Enhancing Medical Image Analysis through Deep Learning: A Comprehensive Study on Classification, Segmentation, and Multitask Learning

Susmita Ghosh
Electronics and Communication Sciences Unit
Indian Statistical Institute

Supervised by
Prof. (Dr.) Swagatam Das
Electronics and Communication Sciences Unit
Indian Statistical Institute

A thesis submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
in Computer Science

May, 2025

This is to certify that the thesis entitled

Enhancing Medical Image Analysis through Deep Learning: A Comprehensive Study on Classification, Segmentation, and Multitask Learning

submitted by

Susmita Ghosh

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Computer Science

has been examined and approved.

Prof. (Dr.) Swagatam Das
Supervisor
Electronics and Communication Sciences Unit
Indian Statistical Institute, Kolkata

To my family.

Acknowledgements

This thesis would not have been possible without the guidance, support, and encouragement of many people, and I am deeply grateful to all those who have contributed to this journey.

First and foremost, I would like to express my sincere gratitude to my advisor, Prof. Swagatam Das, for his invaluable guidance, insight, and patience throughout this research. His expertise and encouragement have been instrumental in helping me navigate challenges and deepen my understanding of my research field. Without his support, this work would not have reached its full potential.

I am also profoundly grateful to the faculty members of the Electronics and Communication Sciences Unit, whose knowledge and dedication have greatly influenced my academic journey. I am especially grateful for the supportive and collaborative environment they foster, which has been essential to my growth as a researcher. I would like to extend my gratitude to the staff of the Electronics and Communication Sciences Unit, as well as the teams in the Dean's and Director's offices, for their steadfast support and assistance. I owe a special thanks to my fellow researchers who offered both intellectual and personal support. Their camaraderie, feedback, and collaboration made this journey far more enjoyable and enlightening. I would also like to extend my heartfelt gratitude to my co-author, whose expertise, collaboration, and dedication have greatly enriched this work.

A heartfelt thanks to my family for their unwavering love, boundless patience, and steady encouragement. Your belief in me has been the quiet strength beneath every step of this journey, granting me the courage and conviction to pursue my dreams. To my husband and daughter, who showed enduring patience and understanding during the many hours I was absorbed in this work, thank you for

your gentle reminders of love and warmth that anchored me through it all. This journey is as much yours as it is mine.

To everyone who contributed to this thesis in ways big and small, thank you.

Susmita Ghosh

Susmita Ghosh
Author
Electronics and Communication Sciences Unit
Indian Statistical Institute, Kolkata

Abstract

Medical image analysis has become indispensable for accurate diagnosis and treatment planning. However, despite advances in deep learning, several critical challenges persist, ranging from more efficient models to the integration of multiple tasks within a unified framework. This thesis addresses these challenges by proposing innovative deep learning architectures that enhance medical image classification, segmentation, and multitask learning. At the heart of this research is the goal of developing models that deliver high performance and tackle the nuanced complexities of medical data. Existing classification models often overlook valuable information hidden in the spectral domain of images. I address this by integrating spatial and spectral features, demonstrating their complementary power to detect diseases such as COVID-19 from chest radiographs. This approach facilitates a more holistic understanding of medical images, improving the accuracy and reliability of diagnostic systems. To further enhance image classification, I explore hybrid architectures that combine convolutional and transformer-based models. These models leverage the strengths of both architectures, capturing fine-grained visual details and long-range dependencies. This significantly improves various medical imaging datasets, offering deeper interpretability and superior classification accuracy, particularly in complex diagnostic scenarios. Moving beyond classification, I tackle the fundamental challenge of segmenting complex and irregular regions within medical images, where traditional deep learning models often struggle. To overcome this, I introduce a novel segmentation framework that combines the power of deep neural networks with trainable morphological operations. This leads to a more precise delineation of regions of interest, even in challenging clinical scenarios, setting a new benchmark for medical image segmentation. One of the most pressing issues in medical imaging is the inefficiency of current multitask learning models, which often require vast computational resources and struggle to generalize across different tasks. I present a lightweight multitask learning framework that excels at both segmentation and classification, particularly in breast tumor analysis. Using novel morphological attention mechanisms and the sharing of task-specific knowledge, proposed model significantly reduces computational complexity while improving performance. Importantly, this framework demonstrates versatility across var-

ious medical imaging domains, from gland segmentation and malignancy detection in histology images to skin lesion analysis, demonstrating its robustness and applicability in real-world settings. Altogether, this thesis offers solutions to some of the most pressing problems in medical image analysis, providing models that are not only more accurate but also computationally efficient, making them suitable for deployment in clinical practice.

Contents

List of Figures

List of Tables

List of Notations

H	Height of image.
W	Width of image.
C	Number of channels in image or feature map.
$\mathbb{X}$	A set of N images
X	An image of dimension $H \times W \times C$ belonging to $\mathbb{X}$.
N_c	Number of classes in classification task.
$\mathcal{Y}$	Set of class labels $\{y_1, y_2, \ldots, y_{N_c}\}$ for classification task.
$\hat{y}$	Class prediction where $\hat{y} \in \mathcal{Y}$.
N_s	Number of classes in segmentation task.
$\mathbb{Z}$	A set of masks corresponding to N images in $\mathbb{X}$.
Z	A mask of dimension $H \times W \times N_s$ belonging to $\mathbb{Z}$.
$\hat{Z}$	Predicted mask of dimension $H \times W \times N_s$ belonging to $\mathbb{Z}$.
$f_\theta(\cdot)$	Learning function parameterized by θ for classification task.
$g_\theta(\cdot)$	Learning function parameterized by θ for segmentation task.
$\mathcal{L}_{classification}$	Loss for classification task.
$\mathcal{L}_{segmentation}$	Loss for segmentation task.
N_{En}	Number of encoder stages.
N_{De}	Number of decoder stages.
K_{SE}	Size of structuring elements.
W_{SE}	Structuring element of size $K_{SE} \times K_{SE} \times C$.
S	Stride.
r	Dilation rate.
$\mathrm{DD}^r(f)$	A functional used to perform *Dilation* morphological operation with dilation rate r.
$\mathrm{DE}^r(f)$	A functional used to perform *Erosion* morphological operation with dilation rate r.
$\mathrm{DO}^r(f)$	A functional used to perform *Opening* morphological operation with dilation rate r
$\mathrm{DC}^r(f)$	A functional used to perform *Closing* morphological operation with dilation rate r
$\mathrm{DG}^r(f)$	A functional used to perform *Gradient* morphological operation with dilation rate r

$ReLU(f)$	A functional used to perform Rectified Linear Unit.
$BN(f)$	A functional used to perform Batch normalization.
$Conv2D(f)$	A functional used to perform 2D convolution operation.
$DropOut(f)$	A functional used to perform dropout operation.
h_0	Half band low pass filter.
h_1	Half band high pass filter.
C_{ij}	Number of samples of i^{th} class predicted as j^{th} class.
Q	Query vector in multihead self attention.
V	Value vector in multihead self attention.
K	Key in vector multihead self attention.
d_k	length of Q, K and V vector in multihead self attention.
W_q	Weight matrix for query.
W_v	Weight matrix for value.
W_k	Weight matrix for key.
$\phi_{En}(\cdot)$	Set of weights for the encoder of UNet.
$\phi_{De}(\cdot)$	Set of weights for the decoder of UNet.
$\phi_{MSMM_i}(\cdot)$	Set of weights for the i^{th} MSMM of UNet.
K_{SE}	Structuring elements of dimension $K \times K \times C$.
$\Vert_{c=1}^{C} \cdot$	Channel-wise concatenation of features.
η	Learning rate.
W_{SMA}	Weight matrix for spatial morphological attention block
W_{CA}	Weight matrix for channel attention block
$\odot$	Element-wise multiplication.
λ	Weight for segmentation loss.
σ	Nonlinear activation function.
α_c	Weighing Factor for class c
$p_{t,c}$	Probability for class c.
η	Degree of down-weighting of easy-to-classify pixels.

List of Abbreviations and Acronyms

ACSA	Average Class Specific Accuracy.
ACSF	Average Class Specific F1-Score.
ACSP	Average Class Specific Precision.
ACSR	Average Class Specific Recall.
AUROC	Area Under Receiver Operating Characteristics.
BCELoss	Binary Cross Entropy Loss.
BN	Batch Normalization.
C-SMA	Channel and Spatial Morphological Attention.
CA	Channel Attention.
CBAM	Convolutional Block Attention Module.
CNN	Convolutional Neural Network
DCT	Discrete cosine Transform.
DS	Dice Score
DS_{object}	Object-level Dice Score.
DWT	Discrete Wavelet Transform.
ES-MTL	Encoder Shared Multitask Learning.
FC	Fully Connected Layer.
fn	False Negative.
fp	False Positive.
GT	Ground Truth.
HD	Hausdorff Distance.
HD95	95th percentile of Hausdorff Distance.
IoU	Intersection over Union.
IR	Imbalance Ratio.
MADF	Mean Absolute Deviation of F1-score.
MHSA	Multihead Self Attention.
MLP	Multi-Layer Perceptron.
MSMCM	Multi-scale Morphological Closing Module.
MSMGM	Multi-scale Morphological Gradient Module.
MSMM	Multi-scale Morphological Module.
MSMOM	Multi-scale Morphological Opening Module.

MTL	Multitask Learning.
PA	Pixel Accuracy.
ReLU	Rectified Linear Unit.
ReMixViT	Vision Transformer with reordered residual block and MLP-Mixer block.
ResViT	Hybrid architecture.
ReViT	Vision Transformer with MLP-Mixer block.
ReViT	Vision Transformer with reordered residual block.
ROI	Region of Intereset.
RRI	Ratio of ROIs to image
SE	Structuring Element.
SMA	Spatial Morphological Attention.
SVM	Support Vector Machine.
tn	True Negative.
tp	True Positive.
ViT	Vision Transformer.

Chapter 1

Foundations of Deep Learning in Medical Imaging: A Prelude

Summary

Medical imaging has become an indispensable tool in modern healthcare, providing detailed insights that facilitate early diagnosis, treatment planning, and patient monitoring. Deep learning, particularly in image classification and segmentation, has emerged as a powerful approach to automate and improve the interpretation of medical images, thus supporting clinical decision-making. This thesis is dedicated to advancing the application of deep learning techniques in medical image analysis, specifically focusing on disease detection and region of interest (ROI) segmentation. The work is structured into four contributory chapters, each addressing a critical aspect of medical image processing. The first two contributory chapters explore medical image classification, mainly focusing on disease detection. These chapters develop and evaluate state-of-the-art methods to classify medical images to identify various diseases accurately. The third chapter focuses on image segmentation, a process that delineates regions of interest within medical images. Accurate segmentation is vital to quantifying and analyzing anatomical structures and pathological areas, which are critical in treatment planning and monitoring. The methods proposed in this chapter aim to improve the precision and reliability of medical image segmentation. The final contributory chapter introduces a multi-task learning framework that combines disease detection and ROI segmentation. Multi-task learning leverages shared representations and complementary tasks to improve the model's overall performance. This chapter demonstrates how integrating these tasks can lead to more robust and efficient models, providing a comprehensive tool for medical image analy-

sis. Through these contributions, this thesis aims to push the boundaries of automated medical image analysis, offering novel solutions to challenges in disease detection, image segmentation, and integration of these tasks. The results of this work have the potential to significantly impact clinical practices, leading to better patient care and more informed decision-making.

1.1 Background and Motivation

1.1.1 Overview of Medical Imaging

Medical imaging is indispensable in modern healthcare and pivotal in disease diagnosis, treatment planning, and patient monitoring. It allows clinicians to visualize the internal structures and functions of the body non-invasively, facilitating the detection of abnormalities and enabling informed medical decisions. Medical imaging techniques enable the identification of conditions ranging from infections to tumors and cardiovascular diseases. The ability to diagnose diseases at an early stage significantly improves patient outcomes, as it allows timely intervention and treatment. Accurate imaging is essential for designing effective treatment plans. For example, in cancer care, imaging guides radiation therapy by delineating tumor boundaries, ensuring that radiation is precisely targeted to the affected area while protecting healthy tissues. Similarly, in surgery, imaging informs the surgeon about the exact location and extent of the pathology, reducing the risk of complications and improving surgical outcomes. Medical imaging is also vital for monitoring the progression of diseases and the effectiveness of treatments. For example, repeated imaging allows clinicians to track changes in the size or shape of a tumor. Thus, medical imaging is a cornerstone of patient care, supporting clinicians in making accurate diagnoses, optimizing treatment strategies, and improving overall patient outcomes.

Medical imaging encompasses diverse techniques, each serving distinct purposes in healthcare. Figure 1.1 shows a few examples of medical images collected through various imaging techniques. X-ray imaging is one of the most common methods, producing 2D images of the body's internal structures, and is often used to detect fractures or lung conditions. Ultrasound uses high-frequency sound waves to create real-time images of soft tissues and organs, which makes it invaluable in obstetrics, cardiology, and abdominal imaging. Dermoscopy is a specialized technique used in dermatology to examine skin lesions and diagnose conditions such as melanoma by providing magnified images of the skin's surface. Histopathology involves the microscopic examination of biopsied tissue samples, allowing a detailed analysis of cellular structures to diagnose diseases such as cancer. Magnetic Resonance Imaging (MRI) uses strong magnetic fields and radio waves to generate detailed images of soft tissues, including the brain, muscles, and joints. It is beneficial in neurological and musculoskeletal

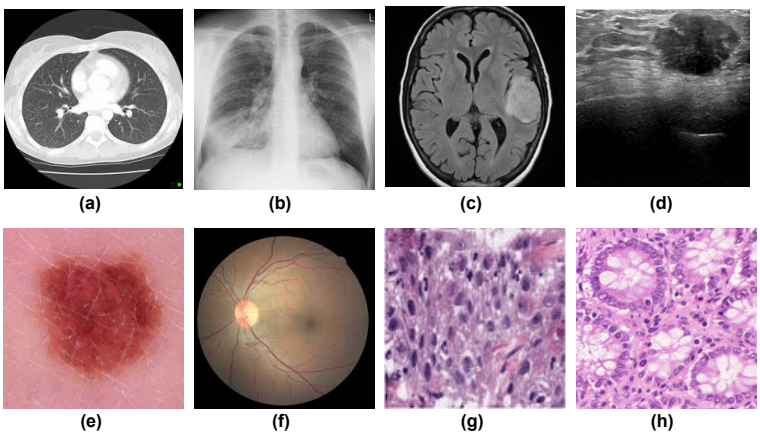

Figure 1.1: Examples of various medical images — (a) CT scan of lung, (b) chest x-ray, (c) Brain MRI, (d) Ultrasound images of breast, (e) Dermoscopic image, (f) Fundus image, (g) histological image of colorectal tissue, and (h) additional histological image of colorectal tissue.

studies. Mammography is a specialized X-ray technique designed to detect breast cancer by capturing detailed images of breast tissue. Fundus photography captures images of the interior surface of the eye, including the retina, optic disc, and blood vessels. It is essential to diagnose and monitor conditions such as diabetic retinopathy and glaucoma. Each imaging technique plays a critical role in the early detection, diagnosis, and management of various medical conditions, contributing to improved patient care across different specialties.

1.1.2 Deep Learning in Medical Image Analysis

Machine learning has transformed medical image analysis by automating and improving tasks traditionally reliant on expert interpretation. Key tasks include disease identification, tumor and lesion delineation, tissue segmentation at the microscopic level, disease progression prediction, image registration, etc (Figure 1.2). In traditional approaches, features are manually extracted from medical images and fed into machine learning algorithms for diagnosis. Deep learning, however, enables fully automated frameworks, where models learn the relevant features during training, resulting in more accurate disease detection, classification, and segmentation.

At the forefront of deep learning models are Convolutional Neural Networks (CNNs), which have revolutionized image processing due to their ability to capture spatial hierarchies

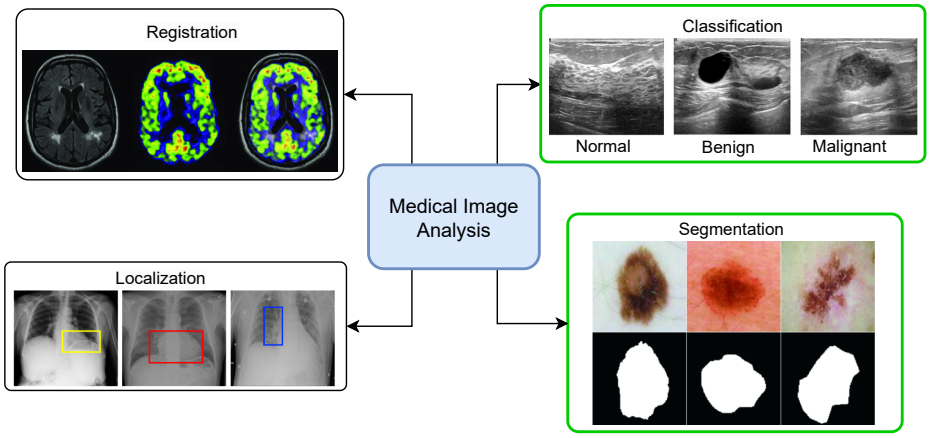

Figure 1.2: Different application of deep learning algorithms on medical images, with tasks highlighted in green boxes representing the focus of this thesis.

in images automatically. CNNs are designed to recognize patterns through convolutional layers that scan an image, identifying essential features like edges, textures, and complex shapes, such as tumors or lesions, as the layers deepen. This hierarchical learning approach allows CNNs to effectively handle the intricate details of medical images, learning from pixel-level information to more abstract features.

Deep learning methods have excelled in identifying abnormalities like tumors and lesions, often achieving higher accuracy than human experts by capturing pixel values and spatial context. These models provide consistent, reliable results, reducing the variability seen in human interpretation. Deep learning has become an indispensable tool in modern medical imaging, advancing precision medicine and enhancing healthcare quality.

In this thesis, I primarily address the detection of underlying medical conditions in clinical images and delineate regions of interest (ROIs) within these images. Detecting medical conditions is framed as a classification problem, where each class corresponds to a distinct medical condition. The goal is to assign a given input image to one of several predefined categories corresponding to the presence or absence of a specific medical condition. Additionally, segmentation of ROIs in medical imaging is posed as a pixel-wise classification problem by treating each pixel in the image as an individual data point that needs to be classified into one of several classes, such as the ROI (e.g., a tumor, lesion, or organ) or the background. In the following sections, I provide a detailed definition of the image classification and segmentation

tasks, specifically in the context of medical image analysis.

