
A Comparative Study of Deep Learning Models for Hepatic Steatosis Detection and Grade Classification

Dissertation submitted to

Visvesvaraya National Institute of Technology, Nagpur

in fulfilment of the requirement for the award of degree in

Master of Technology in Communication System Engineering

by

Pooja Wadhvane (MT23CMN003)

under the guidance of

Prof. Saugata Sinha



**Department of Electronics and Communication
Visvesvaraya National Institute of Technology
Nagpur 440 010 (India)**

June 2025

Department of Electronics and Communication Engineering
Visvesvaraya National Institute of Technology, Nagpur



DECLARATION

I, **Pooja Wadhvane**, hereby declare that this dissertation titled “*A Comparative Study of Deep Learning Models for Hepatic Steatosis Detection and Grade Classification*” is carried out by me under the guidance of Dr. Saugata Sinha in the **Department of Electronics and Communication Engineering** of Visvesvaraya National Institute of Technology, Nagpur. The work is original and has not been submitted earlier whole or in part for the award of any degree/diploma at this or any other Institution / University.

Pooja Wadhvane

Date: _____

CERTIFICATE

This is to certify that the dissertation titled “*A Comparative Study of Deep Learning Models for Hepatic Steatosis Detection and Grade Classification*” is submitted by **Pooja Wadhvane (MT22CMN003)** in partial fulfillment of the requirements for the award of **Master of Technology in Communication System Engineering**, VNIT Nagpur. The work is comprehensive, complete, and fit for final evaluation.

Dr. Ashwin Kothari

Professor and Head of Department
Electronics and Communication
Engineering
VNIT Nagpur

Prof. Saugata Sinha

Professor
Electronics and Communication
Engineering
VNIT Nagpur

Non-Plagiarism Certificate

I, **Pooja Wadhvane (MT23CMN003)** hereby declare that I am the sole author of the thesis entitled "**A Comparative Study of Deep Learning Models for Hepatic Steatosis Detection and Grade Classification**" and that neither any part of the thesis nor the whole of the thesis has been submitted to any University or Institution for obtaining any degree / diploma / academic award.

Certified that the M.Tech Dissertation titled "**A Comparative Study of Deep Learning Models for Hepatic Steatosis Detection and Grade Classification**" submitted by **Pooja Wadhvane**, Enrollment No. **MT23CMN003**, ID No. **30330** has been checked for plagiarism using **TURNITIN** software and the overlap is found within the prescribed limit. A summary of the report is given below:

PLAG_REPORT.png

Date:

Pooja Wadhvane

Place:

MT23CMN003

ACKNOWLEDGEMENTS

I express my sincere gratitude to my supervisor, Dr. Saugata Sinha, for his invaluable guidance, constant encouragement, and insightful feedback throughout the course of this project. I extend my heartfelt thanks to the faculty, staff, and Prof. Ashwin Kothari, H.o.D of the **Department of Electronics and Communication Engineering**, *Visvesvaraya Institute of Technology*, Nagpur for providing the necessary academic environment and resources that made this project possible.

I extend my thanks to my peers and friends for their continuous support and constructive discussions, which greatly enriched this work. Lastly, I am deeply indebted to my family for their unwavering encouragement and understanding throughout my academic journey. This project has been a valuable learning experience and a stepping stone towards my future endeavors.

Pooja Wadhvane

MT23CMN003

ABSTRACT

A hallmark of non-alcoholic fatty liver disease, hepatic steatosis necessitates precise identification and grading for efficient therapeutic management. This study compares eight deep learning models UNet, ResUNet++, RAPUNet, MIFONet, DeepLabV3+, MedSAM, HarDNet-MSEG and DUCK-Net — for segmenting lipid droplets and classifying steatosis grades in H&E stained liver biopsy images.

Using the dataset from Heinemann et al.(2019), accessed via the OSF repository(<https://osf.io/p48rd/>), whole- slide images were preprocessed into 352x352 patches. Stain normalization was applied to regularize color variations and data augmentation to enhance dataset diversity. Models were trained on an NVIDIA RTX A5000 GPU, leveraging architectures from lightweight convolutional networks to transformer-based frameworks. Segmentation performance was evaluated using Dice and IoU metrics, while classification accuracy and F1-score assessed grade prediction (none, mild, moderate, severe).

The models demonstrated varying performance, with lightweight models like RAPUNet and HarDNet-MSEG showing efficiency for resource-constrained settings, and transformer-based models like MedSAM offering robust feature extraction.

Contents

Declaration	i
Certificate	ii
Non-Plagiarism Certificate	iii
Acknowledgements	v
Abstract	vi
List of Figures	x
List of Tables	xi
1 Introduction	1
1.1 Background	1
1.2 Problem Statement	2
1.3 Objectives	3
1.4 Scope	3
1.5 Limitations	4
2 Literature Review	5
2.1 Traditional Segmentation Methods	5
2.2 Deep Learning in Medical Imaging	6
2.3 Gaps in Existing Research	6
2.4 Deep learning Models for Binary Segmentation in Medical Image Analysis	6
2.4.1 Encoder-Decoder Convolutional Neural Network (CNN)-Based Models	7
2.4.2 Dilated (Atrous) Convolution-Based Models	7
2.4.3 Attention-Based Models	9
2.4.4 Transformer-Based and Hybrid CNN-Transformer Models	10
2.4.5 Multi-Scale and Pyramid Feature Fusion Models	11

2.4.6	Lightweight and Efficient Models	12
2.4.7	Prompt-Based and Interactive Segmentation Models	14
2.4.8	Conclusion and Summary	15
3	Methodology	17
3.1	Dataset	17
3.2	Preprocessing	17
3.3	Training	18
3.4	Deep Learning Models	19
3.4.1	Encoder Decoder based Model	20
3.4.1.1	UNet[25]:	20
3.4.1.2	ResUNet++:	22
3.4.2	Attention Based Models	25
3.4.2.1	MIFONet:	25
3.4.2.2	RAPUNet:	28
3.4.3	Transformer Based	31
3.4.3.1	MedSAM	31
3.4.4	Multi-Scale Fusion Models	33
3.4.4.1	HarDNet-MSEG [14]	33
3.4.5	Dilated Convolution Based	36
3.4.5.1	DeepLabV3+	36
3.4.6	Light Weight Models	39
3.4.6.1	DUCK-Net[27]:	39
3.5	Optimizer and loss functions	43
3.5.1	Adam Optimizer	43
3.5.2	BCEDiceLoss	43
3.5.3	BCELogitsLoss	44
3.6	Evaluation Metrics for Binary Segmentation	44
3.6.1	Accuracy	45
3.6.2	Precision (Positive Predictive Value)	45
3.6.3	Recall (Sensitivity or True Positive Rate)	45
3.6.4	F1 Score	45

3.6.5	Dice Coefficient (Sørensen–Dice Index)	46
3.6.6	Intersection over Union (IoU or Jaccard Index)	46
4	Results and Discussion	47
4.1	Experimental Setup	47
4.2	Results	47
4.2.1	Segmentation Results	47
4.2.2	Classification Results	48
4.2.3	Comparison of ground truth with predicted images	48
4.2.4	Classification using Confusion Matrix	57
4.3	Discussion	61
5	Conclusion and Future Work	62
5.1	Conclusion	62
5.2	Future Work	63
	Bibliography	67

List of Figures

3.1	UNet Architecture	20
3.2	RAPUNet architecture[20]	29
3.3	HarDNet-MSEG framework[14]	34
3.4	A single DUCK block[10]	40
3.5	DUCK-Net architecture[10]	41
4.1	Prediction results of all 4 grades of hepatic steatosis using UNet model . .	49
4.2	Prediction results of all 4 grades of hepatic steatosis using ResUNet++ model	50
4.3	Prediction results of all 4 grades of hepatic steatosis using MIFONet model	51
4.4	Prediction results of all 4 grades of hepatic steatosis using RAPUNet model	52
4.5	Prediction results of all 4 grades of hepatic steatosis using MedSAM model	53
4.6	Prediction results of all 4 grades of hepatic steatosis using HarDNet-MSEG model	54
4.7	Prediction results of all 4 grades of hepatic steatosis using DeepLabV3+ model	55
4.8	Prediction results of all 4 grades of hepatic steatosis using DUCK-Net model	56
4.9	Confusion matrix for fat Percentage grades using using UNet model . . .	57
4.10	Confusion matrix for fat Percentage grades using using ResUNet++ model	57
4.11	Confusion matrix for fat Percentage grades using using MIFONet model .	58
4.12	Confusion matrix for fat Percentage grades using using RAPUNet model .	58
4.13	Confusion matrix for fat Percentage grades using using MedSAM model .	59
4.14	Confusion matrix for fat Percentage grades using using HarDNet-MSEG model	59
4.15	Confusion matrix for fat Percentage grades using using DeepLabV3+ model	60
4.16	Confusion matrix for fat Percentage grades using using DUCK-Net model	60

List of Tables

2.1	Summary of Deep Learning Model Categories for Hepatic Steatosis Segmentation	16
3.1	Summary of RAPUNet architectural components	31
3.2	Summary of HarDMSEG Backbone Configuration	35
4.1	Performance Comparison of Deep Learning Models	47
4.2	Grade Classification Comparison of Deep Learning Models	48

CHAPTER 1

INTRODUCTION

1.1 Background

Hepatic steatosis also known as fatty liver disease is the abnormal buildup of fat within liver cells (hepatocytes). This condition is influenced by various factors including obesity, metabolic disorders like diabetes, and excessive alcohol intake. Alcoholic fatty liver disease (AFLD) and non-alcoholic fatty liver disease (NAFLD) are the two main categories into which it is typically divided. NAFLD alone impacts nearly a quarter of the world's population [31]. Without timely detection and management, fatty liver can develop into more serious liver diseases such cirrhosis, fibrosis, non-alcoholic steatohepatitis (NASH), or even liver cancer, increasing the risk of serious health complications. It is imperative to accurately detect and grade hepatic steatosis in order to prevent complications, monitor the disease, and initiate early intervention.

The gold standard for identifying and classifying hepatic steatosis at this time is liver biopsy. This technique provides a direct microscopic view of liver tissue, often stained using hematoxylin and eosin (H&E) for histopathological analysis. These stained images, typically stored in high resolution formats like PNG, allow pathologists to visually assess the proportion of fat-laden hepatocytes, enabling them to classify steatosis severity into categories such as none, mild, moderate, or severe[2]. To increase diagnostic consistency and efficiency, automated and objective techniques are required because manual biopsy image processing is time-consuming, subjective, and subject to inter-observer variability.

To address these challenges, deep learning has become a powerful tool in the field of medical image analysis, particularly for histological images due to its ability to learn complex patterns directly from raw data. Classical deep learning models, such as Fully Convolutional Networks (FCN), U-Net, and DeepLab, have demonstrated remarkable success in segmenting histological structures, while convolutional neural networks (CNNs) like ResNet excel in classification tasks [25, 8]. For hepatic steatosis, these models can automate the segmentation of fat droplets and classify steatosis severity, reducing diagnostic errors and enabling scalable analysis. The use of liver biopsy images in PNG format, as sourced from a research paper, provides a high-resolution dataset ideal for training such models, though challenges like limited dataset size and staining variability must be addressed.

Despite these advancements, applying deep learning to liver biopsy images presents challenges, including limited dataset sizes, variability in staining, and computational complexity. Classical models like U-Net are well-suited for small datasets, while modern architectures like vision transformers (ViT) may offer superior performance but require significant computational resources. A comparative study of these models is essential to identify the optimal approach for hepatic steatosis detection and grading, balancing accuracy, robustness, and clinical applicability. This thesis leverages a dataset of liver biopsy images to evaluate deep learning models, aiming to contribute to the development of automated diagnostic tools for hepatic steatosis.

1.2 Problem Statement

Hepatic steatosis detection through manual interpretation of medical imaging is laborious and subject to inter-observer variability. Convolutional neural networks (CNNs) are one type of deep learning model that has demonstrated promise in automating medical picture interpretation. To find the best model for hepatic steatosis detection and grade categorisation, it is necessary to methodically examine various deep learning architectures.

1.3 Objectives

The main goal of this thesis is to examine the effectiveness of deep learning models for the automated detection and grade classification of hepatic steatosis using liver biopsy images. By comparing classical and modern deep learning architectures, the study seeks to identify the most effective approach for clinical application. The specific objectives are as follows:

- **Detection of Hepatic Steatosis:** To assess deep learning models for precisely recognising fat droplets suggestive of hepatic steatosis in histological liver biopsy pictures.
- **Grade Classification:** To implement models that classify the severity of hepatic steatosis into clinically relevant grades (none, mild, moderate, severe) based on the proportion of fat-laden hepatocytes, ensuring alignment with histopathological standards.
- **Comparative Analysis:** To evaluate the effectiveness of traditional deep learning models (e.g., U-Net, FCN, DeepLab) and modern architectures (e.g., vision transformers) in terms of segmentation accuracy, classification performance, computational efficiency, and robustness to staining variability.
- **Clinical Applicability:** To assess the potential of these models for integration into clinical workflows, providing insights into their practical utility for pathologists in diagnosing and grading hepatic steatosis.

1.4 Scope

This study focuses on the application of deep learning models to hepatic steatosis detection using publicly available medical imaging datasets. This thesis does not address other liver-related conditions (e.g., fibrosis, hepatitis) or non-histological diagnostic methods. The scope is limited to transformer based and UNet based architectures and does not include other machine learning techniques or invasive diagnostic methods.

1.5 Limitations

While this study aims to provide a comprehensive evaluation of deep learning models for hepatic steatosis detection and grading, several limitations must be acknowledged:

1. **Limited Dataset Size:** The dataset, sourced from a research paper, may contain a limited number of liver biopsy images, which is common in histological studies due to the invasive nature of biopsies. This constraint increases the risk of overfitting and may limit model generalization.
2. **Staining and Image Variability:** Histological images often exhibit variations in staining (e.g., H&E intensity) and imaging conditions (e.g., microscope magnification), which can affect model performance. Preprocessing techniques may mitigate this but cannot fully eliminate variability.
3. **Lack of External Validation:** The study relies on a single dataset from a research paper, which may not represent diverse patient populations or imaging protocols. Without external validation, the generalizability of the models to other datasets remains uncertain.
4. **Computational Constraints:** Modern architectures like vision transformers require significant computational resources, which may pose challenges for training and deployment, particularly in resource-constrained clinical settings.
5. **Absence of Clinical Integration:** While the study evaluates model performance, it does not include real-world clinical testing with pathologists, limiting insights into practical implementation challenges, such as integration with existing diagnostic workflows.
6. **Annotation Dependency:** The quality of annotations determines the segmentation and classification accuracy provided in the dataset (e.g., pixel-level masks for fat droplets or steatosis grades). Inaccurate or incomplete annotations may compromise model performance.

CHAPTER 2

LITERATURE REVIEW

2.1 Traditional Segmentation Methods

Until deep learning emerged, conventional methods for analyzing histological liver images relied on handcrafted features and classical image processing techniques. These methods include:

- **Thresholding:** Otsu's method [23] segments fat droplets based on intensity thresholds, effective for high-contrast H&E images but sensitive to staining variations.
- **Region-Based Methods:** Region growing [1] and watershed algorithms [?] delineate fat droplets by grouping similar pixels, but they struggle with noise and require manual seed selection.
- **Clustering:** Fuzzy C-Means (FCM) clustering [3] groups pixels into fat and non-fat regions, offering robustness to noise but requiring predefined cluster numbers.
- **Active Contours:** Models like snakes [17] evolve curves to segment fat droplets, effective for clear boundaries but sensitive to initialization and noise.

These methods, while interpretable, are limited by their reliance on manual feature engineering and poor performance on variable histological images. For hepatic steatosis, they struggle with the complex textures of fat droplets and staining artifacts, highlighting the need for automated, data-driven approaches.

2.2 Deep Learning in Medical Imaging

With deep learning's introduction, automatic liver segmentation and hepatic steatosis has seen significant improvements over classical machine learning techniques. Medical image analysis presently makes extensive use of Convolutional Neural Networks (CNNs) because of their capacity to directly learn hierarchical features from input. The goal of early DL models was to improve accuracy, generalizability, and computational efficiency for liver segmentation and fat quantification.

Deep learning, particularly CNNs, has revolutionized medical image analysis. Models like ResNet [12], VGG [26], DenseNet [15], and EfficientNet [28] have been successfully utilised for tasks including disease classification, organ segmentation, and tumour identification. Recent studies have explored CNNs for hepatic steatosis detection, but comparative analyses across multiple architectures are limited.

2.3 Gaps in Existing Research

While previous studies have demonstrated the efficacy of individual CNN models, there is a lack of comprehensive comparisons across multiple architectures for hepatic steatosis detection and grading. Additionally, the impact of data preprocessing and augmentation on model performance remains underexplored.

