



Degree Project in Medical Engineering
Second cycle, 30 credits

Quantifying Fibrosis for Staging Metabolic Dysfunction-Associated Steatotic Liver Disease from Histopathology Images using Deep Learning

HELENA CHAMBERLAIN ALVARADO

Quantifying Fibrosis for Staging Metabolic Dysfunction-Associated Steatotic Liver Disease from Histopathology Images using Deep Learning

HELENA CHAMBERLAIN ALVARADO

Master's Programme, Medical Engineering

Supervisor: Francisco J. Peña

Co-Supervisor: Fahim Ebrahimi (University Hospital Basel, Switzerland)

Reviewer: Jayanth Raghothama

Examiner: Matilda Larsson

A thesis conducted at the Department of Biomedical Engineering and Health Systems, KTH Royal Institute of Technology

Swedish title: Kvantifiering av fibros för stadiindelning av metabolisk dysfunktion-associerad steatotisk leversjukdom från histopatologiska bilder med hjälp av djupinlärning

TRITA-CBH-GRU-2026:055

Stockholm, Sweden 2026

www.kth.se

Abstract

Metabolic dysfunction-associated steatotic liver disease (MASLD) is a significant and growing clinical challenge that can progress to end-stage liver disease and liver-related mortality. This progression is closely linked to the accumulation of liver fibrosis, making accurate fibrosis staging critical for prognosis. While fibrosis primarily involves the deposition of collagen fibers around vessels, current machine learning-based methods often overlook the spatial relationship between these microstructures or rely on simplistic spatial proximity rules. This limitation restricts the predictive performance of automated staging tools. To address this, we propose a biologically informed, topology-aware graph construction framework where nodes and edges explicitly model vessel-collagen architecture. From these graphs, biologically meaningful feature embeddings are extracted and used to train an advanced graph neural network (GNN) model known as Graph isomorphism networks with edge features (GINE). Using a dataset of 120 slides, covering the full spectrum of fibrosis, we systematically compare our approach with alternative graph construction strategies and benchmark it against a baseline patch-based model and a state-of-the-art (SOTA) graph-based model. Our results demonstrate that, despite challenges in vessel segmentation, the proposed graph construction captures biologically consistent patterns, including a transition from a sparse to an interconnected fibrotic network, and yields improved classification performance compared to purely spatial-based graphs, doubling the accuracy from 0.24 to 0.48. Furthermore, while the baseline and SOTA models struggle with subtle transitions of intermediate fibrosis stages, our framework yields a significantly higher quadratic weighted kappa of 0.44 (an increase of 0.16 over SOTA) and accuracy of 0.37 (an increase of 0.11 over SOTA), demonstrating consistent performance across all stages. These findings highlight the importance of incorporating domain-specific biological knowledge into model design to enhance predictive performance in MASLD fibrosis staging.

Keywords: MASLD, Liver fibrosis, Digital pathology, Graph Neural Networks, Topology-aware learning, Biologically informed modeling

Sammanfattning

Metabolisk dysfunktion-associerad steatotisk leversjukdom (MASLD) är ett betydande och växande kliniskt problem som kan utvecklas till leversjukdom i slutstadiet och leda till dödsfall på grund av leversjukdom. Denna utveckling är nära kopplad till uppkomsten av leverfibros, vilket gör en korrekt stadiindelning av fibrosen avgörande för prognosen. Även om fibros i första hand innebär att kollagenfibrer avsätts runt blodkärlen, bortser dagens maskininlärningsbaserade metoder ofta från det rumsliga sambandet mellan dessa mikrostrukturer eller förlitar sig på förenklade regler för rumslig närhet. Denna begränsning påverkar både den prediktiva prestandan och den kliniska tolkbarheten hos automatiserade verktyg för stadiindelning. För att lösa detta föreslår vi ett biologiskt underbyggt ramverk för grafkonstruktion med hänsyn till topologi, där noder och kanter uttryckligen modellerar arkitekturen hos blodkärl och kollagen. Från dessa grafer extraheras biologiskt meningsfulla egenskapsinbäddningar som används för att träna en avancerad Graph Neural Network (GNN)-modell kallas nätverk för grafisomorfism med kantegenskaper (GINE). Med hjälp av en dataset bestående av 120 bilder, som täcker hela spektrumet av fibros, jämför vi systematiskt vår metod med alternativa strategier för grafkonstruktion och utvärderar den mot en referensmodell baserad på patch-metoden samt en grafbaserad modell som representerar den senaste tekniken (SOTA). Våra resultat visar att den föreslagna grafkonstruktionen, trots utmaningarna med kärlsegmentering, fångar upp biologiskt konsistenta mönster - däribland en övergång från ett gles till ett sammanlänkat fibrotiskt nätverk - och ger bättre klassificeringsresultat jämfört med rent rumsbaserade grafer, vilket innebär att noggrannheten fördubblas från 0.24 till 0.48. Dessutom, medan basmodellen och de SOTA har svårt att hantera subtila övergångar mellan mellanliggande fibrosstadier, uppnår vårt ramverk ett betydligt högre kvadratisk vägt kappa-värde på 0.44 (en ökning med 0.16 jämfört med SOTA) och en träffsäkerhet på 0.37 (en ökning med 0.11 jämfört med SOTA), vilket visar på en jämn prestanda i alla stadier. Dessa resultat understryker vikten av att integrera områdesspecifik biologisk kunskap i modellutformningen för att förbättra både förutsägbarheten och tolkbarheten vid stadiindelning av fibros vid MASLD.

Nyckelord: MASLD, Leverfibros, Digital Patologi, Grafiska Neuronnätverk, Topologibevittnat Lärande, Biologiskt Underbyggd Modellering

Acknowledgments

First and foremost, I would like to thank my family and friends, whose support and advice made this learning journey possible. I sincerely thank my supervisor, Francisco J. Peña, and Dr. Fahim Ebrahimi for relying on me and allowing me to contribute to their research. Their valuable guidance and feedback have been essential for this project. I would also like to thank the University Hospital Basel for providing the whole dataset. I also want to express my gratitude to the whole Learning Beyond Supervision Lab for their support and advice during my master thesis. Finally, but not less importantly, I offer special thanks to all my professors at KTH to whose time and lectures I am indebted.

Contents

1	Introduction	1
2	Background	4
2.1	MASLD: Spectrum and Pathogenesis	4
2.2	Liver fibrosis	4
2.2.1	Mechanism and Topography	4
2.2.2	Current Diagnosis and Gold Standard	6
2.3	Computational Histopathology and Related Work	6
2.3.1	Multiple Instance Learning for Liver Fibrosis	7
2.3.2	Graph Neural Networks for Liver Fibrosis	9
3	Methodology	12
3.1	Data collection and Preprocessing	12
3.2	Proposed Framework: from Graph Construction to Deep Learning Model Implementation	13
3.2.1	Collagen and Vessel Mask Construction	13
3.2.2	Graph Construction	14
3.2.3	Graph Embedding	16
3.2.4	GNN Architecture and Training Procedure	17
3.3	Baseline and State-of-the-Art Models Implementation	18
3.4	Evaluation metrics	19
4	Results and Analysis	20
4.1	Graph Construction Framework Results	20
4.1.1	Performance of Vessel Segmentation	20
4.1.2	Graph interpretation on WSIs	22
4.2	GNN implementation Results	23
4.2.1	Hyperparameter Optimization and Architectural Design Evaluation	23
4.2.2	Comparative Study on Graph Construction	24
4.2.3	Comparative study on different GNN architectures	25
4.3	Comparison With Baseline and State-of-the-Art Models for fibrosis staging	26
5	Discussion	27
5.1	Graph Construction Pipeline	27
5.1.1	Vessel-Collagen Segmentation	27
5.1.2	Graph interpretation on WSIs	27
5.1.3	Model implementation	28
5.2	Comparison With Baseline and State-of-the-Art Models	28
5.3	Ethical Considerations and Clinical Implications	29
5.4	Limitations	29
5.5	Future work	30
6	Conclusion	31

Bibliography 32

Appendix 40

List of Figures

1	Illustration of fibrosis staging from liver biopsy images using common ML and DL methods and the proposed graph-structured representation	2
2	Liver fibrosis stages (METAVIR standard)	5
3	Automated collagen segmentation pipeline	7
4	Overview of the methodology of the project	12
5	Implementation of automated vessel segmentation and classification pipeline	13
6	Graph Construction Pipeline	15
7	Representative figure of vessel detection	20
8	Representative figure of vessel classification	21
9	Representative example of the generated biologically-informed topology-aware graph	22
10	Distribution of key topological and biological features across fibrosis stages with the proposed graph construction	23
11	Representative hyperparameter optimization and architectural comparative results	24
12	Summary of main steps for proposed framework	40
13	Confusion matrix summarizing classifier performance for vessel segmentation.	42
14	Confusion matrix summarizing classifier performance for vessel classification.	42
15	Representative figure of vessel detection using normalized and non-normalized data	44
16	t-SNE plot of graph representations	45

List of Tables

1	Description of Node Features	16
2	Description of Edge Features	17
3	Hyperparameter Search Space for GINE Model Optimization . .	18
4	Performance of the evaluated classifiers for vessel segmentation	20
5	Performance of vessel classification	21
6	Effectiveness of the proposed graph construction pipeline	25
7	Effectiveness of the proposed GNN architecture	25
8	Fibrosis staging performance of the different DL methods on the test set	26
9	ML classification performance with and without normalized data	44

List of Abbreviations

AI	Artificial intelligence
AttenMIL	Attention-based MIL
CE	Cross-Entropy Loss
CNN	Convolutional neural network
CPA	Collagen proportionate area
CAD	Computer-Aided Diagnosis
DL	Deep learning
GAT	Graph attention networks
GCN	Graph convolutional networks
GIN	Graph isomorphism networks
GINE	Graph isomorphism networks with edge features
GNN	Graph neural networks
H&E	Hematoxylin and Eosin
HCC	Hepatocellular carcinoma
HSV	Hue-Saturation-Value
IBMIL	Interventional Bag Multi-Instance Learning
JK-Net	Jumping Knowledge Network
KNN	K-Nearest Neighbor
LR	Learning Rate
ML	Machine learning
MASH	Metabolic dysfunction-associated steatohepatitis
MASL	Metabolic dysfunction-associated steatotic liver
MASLD	Metabolic dysfunction-associated steatotic liver disease
MIL	Multiple instance learning
MLP	Multilayer perceptron
QWK	Quadratic Weigthed Kappa
SMOTE	Synthetic minority oversampling technique
GB-SMOTE	Gradient Boosting with SMOTE
SR	Sirius red
SOTA	State-Of-The-Art
OD	Optical density
XGBoost	Extreme Gradient Boosting
WD	Weight Decay
WSI	Whole slide images

1 Introduction

Metabolic dysfunction-associated steatotic liver disease (MASLD) has rapidly become one of the most prevalent chronic liver diseases worldwide, affecting over one-third of the adult population and representing an increasing burden on healthcare systems [1]. Its clinical course is highly variable and can progress to more severe liver diseases, including fibrosis, cirrhosis, and, ultimately, hepatocellular carcinoma (HCC) [2].

Among the various pathological features of MASLD, liver fibrosis is the strongest predictor of disease progression and liver-related morbidity and mortality [3]. Therefore, accurate fibrosis staging is essential for prognosis, treatment, and monitoring disease progression [4]. Currently, fibrosis staging relies on histological examinations of liver biopsy specimens by expert pathologists. However, this process presents several limitations, including high rates of inter-observer variability [5].

Artificial intelligence (AI) has emerged as a promising tool for achieving a more reliable, scalable, and standardized fibrosis assessment. In this setting, biopsy specimens are digitized as whole slide images (WSIs) and analyzed using machine learning (ML) models like multiple instance learning (MIL) or deep learning (DL) algorithms like graph neural networks (GNNs) [6]. MIL can handle weakly labeled slides and large datasets, but treats each WSI as a collection of independent patches, missing spatial relationships between tissue structures that are essential for staging [7, 8]. GNNs explicitly model spatial dependencies and tissue architecture. However, their performance depends heavily on a high-quality graph construction process, which primarily relies on geometric proximity (e.g., k-nearest neighbors (KNN) or distance-based connectivity) rather than biologically meaningful relationships [7]. This reliance potentially limits their ability to capture the true structural mechanisms underlying fibrosis progression [9].

To address these limitations, this project aims to develop and evaluate a biologically informed, topology-aware DL framework for automated fibrosis staging in MASLD using Sirius Red (SR)-stained WSIs. The proposed approach explicitly models the spatial relationship between vessels and collagen by constructing graphs that treat vessels as nodes and collagen fibers as edges. By combining segmentation-derived masks, a biologically motivated graph construction, and a novel GNN architecture called Graph Isomorphism Network with Edge features (GINE), this study investigates whether modeling tissue architecture improves fibrosis classification compared to patch-based MIL approaches. Additionally, it evaluates the advantages of using biologically informed representations and GINE over geometry-based constructions and standard GNN models commonly used in current literature.

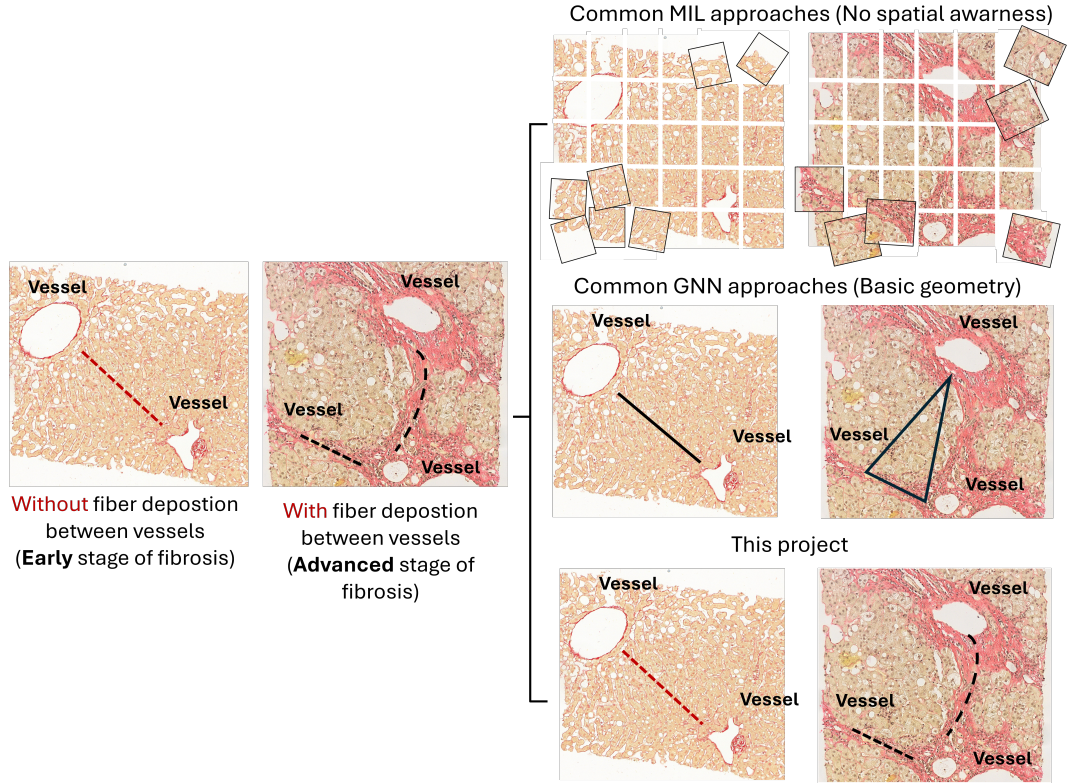


Figure 1: Illustration of fibrosis staging from liver biopsy images using common ML and DL methods and the proposed graph-structured representation. (Right) Liver biopsy image samples with different fibrosis stages. (Left) Concept illustration of commonly used ML and DL methods: (top) MIL considers independent tiles losing spatial information, (middle) standard GNNs depict tissue with basic geometry missing biological information, and (bottom) the proposed graph structured representation where vessels are treated as nodes and the collagen fibers are treated as edges of the graph.

