Journal of Pathology Informatics 23 (2026) 100712

Contents lists available at ScienceDirect

Journal of Pathology Informatics

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

# Weakly supervised deep learning distinguishes alcohol-associated from metabolic dysfunction-associated steatohepatitis on H&E whole-slide images

Chady Meroueh [a,b,*], Samar H. Ibrahim [c,d], Hamid Tizhoosh [b], Yung-Kyun Noh [e], Iljung Kim [e], Peter Lucas [a], Vijay H. Shah [d]

[a] Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN, USA
[b] Department of Artificial Intelligence and Informatics, Mayo Clinic, Rochester, MN, USA
[c] Division of Pediatric Gastroenterology and Hepatology, Mayo Clinic, Rochester, MN, USA
[d] Division of Gastroenterology and Hepatology, Mayo Clinic, Rochester, MN, USA
[e] Department of Computer Science, Hanyang University, Seoul, Republic of Korea

ARTICLE INFO

*Keywords:*
Steatohepatitis
MASLD
ALD
Multiple instance learning
Gated attention
Weak supervision

ABSTRACT

Distinguishing alcohol-associated steatohepatitis (ASH) from metabolic dysfunction-associated steatohepatitis (MASH) is challenging given the absence of pathognomonic differentiators. We developed and evaluated a weakly supervised deep learning model, as a single-institution proof-of-concept study, to test whether routine hematoxylin and eosin (H&E) whole-slide images (WSIs) of liver biopsies contain sufficient morphological information to predict steatohepatitis etiology at the slide level.

A retrospective cohort of 1147 WSIs was assembled (train set: 1007, holdout test set: 140). Models were trained using 5-fold patient-level cross-validation at $\times 20$ and $\times 40$ magnification. Model interpretability was assessed through attention-based clustering with blinded pathologist review of high-attention patches.

The $\times 20$ model achieved a mean area under the receiver operating characteristic of $0.86 \pm 0.01$ on the test set, with balanced accuracy of 0.80. The $\times 40$ model performed comparably. Specificity was high (0.93 at $\times 20$), with an ASH sensitivity of 0.67 at $\times 20$. Fibrosis-stratified showed preserved performance at advanced fibrosis (stage $\geq 3$) with balanced accuracy of 82.5% and ASH sensitivity of 77.4%. Attention-based clustering localized ASH-enriched regions to active injury patterns including Mallory–Denk bodies, neutrophilic inflammation, cholestatic change, and pericellular fibrosis, whereas MASH-enriched regions showed steatosis with lower inflammatory activity.

Weakly supervised deep learning applied to routine liver H&E WSIs can discriminate ASH from MASH with performance preserved at advanced fibrosis. The comparable performance of $\times 20$ and $\times 40$ magnification, and the alignment of model attention with established histological features, support the use of routine morphology as a decision-support input in cases with incomplete or conflicting clinical histories, particularly when the etiological distinction has the greatest implications for management.

*Abbreviations:* ALD, alcohol-associated liver disease; ASH, alcohol-associated steatohepatitis; AUPRC, area under the precision-recall curve; AUROC, area under the receiver operating characteristic; AUDIT-C, Alcohol Use Disorders Identification Test-Consumption; BMI, body mass index; CI, confidence interval; CRN, Clinical Research Network (NASH CRN: non-alcoholic steatohepatitis activity score by the Clinical Research Network, NIH-validated histological scoring system measuring disease activity via NAS [0–8] and fibrosis stage [0–4]); CV, cross-validation; DICOM, Digital Imaging and Communications in Medicine; ECE, expected calibration error; F1, F1 score; H&E, hematoxylin and eosin; HDBSCAN, hierarchical density-based spatial clustering of applications with noise; HDF5, Hierarchical Data Format version 5; IQR, interquartile range; MASLD, metabolic dysfunction-associated steatotic liver disease; MASH, metabolic dysfunction-associated steatohepatitis; MetALD, metabolic dysfunction-associated alcohol-related liver disease; MIL, multiple instance learning; MoCo, Momentum Contrast; NAS, non-alcoholic steatohepatitis activity score; NAFLD, non-alcoholic fatty liver disease; NPV, negative-predictive value; PCA, principal component analysis; PPV, positive-predictive value; ROC, receiver operating characteristic curve; ROI, region of interest; SLD, steatotic liver disease(s); SMD, standardized mean difference; SVS, Aperio whole-slide image file format; UMAP, uniform manifold approximation and projection; WSI, whole-slide image.

* Corresponding author at: Department of Laboratory Medicine and Pathology, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, USA.
*E-mail address:* meroueh.chady@mayo.edu (C. Meroueh).

http://dx.doi.org/10.1016/j.jpi.2026.100712
Received 7 June 2026; Received in revised form 15 July 2026; Accepted 25 August 2026
Available online 29 August 2026
2153-3539/© 2026 The Author(s). Published by Elsevier Inc. on behalf of Association for Pathology Informatics. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Downloaded for Anonymous User (n/a) at Hanyang University from ClinicalKey.com by Elsevier on October 14, 2026. For personal use only. No other uses without permission. Copyright ©2026. Elsevier Inc. All rights reserved.

## Introduction

Steatotic liver diseases (SLDs) have emerged as the most clinically significant and rapidly expanding causes of chronic liver disease worldwide. In the United States, alcohol-associated liver disease (ALD) accounts for approximately 32% of liver transplants, with 67.3% of all liver transplantations due to metabolic dysfunction-associated steatotic liver disease (MASLD) and ALD, underscoring the substantial morbidity associated with steatotic injury.[1–3] SLDs also contribute immensely to the burden of hepatocellular carcinoma (HCC) (35%–55%), liver-related (70%–90%), and all-cause mortality (8%–15%).[1,4,5] MASLD prevalence is estimated to exceed 30% globally, paralleling the global rise in obesity and type 2 diabetes; within MASLD, a substantial subset progresses to steatohepatitis (MASH), the phenotype most closely tied to fibrosis progression and adverse long-term liver outcomes.[6,7] By contrast, ALD affects an estimated 2%–2.5% of the general population, yet, it accounts for a disproportionately higher share of cirrhosis-related mortality and remains a leading indication for liver transplantation.[8,9] In the United States, alcohol use disorder prevalence increased by approximately 50% over the past decade in national survey analyses.[10] Due to the notable increase in both metabolic syndrome and alcohol use, SLD have evolved to include a mixed-etiology category, metabolic dysfunction-associated alcohol-related liver disease (MetALD). MetALD has been introduced for patients meeting MASLD criteria with increased alcohol intake, reflecting a growing clinical reality that many patients do not fit into “pure” metabolic or alcohol-associated categories.[11,12]

For clinicians, etiological attribution in steatohepatitis dictates management priorities, counseling, and eligibility for therapies and clinical trials. Sustained abstinence is the current cornerstone intervention in ASH, whereas diet and exercise are foundational in MASH, in combination with the recently approved medications semaglutide and resmetirom.[13,14]

In general, steatohepatitis is a pattern diagnosis, not an etiological verdict, with histopathological elements such as macrovesicular steatosis, hepatocellular ballooning, lobular inflammation, present in all SLDs.[15] Histopathologically, ASH is more often characterized by a neutrophil-rich lobular infiltrate, satellitosis around ballooned hepatocytes, more conspicuous Mallory–Denk bodies, and accompanying lesions such as canalicular cholestasis, ductular reactions, phlebosclerosis, or sclerosing hyaline necrosis,[16,17] whereas MASH exhibits a mixed mononuclear lobular infiltrate with progressive zone 3 perisinusoidal/pericellular fibrosis in a “chicken-wire” configuration.[18] Even with these histopathological findings, there are no pathognomonic features that can accurately differentiate between ASH and MASH. This limitation is accentuated by under-reported alcohol consumption, and the ubiquitous nature of cardiometabolic risks.

Machine learning-based histopathological analysis has proven effective across a range of diagnostic applications, such as tumor detection and grading, identification of metastasis, and artificial intelligence (AI)-derived NASH CRN scores in liver biopsy assessment.[19–21] Weakly supervised approaches are instrumental in that they learn directly from slide-level labels, obviating the need for time-consuming human annotation and its potential to introduce bias.[22] It is well recognized that routine histological slides contain information beyond what pathologists can explicitly score, a concept exemplified by the ability to predict molecular alterations and biomarkers based on WSI.[23,24] In liver disease, AI-assisted systems have already entered clinical use and large collaborative programs for MASH drug development and are beginning to link digital morphology with spatially resolved molecular data. Examples include fibrosis quantification for assessment of treatment response,[25] and automated quantification of steatosis, ballooning, inflammation, and fibrosis,[26–28] as well as AI-based enrollment criteria and endpoint assessment.[29] In summary, AI application in digital pathology extends beyond quantifying known markers of liver injury to uncover latent signals embedded within routine WSI.

Hence, the aim of the current manuscript is to determine whether steatohepatitic liver biopsies alone contain sufficient information to support etiological inference. Specifically, can ASH and MASH be distinguished based on H&E WSI alone? To test this hypothesis, we assembled a cohort of steatohepatitis WSI comprising 1007 slides for model development and 140 slides reserved as an independent holdout cohort. Cases with other liver diseases were excluded, and patient-level partitioning was enforced to prevent information leakage.

This study is designed as a proof of concept rather than a substitute for clinical diagnosis. If etiological signals can be detected reproducibly in routine liver histology, such models may help in research settings by triaging cases for targeted clinicopathological review, and in the clinical setting by highlighting diagnostically discordant cases. Beyond classification performance, a longer-term objective is to identify quantitative histological attributes that drive model predictions, and to translate these features into biologically grounded descriptors of SLDs.

## Materials and methods

### *Study design and task definition*

This retrospective, single-institution study evaluated whether hematoxylin and eosin (H&E) whole-slide images (WSIs) of liver biopsies with histological steatohepatitis contained sufficient morphological information to infer the etiology of steatohepatitis using weakly supervised machine learning with slide-level binary classification between alcohol-associated steatohepatitis (ASH; class 1) and metabolic dysfunction-associated steatohepatitis (MASH; class 0). This retrospective study was approved by the Institutional Review Board (IRB #20-253), including only patients with research authorization. Fig. 1 represents an illustration of our workflow.

### *Cohort identification and clinical labeling*

Using an institutional pathology database search, we selected patients who underwent liver biopsy or resection with a histological diagnosis of steatohepatitis (Fig. 1), and excluded patients with other or additional liver diseases, including viral hepatitis, autoimmune hepatitis, hereditary hemochromatosis, Wilson's disease, drug-induced liver injury, malignancy, prior systemic chemotherapy, non-native liver, or isolated steatosis. We subsequently performed a manual review of the electronic medical record, and assigned a patient-level etiological reference label based on the hepatologist's documented diagnosis. For example, hepatologists classified ALD based on chart-documented excessive alcohol intake elicited during routine clinical assessment, including AUDIT-C review, in the absence of metabolic syndrome. Patients who had metabolic syndrome without excessive alcohol intake were classified as MASLD by the treating hepatologist. Patients who were identified as having MASLD with increased alcohol intake were classified as MetALD and were excluded from the current cohort ($n = 47$; Fig. 1B). Within this manuscript, we used the histopathology terminology: ASH for alcohol-associated steatohepatitis and MASH for metabolic dysfunction-associated steatohepatitis. The clinicopathological characteristics of the patients are summarized in Table 1. The NAS CRN scores and fibrosis stage were manually extracted from the pathology reports.