2.4 Deep learning Models for Binary Segmentation in Medical Image Analysis

Binary segmentation is a fundamental task in medical imaging, particularly in applications like hepatic steatosis detection, where the objective is to delineate fat-affected regions from normal tissue. Many deep learning models have been presented over time, and they can be generically categorised into different architectural types. Each category reflects different strategies for learning spatial, contextual, and semantic information crucial for accurate

segmentation. The following subsections describe these categories in detail.

2.4.1 Encoder-Decoder Convolutional Neural Network (CNN)-Based Models

At the core of many medical picture segmentation methods are encoder-decoder structures. In order to reconstruct the pixel-wise segmentation map, these models usually comprise an expanding path (decoder) after a contracting path (encoder) to capture contextual information.

One of the most prominent examples is U-Net[25], originally proposed for biomedical image segmentation. Its encoder and decoder layers have skip connections, which protect spatial information that is lost during downsampling. By utilising both fine-grained spatial characteristics and high-level semantic information, this design allows the network to greatly increase segmentation accuracy. Variants like **U-Net++**, **ResUNet**[33] and **Attention U-Net** enhance performance by introducing nested skip connections, residual blocks, or attention mechanisms, respectively.

These encoder-decoder models are widely adopted in medical imaging due to their robustness and relatively low computational cost. They work especially well in situations when there is a dearth of labelled data, which is a frequent problem in the medical field. Their interpretability and strong performance in segmenting lesions or organs with well-defined boundaries make them a popular choice.

However, despite their strengths, these architectures may struggle to capture complex spatial dependencies and long-range contextual relationships within images. Recent research explores combining encoder-decoder models with transformer modules or graph-based methods to address these limitations.

2.4.2 Dilated (Atrous) Convolution-Based Models

Dilated (atrous) convolution has become an influential approach in deep learning models for semantic segmentation, particularly in medical imaging tasks such as hepatic steatosis

detection and grading. This technique expands the receptive field of convolutional layers by introducing gaps between kernel elements, controlled by a dilation rate. This enables extensive multi-scale contextual information to be captured by the network, without sacrificing spatial resolution or increasing the number of model parameters. The effective receptive field is given by:

$$\text{Receptive Field} = (k - 1) \times r + 1$$

where k is the kernel size. This property is particularly advantageous for hepatic steatosis analysis, where both local texture (early-stage microvesicular fat) and global context (diffuse macrovesicular fat) must be captured.

DeepLabv3 [7], a prominent model employing dilated convolutions integrates Atrous Spatial Pyramid Pooling (ASPP) to simultaneously extract features at various scales using different dilation rates. This multi-scale representation enhances the model's capacity to handle the heterogeneity inherent in steatotic liver tissue. **DeepLabv3+** [9] further extends this architecture by adding a lightweight decoder module that refines segmentation boundaries, improving the accuracy of fat droplet delineation, which is crucial for distinguishing subtle grade differences in hepatic steatosis.

The strengths of dilated convolution-based models lie in their effective multi-scale context aggregation, spatial resolution preservation, and improved boundary precision. These characteristics make them well-suited for segmenting diffuse and irregular steatotic regions in high-resolution histopathological images. However, they also involve greater computational complexity and demand significant processing resources, which can pose challenges for clinical deployment. Moreover, the relatively small size of annotated medical datasets necessitates the use of transfer learning and data augmentation to avoid overfitting and achieve reliable performance.

In summary, dilated convolution-based architectures like DeepLabv3 and DeepLabv3+ represent state-of-the-art techniques in hepatic steatosis segmentation and grading, balancing the need for contextual understanding with detailed spatial accuracy

2.4.3 Attention-Based Models

Attention-based models have gained significant traction in deep learning, specifically in the segmentation of medical images, because of their capacity to dynamically concentrate on the areas of an image that are most essential. Unlike traditional convolutional architectures that treat all spatial features equally, attention mechanisms selectively weight feature maps or spatial locations, enhancing important structures while suppressing irrelevant background information.

In segmentation tasks, attention modules are often integrated within encoder-decoder architectures (such as Attention U-Net), where they guide the decoder to focus on salient regions identified during encoding. This selective focus improves the model’s ability to delineate complex and subtle anatomical boundaries, which is crucial for accurately identifying lesions or pathological regions such as steatotic areas in liver histology.

Attention mechanisms can take various forms, including spatial attention that emphasizes specific image locations, channel attention that recalibrates feature responses across channels, or a combination of both. By incorporating these mechanisms, attention-based models can better handle challenges such as noise, variability in tissue appearance, and the presence of small or diffuse pathological features.

In this study, we employ **RAPUNet**[20] and **MIFONet** [16] —two advanced attention-based architectures. RAPUNet leverages residual attention modules within a pyramid U-Net structure to capture multi-scale features, improving precision in delineating small fat vacuoles. MIFONet integrates multi-scale information fusion with attention gates, enhancing robustness to staining variability commonly encountered in liver histology images.

The main advantages of attention-based models are improved segmentation accuracy, enhanced interpretability through attention maps, and robustness to heterogeneity in medical images. However, the added complexity may increase computational cost and require careful tuning to avoid overfitting, particularly when using small datasets for training, as is common in medical imaging.

Overall, attention-based architectures represent a powerful advancement in medical image

segmentation, enabling more precise and reliable detection and classification of pathological conditions like hepatic steatosis.

2.4.4 Transformer-Based and Hybrid CNN-Transformer Models

Transformer architectures were first created for natural language processing, but because of their capacity to collect global contextual information and represent long-range interdependence, they have lately been modified for medical image segmentation. Transformers use self-attention processes that allow the model to comprehend relationships throughout the entire image, in contrast to standard convolutional neural networks, which mainly concentrate on local receptive fields, improving segmentation performance especially in cases with complex or diffuse anatomical structures.

Convolutional layers' local feature extraction skills and transformers' global context modelling capabilities are combined in hybrid CNN-Transformer models. Typically, these models use CNN encoders to extract fine-grained spatial features and transformer blocks to capture high-level semantic relationships. Notable examples include TransUNet [6], which integrates a CNN backbone with a transformer encoder and a U-Net style decoder, effectively balancing spatial detail preservation with global context modeling. **MedSAM** [21], derived from the Segment Anything Model, leverages vision transformers (ViT) for prompt-based segmentation, enabling flexible automated and semi-supervised workflows for fat droplet detection in PNG images. Similarly, Swin-Unet [4] leverages the hierarchical and shifted window mechanism of the Swin Transformer within a U-shaped encoder-decoder architecture, enabling efficient multi-scale feature representation while maintaining manageable computational complexity. This thesis trains the MedSAM model on H&E-stained liver biopsy images and compares its segmentation performance and computational efficiency against other deep learning models for hepatic steatosis detection and grade classification, highlighting the advantages of large-scale vision foundation models in histological analysis.

Other models such as UNETR [11] utilize pure transformer encoders combined with CNN decoders, while CoTr [30] incorporates convolutional operators alongside transformer mod-

ules to enhance segmentation accuracy. These hybrid architectures demonstrate significant improvements in segmenting complex medical images, including hepatic steatosis regions, by effectively modeling both fine local details and long-range spatial dependencies.

Despite their impressive performance, transformer-based models typically require substantial computational resources and extensive training data sets to prevent overfitting, which can be a limitation in medical imaging domains where annotated datasets are often scarce. Transfer learning and data augmentation strategies are commonly employed to mitigate these challenges.

2.4.5 Multi-Scale and Pyramid Feature Fusion Models

Multi-scale and pyramid feature fusion models are designed to capture features at different spatial resolutions, which is crucial for segmenting complex and heterogeneous patterns such as fat vacuoles in H&E-stained liver biopsy images. These models aim to preserve both high-level contextual information and fine-grained specifics by integrating features extracted at multiple scales. In medical image segmentation, where target structures often vary widely in size and shape, this approach improves robustness and accuracy.

Architectures like Feature Pyramid Networks (FPN) and models integrating spatial pyramid pooling (e.g., PSPNet[34]) use hierarchical feature representations to enhance semantic segmentation. In a top-down fashion, FPN combines high-resolution, spatially exact features with low-resolution, semantically powerful ones, helping to refine boundaries. PSPNet integrates context data from different spots using pyramid pooling modules, improving segmentation of objects with variable scales.

For this research, we utilise **HarDNet-MSEG** [14], a multi-scale segmentation model that integrates efficient feature fusion with the lightweight yet powerful **HarDNet** [5](Harmonic DenseNet) backbone. HarDNet optimizes the trade-off between accuracy and computational cost by using harmonically spaced dense connections, significantly reducing redundancy in feature propagation. HarDNet-MSEG leverages this backbone and incorporates hierarchical feature aggregation to enhance its ability to detect both small, sparse fat droplets and large steatotic regions in H&E-stained liver biopsy images. The multi-scale

architecture enables the model to learn rich spatial features across different resolutions, improving segmentation accuracy in the presence of staining variability and complex tissue morphology.

In histopathological analysis of hepatic steatosis, models using such fusion strategies can better identify both small, dispersed fat droplets and large confluent steatotic regions. These methods offer improved localization and boundary delineation, which are critical for accurate fat quantification and grading. However, the increased architectural complexity and memory consumption may limit their application in resource-constrained environments or when training data is scarce, necessitating the use of pretraining or transfer learning techniques.

2.4.6 Lightweight and Efficient Models

Lightweight and efficient neural networks are increasingly important in deep learning, especially in medical imaging where hardware limitations and the need for real-time analysis constrain model design. These models strive to minimize computation, memory footprint, and latency while preserving diagnostic-level accuracy. Their evolution is largely influenced by the demand for deployment on mobile, embedded, or clinical systems with limited computational capacity.

Building such models typically involves architectural optimizations. One method is network pruning, which discards redundant weights or filters that have negligible impact on output, thereby shrinking the model and accelerating inference. Another technique is quantization, where high-precision operations are approximated using lower-bit representations (like 8-bit integers), resulting in faster execution with little degradation in performance.

Beyond pruning and quantization, architectural innovations play a central role. Dilated convolutions, grouped convolutions, and depthwise separable convolutions are frequently used to maintain representational power while reducing the number of operations. Efficient encoder-decoder structures, residual connections, and attention mechanisms are also adapted in a more compact form to retain high-level semantic features without significantly increasing the model's footprint.

Early downsampling is another important idea, which lowers the computational effort by reducing the spatial resolution of feature maps early in the network’s development. This is typically combined with specialized modules such as pyramid pooling, context aggregation, or feature fusion units, which are designed to maximize context-awareness while minimizing resource demands.

DUCK-Net[27], for example, incorporates dilated convolutions and compact residual blocks to achieve effective multi-scale feature representation with minimal computational overhead. **ENet**[24] and **ESPNet**[22] are explicitly designed for real-time semantic segmentation, utilizing strategies such as early downsampling and efficient spatial pyramid modules to balance speed and accuracy. **BiSeNet**[32] employs a dual-branch architecture that separates spatial detail from contextual understanding, enabling both fast and precise segmentation. Collectively, these models demonstrate how lightweight architectural innovations can fulfill the dual requirements of computational efficiency and clinical applicability, particularly in tasks such as hepatic steatosis detection and grading.

In this study, DUCK-Net is selected as the primary model and is trained on liver biopsy images. Its effectiveness is assessed and contrasted with that of other cutting-edge segmentation models to assess its effectiveness for hepatic steatosis analysis.

In the medical domain, where high-resolution images and detailed structural information are common, designing lightweight models poses additional challenges. Yet, achieving this balance is crucial — especially for real-time diagnostic tools, telemedicine applications, and systems used in remote or under-resourced healthcare settings. Lightweight architectures must be carefully tailored to ensure that critical clinical features are preserved and accurately segmented or classified.

Overall, lightweight and efficient models represent a pragmatic direction in deep learning, aiming to close the distance between cutting-edge performance and real-world usability. Their integration into medical imaging workflows offers a path toward scalable, accessible, and fast diagnostic solutions, without the heavy resource demands of traditional deep architectures.

2.4.7 Prompt-Based and Interactive Segmentation Models

Prompt-based and interactive segmentation models represent a paradigm shift in medical image analysis by integrating human input into the model’s decision-making process. Unlike fully automated segmentation methods, these models are designed to respond to user-provided cues or prompts—such as text instructions, bounding boxes, scribbles, or points—to direct and improve segmentation results. This interactive approach is particularly valuable in medical imaging, where anatomical variability, low contrast, and ambiguous boundaries often challenge conventional models.

The core idea behind these models is to combine machine efficiency with human expertise, allowing clinicians or annotators to influence segmentation results in real time. By incorporating prompts as additional input channels or conditioning mechanisms, the models can adapt their predictions based on spatial guidance, reducing the reliance on large-scale annotated datasets and increasing flexibility across tasks and modalities.

Recent advances have leveraged transformer-based architectures and foundation models pre-trained on diverse datasets, enabling generalization to unseen anatomical regions and imaging modalities. These models, such as the Segment Anything Model (**SAM**) [19], can generate high-quality segmentations from minimal input, making them particularly promising for medical applications where annotation manually takes a lot of time and calls for subject knowledge. SAM introduced a prompt-based segmentation framework for general-purpose images, utilising a backbone of Vision Transformers (ViT) to process various kinds of user prompts, including boxes, masks, and points— enabling adaptable and high-fidelity segmentation across domains. Its medical adaptation, **MedSAM** [21], extends this architecture by enhancing the ViT-based encoder and prompt-conditioned decoder to better suit the structural complexity of medical images. MedSAM supports semi-supervised learning and interactive clinical annotation workflows, offering accurate segmentation of histological structures such as fat droplets in PNG-formatted liver biopsy images. By facilitating guided segmentation with minimal supervision, MedSAM exemplifies the potential of prompt-based models to accelerate and refine medical image analysis tasks

In medical image segmentation, prompt-based models are especially advantageous for tasks that demand precise localization, such as lesion boundary delineation or fine-grained tissue classification. By enabling semi-automated workflows, these models help reduce the manual annotation burden while preserving expert oversight and control. This interactive framework makes them highly suitable for clinical environments, where iterative refinement and expert validation are essential.

Overall, prompt-based and interactive segmentation models represent a hybrid approach that synergizes the scalability and speed of artificial intelligence with the nuanced judgment of human experts. As these models continue to evolve, they hold great promise for accelerating annotation processes, enhancing segmentation accuracy, and fostering greater trust and acceptance of AI-assisted tools within medical practice.

2.4.8 Conclusion and Summary

Each category of segmentation models brings unique strengths and trade-offs. Encoder-decoder CNNs remain the backbone of many medical image segmentation techniques due to their performance and usability. Attention and transformer-based models improve context understanding and are more suited for complex or large-scale datasets. Meanwhile, lightweight and promptable models address the growing demand for real-time, flexible, and interactive medical tools. Knowing these classifications facilitates the choice of suitable architectures for tasks like hepatic steatosis segmentation and classification, where image texture, resolution, and annotation availability vary significantly.

Table 2.1: Summary of Deep Learning Model Categories for Hepatic Steatosis Segmentation

Category	Strengths	Limitations	Examples
Encoder-Decoder Models	- Good spatial localization - Easy to train with small datasets	-May lose global context, -Sensitive to small structures	U-Net, U-Net++, ResUNet, ResUNet++ SegNet
Attention-Based Models	- Focuses on important regions - Robust to noise and variability	- Increases model complexity - Higher computational cost	Attention U-Net, MIFONet, RAPU-Net
Transformer-Based Models	- Captures long-range dependencies - Handles complex textures	- Requires large datasets - Computationally expensive	TransUNet, Swin-Unet, MedSAM, UNETR
Multi-Scale Fusion Models	- Handles variable object sizes - Improves boundary segmentation	- May introduce redundancy - Can be harder to optimize	HarDNet-MSEG, HRNet, PSPNet, RAPUNet
Lightweight Models	- Fast and efficient - Ideal for real-time and low-resource environments	- May trade off some accuracy - Less suited for very complex tasks	DUCK-Net, ENet, ESPNet, BiSeNet
Prompt-Based / Interactive Models	- Adaptable and generalizable - Requires minimal labeled data	- Needs prompt input - May not be fully autonomous	SAM, MedSAM, RITM, iGPT

CHAPTER 3

METHODOLOGY

3.1 Dataset

This study utilizes a dataset comprising of 400 images from a liver biopsy stained with haematoxylin and eosin (H&E), originally compiled by Heinemann et al. (2019) [13] and made accessible to the general public via the Open Science Framework (OSF) repository <https://osf.io/p48rd/>[13]. The dataset is partitioned into testing, training and validation subsets in proportions of 20%, 60%, and 20%, respectively. This dataset is well-suited for the current investigation due to its high-quality histological imaging and its established clinical relevance for hepatic steatosis evaluation.

