Deep Learning with Convolutional Neural Network for Differentiation of Liver Masses at Dynamic Contrast-enhanced CT: A Preliminary Study[1]

Koichiro Yasaka, MD, PhD
Hiroyuki Akai, MD, PhD
Osamu Abe, MD, PhD
Shigeru Kiryu, MD, PhD

Purpose: To investigate diagnostic performance by using a deep learning method with a convolutional neural network (CNN) for the differentiation of liver masses at dynamic contrast agent–enhanced computed tomography (CT).

Materials and Methods: This clinical retrospective study used CT image sets of liver masses over three phases (noncontrast-agent enhanced, arterial, and delayed). Masses were diagnosed according to five categories (category A, classic hepatocellular carcinomas [HCCs]; category B, malignant liver tumors other than classic and early HCCs; category C, indeterminate masses or mass-like lesions [including early HCCs and dysplastic nodules] and rare benign liver masses other than hemangiomas and cysts; category D, hemangiomas; and category E, cysts). Supervised training was performed by using 55 536 image sets obtained in 2013 (from 460 patients, 1068 sets were obtained and they were augmented by a factor of 52 [rotated, parallel-shifted, strongly enlarged, and noise-added images were generated from the original images]). The CNN was composed of six convolutional, three maximum pooling, and three fully connected layers. The CNN was tested with 100 liver mass image sets obtained in 2016 (74 men and 26 women; mean age, 66.4 years ± 10.6 [standard deviation]; mean mass size, 26.9 mm ± 25.9; 21, nine, 35, 20, and 15 liver masses for categories A, B, C, D, and E, respectively). Training and testing were performed five times. Accuracy for categorizing liver masses with CNN model and the area under receiver operating characteristic curve for differentiating categories A–B versus categories C–E were calculated.

Results: Median accuracy of differential diagnosis of liver masses for test data were 0.84. Median area under the receiver operating characteristic curve for differentiating categories A–B from C–E was 0.92.

Conclusion: Deep learning with CNN showed high diagnostic performance in differentiation of liver masses at dynamic CT.

Online supplemental material is available for this article.

[1] From the Department of Radiology, The University of Tokyo Hospital, 7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan 113-8655. Received March 26, 2017; revision requested May 30; revision received June 5; accepted July 13; final version accepted August 14. **Address correspondence to** S.K. (e-mail: *kiryu-tky@umin.ac.jp*).

Hepatocellular carcinoma (HCC) is the second most frequent cause of malignancy-related death worldwide (1). In addition to HCCs, several types of masses arise in the liver, including malignant masses such as intrahepatic cholangiocellular carcinomas, and benign masses such as hemangiomas and cysts. The liver is also a target for metastasis from many types of malignant tumor. For the differential diagnosis of these liver masses, dynamic computed tomography (CT) provides useful information (2–5). From arterial phase images, vascularity and the contrast agent enhancement pattern of the liver masses can be assessed (2–4). The degree of contrast enhancement in delayed phase imaging (5) is also useful for differential diagnosis. With careful evaluation of these images, diagnoses can be reached in many patients. However, evaluations of such images are generally subjective and are possibly affected by radiologists' experience to an extent (6).

Advances in Knowledge

- Differential diagnosis of liver masses (classified into the following five categories: category A, classic hepatocellular carcinomas [HCCs]; category B, malignant liver tumors other than classic and early HCCs; category C, indeterminate masses or mass-like lesions [including early HCCs and dysplastic nodules] and rare benign liver masses other than hemangiomas and cysts; category D, hemangiomas; and category E, cysts) at dynamic contrast agent–enhanced CT (by using a combination of noncontrast agent–enhanced, arterial, and delayed-phase images) is possible by using deep learning with a convolutional neural network.
- Use of the models enabled the classification of liver masses to the five categories with sensitivity of 0.71, 0.33, 0.94, 0.90, and 1.00 for category A, B, C, D, and E, respectively, and with an overall accuracy of 0.84.
- Use of the models enabled differentiation of categories A–B from C–E, with an area under the receiver operating characteristic curve of 0.92.

Recently, deep learning with convolutional neural networks (CNNs) has been gaining attention with respect to pattern recognition of images and as a strategy for artificial intelligence (7–9). A neural network is one of the methods used for machine learning. In deep learning, neural networks are composed of several layers. Convolutional layers, in which images are processed with several types of filter, are known to be effective for pattern recognition of images (10–12). Whereas conventional machine-learning algorithms require features from images to be extracted in advance of learning, application of convolutional layers allows the image itself to be used during the learning process (7). Therefore, deep learning with CNNs enables all of the information contained in the images to be used, though this has been limited according to the feature parameters selected in conventional machine learning. Deep learning with CNNs is reported (8,9,13–15) to achieve a good performance in the pattern recognition of images. Therefore, this method has the potential to differentiate liver masses without depending on the experience of the radiologist.

In this study, we aimed to investigate the diagnostic performance of deep learning with a CNN for the differentiation of liver masses on dynamic contrast agent–enhanced CT images.

Materials and Methods

This retrospective clinical study was approved by our institutional review board and the requirement for written informed consent was waived.

Outline

This study consisted of two stages: a training stage, in which deep learning models were built by using training CT image sets (obtained from January 2013 to December 2013), and a test stage to examine the accuracy of the models by using test CT image sets (obtained from January 2016 to June 2016). For training, we undertook supervised learning with a CNN by using CT images (unenhanced, arterial, and delayed phase images, and combinations of these) focused in on liver masses as the input data, and included five categories (described in detail below) as teaching data. A flowchart of the outline of this study is described in Figure 1.