To reach this objective, this project is organized around the following sub-objectives:

1. Preprocess available WSIs to ensure a consistent and representative dataset.
2. Develop and implement a pipeline to construct biologically informed, topology-aware graphs that explicitly encode vascular and collagen structures.
3. Design and train an advanced GNN architecture and benchmark it against a baseline model (max-pooling MIL [10]) and state-of-the-art (SOTA) graph-based model (GCN for fibrosis [8]).
4. Evaluate using standard performance metrics, and cross-validation to ensure robustness and generalizability.
5. Synthesize the findings into key insights and implications and reflect on ethical challenges and future directions for the use of AI in hepatology.

In achieving this objective, this work introduces several contributions to the field of computational pathology and MASLD. First, it presents a novel and data-efficient vessel segmentation and classification pipeline for SR-stained WSIs,

providing a practical alternative to simplistic color-based methods or computationally expensive U-Nets. Second, it departs from conventional geometry-based spatial graphs by implementing a biologically informed graph construction pipeline that reflects the true histopathological progression of fibrosis. Third, to the author’s knowledge, this study represents the first application of GINE for liver fibrosis staging, which leverages both node and edge features to capture the structural relationship between vessels and collagen networks. Finally, through benchmarking against patch-based MIL and GNN SOTA model, this thesis provides evidence regarding the value of topological context ¹ in automated MASLD staging.

This study deliberately focuses on fibrosis staging in SR-stained WSIs and does not consider additional stains. The proposed framework models vessel-collagen structural relationships, as these are central to fibrosis progression and can be reliably extracted from SR-stained WSIs, while excluding other histopathological features (e.g., inflammatory infiltrates, hepatocellular ballooning) and clinical data. Furthermore, the approach prioritizes structural topology over cellular morphology. These choices were made to reduce complexity and ensure time and methodological feasibility.

The outline of this thesis is as follows. Chapter 2 reviews relevant literature, presenting current knowledge, methodologies, and gaps that this work aims to address. Chapter 3 describes the materials and methods employed. Chapter 4 presents the results, with a detailed analysis of the findings. Chapter 5 discusses the implications of the results in the context of existing research, addresses limitations, and suggests directions for future work. Finally, Chapter 6 concludes the thesis by summarizing the main contributions while placing it in the broader academic context.

¹In this thesis, the term “context” refers to the spatial and topological relationships between different structures in the liver tissue, e.g. how collagen fibers distribute and connect portal tracts and central veins, as oppose to its use in NLP/sequence models where context refers to information within an attention window.

2 Background

2.1 MASLD: Spectrum and Pathogenesis

MASLD, formerly known as non-alcoholic fatty liver disease, is a leading cause of chronic liver disease worldwide [11]. Its prevalence is estimated at 38% of the adult population and continues to rise with the global increase in obesity, metabolic syndrome, and type 2 diabetes [1]. By 2040, its prevalence in adults is expected to reach an alarming 55%, placing a significant burden on health-care systems [12].

MASLD is characterized by hepatic steatosis, defined as the accumulation of fat in more than 5% of liver tissue, in the presence of one or more cardiometabolic risk factor(s) (e.g. obesity, insulin resistance, type 2 diabetes, hypertension, etc.) and in the absence of significant alcohol consumption [13]. MASLD encompasses a broad spectrum of diseases, ranging from benign steatosis, known as metabolic dysfunction-associated steatotic liver (MASL), to metabolic dysfunction-associated steatohepatitis (MASH), a progressive necroinflammatory form of MASL that can advance to fibrosis, cirrhosis and, ultimately, HCC.

The pathogenesis of MASLD is multifactorial and results from complex interactions between genetic susceptibility, metabolic dysfunction, and environmental factors. As a result, clinical progression is highly variable and often unpredictable: 25% of patients progress to MASH and of those one-quarter progresses to advanced fibrosis within 8 to 10 years [14, 15]. Despite this variability, liver fibrosis consistently emerges as the strongest histological predictor of disease severity, long-term outcomes, and liver-related mortality [3]. This highlights the importance of accurately assessing and quantifying fibrotic tissue in clinical practice [6, 16, 17].

2.2 Liver fibrosis

As mentioned, fibrosis is an indicator of MASLD progression and is the strongest predictor of adverse clinical outcomes. Beyond its presence, the spatial distribution and organization of fibrotic tissue provide critical information for disease staging and prognosis [3]. Thus, it is essential to understand how fibrosis develops and manifests across the liver. This motivates a closer examination of its underlying mechanisms, structural patterns, and current diagnosis practices.

2.2.1 Mechanism and Topography

Liver fibrosis is defined as the pathological accumulation of extracellular matrix components, mainly collagen, in response to chronic liver injury and sustained inflammation [18]. Fibrosis progresses through distinct histological stages (ob-

servable structural changes under the microscope) commonly categorized from F0 to F4 in the METAVIR staging system [19]. This system reflects the extent of liver tissue remodeling, ranging from mild perisinusoidal or portal fibrosis (F1), where collagen accumulates around small blood vessels, to cirrhosis (F4), which is characterized by widespread scarring and loss of normal liver structure. Accordingly, METAVIR is an inherently semi-quantitative scoring tool, which relies on architectural patterns, spatial distribution, and collagen connections rather than the absolute volume or density of collagen to determine the fibrosis stage [20,21].

In MASLD, fibrosis emerges as a maladaptive wound-healing mechanism that disrupts the anatomy and architecture of the liver. In the first stage (F1), collagen begins to accumulate within the perisinusoidal spaces (region between hepatocytes and small capillary vessels) of the central lobule (area around the central vein of a liver lobule) or in the periportal areas (near the portal veins), keeping the portal and central veins (two different vasculature structures) anatomically isolated. As fibrosis progresses to F2, dense collagen networks form around the perisinusoidal and periportal areas and the first bridging septas (thick bands of fibrotic tissue that physically interconnect different hepatic zones, including portal-to-portal and portal-to-central vein links) start to extend into the hepatic parenchyma (functional tissue of the liver). In F3, numerous bridging septa form, advancing liver remodeling. These connections ultimately lead to cirrhosis (F4), where widespread structural changes disrupt the liver’s architecture, impede normal blood flow, and severely compromise its function [19, 22, 23]. Figure 2 illustrates the different fibrosis stages following the METAVIR standard in real liver biopsy samples.

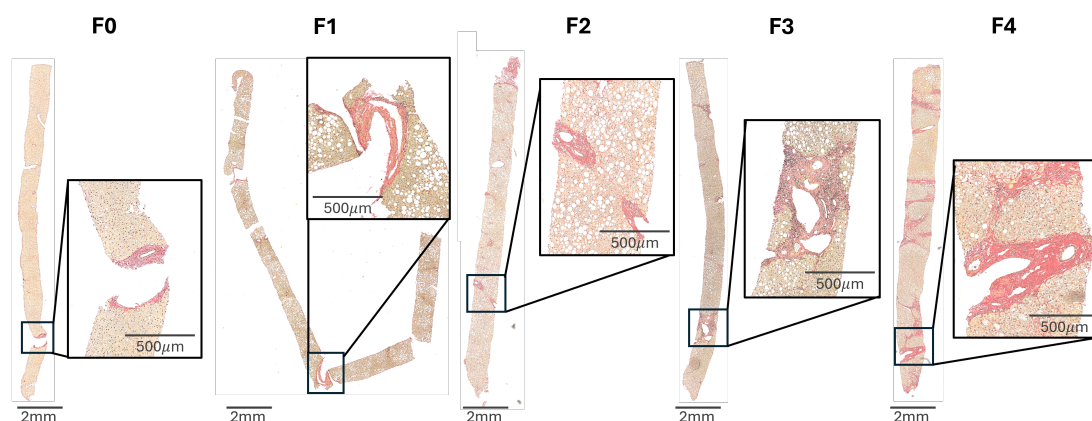


Figure 2: Liver fibrosis stages (METAVIR standard). Healthy liver (F0) shows little collagen (red stain) mostly present in portal tract walls. In the initial stages (F1-F2) a network of fine collagen strands starts appearing in the parenchyma. In the higher stages (F3-F4) the pathological fibers form “bridges” connecting portal tracts and central veins. All histological sections are stained with SR, showing color variability from the staining procedure.

Overall, fibrosis is the strongest predictor of disease severity, making histological staging essential for establishing prognosis, informing therapeutic strategies, and facilitating longitudinal patient monitoring [7].

2.2.2 Current Diagnosis and Gold Standard

In current clinical practice, the gold standard for staging fibrosis remains the manual evaluation of liver biopsies by expert pathologists, typically using specialized histochemical stains such as SR [2, 5]. SR selectively binds to collagen, making it appear red under light microscopy and enabling the semi-quantitative assessment of fibrotic deposition [24]. Nonetheless, despite its widespread use, manual histopathological assessment presents several limitations. The spatial heterogeneity of fibrosis and inflammation can lead to sampling bias and result in under- or over-staging of disease severity. Additionally, the assessment heavily relies on subjective visual interpretation, leading to considerable inter-observer variability [5]. This variability is further aggravated as the METAVIR staging system is a semi-quantitative tool that scores fibrosis based on visual patterns rather than precise measurements [21]. Moreover, manual examination limits reproducibility and places an increasing diagnostic burden on healthcare systems as disease prevalence continues to rise [5]. Furthermore, variations in staining intensity and color further complicate interpretation, hindering reliable, reproducible, and scalable quantification of fibrosis severity [25, 26]. Overall, these limitations highlight the need for more objective, efficient, and scalable assessment strategies.

2.3 Computational Histopathology and Related Work

The above limitations of manual histopathological assessment motivated the transition towards computer-aided diagnosis (CAD) systems, which aimed to support pathologists by providing objective, reproducible, and scalable quantitative assessments of liver fibrosis from WSIs [27]. WSIs are digitalized representations of tissue samples captured at high magnification, resulting in gigapixel-scale images that preserve fine morphological details. These high-resolution images enable pathologists to visualize fine-grained abnormalities with a level of detail comparable to traditional microscopy [28].

In the simplest and earliest CAD approaches, quantitative descriptors of collagen (e.g. collagen proportionate area (CPA), texture, length, thickness, orientation, spatial distribution, etc.) were extracted from WSIs and analyzed for fibrosis staging [29, 30]. For instance, in a previous master thesis, Y. Pan [31] proposed a pipeline for collagen segmentation based on color heuristic and clustering (illustrated in Figure 3) from which different collagen descriptors were obtained for fibrosis staging.

However, these approaches had limited capacity to capture complex tissue heterogeneity and higher-order spatial organization, affecting their ability to detect advanced fibrosis in which large-scale architectural remodeling of liver tissue occurs [32]. Consequently, more recent CAD approaches have focused on ML and DL methods to obtain a fully automated, reproducible, and scalable assessment of histological features and fibrosis staging [5, 6]. In practice, the most popular architectures are MIL and GNNs [8].

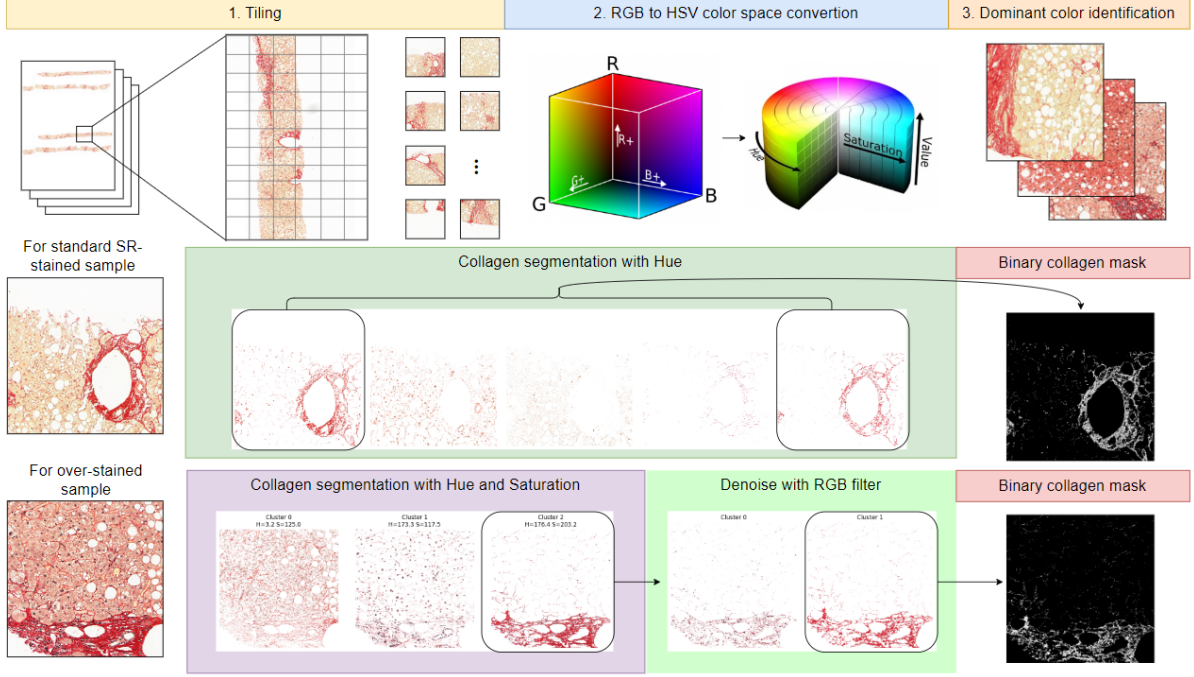


Figure 3: Automated collagen segmentation pipeline. (1) WSIs are cropped, (2) follow a RGB-to-HSV transformation and (3) the dominant tissue colors are identified via heuristic clustering to establish a baseline for segmentation. For standard SR-stained samples, Hue-based clustering isolates the collagenous signal to generate a binary collagen mask. For over-stained samples, it integrates both Hue and Saturation parameters removing dark color nuclei noise to generate a binary collagen mask. Figure borrowed with permission from [31].

2.3.1 Multiple Instance Learning for Liver Fibrosis

MIL has emerged as a popular weakly supervised framework for WSI analysis, particularly in settings where only slide-level labels are available. In this paradigm, each WSI is treated as a bag containing multiple patches or instances. While the bag is associated with a global label (e.g. fibrosis stage), the individual instances are unlabeled. Therefore, the learning objective is to infer slide-level predictions without explicit supervision at the instance-level [33].

Most MIL architectures follow a three-stage pipeline. First, an instance-level feature extractor $f(\cdot)$ encodes each image patch into a latent feature representation. Second, an aggregation or pooling function $\sigma(\cdot)$ combines the set of instance features into a single bag-level representation. Finally, a downstream classifier $g(\cdot)$ maps the aggregated representation to a slide-level prediction [34].

To exemplify, in its simplest, binary, and classical formulation by Dietterich et al. [35], a WSI bag $X = \{(x_1, y_1), \dots, (x_n, y_n)\}$, where y_i denotes the unknown latent label of instance x_i , is classified as positive if at least one instance within the bag exhibits pathological characteristics. This is formally expressed as,

$$\hat{Y} = \begin{cases} 0, & \text{if } \sum_{i=1}^n y_i = 0 \\ 1, & \text{otherwise} \end{cases} \quad (1)$$

Campanella et al. [10] operationalize this binary formulation via max-pooling, where the bag label is determined by the single most discriminative instance,

$$\hat{Y} = \arg \max_i p(y_i = 1 \mid x_i; \theta) \quad (2)$$

where θ denotes the learnable parameters of the instance-level classifier. This framework can be extended to multi-class settings by first assigning each instance a probability vector $p_i = p(y_i = k \mid x_i; \theta)_{k=1}^N \in \mathbb{R}^N$ over all N classes and then assigning the bag prediction the class of the globally most confident instance:

$$\hat{Y} = \arg \max_k \max_i p(y_i = k \mid x_i; \theta) \quad (3)$$

This is equivalent to selecting the instance-class pair (i^*, k^*) with the highest probability across all instances and classes simultaneously so that the final bag-level prediction is $\hat{Y} = k^*$ and i^* the selected representative instance. This max-pooling MIL approach has been adopted for multi-class grading problems using usually softmax over the instance logits as probability [10, 36]. However, while it is effective for tasks driven by rare but highly discriminative regions (e.g., tumor detection), max-pooling tends to be sensitive to noise and ignores contributions from other informative instances. To address this issue, attention-based MIL (AttenMIL) [37] replaces max-pooling with a learnable weighted aggregation of instance embeddings to integrate information from multiple diagnostically relevant regions. This formulation has demonstrated improved performance in settings characterized by heterogeneous and spatially distributed tissue patterns as in the case of fibrosis staging. Subsequent studies have further refined this attention-based paradigm, for instance ReMix [38] introduced a data augmentation strategy to increase instance-level diversity during training, leading to improved generalization performance in MIL-based models. In addition, Interventional Bag Multi-Instance Learning (IBMIL) [39] further enhanced feature representations by explicitly mitigating the influence of confounding factors, resulting in more robust and discriminative instance embeddings.

