J Toxicol Pathol 2023; 36: 21–30

Original Article

Application of convolutional neural network for analyzing hepatic fibrosis in mice

Hyun-Ji Kim[1,2†], Eun Bok Baek[2†], Ji-Hee Hwang[1], Minyoung Lim[1], Won Hoon Jung[3], Myung Ae Bae[3], Hwa-Young Son[2*], and Jae-Woo Cho[1*]

[1] Toxicological Pathology Research Group, Department of Advanced Toxicology Research, Korea Institute of Toxicology, 141 Gajeong-ro, Yuseong-gu, Daejeon 34114, Republic of Korea
[2] College of Veterinary Medicine, Chungnam National University, 99 Daehak-ro, Yuseong-gu, Daejeon 34134, Republic of Korea
[3] Therapeutics & Biotechnology Division, Korea Research Institute of Chemical Technology, 141 Gajeong-ro, Yuseong-gu, Daejeon 34114, Republic of Korea

Abstract: Recently, with the development of computer vision using artificial intelligence (AI), clinical research on diagnosis and prediction using medical image data has increased. In this study, we applied AI methods to analyze hepatic fibrosis in mice to determine whether an AI algorithm can be used to analyze lesions. Whole slide image (WSI) Sirius Red staining was used to examine hepatic fibrosis. The Xception network, an AI algorithm, was used to train normal and fibrotic lesion identification. We compared the results from two analyses, that is, pathologists' grades and researchers' annotations, to observe whether the automated algorithm can support toxicological pathologists efficiently as a new apparatus. The accuracies of the trained model computed from the training and validation datasets were greater than 99%, and that obtained by testing the model was 100%. In the comparison between analyses, all analyses showed significant differences in the results for each group. Furthermore, both normalized fibrosis grades inferred from the trained model annotated the fibrosis area, and the grades assigned by the pathologists showed significant correlations. Notably, the deep learning algorithm derived the highest correlation with the pathologists' average grade. Owing to the correlation outcomes, we conclude that the trained model might produce results comparable to those of the pathologists' grading of the Sirius Red-stained WSI fibrosis. This study illustrates that the deep learning algorithm can potentially be used for analyzing fibrotic lesions in combination with Sirius Red-stained WSIs as a second opinion tool in non-clinical research. (DOI: 10.1293/tox.2022-0066; J Toxicol Pathol 2023; 36: 21–30)

Key words: artificial intelligence, hepatic fibrosis, Xception network, digital pathology

Introduction

The prevalence of nonalcoholic fatty liver disease (NAFLD) is rapidly increasing worldwide, and it is now the most common liver disorder in the Western world[1]. NAFLD is characterized by excess fat deposition (steatosis) in the liver[2]. NAFLD can progress to non-alcoholic steatohepatitis (NASH), a disease in which the liver is additionally affected by varying degrees of cell death, inflammation, and collagen deposition[3]. NASH has evolved to be a potential target for therapeutics owing to its distinct lesion of hepatic fibrosis that may affect cardiovascular comorbidity, malignancy, and mortality[4]. Therefore, the discovery of antifibrotic therapeutics has gained considerable attention for NASH treatment[5].

Animal models of liver disease are generally required to study the efficacy of novel compounds in nonclinical research. Frequently used models are based on rats or mice, where pathophysiology comparable to NAFLD/NASH is induced by a high-fat diet or substances, such as CCL_4[6]. In an efficacy study using animal models of NAFLD/NASH, frequently used evaluations were biochemical parameters, quantitative image analysis, and examination of histopathological sections[7]. Human-based histopathological examination, which is widely accepted, has several drawbacks. First, it relies on expert pathologists, who are in-demand[8]. The task is time-consuming and can be exhaustive, which may affect the performance. Second, the produced results have low reproducibility, that is, they exhibit inherent variability between different pathologists and the same pathologist[9]. This could limit the comparability of the results. Lastly, because of the complex pattern of NAFLD/NASH, a more detailed and subdivided grading system is needed to monitor histopathological changes in NAFLD/NASH[10]. This can limit the

Received: 22 June 2022, Accepted: 7 September 2022
Published online in J-STAGE: 13 October 2022
*Corresponding authors: JW Cho (e-mail: cjwoo@kitox.re.kr); HY Son (e-mail: hyson@cnu.ac.kr)
†These authors contributed equally to this work.
(Supplementary material: refer to PMC https://www.ncbi.nlm.nih.gov/pmc/journals/1592/)

usefulness of pathological assessment in clinical practice.

Recent advances in deep learning, particularly in convolutional neural networks (CNNs), a type of deep learning used in image recognition, have revolutionized image analysis[11]. CNNs generally have three layers: convolution, pooling, and fully-connected[12]. In Xception, an extreme model of Inception, the cross-channel correlations and spatial correlations are mapped more effectively than those in inception. Since the hypothesis of Xception architecture is that cross-channel correlations and spatial correlations are adequately detached, which is more advisable, Chollet assumed that the entire detaching of correlations is possible, and it is a stronger hypothesis than that of Inception[13]. Some studies have applied deep learning methods to detect lesions of interest in pathology image samples derived from animal studies for drug discovery[14]. Studies have successfully classified lesions or performed pathology scoring using rodent models, such as pulmonary pathology related to pulmonary tuberculosis[15], hyperplasia[16], vacuole quantification in the liver for non-alcoholic fatty liver disease (NAFLD)[17], and non-alcoholic steatohepatitis (NASH) scoring[7]. These studies mainly used hematoxylin and eosin (H&E)-stained slides[18], whereas few studies have used other staining solutions. Sirius Red staining is favorably used for the clear discrimination of fibrotic lesions. Therefore, in the current study, we applied an image classification method using a deep learning algorithm to confirm the successful analysis of fibrotic lesions in mice by Sirius Red staining[19].

Materials and Methods

Animal

Six-week-old male C57BL/6 mice were used for this study. Animals were procured from Orient Bio, Inc. (Seongnam-si, Gyeonggi-do, Republic of Korea). All experimental procedures were performed in accordance with the regulations of the Institutional Animal Care and Use Committee (Authorization No.2020-7A-02-02) of the Korea Research Institute of Chemical Technology (Daejeon, Republic of Korea). The animals had access to filtered, ultraviolet light-irradiated municipal tap water *ad libitum*. Drinking water was analyzed every 6 months for the presence of specified contaminants at the Daejeon Regional Institute of Health and Environment (Daejeon, Republic of Korea).