3.2 Preprocessing

The H&E-stained liver biopsy images from Heinemann et al. (2019) [13], accessed via the OSF repository (<https://osf.io/p48rd/>), were preprocessed to prepare them for deep learning model training. Whole-slide images were divided into 352x352 pixel patches to match the input requirements of the models, with corresponding binary masks extracted for segmentation tasks. Patches and masks were resized to 352x352 using bilinear and nearest-neighbor interpolation, respectively, to preserve image quality and binary values. Stain normalization was applied using Macenko’s method to standardize color variations in H&E images, ensuring consistent hematoxylin and eosin staining across samples. To address the limited dataset size, data augmentation was employed, including random hor-

horizontal and vertical flips, rotations up to 30 degrees, scaling between 0.8 and 1.2 times, brightness and contrast adjustments of $\pm 10\%$, and Gaussian noise, applied consistently to images and masks. Pixel intensities were standardised to the interval $[0, 1]$ by dividing by 255. For segmentation, annotations were formatted as single-channel binary masks (0 for background, 1 for fat droplets). For the task of grade classification, each image was assigned an integer label indicating the severity of hepatic steatosis, as derived from the associated metadata. The grading was defined as follows: Grade 0 (none) for fat content less than 5%, Grade 1 (mild) for fat content between 5% and 33%, Grade 2 (moderate) for fat content between 33% and 66%, and Grade 3 (severe) for fat content greater than 66%. This classification schema facilitates a standardized assessment of fatty infiltration severity across the dataset. Care was taken to ensure a balanced representation of all steatosis grades and diverse fat droplet structures in each subset, thereby enabling robust model training and fair evaluation.

Images are preprocessed to enhance quality and consistency:

- **Resizing:** Images are resized to 352x352 to fit the model.
- **Normalization:** Pixel values are scaled to $[0,1]$.
- **Data Augmentation:** Techniques such as rotation, flipping, ColorJitter, RandomAffine, Resize and Normalize are used to increase dataset diversity and prevent overfitting.

3.3 Training

Ground truth annotations were prepared for the 400 H&E-stained liver biopsy images from Heinemann et al. (2019)[13] to make deep learning model training effective for hepatic steatosis segmentation and grading. Manual polygonal annotations of fat droplets were performed using the VGG Image Annotator (VIA), with the resulting annotations saved in JSON format. These JSON files were then processed via a custom Python script to generate binary masks of size 352x352 pixels, where the background pixels were labeled as 0 and fat droplet pixels as 1. The dataset was divided into training (60%, 240 images), validation

(20%, 80 images), and testing (20%, 80 images) subsets, ensuring a balanced distribution of steatosis grades and fat droplet morphologies.

Preprocessing steps included resizing images to 352×352, applying Macenko’s stain normalization, and implementing data augmentation techniques as detailed in the preprocessing section. The segmentation models—UNet, ResUNet++, MIFONet, HarDNet-MSEG, DeepLabV3, and DUCK-Net—were developed using the PyTorch framework, while RAPUNet was implemented in TensorFlow. An NVIDIA RTX A5000 GPU was used for training all the models.

With a learning rate of 1e-4, segmentation training used a composite loss function that combined binary cross-entropy and dice loss, optimised with the Adam optimiser. Batch sizes varied between 1 and 4 depending on the model, with training epochs ranging from 70 to 200. For the classification task, cross-entropy loss was used to categorize steatosis severity into none, mild, moderate, and severe grades. During training, model accuracy and loss metrics were continuously monitored.

Evaluation of segmentation performance on the test set included metrics such as accuracy, Dice coefficient (F1-score), precision, Intersection over Union (IoU), and recall. Fat percentages were computed from both the ground truth and predicted masks to determine steatosis grades, and classification results were further analyzed using confusion matrices to assess true positives, false positives, and other prediction errors.

3.4 Deep Learning Models

The following architectures of are implemented:

3.4.1 Encoder Decoder based Model

3.4.1.1 UNet[25]:

For biomedical image segmentation, UNet is a popular convolutional neural network (CNN) architecture utilised due to its ability to learn precise localization while preserving contextual information. The UNet model implemented in this study follows the traditional skip-connected encoder-decoder architecture.

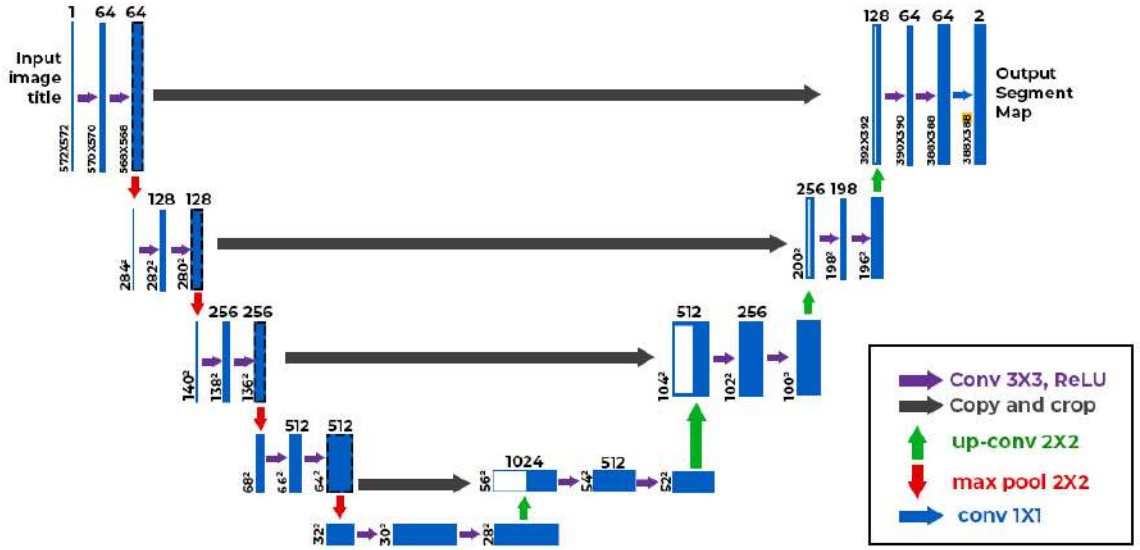


Figure 3.1: UNet Architecture

Model Description: The model accepts RGB input images with 3 channels and produces a single-channel output representing the binary segmentation mask. The architecture consists of the following main components:

1. Encoder (Contracting Path)

Spatial features (such as edges, textures, and patterns) are gradually captured by the encoder path by downsampling the input image through repeated convolutional blocks followed by max-pooling layers.

Structure:

- **Input:** 352x352x3 RGB image.

-
- **Convolution Blocks:** Each block has two 3x3 convolutions (with padding to preserve spatial dimensions), each followed by ReLU activation to increase the number of feature channels (e.g., $3 \rightarrow 64$, $64 \rightarrow 128$).
 - **Max-Pooling:** 2x2 max-pooling with stride 2 halves spatial dimensions (e.g., $352 \times 352 \rightarrow 176 \times 176$) deepening feature representations.
 - **Levels:** Typically 4 encoder levels, reducing spatial size to 22×22 while increasing channels to 512 at the bottleneck.

2. Bridge

The bottleneck of the system is a 1024-channel convolutional block that acts as a link between the encoder and decoder. This layer captures high-level abstract features before reconstruction begins.

3. Decoder (Expanding Path)

The decoder creates a 352×352 binary mask by upsampling the feature maps in order to restore spatial resolution.

Structure:

- **Up-Convolution:** 2x2 transposed convolutions (stride 2) double spatial dimensions (e.g., $22 \times 22 \rightarrow 44 \times 44$) and halve channels.
- **Skip Connections:** Doubles the channels (e.g., $512 + 512 \rightarrow 1024$) by concatenating encoder features with upsampled decoder features.
- **Convolution Blocks:** Two 3x3 convolutions with ReLU reduce channels and refine features.
- **Levels:** Mirrors the encoder (4 levels), restoring the original 352×352 resolution.

4. Output Layer

A 1x1 convolution is the last output layer, which condenses the feature map to a single

channel, representing the raw segmentation logits. Since BCEWithLogitsLoss is used during training, a sigmoid activation is omitted in the model and applied implicitly during loss calculation.

3.4.1.2 ResUNet++:

ResUNet++ is an advanced deep learning architecture designed for semantic segmentation tasks, particularly effective in medical imaging. It builds upon the foundational U-Net and ResUNet designs by incorporating several key innovations: **residual connections**, **(SE) squeeze-and-excitation blocks**, **(ASPP) Atrous Spatial Pyramid Pooling**, and **attention mechanisms**. These enhancements improve both the feature representation capacity and segmentation accuracy, especially in complex or low-contrast imaging scenarios.

1) Encoder (Downsampling Path)

Hierarchical feature representations are taken from the input image by the encoder. In ResUNet++, it consists of an initial convolutional layer followed by three residual convolutional blocks, each integrated with Squeeze-and-Excitation (SE) modules. The primary goal is to gradually increase the depth (number of feature channels) while decreasing the spatial resolution, enabling the network to learn increasingly abstract and discriminative features.

1. Input Layer

The encoder starts with two 3×3 convolutional layers:

- Each convolution uses padding to preserve spatial dimensions.
- To stabilise training and add non-linearity, Batch Normalisation and ReLU activation usually come after these layers.

This initial block prepares the input (e.g., $352 \times 352 \times 3$ image) for deeper processing.

2. Residual Convolutional Blocks

Each of the three main stages in the encoder contains a residual block, which is a set of

layers designed to learn feature transformations while maintaining efficient gradient flow through skip (identity) connections.

Structure of a Residual Block:

- Two or more 3×3 convolutions.
- Batch Normalization and ReLU activation after each convolution.
- A shortcut connection immediately connects the block's input to its output.
- This allows residual functions to be learnt by the network, i.e., changes from the input, rather than full transformations, improving training stability and convergence.

3. Squeeze-and-Excitation (SE) Modules

After each residual block, a SE block recalibrates channel-wise feature responses:

- Reduces each feature map to a single value (one per channel) using global average pooling..
- Passes through a bottleneck structure: a small fully connected (FC) network with a reduction ratio (e.g., $r = 16$).
- Enhances significant features and suppresses unimportant ones by multiplying the weights that are output for each channel by the original feature map.

4. Progressive Downsampling

- At each residual block's end (or as part of it), downsampling is performed using Max pooling or strided convolutions (typically $\text{stride}=2$), which reduces the spatial dimensions to half.
- Simultaneously, more filters (channels) are used to capture increasingly intricate patterns. The progression of feature depths is typically: $32 \rightarrow 64 \rightarrow 128 \rightarrow 256 \rightarrow 512$.

2) Bridge (Bottleneck)

At the base of the U-shaped construction is a bridge that connects the encoder and decoder. It includes:

- A residual block for deep feature extraction.
- Arous Spatial Pyramid Pooling (ASPP) is a module that uses several parallel convolutions with varying rates of dilation (e.g., 1, 6, 12, 18). This improves the model's receptive field and gathers multi-scale contextual information without adding to its computational complexity.

3) Decoder (Upsampling Path)

The decoder gradually raises the resolution to recreate spatial information from high-level features. It mirrors the encoder and includes:

- **Upsampling layers:** Usually transposed convolutions or bilinear upsampling followed by convolution.
- **Residual Decoder Blocks:** These are similar to the encoder's residual blocks but adapted to the decoding context.
- **Attention Gates:** Introduced at each skip connection, attention gates selectively highlight relevant features from the encoder before merging them with decoder features. This prevents the decoder from being overwhelmed by irrelevant information.

4) Skip Connections

ResUNet++ employs skip connections between matching encoder and decoder stages, just like U-Net. These connections are enhanced with attention gates, allowing the model to provide the decoder with only the most pertinent spatial characteristics, thereby improving segmentation accuracy.

5) Output Layer

A sigmoid or softmax activation, depending on the task, follows a 1×1 convolution in the last output layer, which reduces the feature maps to the number of output classes that are desired (often 1 for binary segmentation).

3.4.2 Attention Based Models

3.4.2.1 MIFONet:

MIFONet (Multi-scale Interactive Feature Optimization Network) is a deep learning architecture designed to address the challenges of biomedical image segmentation, especially the requirement to capture both more general contextual information and fine-grained local data. Its architecture integrates multi-scale feature extraction, attention mechanisms, and interactive fusion modules to enhance the segmentation of complex structures, such as fat droplets in liver biopsy images.

1. Encoder: Hierarchical Feature Extraction

The encoder comprises five convolutional blocks that progressively downsample the input and increase the receptive field to capture both low-level textures and high-level semantic structures. Each block includes:

- Two 3×3 convolutional layers, each followed by batch normalization and ReLU activation.
- A 2×2 max-pooling layer to reduce spatial resolution by half.

The number of filters doubles at each level to accommodate increased abstraction with block 1 having 64 filters to block 5 which has 1024 filters.

This encoding scheme allows the model to extract increasingly abstract features relevant to detecting and segmenting fat droplets..

2. Multi-Level Feature Fusion (Decoder)

The decoder reconstructs high-resolution segmentation masks using a top-down fusion strategy:

Each decoder block:

-
- Receives high-level (semantic) upsampled features from the previous decoder layer (upsampled).
 - Integrates corresponding encoder outputs via skip connections (low-level spatial detail).
 - Concatenates and processes the combined features with two 3×3 convolutional layers followed by CBAM attention.

Four decoder blocks mirror the encoder:

- Decoder 4: combines features from encoder5 and encoder4.
- Decoder 3: fuses Decoder4 and Encoder3 features.
- Decoder 2: combines Decoder3 and Encoder2 features.
- Decoder 1: integrates Decoder2 and Encoder1 features.

This hierarchical fusion ensures detailed spatial localization is preserved while progressively integrating deeper semantic context.

3. Attention Mechanism: CBAM

Each decoder block incorporates a Convolutional Block Attention Module (CBAM)[29], which enhances representational capacity by learning both channel-wise and spatial attention:

- **Channel Attention:** Prior to two fully linked layers, important weights are calculated using both average and maximum pooling.
- **Spatial Attention:** Learns spatial significance via a convolution over the concatenation of channel-wise average and max pooled maps.

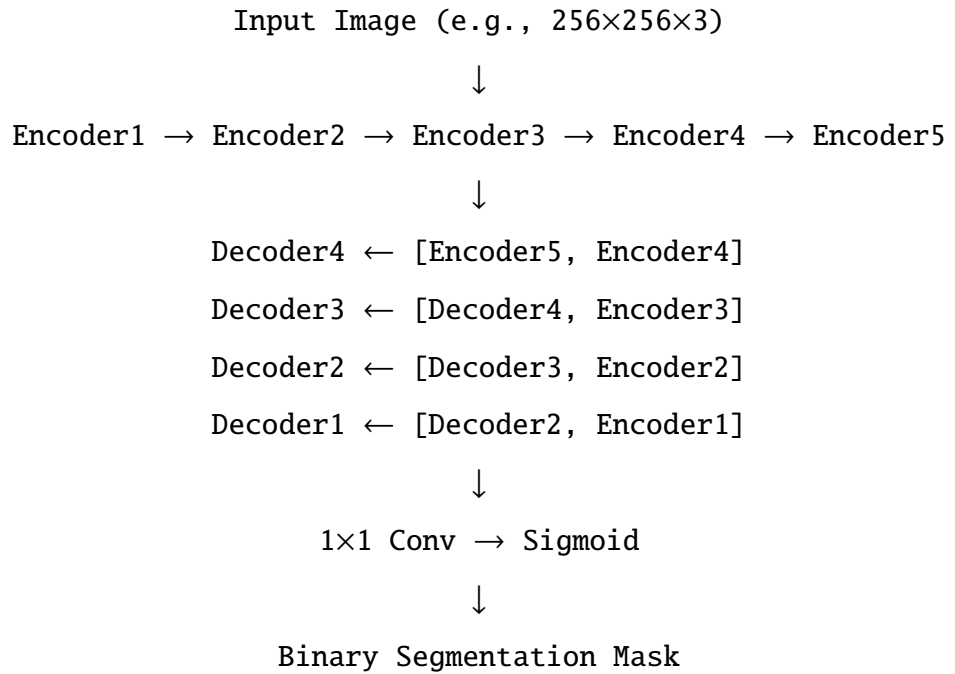
When combined, these processes increase segmentation accuracy by assisting the model in suppressing background noise and concentrating on pertinent features.

4. Output Layer

The final decoder output is passed through:

- A 1×1 convolutional layer that reduces the number of channels to one (binary classification).
- A sigmoid activation function that outputs pixel-wise probabilities in the range $[0, 1]$.

Architecture Summary



Advantages

- Attention-guided feature fusion improves focus on relevant regions and suppresses noise.
- Multi-scale fusion ensures preservation of fine spatial detail while incorporating semantic context.
- Modular design allows for integration with deeper or pretrained backbones if desired.

-
- Lightweight yet effective for medical image segmentation tasks such as hepatic steatosis detection.

