dilation above the obstruction; (ⅱ) enhancement of the intraluminal soft-tissue masses at the arterial phase; and (ⅲ) rarely thickening of the bile duct wall[2, 10, 14]. Based on this fact, doctors can diagnose BDTT by identifying dilated bile ducts (DBDs) near intrahepatic tumors. DBDs caused by BDTT usually present as dark linear structures on CT images (Fig. 1a). DBDs occupy only a small part of the CT image, and these inconspicuous structures tend to be easily ignored if doctors lack sufficient awareness of their appearance, resulting in a high missed diagnosis rate in practice. To the best of our knowledge, no automatic method has been developed to identify DBDs, despite its importance in the accurate preoperative diagnosis of BDTT.

The coming era of big data and powerful computer hardware have promoted the development of artificial intelligence (AI), which enables the mining of disease information in
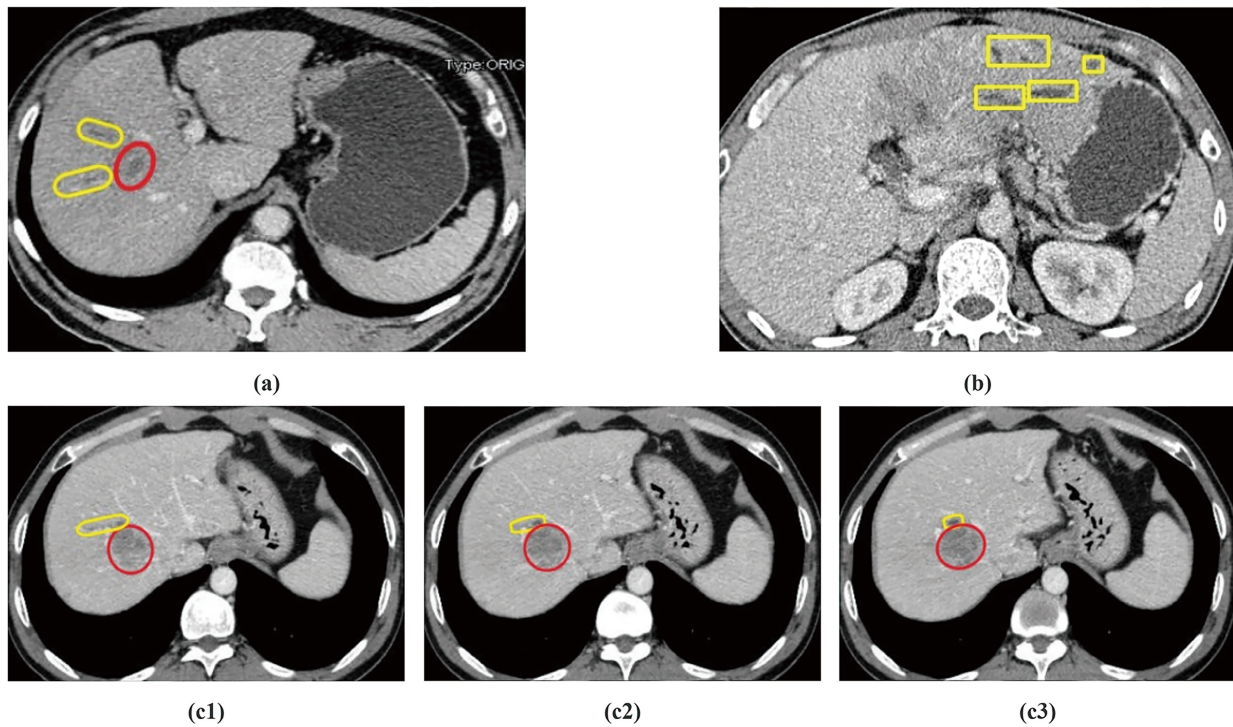
Fig. 1. (a) A CT image of an HCC patient with BDTT. (b) A CT image with four labeled bounding boxes for DBDs. (c1)–(c3) Three consecutive CT images of an HCC patient with BDTT. In all images, tumors were marked in ellipses and DBDs caused by BDTT were marked in capsules.

medical images such as CT and MRI images. Many AI algorithms have been developed to effectively utilize the disease information contained in medical images. Among these AI methods, deep learning methods have shown their competitiveness in many fields. As one of the most representative deep learning methods, convolutional neural networks (CNNs) have been widely applied in image recognition, providing state-of-the-art results, such as in the fields of image classification, semantic segmentation, and object detection[15]. Based on CNNs, more advanced neural networks have been proposed, such as fully convolutional networks[16], generative adversarial networks[17], and variational autoencoders[18]. Many researchers applied these neural networks in medical fields and achieved promising results in identifying bone fractures[19] and tumor regions[20], screening high-risk subjects of a specific disease[21], locating biomarkers[22], segmenting tissues and organs[23–26].

Many neural network methods have been developed specifically for object detection. Compared with traditional classification methods, object detection methods can identify objects and locate them. Some well-developed object detection neural networks include Faster R-CNN[27], Reppoints[28], and Foveabox[29]. These object detection methods have shown broad application prospects in clinical diagnosis. For example, Thian et al.[30] applied Faster R-CNN to locate the fracture area on wrist radiographs and achieved sensitivities as high as 91.2% and 96.3% based on forward images and lateral images, respectively. Their object detection method outperformed traditional CNN methods, such as VGG network[31] and Inception V3 network[32]. Boot and Irshad[33] combined Faster R-CNN with Unet[23] to locate and segment breast

tumors and achieved a micro average F1 score of 0.81 and a macro average F1 score of 0.84.

Our study aims to develop an AI pipeline for the preoperative diagnosis of BDTT by identifying DBDs on CT images. To the best of our knowledge, this is the first computer-aided method developed for diagnosing BDTT. Based on a total number of 32 HCC patients with 2354 CT images, the developed pipeline has an area under the ROC curve as high as 0.94, making it a powerful tool in diagnosing BDTT. The proposed AI pipeline can also locate the DBDs caused by BDTT.

2   Materials and methods

2.1   Study data

From April 2010 to December 2019, CT images of 16 HCC patients with BDTT were collected from four hospitals that are well-known for the diagnosis and treatment of hepatobiliary diseases in China. During the same period, 50 patients who were pathologically confirmed as HCC without BDTT were randomly selected. Our study has a relatively small sample size due to the rare incidence of BDTT. The numbers of cases are generally small in most BDTT studies; for example, only 22[2], 14[3], and 7[13] HCC patients with BDTT were collected in previous studies.