1.1.3 The Classification Problem

In medical imaging, a classification problem involves assigning a label or category to an entire medical image (or a specific region within the image) based on its content. The goal is to determine the presence or absence of a particular disease or condition.

Let $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$ represent a medical image, where H and W are the height and width of the image, respectively, and C is the number of channels (e.g., 3 for RGB images). The classification problem can be defined as learning a function $f_\theta(\mathbf{X})$ parameterized by θ, that maps the input image $\mathbf{X}$ to a discrete set of class labels $\mathcal{Y} = \{y_1, y_2, \ldots, y_{N_c}\}$, where N_c is the number of possible classes (e.g., *disease* or *no disease*).

Mathematically, the classification problem is expressed as:

$$\hat{y} = f_\theta(\mathbf{X}),$$

where $\hat{y} \in \mathcal{Y}$ is the predicted label for the input image $\mathbf{X}$.

The objective is to learn the parameters θ that minimize a loss function $\mathcal{L}_{\text{classification}}(\hat{y}, y)$, where y is the actual label of the image.

1.1.4 The Segmentation Problem

In medical imaging, a segmentation problem involves partitioning an image into meaningful regions, typically by assigning a label to each pixel in the image. The goal is to identify and delineate specific anatomical structures or pathological areas (e.g., tumors) within the image.

Let $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$ again represent a medical image, and let $\mathbf{Z} \in \mathcal{Z}^{H \times W}$ represent the corresponding segmentation mask, where each element $\mathbf{Z}(i, j)$ in the mask belongs to a set of labels $\mathcal{Z} = \{z_1, z_2, \ldots, z_{N_s}\}$, and N_s is the number of classes for segmentation (e.g., '*background*, *tumor*, *organ*).

The segmentation problem can be defined as learning a function $g_\theta(\mathbf{X})$ parameterized by θ, that maps the input image $\mathbf{X}$ to a segmentation mask $\mathbf{Z}$.

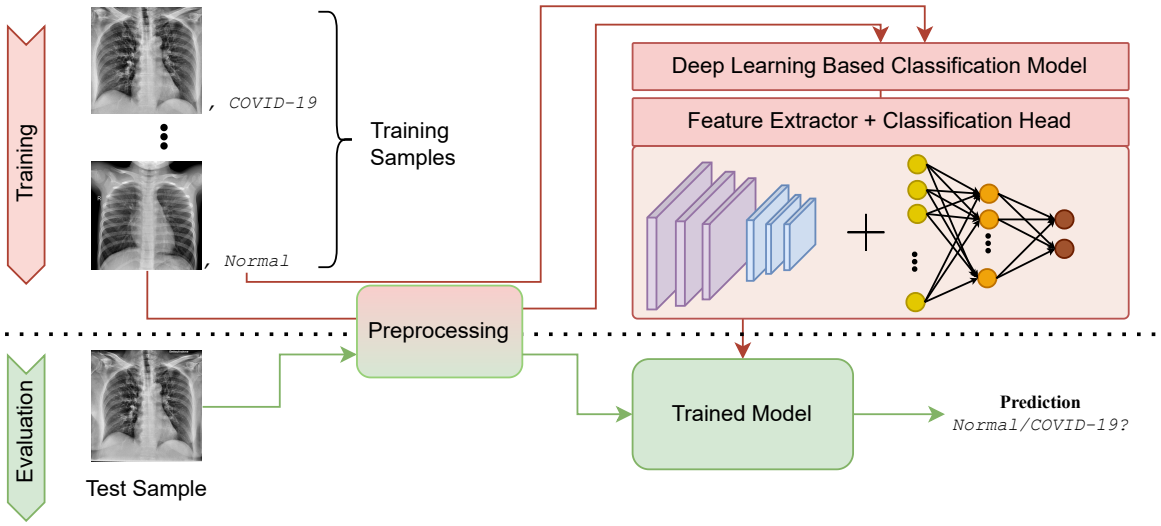

Figure 1.3: A general framework for deep learning-based classification in medical image analysis, outlining the stages of data preprocessing, model training using the training dataset, and final classification. Additionally, the framework illustrates the use of a separate test dataset for evaluating the performance of the trained model.

Mathematically, the segmentation problem is expressed as:

$$\hat{\mathbf{Z}} = g_\theta(\mathbf{Z}),$$

where $\hat{\mathbf{Z}} \in \mathcal{Z}^{H \times W}$ is the predicted segmentation mask for the input image $\mathbf{X}$.

The objective is to learn the parameters θ that minimize a loss function $\mathcal{L}_{\text{segmentation}}(\hat{\mathbf{Z}}, \mathbf{Z})$, where $\mathbf{Z}$ is the true segmentation mask.

1.1.5 Challenges in medical image analysis using deep learning algorithms

Deep learning has revolutionized medical image analysis, offering unprecedented capabilities in detecting, classifying, and segmenting complex medical images with high accuracy. Its ability to automatically learn and extract relevant features from vast data has made it a powerful tool in various clinical applications. However, despite these advancements, significant challenges remain that need to be addressed to realize its potential in healthcare fully. The challenges faced while using deep learning algorithms for medical image classification and segmentation problems are listed below.

- **Data Scarcity:** Medical image datasets are typically limited in size due to strin-

gent ethical regulations and privacy concerns. While classification tasks require image-level labels, segmentation demands pixel-level annotations, which are significantly more labor-intensive. Expert clinicians or radiologists must manually annotate each pixel, which is costly, time-consuming, and prone to human error. This scarcity of labeled data poses a significant challenge for training both classification and segmentation models, hindering their ability to accurately identify complex anatomical structures and generalize effectively across diverse medical conditions.

- **Class Imbalance:** Data imbalance is a common medical image analysis issue for classification and segmentation tasks. In classification, certain medical conditions or disease states may be rare, resulting in significantly fewer examples than more common ones. This imbalance makes models biased toward the majority class, potentially overlooking critical but infrequent cases. In segmentation, the problem is even more pronounced, as regions of interest (ROIs), such as tumors or lesions, often occupy only a small portion of the image compared to the background. This pixel-level imbalance makes it difficult for models to learn fine-grained details of ROIs, leading to poor performance in identifying smaller or less frequent anatomical structures. Effective handling of this imbalance is crucial for improving the accuracy and reliability of deep learning models in medical imaging.

- **High Intra-class Variability and Low Inter-class Variability:** High intra-class variability refers to the diverse appearance of the same medical condition across different patients or imaging conditions. This could arise due to variations in anatomy, image acquisition protocols, or disease progression, making it difficult for models to recognize a condition within the same class consistently. On the other hand, low inter-class variability refers to the similarity between different medical conditions, where distinct diseases or healthy tissue may exhibit overlapping visual characteristics. This makes it challenging for models to distinguish between classes, potentially leading to misclassification. These challenges complicate accurate diagnosis, requiring sophisticated models to capture subtle differences and generalize across varied patient data. Additionally, domain shift — such as changes in image contrast or scanner characteristics during deployment that differ from those seen during training — can exacerbate intra-class

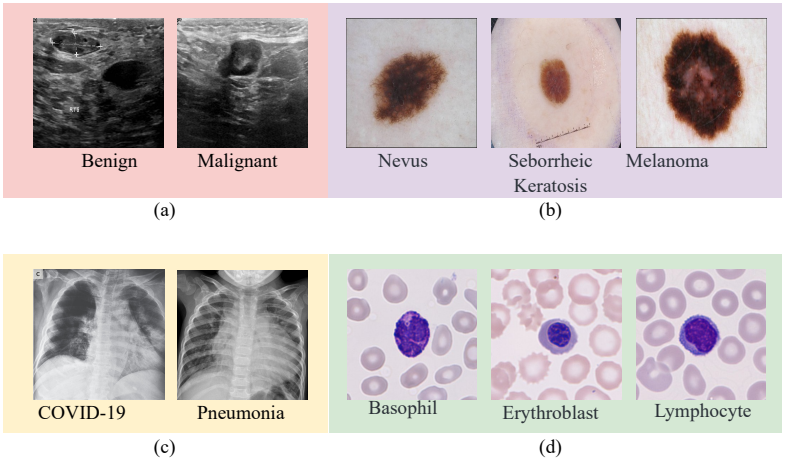

Figure 1.4: Examples of various medical conditions with subtle differences for different modalities including (a) ultrasound images of benign and malignant breast tumors, (b) dermoscopic images of different types of skin lesions, (c) chest x-ray images of COVID-19 and Pneumonia affected patients, and (d) various microscopic peripheral blood cell images.

variability and hinder generalization. Addressing such shifts is essential for reliable real-world performance.

- **Computational Resources:** Medical image classification and segmentation tasks pose significant computational demands, which vary between training and deployment phases. During training, high-resolution medical images, especially with pixel-level annotations for segmentation, require substantial memory, processing power, and storage. This often necessitates advanced GPUs and parallel processing infrastructure, which may not be accessible to all research or clinical institutions. During deployment, the challenge shifts to achieving real-time or near-real-time inference with limited hardware resources—particularly in point-of-care settings or low-resource environments. Balancing model accuracy with computational efficiency is essential for practical and equitable deployment of deep learning models in medical imaging.

In addition to the challenges encountered in medical image classification and segmentation, specific challenges are unique to the task of medical image segmentation.

- **Complex Anatomical Structures:** Medical image segmentation faces significant challenges due to the complex anatomical structures in clinical images. These struc-

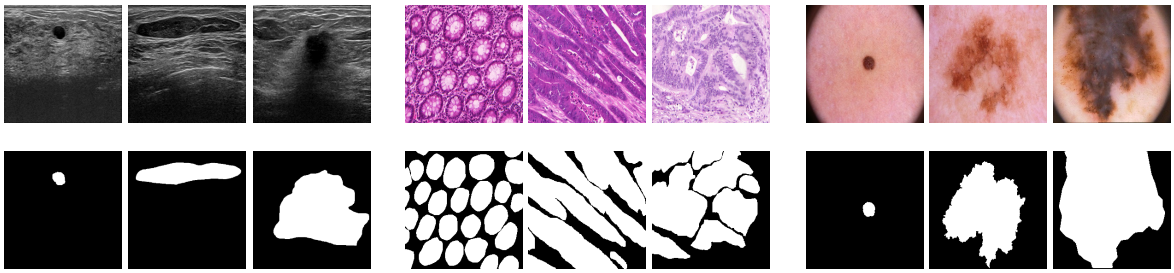

Figure 1.5: Visualization of challenges present in medical image segmentation tasks.

tures, such as organs, tissues, or pathological regions like tumors, often exhibit irregular shapes and sizes, making their boundaries challenging to delineate.

- **Variability in ROI Size:** The regions of interest in medical images can vary significantly in size, ranging from tiny lesions to large lesions (depicted in Figure 1.5), which adds complexity to the segmentation task.

- **Low Contrast and Noise:** Medical images frequently suffer from low contrast, making distinguishing between the target ROI and the background hard. Additionally, noise or imaging artifacts can further obscure the boundaries.

- **Boundary Ambiguity:** The delineation of anatomical structures in medical images is often complicated by blurred or indistinct edges, particularly in the presence of disease progression or texture-altering conditions. This ambiguity not only challenges precise segmentation but also contributes to inter-observer variability during manual annotation.

1.2 Advances in Deep Learning for Medical Image Classification and Segmentation: A Comprehensive Survey

1.2.1 Medical image classification

This literature survey aims to explore the evolution of classification techniques, focusing on deep learning-based approaches.

1.2.2 Traditional Image Classification Methods

Before the advent of deep learning, traditional image classification relied heavily on feature-based approaches. These methods typically involved manually crafting features that could capture the essential characteristics of medical images. Techniques such as texture analysis (Nailon, 2010), shape analysis (Duncan and Ayache, 2000), and statistical methods (Latif et al., 2018; Heimann and Meinzer, 2009) were employed to extract features that were then fed into classical machine learning models like Support Vector Machines (Miranda et al., 2016; Camlica et al., 2015), k-Nearest Neighbors (Zhuang et al., 2020), and decision trees (Azar and El-Metwally, 2013).

While these handcrafted features have been effective in specific applications, they are inherently limited by the need for domain expertise and the difficulty of capturing complex, high-dimensional relationships within the data. This limitation often leads to models specific to particular tasks or datasets, lacking generalization. Once features are extracted from medical images, machine learning algorithms are employed to classify the images based on these features. Each of these algorithms has its strengths and weaknesses, and the quality of the extracted features heavily influences their performance in medical image classification tasks. For instance, SVMs are known for their robustness in high-dimensional spaces, making them well-suited for medical imaging, where the number of features can be large. However, they require careful tuning of parameters such as the kernel function and regularization term.

1.2.3 Deep Learning Based Image Classification Methods

Convolutional Neural Networks (CNNs) have become a cornerstone in computer vision, owing to their ability to automatically and adaptively learn spatial hierarchies of features from input images. Introduced in the late 1980s and popularized in the 2010s, CNNs have revolutionized various image-related tasks, including image classification, object detection, and segmentation.

1.2.3.1 Fundamental Components of CNN

The architecture of CNNs typically includes convolutional layers, pooling layers, nonlinear activation layers, and fully connected layers. The convolutional layers apply filters to the in-

put image to create feature maps through convolution operations that can be mathematically expressed as follows.

$$\mathbf{I}'(x,y) = \sum_{i=-k}^{k} \sum_{j=-k}^{k} \mathbf{I}(x-i, y-j) \cdot \mathbf{K}(i,j), \tag{1.1}$$

where, $\mathbf{I}(x,y)$ represents the input image at pixel position (x,y), while $\mathbf{I}'(x,y)$ is the output feature map obtained after applying the convolution at the same position (x,y). The kernel $\mathbf{K}(i,j)$ is a convolutional filter with dimensions $(2k+1) \times (2k+1)$, where odd dimensions are commonly used in practice to ensure a well-defined central pixel and symmetric receptive field. While filters with even dimensions (e.g., 2×2) can also be employed, they lack a central element, which may introduce asymmetry in the convolution operation. Hence, odd-sized kernels are the standard choice in most convolutional neural network designs. The variables i and j represent the relative indices of the kernel, centered at position (x,y), where k is the half-size of the kernel. This equation represents a sliding window operation, where the kernel $\mathbf{K}$ is applied to the neighborhood of each pixel in the input image $\mathbf{I}$, and the weighted sum of these pixels forms the output at each position. When applied to images, convolutional operations enable the detection of patterns by emphasizing specific local features while reducing noise or irrelevant information. This is particularly beneficial in object recognition, image segmentation, and classification, where objects' spatial relationships and structure need to be preserved and efficiently learned.

In Convolutional Neural Networks, pooling layers are used to downsample the feature maps produced by convolutional layers. The core idea behind pooling is to reduce the spatial resolution of the feature maps while preserving the depth information. This dimension reduction decreases the number of parameters and computation required, making the network more efficient and less prone to overfitting. Mathematically, for a given pooling window of size $k \times k$, the output value at position (i,j) is computed as,

$$Y_{i,j} = f_p\left(\{X_{m,n} \mid (m,n) \in \mathcal{W}_{i,j}\}\right) \tag{1.2}$$

where,

$$f_p(\cdot) = \begin{cases} \max(\cdot) & \text{for max pooling} \\ \min(\cdot) & \text{for min pooling} \\ \frac{1}{|\mathcal{W}_{i,j}|}\sum(\cdot) & \text{for average pooling.} \end{cases} \tag{1.3}$$

In equation 1.2, $X_{m,n}$ represents the values within the pooling window $\mathcal{W}_{i,j}$ of size $k \times k$ centered at position (i, j), defining the neighborhood from which the values are pooled, and $|\mathcal{W}_{i,j}|$ represents the number of elements in the pooling window, typically k^2. This approach helps retain the most prominent features of the input while discarding less important details, which contributes to a reduction in the computational load and helps control overfitting. Pooling also introduces translation invariance, making the model more robust to small shifts and distortions in the input data.

After each convolution operation, an activation function is applied to the resulting feature maps to enable the network to learn complex patterns and relationships in the data. Without nonlinear activation functions, CNNs would be limited to learning only linear transformations, significantly reducing their expressiveness and capability. ReLU is the most commonly used activation function in CNNs, given by the following equation.

$$\text{ReLU}(x) = \max(0, x) \tag{1.4}$$

It sets all negative values in the feature map to zero, allowing the model to focus on positive activations. ReLU is computationally efficient and helps mitigate the vanishing gradient problem—an issue in deep networks where gradients become exceedingly small as they are backpropagated through layers, especially when using activation functions like sigmoid or tanh. This leads to minimal weight updates in early layers, hindering learning. By maintaining positive gradients for positive activations, ReLU alleviates this issue and enables effective training of deeper networks (Bengio et al., 1994).

In convolutional neural networks (CNNs), the sigmoid and softmax activation functions are primarily used in the final layer of CNNs for binary and multi-class classification. The sigmoid function maps input to values between 0 and 1, making it ideal for probabilistic outputs. It is defined as $\text{Sigmoid}(x) = \frac{1}{1+e^{-x}}$, but can suffer from vanishing gradients in deep

networks. The softmax activation function converts logits into a probability distribution across classes, ensuring the sum of the output to 1. It is defined as $\text{Softmax}(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}$, where x_i represents the logit for class i, and n is the total number of classes.

The fully connected (FC) layers in CNNs is typically placed near the network's end and integrates the high-level features extracted by earlier convolutional and pooling layers. Each neuron in the FC layer is densely connected to every neuron in the previous layer, enabling it to learn complex relationships between features. This layer is crucial for transforming feature maps into the final output, such as class scores in classification tasks.

The output of a neuron in a fully connected layer is given by

$$y_i = \sigma\left(\sum_{j=1}^{n} W_{ij} x_j + b_i\right), \tag{1.5}$$

where y_i represents the neuron's output, $\sigma(\cdot)$ denotes the activation function applied (such as ReLU or Softmax), W_{ij} is the weight connecting the j-th input to the i-th neuron, x_j is the j-th input, b_i is the bias term for the i-th neuron, and n is the number of inputs. This equation illustrates how the neuron's output is computed by summing the weighted inputs, adding a bias, and applying an activation function to introduce non-linearity.

