Received: 8 December 2023 | Accepted: 5 April 2024

DOI: 10.1097/HEP.0000000000000904

ORIGINAL ARTICLE

OPEN

Deep learning and digital pathology powers prediction of HCC development in steatotic liver disease

Takuma Nakatsuka[1] | Ryosuke Tateishi[1] | Masaya Sato[1,2] | Natsuka Hashizume[3] | Ami Kamada[3] | Hiroki Nakano[3] | Yoshinori Kabeya[3] | Sho Yonezawa[3] | Rie Irie[4] | Hanako Tsujikawa[4] | Yoshio Sumida[5] | Masashi Yoneda[5] | Norio Akuta[6] | Takumi Kawaguchi[7] | Hirokazu Takahashi[8] | Yuichiro Eguchi[8,9] | Yuya Seko[10] | Yoshito Itoh[10] | Eisuke Murakami[11] | Kazuaki Chayama[12,13,14] | Makiko Taniai[15] | Katsutoshi Tokushige[15] | Takeshi Okanoue[16] | Michiie Sakamoto[4] | Mitsuhiro Fujishiro[1] | Kazuhiko Koike[1,17]

[1]Department of Gastroenterology, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan

[2]Department of Clinical Laboratory Medicine, The University of Tokyo, Tokyo, Japan

[3]RWD Analytics, Healthcare & Life Science, IBM Japan Ltd., Tokyo, Japan

[4]Department of Pathology, Keio University School of Medicine, Tokyo, Japan

[5]Department of Internal Medicine, Division of Hepatology and Pancreatology, Aichi Medical University, Aichi, Japan

[6]Department of Hepatology, Toranomon Hospital and Okinaka Memorial Institute for Medical Research, Tokyo, Japan

[7]Department of Medicine, Division of Gastroenterology, Kurume University School of Medicine, Fukuoka, Japan

[8]Liver Center, Saga University Hospital, Saga, Japan

[9]Loco Medical General Institute, Saga, Japan

[10]Department of Molecular Gastroenterology and Hepatology, Kyoto Prefectural University of Medicine Graduate School of Medical Science, Kyoto, Japan

[11]Department of Gastroenterology and Metabolism, Graduate School of Biomedical and Health Sciences, Hiroshima University, Hiroshima, Japan

[12]Collaborative Research Laboratory of Medical Innovation, Graduate School of Biomedical and Health Sciences, Hiroshima University, Hiroshima, Japan

[13]RIKEN Center for Integrative Medical Sciences, Yokohama, Japan

[14]Hiroshima Institute of Life Sciences, Hiroshima, Japan

[15]Department of Internal Medicine, Institute of Gastroenterology, Tokyo Women's Medical University, Tokyo, Japan

[16]Department of Gastroenterology, Saiseikai Suita Hospital, Suita, Osaka, Japan

[17]Department of Hepatobiliary and Pancreatic Medicine, Kanto Central Hospital, Tokyo, Japan

Abbreviations: ALD, alcohol-associated liver disease; CNN, convolutional neural networks; DL, deep learning; H&E, hematoxylin and eosin; ML, machine learning; SLD, steatotic liver disease; WSI, whole-slide image.

Supplemental Digital Content is available for this article. Direct URL citations are provided in the HTML and PDF versions of this article on the journal's website, www.hepjournal.com.

Correspondence
Ryosuke Tateishi, Department of Gastroenterology, Graduate School of Medicine, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8655, Japan.
Email: tateishi-tky@umin.ac.jp

Abstract

Background and Aims: Identifying patients with steatotic liver disease who are at a high risk of developing HCC remains challenging. We present a deep learning (DL) model to predict HCC development using hematoxylin and eosin-stained whole-slide images of biopsy-proven steatotic liver disease.

Approach and Results: We included 639 patients who did not develop HCC for ≥7 years after biopsy (non-HCC class) and 46 patients who developed HCC <7 years after biopsy (HCC class). Paired cases of the HCC and non-HCC classes matched by biopsy date and institution were used for training, and the remaining nonpaired cases were used for validation. The DL model was trained using deep convolutional neural networks with 28,000 image tiles cropped from whole-slide images of the paired cases, with an accuracy of 81.0% and an AUC of 0.80 for predicting HCC development. Validation using the nonpaired cases also demonstrated a good accuracy of 82.3% and an AUC of 0.84. These results were comparable to the predictive ability of logistic regression model using fibrosis stage. Notably, the DL model also detected the cases of HCC development in patients with mild fibrosis. The saliency maps generated by the DL model highlighted various pathological features associated with HCC development, including nuclear atypia, hepatocytes with a high nuclear-cytoplasmic ratio, immune cell infiltration, fibrosis, and a lack of large fat droplets.

Conclusions: The ability of the DL model to capture subtle pathological features beyond fibrosis suggests its potential for identifying early signs of hepatocarcinogenesis in patients with steatotic liver disease.

INTRODUCTION

With the global obesity epidemic, NAFLD has affected approximately one-fourth of the global population and is now the leading cause of chronic liver diseases.[1,2] NASH is a progressive form of NAFLD that can lead to fibrosis, cirrhosis, and HCC.[3,4] NASH is characterized by fat accumulation in the liver, hepatic lobular inflammation, hepatocellular ballooning, and insulin resistance.[5,6] Since only a small percentage of patients with NAFLD develop HCC,[7,8] identifying patients at a high risk of HCC remains an important clinical need in prioritizing treatment, eligibility for clinical trials, and HCC surveillance.

Liver biopsy in patients with suspected NAFLD/NASH allows confirmation of NASH diagnosis, grading, and staging. The degree of liver fibrosis is the strongest predictor of long-term outcomes such as liver-related mortality, liver failure, and development of HCC in patients with NAFLD.[9,10] However, up to 50% of cases of NAFLD-driven HCC occur in patients without cirrhosis, possibly due to its unique nature of arising from lipotoxicity-mediated chronic inflammation[11,12]; thus, factors other than fibrosis should not be underestimated.

In recent years, digital image analysis based on deep learning (DL) and other forms of machine learning (ML) algorithms has shown promise for improving the reliability of the histological evaluation of NASH.[13,14] This technology has also enabled the identification of novel histological features associated with clinical disease progression and indicators of fibrosis severity based on fibrosis patterns distinct from those of the conventional collagen proportionate area.[15] Furthermore, owing to its ability to analyze complex patterns and features of medical images, the DL algorithm may identify early signs of hepatocarcinogenesis that have previously been overlooked, other than liver fibrosis.[16,17] However, no study has examined the power of DL in predicting HCC development directly from the histopathological images of patients with NAFLD.

The coexistence of mild alcohol consumption and metabolic abnormalities in patients with fatty liver is

common in clinical practice making it difficult to determine which is the primary cause of the fatty liver. To address this issue, a new fatty liver disease nomenclature, steatotic liver disease (SLD), was introduced following an international consensus.[18,19] The name chosen to replace NAFLD was metabolic dysfunction–associated steatotic liver disease, and metabolic dysfunction–associated steatotic liver disease with mild alcohol consumption was named MetALD. Therefore, our study aimed to develop and validate a DL model that can predict HCC development in patients with SLD, using hematoxylin and eosin (H&E)-stained whole-slide images (WSIs) of liver tissue obtained through needle biopsy, including those with mild alcohol consumption.

