ORIGINAL ARTICLE

MAYO CLINIC PROCEEDINGS: DIGITAL HEALTH

MAYO CLINIC

Deep Learning–Based Prediction of Hepatic Decompensation in Patients With Primary Sclerosing Cholangitis With Computed Tomography

Yashbir Singh, PhD; Shahriar Faghani, MD; John E. Eaton, MD; Sudhakar K. Venkatesh, MD; and Bradley J. Erickson, MD, PhD

Abstract

Objective: To investigate a deep learning model for predicting hepatic decompensation using computed tomography (CT) imaging in patients with primary sclerosing cholangitis (PSC).

Patients and Methods: Retrospective cohort study involving 277 adult patients with large-duct PSC who underwent an abdominal CT scan. The portal venous phase CT images were used as input to a 3D-DenseNet121 model, which was trained using 5-fold crossvalidation to classify hepatic decompensation. To further investigate the role of each anatomic region in the model's decision-making process, we trained the model on different sections of 3-dimensional CT images. This included training on the right, left, anterior, posterior, inferior, and superior halves of the image data set. For each half, as well as for the entire scan, we performed area under the receiving operating curve (AUROC) analysis.

Results: Hepatic decompensation occurred in 128 individuals after a median (interquartile range) of 1.5 years (142-1318 days) after the CT scan. The deep learning model exhibited promising results, with a mean ± SD AUROC of 0.89±0.04 for the baseline model. The mean ± SD AUROC for left, right, anterior, posterior, superior, and inferior halves were 0.83±0.03, 0.83±0.03, 0.82±0.09, 0.79±0.02, 0.78±0.02, and 0.76±0.04, respectively.

Conclusion: The study illustrates the potential of examining CT imaging using 3D-DenseNet121 deep learning model to predict hepatic decompensation in patients with PSC.

 ■ Mayo Clin Proc Digital Health 2024;2(3):470-476

From the Department of Radiology (Y.S., S.F., S.K.V., B.J.E.) and Division of Gastroenterology & Hepatology (J.E.E.), Mayo Clinic, Rochester, MN.

Primary sclerosing cholangitis (PSC) is a chronic, cholestatic liver disorder characterized by inflammation and fibrosis of the extra and intrahepatic bile ducts.[1,2] Primary sclerosing cholangitis can progress to cirrhosis and lead to complications such as portal hypertension and hepatic decompensation.[3,4] Surrogate markers that can stratify the risk of hepatic decompensation are important for clinical practice and have the potential to aid in the conduct of clinical trials.

In the dynamic health care landscape, convolutional neural networks (CNNs), a premier deep learning technique, have emerged as an effective tool for medical image analysis. These networks can predict a clinical diagnosis or state by identifying and combining low-level and high-level image features, facilitating the extraction of intricate details from medical images.[5,6]

In the 2023 study by Singh et al,[7] a 3-dimensional (3D) CNN, capable of processing 3D input data, was used. Computed tomography (CT) images of adrenal masses were input into a 3D-DenseNet121 model to classify them as either adrenocortical carcinoma or lipid-poor adrenal adenomas.[7]

The integration of 3D CNNs in medical diagnostics is still nascent, with no reported studies applying an imaging deep learning model to predict hepatic decompensation. Therefore, our research aimed to pioneer the development of a CT-based 3D CNN model for this purpose. We hypothesized that a

CNN can be meticulously trained to differentiate decompensation using CT images, providing a novel approach to managing PSC.