1.2.3.2 Evolution of CNNs in medical image classification tasks:

Early Developments in CNNs: The evolution of CNNs in medical image classification tasks has transformed the field of medical image analysis, enabling automated and accurate diagnosis. Initially, traditional machine learning techniques required handcrafted features, but the emergence of CNNs revolutionized this approach by automatically learning spatial hierarchies of features from raw images. In the early stages of applying CNNs to medical image classification, several architectures like LeNet, AlexNet achieved remarkable success in various tasks, including tumor detection (Rouhi et al., 2015; Pan et al., 2015), anatomical organ detection (Cho et al., 2015), and disease classification (Li et al., 2014; Ciompi et al., 2015), development of different computer-aided diagnostic systems (Roth et al., 2015).

Deeper and More Complex CNN Architectures: Deeper and more complex CNN architectures, such as VGGNet and ResNet, significantly advanced medical image classifica-

tion by enabling deeper feature extraction. VGGNet employed many convolutional layers to capture fine details, effectively detecting tumors (Majib et al., 2021) and diseases (Kaur and Gandhi, 2019). However, its depth also increased computational requirements. ResNet introduced residual connections to solve the vanishing gradient problem, allowing networks to be much deeper without performance degradation. This innovation improved tasks like cancer detection (Sarwinda et al., 2021; Jiang et al., 2019). Later architectures like DenseNet and Inception Networks further enhanced efficiency by promoting feature reuse and multi-scale processing, which proved helpful in tasks requiring both global and local feature recognition in medical imaging (Huang et al., 2020; Gaur et al., 2023). These advancements have enabled CNNs to achieve high accuracy in complex medical tasks but still face computational costs and data availability challenges.

Multi-Scale Learning with CNNs: Initially, traditional CNN architectures focused on fixed-size receptive fields, which were adequate for learning local patterns but struggled with multi-scale variations common in medical images, such as tumors of varying sizes or anomalies scattered across different regions. The evolution of deep learning introduced multi-scale learning as an enhancement in CNNs to overcome this challenge. By incorporating multiple levels of feature extraction—through techniques like dilated convolutions, multi-resolution filters, and image pyramids—multi-scale CNNs allowed for better detection of both small-scale and large-scale patterns. This capability significantly improved the performance of CNNs in complex medical imaging tasks, such as lung nodule detection Shen et al. (2015), brain tumor detection (Rajasree et al., 2021), macular optical coherence tomography (Rasti et al., 2017), classification of mammograms (Xie et al., 2020a) and so on.

CNN Ensemble Techniques: Due to the complexity and variability in medical imaging data, such as different imaging modalities, patient demographics, and noise, a single CNN model often struggled with generalization. Ensemble learning emerged as a solution to address these limitations by combining multiple CNNs to form a more robust and accurate model. By leveraging the diversity of various models, each trained with different architectures, hyperparameters, or data subsets, ensemble methods improved the ability to capture a broader range of features and reduced the risk of overfitting. In medical image classification tasks like disease detection and anomaly identification, ensemble CNNs demonstrated improved performance by integrating the strengths of individual models. For example, de-

tection of breast cancer (Yang et al., 2019), classification of lung nodules and diagnosis of retinal disease (Fu et al., 2018) and multiple other related tasks (Kumar et al., 2016; Zhang et al., 2019b) benefited from the aggregation of predictions, resulting in higher accuracy and robustness.

Transfer Learning in Medical Image Classification: Deep CNNs (VGG, ResNet, Inception, etc) often lack generalizability in medical image classification tasks due to the limited availability of annotated medical data. Unlike general image datasets like ImageNet, medical datasets are typically smaller and less diverse, leading to overfitting and reduced performance on unseen data. Transfer learning emerged as a solution to these challenges, allowing models pre-trained on large-scale datasets to be fine-tuned for specific medical tasks. This approach leverages features learned from non-medical images and applies them to medical domains, significantly improving performance even with limited training data. Transfer learning not only improves model generalization but also accelerates the training process by starting from a robust pre-trained model, which would otherwise require vast amounts of labeled medical data. In various medical imaging classification tasks, transfer learning has proven instrumental in improving feature extraction and classification accuracy (Li et al., 2017; Abbas et al., 2020; Kim et al., 2022; Kora et al., 2022).

1.2.3.3 Introduction of Transformer and Hybrid Architectures:

Transformer-based architectures (Vaswani et al., 2017) have emerged as powerful alternatives to traditional CNNs in medical image classification, primarily due to their ability to model global spatial relationships and handle long-range dependencies. Unlike CNNs, which rely on local receptive fields and convolutional operations to capture features, transformers use self-attention mechanisms to process input data holistically, allowing for more comprehensive feature extraction and the ability to relate distant regions in an image.

Vision Transformers (ViTs): Vision Transformers (Dosovitskiy et al., 2020), initially designed for natural image classification, has been adapted for medical image analysis with promising results. In ViTs, an image is split into fixed-size patches, and each patch is treated as a token, similar to how words are handled in natural language processing. These patches are then passed through multiple layers of self-attention, allowing the model to capture both local and global relationships across the entire image. This capability is beneficial in medical

imaging tasks, where subtle and spatially distant patterns, such as lesions or tumors, might need to be detected. The ViT architecture has shown competitive performance in skin lesion classification, breast cancer detection, and lung disease identification (Manzari et al., 2023; Almalik et al., 2022; Ma et al., 2022). Its strength lies in its ability to model the entire image's context, making it more effective in scenarios where spatial dependencies are crucial for accurate diagnosis.

Hybrid CNN-Transformer architectures: To combine the strengths of both CNNs and transformers, hybrid architectures have been developed, integrating the local feature extraction capabilities of CNNs with the global attention mechanisms of transformers. In these architectures, CNN layers are typically used at the initial stages to capture fine-grained local features, while transformer layers handle the broader spatial context. This synergy between CNNs and transformers is particularly effective in complex medical imaging tasks, such as multi-scale tumor detection, where both local details and global structures are important (Wu et al., 2023b; Dalmaz et al., 2022; Yuan et al., 2023).

Despite their potential, transformer-based models face challenges, particularly the need for large datasets to train self-attention mechanisms effectively. Transformers typically require more data and computational resources than CNNs, making them harder to implement in scenarios with limited labeled medical data. However, the use of transfer learning, where transformers are pre-trained on large-scale datasets and fine-tuned for specific medical tasks, has mitigated some of these challenges. In future, advancements in model efficiency and more sophisticated hybrid architectures are expected to further drive the adoption of transformers in medical image classification, opening up new possibilities for improved diagnostic accuracy.

1.2.4 Medical image segmentation

The evolution of deep learning in medical image segmentation has been marked by significant advancements in both algorithmic techniques and their application across various medical imaging modalities. A detailed examination of the pivotal developments and achievements in this field is presented below.

Early Methods and Pre-Deep Learning Era: Before the advent of deep learning, medical image segmentation was primarily accomplished using traditional methods such as thresholding, region-growing, edge detection, and manual segmentation. These methods were

often limited by their dependence on handcrafted features and were generally specific to particular tasks or imaging modalities. The performance of these traditional methods was usually suboptimal, especially in complex medical images with varying intensity distributions and noise.

Emergence of Convolutional Neural Networks (CNNs): The breakthrough for deep learning in medical image segmentation began with the application of Convolutional Neural Networks (CNNs), which were initially successful in image classification tasks. The introduction of CNNs, such as AlexNet, VGGNet, and ResNet, revolutionized image analysis by automating feature extraction and enabling end-to-end learning. This success quickly translated into segmentation tasks, where CNNs were adapted to create pixel-wise predictions, a fundamental requirement in segmentation.

A significant breakthrough in medical image segmentation came with the introduction of Fully Convolutional Networks (FCNs) by Long et al. in 2015 (Long et al., 2015). FCNs replaced fully connected layers with convolutional layers, allowing the network to produce spatially dense output maps suitable for segmentation. The original FCN architecture was adapted for medical imaging, enabling end-to-end training and pixel-level classification. The key innovation was using upsampling layers (deconvolutions) to transform low-resolution feature maps back to the original image size, facilitating precise segmentation.

The introduction of U-Net in 2015 Ronneberger et al. (2015) marked a significant advancement in medical image segmentation. U-Net introduced a symmetric encoder-decoder architecture with skip connections that link corresponding layers in the encoder and decoder. The encoder or contracting path captures contextual features through a series of convolutional layers followed by ReLU activations and max-pooling, progressively reducing spatial dimensions while extracting higher-level features. The most abstracted features are processed at the network's bottleneck without further downsampling. The decoder or expanding path mirrors the encoder, using transposed convolutions for upsampling and concatenating corresponding layers from the encoder via skip connections. These skip connections let the network retain fine-grained spatial information lost during downsampling, leading to more accurate segmentation. The final output layer typically consists of a 1×1 convolution, which is used to map the high-level feature maps to the desired number of output classes. This 1×1 convolution performs a per-pixel classification, assigning each pixel a probability of belonging

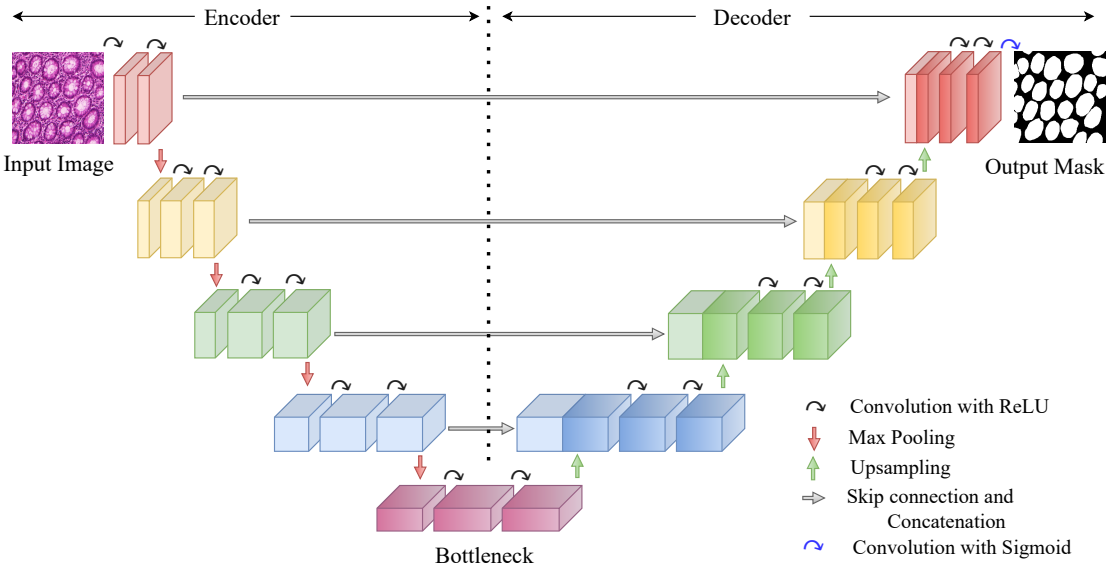

Figure 1.6: An illustration of the U-Net architecture, showcasing its encoder-decoder framework for medical image segmentation

to a particular class. The 1×1 kernel allows the model to preserve spatial resolution while reducing the depth of the feature maps to match the number of output classes. This design ensures that the output is a segmentation map where each pixel is classified independently. While the 1×1 convolution is the standard approach, some U-Net variants may experiment with different configurations, such as using a larger kernel or additional processing layers like dilated convolutions or post-processing refinements, to incorporate broader context or to refine the segmentation output. U-Net's key innovations include its skip connections, which integrate low-level and high-level features, and its efficiency with small datasets, making it ideal for medical imaging.

Following the success of U-Net, several architectures have built upon its encoder-decoder framework by incorporating more sophisticated designs and strategies to enhance performance. V-Net (Milletari et al., 2016) extended U-Net for 3D medical image segmentation, particularly for volumetric data like MRI and CT scans. Using 3D convolutions, V-Net captured spatial dependencies across three dimensions, making it suitable for segmenting organs and other 3D structures. SegNet (Badrinarayanan et al., 2017) introduced a max-pooling indices-based upsampling approach, focusing on computational efficiency while maintaining segmentation accuracy, making it ideal for resource-constrained applications. In 2017,

DenseNet's (Huang et al., 2017) dense connectivity concept was adapted into a segmentation architecture, which improved gradient flow and feature reuse. Recognizing the importance of multi-scale learning, architectures like MultiRes U-Net (Ibtehaz and Rahman, 2020) integrated multi-scale feature extraction and residual connections to handle the segmentation of varied structures like blood vessels and tumors. The DeepLab series (Chen et al., 2017a,b, 2018a) introduced atrous convolutions and spatial pyramid pooling to better capture multi-scale context, with DeepLabv3+ (Chen et al., 2018a) becoming a popular choice in medical image segmentation due to its ability to handle fine details while considering global context.

Recently, attention mechanisms have been increasingly integrated into segmentation networks to improve the model's ability to focus on relevant areas of the image. These mechanisms enable the model to give more weight to important regions, thereby improving segmentation accuracy, especially in complex or cluttered images. Attention U-Net (Oktay, 2018) is a prominent example where attention gates were incorporated into the U-Net architecture, leading to improved performance in medical image segmentation tasks. Table 1.2.4 lists the key modifications to the U-Net architecture to improve segmentation performance in medical imaging tasks. These enhancements incorporate advanced techniques like multi-scale learning, residual connections, and hybrid architectures to achieve better accuracy and efficiency across various datasets.

Table 1.1: A list of recent deep learning-based architectures developed for various medical image segmentation tasks, highlighting modifications and enhancements made to the baseline U-Net model to improve performance in handling complex medical images

Architecture	**Remarks**	**Applied Field**
MALUNet (Ruan et al., 2022)	Integrating four lightweight modules –DGA, IEA, CAB, and SAB that optimize global and local feature extraction, channel attention, spatial attention mechanisms, and computational efficiency.	Skin Lesion Segmentation
MSS-UNet (Zhu et al., 2024)	Integrating double-spatial-shift MLP module and multi-spatial-shift external attention module into a convolution-based encoder-decoder architecture	Skin Lesion Segmentation.
MHorUNet (Wu et al., 2024)	Combining higher-order spatial interactions with recursive gate convolution and multi-stage dimensional fusion in skip connections.	Skin Lesion Segmentation

PDF-UNet (Iqbal and Sharif, 2023)	Combining a Data Expansion Network (DEN), Probability Map Generator (PMG), and a modified PDF-UNet, resulting in significant DSC gains in semi-supervised approach over traditional UNet models.	Breast Lesions Segmentation in Ultrasound Images
Sharp attention UNet(Khaledyan et al., 2023)	Introducing a depthwise convolution with a sharpening filter in the skip connections, resulting in more apparent feature maps without adding extra parameters.	Breast Lesions Segmentation in Ultrasound Images
AAU-Net (Chen et al., 2022a)	Integrated hybrid adaptive attention module consists of a channel self-attention block and a spatial self-attention block to capture more features under different receptive fields.	Breast Lesions Segmentation in Ultrasound Images
HAU-Net (Zhang et al., 2024)	Hybrid CNN-transformer framework from both the long-range dependency of transformers and the local detail representation of CNNs.	Breast Lesions Segmentation in Ultrasound Images
NU-net (Chen et al., 2023)	A nested U-Net with varying depths, multi-output layers, and short connections to better capture tumor features and long-range correlations.	Breast Lesions Segmentation in Ultrasound Images
SMU-Net (Ning et al., 2021)	integrating a style-matching mechanism into UNet that transfers informative features from full-modality MRI to missing-modality MRI.	Breast Lesions Segmentation in Ultrasound Images
MDF-Net Qi et al. (2023)	a Two-stage network using multi-scale feature selection and refinement, deep supervision at all scales, and optimized fusion to handle noise and variability.	Breast Lesions Segmentation in Ultrasound Images
CR-Unet (Li et al., 2019)	Integrating a spatial RNN to capture multi-scale and long-range contexts, coupled with deep and self-supervision strategies.	Ovary and Follicle Segmentation in Ultrasound Images
MA-Unet (Cai and Wang, 2022)	Incorporating attention gates (AGs) to align features across skip connections, model remote dependencies, and integrate multi-scale global context, achieving better performance with fewer parameters.	esophageal and esophageal cancer segmentation, lung structure segmentation
SA-UNet (Guo et al., 2021)	Integrating a spatial attention module for adaptive feature refinement and using structured dropout convolutional blocks to prevent overfitting, achieving state-of-the-art performance with fewer annotated samples.	Retinal Vessel Segmentation
RIC-Unet (Zeng et al., 2019)	Integrating residual blocks, multi-scale, and channel attention mechanisms into a U-Net architecture, achieving higher accuracy than traditional methods and other CNN approaches.	Nuclei Segmentation in Histology Images
TransUNet (Chen et al., 2021)	Integration of a Transformer-based encoder with a U-Net-style decoder, combining global context modeling with precise spatial localization for enhanced medical image segmentation.	cardiac image segmentation and multiple abdominal organ segmentation
BioNet (Zhang et al., 2019a)	Introduction of bi-directional skip connections that recurrently reuse existing building blocks, enhancing U-Net's performance across various tasks without adding extra parameters or increasing model complexity.	MonuSeg and TNBC
RA-UNet (Zhu et al., 2022)	Integration of 3D hybrid residual attention modules into a U-Net architecture, allowing for adaptive attention-aware feature extraction in deep networks.	liver and tumor in CT scans

Swin-Unet (Cao et al., 2022)	Integration of Transformer-based U-shaped Encoder-Decoder, using Swin Transformers with shifted windows to capture local and global features, and a patch expanding layer for effective up-sampling, enhancing segmentation performance.	cardiac image segmentation and multiple abdominal organ segmentation
LeVIT-UNET (Graham et al., 2021)	Incorporate LeViT Transformer module into the U-Net architecture, resulting in better trades off the accuracy and efficiency of the Transformer block.	cardiac image segmentation and multiple abdominal organ segmentation
MedT (Valanarasu et al., 2021)	Gated axial-attention model with a control mechanism in the self-attention module, along with a Local-Global (LoGo) training strategy, enhancing the model's ability to capture both global and local features.	brain ventricles, gland and tissue segmentation
MISSFormer (Huang et al., 2021)	Hierarchical encoder-decoder design featuring an enhanced transformer Block to improve both long-range dependencies and local context, and an Enhanced Transformer Context Bridge to effectively fuse multi-scale features.	Cardiac image segmentation and multiple abdominal organ segmentation
UCTransNet (Wang et al., 2022)	Replaced U-Net's traditional skip connections with the CTrans module, which includes Channel Cross fusion with Transformer and Channel-wise Cross-Attention sub-modules, enhancing multi-scale context modeling and resolving semantic gaps.	Gland, tissue, and multiple abdominal organ segmentation
Attention-UNet (Oktay, 2018)	Integration of attention gates into UNet, enabling automatic focus on target structures without external localization modules.	CT abdominal datasets for multi-class image segmentation
DCA (Ates et al., 2023)	Improved skip-connections through sequential channel and spatial cross-attention, addressing semantic gaps.	Multiple datasets (MonuSeg, Synapse, GlaS, CVC-ClinicDB, Kvasir-SEG)
MultiRes-UNet (Ibtehaz and Rahman, 2020)	Enhances U-Net by integrating multi-scale convolutions and factorizing larger operations into efficient blocks, reducing memory demands while improving feature extraction and segmentation accuracy.	Multiple datasets (MonuSeg, Synapse, GlaS, CVC-ClinicDB, Kvasir-SEG)

1.2.5 Multi-task learning in medical image analysis

In medical imaging, classification is typically used to diagnose or grade diseases, while segmentation helps delineate pathological regions or anatomical structures. These tasks are inherently related, as segmentation can provide spatial information that aids in classification, and classification can guide the model to focus on disease-relevant areas for segmentation. MTL leverages this interdependence by sharing representations across tasks, leading to more informative feature extraction and improved overall performance. It also enhances efficiency by simultaneously learning both tasks in a single model, reducing computational costs. In medical applications, where labeled data is often limited, MTL offers a significant advantage

by utilizing the shared knowledge between tasks to improve generalization. Despite challenges such as task imbalance and data heterogeneity, MTL has demonstrated its potential to provide more accurate, robust, and interpretable models for critical applications in medical image analysis, such as disease diagnosis and detection of regions of interest. In the following section, I explore the advancements and applications of MTL in medical image classification and segmentation, focusing on how these two tasks can benefit from shared learning.

