Radiology: Artificial Intelligence

ORIGINAL RESEARCH

Deep Learning for the Detection, Localization, and Characterization of Focal Liver Lesions on Abdominal US Images

*Hind Dadoun, MS** • *Anne-Laure Rousseau, MD** • *Eric de Kerviler, PhD* • *Jean-Michel Correas, PhD* • *Anne-Marie Tissier, MD* • *Fanny Joujou* • *Sylvain Bodard, MD* • *Kemel Khezzane* • *Constance de Margerie-Mellon, MD* • *Hervé Delingette, PhD* • *Nicholas Ayache, PhD*

From the Université Côte d'Azur, Inria, Epione Team, Sophia Antipolis, 2004 Route des Lucioles, 06902 Valbonne, France (H. Dadoun, H. Delingette, N.A.); Department of Vascular Surgery, Georges Pompidou European Hospital APHP, Université de Paris, Paris, France (A.L.R.); NHance.ngo, Saint Germain en Laye, France (A.L.R.); Department of Radiology, Hôpital Saint Louis APHP, Université de Paris, Paris, France (E.d.K., F.J., K.K., C.d.M.M.); and Department of Adult Radiology, Université de Paris and Université de l'Hôpital Necker, Paris, France (J.M.C., A.M.T., S.B.). Received April 23, 2021; revision requested May 26; revision received February 8, 2022; accepted February 16. **Address correspondence to** H. Dadoun (e-mail: *hind.dadoun@inria.fr*).

Supported in part by the French government through the 3IA Côte d'Azur Investments in the Future project managed by the National Research Agency (ANR) (reference no. ANR-19-P3IA-0002).

* H. Dadoun and A.L.R. contributed equally to this work.

Conflicts of interest are listed at the end of this article.

Radiology: Artificial Intelligence 2022; 4(3):e210110 • https://doi.org/10.1148/ryai.210110 • Content codes: AI GI OI

Purpose: To train and assess the performance of a deep learning–based network designed to detect, localize, and characterize focal liver lesions (FLLs) in the liver parenchyma on abdominal US images.

Materials and Methods: In this retrospective, multicenter, institutional review board–approved study, two object detectors, Faster region-based convolutional neural network (Faster R-CNN) and Detection Transformer (DETR), were fine-tuned on a dataset of 1026 patients (n = 2551 B-mode abdominal US images obtained between 2014 and 2018). Performance of the networks was analyzed on a test set of 48 additional patients (n = 155 B-mode abdominal US images obtained in 2019) and compared with the performance of three caregivers (one nonexpert and two experts) blinded to the clinical history. The sign test was used to compare accuracy, specificity, sensitivity, and positive predictive value among all raters.

Results: DETR achieved a specificity of 90% (95% CI: 75, 100) and a sensitivity of 97% (95% CI: 97, 97) for the detection of FLLs. The performance of DETR met or exceeded that of the three caregivers for this task. DETR correctly localized 80% of the lesions, and it achieved a specificity of 81% (95% CI: 67, 91) and a sensitivity of 82% (95% CI: 62, 100) for FLL characterization (benign vs malignant) among lesions localized by all raters. The performance of DETR met or exceeded that of two experts and Faster R-CNN for these tasks.

Conclusion: DETR demonstrated high specificity for detection, localization, and characterization of FLLs on abdominal US images.

Supplemental material is available for this article.

The accurate detection and assessment of focal liver lesions (FLLs) is a critical public health issue because of the increased incidence of primary hepatic malignant lesions. Liver cancer is the third leading cause of cancer-related deaths worldwide (1), and hepatocellular carcinoma is the primary type affecting adults (2). Typically, these lesions are discovered in patients with stage 3 liver failure or with other cancers, such as colorectal cancer. These lesions are also detected incidentally during abdominal imaging studies (3). Noncontrast US is among the most commonly used modalities for investigating FLLs during the screening stage for patients at high risk of malignancy. However, performing US and interpreting US images are difficult and examiner-dependent tasks, and the number of trained operators is limited (3). In resource-limited countries in particular, health care providers identify lack of training as the main limitation to the use of US (4). Use of a computer-aided tool could facilitate the detection of more early-stage malignant lesions, with the potential for increased differential diagnosis as well as more efficient and cost-effective treatment (5,6). Such a tool could assist nonexpert caregivers in performing an adequate assessment of the liver.

Previous studies have shown promising results with machine learning methods for the diagnosis of FLLs on US images (7–10). A previous study classified malignant and benign lesions in 95 3-minute cine clips using automatically extracted B-mode and contrast-specific features on a support vector machine classifier (7). That study showed results comparable to those of experts with more than 15 years of experience. Another study showed improved performance with the use of sparse representation-based feature extraction methods on multimodal US images in 111 patients (8). A pretrained convolutional neural network with an added attention module of the region of interest was evaluated in 367 two-dimensional B-mode US images, but the study

Abbreviations

AUC = area under the receiver operating characteristic curve, DETR = Detection Transformer, DICOM = Digital Imaging and Communications in Medicine, FLL = focal liver lesion, FN = false-negative finding, FP = false-positive finding, MCC = Matthews correlation coefficient, PACS = picture archiving and communications system, PPV = positive predictive value, R-CNN = region-based convolutional neural network, TN = true-negative finding, TP = true-positive finding

Summary

Two pretrained deep neural networks (Detection Transformer and Faster region-based convolutional neural network) fine-tuned on a dataset of abdominal US images showed satisfactory performance for the detection, localization, and characterization of focal liver lesions.