### *Development and holdout cohorts*

The final dataset comprised 1147 WSIs from 1095 unique patients (October 13, 1999–December 31, 2024), divided into a training cohort (1007 WSIs, 955 unique patients, 725 MASH, and 282 ASH cases), and holdout testing cohort (140 WSIs, 140 unique patients, 97 MASH, and 43 ASH cases), ensuring no patient existed in both the training and testing datasets. The training cohort was divided into five folds ($k = 5$); The model was trained using 5-fold cross-validation, where in each of the five iterations, four folds (80% of data) were used for training and one fold (20% of data) was used for validation, ensuring each fold contained the same proportion of ASH and MASH classes, with assignment at the patient level to prevent data leakage (all slides from a given patient fell within the same fold). The independent holdout cohort was strictly separated at the patient level, with one slide per patient and no patient overlap with the



Downloaded for Anonymous User (n/a) at Hanyang University from ClinicalKey.com by Elsevier on October 14, 2026. For personal use only. No other uses without permission. Copyright ©2026. Elsevier Inc. All rights reserved.

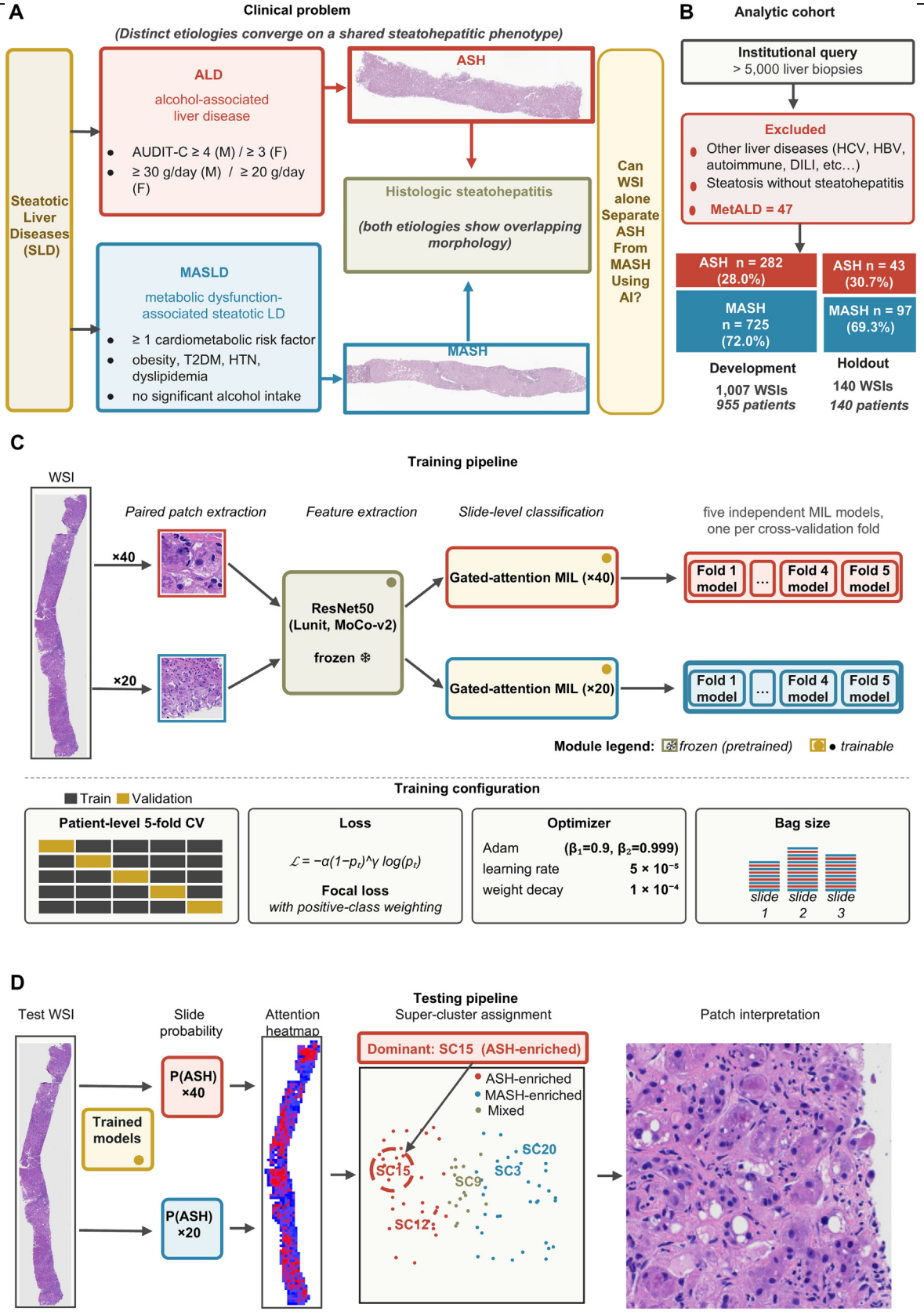



**Fig. 1.** Study design and analytic workflow. ASH and MASH are clinically distinguished by exposure history but converge on a histopathological phenotype that pathologists cannot reliably separate by morphology alone. We tested whether routine H&E whole-slide images contained sufficient information to discriminate the two etiologies at the slide level using weakly supervised deep learning. (A) Clinical problem. Alcohol-associated liver disease (ALD) and metabolic dysfunction-associated steatotic liver disease (MASLD) are defined clinically by increased alcohol intake for ALD, and by one or more cardiometabolic risk factors for MASLD. The steatohepatitic forms of each disease (ASH and MASH) share macrovesicular steatosis, hepatocellular ballooning, and lobular inflammation, with no pathognomonic histological features. (B) Analytic cohort. A



Downloaded for Anonymous User (n/a) at Hanyang University from ClinicalKey.com by Elsevier on October 14, 2026. For personal use only. No other uses without permission. Copyright ©2026. Elsevier Inc. All rights reserved.

development cohort. This cohort was not used for model training, threshold selection, or model selection. The H&E-stained liver biopsy slides were digitized using multiple scanners across clinical sites and time periods, including Leica AT Turbo, Pramana HT40, and Leica GT450 platforms. These sites belong to a single institution and follow uniform histological, staining, and scanning protocols; the cohort is therefore single-center despite spanning multiple physical sites. We intentionally preserved scanner heterogeneity to reflect real-world variation, with some slides scanned using separate scanner profiles. Stain normalization and color standardization were not performed beyond model-specific RGB normalization during feature extraction.

*Patch extraction and paired magnification sampling*

WSIs were annotated using QuPath v0.6[30] to delineate analyzable tissue regions, and to select a single section if multiples exist. For each slide, 224 × 224-pixel paired patches were extracted at ×40 and ×20 from the same (x, y) grid coordinates, ensuring exact spatial alignment between magnifications. Tissue-quality filters and patch-count statistics are provided in Supplementary Methods.

*Tile feature extraction using a frozen self-supervised ResNet50 backbone*

Each patch was converted into a fixed-length embedding using a publicly available histopathology-pretrained ResNet50 model provided by Lunit Inc. (Seoul, South Korea),[31] pretrained by self-supervised learning with Momentum Contrast (MoCo v2).[32] The backbone was frozen (used purely as a feature extractor) and the classification layer was deleted; each patch yielded a 2048-dimensional embedding from the global pooling layer, stored per slide in HDF5 files for downstream MIL training and inference. Patches were normalized using color normalization parameters recommended by Lunit Inc. for this specific model and are provided in Supplementary Methods.

*Slide-level classification with gated-attention multiple instance learning (MIL)*

Slide-level classification was performed using gated-attention MIL,[33] which learns a per-patch attention weight via a tanh-sigmoid gating mechanism and produces an attention-weighted slide embedding for a linear classifier. These attention weights reflected each patch's proportional contribution to the prediction and were visualized as heatmaps over the slide to identify morphologically discriminative regions. The complete mathematical formulation, architecture dimensions, and hyperparameters are provided in Supplementary Methods.

*Model training and cross-validation*

We trained gated-attention MIL models for ×20 and ×40 feature sets using PyTorch Lightning version 2.6.0,[34,35] with 5-fold patient-level cross-validation and a batch size of one slide per step, consistent with standard MIL practice. To address the 2.57:1 MASH-to-ASH class imbalance, we used focal loss with positive-class weighting.[36] Full training hyperparameters, including random seeds, software versions, and the compute environment are also reported there (see the Supplementary Methods sections on Training hyperparameters and Compute environment), together with the patch-filtering rules (Patch extraction parameters), providing the complete set of parameters needed to reimplement the pipeline.

For each slide, the MIL model generated a raw score that was converted to a probability via the sigmoid function. A decision threshold was selected from cross-validation out-of-fold predictions; full threshold selection details are provided in Supplementary Methods.

For holdout evaluation, we combined the 5-fold-specific checkpoints of each magnification into a single ensemble by averaging their per-slide ASH probabilities, yielding one probability per holdout slide per magnification. We applied a single decision threshold per magnification, fixed before holdout scoring as the F1-optimal threshold on the pooled cross-validation out-of-fold predictions (×20 = 0.481, ×40 = 0.379), so that no threshold was tuned on the holdout cohort. In 5-fold cross-validation, every development slide is scored once, by the single fold-model that did not use it for training (its out-of-fold prediction). Pooling these across the five folds gives one prediction per development slide, which we use for pooled cross-validation out-of-fold predictions. The F1-optimal threshold was found by sweeping candidate values from 0.01 to 0.99 in 0.01 increments against these pooled predictions, choosing the value that maximized the binary F1 score for ASH. Threshold-dependent metrics, including sensitivity (recall for ASH), specificity (recall for MASH), accuracy, balanced accuracy, positive-predictive value (PPV), negative-predictive value (NPV), and F1 score, and the confusion matrices were computed from this single ensemble at that fixed threshold. We computed the threshold-independent metrics (AUROC and AUPRC) from the per-slide probabilities of the 5-fold-models and report them as the mean ± SD across the 5-fold-models, because these rank-based metrics are threshold-independent and differed from the ensemble value by less than 0.01.

*Bootstrap confidence intervals*

For the independent holdout cohort, 95% CIs for all classification metrics were estimated by stratified bootstrap resampling (B = 2000 replicates), with full resampling details in Supplementary Methods. Calibration was assessed using expected calibration error (ECE)[37]; post hoc temperature scaling was evaluated as a secondary analysis with detailed methods and results in the Supplementary Material.