Table 1.2: Overview of multi-task learning approaches in deep learning for medical imaging, highlighting architectures designed to simultaneously handle tasks like segmentation and classification, leveraging shared representations for enhanced performance across multiple tasks.

Architecture	**Remarks**	**Applied Field**
Multi-Task Learning Framework (Chen et al., 2018b)	Introducing a Feature Passing Module and Gated function to have controlled communication between the tasks by selectively transferring features between the classification and segmentation branches.	skin lesion segmentation and classification tasks
Multi-Task Learning Framework (Song et al., 2020)	Comprises a feature pyramid network (FPN), a region proposal network (RPN), and three separate convolutional subnets, each dedicated to classification, detection, and segmentation tasks.	skin lesion detection, segmentation and classification tasks
SAH-MTL (Zhang et al., 2021a)	uses soft and hard attention mechanisms and attention-gated units to focus on lesion regions, improving both tasks.	Breast tumor segmentation and classification
Multi-Task Learning Framework (Zhou et al., 2021)	Features an encoder-decoder for segmentation, a lightweight multi-scale network for classification, and an iterative training strategy that refines feature maps.	Breast tumor segmentation and classification in automated breast ultrasound (ABUS) images
Multi-Task Learning Framework (Gao et al., 2020)	Employs feature transfer to leverage combined features from different tasks early in training, enhancing generalizability.	Breast tumor segmentation and classification
MTL-COSA (Xu et al., 2022)	Features a regional attention (RA) module that uses predicted probability maps to guide the classifier in learning important features from tumor, peritumoral, and background regions, improving overall performance.	Breast tumor segmentation and classification
MTU (Dabass et al., 2022)	Hybrid Convolutional Learning Units for advanced feature extraction and Attention Learning Units for enhanced feature focus.	Histopathology gland segmentation and cancer detection

Cerberus (Graham et al., 2023)	Features a novel sampling mechanism that integrates data from multiple sources and leverages task correlations for more accurate results.	identification of various tissue regions
Multi-Task Neural Network (Xue et al., 2019)	multi-task neural network combines deep learning and radiomic features to perform glioma subtyping and multi-region segmentation simultaneously.	Glioma Segmentation and IDH Genotyping
Multi-Task Learning Framework (Cheng et al., 2022)	Hybrid CNN-Transformer encoder addresses task correlation and heterogeneity by extracting shared spatial and global features.	Glioma Segmentation and IDH Genotyping
Multi-Task Learning Framework (Foo et al., 2020)	To address the lack of ground-truth lesion segmentation masks, a semi-supervised learning process is introduced to generate segmentation masks across datasets	DR grading and lesion segmentation
Multi-Task Learning Framework (Wang et al., 2021b)	The model integrates image super-resolution (ISR), lesion segmentation, and DR grading tasks. It employs a task-aware loss function to guide ISR towards pathological regions, improving subsequent lesion segmentation and DR grading	DR grading and lesion segmentation
DSI-Net (Zhu et al., 2021)	DSI-Net includes the Lesion Location Mining (LLM) module for improved lesion classification and the Category-Guided Feature Generation (CFG) module to enhance segmentation using classification features	classification and segmentation of wireless capsule endoscope (WCE) images
MA-MTLN (Zhang et al., 2021b)	The network uses scale-aware and task-aware attention, along with visual and spatial attention, to enhance feature learning.	Tumor segmentation and lymph node classification

1.3 Motivation, Scope, and Research Objectives

1.3.1 Motivation

The rapid advancement of medical imaging technologies has significantly transformed healthcare, providing clinicians with powerful tools to diagnose and monitor various medical conditions. However, the increasing complexity and volume of medical images have introduced new challenges in accurately and efficiently interpreting these images. Traditional manual analysis is time-consuming, prone to human error, and requires high expertise, which may not always be available, particularly in resource-limited settings.

In response to these challenges, integrating machine learning and deep learning techniques

into medical image analysis has emerged as a promising solution. Despite the progress, existing methods still face critical limitations. For instance, conventional classification models often overlook the spatial and spectral domains of medical images, which contain valuable diagnostic information. Furthermore, while Vision Transformers (ViTs) have shown promise in image classification, their performance in medical imaging is hindered by limited data and the inherent complexities of clinical images.

Medical image segmentation, another crucial task, is essential for accurately identifying regions of interest (ROIs) within images. However, traditional segmentation methods struggle with the irregular shapes and varying sizes of ROIs, as well as the low contrast and intricate details present in medical images. Morphological operations have been utilized in image processing, but their rigid nature limits their adaptability to the diverse and complex structures found in clinical images.

This thesis is motivated by the need to overcome these challenges by developing advanced machine learning and deep learning frameworks tailored to the specific demands of medical image analysis. The goal is to create robust, interpretable, and efficient solutions that can enhance diagnostic accuracy, improve segmentation performance, and ultimately support healthcare professionals in making informed decisions. The proposed methodologies aim to bridge the gap between existing techniques and the practical requirements of medical imaging, offering a significant step forward in developing reliable and scalable AI-driven diagnostic tools.

1.3.2 Scope

This thesis explores developing and applying advanced deep learning techniques for medical image analysis, focusing on enhancing the accuracy, efficiency, and interpretability of image classification and segmentation tasks. The work covers several critical areas within the field, including the integration of spatial and spectral domain information for improved disease detection, the modification and hybridization of Vision Transformer (ViT) architectures to address the complexities of medical images better, the development of novel morphological modules for more effective segmentation, and the implementation of multi-task learning strategies to optimize performance across multiple tasks. The scope of this thesis encompasses the design, implementation, and evaluation of cutting-edge deep learning models that aim to

advance the field of medical image analysis. By integrating classification and segmentation tasks and focusing on performance and interpretability, this work contributes to developing more accurate, efficient, and reliable diagnostic tools for clinical use.

1.3.3 Primary Objectives

- Explore the medical images in the spectral domain towards diagnostic solutions.
- Design a novel, lightweight segmentation architecture that can effectively delineate intricate and subtle regions of interest in medical images while simultaneously addressing the irregularities inherent in medical data by integrating advanced architectural innovations.
- To develop efficient multi-task learning frameworks that simultaneously perform classification and segmentation while reducing computational complexity.

1.4 Thesis Organization and Chapter-wise Contributions

Figure 1.7 depicts the outline of the thesis. Among the various tasks related to medical image analysis with deep learning algorithms, I mainly focus on developing novel classification and segmentation algorithms. The first chapter is introductory. The second chapter investigates the contribution of medical images in the spatio-spectral domain towards disease detection tasks. The third and fourth chapters describe two novel architectures designed for classification and segmentation tasks, respectively. The fifth chapter deals with the problem of multi-task learning corresponding to medical image classification and segmentation tasks simultaneously. The conclusion of the thesis is presented in the sixth chapter.

1.4.1 Contributions of Chapter 1

This chapter overviews the challenges and opportunities in deep learning for medical image analysis, focusing on classification and segmentation tasks. It sets the stage for the subsequent contributory chapters by outlining the thesis's motivation, research gaps, and objectives.

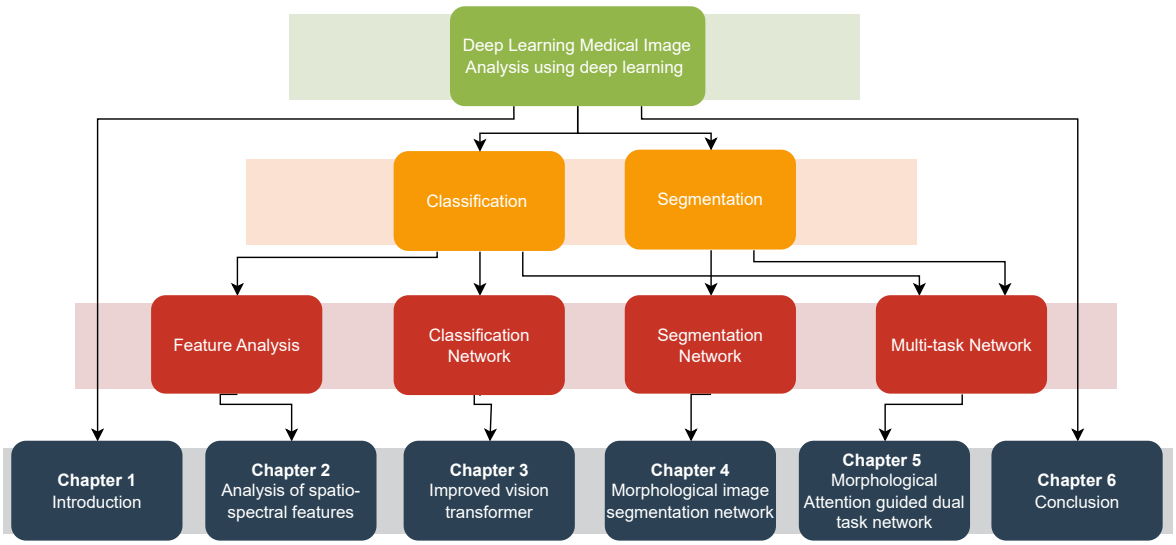

Figure 1.7: Outline of the thesis.

1.4.2 Contribution of chapter 2

This chapter contributes to developing an automated diagnostic solution to detect COVID-19 infections using chest radiographs. The approach integrates machine learning and computer vision algorithms to address the limitations of conventional methods, particularly the lack of consideration for the spatial aspects of medical images in previous studies.

The chapter's primary contribution is introducing a novel method that combines spatial and spectral domain information to enhance disease detection capabilities. The method comprises three stages: Feature Extraction, Dimensionality Reduction via Projection, and Prediction. Initially, the images are transformed into spectral and spatio-spectral domains using Discrete Cosine Transform (DCT) and Discrete Wavelet Transform (DWT), two powerful image processing techniques. These transformations enable extracting features from spatial, spectral, and spatio-spectral domains, which are then projected into a lower dimension through a Convolutional Neural Network (CNN).

The projected features are fed into a Multilayer Perceptron (MLP) for final prediction. Integrating these three types of features results in superior performance compared to using any single feature type alone, highlighting the complementary nature of the spectral domain information in characterizing medical conditions.

Furthermore, saliency maps corresponding to different medical conditions validate the

reliability of the proposed method. The study is extended to identify various medical conditions across diverse medical image datasets, demonstrating the robustness and efficiency of the combined feature approach. The proposed method shows potential as a generalized and robust solution for computer-aided disease detection in medical imaging.

1.4.3 Contribution of chapter 3

This chapter presents the development of a modified Vision Transformer (ViT) architecture, termed ReMixViT, specifically designed to address the challenges of medical image classification, such as limited training samples and complex image attributes. The chapter contributes by incorporating an efficient MLP-Mixer layer and reordering residual blocks within the encoder block of the ViT, thereby improving feature mixing and enhancing the model's generalization ability.

In addition to ReMixViT, the chapter introduces two hybrid architectures, ResReMixViT and ResReMixViT+, which integrate a Convolutional Neural Network (ResNet50) with ReMixViT encoder blocks. These hybrids consider feature maps of single and multiple scales, respectively, further boosting classification performance.

The proposed architectures are rigorously evaluated using six diverse medical imaging datasets encompassing various modalities and medical conditions. Comparative studies reveal that ReMixViT and the hybrid models significantly outperform both vanilla ViT models and other hybrid models that utilize ViT encoder blocks, with observed improvements of 4.62% and 3.08% in the F1-score performance metric.

Moreover, when combined with data augmentation techniques, the hybrid architectures surpass state-of-the-art hybrid networks. Beyond performance metrics, this chapter also enhances the interpretability of the AI system by providing visual explanations through attention maps and gradient flow analysis, aiding medical practitioners in drawing inferences from an explainable AI perspective.

Additionally, the chapter demonstrates the adaptability of the proposed modifications to other vision transformer architectures, resulting in enhanced performance across different models.

1.4.4 Contribution of chapter 4

This chapter introduces an innovative approach to medical image segmentation, addressing the complexities inherent in clinical images, such as irregular shapes, varying sizes, low contrast, and intricate details. The chapter's main contribution is the development of three specialized modules grounded in morphological operations: the Multi-scale Morphological Closing Module, the Multi-scale Morphological Opening Module, and the Multi-scale Morphological Gradient Module. Unlike traditional morphological operations, these modules involve learning structuring elements through a training process, allowing the model to adapt effectively to the irregular shapes of Regions of Interest (ROIs).

Furthermore, introducing dilation rates within these morphological operations accommodates the diverse range of ROI sizes found in clinical images. The proposed modules are integrated into a lightweight framework called Morph-UNet, which synergizes deep neural networks with multi-scale morphological operations for enhanced segmentation performance.

The efficacy of this approach is rigorously validated across diverse medical imaging datasets covering various modalities, conditions, and ROI proportions. Extensive experimentation using widely recognized segmentation metrics demonstrates the model's superiority over thirteen state-of-the-art segmentation methods and baseline models. This work represents a significant advancement in medical image segmentation, offering a robust and adaptable solution for clinical applications.

1.4.5 Contribution of chapter 5

This chapter presents a novel multi-task learning framework specifically designed to address both classification and segmentation tasks within the realm of medical image analysis. The proposed framework is built on the premise that learning shared representations for these tasks can significantly improve overall performance. By jointly optimizing for classification and segmentation, the framework leverages the inherent relationships between these tasks, leading to more accurate and robust outcomes. Extensive experiments demonstrate that the multi-task learning approach not only outperforms traditional single-task methods but also generalizes well across various medical imaging modalities. The rigorous evaluation highlights the framework's potential to enhance diagnostic accuracy and efficiency in medical

image analysis.

1.4.6 Contribution of chapter 6

This chapter summarizes the key findings and contributions of the thesis, reflecting on the advancements made in the fields of classification, segmentation, and multi-task learning for medical image analysis. It also discusses potential future research directions based on the work presented in the thesis.

1.5 Significance of the Study

This thesis represents a significant advancement in the field of medical image analysis, addressing several critical challenges that have historically limited the effectiveness of deep learning based diagnostic tools in clinical practice. The potential impacts of this research are multifaceted in the field of medical image analysis.

Improvement in Segmentation Accuracy: The proposed morphological modules—Multi-scale Morphological Closing, Opening, and Gradient Modules—introduce a novel approach to segmenting regions of interest (ROIs) in medical images. By learning structuring elements and adapting dilation rates, this method can effectively handle the irregular shapes and varying sizes of ROIs commonly seen in clinical images. This advancement could lead to more accurate segmentation, which is crucial for diagnosing and planning treatment for various conditions.

Enhanced Diagnostic Accuracy: Integrating spatial and spectral features for disease detection, as emphasized in Chapter 2, highlights the importance of a multi-domain approach. By incorporating Discrete Cosine Transform (DCT) and Discrete Wavelet Transform (DWT) into the feature extraction process, I have demonstrated the potential to improve diagnostic accuracy for conditions such as COVID-19. This approach could be extended to other diseases, leading to more reliable and early detection of various medical conditions.

Interpretability of the models: Using visual explanations, such as attention maps and gradient flow in the ReMixViT architecture, enhances the interpretability of deep learning based diagnostic systems. This is crucial for gaining the trust of healthcare professionals, as they need to understand and validate the decisions made by deep learning models. By

improving the interpretability of these models, I aim to facilitate their adoption in clinical settings, where transparency is essential.

Efficient and Generalizable Diagnostic Solutions: The development of Morph-UNet and hybrid architectures like ResReMixViT+ provide lightweight and efficient medical image segmentation and classification solutions. With their improved performance across diverse datasets, these models could be generalized to various medical imaging tasks, offering scalable solutions that can be deployed across different healthcare systems.

Contribution to Multi-task Learning in Medical Imaging: By focusing on multi-task learning, where classification and segmentation are integrated, this research addresses the need for holistic analysis in medical imaging. This approach improves performance in individual tasks and enhances the overall efficiency of medical image analysis pipelines. It could pave the way for more comprehensive diagnostic tools that provide richer insights from a single model.

Broader Impact on Healthcare: The efficient methods and lightweight architectures proposed in this research could seamlessly integrate into clinical workflows, offering significant value in resource-limited settings. These innovations can accelerate diagnoses, improve patient outcomes, and lower healthcare costs by aiding radiologists and other healthcare professionals in making more accurate and efficient decisions.