Key Points

- The vision transformer network Detection Transformer (DETR) showed higher performance for all tasks, compared with the Faster region-based convolutional neural network.
- For the detection of focal liver lesions (FLLs) in the liver parenchyma, DETR matched or exceeded the performance of two experts, achieving a specificity of 90% (95% CI: 75, 100) and a sensitivity of 97% (95% CI: 97, 97).
- For the localization of FLLs, the performance of DETR was comparable to that of two experts, achieving a positive predictive value of 77% (95% CI: 70, 84) and a sensitivity of 84% (95% CI: 77, 89).
- For the characterization of FLLs (benign vs malignant), the performance of DETR was better than that of all other raters, achieving a specificity of 81% (95% CI: 67, 91) and a sensitivity of 82% (95% CI: 62, 100).

Keywords

Computer-aided Diagnosis (CAD), Ultrasound, Abdomen/GI, Liver, Tissue Characterization, Supervised Learning, Transfer Learning, Convolutional Neural Network (CNN)

assumed that only one subtype of lesions could be present in the liver (10). The network achieved an area under the receiver operating characteristic curve (AUC) score of 0.94 for FLL detection and 0.916 for FLL characterization over threefold cross-validations. Finally, Yang et al (9) constructed a large multicentric dataset of 24 343 B-mode US images along with radiomics signatures derived from FLLs and liver, with ultrasonic features including posterior acoustic enhancement, echogenicity, shape of the fan, and clinical parameters. Diagnostic performances were verified using external validation and were compared with that of contrast-enhanced CT and MRI, as well as that of radiologists, with an AUC of 0.924 for classification of malignant versus benign FLLs.

Most of these previous studies used small datasets, however, with a large class imbalance between benign and malignant lesions (80% and 20%, respectively, in the study by Yang et al [9], for example). To the best of our knowledge, none of the methods mentioned above localized lesions in the liver or assigned a specific characterization for each lesion. Our study explores the use of deep learning–based networks for the detection, localization, and characterization of FLLs on noncontrast US images to assist nonexpert caregivers in screening patients at high risk of malignancy.

Materials and Methods

Study Design

Institutional review board approval (no. IRB00011591) was obtained for this retrospective, multicenter study, and informed consent was waived. Data for this pilot study were obtained in collaboration with a clinical data warehouse registered under the number 1980120. US images were extracted from the picture archiving and communication system (PACS) of two university hospitals, center 1 (Necker Hospital, Paris) and center 2 (Saint Louis Hospital, Paris). Only adult patients aged 18 years and older were selected. Patients with liver parenchyma containing lesions were included if they met the following criteria: Lesions were visible on the US images as decided unanimously by an adjudication panel, patients had not previously received local therapy, and a definite pathologic diagnosis of lesions was obtained (cyst, angioma, focal nodular hyperplasia, adenoma, metastasis, or hepatocellular carcinoma). Patients without lesions in the liver parenchyma were selected in case of a definite absence of pathologic diagnosis. All images from abdominal US performed at centers 1 and 2 between 2014 and 2019 were selected. For the training and development set, images from examinations performed between 2014 and 2018 were selected. For the test set, images from examinations performed in 2019 were selected, provided that the corresponding patients had not undergone any examination between 2014 and 2018. Extremely magnified images and images obtained using degraded mode simultaneously with contrast-enhanced US images were excluded.

Data Acquisition

Data collected retrospectively in center 1 consisted exclusively of multivendor US images with FLLs (four vendors). Data collected retrospectively in center 2 consisted of multivendor images of healthy liver parenchyma and FLLs (two vendors).

Data Preprocessing

The Digital Imaging and Communications in Medicine (DICOM) red-green-blue images extracted from the PACS were converted to JPEG format. They were processed using a deidentification tool to remove identifying data and metadata. This tool uses the DICOM US attributes to remove the upper band of the image before converting it to JPEG format. A parametric method was used to detect the US fan area and remove annotations and other machine-dependent characteristics outside this region (11). In addition, biometric measurements present inside the region were removed using existing inpainting methods (Fig 1) (11). This enables networks to be trained on raw data with no manually added annotations by examiners and therefore may be more suitable for real-time use by nonexpert examiners.

Determination of the Ground Truth

For each center, the diagnosis associated with each US image was assigned by two radiologists (J.M.C., Pierre Bourrier, MD), each with more than 15 years of experience in US image

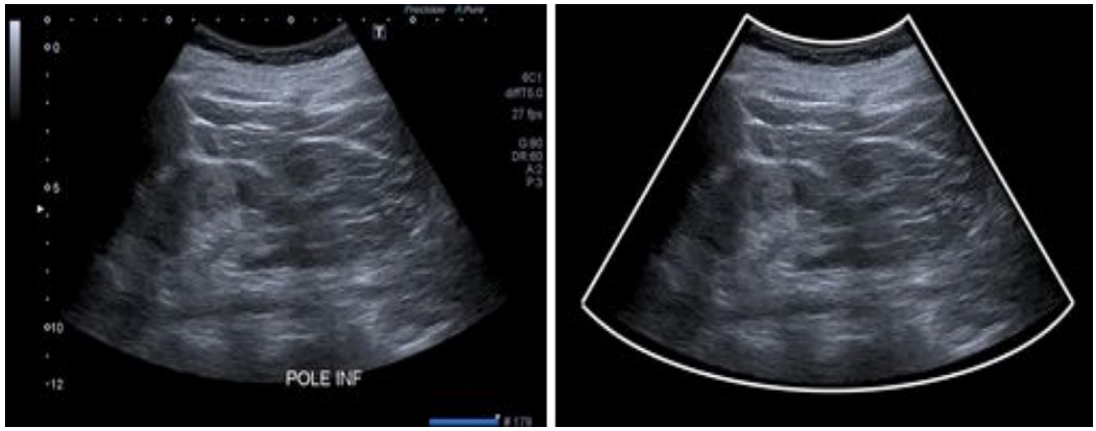

Figure 1: Left: Original image. The US fan area is limited to a conic section, and several text and graphic elements are present. Right: Image after preprocessing. The white lines delimiting the US fan area are automatically detected, and all graphic and text elements have been removed and replaced by a plausible intensity value. (Reprinted, with permission, from reference 11.)