Sixteen HCC patients with BDTT were denoted as the case group, and 50 HCC patients without BDTT were denoted as the control group. All patients underwent preoperative serological examinations, preoperative imaging examination, collection of demographic information (e.g., gender and age), tests for hepatitis B virus infection and assessment of liver

(reserve) function, surgical resection, postoperative histopathological, and immunohistochemical examinations. The diagnosis of HCC with BDTT was based on postoperative histopathological examination with two experienced pathologists.

2.2 Case-control matching and descriptive statistical analyses

Each HCC patient with BDTT was matched with an HCC patient without BDTT through propensity score matching (PSM) to minimize the influence of potential confounders in the diagnosis of BDTT. Specially, propensity scores were calculated using the logistic regression model based on covariates including age x_a, gender x_s, and the maximum tumor size x_m. Let Y denote a label indicating whether a patient is in the case group (1 = yes; 0 = no), X denote the set of the three covariates, then the propensity score $P(Y = 1|X)$ is

$$P(Y = 1|X) = \frac{e^{w_1 x_a + w_2 x_s + w_3 x_m + b}}{1 + e^{w_1 x_a + w_2 x_s + w_3 x_m + b}},$$

where w_1, w_2, w_3, b are the parameters of the logistic regression model, which are estimated by the maximum likelihood principle. Finally, each patient in the case group was matched with a patient in the control group with the closest propensity score. PSM was implemented in the R package Matchit (version 4.3.3)[34]. The 16 control patients after PSM are denoted as the matched control group.

Continuous variables were expressed as the mean ± standard deviation and the categorical variables were expressed as their numbers. The Wilcoxon Mann-Whitney test and Fisher's exact test were adopted to compare continuous variables and categorical variables, respectively, between the case group and the matched/unmatched control group. The Wilcoxon Mann-Whitney test was also adopted to compare the proportions between the case group and the matched control group. The DeLong test implemented in the R package pROC[35] was adopted to compare the AUC values between two ROC curves. All statistical analyses were performed using R version 4.0.3[36]. $P < 0.05$ was considered statistically significant.

2.3 The proposed BDTT diagnosis pipeline

The proposed AI pipeline for diagnosing BDTT is displayed in Fig. 2, which consists of three steps: Step1, image filtering, preprocessing, and label making (Section 2.4); Step 2, DBD identification (Section 2.5); and Step 3, patient-level diagnosis of BDTT (Section 2.7).

2.4 Image filtering, preprocessing, and label making

Some raw CT images have grayscale JPEG (joint photographic experts group) format, while some have DICOM (digital imaging and communications in medicine) format. To remove potential confounding effects due to the format, it is necessary to convert the raw CT images so that they have the same format. First, all the DICOM files were converted to lossless grayscale JPEG images. Then, the resulting grayscale images were converted to pseudo-RGB images by tripling their channels with each channel having the same pixel value so that they could serve as inputs of an object detection neural network.

With CT scans, each patient's images can be divided into four phases: plain scan phase, arterial phase, portal venous phase, and delayed phase. The brightness contrast between DBDs and normal tissues is most obvious in the portal venous phase (Fig. 3). Note that images in the portal venous phase comprise a series of consecutive abdominal scan slices, and the liver region generally occupies the middle part of those slices. Furthermore, intrahepatic DBDs usually present near tumor regions. Therefore, after format conversion, only those CT images in the portal venous phase with tumors present were retained. Such filtering could balance data between the case group and the matched control group. The image filtering procedure was reviewed by an experienced radiologist.

Since CT images of the case group were provided by four hospitals, their sizes vary to some extent due to different scanning equipment. Therefore, all images were resized to have a unified size and noninformative parts were discarded.

After filtering and preprocessing, the heterogeneity of CT images from different centers was largely eliminated, and the main difference between images of the case group and the matched control group was whether DBDs were present. Sev-
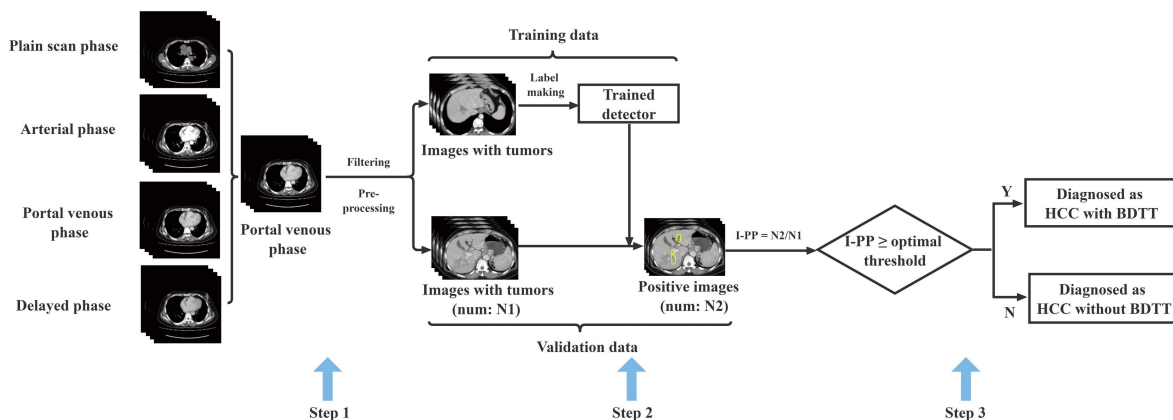
Fig. 2. The proposed AI pipeline for diagnosing BDTT. First, CT images in the portal venous phase with tumors were selected, then the selected images were center-cropped and resized to a unified size. After filtering and preprocessing, the trained DBD detector was applied to identify DBDs on the resulting images. Finally, a patient was diagnosed as HCC with BDTT if the image-level positive proportion (I-PP) was greater than the optimal threshold.
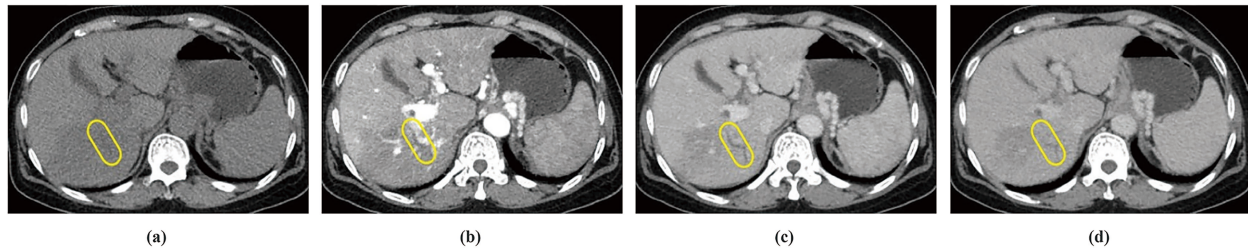
Fig. 3. CT images of an HCC patient with BDTT at the same location from (a) plain scan phase, (b) arterial phase, (c) portal venous phase, and (d) delayed phase. DBDs caused by BDTT were marked in capsules.