Animal experiment

A total of 33 mice were used in this study, with at least 10 animals in each group. All animals were allowed to acclimatize for 7 days to adjust to the laboratory environment. During 9 weeks post the acclimation period (16-week-old male C57BL/6 mice), animals were provided with a diet *ad libitum*. The selection criteria for animals in this study were based on the adequacy of their body weight. Mice in the positive control group, high-fat diet with CCL_4 treatment (HFDC) group, and vehicle control (VC)-administered group (HFDC+VC group) were intraperitoneally (i.p.) administered 3 mL/kg carbon tetrachloride (CCL_4) (Cat. # 289116-1L; Sigma-Aldrich, St. Louis, MO, USA) in corn oil (Cat. # C8267-500mL; Sigma-Aldrich) solution (1:49 corn oil) twice per week for 4 weeks. After 2 weeks of CCL_4 treatment, sodium carboxymethyl cellulose (CMC, vehicle) (Cat. # 37142-02; Kanto Chemical Co., Inc., Tokyo, Japan) was administered daily by oral gavage for two weeks. Mice in the test group, high-fat diet with CCL_4 treatment (HFDC) group, and elafibranor (ELA)-treated group (HFDC+ELA group) were intraperitoneally administered a corn oil solution containing 3 mL/kg (1:49 corn oil) CCL_4 at the same dose and frequency as that in the HFDC+VC group. After 2 weeks of CCL_4 treatment, ELA (Cat. # 561377; MedKoo Biosciences, Inc., Morrisville, NC, USA) with CMC was administered daily by oral gavage for the last 2 weeks. Treated male C57BL/6 mice were fed a high-fat diet (HFD, trans-fat 60 kcal%) (#D12492; Research Diets, Inc., New Brunswick, NJ, USA) *ad libitum*, whereas the control group was fed a normal diet (4.5% fat, 20.12% protein, 3.5% fiber; Purina, Gunpo, Republic of Korea). Animals were deprived of food on the day of sacrifice. After sacrifice, liver tissues were collected from each animal, preserved in 10% neutral buffered formalin, and supplied to the Korea Institute of Toxicology (Daejeon, Republic of Korea) for subsequent histopathological examination.

Tissue special staining

Sirius Red staining was performed to analyze the collagen fibers in the liver. The Picro Sirius Red stain kit (connective tissue stain) (ab150681) was purchased from Abcam PLC, Cambridge, United Kingdom. The staining procedure was performed according to the following protocol: the initial step involved the removal of wax and hydrating paraffin sections. After staining the nuclei with hematoxylin solution (Cat. # H3136; Sigma-Aldrich) for 8 min, the slides were washed under running water for 10 min. The Picro Sirius red solution required a complete hour for staining, as shorter times could lead to poor results. After washing the slides twice with acidified water, most of the water was manually removed, and the samples were dehydrated thrice using 100% ethanol. Finally, the sample was rendered transparent using xylene and mounted on resinous medium. Following Sirius Red staining, collagen appeared red under light microscopy, indicating that the staining was highly specific for collagen.

Whole slide image sample

To obtain a whole slide image (WSI) sample, mice were treated with CCL_4 to induce fibrosis[20], while ELA[21] was used as an antifibrotic therapeutic. The liver was collected after sacrificing the animals, embedded in a paraffin block, and stained using Sirius Red stain. Slides were digitized using an Aperio Scanscope XT (Leica Biosystems, Wetzlar, Germany) with a 20× objective under bright-field illumination. The scan resolution was 0.5 μm per pixel, and the images were saved as TIFF stripes with JPEG2000 image compression. A total of 33 Sirius Red-stained WSIs of mouse hepatic fibrosis were used in this study.

Data preparation for model training and WSI analysis

There were two types of datasets; one was used during the training stage of the model, and the other was used for analyzing the model. The previous data were divided into three sets. First, a training dataset was used to train an algorithm for the input. Second, a validation dataset was used to tune the final model when it was overfitted. Finally, the test dataset was used to evaluate how well the algorithm was trained. The last dataset, called the analyzing dataset, was used to analyze the WSI for practical use.

Mouse hepatic slides (n=33) were scanned using an Aperio Scanscope XT, and 33 WSIs were produced. We magnified the obtained WSI 20× using the Aperio ImageScope ver. 12.4.0.5043 (Leica Biosystems, Buffalo Grove, IL, USA) and captured images in multiples of 128 × 128 using AlCapture (ESTsoft, Seoul, Republic of Korea). The captured images were cropped to 128 × 128 pixel-size images using the PhotoScape v3.7 (MOII Tech, Seoul, Republic of Korea) program.

Cropped images from 13 out of 33 WSIs were classified as normal or fibrotic by a pathologist. The classified data were split into the training, validation, and test sets (Table 1). The total number of WSIs used during the model's training stage was 13, of which five were from the control group, four from the HFDC+VC group, and four from the HFDC+ELA group. For the data distribution for the test, 1,000 images each from fibrosis and normal tissue groups were selected randomly using the "random.sample" program within the Python package. The training and validation datasets were divided in an 8:2 ratio using the TensorFlow addons package.

Twenty of the 33 WSIs were used for analysis, of which five were from the control group, nine were from the HFDC+VC group, and six were from the HFDC+ELA group. Cropped images were used to analyze WSIs using deep learning.

Algorithm

To implement the deep learning algorithm for screening lesions from WSIs, we selected a classification algorithm. The classification algorithm is the basic model applied in machine vision, and its performance is suitable for fast and precise prediction[12]. Among the several classification networks available in open-source deep-learning frameworks, we used the Xception network. The Xception network has advantages over the Inception-v3 module along the same parameter count as Inception-v3[13]. Network performance is not only precise but also involves fewer calculations compared to other algorithms[22]. Moreover, previous studies have demonstrated better performance in the Intel Image Classification Challenge dataset[23] and clinical studies[24–26] using Xception compared with that obtained using other deep learning algorithms. Therefore, we assumed that Xception would be suitable for screening lesions in WSIs.

All the procedures related to the stages of the training of the model to classify mouse hepatic fibrosis were performed using the TensorFlow 2.1.0 package, installed with Python version 3.7, the requirements of which were met in this study. A single NVIDIA RTX 2080 super GPU was used for all calculations in the training and inference tests.

After training the model with 13 WSIs, we analyzed 20 WSIs using the trained model. Twenty WSIs were cropped in the same flow of preparation data for the model training. To obtain the correlation results between the other analyzed methods, we classified normal and fibrosis images using the 20 WSIs generated from each slide with the help of the trained model and computed the proportion of the number of images classified as fibrosis out of the total cropped images for each WSI (Fig. 1A).