interpretation, by cross-referencing with other imaging modalities (contrast-enhanced US, CT, or MRI, or biopsy, when available) and clinical files. The final diagnosis was used for the characterization task (liver parenchyma and characterization of FLLs) and did not include the number of lesions in the US image or their localization. To determine the ground truth boxes for the localization task (ie, boxes around the liver and FLLs), an adjudication panel was used as an external standard of reference. The panel consisted of four physicians (including A.L.R., C.d.M.M., and Mariama Bah, MD) who either are radiologists or hold a national diploma in US imaging and four sonographers (including K.K.) who hold a national diploma in US imaging from six different health institutions and who each have more than 3 years of experience. The annotators, who worked on a custom annotation platform, were first asked to localize the liver and classify it as a "homogeneous liver" (ie, without lesions) or as a "liver with lesion(s)" with respect to the final diagnosis associated with the image. The selection of liver with lesion(s) led to the second task (ie, localization of lesions). Each lesion was required to be classified in accordance with the final diagnosis, with two possibilities: benign lesion(s) or malignant lesion(s). Benign lesions were further subclassified as cyst, angioma, focal nodular hyperplasia, or adenoma. Malignant lesions were further subclassified as metastasis or hepatocellular carcinoma. Examples of such annotations are shown in Figure 2. Each image was annotated by two experts and at least once by a physician. During this phase, in case of a disagreement between two annotators for the localization task, four additional annotators analyzed the questionable image. If annotation of the image was not unanimous between the additional annotators, the image was excluded from the study.

Data Partitions

A total of 1026 patients (n = 2551 images) met the inclusion criteria for the training and development set. This set was randomly split into two subsets containing approximately the same proportions for each class, with 80% and 20%, respectively, for training and development. The objective of the development set is to choose the network settings that achieve the best performance on images the network has never seen. A total of 48 additional patients (n = 155 images) met the inclusion criteria for the test set. All images and clinical reports were deidentified within the centers, and no demographics information from the study population was retained.

Models.— Two object detection networks were used. Detection Transformer (DETR) is an end-to-end object detection vision transformer network (12) and is more suitable for real-time application (13). Faster region-based convolutional neural network (Faster R-CNN) is a two-stage object detection network that showed more robust performance on natural image datasets (14). Figure 3 compares the workflows of both networks. The networks were trained for 100 epochs, after which the validation loss stopped decreasing and the overfitting occurred. Stochastic gradient descent with Nesterov momentum (15) was used for optimization (Appendix E1 [supplement]), with a learning rate of 0.0048, a batch size of 4, a momentum of 0.9, and a weight decay of 0.0001. For the first epoch, a warm-up schedule was applied (16). For the rest of the training, the learning rate of each parameter group was decreased by 0.9 once the number of epochs reached 75 and 90.

Data augmentation.— To improve the performance of the networks, a data augmentation strategy based on the Google Brain team bounding box augmentation policies was applied (17). The data augmentation scheme is based on a learned set of specific subpolicies that were proven to improve generalization performance (Appendix E1 [supplement]). These subpolicies consist of a list of intensity, geometric, and bounding box operations (rotation, cutout, sharpness, translation, contrast, brightness, solarization, shear) applied to the image or to a specific object inside the bounding box. During training, one of those policies is randomly selected and then applied to the current image as reported by Zoph et al (17).

Model evaluation.—Performances of all raters (expert and nonexpert caregivers, networks) were compared against the ground truth. The nonexpert, an emergency physician, had

