



Department: Electronics and Communications

DOCTORAL THESIS

3rd Cycle Doctoral (D-LMD)

Presented by

Hanine LAHMAR

With a view to obtaining the doctoral diploma in 3rd Cycle Doctoral (D-LMD)

Branch: Electronics

Specialty: Multidimensional Analysis and Machine Learning for Imaging

Topic

Automatic Segmentation Medical Image Using Deep Learning

Supported, on 03/10/ 2026, before the jury composed of:

Last and first name	Grade	Institution of affiliation	Designation
Dr. Amel Baha Houda ADAMOU-MITICHE	Professeur	University of Djelfa	President
Dr. Hilal NAIMI	Professeur	University of Djelfa	Supervisor
Dr. Mohamed DJEMAI	MCA	University of Djelfa	Examiner
Dr. Kamal SEHAIRI	MCA	Higher Normal School of Laghouat	Examiner

Acknowledgments

I would like to begin by expressing my deepest gratitude to Allah Almighty, the Most Gracious, the Most Merciful, who bestowed upon me the strength, courage, and determination necessary to complete this humble work.

I would like to express my sincere appreciation to my supervisor, Pr. Hilal NAIMI. He has always been a great support and guide throughout my studies. He was always available to listen to me, offer valuable advice, and guide my research in the right direction.

I also extend my sincere thanks and gratitude to the members of the examination committee for the honor they have bestowed upon me by agreeing to discuss and evaluate this research, chaired by Pr. Amel Baha Houda ADAMOU-MITICHE from the University of Djelfa, as well as to the examiners, namely: Dr Mohamed DJEMAI from the University of Djelfa, and Dr Kamal SEHAIRI from Higher Normal School of Laghouat, for their evaluation of this work. I would like to thank and appreciate everyone who contributed directly or indirectly to this research.



Dedication

I would like to express my sincere gratitude to my dear parents for their unwavering support, patience, and constant encouragement throughout my academic journey. Without their guidance and prayers, I would not have reached this point. I also thank my brothers and sisters for their support and encouragement every step of the way. And I must especially thank the inner child within me, who held fast to her dreams, never gave up in the face of adversity, and believed that hard work and perseverance are the path to success. She was my source of strength to continue striving for progress, and success

Abstract

The healthcare sector is witnessing rapid development in the field of artificial intelligence-powered medical diagnosis, with deep learning (DL) techniques gaining increasing attention in medical image analysis. The importance of this field lies in its ability to improve diagnostic accuracy, speed up the analysis process, and provide innovative solutions to the challenges faced by physicians in manual interpretation of medical images.

As part of this work, two enhanced Unet-based models were developed to improve the semantic segmentation accuracy of medical images. The first model deepens skip connections using additional convolutional layers to better preserve fine spatial detail without increasing computational cost. The second model uses a robust VGG19-based encoder, pre-trained on the ImageNet dataset, coupled with an enhanced decoder featuring deeper convolutional blocks for improved spatial resolution. Additionally, a preprocessing stage was introduced to investigate the impact of image enhancement on segmentation performance. Gaussian and median filters were applied separately for noise reduction, while CLAHE was employed for contrast enhancement. The experimental results showed an improvement in segmentation accuracy, proving the effectiveness of these architectural and preprocessing innovations.

To address the fundamental limitations of traditional convolutional architectures, particularly in capturing long-range dependencies, this work introduces a new model termed FGHQA-Unet, an innovative segmentation framework built upon the core Unet architecture with the integration of the Feature Global Hybrid Query Attention (FGHQA) mechanism. This mechanism relies on a hybrid attention framework that combines local and global representations, while employing a hybrid query representation that enhances the model's ability to model fine-grained contextual relationships within medical images. Through this integration, the model overcomes the constraints associated with exclusive reliance on traditional convolutions by improving feature representation through the simultaneous fusion of precise spatial information and global context, thereby achieving an effective balance between representational accuracy and computational efficiency.

In addition, the adopted preprocessing strategy contributes to improving input data quality, reducing noise, and enhancing computational efficiency, which positively reflects on the model's performance during both training and inference. The FGHQA-Unet architecture is designed to ensure a balance between computational accuracy and operational efficiency while maintaining the ability to extract fine-grained features in complex medical environments. The proposed model was evaluated on three benchmark medical segmentation datasets, BraTS, KiTS, and LiTS, where experimental results demonstrated its superiority over recent advanced methods in terms of segmentation accuracy and computational efficiency, confirming its effectiveness and practicality for DL-based medical imaging applications.

Keyword: Automatic segmentation, Medical imaging, Deep learning, Convolutional neural networks (CNN), Transformers, Hybrid models

Résumé

Le secteur de la santé connaît un développement rapide dans le domaine du diagnostic médical établi sur l'intelligence artificielle, les techniques d'apprentissage profond gagnant en importance dans l'analyse d'images médicales. L'importance de ce domaine réside dans sa capacité à améliorer la précision diagnostique, à accélérer le processus d'analyse et à apporter des solutions innovantes aux défis rencontrés par les médecins lors de l'interprétation manuelle des images médicales.

Dans le cadre de ce travail, deux modèles Unet améliorés ont été développés afin d'optimiser la précision de la segmentation sémantique des images médicales. Le premier modèle approfondit les connexions résiduelles grâce à des couches convolutionnelles supplémentaires, préservant ainsi les détails spatiaux fins sans augmenter la charge de calcul. Le second modèle utilise un encodeur robuste basé sur VGG19, pré-entraîné sur le jeu de données ImageNet, associé à un décodeur amélioré intégrant des blocs convolutionnels plus profonds pour une meilleure résolution spatiale. De plus, une étape de prétraitement a été introduite afin d'étudier l'impact de l'amélioration des images sur les performances de la segmentation. Les filtres Gaussian et median ont été appliqués séparément pour la réduction du bruit, tandis que la méthode CLAHE a été utilisée pour l'amélioration du contraste. Les résultats expérimentaux ont montré une amélioration de la précision de la segmentation, confirmant ainsi l'efficacité des approches architecturales et de prétraitement proposées.

Afin de remédier aux limitations fondamentales des architectures convolutionnelles traditionnelles notamment en ce qui concerne la capture des dépendances à longue portée, ce travail propose un nouveau modèle nommé FGHQA-Unet, un cadre innovant de segmentation fondé sur l'architecture de base d'Unet et intégrant le mécanisme Feature Global Hybrid Query Attention (FGHQA). Ce mécanisme repose sur un cadre d'attention hybride qui combine des représentations locales et globales, tout en employant une représentation hybride des requêtes (queries) qui renforce la capacité du modèle à modéliser des relations contextuelles fines au sein des images médicales. Grâce à cette intégration, le modèle dépasse les contraintes liées à l'utilisation exclusive des convolutions classiques en améliorant la représentation des caractéristiques par la fusion simultanée d'informations spatiales précises et du contexte global, permettant ainsi d'atteindre un équilibre efficace entre précision représentationnelle et efficacité computationnelle.

Par ailleurs, la stratégie de prétraitement adoptée contribue à améliorer la qualité des données d'entrée, à réduire le bruit et à accroître l'efficacité computationnelle, ce qui se reflète positivement sur les performances du modèle lors de l'entraînement et de l'inférence. L'architecture FGHQA-Unet a été conçue de manière à garantir un équilibre entre précision computationnelle et efficacité opérationnelle, tout en conservant la capacité d'extraire des caractéristiques fines dans des environnements médicaux complexes. Le modèle proposé a été évalué sur trois jeux de données de référence en segmentation médicale, BraTS, KiTS et LiTS, où les résultats expérimentaux ont démontré sa supériorité.

orité par rapport aux méthodes récentes de l'état de l'art, tant en termes de précision de segmentation que d'efficacité computationnelle, confirmant ainsi son efficacité et sa pertinence pour les applications d'imagerie médicale basées sur l'apprentissage profond.

Mots clés: Segmentation automatique, imagerie médicale, apprentissage profond, réseaux de neurones convolutifs (CNN), transformeurs, modèles hybrides

ملخص

يشهد قطاع الرعاية الصحية تطوراً سريعاً في مجال التشخيص الطبي المدعوم بالذكاء الاصطناعي، مع تزايد الاهتمام بتقنيات التعلم العميق في تحليل الصور الطبية. تكمن أهمية هذا المجال في قدرته على تحسين دقة التشخيص، وتسريع عملية التحليل، وتقديم حلول مبتكرة للتحديات التي يواجهها الأطباء عند التفسير اليدوي للصور الطبية.

في إطار هذا العمل، طُوِّر نموذجان مُحسَّنان قائمان على شبكة Unet لتحسين دقة التجزئة الدلالية للصور الطبية. يتميز النموذج الأول وصلات التخطي باستخدام طبقات التفاضلية إضافية للحفاظ على التفاصيل المكانية الدقيقة دون زيادة في التكلفة الحسابية. أما النموذج الثاني، فيستخدم مُشَقِّراً قوياً قائماً على شبكة VGG19، مُدَرَّباً مسبقاً على مجموعة بيانات ImageNet، مقترناً بمفكك مُحسَّن يتميز بكتل التفاضلية أعمق لتحسين الدقة المكانية. بالإضافة إلى ذلك، تم استحداث مرحلة للمعالجة الأولية لدراسة تأثير تحسين الصورة على أداء عملية التجزئة؛ حيث تم تطبيق مرشحات غاوس والوسيط بشكل منفصل لتقليل الضوضاء، بينما استُخدمت تقنية CLAHE لتحسين التباين. وقد أظهرت النتائج التجريبية تحسناً في دقة التجزئة، مما يثبت فعالية هذه الابتكارات المتعلقة بالبنية ومعالجة البيانات الأولية.

ولمعالجة القيود الجوهرية في البنى الالتفافية التقليدية، لا سيما فيما يتعلق بالنقاط العلاقات بعيدة المدى، يقدم هذا العمل نموذجاً جديداً أطلق عليه FGHQA-UNet، وهو إطار مبتكر للتجزئة يعتمد على البنية الأساسية لـ U-Net مع دمج آلية الانتباه (FGHQA) Feature Global Hybrid Query Attention. تعتمد هذه الآلية على تمثيل هجين للاستعلامات (Queries) يدمج بين الخصائص المحلية والعالمية، مما يعزز قدرة النموذج على نمذجة العلاقات السياقية الدقيقة داخل الصور الطبية. ومن خلال هذا التكامل، يتجاوز النموذج القيود المرتبطة بالاعتماد الحصري على الالتفافات التقليدية، حيث يتم تحسين تمثيل الميزات عبر دمج المعلومات المكانية الدقيقة مع السياق العالمي بصورة متزامنة، بما يحقق توازناً فعالاً بين الدقة التمثيلية والكفاءة الحسابية. كما تسهم استراتيجية المعالجة المسبقة المعتمدة في رفع جودة البيانات المدخلة وتقليل الضوضاء وتحسين الكفاءة الحسابية، مما ينعكس إيجاباً على أداء النموذج أثناء التدريب والاستدلال. وقد تم تصميم بنية FGHQA-UNet بما يضمن تحقيق توازن بين الدقة الحسابية والكفاءة التشغيلية مع الحفاظ على القدرة على استخراج السمات الدقيقة في البيانات الطبية المعقدة. وقد تم تقييم النموذج المقترح على ثلاث مجموعات بيانات معيارية للتجزئة الطبية هي BraTS و KiTS و LiTS، حيث أظهرت النتائج التجريبية تفوقه على أحدث الأساليب المنافسة من حيث دقة التجزئة والكفاءة الحسابية، مما يؤكد فعاليته وجدواه في تطبيقات التصوير الطبي المعزز بالتعلم العميق.

الكلمات المفتاحية: التجزئة التلقائية، التصوير الطبي، التعلم العميق، الشبكات العصبية التلافيفية (CNN)،

المحولات، النماذج الهجينة

List of Abbreviations

AI	Artificial Intelligence
ACM	Active Contour Model
BN	Batch Normalization
BraTS	Brain Tumor Segmentation
CLAB	Convolutional Linear Attention Block
CNN	Convolutional Neural Network
CoC	Cluster of Clusters
CoTr	Contextual Transformer
CT	Computed Tomography
CUTS	Contrastive Unpaired Translation for Segmentation
CLAHE	Contrast Limited Adaptive Histogram Equalization
DCB	Depthwise Convolution Bottleneck
DL	Deep Learning
DNNs	Deep Neural Networks
DNA	Deoxyribonucleic Acid
DSC	Depthwise Separable Convolutional
DXA	Dual-energy X-ray Absorptiometry
ELU	Exponential Linear Unit
ET	Enhancing Tumor
FCNs	Fully Convolutional Networks
FPNs	Feature Pyramid Networks
FGHQA	Feature Global Hybrid Query Attention
GLCM	Gray-Level Co-occurrence Matrix
HD95	95th Percentile Hausdorff Distance
HU	Hounsfield Unit
IoU	Intersection Over Union
ISBI	International Symposium on Biomedical Imaging
KiTS	Kidney Tumor Segmentation Challenge
LiTS	Liver Tumor Segmentation Challenge
LGG	Lower-Grade Gliomas
LReLU	LeakyReLU
LSTM	Long Short-Term Memory
MCPM	Meta-learning-based Confidence Prediction Mask
MedT	Medical Transformer
ML	Machine Learning
MLP	Multilayer Perceptron

MSA	Multi-Head Self-Attention
MRI	Magnetic Resonance Imaging
mRNA	Messenger Ribonucleic Acid
MHSA	Multi-Head Self-Attention
NLP	Natural Language Processing
NMS	Non-Maximum Suppression
OCT	Optical Coherence Tomography
PET	Positron Emission Tomography
PReLU	Parametric ReLU
R-CNN	Region-Based Convolutional Neural Network
RCC	Renal Cell Carcinoma
ReLU	Rectified Linear Unit
RNA	Ribonucleic Acid
RPPA	Reverse Phase Protein Array
SNNs	Self-Normalizing Neural Networks
SVMs	Support Vector Machines
Swin	Shifted Window
TC	Tumor Core
TL	Transfer Learning
TERT	Telomerase Reverse Transcriptase
VGG	Visual Geometry Group
WHO	World Health Organization
WT	Whole Tumor

Contents

Contents	XII
List of Figures	XIV
List of Tables	XV
General introduction	1
I Previous studies	3
I.1 Introduction	4
I.2 Introduction to Medical Image Segmentation	4
I.3 Traditional Image Segmentation Techniques	5
I.3.1 Pixel-Based Image Segmentation	5
I.3.2 Contour-Based Image Segmentation	6
I.3.3 Region-Based Image Segmentation	7
I.4 Methodological developments in computational modeling for medical im- age segmentation.	8
I.4.1 DL for Medical Image Segmentation	8
I.4.1.1 Key DL Architectures for Segmentation	9
I.4.1.1.1 CNN-based Models	9
I.4.1.1.2 Transformer-based Models	11
I.4.1.1.3 Hybrid CNN-Transformer models	12
I.5 Comparative Analysis of Medical Image Segmentation Methods	12
I.5.1 Evaluation Metrics for Medical Image Segmentation	13
I.5.2 Task-Oriented Performance Analysis	14
I.5.2.1 Tissue Heterogeneity in Medical Image Segmentation	14
I.5.2.2 Segmentation large Organ and tumors	14
I.5.2.3 Fine Structure Segmentation	15
I.5.3 Synthesis and Strategic Recommendations	15
I.6 Overview of Datasets Used in Medical Image Segmentation	15
I.7 Conclusion	17
II Principles and Applications of Medical Imaging Modalities	18
II.1 Introduction	19
II.2 Types of medical imaging	19
II.2.1 X-Rays	20
II.2.2 Computed Tomography (CT)	21
II.2.3 Medical Resonance Imaging (MRI)	22
II.2.4 Colonoscopy	24
II.2.5 Positron Emission Tomography (PET)	25

II.2.6	Ultrasound Images	26
II.2.7	Dermoscopy	27
II.2.8	Microscopic Images	28
II.2.9	Optical Coherence Tomography (OCT)	29
II.3	Medical Imaging Modalities: Principles, Applications, and Comparative Analysis.	30
II.4	Conclusion	31
III Hybrid DL Architectures for Medical Image Segmentation: Integrating CNNs and Transformers		32
III.1	Introduction	33
III.2	DL in Computer Vision and Healthcare	33
III.3	Neural Networks	34
III.4	Convolutional Neural Network (CNN)	35
III.4.1	Principal layers	36
III.4.1.1	Convolutional layer	36
III.4.1.2	Pooling Layer	37
III.4.1.3	Fully-Connected Layer (FCL)	37
III.4.2	Supportive Components	38
III.4.2.1	Activation Functions	38
III.4.2.2	Batch Normalization	39
III.4.2.3	Dropout	40
III.5	Applications of CNN in Medical Image Analysis	41
III.5.1	Segmentation	41
III.5.1.1	Types of Image Segmentation	43
III.5.1.1.1	Semantic segmentation:	43
III.5.1.1.2	Instance segmentation:	43
III.5.1.1.3	Panoptic segmentation:	43
III.5.1.2	Segmentation in medical image:	43
III.5.1.2.1	Unet architecture:	44
III.6	Transformer for medical image analysis	45
III.6.1	ViTs and Their Adaptation to Medical Imaging	46
III.6.2	Transformer-based Models in Medical Image Segmentation	47
III.6.2.1	Hybrid Architectures Combining CNNs and Transformers	47
III.7	Conclusion	49
IV Improving medical semantic segmentation using modified Unet models.		50
IV.1	Introduction	51
IV.2	Dataset	51
IV.3	Experiment 01	52
IV.3.1	Proposed DCB-Unet model	53
IV.3.2	Quality metrics	54
IV.3.2.1	Dice Similarity Coefficient	54
IV.3.2.2	Accuracy	54
IV.3.2.3	IoU coefficient	55
IV.3.2.4	Loss function	55
IV.3.2.5	Training and Implementation Settings	55
IV.3.3	Experimental results and discussion	55

IV.3.3.1	Results	55
IV.3.3.2	Comparison with previous studies	57
IV.4	Exepriment 02:	59
IV.4.1	The modified VGG19-Unet model	61
IV.4.1.1	VGG19 Architecture	61
IV.4.1.2	Depthwise Separable Convolutions (DSC)	61
IV.4.1.3	VGG19-Unet modified	62
IV.4.2	Experimental Methodological Framework	64
IV.4.3	Results and performance evaluation of the modified VGG19-Unet	64
IV.4.4	Preprocessing and Hybrid Model Methodology	68
IV.4.4.1	Filtering Techniques for Noise Reduction	68
IV.4.4.2	Contrast Enhancement Using CLAHE	70
IV.4.4.3	Performance Evaluation with Preprocessed Data	71
IV.4.5	Discussion of the results and comparison with previous studies	74
IV.5	Conclusion	75
V	A Hybrid Unet with FGHQA for Medical Image Segmentation	77
V.1	Introduction	78
V.2	Datasets	78
V.2.1	BraTS2020 Dataset	78
V.2.2	KiTS19 Dataset	79
V.2.3	LiTS17 Dataset	80
V.3	Preprocessing data	81
V.3.1	BraTS preproccesing	81
V.3.2	KiTS19 and LiTS17 preproccesing	83
V.4	Proposed FGHQA-Unet Model	84
V.5	Evaluation Metrics	87
V.5.1	Training and Implementation Settings	88
V.6	Results and discussion	88
V.6.1	Experiment 01	88
V.6.1.1	Results	88
V.6.1.2	Discussion	92
V.6.2	Experiment 02	93
V.6.2.1	Results	93
V.6.2.2	Discussion	96
V.6.3	Experiment 03	98
V.6.3.1	Results	98
V.6.3.2	Discussion	101
V.7	Conclusion	102
	General conclusion	107

List of Figures

II.1	Illustrative examples of diverse medical imaging modalities in clinical practice	19
II.2	Projectional radiography	20
II.3	Computerized Axial Tomography Scan	22
II.4	MRI Example: 2D Magnetic Resonance Angiography (MRA) of Brain Blood Vessels	23
II.5	Diagram of the Colonoscopy	24
II.6	Positron Emission Tomography scans	25
II.7	Example of Abdominal Ultrasound Scan	26
II.8	Diagnosis for skin cancer with dermoscopy	27
II.9	Scanning Electron Microscope (SEM) Components	28
II.10	OCT Machine Components and Eye Anatomy Overview	30
III.1	Schematic representation Neural Network Architecture	33
III.2	An artificial neuron, or node, is the cornerstone of DNNs	34
III.3	Artificial feedforward multilayer neural network	35
III.4	The basic architecture of CNN	36
III.5	Example of a 3×3 convolution operation with a stride length of 1 and no padding	36
III.6	The Most Popular Pooling Strategies in CNN Networks	37
III.7	Step-by-Step Batch Normalization Process.	40
III.8	Example of apply dropout in a neural network.	40
III.9	Conceptual Illustration of Image Segmentation.	42
III.10	Example of segmentation in various medical images.	42
III.11	Examples of types of Image Segmentation	44
III.12	The Unet architecture is illustrated here with an example for a resolution of 64x64 pixels.	45
IV.1	The histogram represents the types of analyses and the number of samples corresponding to each.	52
IV.2	FLAIR MRI images: Original image without tumor and with tumor, ground truth image without tumor and with tumor.	52
IV.3	Proposed DCB-Unet model.	53
IV.4	Depthwise convolution bottleneck.	54
IV.5	(a) Accuracy plot for proposed model (b) IoU coefficient plot for proposed model (c) Dice coefficient plot for proposed model (d) Loss plot for proposed model	56
IV.6	Ground truth image and segmented image	57
IV.7	Bar graph showing the training, validation, and testing parameters. . . .	58
IV.8	(a) IoU coefficient plot for Kumar et al. (b) DSC coefficient plot for . . .	59

IV.9	(a) IoU coefficient plot for Sathish et al. (b) DSC coefficient plot for . . .	60
IV.10	(a) IoU coefficient plot for. (b) DSC coefficient plot for	60
IV.11	(a) IoU coefficient plot for Joshi et al. (b) DSC coefficient plot for	60
IV.12	(a) IoU coefficient plot for proposed approach. (b) DSC coefficient plot for the proposed approach	61
IV.13	Architecture Diagram of VGG19 CNNs	62
IV.14	A diagram showing the difference between natural and deep convolution .	63
IV.15	The architecture of the modified VGG19-Unet model	63
IV.16	Segmentation performance curves: (a) Dice, (b) IoU, and (c) loss.	65
IV.17	The original image and their corresponding segmented representations. . .	67
IV.18	FLAIR MRI images after application of the Gaussian filter.	69
IV.19	FLAIR MRI images after application of the Median filter.	70
IV.20	FLAIR MRI images after application of the CLAHE.	72
IV.21	Segmentation performance curves under different preprocessing techniques: (a) Dice, (b) IoU, and (c) The loss.	73
IV.22	FLAIR MRI images after application of the CLAHE.	74
V.1	BraTS2020 data samples: Display of brain imaging using different MRI techniques (FLAIR, T1, T1ce, T2) with segmentation masks (Mask). . . .	79
V.2	CT data sample demonstration: normal slices and segmentation masks for kidneys and tumors.	79
V.3	CT data sample demonstration: normal slices and segmentation masks for Liver and tumors	80
V.4	The proposed pretreatment of magnetic resonance images. Cutting sizes (above) original size: 240×240 ; after cutting: 128×128 . Multimedia photo collection (below), including T1, T1ce, T2, Flair	81
V.5	A Schematic of the Volumetric to 2D Slice Conversion Pipeline for Medical Image Analysis: (a) for the LiTS dataset, (b) for the KiTS dataset	83
V.6	(a) represents the enhanced images of CT slices of the kidney, (b) repre- sents the enhanced images of the liver using different HU window ranges .	84
V.7	An overview of the FGHQA-Unet architecture is presented. We derive sequence representations from various layers within the transformer and integrate them with the decoder through skip connections.	85
V.8	Proposed FGHQA Hybrid Transformer Encoder Block.	86
V.9	Mean $\pm$ SD of Segmentation Metrics for Selected Slices	89
V.10	Visual comparison of multi-channel radiomics and model predictions in brain tumor segmentation	90
V.11	Graphs of segmentation performance metrics for different tumor regions: (a) dice coefficient, (b) Hausdorff distance 95, and (c) Intersection over union	91
V.12	Mean $\pm$ SD of Segmentation Metrics for Selected Slices	94
V.13	Visualized segmentation samples of KiTS19	95
V.14	Graphs of segmentation performance metrics for Kidney and tumor: (a) dice Score, (b) Hausdorff distance 95, and (c) IoU Score.	96
V.15	Mean $\pm$ SD of Segmentation Metrics for Selected Slices	99
V.16	Visualized segmentation samples of LiTS17	100
V.17	Graphs of segmentation performance metrics for different tumor regions: (a) dice Score, (b) Hausdorff distance 95, and (c) IoU Score.	101

List of Tables

I.1	Traditional Segmentation Methods: Capabilities and Clinical Constraints .	13
I.2	DL Architectures for Medical Segmentation: Performance and Trade-offs .	13
I.3	Common Evaluation Metrics for Medical Image Segmentation	14
I.4	Summary of publicly available datasets for AI-based medical image seg- mentation	16
II.1	Comparison of Diagnostic Imaging Technologies	30
III.1	Summary of activation function	39
IV.1	Displaying previous works in last years	56
IV.2	Comparison of the previous segmentation methodology with the proposed model	59
IV.3	Performance Analysis	64
IV.4	Effect of Preprocessing on VGG19-Unet Performance	72
IV.5	Comparative performance analysis of the proposed modified VGG19-UNet model against advanced methods	75
V.1	distribution of CT value ranges across primary tissues in abdominal CT scans	84
V.2	Brain tumor segmentation test result with different evaluation metrics on BraTS 2020	89
V.3	The segmentation results for our approach and the methods we compared .	92
V.4	Comparison of different models in terms of segmentation performance and computational complexity.	93
V.5	Kidney and tumor segmentation result with different evaluation metrics on KiTS19	94
V.6	The segmentation results for our approach and the methods we compared .	97
V.7	The segmentation results for our approach and the methods we compared.	97
V.8	Liver and tumor segmentation result with different evaluation metrics on LiTS17	98
V.9	The segmentation results for our approach and the methods we compared .	102
V.10	The segmentation results for our approach and the methods we compared.	102

General introduction

Medical image segmentation has become one of the fundamental tasks in modern medical image analysis due to its essential role in identifying anatomical structures and pathological regions with high precision. Accurate segmentation provides detailed quantitative information that supports disease diagnosis, treatment planning, surgical guidance, and longitudinal monitoring of disease progression. Over the past decade, the rapid advancement of deep learning, particularly convolutional neural networks (CNNs) and, more recently, hybrid architectures integrating attention mechanisms and Transformer-inspired models, has significantly improved segmentation performance by enabling automatic extraction of hierarchical image representations directly from medical data. These developments have established deep learning as the dominant paradigm for medical image segmentation and have substantially outperformed conventional image processing and machine learning techniques [1–3].

Despite these remarkable advances, developing segmentation models that are simultaneously accurate, robust, and computationally efficient remains a significant challenge. Medical images exhibit considerable variability in terms of imaging modalities, anatomical structures, tumor morphology, making it difficult for a single segmentation strategy to perform consistently across diverse clinical scenarios. Existing approaches often encounter a trade-off between preserving fine spatial information, learning discriminative feature representations, modeling long-range contextual dependencies, and maintaining acceptable computational complexity. Consequently, improving medical image segmentation requires more than optimizing a single component of the segmentation pipeline. Instead, it demands the complementary integration of architectural innovation, efficient feature learning, high-quality data representation, and practical computational efficiency to address the diverse challenges encountered across different clinical scenarios [4–6].

Although DL has significantly advanced medical image segmentation, most existing studies have primarily focused on improving segmentation performance through architectural modifications or by maximizing segmentation accuracy. Considerable progress has therefore been achieved in feature extraction, contextual modeling, and network optimization. However, comparatively limited attention has been paid to the practical diversity of clinical applications, where segmentation models are expected to operate under different computational resources, diagnostic objectives, and precision requirements. Consequently, the suitability of a segmentation model depends not only on its segmentation accuracy but also on its ability to satisfy the specific demands of the intended clinical scenario. These observations indicate the need for a broader research perspective that investigates complementary segmentation strategies capable of addressing different clinical and computational requirements.

To address these challenges, this thesis aims to develop and evaluate three comple-

mentary deep learning-based segmentation models, each designed to investigate a different aspect of medical image segmentation. Rather than relying on a single architectural improvement, the proposed research explores multiple design strategies, including enhanced spatial information transfer, efficient feature extraction with reduced computational complexity, and hybrid local-global contextual modeling. In addition, the study investigates the contribution of data preprocessing and data selection strategies to segmentation performance. Through these complementary directions, the thesis seeks to demonstrate that robust and accurate medical image segmentation is achieved through the integration of architectural innovation, effective data representation, and computational efficiency, thereby providing practical solutions applicable across different tumor types, imaging modalities.

This thesis makes several scientific contributions to the field of medical image segmentation through the development and comprehensive evaluation of three complementary deep learning-based models. The first contribution consists of enhancing the U-Net architecture by replacing its conventional skip connections with Depthwise Convolution Bottleneck (DCB) modules, thereby improving spatial information transfer while preserving computational efficiency. The second contribution introduces an enhanced VGG19-based segmentation model that combines a pre-trained encoder with an optimized decoder based on depthwise separable convolutions, reducing the number of trainable parameters while maintaining high segmentation performance. This contribution also includes a systematic investigation of different preprocessing techniques, demonstrating the superiority of CLAHE over Gaussian and Median filtering for enhancing input image quality. Finally, the thesis proposes FGHQA-UNet, a novel hybrid segmentation framework featuring a hybrid encoder that integrates convolutional neural networks with attention mechanisms to effectively combine local feature extraction and global contextual modeling. This framework is further complemented by a tumor-guided slice selection strategy based on tumor pixel count and tumor density, enabling the model to focus on the most informative slices while improving both training efficiency and segmentation performance.

This thesis is organized into five chapters. Chapter 1 provides the theoretical background of medical image segmentation, reviewing its evolution from conventional image processing techniques to modern deep learning approaches. Chapter 2 introduces the main medical imaging modalities, benchmark datasets, and evaluation metrics relevant to tumor segmentation. Chapter 3 presents the fundamental deep learning architectures employed in medical image segmentation, with particular emphasis on convolutional neural networks, attention mechanisms, and hybrid CNN-Transformer models. Chapter 4 presents two enhanced U-Net-based segmentation models, including a DCB-based architecture for improved spatial information transfer and an enhanced VGG19-based model incorporating an optimized decoder and preprocessing techniques. Finally, Chapter 5 introduces the proposed FGHQA-UNet framework, which combines a hybrid encoder, attention mechanisms, and a tumor-guided slice selection strategy, followed by a comprehensive experimental evaluation on the BraTS, LiTS, and KiTS benchmark datasets.

CHAPTER I

PREVIOUS STUDIES

Contents

I.1	Introduction	4
I.2	Introduction to Medical Image Segmentation	4
I.3	Traditional Image Segmentation Techniques	5
I.3.1	Pixel-Based Image Segmentation	5
I.3.2	Contour-Based Image Segmentation	6
I.3.3	Region-Based Image Segmentation	7
I.4	Methodological developments in computational modeling for medical image segmentation.	8
I.4.1	DL for Medical Image Segmentation	8
I.4.1.1	Key DL Architectures for Segmentation	9
I.4.1.1.1	CNN-based Models	9
I.4.1.1.2	Transformer-based Models	11
I.4.1.1.3	Hybrid CNN-Transformer models	12
I.5	Comparative Analysis of Medical Image Segmentation Methods	12
I.5.1	Evaluation Metrics for Medical Image Segmentation	13
I.5.2	Task-Oriented Performance Analysis	14
I.5.2.1	Tissue Heterogeneity in Medical Image Segmentation	14
I.5.2.2	Segmentation large Organ and tumors	14
I.5.2.3	Fine Structure Segmentation	15
I.5.3	Synthesis and Strategic Recommendations	15
I.6	Overview of Datasets Used in Medical Image Segmentation	15
I.7	Conclusion	17

I.1 Introduction

Medical image segmentation represents a critical task in modern healthcare, enabling precise delineation of anatomical structures and pathological regions for improved diagnosis and treatment planning. This chapter provides a comprehensive review of the field’s evolution, from early traditional methods to contemporary DL approaches. We examine the fundamental techniques that have shaped medical image segmentation, including pixel-based, contour-based, and region-based methods, before exploring the revolutionary impact of CNNs and transformer architectures. The review systematically analyzes each method’s technical foundations, clinical applications, and performance characteristics while highlighting key datasets. By synthesizing this extensive body of research, we aim to provide readers with a clear understanding of current capabilities and limitations, setting the stage for informed discussions about future directions in medical image analysis.

I.2 Introduction to Medical Image Segmentation

Automatically recognizing semantic concepts in numeric images, such as a specific organ or disease in a medical image or a specific target object in a video sequence, is a computer vision research subject of paramount importance due to its theoretical and methodological challenges as well as the breadth of its applications [7]. Scene-understanding algorithms would enable machines to extract very valuable information from visual data, with an enormous potential to enable powerful predictive models, to provide new insights, and to guide decisions. For instance, semantic segmentation, which is the main focus of this thesis, is of pivotal importance in various image interpretation systems. This difficult and cumbersome computational task aims at automatically assigning a semantic label (or class) to every pixel in an image, for example “car” or “person” in a street scene, acquired by a standard camera, or a tumor structure in a medical Magnetic Resonance Imaging (MRI). In medical imaging, challenges such as variability in anatomy across patients, imaging noise, and low contrast between tissues further complicate segmentation tasks [7,8].

In medical imaging, segmentation information directly translates into quantitative measures for disease diagnosis, treatment, and follow-up, as well as surgery planning, radiology reporting, and healthcare research and practices. Powerful and accurate segmentation algorithms have the potential to disrupt significant application areas of medical imaging, such as cardiology, neuroscience, and oncology. More generally, the problem is of pivotal importance in a breadth of application areas, such as autonomous driving, surveillance, remote sensing, industrial inspections, and web-based commercial platforms [3,4].

From an algorithmic perspective, visual recognition problems, including semantic segmentation, are very challenging and domain-dependent because of the diversity, variability, and complexity of real-world scenes [9]. The current overwhelming growth of large-scale image data acquired and stored every day, coupled with the recent computational and algorithmic breakthroughs in DL research, has produce unprecedented advances in visual-recognition performances in the last decade, triggering wide interest in both academia and industry. There is a clear consensus that, in the next decades, such algorithms will disrupt strategic areas such as health care, security, autonomous systems and robotics, remote sensing, manufacturing, and social media, among many others. However, this also raises substantial computational and algorithmic challenges, which call

for bringing current image interpretation algorithms to a whole new level of automation, scalability, and accuracy [10].

I.3 Traditional Image Segmentation Techniques

Before the advent of machine learning (ML) and DL techniques, image segmentation relied on a set of traditional methods that directly address basic image properties, such as pixel intensity, contrast, and continuity [11]. These methods can be classified into three main categories based on the underlying criteria underlying the segmentation process:

I.3.1 Pixel-Based Image Segmentation

Pixel-based image segmentation methods focus primarily on thresholding and clustering techniques. Because thresholding methods rely heavily on the number of thresholds selected and the choice of parameters, their resulting segmentation results are often inaccurate. In contrast, clustering methods are more popular and widespread due to their robustness and better ability to handle image variations.

Clustering techniques predominantly include hierarchical clustering [12], c-means clustering [13], fuzzy c-means clustering [14], density clustering [15], and spectral clustering [16]. The hierarchical clustering algorithm leverages the similarity among pixels to perform hierarchical decomposition. Although this algorithm is straightforward and easy to implement, the resulting segmentation outcomes are contingent upon the construction of the tree and the selection of thresholds; therefore, image segmentation techniques based on hierarchical clustering are not frequently employed.

Pixel-based segmentation techniques are an important approach in medical image analysis, treating each pixel as an independent unit of analysis. This idea first emerged in 2009, developed by Sladoje and Lindblad [17], and focused on improving segmentation accuracy by re-evaluating boundary pixels after initial segmentation, using gray-level information to estimate each pixel's affinity to the object or background. This approach utilized a complementary strategy to crisp segmentation methods, allowing for more accurate estimates of quantitative features, especially in histological images.

However, this approach was limited in handling topological complexities or subtle variations in intensity. To address these shortcomings, the Gradient-Based Seeded Region Growing (GBSRG) technique was developed by G. N. Rai and T. R. Nair in 2010 [18]. It focused on improving the homogeneity criterion by incorporating a density gradient, allowing for more accurate boundary definition, especially for complex structures such as blood vessels in computed tomography (CT) images. However, this method relied on initial seeds and remained sensitive to noise and local variations, highlighting the need for more robust approaches. In 2011, a generalized classification-based segmentation system using dense neighbors and ML was proposed by Chen et al [19]. which used supervised approaches such as Support Vector Machine (SVMs) with local features (neighborhoods) and incorporated multiscale voting between the results to improve accuracy.

This approach demonstrated high flexibility across different image types (MRI, tissue images, and nuclei images), representing a step toward generalization and reliance on labeled data. However, it required large amounts of training data and complex fine-tuning processes. A major breakthrough came in 2012 with the introduction of pixel-based ML techniques by K.Suzuki [20], particularly the MTANN model, which enabled the training of massive neural networks on sub-angles of images, predicting individual pixel

values or pattern probabilities. This allowed overcoming the inaccurate segmentation and estimation errors that plague traditional methods. This strategy allowed for the direct integration of spatial context into the classification process, enhancing performance on complex medical images. The technology has been in use by researchers for many years [21–25].

I.3.2 Contour-Based Image Segmentation

Contour-based image segmentation aims to achieve segmentation by identifying the boundaries of each region within an image. Initially, this approach primarily relied on image gradient computations, utilizing well-known operators such as Sobel, Robert, and Canny. However, converting the image gradient into a closed contour presents significant challenges [26].

Consequently, traditional edge detection often employs a thresholding method to derive contour detection outcomes tailored to specific task requirements. With advancements in the application of partial differential equations in image processing, the active contour model (ACM) has gained widespread adoption for image segmentation [27]. Segmentation based on the ACM typically utilizes the energy functional method, which fundamentally involves representing the target edge with a continuous curve and defining an energy function that incorporates the edge curve as its independent variables. Therefore, the task of image segmentation is reframed as the process of determining the minimum value of the energy function [28].

In the continuous development of medical image segmentation techniques, research papers spanning the late 1990s to the early 2010s demonstrate a clear gradual improvement in accuracy and reliability, driven by the need to overcome limitations of medical images such as noise and edge blurring. This process began with the application of the ACM to MRIs to extract the cranial cavity. Matsumoto et al. [29], used a series of preparatory steps, including extracting the brain vertex from the histogram, thresholding, and then morphological operations to refine the contour, before applying the ACM. The goal was to achieve automatic brain extraction, but the strategy relied heavily on preprocessing, making it sensitive to variability in image quality. With the growing need to incorporate shape knowledge to improve segmentation, M.Rousson and N.Paragios [30], developed a framework that combined shape priors with level set models. They used a shape model based on training data to constrain boundary evolution according to accepted anatomical shapes, enhancing accuracy, especially in the presence of noise or distortions.

However, this approach relied on global transformations, which limited its ability to handle multiple objects or local shape variations. To address this, S.Lankton and A.Tannenbaum [31], introduced the concept of local region-based segmentation. The model focused on image features within a specific range around each point, using a local symmetry function to improve the discrimination between regions of similar intensity, resulting in significant improvements in segmentation accuracy, despite increased computational cost. Building on these developments, Fu et al. [32], developed a multilevel model based on level set models for complex image segmentation, using two level set functions to represent multi-object interactions, while incorporating edge information as an additional stopping force, allowing for accurate border capture even in low-contrast images. Finally, methods have evolved toward more advanced statistical models, Wei et al. [33], using a Gaussian mixture model to segment brain tissue in CT images, with performance quantitatively evaluated using the overlap rate criterion, reflecting a shift from

visual to precise quantitative assessment. Thus, this series of studies demonstrates how each technique, while successful in a specific context, has revealed limitations that have prompted researchers to develop more adaptive and intelligent methods. This confirms that progress in this field is not a technical luxury but rather an imperative to meet the demands of accuracy and reliability in medical diagnosis and treatment [34–37].

I.3.3 Region-Based Image Segmentation

Region-based image segmentation begins by initializing the seeds, subsequently identifying the approximate pixels that exhibit similarity to these seeds. These identified pixels are typically classified as belonging to a specific class, and the seeds are updated iteratively until convergence is achieved.

The primary techniques used in region-based image segmentation include region growing [38], region splitting and merging [39], and super pixel [40]. The objective of region growing is to extract connected regions within an image based on predefined growth criteria, which may include similar grayscale values, edge characteristics, or texture information [41]. The fundamental concept involves initiating the process from one or more seeded pixels and applying the growth rule to iteratively merge adjacent pixels into the growing region until all pixels are incorporated. Although the region growing technique is not frequently utilized in isolation, it serves as a crucial step in image analysis, particularly for delineating small and simple regions, such as tumors and lesions, within an image.

Image segmentation techniques have undergone a fundamental shift from basic models that process images pixel by pixel to more sophisticated models that focus on region analysis. D. Mumford and J. Shah [42], contribution in 1989 represented a turning point. Its primary goal was to segment an image by dividing it into smooth regions. The model’s strategy is based on minimizing a power function that determines the quality of the segmentation, and this method has proven its ability to perfectly model the image as a “cartoon.” The Chan-Vese model, published by Tony F. Chan and Luminita A. Vese [43], responded to a pressing need to segment objects that lack clear edges. Building on the Mumford-Shah approach, this model introduced a new strategy known as “edgeless active contours,” which uses a level set method, enabling it to detect boundaries without relying on gradients.

Later, the need for more efficient and faster methods emerged, and “Graph-Based Segmentation” was developed. It was introduced by Pedro F. Felzenszwalb and Daniel B. Huttenlocher [44], aiming to provide a fast and efficient segmentation method. This model’s strategy is based on representing the image as a graph and is characterized by its ability to preserve detail in low-contrast image regions.

The most recent development is the Contrastive Unpaired Translation for Segmentation (CUTS) framework developed by Pradyumna B. Sharma et al [45], which presented an unsupervised DL model specifically designed to address the lack of labeled data. This framework achieved impressive results, delivering performance comparable to large models trained on large labeled datasets [46–49].

I.4 Methodological developments in computational modeling for medical image segmentation.

ML techniques formed the cornerstone of the development of early approaches for automated medical image segmentation [11]. These methods relied on a two stage model, beginning with manual feature engineering, followed by the application of supervised classification algorithms. In the first stage, researchers focused on extracting specific numerical features from the image, presumably relevant for distinguishing target anatomical structures such as tumors or organs from the surrounding background. These features included basic pixel intensity indices, texture measures derived from the Gray Level Co-occurrence Matrix (GLCM), and geometric properties such as shape, area, and perimeter [50].

After extracting these features, they were used as inputs to feed classical ML algorithms. Among these algorithms, SVMs emerged as one of the most popular models [51], seeking to identify a geometric boundary in a high-dimensional feature space that maximizes the distinction between "target" and "background" classes. Random Forest algorithms have also played a prominent role, constructing an ensemble of decision trees that aggregate their decisions through a voting mechanism to determine the classification of each pixel or region [52]. Alongside supervised models, unsupervised approaches have also emerged, such as the K-means algorithm, which relies on clustering pixels with similar properties, such as intensity or color, into homogeneous groups without the need for pre-labeled training data [53].

Despite the relative success of these methods in limited contexts, particularly in high contrast, low-noise images, they have faced fundamental challenges that have negatively impacted their generalizability and practicality. The most prominent of these challenges is the fragility of models resulting from an over-reliance on manual feature engineering, which requires extensive redesign whenever image acquisition conditions change (such as modifications to imaging protocols or acquisition devices). Furthermore, manually designed features have proven insufficient in capturing the structural and functional complexity of medical images, particularly the non linear contextual relationships that characterize real clinical data. Manual feature extraction was time-consuming and required extensive clinical and technical expertise, limiting the generalizability of these approaches across diverse datasets [54, 55].

These limitations were the primary motivation for moving to a more advanced and effective approach, DL. This approach has revolutionized medical image segmentation by integrating feature extraction and classification into a unified framework, enabling the model to learn discriminative representations directly from raw data without human intervention. Due to its hierarchical architecture, a deep neural network can learn a gradual sequence of features, from primitive patterns such as edges and spots to complex representations reflecting fine anatomical structures. This shift has led to unprecedented improvements in segmentation accuracy, reliability of results, and the models' ability to generalize across diverse clinical contexts [56, 57].

I.4.1 DL for Medical Image Segmentation

Given the rapid advancements in imaging technologies and the accessibility of extensive medical datasets, the demand for precise, efficient, and automated segmentation techniques has reached an all-time high. DL, which is a branch of Artificial Intelligence (AI), has surfaced as a revolutionary method, providing unparalleled abilities in the processing

and interpretation of intricate medical images [58].

Unlike conventional approaches that rely on manually engineered features and rigid algorithms, DL models, particularly CNNs and architectures like Unet, automatically learn hierarchical features directly from raw imaging data (MRI, CT, or ultrasound), eliminating the need for labor-intensive feature design. This adaptability allows DL to generalize across diverse imaging modalities, noise levels, and anatomical variations, whereas traditional methods often fail under such variability.

I.4.1.1 Key DL Architectures for Segmentation

With the advancement of DL techniques, medical image segmentation has witnessed rapid progress in the last decade, with the emergence of specialized neural network architectures that have overcome the limitations of traditional approaches in handling spatial and anatomical complexity. These developments began with the adoption of deep convolutional networks, which efficiently exploited the spatial structure of images, and then evolved into more advanced architectures that integrated attention mechanisms, transformers, and probabilistic models to improve accuracy and reliability [59]. Since the late 2000s, architectures such as Unet and V-net have become cornerstones of the field, paving the way for the development of hybrid models capable of learning from image data in deeper and more adaptive ways. This technological acceleration demonstrates how model architecture is no longer simply an implementation tool but a key driver of innovation in segmentation accuracy, especially in light of ongoing challenges such as the scarcity of high quality data, variability across sources, and unclear boundaries of anatomical structures. Below, we review the most prominent of these architectural architectures that have marked a turning point in the development of this field.

I.4.1.1.1 CNN-based Models In 1991, W. Zhang et al. [60], presented one of the first attempts to use CNNs in medical image segmentation in a paper titled "Image processing of human corneal endothelium based on a learning network." The study analyzed images of the human corneal endothelium taken with a reflectance microscope. The researchers developed a space-invariant neural network capable of generalizing recognition patterns across the entire image. The network was trained on portions of the original images with hand-drawn cell boundaries as reference models. After training, it accurately extracted boundaries from images not used during the learning process, demonstrating its ability to generalize. The team also analyzed the network's internal representations to understand how it processed data and distinguished between shapes and backgrounds. This study marked an important early step in paving the way for the use of CNNs in medical image segmentation. Although the first application of CNN in medical image segmentation appeared in 1991, limited computing power, limited digital medical data, and the absence of effective DL frameworks and training techniques delayed the emergence of powerful models like Unet until 2015, following the significant advances that followed the AlexNet revolution in 2012.

Ciresan et al. [61], presented a deep CNN to the International Symposium on Biomedical Imaging (ISBI) challenge for segmenting neural structures in electron microscope images, aiming to achieve accurate pixel-by-pixel classification. Their strategy relies on a "pixel-wise classification" approach using a multilayer network trained on local windows (patches), relying on computational acceleration via graphics processing units to achieve efficient training and inference. The results demonstrated significant outper-

formance over traditional methods such as simple filtering or second-person observations, achieving the lowest pixel-wise error and Rand error rates. However, this method revealed major drawbacks, namely the slowness resulting from processing each window separately and the computational duplication resulting from overlap between windows, highlighting the need for a more efficient and comprehensive model. Here comes the development of Long et al. [62], in their paper on "Fully convolutional networks for semantic segmentation", which restructured classifier networks such as AlexNet and Visual Geometry Group (VGG) to transform them into models capable of directly producing dense segmentation maps. Their strategy is based on replacing fully connected layers with convolutional layers while introducing a "skip connections" mechanism that combines features from deep layers (with high context) with those from shallow layers (with high spatial resolution), significantly improving segmentation accuracy.

The results demonstrated that Fully Convolutional Networks (FCNs) outperform previous models in natural and medical image segmentation tasks, reducing time and achieving outstanding performance on metrics such as mean IoU. However, FCNs still suffer from limitations in adapting to the volume of 3D medical data and the lack of labeled data. To address these challenges, Ronneberger et al. [63], introduced the Unet model, specifically designed for medical images. It relies on a symmetrical architecture contracting and expansive paths with strong skip connections and can be successfully trained on a very small number of images thanks to highly efficient data augmentation. Unet outperformed sliding-window CNNs in the ISBI challenge, becoming a gold standard in the field. With the growing need to process 3D data such as MRI or CT scans, 3D models have been developed. Milletari et al. [64], introduced the V-net model, which extends the Unet concept to the 3D and uses a Dice-based loss function to address class imbalance, improving segmentation performance across the entire volume. Çiçek et al. [65], presented a 3D Unet model trainable from scratch on sparsely annotated volumes, with support for processing large volumes sequentially seamless tiling, allowing practical application on massive medical images without memory constraints [66–69].

Blendowski et al. [70], focused on learning 3D contextual representations from unlabeled CT data using a self-learning task based on predicting displacements between planar patches along three axes. This task was combined with an encoder-decoder network to produce a heatmap that improved prediction accuracy, forcing the model to understand the spatial structure of the body. This approach demonstrated significant superiority in a few-shot segmentation task involving six major organs.