*Attention-based clustering and pathologist review*

To characterize the histopathological content of regions prioritized by the model, we performed a secondary interpretability analysis on the independent holdout cohort. For each holdout slide, consensus high-attention patches were identified as the top 5% of patches ranked by percentile-mean attention rank across the 5-fold-specific checkpoints; this rank-aggregation approach is robust to fold-specific outliers and is described in detail in Supplementary Methods. Patch embeddings were dimensionality-reduced and clustered to identify fine-grained morphological groups that were then consolidated into 20 super-clusters at ×20 (full clustering parameters in Supplementary Methods). A board-certified pathologist (CM) reviewed 10 representative highest-attention patches per ×20 super-cluster descriptively in a blinded manner, without knowledge of ground-truth class, prediction outcome group, or enrichment assignment; the purpose was descriptive interpretability rather than diagnostic

query of >5000 institutional liver biopsies was filtered to exclude other liver diseases, isolated steatosis, and MetALD cases ($n$ = 47), then split at the patient level into a development cohort (1007 WSIs from 955 patients; 725 MASH, 282 ASH) and an independent holdout cohort (140 WSIs from 140 patients; 97 MASH, 43 ASH). (C) Training pipeline. Paired ×20 and ×40 patches (224 × 224 pixels) were extracted at the same coordinates after manual tissue annotation in QuPath. Each patch was converted to a 2048-dimensional feature vector using a frozen ResNet50 backbone (Lunit, pre-trained with MoCo-v2). Two parallel gated-attention multiple-instance learning (MIL) branches, one per magnification, were trained independently with focal loss (α-weighted, γ = 2), Adam optimizer (learning rate $5 \times 10^{-5}$, weight decay $1 \times 10^{-4}$), and patient-stratified 5-fold cross-validation, yielding 5-fold-specific models per magnification. (D) Inference and interpretability pipeline. Each fold-model produced a slide-level ASH probability and a patch-level attention heatmap. The top 5% highest-attention patches per slide were embedded with UMAP and grouped into super-clusters. A pathologist then reviewed representative patches from each super-cluster, blinded to ground-truth class, prediction outcome, and super-cluster enrichment, to determine whether model attention corresponded to known histopathological differentiators of ASH and MASH.



Downloaded for Anonymous User (n/a) at Hanyang University from ClinicalKey.com by Elsevier on October 14, 2026. For personal use only. No other uses without permission. Copyright ©2026. Elsevier Inc. All rights reserved.

**Table 1**
Baseline clinicopathological characteristics of the development and independent holdout cohorts.

| Characteristic | Development cohort | Holdout cohort | Abs SMD |
|---|---|---|---|
| *Cohort size* | | | |
| Unique patients, *n* | 955 | 140 | |
| Unique biopsies, *n* | 985 | 140 | |
| Slides, *n* | 1007 | 140 | |
| MASH label, *n/N* slides (%) | 725 (72%) | 97 (69.3%) | 0.061 |
| ASH label, *n/N* slides (%) | 282 (28%) | 43 (30.7%) | 0.061 |
| *Demographics* | | | |
| Age at biopsy, years | 54.6 [44.6–63.7]; $n = 983$ | 53.0 [38.6–64.1]; $n = 140$ | 0.222 |
| BMI, kg/m$^2$ | | | |
| MASH | 34.1 [29.6–40.3]; $n = 709$ | 35.5 [30.4–39.7]; $n = 97$ | 0.052 |
| ASH | 27.6 [23.8–32.8]; $n = 275$ | 25.7 [22.5–31.2]; $n = 43$ | 0.213 |
| *Sex, n/N patients (%)* | | | |
| Female | 566 (59.3%) | 80 (57.1%) | 0.043 |
| Male | 389 (40.7%) | 60 (42.9%) | 0.043 |
| *Race, n/N patients (%)* | | | |
| White | 851 (89.1%) | 128 (91.4%) | 0.078 |
| Black | 23 (2.4%) | 2 (1.4%) | 0.071 |
| Asian | 24 (2.5%) | 3 (2.1%) | 0.025 |
| American Indian/Alaska Native | 12 (1.3%) | 3 (2.1%) | 0.069 |
| Native Hawaiian/Pacific Islander | 2 (0.2%) | 1 (0.7%) | 0.075 |
| Other | 11 (1.2%) | 2 (1.4%) | 0.025 |
| Unknown/Declined | 32 (3.4%) | 1 (0.7%) | 0.188 |
| *Ethnicity, n/N patients (%)* | | | |
| Not Hispanic or Latino | 806 (84.4%) | 127 (90.7%) | 0.192 |
| Hispanic or Latino | 83 (8.7%) | 12 (8.6%) | 0.004 |
| Unknown/Declined | 66 (6.9%) | 1 (0.7%) | 0.328 |
| *Pathology (extracted report fields; biopsy-level)* | | | |
| NAS steatosis score | $n = 817$ | $n = 139$ | 0.069 |
| 0 | 61 | 13 | 0.068 |
| 1 | 356 | 66 | 0.079 |
| 2 | 241 | 31 | 0.165 |
| 3 | 159 | 29 | 0.035 |
| Median [IQR] | 1 [1–2] | 1 [1–2] | 0.069 |
| NAS lobular inflammation score | $n = 512$ | $n = 97$ | 0.074 |
| 0 | 53 | 10 | 0.001 |
| 1 | 389 | 78 | 0.108 |
| 2 | 62 | 7 | 0.166 |
| 3 | 8 | 2 | 0.037 |
| Median [IQR] | 1 [1–1] | 1 [1–1] | 0.074 |
| NAS ballooning score | $n = 540$ | $n = 129$ | 0.065 |
| 0 | 46 | 10 | 0.028 |
| 1 | 371 | 95 | 0.109 |
| 2 | 123 | 24 | 0.103 |
| Median [IQR] | 1 [1–1] | 1 [1–1] | 0.065 |
| *Fibrosis stage* | | | |
| MASH | $n = 702$ | $n = 97$ | 0.122 |
| 0 | 90 | 5 | 0.270 |
| 1 | 153 | 23 | 0.046 |
| 2 | 145 | 22 | 0.049 |
| 2.5 | 29 | 7 | 0.134 |
| 3 | 107 | 13 | 0.053 |
| 3.5 | 17 | 7 | 0.225 |
| 4 | 161 | 20 | 0.056 |
| Median [IQR] | 2 [1–3.5] | 2 [1–3.5] | 0.122 |
| ASH | $n = 270$ | $n = 43$ | 0.291 |
| 0 | 10 | 1 | 0.081 |
| 1 | 34 | 4 | 0.106 |
| 2 | 54 | 5 | 0.231 |
| 2.5 | 9 | 2 | 0.067 |
| 3 | 49 | 3 | 0.342 |
| 3.5 | 13 | 9 | 0.496 |
| 4 | 101 | 19 | 0.138 |
| Median [IQR] | 3 [2–4] | 3.5 [2.5–4] | 0.291 |

Abbreviations: ASH, alcohol-associated steatohepatitis; MASH, metabolic dysfunction-associated steatohepatitis; NAS, non-alcoholic steatohepatitis activity score; SMD, standardized mean difference. Values are median [IQR] unless otherwise indicated. Sex, race, and ethnicity were summarized at the patient level. Age at biopsy, BMI, and extracted pathology variables were summarized at the biopsy level using one record per unique pathology specimen within each split. Slide counts are shown separately. Intermediate fibrosis values such as Fibrosis stage 2–3, encoded as 2.5, or stage 3–4 (3.5) were retained as extracted. Missing values indicate that the original pathology report did not explicitly state the feature. Cohort comparability was evaluated using absolute standardized mean differences (SMDs), which quantify the magnitude of differences between groups independent of sample size. In contrast, *p*-values are highly sensitive to cohort size and may indicate statistical significance for clinically negligible differences or fail to detect meaningful differences in smaller samples. SMD thresholds were interpreted as follows: <0.1 negligible, 0.1–0.2 small, 0.2–0.5 moderate, and >0.5 large imbalance.



Downloaded for Anonymous User (n/a) at Hanyang University from ClinicalKey.com by Elsevier on October 14, 2026. For personal use only. No other uses without permission. Copyright ©2026. Elsevier Inc. All rights reserved.

relabeling. Detailed clustering workflow, representative patch inventories, and supplementary galleries are provided in the Supplementary Material.

## Results

### *Gated-attention MIL discriminates MASH from ASH with strong cross-validation performance*

A gated-attention MIL was trained on 1007 steatohepatitis WSIs drawn from a single institution (development cohort cases: 725 MASH, 282 ASH; from 955 unique patients) and then evaluated on an independent 140 WSIs (holdout test cases: 97 MASH, 43 ASH; from 140 unique patients). All models achieved strong cross-validation discrimination, with the ×20 model achieving a mean area under the receiver operating characteristic (AUROC) of 0.89 ± 0.02 (Fig. S1) and a mean area under the precision-recall curve (AUPRC) of 0.82 ± 0.04. The ×40 model achieved a mean AUROC of 0.88 ± 0.03 and a mean AUPRC of 0.80 ± 0.05 (Table 2). Balanced accuracy was 0.81 at ×20 and 0.82 at ×40. MASH specificity was high at ×20 (0.94), whereas the ×40 models traded specificity (0.88) for higher ASH sensitivity (0.76 versus 0.67 at ×20) at its lower operating threshold. At ×20, the dominant error mode (the most frequent pattern of misclassification, identifying where the model systematically failed rather than where it failed sporadically) was under-calling ASH rather than over-calling it, consistent with the approximately 2.5:1 MASH-to-ASH class imbalance. Routine H&E morphology therefore carried sufficient signal for reproducible etiological discrimination at the slide level, with the asymmetry between MASH and ASH recall establishing the principal error pattern that subsequent analyses sought to characterize and explain.

### *Performance is preserved on the independent holdout cohort*

The 5-fold-models from each magnification were combined into a single ensemble by averaging their per-slide ASH probabilities, and metrics were computed from this ensemble at the cross-validation-frozen F1 threshold without retraining or holdout threshold reoptimization. On the independent holdout cohort ($N$ = 140), the AUROC was 0.86 ± 0.01 at ×20 and 0.84 ± 0.02 at ×40, with a balanced accuracy of 0.80 at ×20 and 0.79 at ×40 (Table 2, Fig. 2). Specificity remained high (0.93 at ×20 and 0.89 at ×40), whereas ASH sensitivity was 0.67 at ×20 and 0.70 at ×40. The holdout metrics with full 95% bootstrap confidence intervals are reported in Table 2. Confusion matrices in Fig. 2 were computed from the single ensemble at the cross-validation-frozen F1 threshold and were not reoptimized on the holdout cohort. The minimal deltas between the cross-validation and holdout metrics supported the robustness of the learned morphological signal and argued against overfitting to scanner-specific artifacts, particularly given that the development cohort and holdout cohort were scanned on a mixture of three different platforms (Leica AT Turbo, Pramana HT40, and Leica GT450). The preserved discrimination across scanner platforms, together with the reproduction of the MASH-versus-ASH recall asymmetry seen in cross-validation, indicated that the model's behavior was driven by the learned morphological signal rather than by cohort- or scanner-specific artifacts.

