Liver Fibrosis Stage Classification and Segmentation

Ahmad Al Hassan
Biomedical Engineering
Lebanese International University
Tripoli, Lebanon
92110021@students.liu.edu.lb

Marwan Alhawat
Biomedical Engineering
Lebanese International University
Tripoli, Lebanon
51630914@students.liu.edu.lb

Ahmad Diab
Biomedical Engineering
Lebanese International University
Tripoli, Lebanon
ahmad.diab@liu.edu.lb

Lara Hamawy
Biomedical Engineering
Lebanese International University
Tripoli, Lebanon
lara.hamawy@liu.edu.lb

2025 Eighth International Conference on Advances in Biomedical Engineering (ICABME) | 979-8-3315-5389-0/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICABME66883.2025.11211652

***Abstract*— Liver fibrosis results from chronic liver injury. It is a progressive condition marked by excessive extracellular matrix deposition that disrupts hepatic architecture and function. Accurate staging of fibrosis is critical for clinical decision-making, yet the current gold standard is liver biopsy which is normally associated with risks due to its invasive nature. Some of these risks are bleeding and sampling error. Ultrasound imaging, a widely used non-invasive modality, offers real-time liver visualization but often suffers from operator dependency and limited diagnostic specificity. This paper proposes a deep learning framework for automatic liver fibrosis classification and weakly supervised segmentation using ultrasound images. The classification part leverages transfer learning, by using trained models and building up on them to achieve our desired response. To enhance model transparency and clinical trust, Gradient-weighted Class Activation Mapping (Grad-CAM) is employed to create class-specific attention maps, which are used for effectively segmenting and highlighting regions relevant to fibrosis severity. Using various transfer learning models, EffiecientNet-b0 yielded the best results achieving a recall of 98.4%, accuracy of 98.84%, precision of 98.4%, and F1-Score of 98.2%. This model was further tested on self-taken private liver ultrasound images, and showcased a robust performance. Finally, this interpretable and scalable system offers a promising alternative to invasive diagnostics, advancing non-invasive liver disease assessment in clinical practice.**

Keywords—Liver Fibrosis, Ultrasound Images, Deep Learning (DL), Classification, Segmentation Grad-CAM, Convolutional Neural Networks (CNN).

I. INTRODUCTION

The adult human liver is the largest internal organ, weighing approximately 1.5 kilograms[1]. It is situated in the abdomen, in the upper right area. It is divided anatomically into right and left, both of which receive blood from the portal vein and the hepatic artery. Structurally, it is composed of microscopic units called lobules, each containing hepatocytes, the liver's main functional cells.

The liver performs a wide array of vital functions, including metabolism, detoxification, bile production, storage of essential vitamins, protein synthesis, and immune regulation. However, damage from conditions such as viral hepatitis, alcohol abuse, or non-alcoholic fatty liver disease (NAFLD) can severely impair these functions. This often leads to the development of liver fibrosis, cirrhosis, or even hepatocellular carcinoma.

Liver fibrosis is characterized by the excessive build-up of extracellular matrix proteins including collagen which is generally associated with chronic liver diseases[2]. It reflects the liver's wound-healing mechanism when subjected to sustained insults, such as hepatitis B or C infection, toxic exposures, autoimmune disorders, alcohol-related liver disease, or non-alcoholic fatty liver disease (NAFLD). Over time, fibrosis disrupts the liver's normal architecture and function. If left untreated, it may eventually advance to cirrhosis, liver failure, or even liver cancer.

Assessing liver fibrosis is crucial for guiding treatment in patients suffering from chronic liver diseases. While several non-invasive techniques have been introduced, including serum biomarkers and elastography-based imaging techniques such as FibroScan and magnetic resonance elastography (MRE), each comes with limitations in sensitivity, specificity, or accessibility. Conventional ultrasound remains a widely used modality due to its safety, affordability, and real-time imaging capabilities, but its diagnostic accuracy is often limited by operator dependency and subjective interpretation. Consequently, there has been a growing demand for more objective, accurate, and scalable tools for fibrosis staging. Although liver biopsy is still considered the benchmark in determining liver fibrosis, it carries significant risks, such as bleeding and sampling errors due to its invasive nature. In modern days, deep learning (DL) has emerged as a powerful alternative, transforming medical image analysis[3].

