Semi-Supervised Ultrasound Framework with Prototype Regularization for Five-Stage Liver Fibrosis Assessment

BY

Md. Shahriar Shakil
ID: 242-25-031

This Report Presented in Partial Fulfillment of the Requirements for
The Degree of Masters of Science in Computer Science and Engineering

Supervised By

Dr. Md. Zahid Hasan
Associate Professor
Department of CSE
Daffodil International University

Co-Supervised By

Dr. Abdus Sattar
Associate Professor
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

APPROVAL

This Project titled "**Semi-Supervised Ultrasound Framework with Prototype Regularization for Five-Stage Liver Fibrosis Assessment**", submitted by **Md. Shahriar Shakil**, ID No: **242-25-031** to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of M.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 13-09-2025.

BOARD OF EXAMINERS

Dr. S.M Aminul Haque — **Chairman**
Professor & Associate Head
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Ms. Nazmun Nessa Moon — **Internal Examiner**
Associate Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

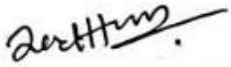
Dr. Md Alamgir Kabir — **Internal Examiner**
Assistant Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Mr. Sadat Hossain — **External Examiner**
Data Scientist,
Risk Management Division,

DECLARATION

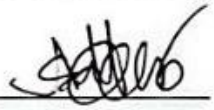
I hereby declare that this research has been done by me under the supervision of **Dr. Md. Zahid Hasan, Associate Professor, Department of CSE,** Daffodil International University. I also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:



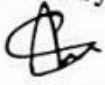
Dr. Md. Zahid Hasan
Associate Professor
Department of CSE
Daffodil International University

Co-Supervised by:



Dr. Abdus Sattar
Associate Professor
Department of CSE
Daffodil International University

Submitted by:



Md. Shahriar Shakil
ID: 242-25-031
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, I express my heartfelt thanks and gratitude to Almighty Allah for His divine blessing, which makes it possible to complete the final year project/internship successfully.

I am grateful and wish to express my profound indebtedness to **Dr. Md. Zahid Hasan, Associate Professor**, Department of CSE, Daffodil International University, Dhaka, deep knowledge & keen interest in the field of Machine Learning to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, and reading many inferior drafts and correcting them at all stages have made it possible to complete this project.

I would like to express my heartfelt gratitude to **Dr. Sheak Rashed Haider Noori**, **Head of the** Department of CSE, for his kind assistance in completing our project, as well as to the other faculty members and staff of the CSE department at Daffodil International University.

Finally, I must acknowledge with due respect the constant support and patience of my parents.

ABSTRACT

Liver fibrosis represents a gradual replacement of normal hepatic parenchyma with scar tissue, and accurate staging (F0–F4) is essential for guiding surveillance, therapeutic decisions, and specialist referral. Although biopsy remains the historical gold standard, it is invasive, costly, and prone to sampling variability. Ultrasound (US), by contrast, is non-invasive, broadly accessible, and suitable for repeated monitoring in chronic liver disease; however, intrinsic challenges such as depth-dependent attenuation, speckle noise, and inter-scanner heterogeneity complicate automated staging. In this study, developed an explainable semi-supervised learning framework for five-stage fibrosis classification using heterogeneous US data from tertiary-care hospitals, capturing real-world class imbalance with predominance of F0 and F4 cases. The framework integrates Mean Teacher consistency learning with prototypical loss to enhance stage-aware embeddings, while a class-balanced focal loss addresses long-tailed distributions. Under limited labeling budgets (5–15% per class), the approach achieves robust performance (test accuracy 93.46%, macro-F1 91.86%, $\kappa = 0.945$, ROC-AUC 0.999), with Grad-CAM highlighting clinically meaningful intraparenchymal features. These results demonstrate the potential of a prototype-regularized semi-supervised pipeline to deliver accurate, interpretable fibrosis staging from routine US, while markedly reducing annotation demands. Future directions include cross-center validation, calibration, and prospective clinical evaluation.

Keywords:

Liver fibrosis staging; Ultrasound imaging (B-mode); Medical image analysis; Semi-supervised learning; Mean Teacher; Prototypical networks; Class-balanced focal loss; Logit adjustment; Ensemble inference; Explainable AI (Grad-CAM)

.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 Introduction

Liver fibrosis, a critical stage in chronic liver disease, involves excessive deposition of extracellular matrix (ECM) proteins in hepatic tissue. If left unmanaged, fibrosis can progress to cirrhosis (stage F4), hepatocellular carcinoma, and eventual liver failure, representing a major global health concern [1]. Accurate and timely staging of liver fibrosis into five clinically accepted grades (F0–F4) is crucial for guiding treatment, determining prognosis, and evaluating therapeutic outcomes [2].

Traditionally, liver biopsy has been the gold standard for fibrosis staging; however, it is invasive, prone to complications (e.g., bleeding, infection), and affected by inter-observer variability and sampling errors [2]. While imaging-based approaches offer safer, non-invasive alternatives, their interpretation often relies on subjective visual analysis, which lacks reproducibility and scalability [1]. These limitations underscore the need for automated, accurate, and interpretable fibrosis staging systems that can support clinical decision-making [2] [1].

Recent advances in artificial intelligence (AI), especially deep learning (DL), have shown promise in medical image analysis, including classification, segmentation, and prognosis [1] [2] [3]. Convolutional neural networks (CNNs) have demonstrated strong potential for recognizing fibrosis-related patterns in imaging modalities such as ultrasound, CT, and MRI [1] [2]. In particular, deep learning applied to ultrasound imaging—a widely accessible, real-time, and non-invasive tool—has attracted increasing attention due to its clinical practicality [2].

Despite these advances, supervised deep learning methods rely heavily on large, annotated datasets. In medical domains like liver fibrosis staging, data annotation demands expert radiologists and hepatologists, making it resource-intensive [4]. Moreover, fibrosis datasets are often imbalanced: extreme stages (F0 and F4) are overrepresented, while intermediate stages (F1–F3)—which are clinically most actionable—are

underrepresented. This imbalance leads to biased learning, misclassification, and poor generalization [5].

To reduce dependency on annotated data, semi-supervised learning (SSL) has emerged as a promising alternative [4] [6] [7]. SSL leverages a small set of labeled samples alongside a larger pool of unlabeled data, improving generalization while minimizing annotation costs [4] [8]. Among SSL techniques, the Mean-Teacher framework has gained popularity due to its simplicity and effectiveness [8]. This framework trains a student model using both labeled data and pseudo-labels generated by a teacher model (an exponential moving average of the student). It encourages consistency under augmented views, improving learning stability [8].

However, SSL in medical imaging is challenged by pseudo-label noise, severe class imbalance, and low feature separability—particularly for subtle morphological changes in ultrasound images across fibrosis stages. Over-confident pseudo-labels may amplify confirmation bias [4]; class imbalance skews the model towards majority stages [5]; and similar appearances across F1–F3 stages hinder discriminative learning [1] [2].

Recent studies have sought to address these limitations via self-correcting pseudo-labeling [4], diverse teacher strategies [6], transfer learning across imaging domains [7], and generative diffusion-based explainability [3]. Additionally, explainable AI (XAI) methods such as Grad-CAM have enabled visual interpretation of model predictions by localizing decision-relevant regions, which is essential for clinical adoption [9].

Despite these contributions, no current approach integrates adaptive pseudo-label filtering, class-balanced optimization, and prototype-aware feature learning into a unified framework for liver fibrosis staging using ultrasound images[10]. This thesis proposes a novel explainable semi-supervised learning framework tailored for liver fibrosis classification. The proposed framework:

- Reduces reliance on annotated datasets,
- Improves mid-stage classification performance through prototype regularization,
- Handles imbalance via class-aware focal loss and augmentation, and

- Provides interpretable outputs through Grad-CAM and prototype-based visualizations.

1.2 Motivation

Clinically, fibrosis staging is essential for managing chronic liver disease. While cirrhosis (F4) is irreversible, intermediate stages (F1–F3) are potentially reversible if detected and treated early. However, accurate differentiation among these stages in ultrasound images is difficult due to subtle visual cues and under-representation in training data. Misclassification at these stages may lead to delayed intervention and disease progression [1].

Methodologically, annotated ultrasound datasets for fibrosis staging are scarce, requiring domain expertise and manual effort. In contrast, large volumes of unlabeled ultrasound data are routinely generated in clinical practice but remain underutilized. Semi-supervised learning provides a scalable solution by extracting useful patterns from unlabeled data without incurring high annotation costs [8].

Another major challenge is class imbalance: most datasets contain disproportionately high samples of F0 and F4, leading to overfitting to majority stages and low accuracy for minority (F1–F3) stages. Techniques like class-balanced focal loss and data augmentation methods (e.g., MixUp, CutMix) have shown promise in other tasks but are underexplored within SSL frameworks for multi-class fibrosis staging [5].

Finally, model explainability is essential for clinical acceptance. Without transparency, clinicians may be reluctant to trust black-box AI predictions. Techniques like Grad-CAM provide intuitive visual feedback, while prototype-based learning enhances interpretability by mapping class-representative features.

This thesis is motivated to address these limitations through a novel, explainable, and data-efficient approach to ultrasound-based liver fibrosis staging [3].

1.3 Problem Statement

Despite growing interest in deep learning for liver fibrosis staging, current solutions are limited by:

- Scarcity of labeled ultrasound datasets, increasing dependency on costly expert annotation.
- Class imbalance, leading to misclassification of intermediate stages.
- Lack of unified frameworks integrating SSL, class-aware learning, and explainability.

1.4 Rationale of the Study

This study is founded on the dual necessity of enabling label-efficient learning and delivering clinically interpretable models. While supervised learning remains effective, it is impractical for ultrasound-based fibrosis staging due to limited labeled data and annotation constraints.

This research proposes a semi-supervised learning framework enhanced with:

- Adaptive Pseudo-Label Thresholding using cosine scheduling to mitigate confirmation bias,
- Class-Balanced Focal Loss combined with MixUp/CutMix to reduce overfitting and balance stage distribution,
- Prototypical Regularization to improve class separability and mid-stage discrimination in feature space,
- Grad-CAM visualization and prototype explanation to support clinical interpretation and transparency.

By integrating these components into a single pipeline, the study aims to deliver a model that is not only accurate but also robust, efficient, and explainable for real-world deployment.