WSI evaluation by pathologists and a researcher

Three certified toxicological pathologists independently graded each WSI, with no additional information. To assess model performance, the calculated mean grades of each WSI assigned by the three pathologists were used. All the pathologists assigned a grade of 0 to the control group. Generally, the standard score of fibrosis for each slide is used in Good Laboratory Practice institutions, and the results are presented in six grades according to the severity of the lesion: 0 (normal), 1 (minimal), 2 (slight), 3 (moderate), 4 (marked), and 5 (severe). Representative views of the fibrosis score are shown in Supplementary Fig 1. In this study, no severe fibrotic lesions (fibrosis score of 5) were observed. We used a measurement tool to obtain perceptible data to measure the severity of fibrosis in each WSI. An annotated area analyzer was used to compute the area of interest. Aperio ImageScope was used as the annotation program. Subsequently, the annotated red-stained fibrotic area was computed for each WSI. The fibrosis ratio was computed from the fibrosis-annotated area over the entire liver. Because fibrosis is shaped such that it creates a portal along a vein, a separate annotation was performed to exclude empty spaces due to blood vessels to compute the actual true area value. Each annotation layer was set apart in four different colors (Fig. 1B). The green layer represents fibrosis, and the blue layer represents the blood vessel area included in the fibrosis annotation area, which may affect the calculation of the true fibrosis area. Similarly, the purple layer indicates the whole area of the liver, and the pink layer displays the blood vessel area inside the liver, which can influence the computation of the true whole liver area. The proportion of the annotated fibrosis area for each liver was computed as follows:

Table 1. Number of Cropped Images for the Model Training Sets

	Number of cropped images	
Test set	Normal 1,000 Fibrosis 1,000	
Training, validation set (8:2)	Training set 19,248	Normal 12,548 Fibrosis 6,700
	Validation set 4,813	Normal 3,138 Fibrosis 1,675

Annotated fibrosis area=

$$\frac{true\ fibrosis\ area\,(green\ layer\ area - blue\ layer\ area)}{true\ liver\ area\ \ (purple\ layer\ area - pink\ layer\ area)}$$

Evaluation of model performance

Model performance was assessed using two different approaches: one by analyzing image tiles obtained from one complete WSI and the other by comparing the correlation between the fibrosis ratio using the trained model, the mean pathologist score, and the annotation data. Since all the cropped images employed for training the model were not used slide by slide but used without distinction, it may be hard to describe the accuracy of the deep learning method at the slide level. To overcome this issue, we performed an inference test targeting a slide to confirm the accuracy of the model when assessing it slide-by-slide. We chose the liver slide of one positive control animal, HFDC+VC_02, and categorized cropped images into normal and fibrosis categories. We defined the cropped images sorted by the researcher as true-classified answers. Subsequently, each label of the image classified by the researcher was compared with the images classified by the trained model, and normalization was performed to examine the relationship between the three results in the common range. Scaling the range of data makes them lucid and comparable. As the min-max normalization[27] is sensitive to outliers, we examined outliers in the data using the ROUT method[28], a method for discerning outliers when data are nonlinear, with the help of Prism 8 (GraphPad Software, San Diego, CA, USA). We designated the Q value (the value of the basis for eliminating outliers) of the ROUT method as 1, as recommended by Motulsky *et al.*[28], and no outliers were identified. We normalized all analyzed data of each slide using the max normalization, which is the simplest normalization technique. We rescaled the ratios and range of the average grade from 0 (minimum value) to 1 (maximum value). The similarity verification between the average fibrosis grade assigned by the three pathologists, fibrosis ratio of deep learning, and fibrosis area ratio of annotation was performed to evaluate the performance of the trained model. Spearman's correlation test was used following the Shapiro–Wilk test to prove the similarity between deep learning and annotation, as well as pathologist grades.

Statistics

Prior to applying the appropriate statistical methods, the Shapiro–Wilk normality test was performed as the n value of each group was less than 30. Since none of the groups passed, the nonparametric Mann–Whitney test (two-tailed, confidence intervals 95%) was performed for each pair of groups[29]. For all cases, a p value less than 0.05 was considered statistically significant. For the correlation test, nonparametric correlation (Spearman's correlation) was used. All statistical analyses were performed using Prism 8 software (GraphPad Software).

Results

Training, validation, and test

To screen for hepatic fibrosis lesions, 24,061 image tiles were used for training and validation. The image and batch sizes for both training and validation were 128 × 128 and 50, respectively, and were trained for 15 epochs using the Xception network. The training accuracy exceeded 0.9950, and the validation accuracy exceeded 0.9900 after the first epoch (Fig. 2A). Both accuracies reached almost one in the first epoch while maintaining high accuracy until the end of the training, as well as high validation accuracy. This accuracy was higher than that reported in a previous study by *Heinemann et al.*, which aimed to classify fibrosis, ballooning, inflammation, and steatosis in NASH models with CNN, which was 88.5% and 86.3% for classifying fibrosis training and validation with 4,251 and 465 image tiles, respectively[7]. The losses for training and validation decreased steeply after the second epoch and remained stable until the final epoch (Fig. 2B).

The model prediction test was performed using the test dataset, which was divided before training. The results showed that one error was observed for 2,000 image tiles using a confusion matrix (Fig. 2C). The one error was from data that deep learning predicted to be normal but was identified as fibrosis. The confusion matrix shows the precision value, which indicates the percentage of true fibrosis out of all the predicted fibrosis. It computed as 999/(999+0)=1. The recall value, which is designated out of the total fibrosis, shows the percentage of the predicted fibrosis. It computed as 999/(999+1)=0.999. From the above calculation, we can obtain the F1 score, which shows the harmonic mean of the precision and recall. The F1 score is 0.999. Because the precision, recall, and F1 score were close to 1 (highest), we can assume that the accuracy of the deep learning model was high. Receiver operating characteristic (ROC) analysis illustrated the classification ability between normal and fibrosis. When the false positive rate is low, which is predicted as fibrosis, but the true value is normal and the true positive rate (synonym for recall) is high, the model's proficiency is classified as high (Fig. 2D).

Model performance confirmation

As we are not contemplating the correct answer for each cropped image of a WSI, we compared the human classified dataset and deep learning classified dataset of the positive control slide, HFDC+VC_02. From the confusion matrix, the precision value was computed to be 1,239/(1,239+225)=0.846. The recall value was computed as 1,239/(1,239+51)=0.96. From these outcomes, we obtained an F1 score, which was computed as 2 × ((0.846 × 0.960)/(0.846+0.960))=0.899. The confusion matrix derived from the prediction suggested that the trained model showed a high performance in predicting fibrosis (Fig. 3A). The ROC curve obtained from the true and false positive rates illustrates the classification ability between normal tissue and fibrosis. The X-axis, which depicts the false positive rate,

