Journal of Pathology Informatics 16 (2025) 100416

Contents lists available at ScienceDirect

Journal of Pathology Informatics

journal homepage: www.elsevier.com/locate/jpi

Liver fibrosis classification on trichrome histology slides using weakly supervised learning in children and young adults

Mahdieh Shabanian [a,1], Zachary Taylor [b,1], Christopher Woods [b,c], Anas Bernieh [c], Jonathan Dillman [b,e], Lili He [b,d], Sarangarajan Ranganathan [c,d], Jennifer Picarsic [b,c,d,f], Elanchezhian Somasundaram [b,e,*]

[a] University of Utah, Biomedical Informatics Department, Salt Lake City, UT, United States
[b] Cincinnati Children's AI Imaging Research (CAIIR) Center, Cincinnati, OH, United States
[c] Cincinnati Children's Hospital Division of Pathology, University of Cincinnati College of Medicine, Cincinnati, OH, United States
[d] Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH, United States
[e] Department of Radiology, Cincinnati Children's Hospital Medical Center, University of Cincinnati College of Medicine, Cincinnati, OH, United States
[f] Department of Pathology, University of Pittsburgh School of Medicine, UPMC Children's Hospital, Pittsburgh, PA, United States

ARTICLE INFO

Keywords:
Deep learning
Liver fibrosis
Trichrome
Pediatrics
METAVIR
Ishak

ABSTRACT

Background: Traditional liver fibrosis staging via percutaneous biopsy suffers from sampling bias and variable inter-pathologist agreement, highlighting the need for more objective techniques. Deep learning models for disease staging from medical images have shown potential to decrease diagnostic variability, with recent weakly supervised learning strategies showing promising results even with limited manual annotation.
Purpose: To study the clustering-constrained attention multiple instance learning (CLAM) approach for staging liver fibrosis on trichrome whole slide images (WSIs) of children and young adults.
Methods: This is an ethics board approved retrospective study utilizing 217 trichrome WSI from pediatric liver biopsies for model development and testing. Two pediatric pathologists scored WSI using two liver fibrosis staging systems, METAVIR and Ishak. Cases were then secondarily categorized into either high- or low-stage liver fibrosis and used for model development. The CLAM pipeline was used to develop binary classification models for histological liver fibrosis. Model performance was evaluated using area under the curve (AUC), accuracy, sensitivity, specificity, and Cohen's Kappa.
Results: The CLAM models showed strong diagnostic performance, with sensitivities up to 0.76 and AUCs up to 0.92 for distinguishing low- and high-stage fibrosis. The agreement between model predictions and average pathologist scores was moderate to substantial (Kappa: 0.57–0.69), whereas pathologist agreement on the METAVIR and Ishak scoring systems was only fair (Kappa: 0.39–0.46).
Conclusions: CLAM pipeline showed promise in detecting features important for differentiating low- and high-stage fibrosis from trichrome WSI based on the results, offering a promising objective method for liver fibrosis detection in children and young adults.

Introduction

Chronic liver disease progression is indicated by liver fibrosis and is traditionally diagnosed through percutaneous liver biopsies. Even though this method of diagnosis is standard, it has several challenges that include sampling errors, invasiveness, and variability in interpretation by pathologists. These challenges are compounded in pediatric and young adult patients because of the need for age specific considerations both in diagnosis and treatment, emphasizing the need for more precise diagnostic methods.

In liver biopsy pathology, staining with Masson trichrome stain[1] is routinely performed to visually detect collagen deposition (i.e., blue staining). Based on the extent and pattern of abnormal collagen deposition, different staging systems have been implemented to assess the extent of disease. All systems offer a structured framework for staging liver fibrosis, using numerical scores that assess a continuum from no fibrosis to low-stage fibrosis (e.g., mild portal fibrosis) to high-stage fibrosis (e.g., bridging fibrosis and cirrhosis), which is likely to be clinically meaningful. The two most frequently used fibrosis scoring systems are the METAVIR (i.e., scale F0–F4