1.5 Research Questions

This study is guided by the following research questions, which collectively aim to evaluate the effectiveness and robustness of the proposed explainable semi-supervised framework for liver fibrosis staging:

- RQ1: Does integrating adaptive pseudo-label thresholding within the Mean-Teacher framework improve five-stage fibrosis classification compared to a supervised baseline?
- RQ2: Do prototype-based regularization strategies enhance feature separability and improve classification of intermediate stages (F1–F3)?
- RQ3: To what extent do class-balanced focal loss and data augmentations (MixUp, CutMix) improve generalization and mitigate class imbalance?
- RQ4: Does soft-voting across K-fold ensembles yield more stable and robust performance compared to single model evaluations?

1.6 Project Management and Finance

This research project is planned for completion within a six-month timeframe, divided into key phases including literature review, dataset preparation, methodology development, model training, experimental evaluation, and thesis writing. The project primarily relies on institutional resources such as GPU clusters, library access, and open-source software, thereby keeping financial requirements minimal. Table 1.1 shows the overview of the Project Management Plan.

Table 1.1: Project Management Plan

Phase	Timeline	Resources
Literature Review & Dataset Setup	Month 1	Digital libraries, datasets
Methodology Design & Implementation	Months 2–3	Python, PyTorch, GPUs
Training & Evaluation	Months 3–5	Institutional HPC resources
Reporting & Submission	Month 6	Writing tools, supervision

1.7 Report Layout

The remainder of this report is structured as follows. Chapter 2 reviews literature on liver fibrosis staging, semi-supervised learning, class imbalance strategies, and explainable AI in medical imaging. Chapter 3 presents the proposed methodology, including the Mean-Teacher framework, adaptive pseudo-label thresholding, and prototypical regularization. Chapter 4 describes the experimental outcomes of the study and provides a detailed analysis. Chapter 5 discusses the sustainability plan, societal and environmental impacts, and ethical considerations. Finally, Chapter 6 concludes the present investigation and outlines a strategy for subsequent endeavors.

CHAPTER 2
BACKGROUND

2.1 Preliminaries

Liver fibrosis represents the pathological accumulation of extracellular matrix components in response to chronic liver injury, often progressing to cirrhosis and hepatocellular carcinoma. The fibrosis staging system, typically categorized from F0 (no fibrosis) to F4 (cirrhosis), is critical for assessing disease severity, determining therapeutic pathways, and evaluating patient prognosis. Among imaging modalities, ultrasound is widely favored for its cost-effectiveness, safety, and accessibility, particularly in low-resource settings. However, the modality's inherent low contrast and dependence on operator skill complicate consistent and accurate fibrosis assessment, particularly at intermediate stages.

Deep learning, especially convolutional neural networks (CNNs), has demonstrated superior performance in image-based diagnostics, including liver pathology. These models autonomously learn spatial hierarchies from imaging data, enabling fibrosis-related pattern recognition that might be imperceptible to the human eye. Yet, CNNs typically require extensive labeled data, which is difficult to collect in the clinical domain due to the labor-intensive nature of expert annotations [1], [5].

Semi-supervised learning (SSL) addresses this limitation by enabling the use of both labeled and unlabeled data to train predictive models. A prominent example is the Mean-Teacher framework, which enforces consistency between a student model and a teacher model maintained via exponential moving averaging (EMA) [8].This method has shown promise across various medical imaging applications but remains underutilized for liver fibrosis detection using ultrasound.

One persistent issue in fibrosis classification is class imbalance. Datasets often contain an overrepresentation of F0 and F4 images, with relatively few samples for intermediate stages (F1–F3), leading to skewed learning and underperformance on the most clinically actionable categories. To alleviate this, loss functions such as class-balanced focal loss

have been developed, taking into account the effective number of samples per class [5]. Complementary to this, data augmentation techniques such as MixUp and CutMix have been applied to improve feature diversity and reduce overfitting in imbalanced scenarios [10].

Explainability is also paramount in medical AI, especially for high-stakes tasks such as fibrosis staging. Grad-CAM (Gradient-weighted Class Activation Mapping) is a well-established interpretability tool that highlights important image regions contributing to a model's decision, providing clinicians with visual cues that align with their diagnostic reasoning [9]. Further advances in explainable AI (XAI), including prototype-based learning and interpretable diffusion models, have extended these capabilities, offering a pathway toward more trustworthy and transparent deep learning models [3].

2.2 Related works

The application of deep learning and, more recently, semi-supervised learning (SSL) to liver fibrosis classification has gained considerable attention in medical imaging research. Among imaging modalities, ultrasound has emerged as a preferred tool due to its non-invasive nature, cost-effectiveness, and wide clinical availability. Consequently, ultrasound-based fibrosis staging has become a focal point for developing automated, label-efficient, and interpretable diagnostic systems.

Early work in this direction was carried out by Meng et al. [11], who employed transfer learning and a fully connected CNN (FCNet) to classify liver fibrosis stages from ultrasound data. Their study, which reported accuracy, sensitivity, and specificity, established an important baseline for the feasibility of deep learning in ultrasound fibrosis classification. Building on this, Xue et al. [12] introduced a multimodal ultrasound radiomics approach that combined transfer learning with handcrafted radiomics features. Their model improved staging performance and provided robust benchmarks in terms of AUC, accuracy, and F1-score. More recently, Joo et al. [13] developed a CNN-based

framework for fibrosis classification using a heterogeneous, multi-center ultrasound dataset. This work is particularly significant, as it directly addressed challenges of scanner and institutional variability. The dataset introduced in their study is also utilized in the present research, enabling direct comparisons with their baseline results.

Beyond ultrasound, other imaging modalities have also been explored. Yin et al. [14] investigated CT-based fibrosis staging using CNNs, incorporating Grad-CAM for visual interpretability, though their model was limited by the availability of annotated CT data. Li et al. [2] advanced MRI-based approaches with a multi-view deep learning framework that integrated uncertainty modeling and ensemble methods, thereby enhancing classification robustness and interpretability. While these works highlight modality-specific progress, they reinforce the broader importance of explainable methods in fibrosis staging.

SSL has emerged as a promising avenue to overcome the high costs of medical image annotation. Miao et al. [4] proposed SC-SSL, a framework that improves pseudo-label reliability through self-correcting collaborative training. Similarly, Zhao et al. [6] introduced an alternate diverse teaching strategy employing multiple teacher models to reduce confirmation bias in SSL. Wang et al. [7] developed TransMedSeg, a transferable SSL framework that improved cross-domain generalization, while Kaleta et al. [3] combined diffusion modeling with SSL in Mediffusion to generate label-efficient and interpretable predictions. Complementary innovations such as prototype-based regularization [10] and advanced loss functions like class-balanced focal loss [5] have further addressed challenges of class imbalance and feature separability in medical imaging datasets.

Ultrasound-specific SSL research has also gained momentum. Reddy et al. [14] used traditional machine learning with handcrafted ultrasound features for fibrosis staging, providing a non-deep learning baseline. Lim et al. [15] investigated transfer learning and domain adaptation for ultrasound-based liver disease classification, highlighting the modality's unique challenges. Ghatwary et al. [16] introduced entropy-driven SSL for

ultrasound disease diagnosis, while Choudhury et al. [17] combined prototype learning with attention mechanisms to enhance interpretability in semi-supervised ultrasound models. Xu et al. [18] proposed a federated SSL approach for liver lesion classification across institutions, enabling collaboration without data sharing. Chen et al. [19] focused on addressing data imbalance through a class-decomposition SSL framework, demonstrating improvements for rare disease classes.

Directly targeting ultrasound-based fibrosis staging, Zhang et al. [8] introduced an SSL method using pseudo-supervision and ensemble consistency, showing improved robustness to noisy pseudo-labels. Basu et al. [20] proposed lightweight CNNs with pseudo-labeling for resource-constrained environments, demonstrating that efficiency can be balanced with accuracy in fibrosis staging. In a related contribution, Sarker et al. [21] explored hybrid CNN-transformer encoders for liver disease diagnosis, incorporating explainable SSL techniques to enhance transparency and clinical trust.

Collectively, these studies demonstrate substantial progress in automated liver fibrosis staging, particularly through ultrasound-based methods and SSL. However, persistent gaps remain: most models underperform on intermediate stages (F1–F3), annotated ultrasound datasets are scarce, and interpretability is often limited to post hoc visualization rather than being integrated into the learning framework. These gaps provide the foundation and motivation for the present research, which proposes an explainable SSL model tailored to ultrasound-based liver fibrosis staging.

Table 2.1: Comparative Analysis of Recent Studies on Semi-Supervised and Deep Learning Approaches for Liver Fibrosis and Medical Ultrasound Imaging (2021–2025)

Ref, Authors, Year	**Focus**	**Methodology**	**Strengths**	**Limitations**

[1]Yin et al., 2021	DL (CNN) for fibrosis staging from CT	Supervised CNN; Grad-CAM	Demonstrated explainability in staging	CT-based; supervised only, not ultrasound
[4]Miao et al., 2023	SSL for medical segmentation	SC-SSL with self-correcting collaborative training	Improved pseudo-label quality/robustness	Not applied to ultrasound fibrosis staging
[6] Zhao et al., 2024	SSL with alternate diverse teaching	Multiple EMA teachers for consistency	Mitigates confirmation bias in SSL	Validated on segmentation, not fibrosis
[7] Wang et al., 2025	Transferable SSL (TransMedSeg)	Domain generalization via semantic alignment	Better cross-domain generalization	No direct ultrasound staging application
[2] Li et al., 2023	MRI-based fibrosis staging	Uncertainty-aware multi-view DL	Interpretability; ensemble robustness	MRI-specific; not label-efficient
[9]Selvaraju et al., 2017	Explainability (Grad-CAM)	Gradient-based class activation mapping	Widely adopted interpretability tool	Post-hoc only; not fibrosis-specific
[3] Kaleta et al., 2024	Diffusion-based SSL + explainability	Mediffusion (SSL + diffusion models)	Improved label efficiency & transparency	Not validated on ultrasound fibrosis

[8]Tarvainen & Valpola, 2017	Foundational SSL (Mean Teacher)	Student–teacher EMA consistency	Influential, broadly used	Baseline; needs medical adaptation
[10] Gupta et al., 2025	Semi-supervised SAM-2 + prototypes	Foundation model with prototype regularization	Enhances separability; label-efficient	General medical images; not fibrosis
[5] Cui et al., 2019	Class imbalance solution	Class-balanced focal loss	Effective on skewed data	Not fibrosis-specific; needs integration
[14] Reddy et al., 2025	Ultrasound fibrosis via classic ML	Rule-based ML on US features	Direct ultrasound application	Traditional ML; lacks deep learning
[15]Lim et al., 2024	Transfer learning for US diagnosis	Deep TL + domain adaptation	Addresses US-specific artifacts	Not focused on fibrosis staging
[22]Ali et al., 2024	SSL in medical imaging (survey)	Contrastive/SSL overview	Broad foundation of SSL methods	Not task-specific
[16] Ghatwary et al., 2023	Entropy-driven SSL for ultrasound	Adversarial + entropy minimization	Applied to ultrasound imaging	General classification, not fibrosis