In summary, this thesis advances the state-of-the-art in medical image analysis. It lays the groundwork for developing more accurate, interpretable, and widely applicable deep learning based diagnostic tools. The significance of this research lies in its potential to improve patient care by enabling more reliable and efficient analysis of medical images across various clinical domains.

1.6 Summary of the Introduction

In the introduction, the significance of this research is established by highlighting the critical role of medical image analysis in healthcare, particularly in improving diagnostic accuracy and patient outcomes. The challenges posed by complex distinguishing patterns, irregular shapes, varying sizes of ROIs, and the need for interpretable deep learning based models are emphasized, underscoring the necessity for innovative approaches in this field. This research

addresses these challenges through the development of advanced methods for classification, segmentation, and multi-task learning, offering efficient and generalizable solutions in the subsequent chapters.

Chapter 2

A Deep Learning Framework Integrating the Spectral and Spatial Features for Image-assisted Medical Diagnostics

Summary

This chapter focuses on developing a computer-aided disease detection system to streamline the manual diagnostic process, with a particular emphasis on detecting COVID-19 infections from chest radiographs. Motivated by the COVID-19 outbreak, the proposed method leverages machine learning and computer vision algorithms to provide an automatic diagnostic solution using widely accessible medical imaging infrastructure. Unlike previous studies, which have focused mainly on spatial features, this work investigates the complementary role of spectral-domain information in medical images to improve disease detection.

The method follows a three-stage process: feature extraction, dimensionality reduction, and prediction. First, images are transformed into the spectral and spatio-spectral domains using Discrete Cosine Transform (DCT) and Discrete Wavelet Transform (DWT), which capture additional information beyond the spatial domain. Features from these domains, along with spatial features, are projected into a lower-dimensional space through a Convolutional Neural Network (CNN). These projected features are then fed into a Multilayer Perceptron (MLP) for final prediction.

Results demonstrate that the combination of spatial, spectral, and spatio-spectral features significantly improves diagnostic performance, revealing complementary information in the

spectral domain that enhances disease characterization. Saliency maps further validate the reliability of this approach, highlighting the model's ability to capture relevant medical conditions. The study also extends to other medical imaging datasets, illustrating the effectiveness and generalizability of the proposed method as a robust, multi-purpose diagnostic tool.

2.1 Introduction

2.1.1 Overview

The proper diagnosis of any medical condition plays a vital role in effective treatment and also in the prevention of any infectious disease to spread out. Various machine learning based diagnostic solutions have been proposed to ease such a process of manual diagnosis that requires domain expertise and long training time(Suzuki, 2017). The recent outbreak of Coronavirus disease 2019 (COVID-19) is the third significant Coronavirus outbreak in less than 20 years. In this context, a computer-based diagnostic solution that leverages readily available infrastructure—even in rural areas across the globe—is essential for improving healthcare accessibility. According to (Cozzi et al., 2020), a chest radiograph of a COVID-19 infected person exhibits 'patchy or diffuse reticular–nodular opacities and consolidation, with basal, peripheral and bilateral predominance'. Thus, the readily and widely available infrastructure for X-rays may be utilized for primary and immediate assessment for detecting COVID-19 infection.

I aim to develop an automatic model using machine learning algorithms that would aid the clinician as an adjunct tool for diagnosing COVID-19 infection. I pose this COVID-19 detection problem as a three-class classification paradigm where the classes are *Normal*, *Pneumonia* (non-COVID), and *COVID-19* by utilizing deep learning algorithms. The reason behind choosing *Pneumonia* as one of the classes is that Pneumonia and COVID-19 can be easily confused. Both non-COVID-19 Pneumonia and COVID-19 may exhibit ground glass patterns in the lung due to lung infiltration and consolidation. Over the last decade, deep learning, a subfield of machine learning, has gained popularity in assisting as a diagnostic aid. The successful application of deep learning can be found in diagnostic assessment of different biomedical conditions such as arrhythmia detection (Acharya et al., 2017), skin cancer classification (Esteva et al., 2017), breast cancer detection (Celik et al., 2020), brain disease

classification (Talo et al., 2019), Pneumonia detection from chest X-ray images (Rajpurkar et al., 2018) and lung segmentation (Souza et al., 2019). These studies on the application of deep learning algorithms on medical imaging data show the efficiency of a deep learning algorithm in expressing complex patterns that are even difficult to capture in untrained eyes. These studies motivate us to exploit deep learning algorithms for characterizing such intricate, differentiating patterns to identify the underlying medical condition.

2.1.2 Background

Few recent studies have endeavored to detect COVID-19 infection from chest X-ray images with the help of deep learning algorithms. Wang and Wong (Wang et al., 2020) have proposed a Convolutional Neural Network (CNN) based architecture built using generative synthesis, referred to as COVID-net, which is trained on 13,975 CXR images across 13,870 persons belonging to the categories of *Normal*, *Pneumonia*, and *COVID-19* class. They have achieved an overall test accuracy of 93.3% with sensitivity and precision to *COVID-19* class of 91.0% and 98.9%, respectively. In another study (Oh et al., 2020), authors have presented a two-stage network to classify four classes, namely *Normal*, *Bacterial*, *Tuberculosis* (TB), and *Viral Pneumonia/COVID-19*. They first trained an extended fully convolutional (FC)-DenseNet103 for image segmentation purposes; thereafter, a patched-based CNN was trained by the segmented 403 lung images. The proposed method yields an accuracy of 88.9% with a specificity of 96.4% on the test data comprising 99 samples. Their study is further extended to three-class (*Normal*, *Pneumonia*, and *COVID-19*) classification, which yielded an accuracy of 91.9%. Though this study achieved high sensitivity for *COVID-19* class (100%), the low precision value (76.9%) for *COVID-19* class is not appropriate for any practical scenario. The DarkNet model was implemented using seven convolutional layers and various filterings on each layer for automatic detection of COVID-19 using the raw chest X-ray images (Ozturk et al., 2020). The model aimed at providing correct diagnostic predictions for binary classification (COVID vs. No-Findings) and multi-class classification (COVID vs. No-Findings vs. *Pneumonia*). This system yielded a classification accuracy of 98.08% for binary classes and 87.02% for multi-class cases for a dataset of a limited number of samples. The performances of various neural architectures for detecting COVID-19 infection from chest X-rays were evaluated in (Apostolopoulos and Mpesiana, 2020). Their results indicated that deep neural

networks aided with X-ray imaging could detect prominent biomarkers pertinent to COVID-19 infection, while the best accuracy, sensitivity, and specificity obtained were reported as 96.78%, 98.66%, and 96.46% respectively. Another study (Afshar et al., 2020) claimed the superiority of COVID-CAPS, a modeling framework based on Capsule Networks with fewer trainable parameters over CNN-based models for COVID-19 detection. COVID-CAPS obtained an Accuracy of 95.7%, Sensitivity of 90%, Specificity of 95.8%, and Area Under the Curve (AUC) of 0.97 for binary classifications. Another study (Ismael and Şengür, 2021) proposed a CNN model with an end-to-end training process for classifying among the chest x-ray images of *Normal* and *COVID-19* classes. Integration of deep learning-based features extractor with Support Vector Machine (SVM) based classifier has achieved an accuracy of 92.6%. However, the binary classification performances were reported on minimal samples. Togaçar et al. (Toğaçar et al., 2020) employed the Fuzzy Color technique for preprocessing the chest X-ray images followed by an image stacking operation to eliminate the existing noises. The integration of deep learning architectures with SVM classifier led to an accuracy of 99.27% for three-class (*Normal*, *Pneumonia* and *COVID-19*) classification. However, the proposed method was validated on a dataset of a total of 458 chest radiograph images. Another similar study (Gupta et al., 2021) showed that the fusion of five deep learning models via integration stacking achieved 99.08% accuracy on limited samples.

2.1.3 Motivation

The body of literature related to COVID-19 infection identification from chest X-rays shows the efficient application of deep learning algorithms. Few of the above-mentioned studies have achieved excellent performances, yet validation on a larger dataset is necessary. All the mentioned studies have investigated the spatial domain characteristics of chest X-rays. To the best of my knowledge, no studies on COVID-19 from chest X-rays have reported investigating the spectral characteristics of the same. This motivates us to explore the unexplored spectral aspect of chest X-rays for COVID-19 infection detection. I aim to validate the hypothesis that the presence of complementary information in the spectral and spatial domain of chest X-ray will improve disease detection ability. I employ two popular tools for transforming images in spectral and spectral as well as spatial domains namely, Discrete Cosine Transform (DCT) and Discrete Wavelet Transform (DWT) to study the spatio-spectral characteristics

of the chest X-ray images. The idea behind the usage of DCT and DWT transformation is to capture any information that is complementary to spatial information and might aid in discriminating among the three classes considered in this study. The DCT decomposes the image into several spectral sub-bands with a cosine function as a basis function. In contrast, DWT has the advantage of assimilating both spatial and spectral domains simultaneously. Hence, chest radiograph images are studied in three different domains concurrently, namely, pixel, DCT, and DWT, to bring about the potential of each of the fields in characterizing the considered medical conditions. The patterns that characterize different medical conditions are complex and intricate. State-of-the-art CNN shows notable efficiency in modeling complex patterns in the domain of image classification. Hence, I employ ResNet50, a state-of-the-art CNN architecture, to extract features from the pixel, DCT, and DWT domain images. The features extracted in these three domains are integrated, and the final class prediction is performed using a multilayer perceptron (MLP) network. The detailed description of my proposed method is given in Section 2.2. Medical imaging datasets often have a limited number of samples for comparatively uncommon disease classes. This class imbalance usually leads to inferior performances corresponding to the minority class. Relevant techniques, such as class weight and unbiased validation performance metrics, have been employed to prevent such undesired outcomes and are discussed in Section 2.2 in detail.

Furthermore, I extend my experiment to six other medical imaging datasets. These data sets incorporate a variety of imaging techniques, the number and types of diseases to be detected, and the class imbalance. This extended study aims to validate the hypothesis of the presence of complementary information in the spectral and spatial domain of medical images. Overall, this study indicates the proposed method's generalization ability toward a diagnostic solution using medical images.

2.1.4 Contributions

The contributions of this study are listed below.

- I propose an automatic computer vision and machine learning-based diagnostic solution for medical images developed on the complementary knowledge of the spatial and spectral domain.

- The proposed method is validated by carrying out experiments on eight diverse medical imaging datasets, suggesting its robustness and capability of generalization.
- The classification performances, along with saliency maps, demonstrate the fusion of spatial, spectral, and spatio-spectral domain features enhances the disease detection capability of the classification model.
- Analyzing the classification performance on the COVID-19 dataset reveals that the performance of the proposed method is unbiased to age and gender factors.

The rest of the chapter is structured as follows. Along with the detailed description of the datasets, the methodology used in this study is presented in Section 2.2. The experimental results and relevant discussion are given in Section 2.3 and Section 2.4 respectively.

2.2 Proposed Method

2.2.1 Discrete Cosine Transform

The Discrete Cosine Transform (DCT) of an image is a real transformation that transforms the image from spatial domain to frequency domain by linear combinations of weighted basis functions pertinent to its frequency components (Ahmed et al., 1974). DCT of an image X of dimension $N \times N$ is given by the following equation.

$$\begin{aligned} \mathrm{DCT}_x(u,v) = \frac{2}{N} C(u)C(v) \sum_{m=0}^{N-1} \sum_{n=0}^{N-1} X(m,n) \\ \cos[\frac{\pi(2m+1)u}{2N}] \cos[\frac{\pi(2n+1)v}{2N}] \end{aligned} \tag{2.1}$$

where, $X(m,n)$ denotes the pixel value X in (m,n) coordinate, $u = 0, ..., N-1$, $v = 0, ..., N-1$ and $C(k) = \begin{cases} \frac{1}{\sqrt{2}} & if\ k = 0 \\ 1 & otherwise. \end{cases}$

In this study, images are of dimension $224 \times 224 \times 3$. DCT is applied to the images by considering each channel separately over the segmented patches of size 8×8. Thus, the dimension of the DCT images is the same as the input images.

2.2.2 Discrete Wavelet Transform

The discrete wavelet transform (DWT) uses multiresolution filter banks to perform the wavelet analysis (Shensa, 1992). It represents the signal in terms of the wavelet coefficients from which it is possible to reconstruct the original signal once again. The signal is represented in various frequency bands by the wavelet coefficients. There can be several ways to process these coefficients, thus endowing DWT with attractive properties over linear filtering.

The DWT is applied to the image X of size $(N \times N)$ to achieve four decomposed subband images of $(N/2 \times N/2)$ dimension. The process includes the the application of a set of half-band low-pass and high-pass filters to the rows of the image,, followed by the decimation by a factor of 2. The same procedure is applied to the two subband images obtained in the previous step, but this time, it is along the columns. Thus, it results in decomposed images in four different frequency bands, which can be mathematically expressed as follows,

$$\mathrm{LL}(x,y) = \sum_m \sum_n h_0(m-2x)h_0(n-2y)X(m,n). \tag{2.2}$$

$$\mathrm{HL}(x,y) = \sum_m \sum_n h_0(m-2x)h_1(n-2y)X(m,n). \tag{2.3}$$

$$\mathrm{LH}(x,y) = \sum_m \sum_n h_1(m-2x)h_0(n-2y)X(m,n). \tag{2.4}$$

$$\mathrm{HH}(x,y) = \sum_m \sum_n h_1(m-2x)h_1(n-2y)X(m,n). \tag{2.5}$$

Here, h_0 and h_1 are half-band low and high pass filters, respectively. Thus, the LL subband approximates the image, i.e., the low-frequency content of the image. In contrast, three other bands, i.e., LH,HL, and HH subbands, contain the details, i.e., the high-frequency content of the image. The four subband images of size $N/2 \times N/2$ are arranged in the manner as shown in Figure 2.1 to form one image of size $N \times N$. Similar to DCT, DWT is also applied individually on each of the channels of images, and the transformed image dimension is the same as the input image dimension.

2.2.3 Machine Learning Framework

The block diagram presented in Figure 2.1 outlines the complete framework for the proposed diagnostic system. The proposed system consists of three stages — projection by using DCT and DWT, complex feature extraction, and prediction. In the first stage, the images are projected to spectral as well as spatio-spectral domains by the application of DCT and DWT, respectively, as described in Section 2.2.1 and Section 2.2.2 respectively. Hence, at the end of this stage, I get three types of images of equal dimension in three different domains — pixel, DCT, and DWT (shown in Figure 2.1).

The objective of the second stage is to extract the features from the output images of the previous stage. The underlying pattern that distinguishes one medical condition from another is intricate in nature. Convolutional neural networks (CNN) can extract complex features from input images. ResNet50 is one of the widely used CNN architectures in the domain of image categorization. Thus, I choose to employ ResNet50 to extract complex features. Moreover, the reasonable number of trainable parameters of ResNet50 makes it suitable for medical image classification tasks where there are limited training samples available. Three separate ResNet50 models are trained for three different types of images. The output feature is extracted from an intermediate layer of trained ResNet50 models, which results in complex features of dimension 1024. For training each of the ResNet50 models, the categorical cross-entropy loss function is optimized with the Adam optimization algorithm with a learning rate of 0.0001 for 300 epochs with a batch size of 32. The model with the best validation ACSA (defined in Section 2.3.2) is chosen for feature mapping. ACSA is chosen for this purpose as this measure is unbiased to class imbalance present in the training data.

The third stage comprises feature-level fusion followed by final prediction using a classification head. Three types of feature vectors, each of dimension 1024, are combined in all seven possible combinations by the concatenation operation. The dimension of the concatenated feature vector is presented in Table 2.1. The resulting feature vector is fed to a classification head for final prediction. The Multilayer Perceptron (MLP) is chosen as the classification head because of its efficiency in modeling complicated nonlinear relationships between input and output vectors. The MLP network comprises one input layer, two hidden layers with 256 and 64 nodes, and one output layer. The leaky ReLU activation function is applied after each

Table 2.1: All possible combinations of pixel, DCT, and DWT features and corresponding integrated feature vector's dimension

Combination Type	Dimension of Feature Vector
pixel	1×1024
DCT	1×1024
DWT	1×1024
pixel+DCT	1×2048
pixel+DWT	1×2048
DCT+DWT	1×2048
pixel+DCT+DT	1×3072

layer except for the last layer, which uses softmax as the activation function. To train the MLP network, the mean square error loss function is optimized with the Adam optimization algorithm with a learning rate of 0.0002 for 300 epochs with a batch size of 32. The model with the best training ACSA is evaluated with the test data.

The effect of class imbalance present in the datasets is handled by incorporating the penalty factor of class weights computed as $N/(cN_i)$ to the loss function while training the ResNet50 model and MLP model, where c is the number of classes, N and N_i is the number of total samples and the number of samples that belong to i^{th} class respectively.

All the experiments (Python scripts) are executed using Keras with TensorFlow as back-end on a computer with Intel core i5 processor running at 2.40 GHz using 16 GB of RAM and NVIDIA GeForce RTX 2060 GPU with 6 GB RAM. The code is available at this GitHub repository https://github.com/SusmitaSenGhosh/COVID-detection-from-X-ray-using-DWT-DCT-.