PATIENTS AND METHODS

Description of Patients and Imaging Methods

This study was conducted in strict accordance with the principles of the Declaration of Helsinki and adhered to the highest ethical standards for research involving human patients. Both the study protocol and the informed consent procedure received approval from the institutional review boards of Mayo Clinic, Rochester, and patient consent was obtained. The inclusion criteria were as follows: (1) diagnosis of large-duct PSC based on standard criteria[8] and (2) availability of a CT study of the abdomen acquired during the portal venous phase. Weight-based volume of intravenous contrast was injected at 2-3 mL/s, and a fixed delay of 70 seconds for portal venous phase was obtained. A total of 277 patients with PSC were included. This retrospective cohort study, spanning from June 19, 1993, to March 26, 2021, used CT studies performed for clinical indications, aiming to demonstrate the potential of deep learning in predicting hepatic decompensation in patients with PSC, with magnetic resonance imaging typically used during follow-ups at our institute. Individuals were excluded if any of the following occurred before or at the time of the CT scan: hepatic decompensation, cholangiocarcinoma or liver transplantation. Patients were followed up from the baseline CT scan to time of hepatic decompensation, date of past follow-up, or liver transplantation (whichever occurred earlier). Hepatic decompensation was defined as the development of ascites, hepatic encephalopathy, or variceal hemorrhage. Continuous variables were reported as median (interquartile ranges) unless otherwise specified.

Model, Initialization, and Training

Leveraging the Scikit-learn toolkit, the data set was divided into 5 folds, stratified by class—decompensation and no decompensation, ensuring that the division was executed at the patient level to prevent information leakage.[9] The robustness of the model was subsequently gauged through 5-fold crossvalidation to scrutinize its resilience to variations in the data.[9,10] Computed tomography images were resampled to achieve $1\times1\times1$-mm voxel dimensions, using trilinear interpolation. This is crucial for maintaining consistency and uniformity in voxel sizes, allowing for a reliable analysis. The voxels were normalized to zero mean and unit SD, and images were zero-padded to $480\times480\times500$ voxels, aligning with the size of the more considerable liver observed (Figure 1). Data augmentation techniques were deployed to mitigate the risk of overfitting and enhance the model's generalizability, as delineated in Table 1. The distinction between decompensation and no decompensation images was achieved by deploying a 3D-DenseNet121 classifier from the MONAI package.[11–13] DenseNet, a type of CNN, uses convolutions to extract pivotal information, interconnecting each layer to every other layer in a feed-forward fashion, thereby enhancing the extraction of intricate patterns within the data.

Each dense block within DenseNet contained a bottleneck layer before each convolution layer, serving to reduce the number of feature mappings and thus increasing parameter efficiency and mitigating model complexity.[11] This structure is instrumental in reducing the probability of overfitting. With a data set comprising 277 patients, the AdamW optimizer was used with a batch size of 4 and a learning rate of 1×10^{-3}, modulated by a cosine annealing learning rate scheduler for 300 epochs.

Our model was trained using weighted cross-entropy as the loss function, with weights inverse-to-class ratios assigned to decompensation and no decompensation classes, respectively. The model's area under the receiving operating curve (AUROC) was computed for each fold, and mean and SD of AUROC were reported for each fold.

In the context of the model described, the concept of "folds" relates to a technique known as crossvalidation, commonly used in machine learning (ML) to assess the performance of a model. Imagine a data set that is intended for training and testing a model. Instead of using the same data for both training and testing, which can lead to biased results, the data are divided into folds.