The strategy evolved the following year with Wang et al. [23], who moved from feature learning to improving the training process itself to address the problem of corrupted or inaccurate labels, a common problem in medical data due to variability between labelers. These researchers proposed a meta-learning framework known as Meta-learning based Confidence Prediction Mask (MCPM), in which a small metamask network was trained to analyze the loss map generated by the segmentation model, identify "suspicious" pixels, and then reduce their weights during training. This was accomplished through an alternating training cycle using a small, clean dataset (meta set) to guide updates. This model demonstrated exceptional performance on the LIDC-IDRI and LiTS datasets, with clear outperformance, especially when label noise was high (up to 80%), highlighting its ability to build more flexible and reliable models.

In 2021, the innovation expanded with Windsor et al. [71], who focused on leveraging multiple imaging modalities of the same patient, such as MRI and dual-energy X-ray absorptiometry (DXA), to enable automatic segmentation transfer. The researchers relied

on a self-learning and contrastive learning framework to train a model to match corresponding regions between modalities without the need for manual position pairings. After achieving accurate alignment, the exact segmentations extracted from the DXA images were used to transfer them to the MRI images, providing accurate training data without the need for manual labeling. This approach has achieved very high accuracy in transferring segments of the pelvic cavity, opening the way for building large, automatically labeled datasets [72–75].

I.4.1.1.2 Transformer-based Models 2021 marked a fundamental shift in the field of medical image segmentation, with new challenges emerging related to understanding global contexts in complex images and overcoming the limitations imposed by CNN-based models, which often suffer from a bias toward local information. This shift began with Chen et al. [76], who introduced the TransUNet model, aiming to combine the representational power of transformers, which can model long-range dependencies, with the fine-grained architecture of a Unet. Their strategy relied on using an encoder based on a transformer model to capture global context from multi-organ CT images, followed by a convolutional decoder that reconstructs the image in fine detail using skip connections. This model demonstrated significant outperformance over CNN-based methods such as V-net and Unet, achieving a higher average dice, demonstrating that incorporating global context improves segmentation accuracy.

As the same year progressed, ideas evolved toward building entirely pure transformation-based architectures, without relying on convolutional operations. This was achieved with Cao et al. [77], who introduced Swin-Unet, the first pure U-shaped architecture for medical segmentation. The goal was to completely eliminate the local bias of CNNs, replacing all convolutional layers with Swin Transformer blocks, which enable progressive attention from local to global. The researchers also developed a patch expanding layer to replace the transposed convolution operations in zooming, preserving the network structure of features. The model demonstrated superior results to TransUNet on both the Synapse and ACDC sets, while reducing unwanted phenomena such as over segmentation, confirming the ability of pure transformations to produce smoother and more accurate predictions.

In the same year, this trend expanded toward improving the interaction of multi-scale information within transformational architectures. Lin et al. [78], introduced DS-TransUNet, which incorporates a cross-attention fusion mechanism to improve feature fusion across scales. The model focused on enhancing the interaction between multi-scale branches in a Unet-like architecture, using mutual attention to focus it on regions of interest across scales. This model achieved exceptional performance on polyp segmentation tasks, outperforming advanced methods, demonstrating that improving the dynamics of feature fusion within transformations can enhance generalization and accuracy.

As 2022 approaches, the need has evolved from optimizing the structure to exploiting multiple sources of information, such as text, to improve the contextual guidance of the model. Here, Li et al. [79], introduced the LViT-T model, which integrates textual information (such as clinical reports) with medical images into a common transformation-based encoder. The strategy relied on integrating textual and visual representations in the same space, allowing the model to use linguistic context to understand and guide the segmentation process, especially in complex or ambiguous cases. The model showed significant improvement in segmentation accuracy and reduced mis-segmentation compared to unimodal methods, opening a new door towards multimodal segmentation [80–83].

I.4.1.1.3 Hybrid CNN-Transformer models The field of medical image segmentation has witnessed rapid development with the transition from CNNs to the integration of transformer models. Li et al. [84], began by introducing a TFCN model that incorporates a Convolutional Linear Attention Block (CLAB) within the FC-DenseNet architecture, aiming to enhance information flow across skip connections using spatial and canal attention. This model achieved significant improvements over Unet and TransUNet, achieving an average dice of 83.72%, but it revealed the need for more comprehensive contextual representations, especially in images with fine details.

Hence, the FAT-Net model introduced by Wu et al. [85], emerged, which adaptively integrates transformations with convolutional features to improve skin lesion segmentation. This model demonstrated improved performance over ensembles such as ISIC but remained limited in efficiently incorporating multi-scale context. Consequently, the need for more sophisticated structures for managing attention across different axes arose, leading to the emergence of MaxViT-UNet, developed by Khan et al. [86], which is based on the MaxViT architecture and uses multi-axis attention to efficiently integrate local and global attention, with linear complexity. This architecture was incorporated into the UNet framework, resulting in its outperforming models such as Swin-UNet and TransUNet in the MoNuSeg and MoNuSAC datasets, achieving improved dice averages and a lower HD95 metric, reflecting its ability to balance accuracy with computational efficiency.

Despite these advances, challenges remain in integrating fine-grained edge features with broad contextual representations, especially in low-contrast medical images. To address this, Manzari et al. [87], introduced the BEFUnet model, which explicitly focuses on integrating "body" features resulting from transformations with "edge" features extracted via specialized convolutional units. BEFUnet utilizes two differential multilevel fusion and cross-axis feature fusion units, as well as the concept of pixel difference convolution to enhance edge sensing. This design outperforms advanced models such as FAT-Net and Swin-UNet, with significant improvements in dice and HD95 metrics, especially in complex tissue segmentation [88–91].

I.5 Comparative Analysis of Medical Image Segmentation Methods

A comprehensive evaluation of medical image segmentation techniques necessitates moving beyond a mere enumeration of strengths and weaknesses. This section provides a systematic, multi-faceted comparison based on quantifiable performance metrics across diverse clinical scenarios. We evaluate not only the accuracy of various models but also their practical viability concerning interpretability, computational efficiency, and integration into clinical workflows. This holistic perspective is essential for understanding the real-world applicability and limitations of these technologies in medical practice.

Medical image segmentation methods are rapidly evolving, with a focus on striking a balance between accuracy and practicality. While current models excel in segmentation accuracy, challenges such as high computing requirements and scarcity of labeled data remain barriers to their clinical application. Integration of different techniques is emerging as a promising solution to address their respective limitations, but it increases model complexity. This progress demonstrates the importance of flexible solutions that take into account the diversity of diagnostic tasks and available infrastructure while maintaining interpretability for clinical decisions. The two tables, I.1 and I.2, provide a comparative

Previous studies

analysis of the most prominent medical image segmentation methods.

Table I.1: Traditional Segmentation Methods: Capabilities and Clinical Constraints

Category	Strengths	Weaknesses	Clinical Applicability
Pixel-Based	- Fast for simple structures (thresholding) - No training data required	- Poor accuracy for complex anatomies - Sensitive to noise/artifacts	Limited to preprocessing tasks (skull stripping)
Contour-Based	- Accurate boundaries (e.g., active contours) - Incorporates shape priors	- Requires manual initialization - Computationally expensive	Cardiac/MRI segmentation (with manual tuning)
Region-Based	Handles texture/heterogeneity (region growing)	- Seed-point dependency - Over-segmentation issues	Tumor segmentation (small ROIs)

Table I.2: DL Architectures for Medical Segmentation: Performance and Trade-offs

Approache	Strengths	Weaknesses	Best For
CNN-Based	- High accuracy with small datasets - Efficient for 3D volumes	- Local receptive fields limit global context - Requires large annotated datasets	Multi-organ CT/MRI segmentation
Transformer-Based	- Captures long-range dependencies - Better generalization across modalities	- High computational cost - Complex hyperparameter tuning	High-resolution MRI/polyp segmentation
Hybrid CNN-Transformer	- Balances local/global features - Improved edge detection	- Increased architectural complexity - Slower inference than CNNs	Low-contrast images (liver tumors)

I.5.1 Evaluation Metrics for Medical Image Segmentation

A robust comparative analysis requires an expanded set of evaluation criteria beyond conventional metrics such as the dice and HD95. The following table I.3 presents a comprehensive framework for assessing segmentation methods:

Table I.3: Common Evaluation Metrics for Medical Image Segmentation

Criterion	Description	Clinical Relevance
Dice	Quantifies spatial overlap between predicted segmentation and ground truth	Primary metric for overall segmentation accuracy
HD95	Measures the 95th percentile of maximum boundary distance between prediction and ground truth	Critical for radiation therapy and surgical planning requiring precise boundaries
Precision (Positive Predictive Value)	Proportion of correctly predicted positive pixels among all positive predictions	Essential for reducing false positives that may lead to unnecessary interventions
Recall (Sensitivity)	Proportion of actual positive pixels correctly identified	Crucial for minimizing false negatives that could result in missed diagnoses
IoU (Intersection over Union)	Measures the ratio of overlap between the predicted segmentation and ground truth to their combined area (union).	widely used to evaluate the accuracy of delineated anatomical structures and lesions.

I.5.2 Task-Oriented Performance Analysis

The optimal segmentation approach varies significantly depending on specific clinical requirements and challenges:

I.5.2.1 Tissue Heterogeneity in Medical Image Segmentation

Clinically, tumor segmentation remains challenging due to high tissue heterogeneity, diffuse boundaries, and substantial morphological variability across cases. Performance analysis reveals that hybrid CNN-Transformer architectures currently offer the best balance, leveraging CNNs' sensitivity to local features while integrating Transformers' ability to model global context, resulting in more accurate and robust segmentations. Pure Transformer models, while powerful in capturing long-range dependencies, tend to over-segment regions with similar textures due to their limited local discrimination. Meanwhile, traditional methods, such as region growing, often fail in such complex scenarios, as they rely heavily on homogeneous intensities and precise seed point initialization. Therefore, hybrid architectures are recommended as the most effective solution for tackling the nuanced demands of tumor segmentation in clinical imaging.

I.5.2.2 Segmentation large Organ and tumors

The segmentation of large organs presents a clinical challenge characterized by volumetric data with generally conserved anatomical structure yet considerable inter-patient shape variation. Performance analysis shows that Unet variants remain highly competitive, offering an optimal blend of accuracy, computational efficiency, and architectural maturity, making them well-suited for routine clinical deployment. While transformer-based models achieve strong segmentation performance by capturing global context, their computational overhead and architectural complexity often exceed practical requirements for this task. Traditional methods, such as active contours with shape priors, can be effective

for organs with stable morphology but are sensitive to parameter tuning and less adaptable to anatomical variability. Therefore, Unet architectures are recommended as the most balanced and pragmatic solution for large organ segmentation across diverse clinical applications.

I.5.2.3 Fine Structure Segmentation

Segmenting small-scale anatomical structures presents a clinical challenge due to their complex morphology and low contrast against surrounding tissues, making precise delineation difficult. Performance analysis indicates that CNN-based architectures enhanced with attention mechanisms excel in this domain, as they effectively highlight fine details and adaptively focus on relevant regions. In contrast, transformer-based models, while powerful for global context, often struggle with minute structures due to their reliance on patch-based tokenization, which can blur or overlook fine spatial details. Traditional contour-based methods are generally unsuitable for such tasks, as they fail to handle the intricate, branching geometries typical of small structures. Therefore, enhanced CNNs incorporating attention gates are recommended as the most effective and reliable approach for accurate fine-structure segmentation in medical imaging.

I.5.3 Synthesis and Strategic Recommendations

The selection of an appropriate segmentation methodology involves balancing three critical dimensions: accuracy, computational efficiency, and practical viability.

For research applications where maximal accuracy is the primary objective without computational constraints, advanced transformer or sophisticated hybrid models represent the optimal choice. In clinical deployment scenarios requiring real-time performance or operation on resource-constrained hardware, optimized Unet variants provide the most practical and robust solution. For specialized applications with limited annotated data, enhanced traditional methods or weakly-supervised learning approaches may offer the most viable pathway.

Future research should focus not only on incremental improvements in accuracy metrics but also on enhancing the efficiency, interpretability, and clinical integration capabilities of these advanced segmentation systems. The ultimate goal remains the translation of technical advancements into tangible improvements in patient care and clinical workflows.

This comprehensive analysis provides a framework for method selection based on specific clinical requirements and operational constraints, supporting informed decision-making in both research and clinical settings.

I.6 Overview of Datasets Used in Medical Image Segmentation

Medical image segmentation relies on diverse datasets to train, validate, and test models. These datasets vary by imaging modality (MRI, CT, ultrasound, etc.), anatomical focus (brain, heart, liver, etc.), and clinical task (tumor detection, organ segmentation, lesion analysis). Below is table I.4, a summary of key datasets and their characteristics:

Previous studies

Table I.4: Summary of publicly available datasets for AI-based medical image segmentation

Category	Dataset Name	Modality	Key Features	Target Application
Brain Imaging	BraTS	MRI (T1, T2, FLAIR, T1ce)	Multi-institutional, glioma segmentation	Brain tumor analysis
	IXI	MRI (T1, T2, PD)	Healthy brain anatomy reference	Neuroimaging research
	LGG	MRI (FLAIR)	Preprocessed low-grade glioma MRI scans with pixel-level tumor annotations for segmentation tasks	Brain tumor segmentation and treatment planning
	ADNI	Longitudinal MRI	Alzheimer's progression tracking	Neurodegenerative studies
Abdominal and Pelvic	LiTS	3D CT	Liver and liver tumor segmentation	Hepatic surgery planning
	KiTS	3D CT	Kidney and kidney tumor segmentation	Nephrology applications
	CHAOS	CT/MRI	Multi-organ (liver, kidney, spleen) segmentation	Abdominal organ analysis
Cardiac Imaging	ACDC	Cine-MRI	Whole heart structures segmentation	Cardiovascular studies
	MM-WHS	CT/MRI	Multi-modality whole heart segmentation	Cardiac structure analysis
Pulmonary Imaging	LIDC-IDRI	CT	Lung nodule annotations (multi-rater)	Lung cancer detection
	LUNA16	CT	Focused nodule detection (LIDC-IDRI subset)	Pulmonary nodule analysis
Histopathology	MoNuSeg	H and E-stained microscopy	Nuclei segmentation	Cancer diagnostics
	GlaS	Histology images	Gland segmentation in colon tissue	Colorectal cancer research
Dermatology	ISIC Archive	Dermoscopy	Largest skin lesion dataset (melanoma focus)	Skin cancer detection
Endoscopy	Kvasir-SEG	Endoscopy	Polyp segmentation in GI tract	Gastrointestinal screening
	ETIS-LaribPolypDB	Endoscopy	Smaller polyp detection dataset	Colonoscopy research
Ultrasound	CAMUS	Echocardiography	Left ventricle segmentation	Cardiac ultrasound analysis

	BUSI	Breast ultra-sound	Benign/malignant tumor classification	Breast cancer detection
--	------	--------------------	---------------------------------------	-------------------------

I.7 Conclusion

This review highlights the evolution of medical image segmentation from traditional methods to advanced DL approaches. While CNNs set new benchmarks, transformer models show superior potential with global context understanding. Key challenges remain in data requirements, interpretability, and clinical implementation. Future works must focus on developing efficient, accurate, and explainable models with real-time capabilities. Standardized evaluation and multimodal integration will be essential to translate these advances into tangible clinical benefits for improved patient outcomes.

CHAPTER II

PRINCIPLES AND APPLICATIONS OF MEDICAL IMAGING MODALITIES

Contents

II.1	Introduction	19
II.2	Types of medical imaging	19
II.2.1	X-Rays	20
II.2.2	Computed Tomography (CT)	21
II.2.3	Medical Resonance Imaging (MRI)	22
II.2.4	Colonoscopy	24
II.2.5	Positron Emission Tomography (PET)	25
II.2.6	Ultrasound Images	26
II.2.7	Dermoscopy	27
II.2.8	Microscopic Images	28
II.2.9	Optical Coherence Tomography (OCT)	29
II.3	Medical Imaging Modalities: Principles, Applications, and Comparative Analysis.	30
II.4	Conclusion	31

II.1 Introduction

Medical imaging has revolutionized modern healthcare by allowing clinicians to visualize the human body's internal structures with remarkable precision. From diagnosing diseases to guiding surgical procedures, these technologies offer priceless advice without invasive techniques. This chapter explores the fundamental types of medical imaging, including X-rays, CT, MRI, ultrasound, and nuclear medicine techniques like PET. Each modality offers unique advantages, leveraging different physical principles to capture anatomical or functional details. Understanding these tools' strengths, limitations, and applications is essential for optimizing patient care and advancing medical practice.

II.2 Types of medical imaging

Medical image analysis plays a crucial role in tracking the clinical progression of diseases and assessing patient responses to specific treatments, thereby aiding in the subsequent planning of patient care. The advancement of medical imaging has significantly accelerated the diagnosis and treatment of various health conditions. Typically, a range of medical imaging modalities exists that utilize diverse techniques to generate images for various objectives. Figure II.1 illustrates examples of various medical images. In contemporary clinical practice, imaging modalities such as X-ray, CT, MRI, and ultrasound are frequently employed. These medical images can either assess multiple organs (MRI and CT) or focus on specific organs (retinal images, mammograms, dermoscopy, and colonoscopy). The amount of data generated by each modality also differs [92]. Each imaging modality presents unique benefits for various applications and diagnostic goals, as well as limitations in depicting the internal structures and functions of the body's numerous organs. X-ray, CT, MRI, ultrasound, dermoscopy, colonoscopy, microscopic images, and optical coherence tomography (OCT) are among the most commonly utilized medical imaging modalities that will be examined in this chapter. Furthermore, we will explore the fundamental principles of operation and their applications.

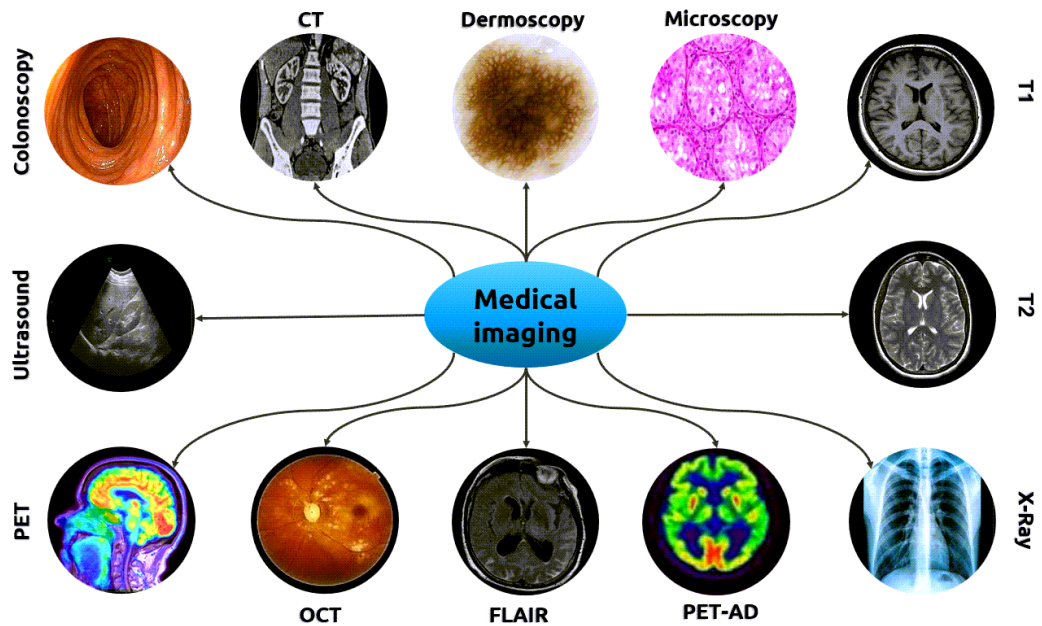


Figure II.1: Illustrative examples of diverse medical imaging modalities in clinical practice

II.2.1 X-Rays

In the realm of medicine, X-ray imaging is utilized to create intricate images of internal anatomical structures, such as bones, which assist in the diagnostic process. X-rays, categorized as electromagnetic waves, can penetrate various solid materials, including the human body. Among the prevalent imaging techniques, X-rays are often employed to visualize internal bodily structures [93]. The process entails directing a concentrated beam of X-rays towards the specific area of interest within the body. As the X-rays pass through the body, they are detected by a sensitive device, such as film or a digital sensor, as shown in Figure II.2. Different types of tissues within the body respond to X-rays in varying ways, softer tissues, such as muscles and organs, absorb fewer X-rays, leading to darker shades on the resulting image, whereas denser tissues, like bones, absorb more X-rays, appearing as white regions on the image. To produce a clear image, it is essential to use an appropriate amount of radiation and ensure that the body part is positioned correctly to the X-ray beam.

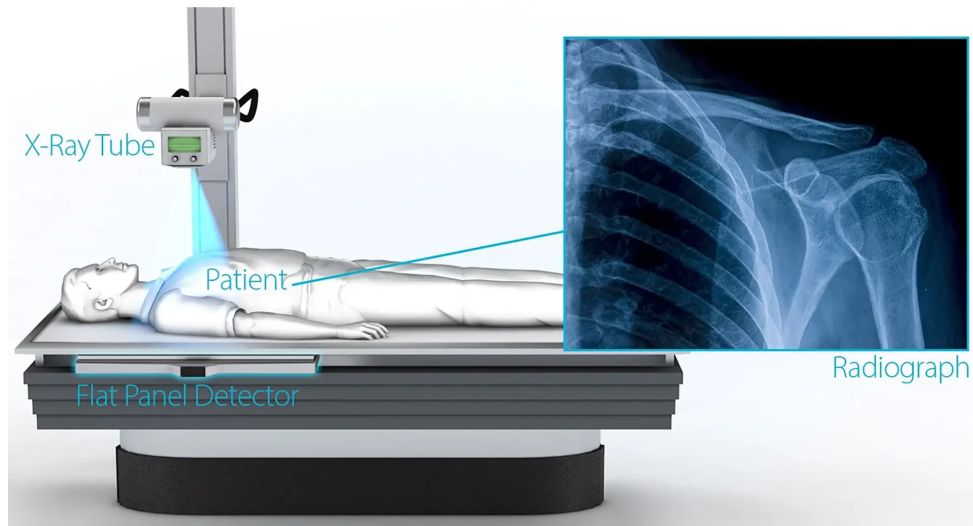


Figure II.2: Projectional radiography [94]

To obtain a more precise representation of the body's 3D structure, multiple images of the same area are captured from various angles and combined to form a 2D image [95, 96]. A radiologist, who is a medical expert specializing in the analysis of medical images, concentrates on interpreting these images to diagnose a range of conditions, including fractures, cancers, and other health issues, by deciphering the visual data presented. An X-ray is a form of electromagnetic radiation generated within a glass tube that has been evacuated. A cathode and an anode establish a voltage difference, which accelerates electrons through the vacuum of the tube towards the rotating tungsten anode. Upon striking the anode, these electrons produce both X-rays and heat. X-rays are generated solely when a voltage difference is applied between the cathode and anode, making the X-ray tube inactive until it is energized by the medical radiation technologist. Various imaging modalities, such as X-ray, CT, fluoroscopy, and angiography, utilize this physical on/off mechanism. X-rays are frequently used to obtain images of bones, primarily to evaluate for fractures or other irregularities [97].

Furthermore, dentists and orthodontists employ X-rays to acquire comprehensive images of teeth. In addition, X-rays play a vital role in identifying tumors or irregularities in bones. They are also utilized to assist surgeons during operations. X-ray images of

the breast are captured during mammography to detect and evaluate irregularities such as tumors or microcalcifications [98]. This technique is essential for the early detection and prevention of breast cancer, as it can identify potential cancer indicators before they are palpable during a physical examination. X-rays are employed not only for imaging bones but also for the chest and lungs to diagnose various conditions, including pneumonia, lung cancer, and emphysema. Beyond identifying gallstones, kidney stones, and intestinal blockages, X-ray imaging can assess abdominal organs like the liver, spleen, and kidneys. Due to its widespread availability, rapid execution, and cost-effectiveness, X-ray imaging is a commonly preferred diagnostic technique. It is a noninvasive procedure, eliminating the necessity for incisions or injections, thereby providing a relatively safe and comfortable experience for patients. However, one limitation of X-ray imaging is its low tissue contrast, which makes it more challenging to differentiate between malignant and healthy tissues compared to other imaging modalities. Additionally, X-ray imaging only offers a 2D representation of the body's 3D structure, which may hinder the visualization of certain internal components, particularly in hard-to-reach areas. The absence of depth in the 2D image can obscure some features of internal structures, complicating the accuracy of diagnoses [99, 100].

II.2.2 Computed Tomography (CT)

The computerized tomography CT scan, creates detailed images of the internal structures of the body by integrating X-ray technology with computer processing [101]. In contrast to conventional X-rays, CT scans yield cross-sectional images, providing a unique viewpoint of the body. These scans are noninvasive, guaranteeing a painless experience for the patient. By directing X-rays at the body from multiple angles and assessing the intensity of the X-rays as they traverse the body using detectors (This is what Figure II.3 shows), CT scans can deliver high-resolution images of internal anatomy [102].

CT scans utilize a multi-slice detector, which is a specialized type of X-ray detector capable of capturing numerous images at once from various angles, thereby providing detailed cross-sectional views of the body. To enhance image quality, two additional techniques are employed in CT scans: spatial filtering, which minimizes noise and enhances contrast, and multi-energy imaging, which uses X-rays of different energies to capture distinct data and further improve contrast [103].

The images produced by CT scans can be displayed in multiple formats, such as cross-sectional slices, 3D images, and even virtual reality representations, all thanks to computer processing. Radiologic technologists perform CT scans, while radiologists, who are medical professionals trained in interpreting images, analyze the results. A doughnut-shaped machine moves the patient's table into its center during a CT scan. The internal X-ray tube of the scanner rotates to emit a series of X-ray beams at various angles as it encircles the patient. As the X-ray light passes through the body, it creates a pattern of attenuation, being absorbed at different rates by various tissues. An array of detectors located opposite the X-ray tube captures the pattern of attenuated X-rays and converts them into electrical impulses. This data from the electrical signals is then sent to a computer, where sophisticated algorithms are employed to reconstruct the information into a 3D image [104].

CT scans are utilized to improve the assessment of soft tissues and complex regions of the image that may not be distinctly visible through conventional X-rays. They are frequently used to visualize and evaluate blood vessels, internal organs, bones, the brain,

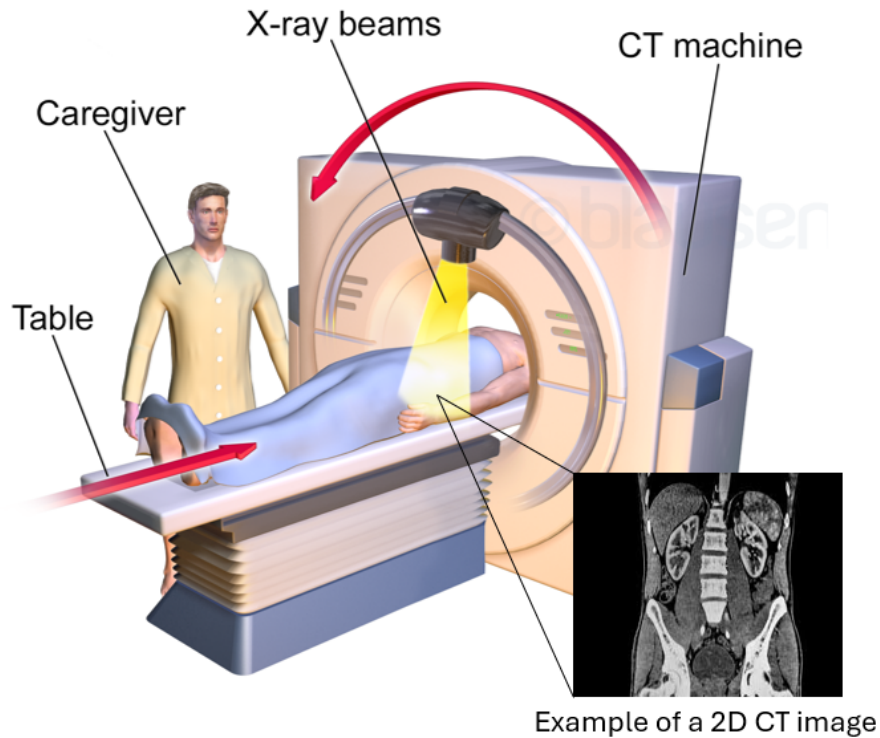


Figure II.3: Computerized Axial Tomography Scan [105]

neck, spine, and chest. CT scans assist healthcare professionals in various diagnostic activities, including the identification of tumors, assessment of fractures, and monitoring the effects of cancer treatment on patients [106]. The high resolution of CT imaging is a significant advantage, as it enables physicians to observe anatomical details such as bone structure, organ functionality, and blood vessel configuration.

This capability makes it advantageous for diagnosing a wide range of diseases. Besides its role as a diagnostic instrument, CT imaging technology is widely accessible in environments such as hospitals, clinics, and physicians' offices. Furthermore, CT imaging is rapid and efficient, with data typically becoming available in a brief period. A primary concern regarding CT imaging is the exposure to ionizing radiation, which can elevate the risk of cancer and other health complications, particularly with repeated scans or in pregnant women. Additionally, CT scanners represent a substantial financial investment, which may be prohibitive for some healthcare facilities. Consequently, certain patients may find it challenging to access this type of imaging. The benefits and drawbacks of CT scans must be carefully evaluated, and alternative imaging techniques should be considered whenever feasible [107].

II.2.3 Medical Resonance Imaging (MRI)

MRI is akin to CT scans, yet it provides enhanced image quality, yielding intricate cross-sectional views of bodily structures. Like CT scans, MRIs are noninvasive and safe, as the magnetic fields and radio waves employed do not have harmful effects on patients. MRI operates by utilizing a strong magnetic field to align the hydrogen nuclei within the body, which facilitates the creation of accurate images of internal structures. This process generates a subtle magnetic moment that can be harnessed to deliver precise depictions of the body's internal anatomy [108].

Initially, a weak radio signal is emitted from the reoriented hydrogen nuclei due to the application of radiofrequency pulses. A detector captures these signals, and following analysis by a computer, clear images of the body's internal mechanisms are produced (Check out Figure II.4). When patients enter the MRI machine, the powerful magnetic field causes the hydrogen atoms in their bodies to realign. Subsequently, a weak radio signal is generated by the aligned hydrogen atoms after radio waves are transmitted through the body. An antenna detects the radio waves, and the data is sent to a computer for further processing. Sophisticated algorithms process this data to construct a 3D model of the body's internal anatomy. Hydrogen atoms, which exist in every cell of the body, are present in varying concentrations depending on the type of tissue. An MRI can produce axial, sagittal, and coronal images, providing a comprehensive 3D representation of the patient's anatomy [109].

MRA (magnetic resonance angiography)

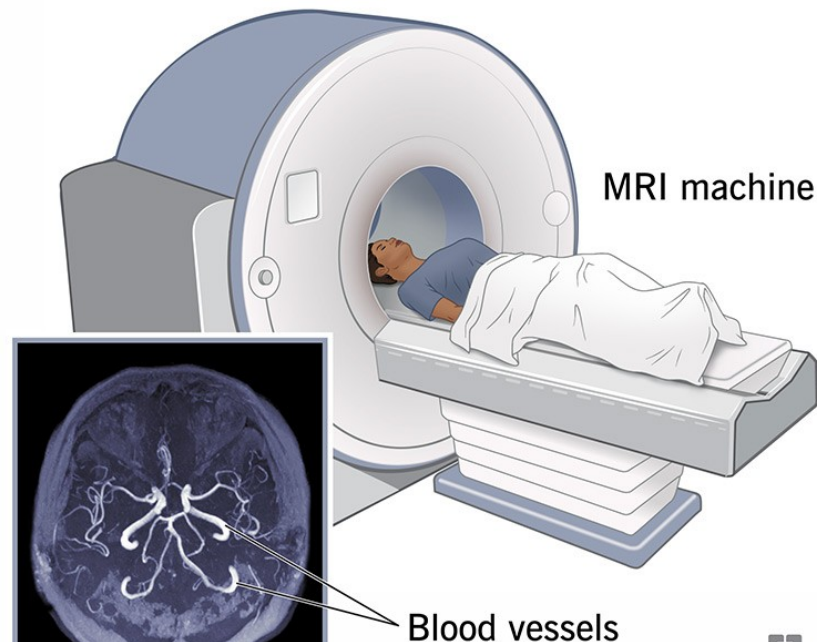


Figure II.4: MRI Example: 2D Magnetic Resonance Angiography (MRA) of Brain Blood Vessels [110]

MRI is utilized to examine the internal structure of the brain, spinal cord, bones, heart, blood vessels, and various other internal organs. MRI is preferred over other imaging techniques such as X-rays and CT scans because it does not expose patients to harmful ionizing radiation [111]. These devices can also capture soft tissues, including muscles, tendons, and ligaments, with remarkable clarity. They are capable of producing images from multiple planes, offering a more detailed perspective of internal anatomy. Furthermore, one of the significant advantages of MRI is its capacity to provide microscopic resolution without inflicting any harm or discomfort to the patient. However, it is important to note that MRI scanners are sensitive to metal, therefore, individuals with certain implants, such as pacemakers, cannot undergo MRI scans. Additionally, the high cost of these scanners can be a substantial financial burden, rendering them inaccessible for some facilities. The duration of examinations may also be longer in comparison to X-rays or CT scans [112].

II.2.4 Colonoscopy

Virtual colonoscopy, commonly known as CT colonoscopy, enables healthcare providers to assess the large intestine for colon or rectal cancer, as well as for growths termed polyps, as evidenced by a study conducted by Nguyen and Lee [113]. The fundamental concept of colonoscopy involves the insertion of a colonoscope into the rectum and along the colon. This colonoscope is a long, flexible tube that is fitted with a light source and a camera at its end. It facilitates real-time visualization of the colon lining, allowing the physician to identify any irregularities such as polyps, ulcers, inflammation, or tumors. Throughout the procedure, the physician may also carry out therapeutic actions, including the removal of polyps or the collection of tissue biopsies, if deemed necessary. Figure II.5 illustrates an example of a colonoscopy.

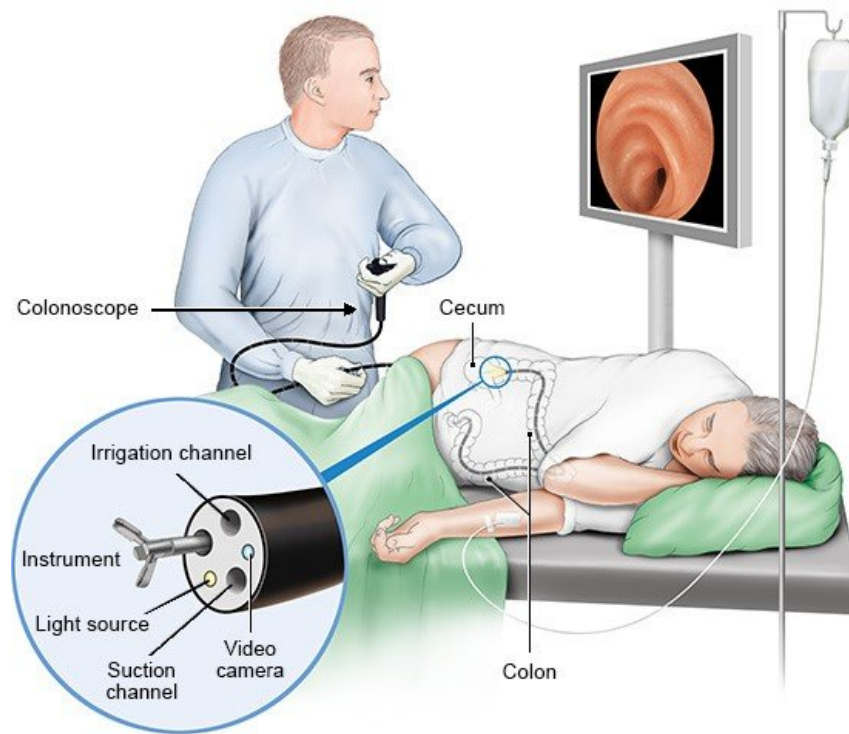


Figure II.5: Diagram of the Colonoscopy [114]

Colonoscopy serves as a potent diagnostic instrument for the detection of colorectal cancer, inflammatory bowel diseases (such as ulcerative colitis and Crohn's disease), and various other conditions that impact the colon and rectum. This procedure enables the identification and excision of polyps, which are abnormal growths that have the potential to progress into cancer over time. Colonoscopy boasts a high level of diagnostic precision for colorectal cancer and other gastrointestinal disorders, thereby assisting in the formulation of suitable treatment strategies [115]. It is important to note that a colonoscopy is an invasive procedure that necessitates the insertion of a flexible tube into the rectum and colon, which may result in discomfort or pain for the patient. Therefore, it is crucial to engage in discussions regarding the risks, benefits, and alternatives of colonoscopy with a healthcare professional to ensure an informed decision is made regarding the procedure, taking into account individual circumstances and medical history.

II.2.5 Positron Emission Tomography (PET)

A PET scan generates a 3D representation of the internal structures of the body [116]. Physicians utilize PET scans to observe and measure the metabolic and functional processes occurring within the body. This imaging technique is commonly used in the medical domain for the identification, categorization, and monitoring of a variety of conditions, such as cancer and neurological disorders. The operation of a PET scan involves the detection of radiotracers that emit positrons. These radiotracers are compounds that are introduced into the body and release positrons, which are particles with a positive charge. When a positron encounters an electron in the body, they annihilate each other, resulting in the production of two gamma rays that travel in opposite directions. The PET scanner detects these gamma rays and utilizes them to create 3D images that illustrate the distribution of the radiotracer within the body. PET scans are extensively utilized in oncology to monitor tumor development, assess treatment effectiveness, and identify metastases. They serve multiple functions, particularly in producing high-resolution images of the brain and predicting the progression of cancer.

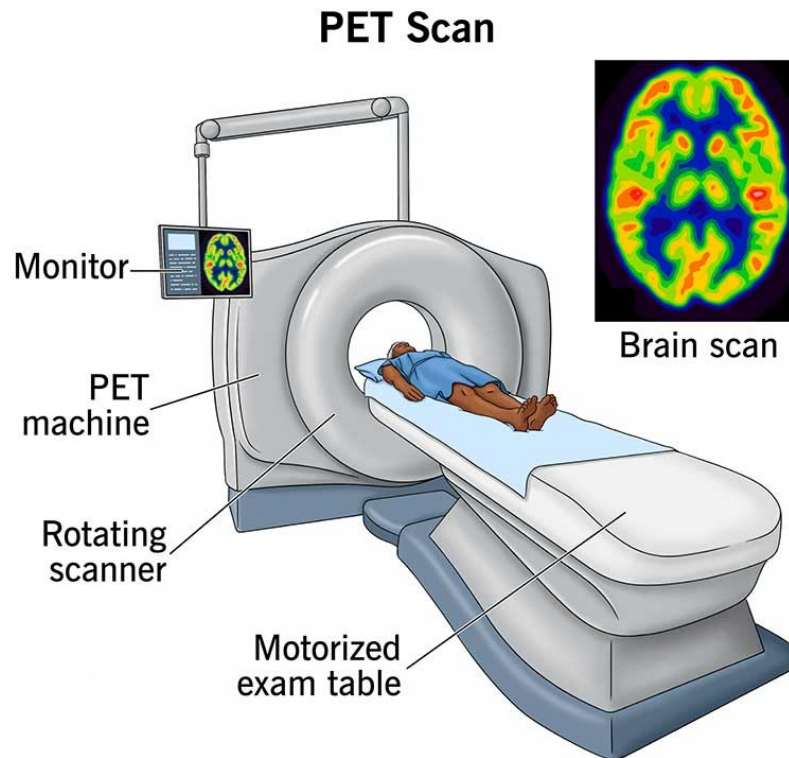


Figure II.6: Positron Emission Tomography scans [117]

PET scans are frequently used for patients who have already been diagnosed with cancer, as they can accurately depict the extent of cancer dissemination and evaluate the success of chemotherapy. Furthermore, PET scans assist in presurgical planning for operations involving the brain or heart [118]. Additionally, conditions such as Alzheimer's disease, epilepsy, and Parkinson's disease can be diagnosed through PET scans, as they offer valuable insights into alterations in brain function. In Figure II.6, examples of PET scans for normal conditions, mild cognitive impairment, and Alzheimer's disease are illustrated, respectively. PET scans are capable of identifying functional and metabolic

changes in tissues prior to the emergence of structural abnormalities that can be detected by other imaging modalities. They provide a comprehensive evaluation of the entire body, facilitating the identification of metastatic disease or the assessment of multiple organ systems at once. In contrast to other imaging techniques, PET scans deliver information regarding the metabolic activity of tissues, which is crucial for accurate diagnosis and treatment planning.

II.2.6 Ultrasound Images

Ultrasound, commonly referred to as a sonogram, employs high-frequency sound waves to visualize internal bodily structures. This medical imaging modality involves a transducer that emits high-frequency sound waves and captures the returning echoes from internal structures to generate images (Figure II.7). Some of the sound waves transmitted into the body are reflected back to the transducer as they traverse a boundary separating different types of tissue. The transducer collects these echoes, converts them into electrical signals, and sends them to a computer for analysis, resulting in high-resolution images of the internal structures of the body. Depth information is acquired by measuring the time delay between the emitted ultrasound wave and the reflected wave from each layer of the sample, facilitating the creation of 2D or 3D images [119].

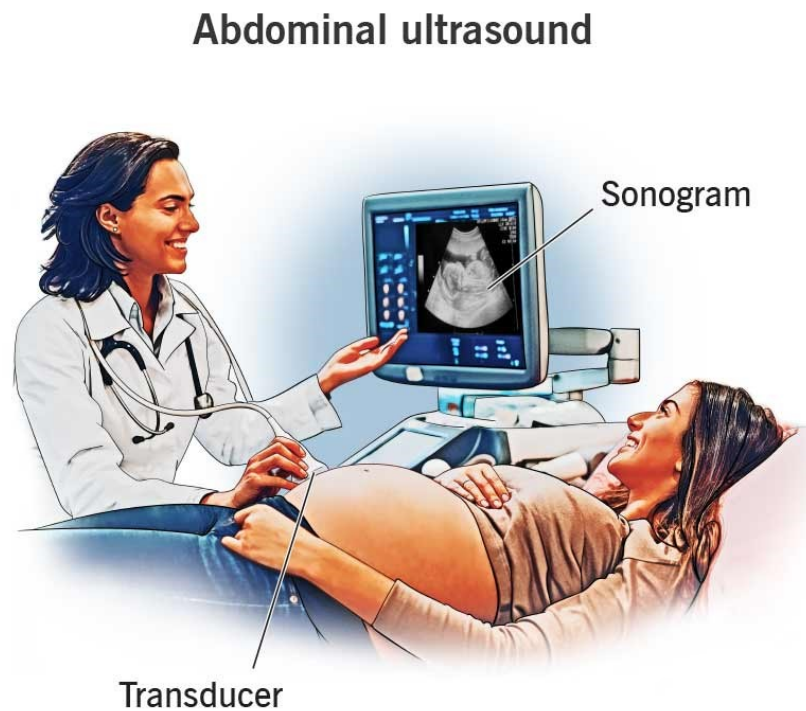


Figure II.7: Example of Abdominal Ultrasound Scan [120]

Ultrasound is often employed to track the development of fetuses, offering real-time imaging capabilities. While it is generally not suitable for visualizing bones or air-filled tissues such as the lungs, it is widely utilized in the assessment of abdominal, vascular,

and thyroid conditions. In the context of abdominal imaging, US is frequently used to investigate the organs within the digestive system, including the gallbladder, pancreas, spleen, and kidneys. As a medical imaging modality, Ultrasound presents several advantages. The portability of the equipment allows for imaging procedures to be conducted in various locations, not limited to hospital settings. Additionally, the use of Ultrasound provides the advantages of rapid imaging and the absence of radiation exposure [121].

Since it does not employ potentially harmful ionizing radiation, it is often preferred over alternative imaging techniques such as X-rays and CT scans. The ability to generate real-time images also facilitates the dynamic assessment of internal structures and their functions. However, Ultrasound imaging is not without its limitations. A significant drawback is its inability to visualize structures that are beneath the surface or obscured by opaque materials (such as bone or gas). Furthermore, artifacts caused by bone or air echoes can distort images, complicating the identification of internal structures. The quality of Ultrasound images and the accuracy of diagnoses may also be influenced by the skill level of the Ultrasound technician [122].

II.2.7 Dermoscopy

Dermoscopy is a non-invasive imaging method employed to assess skin lesions. It is also referred to as dermatoscopy or epiluminescence microscopy [123]. This technique involves examining the skin using a dermatoscope, a portable device equipped with magnifying lenses and a light source. Dermoscopy enables dermatologists to view skin structures and patterns that are not discernible to the naked eye, thereby aiding in the diagnosis and treatment of various skin disorders, particularly melanoma and other types of skin cancer (as shown in Figure II.8). The underlying principle of dermoscopy is the application of polarized or non-polarized light to illuminate the skin and enhance the visibility of both surface and subsurface features.

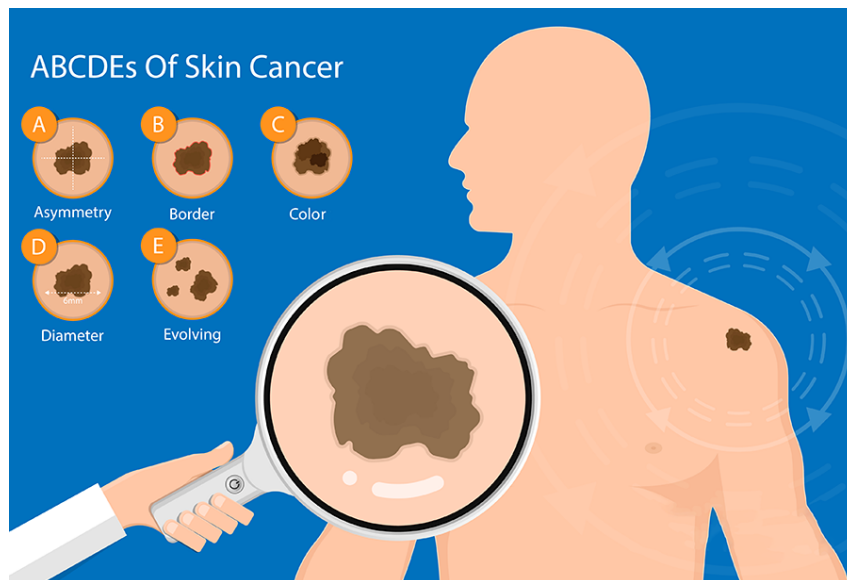


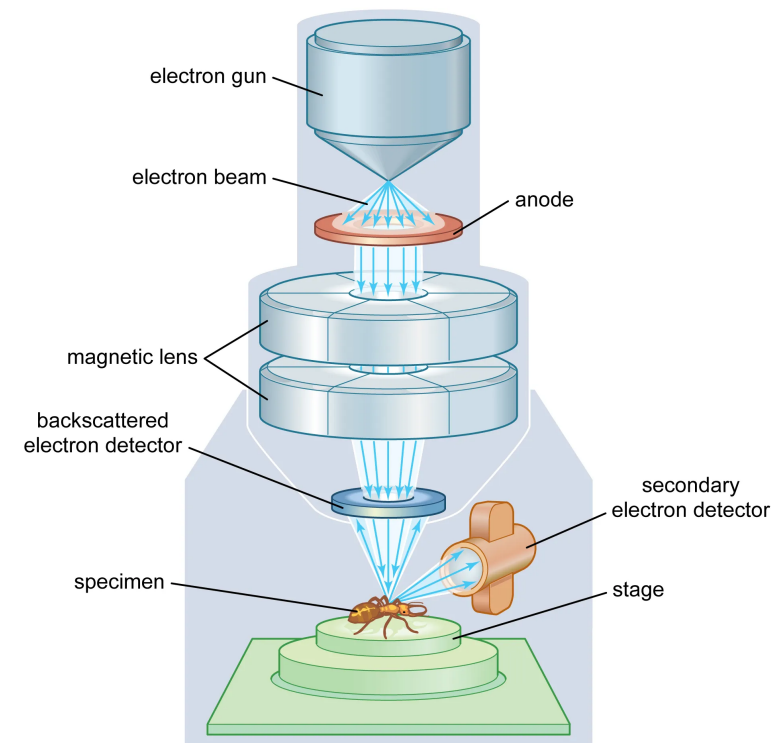
Figure II.8: Diagnosis for skin cancer with dermoscopy [124]

The dermatoscope effectively removes the surface reflection of the skin, providing a clearer perspective of the deeper layers. Additionally, dermoscopy employs liquid immersion or contact dermoscopy, where a gel or oil is applied to the skin to eliminate the air gap

between the dermatoscope and the skin, further enhancing image quality. Dermoscopy is vital for the early detection of melanoma, the most lethal form of skin cancer. It assists dermatologists in evaluating the morphological features of pigmented skin lesions and in distinguishing between benign and malignant lesions [125]. Furthermore, it aids in differentiating various types of skin lesions and in formulating suitable treatment strategies. Dermoscopy can also be utilized to track changes in skin lesions over time, especially in patients with a history of skin cancer or atypical moles. It helps in recognizing any signs of progression or regression, which is crucial for making informed decisions regarding further management. Dermoscopy enhances the diagnostic precision of skin lesions, particularly in differentiating benign from malignant lesions. This noninvasive procedure does not necessitate any incisions or tissue sampling. It can be conducted swiftly and conveniently in an outpatient environment. Moreover, dermoscopy can diminish the necessity for unnecessary biopsies of benign lesions, resulting in cost savings [126].