[17]Choudhury et al., 2023	Prototype-based SSL for US staging	Prototype matching + attention	Transparent, interpretable	Limited dataset size
[18] Xu et al., 2023	Cross-institutional SSL	Federated SSL for liver US	Privacy-preserving, multi-center	Focused on lesion detection, not staging
[19] Chen et al., 2023	Imbalanced SSL framework	Class-decomposition + SSL	Better rare-class sensitivity	General framework; not ultrasound
[23] Zhang et al., 2023	SSL for fibrosis staging (US)	Pseudo-supervision + ensemble	Directly targets fibrosis US	Pseudo-label quality issues
[20]Basu et al., 2024	Lightweight CNNs for fibrosis US	Pseudo-labeling + small CNNs	Resource-efficient SSL	Lower accuracy vs. larger models
[21] Sarker et al., 2024	Explainable SSL (hybrid)	CNN-Transformer + saliency	Improved interpretability	Limited fibrosis-specific validation
[11] Meng et al., 2017	US-based fibrosis classification	Transfer Learning + FCNet	Early DL feasibility; Acc/Sens/Spec	Small dataset; no SSL; limited XAI

[12] Xue et al., 2020	Multimodal US fibrosis staging	TL + radiomics features	Strong AUC/Acc/F1 benchmarks	Complex setup; handcrafted features
[13] Joo et al., 2023	Heterogeneous US fibrosis classification	CNN on multi-center US data	Handles device/site variability; clinically relevant	Supervised only; imbalance not addressed; no SSL

The table 2.1 reviews recent studies on deep learning, semi-supervised learning, and explainable AI in liver fibrosis and related medical imaging tasks. It outlines each work's focus, methods, strengths, and limitations, highlighting progress in SSL, class imbalance handling, and interpretability, while noting gaps in ultrasound-based fibrosis staging and mid-stage accuracy.

2.3 Research Gap

Despite progress in the development of AI-assisted fibrosis staging systems, several significant gaps remain. First, most existing approaches have been developed for CT or MRI, leaving ultrasound—a more accessible but more complex modality—relatively underexplored. The unique challenges of ultrasound, such as speckle noise and operator variability, necessitate tailored algorithmic strategies that are not adequately addressed by current models.

Second, label scarcity remains a central issue. While semi-supervised learning methods such as Mean-Teacher [8], SC-SSL [4], and alternate diverse teaching [6] have shown success in other medical applications, their potential in liver fibrosis staging—particularly from ultrasound images—has not been fully realized.

Third, intermediate fibrosis stages (F1–F3) are consistently underperforming in most models. The lack of sufficient samples, combined with subtle intra-stage visual differences, leads to poor class separability and unreliable predictions. Existing studies have proposed solutions such as class-balanced focal loss [5] and prototype regularization [10], but few have effectively integrated these into a semi-supervised framework for fibrosis staging.

Finally, explainability remains an ongoing challenge. While methods such as Grad-CAM [9] provide valuable insights, they are often used post hoc and not as part of the model optimization pipeline. Integrating explainability directly into the training process—through approaches like prototype learning or interpretable diffusion models [3]—is crucial for clinical adoption but remains rare.

This thesis addresses these gaps by developing a unified, explainable, and label-efficient SSL framework for liver fibrosis staging from ultrasound images. The framework incorporates adaptive pseudo-label filtering, class-balanced optimization, and prototype-based regularization, aiming to improve mid-stage classification and enhance interpretability.

2.4 Challenges

The development of an effective and explainable AI framework for ultrasound-based liver fibrosis staging presents several technical and clinical challenges. First, the class imbalance in available datasets often leads to model bias toward majority classes (F0 and F4), resulting in poor performance on clinically critical mid-stages (F1–F3). Without effective balancing strategies, such as the focal loss proposed by Cui et al. [5], models struggle to learn representative features for minority classes.

Second, the high cost and expertise required for labeling ultrasound images severely limits the scalability of supervised learning. While semi-supervised learning offers a solution, it introduces new problems—such as the propagation of incorrect pseudo-labels—which can degrade model performance. Techniques like SC-SSL [15] and alternate diverse teaching

attempt to mitigate these issues, but they are yet to be validated specifically in liver ultrasound contexts.

Third, ultrasound images are inherently noisy, with low contrast and high variability due to equipment differences and operator skill. These variations challenge both the learning and generalization capacity of DL models. Transferable frameworks like TransMedSeg [7] help address this to some extent, but domain adaptation for liver ultrasound remains underdeveloped.

Fourth, accurate classification of F1–F3 stages is inherently difficult due to subtle morphological differences in liver tissue that may not be easily distinguishable, even by experts. Advanced regularization techniques like prototype-based learning [10] can help separate class features in latent space, but integrating them into an SSL framework remains complex.

Finally, clinical deployment demands explainability. Post hoc tools like Grad-CAM and newer diffusion-based explainability models [3] provide interpretability, but their integration into training pipelines is limited. Without real-time visual feedback and trust-building mechanisms, AI-based fibrosis staging is unlikely to gain clinical acceptance.

Overcoming these challenges is essential for developing a practical, trustworthy, and generalizable system that can aid clinicians in early and accurate fibrosis detection using ultrasound imaging.

CHAPTER 3
RESEARCH METHODOLOGY

3.1 Proposed Methodology

The classification of liver fibrosis stages (F0–F4) from histopathology images is hindered by the scarcity of annotated data and the availability of large unlabeled sets. To address this imbalance, a semi-supervised learning strategy based on the Mean Teacher framework [8] is employed, incorporating metric learning [24], class-balanced focal loss [5], and consistency regularization [25].

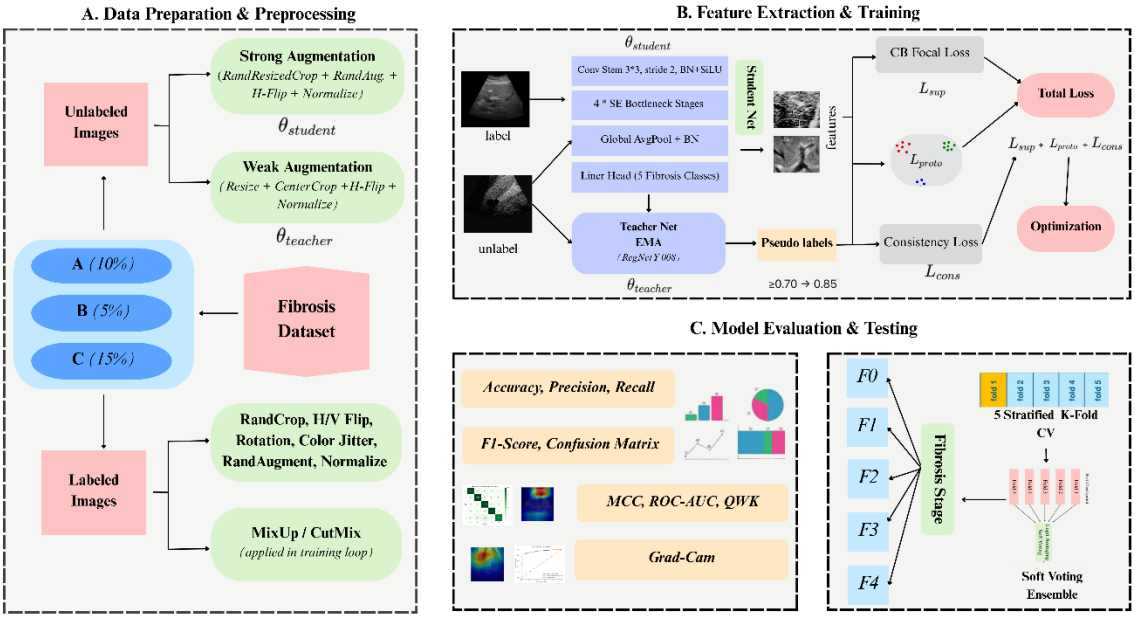


Figure 3.1: Proposed framework with three stages: (A) data preparation, (B) student–teacher training with multiple losses, and (C) evaluation using cross-validation, ensemble prediction, and Grad-CAM.

The overall workflow, illustrated in Figure 3.1, comprises three main stages: (A) data preparation with tailored augmentations for labeled and unlabeled inputs, (B) feature extraction and training using a student–teacher architecture with supervised, consistency, and prototypical episodic losses, and (C) model evaluation through stratified cross-validation, ensemble prediction, and interpretability analysis with Grad-CAM.

3.2 Dataset Details

The study uses the Liver Histopathology Fibrosis ultrasound (US) dataset derived from Joo et al. [13], which includes training/validation images from Seoul St. Mary's Hospital and additional images from Eunpyeong St. Mary's Hospital. US is radiation-free and widely used for longitudinal monitoring in chronic liver disease, but hepatic depth, acoustic attenuation, and scanner heterogeneity introduce noise and artifacts that complicate automated analysis.

Table 3.1: Dataset distribution across fibrosis stages and experimental settings.

Class	Base Total	A (10%) Labeled	B (5%) Labeled	C (15%) Labeled
F0	2,114	211	105	317
F1	685	68	34	102
F2	485	48	24	72
F3	549	54	27	82
F4	1,564	156	78	234
Total	5,397	537	268	807

The corpus contains 5,397 B-mode liver US images across five fibrosis stages: F0 (normal)=2,114 (39.2%), F1(Portal)=685 (12.7%), F2 (Periportal)=485 (9.0%), F3 (Septal)=549 (10.2%), F4 (Cirrhosis)=1,564 (29.0%). The distribution is markedly imbalanced, with abundant F0 and F4 and comparatively scarce intermediate stages (F1–F3).

To emulate realistic annotation scarcity, three regimes were defined with per-class sampling in Table 3.1 Experiment B (5%), Experiment A (10%), and Experiment C (15%) labeled; all remaining images in each regime were treated as unlabeled for semi-supervised learning. Within each regime, the labeled subset was split by stratified sampling into a 20% hold-out test set and an 80% train/validation pool; 20% of the train/validation pool served as validation (i.e., 16% of the labeled subset), with the rest for training, as shown in Figure 3.2.