Teaching Data

An abdominal radiologist (K.Y., with 7 years of imaging experience) searched our picture archiving and communication system for dynamic CT studies performed for the evaluation of liver lesions. We included the following five categories of liver masses or mass-like lesions (hereafter, we will refer to these as liver masses unless otherwise specified) of any size that were diagnosed on the basis of the criteria described in the next subsection: category A, classic HCCs; category B, malignant liver tumors other than classic and early HCCs (eg, intrahepatic cholangiocellular carcinomas, combined hepatocellular-cholangiocarcinomas, and liver metastases); category C, indeterminate masses or mass-like lesions (eg, early HCCs and dysplastic nodules) and rare benign liver masses other than hemangiomas and cysts; category D, liver hemangiomas; and category E, cysts. The

https://doi.org/10.1148/radiol.2017170706

Content code: GI

Radiology 2018; 000:1–10

Abbreviations:
CNN = convolutional neural network
HCC = hepatocellular carcinoma
JPEG = Joint Photographic Experts Group

Author contributions:
Guarantors of integrity of entire study, K.Y., S.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, K.Y., O.A.; clinical studies, K.Y., O.A.; statistical analysis, K.Y., O.A.; and manuscript editing, all authors

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

Figure 1

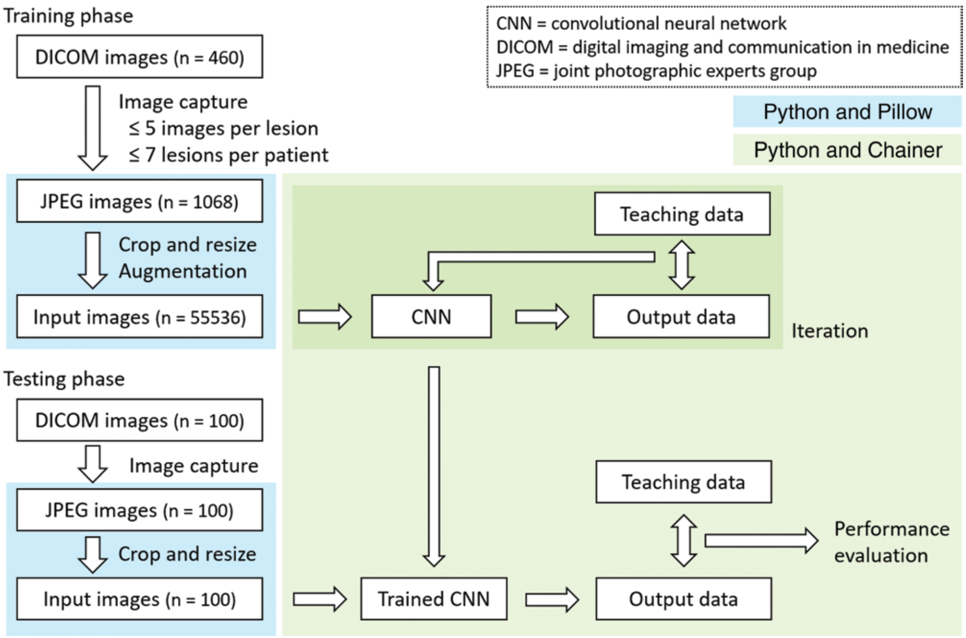

Figure 1: A flowchart of the study process from image capture to training and testing phase with CNN.

following CT image sets were excluded: those with prominent artifacts, those of liver masses treated with transarterial chemoembolization therapy or systemic chemotherapy, and those of liver masses in patients younger than 20 years.

Reference Standard

Liver masses were classified into five categories by referencing radiologic reports made by experienced radiologists and adhering to the following criteria: histopathologic evaluation after surgery or biopsy (categories A–C); imaging criteria (category A, early enhancement and delayed washout; category C [hypervascular lesions], early enhancement and isoattenuation on delayed phase images; category C [hypovascular lesions], hypoattenuation during the arterial phase and isoattenuation or low attenuation during the delayed phase; category D [hemangiomas], peripheral nodular enhancement during the arterial phase and a centripetal filling enhancement pattern during the delayed phase, or uniform intense enhancement [close to attenuation of the aorta] during the arterial phase and high attenuation during the delayed phase; and category E [cysts], water attenuation and no apparent contrast enhancement); and a combination of clinical information and imaging criteria (category B [metastases], a known history of malignancy in other organs and emergent or enlarged masses). For liver masses diagnosed with a combination of clinical information and imaging criteria, the median follow-up time was 98 days (range, 61–136 days) for the training data group and 65 days (range, 28–238 days) for the test data group. The details of diagnoses for each category and distribution of reference standards is described in Table 1.

Image Data: Input Data

For the training phase, up to seven lesions per patient were included in patients with several liver masses. To ensure the model was robust with respect to slight differences in the table position among liver masses, image sets with up to five different table positions were included per lesion. From 460 patients, a total of 1068 image sets (there were 240, 121, 320, 207, and 180 masses for category A, B, C, D, and E, respectively), imaged in 2013, were used for training purposes.

For the test phase, we used CT image sets of 100 liver masses of 100 patients (74 men and 26 women; mean age, 66.4 years ± 10.6 [standard deviation]; mean mass size, 26.9 mm ± 25.9) that were obtained in 2016. The number of patients for test phase was determined so that the ratio of the number of patients (training [ie, 460 patients]/test [ie, 100 patients]) would be about 90:10 to 80:20. The number liver masses for category A, B, C, D, and E for the test was 21, nine, 35, 20, and 15, respectively. For patients who had several liver masses, the image of the largest liver mass was obtained. Image sections that included the largest diameter of liver masses were included. Mean mass size for category A, B, C, D, and E was 38.9 mm ± 32.2, 33.7 mm ± 16.9, 16.4 mm ± 8.3, 27.3 mm ± 21.0, and 30.2 mm ± 42.6, respectively.

The rationale for determining the number of patients in each category is described in Appendix E1 (online).