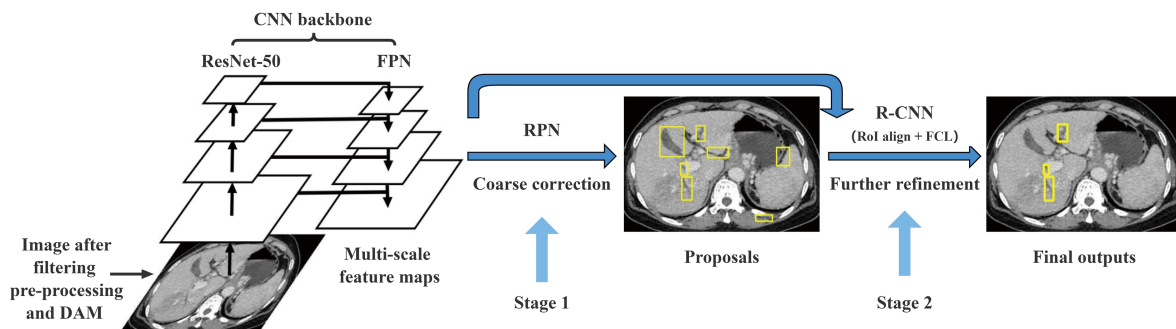eral examples of preprocessed CT images with tumors and DBDs are presented in Fig. 1, which shows that most DBDs can appear on several consecutive images in the portal venous phase since CT images are usually generated by continuous cross-section scanning of the human body (Fig. 1 c1–c3). Finally, DBDs were labeled by an experienced radiologist using the Labelimg software[37], which extracted the location information of DBDs, i.e., the coordinates of the upper left and lower right corners of the bounding boxes as XML files. The resulting CT images combined with the label information were used to train an object detection neural network.

2.5   Object detection

Faster R-CNN[27], a popular two-stage neural network designed for object detection, was adopted to identify DBDs. Faster R-CNN consists of four components, namely data augmentation module (DAM), CNN backbone, region proposal network (RPN), and region CNN (R-CNN). Compared with one-stage object detection methods such as YOLO[38] and SSD[39], Faster R-CNN tends to be more accurate, albeit at a higher computational cost[40].

The flowchart of Faster R-CNN for identifying DBDs is shown in Fig. 4. To enrich data, multiple data augmentation techniques (e.g., random shift, random rotation, etc.) were implemented in the DAM component during the training process. Meanwhile, transfer learning was adopted in this study, i.e., the Faster R-CNN model was pretrained on the Microsoft COCO dataset[41] with ResNet-50[42] combined with a feature pyramid network (FPN)[43] as the CNN backbone for extracting feature maps. By combining FPN within the backbone, multiscale feature maps were introduced, and strong semantic features were propagated through a top-down pathway; thus, the localization ability could be greatly enhanced.

FPN was shown to be very effective in detecting small objects[44, 45]. After feature extraction, RPN and R-CNN completed the identification of DBDs in two stages sequentially. First, RPN was used to discriminate regions that tend to contain DBDs from the background as proposals and correct their coordinates. Then, R-CNN was used to classify the proposals and further refine the coordinates of these proposals. Specifically, in R-CNN, RoI Align was used to extract uniform features from the proposals, and fully connected layers (FCL) were used for classification and further localization. During the training phase, the total loss consisted of classification losses L_{cls} and localization losses L_{loc}, which were produced by RPN and R-CNN, respectively:

$$L_{total} = \lambda_1 L_{cls\_RPN} + \lambda_2 L_{loc\_RPN} + \lambda_3 L_{cls\_RCNN} + \lambda_4 L_{loc\_RCNN},$$

where λ_1, λ_2, λ_3, λ_4 are hyperparameters for tuning the weights of different losses. Model configurations and training details of Faster R-CNN are included in the supporting information (refer to Appendix A for model configurations and Appendix B for training details).

2.6   Evaluation metrics for DBD identification

Due to limited data, four fold cross validation was used to evaluate the model's performance. Specifically, the 16 case patients were evenly and randomly divided into four subsets. Each of the four subsets was used for validation, and the remaining three subsets were used for training. Thus, this training process was repeated four times.

The DBD-level true positive rate (D-TPR) was used as an evaluation metric to assess the model's performance in identifying DBDs. For each patient, D-TPR was defined as the ratio of the number of successfully detected DBDs to the number of true DBDs:

Fig. 4. Flow chart of Faster R-CNN. The pre-processed CT images were first transformed via DAM, then CNN backbone extracted multiscale feature maps from the transformed image. Based on the extracted features, at the first stage, RPN output proposal regions which tend to contain DBDs, then at the second stage, R-CNN further classified the proposal regions and refined their coordinates as the final outputs.

$$D\text{-}TPR = \frac{\#\{\text{ successfully detected DBDs}\}}{\#\{\text{ true DBDs}\}}.$$

Specifically, each DBD was considered to be successfully detected if at least one bounding box output by Faster R-CNN overlapped with the DBD.

In each of the four training processes, the image-level positive proportion (I-PP) was defined as the proportion of images with at least one detected DBD for each patient in both the case validation set and the matched control group:

$$I\text{-}PP = \frac{\#\{\text{ positive images}\}}{\#\{\text{ all images}\}}.$$

I-PPs for each patient in the matched control group were averaged over the four training processes.