Nonetheless, despite their conceptual simplicity, clinical-grade performance, and widespread use, conventional patch-based MIL methods remain limited in their ability to capture spatial distribution and structural relationships between tissue components, which are critical for fibrosis staging with the METAVIR system where disease progression depends on both the presence and spatial distribution of pathological regions [7, 8]. Moreover, most of the MIL studies for fibrosis classification extract tile features with a ResNet pre-trained on ImageNet, which differ widely from WSIs, and commonly simplify the problem to binary operation, e.g. differentiating between early or late stage, or the presence or absence of cirrhosis, etc.

2.3.2 Graph Neural Networks for Liver Fibrosis

GNNs have also gained increasing attention in digital histopathology as they provide an effective and flexible framework to analyze data that can be represented as graphs. In a graph, each node represents a meaningful unit (e.g., a tissue patch), while edges capture relationships between nodes, such as spatial proximity or similarity [40, 41].

The construction of graph representations from WSIs is a critical step in GNN-based approaches, as the choice of nodes and edges directly influences the ability of the model to capture meaningful tissue structure. In practice, nodes are defined at different biological scales, giving rise to patch-graphs, cell-graphs, or tissue-graphs, depending on the chosen granularity [9]. Patch-graphs are the simplest approach, in which each image tile becomes a node, and edges connect spatially adjacent nodes [41]. While straightforward, it leads to very large graphs and might fail to capture long-range tissue interactions, such as collagen bridges connecting distant regions in advanced fibrosis. Other studies as the one proposed by M. Wojciechowska et al. [8] and XFibrosis [7] constructed graphs by first segmenting a region of interest (e.g. vessel or collagen), using them as nodes and connecting them based on spatial proximity. However, these approaches often rely on either data-intensive segmentation models, such as U-Net [8], which require large annotated datasets, or simpler color-based methods [7] that struggle to distinguish between visually similar structures, such as fat droplets and vascular regions. More importantly, proximity-based connections do not necessarily reflect the underlying biology of fibrosis progression. As mentioned, fibrotic remodeling occurs through collagen bridges that link vascular structures across the tissue. Consequently, purely geometric connections are unlikely to capture this true structural relationship, potentially limiting performance as hypothesized in [9].

These graph representations are then analyzed using a GNN. Most GNN architectures follow a recursive neighborhood aggregation (or message passing) scheme, in which each node iteratively aggregates information from its neighboring nodes to update its representation. Thus, in the first iteration or layer, a node’s embedding encodes its initial feature vector, $h_v^{(0)} = x_v$, and after k iterations it encodes both its own features and the structural information within its k -hop neighborhood [42, 43]. Formally, the k -th layer of a GNN can be expressed as:

$$a_v^{(k)} = \text{AGGREGATE}^{(k)}(\{h_u^{k-1} : u \in N(v)\}), \quad (4)$$

$$h_v^{(k)} = \text{COMBINE}^{(k)}(h_v^{(k-1)}, a_v^{(k)}) \quad (5)$$

where $a_v^{(k)}$ denotes the aggregated neighborhood information for node v at iteration k , $N(v)$ is the set of nodes adjacent to v , and $h_v^{(k)}$ is the updated feature vector (or embedding) of node v after the k -th iteration. To obtain a representation of the entire graph, node features from the last iteration are aggregated using graph-level pooling operations, such as summation or mean pooling [44], as follows:

$$h_G = READOUT(\{h_v^k \mid v \in G\}) \quad (6)$$

where h_G is the representation vector of the entire graph G . The choice of $AGGREGATE^{(k)}(\cdot)$, $COMBINE^{(k)}(\cdot)$ and $READOUT(\cdot)$ is also critical.

Different designs have been proposed to capture different types of local structure and interactions [45]. The most widely used architectures for liver fibrosis are Graph Convolutional Networks (GCNs) and Graph Attention Networks (GATs) [46, 47]. GCNs perform neighborhood aggregation by applying an element-wise mean pooling followed by a non-linear transformation, resulting in a computationally efficient and smooth averaging of neighboring features [46]. GCNs have achieved strong performance on fibrosis staging, but struggle to distinguish nodes with similar neighborhood statistics due to their limited expressive power. GATs introduce an attention mechanism that learns adaptive, data-dependent weights for neighboring nodes. By assigning different levels of importance to different neighbors, GATs can better capture heterogeneous local interactions and are well-suited for graphs in which neighboring nodes contribute unequally to a node's representation [47].

Nonetheless, there exist novel and more powerful architectures such as Graph Isomorphism Networks (GINs) [45] and GINes [48], which have not yet been fully studied for fibrosis staging. GIN use summation as a combination function followed by a multilayer perceptron (MLP), which is a feedforward neural network with single or multiple fully connected hidden layers and non-linear activation layers [49], to distinguish different graph structures more effectively [50].

$$h_v^{(k)} = MLP^{(k)} \left((1 + \epsilon^{(k)}) \cdot h_v^{(k-1)} + \sum_{u \in N(v)} \text{ReLU}(h_u^{(k-1)}) \right) \quad (7)$$

In the above formula, ϵ is a learnable or fixed value that controls the relative importance of a node's own features compared to its neighbors' features, and $ReLU$ is a non-linear activation function. Similarly, GINE [48] further extend GIN's framework by explicitly incorporating edge attributes into the message-passing process as:

$$h_v^{(k)} = MLP^{(k)} \left((1 + \epsilon^{(k)}) \cdot h_v^{(k-1)} + \sum_{u \in N(v)} \text{ReLU}(h_u^{(k-1)} + e_{u,v}) \right) \quad (8)$$

where $e_{u,v}$ is the edge's feature vector connecting the nodes u and v . This extension is particularly relevant in applications where relationships between nodes carry semantic meaning, such as spatial distances or interaction types, allowing the model to jointly consider node features, edge features, and graph topology.

Overall, motivated by the limitations of modern DL approaches, this work explores a biologically informed graph-based representation of WSIs for fibrosis staging. Rather than constructing graphs solely from geometric proximity, the

proposed approach exploits domain knowledge of fibrosis progression by explicitly modeling the structural relationship between vessels and collagen deposition. Specifically, segmentation-derived masks of vessels and collagen are used to construct vessel-centered graphs in which collagen fibers define the connectivity between vascular structures. This representation aims to capture fibrotic bridging patterns that characterize disease progression and to enable a more advanced GNN, namely GINE, to learn from biologically meaningful tissue topology.

3 Methodology

An overview of the methodology of this thesis is shown in Figure 4. The pipeline begins with WSIs of liver tissue, which undergo vessel and collagen segmentation to construct a vessel-collagen centered graph. This graph data is then fed into a tuned GINE model. The model’s performance for fibrosis staging is evaluated in different studies and benchmarked against baseline and SOTA models.

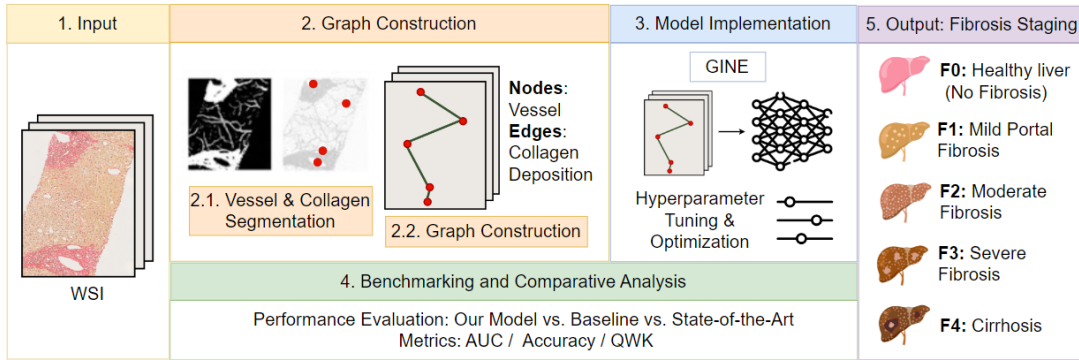


Figure 4: Overview of the methodology of this project. (1) WSIs serve as the primary inputs. (2) Graph Construction, involving vessel and collagen segmentation to define nodes (vessels) and edges (collagen deposition). (3) Model Implementation, involving hyperparameter tuning for the GINE architecture. (4) Benchmarking and Comparative Analysis, evaluating performance against baseline and SOTA models using area under the curve (AUC), accuracy, and quadratic weighted kappa (QWK) metrics for (5) fibrosis classification.

3.1 Data collection and Preprocessing

A total of 120 WSIs from MASLD patients were provided by the University Hospital Basel, Switzerland. All liver tissue sections were stained with SR and scanned at 40x magnification. The slides were graded according to the METAVIR fibrosis staging system (F0–F4) by two senior pathologists, with consensus reached in cases of disagreement. The dataset consisted of 6 WSIs classified as F0, 29 as F1, 39 as F2, 19 as F3, and 27 as F4.

WSIs were divided into non-overlapping 2048×2048 pixel tiles. Tiles with less than 5% tissue area, previously identified via Hue-Saturation-Value (HSV) thresholding ($S > 30$, $V < 230$), were discarded to remove background-dominated regions. Edge tiles were zero-padded to maintain consistent dimensions and avoid computational bias.

3.2 Proposed Framework: from Graph Construction to Deep Learning Model Implementation

This section describes the proposed framework, from vessel and collagen segmentation to graph representation and GNN-based training.

3.2.1 Collagen and Vessel Mask Construction

A dedicated pipeline was implemented to segment vessel and collagen microstructures. These masks are essential for isolating the specific biological structures required for later graph representations.

Collagen detection was performed using the pipeline described in [31]. For vessel segmentation, a custom pipeline was developed to segment vessels and classify them into portal tracts and central veins (Figure 5). These strategies were designed to determine whether incorporating vessel-type information improves staging performance.

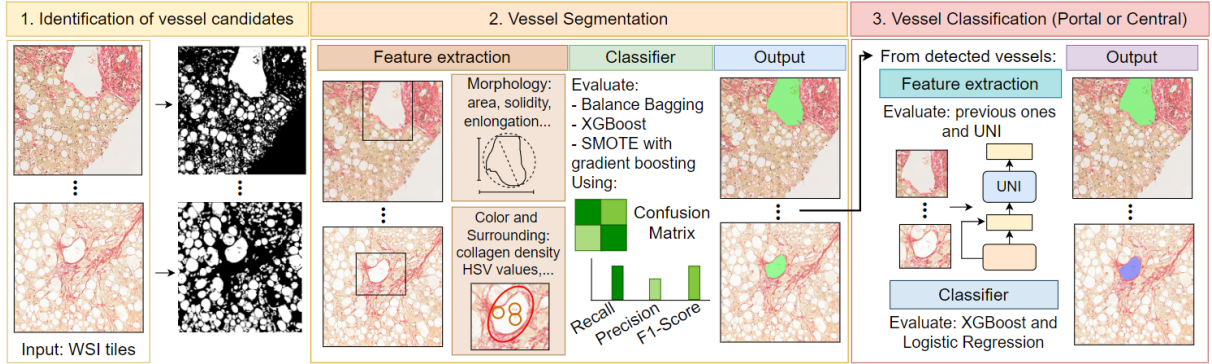


Figure 5: Automated vessel segmentation and classification pipeline. (1) Identification: Vessel candidates are detected as holes within WSI tissue tiles. (2) Segmentation: Morphological, color, and contextual features are extracted to train ML classifiers (Balanced Bagging, XGBoost, and SMOTE-Gradient Boosting) for vessel segmentation. The optimal model is selected via F1-score and confusion matrix metrics to yield segmented vessels (green). (3) Classification: Segmented vessels are further classified into portal tracts (green) or central veins (blue) by testing the engineered features and UNI embeddings with XGBoost or Logistic Regression.

In tissue tiles, lumens (vessels, lipid droplets, and artifacts) appear as white, closed contours against the pink and red stained tissue matrix. To isolate these candidates, morphological operations and Otsu’s thresholding were applied to the tissue mask. Artifacts smaller than $50 \mu m^2$ (approximate size of a red blood cell [51]) were also removed.

For (2), i.e. vessel segmentation, 24 morphological (e.g., area, circularity, elongation) and color-based features, previously proposed in [30, 31, 52], were extracted to characterize each contour and its surrounding microenvironment. These 24 features per contour were fed as a single row to different ML classifiers (Balanced Bagging, Gradient Boosting with SMOTE (GB-SMOTE), and XGBoost), meaning each contour detected in the 2048×2048 tile was represented as one feature vector, and the classifier predicted whether that contour

was a vessel or not. These ML models were previously trained on 200 annotated WSI tiles using 5-fold grouped cross-validation for hyperparameter tuning. Additional information about the ML classifiers can be found at the Appendix B. Model performance was evaluated with the confusion matrix, precision, recall, and F1-score.

For (3), i.e. vessel-type classification, the detected vessels in the previous step were further categorized as portal or central veins. To do so, for each vessel new features were extracted using either the above 24 hand-crafted features or embeddings from the UNI foundation model, a pretrained vision transformer encoder for histopathology [53]. In particular, for the UNI-based features, a 224×224 pixel patch was cropped from the original SR-stained tile centered on each detected vessel, with a context factor of $1.5 \times$ the bounding box size to include surrounding tissue. The patch followed UNI’s pre-processing protocol, and passed through the frozen UNI ViT-L/16 encoder to produce a 1024-dimensional embedding per vessel. Eight additional shape features were added, producing a 1032-dimensional input vector per vessel for the downstream classifier. Both feature representations were used to train XGBoost and Logistic Regression classifiers, with hyperparameter tuning performed using cross-validation. The performance of the feature extraction and classifier models was evaluated as in (2) and compared against a baseline approach, which directly segmented portal and central veins from scratch.

3.2.2 Graph Construction

The vessels and collagen fibers were then transformed into a graph representation as shown in Figure 6. This graph representation, including all intermediate steps, was established through a multidisciplinary consensus between an AI expert and a hepatologist. From a clinical and biological perspective, collagen fibers physically extend and interconnect vascular structures during fibrosis. Thus, mapping vessels as nodes and collagen fibers as edges naturally capture these pathological tissue bridges. Moreover, two graph types were created: a single-class that ignores vessel type, and a two-class that later includes vessel-type information in the nodes and edges.