* Corresponding author at: Cincinnati Children's Hospital Medical Center, Radiology Department, 3333 Burnet Avenue, MLC 5033, Cincinnati, OH 45229, United States.
E-mail address: Elanchezhian.Somasundaram@cchmc.org (E. Somasundaram).
[1] Mahdien Shabanian and Zachary Taylor are co-first authors

http://dx.doi.org/10.1016/j.jpi.2024.100416
Received 29 September 2024; Received in revised form 6 December 2024; Accepted 6 December 2024
Available online 11 December 2024

with no fibrosis = F0 to cirrhosis = F4) and Ishak (i.e., scale 0–6 with no fibrosis = 0 to cirrhosis = 6).[2] Whereas these assessments are based on morphological features observed in trichrome-stained tissue sections, routine pathology assessments are still fraught with intra- and inter-observer variability.[3] Thus, a more reliable and consistent method for detecting and measuring the quantity and extent of fibrosis on whole slide images (WSIs) has been long sought after.[4–6]

Computational pathology has heralded a new era in disease diagnosis aid and management and has offered tools that can increase pathologists' ability to improve diagnostic accuracy and efficiency. Deep learning[7] has enabled the analysis of complex biomedical images in a more detailed and consistent manner than achieved by human experts alone.[8–10] However, the effectiveness of initial deep learning models based on supervised learning are limited by their need for large, extensively annotated datasets. Obtaining well-curated annotated data has practical challenges, even more evident in pediatric cohorts with fewer available datasets. Especially for liver fibrosis quantification, previous methods[11] have relied on pathologists' annotation of the fibrosis regions in the WSI, a time-consuming process.

Recently, techniques to learn visual representations without relying heavily on annotated data have become popular. These techniques can be broadly classified into semi-supervised, self-supervised, or weakly supervised techniques. Semi-supervised learning[12] combines labeled and unlabeled data to iteratively improve model performance without requiring a large quantity of manual annotations. Self-supervised learning[13] techniques involve designing learning objectives that do not require annotation, such as masking image patches and requiring the model to learn the masked patch resulting in models that learn useful features which can be adapted to downstream tasks with a small, annotated dataset. Weakly supervised learning[14] derives concepts from both self- and semi-supervised approaches and uses higher level annotations for supervision, such as using slide-level labels without requiring pixel level annotations.

The clustering-constrained attention multiple instance learning (CLAM) pipeline[15] uses the weakly supervised learning approach to perform multi-instance classification of WSI with slide-level labels and was originally introduced for cancer classification and sub-typing tasks with impressive results. The CLAM pipeline can also identify sample regions in the WSI that are important for a given classification, enabling a mechanism to verify model predictions.

In this work, we explore the feasibility of using the CLAM pipeline to develop a classifier model for liver fibrosis staging on trichrome WSI from pediatric patients. Specific objectives include to: (1) establish pediatric pathologist agreement in liver fibrosis staging of trichrome WSI in pediatric patients using two scoring systems, METAVIR and Ishak; and (2) explore the performance of the CLAM pipeline to perform binary classification of trichrome WSI into high- and low-stage fibrosis categories and compare against pathologist's performance.

Materials and methods

This was an institutional review board approved, retrospective study. The need for participant informed consent was waived.

Dataset

Patients who underwent liver biopsies for the diagnosis of liver disease between January 2011 and December 2020 were included in this study. Trichrome staining was performed for all liver biopsies, and all trichrome-stained slides were scanned using the Leica Aperio® AT2 scanner. WSI was performed with a 40× objective lens at a maximum scanning resolution of 0.25 μm/pixel. The scanned images, saved in SVS file format, were reviewed by study pathologists using the Leica Imaging Management System,[16] eSlide Manager, on a high-definition monitor with a resolution of 1920 × 1080 pixels. Scanned images were stored and exported to an external storage device for additional image analysis. 227 pediatric percutaneous liver biopsies were separately reviewed by the two fellowship-trained pediatric hepatopathologists, blinded to one another, during this manual annotation process.

Manual scoring

Manual fibrosis staging from scanned trichrome WSI was performed by study pathologists (A.B.—12 years of experience and S.R.—39 years of experience) using two staging systems, METAVIR (5 fibrosis levels, F0–F4) and Ishak (7 fibrosis levels, 0–6). Images that were inadequate for proper diagnosis (e.g., due to poor staining) were flagged during manual annotation and were excluded from additional image analysis.