In the pursuit of supporting clinical decision-making and enhancing diagnostic efficiency, this paper proposes a robust deep learning classification and segmentation architecture for liver fibrosis staging based on ultrasound images. The proposed system aims to reduce diagnosis time by automating the analysis process. Therefore, aiding physicians in making informed and timely therapeutic decisions while maintaining high levels of accuracy and consistency.

The primary objectives of this paper center around developing a comprehensive and interpretable deep learning framework for liver fibrosis analysis using ultrasound images. The first one is to augment and balance the data to enhance its generalizability and avoid biases. Second, we aim to leverage the power of transfer learning by fine-tuning state-of-the-art pre-trained models, such as DenseNet, ResNet, and EfficientNet, for the liver fibrosis classification task. We also aim to use Grad-CAM to enhance interpretability. Pseudo-masks based on Grad-CAM will be generated and used to train a segmentation model utilizing attention U-net. The last part involves single-handedly taking images of our own livers and offering them as inputs to our proposed model for further testing.

979-8-3315-5389-0/25/$31.00 ©2025 IEEE

The rest of the paper is divided as follows: in section II we'll dive into related work highlighted in other papers. Section III will present our pipeline including the data, block diagram, preprocessing, online augmentation, balancing, classification, and segmentation. Next, section IV is dedicated to showcase our results. In section V we'll discuss the obtained results. Finally, section VI will serve as a conclusion.

II. Related Work

Li-Yun Xue et al. (2019) proposed a model that's based on transfer learning. It combines information from grayscale ultrasound (GM), elastogram modality (EM), and liver stiffness measurement (LSM) to classify liver fibrosis stages. Using the Inception-V3 architecture pretrained on ImageNet, the model successfully obtained high results, with AUCs of 0.950, 0.932, and 0.930 for classifying stages S4, ≥S3, and ≥S2, respectively. The combination of GM and EM provided the best performance, outperforming models using LSM or single-modality inputs[4].

Hyun-Cheol Park et al. (2024) proposed and compared several transfer learning models for classifying liver fibrosis using ultrasound images. The study utilized a dataset of 7,920 ultrasound images collected from 933 patients who had undergone either liver biopsy or hepatectomy. Liver fibrosis staging was assessed based on the pathology results employed by the METAVIR scoring system. VGGNet, ResNet, DenseNet, EfficientNet, and Vision Transformer (ViT) were all used to predict the METAVIR fibrosis stage across all five classes. All five models demonstrated comparable mean AUC values: 96%, 96%, 95%, 96%, 95% forVGGNet, ResNet, DenseNet, EfficientNet, and ViT respectively. They all achieved an accuracy of 94% and 96% specificity. Regarding sensitivity, EfficientNet achieved the highest mean value at 85%, while the remaining models resulted in marginally lower sensitivities ranging from 82% to 84%[5].

Haiming Ai et al. (2024) designed a deep learning framework for liver fibrosis assessment using ultrasound backscattered radiofrequency (RF) signals. The study addressed limitations in prior methods that relied solely on time-domain RF information and manual liver region of interest (ROI) selection. Their approach involved a two-stage architecture: a 2D CNN incorporating U-Net and Attention U-Net architecture was used to automatically segment the liver's region of interest (ROI), followed by a 1D CNN composed of four 1D layers for classification. The amplitude, phase, and power of the segmented RF signals were passed as inputs to the 1D CNN. A sliding window technique was also implemented to normalize and augment the ROI for better generalizability. Data were collected from two cohorts: Group A (613 participants) for segmentation and Group B (237 participants) for staging, with liver biopsy serving as the ground . In the Group A test set, U-Net and Attention U-Net achieved Dice scores of 95.05% and 94.68%, respectively. In Group B, the 1D CNN achieved the highest performance using the ROI phase spectrum for classifying stages ≥F1 (AUC: 95.7%, accuracy: 89.19%), ≥F2 (AUC: 80.8%, accuracy: 83.34%), and ≥F4 (AUC: 87.6%, accuracy: 85.71%), and using power spectrum signals for ≥F3 (AUC: 72.9%, accuracy: 77.14%)[6].