II.2.8 Microscopic Images

Microscopic imaging pertains to the utilization of microscopes for the visualization and examination of objects at a microscopic scale (Figure II.9). This process involves magnifying objects that are imperceptible to the unaided eye, thereby facilitating a thorough analysis of their structure and characteristics. Both science and medicine depend significantly on microscopic imaging for various applications, including the investigation, diagnosis, and assessment of minute materials, organisms, and other specimens [127].



© Encyclopædia Britannica, Inc.

Figure II.9: Scanning Electron Microscope (SEM) Components [128]

The underlying principle of microscopic imaging is founded on the interaction of light or electrons with the sample under observation. Various types of microscopes are

employed for imaging purposes, such as optical and electron microscopes. Tissue samples are frequently acquired through biopsy for analysis, with small portions of the tissue subsequently stained and colored to highlight cellular features. Counterstains are utilized to improve the color, visibility, and contrast of the resulting images. These images are often used in cancer detection, with attributes like cell size, shape, and distribution being closely examined [129].

Microscopic imaging plays a vital role in medical diagnostics and pathology, enabling the analysis of biopsy samples, the detection of cellular irregularities, and the diagnosis of various diseases and conditions. Techniques in microscopic imaging, such as optical microscopy, permit nondestructive examination of samples. This characteristic allows for the observation of samples without modifying their structure or composition, thus facilitating further analysis or experimentation. Advanced microscopic imaging instruments, such as electron microscopes, can be costly to purchase and maintain. Each microscopic imaging method has its own specific limitations regarding resolution, depth of field, and contrast. It is essential to select the appropriate imaging technique for a particular application to achieve accurate and dependable results [130].

II.2.9 Optical Coherence Tomography (OCT)

OCT is a noninvasive imaging technique that employs light waves to generate intricate cross-sectional images of biological tissues. The underlying principle of OCT is founded on the interference of light waves. This method utilizes low-coherence interferometry, wherein a near-infrared light beam produced by a light source is split into two distinct paths: a reference arm and a sample arm. The reference arm contains a mirror, while the sample arm directs the light towards the tissue being examined. The light that is reflected back from both arms is recombined, leading to interference between the reference and sample beams. The resulting interference pattern is detected and analyzed to reconstruct a depth-resolved image of the tissue [131, 132]. By measuring the time delay of the reflected light, OCT can determine the depth of various tissue layers or structures within the sample, which is subsequently used to create a detailed cross-sectional image of the tissue.

OCT is widely utilized in the diagnosis and treatment of various ocular conditions, including macular degeneration, diabetic retinopathy, glaucoma, and retinal detachment. It offers detailed images of the retinal layers, facilitating early detection and monitoring of eye diseases. Figure II.10 illustrates an example of OCT machine components and eye anatomy overview. In the field of cardiology, OCT is employed to visualize and evaluate coronary arteries [133]. It aids in diagnosing and characterizing coronary artery disease, identifying plaque accumulation, and guiding procedures such as stent placement.

OCT delivers detailed, high-resolution images of tissue structures, allowing for the visualization of minute details and abnormalities. It provides real-time imaging, enabling immediate evaluation and assessment. The penetration depth of OCT is restricted to a few millimeters, contingent upon the type of tissue being imaged. This limitation constrains its application in imaging deeper tissues or structures. Advanced OCT systems can be costly, which may limit their availability in certain healthcare environments. Nevertheless, the technology has gradually become more accessible over time [135].

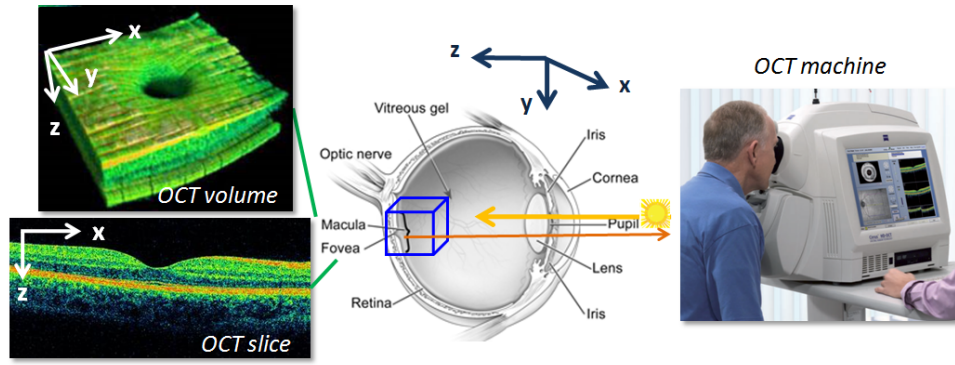


Figure II.10: OCT Machine Components and Eye Anatomy Overview [134]

II.3 Medical Imaging Modalities: Principles, Applications, and Comparative Analysis.

Medical imaging is a cornerstone of modern diagnostics and treatment planning, offering a non-invasive window into the human body's structure and function. However, no single imaging technique is universally ideal; each modality has distinct physical principles, clinical applications, strengths, and limitations. This table II.1 provides a systematic comparison of the primary medical imaging technologies.

Table II.1: Comparison of Diagnostic Imaging Technologies

Feature	Depth Penetration	Primary Use	Data Output	Advantages	Limitations	Contraindications	Scan Time
X-ray	Deep (bones)	Bones, chest	2D projection	Fast, cheap	Poor soft-tissue	Pregnancy (relative)	Seconds
CT	Deep (whole body)	Trauma, tumors	3D cross-sections	Detailed anatomy	High radiation	Pregnancy	<5 mins
MRI	Deep (soft tissue)	Brain, joints, MSK	3D multiplanar	No radiation	Long scan time	Metallic implants	15–60 mins
Ultrasound	Superficial–moderate	Pregnancy, heart	Real-time 2D/3D	Real-time, safe	Operator-dependent	None	Minutes
PET	Moderate (whole body)	BCancer, metabolism	3D metabolic maps	Metabolic activity	High cost	Pregnancy	30–90 mins
Colonoscopy	Surface–lumen (GI)	GI polyps, cancer	Real-time video/2D	Direct visualization	Invasive, prep needed	Bowel perforation risk	20–60 mins
Dermoscopy	Surface (skin)	Skin lesions, melanoma	2D magnified images	Non-invasive, portable	Limited to surface	None	<5 mins
Microscopic Imaging	Surface–thin sections	Biopsies, pathology	2D/3D high-res images	Cellular-level detail	In vitro only	None	Minutes–hours
OCT	Shallow (1–3 mm)	Retina, coronary, skin	3D cross-sections	Non-contact, high speed	Limited depth	None	Seconds–minutes

II.4 Conclusion

In conclusion, this chapter emphasizes the pivotal role medical image analysis plays in enhancing diagnostic accuracy and monitoring patients' clinical progress. We have reviewed in detail basic imaging techniques such as X-ray, CT, MRI, and ultrasound, explaining the distinct characteristics and clinical applications of each. Despite significant technical advances, the field still faces challenges related to the complexity of data interpretation, the massive volume of information generated by these devices, and some technical limitations in imaging fine structures. At the same time, these challenges represent opportunities for research and development, particularly in the area of integrating artificial intelligence techniques to improve analysis, develop more specialized devices, and innovate effective solutions for medical data management.

CHAPTER

III

HYBRID DL ARCHITECTURES FOR MEDICAL IMAGE SEGMENTATION: INTEGRATING CNNs AND TRANSFORMERS

Contents

III.1	Introduction	33
III.2	DL in Computer Vision and Healthcare	33
III.3	Neural Networks	34
III.4	Convolutional Neural Network (CNN)	35
III.4.1	Principal layers	36
III.4.1.1	Convolutional layer	36
III.4.1.2	Pooling Layer	37
III.4.1.3	Fully-Connected Layer (FCL)	37
III.4.2	Supportive Components	38
III.4.2.1	Activation Functions	38
III.4.2.2	Batch Normalization	39
III.4.2.3	Dropout	40
III.5	Applications of CNN in Medical Image Analysis	41
III.5.1	Segmentation	41
III.5.1.1	Types of Image Segmentation	43
III.5.1.1.1	Semantic segmentation:	43
III.5.1.1.2	Instance segmentation:	43
III.5.1.1.3	Panoptic segmentation:	43
III.5.1.2	Segmentation in medical image:	43
III.5.1.2.1	Unet architecture:	44
III.6	Transformer for medical image analysis	45
III.6.1	ViTs and Their Adaptation to Medical Imaging	46
III.6.2	Transformer-based Models in Medical Image Segmentation	47
III.6.2.1	Hybrid Architectures Combining CNNs and Trans- formers	47
III.7	Conclusion	49

III.1 Introduction

In this chapter, we highlight the crucial function of DL in the development of medical image segmentation solutions. We focus on contemporary methodologies that contribute to enhancing the precision of identifying anatomical and pathological regions in numerous imaging datasets. We examine key developments in, with a particular focus on hybrid architectures that combine convolutional networks and transformer models. We explore the potential of these architectures to achieve a paradigm shift in diagnostic accuracy and to improve clinical outcomes.

III.2 DL in Computer Vision and Healthcare

The DL has become the mainstay of computer vision, revolutionizing how machines perceive and interpret visual data [136, 137]. Computer vision is the field of AI that enables computers to interpret and understand digital images or videos similarly to humans [138]. Using DL, particularly through the use of CNNs, computer vision systems have achieved remarkable performance in tasks such as image classification, object detection, segmentation, and more [139, 140].

The DL, as a branch of ML, has led to many breakthroughs in the fields of computer vision. This type of learning relies mainly on neural networks that extract features from the input data and correlate them with the output [141, 142]. A neural network consists of information processing units known as neurons, which are connected to each other to form the network. If several layers of these neurons are arranged, it is called DL the schematic diagram represented in Figure III.1 shows the structure of a neural network. Recent success in this field has been made possible by the availability of increasing computing power capable of handling the complex calculations needed to train DNNs [143].

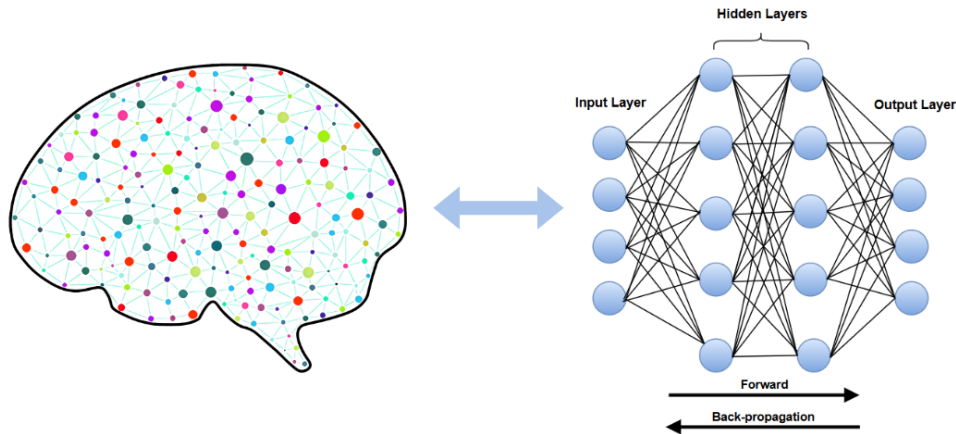


Figure III.1: Schematic representation Neural Network Architecture

Medical image processing and healthcare systems are vital areas receiving increasing priority for AI applications. This interest is underscored by the remarkable growth of the healthcare sector, which continues to attract significant investment and technological advancements. This sector also represents a fertile environment for advanced technological applications, which effectively contribute to improving the efficiency of diagnosis and treatment by automating complex and repetitive operational processes. DL algorithms are highly flexible across various use cases, as each algorithm can perform a specific function

while maintaining sufficient flexibility to suit various applications. In this context, we will review some important DL models used in medical image processing [140].

III.3 Neural Networks

Neural networks represent a category of learning algorithms that underpin most DL methodologies [144]. Neural networks consist solely of a series of arithmetic operations that transform an input (x_i) into an output (y_i). The weights (W_i) and linear operators multiply the preceding values of the interconnected input and output to the neuron. The function (f) within the hidden layer can take on any form. In a basic scenario, featuring a single neuron in the hidden layer, the input value (X_1) is multiplied by the initial weight (W_1), and the resulting value $X_1 \times W_1$ is subsequently directed to the function (f) of the neuron. The outcome is determined by multiplying the result of that function by (W_2). During the training of a neural network, it receives inputs ($X_1, X_2, \dots, X_n$), and the result is computed through mathematical processes. The output value (Y_i) is then compared to the known actual value of y ; the weights (W_i) are adjusted slightly to bring the output value closer to the known true value of y . A simple example of this is depicted in figure III.2 [145].

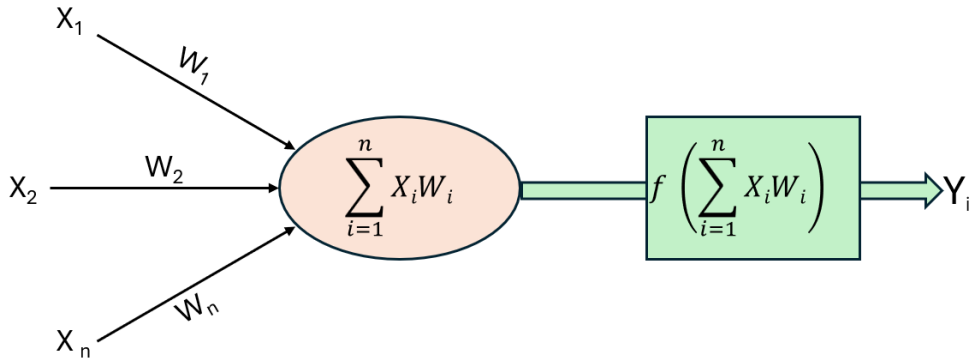


Figure III.2: An artificial neuron, or node, is the cornerstone of DNNs

Each input (x_i) is associated with a specific weight w_i within an artificial neuron or node. The cumulative sum of all weighted inputs (x_i, w_i) is subsequently processed through a nonlinear activation function (f), which converts the neuron's pre-activation level into an output (y_j). The resulting output (y_j) is then utilized as an input for the node in the subsequent layer as illustrated in Figure III.3. There exists a variety of activation functions, each offering a distinct mapping from a preactivation level to an output value. The rectifier function is the most commonly activated function (neurons utilizing it are known as rectified linear units (ReLU)), followed by the hyperbolic tangent function, the sigmoid function, and the softmax function. the softmax function is often employed in the output layer to determine the probabilities of multiclass labels.

A feedforward multilayer neural network, which is also referred to as a multilayer perceptron, consists of two classes. Each node in one layer is connected to every neuron in the following layer, forming a fully connected network. The weighted sum of the inputs for each neuron j in the first hidden layer is processed by a nonlinear function. The resulting output (y_j) from this transformation serves as the input for the second hidden layer. Subsequently, the information is conveyed to the output layer through the network [146].

Hybrid DL Architectures for Medical Image Segmentation: Integrating CNNs and Transformers

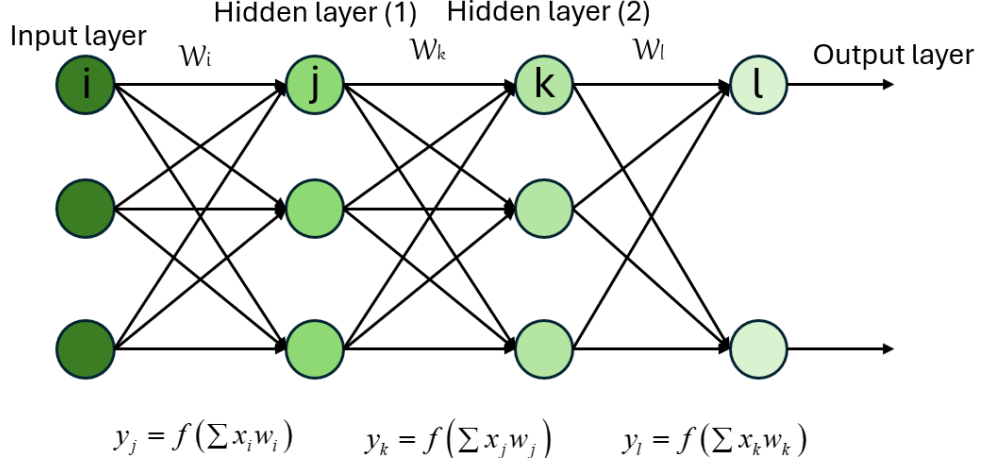


Figure III.3: Artificial feedforward multilayer neural network

III.4 Convolutional Neural Network (CNN)

In traditional multi-layer neural networks, inputs are consistently represented in vector format. However, in the case of (medical) images, the structural or configurational information present among adjacent pixels or voxels serves as an additional source of data. Consequently, the process of vectorization inevitably compromises such structural and configurational details within images. A CNN, which typically incorporates convolutional layers interspersed with pooling (or sub-sampling) layers, followed by FCL akin to a standard multi-layer neural network, is specifically designed to more effectively leverage this spatial and configurational information by accepting 2D or 3D images as input. In contrast to conventional multi-layer neural networks, a CNN utilizes extensive weight-sharing to minimize the degrees of freedom within the models. Furthermore, a pooling layer contributes to reducing computation time while progressively establishing spatial and configurational invariance [147].

CNNs are prevalent learning models characterized by three main features: parameter sharing, sparse interactions, and equivalent representations. A standard CNN is structured with three layers: the input, the hidden, and the output layers. The hidden layer comprises the convolutional layer, pooling layer, and FCNs. The architecture of the CNN is illustrated in figure III.4. Filters utilized in the convolutional layer extract high-level features from the input image. The purpose of the convolution layer is to learn the weights of the convolutional filters or kernels that execute the convolution operation on images. The CNN model transforms the image into a sequence through the convolution operation.

$$C_j^l = f \left(\sum_{i \in M_j} C_i^{l-1} \times W_{ij}^l + b_j^l \right) \quad (\text{III.1})$$

The equation shows how output feature map C_j^l at layer l is computed. Each output comes from input feature maps C_i^{l-1} of the previous layer, combined through convolution kernels W_{ij}^l . The set M_j defines which inputs contribute to output j . A bias term b_j^l is added, and the result passes through a nonlinear activation function $f(\cdot)$, such as ReLU.

Hybrid DL Architectures for Medical Image Segmentation: Integrating CNNs and Transformers

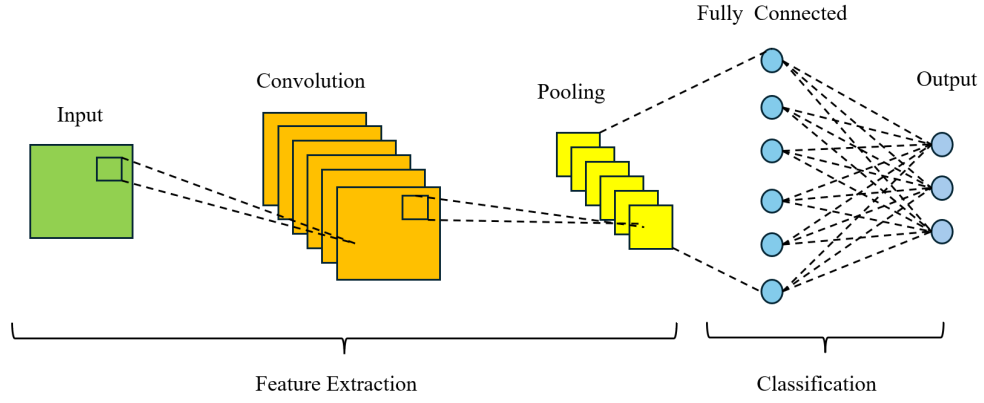


Figure III.4: The basic architecture of CNN

This operation entails sliding a small window across the image. At each position, the elements within the window are multiplied by the corresponding elements of the image, and the results are summed to produce a scalar value. The outcome of the convolution operation is another 2D grid, referred to as the activation map, which is also commonly known as the feature map [148].

III.4.1 Principal layers

III.4.1.1 Convolutional layer

This layer is responsible for determining the output of neurons that are linked to local regions of the input by calculating the scalar product of their weights with the corresponding region of the input volume. As the fundamental component of a CNN, this layer is where most of the processing occurs. It necessitates input data, a filter, and a feature map, among other elements. The layer consists of a set of learnable filters, with the kernel size of each filter typically being 3×3 or $3 \times 3 \times 3$, depending on whether the input is a 2D or 3D image, respectively [149]. Figure III.5 shows an example of a 3×3 convolution operation that has a stride length of 1 and no padding. In this example, the input data is combined with a 3×3 convolution filter to create output data, which is also known as a feature map.

5	2	7	1	4	0
3	9	4	6	2	5
1	6	8	3	5	1
7	2	5	9	0	3
4	8	1	2	7	6
0	5	3	4	1	2

*

2	-1	0
0	1	3
-2	4	1

=

33	28	20	17
43	29	42	12
28	35	18	25
19	15	32	23

Figure III.5: Example of a 3×3 convolution operation with a stride length of 1 and no padding

Hybrid DL Architectures for Medical Image Segmentation: Integrating CNNs and Transformers

III.4.1.2 Pooling Layer

The pooling layer serves as a fundamental component in CNNs. It performs a down-sampling operation that primarily contributes to the reduction of the training data's dimensionality and significantly decreases the number of parameters in the input. Furthermore, the pooling layer is linked to the reduction of both the height and width of the feature maps within CNNs. It operates on each feature map to produce a new feature map that displays enhanced features [150]. There are three different pooling strategies in CNN architectures, Shown in Figure III.6:

Max pooling: The pooling layer extracts the most significant features from the preceding feature maps, resulting in a new feature map. The max pooling layer selects the maximum element from the designated filter region.

Average pooling: This technique calculates the mean value of all elements within the filter region. Compared to max pooling, average pooling produces features that are generally less prominent. Furthermore, global pooling is an additional layer found in some deep networks. The pooling layer does not learn parameters and therefore does not involve weight updates during training, which makes it easy to integrate within convolutional neural network architectures.

Min pooling: layer extracts the least prominent features from the input feature maps by selecting the minimum value within each filter region. This operation suppresses high-intensity activations while preserving low-intensity signals, making it particularly useful for highlighting subtle patterns or suppressing noise in the data.

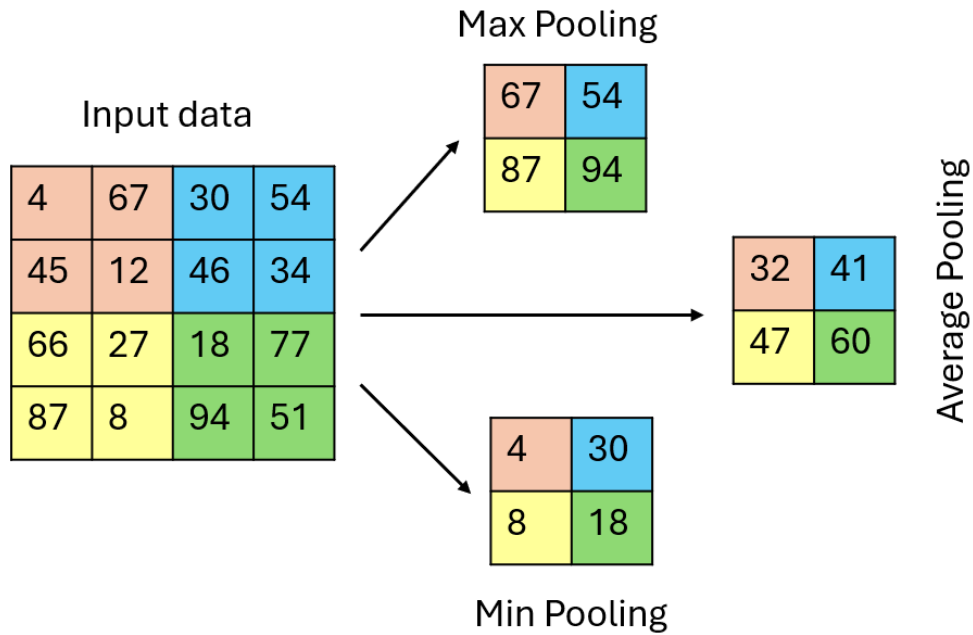


Figure III.6: The Most Popular Pooling Strategies in CNN Networks

III.4.1.3 Fully-Connected Layer (FCL)

The FCL consists of neurons where each neuron is connected to all activations from the previous layer. It combines the extracted features into a high-level representation that is mainly used for classification tasks. In classification networks, the softmax function is commonly used in the output layer to convert raw outputs into probability values between

Hybrid DL Architectures for Medical Image Segmentation: Integrating CNNs and Transformers

0 and 1. Other activation functions such as sigmoid, tanh, and ReLU are typically used in different parts of neural networks depending on the architecture and the role of each layer. In modern CNN architectures, Fully Connected Layers are often reduced or completely replaced by alternative techniques such as global average pooling or 1×1 convolutional layers, especially in more recent models and in segmentation networks, in order to preserve spatial information and reduce the number of parameters.

III.4.2 Supportive Components

Supporting components in CNNs are essential for their efficiency and stability. Although convolutional and pooling layers form the backbone of feature extraction, the overall effectiveness of the network depends on specialized supporting components working together to ensure training stability, reduce overfitting, and improve the model's generalization ability [151].

III.4.2.1 Activation Functions

The activation function is crucial in the training of neural networks because it introduces the essential nonlinearity that allows the model to learn intricate representations. Although various activation functions, such as Sigmoid and Hyperbolic Tangent, are theoretically equivalent in that they can approximate any arbitrary continuous functions [152], they often exhibit different behaviors in practical applications. In fact, during the initial stages of neural network development.

Sigmoid was the predominant activation function employed in neural networks. However, over time, it became widely recognized that its limited derivative could result in the vanishing gradient problem. In contrast, ReLU proved to be much more effective, as it maintains a derivative of one for all positive inputs. As a result, it has emerged as the most commonly used activation function in neural networks [153]. The success of ReLU has prompted extensive research aimed at discovering an activation function that performs as well as or better than ReLU, whether through human efforts [154–157] or automated methods [158]. Unfortunately, none have yet achieved the same level of popularity as ReLU, primarily due to its inherent simplicity. There are seven basic activation functions.

The ReLU function outputs the input directly if it is positive and zero otherwise. ReLU is computationally efficient, avoids vanishing gradients for positive inputs, and has become the default choice for hidden layers in DL. The most important activation functions are summarized in the following table III.1.

Hybrid DL Architectures for Medical Image Segmentation: Integrating CNNs and Transformers

Table III.1: Summary of activation function

Name	Equation	Additional information
Rectified linear unit (ReLU)	$f(x) = \max(0, x)$	The ReLU function outputs the input directly if it is positive and zero otherwise. ReLU is computationally efficient, avoids vanishing gradients for positive inputs, and has become the default choice for hidden layers in DL [159]
LeakyReLU (LReLU)	$f(x) = \max(\alpha x, x)$	The Leaky ReLU modifies ReLU by allowing a small negative slope (αx , where α is typically 0.01) for negative inputs instead of zero. This prevents the dying ReLU issue while maintaining computational efficiency. The function is defined piecewise: $\text{LeakyReLU}(x) = x$ if $x > 0$, else αx [160]
Parametric ReLU (PReLU)	$f(x) = \max(0, x) + \alpha \times \min(0, x)$	PReLU generalizes Leaky ReLU by making the negative slope α a learnable parameter during training. This allows the network to adaptively determine the best slope for negative inputs, improving performance in some cases [161].
Exponential Linear Unit (ELU)	$f(x) = \max(0, x) + \min(0, \alpha(e^x - 1))$	The ELU function smooths the transition for negative inputs, helping avoid dead neurons while maintaining strong gradient signals. It often performs better than ReLU in deeper networks but is slightly more computationally expensive [155]
Scaled Exponential Linear Unit (SELU)	$\text{SELU}(x) = \begin{cases} \lambda x & \text{if } x > 0 \\ \lambda \alpha (e^x - 1) & \text{if } x \leq 0 \end{cases}$	SELU is a self-normalizing variant of ELU that automatically scales its outputs to maintain stable gradients in deep networks. It uses fixed constants $\lambda \approx 1.0507$ and $\alpha \approx 1.6733$ to ensure normalization, making it useful in self-normalizing neural networks (SNNs) without batch normalization [162]
Sigmoid	$f(x) = \frac{1}{1 + e^{-x}}$	A sigmoid is a smooth, S-shaped curve that maps any real-valued input to an output between 0 and 1. It introduces non-linearity into neural networks while ensuring outputs are interpretable as probabilities, making it particularly useful for binary classification tasks in the output layer [163]
Swish	$f(x) = \frac{x}{1 + e^{-x}}$	The Swish activation function is a smooth, non-linear activation function that has gained popularity as a powerful alternative to traditional functions like ReLU. Introduced by Google researchers in 2017, Swish is defined as the product of the input and its sigmoid activation is the logistic sigmoid function [158]

III.4.2.2 Batch Normalization

Training DNNs presents challenges due to the changing distribution of inputs at each layer throughout the training process, as the parameters of preceding layers are adjusted. This variability can hinder the training efficiency, necessitating lower learning rates and meticulous parameter initialization, which complicates the training of models that utilize saturating nonlinearities. This issue is referred to as internal covariate shift, and it can be mitigated by normalizing the inputs of each layer [164].

Hybrid DL Architectures for Medical Image Segmentation: Integrating CNNs and Transformers

BN is a method employed in DL to enhance the stability of neural network training by normalizing layer inputs during the training phase. The technique begins by calculating the mean and variance of the inputs over the current mini-batch [165]. Each input is then normalized by subtracting the batch mean and dividing by the square root of the batch variance, with a small constant added to avoid division by zero. Following normalization, the values are modified using two learnable parameters, γ and β , which allow the layer to adjust the output distribution as necessary, Figure III.7 illustrates these steps in detail. This procedure guarantees that each layer receives inputs with a uniform distribution, thereby diminishing internal covariate shift and facilitating quicker and more stable training. During the inference phase, BN relies on running averages of the mean and variance rather than batch statistics to maintain consistent performance across varying input batches [166].

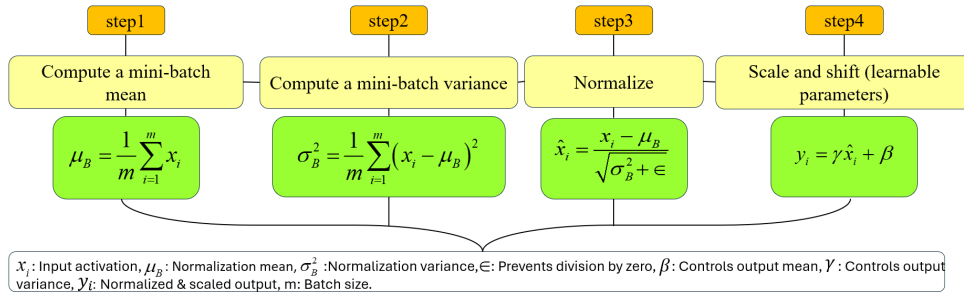


Figure III.7: Step-by-Step Batch Normalization Process.

III.4.2.3 Dropout

Dropout serves as a regularization method employed during the training phase to mitigate overfitting. The fundamental concept involves training various models and utilizing the average of their predictions to enhance the generalization capabilities of a network. This process entails the random removal of neurons at a specified dropout rate, pd , throughout the training phase, allowing the network's weights to be adjusted based on the varying connectivity of the neurons within the network. It is important to note that the dropout rate is a numerical value ranging from 0 to 1. During the training phase, the weights of each neuron are scaled by the probability of retention (pr), $pr = 1 - pd$, to estimate the average predictions from the different trained models. This approach has the potential to enhance the learning outcomes of neural networks [167]. An illustration of dropout application within a neural network can be found in Figure III.8.

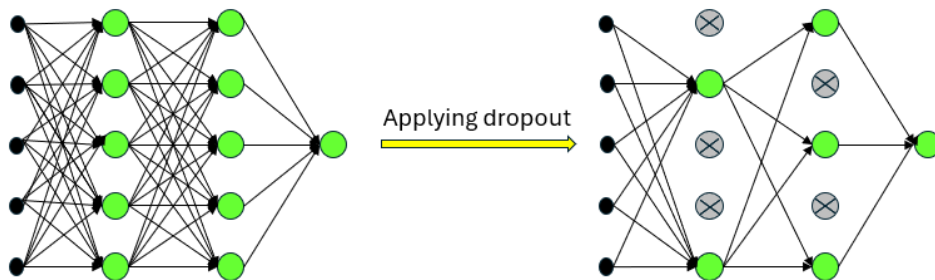


Figure III.8: Example of apply dropout in a neural network.

III.5 Applications of CNN in Medical Image Analysis

CNNs have broad and diverse applications in medical image analysis, demonstrating their high efficiency in diagnosing diseases and interpreting radiological images with accuracy that surpasses humans in some cases. These networks rely on the principle of hierarchical learning, which enables them to extract features ranging from simple edges to complex patterns in medical images. In the field of diagnostic radiology, CNNs excel at analyzing various types of medical images. They are used in analyzing X-ray images to detect pneumonia and tuberculosis, and they also excel at analyzing CT and MRI scans to detect cancerous tumors and determine their margins with millimeter-level accuracy [168].

In the field of pathology, these networks demonstrate high efficiency in analyzing microscopic slides, where they can be used for automated cancer cell counting by detecting or segmenting individual cells in microscopic images. They also play a vital role in analyzing ultrasound images to detect fetal abnormalities and heart disease. CNNs provide an integrated solution that includes precise image segmentation using algorithms such as Unet to identify tumor margins, precise classification using ResNet to distinguish between healthy and diseased tissue, and early detection of abnormalities using Faster Region-Based Convolutional Neural Network (R-CNN), making them an indispensable tool for accurate medical diagnosis [169, 170].

III.5.1 Segmentation

Segmentation simply means breaking a image into connected parts (or smaller sections) where the individual elements (like pixels or voxels) have similar features. This grouping of voxels, or the division of volumes, serves as a preliminary step in recognizing these areas as distinct objects, which can then be classified or labeled as anatomical structures with associated physiological attributes. Segmentation provides the framework through which we organize raw medical imaging data; this framework is utilized when we subsequently visualize the relevant anatomy or pathology, as well as when planning interventions and treatments for the medical condition at hand. Just as there are numerous methods for interpreting an image (including statistical analysis, differential geometry of intensity fields, or partial differential equations, among others), there exists a variety of approaches to the segmentation problem, each stemming from different perspectives [171].

Segmentation is a fundamental analytical technique that takes various forms and applications depending on the domain of knowledge. In natural language processing and text analysis, it manifests itself through mechanisms such as tokenization and thematic text segmentation [172]. In networking and information technology, it plays a prominent role through the concepts of network subnetting, security segmentation, or market segmentation in business contexts [173]. In data science and artificial intelligence, segmentation methods rely heavily on clustering algorithms to discover patterns, while in database management, it is used to improve performance and querying through table partitioning [174].

In the field of image processing and computer vision, various techniques are used to segment an image into regions with similar properties. Traditional approaches include threshold segmentation, such as Otsu's method, which relies on selecting an optimal threshold value to separate important parts of an image [175]. Region-based techniques are also used, such as the "region growing" model, which starts from seed points and then expands based on the similarity between pixels [176]. Additionally, edge-based methods,

Hybrid DL Architectures for Medical Image Segmentation: Integrating CNNs and Transformers

such as Canny or Sobel filters, are used to detect boundaries between regions [177]. In recent times, DL has become more prevalent, with neural networks such as Unet and Mask R-CNN being used to segment images with high accuracy [178].

Image segmentation is the process of dividing a digital image into interconnected and homogeneous parts or regions based on a specific characteristic, such as color, shape, or composition [179]. The main goal of this process is to separate and distinguish objects and elements in the image for easier processing or analysis later. After segmentation, the system or specialist can clearly identify structures, objects, or regions of interest within the image, supporting several fields such as medical diagnosis, autonomous driving, and satellite image analysis.

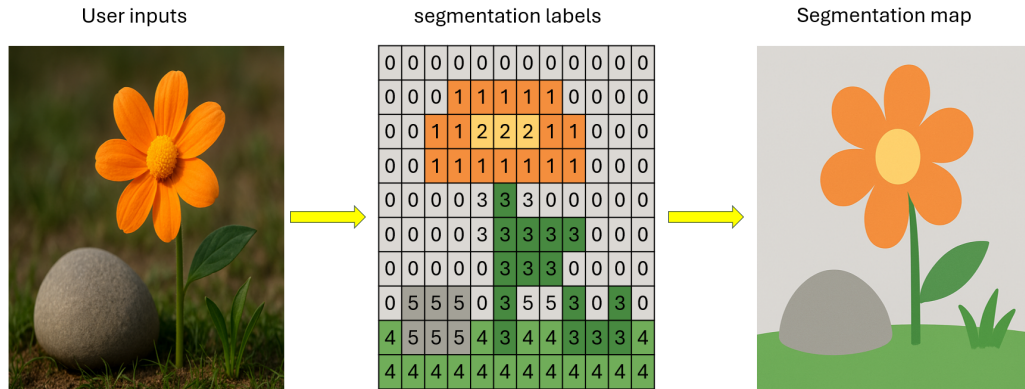


Figure III.9: Conceptual Illustration of Image Segmentation.

The computer processes the image by converting it into numbers or pixels, then applies algorithms to process and analyze the data contained within it, such as segmenting the image into parts and recognizing the shapes or elements within it (Figure III.9 shows an example of this). This makes the image clearer and extracts information from it, such as identifying the locations of tumors and organs in medical images (check out Figure III.10).

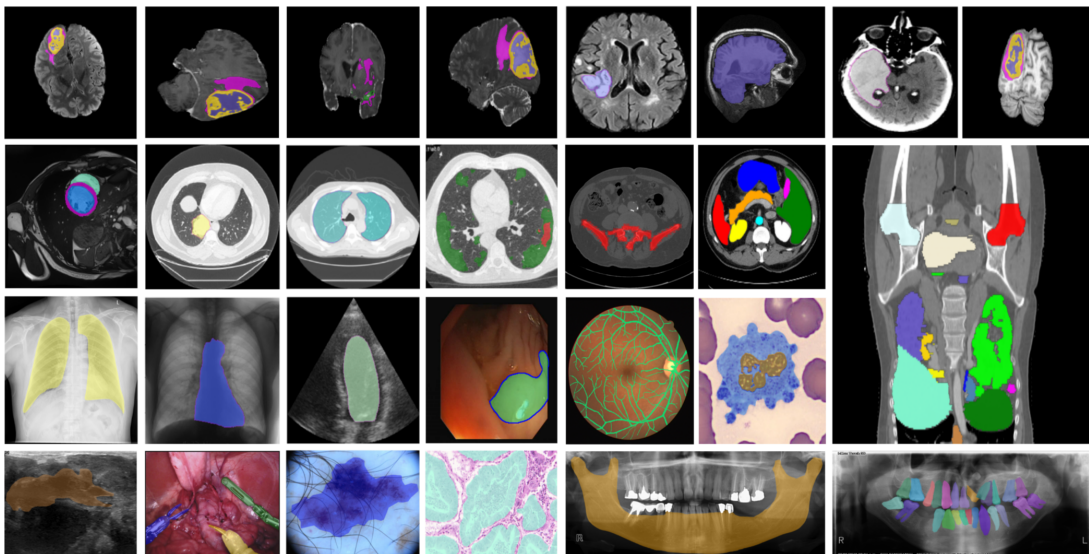


Figure III.10: Example of segmentation in various medical images [180].

Hybrid DL Architectures for Medical Image Segmentation: Integrating CNNs and Transformers

III.5.1.1 Types of Image Segmentation

III.5.1.1.1 Semantic segmentation: Semantic segmentation represents a vital function within the realm of computer vision, entailing the classification of each pixel in an image into designated categories, Figure 11 illustrates this. In contrast to more straightforward tasks such as image classification, which assigns a singular label to an entire image, or object detection, which delineates bounding boxes around various objects, semantic segmentation offers a comprehensive, pixel-level interpretation of the scene [181]. This meticulous analysis is essential for applications necessitating accurate localization and classification of objects, including autonomous driving, medical imaging, and augmented reality.

III.5.1.1.2 Instance segmentation: Instance segmentation merges the capabilities of object detection with semantic segmentation [182]. While object detection identifies and localizes using bounding boxes, and semantic segmentation classifies each pixel, instance segmentation goes further by distinguishing between individual object instances within the same class. This means that in a scene containing multiple objects of the same type, such as several cars in a parking lot, instance segmentation not only recognizes all pixels belonging to cars but also differentiates each car as a separate entity, Figure 11 illustrates this. This fine-grained understanding is crucial for applications requiring precise object separation and identification.

III.5.1.1.3 Panoptic segmentation: Panoptic segmentation represents the next evolution in computer vision by combining the strengths of both semantic and instance segmentation into a single, unified framework [183]. The term panoptic, derived from Greek roots meaning all-seeing, perfectly captures this technique's ambitious goal: to provide machines with a comprehensive understanding of visual scenes that matches human perception. Unlike traditional segmentation methods that treat objects and backgrounds separately, panoptic segmentation simultaneously addresses two critical aspects of scene interpretation classifying every pixel in an image while also distinguishing between individual object instances, Figure III.11 illustrates this. This dual approach creates a complete picture where each pixel receives either a "stuff" label (for amorphous, uncountable regions like sky, road, or grass) or a "things" label (for distinct, countable objects like cars, pedestrians, or traffic signs) [63].

III.5.1.2 Segmentation in medical image:

Automatic segmentation of medical images using CNNs has revolutionized the field of diagnostic imaging, enabling the identification of anatomical and pathological regions with high accuracy and without manual intervention. Specialized architectures such as Unet, V-net, and DeepLab have contributed to this progress through their superior ability to capture hierarchical features in medical imaging data while maintaining spatial accuracy through encoder and decoder structures supported by skip connections [185, 186].

A major challenge facing the development process was the scarcity of labeled medical data, which was overcome by adopting multiple strategies such as data augmentation, transfer learning, and specialized loss functions (such as Dice Loss) to address class imbalances [187]. This challenge has driven researchers to develop more advanced models to enhance automation and accuracy, such as models based on 3D convolutions to capture

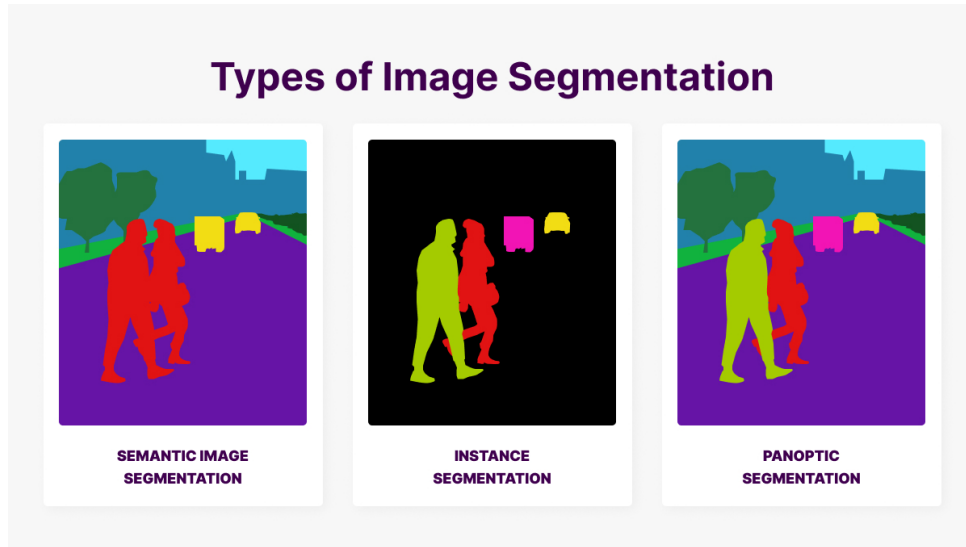


Figure III.11: Examples of types of Image Segmentation [184]

the full volumetric context of medical data, as well as the integration of attention mechanisms and transformer-CNN hybrids to more effectively model long-range relationships within images [188].

These models have diverse applications in vital areas such as tumor and organ segmentation and vascular structure analysis, directly contributing to improved diagnostics, treatment planning, and surgical guidance. For example, it has become possible to perform liver segmentation in CT images or detect brain lesions in MRI images automatically and with high accuracy, enhancing workflow efficiency in clinical settings [189].

III.5.1.2.1 Unet architecture: The Unet architecture initially presented by Ronneberger et al. [63], has emerged as the leading fully CNN for segmentation in medical imaging. Most segmentation challenges are addressed using the Unet or its various adaptations; nevertheless, the fundamental architecture is preserved across these modifications.

This architecture bears resemblance to encoder-decoder frameworks [190, 191]. It features a contracting pathway where multiple convolutional layers are applied sequentially, accompanied by pooling operations that diminish the dimensions of the feature maps while increasing the number of feature maps. This pathway enables the network to identify intricate patterns across different scales. Upon reaching the 'bottleneck' (the final convolutions of the contracting pathway), the feature maps transition through the up-sampling pathway, which employs up-convolutions in succession to restore the dimensions lost during the contracting phase. Each up-sampled feature stack is also concatenated with the output from the contracting pathway at the corresponding level, referred to as skip connections. This process allows the network to recover spatial information that may have been lost during the contracting phase. Following concatenation, additional convolutions are performed, followed again by up-convolutions, repeating this cycle until the original input dimensions are achieved.

The original Unet was characterized by the use of convolutions without any padding applied to the borders of the convolution inputs. Consequently, in the concatenation phase, the input from the contracting pathway had to be cropped to align with the spatial dimensions of the up-sampled feature maps. This limitation also resulted in the final segmentation not matching the original size of the image, representing only a centered

Hybrid DL Architectures for Medical Image Segmentation: Integrating CNNs and Transformers

region of the image. To maintain the same dimensions, it is advisable to apply padding before each convolution. Figure III.12 illustrates a schematic representation of the general architecture of the Unet.

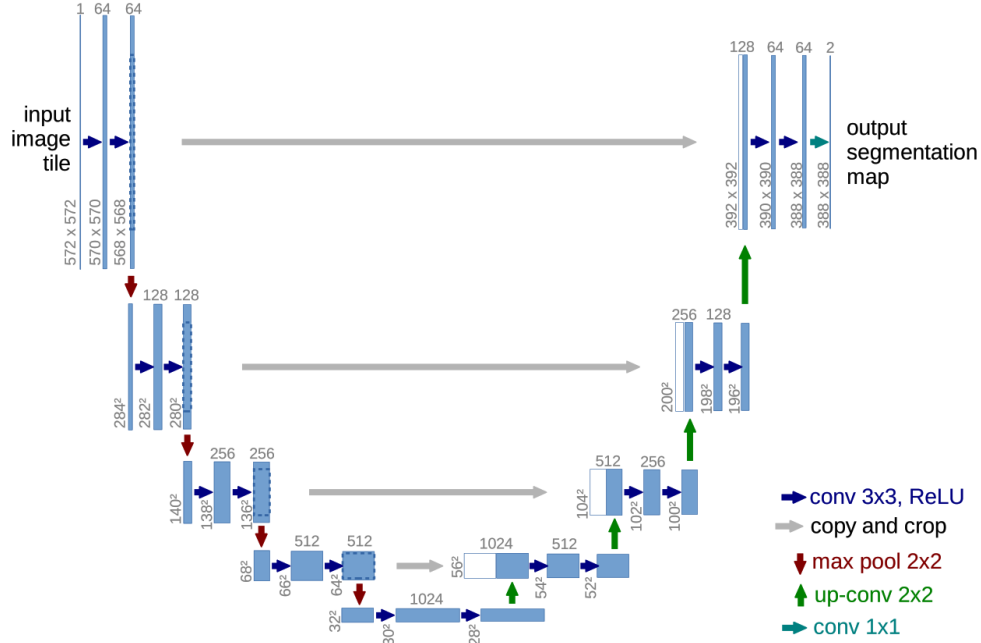


Figure III.12: The Unet architecture is illustrated here with an example for a resolution of 64x64 pixels [63].

The fundamental architecture of Unet can be readily enhanced through the incorporation of additional operations, and a majority of the Unet architectures that have been developed exhibit some level of modification. Initially, the Unet was designed for processing 2D images, utilizing 2D convolutions of size 3×3 , 2D max-pooling operations, and 2D up-convolutions. The activation function employed was the ReLU [192], which ranks among the most frequently utilized activation functions in neural networks, while the pooling operation consists of max-pooling with a 2×2 kernel, resulting in a halving of the size at each downsampling stage. Any of these components can be replaced with other analogous operations, or additional operations may be integrated to address the specific problem at hand; for instance, it is common practice to incorporate a batch normalization operation following each convolution [193]. Ultimately, the final layer responsible for producing the segmentation typically employs a 1×1 convolution with a sigmoid activation function (in scenarios where only a single element is to be segmented) or a softmax activation function (in cases involving multi-class segmentation tasks).

III.6 Transformer for medical image analysis

To effectively address the capture of long-term dependencies, attention-based transformer models have emerged as a compelling alternative, demonstrating the capability to completely supplant traditional convolutions in DNNs by processing sequences of image patches, thus leading to the development of Vision Transformers (ViTs) [194]. Additionally, recent studies suggest that the prediction errors generated by ViTs align more closely with human judgments compared to those produced by CNNs [195–198], which has gener-

Hybrid DL Architectures for Medical Image Segmentation: Integrating CNNs and Transformers

ated considerable interest within the medical imaging field to implement these models for medical image analysis tasks.

