Deep learning helps discriminate between autoimmune hepatitis and primary biliary cholangitis

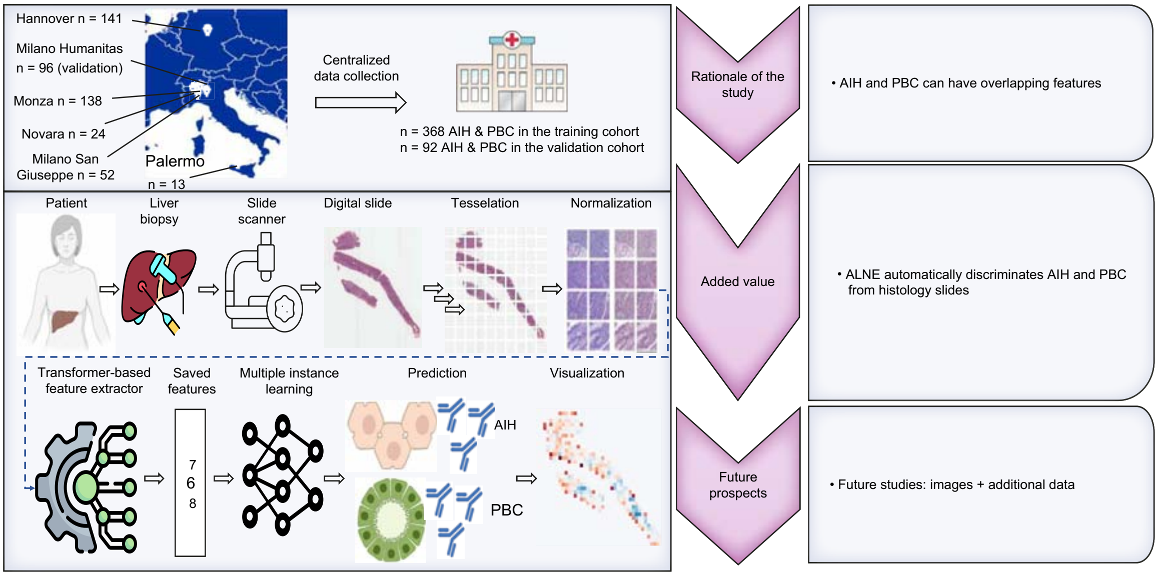
Authors

Alessio Gerussi, Oliver Lester Saldanha, Giorgio Cazzaniga, …, Pietro Invernizzi, Marco Carbone, Jakob Nikolas Kather

Correspondence

pietro.invernizzi@unimib.it (P. Invernizzi), jakob-nikolas.kather@alumni.dkfz.de (J.N. Kather).

Graphical abstract



Hannover n = 141
Milano Humanitas n = 96 (validation)
Monza n = 138
Novara n = 24
Milano San Giuseppe n = 52
Palermo n = 13
Centralized data collection
n = 368 AIH & PBC in the training cohort
n = 92 AIH & PBC in the validation cohort
Patient
Liver biopsy
Slide scanner
Digital slide
Tesselation
Normalization
Transformer-based feature extractor
Saved features
768
Multiple instance learning
Prediction
AIH
PBC
Visualization
Rationale of the study
Added value
Future prospects

- AIH and PBC can have overlapping features
- ALNE automatically discriminates AIH and PBC from histology slides
- Future studies: images + additional data

Highlights:

- ALNE is a transformer-based deep learning model that can accurately distinguish autoimmune hepatitis and primary biliary cholangitis.
- ALNE generates accurate predictions relying only on H&E slides without the support of any human annotation.
- The ALNE model demonstrated robust performance even with the diversity in scanning and digitizing techniques.
- In comparison with ALNE, general pathologists show poor inter-observer agreement when challenged with the same clinical task.

Impact and implications:

This study demonstrates the significant potential of the autoimmune liver neural estimator model, a transformer-based deep learning system, in accurately distinguishing between autoimmune hepatitis and primary biliary cholangitis using digitized liver biopsy slides without human annotation. The scientific justification for this work lies in addressing the challenge of differentiating these conditions, which often present with overlapping features and can lead to therapeutic mistakes. In addition, there is need for quantitative assessment of information embedded in liver biopsies, which are currently evaluated on qualitative or semi-quantitative methods. The results of this study are crucial for pathologists, researchers, and clinicians, providing a reliable diagnostic tool that reduces inter-observer variability and improves diagnostic accuracy of these conditions. Potential methodological limitations, such as the diversity in scanning techniques and slide colorations, were considered, ensuring the robustness and generalizability of the findings.

https://doi.org/10.1016/j.jhepr.2024.101198

Deep learning helps discriminate between autoimmune hepatitis and primary biliary cholangitis

Alessio Gerussi[1,2,†], **Oliver Lester Saldanha**[3,4,†], Giorgio Cazzaniga[5], Damiano Verda[6], Zunamys I. Carrero[3], Bastian Engel[7,8], Richard Taubert[7,8], Francesca Bolis[1,2], Laura Cristoferi[1,2], Federica Malinverno[1,2], Francesca Colapietro[9,10], Reha Akpinar[9,11], Luca Di Tommaso[9,11], Luigi Terracciano[9,11], Ana Lleo[9,10], Mauro Viganó[12], Cristina Rigamonti[13], Daniela Cabibi[14], Vincenza Calvaruso[15], Fabio Gibilisco[16,17], Nicoló Caldonazzi[18], Alessandro Valentino[19], Stefano Ceola[5], Valentina Canini[5], Eugenia Nofit[1,2], Marco Muselli[6], Julien Calderaro[20,21,22], Dina Tiniakos[23,24], Vincenzo L'Imperio[5], Fabio Pagni[5], Nicola Zucchini[5], Pietro Invernizzi[1,2,*], Marco Carbone[2,25,‡], Jakob Nikolas Kather[3,26,27,*,‡]

JHEP Reports **2025**. vol. 7 | 1–11

Background & Aims: Biliary abnormalities in autoimmune hepatitis (AIH) and interface hepatitis in primary biliary cholangitis (PBC) occur frequently, and misinterpretation may lead to therapeutic mistakes with a negative impact on patients. This study investigates the use of a deep learning (DL)-based pipeline for the diagnosis of AIH and PBC to aid differential diagnosis.

Methods: We conducted a multicenter study across six European referral centers, and built a library of digitized liver biopsy slides dating from 1997 to 2023. A training set of 354 cases (266 AIH and 102 PBC) and an external validation set of 92 cases (62 AIH and 30 PBC) were available for analysis. A novel DL model, the autoimmune liver neural estimator (ALNE), was trained on whole-slide images (WSIs) with H&E staining, without human annotations. The ALNE model was evaluated against clinico-pathological diagnoses and tested for interobserver variability among general pathologists.

Results: The ALNE model demonstrated high accuracy in differentiating AIH from PBC, achieving an area under the receiver operating characteristic curve of 0.81 in external validation. Attention heatmaps showed that ALNE tends to focus more on areas with increased inflammation, associating such patterns predominantly with AIH. A multivariate explainable ML model revealed that PBC cases misclassified as AIH more often had ALP values between 1 × upper limit of normal (ULN) and 2 × ULN, coupled with AST values above 1 × ULN. Inconsistency among general pathologists was noticed when evaluating a random sample of the same cases (Fleiss's kappa value 0.09).

Conclusions: The ALNE model is the first system generating a quantitative and accurate differential diagnosis between cases with AIH or PBC.

Introduction

Autoimmune hepatitis (AIH) and primary biliary cholangitis (PBC) are rare autoimmune liver diseases characterized by poorly defined disease etiology and frequent overlapping inflammatory features that challenge the diagnostic and treatment process.[1] These diseases do not represent a single, uniform condition; rather, they encompass a spectrum with diverse clinical presentations, underlying mechanisms, and outcomes.

Liver biopsy for histopathological assessment is essential in differentiation between PBC and AIH. Liver biopsy is particularly useful when serological tests yield negative or conflicting results, or when there are overlapping features of both conditions. In PBC patients, biopsy often reveals inflammatory changes like interface hepatitis in up to 25% of cases; however, this finding alone does not indicate the presence of AIH and can make it difficult to determine the most appropriate treatment for individual patients.[1] Similarly, recent data reveal that up to 83% of cases of AIH show biliary injury at diagnosis that can resolve over time.[2] Actual PBC–AIH overlap/variant syndrome is considered much rarer though. Misclassification could potentially lead to unnecessary, long-term immunosuppressive or choleretic therapy.