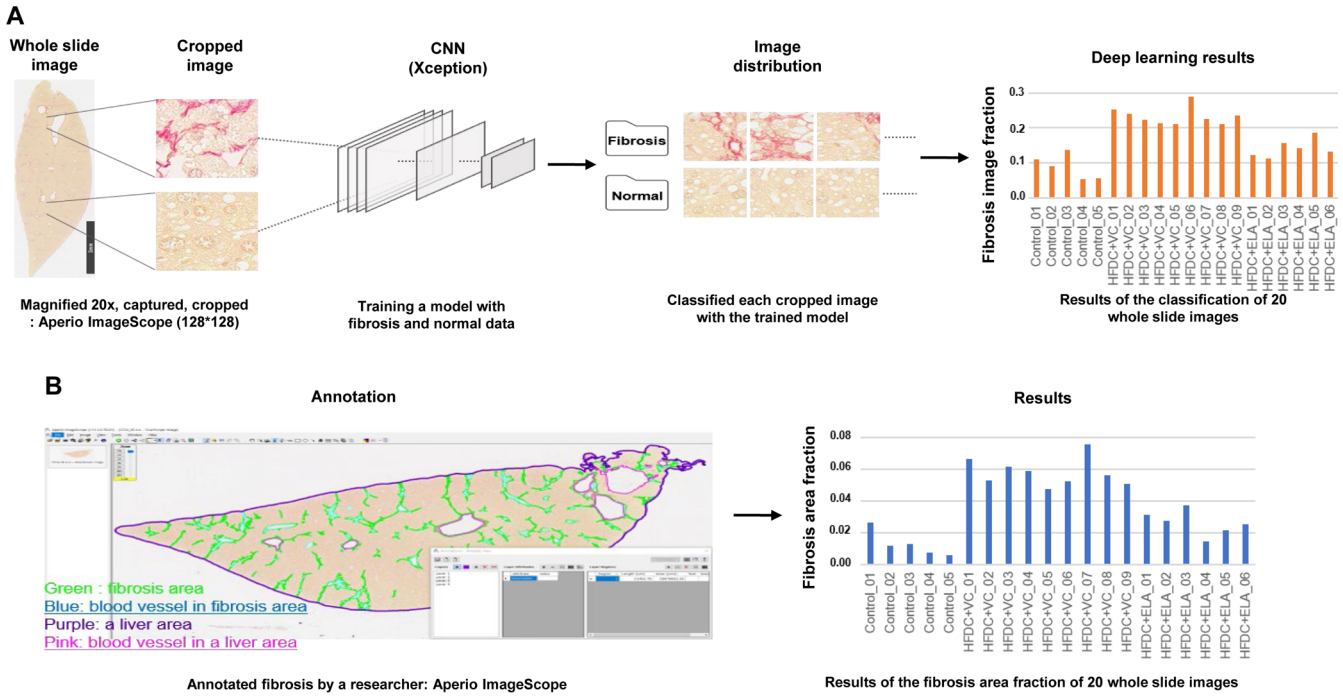

Fig. 1. Flow diagram illustrating the experimental design of the present study. (A) The procedure of data preparation, training, and results for this study. (B) Representative annotated image and results of the analysis. CNN: convolutional neural network.

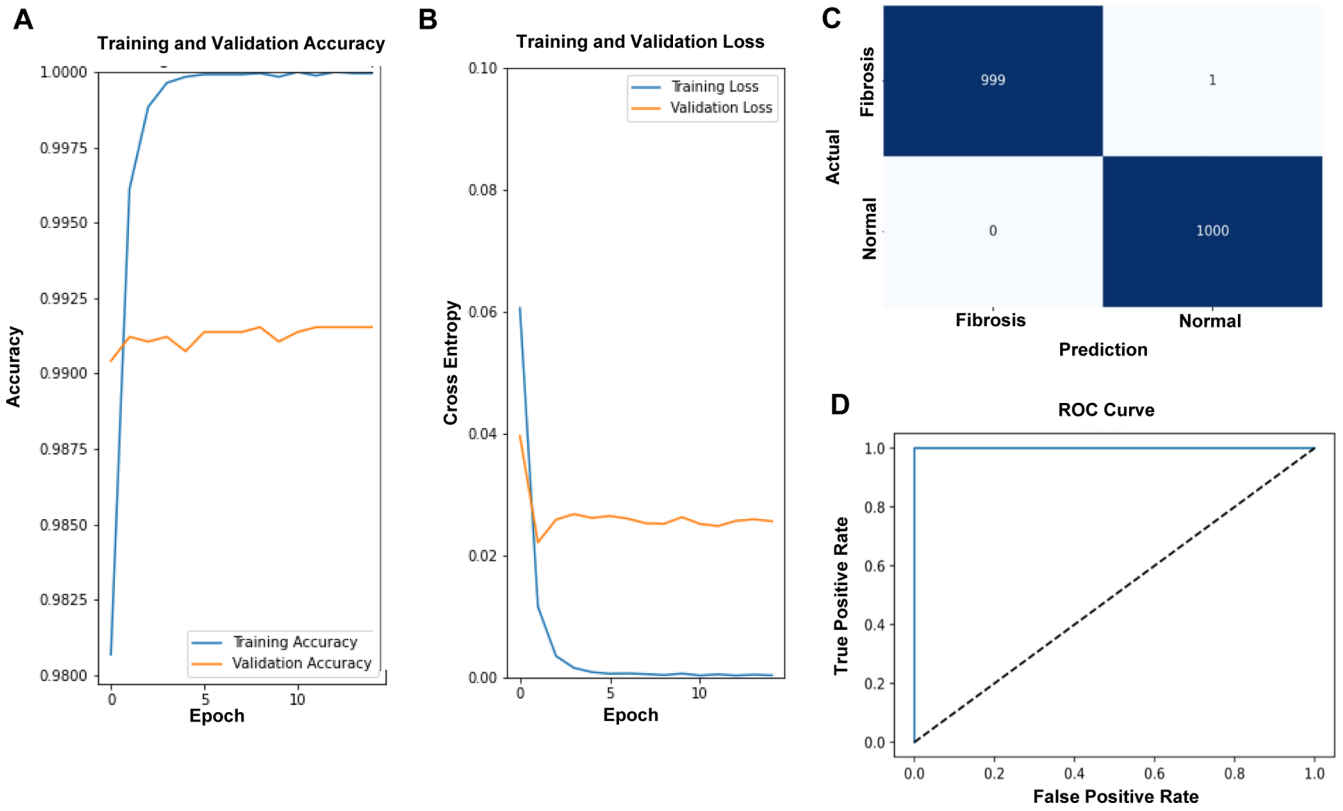

Fig. 2. The result of training, validation, and testing of the model. (A) Accuracy of training (blue line) and validation (orange line) against the number of epochs. (B) Loss for training (blue line) and validation (orange line). (C) Confusion matrix of training and validation data set. (D) Receiver operating characteristic (ROC) curve of the trained model.