### *Comparable performance at ×20 and ×40 magnification, with ×20 retained as the primary model*

Previous work has demonstrated improvements in performance when comparing multiple magnifications, with ×20 generally showing better results.[38] Similarly, we designed our study with a two-magnification comparison (×20 and ×40) to identify the relative contributions of cytological detail (captured at ×40) and architectural context (captured at ×20), because routine histopathological descriptions of ASH and MASH draw on both scales. Across all threshold-independent metrics, the ×20 models matched or slightly exceeded the ×40 models in both cross-validation and holdout evaluation (Fig. S1 and Fig. S2A for cross-validation; Table 2 and Fig. 2 for the holdout cohort). On the holdout cohort, none of the between-magnification differences reached statistical significance, and a paired bootstrap over the 140 slides placed every metric difference within a 95% interval that included zero, including the ASH sensitivity difference of 2.3 points in favor of ×40 (95% CI -14.6 to +9.5 points; Fig. S2B). We therefore treated the two magnifications as comparable and retained ×20 as the primary model, because the interpretability analyses were built on the ×20 models and ×20 carried nominally higher AUROC, AUPRC, specificity, and F1.

To test whether combining the two magnifications offered an advantage over the single-magnification ×20 model, we evaluated nine ensemble configurations on the pooled cross-validation predictions to select the primary configuration before holdout evaluation (Fig. S2A; full definitions in Supplementary Methods). These included the two single-magnification baselines (×20 only, ×40 only), five probability–space combinations (mean of per-fold predicted probabilities, AUROC-weighted mean, soft-OR using per-fold maxima, soft-AND using per-fold minima, and a temperature-recalibrated ×20), and two hard-label rules combining the per-magnification binary calls (OR and AND) without refitting any threshold on the holdout cohort. The OR rule labeled a slide ASH, if either magnification's ensemble called it ASH, maximizing sensitivity at a small specificity cost; the AND rule required both magnifications to agree, maximizing specificity at a sensitivity cost. Probability–space ensembles all clustered within a narrow band of the ×20-only baseline on AUROC (0.887–0.892) and balanced accuracy (0.810–0.828), with the best-performing configuration on sensitivity (weighted_prob) gaining roughly 7 points of ASH sensitivity (67.4–74.1) at the cost of 3 points of specificity

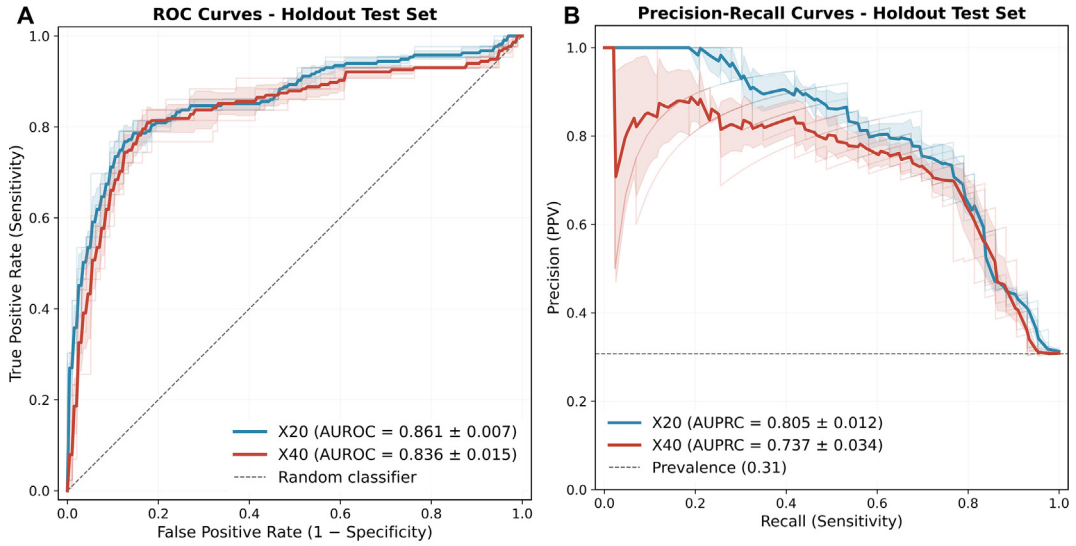
**Table 2**
Comprehensive performance metrics for ×20 and ×40 models in cross-validation and holdout evaluation ($N$ = 140).

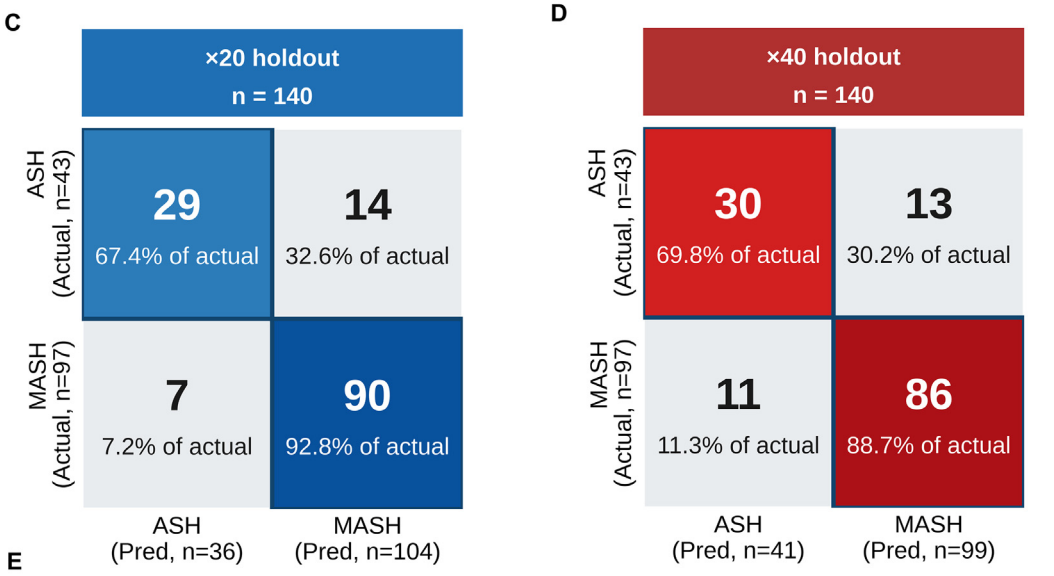
| Metric | ×20 CV | ×20 Holdout | ×40 CV | ×40 Holdout |
|---|---|---|---|---|
| AUROC | 0.888 ± 0.021 | 0.861 ± 0.007 | 0.876 ± 0.030 | 0.836 ± 0.015 |
| AUPRC | 0.817 ± 0.037 | 0.805 ± 0.012 | 0.803 ± 0.046 | 0.737 ± 0.034 |
| F1 score | 0.742 [0.698–0.783] | 0.734 [0.615–0.835] | 0.728 [0.685–0.766] | 0.714 [0.595–0.822] |
| Accuracy | 0.869 [0.848–0.890] | 0.850 [0.786–0.907] | 0.842 [0.818–0.864] | 0.829 [0.764–0.893] |
| Balanced accuracy | 0.809 [0.781–0.838] | 0.801 [0.725–0.874] | 0.816 [0.785–0.842] | 0.792 [0.712–0.869] |
| Sensitivity (ASH recall) | 0.674 [0.618–0.727] | 0.674 [0.531–0.811] | 0.755 [0.702–0.804] | 0.698 [0.553–0.837] |
| Specificity (MASH recall) | 0.945 [0.928–0.961] | 0.928 [0.872–0.977] | 0.876 [0.852–0.900] | 0.887 [0.818–0.949] |
| Precision (PPV) | 0.826 [0.775–0.872] | 0.806 [0.674–0.926] | 0.703 [0.649–0.753] | 0.732 [0.591–0.868] |
| NPV | 0.882 [0.858–0.904] | 0.865 [0.798–0.927] | 0.902 [0.878–0.923] | 0.869 [0.798–0.932] |
| Calibration error | 0.149 ± 0.013 | 0.127 ± 0.009 | 0.150 ± 0.021 | 0.131 ± 0.010 |
| Optimal threshold | 0.481 | 0.481 | 0.379 | 0.379 |
| $N$ (slides) | 1007 | 140 | 1007 | 140 |

Threshold-dependent metrics (F1, accuracy, balanced accuracy, sensitivity, specificity, PPV, and NPV) are computed at the F1-optimal threshold frozen on the cross-validation out-of-fold predictions (×20 = 0.481, ×40 = 0.379); cross-validation values use the pooled out-of-fold predictions (one prediction per slide), and holdout values use a single ensemble (the mean per-slide ASH probability across the 5-fold-models). Threshold-dependent metrics are reported as point estimate [95% bootstrap confidence interval] from 2000 stratified replicates. AUROC, AUPRC, and calibration error are reported as the mean ± SD across the 5-fold-models (threshold-independent or fold-based diagnostics). Positive class: ASH; negative class: MASH.



Downloaded for Anonymous User (n/a) at Hanyang University from ClinicalKey.com by Elsevier on October 14, 2026. For personal use only. No other uses without permission. Copyright ©2026. Elsevier Inc. All rights reserved.




E

| | ×20 holdout | | ×40 holdout | |
|---|---|---|---|---|
| | ASH | MASH | ASH | MASH |
| Sensitivity | 67.4% | 92.8% | 69.8% | 88.7% |
| PPV | 80.6% | 86.5% | 73.2% | 86.9% |
| F1 score | 73.4% | 89.6% | 71.4% | 87.8% |
| **Overall accuracy** | **85.0%** | | **82.9%** | |
| **Balanced accuracy** | **80.1%** | | **79.2%** | |

**Fig. 2.** Holdout performance was preserved across magnifications, with comparable discrimination at ×20 and ×40 and ASH recall the dominant error mode at ×20. The independent holdout cohort ($N$ = 140 WSIs; 97 MASH, 43 ASH) was scored by all 5-fold-specific checkpoints from each magnification without retraining or threshold reoptimization. Performance matched cross-validation, and the error pattern was asymmetric: the model classified MASH with high specificity and under-called ASH, mirroring the 2.5:1 development-cohort class imbalance. (A) ROC curves for the ×20 (blue) and ×40 (red) models, with bold lines showing the mean and shaded bands the ±1 SD envelope across the 5-fold-models. Per-fold spread was narrow at both magnifications. The ×20 model reached a mean holdout AUROC of 0.861 ± 0.007, compared with 0.836 ± 0.015 at ×40. (B) Precision-recall curves for the same comparison. Mean AUPRC was 0.805 ± 0.012 for ×20 and 0.737 ± 0.034 for ×40, with both well above the holdout ASH prevalence of 0.31 (dashed line). The larger AUPRC gap between magnifications, compared with AUROC, indicated that ×20 was more reliable than ×40 in the precision-sensitive regime relevant to minority-class detection. (C, D) Single-ensemble confusion matrices for ×20 (C) and ×40 (D), computed from the mean per-slide ASH probability across the 5-fold-models at the cross-validation-frozen F1 threshold (×20 = 0.481, ×40 = 0.379). The ×20 ensemble classified MASH with 92.8% specificity and ASH with 67.4% sensitivity; the ×40 ensemble reached 88.7% specificity and 69.8% sensitivity. (E) Threshold-dependent metrics by class. Positive-predictive value was 80.6%/73.2% for ASH and 86.5%/86.9% for MASH (×20/×40), and sensitivity differed by roughly 19–25



Downloaded for Anonymous User (n/a) at Hanyang University from ClinicalKey.com by Elsevier on October 14, 2026. For personal use only. No other uses without permission. Copyright ©2026. Elsevier Inc. All rights reserved.