Ground-truth binary labels

In this feasibility study, due to the unbalanced number of cases with scores across the different fibrosis levels in both scoring systems, participants were placed into two groups: (1) non-significant/low-stage fibrosis and (2) clinically significant/high-stage fibrosis. The cut-off points for each staging system were determined through expert consultation with the study pathologists. For both scoring systems, scores less than 2 (<2) (for both METAVIR and Ishak) were defined as non-significant/low-stage fibrosis, and scores greater than or equal to 2 (≥2) (for both METAVIR and Ishak) were defined as clinically significant/high-stage fibrosis, as shown in Fig. 1. Notably, for CLAM model training, participant grouping was performed after averaging the two pathologists' scores from each case for both the METAVIR and Ishak systems. For example, an average Ishak score of 1.5 fell into the clinically non-significant/low-stage group, whereas an average Ishak score of 2.5 fell into the clinically significant/high-stage group. Separate CLAM models were trained using the binary labels generated from each scoring system. To assess the performance of the model at different cut-off points from the pathologist-selected cut-offs, models were also trained using binary labels generated with cut-offs of 2.5 and 3. CLAM models will be referred to as Ishak and METAVIR models, denoting the ground truth used to train them, for this manuscript.

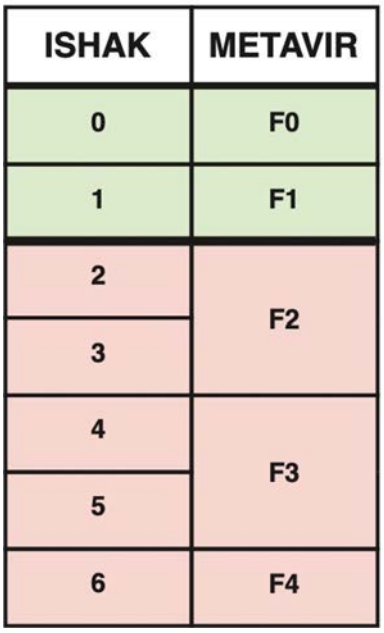

Fig. 1. METAVIR and Ishak scoring systems along with the respective binary grouping. The scores in green represent low-stage fibrosis and the scores in red represent high-stage fibrosis. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Model framework

The CLAM[15] pipeline has four main stages. In the first stage, the WSI is segmented to separate the tissue from background pixels and the tissue regions are patched into square patches of size 256 × 256 pixels. In the second stage, the tissue patches are encoded into a latent feature space, an one-dimensional vector of size 1024, using the CONCH pretrained encoder model[17] that was trained using contrastive learning on over 1.17 million image–caption pairs. In the third stage, the extracted patch level features are used as input to an attention-based network that ranks the patches in terms of their importance for the slide-level labels. This allows aggregation of patch-level information into slide-level representations through a mechanism called attention pooling. Separate branches of the attention network are used to rank the patches for each predicted label such that all patches in the WSI contribute to the model training. Finally, to further improve the learning patch level representations, the slide-level labels and attention scores are used to generated pseudo-patch-level labels to train a clustering layer that can separate patch groups that are highly and weakly important for a particular slide-level prediction. The attention scores can also be used to generate heatmaps to visualize the regions of interest that are most relevant for a given prediction and help with the interpretability of the model.

Training

Models were trained on an NVIDIA DGX server with 80 GB H100 GPUs. A 5-fold cross-validation was performed using a 70%–10%–20% split for training, tuning, and testing, respectively. This approach allowed us to assess the model's ability to learn and generalize new and unseen data. The tuning dataset was used to monitor the training process. The test dataset was constructed to ensure that exams were unique across all folds and that patients with multiple scans were assigned to the same split. The model hyperparameters were selected based on initial experiments to select the best configuration. Table 1 shows the hyperparameter settings used to train both the METAVIR and Ishak models in this work.

Evaluation metrics

Linearly weighted Fleiss' kappa statistics were used to evaluate inter-pathologist agreement for METAVIR and Ishak histological fibrosis staging. Additionally, Cohen's kappa statistics were used to assess agreement after categorizing the individual pathologist fibrosis scores into binary groups —low- and high-stage fibrosis. To further assess inter-observer variability, the standard deviations of the kappa statistics were estimated using a bootstrapping technique. This involved generating 1000 bootstrap samples from the original dataset by sampling with replacement, allowing for the calculation of the variability of the Kappa estimates across these samples.