Fuji et al. (2024) proposed an AI-assisted framework for diagnosing liver fibrosis stages in metabolic dysfunction-associated steatotic liver disease (MASLD) using ultrasonography. ResNet-50 was utilized to classify the liver's surface as rough or smooth, while a U-Net architecture was used for liver segmentation on 214 annotated ultrasound images. ImageJ morphological shape analysis revealed the high correlation between liver fibrosis and MinFeret and Minor axis features. High segmentation accuracy was achieved (IoU = 0.935, F-score = 0.966), while the roughness classification reached 84.6% accuracy. Diagnostic performance was also notable, with AUROC values of 0.722 for ≥F3 and 0.825 for F4 [7].

III. Materials and Methods

A. Data

This dataset is derived from the research conducted by Y. Joo et al. in 2023 in their work entitled *"Classification of Liver Fibrosis From Heterogeneous Ultrasound Image"*. The ultrasound images used for training and testing were sourced from Seoul St. Mary's Hospital. Another testing dataset was sourced from Eunpyeong St. Mary's Hospital. The five stages of fibrosis are: F0, F1, F2, F3, F4.

F0-No Fibrosis: 2114 images; represents healthy liver tissue with no signs of fibrosis. The tissue appears normal, with no scarring or fibrosis present.

F1-Portal Fibrosis: 861 images; indicates fibrosis occurring around portal areas of the liver. Portal fibrosis involves the formation of scar tissue around the portal veins, which are small blood vessels in the liver.

F2-Periportal Fibrosis: 793 images; shows fibrosis around the edges of the liver's portal areas. Periportal fibrosis is characterized by scarring around the liver's boundary regions, often affecting the areas near the portal veins.

F3-Septal Fibrosis: 857 images; features fibrosis that forms bands or septa throughout the liver tissue. Septal fibrosis involves the development of thickened scar tissue that creates bands or partitions within the liver.

F4-Cirrhosis: 1698 images; represents advanced fibrosis leading to cirrhosis. It is the terminal stage of fibrosis, marked by extensive scarring and loss of liver function, which can significantly impact liver health.

Generally, the easiest stages to obtain are normal and cirrhosis. These two classes make up more than 50 % of the dataset. The classes' distribution is 34% for F0, 27% for F4, and 13% for F1, F2, and F3 classes.[8].

B. Model Architecture.

The below figure, Fig. 1., shows the chronological order of our model's architecture. The first step includes preprocessing the data using the appropriate methods. Second, online augmentation and class balancing are done at the same time. The classification part is up next, which uses transfer learning from known classification architectures: DenseNet, ResNet, and EfficientNet, along with the implementation of Grad-CAM. After that, we have the segmentation part followed by the results of both classification and segmentation. Finally, the last step is testing using our own liver's images.

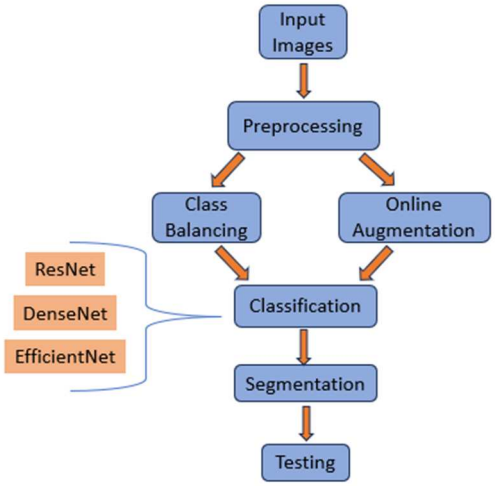

Fig. 1. Model's Architecture.