Nevertheless, the evaluation of liver biopsy remains a qualitative and subjective process, making it susceptible to misclassification and time-consuming. Current classification scores rely on semi-quantitative, subjective metrics and are

* Corresponding authors. Addresses: Division of Gastroenterology and Center for Autoimmune Liver Diseases, Department of Medicine and Surgery, University of Milano - Bicocca, Via Cadore 48, 20900 Monza (MB), Italy. Tel.: +39-039-233-2187 (P. Invernizzi); Else Kroener Fresenius Center for Digital Health, Technical University Dresden, Dresden, Germany (J.N. Kather).
E-mail addresses: pietro.invernizzi@unimib.it (P. Invernizzi), jakob-nikolas.kather@alumni.dkfz.de (J.N. Kather).

† Contributed equally as first authors

‡ Contributed equally as senior authors.

https://doi.org/10.1016/j.jhepr.2024.101198

further complicated by interobserver variability.[3,4] A pivotal study validating the latest PBC-specific grading and staging system, the Nakanuma system, revealed very limited interobserver agreement with a kappa value of 0.110 and a concordance rate of 36.9% for cholangitis. Similarly, for hepatitis activity, agreement was limited, with a kappa value of 0.197 and a concordance rate of 47%.[3]

Deep Learning (DL)-based image analysis can extract quantitative information from complex image data and could offer a solution to this problem. In general, DL has been proposed for a range of questions in liver pathology and has shown promising performance.[5] Convolutional neural networks (CNNs), and more recently transformer neural networks, have shown substantial capability to identify patterns in cancer histological images and can discern molecular classes of tumors without manual annotation, enabling an 'end-to-end' process.[6]

To date, there is increasing evidence supporting the use of DL in liver disease, particularly for metabolic dysfunction-associated steatotic liver disease and hepatocellular carcinoma.[7] Our group has shown that DL can help the classification of hepatocellular-cholangiocarcinomas.[7,8] Yet, there is a lack of data on the application of DL on unannotated liver biopsy slides in the field of autoimmune liver disease.

Our study aims to address this gap by developing and validating a DL model for the automated pre-classification of liver biopsies of AIH and PBC cases. To achieve this, we used digital liver biopsies with pathologist-derived diagnoses as ground truth, evaluating the potential of these models as an assisted system for differentiating between AIH and PBC. In this work, we introduce autoimmune liver neural estimator (ALNE), a state-of-the-art DL approach for the assessment of autoimmune liver diseases in H&E-stained whole-slide images (WSIs), which we have systematically evaluated and released under an open-source license, highlighting its potential as a significant auxiliary tool within the field of hepatopathology.

Patients and methods

Study population

This multicenter study involved a database search for liver biopsy samples from patients diagnosed with AIH and PBC across several institutions. These included the Departments of Pathology at Hannover Medical School, Fondazione IRCCS San Gerardo dei Tintori Monza, Istituto Clinico Humanitas, Ospedale Milano San Giuseppe, Ospedale di Novara, and Policlinico di Palermo, with cases collected from 1997 to 2023. All participating institutions are recognized as expert referral centers for liver diseases and (in part) members of the European Reference Network for rare liver diseases.

In our study, from a total of 368 cases that met the inclusion criteria, 354 were ultimately included in the training cohort; reasons for exclusion are reported in Table 1. The final cohort consisted of 258 cases of AIH and 96 cases of PBC (Table 1 and Fig. S1). An external validation cohort meeting the same inclusion and exclusion criteria of the training cohort was obtained from Istituto Clinico Humanitas, including 92 cases (62 AIH and 30 PBC) (Table 1 and Fig. S2).

The study was conducted in accordance with the Declaration of Helsinki. The institution responsible for the coordination of the study was the University of Milano-Bicocca, coordinator of the Italian PBC National Registry (ClinicalTrials.gov: NCT05151809) and the Italian AIH National Registry (ClinicalTrials.gov: NCT06078098). The study was approved by the University of Milan-Bicocca research ethics committee (study names: PBC322 and AIH Database), the steering committee of the PBC National Registry and the AIH National Registry, and the Research and Development Department of each collaborating hospital. The Ethics Board at the Medical Faculty of Technical University Dresden (BO-EK-444102022) approved the data analysis in this study.

Pathological samples

Glass slides holding formalin-fixed paraffin-embedded tissue sections stained with H&E were prepared at their respective institutions using diverse biopsy protocols and staining approaches, and either digitized in each center with different scanner vendors or sent for scanning at the coordinating center on an Aperio CS2 (Leica Biosystem, Nussloch, Germany) generating .svs files. Slides from collaborating centers were de-identified at their corresponding institutions and received at University of Milano-Bicocca for an initial analysis. Digitized slides were shared by the University of Milano-Bicocca with the Clinical Artificial Intelligence Laboratory at the University of Dresden for deep learning analysis. Our dataset comprising 354 patients and 561 WSIs, with some patients contributing multiple WSIs, was tessellated into image tiles of 244 pixels with 256 microns per pixel. These image tiles were utilized for training the deep learning model on liver biopsy samples.

Ground truth

Liver biopsy samples were locally assigned to either a diagnosis of AIH or PBC based on a dedicated case review performed in each center jointly by the clinical and the pathology team. Only cases related to baseline conditions, before starting immunosuppression or ursodeoxycholic acid (UDCA) were included in the analysis. A definite diagnosis of AIH and PBC was based on established guidelines;[9,10] in addition, the follow up of patients was essential for the clinical team to support the diagnosis based on the response to immunosuppression in AIH cases and UDCA in PBC cases. The slides were reviewed for their quality in each center, and the original diagnosis was confirmed or excluded after revision. The coordinating center has performed a downstream centralized review of the viability of the slides from the pathological point of view. Samples with other diagnoses were not included in the analysis. Only H&E-stained slides that met high standards of technical quality, histological adequacy, and scanning quality, ensuring they were suitable for histological diagnosis of either AIH or PBC. Slides that were too old, with faded staining, or histologically inadequate (i.e. containing fewer than 10 portal spaces) were centrally removed. In addition, any slides that were damaged (such as broken slides) or exhibited significant scanning artifacts were also excluded. No discordant assessments as regards the diagnosis of PBC or AIH were identified during this process. Because our study exclusively considered cases stained with H&E, no quantitative assessment of fibrosis was performed, being outside the scope of the study.

Table 1. Clinico-pathological features of all cohorts.