2.7 Patient-level diagnosis of BDTT

After identifying DBDs, an HCC patient was diagnosed with BDTT if his/her I-PP was larger than a given threshold. Sensitivity, Specificity, receiver operating characteristic (ROC) curve, and area under ROC curve (AUC) were used as diagnostic method evaluation metrics. For each given I-PP threshold, let TP denote the number of correctly diagnosed patients in the case group, FP denote the number of misdiagnosed patients in the control group, TN represent the number of correctly diagnosed patients in the control group, and FN be the number of misdiagnosed patients in the case group. Sensitivity was defined as the proportion of positive cases in HCC patients with BDTT, and Specificity was defined as the proportion of negative cases in HCC patients without BDTT:

$$\text{Sensitivity} = \frac{TP}{TP + FN},$$

$$\text{Specificity} = \frac{TN}{TN + FP}.$$

The ROC curve was then drawn based on Sensitivity vs. Specificity pairs across various I-PP thresholds. The AUC value was defined as the area under the ROC curve. All evaluation metrics were calculated using the R package pROC[35].

2.8 Random forest as an alternative method

Random forest[46] was used as an alternative diagnosis method, which has become one of the most commonly used machine learning algorithms in classification and regression due to its high accuracy and fast training speed. In classification task, random forest ensembles multiple decision trees, and its output category is determined by the mode of categories output by individual trees. The predictors' input into random forest include the serum level of total bilirubin, alanine aminotransferase (ALT), aspartate transaminase (AST), microvascular invasion, tumor number, and the Edmondson-Steiner grade, which have been shown to be potentially associated with BDTT[4, 6, 7, 14].

2.9 Ablation study

To validate the effectiveness of the proposed AI pipeline, the other two commonly used object detection neural networks, Reppoints[28] and Foveabox[29], were adopted as alternative object detectors for identifying DBDs in our proposed pipeline.

Reppoints is a flexible detector that learns a set of points that can adaptively circumscribe objects' patial extent and indicate semantically significant local areas, thus achieving fine-grained localization and classification. Foveabox directly predicts the probability of having a specific object around the center point of each area on the image and calculates its four distances from the four boundaries of the corresponding target. In this study, for a fair comparison, both detectors were also first pretrained on the COCO dataset[41] before detecting DBDs. Meanwhile, the image preprocessing techniques and training mode were the same as those of Faster R-CNN.

3 Results

3.1 Data preprocessing results

Table 1 summarizes the patients' demographic, clinical, and pathological variables in the case group and the matched con-

Table 1. Demographic, clinical, and pathological variables of HCC patients after PSM.

	Case ($n = 16$)	Control ($n = 16$)	P value
Gender (male/female)	13/3	11/5	0.69
Age (years)	52.75 ± 12.84	49.75 ± 9.69	0.65
HBsAg(–/+)	0/16	1/15	1.00
Liver cirrhosis (–/+)	0/16	5/11	0.04*
Total bilirubin (μmol/L)	17.90 ± 6.38	16.61 ± 12.06	0.22
Albumin (g/L)	39.82 ± 5.10	42.68 ± 5.47	0.12
ALT (U/L)	60.81 ± 41.55	68.69 ± 73.06	0.88
AST (U/L)	85.50 ± 69.64	68.69 ± 81.93	0.06
Child-Pugh class (A/B)	15/1	15/1	1.00
AFP (<400 / ⩾400 ng/mL)	5/11	5/11	1.00
Tumor number (single/ multiple)	8/8	13/3	0.14
Maximum tumor size (cm)	8.23 ± 5.54	9.22 ± 5.47	0.68
Microvascular invasion (–/+)	7/9	2/14	0.11
Edmondson-Steiner grade (I-II/III-IV)	4/12	8/8	0.27
The AJCC 8th Staging System (I-II/III-IV)	5/11	5/11	1.00

Abbreviations: HBsAg, hepatitis B surface antigen; ALT, alanine aminotransferase; AST, aspartate transaminase; AFP, a-fetoprotein; AJCC, American Joint Committee on Cancer. *$P < 0.05$.

Table 2. D-TPR of 16 patients in case group (average D-TPR: 0.92).

ID	tDBD	dDBD	D-TPR	ID	tDBD	dDBD	D-TPR
1	5	5	1.00	9	18	15	0.83
2	10	8	0.80	10	11	11	1.00
3	11	11	1.00	11	4	4	1.00
4	11	9	0.82	12	18	17	0.94
5	10	9	0.90	13	9	8	0.89
6	5	5	1.00	14	8	7	0.88
7	7	7	1.00	15	17	14	0.82
8	7	7	1.00	16	16	13	0.81

Abbreviations: tDBD—the number of true DBDs, dDBD—the number of successfully detected DBDs, D-TRP = dDBD/tDBD.
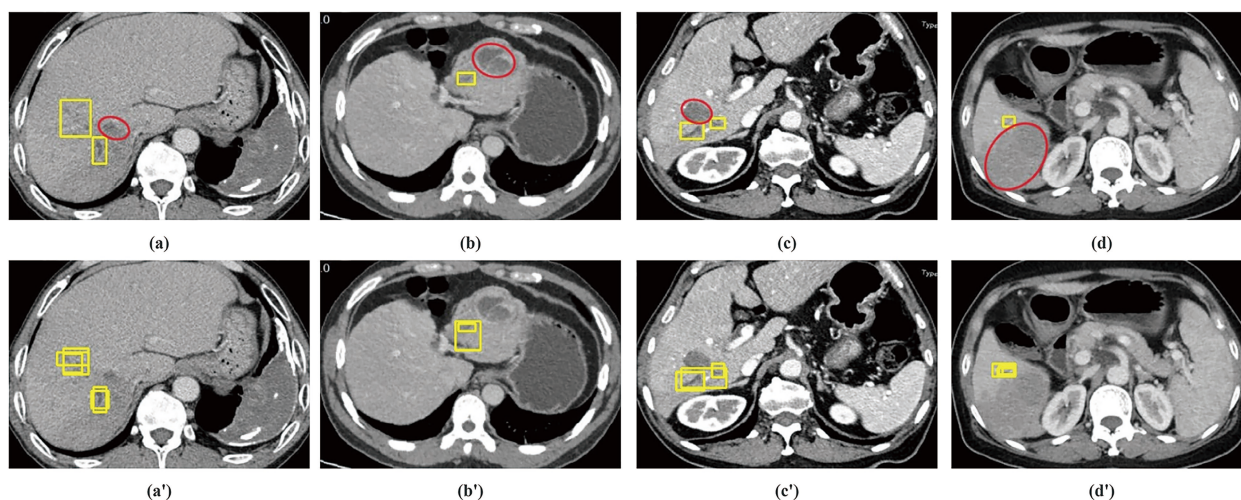
Fig. 5. Several examples of successfully detected DBDs: (a)–(d) were the ground truth annotations of CT images of four different HCC patients with BDTT (tumors were marked in ellipses and DBDs caused by BDTT were marked in rectangles); (a')–(d') were the corresponding bounding boxes output by Faster R-CNN.