3.4.2.2 RAPUNet:

The **Residual and Atrous Convolution in Parallel Unit** or **RAPUNet** model is a hybrid encoder-decoder architecture designed for high-accuracy medical image segmentation. It leverages both transformer-based global contextual understanding and convolution-based local feature refinement through a combination of the CAFormerS18 backbone and custom-designed attention blocks. This hybrid design enables RAPUNet to capture fine-grained boundaries while maintaining a deep semantic understanding of structures in the image.

1. Encoder: CAFormerS18 Backbone

The encoder of RAPUNet is based on CAFormerS18, a lightweight transformer-based model (from the MetaFormer family) pretrained on the ImageNet dataset. Unlike traditional CNN-based backbones, CAFormer improves long-range dependence modelling by integrating feed-forward network (FFN) blocks with self-attention. Feature maps are extracted from four different levels of the CAFormer, corresponding to resolutions of $1/4$, $1/8$, $1/16$, and $1/32$ of the input image. These features are:

- p2: Feature map at $\frac{1}{4}$ resolution,
- p3: Feature map at $\frac{1}{8}$ resolution,
- p4: Feature map at $\frac{1}{16}$ resolution,
- p5: Feature map at $\frac{1}{32}$ resolution.

To match the number of channels needed for additional processing, each of the above features is run through a 1×1 convolution.

2. Initial Feature Extraction

An initial low-level feature map (p1) is extracted by applying a 3×3 convolution with stride 2 to the input image. This serves to downsample the image and capture basic texture and edge information, complementing the semantic features from deeper layers.

3. Residual Attention Pyramid Units (RAPUs)

The architecture utilizes **Residual Attention Pyramid Units (RAPUs)** to process and enhance features at each stage. RAPUs combine multi-scale convolutions, residual connections, and attention mechanisms to extract context-aware features. The hierarchical pipeline proceeds through five stages as follows:

$$t_0 \rightarrow t_1 \rightarrow t_2 \rightarrow t_3 \rightarrow t_4 \rightarrow t_5$$

At each stage, the RAPU processes a combination of downsampled features from the previous RAPU and a skip connection from the backbone feature at the corresponding resolution.

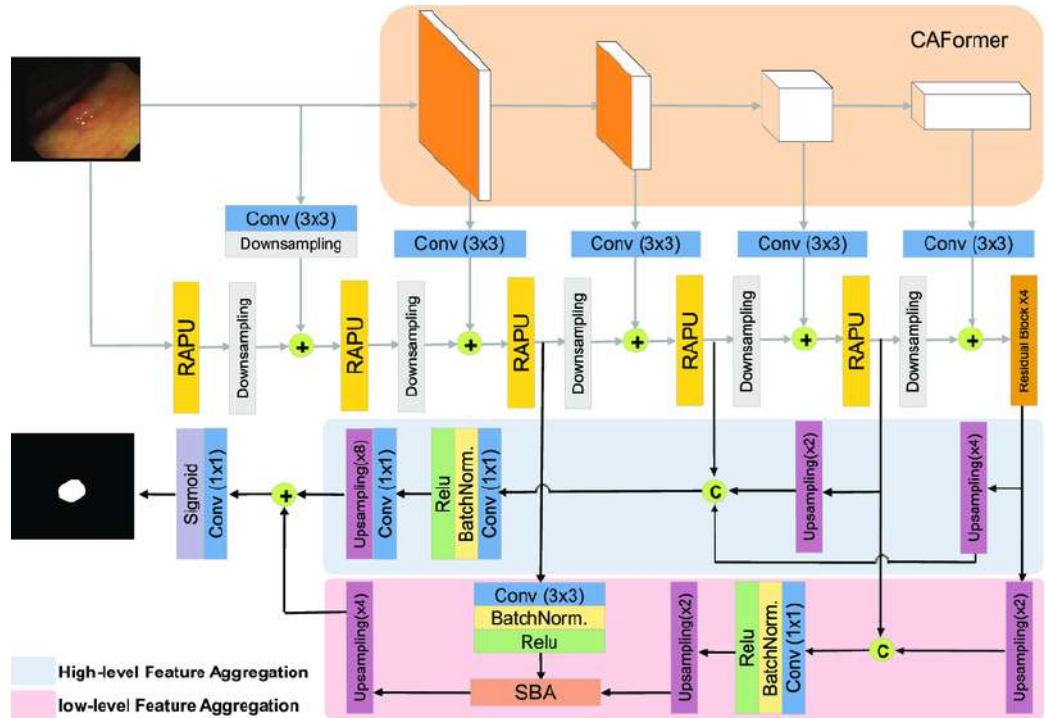


Figure 3.2: RAPUNet architecture[20]

4. Deep Feature Refinement

At the deepest level (t_5), a sequence of residual blocks is applied to further enhance the representation. These blocks help in better capturing the abstract features critical for distinguishing complex regions and structures in the image.

5. Decoder with Dual-Branch Aggregation

RAPUNet employs a **dual-branch decoder** that aggregates features from both deep and shallow layers to generate high-resolution segmentation maps.

- **Mid-Level Aggregation Branch (Output 1):** Combines upsampled features from t_5 and t_4 with t_3 . The fused result is passed through a convolutional module to generate a prediction at 44×44 resolution.
- **Spatial Bilateral Attention (SBA) Branch (Output 2):** Combines t_2 with the upsampled deep features using a custom SBA block, which emphasizes boundary-aware features through spatial attention. This produces a prediction at 88×88 resolution.

Final Prediction

Bilinear interpolation is used to upsample both outputs to the original image resolution:

$$\text{Output1} \uparrow 8 \times + \text{Output2} \uparrow 4 \times$$

The fused result is passed through a 1×1 convolution followed by a sigmoid activation to produce the final binary segmentation mask.

Conclusion

RAPUNet successfully strikes a balance between local accuracy and global context. It can segment medical images accurately thanks to its dual-decoder, attention-enhanced structure, especially when boundary accuracy and structure preservation are crucial. It is computationally efficient and robust due to its mixed use of convolutional blocks and

Component	Functionality
CAFormerS18	Extracts multi-scale features with global context using attention-based mechanisms
RAPU	Refines features via residual attention and multiscale convolution
Residual Blocks	Enhance deep representation and maintain gradient flow
SBA Module	Enhances spatial details and boundaries via attention-based fusion
Dual Decoder Branches	Fuses coarse and fine details for robust prediction

Table 3.1: Summary of RAPUNet architectural components

CAFormer.

3.4.3 Transformer Based

3.4.3.1 MedSAM

This robust transformer-based deep learning system, called the Medical Segment Anything Model (MedSAM), was created for general-purpose medical image segmentation. It expands upon Meta AI’s initial Segment Anything Model (SAM) by customising it for use with medical imaging modalities like histopathology, CT, MRI, and ultrasound.

MedSAM follows a modular encoder-decoder architecture consisting of the following components:

1. Image Encoder

The image encoder is a Vision Transformer (ViT), commonly the ViT-Huge (ViT-H) architecture.

- It receives medical images (e.g., resized to 1024×1024 resolution) and divides them into non-overlapping patches.
- These patches are linearly embedded and passed through transformer layers to produce high-dimensional image embeddings.

Output: A spatial grid of feature embeddings (e.g., $64 \times 64 \times 2560$ for ViT-H).

2. Prompt Encoder

- This module encodes user prompts (e.g., points, bounding boxes, or existing masks) that guide segmentation.
- Point coordinates are converted into learned embeddings with positional encoding.
- Box prompts are similarly encoded.
- It also supports mask inputs by compressing them into latent vectors using convolutional layers.

Output: Prompt tokens used to guide the mask decoder.

3. Mask Decoder

A lightweight decoder that combines:

- Visual tokens from the image encoder
- Prompt embeddings from the prompt encoder

Built using transformer attention blocks followed by upsampling layers.

It produces:

- Segmentation masks (typically 256×256 resolution)
- IoU confidence scores for mask quality estimation

Output: One or more high-quality segmentation masks per prompt, with confidence scores.

4. Output Head

- The output head upsamples the decoder's coarse segmentation to the input image resolution (e.g., 1024×1024).

-
- The binary segmentation map is obtained by applying a sigmoid activation at the end.

3.4.4 Multi-Scale Fusion Models

3.4.4.1 HarDNet-MSEG [14]

The HarDNet-MSEG (Harmonic Dense Network for Medical SEGmentation) is a lightweight, fast, and accurate encoder–decoder architecture designed for binary segmentation tasks, especially in medical imaging. Here’s a structured breakdown of its architecture based on the code:

Backbone: HarDNet-68

The core of HarDMSEG is the HarDNet-68 backbone, a convolutional neural network with 68 layers which are organized into five primary blocks called HarDBlocks. Each HarDBlock employs *harmonic dense connectivity*, where layers connect selectively to earlier layers at distances that are powers of two. This differs from traditional DenseNet, where each layer connects to the preceding layer, leading to fewer redundant connections and thus better efficiency.

Detailed Architectural Components

1. Initial Convolutional Layers

The network accepts a standard 3-channel input image and initially applies two convolutional layers: the first with 32 channels and stride 2 for spatial downsampling, followed by another with 64 channels for feature refinement. Next, downsampling is performed using either max-pooling or depthwise separable convolution.

2. HarDBlocks

The backbone consists of five HarDBlocks with the following configuration:

- **Block 1:** 8 convolutional layers, output channels 128

- **Block 2:** 16 convolutional layers, output channels 256
- **Block 3:** 16 convolutional layers, output channels 320
- **Block 4:** 16 convolutional layers, output channels 640
- **Block 5:** 4 convolutional layers, output channels 1024

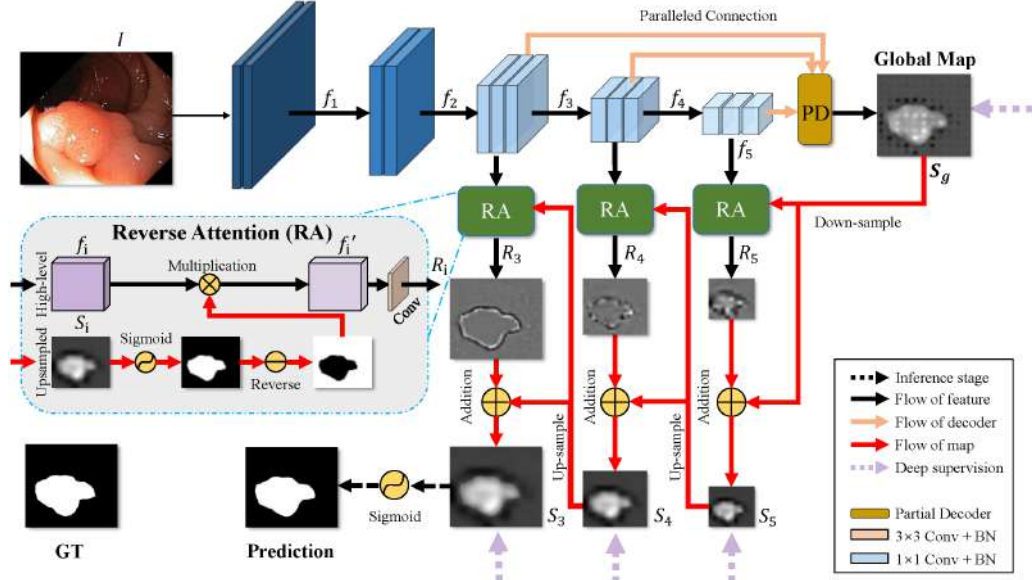


Figure 3.3: HarDNet-MSEG framework[14]

A 3×3 convolutional filter, batch normalisation, and ReLU activation make up each convolution in a block. The growth rate, controlling how many new channels are added per layer, is modulated by a growth multiplier (set to 1.7) to balance capacity and efficiency.

3. Harmonic Dense Connectivity

Layers inside each HarDBlock connect only to preceding layers at indices differing by powers of two (e.g., connections from layer n to layers $n - 1$, $n - 2$, $n - 4$, $n - 8$). This selective connectivity reduces computational complexity and memory usage while preserving effective feature reuse and gradient flow.

4. Downsampling Strategy

The architecture progressively reduces spatial resolution using either max-pooling or depth-wise convolutions with stride 2 after specific HarDBlocks. This multi-scale feature extrac-

tion method is critical for effectively segmenting hepatic fat deposits because it captures both small features and broad context.

5. Multi-Scale Feature Extraction

HarDMSEG extracts feature maps at multiple intermediate stages (after layers 4, 9, 12, and 15) to capture semantic information at different spatial resolutions. These multi-scale features are fused in the segmentation head to improve prediction accuracy and robustness across varying lesion sizes and grades.

Advantages for Hepatic Steatosis Segmentation

The HarDMSEG model’s combination of harmonic dense connectivity and efficient convolutional blocks enables it to:

- Extract rich, multi-scale features with fewer parameters and lower memory footprint.
- Maintain strong gradient flow through selective dense connections, improving training stability.
- Achieve precise segmentation performance, even on limited medical datasets with subtle tissue texture variations.

Summary

Table 3.2 summarizes the key architectural parameters of the HarDMSEG backbone.

Table 3.2: Summary of HarDMSEG Backbone Configuration

Component	Number of Layers	Output Channels
Initial Convolutional Layers	2	32 → 64
HarDBlock 1	8	128
HarDBlock 2	16	256
HarDBlock 3	16	320
HarDBlock 4	16	640
HarDBlock 5	4	1024

Overall, HarDMSEG offers a state-of-the-art balance between segmentation accuracy and model efficiency, making it well-suited for hepatic steatosis detection and grade classification tasks.

3.4.5 Dilated Convolution Based

3.4.5.1 DeepLabV3+

Detailed Architecture of DeepLabV3+ with ResNet-50 Backbone

DeepLabV3+ consists of several core components, including atrous (dilated) convolutional layers, the Atrous Spatial Pyramid Pooling (ASPP) module, a decoder segment, and a ResNet-50 backbone. Each of these elements plays a crucial role in enhancing semantic segmentation performance, particularly in complex medical imaging scenarios such as hepatic steatosis detection, by effectively capturing both local and global contextual features.

1. ResNet-50 Backbone

After processing the input image, the backbone network functions as a feature extractor, creating hierarchical feature maps with increasingly more detailed semantic information.

- **Initial Layers:** The network starts with a 7×7 convolution layer with stride 2, then moves on to a max pooling layer, batch normalisation, and ReLU activation. This captures fundamental visual features while lowering the spatial resolution.
- **Residual Blocks:** ResNet-50 is built from 16 bottleneck residual blocks grouped into four stages:
 - *Conv2_x*: Has 3 bottleneck blocks with 256 output channels.
 - *Conv3_x*: Consists of 4 bottleneck blocks with 512 output channels.
 - *Conv4_x*: Includes 6 bottleneck blocks producing 1024 output channels.

-
- *Conv5_x*: Features 3 bottleneck blocks with 2048 output channels.

Three convolution layers— 1×1 , 3×3 , and 1×1 convolutions—as well as identity skip connections are included in each bottleneck block to promote gradient flow and prevent vanishing gradients.

- **Atrous Convolutions:** To maintain a higher spatial resolution, atrous convolutions replace some strided convolutions in the *Conv4_x* and *Conv5_x* stages, with dilation rates 2 and 4 respectively.

2. Atrous Convolution (Dilated Convolution)

By adding "holes" or spaces to the convolution kernel, atrous convolution broadens the receptive field without lowering feature map resolution or adding more parameters.

- **Purpose:** Without downsampling, it enables the network to capture multi-scale context.
- **Implementation:** In DeepLabV3+, it replaces standard convolutions in the later ResNet blocks with atrous convolutions having controlled dilation rates.
- **Effect:** By doing this, the field of view for context aggregation is expanded while maintaining the spatial resolution of feature maps (output stride of 16 or 8).

3. Atrous Spatial Pyramid Pooling (ASPP)

On the final backbone feature map, the ASPP module applies parallel atrous convolutions with varying dilation rates to extract features at several spatial scales.

- **Branches:**
 1. A 1×1 convolution branch to capture local information.
 2. Three 3×3 atrous convolution branches with dilation rates typically set to 6, 12, and 18, capturing increasing spatial contexts.

3. In order to capture global context, a 1×1 convolution is performed after an image-level pooling branch that uses global average pooling.

- **Concatenation and Fusion:** Outputs from all branches are concatenated and passed through another 1×1 convolution to fuse features and reduce dimensionality.
- **Batch Normalization and Activation:** A ReLU activation function and batch normalisation come after each convolution.

4. Decoder Module

By progressively improving and upsampling the feature maps to the original image resolution, the decoder enhances the segmentation result.