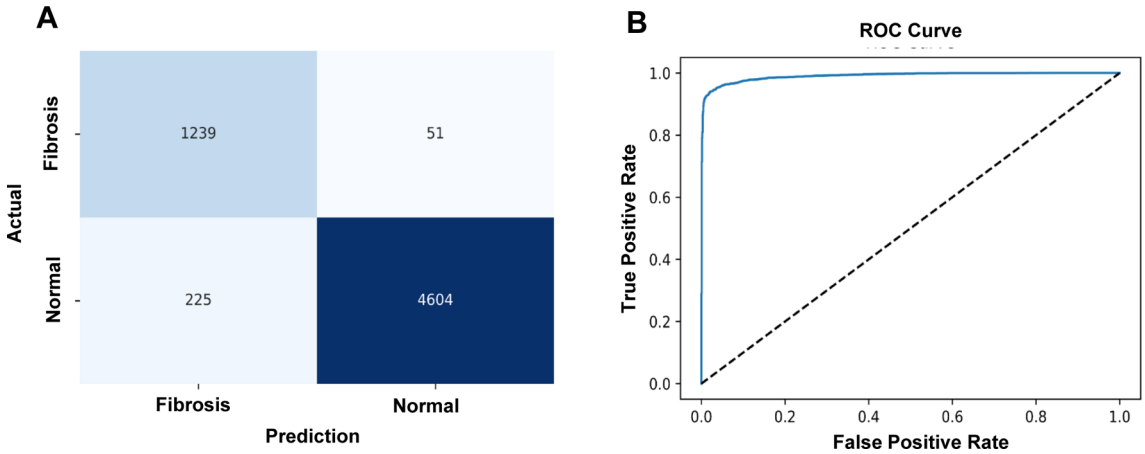

Fig. 3. Model performance in one complete WSI. (A) Confusion matrix of the HFDC+VC_02. (B) Receiver operating characteristic (ROC) curve of the HFDC+VC_02 (positive control) inferred from the slide by the trained deep learning algorithm.

Table 2. Results of Each Analysis

Group	Control					HFDC+VC									HFDC+ELA					
Treatment	None					HFD + CCL_4 +Vehicle									HFD + CCL_4 + ELA					
Animal number	01	02	03	04	05	01	02	03	04	05	06	07	08	09	01	02	03	04	05	06
Deep learning	574	446	813	371	273	1573	1464	1254	1156	1094	1638	1388	1453	1772	1634	1159	1302	1506	1964	1123
Pathologists	0	0	0	0	0	2.667	2	2	2	2	2.667	3	2.333	3	1	1.333	1.333	1.667	1.333	1
Annotation	0.564	0.238	0.315	0.216	0.120	1.662	1.302	1.393	1.281	1.007	1.241	1.963	1.632	1.602	1.699	1.142	1.245	0.632	0.924	0.885

HFDC: high-fat diet with CCL4 treatment; VD: vehicle control; ELA: elafibranor; CCL: carbon tetrachloride.

indicates the prediction of an event when there is no event. Moreover, the Y axis, which depicts the true positive rate, indicates the prediction of an event when there is an event. The low rate of the X-axis and the high rate of the Y-axis indicate high accuracy of the model. The ROC curve obtained from the true- and false-positive rates also proved the high accuracy of the model (Fig. 3B).

Evaluation of deep learning

To evaluate the deep learning model, we compared the grades assigned by the pathologists with the annotation results. The raw data result of deep learning presents the number of WSIs classified as fibrosis, pathologists indicating the average grade of three pathologists, and annotations showing the true fibrosis area of each WSI (Table 2). The unit of true fibrosis area of annotation is square millimeters (mm^2), and the numbers were rounded off to three decimal places. Fibrosis grade by deep learning was determined using the proportion of the number of images classified as fibrosis from the cropped WSI (Fig. 4A). Three certified toxicological pathologists graded each slide prior to testing. Because there may be variations depending on the experience and standards of each pathologist, the average value was used for comparison (Fig. 4B). The true whole fibrosis-annotated area of each WSI was used for annotation (Fig. 4C). Twenty slides were used for the analyzed dataset.

In particular, the differences between the control, HFDC+VC, and HFDC+ELA groups were statistically significant, indicating that the degree of fibrosis was the lowest in the control group, followed by the HFDC+ELA group (Fig. 4D–4F). These results suggest that the deep learning algorithm can differentiate fibrotic areas. Furthermore, all the analysis methods successfully distinguished the severity of fibrosis in each group.

Comparison of the three analyses

The outliers in each analyzed dataset were identified prior to normalization. Since no outliers were found, the min-max normalization was computed to rescale the ranges of each analyzed method's outcome from 0 (minimum value) to 1 (maximum value) (Fig. 5A). To compute and confirm the performance of the deep learning algorithms, the relationships between normalized data from each analyzed method were analyzed using Spearman's correlation coefficient with the Shapiro–Wilk normality test.

The hepatic fibrosis grade predicted by deep learning showed a very high correlation with the pathologist-assigned grade (r=0.9067), and it was the highest among all correlations (Fig. 5B)[30]. The annotated fibrosis area ratio and pathologists' average grades showed a high correlation (r=0.8579) (Fig. 5C). Finally, the annotated fibrosis area ratio and predicted fibrosis ratio were computed as highly correlated (r=0.8346) (Fig. 5D).