2.3 Experimental Results

2.3.1 Dataset Details

I validate my hypothesis primarily on a dataset comprising 15476 chest-ray images from 15279 persons who belong to any of the three categories representing three types of medical conditions, namely —*Normal*, *Pneumonia* and *COVID-19*. Images belonging to *COVID-19* category have been taken from these four sources (Chung, 2020a,b; Cohen et al., 2020; RSNA, 2020) whereas samples of *Pneumonia* and *Normal* class been collected from these sources (Cohen et al., 2020; RSNA, 2018). This dataset is referred to as the COVID-19 dataset. Collecting samples from multiple sources leads to diversity in the types of the images. Thus,

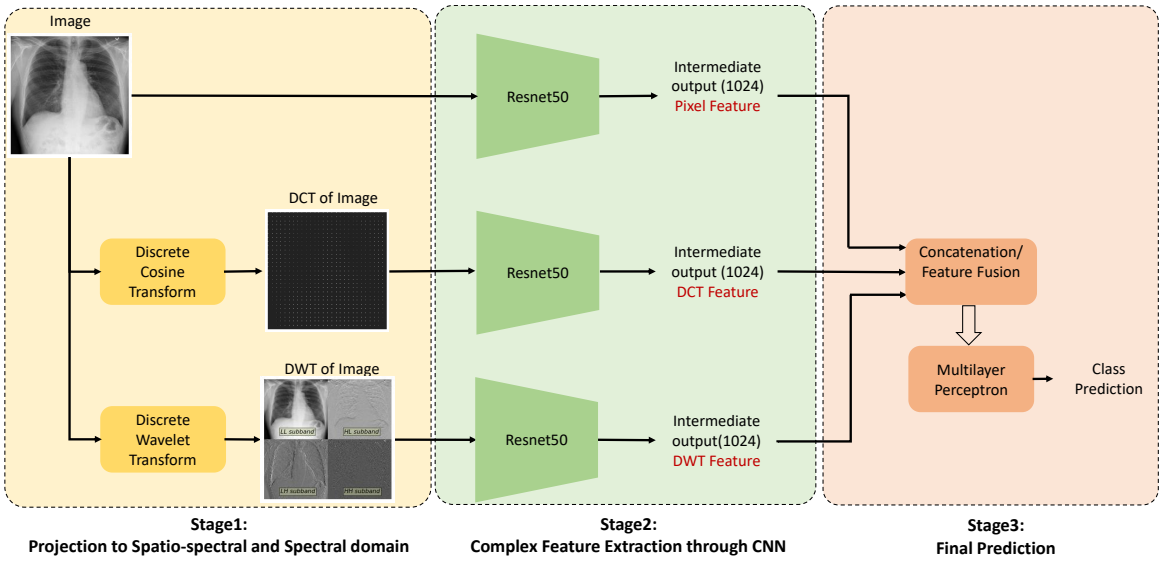

Figure 2.1: Schematic diagram of the classification framework

the experimental results on this dataset are robust and close to a practical scenario. The number of samples and the number of patients that belong to each of the three classes are stated in Table 2.2. Among the 15279 patients considered in this study, the ages and genders of 15029 and 15112 patients are known, respectively. The age distribution of the subjects in each category and gender distribution in each class is shown in Figure 2.2. The dataset suffers from the problem of class imbalance. The imbalance ratio (ρ), defined by the ratio of the number of samples in the majority class to that of the minority class, is 15.45. The chest X-ray images are of different sizes, thus in the preprocessing step, I resized it to the size of $224 \times 224 \times 3$ where the third dimension represents the number of channels. The results reported in this study are yielded by five-fold cross-validation. Out of the five folds, one is kept for testing, another for validation purposes, and the rest of the folds are used to train the model. While partitioning the data into five-folds, it is always ensured that samples from one patient are never grouped into two or more folds to maintain the integrity of the reported result.

Additionally, seven medical imaging datasets of different modalities, e.g., X-ray, histopathology, mammography, etc, and medical conditions are also used for the validation of the hypothesis. The diversity in the number of classes, i.e., the medical conditions, image modalities, and the imbalance ratio, makes the experiment more robust. A detailed description of each

Table 2.2: Number of samples and patients belonging to *Normal*, *Pneumonia* and *COVID-19* class

	Number of Samples	Number of Patients
Normal	8851	8851
Pneumonia	6052	6034
COVID-19	573	394
Total	15476	15279

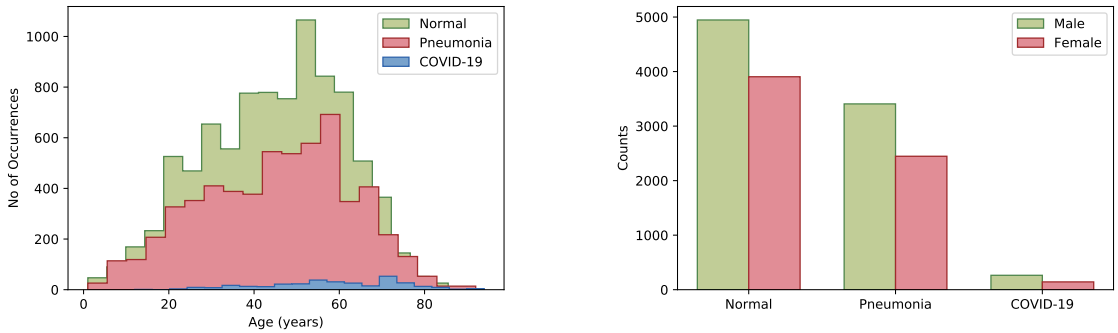

Figure 2.2: Age and gender distribution of persons belong to *Normal*, *Pneumonia* and *COVID-19* classes are presented in the upper and lower panel of the figure, respectively.

of the datasets used in this study is presented in Table 2.3 and additional information about the dataset is provided in A.1 I have used a subset of these original CBIS-DDSM (Sawyer-Lee et al., 2016), DR and BHI (Janowczyk and Madabhushi, 2016) datasets as they comprise a large number of samples. Each dataset is split into training and testing samples maintaining an 80:20 ratio, except for Chestxray1 (Wang et al., 2020) where 100 test samples are provided for each of the three classes.

Table 2.3: Detailed description of medical imaging datasets used in this study.

Dataset	Type of data	Number of classes	Number of Training samples	Imbalance Ratio
Colorectal Histology (CH) (Kather et al., 2016)	Histology images	8	4000	01.00
Diabetic Retinopathy (DR) [1]	Diabetic Retinopathy	2	13052	01.33
BHI (Janowczyk and Madabhushi, 2016)	Histopathology images	2	12952	02.79
Chestxray (Kermany et al., 2018)	Chest x-ray images	2	5232	02.88
CBIS-DDSM (Sawyer-Lee et al., 2016)	Mammography Images	2	8941	06.55
Chestxray1 (Wang et al., 2020)	Chest x-ray images	3	13898	16.84
ISIC18 (Codella et al., 2019)	Dermoscopic images	7	8166	58.86

[1] https://www.kaggle.com/c/diabetic-retinopathy-detection

2.3.2 Performance Metrics

Let **C** be the confusion matrix for n-class classification where element C_{ij} indicates the number of samples of i^{th} class predicted as of j^{th} class. The following metrics presented in terms of the elements of the confusion matrix are used to quantify the classification capability of machine learning models.For notational simplicity, I refer to the F1-score as F1 in all subsequent equations.

$$\text{Sensitivity}_i = \frac{C_{ii}}{\sum_j C_{ij}}, \tag{2.6}$$

$$\text{Precision}_i = \frac{C_{ii}}{\sum_j C_{ji}}, \tag{2.7}$$

$$\text{F1}_i = \frac{2 \cdot \text{Sensitivity}_i \cdot \text{Precision}_i}{\text{Sensitivity}_i + \text{Precision}_i}. \tag{2.8}$$

Along with the above-mentioned performance metrics, three other measurements are also used to quantify the performances of models — average class-specific accuracy/sensitivity (ACSA), average class-specific precision (ACSP), and an average class-specific F1-score (ACSF), respectively indicating sensitivity, precision, and F1-score averaged over all the classes, thus imparting equal priority to all the classes. Hence, these metrics quantify the model's performance without being influenced by the existing class imbalance. The mean absolute deviation of the F1-score (MADF) is also reported along with ACSF to quantify the spread of individual class-wise F1-scores.

$$\text{MADF} = \frac{1}{n}\sum_i |\,\text{F1}_i - \frac{1}{n}\sum_i \text{F1}_i\,|. \tag{2.9}$$

2.3.3 Performance on COVID-19 Dataset

Due to the severity of the COVID-19 outbreak, the primary focus is paid to the elaborated study that is conducted on the COVID-19 dataset (described in Section 2.2). The result obtained from the above-mentioned experiment (Section 2.1) is presented in Table 2.4. The classification performances are quantified by class-wise sensitivity, precision, and F1-score. Additionally, I present sensitivity, precision, and F1-score averaged over the different classes to measure overall performance. They are mentioned as ACSA, ACSP, and ACSF, and these metrics are defined in Section 2.3.2. Table 2.4 compares the performance of the seven feature combinations as listed in Table 2.1. By comparing the performances of three types of features

(pixel, DCT, and DWT) when used individually, pixel and DWT features yield better results than DCT features. However, the fusion of any two types of features shows better performance than any of the features while used solely for classification. Furthermore, the best classification performance is achieved when all three types of features, i.e., pixel, DCT, and DWT, are combined altogether (pixel+DCT+DWT). The corresponding values of ASCA, ACSP, and ACSF are 93.78%, 91.30%, and 92.42%, respectively. It is important to note that the above-mentioned feature combination also yields the best sensitivity to the COVID-19 class, which is 95.28,%, as I aim to detect COVID-19 infection with sensitivity as high as possible. These observations suggest the presence of complementary information in pixel, DCT, and DWT features. To validate the above-said statement from a statistical perspective, the Wilcoxon rank-sum test is performed on the ACSA, ACSP, and ACSF values. Table 2.5 summarises the result obtained from tests, where the null hypothesis is that the performance of three features combined together (pixel+DCT+DWT) is equivalent to that of other combinations of the features. The null hypothesis is rejected when the p-value is less than 0.05, confirming that the samples belong to different distributions, whereas a larger p-value suggests that the two distributions are similar. Furthermore, if test statistics come out to be negative with a p-value less than 0.05, I infer that the performance of pixel+DCT+DWT is superior to that of the other feature combination considered in the test. Table 2.5 shows that the test statistics are never positive, indicating performance of pixel+DCT+DWT is superior (in most of the cases, marked by *) or comparable (in a few cases) with the performance of the rest of the feature combinations. Thus, the outcome indicates that both DCT and DWT features capture information complementary to that of the pixel feature that enhances the discriminating capability among the three classes considered in this study. Hence, my hypothesis of the existence of complementary information in the spatial and spectral domain of medical images is validated for the COVID-19 dataset.

The number of the hidden layers and the number of nodes in hidden layers of the MLP network are similar for all the combinations of feature types. However, the number of nodes in the input layers varies depending on the kind of feature integration. For example, the number of nodes in the input layer of MLP is 1024 for pixel, DCT, and DWT features, 2048 for Pixel+DCT, pixel+DWT, and DCT+DWT feature combination, and 3072 for pixel+DCT+DWT feature combination. Consequently, the number of trainable param-

Table 2.4: The classification performances of all possible combinations of the pixel, DCT, and DWT features in the detection of *Normal*, *Pneumonia* and *COVID-19* classes are presented. Performances are quantified by sensitivity, precision. Average class-specific sensitivity, specificity, and F1-score (ACSA, ACSP, and ACSF) are also reported. All the performance metrics are reported on 5-fold cross-validation. Comparing different feature combinations, the best values of the metrics are marked in bold.

Feature Combinations	Sensitivity(%)			ACSA	Precision(%)			ACSP	F1-Score(%)			ACSF
	Norm.	Pneum.	COVID		Norm.	Pneum.	COVID		Norm.	Pneum.	COVID	
Pixel	94.55 ±0.94	90.26 ±0.65	92.60 ±3.59	92.47 ±1.49	93.89 ±0.53	92.04 ±1.25	84.78 ±3.10	90.24 ±1.40	94.21 ±0.55	91.14 ±0.75	88.44 ±2.79	91.26 ±1.30
DCT	93.15 ±0.70	87.21 ±1.72	80.93 ±4.58	87.10 ±1.21	91.66 ±0.69	89.40 ±0.96	81.18 ±4.86	87.41 ±1.60	92.39 ±0.42	88.27 ±0.68	80.82 ±1.81	87.16 ±0.76
DWT	95.06 ±0.43	89.75 ±0.98	92.15 ±2.38	92.32 ±0.88	93.51 ±0.52	92.79 ±0.56	84.93 ±5.20	90.41 ±1.72	94.28 ±0.39	91.24 ±0.63	88.30 ±3.12	91.27 ±1.14
Pixel+DCT	95.18 ±0.56	90.27 ±0.96	93.25 ±3.41	92.90 ±1.19	93.93 ±0.56	92.91 ±0.74	85.34 ±3.65	90.72 ±1.44	94.55 ±0.49	91.56 ±0.81	89.02 ±02.41	91.71 ±01.06
Pixel+DWT	95.15 ±0.99	**90.91** ±0.64	94.86 ±2.75	93.64 ±0.94	**94.30** ±0.44	93.04 ±1.23	**86.39** ±4.91	91.24 ±1.86	94.72 ±0.58	91.96 ±0.85	90.32 ±2.47	92.33 ±1.08
DCT+DWT	95.31 ±0.55	89.72 ±1.38	93.03 ±1.89	92.69 ±0.63	93.51 ±0.82	93.04 ±0.71	86.45 ±3.21	91.00 ±1.29	94.40 ±0.52	91.34 ±0.87	89.58 ±2.25	91.77 ±0.93
Pixel+DCT +DWT	**95.49** ±0.85	90.58 ±0.80	**95.28** ±2.52	**93.78** ±0.78	94.13 ±0.49	**93.51** ±1.08	86.26 ±4.12	**91.30** ±1.62	**94.80** ±0.53	**92.01** ±0.85	**90.45** ±1.74	**92.42** ±0.86

Table 2.5: Statistics and p-value obtained in Wilcoxon rank-sum test with performance of pixel+DCT+DWT feature combination is compared with that of the rest of the feature combinations. The statistics indicating the superiority of pixel+DCT+DWT feature combinations are marked with *.

	ACSA		ACSP		ACSF	
	Statistics	p-value	Statistics	p-value	Statistics	p-value
Pixel	-5.04*	<0.05	-2.53*	<0.05	-4.37*	<0.05
DCT	-8.61*	<0.05	-2.82*	<0.05	-8.62*	<0.05
DWT	-7.05*	<0.05	-0.25	0.80	-4.82*	<0.05
Pixel-DCT	-4.55*	<0.05	-1.09	0.27	-2.67*	<0.05
Pixel-DWT	-0.12	.90	-8.50*	<0.05	-0.68	0.5
DCT-DWT	-6.38*	<0.05	-0.62	0.53	-3.03*	<0.05

eters also increases with the number of nodes in the input layers. Thus, the question may arise whether the superior performance of the combined pixel+DCT+DWT features is due to the integration of the features or the increased number of trainable parameters of the model. In search of the answer, I performed another experiment, where the considered architecture of the model is the same as it is used for pixel+DCT+DWT features. I construct the new feature vector for pixel by concatenating the pixel feature thrice so that the dimension of the new pixel feature vector becomes 1×3072 vector. A similar operation is applied to DCT and

Table 2.6: The performances of the new pixel, DCT, and DWT features evaluated on the MLP model that has similar architecture as used in the case of pixel+DCT+DWT features. The new pixel, DCT, and DWT feature vectors are constructed by concatenating each type of feature three times.

	Sensitivity(%)			ACSA	Precision(%)			ACSP	F1-Score(%)			ACSF
Feature	**Norm.**	**Pneum.**	**COVID**		**Norm.**	**Pneum.**	**COVID**		**Norm.**	**Pneum.**	**COVID**	
Pixel	94.84 ±1.22	90.10 ±0.54	93.65 ±3.80	92.86 ±1.54	93.96 ±0.23	92.66 ±1.81	81.85 ±4.17	89.49 ±1.69	94.39 ±0.61	91.35 ±0.80	87.30 ±3.28	91.02 ±1.47
DCT	93.08 ±0.72	86.91 ±1.69	83.64 ±3.55	87.88 ±0.84	91.72 ±0.58	89.65 ±1.08	77.69 ±7.11	86.35 ±2.34	92.38 ±0.41	88.24 ±0.66	80.25 ±2.92	86.96 ±1.22
DWT	94.86 ±0.81	90.01 ±1.23	92.42 ±2.82	92.43 ±1.06	93.77 ±0.65	92.62 ±1.01	83.31 ±5.76	89.90 ±1.90	94.30 ±0.42	91.28 ±0.68	87.53 ±3.81	91.04 ±1.37
Pixel+DCT +DWT	**95.49** ±0.85	**90.58** ±0.80	**95.28** ±2.52	**93.78** ±0.78	94.13 ±0.49	**93.51** ±1.08	**86.26** ±4.12	**91.30** ±1.62	**94.80** ±0.53	**92.01** ±0.85	**90.45** ±1.74	**92.42** ±0.86

Table 2.7: Comparison of performance of proposed method with COVID-net ((Wang et al., 2020))

	Sensitivity(%)			ACSA	Precision(%)			ACSP	F1-Score(%)			ACSF
	Norm.	**Pneum.**	**COVID**		**Norm.**	**Pneum.**	**COVID**		**Norm.**	**Pneum.**	**COVID**	
COVID-net	**95.00**	**94.00**	91.00	93.33	90.50	91.30	**98.90**	93.57	92.70	**92.63**	94.79	93.37
Proposed	94.80	91.80	**94.30**	**93.63**	**91.60**	**92.55**	96.92	**93.69**	**93.17**	92.17	**95.59**	93.64

DWT features. The performances of these new feature vectors are evaluated on the above-mentioned MLP. Table 2.6 compares these performances with pixel+DCT+DWT features (last row of Table 2.4). The test-statistics obtained from Wilcoxon rank-sum tests confirm the superiority of the performance of combined features concerning that of single features where the model complexity is kept unchanged.

2.3.3.1 Interpreting Classification Model

In this section, I attempt to understand how information from DCT and DWT domains contributes to enhancing the discriminative potential of the model. For this purpose, the saliency map, i.e., the gradient of the class activation function concerning the input images, is visualized. Thus, the saliency map for a particular class quantifies the amount of change in classification score caused by the slight change in image pixel (Smilkov et al., 2017) indicating the decisive regions in the image. In Figure 2.3(a), the saliency map produced by the proposed model is shown for one image and its corresponding DCT and DWT images from each class. It confirms that all three types of images contribute to decoding the classes. Moreover, the

saliency map of my model for each class validates its reliability as the highlighted regions from each of the images lie in the lung and its surrounding area. It is also noticed that the subband of DWT that represents the higher frequency component of the images does not contribute significantly towards disease detection.