C. Preprocessing

The loaded ultrasound images required multiple preprocessing steps to prepare them for training. First of which was converting the images from BGR to RGB using OpenCV's cv2 library to ensure their compatibility with deep learning architectures. Most DL models expect RGB images[9]. The second step involved using a non-local means denoising which compares pixel patterns of the original images with that of noise to remove it[10]. Next, the images were then resized to 224 × 224 pixels, the most common input shape for convolutional neural network (CNN) architectures. Finally, pixel values were normalized to ensure model's convergence during training. Fig. 2. showcases the effect of preprocessing on a sample image from the dataset.

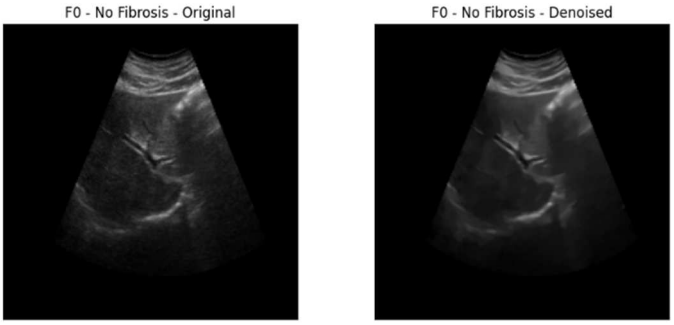

Fig. 2. The effect of preprocessing on a sample image.

D. Augmentation

An on-the-fly augmentation pipeline was built. The training dataset augmentations include spatial transformations such as resizing pixels, random horizontal flipping, and random shifts, scaling, and rotations to enhance spatial variability. To simulate ultrasound-specific characteristics, Gaussian blur and Gaussian noise are applied to mimic focus changes and inherent ultrasound noise. Additional augmentations like random brightness, contrast adjustments, and grid distortion, were implemented to simulate tissue deformation and varying imaging conditions. Finally, all images are normalized using ImageNet mean and standard deviation values to ensure compatibility with pretrained deep learning models. For the validation dataset, only resizing and normalization are applied to maintain consistency during evaluation. Fig. 3. shows the effect of the different augmentation techniques on the same image.

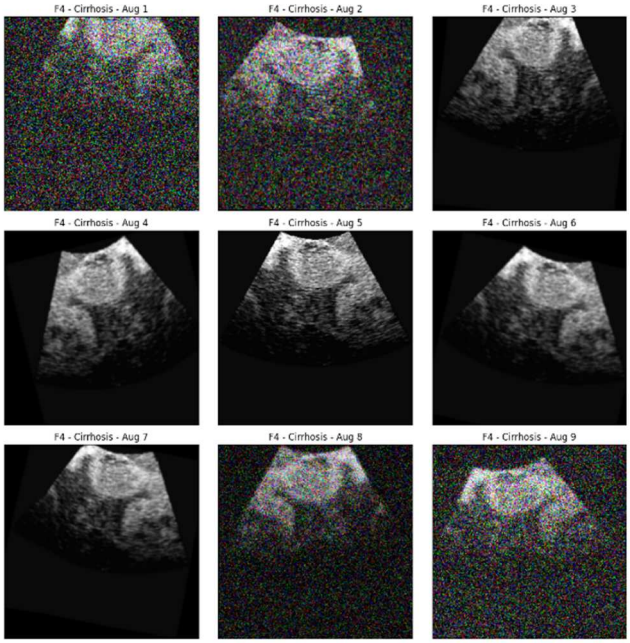

Fig. 3. Augmentations.