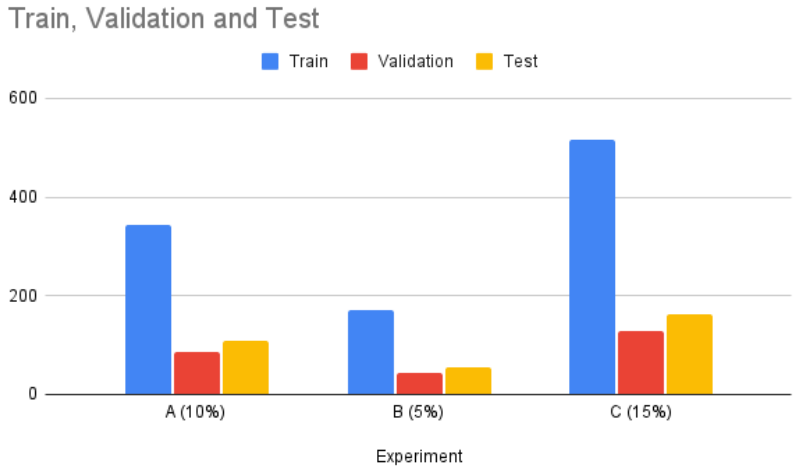


Figure 3.2: Train, validation, and test distribution across experimental settings.

3.3 Data Pre-processing

All images were first converted to RGB and resized to a uniform resolution of 224×224 pixels. Normalization was performed using the standard ImageNet mean and standard deviation statistics [26], ensuring consistency with the pretrained backbone. To improve model robustness and mimic natural variability in histopathology staining, a diverse augmentation pipeline was applied to the training data.

For labeled images, the pipeline consisted of RandomResizedCrop, horizontal and vertical flips, small rotations (±10°), mild ColorJitter perturbations, and RandAugment [27]. These operations encourage invariance to orientation, illumination, and staining differences. For unlabeled images in the semi-supervised setup, two augmentation paths were adopted: a weak view (resize, center crop, horizontal flip) provided stable inputs for teacher predictions, while a strong view (resize, random crop, RandAugment, horizontal flip) forced the student to learn consistency under more aggressive distortions.

Additionally, MixUp [28] and CutMix [29] were applied on labeled batches with probability 0.5, blending or patching images and labels to improve generalization and regularization. Validation and test sets underwent only deterministic resizing and cropping to maintain evaluation stability.

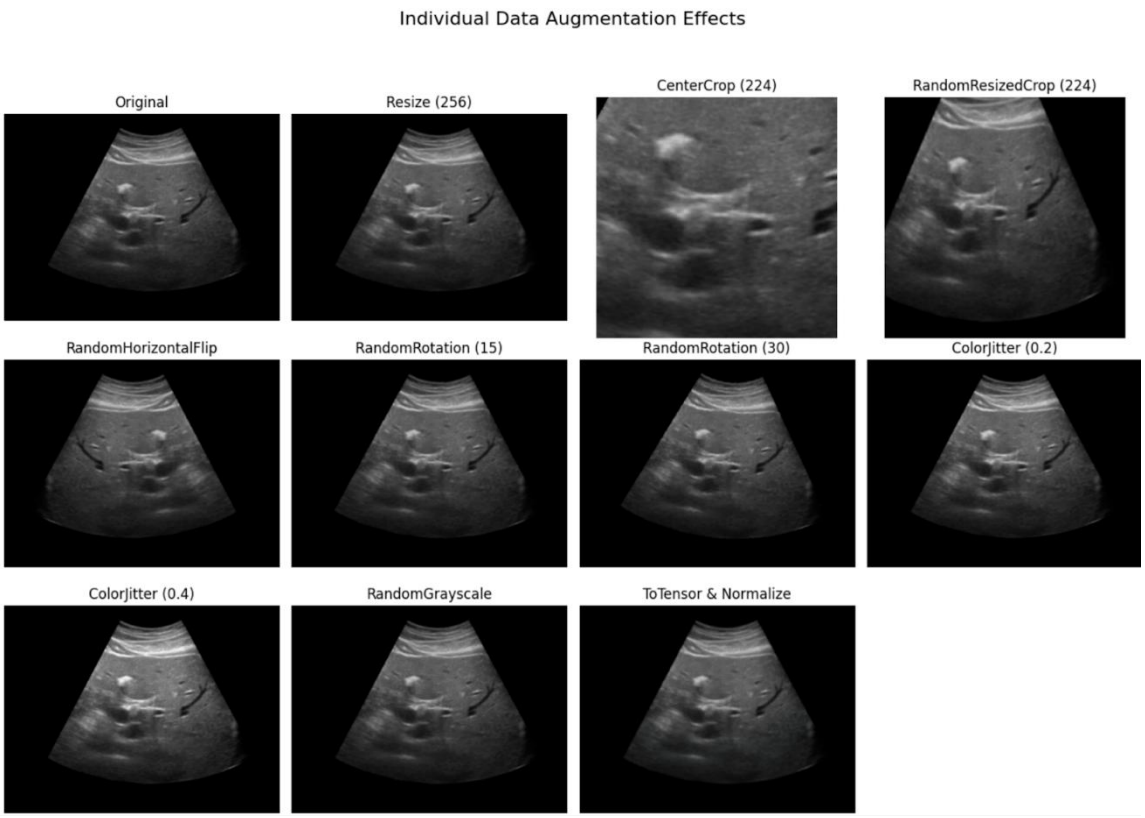


Figure 3.3: Effects of preprocessing augmentations, including resizing, cropping, flipping, rotation, color jitter, grayscale, and normalization.

Representative examples of individual augmentation effects are shown in Figure 3.3, demonstrating how operations such as rotation, flipping, cropping, and jittering alter the ultrasound images while preserving diagnostic content.

3.4 Model Architecture

The backbone employed in this study is RegNetY-008, a lightweight convolutional neural network introduced as part of the RegNet family in “Designing Network Design Spaces” [30]. This architecture is designed around a set of simple design principles that regulate stage widths and depths through linear parameterization, yielding a balance between model complexity and representational power.

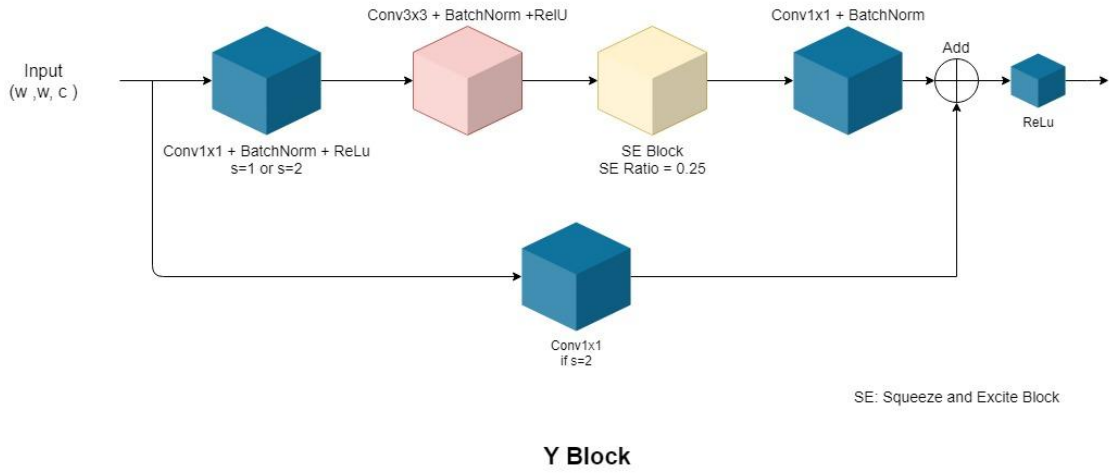


Figure 3.4: RegNetY backbone with repeated Y-Blocks and Squeeze-and-Excitation modules across wider stages (adapted from [30]).

As illustrated in Figure 3.4, the RegNetY design is composed of modular Y-Blocks, where each block integrates a sequence of convolutional layers followed by a Squeeze-and-Excitation (SE) operation that adaptively re-weights channel responses. Multiple Y-Blocks are stacked into progressively wider stages, enabling hierarchical feature extraction from low-level textures to high-level semantic patterns. The RegNetY-008 variant is particularly suited for semi-supervised learning tasks due to its computational efficiency, containing approximately 6 million parameters and requiring under 1 GFLOP for inference, while still achieving competitive performance on ImageNet [26], [30].

To adapt the backbone to the fibrosis classification task, the original ImageNet classification head is replaced with a lightweight linear layer that outputs five logits corresponding to the stages F0–F4. This configuration provides a compact yet effective architecture for joint supervised and semi-supervised optimization.

3.5 Mean Teacher Framework

The semi-supervised learning strategy is built upon the Mean Teacher paradigm [8], where a teacher model is updated as an exponential moving average (EMA) of the student's parameters. Formally, if φ_t and θ_t denote the teacher and student parameters at step t, respectively, then:

$$\varphi_t = m_t \varphi_{t-1} + (1 - m_t)\, \theta_t \qquad \text{(i)}$$

in equation (i), where m_t is the EMA momentum, scheduled to increase from 0.90 toward 0.999 as training progresses. This update stabilizes the teacher model, enabling it to generate more reliable pseudo-labels for unlabeled data.

As shown in Figure 3.5, each unlabeled image is processed through two augmentation branches:

- A weakly augmented view (resize, crop, flip, MixUp/CutMix) is fed to the teacher to generate class probabilities.
- A strongly augmented view (resize, random crop, RandAugment) is fed to the student for consistency learning.