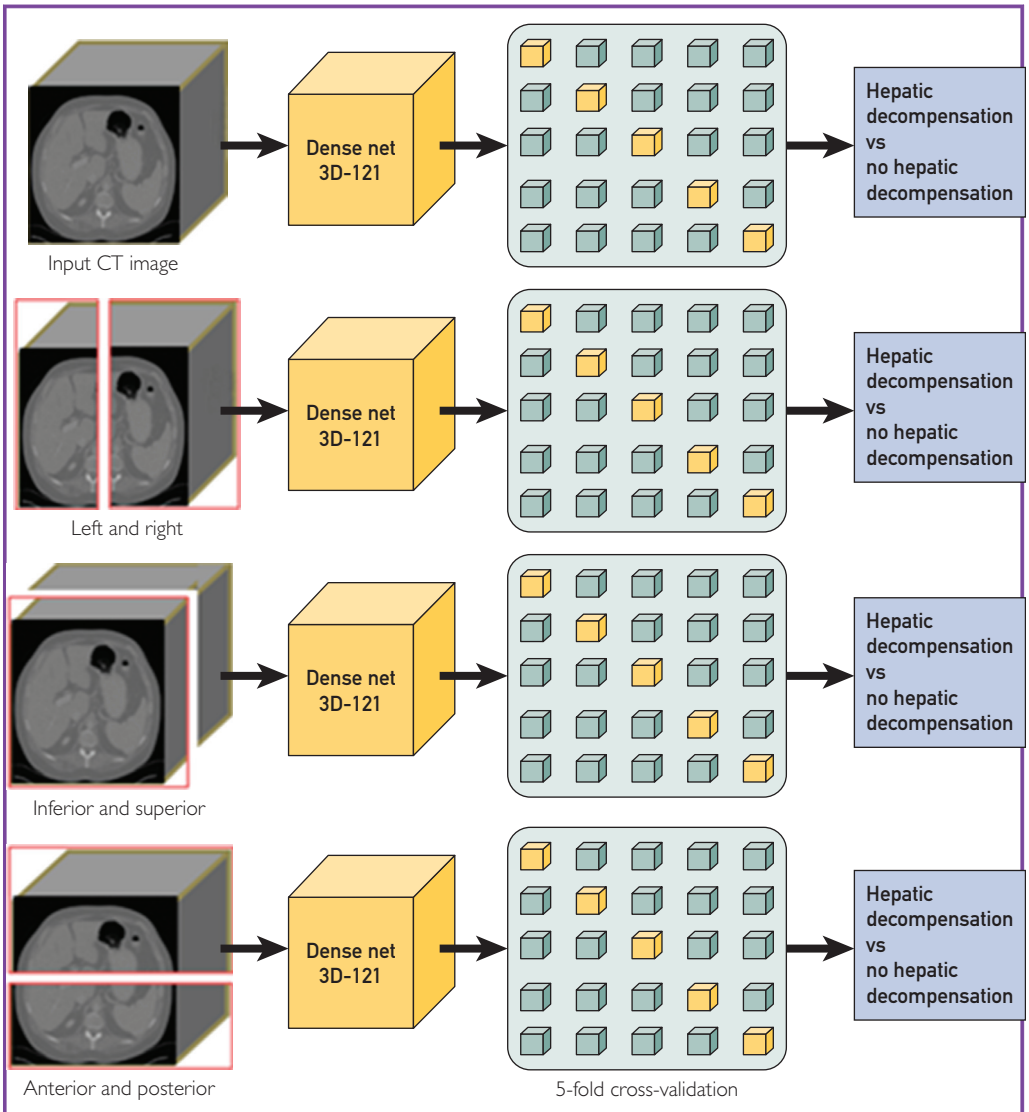

FIGURE 1. A diagram of the pipeline we developed for distinguishing between decompensation and no decompensation is presented in various planes. 3D, 3-dimensional.

Consider dividing the data set into 5 parts, named fold 1, fold 2, fold 3, fold 4, and fold 5. This leads to a process known as 5-fold crossvalidation, where the model undergoes training and testing 5 times. In each iteration, 4 of the 5 folds are used for training the model, although the remaining single fold is used for testing. This rotation ensures that each fold serves as the test set once. For instance, in the first iteration, folds 1-4 might be used for training, and fold 5 for testing. Subsequently, in the second iteration, folds 1-3 and 5 might be used for training, with fold 4 as the test set, and so on. Such a method is advantageous in effectively using all available data for both training and testing, thereby providing a comprehensive evaluation of the model's performance.