E. Balancing

A common issue that arises while dealing with datasets is their imbalances. This occurs when the classes of the given dataset don't have the same number of samples, and this could drastically affect the outcome of the model due to its high impact by introducing bias. Balancing aids in solving this issue by eliminating bias towards the majority class[11]. For this reason, weighted sampler was used to balance the dataset. Class weights are calculated as the inverse frequency of each class, assigning higher sampling probabilities for underrepresented classes. This sampler was applied exclusively during training to ensure balanced class representation. This approach improves model robustness by mitigating the effects of imbalanced class distributions during training. Fig. 8. And Fig. 9. showcases the difference in class data distribution before and after applying class balancing.

F. Classification

The classification task was approached using a CNN specifically designed to differentiate liver fibrosis stages from ultrasound images. Image classification and object recognition tasks can be performed by CNNs which operate on three-dimensional data[12]. The proposed model architecture was built around pre-trained backbone, fine-tuned to extract discriminative features related to fibrosis severity. The used models were: EfficientNet variants (EfficientNet-b0, EfficientNet-b1, EfficientNet-b2, EfficientNet-b3, and EfficientNet-b4), ResNet50, and DenseNet121 CNNs. The model removes the original classification heads of the used pre-trained models, and replaces them with a custom head consisting of dropout, fully connected, batch normalization, and ReLU layers to enhance generalization and mitigate overfitting. Global average pooling is applied to the extracted feature maps before classification. The data was split according to the rule 70%, 15%, and 15% for training, validating, and testing respectively. Training was performed on 100 epochs, with a learning rate of 1e-4, and weight decay of 1e-5. To enhance interpretability, and to make sure the

model is focusing on clinically meaningful regions, Class Activation Maps (CAMs) were generated through Grad-CAM. These visual explanations highlight the image region that the model relied on in making decision. This approach offers valuable insights into the model's reasoning and aligning with medical knowledge. During training, cross-entropy loss was optimized using the Adam optimizer with scheduled learning rate decay to ensure stable convergence.

G. Segmentation

Moving on from the classification model, segmentation masks were automatically generated by thresholding Grad-CAM heatmaps using a threshold value of 0.5, effectively transforming class-discriminative attention maps into pseudo-ground truth masks highlighting fibrotic regions. Image segmentation divides a digital image into clusters of pixels to provide essential information for object detection and other similar tasks. This allows faster and enhanced image processing[12]. This novel approach circumvented the need for labor-intensive manual annotations, enabling a weakly supervised learning strategy grounded in model interpretability. To accurately delineate the spatial extent of fibrosis, an advanced Attention-UNet architecture was employed. This network integrates attention gates within the decoding path to dynamically focus on salient features and suppress irrelevant activations, leading to improved boundary precision and contextual awareness. The model was trained using binary cross-entropy loss between the predicted masks and the Grad-CAM-derived labels, enabling it to learn high-quality segmentations despite the absence of manually annotated data.

H. Testing

The proposed model is designed to be more than just a research topic, but rather to be deployed in real-world scenarios. For this reason, obtained images of our livers were passed to the model as inputs. In this method we'd be able to assess the model's generalizability and its deployment chances.

IV. Results

The augmented and balanced data was then passed on to the classification model to determine its results. Table 1 summarizes the obtained metrics for each transfer learning model.

Table 1. Models' Results.

Model	EfficientNet_b0	EfficientNet_b1	EfficientNet_b2	EfficientNet_b3
Accuracy	0.9884	0.9831	0.9863	0.9842
Precision	0.984	0.978	0.98	0.978
Recall	0.984	0.974	0.978	0.976
F1-Score	0.982	0.976	0.978	0.976
F0-AUC	1	1	1	1
F1-AUC	0.9996	0.9982	0.9986	0.9984
F2-AUC	0.9995	0.9978	0.9965	0.9958
F3-AUC	0.9995	0.9986	0.9990	0.9924
F4-AUC	0.9999	0.9997	0.9999	0.9999

Model	EfficientNet_b4	ResNet50	DenseNet121
Accuracy	0.9863	0.9779	0.9789
Precision	0.98	0.97	0.974
Recall	0.978	0.974	0.968
F1-Score	0.98	0.97	0.972
F0-AUC	1	0.999	1
F1-AUC	0.9999	0.9998	0.9999
F2-AUC	0.9996	0.9999	0.9990
F3-AUC	0.9996	0.9998	0.9999
F4-AUC	0.9998	0.9997	0.9998

Fig. 4. shows the original image, the generated Gradient-weighted Class Activation Mapping (Grad-CAM) heatmaps, and the overlaid heatmaps on the original retinal fundus images, from left to right respectively. The spatial region that had the most contribution in the classification decision is highlighted. Warmer colors (e.g., red and yellow) corresponds to areas of high importance, where the model has focused its attention, while cooler colors (e.g., blue) signify less relevant regions.