trol group. The demographic variables included gender and age; the clinical variables included HBsAg, liver cirrhosis, the serum level of total bilirubin, albumin, ALT, AST, AFP, and Child-Pugh class; the pathological variables included tumor number, maximum tumor size, microvascular invasion, Edmondson-Steiner grade, and AJCC cancer stage. Before PSM, the differences between the case group and the control group were significant in liver cirrhosis ($P = 0.01$), total bilirubin level ($P = 0.01$), albumin level ($P = 0.02$), ALT level ($P = 0.01$), AST level ($P < 0.01$), and microvascular invasion ($P < 0.01$). From Table 1, after PSM, all the variables except liver cirrhosis showed no significant difference between the case group and the matched control group, demonstrating that PSM can effectively control confounding biases.

In total, 1160 and 1194 CT images were retained for the case group and the matched control group, respectively. After labeling the CT images for the case group, there were 756 bounding boxes labeled on the 553 CT images, where 366 (66.2%) contained one bounding box, 172 (31.1%) contained two bounding boxes, 14 (2.5%) contained three bounding boxes, and only 1 (0.2%) contained four bounding boxes (Fig. 1b). The largest annotated bounding box only accounted for less than 6.0% of the whole image, suggesting that

DBDs were inconspicuous and could be easily ignored by the naked eye.

3.2 DBD detection results

Table 2 shows the D-TPRs for the 16 patients in the case group. The average D-TPR value was 0.92 and the range of D-TPR was 0.80–1.00, indicating a high success rate of detecting DBDs. The average D-TPR values were 0.87 and 0.86 for Reppoints and Foveabox, respectively (refer to Table S2 and Table S3 in the supporting information for the D-TPRs of the 16 patients in the case group). Fig. 5 shows several examples of successfully detected DBDs. In these examples, for most DBDs, Faster R-CNN outputs multiple bounding boxes that overlap the DBDs with a certain IoU value. The IoU value is an effective indicator to measure the degree of overlap between bounding boxes (refer to Appendix A in the supporting information for the definition of IoU).

3.3 BDTT diagnosis results

The I-PPs of the patients calculated from the output results of the three different detectors are graphically displayed in Fig. 6 via box plots, which intuitively reflect the difference between the case group and the matched control group. Taking Faster
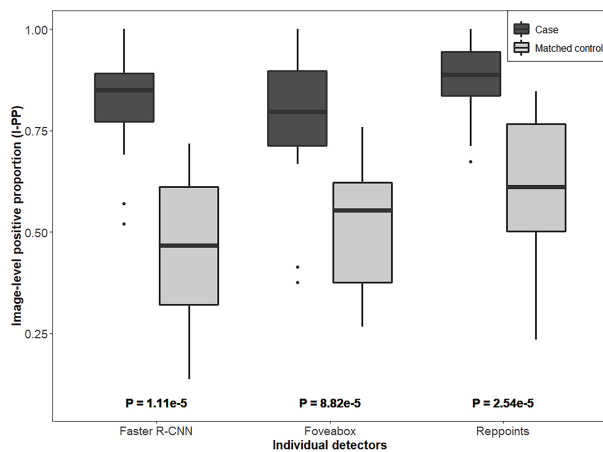
Fig. 6. Box plots of image-level positive proportions (I-PPs) for the case group and the matched control group. I-PPs were calculated from the output results of the three different detectors, respectively, and the P-values for difference comparison were calculated by the Wilcoxon Mann-Whitney test.

Fig. 7. ROC curves of the proposed AI pipeline with Faster R-CNN (solid line), Reppoints (longdash line), Foveabox (dotdash line), and random forest (dotted line). The corresponding AUCs were 0.94, 0.92, 0.89, and 0.71, respectively. The corresponding sensitivities corresponding to the optimal threshold values were 0.81, 0.75, 0.88, and 0.69, respectively. The corresponding specificities corresponding to the optimal threshold values were 1.00, 0.94, 0.88, and 0.75, respectively.
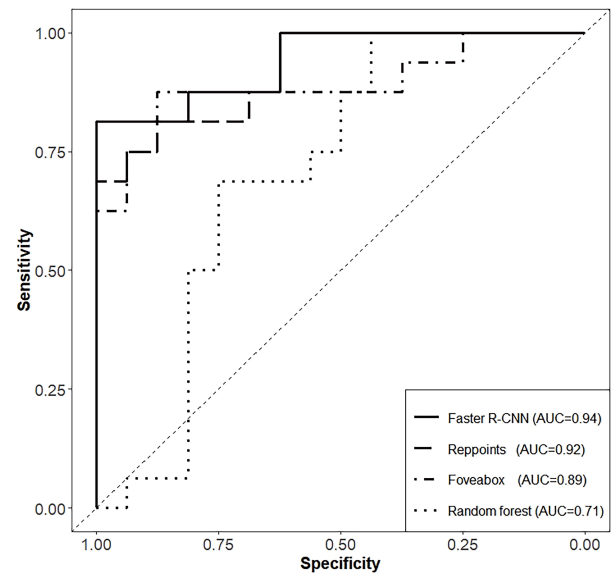
R-CNN as an example, the average I-PP values were 0.82 and 0.47 for the case group and the control group, respectively. Most I-PPs for the case group were equal to or above 0.70 (81.3%), while the majority of I-PPs for the control group were below 0.70 (81.3%), indicating a significant difference between I-PPs of the case group and the matched control group (P-value = 1.11×10^{-5}). Similar results could be observed for Reppoints and Foveabox. Among those HCC patients with BDTT, two patients had relatively low I-PPs. The poor identification results for these patients could be due to their significantly worse image qualities than other patients.