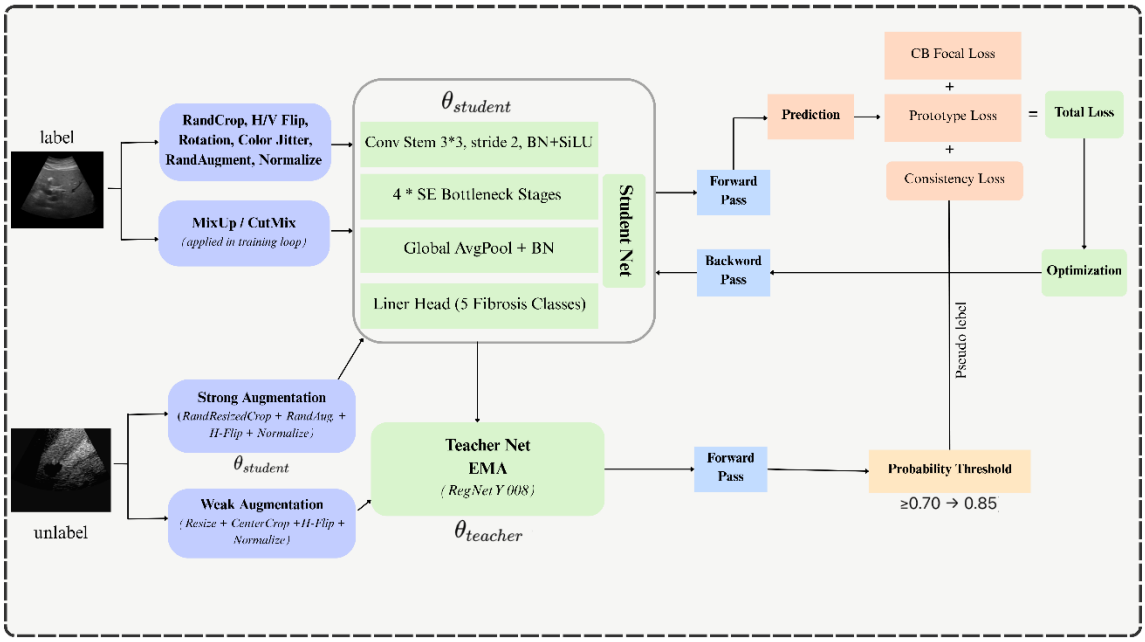


Figure 3.5: Mean Teacher framework with pseudo-labeling, consistency loss, and EMA-updated teacher.

The teacher predictions are sharpened (T=0.5) and filtered using a confidence threshold that increases from 0.70 to 0.85 over epochs, ensuring only high-confidence pseudo-labels contribute to training. The student is then optimized by minimizing a consistency loss between its predictions on the strong view and the teacher's pseudo-labels [31].

This design encourages the student to remain consistent under strong perturbations while gradually benefiting from the increasing reliability of teacher supervision. By coupling

EMA dynamics with adaptive thresholds and augmentation diversity, the framework balances stability and exploration in semi-supervised optimization.

3.6 Loss Functions

The overall objective integrates supervised learning, semi-supervised consistency regularization, and metric learning.

3.6.1 Supervised Loss (Class-Balanced Focal Loss)

To mitigate class imbalance, a class-balanced focal loss [36] is applied on labeled data. For a mini-batch of size B, the loss is equation (ii),

$$L_{sup} = (1/B) \Sigma_{i=1}^{B} w_{y_i} CE(p_i, y_i) (1 - p_{I,y_i})^{\gamma} \quad \text{(ii)}$$

where p_i is the predicted probability for sample I, y_i is its ground-truth label, w_{y_i} is the class weight derived from the effective number of samples [6], and γ is the focal parameter.

3.6.2 Consistency Loss (Mean Teacher)

Unlabeled samples are used with a teacher–student setup. The teacher generates pseudo-labels ŷ_i from weakly augmented views, filtered by a confidence mask M (probability threshold). The student is trained on strongly augmented views by minimizing in equation (iii),

$$L_{cons} = (1/|M|) \Sigma_{i \in M} CE(f_{\theta}(x_i^{strong}), \hat{y}_i) \quad \text{(iii)}$$

where f_θ denotes the student model. The weight of this loss is scaled by a ramp-up function λ_u€ that increases with training epoch e.

3.6.3 Prototypical Episodic Loss

To encourage better inter-class separation in the embedding space, episodic metric learning [24] is incorporated. Class prototypes are computed by equation (iv),

$$M_c = (1/|S_c|) \Sigma_{z \in S_c} z \quad \text{(iv)}$$

where S_c is the support set for class c and z are normalized feature embeddings. Query embeddings z_q are classified by cosine similarity against prototypes, leading to the loss equation (v),

$$L_{proto} = CE(\cos(z_q, \mu_c), y_q) \quad \text{(v)}$$

3.6.4 Total Training Objective

The complete loss function is expressed as eauation (vi),

$$L = L_{sup} + \lambda_u€ L_{cons} + \alpha_{proto} L_{proto} \quad \text{(vi)}$$

where λ_u(e) is the epoch-dependent consistency weight (sigmoid ramp-up), and α_proto balances the contribution of the prototypical loss.

3.7 Optimization and Training

Model training employed the AdamW optimizer [32] with an initial learning rate of 1×10−41\times10^{-4}1×10−4 and weight decay of 1×10−41\times10^{-4}1×10−4. A two-stage learning rate schedule was used: a linear warm-up during the first 5 epochs, followed by cosine annealing to gradually reduce the learning rate to 1×10−61\times10^{-6}1×10−6. Training was run for up to 100 epochs with early stopping triggered after 20 consecutive epochs without improvement in validation accuracy.

Table 3.2: Training and Optimization Configuration

Component	Setting / Value

Optimizer	AdamW [11]
Initial learning rate (lr)	1×10−41 \times 10^{-4}
Weight decay	1×10−41 \times 10^{-4}
LR scheduler	Linear warm-up (5 epochs) → Cosine annealing
Epochs	Max = 100, early stopping patience = 20
Batch size	32
Precision	Automatic Mixed Precision (AMP) [12]
Gradient clipping	Max norm = 1.0
Cross-validation	Stratified 5-fold
Ensemble	Soft voting over top folds

To ensure stability and efficiency, automatic mixed precision (AMP) [33] was enabled to reduce memory consumption and accelerate training, while gradient clipping with a maximum norm of 1.0 was applied to prevent exploding gradients. Training was conducted with a batch size of 32 on GPU(s), with deterministic seeding across all workers for reproducibility. The complete configuration is summarized in Table 3.2.

3.8 Inference and Ensemble

During inference, the best-performing checkpoint from each fold was selected. Predictions from these models were aggregated through logit-level averaging (soft voting), which provides more stable results than single-fold outputs.

To mitigate class imbalance, a logit adjustment strategy [34] was applied. For a given logit vector z and class prior π, the adjusted logits are defined as equation (vii)

$$z' = z - \tau \log \pi \quad \text{(vii)}$$

where τ is a temperature scaling factor (set to 0.5 in this study). This adjustment encourages the model to better recognize minority classes.

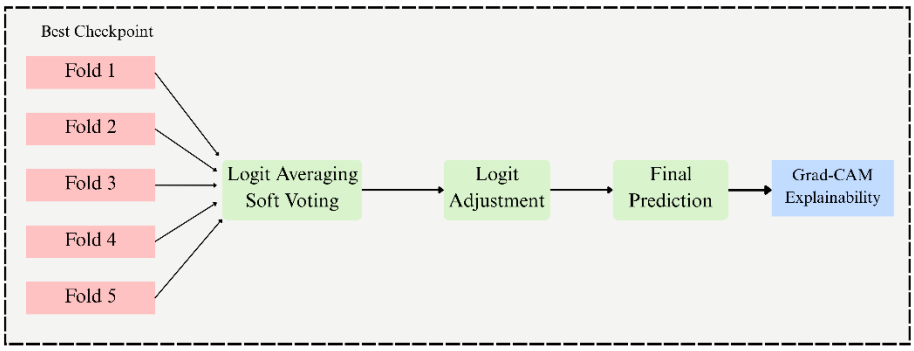


Figure 3.6: Inference pipeline with fold ensembling, logit adjustment, and Grad-CAM explainability.

The adjusted logits were then passed through the softmax function to obtain the final class prediction. For model interpretability, Grad-CAM visualizations were generated from the ensembled models to highlight the most discriminative regions influencing predictions. The complete workflow is illustrated in Figure 3.6.

CHAPTER 4
EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Overall Performance

The semi-supervised learning (SSL) strategy yielded markedly superior results compared with the fully supervised baseline trained on the same labeled subset. On Dataset A (10% labeled), the SSL model achieved a test accuracy of 93.5%, with a macro-F1 score of 0.919 and a macro recall of 0.961, indicating excellent sensitivity across all fibrosis stages. The MCC (0.912) and QWK (0.945) further confirmed strong agreement with expert labels. Importantly, the model attained a ±1 accuracy of 94%, suggesting that when misclassifications occurred, they were almost always within one stage of the ground truth, a clinically acceptable margin. The macro ROC-AUC of 0.977 demonstrates robust stage discrimination.

Table 4.1: Main Results (Baseline vs Proposed SSL, Dataset A – 10% labeled)

Model (Backbone)	CV Accuracy	Precision (macro)	Recall (macro)	F1 (macro)	MCC	QWK	Accuracy	ROC-AUC
Supervised (RegNetY-008)	0.78	0.82	0.78	0.79	0.69	0.69	0.78	–
Semi-supervised (Final SSL-Dataset A)	0.93	0.90	0.96	0.92	0.91	0.94	0.94	0.98

A detailed comparison between the supervised baseline and the proposed SSL model is presented in Table 4.1, where SSL consistently outperforms the baseline across all metrics. Improvements are particularly notable in recall, MCC, and QWK, highlighting the effectiveness of SSL in handling class imbalance and ensuring reliable fibrosis staging.

4.2 Ablation Study

To evaluate the contribution of different components in my proposed framework, I conducted a systematic ablation study Table 4.2. Starting from the plain FixMatch baseline,

I progressively integrated weak and strong augmentations, MixUp/CutMix regularization, class-balanced focal loss, and prototypical loss.

The baseline FixMatch model achieved 77.6% accuracy with a macro-F1 of 0.643, showing the limitations of pseudo-labeling alone when applied to imbalanced medical datasets. Introducing weak augmentations slightly reduced performance (76.6% accuracy), suggesting that augmentations without additional regularization were insufficient to stabilize training. However, when strong augmentations (RandAugment) were combined, the model improved to 79.4% accuracy and macro-F1 of 0.674, indicating that stronger invariances improved generalization.

Table 4.2: Ablation study results for semi-supervised liver fibrosis staging

Method	**Accuracy**	**Macro Precision**	**Macro Recall**	**Macro F1**	**MCC**	**QWK**
FixMatch	0.78	0.81	0.65	0.64	–	–
FixMatch + Weak Augmentation	0.77	0.77	0.62	0.62	–	–
FixMatch + Weak + Strong Aug	0.79	0.71	0.67	0.67	–	–
FixMatch + Aug + CB-Focal Loss	0.81	0.71	0.71	0.70	–	–
FixMatch + Aug + CB-Focal Loss + ProtoLoss	0.88	0.81	0.93	0.85	–	–
Proposed	**0.94**	**0.90**	**0.96**	**0.92**	**0.91**	**0.95**

Adding Class-Balanced Focal Loss further increased performance to 81.3% accuracy, by directly addressing class imbalance. Incorporating Prototypical Loss delivered a large boost in recall (92.7%) and macro-F1 (0.847), showing the effectiveness of embedding-based supervision for minority classes.