(94.5–91.6); overall accuracy did not improve (86.7 versus 86.9) and balanced accuracy improved by 1.9 points (82.8 versus 80.9), reflecting the sensitivity gain (Fig. S2A). The hard-label OR rule produced a similar but smaller sensitivity gain (sensitivity 70.6, specificity 92.5), and the hard-label AND rule moved in the opposite direction (sensitivity 55.7, specificity 97.4). The single-magnification ×40 model was the weakest configuration on most metrics, including AUROC, AUPRC, F1, overall accuracy, and specificity, although it achieved the highest sensitivity of any configuration (75.5%) at its lower operating threshold.

As a secondary technical analysis, we assessed probability calibration before and after post hoc temperature scaling (Fig. S3). Before scaling, both magnifications were systematically underconfident at intermediate predicted probabilities (approximately 0.25–0.50), with per-fold ECE of 0.127 ± 0.009 (×20) and 0.131 ± 0.010 (×40). After per-fold temperature scaling, ECE dropped by approximately 41% (0.075 ± 0.016 (×20) and 0.076 ± 0.006 (×40)), bringing the calibrated models into the range generally considered acceptable for risk-stratification use. The benefit is interpretive because temperature scaling preserved the ranking of cases, with unchanged AUROC and AUPRC, supporting use of the calibrated probabilities as approximate confidence estimates rather than ranking scores alone. The single-magnification ×20 model therefore captured the dominant share of etiological information available at either magnification, with cross-magnification ensembling offering only a sensitivity-for-specificity trade rather than a true performance gain, and calibration adjustments improving the interpretability of predicted probabilities without altering discrimination.

*Fibrosis-stratified performance in the holdout cohort*

The holdout cohort exhibited a higher median extracted fibrosis stage (3.0 [2.0–4.0]) than the development cohort (2.5 [1.0–4.0]), yet, overall holdout performance remained stable (Table 2), indicating that etiological signal detection was not driven solely by fibrosis stage. To assess whether etiological inference varied across fibrosis severity, we performed a stage-stratified secondary analysis of the ×20 holdout predictions. A more granular four-stage analysis (stages 0–1, 2, 3, and 4) was uninformative at the tails because of small ASH counts ($n$ = 5 at stages 0–1), but identified a consistent step in performance between fibrosis <3 and fibrosis ≥3, which motivated the dichotomized presentation used here. The cohort was therefore split into early (stage <3, $n$ = 69; 12 ASH, 57 MASH) and advanced (stage ≥3, $n$ = 71; 31 ASH, 40 MASH) fibrosis groups, with single-ensemble confusion matrices and per-group metrics reported in Fig. 3.

In the early-fibrosis group, MASH classification remained near-perfect (specificity 96.5%), but ASH sensitivity dropped to 41.7% (5 of 12 cases correctly identified), with the model preferentially calling early fibrosis ASH slides as MASH. In the advanced-fibrosis group, ASH sensitivity recovered to 77.4% (24 of 31 cases correctly identified), with a parallel decline in MASH specificity to 87.5%, indicating bidirectional rather than one-sided misclassification. Balanced accuracy also diverged sharply (69.1% early, 82.5% advanced), and F1 for ASH was 52.6% in the early group versus 80.0% in the advanced group. Discriminative performance was therefore preserved at advanced fibrosis stages, where the etiological distinction has the greatest implications for management, whereas the early stage gap was concentrated entirely in ASH under-calling and was directly consistent with the small number of low-fibrosis ASH cases available during training rather than with a true loss of morphological signal. These stage-stratified estimates rest on small per-group ASH counts (12 ASH slides in the early fibrosis group, 31 in the advanced-fibrosis group); the corresponding sensitivities carry wide CIs and should be read as hypothesis-generating rather than as stable subgroup performance estimates.

*Attention-based clustering recovers interpretable histopathological patterns that correspond to known etiological features*

To characterize the histopathological content of patches driving slide-level predictions, we examined the distributions of patches with maximum attention weights and applied unsupervised clustering (HDBSCAN) to the top 5% of highest-attention patches from the 140-slide holdout cohort at ×20 magnification. This yielded 17,520 high-attention patches organized into 116 fine-grained clusters, which we consolidated into 20 super-clusters for descriptive review (Fig. 4). The super-clusters separated into ASH-enriched, MASH-enriched, and mixed groups: ASH-enriched super-clusters accounted for 4083 patches (23.3%), MASH-enriched super-clusters for 7656 patches (43.7%), mixed super-clusters for 2144 patches (12.2%), and 3637 patches (20.8%) remained unassigned as noise. The organization of high-attention patches into etiology-enriched clusters indicated that the model was concentrating on reproducible histopathological findings rather than on stochastic tissue features.

The morphological content of these super-clusters was then reviewed by a pathologist and found to correspond directly to known histopathological differentiators of ASH and MASH. ASH-enriched super-clusters were characterized by active steatohepatitic injury patterns: the highest-purity ASH-enriched clusters (SC15 and SC12) contained regions with marked macrovesicular steatosis, ballooned hepatocytes, Mallory–Denk bodies, cholestatic change, lobular disarray, and pericellular or bridging fibrosis with ductular reaction, recapitulating the classical description of alcohol-associated injury. Another ASH-enriched cluster (SC13) showed the greatest concentration of false-positive patches and included Mallory–Denk bodies, neutrophilic inflammation, cholestasis, and pericellular fibrosis, indicating that some MASH cases contained focal regions morphologically resembling active ASH, consistent with the well-recognized histological overlap between the two etiologies. A more heterogeneous ASH-enriched cluster (SC19) (Fig. 5) contained substantial proportions of both true-positive and false-negative patches and was characterized by relatively bland regions with sinusoidal dilatation, mild macrovesicular steatosis, and mild portal or periportal fibrosis without active inflammation, suggesting that quiescent or less overtly active disease could attract model attention, yet remain insufficient for correct slide-level classification.

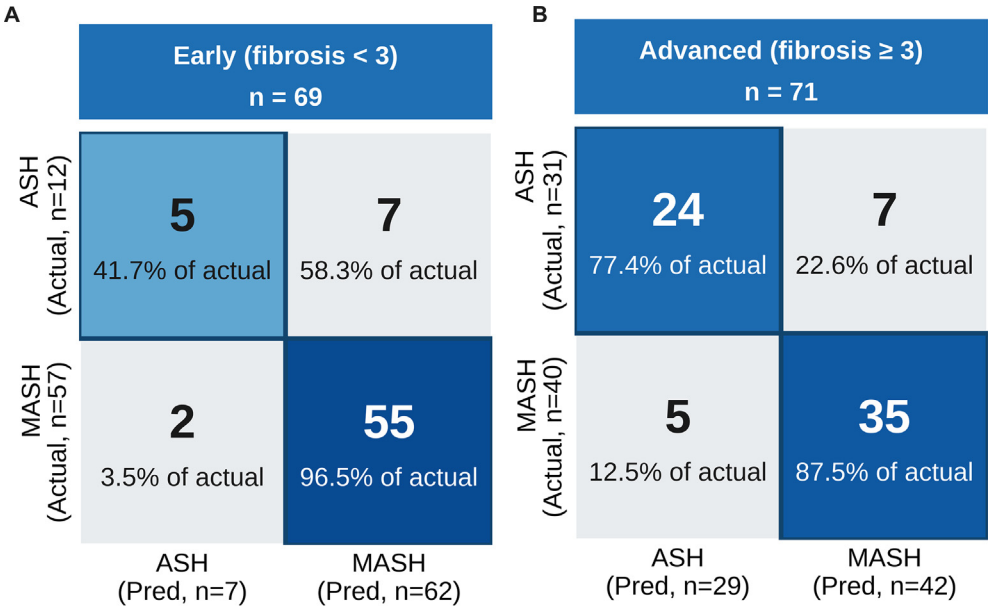
MASH-enriched super-clusters were dominated by lower inflammatory activity or more organized patterns. The highest-purity MASH-enriched clusters (SC20 and SC3) showed near-normal parenchyma or mild steatosis with glycogenated nuclei and little inflammatory activity. SC6, although MASH-enriched, contained marked steatosis with occasional neutrophilic satellitosis and mild pericellular fibrosis, indicating that individual features often associated with ASH were not independently sufficient to drive ASH classification when present outside broader ASH-like parenchymal findings. Mixed super-clusters reflected morphological overlap between etiologies: SC9 showed balanced contributions from true-positive, true-negative, false-positive, and false-negative patches and was characterized by macrovesicular steatosis with portal and periportal fibrosis with relatively organized lobular architecture. Fine-cluster decomposition within some apparently mixed or MASH-enriched super-clusters also revealed embedded false-negative subpopulations, consistent with localized regions of overlapping steatohepatitic morphology.

At ×40 magnification, the same clustering strategy yielded 17,518 high-attention patches, but most were absorbed into a single dominant

points between MASH and ASH within each magnification. Overall accuracy was 85.0% (×20) and 82.9% (×40); balanced accuracy was 80.1% and 79.2%. The ASH-specific recall gap is the principal target for further improvement. AUROC and AUPRC are reported as the mean ± SD across the 5-fold-models (threshold-independent, rank-based). Sensitivity, specificity, balanced accuracy, F1, PPV, NPV, and the confusion matrices (panels C–E) are computed from the single ensemble (mean per-slide probability) at the cross-validation-frozen F1 threshold. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



Downloaded for Anonymous User (n/a) at Hanyang University from ClinicalKey.com by Elsevier on October 14, 2026. For personal use only. No other uses without permission. Copyright ©2026. Elsevier Inc. All rights reserved.