METHODS

Study design and participants

The cohort of patients was enrolled in a nationwide registry study focusing on steatotic liver without heavy alcohol consumption (STEAtotic Liver registry for inves-Tigating clinical outcomes including HCC, STEALTH study), which is a multicenter retrospective cohort study that included patients aged 18 years or older with clinically suspected SLD who underwent liver biopsy between 2003 and 2018. We excluded patients with liver diseases of other etiologies, including alcohol-associated liver disease (ALD) (>60 g/d), viral hepatitis (positive serology for HBsAg or hepatitis C antibody), autoimmune hepatitis, drug-induced liver disease, primary biliary cholangitis, or biliary obstruction. Importantly, we did not exclude patients with SLD who consumed significant amounts of alcohol (≥ 30 g/d in men and ≥ 20 g/d in women), which falls under the MetALD category. The study complied with the human studies guidelines, and was conducted in accordance with the World Medical Association Declaration of Helsinki and the ethical guidelines for epidemiological research of the Japanese Ministry of Education, Culture, Sports, Science, and Technology and the Ministry of Health, Labor, and Welfare. The study protocol was approved by the University of Tokyo Medical Research Center Ethics Committee (Approval No. 2018037NI) and the Institutional Review Board or Ethics Committee of each participating institution. The requirement for individual informed consent was waived due to the retrospective design of the study, and opt-outs were permitted. The study was registered in the University Hospital Medical Information Network (UMIN) Clinical Trial Registry (UMIN-CTR 000049068). All authors had access to the study data and reviewed and approved the final manuscript.

Data set

Of the 2432 cases with SLD collected in the "STEALTH study," those who did not develop HCC for ≥ 7 years of follow-up after biopsy were defined as the non-HCC class, and those who developed HCC <7 years after biopsy were defined as the HCC class. Next, cases from the non-HCC and HCC classes with close biopsy dates (within 1 y) from the same institution were selected on a one-to-one basis to form a paired case group for the discovery set. This pairing was carried out

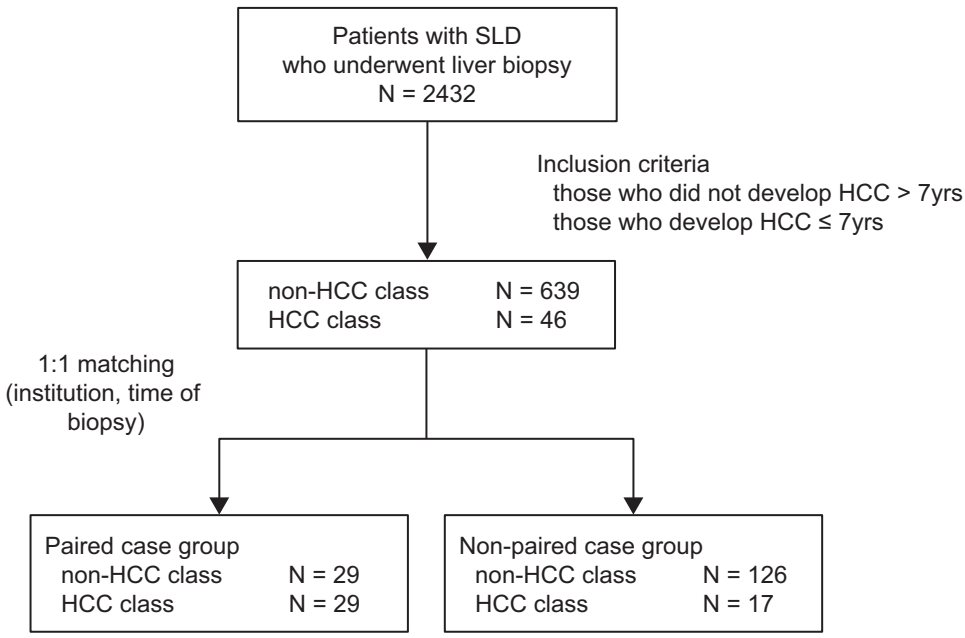

FIGURE 1 Flow diagram of the study population. Abbreviation: SLD, steatotic liver disease.

to avoid learning the features of the staining methods of each institution and the paleness of the color tones that changed over time.[20] Some of the remaining cases were used as the nonpaired case group in the test set (Figure 1).

Deep learning model

Preparation of images

The WSIs of the H&E-stained liver biopsy specimens were randomly cropped into small image patches (256 × 256 pixels) at a ×10 magnification. Images with more than 30% of pixels and a hue of 120–180, saturation of 10–100, and brightness of 180–255 were adopted, and those containing little or no tissue were excluded. The number of image patches obtained from 1 WSI was 400, and 28,000 patch images were used to train the DL model. Supplemental Figure S1, http://links.lww.com/HEP/I430 shows the sample image patches.

Evaluation

Cross-validation. N-fold cross-validation is a widely used technique to evaluate the generalization performance of models.[21] In this method, the data set is divided into N equal parts (folds), with (N−1) of them being used as training data and the remaining one as test data. This process is repeated N times, ensuring each fold is used as test data once. This approach effectively uses the entire data set, allowing for a more accurate evaluation of the model's performance. In our study, we adopted 5-fold cross-validation. This involves dividing the data set into 5 distinct parts, using 4 of them as training data and the remaining one as test data in each iteration. By repeating this process five times, every data piece is used as test data once, ensuring the model's generalization performance is verified. Furthermore, within each cross-validation cycle, ~80% of the selected training data was randomly allocated as the training set, and about 20% as the validation set. The training set included images from 3 types of image folders (normal images, Mixup images, and CutMix images) to ensure data diversity. This diversity is crucial for enabling the model to adapt to various real-world scenarios. The validation set was used to adjust the model's hyperparameters and monitor for overfitting. The test set, in each cycle of cross-validation, was not used for training or validation but for the final evaluation of the model's performance. This allows for a fair and accurate assessment of the model's generalization ability. Through this method, overfitting during model training is avoided, and high performance on actual unknown data is ensured, safeguarding the model's generalization capability.

Following the method described above, the paired case group was used for training and evaluation with a 5-fold cross-validation. To perform a 5-fold cross-validation, the paired cases were equally divided into 5 groups (Figure 2A). Paired patients were included in each group. Four of the 5 groups of cropped images were used to train the classifier, and the remaining group was used as the test set. The training images were randomly divided into the training and validation data sets. The ratios of the number of images in the training and the validation sets were 80% and 20%, respectively. The validation set was not used for training but instead to assess the classifier performance at each epoch during training. An epoch is the training unit and a classifier is trained once with the training set in 1 epoch. The training of a classifier was terminated when the change in the validation loss was <0.05 within 3 epochs. After training, the classifier was evaluated using a test set. Five classifiers were created by repeating this process 5 times, using alternating training, validation, and testing sets. Consequently, all images were used at least once as the test set.

We predicted the probabilities of these patches using the trained classifier and averaged all probabilities as the final probability. Figure 2B shows the detailed procedures and images of the tiled predictions. Areas with red squares were predicted to be the HCC class with at least 60% confidence. On the other hand, areas with blue squares were predicted to be the non-HCC class, with <40% confidence. Areas with gray squares had confidences $\geq 40\%$ and <60%, indicating that they were not unambiguously predicted. The probabilities of all tiles were averaged to form the final probability; if the final probability was ≥ 0.5, the case was judged to be in the HCC class; and if it was 0.50 or less, the case was judged to be in the non-HCC class.

Test with the nonpaired case group. In addition, we evaluated the classifiers using the nonpaired case group. We used a model ensemble, which is a method for creating a better classifier using multiple classifiers in the field of ML. We chose the "Averaging" method, which averaged the predictions of all classifiers (Figure 2C).