To examine how various anatomical regions affect a model's decision-making, we trained it using distinct sections of 3D CT images, focusing separately on the right, left, anterior, posterior, inferior, and superior halves. This approach aimed to identify if specific areas

Mayo Clin Proc Digital Health ■ September 2024;2(3):470-476 ■ https://doi.org/10.1016/j.mcpdig.2024.07.002
www.mcpdigitalhealth.org

TABLE 1. The Parameters Associated With Various Data Augmentation Methods

Random flip	*P*=.5
Random translation	Translate range = 15, 15, 10
Random scaling	Scale range = 0.05, 0.05, 0.05
Random rotation	Rotation range = π/8, π/8, π/8
Random gaussian noise addition	Mean = 0.0, SD = 0.2

TABLE 2. Values for Metrics Obtained During a 5-Fold Stratified Crossvalidation Evaluation of the 3D Densenet121 Classifier (Baseline)

No. of folds	AUC accuracy
Fold 1	0.81
Fold 2	0.87
Fold 3	0.94
Fold 4	0.91
Fold 5	0.92

AUC, area under the receiving operating curve; 3D, 3-dimensional.

were more influential in the model's analyses. We used a geometric approach based on the dimensions of the entire CT volume to divide the images. Specifically, the division was performed as follows—right/left: the division was made along the sagittal plane at the midpoint of the image's width; anterior/posterior: the division was made along the coronal plane at the midpoint of the image's depth; and superior/inferior: the division was made along the axial plane at the midpoint of the image's height.

We then evaluated the model;s performance for each half and the entire scan by calculating the mean AUROC and its SD. This method provided insights into the model's reliance on different anatomical areas, potentially guiding more targeted and efficient diagnostic processes. The training was executed on a cluster of 4 NVIDIA A100 GPUs.

RESULTS

The study included 277 patients with a median duration of follow-up of 1.5 years (142-1318 days). The study patients comprised 63% males, and 75% had concurrent inflammatory bowel disease. Hepatic decompensation occurred in 128 individuals (ascites, n=49; hepatic encephalopathy, n=33; variceal hemorrhage, n=46). Area under the receiving operating curve values for each of the 5 folds were as follows (Table 2)—the baseline model, which constituted the whole liver, reported the following AUROC values: the model yielded AUROC accuracies for folds 1-5 of 0.84, 0.86, 0.78, 0.84, and 0.82 for the left and 0.87, 0.80, 0.80, 0.85, and 0.82 for the right side of the liver, respectively (Table 3).

These comprehensive results underline the model's robust performance, showcasing its potential as a valuable tool in the realm of medical imaging diagnostics, particularly for the nuanced task of predicting hepatic decompensation. The median time of 1.5 years for hepatic decompensation suggests that this cohort likely has advanced liver disease. This relatively short time frame indicates that many patients in the study were already at a more progressed stage of PSC.

It is essential to note that although the AUROC provides valuable insights into the model's performance, additional metrics such as precision, recall, F1 score, and confusion matrices, clinical information, would provide a more comprehensive understanding of the model's applicability and reliability in a real-world setting. Further studies and validations using independent data sets are recommended to affirm the model's efficacy and to potentially fine-tune it for enhanced predictive accuracy in diverse clinical scenarios.

From the plot (Figure 2), we can identify which comparisons are statistically significant ($P<.05$) and which are not. Those tests with bars extending below the red line are considered significant, indicating a statistically meaningful difference between the groups compared in those tests. Comparisons like "baseline vs post," "baseline vs sup," "baseline vs inf," and others, where the *P* value is less than .05, are significant. In contrast, tests with *P* values above the red line, such as "baseline vs right," "baseline vs ant," and others, are not considered statistically significant.

TABLE 3. Accuracy Values for Metrics Obtained During a 5-Fold Stratified Crossvalidation Evaluation of the 3D Densenet121 Classifier for Predicting Hepatic Decompensation Using 6 Different Data Sets

Data set	Left	Right	Anterior	Posterior	Superior	Inferior
Fold 1	0.84	0.87	0.77	0.78	0.77	0.78
Fold 2	0.86	0.80	0.89	0.78	0.82	0.75
Fold 3	0.78	0.80	0.70	0.76	0.76	0.70
Fold 4	0.84	0.85	0.93	0.82	0.78	0.78
Fold 5	0.82	0.82	0.82	0.78	0.78	0.80

3D, 3-dimensional.