C

| | Early fibrosis | | Advanced fibrosis | |
|---|---|---|---|---|
| | ASH | MASH | ASH | MASH |
| Sensitivity | 41.7% | 96.5% | 77.4% | 87.5% |
| PPV | 71.4% | 88.7% | 82.8% | 83.3% |
| F1 score | 52.6% | 92.4% | 80.0% | 85.4% |
| **Accuracy** | **87.0%** | | **83.1%** | |
| **Balanced accuracy** | **69.1%** | | **82.5%** | |

**Fig. 3.** Model prediction was preserved at advanced fibrosis. The holdout cohort was dichotomized by extracted fibrosis stage into early (stage <3, $n = 69$; 12 ASH, 57 MASH) and advanced (stage ≥3, $n = 71$; 31 ASH, 40 MASH) groups. All metrics are reported from the ×20 single ensemble (mean per-slide probability across the 5-fold-models) at the cross-validation-frozen F1 threshold (0.481). The model's behavior was not uniform across fibrosis severity: ASH was systematically under-called at low fibrosis, whereas at advanced fibrosis the two etiologies were separated symmetrically. (A) Early fibrosis group. MASH classification was near-perfect (96.5% specificity), but ASH sensitivity dropped to 41.7% (5 of 12 correctly identified), with the model preferentially calling early fibrosis ASH slides as MASH. The development cohort contained only a small number of low-fibrosis ASH cases, which might have contributed to the asymmetric error, but whether this represents a true biological limit will need further testing. (B) Advanced-fibrosis group. ASH sensitivity recovered to 77.4% (24 of 31 correctly identified), with a parallel decline in MASH specificity (87.5%). The model's errors at advanced fibrosis were bidirectional, likely suggesting a morphological overlap rather than minority-class under-calling. (C) Per-group metrics. Overall accuracy was comparable between groups (87.0% early, 83.1% advanced), but balanced accuracy diverged sharply (69.1% early, 82.5% advanced) because the early group's high MASH-side performance compensated for poor ASH recall. F1 for ASH was 52.6% early and 80.0% advanced. These stratified estimates rest on small per-group ASH counts (12 early fibrosis and 31 advanced-fibrosis ASH slides); the corresponding sensitivities carry wide confidence intervals and require confirmation in larger cohorts.

cluster with the remainder classified as noise (Fig. 4D inset). The model's slide-level decisions were therefore grounded in regions of histopathology that a pathologist would recognize as classical ASH or MASH features, with the failure modes (false-positive ASH calls driven by focal ASH-like morphology in MASH biopsies, and false-negative ASH calls driven by bland background diluting focal ASH features) explained by genuine morphological overlap and by the patch-averaging architecture rather than by abstract statistical artifacts. These interpretability findings derive from a single holdout cohort and a blinded, qualitative review of representative patches; the super-cluster subgroups are correspondingly small and require confirmation in larger, multi-institutional cohorts before they are treated as established feature-etiology associations.

## Discussion

Distinguishing ASH from MASH on routine histopathology is a longstanding challenge because the two share a steatohepatitic phenotype and no individual histological feature is pathognomonic for either etiology. Clinical adjudication relies on detailed exposure history and cardiometabolic profile, both of which may be incomplete and confounded by overlapping risk factors. In this single-institution proof-of-concept study, we tested whether routine H&E liver biopsy WSIs contained a detectable etiological signal in steatohepatitis. A gated-attention MIL framework is well suited to aggregate the subtle, spatially distributed patterns across the biopsy, potentially capturing composite morphological signatures that are not amenable to pathologists' scoring. Our evaluation showed that etiological discrimination was reproducible across 5-fold patient-level cross-validation and preserved on an independent 140-slide holdout cohort, with AUROC of 0.89 in cross-validation and 0.86 on holdout at ×20 magnification. The balanced accuracy and other metrics were also preserved on the independent holdout cohort, which highlighted the importance of the signal learned by the models and argued against overfitting to scanner-specific artifacts, given the heterogeneous scanner mixture



Downloaded for Anonymous User (n/a) at Hanyang University from ClinicalKey.com by Elsevier on October 14, 2026. For personal use only. No other uses without permission. Copyright ©2026. Elsevier Inc. All rights reserved.

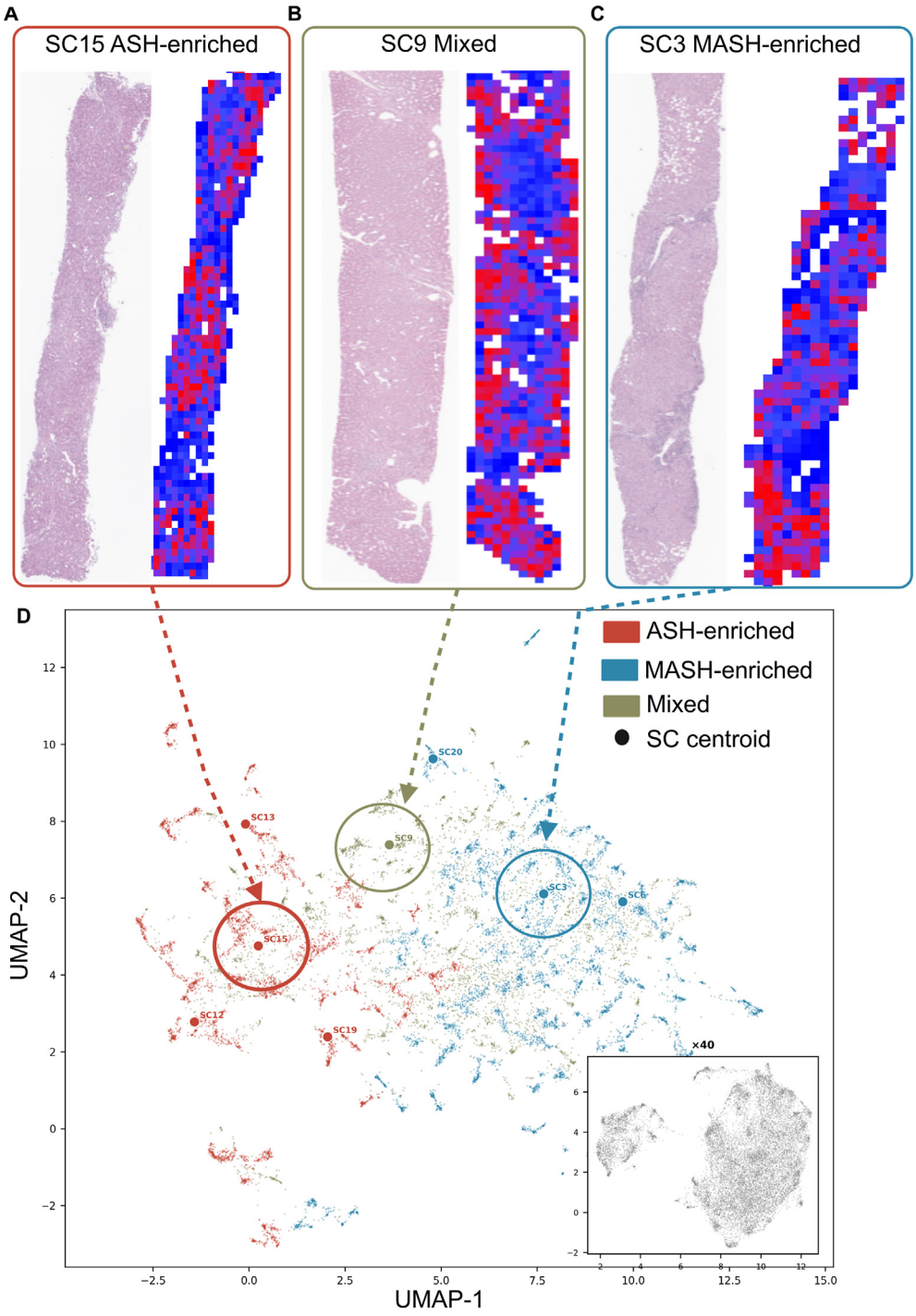



**Fig. 4.** Model attention concentrated on morphologically coherent regions, with ×20 capturing separable etiological structure that ×40 did not. For each holdout slide, we asked whether the regions the model weighted most heavily corresponded to identifiable histopathological patterns rather than arbitrary tissue. Three representative slides illustrate the relationship between attention placement and super-cluster membership; the joint embedding of all high-attention patches resolved into etiology enriched neighborhoods at ×20 but collapsed at ×40. (A–C) For each slide, the H&E thumbnail (left) is paired with its ×20 attention heatmap (right; red = high attention, blue = low attention, normalized within the slide). The dominant super-cluster of each slide's high-attention patches is annotated: (A) an ASH slide whose attention concentrated in SC15 (ASH-enriched), (B) an ASH slide whose attention concentrated in SC9 (mixed), and (C) a MASH slide whose attention concentrated in SC3 (MASH-enriched). (D) The top 5% highest-attention patches per slide across the 140-slide holdout cohort at ×20 magnification (17,520 patches) were embedded with UMAP and colored by super-cluster enrichment (red = ASH-enriched, blue = MASH-enriched, gold = mixed). Filled dots mark super-cluster centroids. The three example slides in panels A–C map to their dominant super-clusters. ASH-enriched and MASH-enriched super-clusters occupied separable regions of the embedding, with mixed super-clusters sitting at the interface. The ×40 inset shows the corresponding UMAP at ×40, which collapsed into a single continuous region with no etiological separation. The greater separability at ×20 supports the ×20 advantage observed quantitatively in the main analysis: discriminative signal in this cohort resides in architectural and spatially distributed patterns rather than in cytological detail. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



Downloaded for Anonymous User (n/a) at Hanyang University from ClinicalKey.com by Elsevier on October 14, 2026. For personal use only. No other uses without permission. Copyright ©2026. Elsevier Inc. All rights reserved.