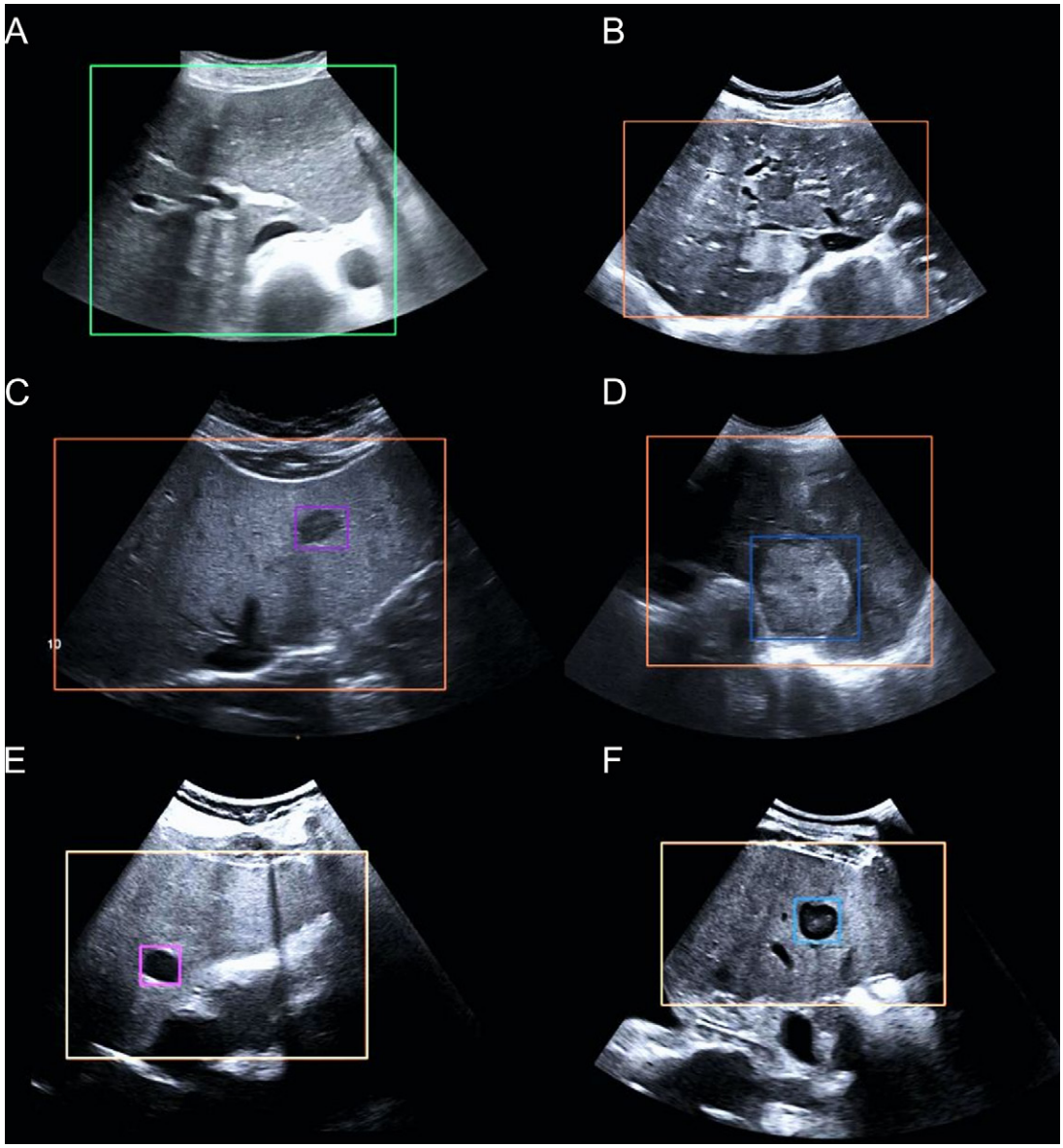

Figure 2: **(A)** A liver without lesions (green box) and **(B)** a liver with lesions (orange box). **(C)** A benign lesion (focal nodular hyperplasia [small purple box]) and **(D)** a malignant lesion (hepatocellular carcinoma [small blue box]). **(C, D)** In this pairing, the benign and malignant lesions have different textures and sizes. **(E)** A benign lesion (cyst [purple box]) with a circular shape and dark pixel intensities. **(F)** A malignant lesion (metastasis [blue box]) with similar characteristics. These images illustrate the difficulty of distinguishing malignant from benign lesions.

5 years of experience with US imaging, both acquiring and interpreting US images. The experts, a radiologist (expert 1, S.B.) and an advanced practice sonographer (expert 2, F.J.), each had a national degree in US imaging with 9 and 8 years of experience, respectively. All three caregivers and the networks were blinded to clinical history and reports. Performances were evaluated in detail with regard to the capability to detect liver parenchyma with FLLs, localize FLLs using bounding boxes, regardless of the label assigned to them, characterize the FLLs correctly localized by all raters as either benign or malignant, and characterize FLLs into benign and malignant subcategories. For each of these tasks and for all raters, the positive predictive value (PPV) (Eq [1]) or the sensitivity (Eq [2]) and the specificity (Eq [3]) are reported. For the localization task, the F1 score (Eq [5]) is reported as well. For binary classification (ie, tasks 1 and 2), the accuracy (Eq [4]) and the Matthews correlation coefficient (MCC) (Eq [6]) metrics are reported as well. For multiclass classification, the macro specificity, sensitivity, and accuracy are used instead. The formulas of all metrics are defined as follows:

$$PPV = \frac{TP}{TP + FP} \quad (1),$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (2),$$