Classifier

In this study, we adopted a DL model based on deep convolutional neural networks (CNNs), EfficientNetB0.[22] The classifier consisted of 9 layers, including 7 convolutional layers and 1 fully connected layer. This has been demonstrated in several tasks that analyze medical images. EfficientNet is a widely recognized network architecture in the field of ML, fundamentally designed to enhance performance efficiently and effectively by using scaling laws to balance the model's size, depth, and resolution. Specifically, EfficientNetB0, the initial model in

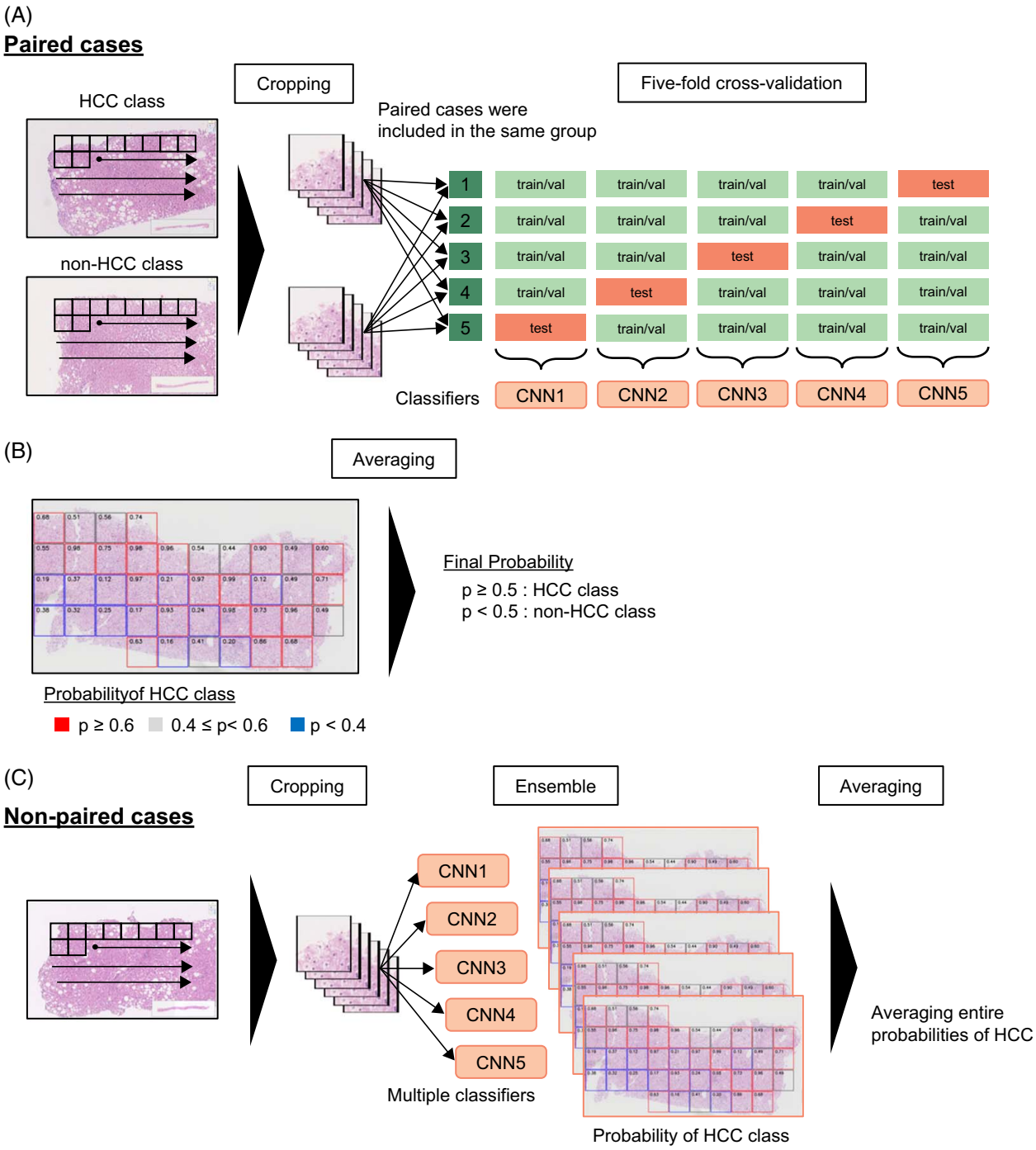

FIGURE 2 Schematic representation of the image-based deep learning model. (A) Method for model development using the paired case group. Patches were cut in order from top left to bottom right in a 256 × 256 size and divided into 5 groups. The cropped images from paired patients were included in the same group. Four of the 5 groups of cropped images were used as training and validation data sets, and the remaining group was used as the test set. After training, the classifier was evaluated using a test set. Five classifiers (CNN1–5) were created by repeating this process 5 times, using alternating training, validation, and testing sets. (B) Prediction of the probabilities of HCC development. The classifier based on deep convolutional neural networks (CNN) calculated the predictive cancerous value for the patches. The predicted probability of the HCC class is shown in each tile, where red represents ≥60%, gray represents between 40% and 60%, and blue represents <40%. The probability of each tile was averaged across the entire image to give the final probability. (C) Method for test with a model ensemble using the nonpaired case group. Each of the 5 classifiers calculated predictions for the test image in a similar way to (A). Those predictions were soft-ensembled into the final predictions of the image.

this architecture series, was chosen for its scalability and efficiency. This model is particularly suitable for processing large image data sets as it achieves high accuracy while minimizing the number of parameters and computational cost. EfficientNetB0 is designed using a unique method called Compound Scaling, which optimizes the

balance between model accuracy and efficiency by scaling the network's width, depth, and the resolution of the input images simultaneously. Furthermore, EfficientNetB0 has proven its versatility and reliability across various medical image analysis tasks. In our research, we used this model to predict the risk of HCC development from liver biopsy images of patients with fatty liver disease. The model was implemented using the PyTorch Lightning framework,[23] enabling efficient training on graphics processing units.

Metrics

The performance of the algorithm was evaluated using 3 indices: accuracy (number of correct predictions divided by the total number of predictions), receiver operating characteristic (ratio of true positives to false positives), and AUROC. The accuracy was calculated using the following formula:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}},$$

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

Tuning

Data augmentation. Data augmentation is used to improve the generalization performance of a model. For example, image data similar to the existing image data were created and added to the training data. This is a useful technique that can generate sufficient training data using a small amount of data without changing the model architecture. In this study, we used data augmentation techniques called Mixup (a weighted linear combination of 2 randomly selected images)[24] and Cutmix (which combines a Mixup with a rectangle masking a portion of a randomly selected image) during the training phase.[25] Specifically, the Mixup and Cutmix images were generated from 256 × 256 image patches and these images were used to train the classifier.

Hyperparameters and optimizers. The classifier was trained using cross-entropy loss with a learning rate of 0.0001 using the AdamW optimizer.[26]

Visualization

Class Activation Mapping is a well-known method for creating saliency maps that highlights the image parts the model focuses on, visualizing the model's inference basis. GradCam ++ was used to create the saliency map.[27] For a CNN with a Global Average Pooling layer, if the model determines that the input image is class c, the classification score Y^c can be calculated using the Global Average Pooling layer feature map A_{ij}^k as follows:

$$Y^c = \sum_k w_k^c \sum_i \sum_j A_{ij}^k,$$

where A_{ij}^k is the activity at position (i, j) in the kth channel.

This equation is transformed as follows:

$$Y^c = \sum_i \sum_j L_{ij}^c,$$

where

$$L_{ij}^c = \sum_k w_k^c A_{ij}^k,$$

L_{ij}^c can be considered as the saliency at position (i, j) of class c. From this, a saliency map can be created.