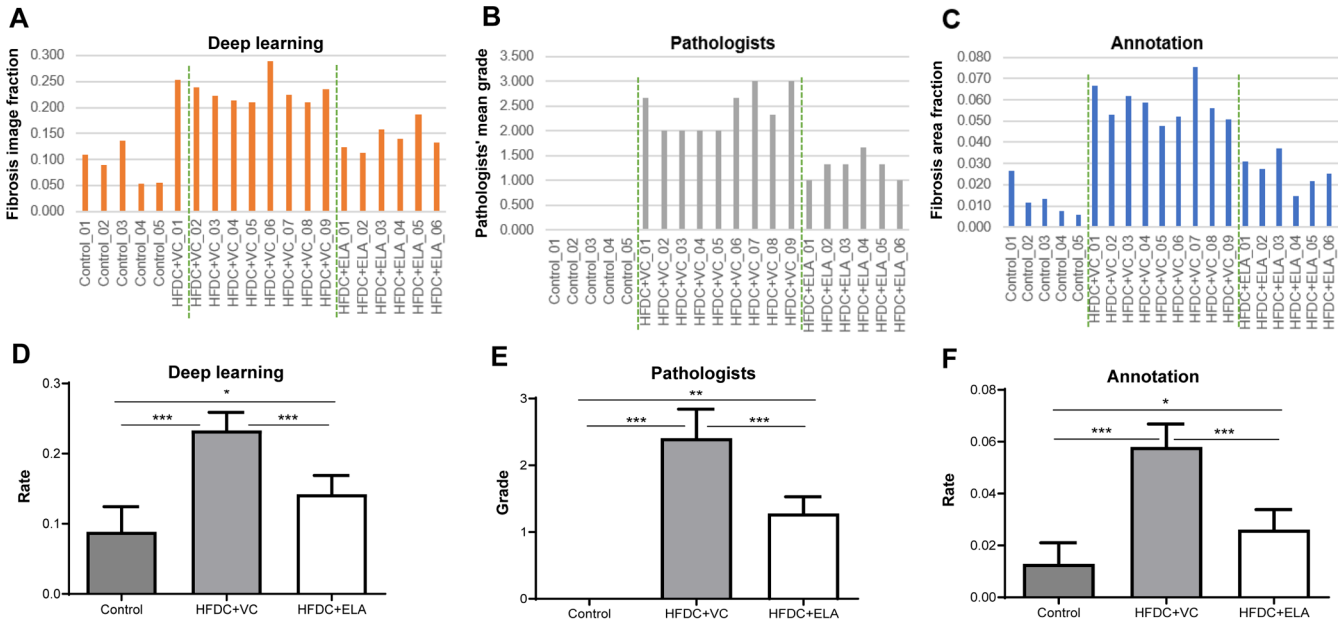

Fig. 4. Fibrosis results and differences between groups of each evaluation method. (A) The rate of the number of images classified as fibrosis to the total number of cropped images for each WSI. (B) The average grade of three pathologists for each whole slide image (WSI). (C) The true whole fibrosis area [(green layer – blue layer)/(purple layer – pink layer)] for each WSI. The green dotted lines separate the groups. (D–F) The difference in the level of fibrosis between the groups of (D) deep learning, (E) pathologists, and (F) annotation, respectively. Data are the mean ± SD, and p values were computed using the Mann–Whitney test; n=5, 9, and 6 per control, HFDC+VC, and HFDC+ELA groups, respectively. Every group showed a statistically significant difference ($p<0.05$) in the level of fibrosis in the different groups in each of the evaluation methods. *$p<0.05$, **p 0.01, ***$p<0.001$.

Discussion

Considering the increasing prevalence of NAFLD and NASH, efforts to discover therapeutics for the treatment of distinct hepatic fibrosis have increased[31]. Before the administration of a test drug to treat lesions, pathological findings from tissue slides interpreted by certified toxicological pathologists provide critical evidence for the drug effect. Recently, deep learning has been implemented in non-clinical studies. The purpose of applying convolution networks in non-clinical studies is to achieve a quantitative and rapid assessment of pathological findings during drug discovery. Therefore, a classification network may be favored over object detection or segmentation, which requires complicated calculations and a longer execution time, as well as a complex annotation procedure. Some studies have applied different classification methods to classify lesions in mouse models. Asay *et al.* used a convoluted network of their own designed layers to discriminate tuberculosis pulmonary lesions using millions of image tiles[15]. Another study used transfer learning with an inception-v3 network to infer NASH scoring with much less data[7]. They showed the possibility of using deep learning for the automatic evaluation of the NASH score from the images in the model mice using transfer learning.

In the present study, we carried out research verifying the implementation of deep learning in the assessment of toxicological pathology in a nonclinical study. This study employed automatic classification of hepatic fibrosis using a deep learning network to evaluate the performance in comparison to pathologists and annotation. The classification algorithm trained using the Xception network showed a high F1 score at the slide level (89.9%). Three certified toxicological pathologists scored the fibrosis. Finally, the fibrosis area was annotated and computed to obtain a quantified fibrosis value. Because the ranges of all analyses are different, we transformed the range of the grades using the min-max normalization method. Using this hepatic fibrosis grade of deep learning, we compared each grade of a slide with the average fibrosis grade assigned by the pathologists and the annotated fibrosis area to confirm the performance of the algorithms. The trained deep learning strongly follows not only the results of the annotation data, which was set as quantitative answer data, but also the mean average pathologist-assigned grade, which is regarded as the gold standard. The results show a strong correlation between each pair of the three. The fibrosis grade computed by annotation showed strong correlations with the pathologist-assigned grades. Pathologists consider not only the context of the section, such as the pattern of the fibrosis bridge between central veins, but also the artifacts in slides when diagnosing hepatic fibrosis. The deep learning model trained in this study showed grades similar to the pathologist-assigned grades compared to those with the annotation, despite the presence of artifacts. Therefore, we consider the results of this study as useful evidence for the ability to use deep learning to distinguish