CT Imaging

CT scanners from two manufacturers (Toshiba Medical Systems, Tochigi, Japan; and GE Healthcare, Waukesha,

Table 1

Diagnostic Details for Each Category of Liver Masses

Parameter	No. of Test Session Patients	No. of Training Session Patients
Category A	21	240
Histopathologic evaluation	9	50
No. of imaging findings	12	190
Category B	9	121
ICC histopathologic evaluation	3	16
Combined HCC and ICC histopathologic evaluation	0	12
Metastasis from colon carcinoma		
Histopathologic evaluation	0	50
Combination of clinical information and imaging criteria	2	22
Metastasis from pancreatic NET		
Histopathologic evaluation	0	9
Combination of clinical information and imaging criteria	2	7
Other histopathologic evaluation	2	5
Category C	35	320
No. of imaging findings of hypervascular lesions	21	140
No. of imaging findings of hypovascular lesions	14	168
Other histopathologic evaluation	0	12
Category D	20	207
No. of imaging findings	20	207
Category E	15	180
No. of imaging findings	15	180

Note.—ICC = intrahepatic cholangiocellular carcinoma, NET = neuroendocrine tumor.

Wis) were used in this study. Axial CT images, acquired during the unenhanced, arterial, and delayed phases, were included. Contrast enhancement materials (Iopamiron, Bayer Yakuhin, Osaka, Japan; Omnipaque, Daiichi Sankyo, Tokyo, Japan; Iopaque, Fuji Pharma, Tokyo, Japan; or Oypalomin, Fuji Pharma) were injected within 30 seconds. The concentration of the contrast materials was determined by the body weight (350 mg of iodine per milliliter for patients who weighed <60 kg and 370 mg of iodine per milliliter for those who weighted >60 kg). The volume of the contrast enhancement materials was determined by multiplying the body weight (in kilograms) by 2, with an upper limit of 100 mL.

The timing for arterial phase scanning was determined by using a bolus-tracking technique. The arterial phase was scanned 15 seconds after CT attenuation of the aorta at the level of the diaphragm had reached 200 HU. The delayed phase was scanned 180 seconds after the contrast material was injected.

All CT examinations were performed with a tube voltage of 120 kVp. For the tube current, an automatic tube current modulation technique was used. Standard deviation or noise index of 13.0 and 11.36 were used for the CT scanners from Toshiba and GE Healthcare, respectively. The examinations were performed with helical mode, and helical pitch was 0.8125:1 for CT imager from Toshiba and 0.984:1 for the CT imager from GE Healthcare. Gantry rotation time was 0.5 second for both CT scanners from Toshiba and GE Healthcare. Detector configuration was 0.5 mm × 80 for Toshiba CT and 0.625 mm × 64 for GE Healthcare CT. Images were reconstructed in the axial plane by using a kernel for the evaluation of soft tissues (FC03 for the Toshiba scanner and Standard for the GE Healthcare scanner).

For the unenhanced images, section thickness was 5 mm and section interval was 5 mm. For contrast-enhanced images, section thickness was 3 mm and the interval was 1.5 mm for Toshiba CT and 2.5 mm and 1.5 mm, respectively, for GE Healthcare CT. The number of patients who were scanned with Toshiba CT and GE Healthcare CT imagers for each category was six and 15 patients, respectively, for category A; three and six patients, respectively, for category B; 11 and 24 patients, respectively, for category C; three and 17 patients, respectively, for category D; and six and nine patients, respectively, for category E. There was no statistically significant difference in the proportion of scanners used among categories in test patient group (P = .545, Fisher exact test).

Image Processing

By using commercial viewing software (Centricity Radiology RA 1000; GE Healthcare), CT images in Digital Imaging and Communications in Medicine format were displayed with a window level and width of 40 HU and 250 HU, respectively, for unenhanced CT, and 45 HU and 290 HU, respectively, for contrast-enhanced CT (arterial and delayed phases). By using image capture function, enlarged images of liver masses were converted to 8-bit Joint Photographic Experts Group (JPEG) format (594 × 644 pixels). Further information regarding the image capturing process is provided in Appendix E2 (online).

The image sets were further processed by using code written in the programming language Python 3.5 (https://www.python.org) and Python imaging library of Pillow 3.3.1 (https://pypi.python.org/pypi/Pillow/3.3.1). Image processing was performed separately for the training and test image sets.

For the training image sets (unenhanced, arterial, and delayed phase images), image processing and data augmentation were performed such that the CNN model became robust against the degree of enlarging, rotation, and parallel shift, as well as to slight differences in the amount of image noise. Through those processes, 52 image sets were generated from one image set, resulting in a total of 55 536 image sets (1068 image sets × 52) that were available for training use. Such data augmentation is commonly performed in deep learning (8). Details of the image processing and data augmentation methods used for the training image sets are provided in Appendix E2 (online).

For each test image set, the central part (500 × 500 pixels) of captured images was cropped. For 10 liver masses, the cropped images included only the liver mass and liver parenchyma, and for the other 90 liver masses, the cropped images included the liver mass, liver parenchyma, and surrounding tissues or organs.

Training and test image sets were scaled down to 70 × 70 pixels by using the resize function with Pillow (https://pypi.python.org/pypi/Pillow/3.3.1).

Deep Learning with a CNN: Training Phase

We performed deep learning by using a computer with a GeForce GTX 1080 (NVIDIA, Santa Clara, Calif) graphics processing unit, a Core i7-6700 K 4.00-GHz (Intel, Santa Clara, Calif) central processing unit, and 64 GB of random-access memory. The Python programming language and Chainer 1.18.0 (http://chainer.org/) framework for neural networks were used for deep learning with a CNN. For training, image sets prepared as per the Image Processing section were provided to the CNN. The output data were compared with the teacher data (five categories), and the error was back propagated to update parameters in the CNN so that the error between the output data and teacher data would be minimal. The CNN comprised several layers (six convolutional layers, three maximum pooling layers, and three fully connected layers) (Fig 2). Details of the CNN are provided in Appendix E3 (online).

The training was performed to obtain five different models (trained with unenhanced, arterial, and delayed phase CT images, hereafter referred to as model triphasic; trained with arterial and delayed phase images, hereafter referred to as model art/del; trained with unenhanced CT images, hereafter referred to as model unenhanced; trained with arterial phase images, hereafter referred to as model arterial; and trained with delayed phase images, hereafter referred to as model delayed). Because there is randomness in initialization, selecting data for a group with a small number of images (known as a minibatch) and dropout of unit in neural network (details of the minibatch and dropout are provided in Appendix E3 [online]), the training sessions were performed five times to obtain each of these models.