The remarkable performance of Transformer models in natural language processing has captured the attention of the vision community, prompting investigations into their potential applications in computer vision challenges. One of the key advantages of transformers is their ability to model long-range dependencies among elements in input sequences, along with facilitating parallel processing of sequences, in contrast to recurrent networks such as Long Short-Term Memory (LSTM) models [199]. Unlike convolutional networks, transformers are designed with minimal inductive biases and are inherently suitable for functioning as set functions. Additionally, the straightforward architecture of Transformers enables the processing of various modalities (such as images, videos, text, and speech) using similar processing units, showcasing remarkable scalability to accommodate enormous networks and extensive datasets [200]. These capabilities have resulted in significant advancements across numerous vision tasks utilizing Transformer networks.

III.6.1 ViTs and Their Adaptation to Medical Imaging

Dosovitskiy et al. [194], introduced the initial pure transformer for computer vision applications, known as the ViT, in 2020. In their work, they illustrated the equivalence between multi-head self-attention linked to a multi-layer perceptron and CNNs by framing image classification as a sequence prediction problem. This approach utilized patch down-sampling and quadratic positional encoding to capture long-range dependencies among image tokens (patches) effectively. Recent literature indicates that these pure or hybrid ViTs have consistently achieved advanced results compared to CNN benchmarks. This trend has been observed across a diverse array of computer vision tasks, including image classification, image reconstruction [201], pixel segmentation [202], image captioning [203], imaging [204], and video applications [205].

CNNs have significantly impacted the domain of medical imaging due to the critical requirements for classification, segmentation, and detection across various imaging modalities [206]. Notably, approximately 90% of all healthcare data consists of instances derived from these diverse medical imaging modalities. This indicates a substantial amount of data is available to enhance effective modeling for clinical diagnosis and decision-making. CNNs have demonstrated exceptional performance due to their capacity to learn spatiotemporal dependencies within images and leverage this ability for the extraction of distinctive representations [82, 207–210]. Nevertheless, convolutional layers possess stationary weights that do not adjust for specific input images, resulting in a limited effective receptive field that constrains their capability to capture long-range dependencies among pixels. The achievements of transformers in natural image processing have prompted researchers to further explore the application of self-attention mechanisms in medical imaging to facilitate long-range dependencies between pixels. Furthermore, transformers have shown performance on par with leading CNNs in medical image classification [211], detection [212], segmentation [82], and reconstruction [213]. Literature indicates that transformers often outperform advanced CNNs; however, the comparative performance of transformers versus CNNs remains a topic of debate, with new modifications to transformer architectures emerging daily to address transformer-related challenges.

III.6.2 Transformer-based Models in Medical Image Segmentation

Transformer-based models have revolutionized medical image segmentation by leveraging their ability to capture long-range dependencies and global context, addressing the limitations of traditional CNNs that rely on localized receptive fields. These models excel at tasks such as organ segmentation, tumor detection, and multimodal image analysis, where understanding relationships across distant pixels is critical. Architectures such as the ViT and Shifted Window (Swin) Transformer have been adapted for medical imaging, along with hybrid approaches such as TransUNet and UNETR that combine transformer encoders and CNN decoders to improve performance while managing computational costs [214].

III.6.2.1 Hybrid Architectures Combining CNNs and Transformers

Hybrid architectures that combine CNNs and transformers have emerged as an effective approach to medical image segmentation, leveraging the complementary strengths of both models [215]. CNNs excel at extracting local features and spatial hierarchies, making them effective for detecting low-level texture and edges in medical scans. However, their limited receptive fields can hinder performance in capturing long-range dependencies, which are essential for segmenting complex anatomical structures (such as tumors with irregular borders). Transformers, on the other hand, use self-Attention mechanisms to model global context, but they often require large datasets and lack the inductive biases that make CNNs effective for image-based tasks [216]. To address these limitations, hybrid models combine CNN-based encoders (for local feature extraction) with transformer-based modules (for global inference) in various configurations:

- **CNN Backbone + Transformer Encoder**

The CNN Backbone + Transformer Encoder architecture is a widely adopted hybrid approach in medical image segmentation, designed to leverage the complementary strengths of CNNs and Transformers. In this setup, a CNN backbone (such as ResNet [217] or EfficientNet [218]) first extracts hierarchical local features from the input image, capturing fine-grained details like edges and textures. These feature maps are then transformed into tokens and fed into a Transformer encoder (e.g., ViT [219] or Swin Transformer [220]), which models long-range dependencies through self-attention mechanisms, enabling the network to understand global context and spatial relationships crucial for accurate segmentation. The CNN's inductive bias for local feature extraction helps mitigate the data-hungry nature of pure Transformers, while the Transformer enriches the feature representations by integrating global information. Popular models like TransUNet and UNETR exemplify this design. TransUNet, for instance, uses a ViT-based encoder to process CNN-extracted feature representations, followed by a Unet-style CNN decoder for precise pixel-wise segmentation. Similarly, UNETR employs a pure Transformer encoder to process 3D medical volumes, combined with a CNN decoder to restore spatial resolution. This hybrid approach excels in tasks like organ and tumor segmentation, where both local precision and global coherence are essential [221, 222].

- **Parallel Feature Extraction** The parallel feature extraction approach concurrently integrates CNNs and transformers, allowing both architectures to process

Hybrid DL Architectures for Medical Image Segmentation: Integrating CNNs and Transformers

input data simultaneously and fuse their complementary features for enhanced segmentation. In this design, a CNN branch extracts localized, high-resolution spatial features ideal for capturing fine details like edges and textures, while a parallel Transformer branch processes the same input using self-attention to model long-range dependencies and global context. The features from both pathways are then combined, often through concatenation or weighted fusion, ensuring the model benefits from both local precision and structural coherence [223].

This strategy is particularly effective in medical imaging, where structures like tumors or organs require both detailed boundary delineation (handled by CNNs) and contextual understanding (provided by Transformers). Models such as Contextual Transformer (CoTr) and Medical Transformer (MedT) adopt this parallel processing approach, where intermediate CNN and Transformer features are dynamically fused at multiple scales to improve segmentation accuracy. However, this approach also introduces challenges, including increased computational overhead due to dual-path processing and the need for sophisticated fusion strategies to optimally merge features [224].

- **Transformer Layers Integrated into CNN Blocks**

Integrating Transformer layers directly into CNN blocks is an emerging hybrid approach that enhances traditional convolutional networks with self-attention mechanisms at specific stages, combining the strengths of both architectures in a unified framework. Unlike sequential or parallel designs, this method embeds Transformer modules such as multi-head self-attention (MHSA) or axial attention within CNN blocks [225], allowing the model to progressively refine feature maps with both local and global receptive fields [226].

In this setup, CNN layers handle low-level feature extraction (edges, textures), while Transformer layers, inserted at intermediate or bottleneck stages, capture long-range dependencies and contextual relationships. A key advantage of this approach is its efficiency: unlike pure Transformers, which require heavy computation for full self-attention, inserting attention selectively into CNN blocks reduces overhead while maintaining performance. Techniques like windowed attention (used in Swin Transformer) or channel-wise attention further optimize computational costs. Additionally, this design retains the inductive biases of CNNs, making it more data-efficient than standalone Transformer models [227, 228].

- **Multi-Scale Transformer-CNN Fusion**

Multi-scale Transformer CNN fusion architectures represent a powerful paradigm in medical image segmentation by systematically integrating convolutional and self-attention features across different spatial resolutions. These models leverage the hierarchical nature of CNNs which naturally capture features at varying scales through pooling and strided convolutions, and enhance them with Transformer-based global context modeling at each level [229].

In this approach, the network processes input images through a multi-stage CNN encoder, where each down-sampling step generates feature maps at progressively coarser resolutions. At critical stages (bottlenecks or skip connections), Transformer modules are inserted to refine the features by capturing long-range dependencies specific to that scale [230]. For instance, low-resolution feature maps benefit from global attention to model anatomical relationships (organ shapes), while

Hybrid DL Architectures for Medical Image Segmentation: Integrating CNNs and Transformers

higher-resolution features use localized or windowed attention to preserve fine details (tumor boundaries). The fused multi-scale features are then aggregated, often via Unet-like decoders or feature pyramid networks (FPNs), to produce precise segmentation masks.

III.7 Conclusion

In this chapter, we reviewed the basic and advanced DL techniques used in medical image segmentation, focusing on CNNs and innovative hybrid frameworks that combine them with transformer models. We highlighted the remarkable efficiency of these models in accurately identifying anatomical structures and pathological regions across various medical imaging techniques. These developments represent a major step toward automated diagnosis, while recognizing the need for further research to enhance computational efficiency and generalize models across diverse datasets.

CHAPTER IV

IMPROVING MEDICAL SEMANTIC SEGMENTATION USING MODIFIED UNET MODELS.

Contents

IV.1	Introduction	51
IV.2	Dataset	51
IV.3	Experiment 01	52
IV.3.1	Proposed DCB-Unet model	53
IV.3.2	Quality metrics	54
IV.3.2.1	Dice Similarity Coefficient	54
IV.3.2.2	Accuracy	54
IV.3.2.3	IoU coefficient	55
IV.3.2.4	Loss function	55
IV.3.2.5	Training and Implementation Settings	55
IV.3.3	Experimental results and discussion	55
IV.3.3.1	Results	55
IV.3.3.2	Comparison with previous studies	57
IV.4	Exepriment 02:	59
IV.4.1	The modified VGG19-Unet model	61
IV.4.1.1	VGG19 Architecture	61
IV.4.1.2	Depthwise Separable Convolutions (DSC)	61
IV.4.1.3	VGG19-Unet modified	62
IV.4.2	Experimental Methodological Framework	64
IV.4.3	Results and performance evaluation of the modified VGG19-Unet	64
IV.4.4	Preprocessing and Hybrid Model Methodology	68
IV.4.4.1	Filtering Techniques for Noise Reduction	68
IV.4.4.2	Contrast Enhancement Using CLAHE	70
IV.4.4.3	Performance Evaluation with Preprocessed Data	71
IV.4.5	Discussion of the results and comparison with previous studies	74
IV.5	Conclusion	75

IV.1 Introduction

Accurate tumor identification is a key challenge in disease diagnosis and treatment planning, as the accuracy of tumor morphological identification is directly linked to the effectiveness of medical intervention. In this context, semantic segmentation technology for medical images emerges as a vital tool for enhancing tumor boundary identification and tissue differentiation. Recent progress in DL theories has made convolutional neural network models, especially the Unet architecture, a key part of solving this difficult problem. In this chapter, our work makes two major contributions to improving the accuracy of medical image segmentation based on the traditional Unet architecture. In the first experiment, we present a model that deepens skip connections by introducing additional convolutional layers, enhancing the flow of fine spatial features without adding significant computational burden. The second experiment relies on constructing a strong encoder using pre-trained weights of the VGG19 network on the ImageNet cluster, supplemented by decoder pipelines augmented by deep convolutional layers to recover fine details. To improve the quality of the input data, hybrid feature maps generated by applying Gaussian and Median filters and the CLAHE technique were combined with the original images, aiming to enhance important edges and texture details.

IV.2 Dataset

The dataset used in this study was obtained from The Cancer Genome Atlas (TCGA), a large-scale cancer research initiative that provides comprehensive datasets for various cancer types. This work employs the low-grade glioma (LGG) dataset, which contains 293 cases, including World Health Organization (WHO) grade II and III tumors.

The LGG dataset includes multimodal MRI scans acquired using three imaging modalities: T1-weighted (pre-contrast), FLAIR, and contrast-enhanced T1-weighted (T1C). In the processed dataset, these three modalities are combined into a single three-channel image ($256 \times 256 \times 3$), allowing the segmentation model to exploit the complementary anatomical and pathological information provided by each MRI sequence. Each image is accompanied by a corresponding ground-truth segmentation mask that delineates the tumor region.

In addition to MRI data, the LGG dataset also provides extensive molecular and genomic information. As illustrated in Figure IV.1, the dataset includes exome sequencing, DNA copy number analysis, mRNA sequencing, microRNA sequencing, DNA methylation analysis, TERT promoter sequencing, and RPPA proteomics. These molecular data have been widely used to identify distinct molecular subtypes of low-grade gliomas and to support molecular classification and personalized treatment strategies .please consult these references [231, 232].

Improving medical semantic segmentation using modified Unet models

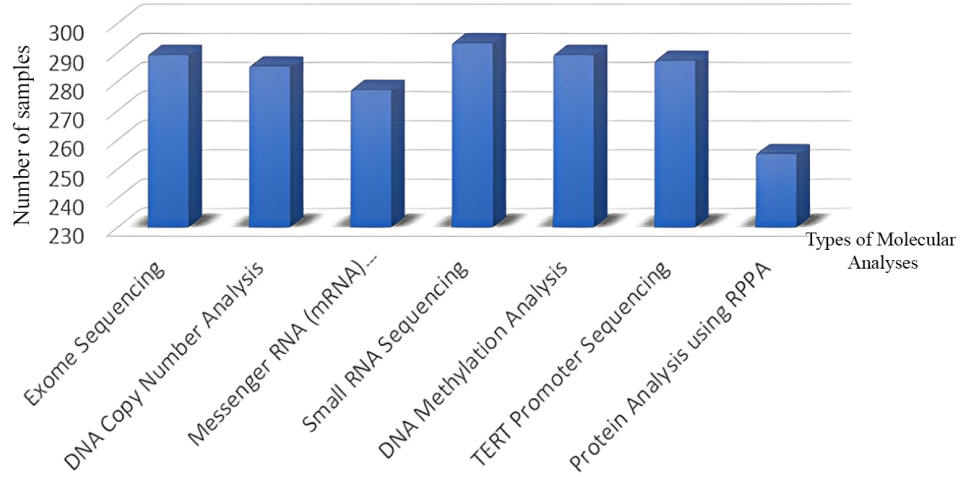


Figure IV.1: The histogram represents the types of analyses and the number of samples corresponding to each.

Since the objective of this research is brain tumor segmentation from MRI, the study is based on the multimodal MRI images and their corresponding segmentation masks. The dataset was split at the patient level, with 80% of the patients allocated to training, 10% to validation, and 10% to testing. The resulting processed dataset consists of 3,929 images, for training, 393 images for validation, and 393 images for testing. Figure IV.2 presents examples of the multimodal MRI images and their corresponding segmentation masks.

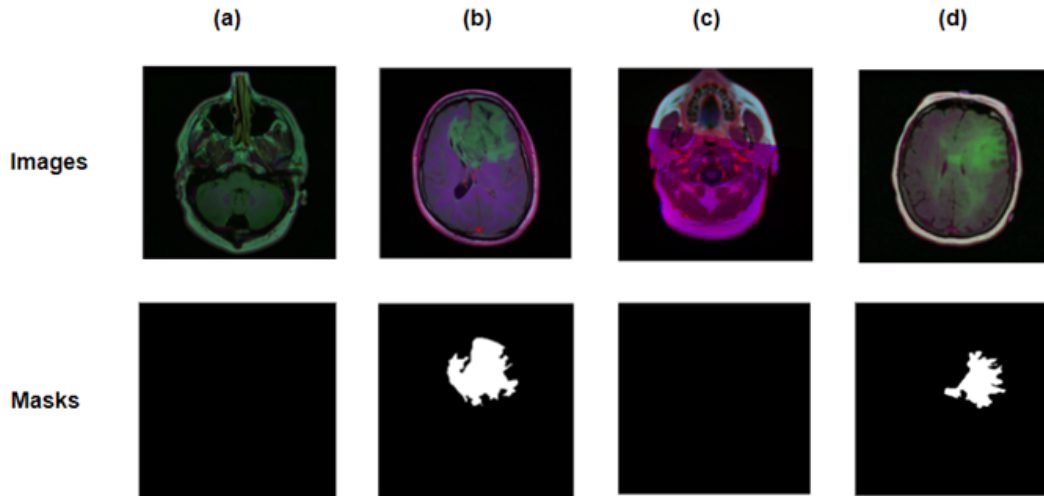


Figure IV.2: FLAIR MRI images: Original image without tumor and with tumor, ground truth image without tumor and with tumor.

IV.3 Experiment 01

The proposed DCB-UNet framework is designed for automatic brain tumor segmentation using multimodal MRI images from the LGG dataset. The model is trained using the training set, while the validation set is employed to monitor the learning process and optimize the model parameters. After training, the final model is evaluated on an independent test set that was not used during training or validation to assess its generalization capability.

Improving medical semantic segmentation using modified Unet models

The proposed architecture consists of three main components: an encoder, bottle-neck bloks, and a decoder. The encoder progressively extracts hierarchical feature representations from the input MRI images, whereas the decoder reconstructs the segmentation map by combining high-level semantic information with fine spatial details. The bottle-neck connects the encoder and decoder and captures deeper feature representations before the reconstruction stage.

The proposed model generates a binary segmentation mask for each input image, where every pixel is classified as either tumor or background. The segmentation performance is evaluated by comparing the predicted masks with the corresponding ground-truth masks using quantitative evaluation metrics, including dice coefficient, IoU, Accuracy.

IV.3.1 Proposed DCB-Unet model

The proposed DCB-UNet is based on the conventional U-Net architecture. The main modification consists of incorporating four Depthwise Convolution Bottleneck (DCB) blocks into the skip connections between the encoder and the decoder, as illustrated in Figure IV.3.

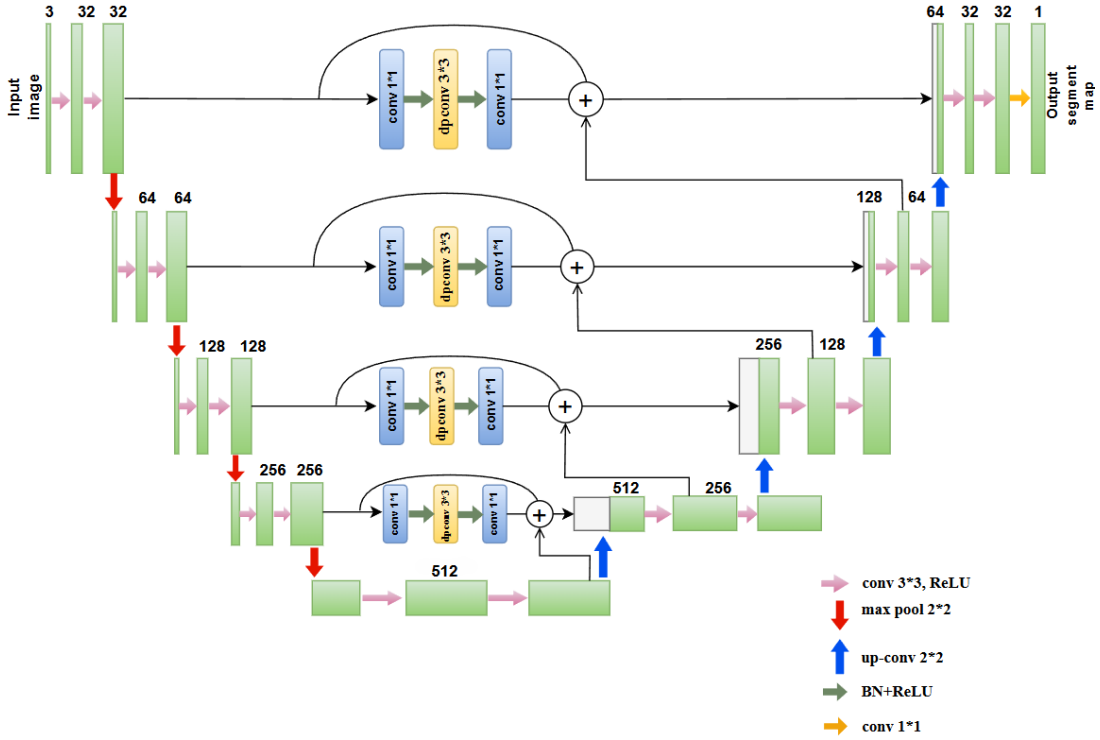


Figure IV.3: Proposed DCB-Unet model.

In the original U-Net, skip connections directly transfer encoder feature maps to the decoder in order to preserve fine spatial information. In the proposed architecture, these feature maps are first refined using the DCB blocks before being fused with the corresponding decoder features. Since the decoder reconstructs the segmentation map based on the information received from the encoder, refining these transferred features provides richer and more discriminative feature representations, leading to improved segmentation performance.

Improving medical semantic segmentation using modified Unet models

Each DCB block consists of a 1×1 convolution, followed by a depthwise 3×3 convolution, another 1×1 convolution, and a residual connection as illustrated in Figure IV.4. The residual connection preserves the original feature information while enabling the block to learn additional discriminative representations. Furthermore, the use of depthwise convolution allows the refinement process to be performed efficiently without significantly increasing the number of parameters or computational complexity.

Finally, the output layer generates a binary segmentation map in which each pixel is classified as either tumor or background.

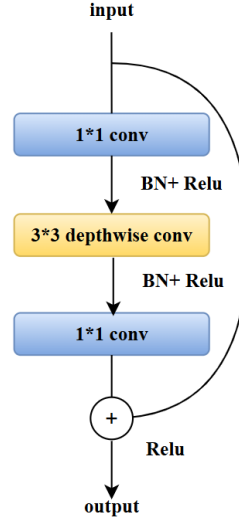


Figure IV.4: Depthwise convolution bottleneck.

IV.3.2 Quality metrics

To evaluate the performance of the proposed model and compare it with existing methods, three widely used evaluation metrics were employed: dice , IoU, and Accuracy. These metrics assess the agreement between the predicted segmentation mask and the corresponding ground-truth mask. Their computation is based on four fundamental quantities: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Here, TP denotes the number of tumor pixels correctly classified as tumor, TN represents the number of background pixels correctly classified as background, FP denotes the number of background pixels incorrectly classified as tumor, and FN represents the number of tumor pixels incorrectly classified as background.

IV.3.2.1 Dice Similarity Coefficient

The dice criterion is evaluation metric that can be employed to assess the precision of segmentation, and it is calculated utilizing the subsequent equation:

$$Dice = \frac{2TP}{2TP + FP + FN} \quad (IV.1)$$

IV.3.2.2 Accuracy

The accuracy of diagnosing or discriminating a particular condition using the binary segmentation approach is the performance metric:

$$accuracy = \frac{TP + TN}{FP + FN + TP + TN} \quad (IV.2)$$

IV.3.2.3 IoU coefficient

The Intersection over Union evaluates the prediction performance of the network through the method of evaluating the overlap level between the ground truth and the prediction, and it is calculated utilizing this equation:

$$IoU = \frac{TP}{TP + FP + FN} \quad (IV.3)$$

IV.3.2.4 Loss function

The *Dice Loss* is a loss function used to measure the error between the predicted output and the ground-truth values. Its main objective is to minimize this error during the training process, thereby improving the model's prediction accuracy. In general, the Dice Loss can be expressed as:

$$L_{Dice} = 1 - \frac{2 \sum_{i=1}^N y_i \hat{y}_i + \epsilon}{\sum_{i=1}^N y_i + \sum_{i=1}^N \hat{y}_i + \epsilon} \quad (IV.4)$$

In this equation, y_i represents the ground-truth label, $\hat{y}_i$ represents the predicted probability, N is the total number of pixels, and ϵ is a small value added for numerical stability.

IV.3.2.5 Training and Implementation Settings

The proposed model was trained using the dice loss function and the Adam optimizer with an initial learning rate of 1×10^{-4} . A `ReduceLROnPlateau` schedule was applied with a reduction factor of 0.1, a patience of 3 epochs, and a minimum learning rate of 1×10^{-7} . A batch size of 32 and a maximum of 100 epochs were used. Early stopping was based on the validation loss with a patience of 20 epochs, and the model corresponding to the best validation loss was saved for model selection.

IV.3.3 Experimental results and discussion

IV.3.3.1 Results

The proposed model effectively segments multimodal MRI images obtained from 110 patients, enabling precise differentiation between tumor and adjacent healthy brain tissue. Each input image consists of three MRI modalities (T1, FLAIR, and T1c) combined as a three-channel image to provide complementary anatomical and pathological information. Figure IV.5 illustrates the successful training and validation outcomes throughout the model's training period.

The quality metrics, including the dice and the IoU coefficients for both training and validation phases, exhibit a nearly simultaneous increase, suggesting that the model is not experiencing overfitting.

The tumor location was accurately identified in the test data, as Figure IV.6 presents several sample images depicting both the ground truth and predicted tumor regions,

Improving medical semantic segmentation using modified Unet models

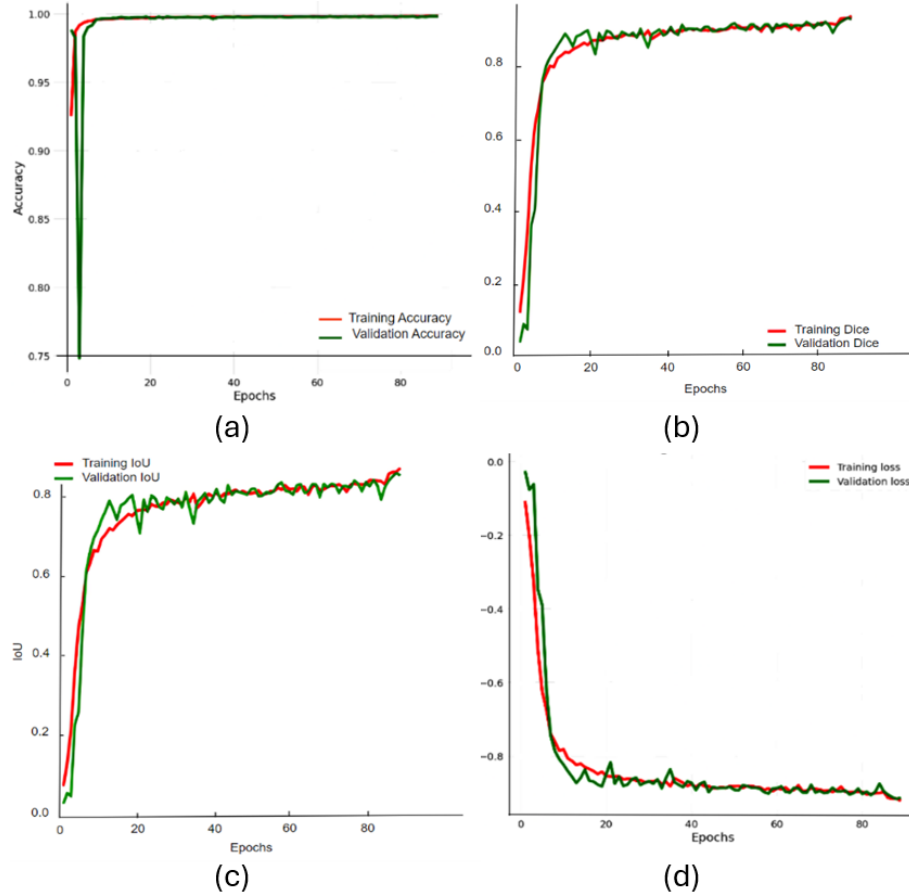


Figure IV.5: (a) Accuracy plot for proposed model (b) IoU coefficient plot for proposed model (c) Dice coefficient plot for proposed model (d) Loss plot for proposed model

derived from the test results, which demonstrate a high degree of similarity between the real and predicted tumor areas. This indicates that the DCB-Unet model is highly compatible with the LGG dataset. Furthermore, it excelled in recognizing tumor tissues and effectively segmenting them from healthy brain tissues.

The model was trained, validated, and then tested. The test data set indicates that the DCB-Unet model can predict the tumor in MRI images and its segmentation with a dice value of about 0.90 and a dice loss of 0.10, as well as successfully distinguish between pixels that were positive in the tumor reality and those that were false positives. It was found that the model successfully identified the majority of positive pixels with an accuracy of 0.99 and 0.83 for the IoU coefficient (table IV.1). We can see that the IoU coefficient performs well in both the training and validation datasets (figure IV.7) with a value above 80%, which means that the model makes a good prediction.

Table IV.1: Displaying previous works in last years

Parameters	Train	Validation	Test
Accuracy	0.9985	0.9981	0.9982
Dice	0.9286	0.9130	0.9083
IoU	0.8673	0.8409	0.8334

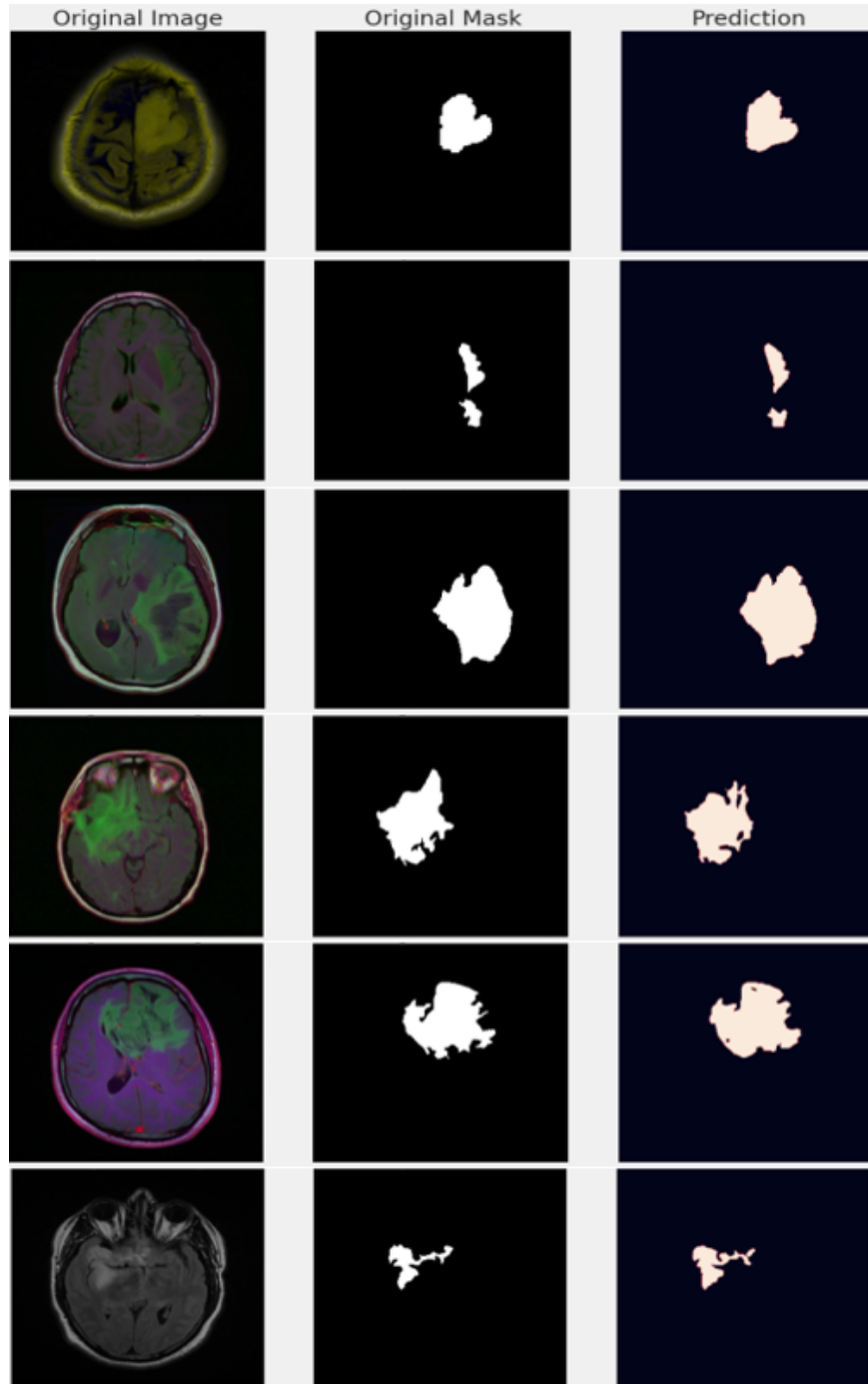


Figure IV.6: Ground truth image and segmented image

IV.3.3.2 Comparison with previous studies

Many researchers have contributed to the advancement of medical image segmentation. Including Kumar et al. [233], used the new ResUnet model to perform multispectral segmentation of MRT images and obtained accuracy, dice, and IoU of 99.80%, 90.56%, and 82.93%, respectively, while Sathish et al. [234], used a model called VGG-Unet, which integrates the weights of the initial model VGG16 in the encoder module with a stand-alone module in the decoder section to get 88.25% dice value and 87.73% IoU coefficient. As for Shomirov et al. [235], they proposed modifying the 3D Unet segmentation model

Improving medical semantic segmentation using modified Unet models

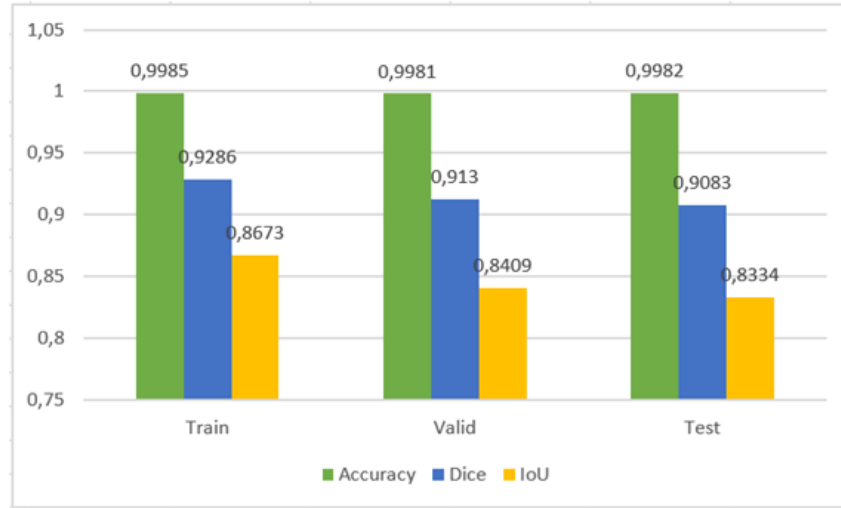


Figure IV.7: Bar graph showing the training, validation, and testing parameters.

based on the local encoder with the generalized loss function that was trained by the Adam optimization algorithm in order to improve performance their model achieved an accuracy 97.24%, and the IoU coefficient is 51.57%. dish Renugadevi et al. [236], employed the U-Net++ architecture for brain tumor segmentation. Their method achieved an accuracy of 98.81% and an IoU of 74.83% during the testing phase

Joshi and Singh [237] employed deep learning (DL) and transfer learning (TL) for brain tumor semantic segmentation based on the U-Net architecture, in semantic segmentation. brain tumors based on the Unet network. Thus, Their method achieved an accuracy of 99.88%, a Dice score of 62.02%, and an IoU of 37.55%. As shown in Table IV.2, although the reviewed studies employ different architectures and methodologies, they all rely on deep learning, particularly CNNs, for brain tumor segmentation. Similarly, the proposed work is based on the U-Net architecture, which was further enhanced by incorporating DCB blocks into its skip connections. All contributions indicate high segmentation accuracy rates, leading to its potential use in clinical applications such as computer-aided diagnosis, treatment planning, as well as monitoring disease progression.

The superiority of the proposed DCB-Unet model was evident, achieving a segmentation accuracy of 99.84%, along with dice and IoU coefficients of 90.83% and 83.34%, respectively. This success can be attributed to the effective strategy of integrating DCB blocks within the network's skip connections, while maintaining the original encoder and decoder modules. By refining the feature maps before they are fused with the decoder, the proposed model provides richer feature representations, enabling more accurate reconstruction of tumor boundaries. Furthermore, the use of depthwise convolution allows efficient feature refinement while introducing only a small increase in computational complexity, resulting in improved segmentation performance.

The performance evaluation of DL-based models depends not only on the final segmentation results but also on the performance of the model during the training and validation phases, as all previous models as well as our proposed model obtained high accuracy in image segmentation as mentioned earlier. We will focus our attention on the variation in the dice and IoU coefficients.

Figures IV.8, IV.9, IV.10, IV.11, and IV.12 compare the training and validation curves of the dice and IoU for the previous studies and the proposed DCB-UNet model. The previous models exhibit larger fluctuations and a wider gap between the training

Improving medical semantic segmentation using modified Unet models

Table IV.2: Comparison of the previous segmentation methodology with the proposed model

References	Dataset	Methodology	Accuracy	Dice	IoU	Loss
Kumar et al. [233]	LGG data	ResNet Backbone Unet model for brain tumor segmentation from FLAIR MRI images.	0.9980	0.9056	0.8293	0.0987
Sathish et al. [234]	EPISURG and clinical DICOM dataset	Proposed VGG-Unet model integrating VGG16 initial weights and a self-attention module in the decoder.	–	0.8825	0.8773	–
Shomirov et al. [235]	BraTS 2020	Modified 3D Unet segmentation model with combined encoder (Unet AE) and generalized loss function (GDL), optimized with ADAM to address class imbalance.	0.9724	–	0.5157	–
Renugadevi et al. [236]	BraTS 2020	Brain tumor segmentation using the Unet++ architecture.	0.9881	–	0.7483	0.2517
Joshi and Singh [237]	BraTS 2020	Modified Unet design with architectural improvements.	0.9938	0.6202	0.3755	0.0179
Proposed approach	LGG data	Enhanced Unet model integrating DCB blocks into skip connections.	0.9984	0.9083	0.8334	0.0916

and validation curves, indicating less stable learning behavior. In contrast, the proposed DCB-UNet demonstrates smoother convergence and closer agreement between the training and validation curves throughout the training process. This behavior suggests that the proposed model generalizes well to unseen data and shows no significant signs of overfitting. Since Dice and IoU are mathematically related overlap-based metrics, they follow similar trends during both the training and validation phases.

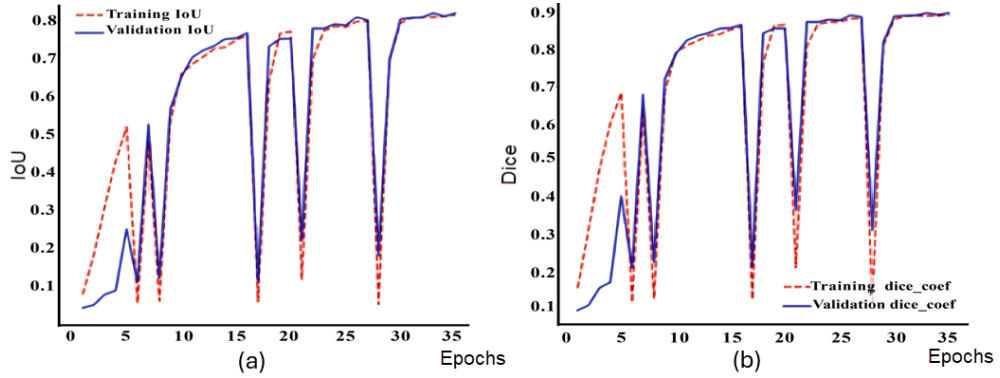


Figure IV.8: (a) IoU coefficient plot for Kumar et al. (b) DSC coefficient plot for [233]

IV.4 Experiment 02:

This experiment investigates the effectiveness of a hybrid approach that combines the powerful feature extraction capability of pretrained deep networks with image preprocessing techniques. Transfer learning enables the model to leverage knowledge learned from the ImageNet dataset, while Gaussian and median filtering are applied to reduce noise and preserve important structural information. In addition, CLAHE is employed

Improving medical semantic segmentation using modified Unet models

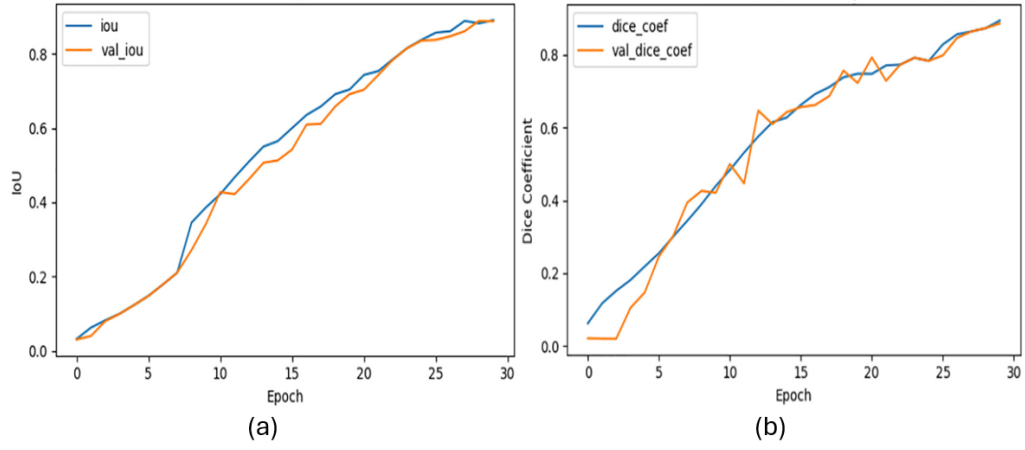


Figure IV.9: (a) IoU coefficient plot for Sathish et al. (b) DSC coefficient plot for [234]

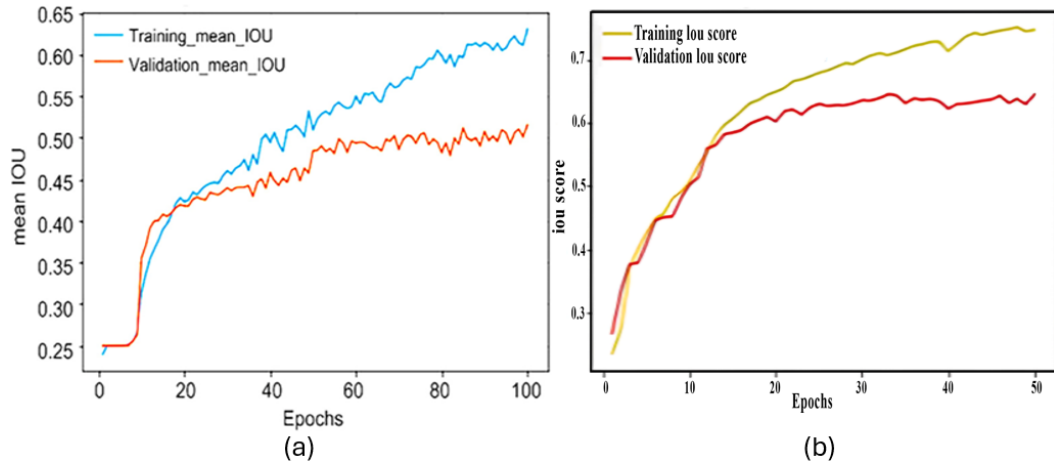


Figure IV.10: (a) IoU coefficient plot for [236]. (b) DSC coefficient plot for [237]

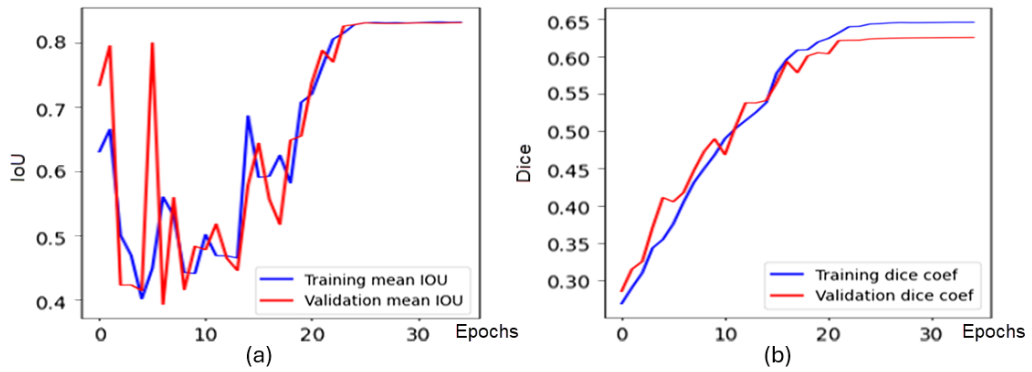


Figure IV.11: (a) IoU coefficient plot for Joshi et al. (b) DSC coefficient plot for [237]

to improve image contrast. This work is based on the hypothesis that integrating deep feature extraction mechanisms with targeted preprocessing enhances the recognition of fine anatomical structures and lesion boundaries in MRI images, which contributes to improving semantic segmentation accuracy by enhancing the quality of features fed into the model.

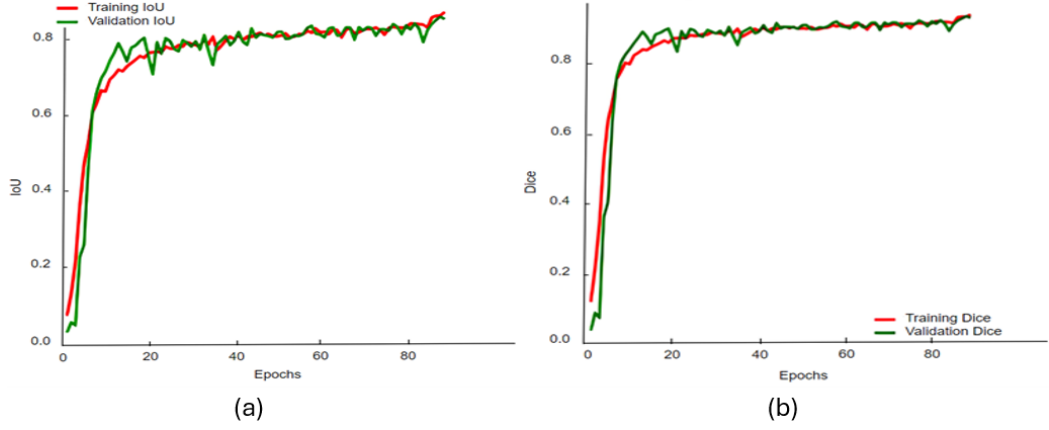


Figure IV.12: (a) IoU coefficient plot for proposed approach. (b) DSC coefficient plot for the proposed approach

IV.4.1 The modified VGG19-Unet model

IV.4.1.1 VGG19 Architecture

VGG19 is a popular deep model in computer vision, developed by the VGG at the University of Oxford. The model consists of 19 consecutive layers, including 16 convolutional layers and 3 fully connected layers [238]. VGG19 is characterized by its use of a small convolutional kernel 3x3 with a stride of 1 enabling high-resolution spatial feature capture via a hierarchical feature extraction process.

To use VGG19 as an encoder in a Unet architecture, its original structure was changed to fit the needs of image segmentation. The fully connected layers were removed, leaving only the convolutional core responsible for representing spatial features. The outputs of the five convolutional blocks were then used as feature extraction points and skip connections to the decoder. The figure IV.13 illustrates this structural adaptation, where each block employs several 3x3 Convolution layers followed by MaxPooling to halve the spatial dimensions while increasing the channel depth gradually. The property maps start at $256 \times 256 \times 64$ and gradually decrease to $16 \times 16 \times 1024$. It should be noted that MaxPooling is not applied in the last block, as it represents the bottleneck (bridge) in the Unet architecture, where the deepest possible spatial representation is maintained before passing these features to the decoder to accurately reconstruct the segmentation map.

IV.4.1.2 Depthwise Separable Convolutions (DSC)

Depthwise separable convolution is a convolutional operation used in CNNs to reduce computational cost and the number of trainable parameters while maintaining competitive performance. This convolutional approach is divided into two primary phases: depthwise convolution and pointwise convolution. During the initial phase, an individual filter is utilized for each input channel independently, producing a distinct feature map for every channel. In the case where the input possesses dimensions denoted as (H, W, C) , with H representing height, W indicating width, and C signifying the number of channels, then deep convolution uses a C filter of size $(K, K, 1)$, resulting in dimensions (H, W, C) . The mathematical formula for deep convolution is expressed in equation 4.