- **Low-Level Feature Extraction:** It takes low-level features from an earlier ResNet stage (typically Conv2_x), which contain fine spatial details but less semantic information.
- **Channel Reduction:** These low-level characteristics are concatenated efficiently by passing them through a 1×1 convolution, which lowers the number of channels (for example, from 256 to 48).
- **Feature Fusion:** The decoder upsamples the ASPP output by a factor of 4 using bilinear interpolation and concatenates it with the processed low-level features.
- **Refinement:** The concatenated feature map is sent through a series of 3×3 convolution layers (usually two) with batch normalization and ReLU activations to refine object boundaries and details.
- **Final Upsampling:** To restore the original input resolution, the improved feature map is upsampled once more (by a factor of 4), producing the final high-resolution segmentation map.

5. Output Layer

-
- Each pixel is given a probability score for each class in the final output, which is usually a pixel-wise classification map.
 - For binary segmentation (e.g., detecting steatosis), a sigmoid activation function is used, while softmax is used for multi-class segmentation.

Summary

The combined architecture allows DeepLabV3+ to capture both global context through atrous convolutions and ASPP, and precise boundary details via the decoder, making it effective for detailed medical image segmentation tasks such as hepatic steatosis detection and grading.

3.4.6 Light Weight Models

3.4.6.1 DUCK-Net[27]:

A deep convolutional neural network specifically engineered for semantic segmentation, DUCK-Net combines innovative convolutional blocks with residual connections to effectively capture multi-scale data. The architecture uses advanced feature aggregation and skip connections in an encoder-decoder fashion, as explained below.

Step-by-Step Architectural Flow

1. Input Layer:

The model receives an input image tensor of shape (B, C, H, W) , where B is the batch size, C is the number of channels (e.g., 3 for RGB), and H, W are spatial dimensions.

2. Initial Downsampling and Feature Extraction:

A sequence of ConvBlock layers is applied to the input, gradually decreasing the spatial resolution by a factor of two at each stage, while simultaneously increasing the number of feature channels. Specifically, five downsampling convolutional layers extract hierarchical

features at different scales:

$$X \xrightarrow{\text{ConvBlock}_1} P_1 \xrightarrow{\text{ConvBlock}_2} P_2 \xrightarrow{\text{ConvBlock}_3} P_3 \xrightarrow{\text{ConvBlock}_4} P_4 \xrightarrow{\text{ConvBlock}_5} P_5,$$

where each P_i represents feature maps at increasingly coarser resolutions and richer semantic content.

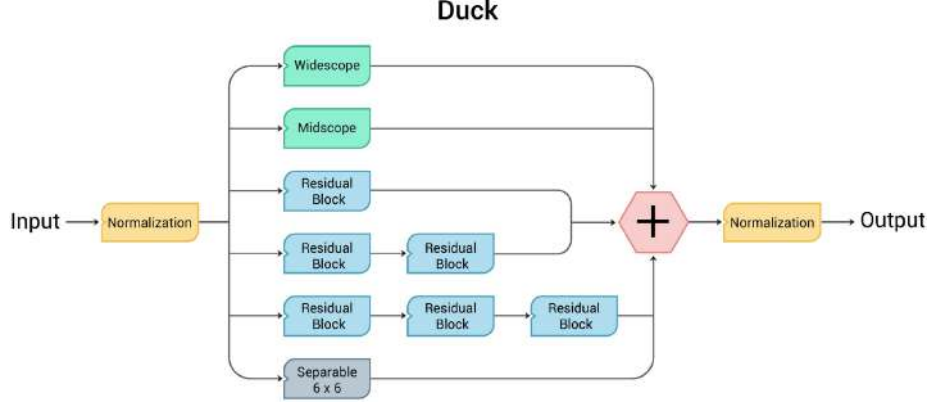


Figure 3.4: A single DUCK block[10]

3. Parallel Multi-Scale Feature Extraction (DUCKBlocks in Encoder):

In parallel to the downsampling path, the input and intermediate feature maps are also processed by DUCKBlock modules. The DUCKBlock is a composite block that aggregates multiple convolutional pathways — including wide, mid, and separable convolutions, as well as residual blocks — to extract multi-scale contextual information:

$$T_0 = \text{DUCKBlock}(X), \quad T_i = \text{DUCKBlock}(P_i + \text{ConvBlock}(T_{i-1})) \quad \text{for } i = 1, \dots, 4.$$

Each DUCKBlock combines several convolutions with different receptive fields and residual connections, enabling the network to concurrently capture local and global spatial characteristics.

4. Bottleneck Residual Processing:

At the deepest layer P_5 , after merging with the corresponding DUCKBlock output T_4 , the combined features pass through two ResidualBlock layers. These residual blocks preserve steady gradient flow while permitting deeper feature extraction, which improves

and enriches the representation:

$$S_5 = \text{ResidualBlock}_1(P_5 + \text{ConvBlock}(T_4)), \quad S'_5 = \text{ResidualBlock}_2(S_5).$$

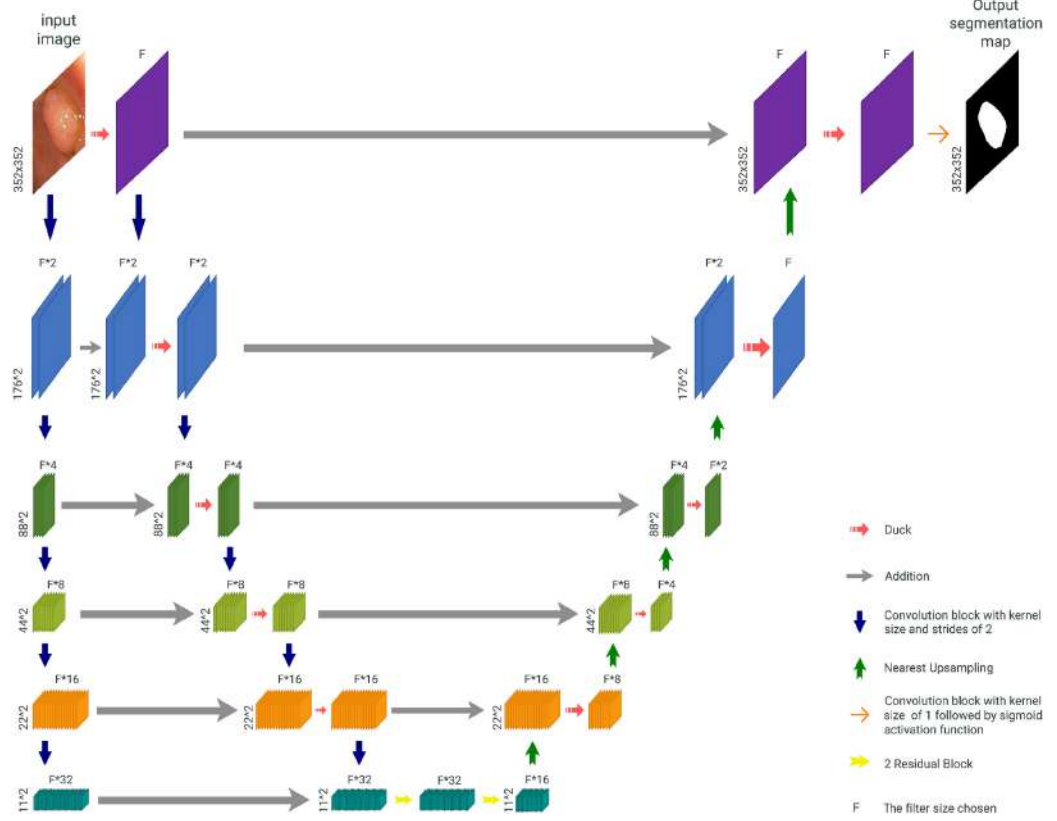


Figure 3.5: DUCK-Net architecture[10]

5. Decoder with Upsampling and Skip Connections:

By progressively upsampling the bottleneck characteristics and combining them with matching encoder features via skip connections, the decoder recreates the high-resolution segmentation map. At each decoder level:

- Nearest-neighbor interpolation is used to upsample the feature map by a factor of two.
- It is added element-wise with the corresponding DUCKBlock output from the encoder side.

-
- The sum is processed through a DUCKBlock to refine the features at that scale.

Formally, for levels $i = 4, 3, 2, 1, 0$:

$$Q_i = \text{DUCKBlock}(\text{Upsample}(Q_{i+1}) + T_i),$$

where $Q_5 = S'_5$.

This design preserves fine spatial details via skip connections while enhancing features with multi-scale contextual information from DUCKBlocks.

6. Final Prediction Layer:

After the final decoder DUCKBlock Q_0 , a 1×1 convolution reduces the channel dimension to the number of segmentation classes, typically 1 for binary segmentation. The sigmoid activation is used to create a probability map at the pixel level:

$$\text{Output} = \sigma(\text{Conv}_{1 \times 1}(Q_0)).$$

The anticipated segmentation mask, which shows the probability that each pixel belongs to the target class, is this output.

Summary

In summary, DUCK-Net begins by extracting features at multiple scales via progressive downsampling and multi-path convolutional blocks. It then refines deep semantic features at the bottleneck using residual blocks. By using multi-scale DUCK-Blocks and skip connections, the decoder gradually upsamples and combines features with encoder representations, culminating in a pixel-wise segmentation mask. The architecture's design ensures effective integration of spatial detail and contextual information, making it well suited for precise semantic segmentation.

3.5 Optimizer and loss functions

3.5.1 Adam Optimizer

The Adam (Adaptive Moment Estimation) optimizer [18] is a popular stochastic gradient descent method combining the benefits of momentum and RMSProp. It adapts learning rates for each parameter using first-order (mean) and second-order (uncentered variance) moments of gradients, making it efficient for deep learning tasks like medical image segmentation. Key features:

Strengths: Fast convergence, robustness to noisy gradients, effective for small datasets (like your histological images), and minimal hyperparameter tuning. **Parameters:** Learning rate (α , e.g., $1e-4$), β_1 (momentum decay, typically 0.9), β_2 (variance decay, typically 0.999), and ϵ (small constant, typically $1e-8$). **Relevance to Your Thesis:** Adam is well-suited for training RAPUNet, MIFONet, HardNet-MSEG, DUCK-Net, and UNet due to their CNN-based architectures, and MedSAM’s ViT backbone, as it handles variable gradient scales in histological image segmentation and classification.

3.5.2 BCEDiceLoss

In medical image segmentation, it is a composite loss that combines Binary Cross-Entropy (BCE) and Dice Loss to balance area overlap with pixel-wise classification.

BCE Component: Calculates the probability differences between the ground truth and forecasted masks at the pixel level, suitable for binary segmentation (0: background, 1: fat droplets).

$$\text{BCE} = -\frac{1}{N} \sum_{i=1}^N [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \quad (3.1)$$

where y_i is the ground truth, $\hat{y}_i$ is the predicted probability, and N is the number of pixels.

Dice Component: Calculates the overlap between the ground truth and anticipated masks,

focusing on region similarity.

$$\text{Dice Loss} = 1 - \frac{2|y \cap \hat{y}| + \epsilon}{|y| + |\hat{y}| + \epsilon} \quad (3.2)$$

where $|y \cap \hat{y}|$ is the intersection, $|y|$ and $|\hat{y}|$ are sums of ground truth and predicted pixels, and ϵ (e.g., $1e-6$) prevents division by zero.

3.5.3 BCELogitsLoss

It is a PyTorch-specific loss function (`torch.nn.BCEWithLogitsLoss`) combining a sigmoid activation and BCE loss in one step, applied to raw model logits (un-normalized scores) rather than probabilities.

Mechanism: Applies sigmoid to logits ($\hat{y}_i = \sigma(z_i)$) before computing BCE, reducing numerical instability.

$$\text{BCELogitsLoss} = -\frac{1}{N} \sum_{i=1}^N [y_i \log(\sigma(z_i)) + (1 - y_i) \log(1 - \sigma(z_i))] \quad (3.3)$$

3.6 Evaluation Metrics for Binary Segmentation

The similarity between the predicted mask and the ground truth mask is measured using a number of measures to assess how well segmentation algorithms perform in medical image analysis. These measurements are especially crucial for activities where precision and border alignment are crucial, such the diagnosis of hepatic steatosis.

Let:

- *TP*: True Positives — correctly predicted positive pixels
- *TN*: True Negatives — correctly predicted negative pixels
- *FP*: False Positives — incorrectly predicted positive pixels
- *FN*: False Negatives — missed positive pixels

3.6.1 Accuracy

The ratio of correctly classified pixels to the total number of pixels.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (3.4)$$

3.6.2 Precision (Positive Predictive Value)

Measures the proportion of predicted positive pixels that are actually correct.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (3.5)$$

3.6.3 Recall (Sensitivity or True Positive Rate)

Indicates the model's ability to detect positive pixels.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3.6)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (3.7)$$

3.6.4 F1 Score

The harmonic mean of precision and recall, providing a balance between the two.

$$\text{F1 Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2TP}{2TP + FP + FN} \quad (3.8)$$

3.6.5 Dice Coefficient (Sørensen–Dice Index)

A measure of statistical significance for determining how comparable two samples are is the Dice Coefficient, sometimes referred to as the Sørensen–Dice index. In the context of binary segmentation, it calculates the degree of overlap between the ground truth mask and the anticipated segmentation. It works especially well at addressing class imbalance by highlighting the foreground (positive) pixels that were accurately predicted.

$$\text{Dice} = \frac{2|P \cap G|}{|P| + |G|} = \frac{2TP}{2TP + FP + FN} \quad (3.9)$$

3.6.6 Intersection over Union (IoU or Jaccard Index)

Measures the intersection over the union of the predicted and ground truth masks.

$$\text{IoU} = \frac{|P \cap G|}{|P \cup G|} = \frac{TP}{TP + FP + FN} \quad (3.10)$$

Because of their ability to withstand imbalanced datasets, Dice and IoU are the most widely employed metrics in medical picture segmentation in practice. For binary segmentation tasks, the Dice coefficient and F1 Score are mathematically equivalent.

Additionally, handling edge cases such as empty masks (where both ground truth and prediction are zero) some metrics may become undefined or misleading. In such scenarios, a common convention is to assign a perfect score (e.g., Dice = 1) to reflect the correct prediction, ensuring a fair and consistent evaluation.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Experimental Setup

The experiments were conducted on a system with an NVIDIA RTX A5000 GPU, 64 GB RAM, and. The dataset was preprocessed and augmented as described in Chapter 3.

4.2 Results

4.2.1 Segmentation Results

The performance of the models is summarized in Table 4.2.

Table 4.1: Performance Comparison of Deep Learning Models

Model	Accuracy (%)	IoU	Dice	Recall	Precision
UNet	96.67	0.8467	0.8883	0.9151	0.9266
ResUNet++	95.17	0.8312	0.882	0.9791	0.8474
RAPUNet	96.08	0.906	0.9396	0.9528	0.9376
MIFONet	96.33	0.7743	0.8144	0.8261	0.8053
DeepLabV3+	96.9	0.8133	0.8519	0.8550	0.9534
DUCK-Net	95.88	0.8654	0.9099	0.9166	0.9398
HarDNet-MSEG	96.95	0.901	0.9368	0.9340	0.9512
MedSAM	96.8	0.7743	0.8144	0.8053	0.8131

4.2.2 Classification Results

Below is the number of correctly classified images by each model for every grade. Also shown is the number of incorrect classifications.