DISCUSSION

We developed an advanced deep learning model tailored to predict hepatic decompensation in patients with PSC using CT imaging data sets (3D-DenseNet121). This approach harnesses the computational power of deep learning to analyze extensive CT imaging data, providing an accurate classification of hepatic decompensation in patients with PSC. Our study represents a proof-of-concept application, demonstrating the potential of deep learning in predicting short-term outcomes

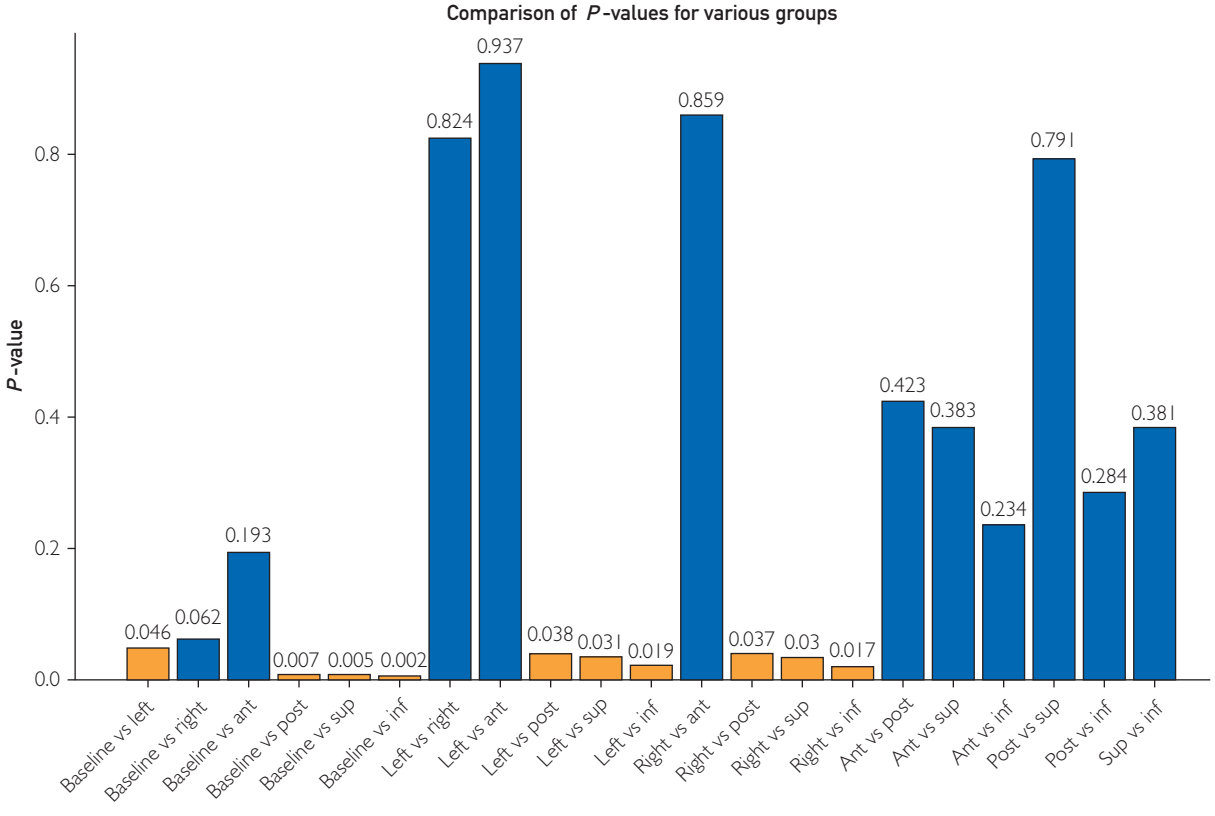

FIGURE 2. A bar plot shows the *P* values for the different group comparisons. Each bar represents a comparison between 2 groups, with the height of the bar indicating the *P* value. Bars are colored green for comparisons where the *P* value is <.05, suggesting statistical significance, and blue for those where the *P* value is ≥.05. A red dashed line at the *P* value of .05 marks the conventional threshold for statistical significance.