Moreover, I inspect the saliency map of chest X-ray images to learn if there is any apparent pattern corresponding to each class and correlate them with existing literature in the medical domain. Four such maps from each class are presented in Figure 2.3(b). While the saliency map of *Normal* class shows a wide variety in the area of chest X-ray that is highlighted, in the case of *Pneumonia* class, the influential pixel cluster around the lower lobes of both or either lung. On the other hand, along with the other lung regions, the upper lobe (bilateral) area of the lung is found to be persistently dominating in the case of *COVID-19* class. The bilateral Ground-glass opacities (GGO) have been reported in COVID-19 chest X-rays, whereas the unilateral and central distribution of GGO has been found in chest X-rays of Pneumonia patients (Ozturk et al., 2020). These characteristics are consistent with the patterns shown in the saliency maps of *COVID-19* and *Pneumonia* classes.

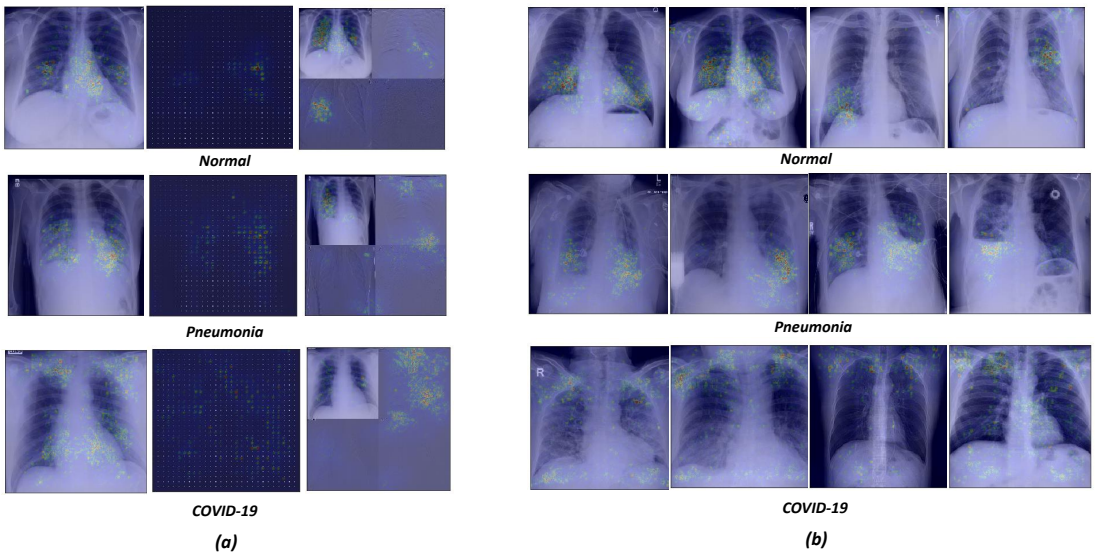

Figure 2.3: (a) Saliency maps corresponding to *Normal*, *Pneumonia* and *COVID-19* classes are presented for chest X-ray image(left panel), DCT image (middle panel) and DWT image (right panel). The highlighted regions play an important role in characterizing the three classes. (b) Saliency maps of chest X-ray images sampled down from three classes are presented.

2.3.3.2 Effect of Age and Gender on Classification Performance

The ages of the patients considered in this study varied from as low as 1 year to as high as 94 years Figure 2.2. To analyze if there is any biasedness of age on the classification performance, I carry out an analysis by clustering samples into five age groups – 0-20 years, 20-40 years, 40-60 years, 60-80 years, and 80-100 years. For each class, I investigated how the performance of the classifiers varies over different age groups. Figure 2.4(b) depicts the fraction of samples of a particular class classified as *Normal*, *Pneumonia*, and *COVID-19* classes. It is observed that over different age groups, the proposed method yields a similar trend, suggesting that the age factor does not influence my classification results.

Furthermore, another similar analysis is executed to investigate the effect of the gender of the patients on the classification result. The outcome of the analysis is shown in Figure 2.4(b), which demonstrates that the gender of the patients does not behave as a factor in classification performance. All the classification performances reported in this section are yielded using the combined features (i.e., pixel, DCT, and DWT).

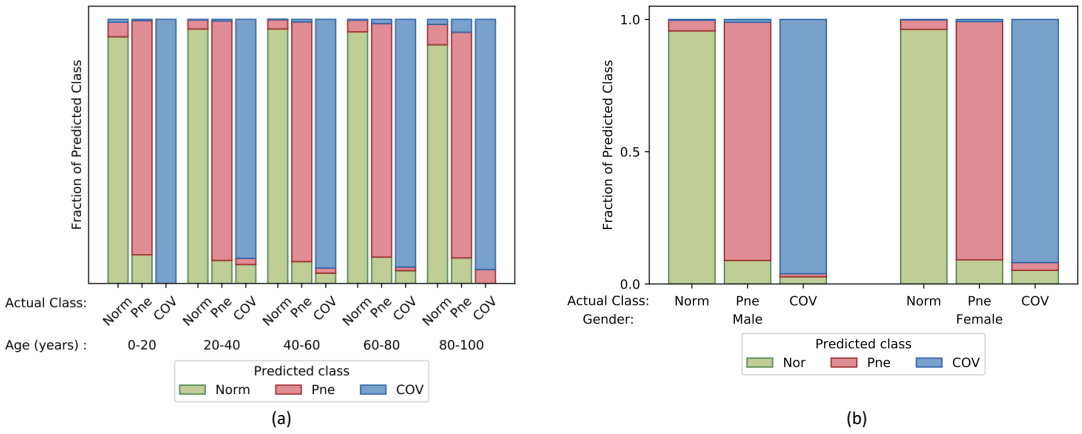

Figure 2.4: Age wise and gender wise classification performance is shown in (a) and (b) respectively. *Norm*, *Pne*, and *COVD-* represent *Normal*, *Pneumonia*, and *COVID-19* respectively. The height of the three stacked-up bars indicates the fraction of a particular class that belongs to three classes. For instance, the green, red, and blue bars in the leftmost bar of the panel (a) indicate the fraction of *Normal* samples (of 0-20 years age group) that is classified as *Normal*, *Pneumonia* and *COVID-19* class.

Table 2.8: The performances of pixel, DCT, DWT, and pixel+DCT+DWT in identifying different medical conditions for different datasets as listed in Table 2.3

	Pixel		DCT		DWT			Combined	
Dataset	**ASCF**	**MADF**	**ASCF**	**MADF**	**ASCF**	**MADF**	**Combination Type**	**ASCF**	**MADF**
BHI	89.85 ±0.28	05.09 ±0.18	87.07 ±0.21	06.33 ±0.18	85.97 ±0.33	07.02 ±0.18	pixel+DCT+DWT, SF	**90.40** ±0.17	**04.81** ±0.10
CBIS_DDSM	92.79 ±0.26	05.36 ±0.21	78.82 ±0.43	15.78 ±0.53	87.05 ±0.26	09.51 ±0.23	pixel+DCT+DWT, FF	**93.00** ±0.48	**05.17** ±0.37
CH	95.32 ±0.32	03.34 ±0.33	92.33 ±0.36	05.07 ±0.42	94.02 ±0.20	04.04 ±0.28	pixel+DWT, FF	**95.62** ±0.23	**02.96** ±0.24
Chestxray1	90.00 ±0.45	**01.04** ±0.21	87.34 ±0.49	00.82 ±0.24	90.02 ±0.45	01.95 ±0.22	pixel+DCT+DWT, FF	**93.64** ±0.43	01.44 ±0.34
Chestxray	72.87 ±2.17	12.44 ±1.37	69.58 ±1.44	14.57 ±0.89	76.18 ±1.51	10.49 ±0.93	pixel+DWT, SF	**76.42** ±1.44	**10.35** ±0.89
DR	69.49 ±0.36	04.24 ±0.30	60.14 ±0.45	06.67 ±1.51	64.73 ±0.67	04.19 ±1.30	pixel+DCT, FF	**69.84** ±0.64	**03.80** ±0.38
ISIC18	75.86 ±0.87	**09.49** ±0.44	64.51 ±0.76	13.93 ±0.82	64.03 ±1.21	15.17 ±0.74	pixel+DCT+DWT, FF	**77.69** ±0.79	09.69 ±0.75

2.3.3.3 Comparison with COVID-net

I compare my result with the state-of-the-art experimental result reported in (Wang et al., 2020). For this purpose, I apply my methodology to the same dataset as mentioned in (Wang et al., 2020). The comparison in terms of sensitivity, precision, and F1-score is presented in table 2.7. The main aim of this study is to detect the COVID-19 class with as high sensitivity as possible and also maintain reasonably high precision at the same time. Though the value of the average class-specific accuracy, precision, and F1-score obtained in the proposed method is comparable with the result presented in (Wang et al., 2020), however, the higher sensitivity to *COVID-19* class of the trained model makes my proposed method more suitable for practical implementations.

2.3.4 Evaluation on Extended Datasets

I validate my proposed method on other medical imaging datasets as described in section 2.2. These datasets comprise different modalities, medical conditions, and class imbalance ratios. Since the number of classes varies in the range from two to eight, instead of showing class-specific metrics, I present average class-specific metrics. Additionally, the mean absolute deviation of the F1-score (MADF) is used to quantify the spread of the individual class-wise F1-score. The lower the value of MADF, the lower the dispersion among the F1-score of each class. Table 2.8 presents the classification performance for seven datasets listed in Table

2.3 for Pixel, DCT, and DWT features individually. The efficiency of these features is also tested by performing feature level fusion (FF) and score level fusion (SF). The best result obtained in feature level fusion or score level fusion is reported in Table 2.8. The detailed results, including precision and accuracy values, are presented in Table A.2. It is observed that the combination of two or three types of features among pixel, DCT, and DWT has surpassed the result obtained while using each of the features separately. Moreover, the lower MADF with a higher ASCF of combined features also indicates superiority in handling class imbalance. Thus, the result obtained in this experiment supports my hypothesis of the presence of complementary information in spatial and spectral domains in medical images that assist in decoding the underlying medical condition.

2.4 Discussion

In this study, I present a novel medical image-based diagnostic solution for the detection of various underlying medical conditions. Along with the spatial information of the medical images, I exploited the less explored spectral-domain information of the medical images. I aspire to enhance disease detection performance by integrating features from spatial, spectral, and spatio-spectral space. The proposed system comprises three stages — conversion of a spatial image into the spectral and spatio-spectral domains, feature mapping from higher to lower dimensions using CNN architecture, and final classification using MLP. The potential and the robustness of the proposed method are demonstrated by using eight diverse medical imaging datasets of different modalities, diseases to be identified, and class imbalance ratio.

The results suggest that the spatial, spectral, and spatio-spectral domain features are solely capable of representing distinguishing characteristics of the underlying medical conditions. However, integrating those three types of features results in a significant increment in classification performance, suggesting that all three types of features possess complementary information. The saliency maps also validate the integrity of the proposed model. A detailed study on COVID-19 shows that my method's performance is unbiased regarding gender and age factors. In comparison to the method proposed in (Wang et al., 2020) for COVID-19 detection, my method achieved significantly higher sensitivity to *COVID-19* class and also marginal improvement in average class-specific accuracy, precision, and F1-score.

Altogether, my study demonstrates a novel approach in identifying various diseases using the medical images that can assist the healthcare worker in the primary screening process.

This study can be extended to other available medical images for generalization purposes. Implementing different deep learning algorithms for feature extraction can be explored to enhance performance. Three separate ResNet50 networks are trained for extracting features from pixel, DCT, and DWT images. As a future research direction for this study, I will consider overcoming this limitation.

Chapter 3

An Improved Vision Transformer Model for Medical Image based Diagnostic Solution

Summary

The Vision Transformer (ViT) has demonstrated promise in image classification tasks, but its performance in medical image analysis is often limited by the complexity of medical images and the scarcity of training samples. To address these challenges, I propose a modified architecture called ReMixViT, which incorporates an efficient MLP-Mixer layer and reorders the residual blocks within the encoder to improve feature interaction and generalization. Furthermore, I extend this work by designing two hybrid architectures, ResReMixViT and ResReMixViT+, which integrate ResNet50 with ReMixViT encoder blocks to handle feature maps of single and multiple scales, respectively. These architectures were evaluated on six diverse medical imaging datasets covering different modalities and conditions.

The comparative analysis shows that ReMixViT and the proposed hybrid models significantly outperform both the standard ViT models and existing hybrid approaches that use ViT encoders. Specifically, the hybrid architectures achieved improvements of 4.62% and 3.08% in the F1-score, demonstrating their effectiveness. When coupled with data augmentation techniques, these models further surpassed other state-of-the-art hybrid networks. In addition to performance metrics, I provide visual explanations through attention maps and gradient flow visualizations, which enhance the interpretability of the proposed model. This not only offers transparency from an explainable AI perspective but also aids medical professionals in

making informed decisions based on model outputs. The extended study also reveals that the proposed modifications can be effectively applied to other ViT-based architectures, yielding similar performance gains across various medical image classification tasks.

3.1 Introduction

3.1.1 Overview

The success rate and effectiveness of any medical treatment are enhanced by the proper diagnosis of the underlying medical condition. The accurate diagnosis of diseases needs rigorous, arduous training. Moreover, the availability of skilled people often does not meet the demand, specifically in rural areas. Applying machine learning algorithms to clinical images to provide primary detection of the underlying medical condition may ease the diagnostic process. However, further assessment by the domain expert is essential as human health is discussed here. The performance of such computer-aided diagnostic systems improves with the evolution of machine-learning algorithms. One necessary aspect of such a diagnostic system is that it should not only predict the right medical condition but also explain its decision as a visual aid. Thus, domain experts can assess the reliability of the system.

The automated diagnostic system faces two significant challenges — extraction of intricate, distinct patterns and scarcity of training data. The convolutional neural network (CNN) based deep learning models dominate the automated categorization of images, including clinical images (Nardelli et al., 2018; Alzubaidi et al., 2022; Shin et al., 2016; Jain et al., 2019; Szegedy et al., 2015; Rocha et al., 2023; Malik et al., 2023; Ghosh et al., 2021) for years. Though the convolution layer is very expressive in extracting complex local features, it lacks in learning the global portrayal of images, i.e., the associations among the local features due to its local receptive field. The view of the receptive field can be broadened at the cost of a heavy computation burden that is far from practical. Transformer, an attention-based deep learning model first coined by Vaswani et al. (Vaswani et al., 2017), has thrived in the natural language processing domain. Though the transformer architecture model entirely discards convolution layers, it has found its potential in computer vision applications (Dosovitskiy et al., 2020; Touvron et al., 2021a; Dai et al., 2021) by overcoming the limitation of inductive convolutional biases. Vision Transformer (ViT) (Dosovitskiy et al., 2020) is the first adapta-

tion of the transformer in image analysis. Afterward, the efficient application of transformer encoder blocks has been reported in various computer vision studies (Wang et al., 2021a; Liu et al., 2021; Ranftl et al., 2021) for classification as well as segmentation purposes. ViT employs the self-attention mechanism of transformers on the patches extracted from input images. Thus, it produces a simply explainable illustration by highlighting the contributing areas of an image. However, it lacks in extracting complex distinguishing patterns from the images. Thus, the subtle characteristics that play an important role in detecting medical conditions from images are not learned by the model efficiently. So, I propose a modified ViT architecture, referred to as ReMixViT, intending to enhance the model's ability to learn to extract delicate and complex patterns. The conventional ViT encoder block comprises a Multihead Self Attention (MSA) residual block and a Multilayer Perceptron (MLP) residual block. I make two changes in the conventional ViT encoder block. Firstly, I replace the MLP residual block with MLP-Mixer (Tolstikhin et al., 2021) block, which is based on MLP operation applied on both channel and spatial locations sequentially. Thus, MLP-Mixer establishes relations among different feature channels and spatial locations defined by the patches of the images. This attribute makes MLP-Mixer more efficient than MLP in extracting features by allowing two-way communication. Secondly, I swap the order of the residual blocks. Thus, MSA is applied to the output of the MLP-Mixer residual block instead of applying MSA to image patches. I anticipate this modification will improve performance as MSA highlights the relations among the complex features modeled by MLP-Mixer. The experimental results support my hypothesis of achieving superior performance with the proposed ReMixViT architecture.

I extend my study by considering a hybrid approach where the ReMixViT encoder blocks are fused with state-of-the-art CNN models. The underlying patterns that characterize one medical condition from another are intricate, so CNN models facilitate complex feature extraction. Subsequently, the extracted features are fed to ReMixViT encoder blocks as patches that operate in a semantic token space, judiciously attending to different patches based on context. In this study, I propose two hybrid models comprising convolutional neural networks and ReMixViT encoder blocks for medical image classification by blending the goodness of both models. The first model (referred to as Res-ReMixViT) is a simple cascaded architecture of CNN and ReMixViT encoder block, whereas the concept of auxiliary classifier (Szegedy

et al., 2016) has been utilized in the second architecture (referred to as Res-ReMixViT+). Here, two ReMixViT encoder blocks as auxiliary classifiers have been attached to the former hybrid model not only to improve its convergence but also to enhance its performance by learning via encoder blocks of different scales.

I validate my hypotheses on six publicly available medical imaging datasets. The diversity of the dataset lies in the modality, type of diseases, number of categories, and class imbalance. Hence, the generalization ability and robustness of the proposed method are established through this study.

The contributions of this chapter are outlined as follows:

- I present ReMixViT, a tailored adaptation of the ViT architecture specifically designed to effectively capture the intricate patterns in medical imaging that indicate underlying medical conditions.
- I introduce two hybrid architectures, Res-ReMixViT and Res-ReMixViT+, which integrate ResNet50 with the proposed ReMixViT encoder blocks in two distinct configurations. These hybrids are designed to harness the strengths of both architectures, leading to enhanced performance.
- To validate the efficacy of ReMixViT and the proposed hybrid architectures, I perform experiments on six medical imaging datasets, each encompassing diverse modalities and medical conditions. This extensive evaluation underscores the potential of the proposed models in medical image classification tasks.
- Beyond performance evaluation, I offer visual interpretations of my proposed models, including attention maps and class activation maps. These visual aids enhance the understanding and interpretability of the models' decision-making processes.
- I further demonstrate the generalization capability of the proposed modifications by successfully adapting them to various vision transformer architectures, achieving significant performance improvements across different models.

3.1.2 Background

The adaptation of the transformer architecture in image classification tasks has paved the way toward the domain of medical imaging. Though the number of parameters in ViT is high compared to CNNs, due to its interpretability of the model's prediction, ViT has been widely accepted in medical image classification and segmentation tasks. In the following two paragraphs, I review the articles using ViT architecture with or without any modifications and the hybrid architectures (combining CNN and ViT) for medical image classification tasks.