Area under the curve (AUC), accuracy, precision-recall area under the curve (PR-AUC), sensitivity and specificity were used to evaluate the diagnostic performance of CLAM models. Agreement between model predictions and pathologist classification was also assessed using Cohen's kappa statistics. Standard deviations for these metrics were derived from the 5-fold cross-validation process. The sensitivity and specificity variation in the model performance when the binary labels are generated using different cut-off values was also plotted.

Table 1
CLAM pipeline[15] hyperparameters used for training the METAVIR and Ishak binary classification models.

Hyperparameter	Selection
Slide-level loss function	Cross-entropy
Weighted sampling of classes	False
Instance-level clustering	True
Instance-level clustering loss function	Support vector machine loss
Number of patches per class	8
Drop-out factor	0.25
Optimizer	Adam
Model type	Clam-mb (Multiple attention branches)

Results

Dataset details

Out of the 227 WSIs reviewed by the pathologists, 10 slides were deemed non-diagnostic and were subsequently excluded. The final dataset included 217 trichrome WSIs from 201 patients. 141 male patients and 76 female patients (mean age = 14.1 years; range: 2–27 years). Study sample demographics are presented in Table 2.

Table 2
Demographic distribution of the 201 patients in the study dataset.

Race	Total	Male	Female	Mean age in years (min-max)
Black	14	5	9	16.3 (12.7–19.0)
Other	24	16	8	11.8 (5.2–18.0)
White	163	110	53	14.3 (2.0–27.4)

Inter-pathologist agreement

Tables 3 and 4 show study pathologist agreement for histological fibrosis classification using the METAVIR and Ishak scoring systems. Using linear weighting, agreement was fair for both the METAVIR and Ishak systems (kappa statistics of 0.38 ± 0.04 and 0.39 ± 0.04, respectively). There was moderate agreement between readers after placing patients into two groups, high- and low-stage fibrosis, when using the METAVIR system (kappa statistic of 0.46 ± 0.06), whereas there was fair agreement when using the Ishak system (kappa statistic of 0.39 ± 0.05).

CLAM model performance

Table 5 presents the 5-fold cross-validation results on the test dataset for the METAVIR and Ishak models, both of which were trained on binary classes generated from the average scores of the two study pathologists. The accuracy, sensitivity, and specificity values were calculated using a threshold value of 0.5 for the predicted probabilities. AUC values were calculated to assess the model's ability to predict high-stage liver fibrosis (class B). The models demonstrated moderate agreement with the average pathologist

Table 3
Distribution of scores among study pathologists using the METAVIR scoring system from 217 liver biopsy trichrome stained samples.

METAVIR score	Pathologist 1	Pathologist 2
F0	13	78
F1	91	37
F2	91	54
F3	19	47
F4	3	1

Table 4
Distribution of scores among study pathologists using the Ishak scoring system from 217 liver biopsy trichrome stained samples.

Ishak score	Pathologist 1	Pathologist 2
0	11	78
1	63	62
2	82	20
3	44	23
4	11	13
5	5	20
6	1	1

Table 5
5-fold cross validation results of the METAVIR and Ishak CLAM models for binary classification (low- vs. high-stage fibrosis).

Model	AUC	Accuracy	PR-AUC	Sensitivity	Specificity	Kappa
METAVIR	0.88 ± 0.07	0.80 ± 0.07	0.83 ± 0.08	0.76 ± 0.16	0.82 ± 0.10	0.57 ± 0.15
Ishak	0.92 ± 0.04	0.85 ± 0.07	0.89 ± 0.03	0.75 ± 0.12	0.93 ± 0.08	0.69 ± 0.15

scores, with a Cohen's kappa statistic of 0.57 ± 0.15 for the METAVIR model and 0.69 ± 0.15 for the Ishak model.

Fig. 2 shows the confusion matrices for the METAVIR and Ishak models, respectively, generated from all the test samples across all folds during 5-fold cross-validation.

Fig. 3 shows the model performance for sensitivity and specificity in distinguishing between low- and high-stage liver fibrosis when the binary cut-off value for grouping the METAVIR and Ishak scores was changed between 2 and 3. The sensitivity decreased for both models, while the specificity increased, indicating that the selected binary cut-off of 2 provides the best balance between sensitivity and specificity.