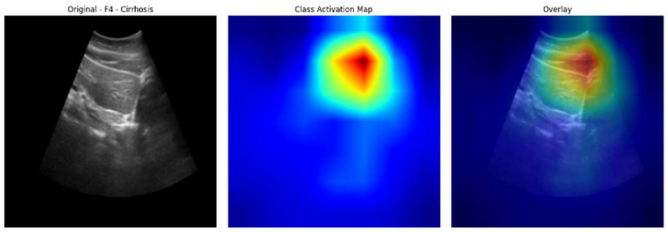

Fig. 4. Grad-CAM.

Fig. 5. Demonstrates the usage of GRAD-CAMs to obtain pseudo masks of the most relevant features detected by the model. White areas represent the areas highlighted by the GRAD-CAM, while the black areas represent the irrelevant features.

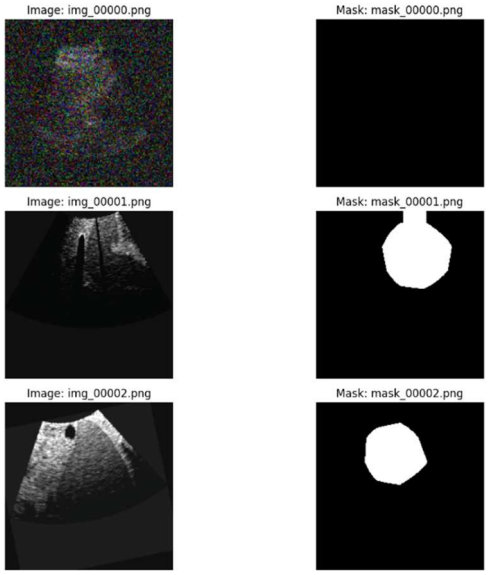

Fig. 5. Generated Masks.

After testing the segmentation model, we obtained an intersection of union (IoU) score of 0.4326, and some of the testing results are shown in Fig. 6.

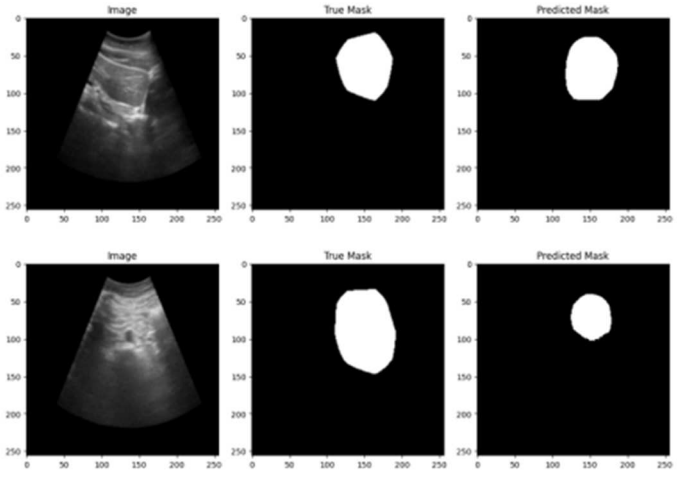

Fig. 6. Segmentation Testing.

Upon, passing our private livers' images to model indicated no fibrosis with a confidence level of 0.84, 0.79, and 0.96 for the images that are passed respectively. Fig. 7. shows the input image along with its processed counterpart with the classification results.