	Training		Validation	
Disease	**PBC**	**AIH**	**PBC**	**AIH**
n	96	258	31	65
Center n (%)				
HAN	7 (7.3)	126 (48.8)	MIL HUM 31 (100.0)	63 (100.0)
HSG	45 (46.9)	92 (35.7)		
MIL SG	12 (12.5)	40 (15.5)		
NOVARA	19 (19.8)	0 (0.0)		
PALERMO	13 (13.5)	0 (0.0)		
Female sex n (%)	88 (91.7)	177 (68.6)	27 (87.1)	43 (68.3)
Age at diagnosis (years)*	52 (45–58)	55 (45–65)	52 (43–58)	60 (51–68)
AST × ULN at diagnosis*	1.28 (0.96–2.21)	4.68 (3.70–24.19)	1.31 (0.94–1.85)	7.69 (3.33–15.85)
ALT × ULN at diagnosis*	1.61 (0.93–2.45)	5.90 (5.90–23.42)	1.46 (1.03–2.22)	8.09 (3.60–22.61)
Total bilirubin × ULN at diagnosis*	0.65 (0.46–0.91)	1.70 (0.81–7.86)	0.80 (0.64–0.96)	1.10 (0.83–2.32)
ALP × ULN at diagnosis*	1.63 (0.97–2.91)	0.97 (0.50–1.60)	1.32 (1.02–2.09)	0.89 (0.64–1.22)
GGT × ULN at diagnosis*	4.68 (2.53–7.45)	2.25 (2.25–4.66)	5.41 (3.74–11.73)	2.92 (1.67–5.76)
Albumin at diagnosis (g/dl)*	4.00 (4.00–4.05)	3.45 (3.00–4.00)	4.00 (4.00–4.50)	3.50 (3.00–4.00)
Sodium at diagnosis (mmol/L)*	140 (138–141)	139 (136–140)	140 (139–142)	139 (138–141)
Creatinine at diagnosis (mg/dl)*	0.70 (0.60–0.80)	0.75 (0.65–0.88)	1.00 (1.00–1.00)	1.00 (1.00–1.00)
INR at diagnosis*	1.06 (1.00–1.10)	1.00 (1.00–1.00)	1.00 (1.00–1.00)	1.00 (1.00–1.00)
PLT at diagnosis ($\times10^3$/μl)*	240 (207–294)	203 (146–256)	246 (222–293)	189 (155–254)
IgG × ULN at diagnosis*	0.81 (0.68–1.01)	1.24 (0.95–1.66)	0.81 (0.71–0.93)	1.04 (0.91–1.26)
ANA (%)				
NA	7 (7.3)	0 (0.0)	5 (16.1)	1 (1.6)
NEG	22 (22.9)	4 (1.6)	8 (25.8)	1 (1.6)
POS	67 (69.8)	254 (98.4)	18 (58.1)	61 (96.8)
AMA (%)				
NA	5 (5.2)	16 (6.2)	1 (3.2)	13 (20.6)
NEG	19 (19.8)	236 (91.5)	6 (19.4)	46 (73.0)
POS	72 (75.0)	6 (2.3)	24 (77.4)	4 (6.3)
gp210 (%)				
NA	28 (29.2)	123 (47.7)	27 (87.1)	62 (98.4)
NEG	51 (53.1)	135 (52.3)	2 (6.5)	1 (1.6)
POS	17 (17.7)	0 (0.0)	2 (6.5)	0 (0.0)
sp100 (%)				
NA	27 (28.1)	122 (47.3)	24 (77.4)	62 (98.4)
NEG	54 (56.2)	136 (52.7)	3 (9.7)	1 (1.6)
POS	15 (15.6)	0 (0.0)	4 (12.9)	0 (0.0)
SMA (%)				
NA	27 (28.1)	13 (5.0)	13 (41.9)	12 (19.0)
NEG	65 (67.7)	127 (49.2)	18 (58.1)	35 (55.6)
POS	4 (4.2)	118 (45.7)	0 (0.0)	16 (25.4)
LKM (%)				
NA	23 (29.1)	3 (1.2)	11 (35.5)	21 (33.3)
NEG	54 (68.4)	234 (94.7)	20 (64.5)	42 (66.7)
POS	2 (2.5)	10 (4.0)	0 (0.0)	0 (0.0)
SLA/LP (%)				
NA	45 (46.9)	79 (30.6)	28 (90.3)	60 (95.2)
NEG	50 (52.1)	174 (67.4)	3 (9.7)	2 (3.2)
POS	1 (1.0)	5 (1.9)	0 (0.0)	1 (1.6)
AST × ULN at 12 months*	0.81 (0.56–1.00)	0.84 (0.66–1.12)	0.82 (0.63–0.86)	0.80 (0.70–1.05)
ALT × ULN at 12 months*	0.76 (0.51–1.20)	0.74 (0.52–1.30)	0.83 (0.63–1.09)	0.67 (0.50–0.86)
Total bilirubin × ULN at 12 months*	0.58 (0.47–0.70)	0.60 (0.38–0.89)	0.76 (0.50–1.11)	0.70 (0.50–0.90)
ALP × ULN at 12 months*	1.10 (0.80–1.80)	0.62 (0.49–0.82)	1.00 (0.71–1.55)	0.41 (0.37–0.50)
GGT × ULN at 12 months*	1.81 (0.96–3.26)	0.80 (0.39–1.63)	2.29 (1.09–3.69)	0.68 (0.37–1.07)
IgG × ULN at 12 months*	0.77 (0.62–0.88)	0.70 (0.58–0.87)	0.70 (0.61–0.85)	0.72 (0.63–0.91)

AIH, autoimmune hepatitis; ALT, alanine aminotransferase; ANA, antinuclear antibodies; AMA, anti-mitochondrial antibodies; AST, aspartate transferase; GGT, gamma-glutamyl transferase; GP210, glycoprotein 210 antibodies; LKM, liver/kidney microsomal antibodies; INR, international normalized ratio; NA, not available; NEG, negative; PBC, primary biliary cholangitis; PLT, platelets; POS, positive; SLA/LP, soluble liver antigen/liver pancreas antibodies; SMA, smooth muscle antibodies; SP100, nuclear antigen antibodies; ULN, upper limit of normal.

*Median values and interquartile ratio (IQR).

Experimental setup

ALNE model

The ALNE model was trained on hundreds of gigapixel WSIs using patient-level labels, promoting scalability to larger datasets while avoiding the need for manual annotation (images did not include local annotations and were categorized using image-level classification labels).[11–13] The model's architecture combines a CTransPath-based self-supervised learning (SSL) framework for feature extraction with an

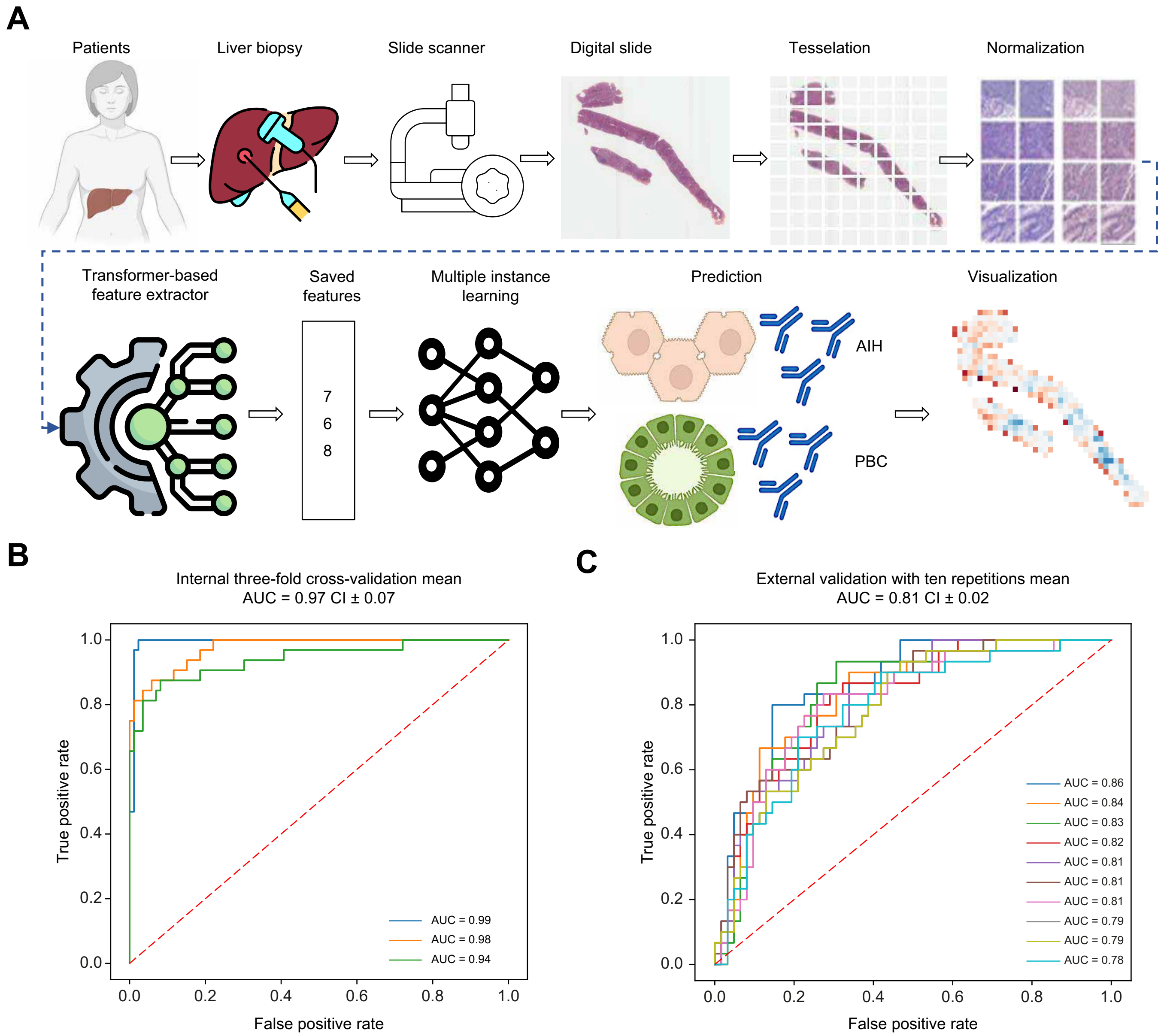


Fig. 1. Schematic of the workflow and results for the ALNE model. (A) Graphical representation of the workflow for classification of PBC and AIH on histopathology WSI data. The process involves transformer-based feature extraction and MIL techniques for effective classification and visualization. (B) Three-fold cross-validated AUROC scores for classifying PBC and AIH on the internal validation cohort utilizing the ALNE model. (C) AUROC scores, based on 10 repetitions, for the classification of PBC and AIH in the external validation cohort using the ALNE model. AIH, autoimmune hepatitis; ALNE, autoimmune liver neural estimator; AUROC, area under the receiver operating characteristic curve; MIL, multiple instance learning; PBC, primary biliary cholangitis; WSI, whole-slide image.