ROC curves for the proposed pipeline with different detectors and the random forest algorithm are displayed in Fig. 7. For the proposed pipeline with Faster R-CNN, the AUC value was 0.94 (95% CI: 0.87, 1.00) and the optimal I-PP threshold value was 0.72, with a corresponding sensitivity and specificity of 0.81 and 1.00, respectively. For random forest, the AUC value was 0.71 (95% CI: 0.51, 0.90), and the optimal probability threshold value was 0.59, with the corresponding sensitivity and specificity values of 0.69 and 0.75, respectively. The AUC of the proposed pipeline with Faster R-CNN is significantly larger than that of the random forest algorithm ($P = 0.03$). In terms of patient-level diagnosis, Reppoints and Foveabox had AUC values of 0.92 (95% CI: 0.83, 1.00) and 0.89 (95% CI: 0.77, 1.00), respectively, which were suggestively larger than that of the random forest (P-values = 0.06 and 0.13, respectively).

In summary, the AI pipeline with Faster R-CNN performs best in the diagnosis of HCC with BDTT.

4　Discussion and conclusions

In this study, an AI pipeline was developed for the accurate preoperative diagnosis of HCC patients with BDTT by automatically identifying DBDs on CT images by the object detection neural network Faster R-CNN. CT images of only 16 HCC patients with BDTT were used in this study due to the rarity of BDTT and the difficulty in obtaining complete ima-

ging data of HCC patients with BDTT. The current study is the first attempt to apply an object detection method to identify DBDs.

The analysis results of the multicenter data demonstrate that the trained Faster R-CNN had promising performance in identifying DBDs. For each patient in the case group, most of their DBDs were successfully detected and nearly half (43.8%) of the patients' DBDs were fully identified. In addition, the bounding boxes output by Faster R-CNN were quite precise (refer to Fig. 5). From the DBD identification result, false negative results (undetected DBDs) arose when their sizes were too small (Fig. 8 a1) or were too inconspicuous (Fig. 8 a2), while the false positive results were mainly due to some confusing structures, such as DBD-like tumor regions (Fig. 8 b1), gaps at the junction between liver and other tissues (Fig. 8 b2), and irrelevant structures outside the liver region (Fig. 8 b3). Nevertheless, most of the false positive findings such as those shown in Fig. 8 b2–b3 were easy to distinguish by doctors.

A significant difference in the identified DBD proportions (I-PPs) was observed between the case group and the matched control group for all three detectors (Fig. 6), which coordinates with the conclusion of Ref. [14]. DBDs can act as biomarkers for BDTT. This motivated us to design a pipeline for diagnosing HCC patients with BDTT (Fig. 2). The proposed pipeline with Faster R-CNN significantly outperformed the traditional machine learning method, random forest, that did not utilize image data (AUCs: 0.94 vs. 0.71), demonstrating the superiority of the proposed pipeline.

The stunning performance of the proposed pipeline could be attributed to two aspects. On the one hand, the two-stage anchor-based object detection neural network Faster R-CNN identified DBDs, which first output object proposals and then
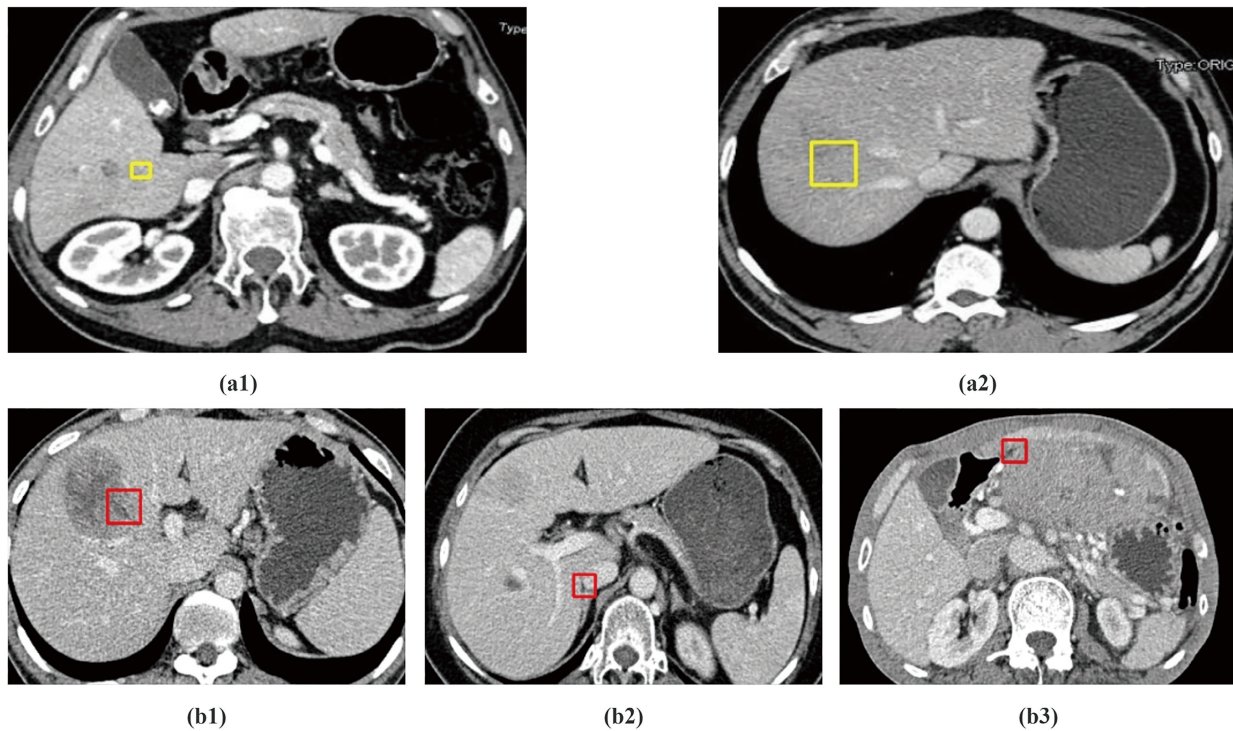
Fig. 8. The first row shows two samples of undetected DBDs: (al) tiny DBD and (a2) inconspicuous DBD. The second row shows three types of dominant false positive bounding boxes: (b1) DBD-like liver tumor region and (b2) gap between liver and other tissues, (b3) irrelevant structures outside the liver region

extracted regional features based on these proposals for detecting DBDs. The design philosophy of looking and thinking twice is consistent with the human vision system and can result in a higher detection accuracy than one-stage network[40]. Meanwhile, our ablation study result indicates that the AI pipeline with Faster R-CNN performs best in the DBD-level detection and patient-level diagnosis among the three detectors. On the other hand, the CT images were first manually filtered and processed, then labeled by an experienced radiologist. The heterogeneity of CT images from different centers was eliminated as much as possible, so the CT image difference between the case group and the matched control group was mainly due to the presence or absence of DBDs. Furthermore, the DBD positions were accurately located, providing a prerequisite for training DBD detectors.