In GradCam++, w_k^c is defined as follows:

$$\alpha_{ij}^{kc} = \frac{\frac{\partial^2 Y^c}{\left(\partial A_{ij}^k\right)^2}}{2\frac{\partial^2 Y^c}{\left(\partial A_{ij}^k\right)^2} + \sum_a \sum_b A_{ab}^k \left\{\frac{\partial^3 Y^c}{\left(\partial A_{ij}^k\right)^3}\right\}}.$$

The method of the benchmark

Logistic regression was used as the benchmark test. Logistic regression is one of the most popular algorithms for solving binary classification problems. Let Y denote the binary response variable of interest and X denote the random variables considered as explanatory variables. The logistic regression model relates the conditional probability $P(Y=1|X)$ to X through

$$P(Y = 1, |, X) = \frac{\exp(\beta_0 + \beta_1 \text{X})}{1 + \exp(\beta_0 + \beta_1 \text{X})},$$

where β_0 and β_1 are regression coefficients, which are estimated by maximum-likelihood from the considered data set. The probability that $Y = 1$ for a new instance is then estimated by replacing the βs by their estimated counterparts and the Xs by their realizations for the considered new instance in the above equation. The new instance is then assigned to class $Y = 1$ if $P(Y = 1) > c$, where c is a fixed threshold and to class $Y = 0$ otherwise. The commonly used threshold is $c = 0.5$, which was also used in this study. We used the fibrosis stage as an explanatory variable because liver fibrosis is a key factor that determines the prognosis of chronic liver diseases, including SLD.[9,10] The benchmark

method for training and evaluation is similar to that used in our image-based classifier. In other words, the cases for training and evaluation matched our image-based classifier with the fibrosis stage-based classifier, and a model ensemble was implemented. The hypothesis that "the coefficient corresponding to a term is zero" was tested using the Wald test. These analyses were performed using the Statsmodels library (version 0.13.2) in the Python software.

Pathological evaluation

Liver histology was assessed according to Kleiner fibrosis stage (0 = none, 1a = mild perisinusoidal, 1b = moderate perisinusoidal, 1c = portal/periportal, 2 = zone 3 and portal/periportal, 3 = bridging fibrosis, and 4 = cirrhosis).[28] The grading of steatosis, ballooning, and lobular inflammation were graded using the NAFLD activity scoring system and the NASH Clinical Research Network criteria.[28,29] The histologic evaluation of the present data set was carried out by 3 board-certified liver pathologists (Rie Irie, Hanako Tsujikawa, and Michiie Sakamoto), who were blinded to the clinical data and patient outcomes, and this was meticulously validated.

Statistical analysis

Data are expressed as medians with the 25th to 75th percentiles or as numbers and percentages. Student *t* test or the Wilcoxon rank-sum test was used to analyze continuous variables. Differences between groups were assessed using the chi-square test or Fisher exact test for categorical data. The accuracy was compared using a proportion Z-test, and the significance of the difference in AUCs was evaluated using the DeLong test. The trend toward a higher probability of increase was evaluated using the Cochran-Armitage trend test. Statistical analyses were performed using R software (version 4.2.3; R Development Core Team), and p values <0.05 were considered significant.

RESULTS

Data set creation

Among 2432 patients with SLD, 639 were in the non-HCC class (those who did not develop HCC ≥ 7 y after biopsy) and 46 were in the HCC class (those who developed HCC <7 years after biopsy). To prevent learning about the differences in staining methods between facilities and the extent of color fading over time, one-to-one pairs were created from the non-HCC and HCC classes from the same institution with a biopsy date difference of <1 year. Finally, 58 cases (29 pairs of non-HCC and HCC) were extracted as paired case groups for the discovery set. Of the remaining nonpaired cases, 126 non-HCC and 17 HCC cases were used as nonpaired case groups for the test set (Figure 1).

Model development from the paired case group

The group of 29 pairs of non-HCC and HCC cases was used to develop and test the image-based DL model. Each pair of non-HCC and HCC cases was matched for both the time and facility at which the liver biopsy was performed. Detailed characteristics of the paired groups are presented in Table 1. The HCC class was significantly older; had higher bilirubin, aspartate aminotransferase, gamma-glutamyltransferase, alpha-fetoprotein, and FIB-4 index; and lower albumin and platelet counts.

The results of the cross-validation with the paired case group are shown in Figure 3A and Table 2. The average accuracy was 81.0% (95% CI: 0.71–0.88) and the AUC was 0.80 (95% CI: 0.69–0.92). The DL model was trained to correctly predict 8 out of 9 cases of HCC development in patients with mild fibrosis (F0-2) (Supplemental Figures S2A and S2B, http://links.lww.com/HEP/I431).

Test with the nonpaired case group

We assessed the robustness of our image-based DL model by testing it against a nonpaired case group. Detailed characteristics of the nonpaired groups are presented in Table 1. The patterns observed in clinical data were comparable to those observed in the paired case group.

The results of the test for the nonpaired case group are shown in Figure 3B and Table 2. The accuracy was 82.3% (95% CI: 0.75–0.88) and the AUC was 0.84 (95% CI: 0.72–0.96). Importantly, in patients with mild fibrosis (F0-2), the DL model was able to correctly classify 3 of 6 patients who developed liver cancer as HCC class (positive predictive value 0.50), while 91 of 100 patients who did not develop liver cancer were correctly classified as non-HCC class (negative predictive value 0.91) (Supplemental Figures S2C and S2D, http://links.lww.com/HEP/I431).

Visualization of pathological features involved in HCC development

Saliency maps were created to investigate where the DL model focused on predicting HCC development in

TABLE 1 Baseline characteristics

	Paired cases					Nonpaired cases				
Variable	**Non-HCC class**		**HCC class**		***p***	**Non-HCC class**		**HCC class**		***p***
Number of patients	29		29			126		17		
Age (y)	59.0	(50.0–65.0)	65.0	(59.0–70.0)	0.008	59.5	(46.5–66.0)	68.0	(65.0–74.0)	< 0.001
Males, n (%)	17	(58.6)	12	(41.4)	0.29	56	(44.4)	7	(41.2)	1.00
BMI (kg/m^2)	27.2	(24.3-29.0)	27	(26.1-30.0)	0.45	26.9	(24.3-29.8)	27.7	(23.9-28.8)	0.68
Diabetes mellitus, n (%)	15	(51.7)	23	(79.3)	0.05	69	(54.8)	15	(88.2)	0.01
Hypertension, n (%)	13	(44.8)	18	(62.1)	0.29	65	(51.6)	13	(76.5)	0.07
Dyslipidemia, n (%)	21	(72.4)	15	(51.7)	0.18	92	(73.0)	8	(47.1)	0.046
Alcohol consumers, n (%)	3	(10.3)	3	(10.3)	1.00	10	(7.9)	2	(11.8)	0.64
Albumin (g/dL)	4.3	(4.0–4.6)	4.0	(3.6–4.2)	0.01	4.4	(4.1–4.7)	4.0	(3.6–4.4)	0.001
Total bilirubin (mg/dL)	0.8	(0.7–1.1)	1.1	(0.9–1.3)	0.03	0.9	(0.7–1.1)	0.9	(0.7–1.4)	0.23
AST (IU/L)	41.0	(30.0–51.0)	61.0	(45.0–72.0)	0.001	43.5	(32.3–68.8)	54.0	(41.0–81.0)	0.13
ALT (IU/L)	48.0	(38.0–76.0)	55.0	(39.0–65.0)	0.89	65.5	(43.3–104.0)	42.0	(37.0–82.0)	0.12
GGT (IU/L)	52.0	(34.0–75.0)	87.0	(45.0–175.0)	0.04	58.5	(37.0–93.8)	142.0	(65.0–209.0)	0.01
Platelet count (×10^4/μL)	18.6	(15.2–22.9)	15.1	(11.1–17.2)	0.008	21.2	(15.9–25.3)	13.9	(11.3–15.2)	< 0.001
AFP (ng/mL)	3.3	(2.8–4.3)	7.7	(5.2–10.4)	< 0.001	4.1	(2.5–5.0)	4.2	(3.4–7.2)	0.18
FIB-4 index	1.76	(1.18–2.31)	4.19	(2.69–4.73)	< 0.001	1.54	(0.91–2.63)	4.20	(3.09–4.53)	< 0.001
Institution, n (%)					1.00					< 0.001
Saga University Hospital	0	(0.0)	0	(0.0)		1	(0.8)	6	(35.3)	
Kyoto Prefectural University of Medicine	0	(0.0)	0	(0.0)		0	(0.0)	2	(11.8)	
Hiroshima University	4	(13.8)	4	(13.8)		29	(23.0)	3	(17.6)	
The University of Tokyo	5	(17.2)	5	(17.2)		17	(13.5)	4	(23.5)	
Tokyo Women's Medical University	3	(10.3)	3	(10.3)		12	(9.5)	0	(0.0)	
Saiseikai Suita Hospital	17	(58,6)	17	(58,6)		67	(53.2)	2	(11.8)	