Figure 2

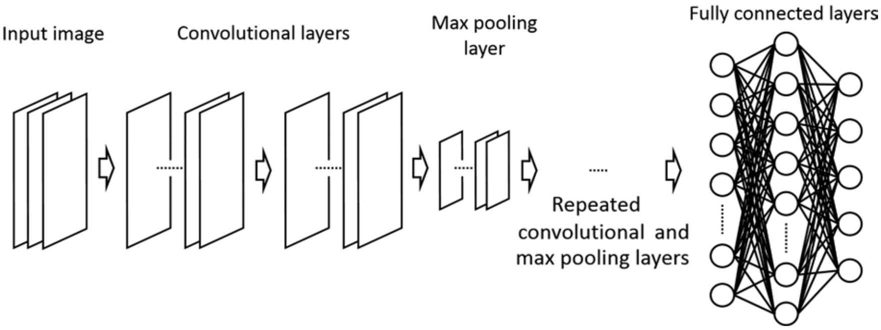

Figure 2: Conceptual image of the CNN used in this study. Images provided to the CNN were processed initially in two convolutional layers and one maximum pooling layer. These layers were then combined three times. The data were finally processed in fully connected layers.

Testing Phase with CNN

After building the models, we examined the accuracy of the trained models in distinguishing among the five liver mass categories by using test CT image sets. These image sets were provided for the CNN. The CNN returns numbers for each category (output data) that can be regarded as probability for each category. The mass category for which the output value was highest was observed as the diagnosis made by the deep learning model.

Statistics

For statistical analyses, the EZR software version 1.33 (http://www.jichi.ac.jp/saitama-sct/SaitamaHP.files/statmedEN.html) (16), which is a graphical user interface for the R software package (version 3.2.2; R Development Core Team, Vienna, Austria), was used. Accuracy, with respect to differentiation of liver masses by using the CNN models, was calculated. The median accuracy among five sessions was compared between five different models with the Kruskal-Wallis test. The Steel test of post hoc multiple comparison was performed by using model triphasic as a control. A *P* value of less than .05 was considered to indicate a statistically significant difference.

Receiver operating characteristic analyses were performed with software (EZR software and pROC package; https://cran.r-project.org/web/packages/pROC/index.html) to calculate nonparametric estimate of the area under the receiver operating characteristic curve for performance of models in discriminating malignant masses (categories A–B) from indeterminate and benign masses (categories C–E). In the receiver operating characteristic analyses, the degree of clinical relevance was scaled, and category A and B (malignant tumors), C (mainly indeterminate masses), D (benign tumors), and E (nontumor lesions) were assigned scores of 4, 3, 2, and 1, respectively.

The sensitivity, specificity, and accuracy for diagnosing each category were calculated by using the test data.

Results

The accuracy of the differential diagnosis with each model is shown in Table 2. The accuracy of model triphasic

Table 2

Accuracy of Classifying Liver Masses into Five Categories by Using CNN Models

Parameter	Model Triphasic	Model Art/Del	Model Unenhanced	Model Arterial	Model Delayed
Training data	0.95 (0.95–0.96)	0.96 (0.96–0.96)	0.96 (0.96–0.96)	0.97 (0.97–0.98)	0.97 (0.96–0.97)
Test data	0.84 (0.82–0.90)	0.81 (0.75–0.86)	0.48 (0.39–0.60)	0.66 (0.63–0.70)	0.71 (0.67–0.74)
P value	NA	.384	.031*	.031*	.030*

Note.—Data are median values; data in parentheses are range. For test data, comparison by using Kruskal-Wallis test among five models followed by the post hoc Steel test by using model triphasic as a control was performed. *P* value for Kruskal-Wallis test was <.001. *P* values for the post hoc Steel test are shown in the table. NA = not applicable.

* Statistically significant difference ($P < .05$).