Improving medical semantic segmentation using modified Unet models

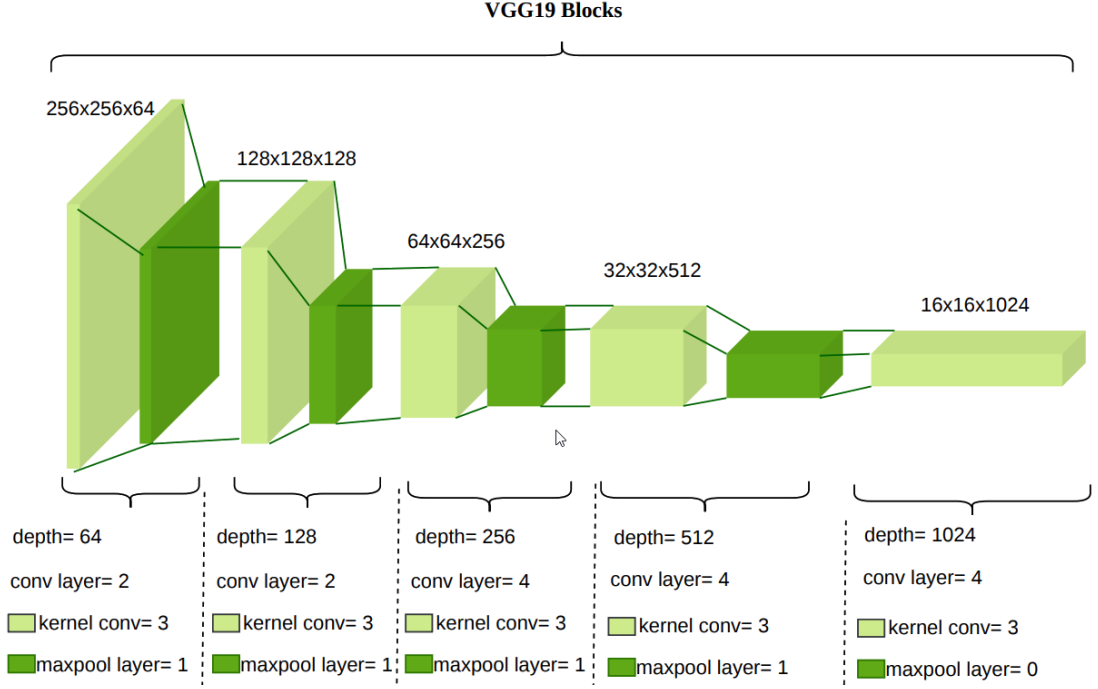


Figure IV.13: Architecture Diagram of VGG19 CNNs

$$G_{i,j,c} = \sum_{m=1}^K \sum_{n=1}^K X_{i+m-1,j+n-1,c} \cdot K_{m,n,c} \quad (IV.5)$$

In the second stage, a pointwise convolution is applied using a filter of size (1, 1, C) to combine information across channels, resulting in dimensions (H, W, C), where C is the number of filters required. The mathematical formula for pointwise convolution is expressed in equation 5. Figure IV.14 shows an illustrative example of the difference between a standard convolution and a Depthwise convolution. The most interesting feature is the reduction of computational cost by reducing the number of parameters while maintaining good performance.

$$O_{i,j,c'} = \sum_{c=1}^C G_{i,j,c} \cdot P_{1,1,c,c'} \quad (IV.6)$$

IV.4.1.3 VGG19-Unet modified

The proposed VGG19-Unet model (Figure IV.15) employs a pretrained VGG19 network as the encoder of the U-Net architecture. The fully connected layers of the original VGG19 are removed, retaining only the convolutional layers to extract rich hierarchical feature representations from the input MRI images. Through successive convolution and MaxPooling operations, the encoder captures both low-level features, such as edges and textures, and high-level semantic information while progressively reducing the spatial resolution of the feature maps [239, 240].

Improving medical semantic segmentation using modified Unet models

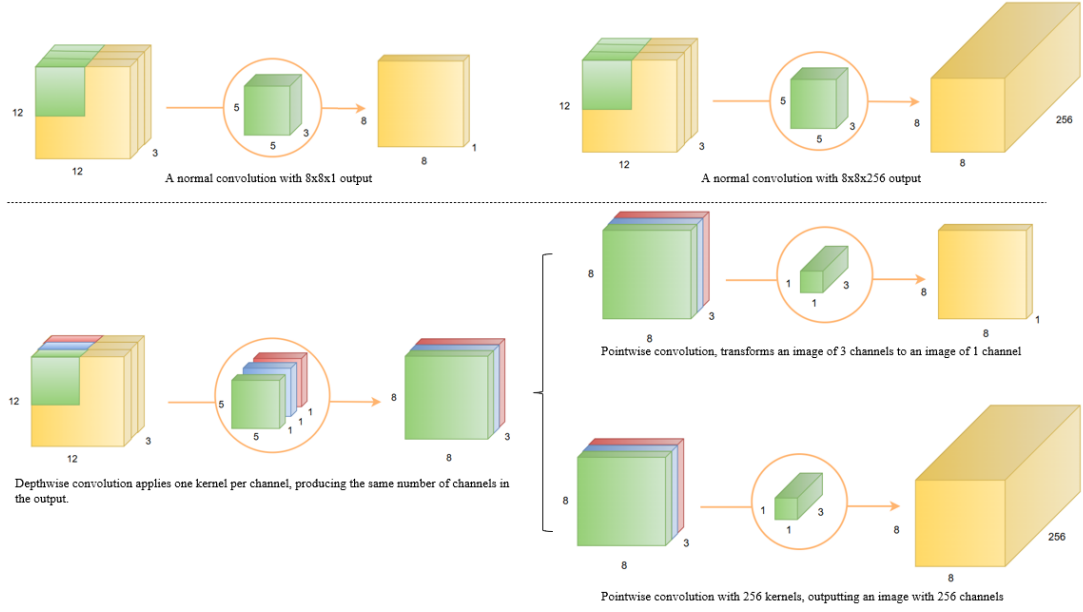


Figure IV.14: A diagram showing the difference between natural and deep convolution

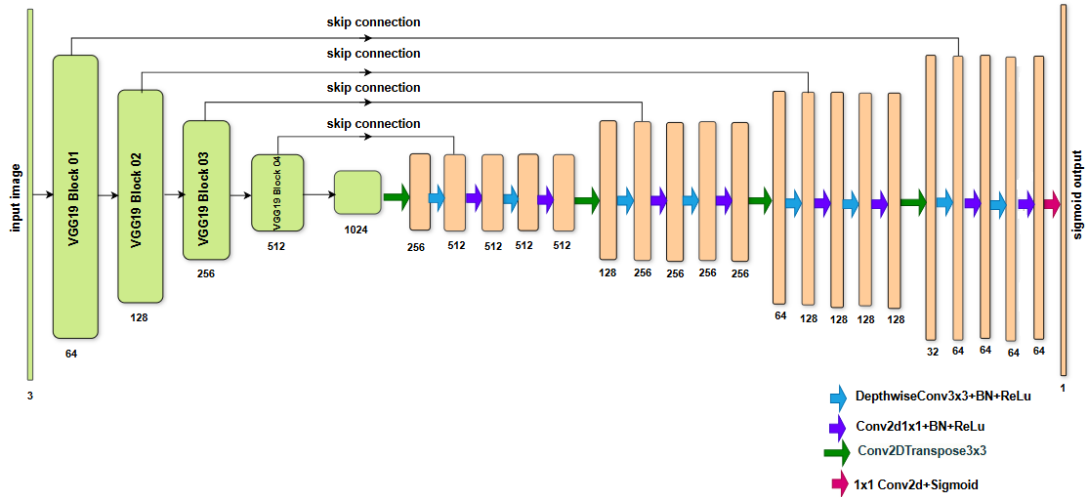


Figure IV.15: The architecture of the modified VGG19-Unet model

In the proposed architecture, each pair of standard convolutional layers is replaced with two DSC layers. Compared with conventional convolutions, DSC significantly reduces the number of trainable parameters and computational complexity. Each DSC layer performs depthwise convolution independently on every input channel, followed by a 1×1 pointwise convolution that integrates information across channels, thereby improving computational efficiency while preserving informative feature representations [241, 242].

The decoder progressively upsamples the extracted features using transposed convolution layers, while the original U-Net skip connections fuse the corresponding encoder and decoder feature maps to preserve fine spatial details and improve localization accuracy. Finally, a 1×1 convolutional layer generates the binary segmentation map by classifying each pixel as either tumor or background [243, 244]. To evaluate the proposed model, the same evaluation metrics used in the previous experiment were adopted, enabling a fair comparison of segmentation performance.

IV.4.2 Experimental Methodological Framework

This study employed a multi-stage methodology to evaluate and develop a tumor segmentation model, focusing on independently examining the impact of each enhancement. Initially, a standard Unet model was tested on the original dataset to determine the model's baseline performance without any additions. Subsequently, a Unet model equipped with a pre-trained on ImageNet VGG19 encoder was evaluated to determine the effect of using a robust, pre-trained encoder on more accurate feature extraction and representation. Finally, DSC blocks were incorporated into the decoder stage and tested separately to study their impact on improving learning efficiency and reducing computational complexity. This staged approach allows for a thorough understanding of each addition's contribution before its integration into the final model.

Based on the results of the previous phases, a modified version of the Unet architecture was derived, resulting in improved performance and fewer parameters. The final version was directly tested on the LGG dataset. The primary objective of this phase was to determine the model's "baseline performance," which serves as a benchmark for evaluating any subsequent improvements. The data was prepared for input into the model through standard preprocessing operations limited to normalization. The dataset was then divided into three sets, training, validation, and testing, to ensure an objective and unbiased evaluation of the model's performance.

The model was trained on the training dataset, where it learned to adjust its internal weights to minimize the difference between its predictions and the ground truth. The final output of the model was a segmentation mask that accurately identified each pixel in the original image as tumor or healthy tissue. To quantitatively and objectively evaluate the model's performance, a set of standard performance metrics were calculated: dice, accuracy, and IoU coefficient. These metrics were calculated by comparing the masks predicted by the model on the test set with ground-truth masks prepared and validated by medical professionals. The results of this initial phase serve as a benchmark against which the value and effectiveness of any subsequent improvements, such as the application of advanced image processing techniques, can be measured. This scientific approach ensures that the added value of any subsequent modifications is conclusively and objectively demonstrated.

IV.4.3 Results and performance evaluation of the modified VGG19-Unet

This section presents the experimental results. Initially, the modified VGG19-Unet was trained and validated on the LGG dataset, and the obtained results were recorded.

The results shown in Table IV.3 illustrate the performance of the unet model with various additions in the semantic segmentation task, based on the dice and IoU metrics, as well as the number of parameters per model.

Table IV.3: Performance Analysis

Model	Dice	IoU	N of parameters (M)
Unet	0.8417	0.7487	31,03
VGG19-Unet	0.8892	0.8036	31,17
Unet-DSC	0.8547	0.7576	22,72
modified VGG19-Unet	0.8921	0.8072	22,85

Improving medical semantic segmentation using modified Unet models

We can observe that the baseline Unet model achieved acceptable performance, with a Dice value of (0.8417) and an IoU value of (0.7487), reflecting its capability to perform the segmentation task with reasonable efficiency. However, when using the VGG19-Unet model, a significant improvement in performance is observed, as the values increased to (0.8892) and (0.8036), respectively. This improvement can be attributed to the ability of the VGG19 network to extract rich and deep feature representations that enhance segmentation accuracy.

In contrast, the Unet model based on Depthwise Separable Convolutions showed a slight improvement in performance compared to the baseline Unet model, achieving a dice of (0.8547) and an IoU of (0.7576), compared to (0.8417) and (0.7487) in the baseline model. Additionally, this model demonstrated a notable reduction in the number of parameters, decreasing to 22,723,905 compared to 31,031,745 in the conventional Unet, indicating higher efficiency in terms of computational complexity and resource utilization.

Finally, the modified VGG19-Unet model achieved the best performance among all models, recording the highest values for both dice (0.8921) and IoU (0.8072), while maintaining a relatively lower number of parameters (22,856,513) compared to the standard VGG19-Unet model. These results indicate that the proposed modifications contributed to improving model accuracy while reducing computational complexity, thereby enhancing its efficiency and suitability for practical applications.

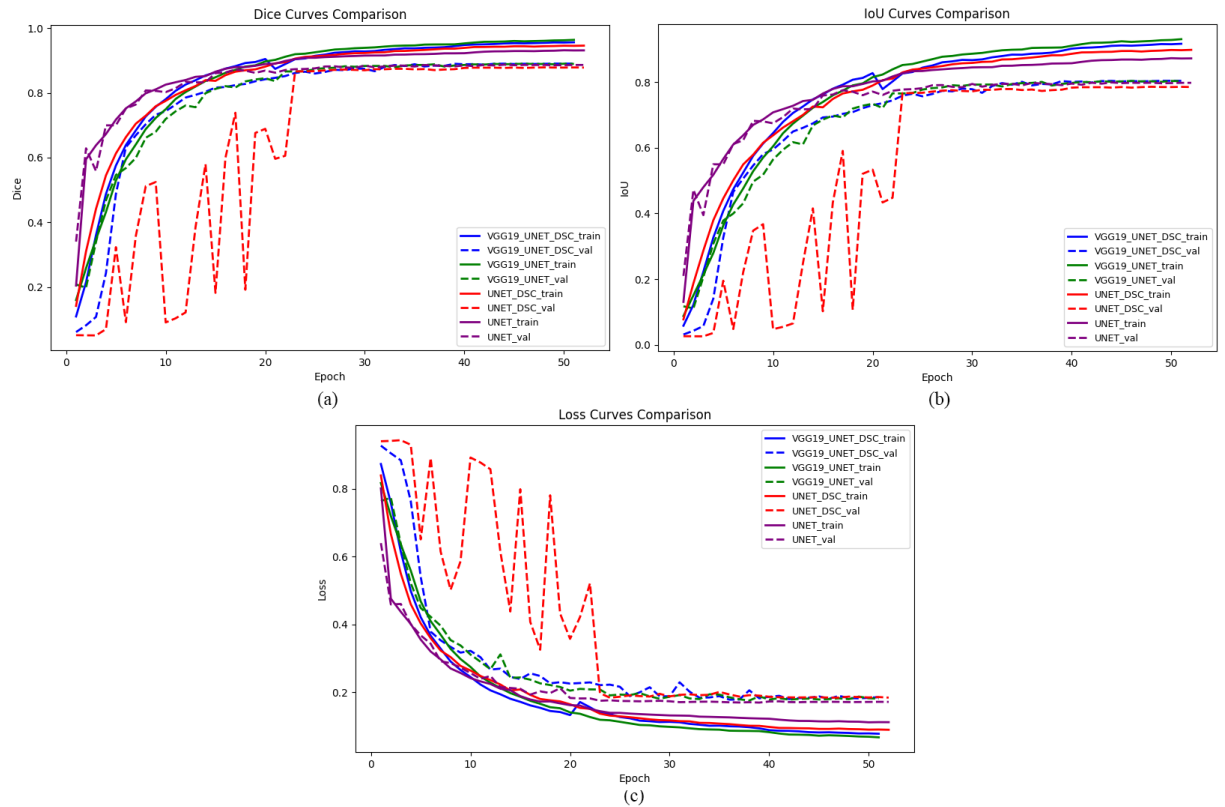


Figure IV.16: Segmentation performance curves: (a) Dice, (b) IoU, and (c) loss.

The performance curves comparing three modifications to the UNet model architecture UNet based on DSC, VGG19-UNet, and VGG19-UNet based on DSC with traditional Unet are shown in Figure IV.16. These curves reveal a clear gradual improvement in learning quality across training epochs. The Dice curve shows that the VGG19-UNet model with DSC achieves the best gradual convergence and is more stable, with both the training

Improving medical semantic segmentation using modified Unet models

and validation dice scores increasing smoothly to significantly higher levels compared to the other models. This behavior reflects the model’s ability to extract distinct and deep spatial feature representations thanks to the integration of a deep feature extractor with separable convolutions, which reduces computational complexity without compromising feature quality.

The VGG19-UNet model without DSC blocks performs well and is close to the best model, but slightly below it, demonstrating the positive impact of using DSC in reducing computational complexity and model parameters while preserving reconstruction quality. In contrast, the conventional UNet exhibits rapid convergence in the early stages of training but stabilizes at lower values, reflecting its limited representativeness compared to the two VGG19-based models.

The most striking behavior is observed in the UNet model with DSC, where we see very sharp fluctuations in the validation curve during the initial training epochs, both in dice and IoU. These fluctuations indicate the instability of the early learning process, with the model abruptly shifting between very weak and better predictions. This behavior is explained by the fact that depthwise separable convolutions, despite their high computational efficiency, reduce the model’s ability to integrate information across channels during the early stages of learning, making the model more sensitive to weight updates when the encoder does not provide sufficiently rich feature representations.

The IoU curves confirm this general behavior, with VGG19-UNet with DSC outperforming in terms of final values and stability, followed by VGG19-UNet, then the conventional UNet, while UNet with DSC continues to exhibit significant instability for a large portion of the training. Since IoU is a more rigorous measure than Dice, this fluctuation suggests that the model was producing unstable masks in the early stages before gradually learning the correct boundary for the objects.

The loss curves clearly support these observations. Both VGG19-UNet and VGG19-UNet with DSC exhibit a smooth and consistent decrease in loss with a very small gap between training and validation, indicating stable learning and excellent generalization. In contrast, UNet with DSC shows large fluctuations in validation loss during the early epochs, further evidence that using separable convolutions without deep representational support can lead to difficulties in stabilizing the numerical optimization process. Overall, the small gap between the training and validation curves in the VGG19-UNet with DSC model reflects a high generalization capability. It can be concluded that the effectiveness of depthwise separable convolutions depends heavily on the strength of the architecture used, and that combining it with a deep feature extractor such as VGG19 achieves the best balance between accuracy, stability, and computational efficiency.

Figure IV.17 presents some examples of segmentation results on brain MRI images, arranged in four columns representing, respectively, the original image, the true mask (ground truth), the model’s predicted mask, and the overlay image between predicted mask, ground truth and the original image. These samples illustrate the visual performance of our proposed model across different cases, including variations in tumor size, location, and shape.

In the rows shown, a wide variety of tumor characteristics can be observed in terms of area, shape, and anatomical location. Some cases represent very small, difficult-to-detect tumors, while others represent larger tumors with well-defined borders. This variety aims to evaluate the model’s ability to handle both small and large tumors within the same experimental framework. Comparing the second and third columns, it is observed that the predicted mask closely matches the true mask in most cases. This is clearly evident

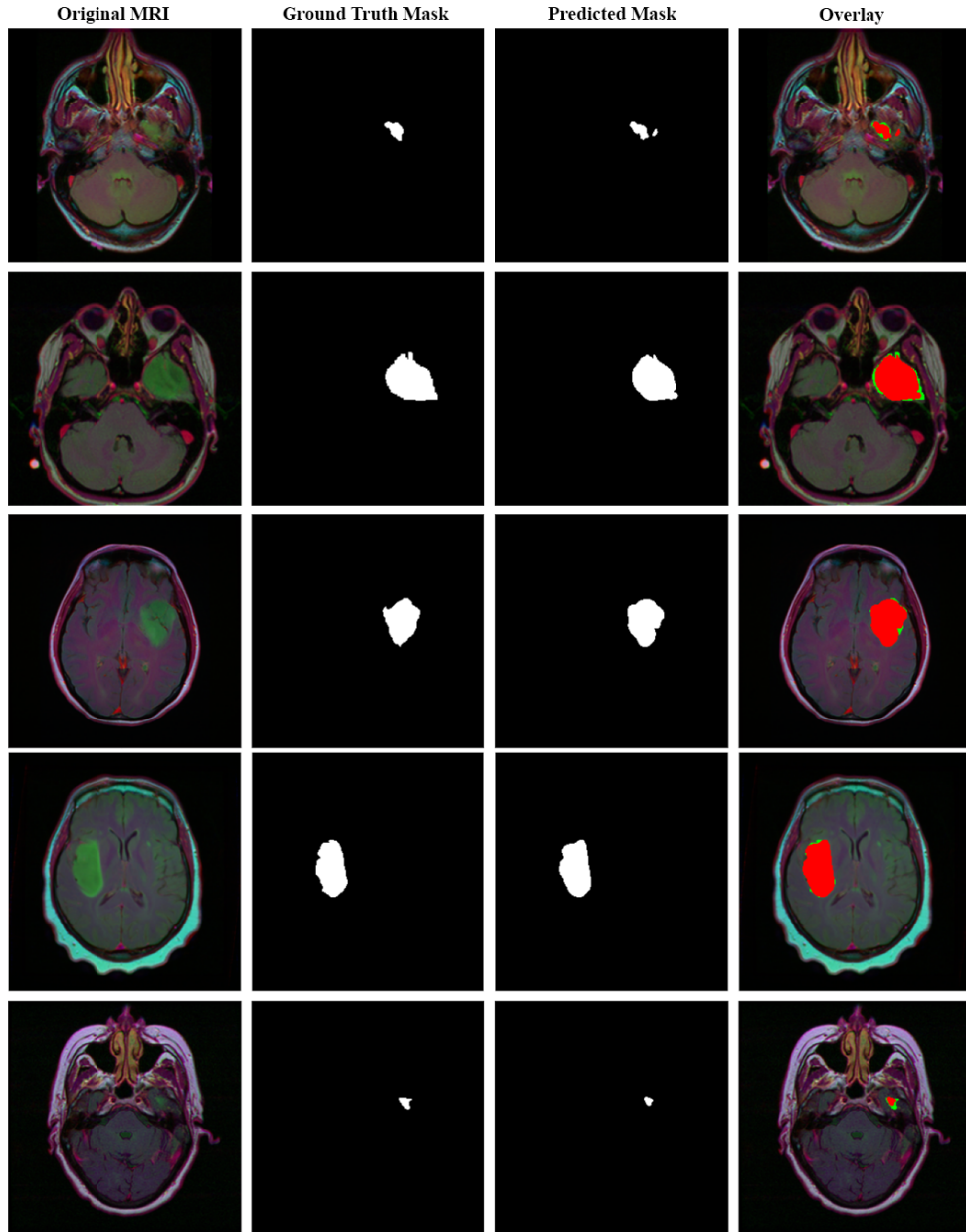


Figure IV.17: The original image and their corresponding segmented representations.

in the second, third, and fourth rows, where the tumor geometries closely resemble the predicted and true images in terms of area, perimeter, and location. This reflects the model's ability to learn a precise spatial representation of tumor boundaries, not just an approximate location.

The overlay images in column 4 demonstrate the accuracy of the overlap between the predicted and actual mask; the red color (prediction) overlaps significantly with the actual area. In cases with large tumors, the overlap is almost complete, indicating high Dice and IoU values for these samples. In cases with very small tumors (as in rows 1 and 5), very slight differences in size or shape are observed, which is expected given the sensitivity of segmentation to small tumors, where even a few pixels significantly affect the quantitative metrics.

These samples also reveal that the model did not produce false positives outside the tumor area against the black background, demonstrating good background-to-pathological

differentiation, a crucial characteristic in medical segmentation applications.

Overall, these qualitative results highlight the model’s ability to achieve accurate and stable segmentation across various sizes and shapes, with a clear capacity to accurately define the anatomical boundaries of tumors. These visual results also confirm what was observed in the quantum curves (dice and IoU), where numerical stability in performance is reflected in the visually predicted quality of the masks.

IV.4.4 Preprocessing and Hybrid Model Methodology

In this study, a modified VGG19-Unet model was employed for the segmentation of LGG. The preprocessing stage focused on enhancing the quality of medical images before inputting them into the model, using two categories of techniques: filters and contrast enhancement methods, specifically CLAHE (Contrast Limited Adaptive Histogram Equalization).

Filters were applied to reduce noise in the images. The Gaussian filter smooths the image and reduces normally distributed noise while preserving the overall structure, whereas the Median filter mitigates impulsive noise while relatively maintaining important edges.

In contrast, the CLAHE technique was applied separately to enhance local contrast within the image by redistributing intensity values over small regions, thereby highlighting fine details and improving the delineation between different tissue types.

The use of these techniques aims to improve the quality of the data fed into the model. Reducing noise through filters, combined with enhancing contrast using CLAHE, is expected to facilitate the model’s learning of distinctive features between healthy tissue and tumor tissue. The impact of this preprocessing will be subsequently evaluated through experimental results.

IV.4.4.1 Filtering Techniques for Noise Reduction

Filters are fundamental tools in image preprocessing, primarily used to reduce noise while preserving essential structural information. Noise in medical images can arise from various sources, including acquisition artifacts and environmental interference, which may adversely affect the performance of segmentation models. Applying appropriate filters helps enhance image quality, facilitating more accurate feature extraction by the network [245] [246].

In this study, two widely used filtering techniques were employed: Gaussian filtering, which effectively reduces Gaussian (normally distributed) noise by smoothing the image, and Median filtering, which is particularly effective for removing impulsive noise while preserving edges. The principles and applications of each filter are described in the following sections.

- a) Gaussian: The Gaussian filter is a linear smoothing filter commonly used in image processing to reduce normally distributed noise while preserving the general structure of the image. It operates by convolving the input image with a Gaussian function, which assigns higher weights to pixels near the center of the kernel and progressively lower weights to pixels farther away. This weighted averaging effectively smooths intensity variations caused by noise, without significantly blurring edges [247] [248]. Mathematically, the two-dimensional Gaussian function is defined as:

$$G(x, y) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right) \quad (\text{IV.7})$$

where x and y are the coordinates of the kernel, and σ determines the standard deviation, controlling the degree of smoothing.

In medical image preprocessing, applying a Gaussian filter helps reduce random noise introduced during image acquisition, yielding cleaner images that facilitate more accurate feature extraction by segmentation models, such as the modified VGG19-Unet used in this study. As is clearly the improvement Figure IV.18.

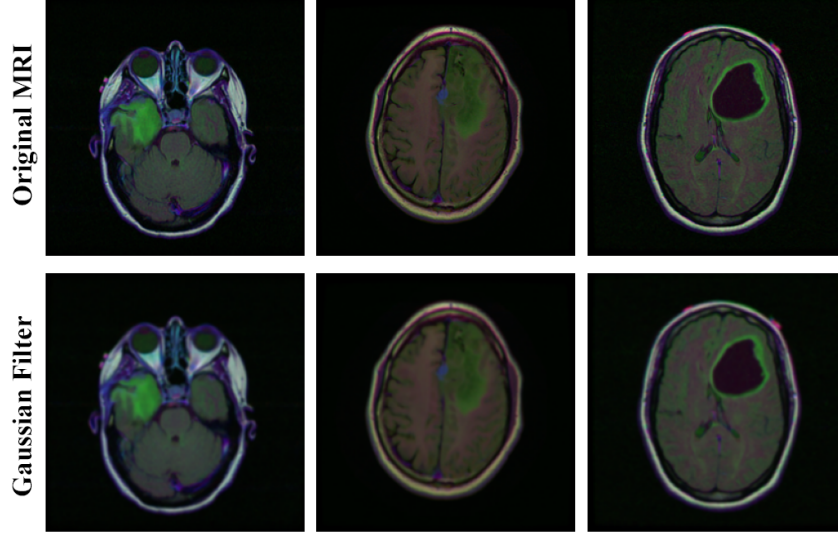


Figure IV.18: FLAIR MRI images after application of the Gaussian filter.

- a) Median: The Median filter is a nonlinear filtering technique widely used to remove impulsive noise (also known as salt-and-pepper noise) from images while preserving important edges. Unlike linear filters that perform weighted averaging, the median filter replaces each pixel value with the median value of its surrounding neighborhood, effectively eliminating extreme intensity values caused by noise [249] [250]. Mathematically, the median filter operates by considering a neighborhood of pixels around each pixel in the image, usually a square or rectangular window. For a given pixel $p(i, j)$ at location (i, j) , the median value $M(i, j)$ of the pixel in the neighborhood is calculated as follows:

$$M(i, j) = \text{median}\{f(i + k, j + l) \mid k, l \in \{-1, 0, 1\}\} \quad (\text{IV.8})$$

where $f(i + k, j + l)$ represents the pixel values within the neighborhood, and the set includes all the pixel values in the window defined by k and l , typically within a 3×3 window around (i, j) . The value $M(i, j)$ replaces the original pixel value $p(i, j)$. This approach is particularly effective in medical imaging, where fine structures and edges must be preserved for accurate segmentation. By maintaining the integrity of edges, the median filter ensures that the boundaries of anatomical structures, such as tumor regions, remain sharp, which is critical for models like the modified VGG19-Unet to accurately learn and delineate relevant features. In this work, a kernel size of 3×3 was employed, providing a

balance between effective noise reduction and preservation of fine image details. While larger kernels can remove more noise, they may also lead to a loss of important structural information, whereas smaller kernels help retain details but may leave some residual noise.

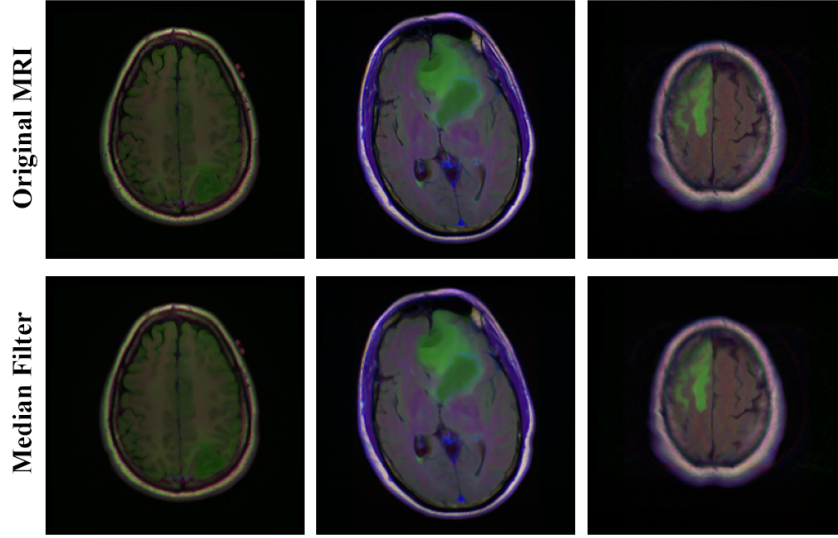


Figure IV.19: FLAIR MRI images after application of the Median filter.

As illustrated in Figure IV.19, the median filter produces a subtle smoothing effect while preserving the structural details and object boundaries of the original image. Although the visual differences are slight, the filter is intended to suppress impulsive noise without degrading important anatomical structures.

IV.4.4.2 Contrast Enhancement Using CLAHE

Contrast enhancement is a crucial preprocessing step in medical image analysis, as it improves the visibility of important structures and facilitates better feature extraction. In this study, Contrast Limited Adaptive Histogram Equalization (CLAHE) was employed to enhance the local contrast of the images [251]. Unlike global histogram equalization, which applies a uniform transformation to the entire image, CLAHE operates on small contextual regions (tiles). It redistributes the intensity values within each region, thereby enhancing local details without over-amplifying noise. Mathematically, CLAHE can be described by the following steps:

1. **Image Division into Tiles:** The image $I(x, y)$ is divided into non-overlapping tiles of size T , where each tile T_i covers a local region of the image. This allows CLAHE to operate on smaller patches, preserving local contrast information.
2. **Local Histogram Computation and Contrast Limiting:** For each tile, a local histogram is first computed to represent the distribution of pixel intensities within that region H_i . CLAHE then limits the contrast by clipping the histogram at a predefined threshold ($T_{\max}$). If any histogram bin exceeds ($T_{\max}$), the excess is redistributed equally among the other bins. The contrast-limited histogram ($H_i^{CL}(k)$) is obtained by the following relation:

$$H_i^{CL}(k) = \min(H_i(k), T_{\max}) \quad (\text{IV.9})$$

The clipped values are then redistributed across the bins to maintain the total histogram sum.

3. Cumulative Distribution Function (CDF) calculation: The normalized CDF is then calculated using the clipped histogram as follows:

$$CDF_i(r) = \frac{1}{N_i} \sum_{k=0}^r H_i^{CL}(k) \quad (\text{IV.10})$$

where $H_i^{CL}(k)$ denotes the frequency of pixels with intensity level k in the i -th tile, N_i is the total number of pixels within the tile, and r denotes the intensity level.

4. Pixel Intensity Mapping: After applying contrast limiting and redistributing the clipped histogram, the normalized CDF is used to map each pixel to a new intensity value according to:

$$I^{\text{new}}(x, y) = (L - 1) \times CDF_i(I(x, y)) \quad (\text{IV.11})$$

where $I(x, y)$ is the original pixel intensity, $I^{\text{new}}(x, y)$ is the enhanced pixel intensity, and L is the total number of gray levels in the image (typically $L = 256$ for an 8-bit image). This mapping ensures that each pixel's intensity is adjusted according to the local contrast enhancement.

Additionally, CLAHE incorporates a contrast-limited mechanism that prevents excessive enhancement, which could otherwise lead to noise amplification [252]. This technique is particularly effective in medical imaging, where subtle intensity differences between tissues are critical. By improving local contrast, CLAHE can enhance the visibility of tumor boundaries and fine structural details, making it easier for the segmentation model to distinguish between healthy and abnormal tissues. In this work, CLAHE was applied as a preprocessing step to improve image quality before model training. As illustrated in Figure IV.20, the application of CLAHE results in clearer contrast and better delineation of regions of interest compared to the original images.

IV.4.4.3 Performance Evaluation with Preprocessed Data

The evaluation of preprocessing techniques on the performance of the modified VGG19-Unet model for image segmentation revealed that all tested methods improved the model's results. As shown in Table IV.4, without preprocessing, the model achieved a Dice score of 0.8921 and an IoU score of 0.8072, indicating strong baseline performance in capturing fine details. Applying Gaussian filtering slightly enhanced these metrics to a dice score of 0.9006 and an IoU score of 0.8197, reflecting its noise reduction benefits. Similarly, Median filtering led to comparable improvements, with a dice score of 0.8997 and an IoU score of 0.8180, though it might compromise some fine details. Notably, CLAHE delivered the best performance, achieving a dice score of 0.9055 and an IoU score of 0.8278, demonstrating the importance of enhanced local contrast for distinguishing fine image features. Overall, these results highlight the significant role of preprocessing in boosting the modified VGG19-Unet model's segmentation accuracy, with CLAHE being the most effective technique.

The dice coefficient curves (as shown in Figure IV.21) illustrate the evolution of segmentation performance across training epochs. A rapid increase is observed during

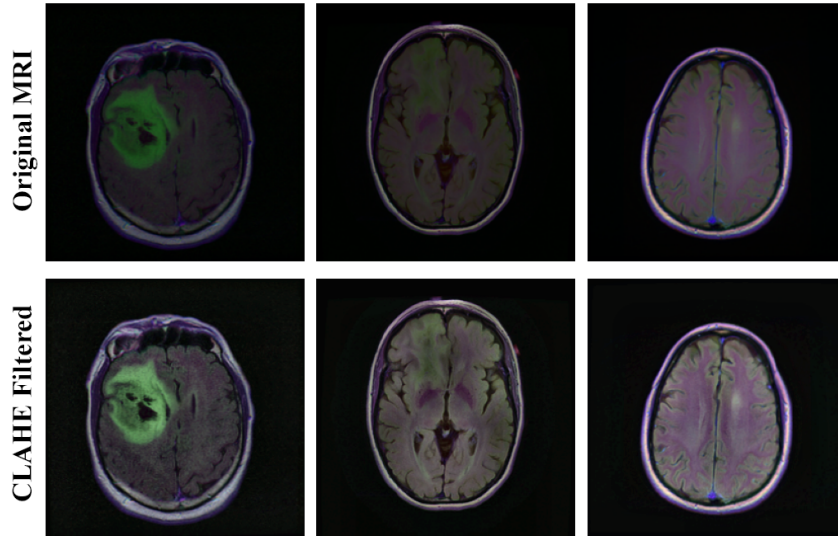


Figure IV.20: FLAIR MRI images after application of the CLAHE.

Table IV.4: Effect of Preprocessing on VGG19-Unet Performance

Model	Preprocessing	Dice	IoU
Modified VGG19-Unet	No	0.8921	0.8072
	With Gaussian	0.9006	0.8197
	With Median	0.8997	0.8180
	With CLAHE	0.9055	0.8278

the first ten epochs. This is followed by a period of slower but continuous improvement until approximately epoch 30, after which the curves reach a stable plateau, reflecting the model's ability to efficiently learn the fundamental features of medical images. In terms of model comparison, the configuration based on CLAHE preprocessing achieved the best performance, recording the highest dice values for both training and validation datasets. This was followed closely by the Gaussian filter, which demonstrated competitive performance with notable stability. In contrast, the median filter yielded relatively lower performance, while the baseline model without preprocessing produced the weakest results, highlighting the importance of enhancing input data quality prior to training.

Regarding the IoU curves, a similar trend to the Dice curves is observed, which is expected due to the strong correlation between the two metrics. The CLAHE-based model again achieved the highest IoU values, indicating a more accurate overlap between predicted and ground truth segmentation masks. The Gaussian filter exhibited competitive performance, whereas the median filter showed moderate results. It is also noteworthy that all models reached a stable performance level after approximately 30 training epochs, suggesting that preprocessing techniques contributed to improving the model's learning efficiency. Furthermore, the close alignment between training and validation curves indicates a good generalization capability.

Concerning the loss curves, all models exhibited a sharp decrease in loss values during the initial training stages, reflecting rapid performance improvement. This decline continued more gradually until reaching a stable level. The model utilizing CLAHE preprocessing achieved the lowest loss values, indicating the best agreement between predictions and ground truth. The Gaussian filter-based model showed very similar performance, while the median filter model demonstrated moderate results. In contrast,

Improving medical semantic segmentation using modified Unet models

the baseline model recorded the highest loss values. Additionally, a small gap between training and validation loss curves was observed; however, it remained within acceptable limits, suggesting the absence of significant overfitting and confirming good generalization performance.

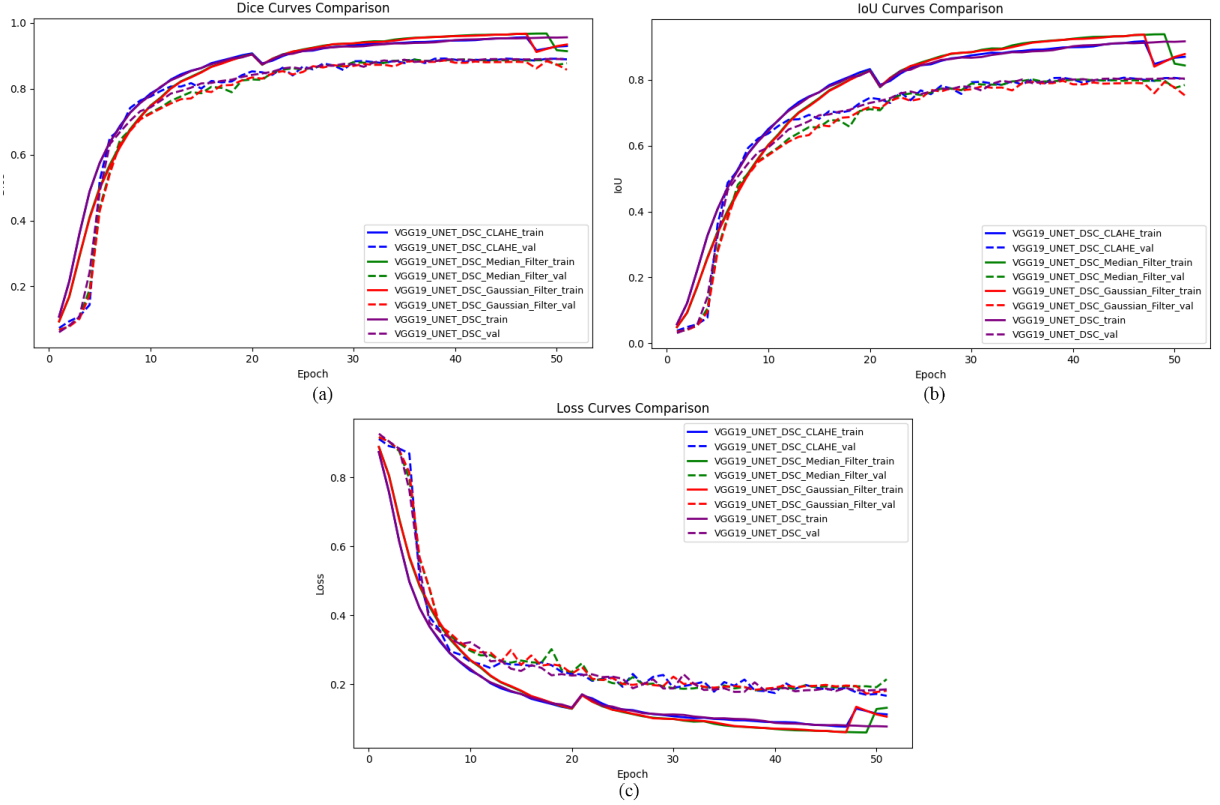


Figure IV.21: Segmentation performance curves under different preprocessing techniques: (a) Dice, (b) IoU, and (c) The loss.

Figure IV.22 presents the segmentation results obtained after applying local contrast enhancement using CLAHE. The first column shows the original MRI image; the second column displays the image after CLAHE processing, followed by the ground truth mask, the predicted mask generated by the model after input enhancement, and finally the overlay between the predicted mask, the ground truth, and the image. This visualization highlights the effect of CLAHE on the clarity of anatomical structures and tumor boundaries and how this improvement is directly reflected in the segmentation quality.

A comparison between the original MRI images and the CLAHE-processed images shows a noticeable improvement in local contrast within brain tissues, particularly at the transition regions between the tumor and the surrounding healthy tissues. This enhancement facilitates the model's ability to learn tumor boundary features, which is evident from the strong agreement between the ground truth masks and the predicted masks across different cases.

Overall, these results indicate that the use of CLAHE not only improves the visual appearance of the images but also increases their segmentation readiness by making the tumor statistically more distinguishable from the background, enhancing boundary definition, and improving the model's ability to accurately represent both large and small tumor regions.

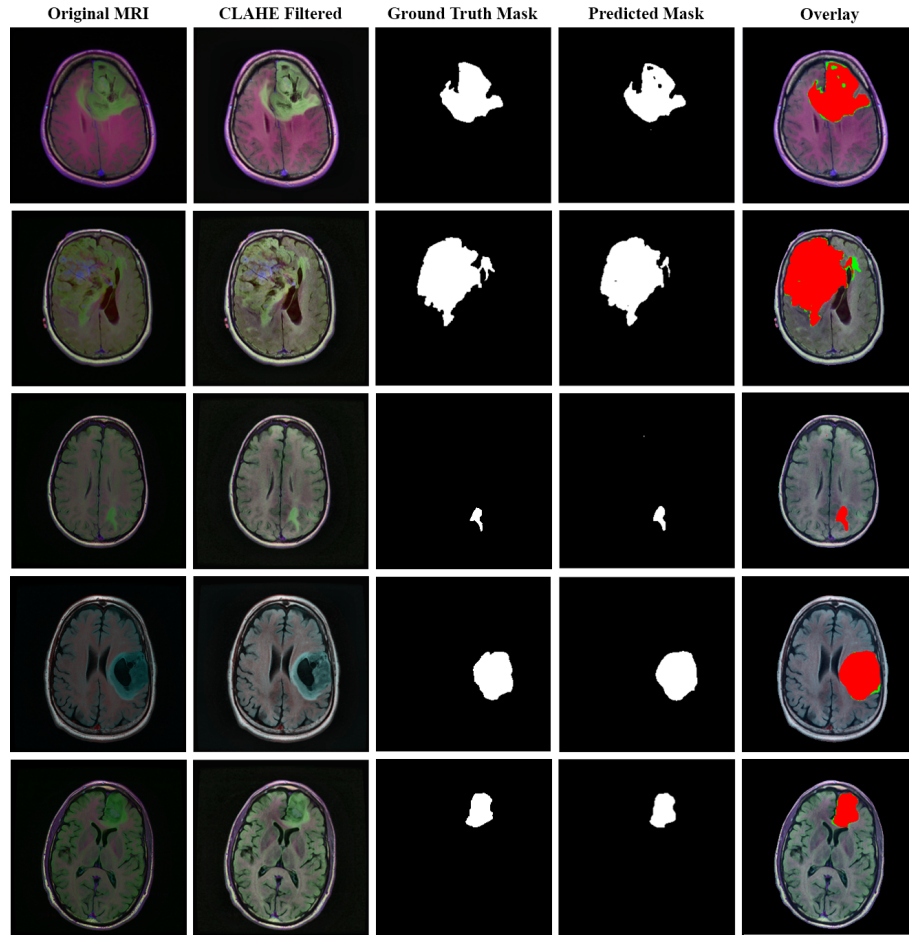


Figure IV.22: FLAIR MRI images after application of the CLAHE.

IV.4.5 Discussion of the results and comparison with previous studies

The results presented in the table IV.5 provide a comprehensive comparison between the proposed approach and several previous studies in the LGG segmentation task, using common performance metrics such as the Dice coefficient, the Intersecting IoU, and accuracy. Overall, this comparison reflects a significant improvement in the performance of modern models, with clear variations in results depending on the model architecture and preprocessing techniques employed.

Analyzing the results of previous studies, we note that Hu et al. and Xie et al. achieved acceptable performance, with dice values of 0.8345 and 0.7585, respectively, but relatively lower IoU values, indicating some shortcomings in tumor boundary accuracy. In contrast, Das et al. achieved a Dice coefficient of 0.9000 and an IoU of 0.8200 using a U-Net-based framework with MRI image preprocessing. Following tumor segmentation, morphological features extracted from the segmented tumors were utilized to predict the genomic subtypes of lower-grade gliomas, highlighting the potential of MRI-derived features for radiogenomic analysis. The study by Ullah et al. also achieved strong results (dice = 0.8900) by employing preprocessing and enhancement techniques, highlighting the crucial role of data quality improvement in enhancing model efficiency.

Upon examining the impact of different preprocessing techniques within the proposed approach, it becomes clear that the type of filter used directly influences the quality

Improving medical semantic segmentation using modified Unet models

of the results. The Gaussian filter performed well (dice = 0.9006, IoU = 0.8197), reducing noise while maintaining relative detail. The median filter, on the other hand, performed slightly worse (dice = 0.8997, IoU = 0.8180), which may be attributed to its smoothing effect, potentially leading to the loss of some fine edges. In contrast, CLAHE achieved the best results (dice = 0.9055, IoU = 0.8278), demonstrating that improving local contrast effectively highlights tumor characteristics and thus enhances segmentation accuracy.

Based on these findings, it can be concluded that improving data quality through preprocessing, along with selecting a robust model architecture, are crucial factors in optimizing the performance of medical segmentation models. The results also confirm that CLAHE technology is one of the most effective methods in this context, given its ability to enhance contrast without negatively affecting fine details, which positively impacts the model's accuracy in identifying tumor boundaries.

Table IV.5: Comparative performance analysis of the proposed modified VGG19-UNet model against advanced methods

References	Dataset	Methodology	dice	IoU
Hu et al [253]	Kaggle LGG	A flexible Transformer model incorporating position-sensitive and information-sensitive attention mechanisms.	0.8179	0.7187
Xie et al [254]	BraTS2020	Brain tumor segmentation using GANs and topological analysis to improve and explain results. Tested on BraTS with improved IoU and dice.	0.7585	0.6373
Das et al [255]	Kaggle LGG	U-Net-based brain tumor segmentation followed by genomic subtype prediction using extracted tumor morphology features	0.9000	0.8200
Ullah et al [256]	Kaggle LGG	Study using preprocessing, augmentation, and model training on Google Colab; highlights radiogenomics linking imaging and genomics for personalized treatment.	0.8900	-
Naser et al [257]	Kaggle LGG	U-Net for LGG tumor segmentation, followed by VGG16 transfer learning	0.8400	-
Proposed approach	Kaggle LGG	The proposed VGG19-Unet was evaluated using the original images and images generated by separately applying Gaussian filtering, Median filtering, and CLAHE. In each case, the preprocessed images were merged with the corresponding original images before being fed into the model.	0.9000	0.8270
	Kaggle LGG + Gaussian		0.9006	0.8197
	Kaggle LGG + median		0.8997	0.8180
	Kaggle LGG + CLAHE		0.9055	0.8278

IV.5 Conclusion

This chapter concludes with a methodological contribution to the field of semantic segmentation of medical images through the proposal of two advanced models based on a Unet architecture. The design process focused on improving and maintaining the flow of detailed spatial features across the network through two parallel approaches: first, by deepening skip connections using slightly optimized layers, and second, by building a robust encoder using a pre-trained VGG19 with an enhanced decoding path. The proposal also included improving the quality of the input data by incorporating Gaussian and median filters into the original image, as well as using CLAHE technique to enhance image

Improving medical semantic segmentation using modified Unet models

quality and highlight edges and fine details.

A comprehensive evaluation of the performance of the two models resulted in a significant improvement in segmentation accuracy, especially in identifying the boundaries of pathological regions, confirming the effectiveness of the proposed improvements. These improvements not only succeeded in raising performance metrics such as dice and IoU but also provided a practical solution that balances computational complexity and accuracy. The results of this chapter pave the way for future research focused on applying these models to more diverse and complex datasets and exploring more sophisticated fusion mechanisms for enhanced features, with the goal of achieving more reliable aids in automated medical diagnosis.

A HYBRID UNET WITH FGHQA FOR MEDICAL IMAGE SEGMENTATION

Contents

V.1	Introduction	78
V.2	Datasets	78
V.2.1	BraTS2020 Dataset	78
V.2.2	KiTS19 Dataset	79
V.2.3	LiTS17 Dataset	80
V.3	Preprocessing data	81
V.3.1	BraTS preprocessing	81
V.3.2	KiTS19 and LiTS17 preprocessing	83
V.4	Proposed FGHQA-Unet Model	84
V.5	Evaluation Metrics	87
V.5.1	Training and Implementation Settings	88
V.6	Results and discussion	88
V.6.1	Experiment 01	88
V.6.1.1	Results	88
V.6.1.2	Discussion	92
V.6.2	Experiment 02	93
V.6.2.1	Results	93
V.6.2.2	Discussion	96
V.6.3	Experiment 03	98
V.6.3.1	Results	98
V.6.3.2	Discussion	101
V.7	Conclusion	102

V.1 Introduction

The CNNs and Transformers have recently emerged as integrated frameworks in the field of computer vision. CNNs are distinguished by their strong ability to extract local features, while Transformers excel at capturing global contextual information. In this chapter, we introduce an improved architecture based on a Unet model with a hybrid encoder, incorporating a Feature Global Extraction Module and a Hybrid Query Attention Module. The chapter provides a detailed discussion of the model’s architectural design, which aims to overcome the limitations of conventional CNN architectures by integrating local feature extraction with global contextual modeling, along with a tumor-guided slice selection strategy to enhance efficiency. Furthermore, the proposed model is evaluated experimentally on the BraTS, KiTS, and LiTS datasets, and the results will be compared with advanced methods to demonstrate the superiority of the proposed approach in terms of both accuracy and efficiency.