attention-based multiple instance learning (attMIL) aggregation method (Fig. 1A). We used the hybrid CNN-transformer model CTransPath[14] as the backbone architecture of an end-to-end image analysis pipeline. CTransPath is pre-trained with SSL on a dataset of weakly labeled histopathological images or clinical data, and functions as an effective joint local and global feature extractor. The CTransPath model was not fine-tuned for our application. Extracted features are then used to train an attMIL model to predict outcomes at the tile level. Following this, it executes patient-wise aggregation and utilizes an attention mechanism to assess the significance of various instances within a bag for weighing their contributions. This process involves merging predictions from multiple tiles that belong to the same patient, resulting in a comprehensive prediction for that patient. If any patient has more than one WSI then all the WSIs of a single patient are considered as a single datapoint. The attMIL has a multilayer perceptron (classifier network) (512 × 256), (256 × 2) with an attention mechanism. This is followed by a hyperbolic tangent (tanh) layer to obtain the tile-wise prediction score, which is aggregated patient-wise for the WSI data. For internal validation, we used a three-fold cross-validation approach, dividing the dataset into three equal parts or 'folds'. In this process, the model was trained on two folds and validated on the remaining one, with this procedure being repeated three times using different folds as the validation set. The cohort was divided at the patient level, ensuring a balanced representation of each diagnosis across all groups. For external validation, the model was trained on the entire internal cohort and was tested on the external validation cohort. The ALNE model was trained to classify cases as AIH or PBC. To enhance model interpretability and introspection, we utilized high-resolution heatmaps to visualize the model's predictions, highlighting regions of morphological diagnostic significance within the liver biopsies.

Analysis for interobserver variability

To compare the performance of the ALNE model with human readers, five general pathologists were recruited; their scoring was evaluated in terms of interobserver variability and the variability between the ALNE model and individual experts. Nineteen liver biopsies from the validation cohort were selected: five best predicted AIH, five best predicted PBC, four worst predicted AIH, and five worst predicted PBC. All pathologists were blinded to pathology reports and previous assessments made on these biopsies. The cases were presented to the pathologists in random order through a digital web-based platform. A week before the challenge the Department of Pathology of the coordinating center organized an online webinar led by the liver pathologist aimed at reviewing the main pathological features of AIH and PBC. To have a fair comparison with the training of the ALNE model, only H&E slides were provided. No timing constraints were imposed on the pathologists.

We assessed the interobserver agreement among pathologists by calculating Cohen's kappa for each pair. Subsequently, we calculated the agreement between each individual expert pathologist and the ALNE model.[15] Lastly, we also calculated the interobserver agreement among pathologists with the Fleiss kappa for multiple categorical variables.[16] Further details about these analyses are provided in the Statistics section.

Model explainability

To interpret and visualize the predictions made by our model, we used attention heatmaps on the external validation dataset. These heatmaps highlight areas interpreted as diagnostically significant for distinguishing the two diseases, thereby also demonstrating the model's generalizability. In each WSI, normalized attention scores were allocated to their respective spatial locations, signifying the model's assessment of the diagnostic importance of different biopsy regions. Regions with higher attention scores are indicative of morphological features with greater diagnostic relevance. It is important to emphasize that these attention heatmaps do not specifically delineate areas of AIH or PBC. Instead, they provide insight into the relative significance each liver biopsy region holds in contributing to the model's predictions.

To improve explainability, the score obtained from the ALNE model was imported into the Rulex software (Rulex Inc., Genoa, Italy) for integration with clinical data. The Rulex Platform (www.rulex.ai) allows the handling of input data from diverse sources, and it can be seamlessly integrated with other tools or custom code.[17] Together with standard classification and regression techniques, the platform provides emphasis on interpretability, while also incorporating logic learning machine (LLM). LLM is an algorithm designed to model problems through intelligible, and possibly overlapping, if-then rules. This algorithm has been pivotal in the development of various applications, both in industrial settings[18] and in research.[19]

Statistics

The primary endpoint for evaluating the performance of the model was the area under the receiver operating characteristic curve (AUROC). To ensure robustness, the experiments were iterated 10 times using different random seeds. Confidence intervals (CIs) were calculated to indicate the plausible range of the true value of our measurements. Wider intervals suggest heightened uncertainty, whereas narrower ones imply greater precision. These intervals are typically set at a predetermined confidence level, such as 95%. Key metrics in the evaluation of diagnostics tests or models include sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). Sensitivity refers to the ability of the test to accurately identify true positive cases, whereas specificity measures the ability to accurately identify true negatives. PPV evaluates the probability that a positive test result is indeed a true positive, and NPV assesses the probability that a negative result is accurate.

Cohen's kappa index was utilized as a statistical measure to evaluate the level of agreement or concordance between two subgroups, either between two pathologists or between AI and pathologists. A kappa value of 1 signifies perfect agreement. In contrast, a value of 0 indicates no agreement beyond chance, whereas negative values imply agreement among raters is less than what would be expected by chance, signifying systematic disagreement. Values falling between these extremes represent varying degrees of agreement.

Contrasting with Cohen's kappa, which measures agreement between two raters, Fleiss's kappa extends this concept to accommodate any fixed number of raters. This makes it particularly suitable for studies involving multiple raters, such as in our case with several pathologists. Fleiss's kappa was utilized to provide a more comprehensive understanding of the agreement among pathologists. Similar to Cohen's kappa, the range of Fleiss's kappa spans from -1, indicating complete disagreement, to 1, signifying complete agreement. A value of 0 in Fleiss's kappa, like in Cohen's, suggests that the agreement is no better than what would be expected by chance.

Results

Assessment of DL-based ALNE model for differentiating PBC and AIH in liver pathology

We evaluated the ability of our DL-based ALNE model to distinguish PBC from AIH in liver pathology via comprehensive experiments encompassing different validation approaches.

Upon training our model with our internal validation cohort, we proceeded to evaluate it by conducting three-fold cross-validation. Our ALNE model demonstrated a remarkable performance, achieving an AUROC of 0.97 (CI: ±0.07) (Fig. 1B). This result robustly validates the model's proficiency in accurately classifying PBC and AIH cases. The ALNE model generates for each prediction a classification score; the closer to 1, the more likely AIH, and the closer to 0, the more likely PBC. The distribution of the score from the ALNE model is shown in Fig. S3 for AIH and PBC classes, respectively.

To assess real-world applicability, we transitioned to external validation experiments. The model maintained strong performance, by achieving an AUROC result at 0.81 (CI: ±0.02), demonstrating minimal variance across ten separate validation runs (Fig. 1C). This consistency emphasizes the reliability of our model in clinical settings. Furthermore, aggregating patient-wise probabilities across ten repetitions provided the following diagnostic accuracy metrics: true positives (TP) = 58, false positives (FP) = 17, true negatives (TN) = 13, and false negatives (FN) = 4. These metrics give rise to a sensitivity of 0.94, specificity of 0.43, PPV of 0.77, and NPV of 0.76. This comprehensive evaluation further supports the ALNE model's effectiveness in discerning PBC from AIH.

To assess the performance of individual cohorts and examine how performance varies with cohort size and data diversity, we conducted a series of experiments, each repeated five times. Specifically, we evaluated the local performance of the ALNE model across different cohorts. For the Hannover Medical School cohort with n = 141, the AUROC was 0.6305 (±0.1172). When training solely on the Fondazione IRCCS San Gerardo dei Tintori Monza cohort with n = 138, the AUROC improved to 0.7505 (±0.1108). Additionally, combining three smaller cohorts – Ospedale Milano San Giuseppe, Ospedale di Novara, and Policlinico di Palermo – with n = 89 yielded an AUROC of 0.7086 (±0.0447). These experiments provide insights into the impact of cohort size and data diversity on model performance. The above results and experiments show the individual cohort experiments provide insights into the impact of cohort size and data diversity on the model performance.

To summarize, our exploration of the ALNE model's predictive capabilities in diagnosing PBC and AIH has yielded highly promising outcomes. Both internal and external validations have consistently demonstrated the model's remarkable accuracy in distinguishing between these two liver diseases.