Note: Data are expressed as the median (25th–75th percentiles) or number (percentages). HCC class: Patients who developed HCC <7 y after biopsy. Non-HCC class: Patients who did not develop ≥ 7 y after biopsy. Alcohol consumers were defined as those who drink ≥ 30 g/d for men and ≥ 20 g/d for women.
Abbreviation: GGT, gamma-glutamyltransferase.

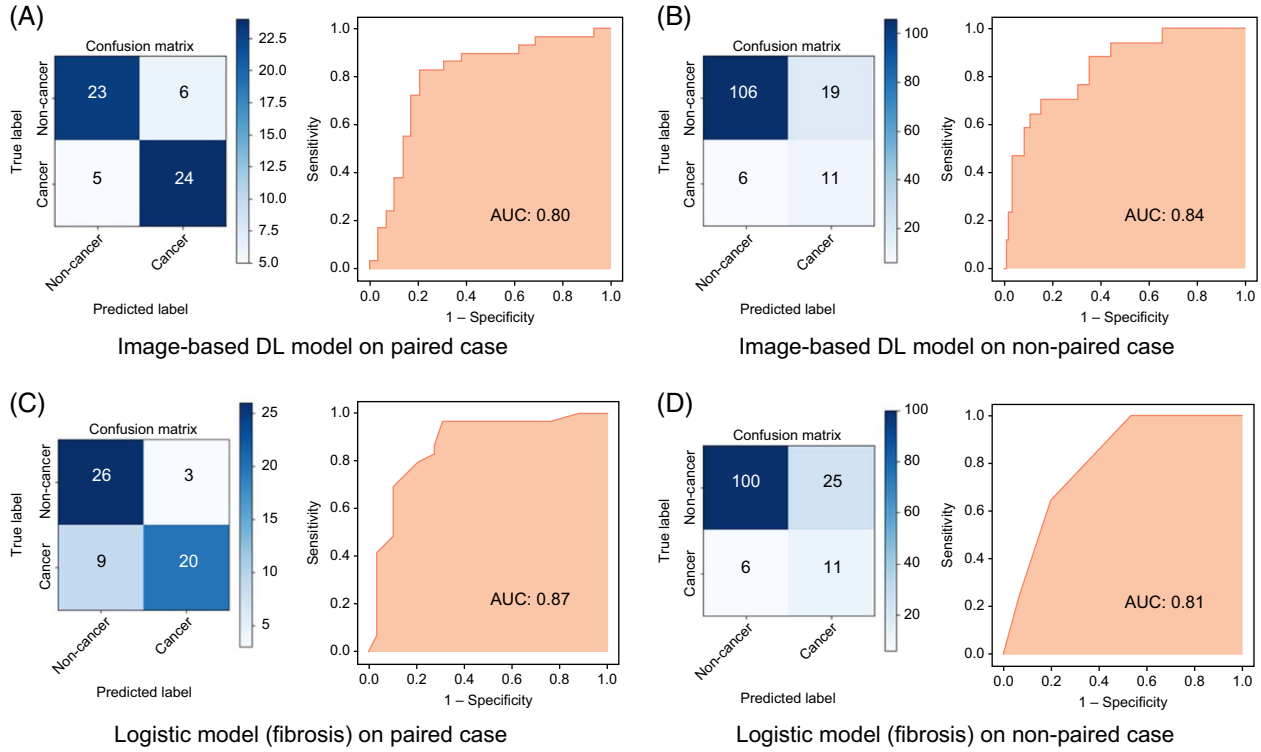

FIGURE 3 Predictive ability for HCC class. The confusion matrix and AUC for each group are shown for (A) predicted results of cross-validation using the image-based DL model on the paired case group; (B) predicted results of ensemble using the image-based DL model on the nonpaired case group as an additional test; (C) predicted results of cross-validation using the logistic regression model on the paired case group; and (D) predicted results of ensemble using the logistic regression model on the nonpaired case group as an additional test.

the images. Figure 4A shows saliency maps of the cropped images predicted to belong to the HCC class. Darker red color indicates more salient predicted tumorigenic features. Saliency tended to be higher in areas with nuclear atypia, hepatocytes with a high nuclear-cytoplasmic ratio, fibrosis, and high infiltration of immune cells. Figure 4B shows the saliency maps of the cropped images predicted to belong to the non-HCC class. The darker the red color, the more salient the predicted nontumorigenic features. The saliency tended to be higher in areas with large fat droplets.

Comparison with logistic regression model (fibrosis stage)

As liver fibrosis is the most useful prognostic factor in chronic liver disease, the pathological fibrosis stage was used as a variable in the benchmark model, logistic regression, and was compared with our image-based DL model. The results of the logistic regression analysis for cross-validation with the paired case group are shown in Figure 3C and Table 2. The accuracy was 79.3% (95% CI: 0.69–0.90) and the AUC was 0.87 (95% CI: 0.77–0.96). It implies that the probability of the HCC class increases according to increased fibrosis stage (coefficient: 1.82, 95% CI: 0.95–2.69, $p < 0.01$). There were no significant differences in accuracy (Z value: 0.23, $p = 0.82$) or AUC ($p = 0.30$) between the image-based DL model and the fibrosis model.

The results of the logistic regression analysis for the test with the nonpaired case group are shown in Figure 3D and Table 2. The accuracy was 78.2% (95% CI: 0.71–0.85) and the AUC was 0.81 (95% CI: 0.68–0.94). There were no significant differences in accuracy (Z value 0.89, $p = 0.37$) or AUC ($p = 0.48$) between the image-based DL model and the fibrosis model.

The detailed pathological findings are shown in Table 3. The HCC class tended to have higher grades of fibrosis, lobular and portal inflammation, and ballooning, and lower steatosis stages. Mallory-Denk bodies and small- and large-cell dysplasias were detected in a higher proportion in the HCC class.