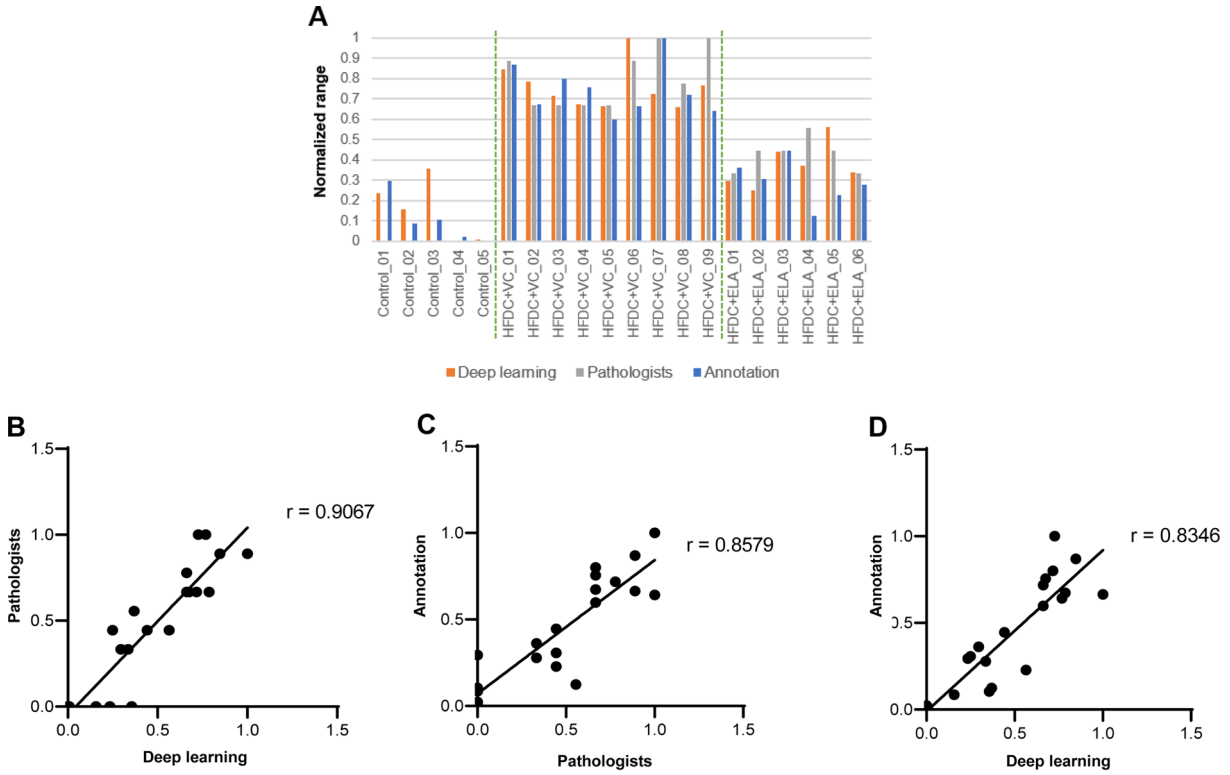

Fig. 5. Comparison and correlations between three analyses. (A) Normalized analyzed data of fibrosis area rate of predicted fibrosis rate by deep learning (orange), average grade by three pathologists (gray), and annotation (blue) for each whole slide image (WSI). No outliers were identified using the ROUT method (with Q* set to 1%). Therefore, we used min-max normalization to rescale the ranges from 0 (minimum value) to 1 (maximum value). The green dotted lines separate the groups. (B–D) Correlation between each analysis method with normalized data. For all graphs, Spearman's correlation coefficient (r) is used. (B) Correlation between normalized predicted fibrosis rate assigned by deep learning and normalized average grade assigned by pathologists. (C) Correlation between normalized average grade by pathologists and fibrosis area rate from the annotation in WSIs. (D) Correlation between normalized predicted images as fibrosis rate by deep learning and normalized fibrosis area rate from the annotation in WSI.
*Q: value of the basis for eliminating outliers.

between normal and fibrotic lesions, based not only on the red color but also on other features that humans would not recognize.

Sirius Red-stained samples can be easily distinguished by their color as a result of collagen staining. Hence, the deep learning algorithm tends to classify reddish-cropped images into a fibrosis folder. Because pathologists can determine the control value and assign a score of 0, deep learning has no standard for control but simply classifies reddish images. The annotation process is also similar to deep learning; the reddish-stained part of the control group was annotated as fibrosis, and hence fibrosis appeared in the control group as well. Nevertheless, Sirius Red staining was used because the color-coded feature of fibrosis is considered to be a big advantage in image classification in deep learning. Therefore, to alleviate the classification of normal misrecognition as fibrosis, adding control WSIs in order to use normal images of the control group as an ideal normal value is needed. Through the examination of another deep learning model before adding two control WSIs from the training stage, the Spearman's correlation coefficient (r) between deep learning and annotation was reduced to 0.0203 and slightly increased (less than 0.01) between deep learning and annotation (Supplementary Fig. 2). This indicates that a wide spectrum of normal values might have affected the performance of the deep learning results.

Computer vision techniques have achieved breakthroughs since the advent of AlexNet, which competed to obtain the highest accuracy in a multi-class classification problem challenge with convolutional networks, as evaluated by the 2012 ImageNet Large Scale Visual Recognition Challenge (ILSVRC)[32]. AlexNet is a deep CNN trained model that is similar to LeNet but different from all the convoluted layers stacked together[33]. The inception that has deepened the depth and widened the width of the network, and advanced the utilization of computing data[34], has fewer parameters compared to AlexNet and won the ILSVRC 2014 competition[33]. Subsequently, Xception, which we used in this study, is inspired by Inception-v3 and computes the cross-channel correlation and spatial correlation independently. Xception performed better than Inception-v3 when using model parameters[13].

Diagnosis by certified toxicological pathologists is believed to be the gold standard for determining the severity

of a lesion; nevertheless, pathologists rely on their personal knowledge and experience for diagnosis. Sometimes, accuracy cannot be guaranteed, and misdiagnosis is inevitable. Notably, deep learning and pathologists showed the highest correlation. From this perspective, we propose that the fibrosis evaluation method in Sirius Red-stained WSIs using the trained model might produce similar results to the pathologists. The trained model can be applied to analyze fibrosis in Sirius Red-stained WSIs as a second-opinion tool for practical use.

Disclosure of Potential Conflicts of Interest: The authors declare no real, perceived, or potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Acknowledgments: This study was supported by the Korea Institute of Toxicology, Republic of Korea (1711159827).