Heatmap visualization

Fig. 4 shows an example of a trichrome-stained WSI categorized into the high-stage fibrosis group (scored METAVIR F4 by both study pathologists). The CLAM heatmap generated from the attention scores for each patched tissue region is calculated by the METAVIR model. The figure also shows eight patches as determined by CLAM model that were highly important in classifying the slide into the clinically significant fibrosis group. The patches accurately reflect areas of blue staining (i.e., histological fibrosis).

Discussion

Staging of histological liver fibrosis from pediatric and young adult liver biopsy specimens by two expert pediatric pathologists showed only fair inter-rater agreement (METAVIR: 0.38 ± 0.04 and Ishak: 0.39 ± 0.04) using weighted Kappa assessment. When the scores were grouped into two categories to represent patients with low-stage and clinically significant fibrosis, the inter-rater agreement slightly increased for the METAVIR system (0.46 ± 0.06, moderate), while remaining unchanged for the Ishak system (0.39 ± 0.05, fair). For comparison, in the original French METAVIR study[18] on chronic hepatitis patients, an interobserver agreement of 0.78 (substantial agreement) was reported for fibrosis scoring between the 10 pathologists that were part of the study. This study consisted of only 30 biopsy specimens from patients with a single liver disease (confirmed chronic hepatitis C) that were stained using H&E (Hematoxylin and Eosin Safran) and Masson Trichrome or Sirius Red. Another study reported an inter-observer weighted Kappa between two pathologists in the range of 0.57 (moderate) to 0.67 (substantial) using the Ishak scoring system.[19] This study had 65 biopsy specimens from patients with mixed liver disease and were stained using Sirius Red. Both the comparison studies did not provide patient demographics information.

Our results based on 5-fold cross-validation demonstrate that AI models built using the weakly supervised CLAM algorithm achieve acceptable performance, with the Ishak model slightly outperforming the METAVIR model in terms of AUC (METAVIR: 0.92 ± 0.04, Ishak: 0.88 ± 0.04). A similar trend was observed for accuracy (METAVIR: 0.80 ± 0.07, Ishak: 0.85 ± 0.07). PR-AUC (METAVIR: 0.83 ± 0.08, Ishak: 0.89 ± 0.05) and specificity (METAVIR: 0.82 ± 0.10, Ishak: 0.93 ± 0.08). However, for sensitivity, the METAVIR model slightly outperformed the Ishak model (METAVIR: 0.76 ± 0.16, Ishak: 0.75 ± 0.12).

Out of 217 total cases, the METAVIR and Ishak models misclassified 43 and 32 cases, respectively. In 17 of the METAVIR misclassified cases and 10 of the Ishak misclassified cases, the two study pathologists also disagreed on the diagnosis. This suggests that a significant portion of the models' errors may stem from cases that are inherently challenging, even for human experts.

A recent study[20] using weakly supervised multi-instance deep learning, a technique similar to the CLAM method, applied to Sirius red stained WSIs reported an AUC and accuracy of 0.87 and 0.79 for binary classification of liver fibrosis severity. These results are in line with those observed in the current investigation.

In terms of agreement between the models and the pathologist-annotated labels measured using the Cohen's Kappa statistic, the METAVIR model showed moderate agreement, achieving a Kappa of 0.57