$$Specificity = \frac{TN}{TN + FP} \quad (3),$$

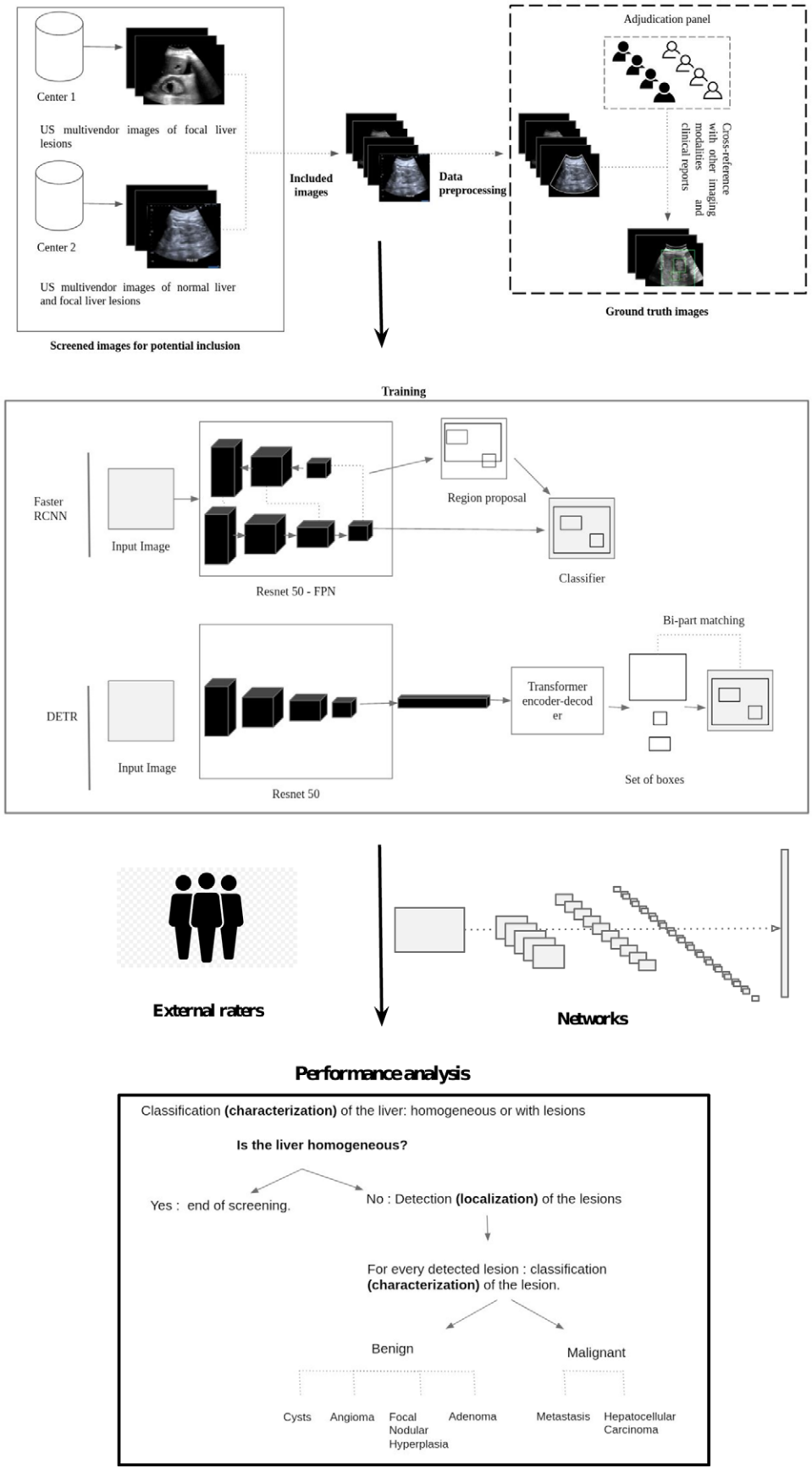

Figure 3: Study workflow. DETR = Detection Transformer, FPN = feature pyramid network, R-CNN = region-based convolutional neural network.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (4),$$

$$F1\ score = 2 \times \frac{PPV \times Sensitivity}{PPV + Sensitivity} \quad (5),$$

and

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (6),$$

where TP is true-positive finding, FP is false-positive finding, TN is true-negative finding, and FN is false-negative finding.

To report all of these metrics, a class-specific threshold was set to select the probabilities output by the networks; the threshold was defined as the value that maximizes the F1 score of each class during the validation step. In this setting, a prediction (box + label) is considered a TP if it has an intersection over union score with a ground truth box greater than 0.3 and the same label as the ground truth box and an FP otherwise. Ground truth boxes with no intersections with predicted boxes are FNs.

Statistical Analysis

Because images in the same patient are not independent, bootstrapping experiments were used, as described by Linnet (18). The test set was resampled to contain only one randomly selected image per patient, and the inference was repeated 1000 times. At each iteration, the resampled test set had the same size as the patient population size (48 patients, 48 images). For accuracy, specificity, sensitivity, PPV, F1 score, and MCC values, we report the mean value across the observations, as well as 95% CIs computed using the 2.5 and 97.5 percentiles of the ranked observations. The sign test was used to report the significant results for all metrics. Given multiple comparisons, the Bonferroni correction method (the six classes are cyst, angioma, focal nodular hyperplasia, adenoma, metastasis, and hepatocellular carcinoma) was used to adjust *P* values (19), and a *P* value less than .05 was considered statistically significant.

Data Availability

All code used for this study is publicly available at *https://gitlab.inria.fr/hdadoun/focal-liver-lesions-us.*

Results

Detection of Lesions in the Liver Parenchyma

Table 1 shows performance in detecting liver parenchyma with or without FLLs in terms of accuracy, specificity, sensitivity, and MCC. The nonexpert caregiver achieved an accuracy of 78% (95% CI: 70, 86), expert 1 achieved a slightly higher accuracy of 80% (95% CI: 74, 86), and expert 2 had the best accuracy score overall, achieving 99% (95% CI: 98, 100). Faster R-CNN achieved an accuracy of 93% (95% CI: 88, 98), and DETR achieved an accuracy of 96% (95% CI: 92, 98). Overall, the networks were consistent in their predictions, with an MCC of 81% (95% CI: 66, 95) for Faster R-CNN and 88% (95% CI: 77, 95) for DETR. Overall, we found no difference between expert 2 and DETR in terms of accuracy, specificity, and sensitivity. In addition, both expert 2 and DETR showed higher accuracy and sensitivity compared with the nonexpert caregiver and expert 1 ($P < .001$).

Localization of Lesions

Table 2 shows performance in the localization of FLLs in terms of PPV, sensitivity, and F1 score. Because there are no TNs for this task, accuracy and specificity are not reported. Expert 2 and the networks achieved comparable results for this task, with a mean PPV of 76% (95% CI: 72, 79) and mean sensitivity of 78% (95% CI: 74, 83). Expert 1 achieved a PPV of 73% (95% CI: 63, 84) and sensitivity of 69% (95% CI: 62, 77). This result was consistent with the detection accuracy of liver parenchyma with FLLs, wherein the performance of expert 1 was lower than that of expert 2 and the networks.

Characterization of Lesions

To report the characterization performance, a subset of lesions that were detected and localized both by the networks and the experts was selected. Table 3 shows performance in the characterization of FLLs in terms of accuracy, specificity, sensitivity, and MCC. DETR achieved the highest accuracy, 81% (95% CI: 68, 94), compared with 61% (95% CI: 50, 71) for the best-performing expert (expert 2). The accuracy achieved by expert 2 was significantly different from that of the networks ($P < .001$) but not from that of expert 1 ($P = .25$), while the accuracy achieved by DETR was significantly different from both experts but not from Faster R-CNN ($P = .18$). Expert 2 achieved the highest sensitivity among all raters (87% [95% CI: 73, 100]) and the lowest specificity (33% [95% CI: 25, 44]). DETR achieved the second highest sensitivity (82% [95% CI: 62, 100]) and the highest specificity (81% [95% CI: 67, 91]) among all raters. Both networks also achieved a higher MCC with the true labels compared to experts: 63% (95% CI: 37, 88) for DETR compared to 25% (95% CI: 4, 50) for expert 2.