V.2 Datasets

In the pursuit of constructing an accurate intelligent model for medical image segmentation, the selection and preparation of data are of paramount importance, as they exert a direct influence on the model’s performance and reliability. Consequently, in this study, a meticulous selection of diverse and exhaustive medical datasets was undertaken, in accordance with the most stringent standards of quality and research ethics. The most popular and widely used datasets by researchers were selected for analysis. Prior to utilization for training and evaluation of the proposed model, all datasets underwent rigorous preprocessing to enhance image quality and eradicate potential biases.

V.2.1 BraTS2020 Dataset

Brain tumors, particularly gliomas, present a significant clinical and research challenge due to their Infiltrative behavior and variable histological grades. Accurate segmentation of different tumor regions from MRI is important of surgical planning and assessment of tumor response to treatment, but the manual process is expensive, time-consuming, and subject to significant inter-observer variability.

To overcome these limitations, the Brain Tumor Segmentation (BraTS) initiative launched an annual challenge to provide a large, high-quality benchmark dataset for the development of automated segmentation algorithms. The BraTS 2020 dataset included data collected from over 30 leading medical and research institutions worldwide, encompassing diverse imaging devices and acquisition protocols. Figure V.1 illustrates representative samples from the BraTS 2020 dataset.

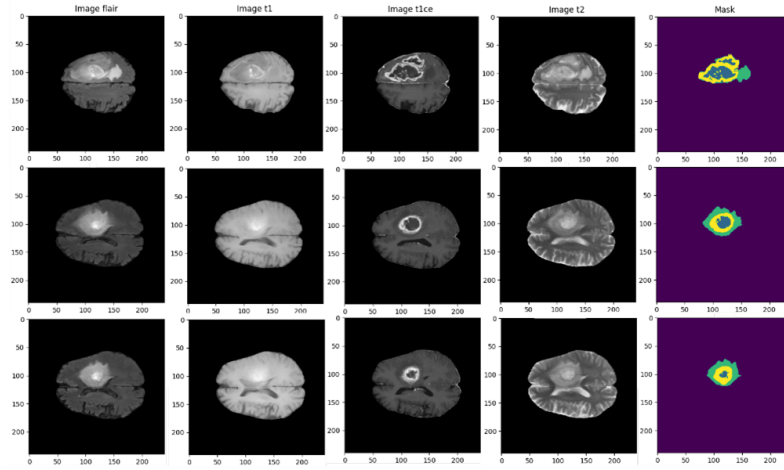


Figure V.1: BraTS2020 data samples: Display of brain imaging using different MRI techniques (FLAIR, T1, T1ce, T2) with segmentation masks (Mask).

V.2.2 KiTS19 Dataset

Figure V.2 shows sample CT images from the KiTS dataset. The KiTS dataset is a public benchmark and a widely recognized gold standard for kidney tumor segmentation in medical AI. It was developed to address the need for a comprehensive collection of CT scans with precise manual annotations of the kidneys and tumors, accompanied by rich clinical data. This resource supports the development of machine learning algorithms for automated segmentation and enables research into links between tumor imaging features and patient outcomes.

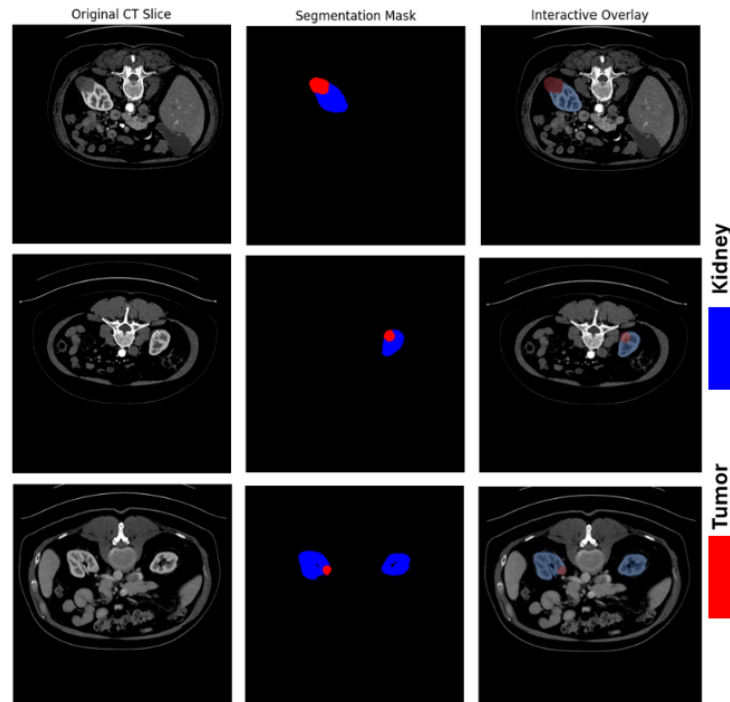


Figure V.2: CT data sample demonstration: normal slices and segmentation masks for kidneys and tumors.

A Hybrid Unet with FGHQA for Medical Image Segmentation

The dataset was built in response to the growing global health burden of Renal Cell Carcinoma (RCC) and a shift in treatment towards partial nephrectomy, which requires precise preoperative planning. It consists of 300 retrospective cases from patients who underwent nephrectomy between 2010 and 2018, meeting specific criteria including the availability of preoperative CT scans.

For the research community, the dataset is partitioned into 210 public cases for training and development and 90 withheld cases for objective final evaluation. Each case includes the CT scan, a segmentation file (with labels: background=0, kidney=1, tumor=2), and a JSON file containing the associated clinical data [258, 259].

V.2.3 LiTS17 Dataset

The LiTS dataset has become a widely recognized benchmark in medical AI for the automated segmentation of the liver and liver tumors from CT scans. Figure V.3 presents sample CT slices and their corresponding segmentation masks. The challenge aims to leverage advanced technologies to support clinicians by enabling faster and more accurate diagnoses, improving treatment planning, and ultimately increasing survival rates among liver cancer patients.

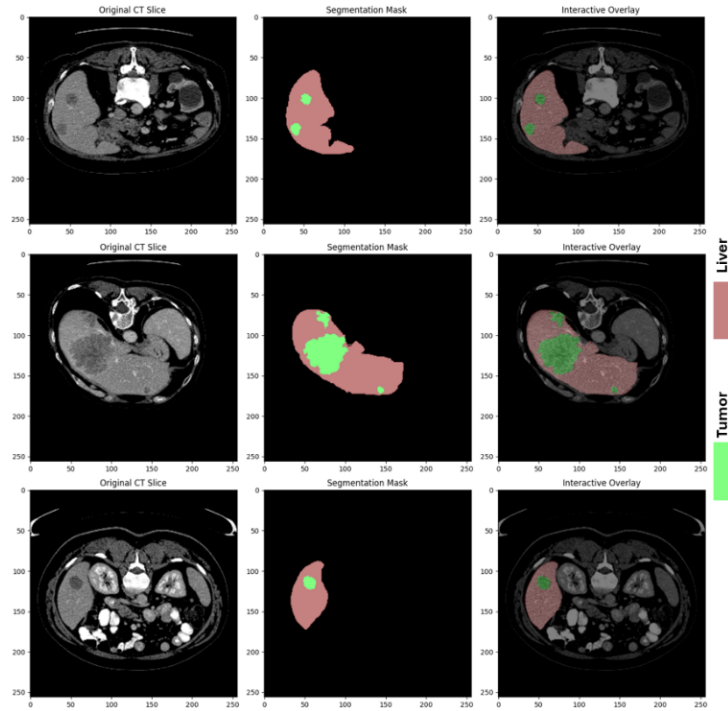


Figure V.3: CT data sample demonstration: normal slices and segmentation masks for Liver and tumors

The LiTS17 challenge provided a large and diverse dataset consisting of 200 abdominal CT images collected from multiple clinical sites. This diversity resulted from the different clinical imaging devices used, which contributed to a more realistic and diverse dataset. The dataset included three-dimensional images with significant variations in spatial resolution, both in slice thickness and in plane resolution, and significant variations in the number of slices per scan. The images also included a wide range of tumor characteristics in terms of size, shape, location, and appearance, making them suitable for developing robust and generalizable algorithms in real clinical settings.

The dataset was split into 130 training cases, which were provided with expert-annotated ground truth segmentations, and 70 test cases, where the ground truth was withheld to allow for an objective and unbiased evaluation of the competing algorithms [260–262].

V.3 Preprocessing data

Preprocessing is a fundamental step in medical image analysis, as it improves image quality, reduces redundant information, and prepares the data for efficient model training. Effective preprocessing can enhance feature representation, reduce computational complexity, and improve the overall performance of deep learning models. In this study, a preprocessing pipeline was designed to prepare the raw images before training the proposed model.

V.3.1 BraTS preprocessing

1. Crop volumes:

Before performing the cropping process, multiple imaging modes, including T1, T2, FLAIR, and T1ce, were first combined to maintain consistency and spatial information between different layers. Because the frame dimensions were large and important information was concentrated in the center of the image with large black areas in the background, the original images and masks were cropped to remove unnecessary artifacts, as shown in Figure V.4. This process speeds up model training by focusing on important brain features and reduces computational resource consumption during training.

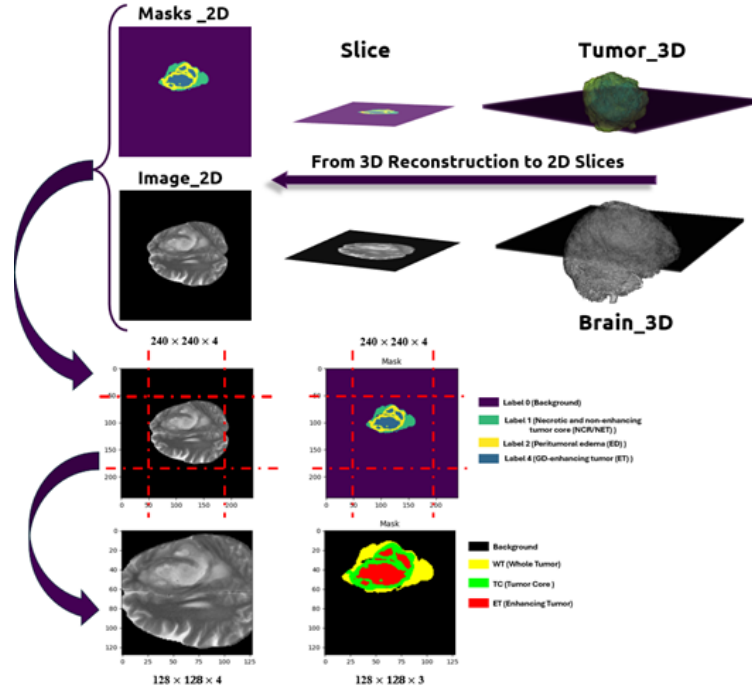


Figure V.4: The proposed pretreatment of magnetic resonance images. Cutting sizes (above) original size: 240×240 ; after cutting: 128×128 . Multimedia photo collection (below), including T1, T1ce, T2, Flair

2. Number of slices:

In many studies dealing with the analysis of three-dimensional medical images, the data are treated as two-dimensional slices to facilitate the study process. However, traditional methods are often based on a constant or random division of segments, which can lead to neglect of areas of clinical importance or increase the computational load without obvious benefit. The researchers additionally examined how the quantity of input slides affects the performance of the model, where the results showed that the Brats data set obtained almost constant performance from five slides and above. In our work, we applied a thoughtful methodology for selecting slices based on such specific criteria as the number of pixels and tumor density as a practical and effective solution. This methodology has several advantages, including increasing diagnostic accuracy by focusing on the most important areas in the image, such as high-density tumors, which enhances the analytical ability of the model to detect pathological features with higher accuracy and without computational burden.

3. Normalization:

To maintain the uniformity of MRI data and enhance the analytical quality, we implement the normalization procedure through min-max normalization. This step addresses the pixel intensity variation between different images caused by different device settings or inspection conditions. min-max normalization rescales pixel values to a fixed range, typically $[0, 1]$. This approach preserves the data's original distribution while making it more suitable for computational processing, particularly for algorithms sensitive to input scale.

$$I_{norm}(x, y) = \frac{I(x, y) - I_{\min}}{I_{\max} - I_{\min}} \quad (\text{V.1})$$

where:

- $I(x, y)$ Original pixel intensity at coordinate (x, y)
- $I_{\min}$ Minimum pixel intensity in the image
- $I_{\max}$ Maximum pixel intensity in the image
- $I_{norm}(x, y)$ Normalized pixel value (scaled to $[0, 1]$)

After loading and normalization, we combine all the images into a multimodal image collection, which can be easily used as model inputs. This procedure synchronizes and processes data from several imaging modalities. Changing between several kinds is the primary method. MRI data is gathered using a variety of modalities, including FLAIR, T1, T2, and T1ce. Transforming and syncing the data across these image types is part of a preprocessing technique known as "intermodal interleaving," which ensures that information from the various modalities can be used to enhance model performance.

4. New labels:

The data has 5 labels, but we don't use labels 0 (background) and 3 (missing). The other labels, 1 (Necrotic and non-enhancing tumor core), 2 (Peritumoral edema), and 4 (GD-enhancing tumor), show different parts of the tumor. By mixing these labels, we can make different tumor areas: ET (Enhancing Tumor) is label 4, TC

(Tumor Core) is labels 1 and 4 together, and WT (Whole Tumor) is labels 1, 2, and 4 combined. We will test how well these new labels (ET, TC, WT) work as shown in Figure V.4.

V.3.2 KiTS19 and LiTS17 preprocessing

Preprocessing the original CT images before feeding them into the network is a main step in training AI-based models. This study aims to improve image quality and thus enhance the segmentation accuracy of kidneys, kidney tumors, liver, and liver tumors in abdominal CT images without requiring tumor type identification. To prepare the data as input for the model, the 3D data was segmented into multiple 2D images with a resolution of 512×512 pixels and then resized to 256×256 pixels, as shown in Figure V.5.

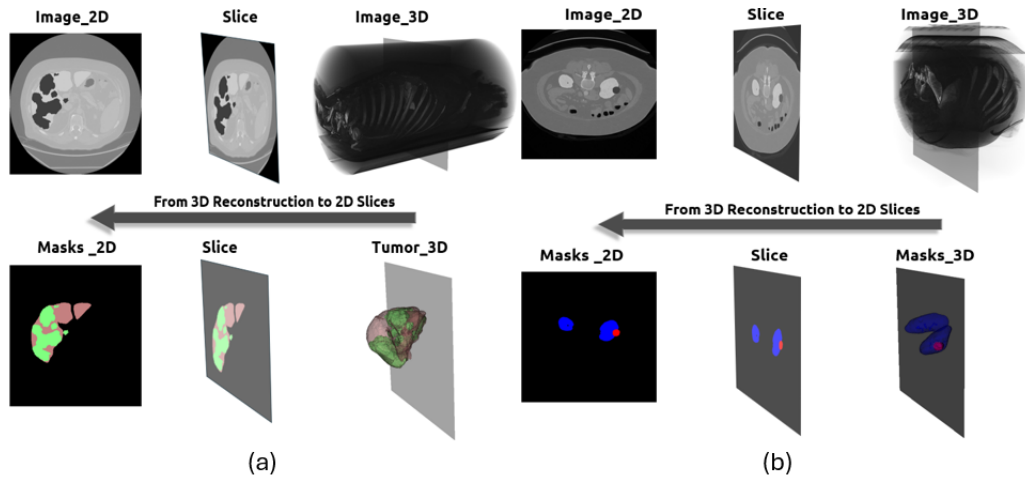


Figure V.5: A Schematic of the Volumetric to 2D Slice Conversion Pipeline for Medical Image Analysis: (a) for the LiTS dataset, (b) for the KiTS dataset

In the subsequent stage of the preprocessing, the image intensity values were reduced for all images within the Hounsfield unit (HU) range of -30 to 300 HU for segmenting the kidneys and renal tumors. Figure V.6 (a) illustrates the difference between the original data and the data within the specified HU range. This facilitated the elimination of the fatty regions surrounding the kidneys from the segmentation process, which are characterized by values less than -30 HU [258]. Moreover, the higher threshold was selected to accentuate the high-density structures and borders that are generally characteristic of cancerous tissue. Subsequently, To ensure data consistency and enhance analysis quality, a normalization process is implemented using min-max normalization.

The range of -50 to 250 HU is considered optimal and most comprehensive for segmenting the liver and its tumors in CT [263]. This range is characterized by its ability to cover all histological patterns (Refer to Figure V.6(b), which illustrates the difference between the original data and the data within the specified HU range.), encompassing normal liver tissue ranging from 50-70 HU (Table V.1) in non-contrast-enhanced scans that is, before the injection of an iodine-based contrast agent and between 100-150 HU after injection. It also covers the diverse characteristics of liver tumors, from arterial enhancement that may exceed 200 HU to late-stage contrast washout that may be as low as 30-80 HU. This wide range ensures that important data are preserved despite individual differences between patients or scan conditions, making it a reliable standard for both clinical applications and artificial intelligence algorithms. This step addresses the

A Hybrid Unet with FGHQA for Medical Image Segmentation

issue of pixel intensity differences between different images resulting from varying device settings or scanning conditions.

Table V.1: distribution of CT value ranges across primary tissues in abdominal CT scans

Organs and Tissues	CT Value (HU)
Liver	50–70
Spleen	35–60
Pancreas	30–55
Kidney	25–50
Bone	>400

Recent studies, including the study of Vu et al. [232], have shown that carefully choosing the number of slices in medical image analysis is a great way to improve both diagnostic accuracy and efficiency in computing. The findings demonstrated that the performance of segmentation for kidney and liver tumors stabilizes significantly when utilizing between five and seven slices, the optimal range adopted in this study.

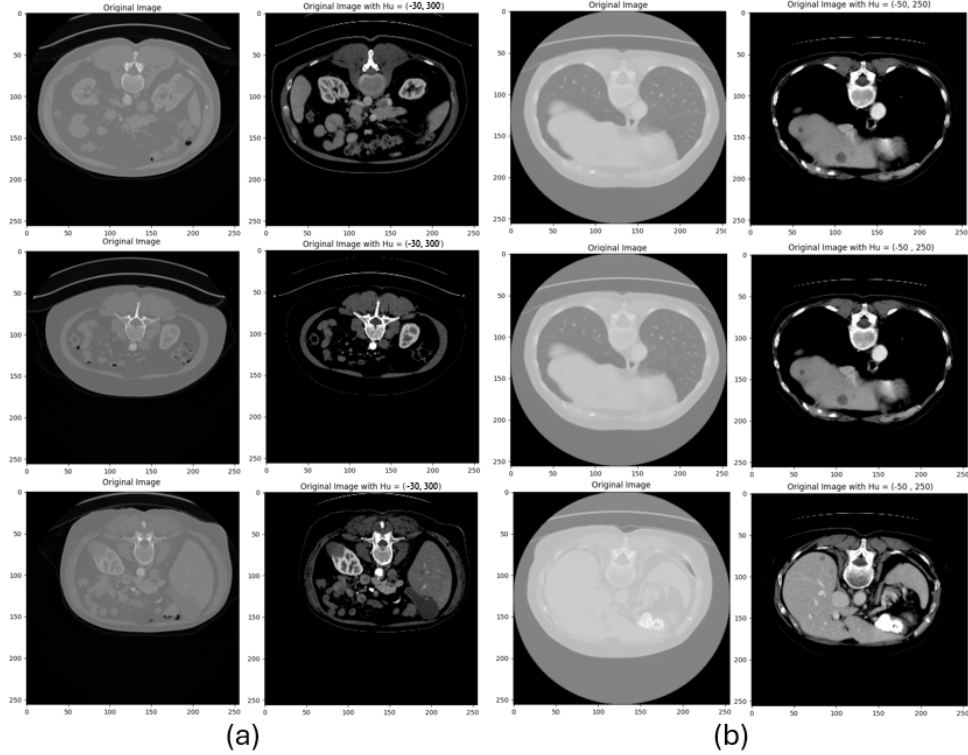


Figure V.6: (a) represents the enhanced images of CT slices of the kidney, (b) represents the enhanced images of the liver using different HU window ranges

This slice selection strategy reduces the computational burden while maintaining the quality of the analysis by focusing on the most informative slices. This guarantees the most effective use of resources while maintaining the precision of the outcomes.

V.4 Proposed FGHQA-Unet Model

Medical image segmentation is a critical process in helping diagnose and treat, but it faces many challenges that make it a complex task. One of the most significant challenges is

A Hybrid Unet with FGHQA for Medical Image Segmentation

the low contrast between tissues, where the intensity of colors or signals in some organs or tumors is similar, making it difficult to accurately define their boundaries. Despite the great success of CNN in extracting local features, it suffers from a lack of understanding of the overall context of the image.

One of the main reasons is the limited receptive field, where each convolutional layer focuses on only a small area, hindering the understanding of relationships between distant parts of the image. Pooling operations, such as max pooling, also lead to the loss of precise location information, weakening the ability to connect distant areas. Transformers came as a revolutionary solution that leverages self-attention mechanisms to bridge this gap, providing a more comprehensive understanding of medical images by leveraging natural language processing (NLP) techniques and applying them to the field of computer vision.

In this work, we propose an improved architecture based on the Unet model (Figure V.7) by designing a hybrid encoder that embeds attention mechanisms inspired by ViT [194]. The input image is transformed into a structured token-based representation, enabling the extraction of rich semantic and contextual features through multi-head attention layers.

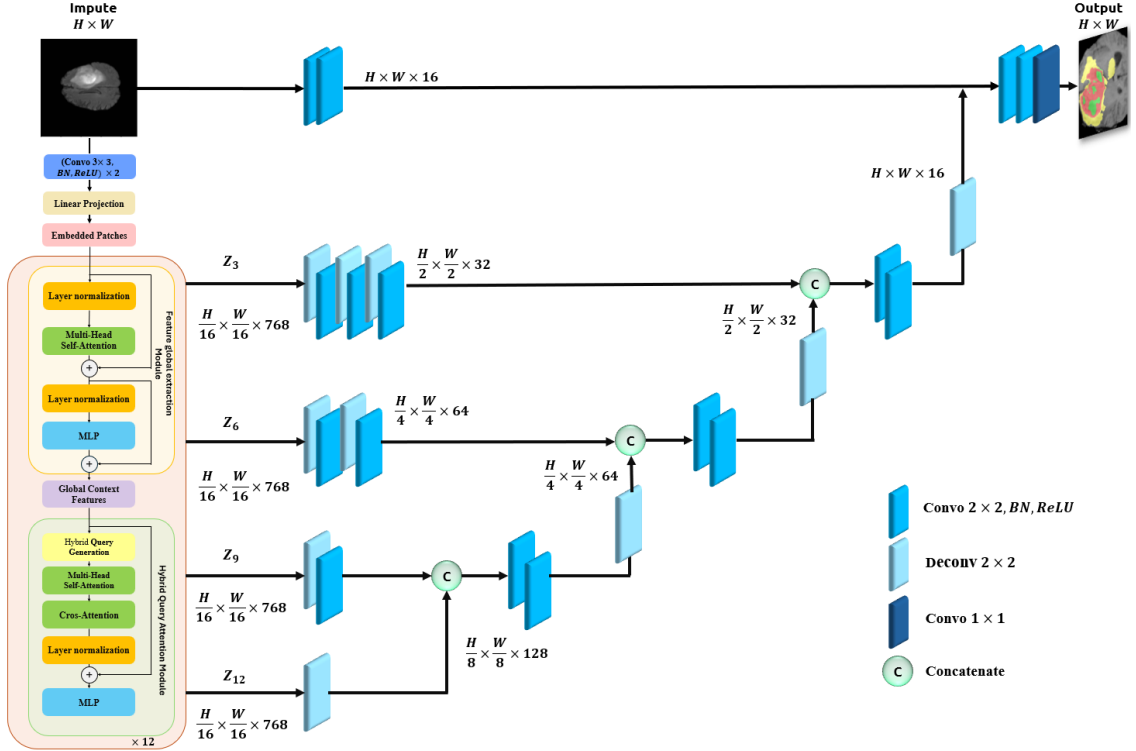


Figure V.7: An overview of the FGHQA-Unet architecture is presented. We derive sequence representations from various layers within the transformer and integrate them with the decoder through skip connections.

A novel hybrid query generation strategy is adopted, combining dynamically generated queries derived from image features with randomly initialized learnable queries. The learnable queries follow the query initialization principle introduced in the DETR architecture [264], where they act as generic queries that gradually evolve their semantic meaning during training.

In addition, a cross-attention mechanism is integrated within the encoder to enable

a guided interaction in which the queries attend to the feature maps. These queries are used not only to explore the most relevant information but also to progressively refine the representations before being passed to the Unet decoder. This design enhances the quality of features fed into the reconstruction stages, thereby improving segmentation accuracy, particularly in scenarios requiring fine-grained contextual understanding, such as medical imaging.

At the same time, the approach maintains high computational efficiency by relying on a simplified architecture without employing a full Transformer model. It also provides better interpretability by associating queries with specific semantic regions within the image.

a) Encoder

The input 2D image $x \in \mathbb{R}^{H \times W \times C}$, with resolution $H \times W$, where C denotes the number of channels, is first processed by convolutional layers to extract local feature representations while preserving fine-grained spatial details. The resulting feature maps are then divided into a sequence of fixed-size patches of dimension $P \times P$. The total number of patches is given by $N = \frac{H \times W}{P^2}$. These patches are transformed into token representations through a linear projection that maps each patch into a fixed-dimensional embedding. A learnable positional embedding is added to each token to preserve spatial information before being fed into the Transformer encoder.

Therefore, the image is represented as a sequence of tokens that encode visual and spatial information, enabling the model to preserve spatial information during processing. After the embedding layer, a stack of attention layers is employed, consisting of Multi-Head Self-Attention (MSA) sublayers and Multi-Layer Perceptron (MLP) blocks, as shown in Figure V.8. These components model global contextual dependencies across different image regions, enriching the feature representations with semantic and contextual information. Building upon the locally extracted features, this module constructs a structured semantic feature space that forms the basis for subsequent inference.

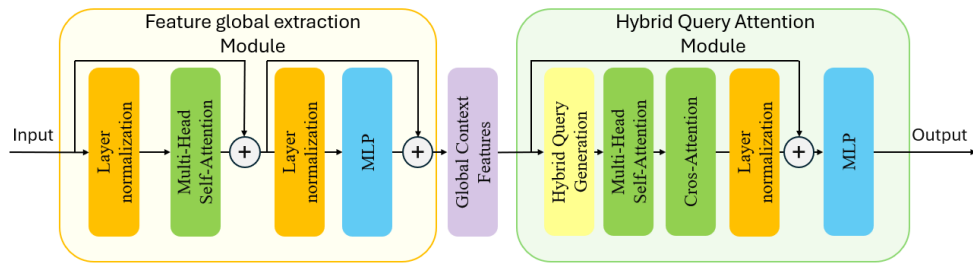


Figure V.8: Proposed FGHQA Hybrid Transformer Encoder Block.

After that, an inference module based on queries is constructed, where the queries are generated using a hybrid strategy. This integration produces a more content-aware initial representation and enhances the semantic alignment between the query space and the extracted features. A clear separation is also introduced between the representation stage and the inference stage. In this phase, the hybrid queries are utilized within a cross-attention mechanism to capture relationships among different entities within the image. Additionally, the encoder supports interaction between the hybrid queries and the extracted features, where the queries are not only used

to interrogate the features but also to progressively refine their own representations, thereby enhancing focus on relevant regions while reducing the influence of noise.

This design achieves a balance between representational and inferential capabilities without adopting a standard Transformer architecture, resulting in higher computational efficiency while maintaining strong performance in fine-grained tasks such as medical image segmentation. After constructing the encoder and completing the image representation stage, the same encoding block is applied sequentially 12 times to progressively enhance the model’s representational capacity and improve the quality of the learned feature representations across layers.

b) Decoder

The decoder in the FGHQA-Unet architecture features a multi-stage hierarchical design that integrates spatial upscaling mechanisms with skip connections to restore the spatial resolution lost during encoding. The process begins by merging multi-level (Z3, Z6, Z9, Z12) feature maps through an integrated mechanism that progressively increases the resolution using 2×2 deconvolution layers while preserving fine details via skip connections that incorporate spatial features from earlier encoding stages. The merged features are further processed through a series of 3×3 convolutional layers with normalization and activation, progressively refining the feature representations at each resolution level. Upon reaching the original resolution, the output is routed through a final $1 \times 1 \times 1$ convolutional layer followed by Softmax to generate accurate pixel-level probability maps, enabling high-precision tumor delineation. The effective integration between the transformer’s ability to capture the global context of tumors and the efficiency of skip connections in preserving fine anatomical details enables the proposed model to achieve accurate and robust tumor segmentation.

V.5 Evaluation Metrics

In this study, the evaluation of the model’s segmentation accuracy was conducted using dice, IoU, and Hausdorff’s 95th-mean-point distance. The definition of the first two coefficients appears and the Loss function in Chapter IV.

HD95 is one of the most widely used metrics for assessing the accuracy of medical image segmentation, especially when comparing the boundaries of a segmented region with the ground truth. This metric is based on the idea of measuring the largest boundary distance while ignoring the largest 5% of the distances to reduce the influence of positional errors or outliers.

$$d_{\text{HD95}}(A, B) = 95^{\text{th}} \text{ percentile of } \left\{ \max \left(\sup_{a \in A} \inf_{b \in B} d(a, b), \sup_{b \in B} \inf_{a \in A} d(a, b) \right) \right\} \quad (\text{V.2})$$

A and B are two sets of boundary points from the segmented output and the ground truth, respectively. $d(a, b)$ denotes the Euclidean distance between a point $a \in A$ and a point $b \in B$. For each point a in A , $\min_{b \in B} \{d(a, b)\}$ computes the minimum distance from a to its nearest point in B , resulting in a list of such distances.

In standard HD, the maximum of these minimum distances is used, making it highly sensitive to outliers. Instead of taking the absolute maximum, HD95 considers the 95th

percentile of the sorted distances. This ignores the worst 5% of outliers, making the metric more reliable in real-world medical imaging scenarios.

V.5.1 Training and Implementation Settings

The proposed model was trained using the dice loss function and the Adam optimizer with an initial learning rate of 1×10^{-4} . A ReduceLROnPlateau schedule was applied with a reduction factor of 0.1, a patience of 3 epochs, and a minimum learning rate of 1×10^{-7} . A batch size of 4 and a maximum of 500 epochs were used. Early stopping was based on the validation loss with a patience of 10 epochs, and the model corresponding to the best validation loss was saved for model selection. A dropout rate of 0.2 was used.

V.6 Results and discussion

To test the performance of our proposed model and evaluate its potential for clinical applications, we conducted three experiments on the most widely used public benchmark datasets for training and evaluating AI-based medical image segmentation models. In the first experiment, the model was tested on the BraTS2020 dataset for brain tumor segmentation, while in the second experiment, it was tested on the KiTS19 dataset for kidney and renal tumor segmentation, and in the third experiment, it was tested on the LiTS17 dataset for liver and hepatic tumor segmentation. The detailed experimental results are presented in the following sections.

V.6.1 Experiment 01

V.6.1.1 Results

To evaluate the performance of the proposed model, we tested it on BraTS2020 brain tumor MRI dataset using different numbers of input slices: 3, 4, 5, 6, and 7 slices were selected based on the number of pixels and tumor density to achieve the best possible coverage of the tumor’s internal structure. Multiple performance metrics were used to evaluate the model, including the Dice, the IoU, and the mean distance between boundaries HD95.

Table V.2 presents the test performance of the proposed model trained separately with 3, 4, 5, 6, and 7 input slices and evaluated using Dice, IoU, HD95, and inference time. The class-wise results indicate that performance varies across the WT, TC, and ET regions, with ET generally representing the more challenging region due to its smaller size and greater heterogeneity. With 3 slices, the model achieved Dice scores of 0.9591, 0.9481, and 0.9102 for WT, TC, and ET, respectively, with corresponding IoU values of 0.9214, 0.9013, and 0.8352. When using 4 slices, the Dice scores were 0.9584, 0.9482, and 0.9151, while the IoU values reached 0.9201, 0.9015, and 0.8436 for WT, TC, and ET, respectively. With 5 slices, the model obtained Dice scores of 0.9595, 0.9561, and 0.9142 and IoU values of 0.9221, 0.9158, and 0.8420 for WT, TC, and ET, respectively. The results obtained with 6 and 7 slices remained generally comparable to those obtained with 5 slices, with only minor variations across the different regions and metrics.

Overall, the effect of the number of input slices does not follow a consistent increasing or decreasing trend. WT and TC maintain relatively high performance across the different configurations, whereas ET exhibits greater variability, particularly in terms of HD95,

A Hybrid Unet with FGHQA for Medical Image Segmentation

which ranged from 2.1164 to 12.7679 mm. These findings suggest that increasing the number of slices beyond 5 does not consistently improve segmentation performance. The inference time also remained below 0.4 s across all configurations, ranging from 0.0898 s with 3 slices to 0.3596 s with 7 slices.

This observation is further supported by the mean and standard deviation analysis in Figure V.9, where the 5-slice configuration achieves the highest mean Dice 0.9152 ± 0.134 , the lowest mean HD95 5.4835 ± 18.5582 mm, and a mean IoU of 0.8626 ± 0.158 , comparable to that obtained with 7 slices. Its standard deviation also indicates relatively lower variability across the three regions. Considering these results together with the increased inference time associated with using more slices, the 5-slice configuration provides the most favorable balance between segmentation performance, consistency, and computational efficiency and was therefore selected for the subsequent experiments.

Table V.2: Brain tumor segmentation test result with different evaluation metrics on BraTS 2020

Slices	Class	Dice	IoU	HD95 [mm]	Inference time (s)
3	WT	0.9591	0.9214	3.2493	0.0898
	TC	0.9481	0.9013	4.8449	
	ET	0.9102	0.8352	12.7679	
4	WT	0.9584	0.9201	4.7057	0.3466
	TC	0.9482	0.9015	3.3375	
	ET	0.9151	0.8436	2.3686	
5	WT	0.9595	0.9221	5.4038	0.3343
	TC	0.9561	0.9158	3.1907	
	ET	0.9142	0.8420	2.8150	
6	WT	0.9611	0.9251	3.8388	0.3573
	TC	0.9564	0.9165	2.6680	
	ET	0.9180	0.8397	2.1164	
7	WT	0.9573	0.9180	3.9023	0.3596
	TC	0.9570	0.9175	3.4978	
	ET	0.9180	0.8471	2.5826	

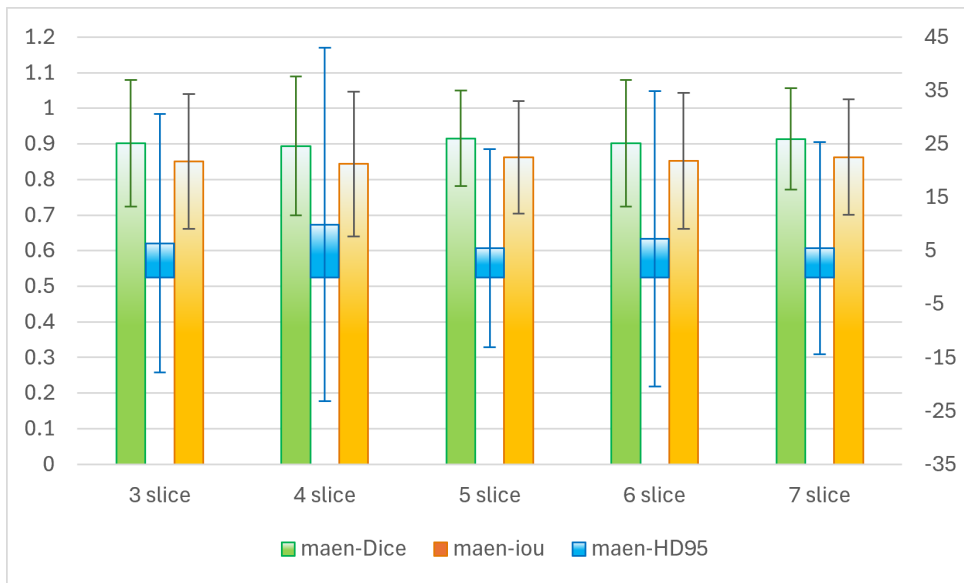


Figure V.9: Mean $\pm$ SD of Segmentation Metrics for Selected Slices

Figure V.10 shows three different brain tumor cases as seen in multiple MRI sequences (FLAIR, T1ce, T1, and T2), along with a comparison between the ground truth mask and the predicted mask generated by the proposed model. Each row of cases shows the original images and the mask overlays to evaluate the model's accuracy in identifying the three tumor regions: WT, TC, and ET. In the first case, a small tumor, the model demonstrates a near-perfect match between the predicted mask and the ground truth across all MRI sequences.

The second case, a medium-sized tumor, also demonstrates high accuracy in predicting all regions without significant deviations. The third case, a large and complex tumor, demonstrates that the model's performance remains satisfactory but is affected by the extent of the edema and the complexity of the shape, with some discrepancies in the definition of the outer edges. The figure indicates that the proposed model successfully identified the tumor boundaries with high accuracy, with a clear match between the actual and predicted masks.

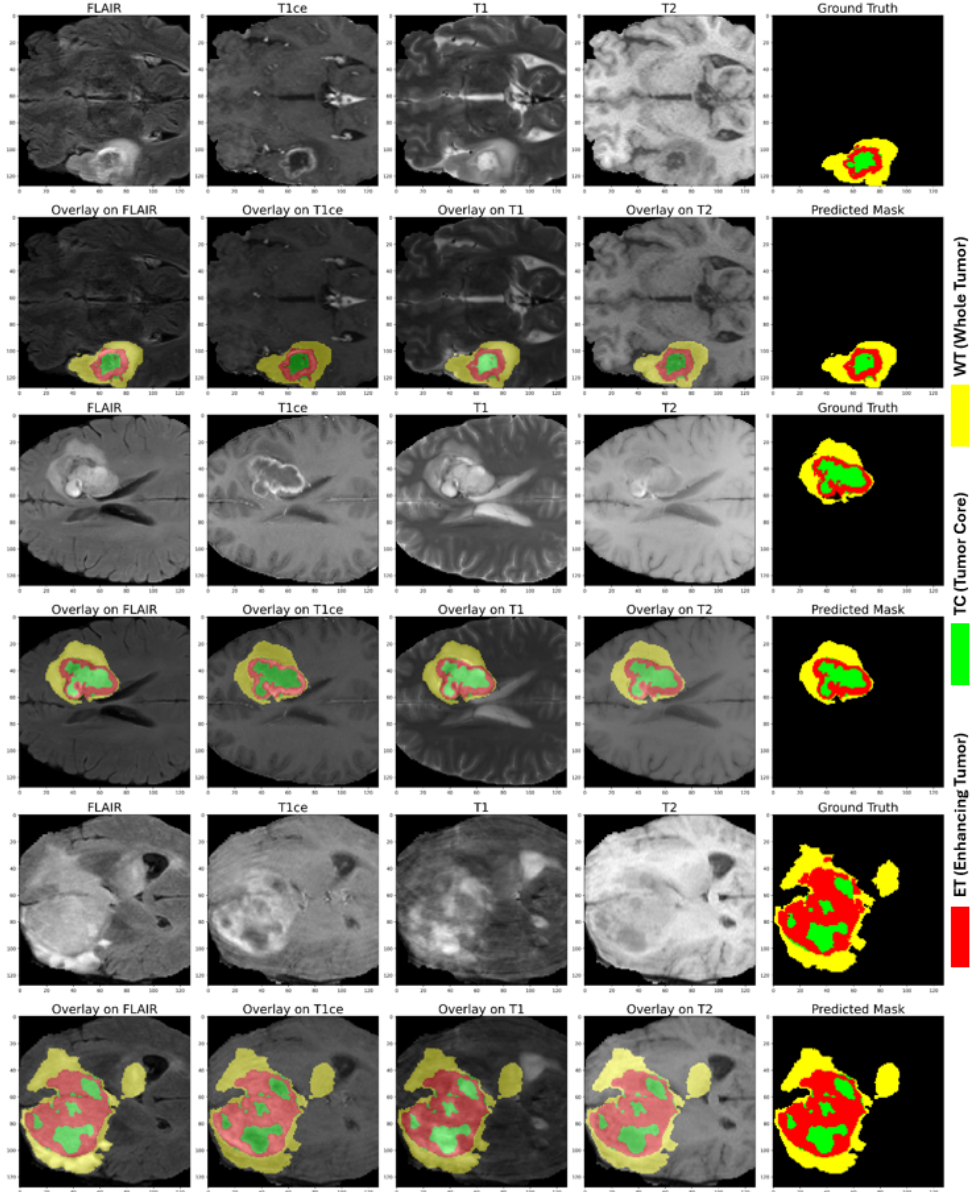


Figure V.10: Visual comparison of multi-channel radiomics and model predictions in brain tumor segmentation

Figure V.11.a shows the evolution of the Dice scores during training and validation for the WT, TC, and ET tumor regions across the epochs. A rapid improvement in performance is observed during the early epochs, with the Dice scores increasing substantially during the first 10–15 epochs. After this initial improvement, the curves exhibit some fluctuations while continuing to improve gradually, followed by a relatively stable performance from approximately the 30th epoch onward, with the Dice scores reaching values above 0.90 for all tumor regions. The training and validation curves remain relatively close throughout the training process, with no clear signs of overfitting. These results indicate that the model achieves stable and consistent segmentation performance across the different tumor regions.

The HD95, shown in Figure V.11.b, shows a sharp decline in values during the early epochs of training (around the first ten epochs), reflecting the rapid improvement in the model's ability to accurately localize the tumor boundaries for the three regions and reduce the large boundary errors present during the early stages of training. After approximately the twentieth epoch, the values continue to decrease with some fluctuations before stabilizing at relatively low levels from around the thirtieth epoch onward, indicating that the model has achieved accurate boundary localization. The training and validation curves are also relatively close, indicating the stability of the learning process and no clear signs of overfitting. These results demonstrate the model's ability to produce accurate segmentation maps with low boundary error.

The IoU coefficient, shown in Figure V.11.c, also follows a pattern similar to the Dice coefficient, showing rapid improvement during the first epochs of training before gradually reaching a plateau after about epoch 30, exceeding 0.85 for the three tumor regions. A clear convergence between the training and validation curves is also observed, reflecting the stability of the learning process and the model's efficiency in achieving accurate and consistent segmentation without clear signs of overfitting.

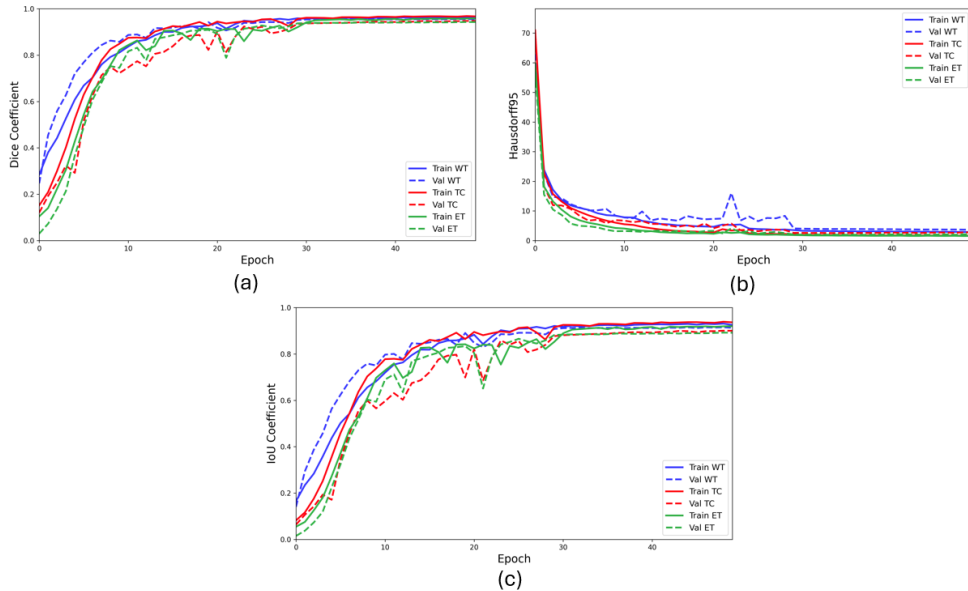


Figure V.11: Graphs of segmentation performance metrics for different tumor regions: (a) dice coefficient, (b) Hausdorff distance 95, and (c) Intersection over union

V.6.1.2 Discussion

The evaluation results showed that FGHQA-UNet performed strongly across most of the considered evaluation metrics (Table V.3) and exhibited stable behavior during both the training and validation phases, with no clear indication of overfitting, as illustrated by the curves in Figure V.11. The comparison with the other models was conducted using the same dataset (BraTS2020) and identical training, validation, and testing splits to ensure a fair comparison. In particular, the model achieved high segmentation performance in terms of the Dice coefficient.

The dice value reached 0.9635 for the WT class, 0.9589 for the TC class, and 0.9195 for the ET class, results that clearly outperform several advanced models such as CFCI-Net, AXUnet, MWG-Unet++, PU-NET, InputSparseMultiOCM, and MUnet. Despite the difficulty of classifying the ET class, the proposed model still outperformed MWG-Unet++ 0.9195 vs. 0.8864. Regarding the HD95 index, which measures the spatial accuracy of defining tumor boundaries, FGHQA-UNet recorded varying values: 5.4038 for the WT class, 3.1907 for the TC class, and 2.8150 for the ET class. Although the WT value is higher than some other models, such as MUnet 3.755, the model outperformed in the TC and ET categories, demonstrating its superior ability to accurately define the internal boundaries of the tumor. It's worth noting that the lack of HD95 data in some models, such as AXUnet and MWG-Unet++, makes comprehensive comparisons difficult, but the available results indicate that FGHQA-UNet outperforms most models in this regard.

Table V.3: The segmentation results for our approach and the methods we compared

Method	dice			HD95 (mm)		
	WT	TC	ET	WT	TC	ET
CFCI-Net [265]	0.9076	0.8404	0.7956	4.377	6.252	20.695
AXUnet [266]	0.9259	0.8681	0.8489	-	-	-
PU-NET [267]	0.905	0.827	0.803	4.638	8.679	3.573
InputSparseMultiOCM [268]	0.866	0.747	0.748	10.1868	11.834	6.142
MWG-Unet++ [269]	0.9210	0.8820	0.8864	-	-	-
MUnet [270]	0.9150	0.8230	0.8350	3.755	6.437	2.421
FGHQA-UNet (Proposed)	0.9595	0.9561	0.9142	5.4038	3.1907	2.8150

The outstanding performance can be attributed to the proposed hybrid encoder within the U-Net architecture, where the model leverages the capability of the FGHQA module to effectively model fine-grained contextual relationships within medical images. This integration contributes to improved segmentation quality, particularly in cases characterized by significant variability in tumor size and shape. However, there remains potential to further enhance the model's ability to represent fine structures with low contrast within the tumor, especially small and marginal regions where boundary delineation is challenging. Such improvements could increase the model's robustness and stability in complex segmentation scenarios, thereby strengthening its suitability for deployment in clinical applications.

Table V.4 presents a fair comparison between FGHQA-UNet and the baseline models under identical training and testing conditions. Since the proposed methodology converts the 3D volumes into 2D slices, five slices were selected from each volume based on the slice-selection experiment described in the methodology. The same five selected slices

A Hybrid Unet with FGHQA for Medical Image Segmentation

Table V.4: Comparison of different models in terms of segmentation performance and computational complexity.

Model	Class	Dice	IoU	HD95 (mm)	Time (s)	GFLOPs	Params (M)
Unet	WT	0.9064	0.8289	8.5969	0.0388	24.1187	31.05
	TC	0.8303	0.7098	11.0334			
	ET	0.8081	0.6780	18.5715			
nnUnet	WT	0.9321	0.8728	6.5032	0.0423	6.5698	12.41
	TC	0.8598	0.7540	9.3847			
	ET	0.8577	0.7508	15.9058			
TransUnet	WT	0.9225	0.8562	9.4903	0.0652	93.98	93.98
	TC	0.8266	0.7045	17.1648			
	ET	0.8363	0.7186	19.3952			
FGHQA-Unet	WT	0.9595	0.9221	5.4038	0.3343	1.4933	27.37
	TC	0.9561	0.9158	3.1907			
	ET	0.9142	0.8420	2.8150			

were used to train and test U-Net, nnU-Net, and TransUNet, ensuring a consistent input setting for all models. FGHQA-Unet achieved the highest Dice for WT, TC, and ET, with values of 0.9595, 0.9561, and 0.9142, respectively. It also obtained the highest IoU values for the three regions, reaching 0.9221, 0.9158, and 0.8420 for WT, TC, and ET, respectively.

Regarding boundary accuracy, FGHQA-Unet achieved the lowest HD95 values for all three regions, with 5.4038 mm for WT, 3.1907 mm for TC, and 2.8150 mm for ET. In terms of computational complexity, FGHQA-Unet required only 1.4933 GFLOPs, compared with 24.1187 GFLOPs for U-Net, 6.5698 GFLOPs for nnU-Net, and 93.98 GFLOPs for TransUNet. However, its inference time was 0.3343, which was higher than U-Net 0.0388 s, nnU-Net 0.0423 s, and TransUNet 0.0652 s. Nevertheless, this inference time remained below half a second, indicating that the proposed model maintains an acceptable inference time despite its higher computational time than the baseline models. Regarding the number of parameters, FGHQA-Unet contained 27.37 M parameters, which was lower than U-Net 31.05 M and TransUNet 93.98 M, but higher than nnU-Net 12.41 M. Overall, FGHQA-Unet maintained a favorable balance between segmentation accuracy, computational efficiency, and inference time under the same experimental conditions.