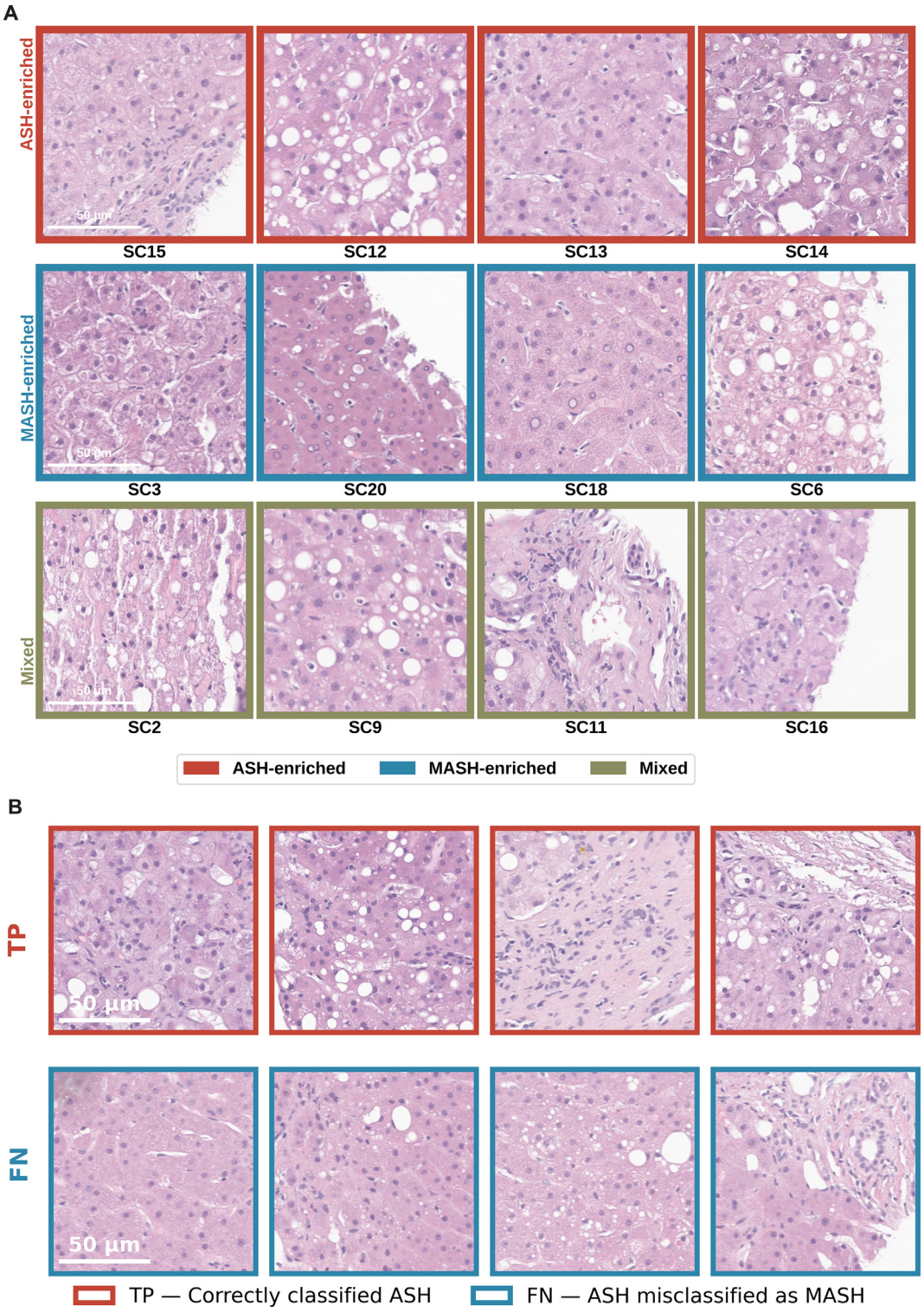



**Fig. 5.** Super-cluster morphology recapitulated established ASH and MASH histopathology, and the dominant failure mode was interpretable rather than random. To establish whether the model's slide-level decisions were grounded in recognizable pathology, we examined the morphological content of every super-cluster. The model's attention concentrated on regions that a pathologist would describe as steatohepatitis features, and the false negatives were explained by the patch-averaging architecture acting on focally distributed disease rather than by a breakdown in feature learning. (A) Representative high-attention patches from the 12 most discriminative super-clusters are grouped by enrichment category. The colored border indicates enrichment: Top row (red, ASH-enriched; ASH fraction $\geq 0.65$): SC15, SC12, SC13, and SC14 contained marked macrovesicular steatosis, ballooned hepatocytes, Mallory–Denk bodies, cholestatic change, neutrophilic inflammation, and pericellular or bridging fibrosis with ductular reaction, recapitulating the classical description of alcohol-associated injury. Middle row (blue, MASH-enriched; ASH fraction $\leq 0.35$): SC3, SC20, SC18, and SC6 showed near-normal parenchyma, mild steatosis with glycogenated nuclei, and lower inflammatory activity. SC6 contained marked steatosis with occasional neutrophilic satellitosis, indicating that individual ASH-like features were not sufficient to drive ASH classification outside a broader ASH-like context. Bottom row (olive, mixed; ASH fraction 0.35–0.65): SC2, SC9, SC11, and SC16 captured genuine morphological overlap between etiologies, including macrovesicular steatosis with portal and periportal fibrosis and organized lobular architecture. The correspondence between super-cluster morphology and established differentiating features supports the conclusion that the model's slide-level decisions were grounded in known histopathology rather than abstract statistical artifacts. (B) SC19 was ASH-enriched



Downloaded for Anonymous User (n/a) at Hanyang University from ClinicalKey.com by Elsevier on October 14, 2026. For personal use only. No other uses without permission. Copyright ©2026. Elsevier Inc. All rights reserved.

spanning the study period and the absence of stain normalization. This learned signal was also biologically meaningful: attention-based clustering with pathologist interpretation localized ASH-enriched regions to active injury patterns (Mallory–Denk bodies, neutrophilic inflammation, cholestatic change, lobular disarray, and pericellular or bridging fibrosis), whereas MASH-enriched regions more often showed organized steatosis, glycogenated nuclei, and lower inflammatory backgrounds. The detection of an etiological signal in the liver biopsies supports the use of routine H&E morphology as an additional, objective input alongside the patient's exposure history and metabolic profile.

The comparable performance of the $\times 20$ and $\times 40$ models, together with the collapse of $\times 40$ high-attention patches into a single dominant cluster under the same unsupervised clustering strategy, indicated that etiological discrimination depended more on architectural and spatially distributed lobular patterns than on isolated cytological detail, and supported our use of the $\times 20$ model as the primary model for the interpretability analyses. These interpretable attention maps and the cluster-level correspondence to established ASH and MASH morphology are supportive of the model attending to recognizable histopathology and provide a patch-level focus for future mechanistic analysis of these highly attended regions. The result that lobular and architectural context outperformed high-resolution cytology has practical implications for clinical deployment: $\times 20$ patches require roughly four times less storage and compute per slide than $\times 40$ patches, lowering the operational barrier to running this type of model on routinely scanned biopsies in centers that do not maintain dedicated high-resolution archives.

Beyond the choice between magnifications, the choice of ensemble configuration is itself a clinical decision that depends on how the model's output is intended to be used. A single $\times 20$ model offered the simplest operational footprint and the highest specificity in our evaluation (0.93 on holdout). High specificity is desirable when the cost of a false-positive ASH call is high, for example, when the prediction is used as a flagging signal that would trigger a chart review, a re-examination of alcohol-use history, or a hepatology consultation. In such workflows, the dominant failure mode to minimize is over-calling ASH in a patient whose true disease is metabolic, because an erroneous etiological label could affect counseling, insurance coding, or eligibility for MASH-specific therapeutics. Conversely, when sensitivity for the minority class is the priority, a cross-magnification OR rule can recover additional ASH cases (sensitivity 70.6 versus 67.4 for $\times 20$ alone in cross-validation) at a small specificity cost, although this requires deploying both $\times 20$ and $\times 40$ models in parallel. High sensitivity is desirable when the cost of missing an ASH case is high, for example, in prospective enrichment for ASH clinical trials, in screening biopsies from patients with documented alcohol use disorder, or in clinical trials where misattribution of ASH cases to a MASH cohort would obscure treatment effects. We therefore recommend the single $\times 20$ model as the default configuration for general decision-support deployment, with cross-magnification OR ensembling reserved for enrichment or screening use cases in which minority-class recall is explicitly preferred over operational simplicity, recognizing that any ensemble that combines two magnifications doubles the storage, compute, and validation burden of a deployed pipeline.

The clinical value of etiological inference from routine histology spans both clinical and research applications. In the clinical setting, in cases with incomplete or conflicting exposure histories, particularly when patient-reported alcohol intake conflicts with the metabolic profile, a morphology-based etiological estimator can serve as a decision-support signal that prompts more detailed clinical review. In the research setting, systematic morphological adjudication may help stratify steatohepatitis cohorts, whose etiological assignment is currently based on clinical documentation alone, which directly affects eligibility for clinical trials. For example, patients failing MASH-specific therapeutics may have underlying histology that supports pure ASH or MASH with increased alcohol intake (MetALD). We emphasize that the present model is designed as a translational proof of concept and not as a replacement for clinical adjudication. The high MASH specificity of the model (0.93 on holdout) supports its potential use as a flagging tool for cases, whose morphology is discordant with the clinical etiological assignment, similar to how discordant-case review has been used in other AI-assisted pathology pipelines.[27,29] The relevant use could be assisted adjudication, in which the model identifies a small number of high-confidence discordant cases that warrant pathologist or hepatologist re-review, leaving the final etiological judgment with the clinician.

The preservation of overall performance on a holdout cohort with a higher median fibrosis stage than the development cohort (3.0 [2.0–4.0] versus 2.5 [1.0–4.0]) indicated that etiological signal detection did not depend solely on fibrosis stage and did not collapse when the background fibrosis distribution shifted. Specificity at $\times 20$ across cohorts (0.93–0.95) was also preserved, confirming that the models did not generate false alarms at a rate that would prohibit downstream clinical or research use. The remaining gap between cross-validation and holdout sensitivity (0.67 at $\times 20$ on holdout, 95% CI 0.53–0.81, versus higher cross-validation values) reflected the sensitivity of minority-class recall to small-sample variation and to systematic clinical differences between the two cohorts, and we plan to revisit this as the system is extended to larger multi-institutional evaluation.

A stage-stratified secondary analysis refined the main conclusion: fibrosis stage modulated the ease of etiological classification without driving it on its own. Dichotomizing the holdout cohort into early (fibrosis $<3$, $n = 69$; 12 ASH) and advanced (fibrosis $\geq 3$, $n = 71$; 31 ASH) groups, the model preserved discrimination at advanced fibrosis (balanced accuracy 82.5%, ASH sensitivity 77.4%) but under-called ASH at early fibrosis (balanced accuracy 69.1%, ASH sensitivity 41.7%). The early stage gap was concentrated entirely in ASH under-calling and was consistent with the small number of low-fibrosis ASH cases available during training, rather than with a true loss of morphological signal. These findings motivate future enrichment of low-fibrosis ASH cases, explicit stage-aware validation, and multimodal approaches that incorporate clinical or molecular information for the earliest or most morphologically equivocal biopsies. Critically, the model performed best in the fibrosis range where the clinical decision between ASH and MASH carries the greatest consequence for management, because patients with advanced fibrosis are those most likely to be considered for liver transplantation, abstinence-conditional listing, or enrollment in disease-modifying trials, all of which depend on accurate etiological attribution.

This study has several limitations. First, the reference standard was clinical rather than biological: we used the treating hepatologist's adjudicated etiological diagnosis as ground truth, which depends on patient reporting of alcohol consumption. Because this reference standard relies on documented clinical adjudication rather than a biological gold-standard, the slide-level labels carry a degree of label noise, most plausibly from under-reported alcohol use that would assign true ASH cases a MASH label. Such misclassification in the training labels would, if anything, bias measured discrimination downward, so the reported performance is likely a

but contained a high proportion of false-negative patches. Eight representative patches are shown, drawn from true-positive predictions (top row, red borders) and false-negative predictions (bottom row, blue borders) within the same super-cluster. The false-negative patches resembled normal liver parenchyma or minimally active steatohepatitis, with sinusoidal dilatation, mild macrovesicular steatosis, and mild portal or periportal fibrosis but no active inflammation. Because the slide-level ASH probability was computed as the average of patch-level probabilities, an ASH slide composed primarily of these morphologically bland patches yielded a low slide-level probability and a false-negative call. The dominant failure mode might be architectural (focal disease diluted by bland background), not a failure of the model to learn discriminative features. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



Downloaded for Anonymous User (n/a) at Hanyang University from ClinicalKey.com by Elsevier on October 14, 2026. For personal use only. No other uses without permission. Copyright ©2026. Elsevier Inc. All rights reserved.

conservative estimate of the true morphological separability. We deliberately excluded MetALD cases from this analytic cohort because mixed etiologies deserve dedicated modeling rather than forced binary assignment; this decision also means that the present model is not designed to handle the most clinically ambiguous cases. Extension to MetALD and to other etiologies such as drug-induced liver injury will be needed for a comprehensive digital-pathology framework for all SLDs. Second, this is a single-institution proof-of-concept study. Scanner heterogeneity within the cohort tests one dimension of robustness but does not replace external validation, because the development and holdout cohorts share a single center, patient population, and set of lab and staining protocols. Prospective, multi-institutional validation remains a prerequisite for clinical deployment. Third, sensitivity on the holdout test dataset, although preserved, was lower than specificity and exhibited larger variance, reflecting the minority-class prevalence of ASH in this cohort and the sensitivity of recall metrics to small sample sizes; larger ASH cohorts will be required to derive clinically meaningful results. Fourth, the interpretability analysis is descriptive and blinded; it is not a quantitative attribution analysis. Attention weights identify the regions the model relies on, and the blinded cluster-level correspondence to established ASH and MASH morphology is supportive evidence that the predictions track recognizable histopathology; neither attention maps nor clustering establishes that those features causally drive the model's decisions. Formal, quantitative feature attribution and correlation with transcriptomic or spatial molecular data will be needed to establish causal explanations, and are the focus of our future work.