DISCUSSION

In this study, we developed a novel DL-based CNN-assisted system to predict HCC development in patients with SLD using an algorithm trained with limited image data (H&E-stained WSI of liver biopsy samples). The predictive ability of the image-based DL model was comparable to that of the logistic regression model using pathological fibrosis stage as an explanatory variable. Furthermore, the DL model can predict HCC

TABLE 2 Predictive performances for HCC development

	Image-based DL model				Fibrosis stage			
	Accuracy (mean)	Accuracy (95% CI)	AUC (mean)	AUC (95% CI)	Accuracy (mean)	Accuracy (95% CI)	AUC (mean)	AUC (95% CI)
Paired cases (model development)	81.0%	(0.71–0.88)	0.80	(0.69–0.92)	79.3%	(0.69–0.90)	0.87	(0.77–0.96)
Nonpaired cases (test set)	82.3%	(0.75–0.88)	0.84	(0.72–0.96)	78.2%	(71–0.85)	0.81	(0.68–0.94)

Abbreviation: DL, deep learning.

development even in patients with mild liver fibrosis. Our findings demonstrate the potential of DL-based image analysis in predicting HCC development in patients with SLD. Studies have shown the promise of DL-based approaches in improving the reliability of the histological evaluation of NASH,[13,14] as well as identifying novel histological features associated with disease progression and fibrosis severity.[15] However, this study specifically examines the utility of DL models in predicting HCC development directly from histopathological images of patients with SLD.

The use of ML in the analysis of pathological images for predicting carcinogenesis and prognosis has attracted significant attention in the scientific community. Numerous studies have explored the potential of ML algorithms, particularly DL techniques, to demonstrate their efficacy and relevance. In the hepatology practice. The combination of DL and digital pathology could lead to transformative changes in the clinical practice of hepatology, including standardization of pathological diagnoses of SLD; personalized risk stratification and efficient allocation of medical resources; provision of remote diagnosis, improving medical access, and reducing regional disparities; and the potential discovery of new insights into the progression and prognosis of SLD. In the context of HCC prognostication, several studies have demonstrated the potential of ML algorithms for predicting HCC prognosis and recurrence based on histopathological images. Saillard et al[17] developed a model using HCC digital slides that outperformed traditional scores in predicting patient survival following surgical resection. Yamashita et al[30] confirmed the effectiveness of ML algorithms in predicting outcomes using HCC digital slides. Lu and Daigle[31] used advanced CNNs for feature extraction from the histopathological slides of HCC. Saito et al[32] achieved promising results in predicting HCC recurrence using handcrafted WSI features. Despite the growing interest in prognostic and recurrence prediction using HCC digital slides, no studies have investigated the predictive potential of HCC development using noncancerous liver tissue.

Progression from chronic liver disease to hepatocarcinogenesis involves continued cell death and regeneration associated with inflammation, leading to an increased potential for carcinogenesis.[33] Liver fibrosis is an indicator of the accumulation of such changes and is an important factor in assessing the hepatocarcinogenic potential of chronic liver diseases, including NAFLD.[9,10] However, in NASH-related liver pathology, multiple changes occur, including not only fibrosis but also inflammation, steatosis, and ballooning.[34] DL algorithms harness the power to analyze complex patterns and features in medical images, allowing them to potentially capture subtle details that may not be easily recognizable by the human eye.[16,17] This can

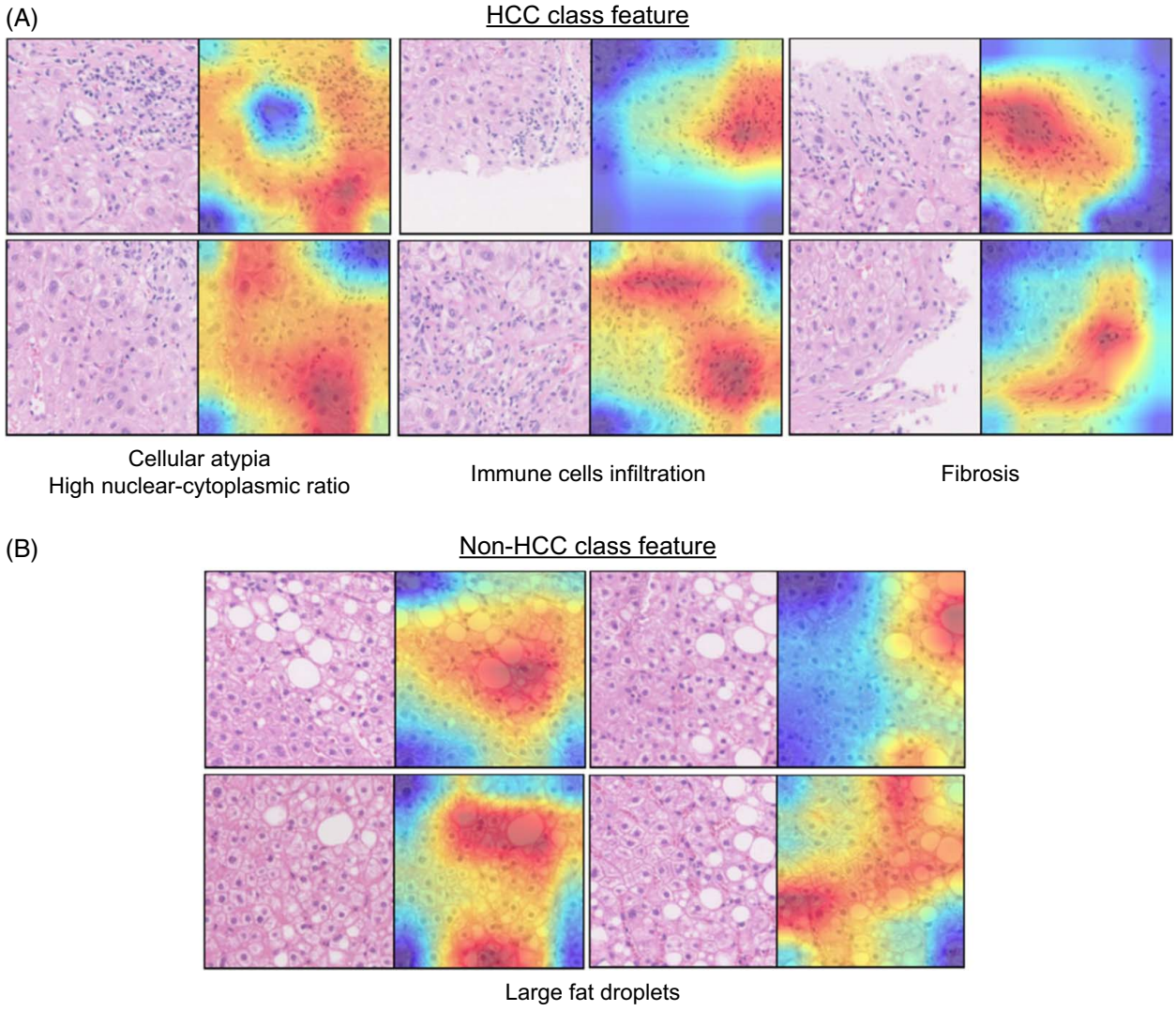

FIGURE 4 Saliency map. (A) Predicted HCC class cropped images (left images) and its saliency maps (right images). The darker the red color, the more salient the predicted tumorigenic feature is. (B) Predicted non-HCC class cropped images (left images) and its saliency maps (right images). The darker the red color, the more salient the predicted nontumorigenic feature is.

enable the DL model to provide a more comprehensive assessment of liver pathology and to identify early signs of liver carcinogenesis that may not be evident from fibrosis staging alone.

In this study, we created saliency maps using the GradCam++ method to visualize the pathological features involved in HCC development. These maps highlight the areas of the image on which the DL model focused during the inference process. The saliency map indicated that the DL model focused on pathological findings other than fibrosis, such as nuclear atypia, hepatocytes with a high nuclear-cytoplasmic ratio, and high infiltration of immune cells, as features which lead to HCC development. Notably, a detailed evaluation of the pathological findings revealed that inflammation, ballooning, Mallory-Denk bodies, and small- and large-cell dysplasia were detected at high rates in the HCC class. The DL model also focuses on steatosis, suggesting a low risk of carcinogenesis. This may reflect the burned-out of hepatic steatosis in patients with a high risk of carcinogenesis.[35] Although the short-term risk of HCC development is higher in patients with advanced fibrosis, the results reiterate that ballooning, inflammation, and nuclear atypia are also important findings in the subsequent long-term exacerbation of SLD. The importance of these pathological findings may have been demonstrated in the present study, as the control group consisted of patients who did not develop HCC for at least 7 years.