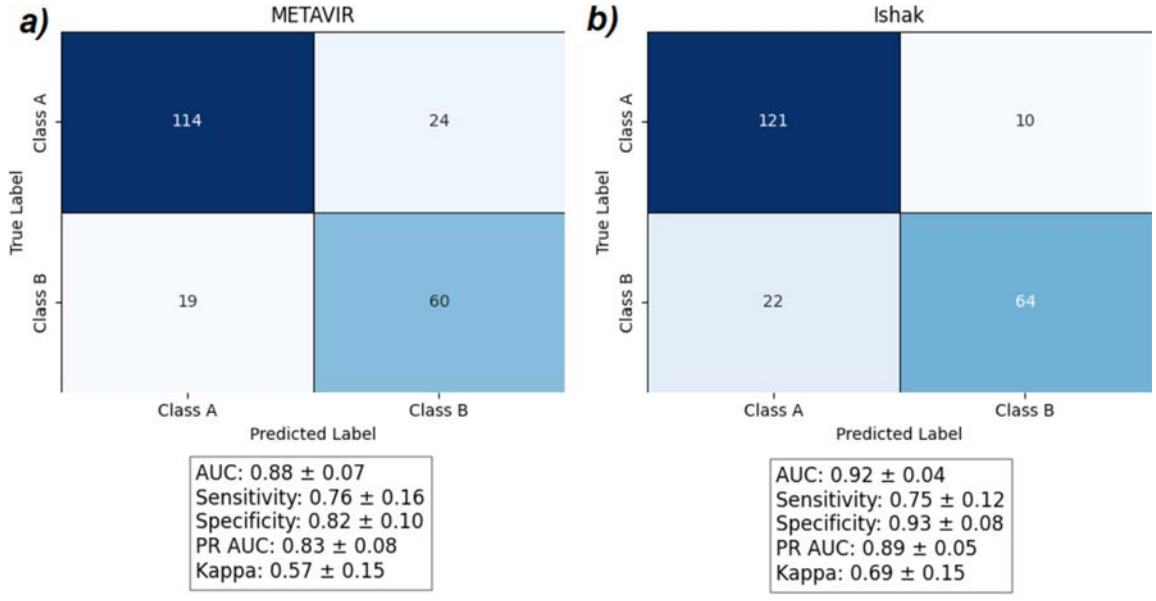

Fig. 2. (a) Aggregated confusion matrix from all the test samples in each fold from cross-validation for the METAVIR model (a) and the Ishak model (b). Class A—low-stage fibrosis; Class B—high-stage fibrosis.

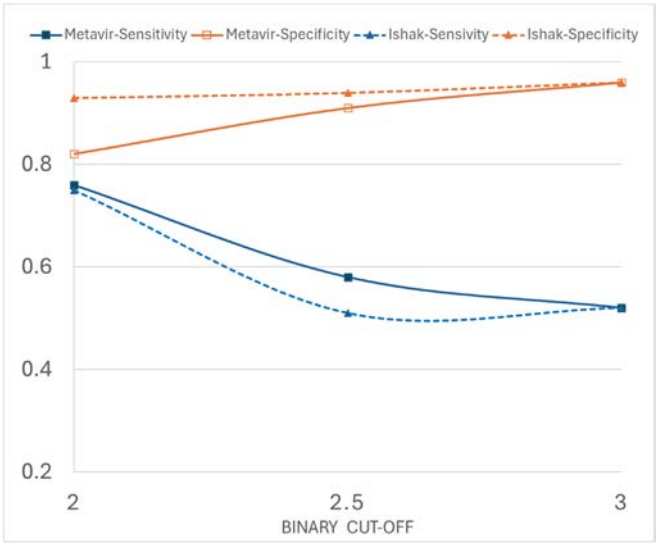

Fig. 3. Sensitivity and specificity of models trained on binary classification of the METAVIR and Ishak scores into low- and high-stage fibrosis at different threshold values.

± 0.15, whereas the Ishak model showed substantial agreement achieving a Kappa of 0.69 ± 0.15. This level of agreement was better than the agreement observed between the two individual pathologist scores, both before and after binary grouping, indicating that the AI models may offer a consistent and potentially valuable adjunct to human assessments, though further refinement is needed for clinical application.

Varying the binary cut-off values led to models with increased specificity; however, this came at the cost of a significant decrease in sensitivity. This suggests that a *prior* cut-off used in this study is optimal for distinguishing between low- and high-stage fibrosis given the characteristics of the dataset. The observed decrease in sensitivity with higher cut-off values is likely due to class imbalance in the dataset, which is exacerbated as the cut-off increases. From a clinical standpoint, a lower threshold is advantageous, as it allows for higher sensitivity in detecting high-stage fibrosis, which may be preferred for ensuring that cases requiring clinical intervention are not missed.

Heatmaps generated by the CLAM algorithm showing regions of attention provide novel insights into the model's working and model transparency. Upon initial investigation of the patches with high attention scores in predicting clinically significant fibrosis, it was noted that many patches have clear patterns of liver fibrosis, but there are few highly attended patches that do not have any noticeable disease patterns. Further analysis of the highly attended eight patches for both groups in a more systematic way is required to quantify the model's learned representations.