Subcharacterization

The subcharacterization performance on the subset of lesions that were detected and localized both by the networks and the experts is also reported in Table 3. DETR achieved the highest accuracy, with 76% (95% CI: 62, 91) compared with 52% (95% CI: 36, 70) for the best-performing expert (expert 1). All raters (experts and networks) showed a comparable specificity for the subcharacterization task, with a mean value of 91% ± 3 (SD) and a mean sensitivity of 59% ± 10. Pairwise comparisons for this task did not show significant differences between all raters (experts and networks).

In summary, for the detection of lesions in the liver parenchyma, no significant difference was found between the best-performing network (DETR) and the best-performing expert (expert 2) in the small test set; for the localization of FLLs, all raters achieved comparable results; for the characterization of FLLs detected and localized by all raters, both networks

Table 1: Detection of Liver Parenchyma with or without Focal Liver Lesions: Performance by Rater

Rater	Accuracy (%)	Specificity (%)	Sensitivity (%)	Matthews Correlation Coefficient (%)
Nonexpert	78*† (70, 86)	80 (58, 100)	77*† (68, 84)	0.51*† (0.31, 0.68)
Expert 1	80*† (74, 86)	74 (50, 92)	82*† (76, 87)	0.51*† (0.31, 0.68)
Expert 2	99 (98, 100)	98 (92, 100)	100 (100, 100)	0.98 (0.95, 1.00)
Faster R-CNN	93 (88, 98)	71 (50, 92)	100 (100, 100)	0.81 (0.66, 0.95)
DETR	96 (92, 98)	90 (75, 100)	97 (97, 97)	0.88 (0.77, 0.95)

Note.—Detection of liver parenchyma with focal liver lesions (37 patients, 100 images) or without focal liver lesion (11 patients, 55 images). Reported results were computed on the bootstrapped test set. Each subset contained only one image per patient. Data in parentheses are 95% CIs. DETR = detection transformer network, R-CNN = region-based convolutional neural network.
* $P \leq .05$ in comparison with DETR.
† $P \leq .05$ in comparison with expert 2.

Table 2: Localization of Focal Liver Lesions: Performance by Rater

Rater	IoU	PPV (%)	Sensitivity (%)	F1 Score
Expert 1	0.72 ± 0.14	73 (63, 84)	69 (62, 77)	71 (64, 79)
Expert 2	0.68 ± 0.12	80 (71, 89)	78 (71, 85)	79 (72, 86)
Faster R-CNN	0.71 ± 0.10	72 (66, 79)	73 (66, 80)	73 (67, 78)
DETR	0.69 ± 0.12	77 (70, 84)	84 (77, 89)	80 (74, 85)

Note.—Analysis based on 37 patients (214 lesions). Reported results were computed on the bootstrapped test set. Each subset contained only one image per patient. Intersection over union (IoU) is shown as the value ± standard deviation. Data in parentheses are 95% CIs. DETR = detection transformer network, PPV = positive predictive value, R-CNN = region-based convolutional neural network.

Table 3: Characterization and Subcharacterization of Focal Liver Lesions: Performance by Rater

Rater for Each Subset	Accuracy (%)	Specificity (%)	Sensitivity (%)	Matthews Correlation Coefficient (%)
Characterization of FLLs as benign or malignant				
Expert 1	59* (47, 70)	79† (62, 92)	40*† (22, 55)	0.20*† (–0.04, 0.44)
Expert 2	61* (50, 71)	33* (25, 44)	87 (73, 100)	0.25* (0.04, 0.50)
Faster R-CNN	76† (64, 90)	72*† (55, 90)	81 (60, 100)	0.53† (0.27, 0.81)
DETR	81† (68, 94)	81 (67, 91)	82 (62, 100)	0.63† (0.37, 0.88)
Subcharacterization of FLLs into six classes				
Expert 1	52‡ (36, 70)	91§ (88, 94)	48§ (38, 61)	NA
Expert 2	50‡ (36, 67)	87§ (84, 92)	54§ (40, 68)	NA
Faster R-CNN	72‡ (54, 91)	0.93§ (88, 98)	70§ (45, 93)	NA
DETR	76‡ (62, 91)	94§ (90, 98)	65§ (50, 80)	NA

Note.—Results reported on the subset of lesions that were localized by all raters on the bootstrapped test set. Each subset contained only one image per patient. Characterization of FLLs as benign or malignant based on 24 patients (48 benign FLLs) and 13 patients (74 malignant FLLs), respectively. Subcharacterization of FLLs into six classes (cyst, angioma, focal nodular hyperplasia, adenoma, metastasis, and hepatocellular carcinoma) based on 37 patients (122 lesions). Data in parentheses are 95% CIs. DETR = detection transformer network, FLL = focal liver lesion, NA = not applicable, R-CNN = region-based convolutional neural network.
* $P < .05$ in comparison with DETR.
† $P < .05$ in comparison with expert 2.
‡ Overall accuracy.
§ Macro average (average accuracy at the class level).

achieved higher performance compared with both experts; and for the subcharacterization of benign and malignant FLLs, pairwise comparisons between raters were not significantly different. More details on the discrimination performance of the networks are reported in Appendix E1 (supplement).

Discussion

In this study, we present a framework for the detection, localization, and characterization of FLLs on B-mode US images. First, we investigated the accuracy of detecting FLLs. Second, we included the FLL localization task with the aim of drawing the examiner's attention to a region of interest. Finally, after FLL localization, we characterized the malignancy of each lesion in the image. The network DETR achieved a specificity of 90% (95% CI: 75, 100) and a sensitivity of 97% (95% CI: 97, 97) for the detection of lesions in the liver parenchyma on the test set. It correctly localized 80% of the lesions, and among the lesions correctly localized by all raters (experts and networks), DETR achieved a specificity of 81% (95% CI: 67, 91) and a sensitivity of 82% (95% CI: 62, 100) for FLL characterization (benign vs malignant).