Figure 3

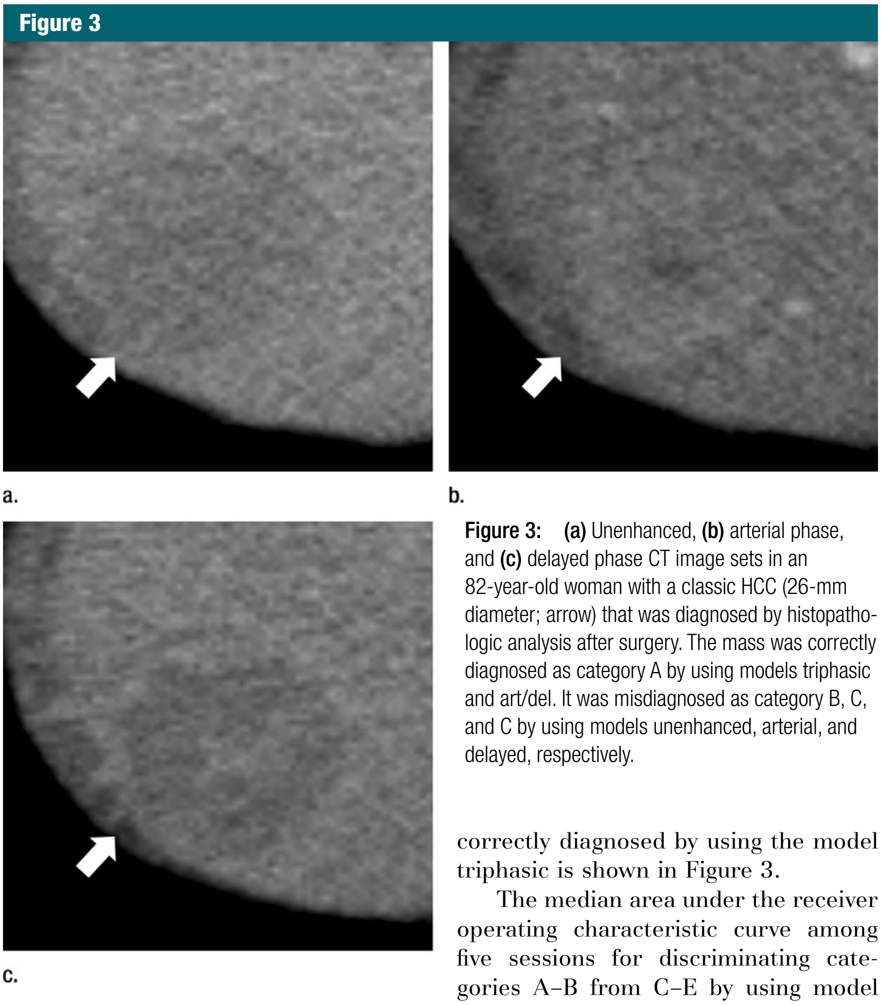

Figure 3: **(a)** Unenhanced, **(b)** arterial phase, and **(c)** delayed phase CT image sets in an 82-year-old woman with a classic HCC (26-mm diameter; arrow) that was diagnosed by histopathologic analysis after surgery. The mass was correctly diagnosed as category A by using models triphasic and art/del. It was misdiagnosed as category B, C, and C by using models unenhanced, arterial, and delayed, respectively.

(median, 0.84 [84 of 100]; range, 0.82–0.90) in test data was significantly higher than that in model unenhanced, model arterial, and model delayed ($P \leq .031$). The accuracy of model art/del was also high (median, 0.81 [81 of 100]; range, 0.75–0.86). A representative image of category A liver mass that was correctly diagnosed by using the model triphasic is shown in Figure 3.

The median area under the receiver operating characteristic curve among five sessions for discriminating categories A–B from C–E by using model triphasic, model art/del, model unenhanced, model arterial, and model delayed with the test data sets was 0.92 (range, 0.89–0.93), 0.88 (range, 0.82–0.89), 0.61 (range, 0.55–0.63), 0.78 (range, 0.77–0.82), and 0.81 (range, 0.79–0.82), respectively. The receiver operating characteristic curves of the models that showed median area under the receiver operating characteristic curve value among the five sessions are shown in Figure 4.

The sensitivity for each mass category, by using the test data, is described in Table 3. All data regarding differential diagnosis for each mass category are described in Tables E1–E5 (online). Model triphasic and model art/del showed moderate to high sensitivity for each mass category except category B. In all models, the sensitivity for diagnosing category B masses was relatively low (range, 0.11–0.33) compared with that for the other categories (Fig 5). By using model arterial, the sensitivity for diagnosing category A and hypervascular lesions was relatively low (<0.60); category A lesions and hypervascular lesions were mainly misdiagnosed as category C (misclassified ratio, 0.43 [nine of 21]) and as category A (misclassified ratio, 0.38 [eight of 21]), respectively. By using model delayed, the sensitivity for diagnosing category A and hypovascular lesions was also relatively low (<0.60); category A lesions and hypovascular lesions were mainly misdiagnosed as category C (misclassified ratio, 0.43 [nine of 21]) and as category A (misclassified ratio, 0.36 [five of 14]), respectively.

Discussion

We investigated whether different types of liver mass could be differentiated at dynamic CT by using models based on deep learning with a CNN. We found that model triphasic and model art/del were useful for classifying liver masses into five categories