Mayo Clin Proc Digital Health ■ September 2024;2(3):470-476 ■ https://doi.org/10.1016/j.mcpdig.2024.07.002
www.mcpdigitalhealth.org

in this context. Our model reported good performance in predicting hepatic decompensation with a median lead time of 1.5 years, highlighting its potential for early risk stratification in patients with PSC.

It is important to note that although our model is PSC-specific, there is a possibility that it may be detecting features common to cirrhosis in general. Future research should investigate whether this model can be applied to other forms of cirrhosis or if it truly captures PSC-specific changes. We speculate that the model is likely picking up on volumetric features and tissue density patterns characteristic of advanced liver disease in PSC, but further investigation is needed to confirm this.

Previous research has already highlighted the utility of ML algorithms and topological data analysis in analyzing laboratory data to predict hepatic decompensation and postliver transplant survival for patients with PSC.[14–16] Furthermore, automated deep learning algorithms have been reported to accurately detect patients with PSC using magnetic resonance cholangiopancreatography images.[17,18] However, the integration of CT imaging data with ML to predict outcomes in patients with PSC has not been previously explored. Our next crucial step involves validating the effectiveness of our predictive model on a large-scale, multicenter cohort. This methodological approach holds promise for not only predicting hepatic decompensation but also early detection of other PSC-related complications, such as cholangiocarcinoma. Moreover, the potential applications of this approach extend to more prevalent chronic liver diseases, like nonalcoholic fatty liver disease, where timely prediction and intervention are equally vital. By advancing the synergy between deep learning and medical imaging, our research offers new horizons for enhancing the management of hepatic complications, ultimately improving patient outcomes and health care practices.

CONCLUSION

The study underscores the potential of deep learning models, especially the 3D-DenseNet121, in predicting hepatic decompensation using CT imaging. This innovative approach could useful for predicting outcomes and intervention strategies for patients with chronic liver diseases, such as PSC. Subsequent investigations are imperative to validate and refine the initial findings, ensuring the model can robustly generalize across varied patient populations and health care settings. This iterative development and validation process will be vital in ensuring the model's efficacy and reliability in real-world clinical applications.

POTENTIAL COMPETING INTERESTS

Dr Venkatesh reports grants from the National Institutes of Health (NIH R01-EB17197, NIH R01-DK132718-A1, NIH R01-DK136731, NIH UH3-AA026887); U.S. Department of Defense PR181303_W81XWH-19-1-0583; and royalties for textbook from Springer-Verlag. Given his role as Editorial Board Member, Dr Erickson had no involvement in the peer-review of this article and has no access to information regarding its peer-review. The other authors report no competing interests.

ACKNOWLEDGMENTS

We would like to thank Mayo Artificial Intelligence laboratory members.

Abbreviations and Acronyms: **AUROC,** area under the receiving operating curve; **CNN,** convolutional neural network; **CT,** computed tomography; **ML,** machine learning; **PSC,** primary sclerosing cholangitis; **3D,** 3-dimensional

Correspondence: Address to Bradley J. Erickson, MD, PhD, Department of Radiology, Mayo Clinic, 200 1st St SW, Rochester, MN 55905 (Twitter: @Slowvak; @MRElastography).

ORCID

Bradley J. Erickson: https://orcid.org/0000-0001-7926-6095

REFERENCES

1. Eaton JE, Talwalkar JA, Lazaridis KN, Gores GJ, Lindor KD. Pathogenesis of primary sclerosing cholangitis and advances in diagnosis and management. *Gastroenterology*. 2013;145(3):521-536. https://doi.org/10.1053/j.gastro.2013.06.052.
2. Ponsioen CY, Chapman RW, Chazouillères O, et al. Surrogate endpoints for clinical trials in primary sclerosing cholangitis: review and results from an International PSC Study Group consensus process. *Hepatology*. 2016;63(4):1357-1367. https://doi.org/10.1002/hep.28256.
3. Singh Y, Jons W, Sobek JD, et al. Betti-number based machine-learning classifier framework for predicting the hepatic decompensation in patients with primary sclerosing cholangitis. In: *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)*. IEEE; 2022:159-162.