The strengths of this study are: First, while the assessment of fibrosis and other findings related to NASH may vary between institutions and pathologists,[36] the DL model can predict the risk of developing HCC using only H&E-stained slides, allowing for risk stratification for HCC without special staining to assess liver fibrosis. Furthermore, it is possible to predict HCC

TABLE 3 Distributions of pathological findings

Variable	All cases				
	Non-HCC class		HCC class		*p*
No. of patients	155		46		
Fibrosis stage, n (%)					< 0.001
0	12	(7.7)	0	(0.0)	
1	66	(42.6)	1	(2.2)	
2	48	(31.0)	14	(30.4)	
3	19	(12.3)	18	(39.1)	
4	9	(5,8)	13	(28.3)	
Steatosis, n (%)					0.008
0	3	(1.9)	3	(6,5)	
1	62	(40.0)	25	(54.3)	
2	53	(34.2)	13	(28.3)	
3	37	(23.9)	5	(10.9)	
Lobular inflammation, n (%)					< 0.001
0	9	(5.8)	1	(2.2)	
1	135	(87,1)	29	(63.0)	
2	11	(7.1)	16	(34.8)	
3	0	(0.0)	0	(0.0)	
Portal inflammation, n (%)					< 0.001
0	4	(2.6)	1	(2.2)	
1	110	(71.0)	18	(39.1)	
2	41	(26.5)	22	(57.8)	
3	0	(0.0)	5	(10.9)	
Ballooning, n (%)					< 0.001
0	56	(36.1)	5	(10.9)	
1	54	(34.8)	17	(37.0)	
2	45	(29.0)	24	(52.2)	
Mallory-Denk bodies, n (%)					0.002
No	97	(62.6)	17	(37.0)	
Yes	58	(37.4)	29	(63.0)	
Small cell dysplasia, n (%)					0.13
No	152	(98.1)	43	(93.5)	
Yes	3	(1.9)	3	(6.5)	
Large cell dysplasia, n (%)					0.08
No	138	(89.0)	36	(78.3)	
Yes	17	(11.0)	10	(21.7)	

Note: Data are expressed as number (percentages). Non-HCC class: Patients who did not develop HCC ≥ 7 y after biopsy. HCC class: Patients who developed HCC <7 y after biopsy. The trend toward a higher probability of increase was evaluated using the Cochran-Armitage trend test.

development from SLD with mild fibrosis, since pathological factors other than fibrosis are also considered. Second, as DL models have been created using H&E-stained slides produced in different years with different staining protocols at different sites, our results have the potential to be generalized across diverse clinical settings and institutions. Finally, our cohort includes ~10% of patients with SLD who consume alcohol (≥ 30 g/d in men and ≥ 20 g/d in women); therefore, the results of this study may be widely applicable in patients with SLD who consume mild amounts of alcohol (ie, MetALD).

This study has some limitations. First, the study was retrospective, which may have introduced bias and limited causal inferences. Prospective studies with larger sample sizes are required to validate our findings. Second, because this study focused on a specific Japanese population of patients with suspected SLD, future studies should aim to validate the model using diverse patient cohorts from around the world. Addressing these limitations through further research and technological advancements will help refine and expand the utility of DL-based models for predicting HCC development in patients with SLD.

In conclusion, our study demonstrated the potential of DL-based image analysis for predicting HCC development in patients with SLD. The DL model revealed good performance in identifying patients at high risk of HCC, which has important implications for clinical decision-making and patient management. Further research is required to validate and refine the DL model and explore its integration into clinical practice to improve the care and outcomes of patients with SLD.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study will be made available by the corresponding author upon reasonable request.

AUTHOR CONTRIBUTIONS

All authors contributed to the study design and data interpretation. Takuma Nakatsuka, Ryosuke Tateishi, Masaya Sato, Mitsuhiro Fujishiro, and Kazuhiko Koike contributed to the study conception. Yoshio Sumida, Masashi Yoneda, Norio Akuta, Takumi Kawaguchi, Hirokazu Takahashi, Yuichiro Eguchi, Yuya Seko, Yoshito Itoh, Eisuke Murakami, Kazuaki Chayama, Makiko Taniai, Katsutoshi Tokushige, and Takeshi Okanoue contributed to data acquisition. Natsuka Hashizume, Ami Kamada, Hiroki Nakano, Yoshinori Kabeya, Sho Yonezawa, Rie Irie, Hanako Tsujikawa, and Michiie Sakamoto contributed to data analysis. Takuma Nakatsuka, Masaya Sato, and Ryosuke Tateishi drafted the manuscript, and all authors revised it critically for important intellectual content. All the authors approved the final manuscript and agreed to be accountable for all aspects of the work, ensuring that questions related to the accuracy or integrity of any part of the work were appropriately investigated and resolved.

ACKNOWLEDGMENTS

The authors thank Editage (www.editage.jp) for English language editing.

FUNDING INFORMATION

This research was supported by the Research Program on Hepatitis from the Japan Agency for Medical Research and Development, AMED, under Grant Number JP23fk0210090 and JP24fk0210149; Health, Labour, and Welfare Policy Research Grants from the Ministry of Health, Labour, and Welfare of Japan under Grant Number 23HC2001.

CONFLICTS OF INTEREST

Yoshio Sumida is on the speakers' bureau for Kowa, MSD, and Taisho. Hirokazu Takahashi received grants from Astellas and Sysmex. The remaining authors have no conflicts to report.

ORCID

Takuma Nakatsuka https://orcid.org/0000–0002–5727–5385
Ryosuke Tateishi https://orcid.org/0000–0003–3021–2517
Yoshio Sumida https://orcid.org/0000–0002–4342–1361
Takumi Kawaguchi https://orcid.org/0000–0002–7064–4325
Hirokazu Takahashi https://orcid.org/0000–0003–1900–4389
Yuya Seko https://orcid.org/0000–0002–3658–5894
Yoshito Itoh https://orcid.org/0000–0001–9890–3635
Eisuke Murakami https://orcid.org/0000–0003–0357–2795
Kazuaki Chayama https://orcid.org/0000–0002–5530–5341
Takeshi Okanoue https://orcid.org/0000–0002–2390–3400
Mitsuhiro Fujishiro https://orcid.org/0000–0002–4074–1140