Figure 4

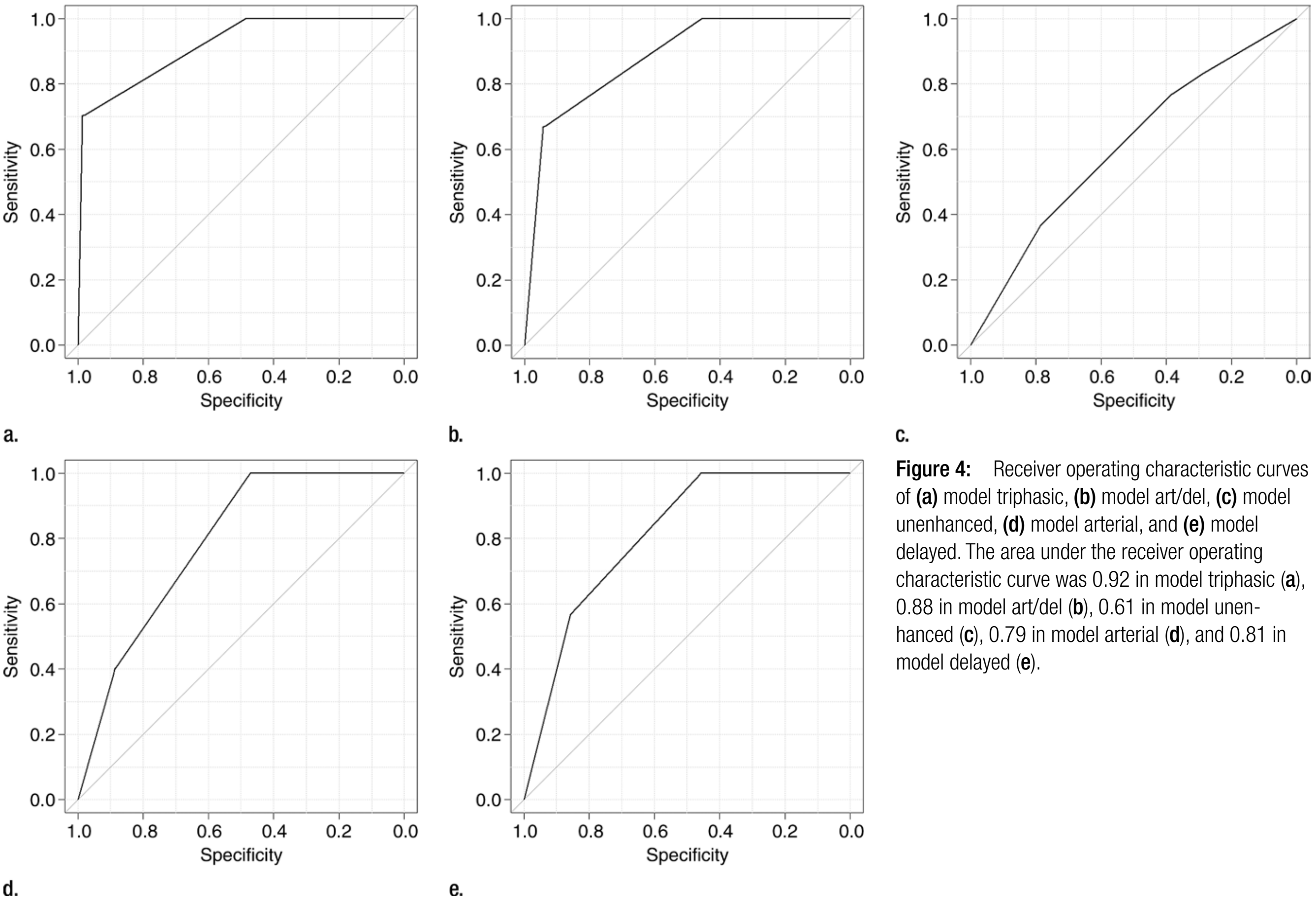

Figure 4: Receiver operating characteristic curves of **(a)** model triphasic, **(b)** model art/del, **(c)** model unenhanced, **(d)** model arterial, and **(e)** model delayed. The area under the receiver operating characteristic curve was 0.92 in model triphasic (**a**), 0.88 in model art/del (**b**), 0.61 in model unenhanced (**c**), 0.79 in model arterial (**d**), and 0.81 in model delayed (**e**).

Table 3

Sensitivity for Each Liver Mass Category by Using Test Data

Parameter	Model Triphasic	Model Art/Del	Model Unenhanced	Model Arterial	Model Delayed
Category A	0.71 [15/21] (0.71–0.90)	0.71 [15/21] (0.43–0.76)	0.33 [7/21] (0.24–0.38)	0.52 [11/21] (0.29–0.67)	0.52 [11/21] (0.52–0.62)
Category B	0.33 [3/9] (0.22–0.44)	0.22 [2/9] (0.11–0.33)	0.22 [2/9] (0–0.33)	0.11 [1/9] (0–0.44)	0.33 [3/9] (0–0.44)
Category C	0.94 [33/35] (0.91–0.97)	0.91 [32/35] (0.86–1.00)	0.54 [19/35] (0.43–0.74)	0.69 [24/35] (0.66–0.77)	0.74 [26/35] (0.63–0.86)
Hypervascular lesions	0.95 [20/21] (0.90–0.95)	0.86 [18/21] (0.81–1.00)	0.62 [13/21] (0.57–0.90)	0.57 [12/21] (0.52–0.76)	0.90 [19/21] (0.71–0.95)
Hypovascular lesions	0.93 [13/14] (0.86–1.00)	1.00 [14/14] (0.93–1.00)	0.43 [6/14] (0.21–0.57)	0.86 [12/14] (0.79–0.93)	0.50 [7/14] (0.50–0.71)
Category D	0.90 [18/20] (0.85–0.95)	0.90 [18/20] (0.65–1.00)	0.25 [5/20] (0.05–0.50)	0.75 [15/20] (0.65–0.95)	0.80 [16/20] (0.75–0.85)
Category E	1.00 [15/15] (1.00–1.00)	1.00 [15/15] (1.00–1.00)	1.00 [15/15] (1.00–1.00)	1.00 [15/15] (0.93–1.00)	1.00 [15/15] (1.00–1.00)

Note.—Data are median values with numerator/denominator in brackets and range in parentheses.

and showed a median accuracy of 0.84 (range, 0.82–0.90) and 0.81 (range, 0.75–0.86), respectively. Use of the model triphasic allowed high discrimination between malignant versus category C and benign masses, with area under the receiver operating characteristic curve of 0.92 (range, 0.89–0.93).

Some studies (17–20) applied conventional machine learning for differentiation of liver masses by placing regions of interest on masses and extracting features such as quantitative texture parameters; they reported good diagnostic performance, with an accuracy of 0.817–1.000. However, their models did not take indeterminate masses or mass-like lesions into account. Dysplastic nodules can appear as hypovascular lesions with