Attention heatmap analysis in AIH and PBC diagnosis using the ALNE model

To better understand how the model achieves its predictions, we generated attention heatmaps, focusing in particular on outliers, that is those cases showing the highest and lowest scores.

In WSIs which were correctly classified as PBC and AIH, we identified highly predictive regions that the model associated with each disease. For instance, Fig. 2A illustrates a portal tract from a case with AIH, characterized by moderate chronic inflammation and interface activity. Similarly, Fig. 2A shows a case with PBC correctly predicted by the model, marked by the absence of lobular inflammation and the presence of a terminal hepatic vein and mild sinusoidal dilatation. These examples highlight that the model tends to focus more on areas with increased inflammation, associating such patterns predominantly with AIH.

The top misclassified cases are reported in Fig. 2B, where two main patterns were identified. First, the ALNE model at times misinterprets marked inflammation in cases with PBC as indicative of AIH. In the validation set, only four cases with AIH with only mild inflammation were misclassified as PBC. While the model accurately detects inflammation, this alone does not always lead to correct disease classification. The attention scores indicate that the model assessed these inflamed areas as diagnostically relevant, yet they may not always provide enough discriminatory power for accurate diagnosis. Nonetheless, these attention scores are valuable for pathologists, as they highlight critical regions within biopsies, potentially reducing both interobserver and intra-observer variability.

In summary, the ALNE model, without reliance on manual annotation, has effectively learned to distinguish between AIH and PBC in liver tissue.

AI and pathologist concordance in distinguishing AIH from PBC

In evaluating the effectiveness of our AI model against general pathologists, the AI predictions demonstrated equivalence or potential superiority to the interpretations made by human readers. Specifically, the average agreement among pathologist pairs was measured with Cohen's kappa as -0.02462, indicating no agreement. Fleiss's kappa value for overall agreement among pathologists was 0.09, reflecting only slight agreement. This is compared to the average agreement between individual pathologists and the ALNE model, which yielded a Cohen's kappa of 0.28, signifying fair agreement (Fig. 3).

These results lead to an important conclusion: generalist pathologists can find themselves at odds when making differential diagnoses between PBC and AIH. This inconsistency reiterates the potential of the ALNE model as a more supportive tool in such diagnostic scenarios.

Clinical characterization of misclassified cases of PBC and AIH in AI diagnosis

To enhance our understanding of the model's predictions and limitations, we conducted a thorough analysis of the misclassified cases, focusing on their clinical characteristics to improve interpretability. Given that only five out of 62 cases with AIH (8%) were misclassified compared to 16 out of 30 PBC cases (53%), our attention was primarily directed toward the latter group.

Initially, we examined whether clinical variables differed between correctly classified and misclassified groups. Although no statistically significant differences were found (Table S1), we observed suggestive trends of higher age at diagnosis and higher aspartate aminotransferase (AST) values in the misclassified PBC group. Multivariate analysis was then conducted using the Rulex LLM, which generated ten explainable if-then rules (Table S2). The Rulex method does not evaluate AST levels solely in correlation with the target variable; it considers interactions between AST and other input variables. Consequently, while the *p* value for AST alone may not appear significant in univariate analysis, its contribution can become significant when combined with other variables in the model.

The most prominent rule for misclassifying PBC patients identified those with alkaline phosphatase (ALP) levels between 1 × ULN and 2 × ULN, combined with AST levels above 1 × ULN, as more likely to be misclassified as having AIH. The second key rule indicated that PBC patients with gamma-glutamyl transferase (GGT) values higher 5 × ULN and AST levels above 1 × ULN were also prone to misclassification as AIH.

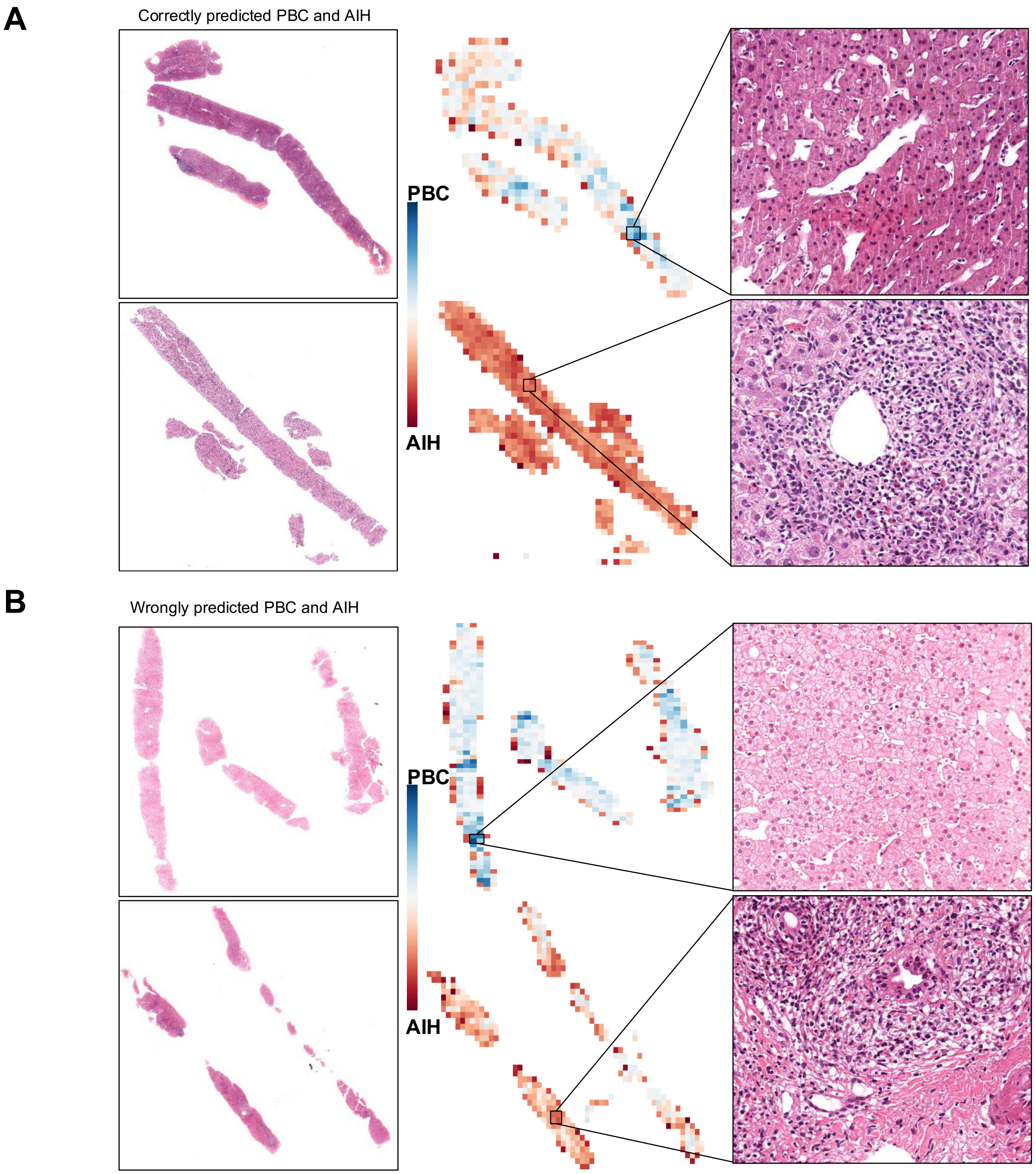

A
Correctly predicted PBC and AIH
PBC
AIH
B
Wrongly predicted PBC and AIH
PBC
AIH

Fig. 2. Visualization of the attention heatmaps for the validation cohort. (A) Correctly predicted instances. Above: terminal hepatic vein with adjacent mild sinusoidal dilatation and no inflammation, indicating PBC. Below: portal tract exhibiting moderate chronic inflammation and interface activity, correctly classified as AIH. (B) Incorrectly predicted instances. Above: parenchymal area without necroinflammatory foci or confluent necrosis, misclassified as PBC. Below: moderate chronic inflammation in an enlarged portal tract with lymphocytic cholangitis and degenerative changes of the bile duct, misclassified as AIH. AIH, autoimmune hepatitis; PBC, primary biliary cholangitis.

Further analysis of clinical records focusing on patients with abnormal AST values (outliers) revealed three interesting cases (Fig. S4). The first case showed pruritus and systemic sclerosis, fluctuating AST and alanine aminotransferase (ALT) levels, positive antinuclear antibodies (ANAs) with a centromeric pattern, and normalized liver enzymes after UDCA therapy. The second case exhibited pruritus, positive ANA with a nuclear rim pattern, elevated immunoglobulin M (IgM), normal immunoglobulin G (IgG), and partial response to UDCA and bezafibrate. The third case, also AMA-negative and gp210 (nucleoporin 210) positive, showed an incomplete response to UDCA, improving after adding obeticholic acid.