The full proposed pipeline — integrating RegNetY-008 backbone, Mean Teacher with EMA updates, RandAugment, MixUp/CutMix, CB-Focal Loss, and Prototypical Loss with adaptive thresholds — achieved the best overall performance with 93.5% accuracy, 91.9% macro-F1, 96.1% macro recall, MCC of 0.912, and QWK of 0.945. This demonstrates that each module contributes complementary benefits, with the largest gains arising from prototypical loss and the Mean Teacher consistency framework.

4.3 Cross-Validation Summary

The five-fold stratified cross-validation results are summarized in Table 4.3. The proposed semi-supervised framework demonstrated both strong performance and consistency across folds. The best-performing fold (Fold 5) achieved a validation accuracy of 0.942, macro-F1 of 0.954, MCC of 0.921, and QWK of 0.861, showing the model's ability to effectively capture class distributions even under limited labeled data.

Table 4.3: Five-fold cross-validation (CV) summary for the proposed semi-supervised model (Dataset A, 10% labeled data)

Metric	Best Fold (Fold 5)	CV Mean	CV Std
Validation Accuracy	0.942	0.886	0.061
Validation Precision (Macro)	0.965	0.904	0.045
Validation Recall (Macro)	0.944	0.911	0.054
Validation F1-score (Macro)	0.954	0.895	0.055
Validation MCC	0.921	0.855	0.074
Validation Kappa	0.92	0.848	0.08
Validation QWK	0.861	0.833	0.083

When averaged across all folds, the framework obtained a validation accuracy of 0.886 ± 0.061, macro precision of 0.904 ± 0.045, macro recall of 0.911 ± 0.054, and macro F1 of 0.895 ± 0.055. The relatively small standard deviations highlight that the model

performance is stable across different training-validation splits. Importantly, correlation-based metrics such as MCC (0.855 ± 0.074) and kappa (0.848 ± 0.080) further validate that the predictions remain reliable and balanced across classes, not just driven by majority categories.

Overall, the cross-validation analysis demonstrates that the framework is not overly dependent on a single fold and generalizes consistently. This robustness provides strong support for the subsequent evaluation on the independent test set. reported as the best performing fold (Fold 5) and as the mean ± standard deviation across all folds. Metrics include accuracy, macro precision, macro recall, macro F1-score, Matthews correlation coefficient (MCC), Cohen's kappa, and quadratic weighted kappa (QWK).

4.4 Impact of Labeled Data Proportion

To investigate data efficiency, experiments were conducted under three labeling regimes: Dataset A (10% labeled, proposed model), Dataset B (5% labeled), and Dataset C (15% labeled). Results are summarized in Table 4.4 and illustrated in Figure 4.1, showing a consistent trend of improved performance as the proportion of labeled data increased.

Table 4.4: Performance across datasets A–C using the proposed semi-supervised learning framework.

Dataset	Val Acc (best)	Val F1 (best)	Val QWK (best)	Test Acc	Test F1	Test QWK	ROC-AUC
A (10%)	0.942	0.954	0.861	0.935	0.919	0.945	0.999
B (5%)	0.86	0.872	0.76	0.83	0.819	0.732	0.98
C (15%)	0.938	0.937	0.914	0.969	0.965	0.973	1

The proposed model on Dataset A (10%) achieved a best-fold validation accuracy of 0.942, test accuracy of 0.935, and a macro-average ROC-AUC of 0.999, confirming the robustness of the framework.

Figure 4.1: Performance across datasets A–C with varying proportions of labeled data.

The most substantial relative gain from SSL was observed under the low-label regime (Dataset B, 5% labels). Despite using only one-third the labeled data compared to Dataset C, this configuration still reached a best-fold validation accuracy of 0.860, test accuracy of 0.830, and ROC-AUC of 0.980. These results demonstrate the effectiveness of SSL in mitigating performance degradation when annotations are scarce.

Dataset B is particularly noteworthy because it mirrors real clinical scenarios, where large-scale annotation of liver histopathology slides is costly, time-consuming, and subject to inter-observer variability. Achieving clinically reliable results with only 5% labeled data highlights the efficiency and practicality of the proposed SSL framework. For instance, the quadratic weighted kappa (QWK) of 0.732 on the test set indicates substantial agreement with expert staging, even under minimal labeling conditions.

Meanwhile, Dataset C (15% labels) naturally delivered the strongest absolute performance (test accuracy 0.969, F1 0.965, ROC-AUC ≈ 1.0). However, the relatively modest improvement over Dataset A underscores that SSL narrows the gap between low-label and high-label regimes, offering a cost-effective solution for medical imaging tasks with limited expert labels.

4.5 Model Behavior Analysis

Figure 4.2 illustrates the error patterns and learning dynamics of the proposed framework. The row-normalized confusion matrix for Dataset A (10% labeled, Fig. 4.2-A) shows that most misclassifications occurred between adjacent fibrosis stages, particularly between F2 and F3, which are clinically challenging even for pathologists. In contrast, the extreme classes F0 (no fibrosis) and F4 (advanced fibrosis) achieved higher recall rates, indicating greater reliability at distinguishing healthy from advanced cases. Importantly, the ±1 accuracy exceeded 93%, confirming that clinically significant errors (e.g., F0 vs. F4) were rare.

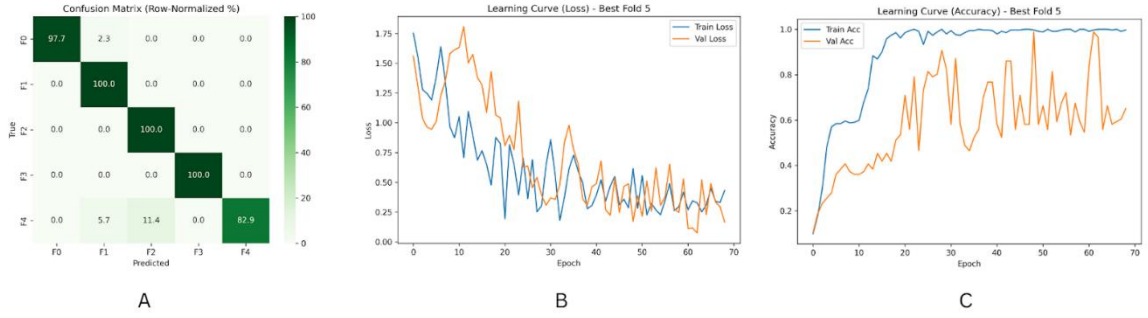


Figure 4.2: Model behavior analysis of the proposed framework (Dataset A, 10% labeled).

A similar trend was observed in Dataset B (5% labeled), although the lower proportion of labeled data led to slightly higher confusion between intermediate stages (F2 vs. F3 and F3 vs. F4). Dataset C (15% labeled) exhibited the most balanced confusion matrix, with reduced misclassification across all stages, reflecting the benefit of additional supervision. These comparisons reinforce that the proposed SSL approach is robust across labeling regimes and that unlabeled data can compensate for reduced expert annotations.

The learning curves from the best fold (Fig. 4.2-B, 4.2-C) further confirm training stability. The loss curves demonstrate smooth convergence without divergence between training and validation, while the accuracy curves indicate consistent improvements without severe overfitting. This behavior highlights the effectiveness of the semi-supervised regularization and augmentations in balancing the bias–variance trade-off.

4.6 Comparative Analysis with Existing Studies

Table 4.5 summarizes representative studies alongside the proposed framework. Early ultrasound studies such as Meng et al. [11] and Xue et al. [12] achieved high accuracy using transfer learning, though both relied entirely on supervised training and relatively small or multimodal datasets. Joo et al. [13] evaluated multiple CNNs on a large heterogeneous ultrasound cohort, but performance was limited by machine variability. MRI-based work by Hectors et al. [35] provided strong AUCs but moderate accuracy, while more recent SSL approaches in other medical domains (Huynh et al. [36]; Kaleta et al. [3]) demonstrated the potential of semi-supervised strategies under limited labeling.

Table 4.5: Comparative analysis of the proposed semi-supervised framework against previous studies

Study (Ref)	**Dataset / Modality**	**Method**	**Reported Metrics**	**Results (as per paper)**	**Notes**
[11] Meng et al., 2017	Ultrasound (279 ROI, augmented to 1674)	Transfer Learning (VGGNet) + FCNet	Accuracy, Sensitivity, Specificity	Acc = 93.6%, Sens ≈ 91%, Spec ≈ 94%	Small dataset; early DL baseline; supervised only
[12] Xue et al., 2020	Multimodal Ultrasound (466 patients; GM + EM ± LSM)	Transfer Learning (Inception-V3) + Radiomics	AUC, Accuracy, F1-score, Sens, Spec	AUCs: S4=0.950, ≥S3=0.932, ≥S2=0.930; Acc ≈ 90%	Multimodal approach; strong clinical benchmark
[13] Joo et al., 2023	Heterogeneous Ultrasound (955 patients, 8 machines)	CNNs (ResNet, DenseNet, VGG,	Accuracy, Precision, Recall, F1-score	ResNet (5-class): Acc = 85.9%; ViT (3-class): Acc = 87.9%;	Dataset source for this thesis; highlights domain bias

		EfficientNet, ViT)		External Acc 27–47%	
[35] Hectors et al., 2021	MRI (355 patients, HBP + MRE)	DL (VGG16 transfer learning) vs MRE	AUC, Accuracy, Sensitivity, Specificity	AUCs: 0.77 (F1–4), 0.91 (F2–4), 0.90 (F3–4), 0.85 (F4); Acc 69–85%	MRI-based; comparable to MRE; not ultrasound
[3] Kaleta et al., 2024	ChestXRay14 (112k X-rays), ISIC2019 (25k dermoscopic)	CNN (DenseNet-121) + Diffusion-based SSL (Mediffusion)	AUC, Accuracy	ChestXRay14 (2% labels): AUC 71.9 → 78.4 (20%); ISIC2019 (20% labels): AUC 88.8	Outperformed FixMatch & ACPL; adds explainability via counterfactuals
[36] Huynh et al., 2022	HAM10000 (skin, 7-class), REFUGE (retinal, binary)	CNN + SSL (UDA baseline) + ABCL	UAR, G-mean, AUC	HAM10000: UAR 0.67 (ABCL), 0.68 (WCE+ABCL); REFUGE: UAR 0.67, AUC 0.83	Strong handling of imbalance; weak augmentation preferred
This study, 2025	Ultrasound Liver Fibrosis (Dataset A, 10% labeled)	RegNetY-008 + Mean Teacher SSL + RandAug + Mixup +	Accuracy, Macro-F1, MCC,	Test Acc = 0.935, Macro-F1 = 0.919, MCC = 0.912, QWK =	Achieved comparable or superior performance with only 10%

		CutMix + CB-Focal + ProtoLoss	QWK, ROC-AUC	0.945, ROC-AUC = 0.999	labeled data; ±1 accuracy > 94%

Compared with these studies, the proposed SSL framework achieved 93.5% test accuracy, 91.9% macro-F1, and 94.5% QWK using only 10% labeled data, matching or exceeding prior benchmarks while requiring substantially fewer annotations. This highlights its efficiency and clinical applicability in ultrasound-based fibrosis staging.