Figure 5

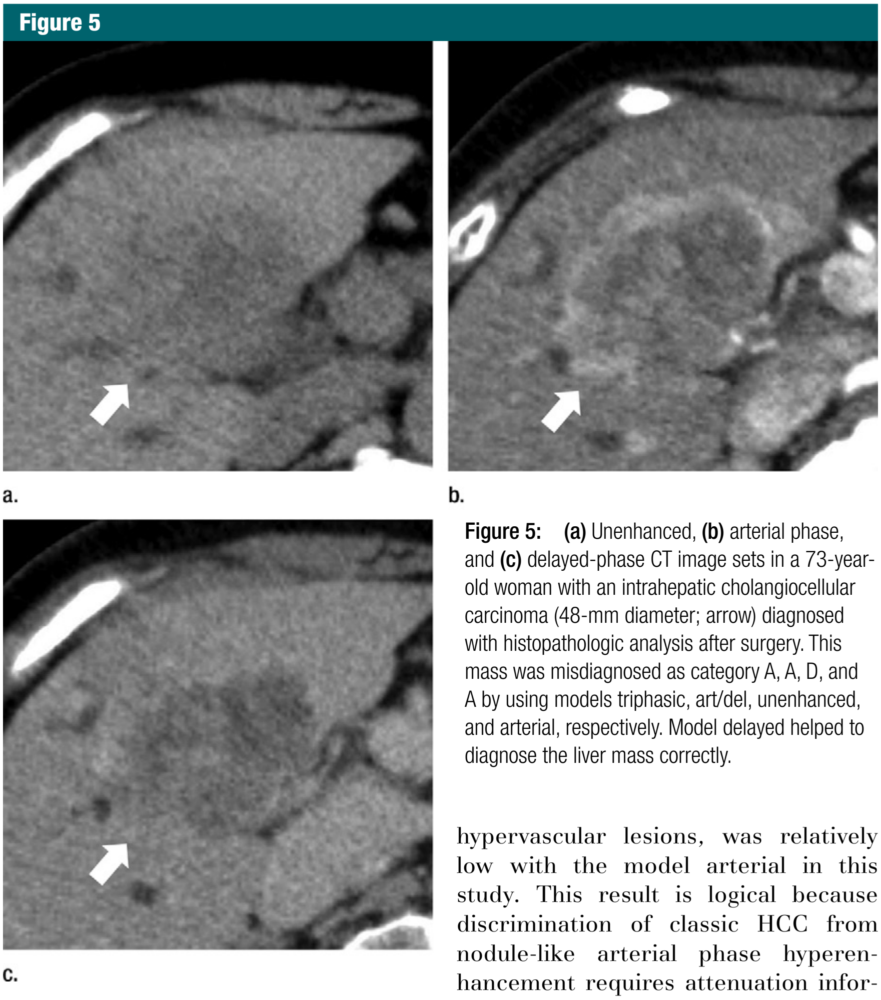

Figure 5: **(a)** Unenhanced, **(b)** arterial phase, and **(c)** delayed-phase CT image sets in a 73-year-old woman with an intrahepatic cholangiocellular carcinoma (48-mm diameter; arrow) diagnosed with histopathologic analysis after surgery. This mass was misdiagnosed as category A, A, D, and A by using models triphasic, art/del, unenhanced, and arterial, respectively. Model delayed helped to diagnose the liver mass correctly.

isoattenuation or low attenuation in the delayed phase (21,22), but early HCC can also show such features (22,23). Therefore, these masses on CT images should be included in models, and we included these as category C masses in this study. Nodule-like arterial phase hyperenhancement, visible only at the arterial phase, can also be observed on dynamic CT images (24). Nodule-like arterial phase hyperenhancement is thought to represent an alteration in perfusion or a small nonmalignant mass. However, such lesions, which were classified as category C in this study, also need to be included in models because some HCCs can manifest with the same imaging findings (24).

Accuracy, with respect to differentiating between category A and hypervascular lesions, was relatively low with the model arterial in this study. This result is logical because discrimination of classic HCC from nodule-like arterial phase hyperenhancement requires attenuation information during the delayed phase. The accuracy for distinguishing category A and hypovascular lesions by using model delayed was also relatively low. This result is also logical, because discrimination of classic HCC and hypovascular lesions requires information regarding vascularity, which can be obtained in the arterial phase. Model triphasic and model art/del classified these indeterminate lesions and classic HCCs as category C and category A with moderate to high sensitivity, and overall accuracy with model triphasic was significantly higher than that in model arterial and model delayed.

Our model requires the radiologists only to focus on the tumors, capture the images, and provide the image to the CNN. Our model also differs from previous studies (17–20) that used conventional machine-learning methods for the differentiation of liver masses in that it does not require complex-shaped regions of interest tracing boundaries of tumors, or circular or elliptical regions of interest within tumors. Our model showed high performance irrespective of what was depicted around the liver masses. This might have been achieved by training with several liver masses in various sections of the liver.

The image-capture process might be subjective; however, our model was built to be robust for this process by training with multiple images with different table positions that were augmented by enlarging with different degrees and by parallel shift.

Complex models, like CNNs, are known to experience the problem of overfitting, which results in suboptimal performance for data not included in the training phase. Because of this, preparing an independent test data set, which was not included in the training process, is necessary to appropriately evaluate the performance of the model (7,25). Another approach would be to divide the whole dataset into training data and test data and repeating calculations with different training and test splits (eg, the K-fold cross-validation and leave-one-out cross-validation), which is commonly used when the number of available datasets is limited (25). We evaluated the performance of the model on the basis of the former method by preparing independent test image sets scanned 3 years after training image sets.

The use of a graphics processing unit is known to enable models to be trained 10–20 times faster, which is one of the reasons for the success of deep learning (7). In this study, we used a graphics processing unit and each CNN training session took about 10–20 minutes, although we did not precisely measure the time elapsed. Once a model has been established, its application to a new lesion does not require complex computations. In this study, the time needed to perform the test, without the graphics processing unit and by using 100 image sets in each session, was about 10 seconds.

Several limitations should be acknowledged regarding this study. First, the sensitivity for diagnosing category B masses was not as good as that for the other categories. The inclusion of background clinical data, such as the presence of malignant diseases or the addition of a function that allowed for comparison with previous CT examinations, would be beneficial when building models to obtain better diagnostic performance for category B masses. We performed this study by using single-section JPEG images, and resized the images to 70 × 70 pixels. Such image processing might have caused loss of information and might have resulted in the relatively low sensitivity for category B masses. We used enlarged JPEG images because they were easy to obtain by using image capture function of the viewer. Conversion of Digital Imaging and Communications in Medicine images to JPEG format would result in information loss; however, we think it would be minimal. Because the window settings used in this study (window level of 40 HU and width of 250 HU for unenhanced CT; and window level of 45 HU and window width of 290 HU for contrast-enhanced CT) allows for evaluations of most abdominal tissues. We used single-section resized (70 × 70 pixels) image sets because otherwise the computational burden would have been too large. The relatively low sample size might also have been one of the reasons for the low sensitivity for category B. Thus, a future investigation that considers these issues is required. Second, we did not intend to build models that are dedicated to liver masses that are difficult to diagnose or models that would predict detailed histologic subtypes. Because the deep learning with CNN model was found to be effective for categorization of many of the liver masses that we encounter in daily clinical practice, we anticipate that studies focusing on those topics will be performed in the future. Third, the histopathologic evaluation was not necessarily obtained for all the liver masses. However, because most benign liver masses are diagnosed with image findings and do not involve surgery, and because some classic HCCs are diagnosed with image findings and are treated with radiofrequency ablation or transarterial chemoembolization, it might be difficult to obtain histopathologic evaluations for all lesions.