V.6.2 Experiment 02

V.6.2.1 Results

The model demonstrated good and relatively stable performance in the kidney and tumor segmentation tasks across the different slice configurations, as indicated by the Dice, IoU, and HD95 values in TableV.5. For the 5-slice configuration, the model achieved a Dice of 0.9403 and IoU of 0.8873 for the kidney, with an HD95 of 6.50026 mm, while the tumor achieved a Dice of 0.9218, IoU of 0.8549, and HD95 of 31.1485 mm. With 6 slices, performance improved overall, with the kidney reaching a Dice of 0.9445 and IoU of 0.8948, and the tumor reaching a Dice of 0.9326 and IoU of 0.8737. The HD95 also improved substantially for the tumor, decreasing to 17.6866 mm, while the kidney HD95 increased slightly to 8.1024 mm. With 7 slices, the kidney performance remained high, with a Dice of 0.9411 and IoU of 0.8888, whereas the tumor Dice and IoU decreased to 0.9043 and

A Hybrid Unet with FGHQA for Medical Image Segmentation

0.8253, respectively, despite an improvement in tumor HD95 to 28.12 mm. The 8-slice and 9-slice configurations did not provide a consistent improvement over 6 slices, with slight fluctuations across the different metrics. This overall trend is also reflected in the mean segmentation analysis shown in Figure V.12, where the 6-slice configuration provides the highest overall mean Dice 0.8736 ± 0.2146 , IoU 0.8191 ± 0.233 and the lowest mean HD95 16.6702 ± 55.6139 mm., together with comparatively lower variability across the kidney and tumor classes. In terms of inference time, the configurations show only small differences, ranging from 0.0881 to 0.0981 s, indicating that the performance improvement obtained with 6 slices is achieved without a substantial increase in inference time. Therefore, considering the class-wise results, the mean $\pm$ standard deviation analysis, and the inference time, the 6-slice configuration provides the most favorable overall balance.

Table V.5: Kidney and tumor segmentation result with different evaluation metrics on KiTS19

Slices	Class	dice	IoU	HD95 [mm]	Inference time (s)
5	Kidney	0.9403	0.8873	6.5002	0.0981
	Tumor	0.9218	0.8549	31.1485	
6	Kidney	0.9445	0.8948	8.1024	0.0975
	Tumor	0.9326	0.8737	17.6866	
7	Kidney	0.9411	0.8888	10.8416	0.0950
	Tumor	0.9043	0.8253	28.1289	
8	Kidney	0.9343	0.8767	12.9415	0.0881
	Tumor	0.9217	0.8548	31.8296	
9	Kidney	0.9382	0.8836	8.4427	0.0906
	Tumor	0.9300	0.8691	33.2719	

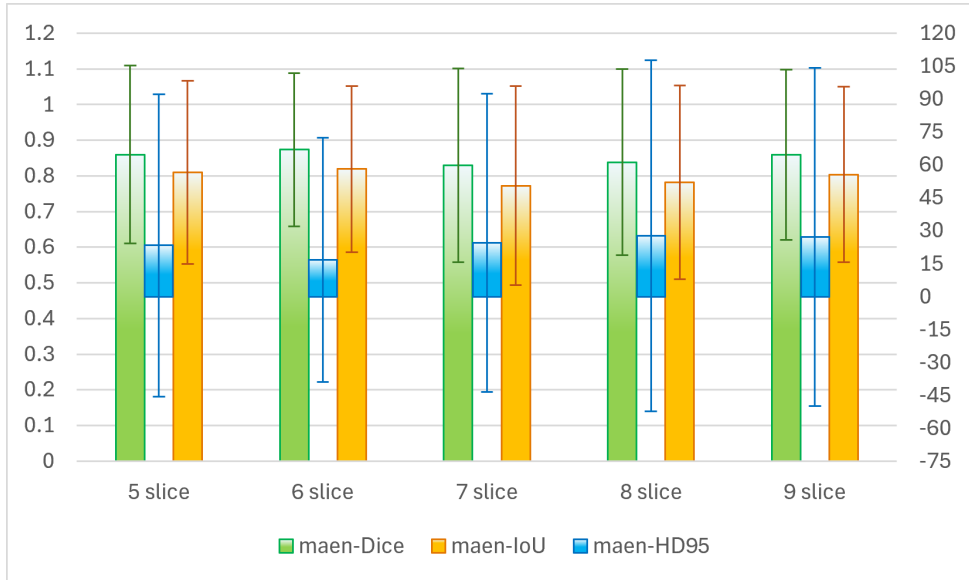


Figure V.12: Mean $\pm$ SD of Segmentation Metrics for Selected Slices

Figure V.13 shows a set of kidney CT test images used to evaluate the model's performance in segmenting both the kidney and its tumors. Each row consists of four columns: the original image, the ground truth mask, the predicted mask, and a overlay

image showing the predicted mask overlaid on the original image. Comparing the ground truth and predicted masks, the model accurately localizes the kidney, while slight variations can be observed in the segmentation of some tumor regions. Overall, the figure demonstrates the model’s strong performance in kidney and tumor segmentation.

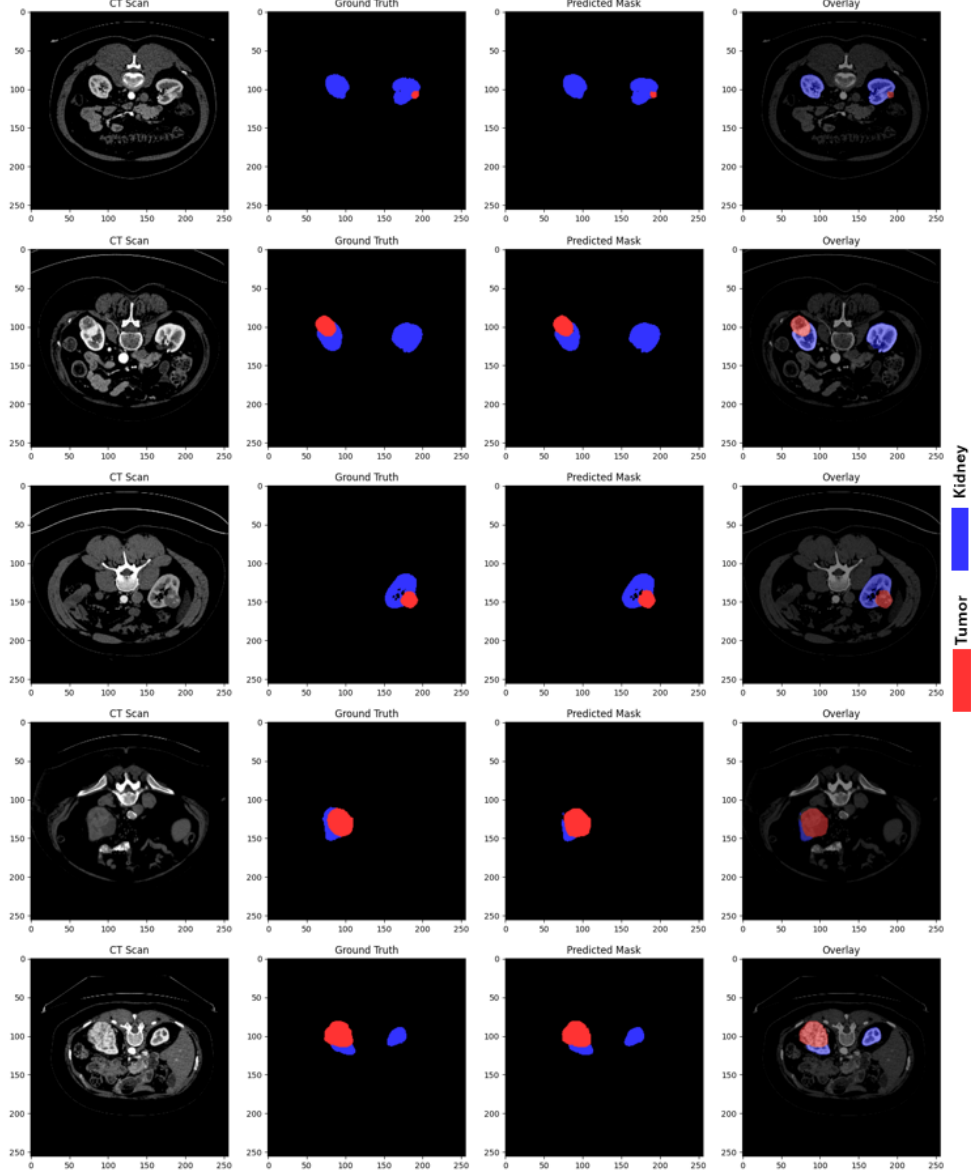


Figure V.13: Visualized segmentation samples of KiTS19

Figure V.14(a) shows the evolution of the Dice values over the training epochs for the kidney and renal tumor segmentation model. The model’s performance improves rapidly during the first 15–20 epochs, with the Dice values increasing sharply for both the training and validation sets for the kidney and renal tumor. Thereafter, the performance gradually stabilizes at high levels (above 0.9), indicating accurate and stable segmentation. It is also observed that the model’s performance in tumor segmentation improves slightly more slowly than in kidney segmentation, which is expected because tumor segmentation is generally more challenging due to the smaller size and greater variability of tumors. There is no clear indication of overfitting, as the training and validation curves remain relatively stable throughout the training process.

Figure V.14(b) shows the HD95 values over the training epochs for both the kidney and renal tumors. The HD95 values decrease progressively as training proceeds. Very high values appear during the first few epochs, then decrease rapidly after about 10 epochs, indicating a substantial improvement in boundary localization for the kidney. For the renal tumor, the HD95 values decrease more gradually during the first 20 epochs. Thereafter, the values gradually stabilize at relatively low levels, with lower values observed for the training sets than for the validation sets. The validation HD95 values remain slightly higher, particularly for the renal tumor, but the curves become relatively stable during the later epochs, indicating a stable learning process without clear signs of severe overfitting.

Figure V.14(c) shows the evolution of the IoU values over the training epochs for both the kidney and renal tumors. The IoU values improve rapidly during the first 20 epochs, with values increasing significantly, especially for the training set. Thereafter, the values gradually stabilize at high levels, with the training IoU reaching approximately 0.95 for both regions, while the validation IoU stabilizes at around 0.88 for the kidney and 0.84 for the tumor. The validation curves remain below the corresponding training curves, particularly for the renal tumor, indicating a moderate performance gap between the training and validation sets. However, both curves become relatively stable during the later epochs, with no clear indication of severe overfitting. The model also achieves better segmentation performance for the kidney than for the tumor on the validation set, which is consistent with the Dice results. The greater difficulty in accurately segmenting tumor regions is expected because tumors are smaller and exhibit greater variability in shape than the kidney

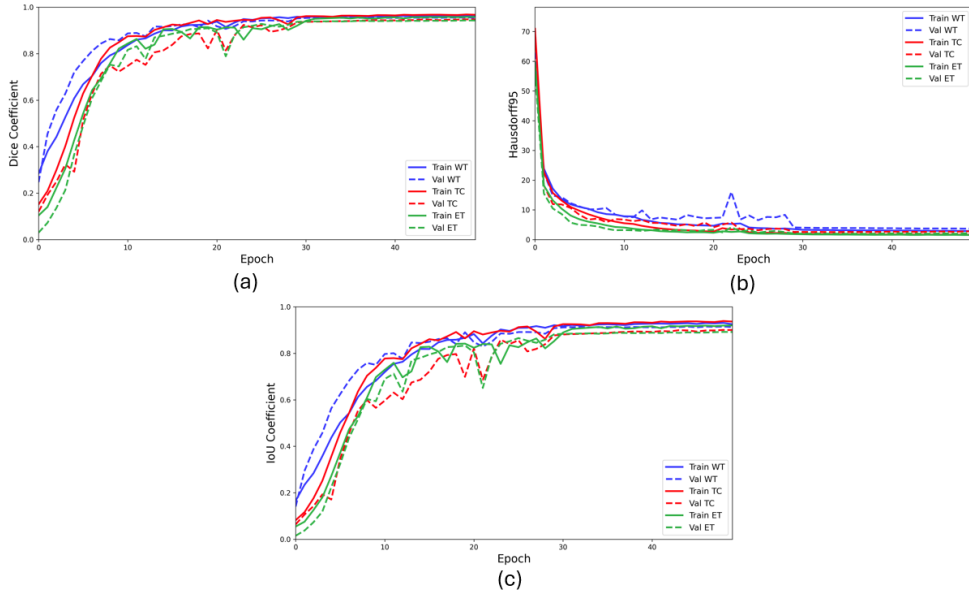


Figure V.14: Graphs of segmentation performance metrics for Kidney and tumor: (a) dice Score, (b) Hausdorff distance 95, and (c) IoU Score.

V.6.2.2 Discussion

By examining the Dice coefficient, a key metric that measures the overlap between the predicted segmentation and the ground truth, we find that the proposed model achieved the best tumor segmentation performance with a Dice score of 0.9274 (see Table V.6), clearly

A Hybrid Unet with FGHQA for Medical Image Segmentation

outperforming all other models, including WPU-net and Inter-CBL, which achieved scores of 0.805 and 0.668, respectively. This significant improvement indicates that FGHQA-Unet was able to capture the tumor region more accurately, which is particularly important given the small size and irregular shape of renal tumors. For kidney segmentation, FGHQA-Unet achieved a Dice score of 0.9445, which was close to the best result (0.948) achieved by Inter-CB, demonstrating strong segmentation performance with only a small difference from the top-performing model.

Table V.6: The segmentation results for our approach and the methods we compared

Method	dice		HD95 (mm)	
	Kidney	Tumor	Kidney	Tumor
Rel-Unet [271]	0.800	0.641	-	-
FR2PAttU-Net [272]	0.948	0.805	-	-
Inter-CBL [273]	0.924	0.668	6.4	21.2
FGHQA-Unet (Proposed)	0.9445	0.89.48	8.1024	17.6866

When analyzing the HD95 metric, which measures the boundary distance between the predicted and ground truth segmentations, FGHQA-Unet achieved values of 8.1024 mm for the kidney and 17.6866 mm for the tumor. FGHQA-Unet demonstrated a substantial improvement in boundary localization, particularly for tumor segmentation, indicating that its predicted boundaries were much closer to the ground truth. However, the Inter-CBL model achieved a slightly lower HD95 value for kidney segmentation (6.4 mm versus 7.79 mm). Nevertheless, this small difference is acceptable given the clear superiority of FGHQA-Unet in tumor segmentation.

Overall, the proposed FGHQA-Unet model strikes a strong balance between prediction accuracy and boundary accuracy, particularly excelling in the challenging tumor segmentation task, where it outperformed all compared models in terms of Dice and HD95. This performance makes it a strong candidate for use in clinical applications, especially in cases where high-precision tumor segmentation is required.

Table V.7: The segmentation results for our approach and the methods we compared.

Model	Class	Dice (%)	IoU (%)	HD95 (mm)	Time (s)	GFLOPs	Params (M)
Unet	Kidney	0.8838	0.7919	37.4481	0.0553	96.208	31.05
	Tumor	0.8351	0.7169	54.3720			
nnUnet	Kidney	0.8924	0.8057	18.9559	0.0476	26.1616	12.41
	Tumor	0.8431	0.7287	20.7399			
TransUnet	Kidney	0.7053	0.5447	35.0762	0.0604	52.2126	93.98
	Tumor	0.7566	0.6085	72.1286			
FGHQA-Unet	Kidney	0.9445	0.8948	8.1024	0.0975	5.8844	27.37
	Tumor	0.9326	0.8737	17.6866			

Table V.7 presents a fair comparison between FGHQA-UNet and the baseline models under identical training and testing conditions. For this comparison, six selected slices per volume were used for training and testing all models, ensuring a consistent input setting. FGHQA-UNet achieved the highest Dice scores for both kidney and tumor segmentation, with values of 0.9445 and 0.9326, respectively. It also obtained the highest IoU values, reaching 0.8948 for the kidney and 0.8737 for the tumor. Regarding boundary accuracy, FGHQA-UNet achieved the lowest HD95 values for both regions, with 8.1024 mm for the kidney and 17.6866 mm for the tumor. Its inference time was 0.0975 s,

which was higher than U-Net (0.0479 s), nnU-Net (0.0492 s), and TransUNet (0.0613 s). Nevertheless, the inference time remained below half a second, indicating an acceptable inference time. Overall, FGHQA-UNet maintained a favorable balance between segmentation accuracy and inference time under the same experimental conditions.

V.6.3 Experiment 03

V.6.3.1 Results

The model demonstrated stable performance in the segmentation task across the different numbers of selected slices, as evidenced by the Dice, IoU, and HD95 values. As shown in Table V.8, increasing the number of slices from six to eight generally improved the segmentation performance. With 6 slices, the Dice and IoU values were 0.9594 and 0.9220 for the liver and 0.9267 and 0.8635 for the tumor, respectively, while the corresponding HD95 values were 7.3785 mm and 36.7762 mm. With seven slices, the Dice and IoU values increased to 0.9656 and 0.9335 for the liver and 0.9337 and 0.8756 for the tumor, accompanied by a substantial reduction in tumor HD95 to 17.9320 mm. The best overall Dice and IoU values were obtained with eight slices, reaching 0.9686 and 0.9391 for the liver and 0.9365 and 0.8806 for the tumor.

Table V.8: Liver and tumor segmentation result with different evaluation metrics on LiTS17

Slices	Class	dice	IoU	HD95 (mm)	Inference time (s)
6	Liver	0.9594	0.9220	7.3785	0.0960
	Tumor	0.9267	0.8635	36.7762	
7	Liver	0.9656	0.9335	5.2995	0.0937
	Tumor	0.9337	0.8756	17.9320	
8	Liver	0.9686	0.9391	5.2982	0.0919
	Tumor	0.9365	0.8806	50.1078	
9	Liver	0.9621	0.9270	6.2649	0.0947
	Tumor	0.9355	0.8789	23.1492	
10	Liver	0.9627	0.9282	6.0663	0.0940
	Tumor	0.9315	0.8719	29.1512	

However, the tumor HD95 increased to 65.0751 mm, indicating that higher overlap accuracy did not necessarily correspond to better boundary localization. The results shown in the figure V.15 further summarize the segmentation performance across the different slice configurations, where the mean values and standard deviations are calculated across all evaluated slices. where the 7-slice configuration achieves the highest mean Dice 0.8866 ± 0.2095 , the lowest mean HD95 12.883 ± 43.5597 mm, and a mean IoU of 0.8371 ± 0.2219 . Beyond eight slices, a slight decline in Dice and IoU was observed, while the overall performance remained relatively stable. In terms of computational efficiency, the inference time remained highly consistent, ranging from 0.0919 to 0.0960 s, with the shortest time recorded for eight slices.

Considering the segmentation performance and inference time together, seven slices provide the most appropriate balance between segmentation quality and computational efficiency. Although eight slices achieved slightly higher Dice and IoU values, the substantial increase in tumor HD95 to 65.0751 mm indicates poorer boundary localization, while the reduction in inference time from 0.0937 s to 0.0919 s is negligible. Moreover,

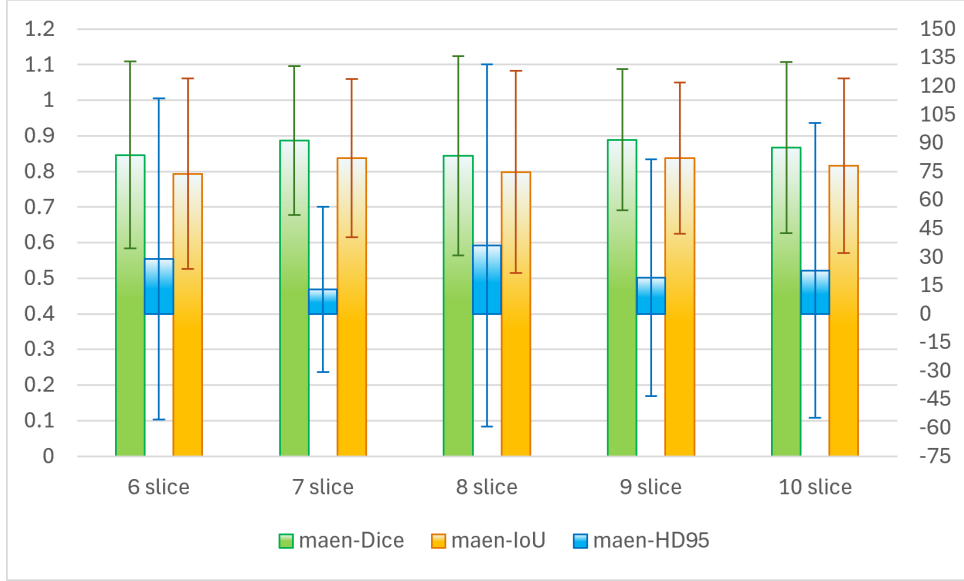


Figure V.15: Mean $\pm$ SD of Segmentation Metrics for Selected Slices

increasing the number of slices to nine or ten did not provide a consistent performance improvement. Therefore, seven slices were selected as the preferred configuration, offering strong segmentation accuracy, better boundary localization, and a low inference time without introducing unnecessary computational cost.

Figure V.16 shows a comparison between the outputs of the proposed medical image segmentation model and the ground truth using CT images of liver tumors. The figure is divided into five rows, each containing four images: the original image, the ground truth showing the liver and tumor regions, the predicted mask generated by the model, and finally an overlay of the predicted mask on the original image. Specific colors are used to distinguish the structures: brown for the liver and green for the tumor. Visual inspection shows that the model achieves highly accurate delineation of both the liver and tumor boundaries, although slight differences can be observed in a few localized regions between the ground truth and the predicted masks.

The curve (a) in Figure V.17 shows the variation in dice values over the training epochs for segmenting both the liver and liver tumors. The model's performance gradually improves as the number of epochs increases, with the dice values rising rapidly during the early stages of training before stabilizing at high values (around 0.97 for the liver and 0.96 for tumors in training, and around 0.95 for the liver and 0.89 for tumors in validation). This indicates that the model learns the underlying features effectively during the initial training stages and then converges to a stable solution. The model also achieves higher segmentation accuracy for the liver than for the tumor, particularly in the validation results, likely due to the greater difficulty of accurately delineating tumor boundaries. Some fluctuations are observed, particularly in the tumor validation curve during the early and middle epochs; however, these gradually diminish and stabilize, indicating stable convergence and effective learning.

The curve (b) in Figure V.17 shows the variation of the HD95 metric over the training epochs for both the liver and hepatic tumors. The HD95 values are initially very high, particularly for the tumors. For the liver, the metric decreases rapidly during the first 20 epochs before stabilizing at relatively low values, indicating that the model effectively learned the liver features and accurately localized its boundaries early in the

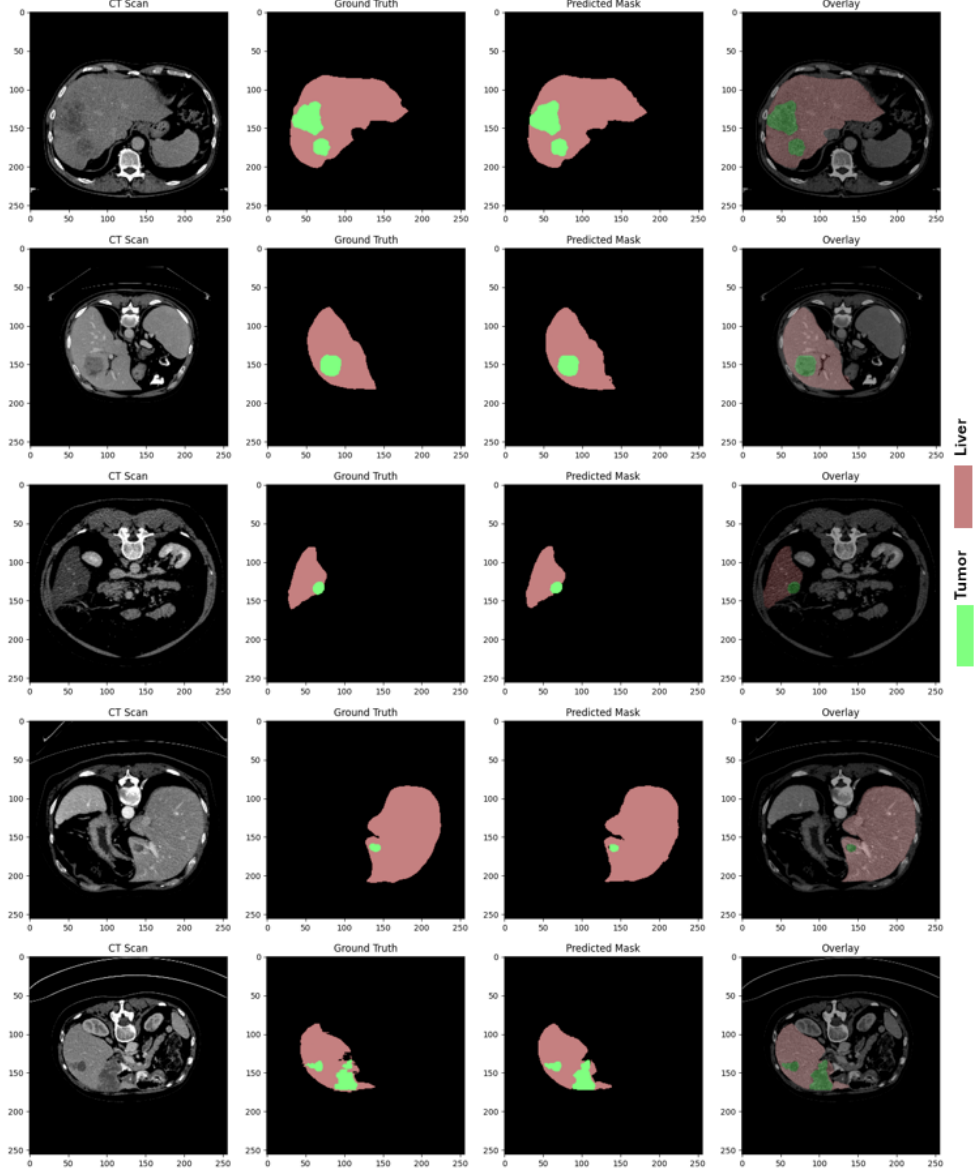


Figure V.16: Visualized segmentation samples of LiTS17

training process. For the hepatic tumors, the decrease is more gradual, with noticeable fluctuations continuing until approximately epoch 30 before the metric gradually stabilizes at values higher than those of the liver. This behavior is expected because hepatic tumors exhibit greater variability in shape, size, and appearance, making accurate boundary delineation more challenging. Nevertheless, the model demonstrates stable performance during the later stages of training, although the validation curve remains more fluctuating and slightly higher than the training curve, particularly for the tumor.

The curve (c) in Figure V.17 shows the variation of IoU values over the training epochs for both liver and hepatic tumor segmentation. The model achieved better segmentation performance for the liver than for the hepatic tumors, with the liver IoU values increasing rapidly during the first 20 epochs before stabilizing at high values of approximately 0.95 for training and 0.92 for validation, indicating effective learning of the liver features. In contrast, the IoU values for the hepatic tumors increased more gradually, remaining relatively low during the early stages of training before improving after approximately epoch 20 and stabilizing at values of around 0.91 for training and 0.81 for

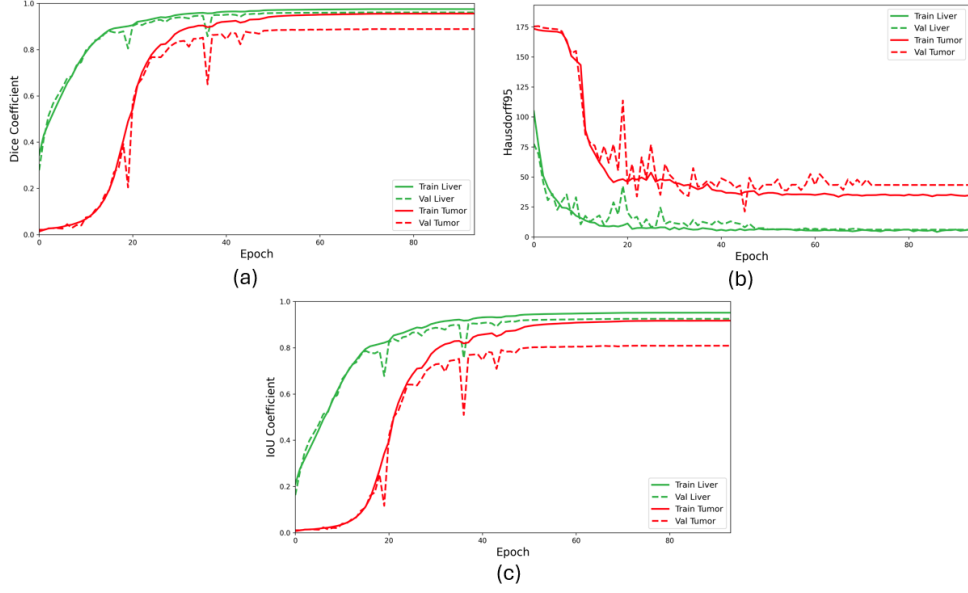


Figure V.17: Graphs of segmentation performance metrics for different tumor regions: (a) dice Score, (b) Hausdorff distance 95, and (c) IoU Score.

validation. This performance difference is likely attributed to the smaller size and greater variability in the shape and appearance of hepatic tumors, which make them more challenging to segment accurately. Nevertheless, the model shows gradual improvement and stable convergence during the later stages of training, although some fluctuations remain in the validation curves, particularly for the tumor.

V.6.3.2 Discussion

The proposed model, FGHQA-Unet, demonstrated excellent performance in the liver and hepatic tumor segmentation task using CT images. Its performance was evaluated against several advanced methods using two key metrics: the Dice coefficient, which measures the degree of overlap between the predicted segmentation and the ground truth, and HD95, which evaluates boundary accuracy by measuring the distance between the predicted and ground-truth boundaries.

In the context of liver segmentation, FGHQA-Unet attained the highest Dice score of 0.9656, slightly outperforming models such as Self-Supervised Swin-Transformer-Based (0.9662) and MA-cGAN (0.9695) (see Table V.9), demonstrating its excellent ability to accurately segment the overall liver region. However, its HD95 value was slightly higher (5.2995 mm) than that of some competing methods, such as MA-cGAN (2.15 mm). This suggests minor boundary deviations in a limited number of regions, while the superior Dice score indicates that the overall segmentation remained highly accurate.

Regarding liver tumor segmentation, the proposed model demonstrated outstanding performance, achieving the highest Dice value of 0.9267 and outperforming all other listed methods, including X-Fuse (0.7876), MA-cGAN (0.7853), and RHEUnet (0.7019). Furthermore, the model achieved superior boundary accuracy, as reflected by its lower HD95 value of 6.3117 mm, compared with 10.23 mm and 12.12 mm for X-Fuse and MA-cGAN, respectively. These results demonstrate the model's excellent ability to accurately segment small and irregular tumors, which are particularly challenging due to their high variability in size, shape, and appearance.

A Hybrid Unet with FGHQA for Medical Image Segmentation

Table V.9: The segmentation results for our approach and the methods we compared

Method	dice		HD95 (mm)	
	Liver	Tumor	Liver	Tumor
Self-Supervised Swin-Transformer-Based [274]	0.9662	0.8647	3.62	4.12
X-Fuse [275]	0.9275	0.7876	3.30	10.23
MA-cGAN [276]	0.9695	0.7853	2.15	12.12
RHEUnet [277]	0.9572	0.7019	-	-
FGHQA-Unet (Proposed)	0.9656	0.9337	5.2995	17.2320

Thus, it can be concluded that the proposed model achieves a good balance between segmentation accuracy and boundary accuracy for both the liver and hepatic tumors, making it a promising approach for clinical applications such as computer-aided diagnosis and surgical planning. Future work may further improve the model’s performance by enhancing the representation of small and irregular tumors, thereby reducing the performance gap between liver and tumor segmentation.

Table V.10: The segmentation results for our approach and the methods we compared.

Model	Class	Dice (%)	IoU (%)	HD95 (mm)	Time (s)	GFLOPs	Params (M)
Unet	Liver	0.8938	0.8080	12.9832	0.0479	96.208	31.05
	Tumor	0.7442	0.5926	52.9264			
nnUnet	Liver	0.9234	0.8578	10.5961	0.0492	26.1616	12.41
	Tumor	0.7954	0.6604	22.4085			
TransUNet	Liver	0.8132	0.6853	17.4518	0.0613	52.2126	93.98
	Tumor	0.0741	0.0385	127.3563			
FGHQA-Unet	Liver	0.9656	0.9335	5.2995	0.0937	5.8844	27.37
	Tumor	0.9337	0.8756	17.9320			

Table V.10 presents a fair comparison between FGHQA-UNet and the baseline models under identical training and testing conditions. For this comparison, seven selected slices per volume were used for training and testing all models, ensuring a consistent input setting. FGHQA-UNet achieved the highest Dice scores for both liver and tumor segmentation, with values of 0.9656 and 0.9337, respectively. It also obtained the highest IoU values, reaching 0.9335 for the liver and 0.8756 for the tumor. Regarding boundary accuracy, FGHQA-UNet achieved the lowest HD95 values for both regions, with 5.2995 mm for the liver and 17.9320 mm for the tumor. Its inference time was 0.0937 s, which was higher than U-Net 0.0479 s, nnU-Net 0.0492 s, and TransUNet 0.0613 s. Nevertheless, the inference time remained below half a second, indicating an acceptable inference time. Overall, FGHQA-UNet maintained a favorable balance between segmentation accuracy and inference time under the same experimental conditions.

V.7 Conclusion

This chapter concludes with the presentation of a novel scientific contribution through the proposed FGHQA-Unet model, which represents an advanced extension of traditional convolutional architectures by integrating the FGHQA mechanism within its architectural design. This hybrid integration achieves an effective balance between global and local

context modeling and generates a semantically guided contextual representation that focuses on important regions.

Moreover, the adopted preprocessing strategy contributes to effective data preparation for model training. A comprehensive evaluation of the proposed model was conducted on the BraTS, KiTS, and LiTS datasets, where experimental results demonstrated its superiority over advanced methods in terms of both segmentation accuracy and computational efficiency, confirming its effectiveness and applicability in medical computer vision scenarios.

General conclusion

Medical image segmentation remains a challenging task due to the large variability in tumor characteristics, the complexity of anatomical structures, and the diversity of medical imaging modalities. These challenges make it difficult to achieve accurate and robust segmentation that can generalize across different clinical scenarios. This thesis addressed these challenges through the development of effective deep learning-based segmentation models that improve segmentation accuracy and robustness by enhancing feature extraction, contextual representation, and feature fusion strategies.

This thesis advances the field of medical image segmentation through the development and evaluation of three complementary deep learning-based models. The first model enhances the conventional U-Net architecture by incorporating a DCB module into the skip connections, enabling more effective propagation and refinement of the spatial information exchanged between the encoder and decoder. The second model introduces an enhanced VGG19-based segmentation framework that integrates a pre-trained VGG19 encoder with a modified decoder employing DSC layers, thereby reducing the number of trainable parameters and computational cost while maintaining competitive segmentation performance. In addition, a comparative study of three preprocessing techniques, namely Gaussian filtering, Median filtering, and CLAHE, was conducted to improve input image quality, identifying CLAHE as the most effective preprocessing technique. Finally, this thesis presents FGHQA-UNet, an enhanced U-Net architecture featuring a hybrid encoder that combines CNN layers with attention mechanisms and a Hybrid Query Attention module to effectively capture both local and global contextual information for accurate medical image segmentation.

The experimental evaluation confirmed the effectiveness of the proposed approaches across multiple medical image segmentation tasks. The DCB-UNet demonstrated a substantial improvement over the baseline U-Net, achieving more than a 6% increase in the Dice coefficient (0.9083 versus 0.8417), confirming the effectiveness of incorporating the DCB module into the skip connections to enhance the transfer and refinement of spatial information between the encoder and decoder, resulting in more accurate tumor boundary delineation. The enhanced VGG19-based model further demonstrated that combining a pre-trained encoder with a modified decoder based on depthwise separable convolutional layers reduced the number of trainable parameters by approximately 26.4%, from 31,031,745 to 22,856,513, while maintaining high segmentation performance. Furthermore, among the evaluated preprocessing techniques, CLAHE provided the best segmentation performance, achieving a Dice coefficient of 0.9055 and an IoU of 0.8278. Finally, the proposed FGHQA-UNet model consistently demonstrated robust and reliable performance across the BraTS, KiTS, and LiTS datasets, confirming the effectiveness of the proposed hybrid architecture for accurate and generalizable medical image segmenta-

tion. Overall, these findings demonstrate that progressively improving spatial information transfer, reducing model complexity, and enhancing contextual representation lead to consistent improvements in segmentation performance across different organs, tumor types, and medical imaging modalities.

From a theoretical perspective, this thesis demonstrates that achieving accurate and robust medical image segmentation requires the integration of complementary deep learning strategies rather than relying on a single architectural enhancement. The findings indicate that preserving and refining spatial information throughout the encoding and decoding process is fundamental for recovering fine anatomical details and accurately delineating complex tumor boundaries. This highlights the importance of enhancing skip connections, not merely as pathways for feature transmission, but as effective mechanisms for maintaining spatial consistency and facilitating the reconstruction of detailed segmentation masks.

The study further confirms the value of transfer learning as an effective approach for medical image segmentation, demonstrating that knowledge acquired from large-scale natural image datasets can be successfully transferred to medical imaging tasks to improve feature extraction, particularly when annotated medical data are limited. In addition, the results show that depthwise separable convolutions provide an efficient alternative to conventional convolutions by substantially reducing the number of trainable parameters while maintaining competitive segmentation performance. These findings demonstrate that a substantial reduction in the number of trainable parameters can be achieved while preserving competitive segmentation performance through an appropriate architectural design.

Another important theoretical outcome of this work is the demonstration that the quality of input data remains a fundamental factor influencing the overall learning process. The comparative evaluation of preprocessing techniques revealed that enhancing image contrast contributes more effectively to segmentation performance than noise reduction alone, highlighting the importance of maximizing the discriminability of anatomical structures before feature extraction. This observation reinforces the close relationship between data quality and the effectiveness of deep learning-based segmentation models.

Finally, this thesis demonstrates that combining the local feature extraction capability of convolutional neural networks with the global contextual modeling ability of transformer-inspired attention mechanisms leads to richer and more informative feature representations than relying on either approach independently. Rather than treating local and global representations as separate sources of information, the proposed framework shows that their integration enables a more comprehensive understanding of medical images, particularly in tumors exhibiting complex morphological and imaging characteristics, irregular morphology, and multiple pathological regions. Furthermore, the study highlights the importance of slices selection by demonstrating that focusing the learning process on the most informative image slices, identified according to tumor pixel count and tumor density, allows the model to concentrate on clinically relevant regions and improves the effectiveness of feature learning. Collectively, these findings provide valuable insights into the design of future medical image segmentation architectures by emphasizing that robust performance is achieved through the complementary integration of efficient feature extraction, enhanced spatial information transfer, contextual modeling, tumor-guided data selection, and high-quality input representations.

Despite the promising results achieved in this thesis, several limitations remain. The proposed models were evaluated exclusively on publicly available benchmark datasets,

which may not fully represent the diversity encountered in real clinical environments. In addition, the evaluation relied solely on quantitative segmentation metrics without validation by clinical experts, leaving the clinical applicability of the generated segmentation masks to be further investigated. Furthermore, although the proposed frameworks demonstrated strong performance across different organs and imaging modalities, their robustness under unseen imaging protocols and independent clinical datasets has not yet been evaluated. Addressing these limitations would provide a more comprehensive assessment of the proposed models and further support their adoption in real-world medical applications.

The findings of this thesis open several promising directions for future research aimed at further improving the robustness, applicability, and clinical adoption of deep learning-based medical image segmentation systems. An important research direction is the validation of the proposed frameworks using multiple independent clinical datasets acquired from different medical institutions to better assess their generalization capability under diverse imaging conditions. Future work may also extend the proposed architectures from 2D slice-based segmentation to full 3D volumetric segmentation, enabling a more comprehensive representation of anatomical structures and tumor morphology. In addition, integrating explainable artificial intelligence techniques into the proposed models would improve the transparency of the decision-making process, allowing clinicians to better understand and trust the generated segmentation results. Finally, incorporating the proposed frameworks into clinical workflow systems has the potential to support real-time surgical guidance, treatment planning, and patient monitoring, thereby facilitating the translation of advanced deep learning research into practical healthcare applications.

BIBLIOGRAPHY

- [1] S. Asgari Taghanaki, K. Abhishek, J. P. Cohen, J. Cohen-Adad, and G. Hamarneh, “Deep semantic segmentation of natural and medical images: a review,” *Artificial intelligence review*, vol. 54, no. 1, pp. 137–178, 2021.
- [2] S. U. Rahman, H. Rahman, F. Din, S. Ullah, and A. Uddin, “An Investigation of Challenges in Automatic Segmentation of Medical Images,” *Bahria University Journal of Information & Communication Technology*, vol. 10, no. 1, pp. 6–13, 2017.
- [3] J. Chaki, “Deep learning in healthcare: applications, challenges, and opportunities,” *Next Generation Healthcare Informatics*, pp. 27–44, 2022.
- [4] N. RajamohanReddy and G. Muneeswari, “Advancing multi-categorization and segmentation in brain tumors using novel efficient deep learning approaches,” *PeerJ Computer Science*, vol. 10, p. e2496, 2024.
- [5] C. Fedele, R. W. Tothill, and G. A. McArthur, “Navigating the challenge of tumor heterogeneity in cancer therapy,” *Cancer discovery*, vol. 4, no. 2, pp. 146–148, 2014.
- [6] P. Gilson, J.-L. Merlin, and A. Harlé, “Deciphering tumour heterogeneity: from tissue to liquid biopsy,” *Cancers*, vol. 14, no. 6, p. 1384, 2022.
- [7] V. Rakhshan, A. H. Okano, Z. Huang, G. Castelnovo, and A. F. Baptista, “Biomedical applications of computer vision using artificial intelligence,” *Computational Intelligence and Neuroscience*, vol. 2022, p. 9843574, 2022.
- [8] R. Birjais, “Challenges and future directions for segmentation of medical images using deep learning models,” in *Deep Learning Applications in Medical Image Segmentation: Overview, Approaches, and Challenges*. Wiley Online Library, 2025, pp. 243–264.
- [9] P. H. O. Pinheiro, “Large-scale image segmentation with convolutional networks,” Ph.D. dissertation, Ecole Polytechnique Fédérale de Lausanne, 2017.
- [10] R. Girasa, “Applications of ai and projections of ai impact,” in *Artificial Intelligence as a Disruptive Technology: Economic Transformation and Government Regulation*. Springer, 2020, pp. 23–67.
- [11] M. Ashraf and M. Zaman, “Revealing historical insights: A comprehensive exploration of traditional approaches in medical image segmentation,” in *Deep Learning*

- Applications in Medical Image Segmentation: Overview, Approaches, and Challenges*. Wiley Online Library, 2025, pp. 65–84.
- [12] W. Yang, X. Wang, J. Lu, W. Dou, and S. Liu, “Interactive steering of hierarchical clustering,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 10, pp. 3953–3967, 2020.
 - [13] T. Lei, X. Jia, Y. Zhang, L. He, H. Meng, and A. K. Nandi, “Significantly fast and robust fuzzy c-means clustering algorithm based on morphological reconstruction and membership filtering,” *IEEE Transactions on Fuzzy Systems*, vol. 26, no. 5, pp. 3027–3041, 2018.
 - [14] T. Lei, P. Liu, X. Jia, X. Zhang, H. Meng, and A. K. Nandi, “Automatic fuzzy clustering framework for image segmentation,” *IEEE Transactions on Fuzzy Systems*, vol. 28, no. 9, pp. 2078–2092, 2019.
 - [15] A. Lotfi, P. Moradi, and H. Beigy, “Density peaks clustering based on density backbone and fuzzy neighborhood,” *Pattern Recognition*, vol. 107, p. 107449, 2020.
 - [16] G. Sun, Y. Cong, Q. Wang, J. Li, and Y. Fu, “Lifelong spectral clustering,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 04, 2020, pp. 5867–5874.
 - [17] N. Sladoje and J. Lindblad, “Pixel coverage segmentation for improved feature estimation,” in *International Conference on Image Analysis and Processing*. Springer, 2009, pp. 929–938.
 - [18] G. Rai and T. Nair, “Gradient based seeded region grow method for ct angiographic image segmentation,” *arXiv preprint arXiv:1001.3735*, 2010.
 - [19] C. Chen, J. A. Ozolek, W. Wang, and G. K. Rohde, “A general system for automatic biomedical image segmentation using intensity neighborhoods,” *International Journal of Biomedical Imaging*, vol. 2011, no. 1, p. 606857, 2011.
 - [20] K. Suzuki, “Pixel-based machine learning in medical imaging,” *International Journal of Biomedical Imaging*, vol. 2012, no. 1, p. 792079, 2012.
 - [21] I. Aganj, M. G. Harisinghani, R. Weissleder, and B. Fischl, “Unsupervised medical image segmentation based on the local center of mass,” *Scientific reports*, vol. 8, no. 1, p. 13012, 2018.
 - [22] O. Salih and S. Viriri, “Skin lesion segmentation using stochastic region-merging and pixel-based markov random field,” *Symmetry*, vol. 12, no. 8, p. 1224, 2020.
 - [23] J. Wang, S. Zhou, C. Fang, L. Wang, and J. Wang, “Meta corrupted pixels mining for medical image segmentation,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 335–345.
 - [24] E. Ahn, D. Feng, and J. Kim, “A spatial guided self-supervised clustering network for medical image segmentation,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 379–388.

- [25] B. Serrano-Antón, M. Insúa Villa, S. Pendón-Minguillón, S. Paramés-Estévez, A. Otero-Cacho, D. López-Otero, B. Díaz-Fernández, M. Bastos-Fernández, J. R. González-Juanatey, and A. P. Muñuzuri, “Unsupervised clustering based coronary artery segmentation,” *BioData Mining*, vol. 18, no. 1, p. 21, 2025.
- [26] M. R. Maire, *Contour detection and image segmentation*. University of California, Berkeley, 2009.
- [27] D. Baswaraj, A. Govardhan, and P. Premchand, “Active contours and image segmentation: The current state of the art,” *Global Journal of Computer Science and Technology*, vol. 12, no. 11-F, 2012.
- [28] Y. Chen, P. Ge, G. Wang, G. Weng, and H. Chen, “An overview of intelligent image segmentation using active contour models,” *Intelligence & Robotics*, vol. 3, no. 1, pp. 23–55, 2023.
- [29] S. Matsumoto, R. Asato, T. Okada, and J. Konishi, “Intracranial contour extraction with active contour models,” *Journal of Magnetic Resonance Imaging*, vol. 7, no. 2, pp. 353–360, 1997.
- [30] M. Rousson and N. Paragios, “Shape priors for level set representations,” in *European Conference on Computer Vision*. Springer, 2002, pp. 78–92.
- [31] S. Lankton and A. Tannenbaum, “Localizing region-based active contours,” *IEEE transactions on image processing*, vol. 17, no. 11, pp. 2029–2039, 2008.
- [32] Z.-l. Fu, M. Ye, Y.-l. Su, Y.-p. Lin, and C.-t. Wang, “A novel active contour model for medical image segmentation,” *Journal of Shanghai Jiaotong University (Science)*, vol. 15, no. 5, pp. 549–555, 2010.
- [33] X. Qian, J. Wang, S. Guo, and Q. Li, “An active contour model for medical image segmentation with application to brain ct image,” *Medical physics*, vol. 40, no. 2, p. 21911, 2013.
- [34] Y. Lu, K. Zheng, W. Li, Y. Wang, A. P. Harrison, C. Lin, S. Wang, J. Xiao, L. Lu, and C.-F. Kuo, “Contour transformer network for one-shot segmentation of anatomical structures,” *IEEE transactions on medical imaging*, vol. 40, no. 10, pp. 2672–2684, 2020.
- [35] E. Upschulte, S. Harmeling, K. Amunts, and T. Dickscheid, “Contour proposal networks for biomedical instance segmentation,” *Medical image analysis*, vol. 77, p. 102371, 2022.
- [36] S. Park and M. Chung, “Cardiac segmentation on ct images through shape-aware contour attentions,” *Computers in Biology and Medicine*, vol. 147, p. 105782, 2022.
- [37] Y. Gu, Y. Liu, Z. Sun, J. Zhu, Y. Xu, and L. Najman, “Shape transformation driven by active contour for class-imbalanced semi-supervised medical image segmentation,” in *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2024, pp. 1966–1973.