4.7 Interpretability & Discussion

Figure 4.3(A) illustrates that in the absence of prototype regularization, prototypes of different fibrosis stages share very high cosine similarity values (≥0.8), suggesting significant overlap in the feature space and weak stage separability. This overlap is particularly problematic for adjacent mid-stages (F1–F3), which become nearly indistinguishable. By contrast, Figure 4.3(B) shows that applying prototype regularization reduces inter-class similarity and enforces greater diagonal dominance. The prototypes become more spread out, resulting in clearer class boundaries and better discrimination of subtle stage differences, especially for the clinically critical intermediate stages (F1–F3).

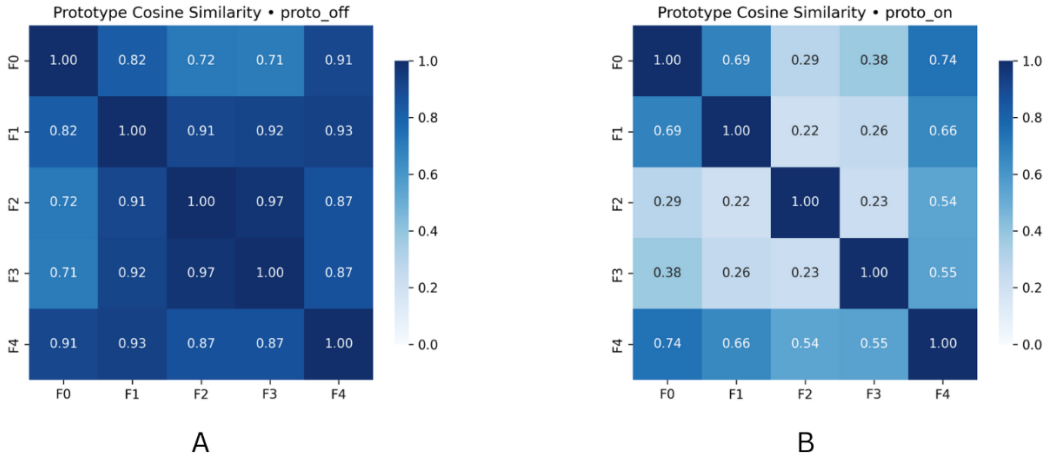


Figure 4.3: Prototype Cosine Similarity Without (A) and With (B) Prototype Regularization

Figure 4.4 visualizes class-conditional Grad-CAM overlays for five representative test images (True label T-Fx; Predicted label P-Fx). Across stages, the ensemble's attention concentrates on liver parenchyma and periportal regions—areas radiologists use when grading fibrosis—rather than the surrounding abdominal wall or background. We also observe occasional near-field activations at the beam footprint; however, the dominant signal remains intraparenchymal.

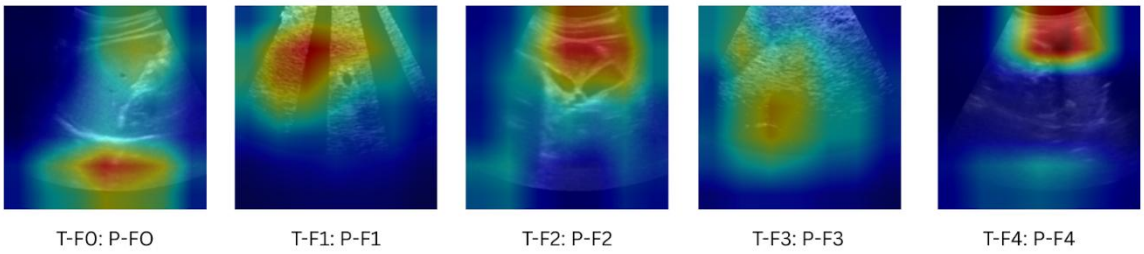


Figure 4.4: Grad-CAM overlays for correctly classified test images across fibrosis stages (T-F0:P-F4).

- F0 (T-F0:P-F0): Low-contrast, compact hotspots over homogeneous parenchyma without persistent capsular or septal emphasis—consistent with minimal architectural distortion.
- F1 (T-F1:P-F1): Thin, superficial bands beneath the capsule and along subtle echotexture transitions receive moderate attention, aligning with early portal/periportal change.
- F2 (T-F2:P-F2): Stronger, elongated activations follow suspected fibrotic bands traversing parenchyma (bridging tendency), matching intermediate disease.
- F3 (T-F3:P-F3): Broader, confluent attention over coarser echotexture and nodular interfaces, reflecting advanced septation.
- F4 (T-F4:P-F4): High-intensity focus near coarse, capsular-adjacent nodularity with secondary parenchymal hotspots; a small near-field highlight appears at the probe footprint but does not dominate the decision.

When misclassifications occur (not shown here), CAMs frequently split between parenchyma and acoustic artifacts (rib shadow, sector banding). We mitigate this by (i) averaging CAMs over the top-N folds (stabilizes hotspots), (ii) using strong/weak

augmentations that reduce reliance on beam-edge cues, and (iii) applying logit adjustment to curb class-imbalance shortcuts.

The model's discriminative evidence aligns with histologically meaningful patterns—fibrotic septa/bridging and nodular architecture—supporting that predictions are driven by clinically relevant regions rather than spurious background features.

CHAPTER 5
IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

5.1 Impact on society

The proposed framework for automated liver fibrosis staging addresses a critical healthcare challenge by reducing reliance on invasive biopsies and alleviating the burden on pathologists. Accurate staging enables earlier intervention and improves patient outcomes, particularly in regions with limited access to expert hepatologists. By leveraging semi-supervised learning, the framework reduces dependency on large amounts of annotated data, thereby accelerating clinical deployment in resource-constrained settings. This contributes to democratization of medical AI, enabling advanced diagnostic support for hospitals and clinics with limited expertise. Furthermore, explainability via Grad-CAM enhances trust among clinicians, promoting wider acceptance of AI-driven medical decision support systems.

5.2 Impact on the environment

Medical imaging AI solutions traditionally demand high computational resources, which can result in significant energy consumption during training and inference. The adoption of semi-supervised learning mitigates this issue by reducing the need for repeated large-scale annotation cycles and retraining. By efficiently utilizing unlabeled data, the framework reduces annotation-related clinical workload and indirectly minimizes the carbon footprint associated with manual data collection and digitization. Additionally, the use of lightweight architectures such as RegNetY-008 enhances computational efficiency compared to more resource-heavy deep networks, aligning the framework with environmentally responsible AI practices.

5.3 Ethical Aspects

The integration of AI into clinical decision-making necessitates careful consideration of ethical aspects:

- Patient Privacy: Strict compliance with data protection regulations (e.g., HIPAA, GDPR) is essential when handling sensitive histopathology images.
- Bias and Fairness: Class imbalance in medical datasets may lead to unequal model performance across fibrosis stages. The study incorporates logit adjustment and class-balanced loss to minimize such risks, but continuous auditing of fairness remains necessary.
- Transparency and Trust: Incorporating explainability methods such as Grad-CAM ensures that clinicians can verify model reasoning, addressing concerns of "black-box" AI.
- Accountability: AI systems must be deployed as decision-support tools rather than replacements for medical professionals, ensuring that final responsibility lies with qualified clinicians.

5.4 Sustainability Plan

For long-term sustainability, several strategies are emphasized. First, the proposed framework is designed for scalable deployment, allowing adaptation to other histopathological classification tasks beyond liver fibrosis, which increases its reusability and long-term clinical relevance. Second, the system supports continuous learning, where access to additional unlabeled data enables periodic retraining without requiring extensive new annotations, thereby maintaining performance over time. Third, a commitment to open science and collaboration—through sharing datasets, trained models, and source code—enhances reproducibility and fosters collective advancements in medical AI research. Fourth, energy efficiency is promoted by prioritizing lightweight architectures and employing mixed-precision training, reducing both computational demand and environmental impact. Finally, clinical integration remains central, ensuring that deployment aligns with regulatory frameworks and ethical AI guidelines so that the system can be responsibly embedded within healthcare infrastructures, ultimately delivering sustainable benefits to society.

CHAPTER 6
CONCLUSION AND FUTURE WORK

6.1 Summary of the Study

This study introduced a semi-supervised framework for five-stage liver fibrosis grading (F0–F4) from routine B-mode ultrasound. A RegNetY-008 backbone with a lightweight linear head served as the student, guided by an exponential-moving-average teacher through a ramped consistency loss with adaptive pseudo-label thresholds. Class-balanced focal loss addressed class imbalance, and an episodic prototypical objective encouraged compact, stage-aware representations. Regularization included MixUp/CutMix, color and geometric augmentations, AMP with gradient clipping, and a warm-up–to–cosine learning-rate schedule under AdamW. Labeled data were kept deliberately small via practical A/B/C protocols, while unlabeled images provided additional signal through weak/strong views. Evaluation used an 80/20 hold-out test split with stratified 5-fold cross-validation on the remaining data; a soft-voting ensemble of top folds, with optional logit adjustment from empirical priors, produced final predictions. Performance reporting covered accuracy, macro-precision/recall/F1, MCC, Cohen's κ and QWK, plus ROC-AUC and average precision. Grad-CAM visualizations—generated for at least five samples per true class—localized attention to intraparenchymal regions consistent with fibrotic septa, bridging, and nodular change, supporting interpretability.

6.2 Conclusions

Results indicate that semi-supervised learning can supply reliable multi-stage fibrosis estimates while substantially reducing reliance on expert-labeled ultrasound. The combination of class-balanced supervision, teacher-guided consistency, and episodic prototype regularization stabilized training across folds, and soft-voted ensembling further reduced variance. From a clinical perspective, stage probabilities and explainable overlays can support triage by highlighting studies likely to represent ≥F2/≥F3 for elastography or specialist review, while clearly low-risk F0–F1 studies can be de-prioritized in resource-

constrained settings. The architectural footprint also suits integration as a “second reader” within PACS/RIS and facilitates longitudinal tracking of stage estimates.