References

1. Masuoka HC, and Chalasani N. Nonalcoholic fatty liver disease: an emerging threat to obese and diabetic individuals. Ann N Y Acad Sci. **1281**: 106–122. 2013. [Medline] [CrossRef]
2. Leoni S, Tovoli F, Napoli L, Serio I, Ferri S, and Bolondi L. Current guidelines for the management of non-alcoholic fatty liver disease: a systematic review with comparative analysis. World J Gastroenterol. **24**: 3361–3373. 2018. [Medline] [CrossRef]
3. Widjaja AA, Singh BK, Adami E, Viswanathan S, Dong J, D'Agostino GA, Ng B, Lim WW, Tan J, Paleja BS, Tripathi M, Lim SY, Shekeran SG, Chothani SP, Rabes A, Sombetzki M, Bruinstroop E, Min LP, Sinha RA, Albani S, Yen PM, Schafer S, and Cook SA. Inhibiting Interleukin 11 signaling reduces hepatocyte death and liver fibrosis, inflammation, and steatosis in mouse models of nonalcoholic steatohepatitis. Gastroenterology. **157**: 777–792.e14. 2019. [Medline] [CrossRef]
4. Lim YS, and Kim WR. The global impact of hepatic fibrosis and end-stage liver disease. Clin Liver Dis. **12**: 733–746, vii. 2008. [Medline] [CrossRef]
5. Lemoinne S, and Friedman SL. New and emerging antifibrotic therapeutics entering or already in clinical trials in chronic liver diseases. Curr Opin Pharmacol. **49**: 60–70. 2019. [Medline] [CrossRef]
6. Tsuchida T, Lee YA, Fujiwara N, Ybanez M, Allen B, Martins S, Fiel MI, Goossens N, Chou HI, Hoshida Y, and Friedman SL. A simple diet- and chemical-induced murine NASH model with rapid progression of steatohepatitis, fibrosis and liver cancer. J Hepatol. **69**: 385–395. 2018. [Medline] [CrossRef]
7. Heinemann F, Birk G, and Stierstorfer B. Deep learning enables pathologist-like scoring of NASH models. Sci Rep. **9**: 18454. 2019. [Medline] [CrossRef]
8. Robboy SJ, Weintraub S, Horvath AE, Jensen BW, Alexander CB, Fody EP, Crawford JM, Clark JR, Cantor-Weinberg J, Joshi MG, Cohen MB, Prystowsky MB, Bean SM, Gupta S, Powell SZ, Speights VO Jr, Gross DJ, and Black-Schaffer WS. Pathologist workforce in the United States: I. Development of a predictive model to examine factors influencing supply. Arch Pathol Lab Med. **137**: 1723–1732. 2013. [Medline] [CrossRef]
9. Rousselet MC, Michalak S, Dupré F, Croué A, Bedossa P, Saint-André JP, Calès P, Hepatitis N. Hepatitis Network 49. Sources of variability in histological scoring of chronic viral hepatitis. Hepatology. **41**: 257–264. 2005. [Medline] [CrossRef]
10. Bedossa P. Pathology of non-alcoholic fatty liver disease. Liver Int. **37**(Suppl 1): 85–89. 2017. [Medline] [CrossRef]
11. Al-Saffar AAM, Tao H, and Talab MA. 2017 International Conference on Radar, Antenna, Microwave, Electronics, and Telecommunications (ICRAMET). Review of deep convolution neural network in image classification. Vol. 26–31. 2017.
12. Krizhevsky A, Sutskever I, and Hinton GE. Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst. **25**. 2012.
13. Chollet F. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Xception: Deep learning with depthwise separable convolutions. pp 1800–1807. 2017.
14. Chang HY, Jung CK, Woo JI, Lee S, Cho J, Kim SW, and Kwak TY. Artificial intelligence in pathology. J Pathol Transl Med. **53**: 1–12. 2019. [Medline] [CrossRef]
15. Asay BC, Edwards BB, Andrews J, Ramey ME, Richard JD, Podell BK, Gutiérrez JFM, Frank CB, Magunda F, Robertson GT, Lyons M, Ben-Hur A, and Lenaerts AJ. Digital image analysis of heterogeneous tuberculosis pulmonary pathology in non-clinical animal models using deep convolutional neural networks. Sci Rep. **10**: 6047. 2020. [Medline] [CrossRef]
16. Pischon H, Mason D, Lawrenz B, Blanck O, Frisk AL, Schorsch F, and Bertani V. Artificial intelligence in toxicologic pathology: quantitative evaluation of compound-induced hepatocellular hypertrophy in rats. Toxicol Pathol. **49**: 928–937. 2021. [Medline] [CrossRef]
17. Ramot Y, Zandani G, Madar Z, Deshmukh S, and Nyska A. Utilization of a deep learning algorithm for microscope-based fatty vacuole quantification in a fatty liver model in mice. Toxicol Pathol. **48**: 702–707. 2020. [Medline] [CrossRef]
18. Srinidhi CL, Ciga O, and Martel AL. Deep neural network models for computational histopathology: a survey. Med Image Anal. **67**: 101813. 2021. [Medline] [CrossRef]
19. Huang Y, de Boer WB, Adams LA, MacQuillan G, Rossi E, Rigby P, Raftopoulos SC, Bulsara M, and Jeffrey GP. Image analysis of liver collagen using sirius red is more accurate and correlates better with serum fibrosis markers than trichrome. Liver Int. **33**: 1249–1256. 2013. [Medline] [CrossRef]
20. Seifert WF, Bosma A, Brouwer A, Hendriks HF, Roholl PJ, van Leeuwen RE, van Thiel-de Ruiter GC, Seifert-Bock I, and Knook DL. Vitamin A deficiency potentiates carbon tetrachloride-induced liver fibrosis in rats. Hepatology. **19**: 193–201. 1994. [Medline] [CrossRef]
21. Tølbøl KS, Kristiansen MN, Hansen HH, Veidal SS, Rigbolt KT, Gillum MP, Jelsing J, Vrang N, and Feigh M. Metabolic and hepatic effects of liraglutide, obeticholic acid and elafibranor in diet-induced obese mouse models of biopsy-confirmed nonalcoholic steatohepatitis. World J Gastroen-

terol. **24**: 179–194. 2018. [Medline] [CrossRef]
22. Canziani A, Paszke A, and Culurciello E. An analysis of deep neural network models for practical applications. arXiv preprint arXiv:1605.07678. 2016.
23. Wu X, Liu R, Yang H, and Chen Z. 2020 2nd International Conference on Information Technology and Computer Application (ITCA). An xception based convolutional neural network for scene image classification with transfer learning. 262–267. 2020.
24. Khan AI, Shah JL, and Bhat MM. CoroNet: A deep neural network for detection and diagnosis of COVID-19 from chest x-ray images. Comput Methods Programs Biomed. **196**: 105581. 2020. [Medline] [CrossRef]
25. Malik S, Singh S, Singh NM, and Panwar N. Diagnosis of COVID-19 using Chest X-ray. International Journal of Informatics. Information System and Computer Engineering. **2**: 55–64. 2021; (INJIISCOM).
26. Jain R, Gupta M, Taneja S, and Hemanth DJ. Deep learning based detection and analysis of COVID-19 on chest X-ray images. Appl Intell. **51**: 1690–1700. 2021. [Medline] [CrossRef]
27. Jain A, Nandakumar K, and Ross A. Score normalization in multimodal biometric systems. Pattern Recognit. **38**: 2270–2285. 2005. [CrossRef]
28. Motulsky HJ, and Brown RE. Detecting outliers when fitting data with nonlinear regression - a new method based on robust nonlinear regression and the false discovery rate. BMC Bioinformatics. **7**: 123. 2006. [Medline] [CrossRef]
29. Ramot Y, Deshpande A, Morello V, Michieli P, Shlomov T, and Nyska A. Microscope-based automated quantification of liver fibrosis in mice using a deep learning algorithm. Toxicol Pathol. **49**: 1126–1133. 2021. [Medline] [CrossRef]
30. Rovai AP, Baker JD, and Ponton MK. Social Science Research Design and Statistics: A Practitioner's Guide To Research Methods and IBM SPSS. Watertree Press, 2013.
31. Angulo P, Kleiner DE, Dam-Larsen S, Adams LA, Bjornsson ES, Charatcharoenwitthaya P, Mills PR, Keach JC, Lafferty HD, Stahler A, Haflidadottir S, and Bendtsen F. Liver fibrosis, but no other histologic features, is associated with long-term outcomes of patients with nonalcoholic fatty liver disease. Gastroenterology. **149**: 389–97.e10. 2015. [Medline] [CrossRef]
32. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, and Bernstein M. Imagenet large scale visual recognition challenge. Int J Comput Vis. **115**: 211–252. 2015. [CrossRef]
33. Aloysius N, and Geetha M. 2017 International Conference on Communication and Signal Processing (ICCSP). A review on deep convolutional neural networks. 0588–0592. 2017.
34. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, and Rabinovich A. Proceedings of the IEEE Conference on Computer Vision and Pattern recognition. Going deeper with convolutions. 1–9. 2015.