Weakly supervised gated-attention MIL applied to routine liver H&E WSIs distinguished ASH from MASH using morphology alone and maintained performance on an independent holdout cohort, with attention-based regions that corresponded to established histopathological features of steatohepatitis. Our approach showed that steatohepatitic injury retained an etiological morphological signature detectable at the slide level, with preserved performance across scanner platforms and advanced fibrosis, and with model attention focused on known histological features. As MASH-directed therapeutics enter clinical practice and as the nomenclature of SLDs evolves to recognize mixed etiologies, the ability to extract a morphology-based etiological signal from a routinely scanned H&E biopsy offers an additional modality that complements clinical adjudication without replacing it, particularly for cases in which the patient's exposure history and metabolic profile point in different directions.

## Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the author(s) used Claude in order to identify grammatical and typographical errors. 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 published article.

## CRediT authorship contribution statement

Using the CRediT (Contributor Roles Taxonomy) framework: **C.M.**: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing - original draft, Writing - review & editing. **S.H.I.**: Resources, Supervision, Writing - review & editing. **V.H.S.**: Resources, Supervision, Writing - review & editing. **H.T.**: Methodology, Supervision, Writing - review & editing. **Y.N.**: Writing - review & editing. **I.K.**: Writing - review & editing. **P.L.**: Supervision, Writing - review & editing.

## Ethics approval and consent to participate

This retrospective study was approved by the Mayo Clinic Institutional Review Board (IRB #20-253). Only patients with research authorization were included.

## Funding

Research reported in this publication was supported by the National Institute on Alcohol Abuse and Alcoholism (NIAAA) of the National Institutes of Health under Award Number K08AA032553 and by the National Institute of Diabetes and Digestive and Kidney Diseases of the National Institutes of Health under Award Number P30DK084567. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

## Declaration of competing interest

The authors declare the following financial interests/personal relationships, which may be considered as potential competing interests:

Chady Meroueh reports financial support was provided by National Institute on Alcohol Abuse and Alcoholism. Chady Meroueh reports financial support was provided by National Institute of Diabetes and Digestive and Kidney Diseases. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.

## Acknowledgments

The authors thank the Pathology Research Core at Mayo Clinic for their contributions to tissue processing, slide preparation, and slide scanning.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jpi.2026.100712.

## Data availability

The whole-slide images and linked clinical data are not publicly available because of patient privacy and institutional restrictions. De-identified derived data and analysis code sufficient to reproduce the reported results are available from the corresponding author on reasonable request, subject to institutional review and applicable data-use restrictions.

## References

1. Kwong AJ, Ebel NH, Kim WR, et al. OPTN/SRTR 2020 annual data report: liver. Am J Transplant 2022;22(suppl 2):204–309.
2. Bertuccio P, Turati F, G C. Global trends and predictions in HCC mortality. J Hepatol;67:302–309.
3. Cholankeril G, Yoo ER, Rb P. Rising rates of hepatocellular carcinoma leading to liver transplantation in baby boomer generation with chronic hepatitis C, alcohol liver disease, and nonalcoholic steatohepatitis-related liver disease. Diseases;5.
4. Asrani SK, Devarbhavi H, Eaton J, et al. Burden of liver diseases in the world. J Hepatol 2019;70:151–171.
5. Singal AG, Lampertico P, Nahon P. Epidemiology and surveillance for hepatocellular carcinoma: new trends. J Hepatol 2020;72:250–261.
6. Younossi ZM, Koenig AB, Abdelatif D, et al. Global epidemiology of nonalcoholic fatty liver disease—meta-analytic assessment of prevalence, incidence, and outcomes. Hepatology 2016;64:73–84.
7. Younossi Z, Anstee QM, Marietti M, et al. Global burden of NAFLD and NASH: trends, predictions, risk factors and prevention. Nat Rev Gastroenterol Hepatol 2018;15:11–20.
8. Idalsoaga F, Kulkarni AV, Mousa OY, et al. Non-alcoholic fatty liver disease and alcohol-related liver disease: two intertwined entities. Front Med (Lausanne) 2020 Aug 20;7:448.
9. Thursz M, Gual A, Lackner C, et al. EASL Clinical Practice Guidelines: management of alcohol-related liver disease. J Hepatol 2018;69:154–181.
10. Grant BF, Chou SP, Saha TD, et al. Prevalence of 12-month alcohol use, high-risk drinking, and DSM-IV alcohol use disorder in the United States, 2001-2002 to 2012-2013: results from the National Epidemiologic Survey on Alcohol and Related Conditions. JAMA Psychiatry 2017;74:911–923.
11. Rinella M, Lazarus J, Ratziu V, et al. A multi-society Delphi consensus statement on new fatty liver disease nomenclature. Hepatology 2023;78(6):1966–1986. https://doi.org/10.1097/HEP.0000000000000520. Epub 2023 Jun 24. PMID: 37363821; PMCID: PMC10653297. NNc, group.
12. Díaz LA, Arab JP, Louvet A, et al. The intersection between alcohol-related liver disease and nonalcoholic fatty liver disease. Nat Rev Gastroenterol Hepatol 2023;20:764–783.
13. Harrison SA, Bedossa P, Guy CD, et al. A phase 3, randomized, controlled trial of resmetirom in NASH with liver fibrosis. N Engl J Med 2024;390:497–509.



Downloaded for Anonymous User (n/a) at Hanyang University from ClinicalKey.com by Elsevier on October 14, 2026. For personal use only. No other uses without permission. Copyright ©2026. Elsevier Inc. All rights reserved.

14. Sanyal AJ, Newsome PN, Kliers I, et al. Phase 3 trial of semaglutide in metabolic dysfunction-associated steatohepatitis. N Engl J Med 2025;392:2089–2099.
15. Kleiner DE, Brunt EM, Van Natta M, et al. Design and validation of a histological scoring system for nonalcoholic fatty liver disease. Hepatology 2005;41:1313–1321.
16. Torruellas C, French SW, Medici V. Diagnosis of alcoholic liver disease. World J Gastroenterol 2014;20:11684–11699.
17. Sakhuja P. Pathology of alcoholic liver disease, can it be differentiated from nonalcoholic steatohepatitis? World J Gastroenterol 2014;20:16474–16479.
18. Brown GT, Kleiner DE. Histopathology of nonalcoholic fatty liver disease and nonalcoholic steatohepatitis. Metabolism 2016;65:1080–1086.
19. Litjens G, Bandi P, Ehteshami Bejnordi B, et al. 1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. Gigascience 2018;7.
20. Liu F, Wei L, Leow WQ, et al. Developing a new qFIBS model assessing histological features in pediatric patients with non-alcoholic steatohepatitis. Front Med 2022;9, 925357.
21. Perincheri S, Levi AW, Celli R, et al. An independent assessment of an artificial intelligence system for prostate cancer detection shows strong diagnostic accuracy. Mod Pathol 2021;34:1588–1595.
22. Campanella G, Hanna MG, Geneslaw L, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med 2019;25:1301–1309.
23. 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;24:1559–1567.
24. 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.
25. Ratziu V, Francque S, Behling CA, et al. Artificial intelligence scoring of liver biopsies in a phase II trial of semaglutide in nonalcoholic steatohepatitis. Hepatology 2024;80:173–185.
26. Bosch J, Chung C, Carrasco-Zevallos OM, et al. A machine learning approach to liver histological evaluation predicts clinically significant portal hypertension in NASH cirrhosis. Hepatology 2021;74:3146–3160.
27. Pulaski H, Mehta SS, Manigat LC, et al. Validation of a whole slide image management system for metabolic-associated steatohepatitis for clinical trials. J Pathol Clin Res 2024;10, e12395.
28. Abdurrachim D, Wee A, Soon G, et al. AI-assisted fibrosis scoring in MASH: exploring pathologist decision-making with an SHG-based AI digital pathology tool. JHEP Rep 2026;8.
29. Iyer JS, Juyal D, Le Q, et al. AI-based automation of enrollment criteria and endpoint assessment in clinical trials in liver diseases. Nat Med 2024;30:2914–2923.
30. Bankhead P, Loughrey MB, Fernandez JA, et al. QuPath: open source software for digital pathology image analysis. Sci Rep 2017;7, 16878.
31. Kang M, Song H, Park S, et al. Benchmarking self-supervised learning on diverse pathology datasets. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023.
32. Chen X, Fan H, Girshick R, et al. Improved baselines with momentum contrastive learning. arXiv preprint 2020.ArXiv 2020 Vol. abs/2003.04297.
33. Ilse M, Tomczak J, Welling M. Attention-based deep multiple instance learning. In: *Jennifer D, Andreas K, eds. Proceedings of the 35th International Conference on Machine Learning*. Proceedings of Machine Learning Research: PMLR; 2018. p. 2127–2136.
34. Paszke A, Gross S, Massa F, et al. In: *Wallach H, Larochelle H, Beygelzimer A, Alch-Buc Fdt, Fox E, Garnett R, eds. PyTorch: An Imperative Style, High-Performance Deep Learning Library, Vol. 32*. 2019.
35. Falcon W, Team TPL. *PyTorch Lightning*. Zenodo. 2025.
36. Lin T-Y, Goyal P, Girshick R, et al. Focal loss for dense object detection. Proceedings of the IEEE International Conference on Computer Vision; 2017.
37. Pakdaman Naeini M, Cooper G, Hauskrecht M. Obtaining well calibrated probabilities using Bayesian binning. Proceedings of the AAAI Conference on Artificial Intelligence, 29. 2015.
38. Boumaraf S, Liu X, Wan Y, et al. Conventional machine learning versus deep learning for magnification dependent histopathological breast cancer image classification: a comparative study with visual explanation. Diagnostics (Basel, Switzerland) 2021;11, 528.



Downloaded for Anonymous User (n/a) at Hanyang University from ClinicalKey.com by Elsevier on October 14, 2026. For personal use only. No other uses without permission. Copyright ©2026. Elsevier Inc. All rights reserved.