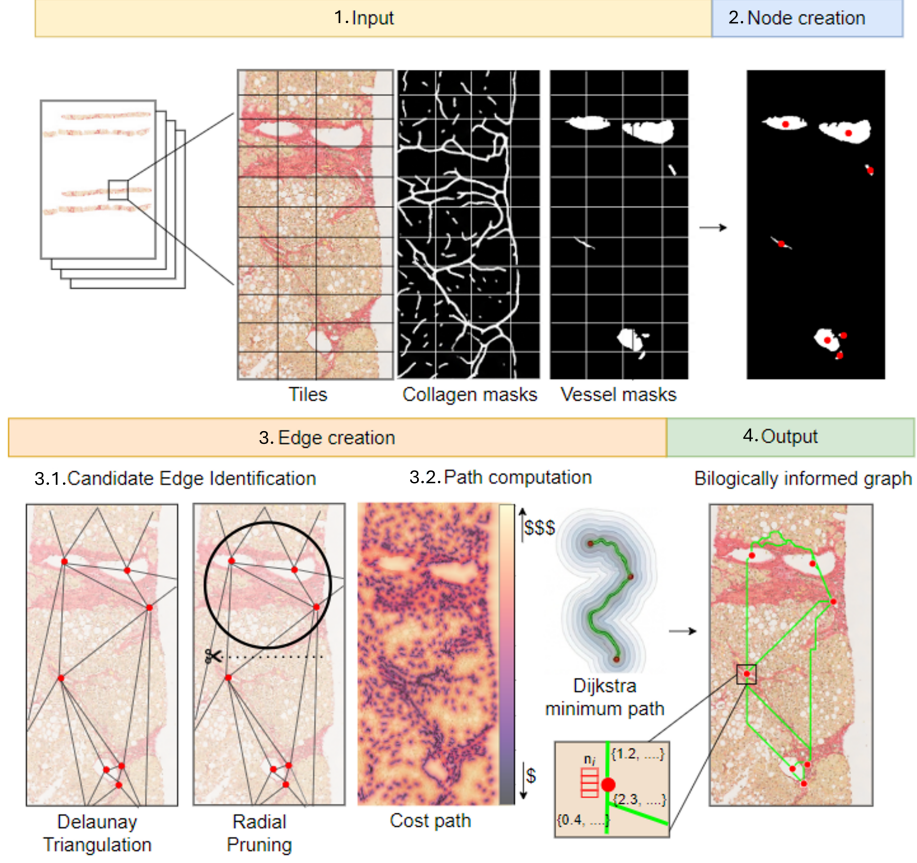


Figure 6: Graph construction pipeline. (1) Input, tiles of WSIs with their collagen and vessel masks. (2) Node creation, nodes are generated by detecting the centroids of individual vessels. (3) Edge creation, (3.1) Candidate neighboring nodes are identified using Delaunay triangulation, and distant connections are removed through radial pruning. (3.2) A cost map/path is computed in which movement outside collagen regions is penalized (\$ indicates cost of pixel). The least-cost path between nodes is then determined using Dijkstra’s algorithm. (4) Output, final graph with features in node and edges (see next Section).

First, nodes were defined as the centroids of the vessels, obtained from the previous segmentation. Vessels that were artificially split due to tiling were merged to avoid duplicate representations; in the two-class vessel graphs merging was restricted to vessels of the same type. Candidate edges were subsequently identified using Delaunay triangulation and pruned using a radial distance threshold to prioritize local anatomical adjacency and remove implausibly long connections. For each remaining candidate edge, the optimal biological path was estimated using Dijkstra’s algorithm [54] on a spatial cost map. Specifically, the traversal cost at pixel x was defined as:

$$C(x) = 1 + \alpha(1 - e^{-\beta d(x)}) \quad (9)$$

where α determined the maximum penalty, β controlled the exponential decay rate and $d(x)$ denoted the Euclidean distance to the nearest collagen pixel. The cost function parameters, $\alpha=150$ and $\beta=0.005$ (determined empirically), encouraged paths to follow collagen-dense regions by exponentially penalizing

distance from collagen pixels. To ensure computational stability, vessel centroids were projected onto the nearest collagen skeleton for path initialization. This projection facilitated plausible starting points without altering the underlying vessel coordinates, mitigating artificial penalties from minor segmentation offsets or bifurcations. Background pixels were assigned infinite cost to constrain paths within the biopsy area.

The single and two-class graphs were evaluated against known biological properties to assess whether the proposed construction captured fibrosis organization patterns and provided reliable inputs for downstream learning.

3.2.3 Graph Embedding

Node and edge embeddings were constructed to capture both morphological and structural information (Table 1 and Table 2). Node morphological features were generated by extracting embeddings using UNI [53]. In particular, for each vessel, a 224×224 RGB patch was extracted directly from the SR-stained WSI tile at the vessel centroid coordinates, each patch followed the original pre-processing steps of UNI and passed through the frozen UNI encoder. The output was a 1024-dimensional feature vector per vessel, capturing morphological characteristics of the vessel and its immediate tissue microenvironment, including collagen texture, lumen shape, and surrounding cellular organization. To preserve global spatial context, these features were integrated with positional encodings based on a vision transformer formulation [7]. In the two-class graph representation, vessel-type information was also added. Moreover, edge embeddings were designed to characterize the collagen network through a multi-dimensional feature vector. These features include geodesic resistance (defined as the edge cost normalized by the Euclidean distance), straightness, mean and standard deviation of fiber thickness, local collagen density, tissue fraction, edge length, and orientation. In the two-class graph representation, edge features also included the types of vessels connected by each edge.

Table 1: Description of Node Features

Feature	Description	Dim.
UNI embedding	Patch-level morphological representation around the vessel centroid extracted by UNI foundation model	1024
Positional encoding	Normalised (x, y) coordinates of the vessel centroid within the WSI	2
Vessel type (two-class only)	Encoding indicating portal vein or central vein	1

Table 2: Description of Edge Features

Feature	Description	Dim.
Biological encoding	Geodesic resistance: total path cost normalised by Euclidean distance (low values = collagen-rich paths). Straightness ratio: ratio of Euclidean distance to path length (values close to 1 = direct collagen path). Mean fiber thickness: mean diameter of the collagen along the path, estimated from distance transform. Std fiber thickness: standard deviation of fiber thickness along the path, reflecting structural heterogeneity. Local collagen density: fraction of pixels around the path that are collagen. Tissue fraction: fraction of the path traversing non-collagen tissue. Orientation entropy: circular variance of path direction angles (high values = tortuous, disorganised path). Edge length: Euclidean distance in μm .	8
Edge type (two-class only)	Encoding of the vessel-type pair connected: portal-portal, portal-central, or central-central	3

Overall, both graphs are undirected, weighted, and simple. Due to the Delaunay-based construction, cycles are present in the graph structure. Connectivity is not guaranteed across all WSIs, as vessel pairs with no viable collagen path are excluded.

3.2.4 GNN Architecture and Training Procedure

All graph features were processed using a GINE. Model evaluation and tuning were performed using cross-validation framework with a held-out test set (15%) and a training set (85%). Within the training set, a 5-fold stratified cross-validation was applied for hyperparameter tuning and architectural selection, where in each fold 80% of the training data was used for training and the remaining 20% for validation.

Within this framework, the hyperparameter search space, summarized in Table 3, was systematically explored. Training-related parameters included the learning rate and learning rate schedulers, optimized via different optimizers with early stopping. Different loss functions, cross-entropy (CE) and focal loss, were evaluated alongside their combinations with an ordinal regularization term (that accounted for the ordered nature of fibrosis) and weight decay. The combined loss function were defined as:

$$L = L_{ce} + \lambda L_{reg} + \lambda_d L_2 \quad \text{or} \quad L = L_f + \lambda L_{reg} + \lambda_d L_2 \quad (10)$$

where L_{ce} is the CE loss, L_f is the focal loss, L_{reg} is the ordinal regularization

term, λ controls the contribution of L_{reg} , and L_2 being the weight decay term.

Furthermore, key architectural components of the GINE model were investigated, including the number of layers, the use of Jumping Knowledge network (JK-Net) (a scheme to combine information from multiple graph layers [43]), dropout rate, hidden dimensionality, and aggregation strategy, to assess their impact on model capacity and performance. The GINE architecture was also compared with other commonly used GNNs in fibrosis analysis (GAT and GCN) and GIN to assess the contribution of edge features. These baseline models were tuned over a reduced hyperparameter search space to ensure fair comparison. Moreover, the single, two-class, and spatial-based KNN graphs were used as inputs to the GINE model to further assess the graph construction process for fibrosis staging.

Table 3: Hyperparameter Search Space for GINE Model Optimization

Hyperparameter	Search Values
Dropout Rate	0, 0.3, 0.5
Hidden Dimensions	128, 256
Readout (Pooling)	Sum, Mean, Max, Attention
Learning Rate (LR)	10^{-5} , 5×10^{-5} , 10^{-4} , 10^{-3}
LR Scheduler	None, ReduceLROnPlateau, CosineAnnealing
Loss Function Types	CE, Focal, CE + ordinal loss, Focal + ordinal loss
Number of GNN Layers	3, 5, 6 (with and without JK-Net)
Optimizer	Adam, AdamW
Ordinal Weight (λ)	0.0, 0.005, 0.05, 0.1
Weight Decay	10^{-3} , 10^{-4} , 10^{-5}

3.3 Baseline and State-of-the-Art Models Implementation

The proposed model was evaluated against a baseline and SOTA model. For the baseline model, the max-pooling MIL approach proposed by Campanella et al. [10] was selected due to its widespread adoption in digital pathology. To align with this study, its original binary formulation was extended to a multiclass setting [36]. Moreover, the WSIs were divided into non-overlapping 512×512 pixel tiles to account for the different magnification (40x in this study vs 20x in the original paper).

For the SOTA model, the GCN framework proposed by Wojciechowska et al. [8] was chosen as it provides a relevant benchmark for evaluating the proposed graph construction strategy and feature representation. The graph construction pipeline and model architecture were preserved, but the collagen segmentation step was replaced with [31] due to the unavailability of the original U-Net. As in the baseline, WSIs were tiled into non-overlapping 512×512 patches.

Both models were evaluated following the recommended hyperparameters and without performing any tuning.

3.4 Evaluation metrics

Following [7, 8], accuracy and area under the curve (AUC) were used to evaluate the DL models. To compute AUC in the multiclass setting, the problem was decomposed into multiple binary classification tasks (e.g. F0 vs. the rest), allowing the assessment of the model’s ability to discriminate each stage. The macro-average AUC was also reported to provide an overall measure of performance. Additionally, the quadratic weighted kappa (QWK) was included to account for the ordinal nature of fibrosis stages, giving larger penalties to more severe misclassifications.

4 Results and Analysis

4.1 Graph Construction Framework Results

This section quantitatively and qualitatively evaluates the vessel segmentation pipeline and resulting graph representation.

4.1.1 Performance of Vessel Segmentation

Table 4 summarizes the performance of the evaluated classifiers for vessel segmentation (step 2 in Figure 5), with representative predictions shown in Figure 7. Balanced Bagging achieved the highest recall (0.83) at the cost of a high false-positive rate, as shown in its low precision (0.45) and high variance (over 0.30 in all metrics). GB-SMOTE provided a more balanced result but it was ultimately outperformed by XGBoost, which achieved the highest F1-score (0.69 ± 0.17).

Table 4: Performance of the evaluated classifiers for vessel segmentation.

Method	F1-score	Precision	Recall
Balanced Bagging	0.42 ± 0.37	0.45 ± 0.38	0.83 ± 0.33
XGBoost	0.69 ± 0.17	0.73 ± 0.25	0.70 ± 0.05
GB-SMOTE	0.66 ± 0.19	0.69 ± 0.25	0.69 ± 0.07

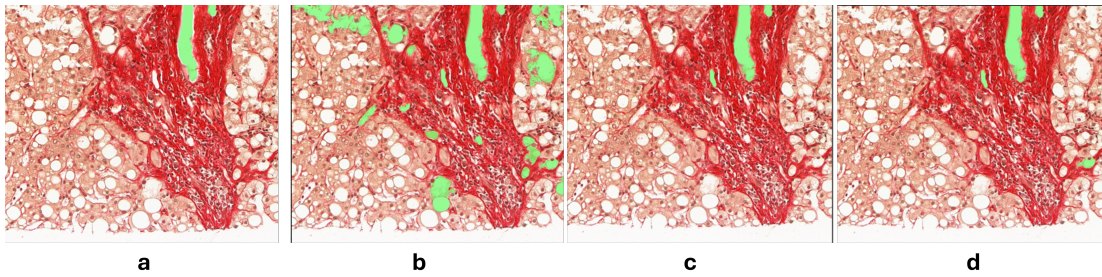


Figure 7: Representative figure of vessel detection. (a) Ground truth vessels overlaid on the tissue tile (b) Detected vessels by Balance Bagging (c) Detected vessels by XGBoost and (d) Detected vessels by GB-SMOTE.

Table 5 summarizes the performance of the different approaches for vessel-type classification (step 3 in Figure 5), while Figure 8 shows a representative prediction from the best-performing approach. The baseline approach, which directly segmented portal tracts and central veins from scratch, achieved very limited performance. The remaining approaches that relied on the vessel segmentation

output for subsequent classification demonstrated a clear advantage. Hand-crafted morphological and color features combined with XGBoost showed improved results except for the minority central vein class (F1-score of 0.68 ± 0.21). In contrast, using UNI embeddings significantly enhanced discriminative power across both classes. While UNI combined with XGBoost showed strong results, UNI with Logistic Regression emerged as the best-performing configuration (F1-score of 0.91 ± 0.03 for portal and 0.80 ± 0.11 for central).

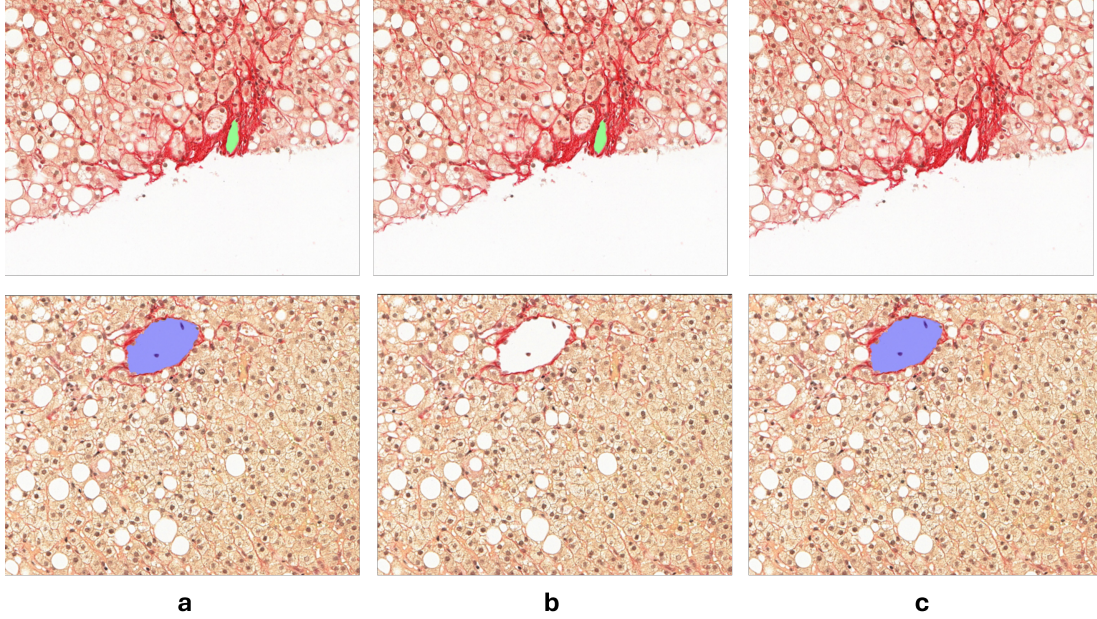


Figure 8: Representative figure of vessel classification using UNI features embeddings and Logistic Regression. (a) Ground truth vessels overlaid on the tissue tile showing in green portal veins and in purple central veins (b) Detected portal veins (c) Detected central veins.