A previous study showed high performance of a deep learning model for the detection of FLL (mean AUC score, 0.935) on B-mode US images, using repeated random cross-validation with approximately the same proportion (70%) of liver parenchyma with lesions as in our study (10). Our contribution for this task lies in the comparison of screening performance with that of nonexpert and expert caregivers. Our analysis led to the conclusion that the performance of the DETR network matches the performance of experts for this task.

To our knowledge, this is the first attempt to automatically localize FLLs on B-mode US images. This step enables the characterization of each lesion individually in addition to assigning global class to the liver (with or without lesions). Because our main objective is to detect FLLs in the areas of interest, we chose an intersection over union threshold of 0.3. Thus, slightly delocalized correct detections were still labeled TPs. We demonstrated that both networks met the performances of experts for this task.

Previous studies have investigated the use of deep learning networks for the characterization of lesions in US imaging (7–10,20). Focusing only on the characterization task implies that either the whole liver is classified as benign or malignant, or that the lesions are first localized by an expert and then classified. This is particularly relevant in the diagnosis task for FLLs, in which additional external clinical information or features from multiphase imaging with the administration of contrast material (ie, contrast-enhanced CT or MRI or contrast-enhanced US) are often used to reach a final diagnosis. Yang et al (9) reported diagnostic performance comparable to that of contrast-enhanced CT using solely B-mode US data, segmented regions of interest, and clinical information. In the screening of FLLs, it is particularly important to localize the lesions during the examination to draw the attention of the examiner when needed and enable the examiner to make a decision regarding follow-up on the basis of their observations and the characterization predicted by the network. Both of our trained networks showed higher performance for this task compared with the experts.

Our performance results for the subcharacterization of benign and malignant lesions are lower than those previously reported for contrast-enhanced US (21,22) and show that the subcharacterization task is much more challenging in noncontrast US, which correlates well with the literature (23). A previous study showed high performance for the subcharacterization task on B-mode US images at an image level (as opposed to the lesion level) (10); although this may be confusing in practice, it offers insight into how performance can be improved. We believe these results could be further improved by adding more weight to the liver features when classifying the lesions, because many experts look at patterns in the liver when making a primary diagnosis on the basis of US images (24). Performance may improve by using US video clips rather than static US images to have more context when making an interpretation. Another interesting approach would be to investigate the use of self-supervised and semisupervised learning, as we have access to thousands of unlabeled images. Previous studies have shown that such approaches could close the gap between strong and weak supervision, for example, in digital pathologic analysis (25).

Our study had several limitations. First, the number of images in the test set was limited, and the study was retrospective. To ensure the generalizability of the network, further investigations with a larger multicentric and multivendor prospective cohort are needed. Second, the reference standard used for localization was based on unanimous annotation by an adjudication panel to avoid interexpert variability. While this enables more robust learning, it makes it difficult to construct a larger dataset and creates bias. A less stringent criterion would be a majority vote, which could be used instead during annotation of the test set. In addition, excluding images of poor quality or those deemed to be uninterpretable likely creates bias and can potentially affect future predictions on images of the same type. Given that US images are operator dependent, quality may vary considerably between operators, as well as during the same examination (eg, the operator may capture images in several planes to best describe the abnormality, some of which become difficult to interpret a posteriori without context). One way to address this issue is to use image quality assessment networks (with access to references that are considered to be of good-quality and poor-quality images) to filter images before applying the FLL screening network. Future studies are needed to analyze how these networks can be included in the method. Third, the experts and networks were blinded to clinical information that is normally available and that can aid in the screening of FLLs (eg, at-risk patients with known cirrhosis or chronic hepatitis B virus infection). The decision to remove clinical information from our study was made to allow analysis of only the information that can be extracted from US images. Finally, given the size of our test set and the imbalance in subcategories, we were not able to complete a comparative analysis between our method and the experts for this task. Nonetheless, it is important to note that the network performed poorly compared with experts for the detection of cysts. We believe this is owing to the lack of training examples for this class, as well as the fact that, for a network that has only seen FLLs, cysts can be confused with the portal vein or the inferior vena cava in the sagittal plane. A potential

improvement to avoid this confusion would be to add annotated examples of the different anatomic structures visible in abdominal US during training.

This study provides new information on the use of deep learning networks for the detection, localization, and characterization of FLLs on B-mode US images. Experiments on a test set and comparison with experts showed that the vision transformer network DETR can aid the examiner by helping focus the examiner's visual attention on areas of interest. This can also provide insights to nonexpert caregivers and facilitate screening of FLLs, with the potential to increase early detection of hepatocellular carcinoma.

Acknowledgments: The authors are grateful to the OPAL infrastructure from Université Côte d'Azur, the radiology team from Saint Louis Hospital of Greater Paris University Hospitals, the radiology team from Necker Hospital, and the French Health Data Hub for providing resources and support. We thank Deepomatic for making available an annotation platform to the NHance project. We thank Guillaume Oules, MS, and Alexandre Dubreucq, MS, both of École Polytechnique, for their help in the design of the NHance project. The authors are grateful for the help provided by the team of the radiology department at Saint Louis Hospital: Veronique Caloc, Claudine Singh, Cedric de Bazelaire, MD, PhD, Damien Bouda, MD, and Pierre Bourrier, MD. The authors greatly appreciated the help that Mariama Bah, MD, and Gregory Khelifi, MD, provided. This work was supported by the clinical research unit of Saint Louis Hospital: Jérome Lambert, MD, PhD, and Claire Montlahuc, MD, PhD.