Our study has limitations. First, it is likely that our relatively small dataset size has prevented us from achieving maximum model performance. In cancer staging tasks reported in the original CLAM study, models were shown to have improved performance upon increasing the dataset size. Second, our dataset was highly unbalanced with regards to severe liver fibrosis (METAVIR F4 or Ishak 5–6). This imbalance was alleviated by placing patients into two groups (individuals with low-stage fibrosis and clinically significant fibrosis). This prevented training the models on the original fibrosis scoring systems. Finally, despite being experts in the field of pediatric liver pathology, there was only fair inter-rater agreement between study pathologists. As our reference standard for training AI models was based on the average METAVIR and Ishak scores from two

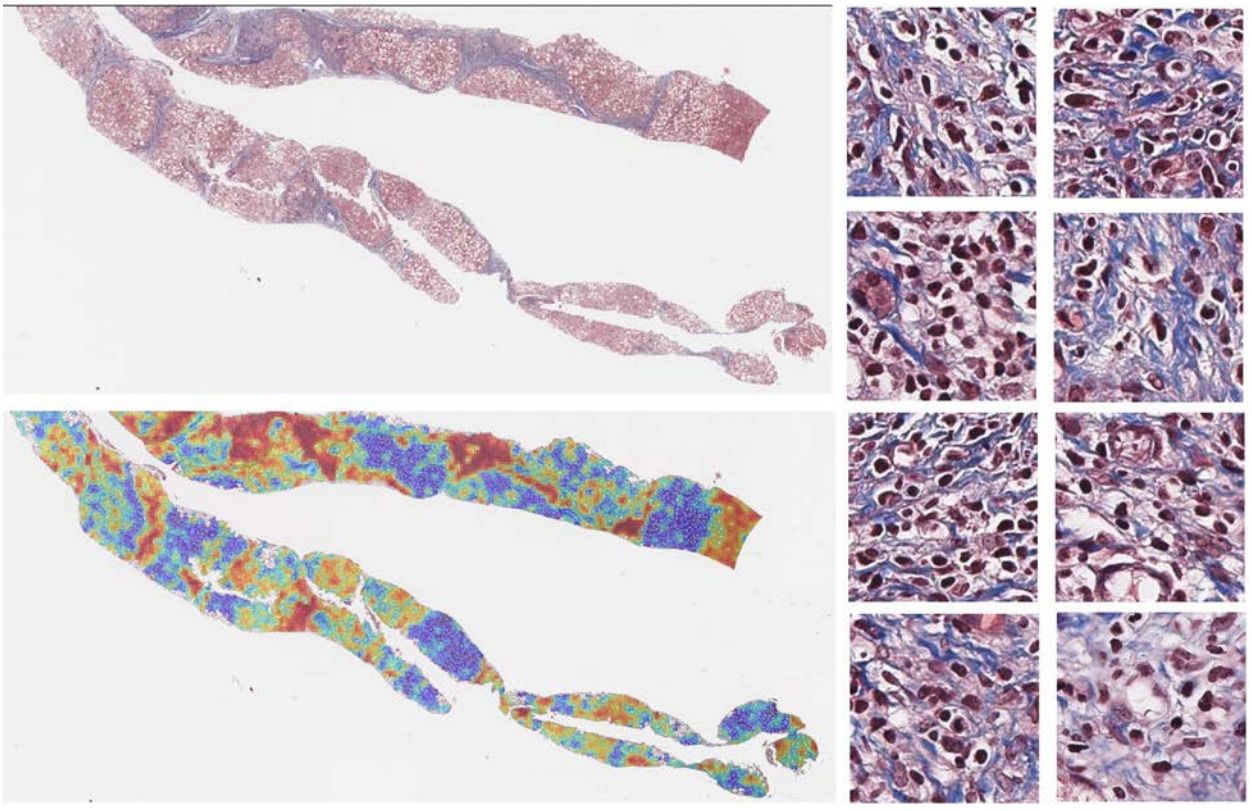

Fig. 4. Sample trichome-stained WSI scored as METAVIR F4 by both study pathologists, along with attention heatmaps, with red indicating regions of high importance and blue indicating regions of low importance in classifying the slide as a case of clinically significant fibrosis. The eight patches on the right determined by the CLAM model indicate the highly important regions in the WSI that contributed to the slide being classified correctly as a clinically significant fibrosis study. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

pathologists, it is conceivable that some of our instances of misclassification are the result of an imperfect reference standard and discrepant pathologist scoring leading to incorrect binary categorization.