Table 5: Performance of vessel classification (portal vs central). Performance metrics are reported as mean $\pm$ standard deviation

Approach	Class	F1-score	Precision	Recall
Baseline	Portal	0.63 ± 0.05	0.64 ± 0.07	0.64 ± 0.04
	Central	0.52 ± 0.11	0.59 ± 0.14	0.55 ± 0.12
Hand-crafted Features + XGBoost	Portal	0.82 ± 0.08	0.92 ± 0.05	0.86 ± 0.05
	Central	0.68 ± 0.21	0.48 ± 0.20	0.54 ± 0.17
UNI + XGBoost	Portal	0.88 ± 0.06	0.93 ± 0.03	0.84 ± 0.12
	Central	0.74 ± 0.07	0.81 ± 0.12	0.73 ± 0.15
UNI + Logistic Regression	Portal	0.91 ± 0.03	0.92 ± 0.05	0.91 ± 0.06
	Central	0.80 ± 0.11	0.77 ± 0.18	0.77 ± 0.10

4.1.2 Graph interpretation on WSIs

With the collagen and vessel masks, biologically informed and topology-aware graphs were constructed as illustrated in Figure 9.

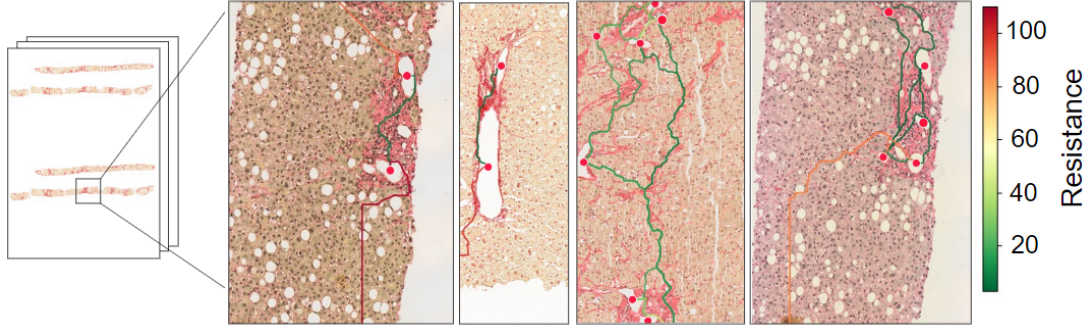


Figure 9: Representative example of the generated biologically-informed topology-aware graphs. Figure shows zoomed-in regions from different WSIs. Vessel centroids are represented as red nodes. The color of the edges indicates the resistance of the computed minimum-cost paths, reflecting the underlying collagen structure.

The resulting graphs varied in size depending on the tile cellularity (density or total number of vessels) and fibrosis stage. Across the dataset, the single-class and two-class graphs contained on average 79 ± 53 nodes and 116 ± 110 edges per WSI. Note that the graph construction was executed on the UPPMAX Bianca cluster using 8 CPU cores and 120 GB of RAM. Graph construction for a single WSI took approximately 15 minutes on average.

To evaluate whether the proposed graph representations, both single and two-class, capture meaningful tissue organization, graph-derived features (edge connectivity, geodesic resistance, collagen density, and clustering coefficient) were analyzed across fibrosis stages. Figure 10 illustrates the results, with the first row corresponding to the single-class graph and the second row to the two-class with vessel type differentiation.

Non-parametric statistical tests (Mann-Whitney U test [55] and Kruskal-Wallis test [56]) revealed significant differences across stages. In both graph representations, the mean edge resistance decreased significantly while the number of edges per node, the mean edge collagen density, and clustering coefficient increased significantly with fibrosis severity.

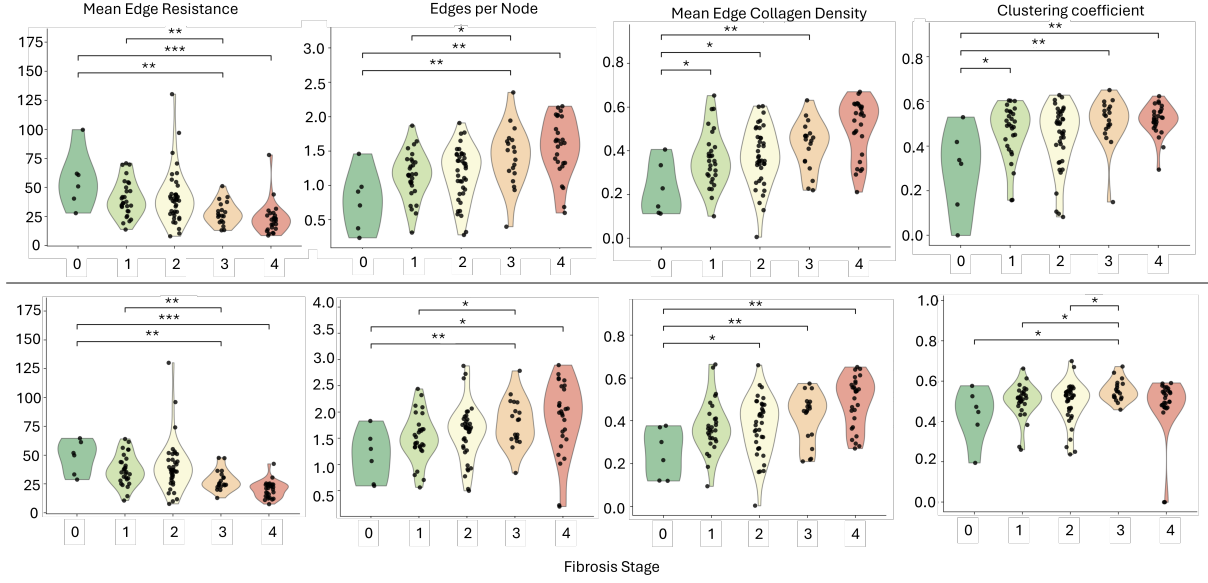


Figure 10: Distribution of topological and biological features across fibrosis stages (x-axis) for the (top) single-class and (bottom) two-class with vessel-type differentiation. From left to right: Mean Edge Resistance, Edges per node, Mean Edge Collagen Density and Clustering Coefficient. Statistical significance was determined via Mann-Whitney U test and Kruskal-Wallis test with post-hoc comparisons indicated by asterisks ($p < 0.05$, * $p < 0.01$, ** $p < 0.001$, *** $p < 0.0001$).

4.2 GNN implementation Results

This section summarizes the key trends observed during hyperparameter tuning, architectural evaluation, and results from different comparative studies.

4.2.1 Hyperparameter Optimization and Architectural Design Evaluation

Architectural and hyperparameter choices significantly impacted model performance as shown in Figure 11. GINE training was executed on the UPPMAX Bianca cluster using 8 CPU cores, 120 GB of RAM, and an NVIDIA A100 GPU, with total training completing in under an hour on average.

Sum pooling consistently outperformed other pooling operations, and deep architectures with six layers using JK-Net scheme achieved better performance. From an optimization perspective, the Adam optimizer with cosine annealing learning rate scheduling produced the most stable convergence behaviour within approximately 40 epochs (triggered by early stopping). Lower learning rates ($5e-5$, $1e-4$) consistently outperformed other values, with a 0.3 dropout rate and lower weight decays ($1e-5$) providing the best balance for regularization. Finally, the combined focal and ordinal regression loss ($\lambda = 0.05$) function outperformed individual loss components.

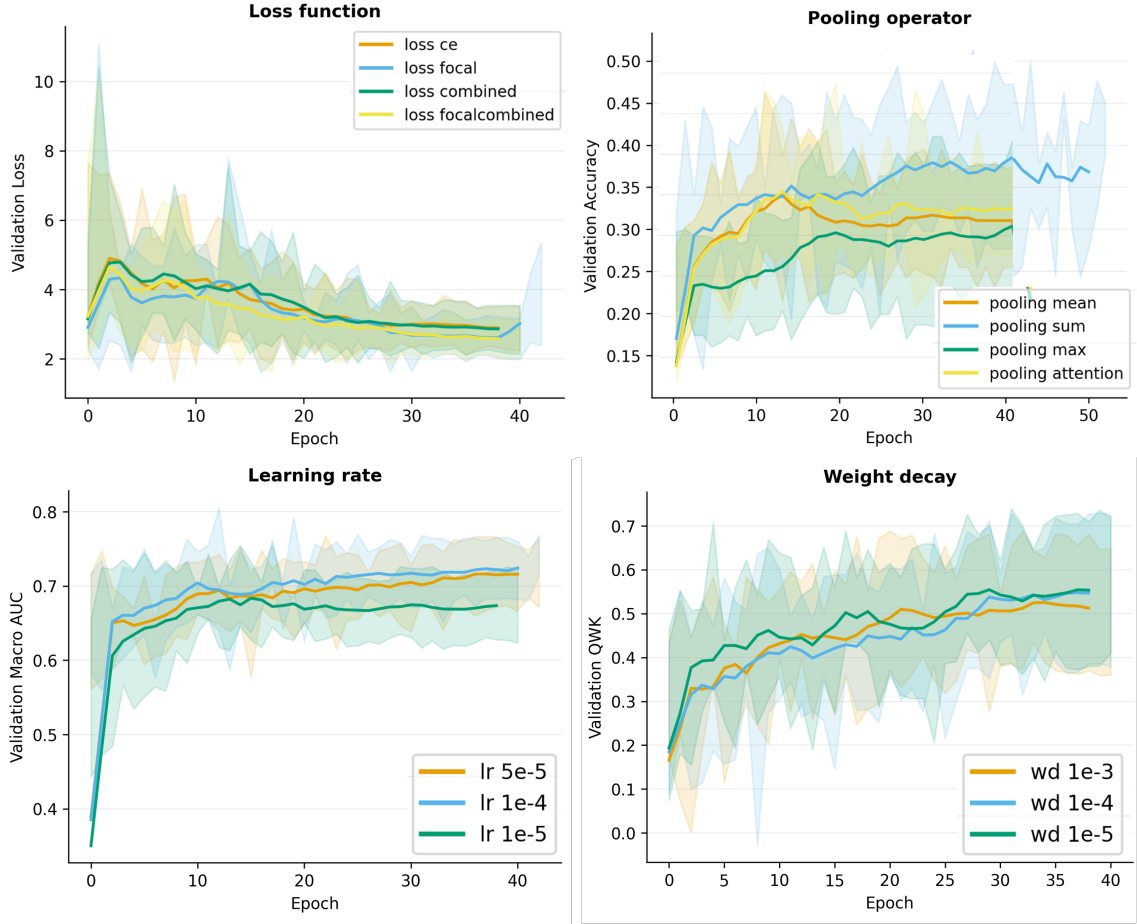


Figure 11: Representative hyperparameter optimization and architectural comparative results. Curves illustrate model training dynamics across different configurations. Solid lines depict the cross-validation mean performance and shaded bands show variance boundaries. The panels show illustrative examples from a broader configuration study. (Top-left) validation loss across different loss functions, (top-right) validation accuracy across graph pooling operators, (bottom-left) validation macro AUC across base learning rates, (bottom-right) validation QWK across weight decay regularization constraints (bottom-right).

4.2.2 Comparative Study on Graph Construction

A comparative study compared the three graph representations for staging. The goal was to evaluate the benefit of adding biological information and distinguishing between vessel types. The results are shown in Table 6.

Table 6: Effectiveness of the proposed graph construction pipeline (including single-class, i.e. no vessel-type differentiation, and two-class, i.e. with vessel-type differentiation) against a baseline proximity-based graph. Performance metrics are reported as mean $\pm$ standard deviation across folds.

Method	AUC ₀	AUC ₁	AUC ₂	AUC ₃	AUC ₄	AUC	Accuracy	QWK
KNN	0.87 ± 0.12	0.60 ± 0.11	0.60 ± 0.10	0.72 ± 0.08	0.69 ± 0.19	0.69 ± 0.05	0.24 ± 0.12	0.32 ± 0.17
Single-class	0.88 ± 0.12	0.71 $\pm$ 0.13	0.66 ± 0.07	0.75 $\pm$ 0.15	0.85 ± 0.10	0.77 ± 0.07	0.37 ± 0.17	0.60 ± 0.14
Two-class	0.88 $\pm$ 0.10	0.66 ± 0.09	0.70 $\pm$ 0.09	0.71 ± 0.16	0.96 $\pm$ 0.03	0.78 $\pm$ 0.04	0.48 $\pm$ 0.05	0.68 $\pm$ 0.10

Overall, the KNN-based baseline underperformed compared to the biologically informed graph constructions, achieving lower macro AUC (0.69 ± 0.05), accuracy (0.24 ± 0.12), and QWK (0.32 ± 0.17). Among the proposed representations, the two-class graph demonstrated a slight performance advantage, achieving higher accuracy (0.48 ± 0.05), QWK (0.68 ± 0.10) and macro AUC (0.78 ± 0.04). Furthermore, the two-class approach exhibited greater stability, as shown by the reduced standard deviation across individual stages compared to the single-class representation.

4.2.3 Comparative study on different GNN architectures

The GINE architecture was also compared against GAT, GCN and GIN. As shown in Table 7, GINE achieved best overall performance with the proposed graph representation. However, the performance differences between GCN, GAT, and GIN were modest and likely within the margin of variability given the small dataset size.

Table 7: Effectiveness of the proposed GNN architecture. Performance metrics are reported as mean $\pm$ standard deviation across folds.

GNN	AUC ₀	AUC ₁	AUC ₂	AUC ₃	AUC ₄	AUC	Accuracy	QWK
GCN	0.89 ± 0.13	0.64 ± 0.11	0.61 ± 0.06	0.63 ± 0.18	0.76 ± 0.08	0.70 ± 0.07	0.28 ± 0.05	0.33 ± 0.12
GAT	0.86 ± 0.12	0.63 ± 0.13	0.60 ± 0.10	0.67 ± 0.13	0.77 ± 0.11	0.71 ± 0.07	0.25 ± 0.02	0.38 ± 0.18
GIN	0.88 ± 0.07	0.68 ± 0.13	0.59 ± 0.05	0.59 ± 0.08	0.75 ± 0.10	0.69 ± 0.05	0.23 ± 0.05	0.36 ± 0.15
GINE	0.89 $\pm$ 0.15	0.71 $\pm$ 0.12	0.67 $\pm$ 0.07	0.77 $\pm$ 0.05	0.89 $\pm$ 0.10	0.78 $\pm$ 0.06	0.47 $\pm$ 0.13	0.60 $\pm$ 0.17

4.3 Comparison With Baseline and State-of-the-Art Models for fibrosis staging

The fully optimized graph construction pipeline and GINE model were evaluated against both a baseline and a SOTA model for benchmarking purposes. Table 8 summarizes the results.

Table 8: Fibrosis staging performance of the different DL methods on the test set.

Method	AUC ₀	AUC ₁	AUC ₂	AUC ₃	AUC ₄	AUC	Accuracy	QWK
MIL	0.47	0.62	0.53	0.52	0.85	0.60	0.21	0.06
GCN	0.79	0.71	0.37	0.65	0.88	0.68	0.26	0.28
Ours	0.65	0.67	0.55	0.56	0.93	0.67	0.37	0.44

The baseline MIL model showed limited performance (macro AUC 0.60, Accuracy 0.21, and QWK 0.06), with strong discrimination only at advanced stages (AUC₄ = 0.85). The GCN-based model improved overall metrics (AUC 0.68, QWK 0.28, accuracy 0.26), effectively identifying early (AUC₀ = 0.79) and late stages (AUC₄ = 0.88), though it struggled significantly with middle stages as F2 (AUC₂ = 0.37). In contrast, the proposed approach achieved the highest accuracy (0.37) and QWK (0.44), while matching the SOTA macro AUC (0.67) and demonstrating consistent performance across all fibrosis stages.

5 Discussion

5.1 Graph Construction Pipeline

This section evaluates the integrated framework, tracing the pipeline from vessel segmentation through graph representation to GINE tuning and fibrosis staging.

5.1.1 Vessel-Collagen Segmentation

For vessel segmentation, XGBoost outperformed Balanced Bagging and GB-SMOTE. The weaker performance of Balanced Bagging likely stems from its random undersampling of the majority class, which prevents the model from capturing the morphological variability of the background, leading to increasing false positives as suggested in [57]. Similarly, SMOTE’s synthetic sample generation of the minority class could struggle to accurately represent the complex vessel patterns, reducing overall performance [58]. In contrast, XGBoost’s iterative focus on misclassified samples allowed it to emphasize rare vessel instances while maintaining a robust representation of the background [59]. However, its final F1 score of 0.69 remains non-ideal, reflecting the inherent difficulty of class imbalances and distinguishing vessels from complex background tissue.