In conclusion, this preliminary study, which used 55 536 image sets (1068 image sets augmented by a factor of 52) to obtain models, indicated that classifying liver masses into five categories (classic HCCs, malignant tumors other than classic and early HCCs, indeterminate masses [including early HCCs and dysplastic nodules] or rare benign masses, hemangiomas, and cysts) can be accomplished with a high degree of accuracy by using a deep learning method with a CNN on dynamic contrast-enhanced CT images. The CNN model could be useful for diagnosing most liver masses that we encounter in daily clinical practice; however, further improvement would be necessary to achieve adequate performance for diagnosis of relatively rare malignant liver masses.

Disclosures of Conflicts of Interest: K.Y. disclosed no relevant relationships. **H.A.** disclosed no relevant relationships. **O.A.** disclosed no relevant relationships. **S.K.** disclosed no relevant relationships.

References

1. Ferlay J, Soerjomataram I, Dikshit R, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer 2015;136(5): E359–E386.
2. Van Hoe L, Baert AL, Gryspeerdt S, et al. Dual-phase helical CT of the liver: value of an early-phase acquisition in the differential diagnosis of noncystic focal lesions. AJR Am J Roentgenol 1997;168(5):1185–1192.
3. Nino-Murcia M, Olcott EW, Jeffrey RB Jr, Lamm RL, Beaulieu CF, Jain KA. Focal liver lesions: pattern-based classification scheme for enhancement at arterial phase CT. Radiology 2000;215(3):746–751.
4. van Leeuwen MS, Noordzij J, Feldberg MA, Hennipman AH, Doornewaard H. Focal liver lesions: characterization with triphasic spiral CT. Radiology 1996;201(2):327–336.
5. Itai Y, Ohtomo K, Kokubo T, et al. CT of hepatic masses: significance of prolonged and delayed enhancement. AJR Am J Roentgenol 1986;146(4):729–733.
6. Blachar A, Federle MP, Ferris JV, et al. Radiologists' performance in the diagnosis of liver tumors with central scars by using specific CT criteria. Radiology 2002;223(2):532–539.
7. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521(7553):436–444.
8. Krizhevsky A, Sutskever I, Hinton G. ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems 25 (NIPS 2012). http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks. Published 2012. Accessed January 31, 2017.
9. Le QV, Ranzato M, Monga R, et al. Building high-level features using large scale unsupervised learning. International Conference on Machine Learning http://icml.cc/2012/papers. Published 2012. Accessed January 31, 2017.
10. Fukushima K, Miyake S. Neocognitron: a new algorithm for pattern recognition tolerant of deformations and shifts in position. Pattern Recognit 1982;15(6):455–469.
11. LeCun Y, Boser B, Denker JS, et al. Backpropagation applied to handwritten zip code recognition. Neural Comput 1989;1(4):541–551.
12. LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc IEEE 1998;86(11):2278–2324.
13. Farabet C, Couprie C, Najman L, Lecun Y. Learning hierarchical features for scene labeling. IEEE Trans Pattern Anal Mach Intell 2013;35(8):1915–1929.
14. Tompson J, Jain A, LeCun Y, Bregler C. Joint training of a convolutional network and a graphical model for human pose estimation. Advances in Neural Information Processing Systems 27 (NIPS 2014). https://papers.nips.cc/book/advances-in-neural-information-processing-systems-27-2014. Published 2014. Accessed January 31, 2017.
15. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. Cornell University Library. http://arxiv.org/abs/1409.4842. Published 2014. Accessed January 31, 2017.
16. Kanda Y. Investigation of the freely available easy-to-use software 'EZR' for medical statistics. Bone Marrow Transplant 2013;48(3): 452–458.
17. Mougiakakou SG, Valavanis IK, Nikita A, Nikita KS. Differential diagnosis of CT focal liver lesions using texture features, feature selection and ensemble driven classifiers. Artif Intell Med 2007;41(1):25–37.
18. Huang YL, Chen JH, Shen WC. Diagnosis of hepatic tumors with texture analysis in nonenhanced computed tomography images. Acad Radiol 2006;13(6):713–720.

19. Gletsos M, Mougiakakou SG, Matsopoulos GK, Nikita KS, Nikita AS, Kelekis D. A computer-aided diagnostic system to characterize CT focal liver lesions: design and optimization of a neural network classifier. IEEE Trans Inf Technol Biomed 2003;7(3):153–162.

20. Ye J, Sun Y, Wang S. Multi-phase CT image based hepatic lesion diagnosis by SVM. Biomedical Engineering and Informatics, 2009. http://ieeexplore.ieee.org/document/5304774. Published 2009. Accessed January 31, 2017.

21. Choi BI, Han JK, Hong SH, et al. Dysplastic nodules of the liver: imaging findings. Abdom Imaging 1999;24(3):250–257.

22. Sano K, Ichikawa T, Motosugi U, et al. Imaging study of early hepatocellular carcinoma: usefulness of gadoxetic acid-enhanced MR imaging. Radiology 2011;261(3):834–844.

23. Takayasu K, Furukawa H, Wakao F, et al. CT diagnosis of early hepatocellular carcinoma: sensitivity, findings, and CT-pathologic correlation. AJR Am J Roentgenol 1995;164(4):885–890.

24. Jha RC, Mitchell DG, Weinreb JC, et al. LI-RADS categorization of benign and likely benign findings in patients at risk of hepatocellular carcinoma: a pictorial atlas. AJR Am J Roentgenol 2014;203(1):W48–W69.

25. Vieira S, Pinaya WH, Mechelli A. Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications. Neurosci Biobehav Rev 2017;74(Pt A):58–75.