Author contributions: Guarantors of integrity of entire study, **H. Dadoun, A.L.R., F.J., S.B., K.K.**; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, **H. Dadoun, A.L.R., S.B., H. Delingette, N.A.**; clinical studies, **A.L.R., A.M.T., S.B.**; experimental studies, **H. Dadoun, A.L.R., F.J., S.B., K.K., N.A.**; statistical analysis, **H. Dadoun, S.B.**; and manuscript editing, **H. Dadoun, E.d.K., J.M.C., S.B., C.d.M.M., H. Delingette, N.A.**

Disclosures of conflicts of interest: **H. Dadoun** No relevant relationships. **A.L.R.** No relevant relationships. **E.d.K.** No relevant relationships. **J.M.C.** No relevant relationships. **A.M.T.** No relevant relationships. **F.J.** No relevant relationships. **S.B.** No relevant relationships. **K.K.** No relevant relationships. **C.d.M.M.** Payment or honoraria from Bracco for lectures, presentations, speakers bureaus, manuscript writing, or educational events. **H. Delingette** No relevant relationships. **N.A.** No relevant relationships.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018;68(6):394–424. [Published correction appears in CA Cancer J Clin 2020;70(4):313.]
2. Craig AJ, von Felden J, Garcia-Lezana T, Sarcognato S, Villanueva A. Tumour evolution in hepatocellular carcinoma. Nat Rev Gastroenterol Hepatol 2020;17(3):139–152.
3. Marrero JA, Ahn J, Rajender Reddy K; American College of Gastroenterology. ACG clinical guideline: the diagnosis and management of focal liver lesions. Am J Gastroenterol 2014;109(9):1328–1347; quiz 1348.
4. Shah S, Bellows BA, Adedipe AA, Totten JE, Backlund BH, Sajed D. Perceived barriers in the use of ultrasound in developing countries. Crit Ultrasound J 2015;7(1):28.
5. Cadier B, Bulsei J, Nahon P, et al. Early detection and curative treatment of hepatocellular carcinoma: A cost-effectiveness analysis in France and in the United States. Hepatology 2017;65(4):1237–1248.
6. Trinchet JC. Hepatocellular carcinoma: increasing incidence and optimized management [in French]. Gastroenterol Clin Biol 2009;33(8-9):830–839.
7. Ta CN, Kono Y, Eghtedari M, et al. Focal Liver Lesions: Computer-aided Diagnosis by Using Contrast-enhanced US Cine Recordings. Radiology 2018;286(3):1062–1071.
8. Yao Z, Dong Y, Wu G, et al. Preoperative diagnosis and prediction of hepatocellular carcinoma: Radiomics analysis based on multi-modal ultrasound images. BMC Cancer 2018;18(1):1089.
9. Yang Q, Wei J, Hao X, et al. Improving B-mode ultrasound diagnostic performance for focal liver lesions using deep learning: A multicentre study. EBioMedicine 2020;56:102777.
10. Schmauch B, Herent P, Jehanno P, et al. Diagnosis of focal liver lesions from ultrasound using deep learning. Diagn Interv Imaging 2019;100(4):227–233.
11. Dadoun H, Delingette H, Rousseau AL, de Kerviler E, Ayache N. Combining Bayesian And Deep Learning Methods For The Delineation Of The Fan In Ultrasound Images. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), 2021; 743–747.
12. Dosovitskiy A, Beyer L, Alexander Kolesnikov, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Poster presented at: International Conference on Learning Representations; May 4, 2021; Vienna, Austria. https://iclr.cc/virtual/2021/poster/3013. Accessed November 22, 2021.
13. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-End Object Detection with Transformers. arXiv 2005.12872 [preprint] https://arxiv.org/abs/2005.12872. Posted May 26, 2020. Accessed April 13, 2021.
14. Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Trans Pattern Anal Mach Intell 2017;39(6):1137–1149.
15. Goyal P, Dollár P, Girshick R, et al. Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour. arXiv 1706.02677 [preprint] https://arxiv.org/abs/1706.02677. Posted June 8, 2017. Accessed April 13, 2021.
16. Sutskever I, Martens J, Dahl G, Hinton G. On the importance of initialization and momentum in deep learning. Proceedings of the 30th International Conference on Machine Learning. PMLR 2013;28(3):1139–1147.
17. Zoph B, Cubuk ED, Ghiasi G, Lin TY, Shlens J, Le QV. Learning Data Augmentation Strategies for Object Detection. arXiv 1906.11172 [preprint] https://arxiv.org/abs/1906.11172. Posted June 26, 2019. Accessed April 13, 2021.
18. Linnet K. Nonparametric estimation of reference intervals by simple and bootstrap-based procedures. Clin Chem 2000;46(6 Pt 1):867–869.
19. Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. BMJ 1995;310(6973):170.
20. Hassan TM, Elmogy M, Sallam ES. Diagnosis of Focal Liver Diseases Based on Deep Learning Technique for Ultrasound Images. Arab J Sci Eng 2017;42(8):3127–3140.
21. Park H, Park JY, Kim DY, et al. Characterization of focal liver masses using acoustic radiation force impulse elastography. World J Gastroenterol 2013;19(2):219–226.
22. 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(3):887–896.
23. Hanna RF, Miloushev VZ, Tang A, et al. Comparative 13-year meta-analysis of the sensitivity and positive predictive value of ultrasound, CT, and MRI for detecting hepatocellular carcinoma. Abdom Radiol (NY) 2016;41(1):71–90.
24. Harvey CJ, Albrecht T. Ultrasound of focal liver lesions. Eur Radiol 2001;11(9):1578–1593.
25. Dehaene O, Camara A, Moindrot O, de Lavergne A, Courtiol P. Self-Supervision Closes the Gap Between Weak and Strong Supervision in Histology. arXiv 2012.03583 [preprint] https://arxiv.org/abs/2012.03583. Posted December 7, 2020. Accessed April 13, 2021.