For vessel-type classification, a cascade approach outperformed direct segmentation. While hand-crafted features sufficed for detection, they lacked the discriminative depth to distinguish portal from central veins. UNI embeddings addressed this by capturing complex contextual cues from the vessel architecture [53]. Logistic Regression’s better performance over XGBoost suggests that UNI projects features into a linearly separable space where complex ensembles may overfit. However, the system remains susceptible to downstream error propagation, as false positives of the previous vessel segmentation are inevitably assigned a vessel-type category.

5.1.2 Graph interpretation on WSIs

The second stage of the pipeline constructs graphs where vessels act as nodes and collagen structures define the edges, effectively capturing structural patterns associated with fibrosis progression. Regardless of vessel-type differentiation, both graph versions show consistent trends across fibrosis stages (e.g., decreased edge resistance, increased connectivity, and higher collagen density), reflecting the biological transition from a sparse isolated network to a dense, interconnected fibrotic network in cirrhosis. Moreover, the significant differences in feature values across stages suggest that these graph descriptors pos-

sess discriminative power to distinguish early from advanced fibrosis.

Importantly, the proposed graph constructions outperformed traditional KNN-based methods, highlighting the limitations of purely geometric approaches. Since KNN graphs rely on spatial proximity, they fail to capture the underlying biological tissue organization, which leads to reduced performance as noted in [9]. Moreover, differentiating vessel-types improved staging likely because it explicitly encoded anatomical patterns related with fibrosis progression. Collagen bridges between distinct vessel-types is a characteristic of fibrosis progression and, by incorporating these features, the model should be better equipped to discriminate between intermediate stages. Overall, these results suggest that integrating domain knowledge into graph construction leads to improved predictive performance.

5.1.3 Model implementation

From an implementation perspective, the success of JK-Net and ordinal loss reflects the need to align the model with biology. JK-Net’s multi-scale aggregation captures both local collagen density and global architectural shifts across layers, which is essential for staging. Moreover, the effectiveness of the combined ordinal loss reflects the inherent structure of the staging system, by penalizing distant stage misclassifications more heavily than adjacent ones. Furthermore, the preference for lower learning rates and weight decay suggest that the model requires a stable optimization path to resolve these high-dimensional topological patterns. Overall, these findings indicate that GNNs should be carefully tuned to respect the topological and biological complexity.

When comparing GNN architectures, GINE outperformed GIN, GCN, and GAT, highlighting the critical role of edge features in fibrosis staging. Unlike GIN, GINE incorporates edge information into the message-passing process, enabling the model to capture collagen fiber relationships and spatial interactions more effectively. This aligns with the biological basis of fibrosis, where the connectivity and structural organization of the fiber network, rather than isolated node properties, define the disease stage [3]. Furthermore, the lower performance of GAT and GCN suggests that GINE is better suited for learning complex graph representations, further supporting its application in fibrosis staging tasks.

5.2 Comparison With Baseline and State-of-the-Art Models

The baseline max-pooling MIL treats each WSI as a collection of independent patches, assigning the bag label based on the instance with the highest predicted class probability [10]. While this strategy identifies cirrhosis (F4) effectively due to its highly distinctive localized patterns, it struggles with intermediate stages where the discriminative information lies in the spatial distribution and extent of fibrosis across the slide. This limitation is reflected in the near-chance accuracy (0.21 vs. 0.20 expected by chance) and low QWK (0.06), indicating that MIL fails to preserve the ordinal relationships between stages.

Incorporating spatial information through graph-based modeling addresses this

limitation. The SOTA GCN model achieved higher macro AUC, QWK, and accuracy than the MIL baseline, supporting the hypothesis that modeling spatial relationships provides critical context beyond independent patches [7, 8]. Nevertheless, its performance remained limited in middle stages ($AUC_2=0.37$), suggesting that spatial connectivity alone is insufficient to identify the subtle transition from portal expansion to early bridging.

The proposed model improved upon these results, matching the SOTA macro AUC and substantially increasing the accuracy (0.37) and QWK (0.44). These results suggest that integrating domain-specific biological knowledge (encoding collagen fibers as edges and preserving vessel identities) enhances staging robustness. Notably, the higher QWK score indicates that the model’s errors are primarily “adjacent” (e.g., mistaking F2 for F1 or F3), aligning closely with clinical practice where intermediate stages often exhibit inter-observer variability. Furthermore, unlike the SOTA GCN, our approach maintained more consistent AUC values across stages. This stability demonstrates that our model captures anatomical hallmarks of progression rather than abstract spatial proximity, providing a more reliable tool for MASLD staging than standard GNN or MIL architectures.

5.3 Ethical Considerations and Clinical Implications

One of the main ethical concerns of the use of AI in medical research is the privacy and confidentiality of patient data [60]. In this project, to ensure all patient information was handled securely, all data was stored and processed on the Bianca UPPMAX cluster [61], a system designed according to the ISO/IEC 27000 series of information security standards for managing sensitive personal data. The opacity of AI models is also another challenge as they often do not provide clear insights into their decision-making processes, complicating clinical implementation and accountability [62]. To address this, human oversight was kept throughout the project, AI models were kept as simple as possible, and interactive plots were designed to explore model decisions and predictions. This ensured that the proposed models relied on clinically relevant features rather than spurious correlations. Finally, regarding liability, this model is intended to be used as a decision-support tool rather than an autonomous diagnostic system. By clearly framing it as a helper to clinicians rather than a replacement, potential legal and ethical risks are minimized while ensuring that the ultimate responsibility lies with the human experts.

5.4 Limitations

Several limitations of this project should be acknowledged. First, the quality of the graph representation is highly dependent on the accuracy of the segmentation vessel pipeline. Errors in segmentation will propagate to the graph construction stage, leading to incorrect node placement or spurious edges. Second, the graph construction process relied on predefined rules (e.g., cost functions and path constraints), which could have introduced bias and limited the model to learn alternative structural relationships. Additionally, the approach

may be sensitive to parameter choices in the preprocessing and graph construction steps. Third, the size of the dataset might limit the generalizability of the model. Histological data are inherently heterogeneous, and the absence of external validation on multi-center datasets restricts the assessment of robustness across different staining protocols and acquisition settings. Finally, while the proposed GNN captures spatial relationships, it still operates on a reduced representation of the WSI. As a result, it potentially overlooks other relevant information outside the modeled structure, which could improve performance.

5.5 Future work

Future work could address these limitations in several directions. First, improving the robustness of the segmentation pipeline, e.g. by introducing new features. These changes could reduce error propagation and improve graph quality. Second, the graph representation could be extended to incorporate additional biological features, such as inflammatory cells, enabling a more comprehensive modeling of the liver. This would allow the framework to better reflect the multifactorial nature of fibrosis progression. Third, instead of relying on predefined graph construction rules, future approaches could explore learnable or adaptive graph construction strategies, where the connectivity between nodes is optimized during training. Fourth, from a data perspective, evaluating the model on larger and more diverse datasets, including multi-center cohorts, would be essential to assess generalizability and clinical applicability. Finally, future work should improve model interpretability by plotting new graphs and carrying out experiments or interviews with hepatologist to enhance trust and facilitate potential clinical adoption.

6 Conclusion

The main objective of this thesis was to develop and evaluate a biologically informed, topology-aware DL framework for automated fibrosis staging in MASLD using SR-stained WSIs. Instead of relying on patch-based MIL methods that ignore spatial context or purely geometric-driven graph construction pipelines, this work explicitly modeled the structural relationship between vessels and collagen fibers (key components of fibrosis progression) to investigate whether incorporating biological domain knowledge could improve staging.

Several sub-objectives were carried out to achieve this. First, 120 WSIs from MASLD patients were successfully preprocessed into tiles. Second, a vessel segmentation and classification pipeline was developed by combining morphological and color features, UNI embeddings, and ML classifiers. Despite moderate segmentation performance (F1-score of 0.69), this pipeline achieved strong vessel classification results (F1-scores of 0.91 and 0.80 for portal and central vein). Thus, it provides a practical and annotation-efficient alternative to simple color-based or U-Net segmentation models, reducing the need for large annotated datasets while still enabling biologically grounded representations. Third, a biologically informed graph construction pipeline was implemented, representing vessels as nodes and collagen fibers as edges. The two graph variants (with and without vessel-type differentiation) were shown to capture significant and biologically consistent trends across stages, substantially outperforming geometric-based graphs by doubling the staging accuracy. This highlights the importance of incorporating biological knowledge into graph design. Fourth, a GINE model was tuned and applied to fibrosis staging for the first time. The model was benchmarked against a max-pooling MIL baseline and a GCN SOTA model. The proposed model showed improved accuracy (an increase of 0.11 over SOTA) and QWK (an increase of 0.16 over SOTA), while matching the SOTA’s macro AUC and demonstrating consistent discrimination across stages.

Given the successful completion of these sub-objectives, it can be concluded that the main objective was achieved. The results demonstrate that incorporating biological domain knowledge into graph-based DL frameworks can produce more robust, data-efficient, and clinically meaningful predictions. This suggests that domain knowledge could compensate for limited data, offering a viable path toward clinically deployable models in hospitals and research settings where large annotated datasets are unavailable. Overall, this work serves as a strong foundation for continued research in graph-based computational pathology and biologically informed AI applications in hepatology.

Bibliography

- [1] P. Le, M. Tatar, S. Dasarathy, N. Alkhouri, W. H. Herman, G. B. Taksler, A. Deshpande, W. Ye, O. A. Adekunle, and A. McCullough, “Estimated burden of metabolic dysfunction–associated steatotic liver disease in us adults, 2020 to 2050,” *JAMA Network Open*, vol. 8, no. 1, p. e2454707, 2025, published Jan 17, 2025. [Online]. Available: <https://doi.org/10.1001/jamanetworkopen.2024.54707>
- [2] D. Q. Huang, V. W. S. Wong, M. E. Rinella *et al.*, “Metabolic dysfunction-associated steatotic liver disease in adults,” *Nature Reviews Disease Primers*, vol. 11, p. 14, 2025. [Online]. Available: <https://doi.org/10.1038/s41572-025-00599-1>
- [3] P. Angulo, D. E. Kleiner, S. Dam-Larsen, L. A. Adams, E. S. Bjornsson, P. Charatcharoenwitthaya, P. R. Mills, J. C. Keach, H. D. Lafferty, A. Stahler *et al.*, “Liver fibrosis, but no other histologic features, is associated with long-term outcomes of patients with nonalcoholic fatty liver disease,” *Gastroenterology*, vol. 149, no. 2, pp. 389–397, 2015.
- [4] Z. Lin, R. Zhang, S. Ren, T. He, H. Mi, W. Jiang, and C. Su, “Global burden of metabolic dysfunction-associated steatotic liver disease from 1990 to 2021 and the prediction for the next 10 years,” *Preventive Medicine Reports*, vol. 59, p. 103248, Sep. 2025.
- [5] N. Pugliese, A. Bertazzoni, C. Hassan, J. M. Schattenberg, and A. Aghemo, “Revolutionizing masld: How artificial intelligence is shaping the future of liver care,” *Cancers*, vol. 17, no. 5, 2025. [Online]. Available: <https://www.mdpi.com/2072-6694/17/5/722>
- [6] Y. Takahashi, E. Dungubat, H. Kusano, and T. Fukusato, “Artificial intelligence and deep learning: New tools for histopathological diagnosis of nonalcoholic fatty liver disease/nonalcoholic steatohepatitis,” *Computational and Structural Biotechnology Journal*, vol. 21, pp. 2495–2501, 2023. [Online]. Available: <https://doi.org/10.1016/j.csbj.2023.03.048>
- [7] C. Yin, S. Liu, F. Lyu, J. Lu, S. Darkner, V. W.-S. Wong, and P. C. Yuen, “Xfibrosis: Explicit vessel-fiber modeling for fibrosis staging from liver pathology images,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 11 282–11 291. [Online]. Available: <https://doi.org/10.1109/CVPR52733.2024.01072>
- [8] M. Wojciechowska, S. Malacrino, N. Garcia Martin, H. Fehri, and J. Rittscher, “Early detection of liver fibrosis using graph convolutional networks,” in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2021*, vol. 12908. Springer, 2021, pp. 217–226.

- [9] S. Brussee, G. Buzzanca, A. M. R. Schrader, and J. Kers, "Graph neural networks in histopathology: Emerging trends and future directions," 2024. [Online]. Available: <https://arxiv.org/abs/2406.12808>
- [10] G. Campanella, M. G. Hanna, L. Geneslaw, A. Miraflor, V. Werneck Krauss Silva, K. J. Busam, E. Brogi, V. E. Reuter, D. S. Klimstra, and T. J. Fuchs, "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *Nature Medicine*, vol. 25, no. 8, pp. 1301–1309, 2019.
- [11] W. K. Chan, K. H. Chuah, R. B. Rajaram, L. L. Lim, J. Ratnasingam, and S. R. Vethakkan, "Metabolic dysfunction-associated steatotic liver disease (masld): A state-of-the-art review," *Journal of Obesity and Metabolic Syndrome*, vol. 32, no. 3, pp. 197–213, Sep. 2023, epub 2023 Sep 13.
- [12] L. Miao, G. Targher, C. D. Byrne, Y. Y. Cao, and M. H. Zheng, "Current status and future trends of the global burden of masld," *Trends in Endocrinology and Metabolism*, vol. 35, no. 8, pp. 697–707, Aug 2024, epub 2024 Feb 29.
- [13] F. Tacke, P. Horn, V. Wai-Sun Wong, V. Ratziu, E. Bugianesi, S. Francque, S. Zelber-Sagi, L. Valenti, M. Roden, F. Schick, H. Yki-Järvinen, A. Gastaldelli, R. Vettor, G. Frühbeck, and D. Dicker, "Easl–easdeaso clinical practice guidelines on the management of metabolic dysfunction-associated steatotic liver disease (masld)," *Journal of Hepatology*, vol. 81, no. 3, pp. 492–542, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0168827824003295>
- [14] M. H. Le, Y. H. Yeo, X. Li, J. Li, B. Zou, Y. Wu, Q. Ye, D. Q. Huang, C. Zhao, J. Zhang, C. Liu *et al.*, "2019 global nafld prevalence: A systematic review and meta-analysis," *Clinical Gastroenterology and Hepatology*, vol. 20, no. 12, pp. 2809–2817.e28, 2021. [Online]. Available: <https://doi.org/10.1016/j.cgh.2021.12.002>
- [15] P. Portincasa, M. Khalil, L. Mahdi, V. Perniola, V. Idone, A. Graziani, G. Baffy, and A. Di Ciaula, "Metabolic dysfunction-associated steatotic liver disease: From pathogenesis to current therapeutic options," *International Journal of Molecular Sciences*, vol. 25, no. 11, p. 5640, May 2024.
- [16] D. M. Miller, K. F. McCauley, and K. J. Dunham-Snary, "Metabolic dysfunction-associated steatotic liver disease (masld): Mechanisms, clinical implications and therapeutic advances," *Endocrinology, Diabetes and Metabolism*, vol. ?, no. ?, p. ?, 2025. [Online]. Available: <https://doi.org/10.1002/edm2.70132>
- [17] R. F. Schwabe, F. Tacke, A. Sugimoto, and S. L. Friedman, "Antifibrotic therapies for metabolic dysfunction-associated steatotic liver disease," *JHEP Reports*, vol. 7, no. 8, p. 101421, 2025. [Online]. Available: <https://doi.org/10.1016/j.jhepr.2025.101421>