- [38] M. G. Oghli, A. Fallahi, and M. Pooyan, “Automatic region growing method using gsmmap and spatial information on ultrasound images,” in *2010 18th Iranian Conference on Electrical Engineering*. IEEE, 2010, pp. 35–38.
- [39] W. Deng, W. Xiao, H. Deng, and J. Liu, “Mri brain tumor segmentation with region growing method based on the gradients and variances along and inside of the boundary curve,” in *2010 3rd International Conference on Biomedical Engineering and Informatics*, vol. 1. IEEE, 2010, pp. 393–396.
- [40] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, “Slic superpixels compared to state-of-the-art superpixel methods,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.
- [41] X. Shan, X. Gong, and A. K. Nandi, “Active contour model based on local intensity fitting energy for image segmentation and bias estimation,” *IEEE Access*, vol. 6, pp. 49 817–49 827, 2018.
- [42] D. Mumford and J. Shah, “Optimal approximations by piecewise smooth functions and associated variational problems,” 1989.
- [43] T. F. Chan and L. A. Vese, “Active contours without edges,” *IEEE Transactions on image processing*, vol. 10, no. 2, pp. 266–277, 2001.
- [44] P. F. Felzenszwalb and D. P. Huttenlocher, “Efficient graph-based image segmentation,” *International journal of computer vision*, vol. 59, no. 2, pp. 167–181, 2004.
- [45] C. Liu, M. Amodio, L. L. Shen, F. Gao, A. Avesta, S. Aneja, J. C. Wang, L. V. Del Priore, and S. Krishnaswamy, “Cuts: A deep learning and topological framework for multigranular unsupervised medical image segmentation,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 155–165.
- [46] A. El Allaoui and M. Nasri, “Medical image segmentation by evolutionary approach and watershed morphology,” *International Journal of Computer Applications*, vol. 975, p. 8887, 2012.
- [47] Z. Zhou, H. Wan, H. Zhang, X. Chen, X. Wang, S. Lili, and T. Zhang, “Segmentation of spontaneous intracerebral hemorrhage on ct with a region growing method based on watershed preprocessing,” *Frontiers in Neurology*, vol. 13, p. 865023, 2022.
- [48] Z. Chen, L. Yao, Y. Liu, X. Han, Z. Gong, J. Luo, J. Zhao, and G. Fang, “Deep learning-aided 3d proxy-bridged region-growing framework for multi-organ segmentation,” *Scientific Reports*, vol. 14, no. 1, p. 9784, 2024.
- [49] M. Shi and I. Hussain, “Improved region-based active contour segmentation through divergence and convolution techniques,” *AIMS Math*, vol. 10, no. 1, pp. 654–671, 2025.
- [50] J. Q. Lim, N. Alias, and F. Johar, “Supervised classification of normal and tumorous brain mr images using machine learning schemes,” 2020.
- [51] W. S. Noble, “What is a support vector machine?” *Nature biotechnology*, vol. 24, no. 12, pp. 1565–1567, 2006.

- [52] H. A. Salman, A. Kalakech, and A. Steiti, “Random forest algorithm overview,” *Babylonian Journal of Machine Learning*, vol. 2024, pp. 69–79, 2024.
- [53] M. Ahmed, R. Seraj, and S. M. S. Islam, “The k-means algorithm: A comprehensive survey and performance evaluation,” *Electronics*, vol. 9, no. 8, p. 1295, 2020.
- [54] H. Yuan, K. Yu, F. Xie, M. Liu, and S. Sun, “Automated machine learning with interpretation: A systematic review of methodologies and applications in healthcare,” *Medicine Advances*, vol. 2, no. 3, pp. 205–237, 2024.
- [55] S. Gupta and S. Gupta, “Feature extraction and feature selection procedures for medical image analysis,” in *Computer-Assisted Analysis for Digital Medicinal Imagery*. IGI Global, 2025, pp. 221–280.
- [56] Y. Xu, R. Quan, W. Xu, Y. Huang, X. Chen, and F. Liu, “Advances in medical image segmentation: A comprehensive review of traditional, deep learning and hybrid approaches,” *Bioengineering*, vol. 11, no. 10, p. 1034, 2024.
- [57] A. Rehman and S. Y. Bhat, “Fundamentals of deep learning models for medical image segmentation,” in *Deep Learning Applications in Medical Image Segmentation: Overview, Approaches, and Challenges*. Wiley Online Library, 2025, pp. 27–63.
- [58] A. K. Sharma, P. K. Tripathi, and S. Sharma, “Role of artificial intelligence in biomedical imaging,” in *Advancements in bio-medical image processing and authentication in telemedicine*. IGI Global, 2023, pp. 17–34.
- [59] K. Ma, J. Zhang, X. Huang, and M. Wang, “Leveraging transformer models to predict cognitive impairment: accuracy, efficiency, and interpretability,” *BMC Public Health*, vol. 25, no. 1, p. 504, 2025.
- [60] W. Zhang, A. Hasegawa, K. Itoh, and Y. Ichioka, “Image processing of human corneal endothelium based on a learning network,” *Applied optics*, vol. 30, no. 29, pp. 4211–4217, 1991.
- [61] D. Ciresan, A. Giusti, L. Gambardella, and J. Schmidhuber, “Deep neural networks segment neuronal membranes in electron microscopy images,” *Advances in neural information processing systems*, vol. 25, 2012.
- [62] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.
- [63] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5–9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.
- [64] F. Milletari, N. Navab, and S.-A. Ahmadi, “V-net: Fully convolutional neural networks for volumetric medical image segmentation,” in *2016 fourth international conference on 3D vision (3DV)*. Ieee, 2016, pp. 565–571.

- [65] . Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, “3d u-net: learning dense volumetric segmentation from sparse annotation,” in *International conference on medical image computing and computer-assisted intervention*. Springer, 2016, pp. 424–432.
- [66] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, “H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes,” *IEEE transactions on medical imaging*, vol. 37, no. 12, pp. 2663–2674, 2018.
- [67] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, and B. Kainz, “Attention u-net: Learning where to look for the pancreas,” *arXiv preprint arXiv:1804.03999*, 2018.
- [68] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, “Unet++: A nested u-net architecture for medical image segmentation,” in *International workshop on deep learning in medical image analysis*. Springer, 2018, pp. 3–11.
- [69] F. Isensee, J. Petersen, A. Klein, D. Zimmerer, P. F. Jaeger, S. Kohl, J. Wasserthal, G. Koehler, T. Norajitra, and S. Wirkert, “nnu-net: Self-adapting framework for u-net-based medical image segmentation,” *arXiv preprint arXiv:1809.10486*, 2018.
- [70] M. Blendowski, H. Nickisch, and M. P. Heinrich, “How to learn from unlabeled volume data: Self-supervised 3d context feature learning,” in *International conference on medical image computing and computer-assisted intervention*. Springer, 2019, pp. 649–657.
- [71] R. Windsor, A. Jamaludin, T. Kadir, and A. Zisserman, “Self-supervised multi-modal alignment for whole body medical imaging,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 90–101.
- [72] K. Zhang and D. Liu, “Customized segment anything model for medical image segmentation,” *arXiv preprint arXiv:2304.13785*, 2023.
- [73] N. Huda and K. R. Ku-Mahamud, “Cnn-based image segmentation approach in brain tumor classification: A review,” *Engineering Proceedings*, vol. 84, no. 1, p. 66, 2025.
- [74] Q. Xia, H. Zheng, H. Zou, D. Luo, H. Tang, L. Li, and B. Jiang, “A comprehensive review of deep learning for medical image segmentation,” *Neurocomputing*, vol. 613, p. 128740, 2025.
- [75] W. Baccouch, S. Oueslati, B. Solaiman, and S. Labidi, “A comparative study of cnn and u-net performance for automatic segmentation of medical images: application to cardiac mri,” *Procedia Computer Science*, vol. 219, pp. 1089–1096, 2023.
- [76] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, “Transunet: Transformers make strong encoders for medical image segmentation,” *arXiv preprint arXiv:2102.04306*, 2021.
- [77] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, “Swin-unet: Unet-like pure transformer for medical image segmentation,” in *European conference on computer vision*. Springer, 2022, pp. 205–218.

- [78] A. Lin, B. Chen, J. Xu, Z. Zhang, G. Lu, and D. Zhang, “Ds-transunet: Dual swin transformer u-net for medical image segmentation,” *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–15, 2022.
- [79] Z. Li, Y. Li, Q. Li, P. Wang, D. Guo, L. Lu, D. Jin, Y. Zhang, and Q. Hong, “Lvit: language meets vision transformer in medical image segmentation,” *IEEE transactions on medical imaging*, vol. 43, no. 1, pp. 96–107, 2023.
- [80] Q. Liu, C. Kaul, J. Wang, C. Anagnostopoulos, R. Murray-Smith, and F. Deligianni, “Optimizing vision transformers for medical image segmentation,” in *ICASSP 2023-2023 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [81] R. Azad, M. Heidari, M. Shariatnia, E. K. Aghdam, S. Karimijafarbigloo, E. Adeli, and D. Merhof, “Transdeeplab: Convolution-free transformer-based deeplab v3+ for medical image segmentation,” in *International Workshop on PRedictive Intelligence In MEDicine*. Springer, 2022, pp. 91–102.
- [82] A. Tragakis, C. Kaul, R. Murray-Smith, and D. Husmeier, “The fully convolutional transformer for medical image segmentation,” in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2023, pp. 3660–3669.
- [83] C. Wei, S. Ren, K. Guo, H. Hu, and J. Liang, “High-resolution swin transformer for automatic medical image segmentation,” *Sensors*, vol. 23, no. 7, p. 3420, 2023.
- [84] Z. Li, D. Li, C. Xu, W. Wang, Q. Hong, Q. Li, and J. Tian, “Tfcns: A cnn-transformer hybrid network for medical image segmentation,” in *International conference on artificial neural networks*. Springer, 2022, pp. 781–792.
- [85] H. Wu, S. Chen, G. Chen, W. Wang, B. Lei, and Z. Wen, “Fat-net: Feature adaptive transformers for automated skin lesion segmentation,” *Medical image analysis*, vol. 76, p. 102327, 2022.
- [86] A. R. Khan and A. Khan, “Maxvit-unet: Multi-axis attention for medical image segmentation,” *arXiv preprint arXiv:2305.08396*, 2023.
- [87] O. N. Manzari, J. M. Kaleybar, H. Saadat, and S. Maleki, “Befunet: A hybrid cnn-transformer architecture for precise medical image segmentation,” *arXiv preprint arXiv:2402.08793*, 2024.
- [88] Z. Gong, A. P. French, G. Qiu, and X. Chen, “Convtransseg: A multi-resolution convolution-transformer network for medical image segmentation,” *arXiv preprint arXiv:2210.07072*, 2022.
- [89] F. Bougourzi, F. Dornaika, C. Distant, and A. Taleb-Ahmed, “D-trattunet: Toward hybrid cnn-transformer architecture for generic and subtle segmentation in medical images,” *Computers in biology and medicine*, vol. 176, p. 108590, 2024.
- [90] Y. Huang, L. Chen, and C. Zhou, “Multi-modal brain tumor segmentation via 3d multi-scale self-attention and cross-attention,” *arXiv preprint arXiv:2504.09088*, 2025.

- [91] D. L. Ferreira, B. A. A. Nunes, X. Zhang, L. C. Gomez, M. Fung, and R. Soni, "Samri-2: A memory-based model for cartilage and meniscus segmentation in 3d mris of the knee joint," *arXiv preprint arXiv:2502.10559*, 2025.
- [92] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annual review of biomedical engineering*, vol. 19, no. 1, pp. 221–248, 2017.
- [93] M.-C. Zdora, "Principles of x-ray imaging," *X-ray Phase-Contrast Imaging Using Near-Field Speckles*, pp. 11–57, 2021.
- [94] C. Hutchison, "How does an x-ray machine work? a quick overview," patientimage, 2023. [Online]. Available: <https://www.patientimage.com/blog/how-does-x-ray-machine-work>
- [95] W. Huda and R. B. Abrahams, "X-ray-based medical imaging and resolution," *American Journal of Roentgenology*, vol. 204, no. 4, pp. W393–W397, 2015.
- [96] K. V. Dhruwdas, "Systems and methods for obtaining 3-d images from x-ray information," Google Patents, feb 2019.
- [97] N. Umadevi and S. N. Geethalakshmi, "Multiple classification system for fracture detection in human bone x-ray images," in *2012 Third International Conference on Computing, Communication and Networking Technologies (ICCCNT'12)*. IEEE, 2012, pp. 1–8.
- [98] R. Jamil, M. Dong, S. Bano, A. Javed, and M. Abdullah, "Precancerous microcalcification detection of breast cancer mammogram images using linear time-invariant filtering wiener method with tophat transformation," 2023.
- [99] A. A. Raut, P. S. Naphade, and S. Maheshwari, "Abdominal radiograph," *Journal of Gastrointestinal and Abdominal Radiology*, vol. 3, no. S 01, pp. S22–S34, 2020.
- [100] E. L. Irede, O. R. Aworinde, O. K. Lekan, O. D. Amienghemhen, T. P. Okonkwo, A. P. Onivefu, and I. H. Ifijen, "Medical imaging: a critical review on x-ray imaging for the detection of infection," *Biomedical Materials & Devices*, pp. 1–45, 2024.
- [101] T. M. Buzug, "Computed tomography," in *Springer handbook of medical technology*. Springer, 2011, pp. 311–342.
- [102] C. Bracco, D. Regge, M. Stasi, M. Gabelloni, and E. Neri, "Principles of ct and mr imaging," in *Nuclear Medicine Textbook: Methodology and Clinical Applications*. Springer, 2019, pp. 187–198.
- [103] S. Tao, J. F. Marsh, A. Tao, G. J. Michalak, K. Rajendran, C. H. McCollough, and S. Leng, "Multi-energy ct imaging for large patients using dual-source photon-counting detector ct," *Physics in Medicine & Biology*, vol. 65, no. 17, p. 17NT01, 2020.
- [104] M. Spahn, "X-ray detectors in medical imaging," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 731, pp. 57–63, 2013.

- [105] J. Essam, “Ct scan concept and operation,” 2019. [Online]. Available: <https://www.linkedin.com/pulse/ct-scan-concept-operation-p1-jean-essam>
- [106] A. H. M. Algofinah, T. M. A. Al Sqoor, H. M. A. Alyami, H. A. H. AlZaaman, M. A. Albaltheen, M. M. M. Almahri, S. M. M. Alsharyah, J. M. H. Alyami, and A. M. A. Alarram, “Comprehensive review of radiological applications in cancer detection and monitoring,” *Journal of Ecohumanism*, vol. 3, no. 8, pp. 5179–5187, 2024.
- [107] J. G. Mainprize, M. J. Yaffe, T. Chawla, and P. Glanc, “Effects of ionizing radiation exposure during pregnancy,” *Abdominal Radiology*, vol. 48, no. 5, pp. 1564–1578, 2023.
- [108] A. Shah and S. Aran, “A review of magnetic resonance (mr) safety: the essentials to patient safety,” *Cureus*, vol. 15, no. 10, 2023.
- [109] D. Pradeep, M. K. Tembhre, A. S. Parihar, and C. Rao, “Magnetic resonance imaging: basic principles and advancement in clinical and diagnostics approaches in health care,” in *Biomedical Imaging Instrumentation*. Elsevier, 2022, pp. 45–66.
- [110] Clevelandclinic, “Mra (magnetic resonance angiography),” 2025. [Online]. Available: <https://my.clevelandclinic.org/health/diagnostics/24024-mra>
- [111] D. Kumar, B. Pratap, N. Boora, R. Kumar, and N. K. Sah, “A comparative study of medical imaging modalities,” *Int. J. Radiol. Sci*, vol. 3, pp. 9–16, 2021.
- [112] J. E. Brown, R. Qiang, P. J. Stadnik, L. J. Stotts, and J. A. Von Arx, “Virtual humans for implantable device safety assessment in mri: mitigating magnetic resonance imaging hazards for implanted medical devices,” *IEEE pulse*, vol. 8, no. 4, pp. 50–53, 2017.
- [113] L. V. Nguyen, N. H. D. Nguyen, H. K. T. Truong, and M. T. T. Giang, “School problems among left-behind children of labor migrant parents: a study in vietnam,” *Health psychology report*, vol. 10, no. 4, p. 266, 2022.
- [114] I. for Quality and G. Efficiency in Health Care (IQWiG, “Colorectal cancer,” 2021. [Online]. Available: <https://www.informedhealth.org/colonoscopy-information-for-women.html>
- [115] S. Ashtari, M. A. Pourhoseingholi, and R. M. Zali, “Overview of diagnostic and treatment colonoscopy function in gastrointestinal diseases,” *J Liver Res Disord Ther*, vol. 2, no. 4, pp. 113–117, 2016.
- [116] E. R. Crandell, *Quizlet flashcards for the first 500 words of the academic vocabulary list*. Brigham Young University, 2017.
- [117] “Pet scan,” clevelandclinic, 2022. [Online]. Available: <https://my.clevelandclinic.org/health/diagnostics/10123-pet-scan>
- [118] F. A. G. Hernández, G. A. A. Audor, N. Q. Serrano, A. F. G. Robayo, J. S. G. Guillén, and N. F. T. García, “Advanced imaging techniques in preoperative planning for brain tumor resection: Evaluating the impact on surgical precision and neurological outcomes,” *Arandu UTIC*, vol. 11, no. 2, pp. 272–291, 2024.

- [119] I. Vrdoljak and L. Dravinac, “The role of ultrasound in medical diagnostics and treatment,” *ENTRENOVA-ENTerprise REsearch InNOVAtion*, vol. 10, no. 1, pp. 163–175, 2024.
- [120] C. Clinic, “Abdominal ultrasound,” 2023. [Online]. Available: <https://my.clevelandclinic.org/health/diagnostics/4994-abdominal-ultrasound>
- [121] T. C. Gerber and R. J. Gibbons, “Weighing the risks and benefits of cardiac imaging with ionizing radiation,” *JACC: Cardiovascular Imaging*, vol. 3, no. 5, pp. 528–535, 2010.
- [122] S. J. Marshall, K. M. Hicks, G. Howatson, L. Cox, and K. Thomas, “B-mode ultrasonography is a reliable and valid alternative to magnetic resonance imaging for measuring patellar tendon cross-sectional area,” *Ultrasound in Medicine & Biology*, vol. 49, no. 2, pp. 578–587, 2023.
- [123] S. Sonthalia, S. Yumeen, and F. Kaliyadan, “Dermoscopy overview and extradiagnostic applications,” in *StatPearls [Internet]*. StatPearls Publishing, 2023.
- [124] K. Warren, “How to protect your skin in the sizzling summer sun,” 2023. [Online]. Available: <https://www.laurelhc.org/tags/371-how-to-protect-your-skin-in-the-sizzling-summer-sun>
- [125] F. I. Valenzuela and M. Hohnadel, “Dermatoscopic characteristics of melanoma versus benign lesions and nonmelanoma cancers,” in *StatPearls [Internet]*. StatPearls Publishing, 2024.
- [126] A. Rizzo, N. Nami, and E. Cinotti, “Dermoscopy for non-melanocytic benign skin tumors,” in *Technology in Practical Dermatology: Non-Invasive Imaging, Lasers and Ulcer Management*. Springer, 2020, pp. 45–53.
- [127] P. Gupta, N. Rai, A. Verma, and V. Gautam, “Microscopy based methods for characterization, drug delivery, and understanding the dynamics of nanoparticles,” *Medicinal Research Reviews*, vol. 44, no. 1, pp. 138–168, 2024.
- [128] J. Irigoyen Otamendi, “Fabrication and characterization of multilayered assemblies based on polyelectrolytes and hybrid systems with carbon nanomaterials for applications in nanofiltration and as smart surfaces,” 2016.
- [129] P. B. Shanthi and K. S. Hareesha, “Automated detection and classification of cervical cancer using pap smear microscopic images: a comprehensive review and future perspectives,” *Engineered Science*, vol. 19, no. 4, pp. 20–41, 2022.
- [130] J. C. Underwood, *Introduction to biopsy interpretation and surgical pathology*. Springer Science & Business Media, 2012.
- [131] U. Najar, V. Barolle, P. Balondrade, M. Fink, C. Boccara, and A. Aubry, “Harnessing forward multiple scattering for optical imaging deep inside an opaque medium,” *Nature Communications*, vol. 15, no. 1, p. 7349, 2024.
- [132] M. Zeppieri, S. Marsili, E. S. Enaholo, A. O. Shuaibu, N. Uwagboe, C. Salati, L. Spadea, and M. Musa, “Optical coherence tomography (oct): a brief look at the uses and technological evolution of ophthalmology,” *Medicina*, vol. 59, no. 12, p. 2114, 2023.

- [133] M. Terashima, H. Kaneda, and T. Suzuki, “The role of optical coherence tomography in coronary intervention,” *The Korean journal of internal medicine*, vol. 27, no. 1, p. 1, 2012.
- [134] Y.-Y. Liu, “Automated macular pathology diagnosis in retinal oct images using multi-scale spatial pyramid and local binary patterns in texture and shape encoding,” 2011. [Online]. Available: https://sites.cc.gatech.edu/grads/y/yliu88/project/macula_pathology/
- [135] J. C. Tang, R. Magalhães, A. Wisniowiecki, D. Razura, C. Walker, and B. E. Aplegate, “Optical coherence tomography technology in clinical applications,” in *Biophotonics and Biosensing*. Elsevier, 2024, pp. 285–346.
- [136] M. Laad, R. Maurya, and N. Saiyed, “Unveiling the vision: A comprehensive review of computer vision in ai and ml,” in *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*. IEEE, 2024, pp. 1–6.
- [137] R. Archana and P. S. E. Jeevaraj, “Deep learning models for digital image processing: a review,” *Artificial Intelligence Review*, vol. 57, no. 1, p. 11, 2024.
- [138] S. Gollapudi and S. Gollapudi, *Artificial intelligence and computer vision*, 2019.
- [139] P. Krishna, S. Kollem, and B. V. D. Rao, “An optimal review of deep learning algorithms for object detection,” in *2024 IEEE Region 10 Symposium (TENSYP)*. IEEE, 2024, pp. 1–6.
- [140] K. Rabie, C. Karthik, S. Chowdhury, and P. K. Dutta, *Deep learning in medical image processing and analysis*, 2023.
- [141] B. J. Erickson, P. Korfiatis, Z. Akkus, and T. L. Kline, “Machine learning for medical imaging,” *Radiographics*, vol. 37, no. 2, pp. 505–515, 2017.
- [142] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, “A survey on deep learning in medical image analysis,” *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.
- [143] S. Srinivas and R. V. Babu, “Deep learning in neural networks: An overview,” *Computer Science*, 2015.
- [144] C. C. Aggarwal, “Neural networks and deep learning,” vol. 10, no. 978, 2018.
- [145] J. G. Meyer, “Deep learning neural network tools for proteomics,” *Cell Reports Methods*, vol. 1, no. 2, 2021.
- [146] S. Vieira, W. H. L. Pinaya, and A. Mechelli, “Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications,” *Neuroscience & Biobehavioral Reviews*, vol. 74, pp. 58–75, 2017.
- [147] A. N. Rachmayani, *Deep Learning for Medical Image Analysis*, 2015.
- [148] . Öztürk, *Convolutional neural networks for medical image processing applications*, 2022.

- [149] T. Magadza and S. Viriri, “Deep learning for brain tumor segmentation: a survey of state-of-the-art,” *Journal of Imaging*, vol. 7, no. 2, p. 19, 2021.
- [150] H. Gholamalinezhad and H. Khosravi, “Pooling methods in deep neural networks, a review,” *arXiv preprint arXiv:2009.07485*, 2020.
- [151] R. Yeh, M. Hasegawa-Johnson, and M. N. Do, “Stable and symmetric filter convolutional neural network,” in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 2652–2656.
- [152] M.-X. Wang and Y. Qu, “Approximation capabilities of neural networks on unbounded domains,” *Neural Networks*, vol. 145, pp. 56–67, 2022.
- [153] S. Sharma, S. Sharma, and A. Athaiya, “Activation functions in neural networks,” *Towards Data Science*, vol. 6, no. 12, pp. 310–316, 2017.
- [154] K. He, X. Zhang, S. Ren, and J. Sun, “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.
- [155] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, “Fast and accurate deep network learning by exponential linear units (elus),” *arXiv preprint arXiv:1511.07289*, vol. 4, no. 5, p. 11, 2015.
- [156] L. Trottier, P. Giguere, and B. Chaib-Draa, “Parametric exponential linear unit for deep convolutional neural networks,” in *2017 16th IEEE international conference on machine learning and applications (ICMLA)*. IEEE, 2017, pp. 207–214.
- [157] S. K. Roy, S. Manna, S. R. Dubey, and B. B. Chaudhuri, “Lisht: Non-parametric linearly scaled hyperbolic tangent activation function for neural networks,” in *International Conference on Computer Vision and Image Processing*. Springer, 2022, pp. 462–476.
- [158] P. Ramachandran, B. Zoph, and Q. V. Le, “Searching for activation functions,” *arXiv preprint arXiv:1710.05941*, 2017.
- [159] Y. Bai, “Relu-function and derived function review,” in *SHS web of conferences*. EDP Sciences, 2022, p. 2006.
- [160] S. Mastromichalakis, “Alrelu: A different approach on leaky relu activation function to improve neural networks performance,” *arXiv preprint arXiv:2012.07564*, 2020.
- [161] T. Jiang and J. Cheng, “Target recognition based on cnn with leakyrelu and prelu activation functions,” in *2019 international conference on sensing, diagnostics, prognostics, and control (SDPC)*. IEEE, 2019, pp. 718–722.
- [162] Y. Chang, X. Wu, and S. Zhang, “Piecewise linear units for fast self-normalizing neural networks,” in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 429–434.
- [163] N. Kyurkchiev and S. Markov, “Sigmoid functions: some approximation and modelling aspects,” *LAP LAMBERT Academic Publishing, Saarbrücken*, vol. 4, p. 34, 2015.

- [164] D. Perera, “Deep learning based human movement classification and scoring,” 2025.
- [165] K. Doshi, “Batch norm explained visually—how it works, and why neural networks need it,” 2022.
- [166] N. Bjorck, C. P. Gomes, B. Selman, and K. Q. Weinberger, “Understanding batch normalization,” *Advances in neural information processing systems*, vol. 31, 2018.
- [167] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting,” *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.
- [168] T. T. Teoh, *Convolutional neural networks for medical applications*. Springer, 2023.
- [169] B. S. Sharvani, G. JayaLakshmi, and K. Bhargavi, “Automated brain tumor diagnosis and localization using cnn and u-net 3d: A radiomics approach for enhanced medical imaging analysis,” in *2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)*. IEEE, 2024, pp. 1–6.
- [170] S. Srimathi, B. Suhashini, K. Vasanthan, N. Amirthavarshini, B. Arunkumar, B. K. Dhanush, and S. Sathesh, “Resnet-101 and faster r-cnn fusion accurate brain tumor detection and categorization,” in *2024 2nd International Conference on Artificial Intelligence and Machine Learning Applications Theme: Healthcare and Internet of Things (AIMLA)*. IEEE, 2024, pp. 1–6.
- [171] I. Analysis and T. S. Yoo, “Insight into images,” *Insight into Images*, 2004.
- [172] I. Pak and P. L. Teh, “Text segmentation techniques: a critical review,” *Innovative Computing, Optimization and Its Applications: Modelling and Simulations*, pp. 167–181, 2017.
- [173] V. V. Baligodugula, A. Ghimire, and F. Amsaad, “An overview of secure network segmentation in connected iiot environments,” *Computing & AI Connect*, vol. 1, no. 1, pp. 1–10, 2024.
- [174] S. S. Chouhan, A. Kaul, and U. P. Singh, “Image segmentation using computational intelligence techniques,” *Archives of Computational Methods in Engineering*, vol. 26, no. 3, pp. 533–596, 2019.
- [175] A. Gurung and S. L. Tamang, “Image segmentation using multi-threshold technique by histogram sampling,” *arXiv preprint arXiv:1909.05084*, 2019.
- [176] A. Chaturvedi, R. Khanna, and V. Kumar, “An analysis of region growing image segmentation schemes,” *International Journal of Computer Trends and Technology (IJCTT)*, vol. 34, no. 1, pp. 46–51, 2016.
- [177] S. H. Dawood, “A survey on various approaches for edge detection,” *Int J Sci Res Sci Eng Technol*, pp. 27–41, 2023.
- [178] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, “Image segmentation using deep learning: A survey,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 7, pp. 3523–3542, 2021.

- [179] J. Chen, H. Shao, and C. Hu, “Image segmentation based on mathematical morphological operator,” *Colorimetry and Image Processing*, 2017.
- [180] F. Bougourzi and A. Hadid, “Recent advances in medical imaging segmentation: A survey,” *arXiv preprint arXiv:2505.09274*, 2025.
- [181] N. Thangarasu, J. Sohal, B. Umamaheswari, J. Singh, R. Manjunatha, and G. C. Shelke, “Semantic segmentation for image understanding,” in *2024 15th International Conference on Computing Communication and Networking Technologies (IC-CCNT)*. IEEE, 2024, pp. 1–6.
- [182] L.-C. Chen, A. Hermans, G. Papandreou, F. Schroff, P. Wang, and H. Adam, “Masklab: Instance segmentation by refining object detection with semantic and direction features,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4013–4022.
- [183] O. Elharrouss, S. Al-Maadeed, N. Subramanian, N. Ottakath, N. Almaadeed, and Y. Himeur, “Panoptic segmentation: A review,” *arXiv preprint arXiv:2111.10250*, 2021.
- [184] N. Malingan, “Image segmentation,” scaler, 2023. [Online]. Available: <https://www.scaler.com/topics/tensorflow/image-segmentation-tensorflow/>
- [185] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, “Deep learning techniques for medical image segmentation: achievements and challenges,” *Journal of digital imaging*, vol. 32, no. 4, pp. 582–596, 2019.
- [186] B. Rajesh, M. Satyanarayana, and P. S. Vasarao, “Automated image segmentation for complex scenes using u-net architecture,” *International Journal of Research and Innovation in Applied Science*, vol. 10, no. 7, pp. 77–83, 2025.
- [187] S. Fayaz, S. Z. Ahmad Shah, N. M. ud din, N. Gul, and A. Assad, “Advancements in data augmentation and transfer learning: A comprehensive survey to address data scarcity challenges,” *Recent advances in computer science and communications (Formerly: Recent patents on computer science)*, vol. 17, no. 8, pp. 14–35, 2024.
- [188] A. Subasi, “Artificial intelligence for 3d medical image analysis,” in *Applications of Artificial Intelligence in Healthcare and Biomedicine*. Elsevier, 2024, pp. 357–375.
- [189] L. Cai, J. Gao, and D. Zhao, “A review of the application of deep learning in medical image classification and segmentation,” *Annals of translational medicine*, vol. 8, no. 11, p. 713, 2020.
- [190] D. Bank, N. Koenigstein, and R. Giryes, “Autoencoders,” in *Machine learning for data science handbook: data mining and knowledge discovery handbook*. Springer, 2023, pp. 353–374.
- [191] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, and L. Bottou, “Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion,” *Journal of machine learning research*, vol. 11, no. 12, 2010.

- [192] X. Glorot, A. Bordes, and Y. Bengio, “Deep sparse rectifier neural networks,” in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2011, pp. 315–323.
- [193] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *International conference on machine learning*. pmlr, 2015, pp. 448–456.
- [194] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, and S. Gelly, “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.
- [195] M. M. Naseer, K. Ranasinghe, S. H. Khan, M. Hayat, F. Shahbaz Khan, and M.-H. Yang, “Intriguing properties of vision transformers,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 23 296–23 308, 2021.
- [196] E. Portelance, M. C. Frank, D. Jurafsky, A. Sordoni, and R. Laroché, “The emergence of the shape bias results from communicative efficiency,” *arXiv preprint arXiv:2109.06232*, 2021.
- [197] R. Geirhos, K. Narayanappa, B. Mitzkus, T. Thieringer, M. Bethge, F. A. Wichmann, and W. Brendel, “Partial success in closing the gap between human and machine vision,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 23 885–23 899, 2021.
- [198] S. Tuli, I. Dasgupta, E. Grant, and T. L. Griffiths, “Are convolutional neural networks or transformers more like human vision?” *arXiv preprint arXiv:2105.07197*, 2021.
- [199] J. Oruh, S. Viriri, and A. Adegun, “Long short-term memory recurrent neural network for automatic speech recognition,” *IEEE Access*, vol. 10, pp. 30 069–30 079, 2022.
- [200] U. Kamath, K. Graham, and W. Emara, *Transformers for machine learning: a deep dive*. Chapman and Hall/CRC, 2022.
- [201] Y. Jiang, S. Chang, and Z. Wang, “Transgan: Two pure transformers can make one strong gan, and that can scale up,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 14 745–14 758, 2021.
- [202] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, and P. H. S. Torr, “Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 6881–6890.
- [203] L. Cheng, W. Wei, F. Zhu, Y. Liu, and C. Miao, “Geometry-entangled visual semantic transformer for image captioning,” *arXiv preprint arXiv:2109.14137*, 2021.
- [204] H.-Y. Zhou, J. Guo, Y. Zhang, L. Yu, L. Wang, and Y. Yu, “nnformer: Interleaved transformer for volumetric segmentation,” *arXiv preprint arXiv:2109.03201*, 2021.

- [205] L. Zhou, Y. Zhou, J. J. Corso, R. Socher, and C. Xiong, “End-to-end dense video captioning with masked transformer,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8739–8748.
- [206] M. Jackson, *The Oxford handbook of the history of medicine*. Oxford University Press, 2011.
- [207] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.
- [208] W. Weng and X. Zhu, “Inet: convolutional networks for biomedical image segmentation,” *IEEE Access*, vol. 9, pp. 16 591–16 603, 2021.
- [209] C. Li, H. Chen, X. Li, N. Xu, Z. Hu, D. Xue, S. Qi, H. Ma, L. Zhang, and H. Sun, “A review for cervical histopathology image analysis using machine vision approaches,” *Artificial Intelligence Review*, vol. 53, no. 7, pp. 4821–4862, 2020.
- [210] C. Yu and E. J. Helwig, “The role of ai technology in prediction, diagnosis and treatment of colorectal cancer,” *Artificial intelligence review*, vol. 55, no. 1, pp. 323–343, 2022.
- [211] Y. Dai, Y. Gao, and F. Liu, “Transmed: Transformers advance multi-modal medical image classification,” *Diagnostics*, vol. 11, no. 8, p. 1384, 2021.
- [212] M. Ghaderzadeh and F. Asadi, “Deep learning in the detection and diagnosis of covid-19 using radiology modalities: a systematic review,” *Journal of healthcare engineering*, vol. 2021, no. 1, p. 6677314, 2021.
- [213] B. Zhou, N. Dey, J. Schlemper, S. S. M. Salehi, C. Liu, J. S. Duncan, and M. Sofka, “Dsformer: A dual-domain self-supervised transformer for accelerated multi-contrast mri reconstruction,” in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2023, pp. 4966–4975.
- [214] S. Naveen and M. R. Kounte, “Optimized convolutional neural network at the iot edge for image detection using pruning and quantization,” *Multimedia Tools and Applications*, vol. 84, no. 9, pp. 5435–5455, 2025.
- [215] C. Li, L.-F. Li, X. Wang, and M.-X. Luo, “Medical image segmentation with hybrid neural networks,” *Physica Scripta*, vol. 100, no. 3, p. 36001, 2025.
- [216] Y. Wang, Y. Deng, Y. Zheng, P. Chattopadhyay, and L. Wang, “Vision transformers for image classification: A comparative survey,” *Technologies*, vol. 13, no. 1, p. 32, 2025.
- [217] S. Targ, D. Almeida, and K. Lyman, “Resnet in resnet: Generalizing residual architectures,” *arXiv preprint arXiv:1603.08029*, 2016.
- [218] B. Koonce, “Efficientnet,” in *Convolutional neural networks with swift for Tensor-flow: image recognition and dataset categorization*. Springer, 2021, pp. 109–123.
- [219] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, “Transformers in vision: A survey,” *ACM computing surveys (CSUR)*, vol. 54, no. 10s, pp. 1–41, 2022.

- [220] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.
- [221] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, “Transunet: Transformers make strong encoders for medical image segmentation,” *arXiv preprint arXiv:2102.04306*, 2021.
- [222] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, “Unetr: Transformers for 3d medical image segmentation,” in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 574–584.
- [223] W. Chen, R. Zhang, Y. Zhang, F. Bao, H. Lv, L. Li, and C. Zhang, “Pact-net: Parallel cnns and transformers for medical image segmentation,” *Computer Methods and Programs in Biomedicine*, vol. 242, p. 107782, 2023.
- [224] X. Ye, J. Pan, J. Chen, and J. Zhang, “A multiscale feature fusion-guided lightweight semantic segmentation network,” *Journal of Field Robotics*, vol. 42, no. 1, pp. 272–286, 2025.
- [225] C. Wu, F. Wu, S. Ge, T. Qi, Y. Huang, and X. Xie, “Neural news recommendation with multi-head self-attention,” in *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, 2019, pp. 6389–6394.
- [226] F. Tian and L. Ma, “A hybrid cnn-transformer model for efficient hyperspectral image reconstruction,” in *2024 IEEE 6th International Conference on Power, Intelligent Computing and Systems (ICPICS)*. IEEE, 2024, pp. 732–736.
- [227] M. Escobar, L. Daza, C. González, J. Pont-Tuset, and P. Arbeláez, “Video swin transformers for egocentric video understanding@ ego4d challenges 2022,” *arXiv preprint arXiv:2207.11329*, 2022.
- [228] L. Meng, M. Goodwin, A. Yazidi, and P. Engelstad, “Deep reinforcement learning with swin transformers,” in *Proceedings of the 2024 8th international conference on digital signal processing*, 2024, pp. 205–211.
- [229] B. Huang, J. Ying, R. Lyu, N. S. Schaadt, B. M. Klinkhammer, P. Boor, J. Lotz, F. Feuerhake, and D. Merhof, “Utnetpara: A hybrid cnn-transformer architecture with multi-scale fusion for whole-slide image segmentation,” in *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2024, pp. 1–5.
- [230] X. Wang, M. Shao, D. Guo, Y. Cui, X. Huang, M. Xia, and C. Bai, “Multi-stage aggregation transformer for medical image segmentation,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [231] C. G. A. R. Network, “Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas,” *New England Journal of Medicine*, vol. 372, no. 26, pp. 2481–2498, 2015.

- [232] M. H. Vu, G. Grimbergen, T. Nyholm, and T. Löfstedt, “Evaluation of multislice inputs to convolutional neural networks for medical image segmentation,” *Medical Physics*, vol. 47, no. 12, pp. 6216–6231, 2020.
- [233] P. S. Kumar, V. P. Sakthivel, M. Raju, and P. D. Sathya, “Brain tumor segmentation of the flair mri images using novel resunet,” *Biomedical Signal Processing and Control*, vol. 82, p. 104586, 2023.
- [234] P. K. Sathish and G. Raju, “Post-operative brain mri resection cavity segmentation model and follow-up treatment assistance,” *International Journal of Online & Biomedical Engineering*, vol. 20, no. 5, 2024.
- [235] A. Shomirov, J. Zhang, and M. M. Billah, “Brain tumor segmentation of hgg and lgg mri images using wfl-based 3d u-net,” *Journal of Biomedical Science and Engineering*, vol. 15, no. 10, pp. 241–260, 2022.
- [236] M. Renugadevi, K. Narasimhan, C. V. Ravikumar, R. Anbazhagan, G. Pau, K. Ramkumar, M. Abbas, N. Raju, K. Sathish, and P. Sevugan, “Machine learning empowered brain tumor segmentation and grading model for lifetime prediction,” *IEEE Access*, vol. 11, pp. 120 868–120 880, 2023.
- [237] M. Joshi and B. K. Singh, “Proportion estimation and multi-class classification of abnormal brain cells,” *Medinformatics*, vol. 1, no. 2, pp. 79–90, 2024.
- [238] K. Kashyap, M. Garg, S. Fargose, and S. Nair, “Dynamic neural style transfer for artistic image generation using vgg19,” *arXiv preprint arXiv:2501.09420*, 2025. [Online]. Available: <https://doi.org/10.48550/arXiv.2501.09420>
- [239] L. Gatys, A. S. Ecker, and M. Bethge, “Texture synthesis using convolutional neural networks,” in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds., vol. 28. Curran Associates, Inc., 2015. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2015/file/a5e00132373a7031000fd987a3c9f87b-Paper.pdf
- [240] D. Im, D. Han, S. Choi, S. Kang, and H.-J. Yoo, “Dt-cnn: An energy-efficient dilated and transposed convolutional neural network processor for region of interest based image segmentation,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 67, no. 10, pp. 3471–3483, 2020.
- [241] A. Baldrich Salvadó, “Smoke plume segmentation of wildfire images,” Master’s thesis, Universitat Politècnica de Catalunya, 2023.
- [242] Y. Cheng, D. Wang, P. Zhou, and T. Zhang, “A survey of model compression and acceleration for deep neural networks,” *arXiv preprint arXiv:1710.09282*, 2017.
- [243] R. Chabiniok, V. Y. Wang, M. Hadjicharalambous, L. Asner, J. Lee, M. Sermesant, E. Kuhl, A. A. Young, P. Moireau, M. P. Nash, D. Chapelle, and D. A. Nordsletten, “Multiphysics and multiscale modelling, data-model fusion and integration of organ physiology in the clinic: ventricular cardiac mechanics,” *Interface Focus*, vol. 6, no. 2, p. 20150083, 2016. [Online]. Available: <https://doi.org/10.1098/rsfs.2015.0083>

- [244] S. Chen, C. Ding, and M. Liu, “Dual-force convolutional neural networks for accurate brain tumor segmentation,” *Pattern Recognition*, vol. 88, pp. 90–100, 2019. [Online]. Available: <https://doi.org/10.1016/j.patcog.2018.11.009>
- [245] S. Ravi and A. M. Khan, “A Study on Noise Filters to Pre-Process Magnetic Resonant Biomedical Images for Segmentation,” *Research Journal of Recent Sciences*, vol. 4, pp. 51–56, 2015.
- [246] A. K. Nugroho, M. E. Wibowo, and H. Soebono, “The Effect of Preprocessing on Skin Lesion Segmentation,” in *2024 8th International Conference on Information Technology (InCIT)*. IEEE, 2024, pp. 455–460.
- [247] F. Cabello, J. León, Y. Iano, and R. Arthur, “Implementation of a fixed-point 2D Gaussian Filter for Image Processing based on FPGA,” in *2015 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*. IEEE, 2015, pp. 28–33.
- [248] P.-Y. Hsiao, S.-S. Chou, and F.-C. Huang, “Generic 2-D gaussian smoothing filter for noisy image processing,” in *TENCON 2007-2007 IEEE Region 10 Conference*. IEEE, 2007, pp. 1–4.
- [249] M. Kirchner and J. Fridrich, “On detection of median filtering in digital images,” in *Media forensics and security II*, vol. 7541. SPIE, 2010, pp. 371–382.
- [250] G. George, R. M. Oommen, S. Shelly, S. S. Philipose, and A. M. Varghese, “A survey on various median filtering techniques for removal of impulse noise from digital image,” in *2018 conference on emerging devices and smart systems (ICEDSS)*. IEEE, 2018, pp. 235–238.
- [251] R. Priyah and S. Kamalakkannan, “Contrast Limited Adaptive Histogram Equalization Approach Contrast Enhancement in Liver Tumor,” in *2025 9th International Conference on Inventive Systems and Control (ICISC)*. IEEE, 2025, pp. 1–7.
- [252] K. Kokufuta and T. Maruyama, “Real-time processing of contrast limited adaptive histogram equalization on FPGA,” in *2010 international conference on field programmable logic and applications*. IEEE, 2010, pp. 155–158.
- [253] Y. Hu, N. Mu, L. Liu, L. Zhang, J. Jiang, and X. Li, “Slimmable transformer with hybrid axial-attention for medical image segmentation,” *Computers in biology and medicine*, vol. 173, p. 108370, 2024.
- [254] R. Xie, L. Jiang, X. He, Y. Pan, and Y. Cai, “A weakly supervised and globally explainable learning framework for brain tumor segmentation,” in *2024 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2024, pp. 1–6.
- [255] U. C. Das, W. Benjapolakul, M. Gupta, C. Somthawinpongsai, T. Vitoonpong, P. Suanpang, S. Butdisuwan, and A. Nanthamornphong, “Brain cancer tumor detection by u-net deep learning algorithm from mri images,” in *2024 12th International Conference on Internet of Everything, Microwave, Embedded, Communication and Networks (IEMECON)*. IEEE, 2024, pp. 1–6.
- [256] W. Ullah, H. Naveed, and S. Ali, “Deep learning for precise mri segmentation of lower-grade gliomas,” *Sustain. Mach. Intell. J.*, vol. 10, pp. 23–36, 2025.

- [257] M. A. Naser and M. J. Deen, “Brain tumor segmentation and grading of lower-grade glioma using deep learning in mri images,” *Computers in biology and medicine*, vol. 121, p. 103758, 2020.
- [258] N. Heller, N. Sathianathen, A. Kalapara, E. Walczak, K. Moore, H. Kaluzniak, J. Rosenberg, P. Blake, Z. Rengel, and M. Oestreich, “The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes,” *arXiv preprint arXiv:1904.00445*, 2019.
- [259] V. Ficarra, G. Novara, S. Secco, V. Macchi, A. Porzionato, R. De Caro, and W. Artibani, “Preoperative aspects and dimensions used for an anatomical (padua) classification of renal tumours in patients who are candidates for nephron-sparing surgery,” *European urology*, vol. 56, no. 5, pp. 786–793, 2009.
- [260] J. D. L. Araujo, L. B. da Cruz, J. O. B. Diniz, J. L. Ferreira, A. C. Silva, A. C. de Paiva, and M. Gattass, “Liver segmentation from computed tomography images using cascade deep learning,” *Computers in biology and medicine*, vol. 140, p. 105095, 2022.
- [261] D. Veiga-Canuto, L. Cerdà-Alberich, C. Sangüesa Nebot, B. Martínez de Las Heras, U. Pötschger, M. Gabelloni, J. M. Carot Sierra, S. Taschner-Mandl, V. Düster, and A. Cañete, “Comparative multicentric evaluation of inter-observer variability in manual and automatic segmentation of neuroblastic tumors in magnetic resonance images,” *Cancers*, vol. 14, no. 15, p. 3648, 2022.
- [262] P. Bilic, P. Christ, H. B. Li, E. Vorontsov, A. Ben-Cohen, G. Kaissis, A. Szeskin, C. Jacobs, G. E. H. Mamani, and G. Chartrand, “The liver tumor segmentation benchmark (lits),” *Medical image analysis*, vol. 84, p. 102680, 2023.
- [263] O. I. Alirr, “Deep learning and level set approach for liver and tumor segmentation from ct scans,” *Journal of Applied Clinical Medical Physics*, vol. 21, no. 10, pp. 200–209, 2020.
- [264] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” in *European conference on computer vision*. Springer, 2020, pp. 213–229.
- [265] D. Chen, B. Zhao, Y. Zhang, and M. Zhao, “Selective complementary feature fusion and modal feature compression interaction for brain tumor segmentation,” *arXiv preprint arXiv:2503.16149*, 2025.
- [266] F. Moodi, F. K. Shoushtari, G. Valizadeh, D. Mazinani, H. M. Salari, and H. S. Rad, “Attention xception unet (axunet): A novel combination of cnn and self-attention for brain tumor segmentation,” *arXiv preprint arXiv:2503.20446*, 2025.
- [267] Y. Azzi, A. Moussaoui, and M.-T. Kechadi, “Pu-net deep learning architecture for gliomas brain tumor segmentation in magnetic resonance images,” *Image Analysis and Stereology*, vol. 42, no. 3, pp. 197–206, 2023.
- [268] M. Akil, R. Saouli, R. Kachouri *et al.*, “Fully automatic brain tumor segmentation with deep learning-based selective attention using overlapping patches and multi-class weighted cross-entropy,” *Medical image analysis*, vol. 63, p. 101692, 2020.

- [269] Y. Lyu and X. Tian, “Mwg-unet++: Hybrid transformer u-net model for brain tumor segmentation in mri scans,” *Bioengineering*, vol. 12, no. 2, 2025.
- [270] L. Yang, Q. Dong, D. Lin, C. Tian, and X. Lü, “Munet: a novel framework for accurate brain tumor segmentation combining unet and mamba networks,” *Frontiers in Computational Neuroscience*, vol. 19, p. 1513059, 2025.
- [271] S. S. Ziaee, F. Maleki, and K. Ovens, “Rel-unet: Reliable tumor segmentation via uncertainty quantification in nnu-net,” in *2025 11th International Conference on Computing and Artificial Intelligence (ICCAI)*. IEEE, 2025, pp. 1–8.
- [272] P. Sun, Z. Mo, F. Hu, F. Liu, T. Mo, Y. Zhang, and Z. Chen, “Kidney tumor segmentation based on fr2pattu-net model,” *Frontiers in oncology*, vol. 12, p. 853281, 2022.
- [273] F. Xu, F. Yang, X. Li, and X. Zhang, “A unified loss for handling inter-class and intra-class imbalance in medical image segmentation,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 8, 2025, pp. 8842–8850.
- [274] Z. Chen, M. Dou, X. Luo, and Y. Yao, “Enhanced liver and tumor segmentation using a self-supervised swin-transformer-based framework with multitask learning and attention mechanisms,” *Applied Sciences*, vol. 15, no. 7, p. 3985, 2025.
- [275] P. Lyu, W. Liu, T. Lin, J. Zhang, Y. Liu, C. Wang, and J. Zhu, “Semi-supervised segmentation of abdominal organs and liver tumor: uncertainty rectified curriculum labeling meets x-fuse,” *Machine Learning: Science and Technology*, vol. 5, no. 2, p. 25047, 2024.
- [276] J. Liao, H. Wang, H. Gu, and Y. Cai, “Liver tumor segmentation method combining multi-axis attention and conditional generative adversarial networks,” *PloS one*, vol. 19, no. 12, p. e0312105, 2024.
- [277] L. Sun, L. Jiang, M. Wang, Z. Wang, and Y. Xin, “A multi-scale liver tumor segmentation method based on residual and hybrid attention enhanced network with contextual integration,” *Sensors*, vol. 24, no. 17, p. 5845, 2024.