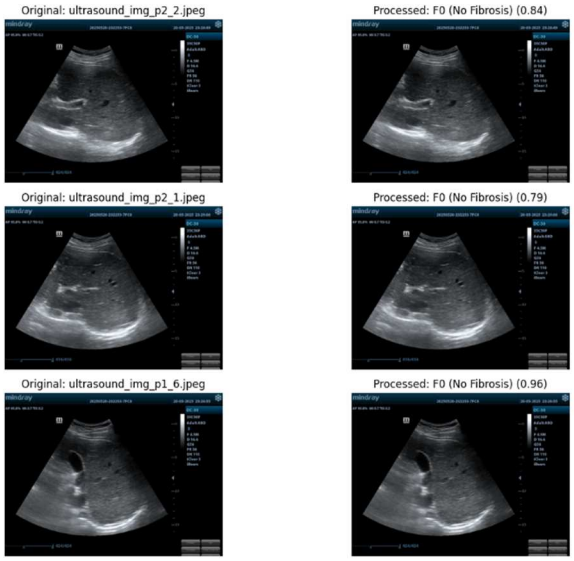

Fig. 7. Testing.

V. Discussion

This study presents a novel weakly supervised approach to liver fibrosis classification and segmentation. All the classification models demonstrated the ability to accurately distinguish liver fibrosis stages, but EfficientNet-b0 emerged as the most accurate with an accuracy of 98.84%, precision of 98.4%, F1-Score of 98.2%, and recall of 98.4%. The use of Grad-CAM offered more than interpretability to the model's decisions, but played a central role in generating meaningful pseudo-masks for fibrosis segmentation, thereby addressing the common challenge of lacking pixel-level annotations in medical imaging datasets.

The integration of Grad-CAM with Attention-UNet enabled effective localization of fibrotic regions, even in the absence of manually labeled segmentation masks. This showcases the feasibility of using class-discriminative attention for weak supervision where proper annotations are not available. Visual inspection of the segmentation outputs confirmed that the Attention-UNet was able to focus on fibrosis-consistent patterns, reinforcing the clinical validity of the approach. The proposed architecture achieved higher metrics than the studies elaborated in the related work chapter.

Despite the promising results, the study has several limitations. Mainly, the absence of real masks obtained from physicians makes the model's performance weaker, due to the absence of a ground truth to compare to. Therefore, the Grad-CAM-generated masks, while informative, may not capture precise anatomical boundaries, potentially introducing noise into the segmentation training process. This was reflected by the wide difference in IU score between this proposed framework and the one proposed by Fujii et al. where they achieved a 0.935 score compared to 0.4326. Additionally, the reliance on a fixed threshold to binarize CAMs could lead to variability across cases. The dataset size and diversity are relatively low; this may also limit the generalizability of the trained models to broader populations or different ultrasound acquisition protocols. Moreover, the proper identification of liver fibrosis stage is critical, as misclassifications hold a serious risk. Underestimating the disease severity could lead

to a delayed intervention which further worsens the case, while overestimations lead to unnecessary interventions. That's why proper classification is paramount.

Evaluating the model on real-world data showcased its ability to generalize well to new clinical data. This indicates the model's robustness, and potential deployment in clinical settings. However, the tested data belong to the same class. It wasn't diverse and didn't cover the entire classes.

Future work could explore more adaptive CAM thresholding techniques, incorporate multi-class segmentation aligned with fibrosis staging, and validate the method on larger, multicenter datasets. Furthermore, comparing this approach with fully supervised segmentation models trained on expert-labeled masks could better quantify the trade-offs between annotation effort and segmentation accuracy.

Overall, this work demonstrates that combining interpretable classification with Grad-CAM-based weak supervision offers a promising direction for scalable and clinically relevant liver fibrosis analysis.

VI. Conclusion and Perspectives

This work introduces a novel and practical weakly supervised framework for the classification and segmentation of liver fibrosis using ultrasound imaging. By leveraging the power of transfer learning, we achieved a robust performance despite the limited dataset size. The classification pipeline demonstrated strong discriminatory power across fibrosis stages, while the segmentation strategy creatively employed Grad-CAM-derived pseudo-labels to overcome the absence of manual annotations.