Table 4.2: Grade Classification Comparison of Deep Learning Models

Model	Grade 0 (<5%)	Grade 1 (5–33%)	Grade 2 (33–66%)	Grade 3 (>66%)	Incorrect Classifications
UNet	16	16	32	7	9
ResUNet++	16	9	26	15	14
RAPUNet	16	11	31	15	7
MIFONet	16	11	32	11	10
DeepLabV3+	16	14	31	11	8
DUCK-Net	16	11	30	15	8
HarDNet-MSEG	16	15	30	15	4
MedSAM	15	14	31	12	8

4.2.3 Comparison of ground truth with predicted images

UNet Model

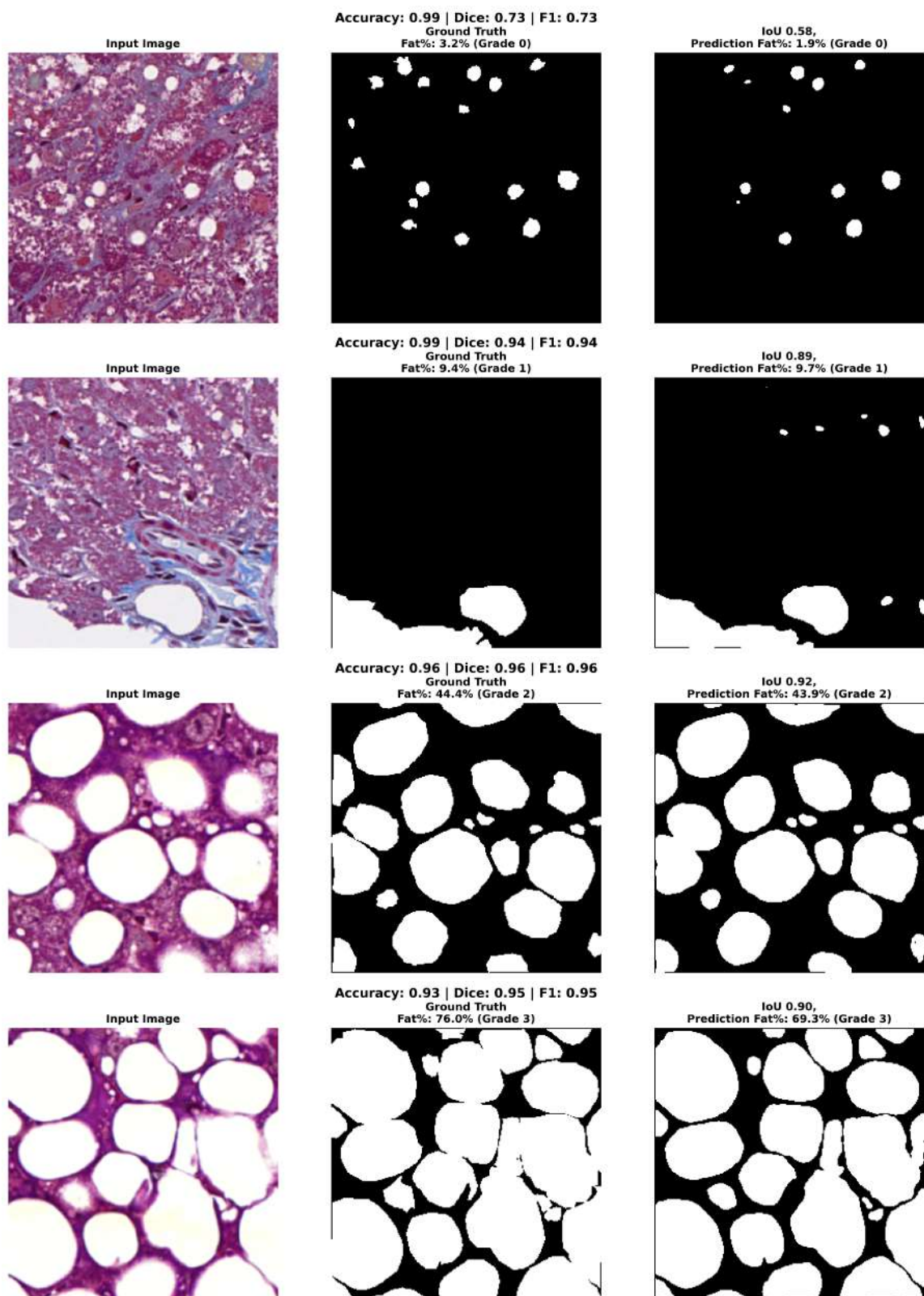


Figure 4.1: Prediction results of all 4 grades of hepatic steatosis using UNet model

ResUNet++ Model

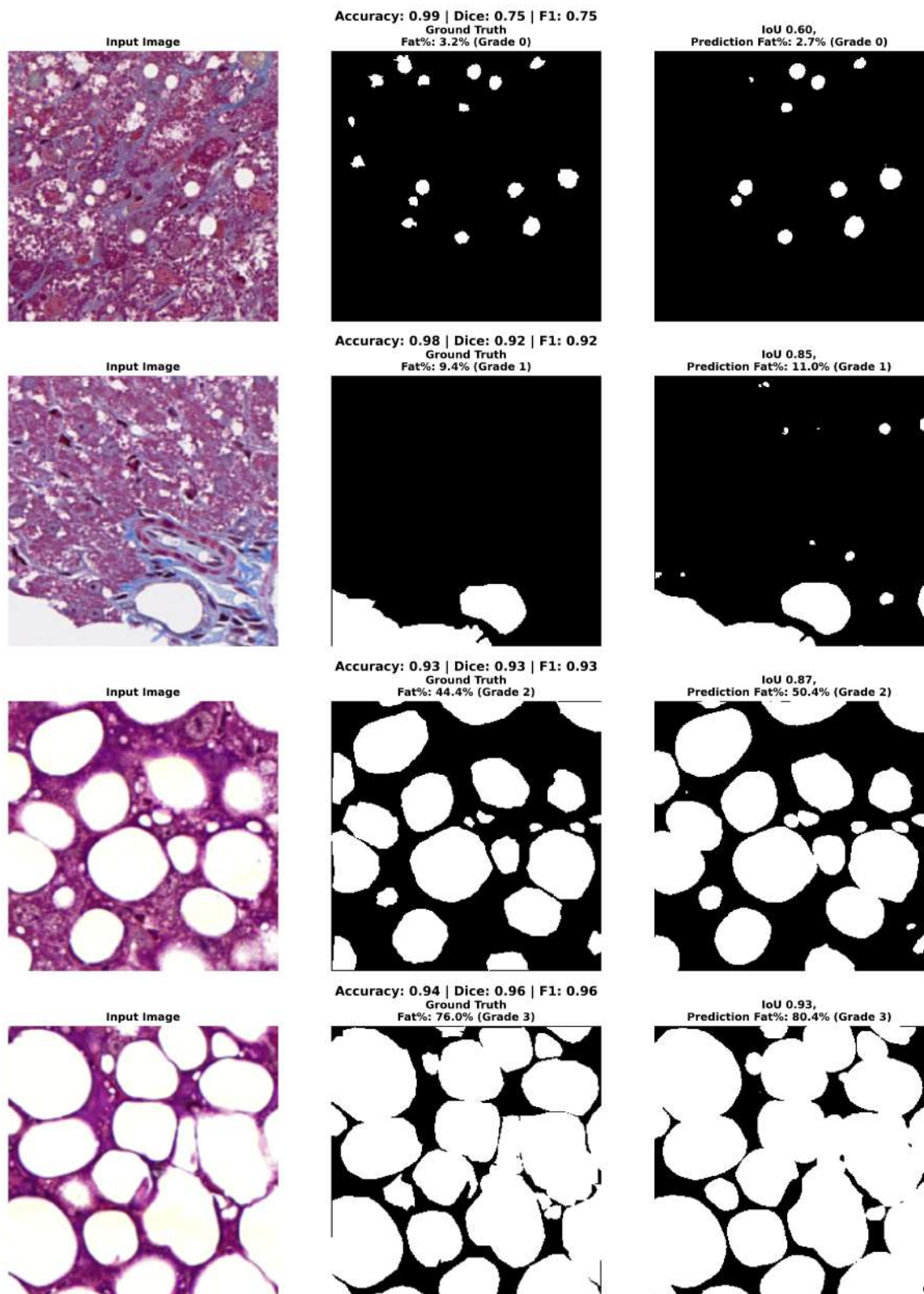


Figure 4.2: Prediction results of all 4 grades of hepatic steatosis using ResUNet++ model

MIFONet Model

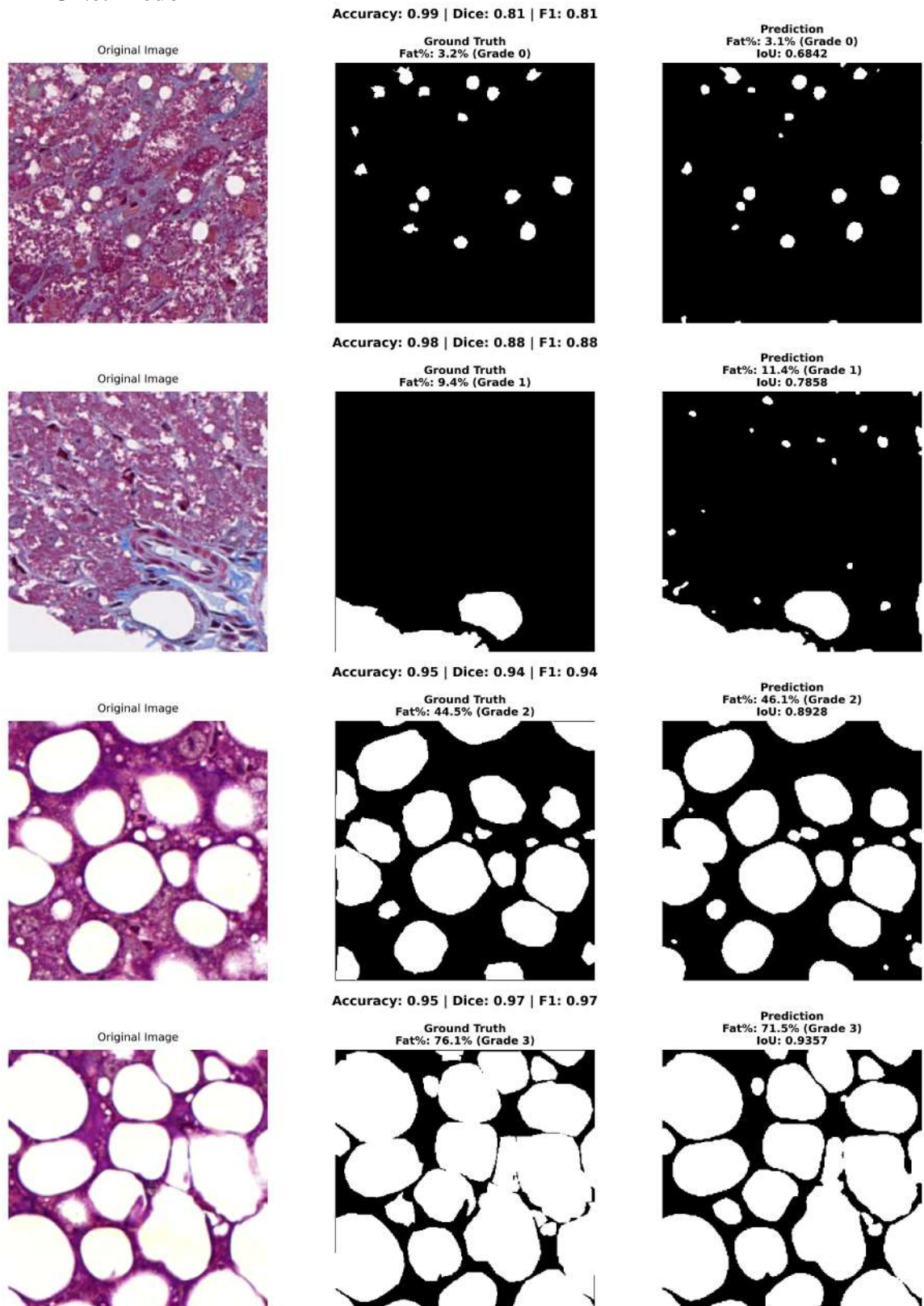


Figure 4.3: Prediction results of all 4 grades of hepatic steatosis using MIFONet model

RAPUNet Model

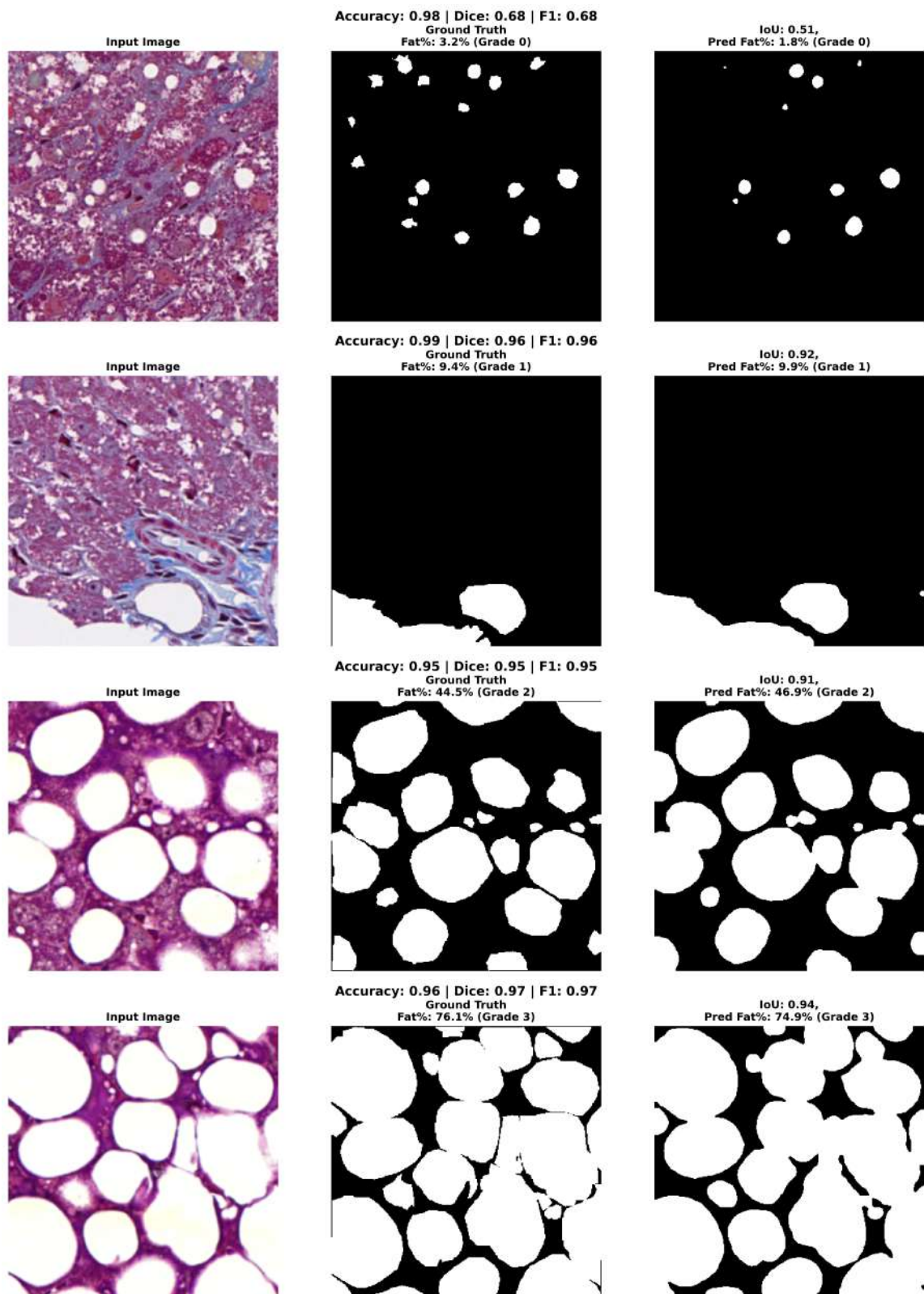


Figure 4.4: Prediction results of all 4 grades of hepatic steatosis using RAPUNet model

MedSAM Model

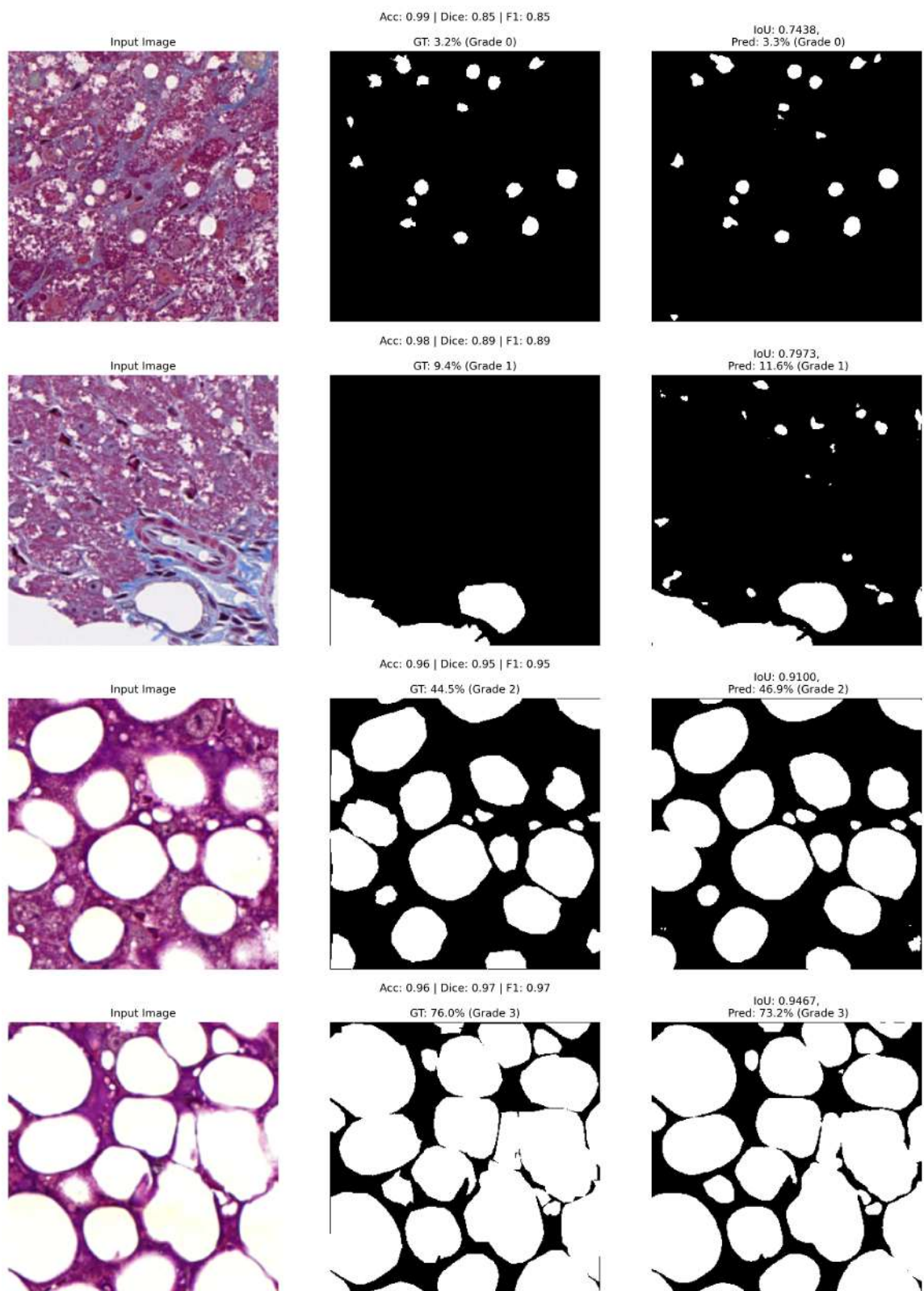


Figure 4.5: Prediction results of all 4 grades of hepatic steatosis using MedSAM model