Although postoperative histopathological review of surgical specimens remains the gold standard for diagnosing and evaluating BDTT, determining the existence of BDTT (and their locations, if any) before surgical resection is of significance. It motivated us to use object detection neural networks instead of conventional classification neural networks. The advantage of using object detection methods is obvious; it is difficult for the classification neural networks to focus on such inconspicuous structures of DBDs, while object detection methods not only identify which images have DBDs but also frame the DBD positions. Since the DBD positions are informative for typing and positioning BDTT, the proposed pipeline also provides useful information for doctors to evaluate the condition of BDTT.

There are several limitations to this study. First, some HCC patients with BDTT have no obvious biliary dilation, and our pipeline produced low I-PP values and could be used to misdiagnose these patients. However, as reported in Ref. [14], 93.3% of HCC patients with BDTT had DBDs. In comparison, only 1.4% of the HCC patients without BDTT had DBDs, which could be caused by other diseases, such as biliary stones and intrahepatic cholangiocarcinoma. In the latter case, the diagnosis by our method alone is not sufficient. As shown in Ref. [9], other characteristics, such as washout during the portal venous phase without a thickened bile duct wall, chronic HBV infection, and higher serum AFP levels, can facilitate the differentiation between BDTT and other diseases.

The current AI pipeline developed in this study can be improved from several aspects. First, since the scope of object detection is the whole image, irrelevant structures outside the liver region can produce false positive findings. Accordingly, the false positive rate can be effectively reduced by segmenting the liver region from CT images in advance. However, pixel-level labels are needed to train an organ segmentation neural network such as Unet[23], but they are much more difficult to obtain than bounding box labels. Second, individual images in the portal venous phase were treated as the research objects in this study, which means that the correlation between the CT images was not utilized. Therefore, the DBD identification accuracy could potentially be improved by mining the correlation among the sequential images in the portal venous phase. Third, the proposed pipeline only utilizes CT images of the HCC patients, which can be combined with other clinical evidence to potentially improve the diagnosis of BDTT. Finally, low-quality images may lower the

detection accuracy, as demonstrated in the previous section. Consequently, the proposed pipeline can be improved by using higher-quality images.

In conclusion, the high accuracies make our AI pipeline a powerful tool in the diagnosis and evaluation of BDTT patients and the localization of DBDs caused by BDTT. The pipeline proposed in this study can be extended to other medical applications, such as the diagnosis of lymph node enlargement.

Supporting information

The supporting information for this article can be found online at https://doi.org/10.52396/JUSTC-2022-0057. The supporting information and codes are available online at https://github.com/Daniel-1997/AI-pipeline-for-diagnosing-BDTT.

Ethics declaration

The present study was approved and reviewed by the Institutional Ethics Committee of all relevant institutions. Informed consent was obtained from the patients or their guardians for their data to be used for research purposes. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article. All methods were performed in accordance with the relevant guidelines and regulations.

Conflict of interest

The authors declare that they have no conflict of interest.

Biographies

Jinming Liu is a postgraduate student under the tutelage of Prof. Hong Zhang at the University of Science and Technology of China. His research interests focus on pattern recognition and natural language processing.

Hong Zhang is a Professor with the University of Science and Technology of China (USTC). He received his Bachalor's degree in Mathematics and Ph.D. degree in Statistics from USTC in 1997 and 2003, respectively. His major research interests include statistical genetics, causal inference, and machine learning.

References

[1] Navadgi S, Chang C C, Bartlett A, et al. Systematic review and meta-analysis of outcomes after liver resection in patients with hepatocellular carcinoma (HCC) with and without bile duct thrombus. *HPB*, **2016**, *18* (4): 312–316.

[2] Lu W, Tang H, Yang Z, et al. A proposed modification for the Barcelona clinic liver cancer staging system: Adding bile duct tumor thrombus status in patients with hepatocellular carcinoma. *The American Journal of Surgery,* **2020**, *220* (4): 965–971.

[3] Meng K W, Dong M, Zhang W G, et al. Clinical characteristics and surgical prognosis of hepatocellular carcinoma with bile duct invasion. *Gastroenterology Research and Practice,* **2014**, *2014*: 604971.

[4] Wang D D, Wu L Q, Wang Z S. Prognosis of hepatocellular carcinoma with bile duct tumor thrombus after R0 resection: A matched study. *Hepatobiliary & Pancreatic Diseases International,* **2016**, *15* (6): 626–632.

[5] Wang C, Yang Y, Sun D, et al. Prognosis of hepatocellular carcinoma patients with bile duct tumor thrombus after hepatic resection or liver transplantation in Asian populations: A meta-analysis. *PLoS One,* **2017**, *12* (5): e0176827.

[6] Shao W, Sui C, Liu Z, et al. Surgical outcome of hepatocellular carcinoma patients with biliary tumor thrombi. *World Journal of Surgical Oncology,* **2011**, *9*: 2.

[7] Rammohan A, Sathyanesan J, Rajendran K, et al. Bile duct thrombi in hepatocellular carcinoma: Is aggressive surgery worthwhile? *HPB,* **2015**, *17* (6): 508–513.

[8] Shiomi M, Kamiya J, Nagino M, et al. Hepatocellular carcinoma with biliary tumor thrombi: Aggressive operative approach after appropriate preoperative management. *Surgery,* **2001**, *129* (6): 692–698.