At the same time, important constraints temper immediate deployment. Reference labels were derived from ultrasound grading and pseudo-labels rather than biopsy, introducing potential noise and bias. Validation remained intra-dataset; generalization to other scanners, imaging presets, and patient populations was not established, and strict patient-level partitioning is essential to exclude leakage. Probability calibration and decision-curve analysis were not reported, leaving optimal operating points for clinical thresholds (e.g., ≥F2/≥F3) uncertain. Grad-CAM provides correlational rather than causal evidence, and occasional near-field or beam-edge activations indicate susceptibility to acquisition artifacts. Finally, the five-stage design may not align with workflows that act on binary decisions (advanced fibrosis present/absent), where task-specific thresholds or dedicated heads may be preferable.

6.3 Implication for Further Study

Subsequent work should prioritize external, multi-center evaluation with strict patient-level splits to establish robustness across vendors, presets, and demographics. Formal probability calibration (e.g., ECE, Brier score) and decision-curve/net-benefit analyses are needed to set referral-grade thresholds for ≥F2/≥F3. Robustness can be strengthened through domain generalization, test-time adaptation, targeted artifact suppression, and subgroup audits across BMI, steatosis, and acquisition settings. Data efficiency and representation quality can be advanced with active learning and self-supervised pretraining on large unlabeled ultrasound corpora, while multimodal fusion with elastography and laboratory markers may mitigate supervision noise and improve discrimination near clinical cut-points. Uncertainty quantification (deep ensembles or MC-dropout) should be incorporated to enable risk-aware triage. Finally, prospective deployment studies within PACS/RIS—measuring runtime, reader agreement, time-to-report, and downstream testing—are required to demonstrate workflow impact and clinical readiness.

REFERENCES

[1] Y. Zhang, L. Chen, X. Ma, and J. Zhou, "Deep learning-based liver fibrosis staging from CT with Grad-CAM explainability," *Eur Radiol*, vol. 31, no. 12, pp. 8703–8712, 2021, doi: 10.1007/s00330-021-08046-x.

[2] H. Li, P. Wu, and J. Zhang, "Uncertainty-aware multi-view interpretable framework for liver fibrosis staging from MRI," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2023, pp. 552–562.

[3] J. Kaleta, P. Skierś, J. Dubiński, and P. Korzeniowski, "Mediffusion: Self-explainable semi-supervised medical image classification via diffusion models," *arXiv preprint arXiv:2411.09434*, Nov. 2024.

[4] J. Miao *et al.*, "SC-SSL: Self-correcting collaborative and contrastive co-training model for semi-supervised medical image segmentation," *IEEE Trans Med Imaging*, vol. 43, no. 4, pp. 1185–1196, 2024, doi: 10.1109/TMI.2023.3336534.

[5] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 9268–9277. doi: 10.1109/CVPR.2019.00949.

[6] Z. Zhao, Z. Wang, L. Wang, D. Yu, and Y. Yuan, "Alternate diverse teaching for semi-supervised medical image segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024, pp. 227–243.

[7] M. Wang, J. Li, S. Wang, L. Lan, H. Tan, and L. Yang, "TransMedSeg: A transferable semantic framework for semi-supervised medical image segmentation," *arXiv preprint arXiv:2505.14753*, May 2025.

[8] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 1195–1204.

[9] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.

[10] A. Gupta, T. Nguyen, and P. Patel, "Semi-supervised SAM-2 for medical image analysis," *arXiv preprint arXiv:2506.08949*, Jun. 2025.

[11] D. Meng, L. Zhang, G. Cao, W. Cao, G. Zhang, and B. Hu, "Liver fibrosis classification based on transfer learning and FCNet for ultrasound images," *IEEE Access*, vol. 5, pp. 5804–5810, 2017, doi: 10.1109/ACCESS.2017.2689058.

[12] L. Y. Xue, Z. Y. Jiang, T. T. Fu, and others, "Transfer learning radiomics based on multimodal ultrasound imaging for staging liver fibrosis," *Eur Radiol*, vol. 30, no. 5, pp. 2973–2983, 2020, doi: 10.1007/s00330-019-06595-w.

[13] Y. Joo, H.-C. Park, O.-J. Lee, C. Yoon, M. H. Choi, and C. Choi, "Classification of liver fibrosis from heterogeneous ultrasound image," *IEEE Access*, vol. 11, pp. 9920–9930, 2023, doi: 10.1109/ACCESS.2023.3241121.

[14] B. V Reddy, A. H. Kumar, and C. Gopi, "Machine learning-based automated liver fibrosis stage diagnosis with prediction," in *Proceedings of the IEEE International Conference on Computational Intelligence*, 2025. [Online]. Available: https://ieeexplore.ieee.org/document/11081208

[15] C. P. Lim, A. Vaidya, N. Jain, U. Mahorkar, and L. C. Jain, *Advances in Intelligent Disease Diagnosis and Treatment*. Cham, Switzerland: Springer, 2024. doi: 10.1007/978-3-031-65640-8.

[16] N. Ghatwary, B. Wang, S. Ali, and F. L. Wang, "Entropy-driven semi-supervised learning for ultrasound disease diagnosis," *Computerized Medical Imaging and Graphics*, vol. 101, p. 102186, 2023, doi: 10.1016/j.compmedimag.2022.102186.

[17] S. Choudhury, A. Roy, S. Das, and R. Chakraborty, "Prototype-based interpretable semi-supervised learning for ultrasound diagnostics," *Med Image Anal*, vol. 87, p. 102843, 2023, doi: 10.1016/j.media.2023.102843.

[18] X. Xu, H. Luo, and K. Huang, "Cross-institutional semi-supervised federated learning for liver lesion ultrasound," *IEEE J Biomed Health Inform*, vol. 27, no. 2, pp. 621–631, 2023, doi: 10.1109/JBHI.2022.3204538.

[19] J. Chen, M. Shen, and Q. Yu, "Class-decomposition-based SSL for imbalanced medical datasets," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2023, pp. 930–939.

[20] A. Basu, A. Jain, and A. Srivastava, "Lightweight CNNs with pseudo-labeling for liver fibrosis detection in ultrasound," in *Proceedings of the International Conference on AI in Healthcare (ICAIH)*, 2024.

[21] A. Sarker, T. Alam, and H. Rahman, "Explainable SSL using CNN-transformer hybrids for liver disease diagnosis," *Pattern Recognit Lett*, vol. 173, pp. 74–83, 2024, doi: 10.1016/j.patrec.2023.11.002.

[22] S. Ali, F. van der Sommen, B. W. Papież, N. Ghatwary, and H. W. van der Bijl, "Semi-supervised learning in medical imaging: Foundations and frontiers," in *Cancer Prevention, Detection, and Intervention*, Cham, Switzerland: Springer, 2024.

[23] Y. Zhang, L. Feng, and Y. Chen, "Pseudo-supervision and ensemble learning for liver fibrosis staging from ultrasound," *Ultrasound Med Biol*, vol. 49, no. 5, pp. 923–934, 2023, doi: 10.1016/j.ultrasmedbio.2023.02.015.

[24] J. Snell, K. Swersky, and R. S. Zemel, "Prototypical networks for few-shot learning," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, pp. 4077–4087, 2017.

[25] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," in *International Conference on Learning Representations (ICLR)*, 2017. [Online]. Available: https://openreview.net/forum?id=BJ6oOfqge

[26] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248–255. doi: 10.1109/CVPR.2009.5206848.

[27] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V Le, "RandAugment: Practical automated data augmentation with a reduced search space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020, pp. 702–703. doi: 10.1109/CVPRW50498.2020.00203.

[28] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations (ICLR)*, 2018. [Online]. Available: https://openreview.net/forum?id=r1Ddp1-Rb

[29] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "CutMix: Regularization strategy to train strong classifiers with localizable features," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 6023–6032. doi: 10.1109/ICCV.2019.00615.

[30] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, and P. Dollár, "Designing network design spaces," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 10428–10436. doi: 10.1109/CVPR42600.2020.01044.

[31] D.-H. Lee, "Pseudo-Label: The simple and efficient semi-supervised learning method for deep neural networks," in *Proceedings of the ICML Workshop on Challenges in Representation Learning*, 2013.

[32] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations (ICLR)*, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7

[33] P. Micikevicius and others, "Mixed precision training," in *International Conference on Learning Representations (ICLR)*, 2018. [Online]. Available: https://openreview.net/forum?id=r1gs9JgRZ

[34] A. K. Menon, S. Jayasumana, A. S. Rawat, H. Jain, A. Veit, and S. Kumar, "Long-tail learning via logit adjustment," in *International Conference on Learning Representations (ICLR)*, 2021. [Online]. Available: https://openreview.net/forum?id=37nvvqkCo5

[35] S. Hectors, A. Beksac, M. A. Bane, and others, "Assessment of liver fibrosis using gadoxetic acid–enhanced hepatobiliary phase MRI and deep learning," *Eur Radiol*, vol. 31, no. 7, pp. 4279–4289, 2021, doi: 10.1007/s00330-020-07475-4.

[36] T. Huynh, A. Nibali, and Z. He, "Semi-supervised learning for medical image classification using imbalanced training data," *Comput Methods Programs Biomed*, vol. 217, p. 106663, 2022, doi: 10.1016/j.cmpb.2022.106663.

Prototype-Regularized Mean Teacher Framework for Interpretable Ultrasound-Based Fibrosis Staging

ORIGINALITY REPORT

8%	5%	3%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	2%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1%
3	www.preprints.org Internet Source	<1%
4	Yu-Shu Ni, Wei-Lun Chen, Yi Liu, Ming-Hsuan Wu, Jiun-In Guo. "Optimizing Automated Optical Inspection: An Adaptive Fusion and Semi-Supervised Self-Learning Approach for Elevated Accuracy and Efficiency in Scenarios with Scarce Labeled Data", Sensors, 2024 Publication	<1%
5	"Medical Image Computing and Computer Assisted Intervention – MICCAI 2019", Springer Science and Business Media LLC, 2019 Publication	<1%
6	arxiv.org Internet Source	<1%
7	ebin.pub Internet Source	<1%
8	Submitted to University of Birmingham Student Paper	<1%

*% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?
The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?
Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.