REFERENCES

1. Younossi Z, Tacke F, Arrese M, Chander Sharma B, Mostafa I, Bugianesi E, et al. Global perspectives on nonalcoholic fatty liver disease and nonalcoholic steatohepatitis. Hepatology. 2019;69:2672–82.
2. Cotter TG, Rinella M. Nonalcoholic fatty liver disease 2020: The state of the disease. Gastroenterology. 2020;158:1851–64.
3. Huang DQ, Singal AG, Kono Y, Tan DJH, El-Serag HB, Loomba R. Changing global epidemiology of liver cancer from 2010 to 2019: NASH is the fastest growing cause of liver cancer. Cell Metab. 2022;34:969–977 e962.
4. Younossi ZM. Non-alcoholic fatty liver disease—A global public health perspective. J Hepatol. 2019;70:531–44.
5. Browning JD, Szczepaniak LS, Dobbins R, Nuremberg P, Horton JD, Cohen JC, et al. Prevalence of hepatic steatosis in an urban population in the United States: Impact of ethnicity. Hepatology. 2004;40:1387–95.
6. Ludwig J, Viggiano TR, McGill DB, Oh BJ. Nonalcoholic steatohepatitis: Mayo Clinic experiences with a hitherto unnamed disease. Mayo Clin Proc. 1980;55:434–8.
7. Simon TG, Roelstraete B, Khalili H, Hagstrom H, Ludvigsson JF. Mortality in biopsy-confirmed nonalcoholic fatty liver disease: Results from a nationwide cohort. Gut. 2021;70:1375–82.
8. Hagström H, Nasr P, Ekstedt M, Hammar U, Stål P, Hultcrantz R, et al. Fibrosis stage but not NASH predicts mortality and time to development of severe liver disease in biopsy-proven NAFLD. J Hepatol. 2017;67:1265–73.
9. Ekstedt M, Hagström H, Nasr P, Fredrikson M, Stål P, Kechagias S, et al. Fibrosis stage is the strongest predictor for disease-specific mortality in NAFLD after up to 33 years of follow-up. Hepatology. 2015;61:1547–54.
10. Angulo P, Kleiner DE, Dam-Larsen S, Adams LA, Bjornsson ES, Charatcharoenwitthaya P, et al. Liver fibrosis, but no other histologic features, is associated with long-term outcomes of patients with nonalcoholic fatty liver disease. Gastroenterology. 2015;149:389–397.e310.
11. Yasui K, Hashimoto E, Komorizono Y, Koike K, Arii S, Imai Y, et al. Characteristics of patients with nonalcoholic steatohepatitis who develop hepatocellular carcinoma. Clin Gastroenterol Hepatol. 2011;9:428–33; quiz e450.
12. Foerster F, Gairing SJ, Muller L, Galle PR. NAFLD-driven HCC: Safety and efficacy of current and emerging treatment options. J Hepatol. 2022;76:446–57.
13. Soon GST, Liu F, Leow WQ, Wee A, Wei L, Sanyal AJ. Artificial intelligence improves pathologist agreement for fibrosis scores in nonalcoholic steatohepatitis patients. Clin Gastroenterol Hepatol. 2023;21:1940–42 e1943.
14. Taylor-Weiner A, Pokkalla H, Han L, Jia C, Huss R, Chung C, et al. A machine learning approach enables quantitative measurement of liver histology and disease monitoring in NASH. Hepatology. 2021;74:133–47.
15. Loomba R, Noureddin M, Kowdley KV, Kohli A, Sheikh A, Neff G, et al. Combination therapies including cilofexor and firsocostat for bridging fibrosis and cirrhosis attributable to NASH. Hepatology. 2021;73:625–43.
16. Yamamoto Y, Tsuzuki T, Akatsuka J, Ueki M, Morikawa H, Numata Y, et al. Automated acquisition of explainable knowledge from unannotated histopathology images. Nat Commun. 2019;10:5642.
17. Saillard C, Schmauch B, Laifa O, Moarii M, Toldo S, Zaslavskiy M, et al. Predicting survival after hepatocellular carcinoma resection using deep learning on histological slides. Hepatology. 2020;72:2000–13.
18. Rinella ME, Lazarus JV, Ratziu V, Francque SM, Sanyal AJ, Kanwal F, et al. A multisociety Delphi consensus statement on new fatty liver disease nomenclature. J Hepatol. 2023;79:1542–56.
19. Rinella ME, Lazarus JV, Ratziu V, Francque SM, Sanyal AJ, Kanwal F, et al. A multisociety Delphi consensus statement on new fatty liver disease nomenclature. Hepatology. 2023;78:1966–86.
20. Tellez D, Litjens G, Bándi P, Bulten W, Bokhorst JM, Ciompi F, et al. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. Med Image Anal. 2019;58:101544.
21. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In Ijcai. 1995;14:1137–45.
22. Tan M, Lw Q Efficientnet: Rethinking model scaling for convolutional neural networks. *International Conference on Machine Learning* Proceedings of Machine Learning Research, MLResearchPress; 2019:6105–14.
23. Falcon W. The PyTorch Lightning team. PyTorch Lightning. 2019. Accessed November 30, 2023. https://doi.org/10.5281/zenodo.3828935https://github.com/PyTorchLightning/pytorch-lightning

24. Zhang H, Cisse M, Dauphin YN, Lopez-Paz D mixup: Beyond empirical risk minimization. In: *International Conference on Learning Representations*; 2018. https://doi.org/10.48550/arXiv.1710.09412
25. Yun S, Han D, Oh SJ, Chun S, Choe J, Yoo Y CutMix: Regularization strategy to train strong classifiers with localizable features. Proceedings of the *IEEE/CVF International Conference on Computer Vision (ICCV)*; 2019:6023–32.
26. Loshchilov I, Hutter F Decoupled weight decay regularization. In: *International Conference on Learning Representations*; 2019. https://doi.org/10.48550/arXiv.1711.05101
27. Chattopadhay A, Sarkar A, Howlader P, Balasubramanian VN Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. Proceedings of the 2018 *IEEE Winter Conference on Applications of Computer Vision (WACV)*; 2018:839–7.
28. Kleiner DE, Brunt EM, Van Natta M, Behling C, Contos MJ, Cummings OW, et al. Design and validation of a histological scoring system for nonalcoholic fatty liver disease. Hepatology. 2005;41:1313–21.
29. Brunt EM, Kleiner DE, Wilson LA, Unalp A, Behling CE, Lavine JE, et al. Portal chronic inflammation in nonalcoholic fatty liver disease (NAFLD): A histologic marker of advanced NAFLD-clinicopathologic correlations from the nonalcoholic steatohepatitis clinical research network. Hepatology. 2009;49:809–20.
30. Yamashita R, Long J, Saleem A, Rubin DL, Shen J. Deep learning predicts postsurgical recurrence of hepatocellular carcinoma from digital histopathologic images. Sci Rep. 2021;11:2047.
31. Lu L, Daigle BJ Jr. Prognostic analysis of histopathological images using pre-trained convolutional neural networks: Application to hepatocellular carcinoma. PeerJ. 2020;8:e8668.
32. Saito A, Toyoda H, Kobayashi M, Koiwa Y, Fujii H, Fujita K, et al. Prediction of early recurrence of hepatocellular carcinoma after resection using digital pathology images assessed by machine learning. Mod Pathol. 2021;34:417–25.
33. Llovet JM, Kelley RK, Villanueva A, Singal AG, Pikarsky E, Roayaie S, et al. Hepatocellular carcinoma. Nat Rev Dis Primers. 2021;7:6.
34. Matteoni CA, Younossi ZM, Gramlich T, Boparai N, Liu YC, McCullough AJ. Nonalcoholic fatty liver disease: A spectrum of clinical and pathological severity. Gastroenterology. 1999;116:1413–9.
35. Vilar-Gomez E, Calzadilla-Bertot L, Wai-Sun Wong V, Castellanos M, Aller-de la Fuente R, Metwally M, et al. Fibrosis severity as a determinant of cause-specific mortality in patients with advanced nonalcoholic fatty liver disease: A multi-national cohort study. Gastroenterology. 2018;155:443–457.e417.
36. Kleiner DE, Brunt EM, Wilson LA, Behling C, Guy C, Contos M, et al. Association of histologic disease activity with progression of nonalcoholic fatty liver disease. JAMA Netw Open. 2019;2:e1912565.

How to cite this article: Nakatsuka T, Tateishi R, Sato M, Hashizume N, Kamada A, Nakano H, et al. Deep learning and digital pathology powers prediction of HCC development in steatotic liver disease. Hepatology. 2025;81:976–989. https://doi.org/10.1097/HEP.0000000000000904