[9] Zhou X, Wang J, Tang M, et al. Hepatocellular carcinoma with hilar bile duct tumor thrombus versus hilar Cholangiocarcinoma on enhanced computed tomography: A diagnostic challenge. *BMC Cancer,* **2020**, *20* (1): 54.

[10] Liu Q Y, Zhang W D, Chen J Y, et al. Hepatocellular carcinoma with bile duct tumor thrombus: Dynamic computed tomography findings and histopathologic correlation. *Journal of Computer Assisted Tomography,* **2011**, *35*: 187–194.

[11] Zeng H, Xu L, Wen J, et al. Hepatocellular carcinoma with bile duct tumor thrombus: a clinicopathological analysis of factors predictive of recurrence and outcome after surgery. *Medicine,* **2015**, *94* (1): e364.

[12] Liu Q, Chen J, Li H, et al. Hepatocellular carcinoma with bile duct tumor thrombi: Correlation of magnetic resonance imaging features to histopathologic manifestations. *European Journal of Radiology,* **2010**, *76* (1): 103–109.

[13] Liu Q Y, Huang S Q, Chen J Y, et al. Small hepatocellular carcinoma with bile duct tumor thrombi: CT and MRI findings. *Abdominal Imaging,* **2010**, *35* (5): 537–542.

[14] Wu J Y, Huang L M, Bai Y N, et al. Imaging features of hepatocellular car-cinoma with bile duct tumor thrombus: A multicenter study. *Frontiers in Oncology,* **2021**, *11*: 723455.

[15] Yue-Hei Ng J, Hausknecht M, Vijayanarasimhan S, et al. Beyond short snippets: Deep networks for video classification. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE, **2015**: 4694–4702.

[16] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE, **2015**: 3431–3440.

[17] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks. *Communications of the ACM,* **2020**, *63* (11): 139–144.

[18] Kingma D P, Welling M. Auto-encoding variational Bayes. **2013**. https://arxiv.org/abs/1312.6114 . Accessed February 1, 2022.

[19] Pranata Y D, Wang K C, Wang J C, et al. Deep learning and SURF for automated classification and detection of calcaneus fractures in CT images. *Computer Methods and Programs in Biomedicine,* **2019**, *171*: 27–37.

[20] Li W, Jia F, Hu Q. Automatic segmentation of liver tumor in CT images with deep convolutional neural networks. *Journal of Computer and Communications,* **2015**, *3* (11): 146–151.

[21] Basu S, Wagstyl K, Zandifar A, et al. Early prediction of alzheimer's disease progression using variational autoencoders. In: Medical Image Computing and Computer Assisted Intervention—MICCAI 2019. Cham: Springer, **2019**: 205–213.

[22] Zhang R, Tan S, Wang R, et al. Biomarker localization by combining CNN classifier and generative adversarial network. In: Medical Image Computing and Computer Assisted Intervention—MICCAI 2019. Switzerland: Springer, Cham, **2019**: 209–217.

[23] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells W, editors. Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015. Switzerland: Springer, Cham, **2015**: 234–241.

[24] Vorontsov E, Tang A, Pal C, et al. Liver lesion segmentation informed by joint liver segmentation. In: 15th International Symposium on Biomedical Imaging (ISBI 2018). Washington, USA: IEEE, **2018**: 1332–1335.

[25] Christ P F, Ettlinger F, Grün F, et al. Automatic liver and tumor segmentation of CT and MRI volumes using cascaded fully convolutional neural networks. **2017**. https://arxiv.org/abs/1702.05970. Accessed January 12, 2022.

[26] Alirr O I. Deep learning and level set approach for liver and tumor segmentation from CT scans. *Journal of Applied Clinical Medical Physics,* **2020**, *21* (10): 200–209.

[27] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **2017**, *39* (6): 1137–1149.

[28] Yang Z, Liu S, Hu H, et al. RepPoints: Point set representation for object detection. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, South Korea: IEEE, **2020**: 9656–9665.

[29] Kong T, Sun F, Liu H, et al. FoveaBox: Beyound anchor-based object detection. *IEEE Transactions on Image Processing,* **2020**, *29*: 7389–7398.

[30] Thian Y L, Li Y, Jagmohan P, et al. Convolutional neural networks for automated fracture detection and localization on wrist radiographs. *Radiology: Artificial Intelligence,* **2019**, *1*: e180001.

[31] Olczak J, Fahlberg N, Maki A, et al. Artificial intelligence for analyzing or-thopedic trauma radiographs: deep learning algorithms—Are they on par with humans for diagnosing fractures? *Acta Orthopaedica,* **2017**, *88* (6): 581–586.

[32] Kim D H, MacKinnon T. Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. *Clinical Radiology,* **2018**, *73* (5): 439–445.

[33] Boot T, Irshad H. Diagnostic assessment of deep learning algorithms for detection and segmentation of lesion in mammographic images. In: Medical Image Computing and Computer Assisted Intervention—MICCAI 2020. Cham: Springer, **2020**: 56–65.

[34] Ho D, Imai K, King G, et al. Matchit: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software,* **2011**, *42* (8): 1–28.

[35] Robin X, Turck N, Hainard A, et al. pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics,* **2011**, *12* (1): 77.

[36] Development CoreR Team. R: A Language and Environment for Statistical Computing. Vienna, Austria, **2013**.

[37] Tzutalin. Labelimg. **2015**. https://github.com/tzutalin/labelImg. Accessed March 20, 2022.

[38] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, **2016**: 779–788.

[39] Liu W, Anguelov D, Erhan D, et al. SSD: Single shot multibox detector. In: Leibe B, Matas J, Sebe N, editors. Computer Vision–ECCV 2016. Cham: Springer, **2016**: 21–37.

[40] Zhang S, Chi C, Yao Y, et al. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, **2020**: 9756–9765.

[41] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: Common objects in context. In: Fleet D, Pajdla T, Schiele B, editors. Computer Vision–ECCV 2014. Cham: Springer, **2014**: 740–755.

[42] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, **2016**: 770–778.

[43] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, **2017**: 936–944.

[44] Pang J, Chen K, Shi J, et al. Libra R-CNN: Towards balanced learning for object detection. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, **2019**: 821–830.

[45] Liu S, Qi L, Qin H, et al. Path aggregation network for instance segmentation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, **2018**: 8759–8768.

[46] Breiman L. Random forests. *Machine Learning,* **2001**, *45*: 5–32.