HardNet-MSEG Model

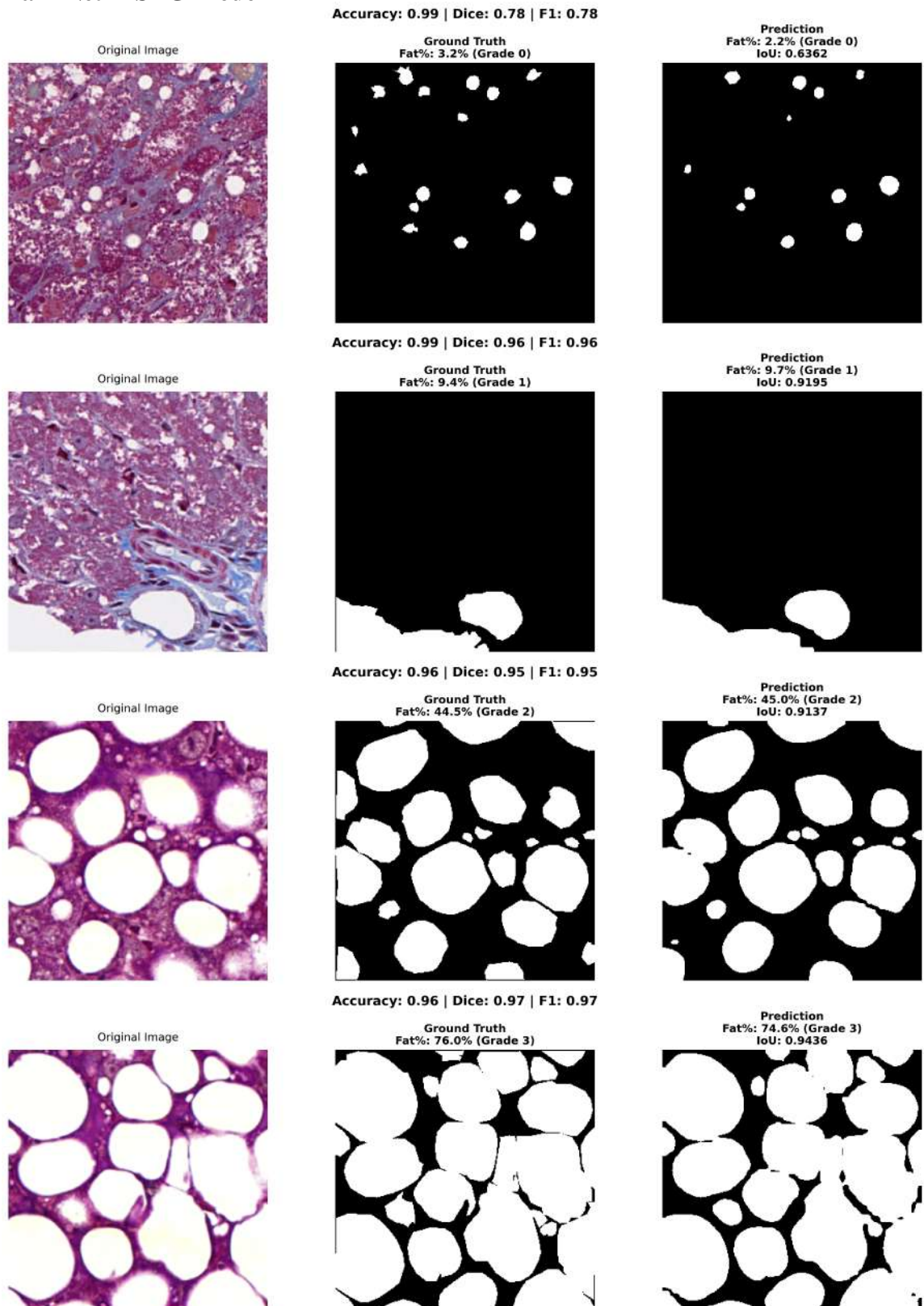


Figure 4.6: Prediction results of all 4 grades of hepatic steatosis using HardNet-MSEG model

DeepLabV3+ Model

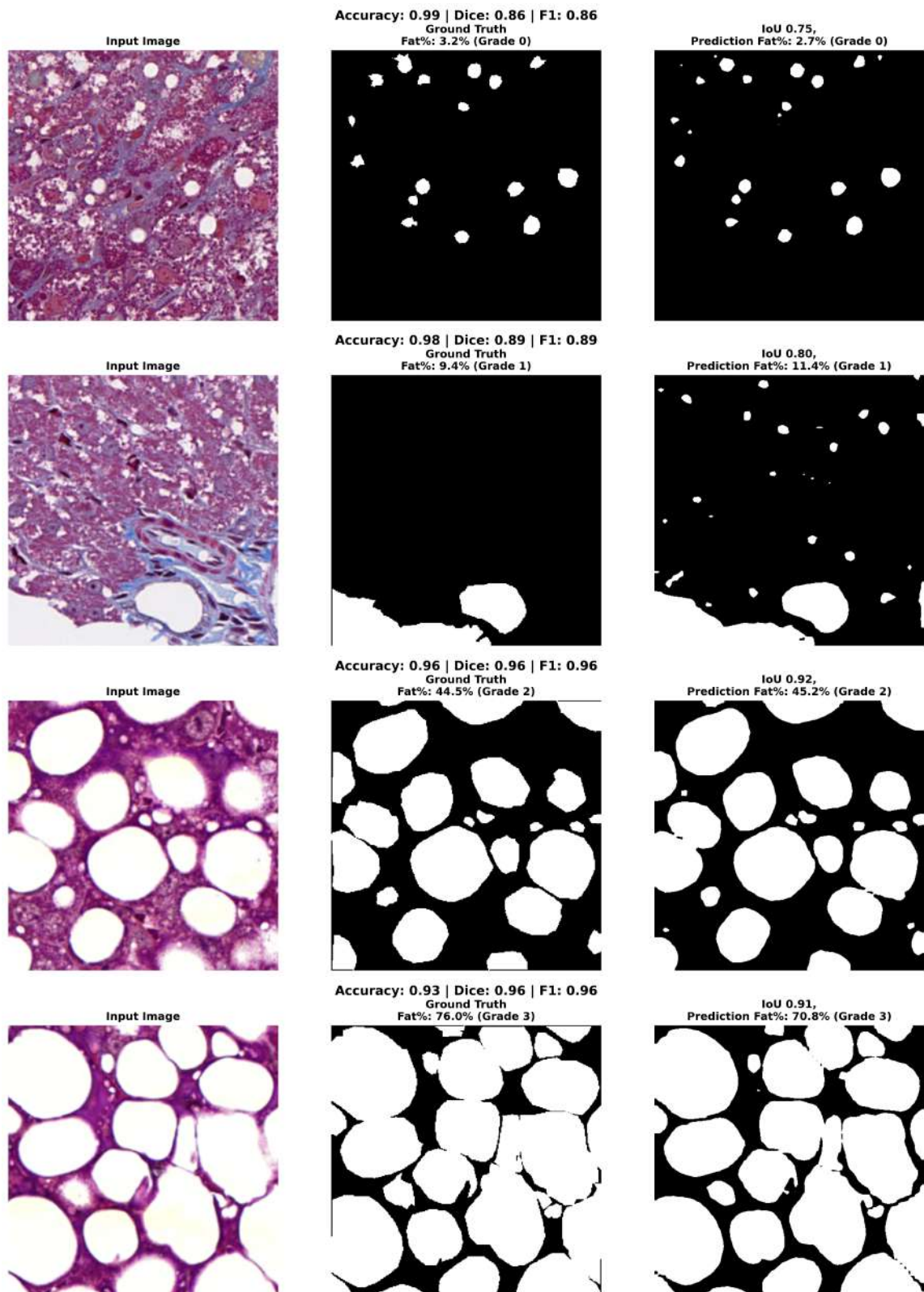


Figure 4.7: Prediction results of all 4 grades of hepatic steatosis using DeepLabV3+ model

DUCK-Net Model

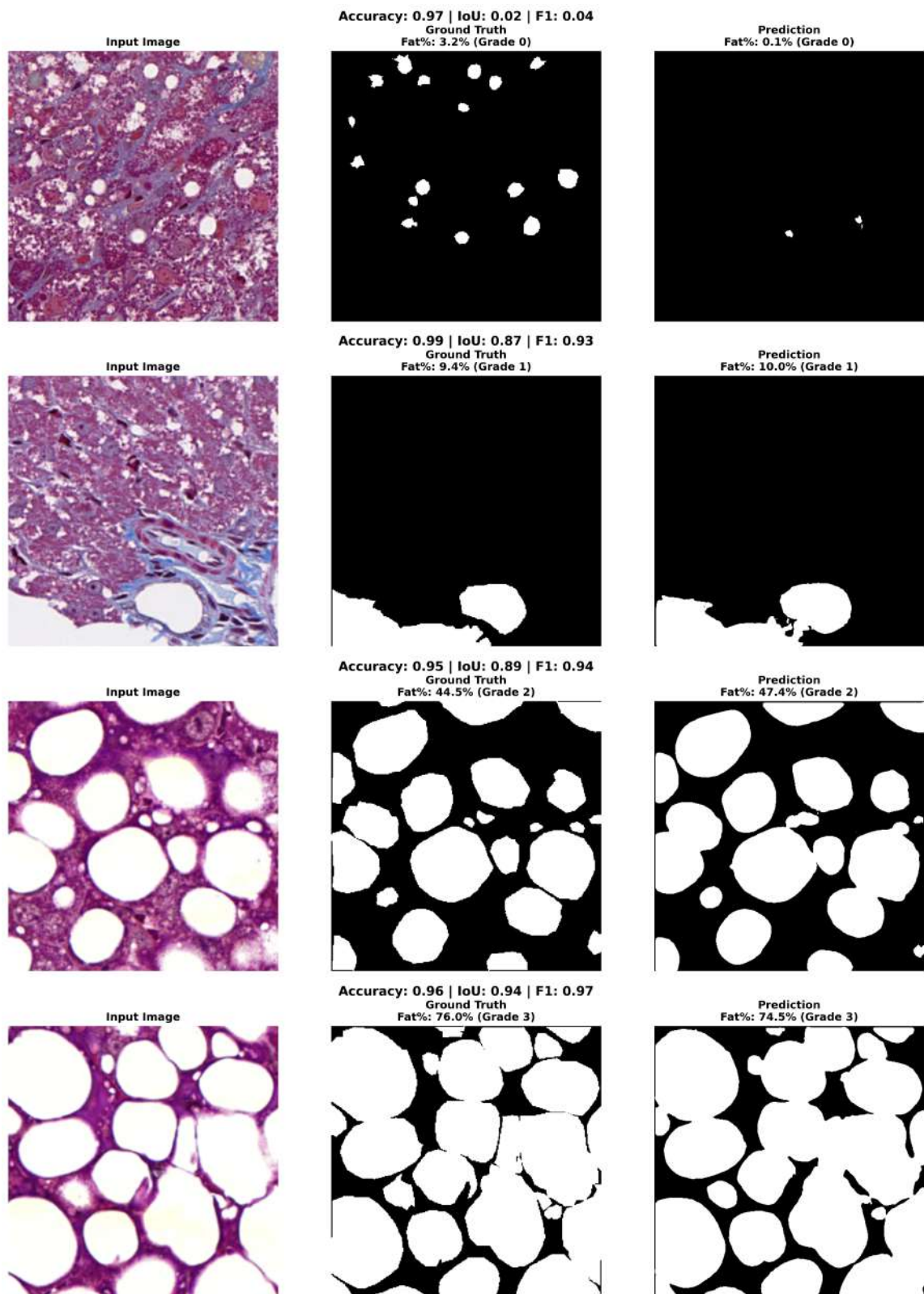


Figure 4.8: Prediction results of all 4 grades of hepatic steatosis using DUCK-Net model

4.2.4 Classification using Confusion Matrix

Confusion Matrix

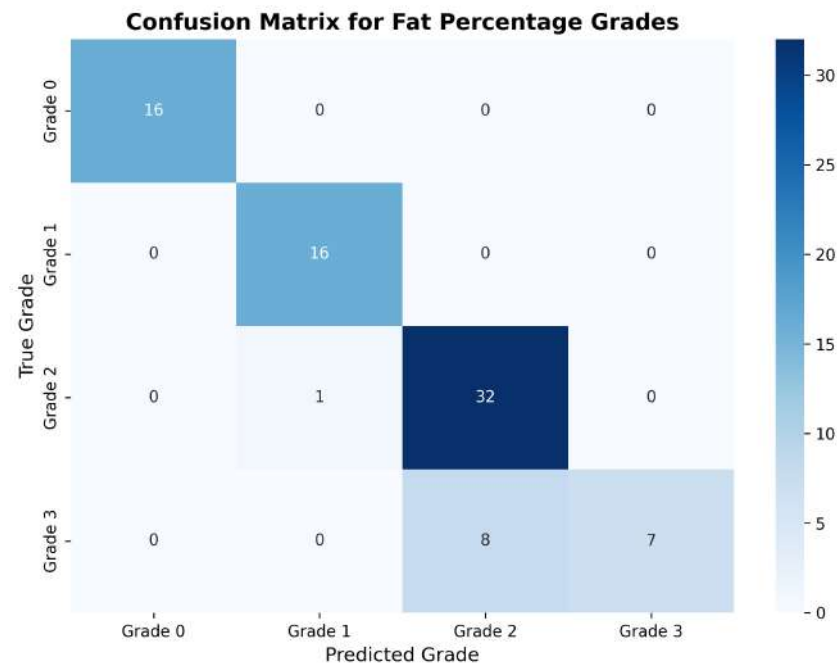


Figure 4.9: Confusion matrix for fat Percentage grades using using UNet model

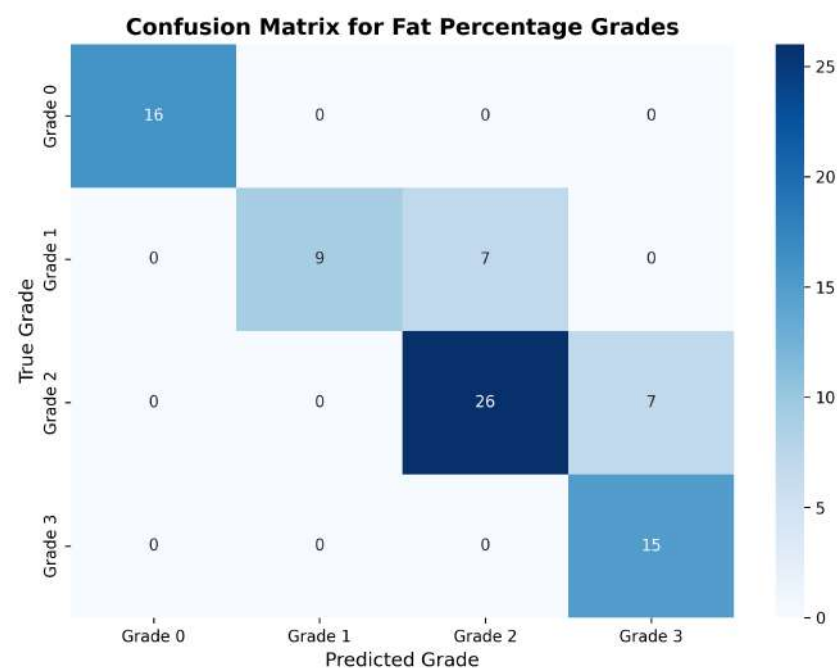


Figure 4.10: Confusion matrix for fat Percentage grades using using ResUNet++ model