This framework not only minimizes reliance on costly expert labeling, but also aligns model attention with clinically relevant features, promoting both trust and interpretability. The use of Attention-UNet further refined segmentation precision, confirming the feasibility of weakly supervised learning in sensitive medical tasks.

Moving forward, this study paves the way for fully automated, scalable, and explainable liver fibrosis assessment systems. Future research will aim to validate this framework on larger, multi-center datasets and explore integration into real-time diagnostic workflows. This will bring us closer to accessible, AI-powered liver disease screening that could save countless lives.

References

[1] K. Wallace, A. D. Burt, and M. C. Wright, "Liver fibrosis," *Biochem. J.*, vol. 411, no. 1, pp. 1–18, Apr. 2008, doi: 10.1042/BJ20071570.

[2] R. Bataller and D. A. Brenner, "Liver fibrosis," *J. Clin. Invest.*, vol. 115, no. 2, pp. 209–218, Feb. 2005, doi: 10.1172/JCI24282.

[3] R. Anteby *et al.*, "Deep learning for noninvasive liver fibrosis classification: A systematic review," *Liver Int.*, vol. 41, no. 10, pp. 2269–2278, 2021, doi: 10.1111/liv.14966.

[4] L.-Y. Xue *et al.*, "Transfer learning radiomics based on multimodal ultrasound imaging for staging liver fibrosis," *Eur. Radiol.*, vol. 30, no. 5, pp. 2973–2983, May 2020, doi: 10.1007/s00330-019-06595-w.

[5] H.-C. Park *et al.*, "Automated classification of liver fibrosis stages using ultrasound imaging," *BMC Med. Imaging*, vol. 24, no. 1, p. 36, Feb. 2024, doi: 10.1186/s12880-024-01209-4.

[6] H. Ai, Y. Huang, D.-I. Tai, P.-H. Tsui, and Z. Zhou, "Ultrasonic Assessment of Liver Fibrosis Using One-Dimensional Convolutional Neural Networks Based on Frequency Spectra of Radiofrequency Signals with Deep Learning Segmentation of Liver Regions in B-Mode Images: A Feasibility Study," *Sensors*, vol. 24, no. 17, Art. no. 17, Jan. 2024, doi: 10.3390/s24175513.

[7] I. Fujii *et al.*, "Artificial Intelligence and Image Analysis-Assisted Diagnosis for Fibrosis Stage of Metabolic Dysfunction-Associated Steatotic Liver Disease Using Ultrasonography: A Pilot Study," *Diagnostics*, vol. 14, no. 22, p. 2585, Jan. 2024, doi: 10.3390/diagnostics14222585.

[8] "Liver Histopathology (Fibrosis) Ultrasound Images." Accessed: June 03, 2025. [Online]. Available: https://www.kaggle.com/datasets/vibhingupta028/liver-histopathology-fibrosis-ultrasound-images

[9] Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar, India., U. Oza*, Prof. P. Kumar, and Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar, India., "Empirical Examination of Color Spaces in Deep Convolution Networks," *Int. J. Recent Technol. Eng. IJRTE*, vol. 9, no. 2, pp. 1011–1018, July 2020, doi: 10.35940/ijrte.B4038.079220.

[10] M. p. k. k. and S. S. Rao, "Image Denoising And Enhancement : A Comparative Study," *Int. J. Eng. Res. Technol.*, vol. 12, no. 12, Dec. 2023, doi: 10.17577/IJERTV12IS120021.

[11] "Introduction to Balanced and Imbalanced Datasets in Machine Learning." Accessed: June 03, 2025. [Online]. Available: https://encord.com/blog/an-introduction-to-balanced-and-imbalanced-datasets-in-machine-learning/

[12] "IBM." Accessed: Aug. 31, 2025. [Online]. Available: https://www.ibm.com/us-en