In conclusion, the use of Rulex ML software revealed that PBC patients misclassified as AIH by the ALNE model, which

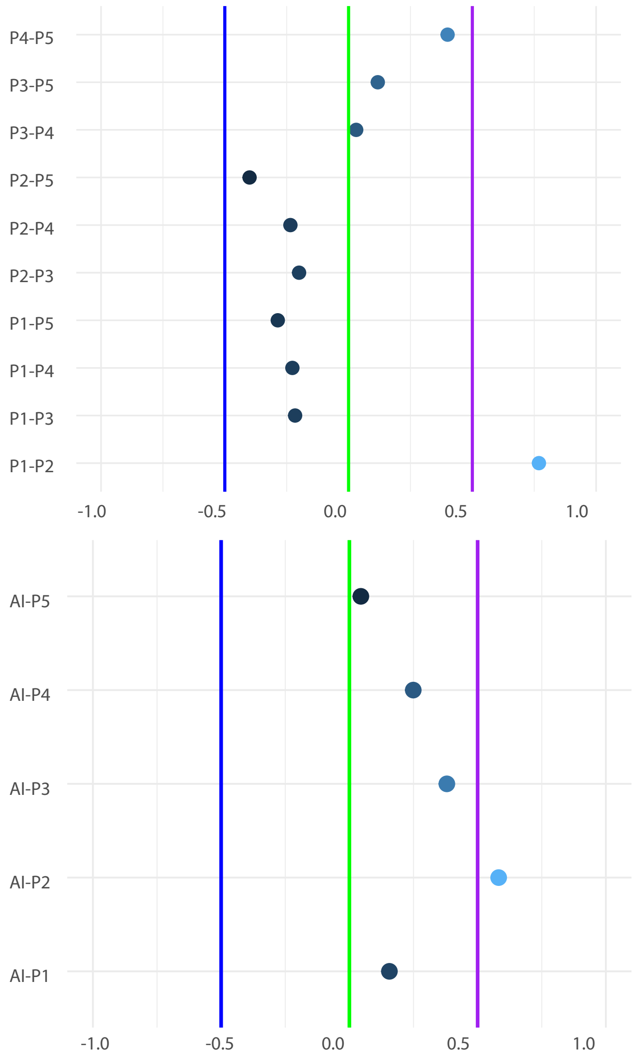


Fig. 3. Evaluating pathologist interobserver variability and AI conformity. The paired agreement among couples of pathologists (at the top) and between each pathologist and the AI model (at the bottom) in classifying AIH and PBC is represented by a variation of the Cohen's kappa index. The Cohen's kappa index is a metric which runs between -1 and 1 and takes into account agreement by chance. The subplot at the top shows the agreement between each pair of pathologists, whereas the agreement between the AI model and each pathologist is shown in the subplot at the bottom. The analysis was performed on a random subset of 19 cases from the validation cohort. For evaluation purposes, the pathologists assessed each case using the H&E slides only. AI, artificial intelligence; AIH, autoimmune hepatitis; PBC, primary biliary cholangitis.

relied only on image data, exhibited specific biochemical patterns in liver enzyme levels.

Discussion

Our study presents the ALNE model, a DL-based tool able to discriminate between AIH and PBC only on the grounds of digital pathology images derived from H&E-stained liver biopsy slides without annotations. Our work used attention heatmaps and the use of explainable AI software to characterize model predictions and open the black box. Our ALNE model demonstrated robust performance even with the diversity in scanning and digitizing techniques, as well as slide staining used across different centers. This adaptability supports the model's applicability in real-world settings, where such variability is common. The ALNE model achieved good performance in external validation cohorts, recognizing areas of inflammation within the biopsy in a self-supervised manner.

The ALNE model achieved excellent performance in discriminating AIH from PBC without the need for manual annotation, which is error-prone and time-consuming. Although studies using AI on manually annotated slides have been published, we should stress that the field of digital and computational pathology is moving toward unsupervised or semi-supervised techniques like the one in our experimental approach.[8,20–23] One of the most pressing reasons behind this change is the progressive shortage of pathologists worldwide in contrast with the rapid accumulation of WSI data and the increasing daily workload of pathologists.[24] The ALNE model, with its visualization techniques, offers a valuable tool in this regard. The rapid scanning and model inference capabilities, combined with attention heatmaps and predictive tiles, could significantly aid in the diagnostic process even in rare, complex, and heterogeneous diseases. Tools such as ALNE could potentially complement the capabilities of pathologists, by viewing and assessing images and also automating tedious tasks.

At the core of the ALNE model resides a cutting-edge technique: transformers. Most studies published between 2018 and 2020 used CNNs as their DL backbone.[25] More recently, transformers have started to replace CNNs.[26] This newer class of neural networks yields a higher accuracy for image classification,[27] and is more robust and explainable.[28] We should also underscore that the model has reached these goals after training with hundreds of cases, whereas other published models in other fields used samples in the order of thousands and even more.[6] Overall, these arguments reinforce the idea that the ALNE model represents an extremely novel and robust AI tool in the field of digital pathology, bringing unprecedented innovation in the area of rare liver diseases.

In addition, our work did not focus only on accuracy, but several approaches were used to understand the process behind prediction and features associated with right and wrong classifications, in line with the call for explainability of AI tools in medicine.[29] Model explainability is crucial to apply DL methods in clinical medicine.[29,30] Some algorithms, such as those using neural networks, are particularly well suited for the analysis of unstructured complex data, for example images or text.[31,32] Yet, being inherently black-box models, they are often difficult to understand even for domain experts. The use of techniques that enhance the interpretability of the model is one strategy to build trust among end-users.[33]

For models using images, the generation of attention heatmaps is one example; they are used to understand how these models focus on different parts of input data (the image) when making predictions or generating outputs.[34] Attention heatmaps were generated and showed relevant areas associated with predictions of AIH or PBC. Heatmaps identify whether the features employed by neural networks are consistent with medical insight of domain experts, both qualitatively and quantitatively. The methodology of our pipeline, based on an end-to-end approach, did not permit us to explicitly quantify inflammation or other pathological elements. Instead, they indicated most indicative regions of AIH or PBC, avoiding the biases introduced by human annotation.

In addition, the use of rule-based models, such as Rulex,[19,35,36] is also gaining traction in medicine thanks to their interpretability. There is increasing interest for blended approaches that utilize black-box models with high levels of accuracy together with other ML models such as Rulex to generate a downstream characterization of the features highlighted by the black-box model.[19] The use of Rulex software did complement the ALNE model in our experimental approach and allowed us to integrate the prediction score generated by the model with clinical variables.. It is important to note that the ALNE model was fully unaware of any clinical variables and achieved its prediction only on the grounds of the information embedded within images. This reinforces the concept that we achieved a fully unbiased classification approach. Our novel approach has allowed us to further understand which clinical scenarios were more challenging for the model, and to plan for future improvements.

The ALNE model has the potential to support general pathologists in cases of PBC and AIH by highlighting areas of inflammation that need further attention. AI-based tools are capable of bringing quantitative assessment of elementary lesions, improving the accuracy of differential diagnosis, and promoting precision medicine principles.[32] Liver biopsy is increasingly less used for the diagnosis of PBC;[9] this trend will reduce the number of cases that are seen by liver pathologists, with potential de-skilling. Instead, AI tools can be trained with a large number of historical slides from digital archives, representing a powerful assistant for future pathologists.[37] In the field of metabolic-dysfunction associated liver disease, there is initial evidence that AI-based models can provide reproducible evaluation of hepatocyte ballooning and lobular inflammation, assisting liver pathologists and standardizing slide scoring in clinical trials.[38] Our work is the first to show the feasibility and accuracy of such approaches in rare liver diseases.