4. Eaton JE, Dzyubak B, Venkatesh SK, et al. Performance of magnetic resonance elastography in primary sclerosing cholangitis. *J Gastroenterol Hepatol*. 2016;31(6):1184-1190. https://doi.org/10.1111/jgh.13263.
5. Erickson BJ, Korfiatis P, Kline TL, Akkus Z, Philbrick K, Weston AD. Deep learning in radiology: does one size fit all? *J Am Coll Radiol*. 2018;15(3 Pt B):521-526. https://doi.org/10.1016/j.jacr.2017.12.027.
6. Indolia S, Goswami AK, Mishra SP, Asopa P. Conceptual understanding of convolutional neural network- a deep learning approach. *Proc Comput Sci*. 2018;132:679-688. https://doi.org/10.1016/j.procs.2018.05.069.
7. Singh Y, Kelm ZS, Faghani S, et al. Deep learning approach for differentiating indeterminate adrenal masses using CT imaging. *Abdom Radiol (NY)*. 2023;48(10):3189-3194. https://doi.org/10.1007/s00261-023-03988-w.
8. Bowlus CL, Arrivé L, Bergquist A, et al. AASLD practice guidance on primary sclerosing cholangitis and cholangiocarcinoma. *Hepatology*. 2023;77(2):659-702. https://doi.org/10.1002/hep.32771.
9. Rouzrokh P, Khosravi B, Faghani S, et al. Mitigating bias in radiology machine learning: 1. Data handling. *Radiol Artif Intell*. 2022;4(5):e210290. https://doi.org/10.1148/ryai.210290.
10. Handelman GS, Kok HK, Chandra RV, et al. Peering into the black box of artificial intelligence: evaluation metrics of machine learning methods. *AJR Am J Roentgenol*. 2019;212(1):38-43. https://doi.org/10.2214/AJR.18.20224.
11. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE; 2017:4700-4708.
12. Moassefi M, Faghani S, Conte GM, et al. A deep learning model for discriminating true progression from pseudoprogression in glioblastoma patients. *J Neurooncol*. 2022;159(2):447-455. https://doi.org/10.1007/s11060-022-04080-x.
13. The MONAI Consortium. Project MONAI. Zenodo. 2020: https://zenodo.org/record/4323059. Accessed July 20, 2024.
14. Eaton JE, Vesterhus M, McCauley BM, et al. Primary sclerosing cholangitis risk estimate tool (PREsTo) predicts outcomes of the disease: a derivation and validation study using machine learning. *Hepatology*. 2020;71(1):214-224. https://doi.org/10.1002/hep.30085.
15. Singh Y, Jons WA, Eaton JE, et al. Algebraic topology-based machine learning using MRI predicts outcomes in primary sclerosing cholangitis. *Eur Radiol Exp*. 2022;6(1):58. https://doi.org/10.1186/s41747-022-00312-x.
16. Andres A, Montano-Loza A, Greiner R, et al. A novel learning algorithm to predict individual survival after liver transplantation for primary sclerosing cholangitis. *PLoS One*. 2018;13(3):e0193523. https://doi.org/10.1371/journal.pone.0193523.
17. Venkatesh SK, Welle CL, Miller FH, et al. Reporting standards for primary sclerosing cholangitis using MRI and MR cholangiopancreatography: guidelines from MR Working Group of the International Primary Sclerosing Cholangitis Study Group. *Eur Radiol*. 2022;32(2):923-937. https://doi.org/10.1007/s00330-021-08147-7.
18. Ringe KI, Vo Chieu VD, Wacker F, et al. Fully automated detection of primary sclerosing cholangitis (PSC)-compatible bile duct changes based on 3D magnetic resonance cholangiopancreatography using machine learning. *Eur Radiol*. 2021;31(4):2482-2489. https://doi.org/10.1007/s00330-020-07323-5.