Conclusion

AI models for histological liver fibrosis classification developed using the CLAM pipeline show promising results, with only minimal variability in performance across validation folds. This stability is crucial when considering the application of the model in clinical settings. Moreover, the level agreement achieved between the CLAM models and study pathologists was similar or higher than that observed between study pathologists. Additional studies in different patient populations and prospectively in the clinical setting would be helpful to further validate our results. Ultimately, such models have the potential to augment pathologist performance, decrease inter-pathologist variability, and positively impact patient outcomes.

Funding sources

This research was supported by a National Institutes of Health R01 grant (#R01-EB030582) and an Academic Research Committee Grant from Cincinnati Children's Hospital Medical Center.

During the preparation of this work the authors used ChatGPT to assist with clarity in writing. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Foot NC. The Masson trichrome staining methods in routine laboratory use. Stain Technol 1933;8(3):101–110.
2. Chowdhury AB, Mehta KJ. Liver biopsy for assessment of chronic liver diseases: a synopsis. Clin Exp Med 2023;23(2):273–285.
3. Rousselet MC, Michalak S, Dupré F, et al. Sources of variability in histological scoring of chronic viral hepatitis. Hepatology 2005;41(2):257–264.
4. Forlano R, Mullish BH, Giannakeas N, et al. High-throughput, machine learning–based quantification of steatosis, inflammation, ballooning, and fibrosis in biopsies from patients with nonalcoholic fatty liver disease. Clin Gastroenterol Hepatol 2020;18(9):2081–2090.[e9].
5. Gawrieh S, Sethunath D, Cummings OW, et al. Automated quantification and architectural pattern detection of hepatic fibrosis in NAFLD. Ann Diagn Pathol 2020;47, 151518.
6. Yu Y, Wang J, Ng CW, et al. Deep learning enables automated scoring of liver fibrosis stages. Sci Rep 2018;8(1), 16016.
7. Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. Nat Med 2019;25(1):24–29.
8. Coudray N, Ocampo PS, Sakellaropoulos T, et al. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nat Med 2018;24(10):1559–1567.
9. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. Nature 2020;577(7788):89–94.
10. Mitani A, Huang A, Venugopalan S, et al. Detection of anaemia from retinal fundus images via deep learning. Nat Biomed Eng 2020;4(1):18–27.
11. Panzeri D, Pagani E, Scodellaro R, et al. Fibrosis detection and quantification in whole slide images through deep learning. SPIE 2023;PC12622.PC126220K.
12. Learning S-S. Semi-supervised learning. CSZ2006 html 2006;5:2.
13. Gui J, Chen T, Zhang J, et al. A survey on self-supervised learning: algorithms, applications, and future trends. IEEE Trans Pattern Anal Mach Intell 2024;46(12):9052–9071.
14. Zhou Z-H. A brief introduction to weakly supervised learning. Natl Sci Rev 2018;5(1):44–53.
15. Lu MY, Williamson DF, Chen TY, Chen RJ, Barbieri M, Mahmood F. Data-efficient and weakly supervised computational pathology on whole-slide images. Nat Biomed Eng 2021;5(6):555–570.
16. Miscroscope Software. https://www.leica-microsystems.com/products/microscope-software/ 2023.
17. Lu MY, Chen B, Williamson DF, et al. A visual-language foundation model for computational pathology. Nat Med 2024;30(3):863–874.
18. Group FMCS, Bedossa P. Intraobserver and interobserver variations in liver biopsy interpretation in patients with chronic hepatitis C. Hepatology 1994;20(1):15–20.
19. Pavlides M, Birks J, Fryer E, et al. Interobserver variability in histologic evaluation of liver fibrosis using categorical and quantitative scores. Am J Clin Pathol 2017;147(4):364–369.
20. Naik SN, Forlano R, Manousou P, Goldin R, Angelini ED. Fibrosis severity scoring on Sirius red histology with multiple-instance deep learning. Biol Imaging 2023;3, e17.