Limitations of the study included the lack of a centralized histological review of digital slides and the absence of control groups reflecting our focus on longitudinal diagnoses rather than purely pathological interpretation. Furthermore, other drawbacks are the unbalanced number of cases between AIH and PBC, the limited sample size of the validation cohort, the presence of some missing data within some clinical tables, and the absence of PBC–AIH mixed phenotypes. It is important to acknowledge that because autoimmune liver diseases are relatively rare compared to more prevalent conditions like cancer, achieving the large sample sizes typical in cancer research remains difficult. This limitation underscores the unique challenges in studying rare diseases.[39] The predominant geographical origin from Western Europe also presents a limitation. Misclassification occurred more frequently for PBC cases with raised AST levels;these cases highlight the need for a dedicated multicenter effort to address this issue in future studies. For broader clinical implementation, we aim to generate a multi-disease liver pathology atlas requiring public/private partnerships. The scarcity of large datasets in autoimmune liver diseases is a challenge for the field, necessitating collaborative efforts for development. We acknowledge that a multimodal approach, which integrates imaging with clinical data such as age, gender, laboratory parameters, and autoantibodies, is likely to enhance diagnostic accuracy. However, incorporating these varied data types into a cohesive model presents significant technical challenges, particularly in terms of model architecture and data integration techniques.[40] These challenges are not unique to our study but are widely acknowledged in the field of medical AI as areas requiring further research and development. Currently, the field lacks a standardized methodology for effectively combining these disparate types of data. Our future work intends to address this gap by developing and testing multimodal models that can leverage both image data and clinical information.

In conclusion, we have presented ALNE, the first transformer-based DL model able to accurately distinguish two rare diseases of the liver, AIH and PBC, relying only on H&E slides without the support of any human annotation. Our ALNE model demonstrated robust performance even with the diversity in scanning and digitizing techniques, as well as slide colorations used across different centers, and in an external validation cohort. Our work presents significant advancements in the field of liver pathology, leveraging the potential of newer, cutting-edge AI technologies also in the differential diagnosis of rare liver diseases.

Affiliations

[1]Division of Gastroenterology, Center for Autoimmune Liver Diseases, European Reference Network on Hepatological Diseases (ERN RARE-LIVER), Fondazione IRCCS San Gerardo dei Tintori, Monza, Italy; [2]Department of Medicine and Surgery, University of Milano-Bicocca, Monza, Italy; [3]Else Kroener Fresenius Center for Digital Health, Medical Faculty Carl Gustav Carus, Technical University Dresden, Dresden, Germany; [4]Department of Diagnostic and Interventional Radiology, University Hospital RWTH Aachen, Aachen, Germany; [5]Department of Medicine and Surgery, Pathology, Fondazione IRCCS San Gerardo dei Tintori, Università di Milano-Bicocca, Monza, Italy; [6]Rulex Innovation Labs, Rulex Inc., Genoa, Italy; [7]Department of Gastroenterology, Hepatology, Infectious Diseases and Endocrinology, Hannover Medical School, Hannover, Germany; [8]European Reference Network on Hepatological Diseases (ERN RARE-LIVER), Hamburg, Germany; [9]Department of Biomedical Sciences, Humanitas University, Pieve Emanuele, Milan, Italy; [10]Division of Internal Medicine and Hepatology, Department of Gastroenterology, IRCCS Humanitas Research Hospital, Rozzano, Milan, Italy; [11]Department of Pathology, IRCSS Humanitas Research Hospital, Rozzano-Milan, Italy; [12]Gastroenterology Hepatology and Transplantation Unit, ASST Papa Giovanni XXIII, Bergamo, Italy; [13]Department of Translational Medicine, Università del Piemonte Orientale, Division of Internal Medicine, AOU Maggiore della Carità, Novara, Italy; [14]Pathology Institute, PROMISE, University of Palermo, Palermo, Italy; [15]Gastrointestinal and Liver Unit, Department of Health Promotion Sciences, Maternal and Infantile Care, Internal Medicine and Medical Specialties, University of Palermo, Palermo, Italy; [16]Department of Pathology, Hospital “Gravina e Santo Pietro”, Caltagirone, Italy; [17]Department of Medical and Surgical Sciences and Advanced Technologies, “G. F. Ingrassia”, University of Catania, Catania, Italy; [18]Department of Diagnostics and Public Health, Section of Pathology, University of Verona, Verona, Italy; [19]Pathological Unit, Niguarda Hospital, Milan, Italy; [20]Université Paris Est Créteil, INSERM, IMRB, Créteil, France; [21]Assistance Publique-Hôpitaux de Paris, Henri Mondor-Albert Chenevier University Hospital, Department of Pathology, Créteil, France; [22]Inserm, U955, Team 18, Créteil, France; [23]Department of Pathology, Aretaieion Hospital, Medical School, National and Kapodistrian University of Athens, Athens, Greece; [24]Translational and Clinical Research Institute, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, UK; [25]Liver Unit, ASST Grande Ospedale Metropolitano Niguarda, Milan, Italy; [26]Medical Oncology, National Center for Tumor Diseases (NCT), University Hospital Heidelberg, Heidelberg, Germany; [27]Department of Medicine 1, University Hospital and Faculty of Medicine Carl Gustav Carus, Technische Universität Dresden, Dresden, Germany

Abbreviations

AI, artificial intelligence; AIH, autoimmune hepatitis; ALNE, autoimmune liver neural estimator; ALP, alkaline phosphatase; ALT, Alanine aminotransferase; AMA, anti-mitochondrial antibodies; ANAs, antinuclear antibodies; AST, aspartate aminotransferase; attMIL, attention-based multiple instance learning; AUROC, area under the receiver operating characteristic curve; CNNs, convolutional neural networks; CTransPath, hybrid model combining CNN with multi-scale swin transformer architecture; DL, deep learning; FN, false negatives; FP, false positives; GGT, gamma-glutamyl transferase; GP210, glycoprotein 210 antibodies; H&E, hematoxylin and eosin; IgG, Immunoglobulin G; IgM, Immunoglobulin M; LKM, liver/kidney microsomal antibodies; LLM, logic learning machine; NPV, negative predictive value; PBC, primary biliary cholangitis; PPV, positive predictive value; SLA/LP, soluble liver antigen/liver pancreas antibodies; SMA, smooth muscle antibodies; SP100, nuclear antigen antibodies; SSL, self-supervised learning; TN, true negatives; TP, true positives; UDCA, ursodeoxycholic acid; ULN, upper limit of normal; WSIs, whole-slide images.

Financial support

The authors thank AMAF Monza ONLUS, AIRCSS, and IPSEN for the unrestricted research grants. This research was partially supported by the grants: Italian MUR Dipartimenti di Eccellenza 2023–2027 (l. 232/2016, art. 1, commi 314–337), Italian MUR PNRR PE06 “HEAL ITALIA – Health Extended ALliance for Innovative Therapies, Advanced Lab-research and Integrated Approaches of Precision Medicine” - Spoke 4 - Precision Diagnostics, Italian MUR PRIN 2022 PNRR P2022H7JYZ “Non-invasive biological and molecular characterization of autoimmune liver diseases and variant syndromes’’.

Conflicts of interest

AG declares consulting services for Ipsen and CAMP4 Therapeutics, and speaker fees from Advanz Pharma. JNK declares consulting services for Owkin, France; DoMore Diagnostics, Norway, Panakeia, UK and Histofy, UK; furthermore, he holds shares in StratifAI GmbH and has received honoraria for lectures by AstraZeneca, Bayer, Eisai, MSD, BMS, Roche, Pfizer, and Fresenius. AL declares consulting fees from Advanz Pharma, GSK, AlfaSigma, Takeda, Ipsen, and Albireo Pharma, and speaker fees from Gilead, Abbvie, MSD, Advanz Pharma, AlfaSigma, GSK, and Incyte. AL declares consulting fees from Advanz Pharma, GSK, AlfaSigma, Takeda, Ipsen, and Albireo Pharma, and speaker fees from Gilead, Abbvie, MSD, Advanz Pharma, AlfaSigma, GSK, and Incyte. MC declares consulting services for Advanz Pharma, Cymabay, GSK, Falk, Ipsen, Albireo, Mirum Pharma, Perspectum, Echosens, Gentic s.p.a. DV works for Rulex, MM is the CEO of Rulex.

Please refer to the accompanying ICMJE disclosure forms for further details.

Authors’ contributions

Study concept and design: AG, OLS, DV, PI, MC, JNK. Acquisition of data: all authors. Analysis and interpretation of data: all authors. Drafting of the manuscript: AG, LS, GC, DV, ZIC, MC, JNK. Critical revision of the manuscript for important intellectual content: all authors. Machine learning analysis: AG, OLS, DV. Statistical analysis: AG. Obtained funding: FP, PI, MC, JNK. Study supervision: DT, VL, FP, NZ, PI, MC, JNK.