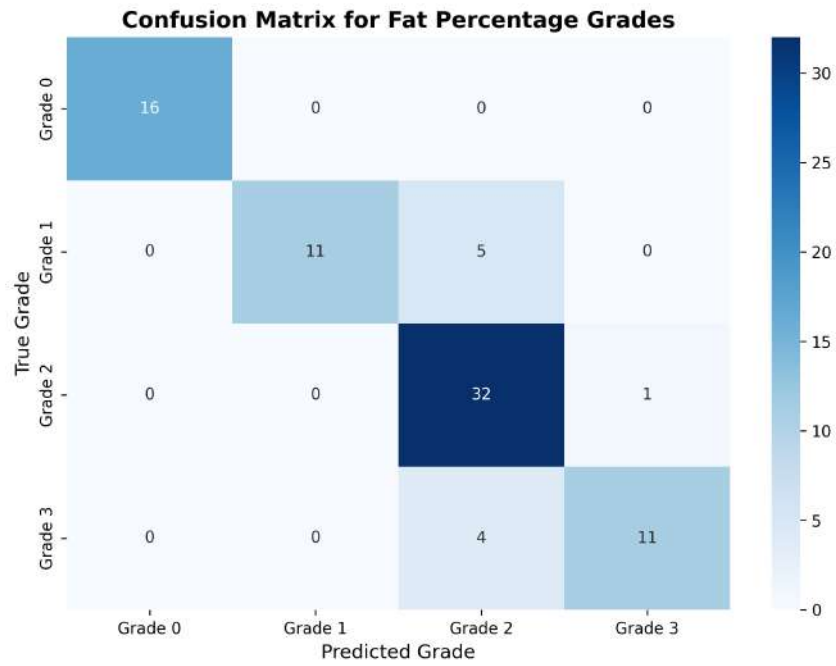


Figure 4.11: Confusion matrix for fat Percentage grades using using MIFONet model

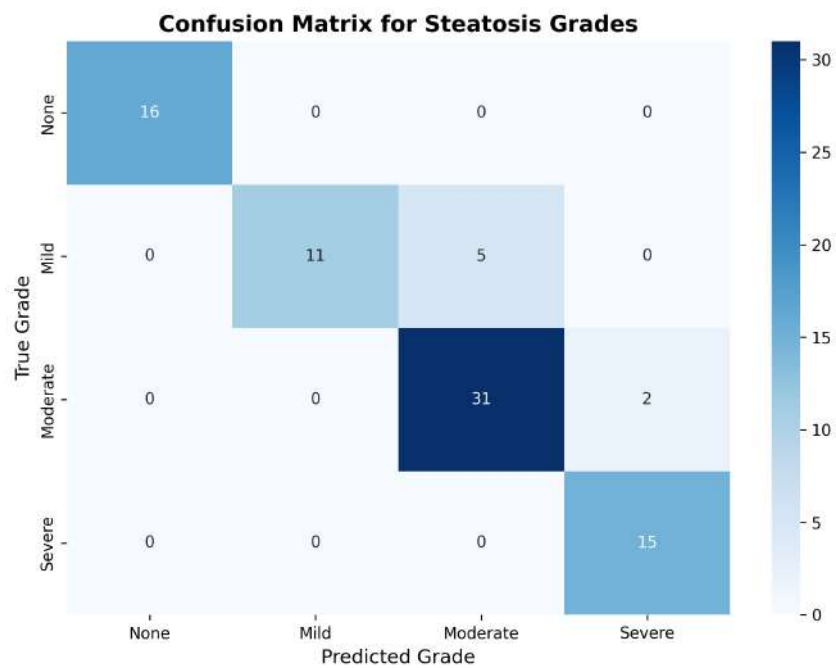


Figure 4.12: Confusion matrix for fat Percentage grades using using RAPUNet model

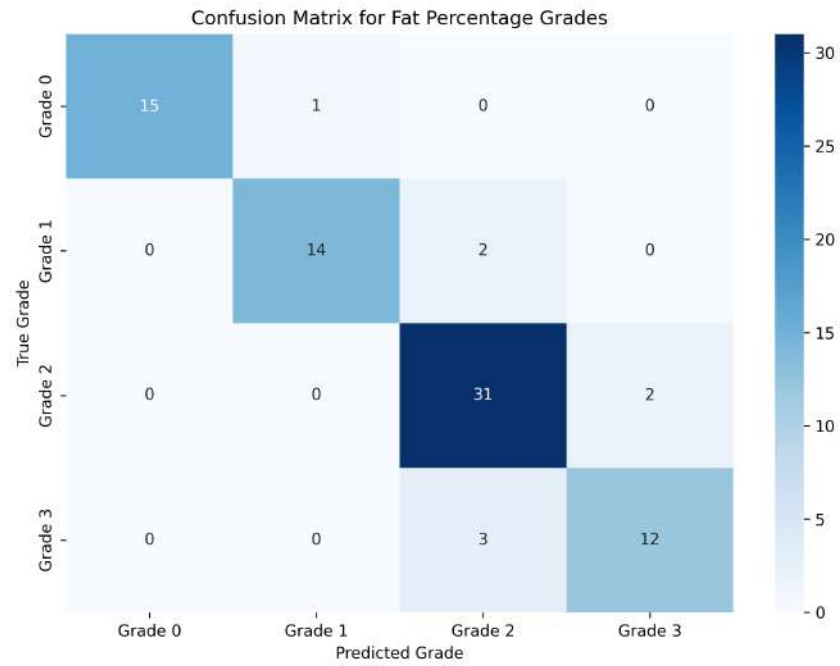


Figure 4.13: Confusion matrix for fat Percentage grades using using MedSAM model

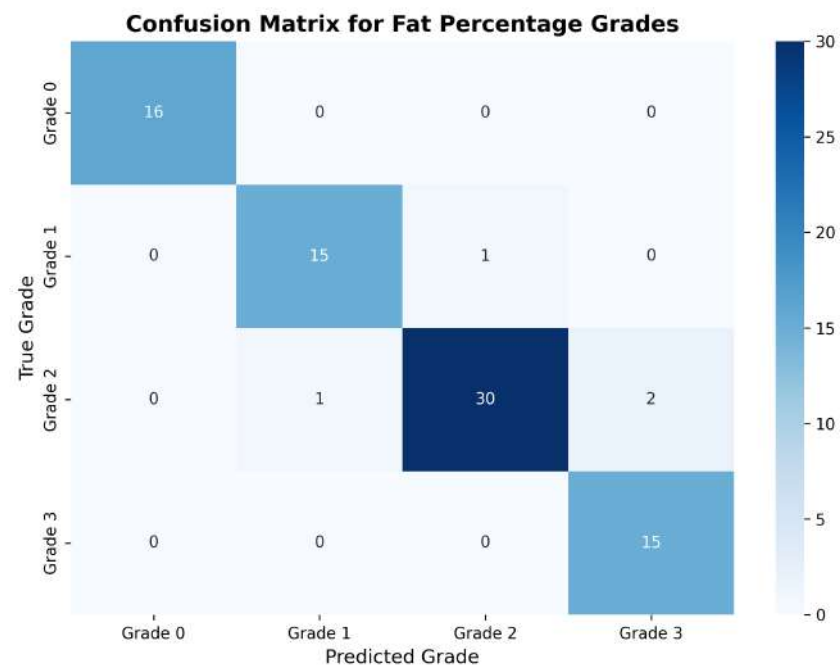


Figure 4.14: Confusion matrix for fat Percentage grades using using HarDNet-MSEG model

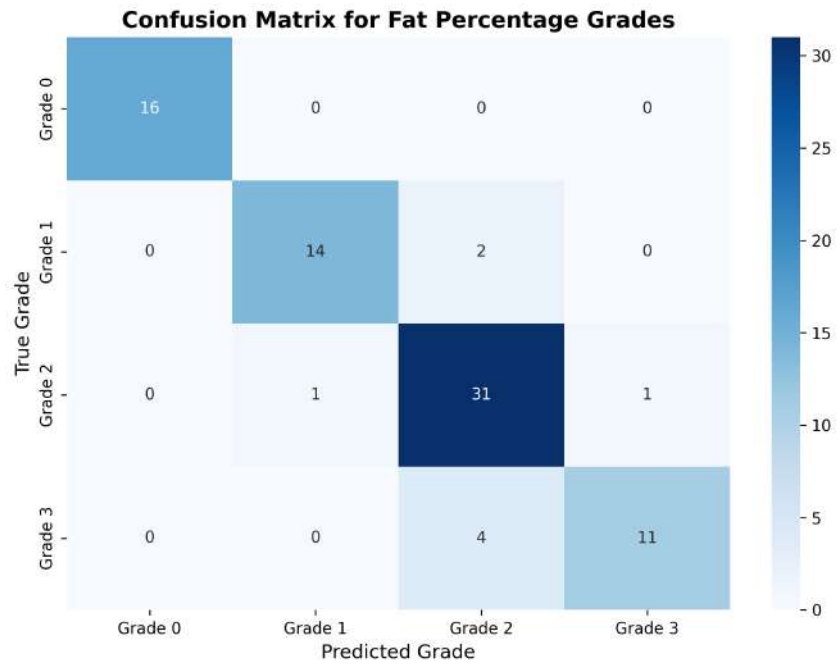


Figure 4.15: Confusion matrix for fat Percentage grades using using DeepLabV3+ model

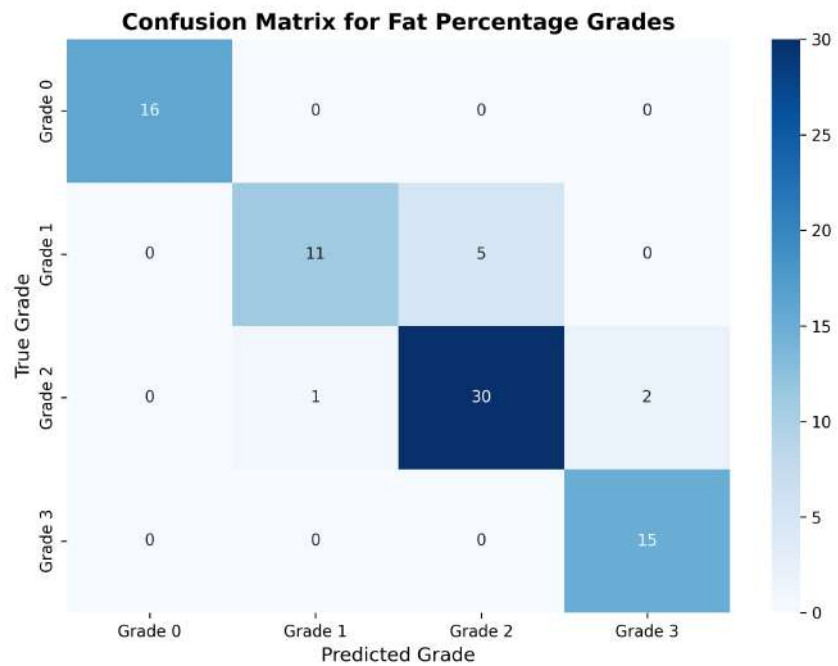


Figure 4.16: Confusion matrix for fat Percentage grades using using DUCK-Net model

4.3 Discussion

HarDNet MSEG model achieves the highest score of 0.9113 for IoU and 0.9481 for Dice being the best model for steatosis detection. Also, With only 4 wrong classifications it is the best model for grade classification of liver steatosis. Behind it is the RAPUNet model with an IOU score of 0.906 and Dice score of 0.9396 along with 7 incorrect classifications. It is also to be observed that grade 0 has almost zero incorrect classifications but poor at calculating iou whereas the iou score of other graded images was significantly better but performed poor at classification.

All these results indicate that HarDNet-MSEG is the best suited model for hepatic steatosis detection due to its balanced architecture.

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 Conclusion

This study conducted a comparative analysis of deep learning models for hepatic steatosis detection and grade classification. The results demonstrated that HarDNet-MSEG is the best suited model for steatosis segmentation, with an IoU of 0.9113, Dice score of 0.9481 for fat droplet segmentation, and a classification accuracy of [96.95%] for steatosis grades (none, mild, moderate, severe) with the lowest number of incorrect grade classifications (4). Other models such as DUCK-Net, RAPUNet, and DeepLabV3+ also delivered competitive results lower computational demands, however they were relatively less consistent in grade classification accuracy compared to HarDNet-MSEG Transformer-based models such as MedSAM excel in capturing complex features. Severe cases of steatosis is detected with more precision compared to mild cases. The use of the VGG Image Annotator for ground truth creation and BCEDiceLoss for segmentation training ensures robust model optimization.

This work contributes to the field by identifying HarDNet-MSEG as an efficient and accurate model for automated hepatic steatosis analysis, with potential applications in clinical diagnostics. However, the study is limited by the dataset's modest size and reliance on manual annotations. These findings underscore the promise of deep learning in liver disease assessment, paving the way for enhanced diagnostic tools.

5.2 Future Work

Future research can explore:

- Integration of multi-modal data (e.g., combining ultrasound and MRI).
- Explore larger datasets (e.g., human biopsies via public repositories).
- Application of advanced architectures like transformers.(e.g., SAM for initial masks)
- Deployment of models in real-time clinical settings.

BIBLIOGRAPHY

- [1] Richard Adams and Laurence Bischof. Seeded region growing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(6):641–647, 1994.
- [2] Pierre Bedossa. Histological assessment of nafld. *Hepatology*, 60(4):1138–1140, 2014.
- [3] James C Bezdek. Models for pattern recognition. In *Pattern recognition with fuzzy objective function algorithms*, pages 1–13. Springer, 1981.
- [4] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pages 205–218. Springer, 2022.
- [5] Ping Chao, Chao-Yang Kao, Yu-Shan Ruan, Chien-Hsiang Huang, and Youn-Long Lin. Hardnet: A low memory traffic network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3552–3561, 2019.
- [6] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.
- [7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages –, 2017.

-
- [8] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Re-thinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2018.
- [9] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 801–818, 2018.
- [10] Razvan-Gabriel Dumitru and Darius Peteleaza. Using duck-net for polyp image segmentation. *Scientific Reports*, 13, 06 2023.
- [11] Amirhossein Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 574–584, 2022.
- [12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. IEEE, 2016.
- [13] Fabian Heinemann, Gerald Birk, and Birgit Stierstorfer. Deep learning enables pathologist-like scoring of nash models. *Scientific reports*, 9(1):18454, 2019.
- [14] Chien-Hsiang Huang, Hung-Yu Wu, and Youn-Long Lin. Hardnet-mseg: A simple encoder-decoder polyp segmentation neural network that achieves over 0.9 mean dice and 86 fps. *arXiv preprint arXiv:2101.07172*, 2021.
- [15] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.
- [16] Qi Jin et al. Mifonet: Multi-scale information fusion network for medical image segmentation. *Medical Image Analysis*, 73:102176, 2021.
- [17] Michael Kass, Andrew Witkin, and Demetri Terzopoulos. Snakes: Active contour models. *International journal of computer vision*, 1(4):321–331, 1988.

-
- [18] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [19] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.
- [20] Hyunnam Lee and Juhan Yoo. Metaformer and cnn hybrid model for polyp image segmentation. *IEEE Access*, 12:133694–133702, 2024.
- [21] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15:654, 2024.
- [22] Sachin Mehta, Mohammad Rastegari, Anat Caspi, Linda Shapiro, and Hannaneh Hajishirzi. Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation. In *Proceedings of the european conference on computer vision (ECCV)*, pages 552–568, 2018.
- [23] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979.
- [24] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello. Enet: A deep neural network architecture for real-time semantic segmentation. *arXiv preprint arXiv:1606.02147*, 2016.
- [25] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241. Springer, 2015.
- [26] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [27] Farhat Sultana et al. Duck-net: A lightweight convolutional neural network for medical image segmentation. *arXiv preprint arXiv:2304.01234*, 2023.

-
- [28] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.
- [29] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19. Springer, 2018.
- [30] Yutong Xie, Jingcheng Zhang, and Chunhua Shen. Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. *arXiv preprint arXiv:2103.03024*, 2021.
- [31] Zobair M. Younossi, Aaron B. Koenig, Dinan Abdelatif, Yousef Fazel, Linda Henry, and Mark Wymer. Global epidemiology of nonalcoholic fatty liver disease—meta-analytic assessment of prevalence, incidence, and outcomes. *Hepatology*, 64(1):73–84, 2016.
- [32] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018.
- [33] Zheng Zhang, Wei Liu, and Ying Wang. Road extraction by deep residual u-net. In *IEEE Geoscience and Remote Sensing Letters*, volume 15, pages 749–753. IEEE, 2018.
- [34] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.