- [18] C. Y. Zhang, W. G. Yuan, P. He, J. H. Lei, and C. X. Wang, "Liver fibrosis and hepatic stellate cells: Etiology, pathological hallmarks and therapeutic targets," *World Journal of Gastroenterology*, vol. 22, no. 48, pp. 10 512–10 522, Dec. 2016.
- [19] P. Bedossa and T. Poynard, "A simple scoring system for assessing liver fibrosis in chronic hepatitis c," *Hepatology*, vol. 24, no. 2, pp. 289–293, 1996.
- [20] M. J. Brol, U. Drebber, J. A. Luetkens, M. Odenthal, and J. Trebicka, "the pathogenesis of hepatic fibrosis: basic facts and clinical challenges"—assessment of liver fibrosis: a narrative review," *Digestive Medicine Research*, vol. 5, no. 0, 2022. [Online]. Available: <https://dmr.amegroups.org/article/view/8412>
- [21] G. Soon and A. Wee, "Updates in the quantitative assessment of liver fibrosis for nonalcoholic fatty liver disease: Histological perspective," *Clinical and Molecular Hepatology*, vol. 27, no. 1, pp. 44–57, 2021.
- [22] C. Zhu, I. Tabas, R. F. Schwabe *et al.*, "Maladaptive regeneration. the reawakening of developmental pathways in nash and fibrosis," *Nature Reviews Gastroenterology & Hepatology*, vol. 18, pp. 131–142, 2021.
- [23] G. Garcia-Tsao, S. Friedman, J. Iredale, and M. Pinzani, "Now there are many (stages) where before there was one: In search of a pathophysiological classification of cirrhosis," *Hepatology*, vol. 51, no. 4, pp. 1445–1449, 2010. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/hep.23478>
- [24] S. Vanderbeck, J. Bockhorst, R. Komorowski, D. E. Kleiner, and S. Gawrieh, "Automatic classification of white regions in liver biopsies by supervised machine learning," *Human Pathology*, vol. 45, no. 4, pp. 785–792, 2014. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0046817713005145>
- [25] M. Z. Hoque, A. Keskinarkaus, P. Nyberg, and T. Seppänen, "Stain normalization methods for histopathology image analysis: A comprehensive review and experimental comparison," *Information Fusion*, vol. 102, p. 101997, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1566253523003135>
- [26] E. M. Brunt, "Nonalcoholic fatty liver disease: pros and cons of histologic systems of evaluation," *International journal of molecular sciences*, vol. 17, no. 1, p. 97, 2016.
- [27] K. Doi, "Computer-aided diagnosis in medical imaging: Historical review, current status and future potential," *Computerized Medical Imaging and Graphics*, vol. 31, no. 4, pp. 198–211, 2007, computer-aided Diagnosis (CAD) and Image-guided Decision Support. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0895611107000262>

- [28] M. G. Hanna, A. Parwani, and S. J. Sirintrapun, "Whole slide imaging: Technology and applications," *Advances in Anatomic Pathology*, vol. 27, no. 4, pp. 251–259, Jul 2020.
- [29] M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot, and B. Yener, "Histopathological image analysis: A review," *IEEE Reviews in Biomedical Engineering*, vol. 2, pp. 147–171, 2009, epub 2009 Oct 30.
- [30] G. E. Courtoy, I. Leclercq, A. Froidure, G. Schiano, J. Morelle, O. Devuyst, F. Huaux, and C. Bouzin, "Digital image analysis of picrosirius red staining: A robust method for multi-organ fibrosis quantification and characterization," *Biomolecules*, vol. 10, no. 11, p. 1585, 2020.
- [31] Y. Pan. (2026) Deployable pipeline for liver fibrosis quantification from routine histology in low-resource settings. [Online]. Available: https://github.com/yuyp4220/NASLD_collagen
- [32] D. Komura and S. Ishikawa, "Machine learning methods for histopathological image analysis," *Computational and Structural Biotechnology Journal*, vol. 16, pp. 34–42, 2018.
- [33] M. Waqas, S. U. Ahmed, M. A. Tahir, J. Wu, and R. Qureshi, "Exploring multiple instance learning (mil): A brief survey," *Expert Systems with Applications*, vol. 250, p. 123893, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0957417424007590>
- [34] M.-A. Carbonneau, V. Cheplygina, E. Granger, and G. Gagnon, "Multiple instance learning: A survey of problem characteristics and applications," *Pattern Recognition*, vol. 77, pp. 329–353, 2018. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0031320317304065>
- [35] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial Intelligence*, vol. 89, no. 1-2, pp. 31–71, 1997.
- [36] M. Y. Lu, D. F. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood, "Data-efficient and weakly supervised computational pathology on whole-slide images," *Nature Biomedical Engineering*, vol. 5, no. 6, pp. 555–570, 2021.
- [37] M. Ilse, J. M. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 80. PMLR, 2018, pp. 2127–2136. [Online]. Available: <https://proceedings.mlr.press/v80/ilse18a.html>
- [38] J. Yang, H. Chen, Y. Zhao, F. Yang, Y. Zhang, L. He, and J. Yao, "Remix: A general and efficient framework for multiple instance learning based whole slide image classification," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2022*. Springer, 2022, pp. 35–45.

- [39] T. Lin, Z. Yu, H. Hu, Y. Xu, and C. W. Chen, “Interventional bag multi-instance learning on whole-slide pathological images,” 2023. [Online]. Available: <https://arxiv.org/abs/2303.06873>
- [40] W. L. Hamilton, R. Ying, and J. Leskovec, “Representation learning on graphs: Methods and applications,” *IEEE Data Engineering Bulletin*, vol. 40, no. 3, pp. 52–74, 2017.
- [41] R. Pereira, M. R. Verdelho, C. Barata, and C. Santiago, “The role of graph-based mil and interventional training in the generalization of wsi classifiers,” 2025. [Online]. Available: <https://arxiv.org/abs/2501.19048>
- [42] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, “Neural message passing for quantum chemistry,” in *Proceedings of the 34th International Conference on Machine Learning*, 2017, pp. 1263–1272.
- [43] K. Xu, C. Li, Y. Tian, T. Sonobe, K.-i. Kawarabayashi, and S. Jegelka, “Representation learning on graphs with jumping knowledge networks,” in *Proceedings of the 35th International Conference on Machine Learning*, 2018, pp. 5453–5462.
- [44] R. Ying, J. You, C. Morris, X. Ren, W. L. Hamilton, and J. Leskovec, “Hierarchical graph representation learning with differentiable pooling,” in *Advances in Neural Information Processing Systems*, 2018.
- [45] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, “How powerful are graph neural networks?” 2019. [Online]. Available: <https://arxiv.org/abs/1810.00826>
- [46] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” in *International Conference on Learning Representations*, 2017.
- [47] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, “Graph attention networks,” 2018. [Online]. Available: <https://arxiv.org/abs/1710.10903>
- [48] W. Hu, B. Liu, J. Gomes, M. Zitnik, P. Liang, V. Pande, and J. Leskovec, “Strategies for pre-training graph neural networks,” 2020. [Online]. Available: <https://arxiv.org/abs/1905.12265>
- [49] L. Piccioni Costa, M. Guerreiro, E. Puchta, Y. Tadano, T. Antonini Alves, M. Kaster, and H. Siqueira, *Multilayer Perceptron*. IEEE Computational Intelligence Society Open Book, 01 2023, p. 105.
- [50] B. Weisfeiler and A. Lehman, “A reduction of a graph to a canonical form and an algebra arising during this reduction,” *Nauchno-Technicheskaya Informatsia*, vol. 9, 1968.
- [51] B. Alberts, *Molecular Biology of the Cell*. W.W. Norton, 2017. [Online]. Available: <https://books.google.se/books?id=2xIwDwAAQBAJ>

- [52] M. Farzi, C. McGenity, A. Cratchley, L. Leplat, P. Bankhead, A. Wright, and D. Treanor, “Liver-quant: Feature-based image analysis toolkit for automatic quantification of metabolic dysfunction-associated steatotic liver disease,” *Computers in Biology and Medicine*, vol. 190, p. 110049, 2025. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0010482525004007>
- [53] R. J. Chen, T. Ding, M. Y. Lu, D. F. Williamson, G. Jaume, B. Chen, A. Zhang, D. Shao, A. H. Song, M. Shaban *et al.*, “Towards a general-purpose foundation model for computational pathology,” *Nature Medicine*, 2024.
- [54] E. W. Dijkstra, “A note on two problems in connexion with graphs,” *Numerische mathematik*, vol. 1, no. 1, pp. 269–271, 1959.
- [55] H. B. Mann and D. R. Whitney, “On a test of whether one of two random variables is stochastically larger than the other,” *The Annals of Mathematical Statistics*, vol. 18, no. 1, pp. 50–60, 1947.
- [56] W. H. Kruskal and W. A. Wallis, “Use of ranks in one-criterion variance analysis,” *Journal of the American Statistical Association*, vol. 47, no. 260, pp. 583–621, 1952.
- [57] M. Galar, A. Fernández, E. Barrenechea, H. Sola, and F. Herrera, “A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches,” *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 42, pp. 463 – 484, 07 2012.
- [58] J. A. Sáez, J. Luengo, J. Stefanowski, and F. Herrera, “Smote–ipf: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering,” *Information Sciences*, vol. 291, p. 184–203, 01 2015.
- [59] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD ’16. ACM, Aug. 2016, p. 785–794. [Online]. Available: <http://dx.doi.org/10.1145/2939672.2939785>
- [60] N. Díaz-Rodríguez, J. D. Ser, M. Coeckelbergh, M. L. de Prado, E. Herrera-Viedma, and F. Herrera, “Connecting the dots in trustworthy artificial intelligence: From ai principles, ethics, and key requirements to responsible ai systems and regulation,” 2023. [Online]. Available: <https://arxiv.org/abs/2305.02231>
- [61] Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX), “Bianca: A sensitive data cluster at UPPMAX,” <https://www.uppmass.uu.se/resources/systems/bianca/>, 2026, accessed: 2026-04-06.

- [62] S. Yu, S. S. Lee, and H. Hwang, “The ethics of using artificial intelligence in medical research,” *Kosin Medical Journal*, vol. 39, no. 4, pp. 229–237, 2024.
- [63] J. Błaszczyński and J. Stefanowski, “Actively balanced bagging for imbalanced data,” in *Foundations of Intelligent Systems*, M. Kryszkiewicz, A. Appice, D. Ślęzak, H. Rybinski, A. Skowron, and Z. W. Raś, Eds. Cham: Springer International Publishing, 2017, pp. 271–281.
- [64] E. C. Gök and M. O. Olgun, “Smote-nc and gradient boosting imputation based random forest classifier for predicting severity level of covid-19 patients with blood samples,” *Neural Computing and Applications*, vol. 33, no. 22, pp. 15 693–15 707, 2021, epub 2021 Jun 11.
- [65] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” *CoRR*, vol. abs/1505.04597, 2015. [Online]. Available: <http://arxiv.org/abs/1505.04597>
- [66] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, “Segment anything,” 2023. [Online]. Available: <https://arxiv.org/abs/2304.02643>
- [67] R. Deng, C. Cui, Q. Liu, T. Yao, L. W. Remedios, S. Bao, B. A. Landman, L. E. Wheless, L. A. Coburn, K. T. Wilson, Y. Wang, S. Zhao, A. B. Fogo, H. Yang, Y. Tang, and Y. Huo, “Segment anything model (sam) for digital pathology: Assess zero-shot segmentation on whole slide imaging,” 2023. [Online]. Available: <https://arxiv.org/abs/2304.04155>
- [68] T. A. A. Tosta, P. R. de Faria, L. A. Neves, and M. Z. do Nascimento, “Color normalization of faded h&e-stained histological images using spectral matching,” *Computers in Biology and Medicine*, vol. 111, p. 103344, 2019. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0010482519302136>
- [69] A. Babic, I. R. Loftin, S. Stanislaw, M. Wang, R. Miller, S. M. Warren, W. Zhang, A. Lau, M. Miller, P. Wu *et al.*, “The impact of pre-analytical processing on staining quality for h&e, dual hapten, dual color in situ hybridization and fluorescent in situ hybridization assays,” *Methods*, vol. 52, no. 4, pp. 287–300, 2010.
- [70] C. C. Bilgin, J. Rittscher, R. Filkins, and A. Can, “Digitally adjusting chromogenic dye proportions in brightfield microscopy images,” *Journal of microscopy*, vol. 245, no. 3, pp. 319–330, 2012.
- [71] G. O. Rolls, N. J. Farmer, and J. B. Hall, *Artifacts in histological and cytological preparations*. Leica Microsystems., 2008.
- [72] M. Macenko, M. Niethammer, J. S. Marron, D. Borland, J. T. Woosley, X. Guan, C. Schmitt, and N. E. Thomas, “A method for normalizing histology slides for quantitative analysis,” in *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, 2009, pp. 1107–1110.

- [73] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley, “Color transfer between images,” *IEEE Computer Graphics and Applications*, vol. 21, no. 5, pp. 34–41, 2001.
- [74] A. Vahadane, T. Peng, A. Sethi, S. Albarqouni, L. Wang, M. Baust, K. Steiger, A. M. Schlitter, I. Esposito, and N. Navab, “Structure-preserving color normalization and sparse stain separation for histological images,” *IEEE Transactions on Medical Imaging*, vol. 35, pp. 1962–1971, 2016. [Online]. Available: <https://api.semanticscholar.org/CorpusID:6994762>
- [75] L. van der Maaten, G. Hinton, and Y. Rachmad, “Visualizing data using t-sne,” *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 11 2008.

Appendix A: Flowchart of the proposed pipeline

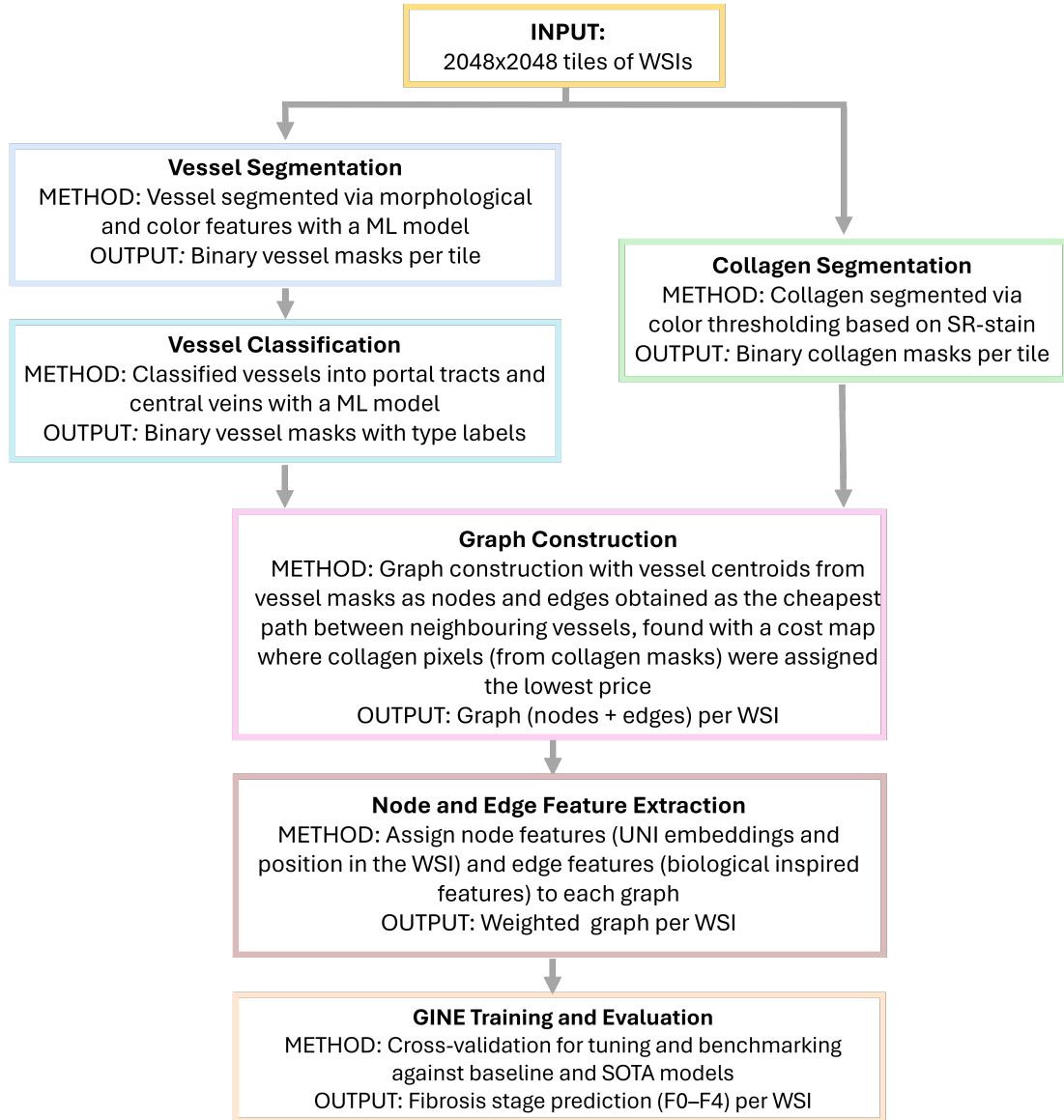


Figure 12: Summary of main steps (within boxes) for proposed framework. Arrows shows the input for each step. The pipeline starts with collagen and vessel segmentation from 2048×2048 WSI tiles, creating binary vessel and collagen masks. Vessel Segmentation undergoes a extra step of classification, separating into portal and central veins the vessels in the vessel masks. From these two masks, a graph is constructed where vessel centroids act as nodes and edges are constructed following the collagen path in between vessels (identified with a cost map). The nodes and edges are assigned features to create a weighted graph. The graphs are fed into a GINE model, which is then tuned and benchmarked against a baseline and SOTA for fibrosis staging.