Data availability statement

Restrictions apply on the availability of the data used in this study, as they were accessed under specific permissions and are not publicly available. Requests for academic use of both raw and processed data should be directed to the corresponding author(s). These requests will then be evaluated on a case-by-case basis by the overseeing consortium. Restrictions apply to the availability of the data, which were used with permission for the current study and cannot be publicly available. Any request for academic use of raw and processed data should be addressed to the corresponding author and will be evaluated by the consortium.

Code availability

All code and scripts to reproduce the experiments of this paper are available at https://github.com/KatherLab/marugoto. The Marugoto is the open-source toolbox to build deep learning workflow. The ALNE model’s training parameters are specified as follows: the model undergoes training for 256 epochs and early stopping (model does not improve for 10 epochs) with a fit_one_cycle learning rate set to 1e-4. During the training process, a MIL bag size of 512 is utilized, while both the training consists of 64 samples and validation with one batch each. These parameters are carefully chosen to optimize the model’s performance and ensure effective learning from the dataset.

Acknowledgements

The authors thank Mariangela Rizzo, Miki Scaravaglio, Matteo Jiang, and Alberto Marini for their support in the creation of the digital archive of AIH and PBC cases at Fondazione IRCCS San Gerardo dei Tintori, Monza, Italy.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used ChatGPT in order to ensure adherence to English standards and the guidelines of the journal. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jhepr.2024.101198.

References

Author names in bold designate shared co-first authorship

[1] Boberg KM, Chapman RW, Hirschfield GM, et al. Overlap syndromes: the International Autoimmune Hepatitis Group (IAIHG) position statement on a controversial issue. J Hepatol 2011;54:374–385.

[2] Verdonk RC, Lozano MF, van den Berg AP, et al. Bile ductal injury and ductular reaction are frequent phenomena with different significance in autoimmune hepatitis. Liver Int 2016;36:1362–1369.
[3] Nakanuma Y, Zen Y, Harada K, et al. Application of a new histological staging and grading system for primary biliary cirrhosis to liver biopsy specimens: interobserver agreement. Pathol Int 2010;60:167–174.
[4] Zen Y, Harada K, Sasaki M, et al. Are bile duct lesions of primary biliary cirrhosis distinguishable from those of autoimmune hepatitis and chronic viral hepatitis? Interobserver histological agreement on trimmed bile ducts. J Gastroenterol 2005;40:164–170.
[5] Nam D, Chapiro J, Paradis V, et al. Artificial intelligence in liver diseases: improving diagnostics, prognostics and response prediction. JHEP Rep 2022;4:100443.
[6] Kather JN, Pearson AT, Halama N, et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. Nat Med 2019;25:1054–1056.
[7] **Zeng Q**, **Klein C**, Caruso S, et al. Artificial intelligence predicts immune and inflammatory gene signatures directly from hepatocellular carcinoma histology. J Hepatol 2022;77:116–127.
[8] Calderaro J, Ghaffari Laleh N, Zeng Q, et al. Deep learning-based phenotyping reclassifies combined hepatocellular-cholangiocarcinoma. Nat Commun 2023;14:1–10.
[9] European Association for the Study of the Liver. EASL Clinical Practice Guidelines: the diagnosis and management of patients with primary biliary cholangitis. J Hepatol 2017;145:167–172.
[10] European Association for the Study of the Liver. EASL clinical practice guidelines: autoimmune hepatitis. J Hepatol 2015;63:971–1004.
[11] Saldanha OL, Loeffler CML, Niehues JM, et al. Self-supervised attention-based deep learning for pan-cancer mutation prediction from histopathology. NPJ Precis Oncol 2023;7:35.
[12] Niehues JM, Quirke P, West NP, et al. Generalizable biomarker prediction from cancer pathology slides with self-supervised deep learning: a retrospective multi-centric study. Cell Rep Med 2023;4:100980.
[13] **Seraphin TP**, **Luedde M**, **Roderburg C**, et al. Prediction of heart transplant rejection from routine pathology slides with self-supervised deep learning. Eur Heart J Digit Health 2023;4:265–274.
[14] Wang X, Yang S, Zhang J, et al. Transformer-based unsupervised contrastive learning for histopathological image classification. Med Image Anal 2022;81:102559.
[15] Hsu LM, Field R. Interrater agreement measures: comments on Kappan, Cohen's Kappa, Scott's π, and Aickin's α. Underst Stat 2003;2:205–219.
[16] Fleiss JL, Levin BA, Paik MC. Statistical methods for rates and proportions, 2003. https://worldcat.org/title/85820133; 2003.
[17] Muselli M. Switching neural networks: a new connectionist model for classification. Proceedings of the 16th Italian conference on neural networks. Berlin, Heidelberg: Springer; 2005. p. 23–30.
[18] **Ferrari E**, **Verda D**, **Pinna N**, et al. Optimizing water distribution through explainable AI and rule-based control. Computers 2023;12:123. https://doi.org/10.3390/computers12060123.
[19] **Gerussi A**, **Verda D**, Bernasconi DP, et al. Machine learning in primary biliary cholangitis: a novel approach for risk stratification. Liver Int 2021:1–13.
[20] Yu G, Sun K, Xu C, et al. Accurate recognition of colorectal cancer with semi-supervised deep learning on pathological images. Nat Commun 2021;12:6311.
[21] Tiu E, Talius E, Patel P, et al. Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. Nat Biomed Eng 2022;6:1399–1406.
[22] Peiris H, Hayat M, Chen Z, et al. Uncertainty-guided dual-views for semi-supervised volumetric medical image segmentation. Nat Mach Intell 2023;5:724–738.
[23] Zhou Y, Chia MA, Wagner SK, et al. A foundation model for generalizable disease detection from retinal images. Nature 2023;622:156–163.
[24] Metter DM, Colgan TJ, Leung ST, et al. Trends in the US and Canadian pathologist workforces from 2007 to 2017. JAMA Netw Open 2019;2:e194337.
[25] Ghaffari Laleh N, Muti HS, et al. Benchmarking weakly-supervised deep learning pipelines for whole slide classification in computational pathology. Med Image Anal 2022;79:102474.
[26] Ghaffari Laleh N, Truhn D, Veldhuizen GP, et al. Adversarial attacks and adversarial robustness in computational pathology. Nat Commun 2022;13:5711.
[27] Liu Z, Lin Y, Cao Y, et al. Swin transformer: hierarchical vision transformer using shifted windows. Proc IEEE/CVF Int Conf Comp Vis 2021:10012–10022.
[28] **Xu H**, **Usuyama N**, Bagga J, et al. A whole-slide foundation model for digital pathology from real-world data. Nature 2024;630:181–188.
[29] Gunning D, Stefik M, Choi J, et al. XAI—explainable artificial intelligence. Sci Robot 2019;4:eaay7120.
[30] Price WN. Big data and black-box medical algorithms. Sci Transl Med 2018;10:eaao5333.
[31] Lecun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–444.
[32] van der Laak J, Litjens G, Ciompi F. Deep learning in histopathology: the path to the clinic. Nat Med 2021;27:775–784.
[33] Salahuddin Z, Woodruff HC, Chatterjee A, et al. Transparency of deep neural networks for medical image analysis: a review of interpretability methods. Comput Biol Med 2022;140:105111.
[34] Preechakul K, Sriswasdi S, Kijsirikul B, et al. Improved image classification explainability with high-accuracy heatmaps. IScience 2022;25:103933.
[35] Gerussi A, Verda D, Cappadona C, et al. LLM-PBC: logic learning machine-based explainable rules accurately stratify the genetic risk of primary biliary cholangitis. J Pers Med 2022;12:1587.
[36] Skotko BG, Macklin EA, Muselli M, et al. A predictive model for obstructive sleep apnea and Down syndrome. Am J Med Genet A 2017;173:889–896.
[37] McGenity C, Randell R, Bellamy C, et al. Survey of liver pathologists to assess attitudes towards digital pathology and artificial intelligence. J Clin Pathol 2023;77:27–33.
[38] Brunt EM, Clouston AD, Goodman Z, et al. Complexity of ballooned hepatocyte feature recognition: defining a training atlas for artificial intelligence-based imaging in NAFLD. J Hepatol 2022:1–12.
[39] **Banerjee J**, **Taroni JN**, Allaway RJ, et al. Machine learning in rare disease. Nat Methods 2023;20:803–814.
[40] Acosta JN, Falcone GJ, Rajpurkar P, et al. Multimodal biomedical AI. Nat Med 2022;28:1773–1784.

Keywords: Computational pathology; Digital pathology; Liver; Rare liver diseases; Autoimmunity; Artificial intelligence.
Received 26 April 2024; received in revised form 14 August 2024; accepted 16 August 2024; Available online 31 August 2024