Appendix B: Models used for Vessel Segmentation

Machine Learning Training, Evaluation and Additional Results

For vessel segmentation (step 2 in Figure 5), a ML approach was used to identify vessels based on morphological and color features. Due to the severe class imbalance, where vessels were rare compared to other cavities (e.g. some tiles presented 2 vessels against 228 fat droplets) only strategies designed to handle imbalanced datasets were considered. These included Balanced Bagging, XGBoost, and GB-SMOTE. Balanced bagging is an ensemble learning method that trains multiple base classifier models on different balanced versions of the data, typically obtained through random undersampling of the majority class [63]. In contrast, XGBoost is an efficient and scalable implementation of gradient boosting that builds an ensemble of decision trees, where each new tree corrects the errors of the previous ones [59]. Finally, SMOTE handles class imbalance by generating artificial samples of the minority class to train a gradient boosting model [64].

More sophisticated segmentation models such as U-Net [65] and Segment Anything Model (SAM), a promptable segmentation model trained on natural images [66], were considered but ultimately not adopted. U-Net was discarded due to limited number of annotated tiles. Moreover, this vessel segmentation problem is better framed as a contour-level classification task rather than pixel-wise segmentation, i.e. candidate cavities are already reliably detected via morphological operations, so the challenge is to classify each detected cavity, not finding its boundaries from scratch. Similarly, SAM was discarded due to the limited number of annotated tiles and, mainly, because its adaptation to segment histological images is challenging as natural and histological images differ widely and vessels are particular small which complicates segmentation. Moreover, SAM requires either manual prompts (points, boxes, or masks) to identify target structures, or an automated prompt generation strategy, neither of which was feasible given the annotation budget of this study [67]. In contrast, the chosen ML classifiers require far fewer samples to generalize, making them better suited to the annotation constraints of this project.

These models were trained on an expert annotated ground truth dataset of 200 WSI tiles, capturing different stages of fibrosis and staining variability. In this context, annotation refers to the manual delineation of contours to identify vessel boundaries. To address class imbalance, geometric data augmentation, specifically rotation and flipping, was applied to the minority class. For each tile, candidate cavities were extracted using morphological operations. For each cavity, 24 morphological and color features were extracted and fed into the classifiers as a single vector. The classifier therefore operated at the contour

level, with one prediction per detected cavity, rather than at the pixel or tile level. Model performance was evaluated using 5-fold group cross-validation framework. To prevent data leakage, image tiles were treated as independent groups, ensuring that patches from the same tile did not span across the training, validation, and testing splits (partitioned at 70%, 15%, and 15%, respectively). Classification probabilities were converted to binary predictions using a decision threshold. This threshold was not fixed but rather optimized within each cross-validation fold by maximizing the F1-score. The mean threshold across folds was selected as final value. Finally, performance was evaluated using the confusion matrix, precision, recall and F1-score. The confusion matrix is illustrated in Figure 13, which correlate with the previously explained results.

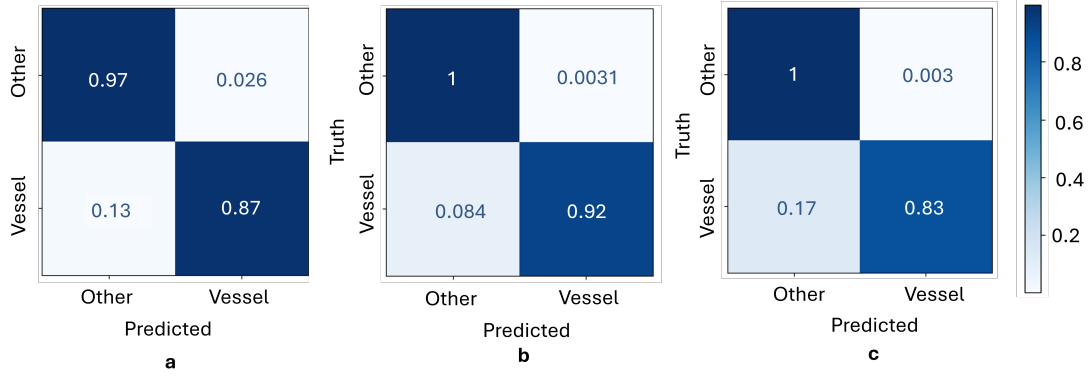


Figure 13: Confusion matrix summarizing classifier performance of (a) Balance Bagging (b) XGBoost and (c) SMOTE with gradient boosting.

Similarly, for vessel classification (step 3 in Figure 5), WSI tiles were never passed directly into the two tested classifiers. Instead, the detected vessels in step 2 were represented either as a 24-dimensional hand-crafted feature vector or as a 1032-dimensional vector combining UNI patch embeddings with shape features. Model performance was evaluated again using 5-fold group cross-validation framework with the same data split as in step 2. The confusion matrix for vessel classification is illustrated in Figure 14, which correlate with the previously explained results.

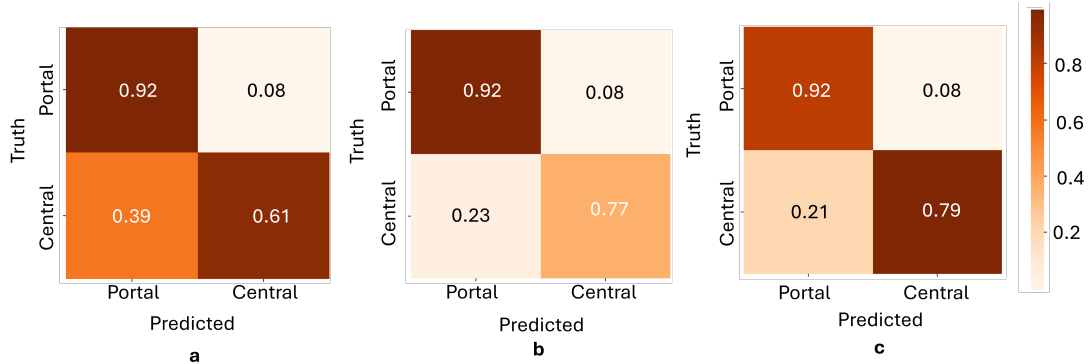


Figure 14: Confusion matrix summarizing classifier performance of (a) Selected features + XGBoost (b) UNI + XGBoost and (c) UNI + Logistic Regression.

Ablation Study on Color Normalization

Histological images are generated through the preparation, staining, and digitization of tissue samples. These steps are known to introduce variability in the color of the resulting WSIs [68]. For example, the fixation step, which prevents tissue degradation, can alter stain intensity depending to the fixative used and its fixation time [69]. Similarly, incomplete tissue dehydration can result in a more opaque appearance due to the residual water droplets [70]. Moreover, the staining process itself can be affected by factors such as the concentration and pH of the staining solution, chemical oxidation, and the duration of staining [71].

While humans have a good tolerance for different stain appearances, studies have shown that color variation is a more significant challenge in digital pathology [68]. To address this issue, several normalization techniques have been proposed, including Macenko [72], Reinhard [73], and Vahadane [74]. However, most of these normalization frameworks were originally developed and validated for Hematoxylin and Eosin (H&E)-stained tissue. For example, Macenko’s approach employs a color deconvolution model in optical density (OD) space, assuming that pixel intensities are a linear combination of multiple stains. The method estimates the dominant stain vectors and projects pixel intensities onto this basis for normalization relative to a reference image [72]. When applied to single-chromogen stains such as SR, this assumption is less directly applicable, making stain vector estimation more sensitive to noise and parameter selection. Consequently, although these frameworks can be adapted for single-chromogen stains, their application outside H&E remains less common and less thoroughly studied.

Nevertheless, in this project, the vessel segmentation pipeline was also evaluated on both non-normalized and normalized data to account for stain variability. Following [7], each tile was normalized using an adaptation of Macenko’s pipeline with a predefined “well-stained” tile serving as the reference, followed by standardization via subtraction of the mean and division by the standard deviation. An ablation study was conducted by training and testing the ML models on both normalized and non-normalized data. The results of the XGBoost (as this one demonstrated superior performance compared to the other ML models) is presented below.

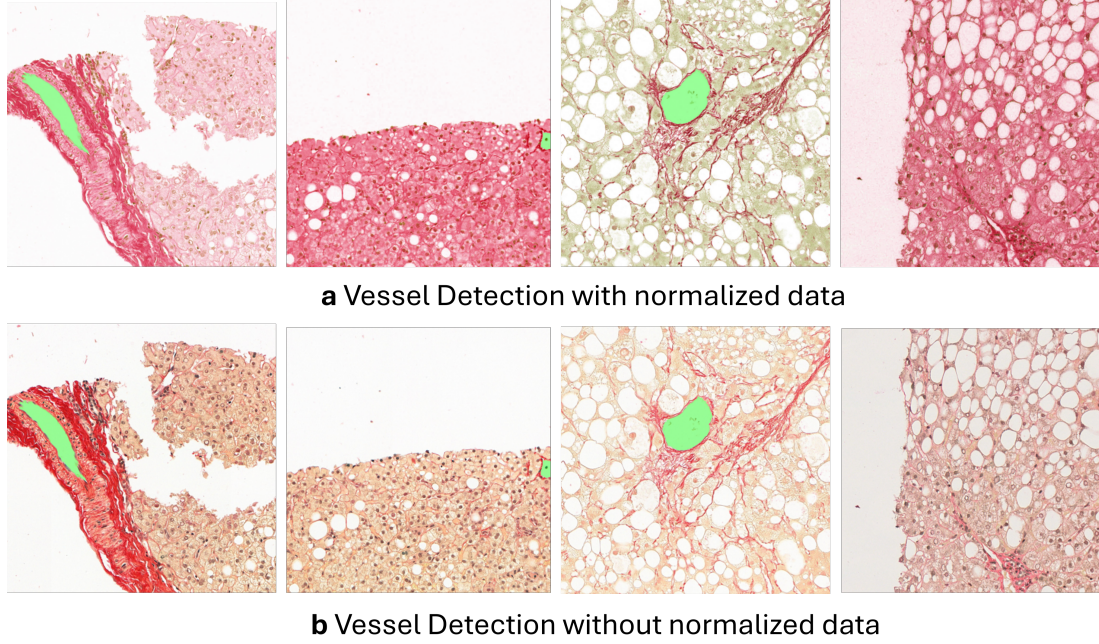


Figure 15: Representative figure of vessel detection using normalized (top) and non-normalized (bottom) data. Detected vessels across tiles of different WSIs with different staining appearance are illustrated in green

Table 9: XGBoost classification performance with and without normalized data

Data	F1-score	Precision	Recall
Normalized data	0.68 ± 0.39	0.70 ± 0.22	0.73 ± 0.41
Non-normalized data	0.69 ± 0.17	0.73 ± 0.25	0.70 ± 0.05

As illustrated in the representative example of Figure 15 and the results summarized in Table 9, the XGBoost model showed a similar performance with and without stain normalization. One possible explanation is that the model was trained on annotated tiles from a wide range of staining appearances, enhancing its robustness to color variability. It should also be noted that the effectiveness of Macenko normalization method is highly dependent on the choice of reference slide, which introduces an additional source of variability. Furthermore, as the dataset was derived from a single center and, hence, scanner, the extent of staining variability was relatively limited. The collagen-segmentation pipeline was also originally designed to account for stain variability and works with non-normalized data as input.

Considering these observations, along with the additional computational cost in terms of time and memory associated with normalization, non-normalized data were used in this project. Nevertheless, for future work involving multi-center datasets or images acquired from different scanning devices, the incorporation of a stain normalization pipeline is likely to be beneficial.

Appendix C: Graph interpretation on WSIs

Graph Representation Quality

The benefits of introducing biology-based graph-structured representations are demonstrated by using t-stochastic neighbor embedding (t-SNE) [75]. This analysis provides a low-dimensional visualization of the learned representations. As shown in Figure 16, both of the proposed graph representations shows certain discriminability across fibrosis stages. In the single-class representation, the advanced stages (depicted in orange and red) tend to accumulate separately to early stages (in green and yellow). Similar behavior is observed in the two-class representation which contains vessel-type information.

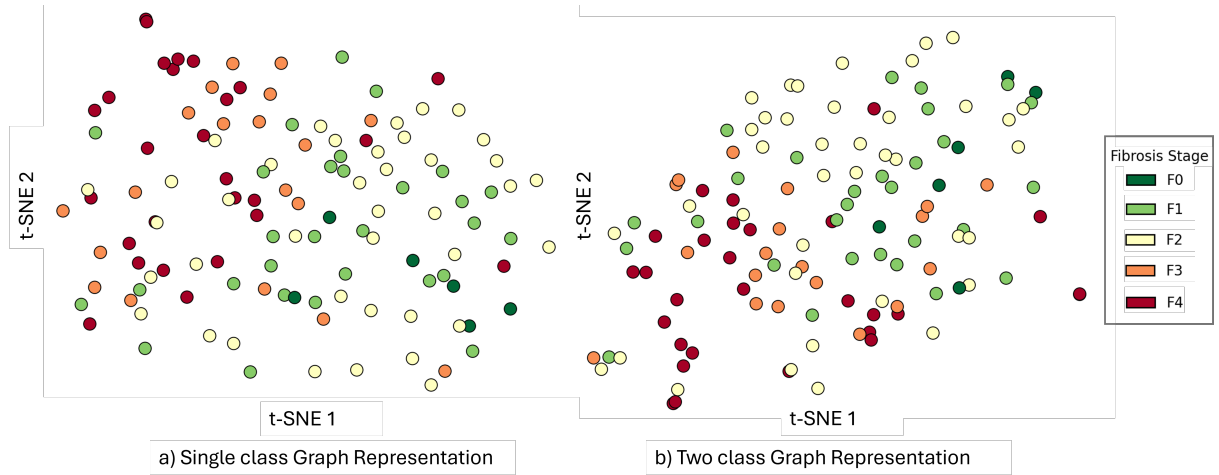


Figure 16: t-SNE plot of the (left) single-class graph representation and (right) two-class representation, showing in colors the ground truth label of each WSI.

This t-SNE visualization suggests that the model should be capable of learning discriminative feature embeddings that separate early and advanced fibrosis stages.

TRITA-CBH-GRU-2026:055
Stockholm, Sweden 2026
www.kth.se