The pure ViT architecture has been employed in various medical image classification tasks — COVID-19 infection detection from Chest radiograph images (Mondal et al., 2021; Park and Ye, 2022), fundus image classification (Yu et al., 2021), skin lesion classification (Xin et al., 2022), Histological subtype detection, tumor detection (Gheflati and Rivaz, 2022), brain stroke assessment (Ayoub et al., 2023). Few studies have reported the assimilation of multiple pure ViT architectures at different scales to incorporate both global and local attention to the clinical images simultaneously. For instance, multi-scale ViT-based architecture for characterizing whole-slide histopathological images (Chen et al., 2022c) and for estimating brain age from brain MRI (He et al., 2021).

Another line of study uses both the global attention of CNN and local connectivity drawn by ViT models. Several such task-specific networks are reported in the following studies. A transformer encoder-decoder based lesion-aware network is proposed to grade the level of diabetic retinopathy from the fundus images in (Sun et al., 2021). He et. al. proposed a hybrid architecture to compute global-to-local attention for estimating brain age from brain MRI (He et al., 2021). GasHis-Transformer, another hybrid model, developed by Chen et al.2022 (Chen et al., 2022b) to diagnose gastric cancer from histopathological images. In another study (Zhao et al., 2022), a ViT-based model has been proposed to classify the cervical cells in the scenario of data scarcity and class imbalance. The integration of the transformer encoder block with CNN architecture is also reported for mammogram classification task (Xia et al., 2023). MedViT proposed by Manzari et al. (Manzari et al., 2023) is claimed to be a robust hybrid architecture that efficiently integrates the local and global connectivity inferred by CNNs and ViTs, respectively. Similarly, another hybrid architecture (Liang et al., 2022) integrates local and global features from CNN and ViT to distinguish COVID-19 CT.

Similarly, all these ViT-based architectures as well as hybrid architectures, excluding Med-ViT (Manzari et al., 2023) are designed for a particular task at hand. The proposed hybrid architectures are not limited to any particular type of task but rather can be generalized to various medical image classification tasks.

3.2 Methods

3.2.1 Vision Transformer

The transformer, introduced as a sequence transduction model, is composed of encoders and decoders. Afterward, it has been adapted in the domain of image classification (Dosovitskiy et al., 2020; Valanarasu et al., 2021), image segmentation. ViT (Dosovitskiy et al., 2020), one of the successful adaptations of the transformer in the image classification task, considers only the encoder blocks of the transformer. As my task is to classify medical images according to the underlying medical condition, I consider ViT to be the base architecture. Before feeding the images into the encoder block, images are decomposed into patches, and patches are converted to embedding vectors by applying the linear transformation. In the next step, the class token is appended to the embedding vector, and a positional encoding is added afterward to preserve the positional information of the patches. Both the class token and the positional encoding are learned while training the model. The encoder block comprises two residual connections, as illustrated in Fig. 3.1(a). The first one surrounds the layer normalization and the Multihead Self-Attention (MSA) block, referred to as the MSA residual block (Fig. 3.2(e)). The second residual connection surrounds the layer normalization and the Multilayer Perceptron (MLP) block, known as the MLP residual block (Fig. 3.2(f)). Layer normalization ensures that the mean and standard deviation of the previous layer output remain approximately 0 and 1, respectively. The key feature that makes the ViT model unique is the use of MSA abandoning convolutions. The attention mechanism computes a weighted sum of the value vectors (V), where the weights are determined by the similarity between a query vector (Q) and a set of key vectors (K). Specifically, the attention score is calculated using the scaled dot-product attention formulation, as shown in equation 3.1. Here, $Q, K \in \mathbb{R}^{n \times d_k}$ and $V \in \mathbb{R}^{n \times d_v}$, where n is the number of tokens (or patches). The softmax operation is applied across the rows of the matrix product QK^T, ensuring that for

each query, the resulting attention weights over all keys sum to 1.

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_K}}\right)V \tag{3.1}$$

Since the vector query, key, and value are derived from the input itself, the name self-attention is justified. In the case of Multi-Head Self-Attention (MSA), instead of computing a single attention score, the input $X \in \mathbb{R}^{n \times d_{\text{model}}}$ (n is the number of token and d_{model} is the embedding dimension of each token) is linearly projected into multiple sets of queries, keys, and values — one for each attention head using learned projection matrices. Attention is then computed independently in each subspace, and the outputs are concatenated and linearly transformed to form the final result. Thus, MSA establishes relations among not only different input vectors representing patches but also different segments within an input patch. The attention score obtained from h different heads are combined by concatenating and linear projection as given by equation 3.2.

$$\text{MSA} = \text{Concat}(\text{SA}_1, \text{SA}_2, \ldots, \text{SA}_h) \cdot W, \quad \text{where } W \in \mathbb{R}^{(h \cdot d_v) \times d_{\text{model}}} \tag{3.2}$$

where $\text{SA}_i = \textit{Attention}(XW_i^q, XW_i^k, XW_i^V)$ is the attention score obtained by i^{th} head and Concat$(\cdot)$ represents column-wise concatenation operation. All the projection matrices $W \in \mathbb{R}^{h \cdot d_v \times d_{\text{model}}}$, $W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_k}$ are learned during the training process. In the Transformer architecture, d_{model} represents the embedding dimension of the model, and for multi-head attention with h heads, each head typically uses $d_k = d_v = \frac{d_{\text{model}}}{h}$. The MLP block consists of fully connected layers, GELU activation layers, and drop-out layers. The MLP block is designed with fully connected layers accompanied by the GELU activation function so that the dimensions of input and output to the MLP blocks remain the same. The patch embedded bock is followed by sequentially connecting several encoder blocks, resulting in ViT architecture. In the context of medical image classification, there is a constraint on the number of trainable parameters of the network due to the scarcity of labeled training data. The appropriate number of encoder blocks in the ViT model is decided by empirical evidence, which is 12 (for a detailed description of the experiment, refer to Section 3.3.3). The total number of trainable parameters of ViT architecture comprising 12 encoder blocks with a 512 number of nodes in the MLP layers is 19.5 M.

3.2.2 MLP-Mixer

MLP-Mixer architecture is proposed by Tolstikhin et al. (Tolstikhin et al., 2021), which does not use either convolution or self-attention for image classification tasks. MLP-Mixer block consists of two sequential sets of MLP blocks with residual connections. However, these two MLP blocks differ in how they are applied to the input matrix. Each of the MLP blocks comprises two fully connected layers, with GELU as an activation function in between and a residual connection. The input matrix is of size $n \times c$ where n and c represent the number of patches and the number of channels, respectively. Before the application of the first MLP block, the input matrix is transposed, and each channel (comprising information from all the patches) is fed to the two fully connected layers sequentially. Moreover, the output from the last fully connected layers of this MLP block is again transposed to maintain its original dimension. Similarly, the second MLP block is applied to the output of the first MLP block. However, the input to the fully connected layers are patches containing information from each of the channels. Thus, the first MLP blocks establish relations among the different patches, hence referred to as token-mixing or patch-mixing. On the other hand, the second MLP block allows commutation between different channels and is referred to as channel-mixing. This attribute of two-way interactions makes MLP-Mixer superior to simple MLP operations in the context of modeling complex attributes.

3.2.3 Proposed Modification over ViT

Characterizing biomedical images for disease identifications involves a complex feature extraction process. Additionally, medical image classification tasks usually have a limited number of training samples, making the classification difficult using a deep learning model with many parameters. While training a ViT model for predicting underlying medical conditions from medical images, I observed that the low test performance is due to low training performance (Figure 3.4). This suggests the inefficiency of the ViT model in modeling complex and subtle distinguishing attributes of medical conditions. Thus, there is a scope to enhance the training as well as test performance by learning those attributes. The encoder block of ViT comprises two residual blocks joined sequentially. The first residual block is made up of a normalization operation followed by MSA, which computes the attention map from the input

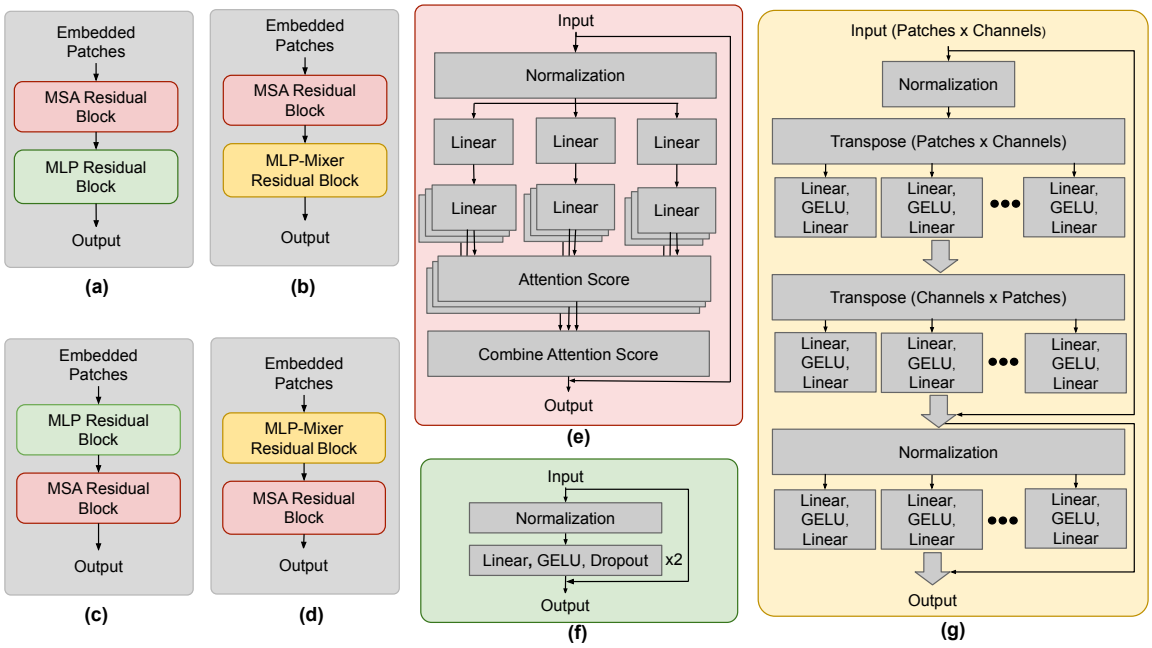

Figure 3.1: The architecture of ViT, MixViT, ReViT, and ReMixViT encoder block is shown in (a), (b), (c), and (d) panels of the figure. (e), (f) and (g) respectively depict the architecture of the MSA residual block, MLP residual block, and MLP-Mixer residual block.

feature vector. The second residual block comprises normalization operation and a multilayer perception that models the underlying complex patterns. Thus, for each encoder block, the process of complex feature mapping is performed after the execution of the MSA operation. Instead, I propose a modification of reversing the sequence of residual blocks. I first perform the operation of complex feature extracting followed by the MSA estimation. The motivation behind such modification is that the application of MSA on the extracted complex feature by MLP residual block instead of patches extracted from original images would lead to superior modeling ability of complex patterns. Secondly, I replace the MLP residual block of the encoder with the MLP-Mixer residual block (shown in Fig. 3.1(g)) as both patch-wise and channel-wise communication might lead to extracting additional information regarding medical condition detection tasks. I refer to this modified ViT architecture as ReMixViT, where R stands for reverse and MM stands for MLP-Mixer. Other than ViT and ReMixViT architectures, I have considered two more architectures, ReViT and MixViT, for the ablation study. ReViT is the modified ViT architecture where the only change is the reverse order occurrence of MLP residual block and MSA residual block. In MixViT architecture, the only

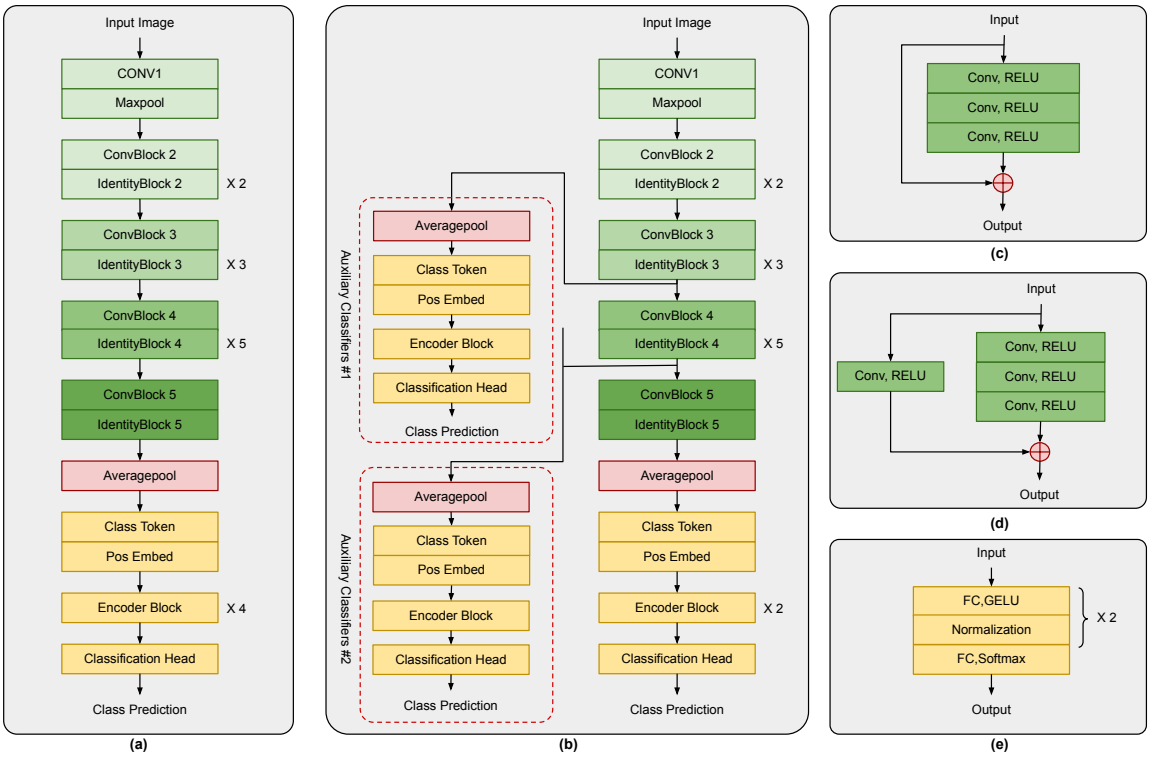

Figure 3.2: Architecture of proposed hybrid models — (a) ResNet-ViT/Resnet-ReMixViT and (b) Res-ReMixViT+. The encoder blocks in both architectures represent the corresponding encoder block of the ViT/ReMixViT model (presented in Fig. 3.1(d)). For example, in the case of ResNet-ViT, the encoder blocks represent ViT encoder blocks. Similarly, in ReMixViT, the encoder block is considered in the case of Res-ReMixViT+. The block diagram of ConvBlock, IndentityBlock, and Classification Head is shown in (c), (d), and (g) panels of the figure, respectively.

difference is the replacement of MLP with the MLP-Mixer layer. The parameters of ReViT and MixViT are similar to ViT and ReMixViT, respectively. For reference, the encoder blocks of ViT, MixViT, ReViT, and ReMixViT are shown in (a), (b), (c), and (d) panels of Figure 3.1 respectively. With a similar number of encoder blocks as in ViT architecture, the number of trainable parameters of ReViT is 19.5 M, while the same for MixViT and ReMixViT is 20.4 M, which is comparable with the considered ViT model for this study.

3.2.4 Architectures of Proposed Hybrid Models

In the context of medical image analysis, I also investigate the potential of two deep neural network architectures that result from two types of hybridization of traditional CNN and

proposed ReMixViT. The purpose of this fusion of these two networks is so that complex features extracted by the CNN architecture are fed to the ReMixViT encoder block to estimate MSA. ResNet50 is considered the backbone of the integrated models in famous CNN architecture. ResNet50 is essentially comprised of cascaded convolutional and identity blocks in a particular fashion. The first architecture, referred to as Res-ReMixViT, is developed by replacing the final identity layer of the ResNet50 architecture series of transformer encoder blocks as shown in Fig. 3.2. Thus, instead of passing the representative vectors projected from image patches, the complex features extracted by the part of the ResNet50 model are directed to the encoder block of ReMixViT through the process of appending class tokens and adding position encoding. Finally, a block of alternative dense layers and batch normalization layers leads to the ultimate class probabilities. I consider a similar hybrid architecture by replacing the ReMixViT encoder block with the ViT encoder block for comparison purposes.

The second architecture is proposed by introducing deep supervision into Res-ReMixViT through the addition of auxiliary classifiers. Specifically, two auxiliary classifiers, each comprising ReMixViT encoder blocks, are attached to intermediate layers of Res-ReMixViT, as illustrated in Fig. 3.2. This deeply supervised design aligns with the concept introduced by Lee et al. (Lee et al., 2015), where intermediate layers receive direct supervision to facilitate improved feature learning and gradient flow. I refer to this enhanced architecture as Res-ReMixViT+. The encoders within these auxiliary classifiers, along with the primary classification head, receive features from different levels with varying patch sizes. Consequently, attention scores are computed from multi-scale feature representations.

The ability of the hybrid models to identify different medical conditions is assessed by comparing their performance with the ResNet50 model. Therefore, while designing the hybrid models, it is ensured that the numbers of parameters of the models to be compared are in a similar range (lies between 25 M and 28 M). As a result, ReMixViT includes four Res-ReMixViT encoder blocks with multi-head self-attention layers with eight heads. On the other hand, Res-ReMixViT+ incorporates two Res-ReMixViT encoder blocks at the final stage and one Res-ReMixViT encoder block within each of the two auxiliary classifiers.

3.2.5 Attention Map

The attention map represents the distribution of the attention weights over the input image as experienced by the trained model. From a trained encoder block, I can extract the attention score for each of the heads of an MSA layer as computed by equation 3.1. This attention score is then averaged over all the heads. For a particular encoder block, the attention weights, i.e., the attention score as computed by equation 3.1 experienced by each head, are averaged, and an identity matrix is added to the mean attention score to consider the effect of residual connection. The resulting attention weight is then recursively multiplied over the number of encoder blocks to get the final attention map.

3.3 Experiment and Analysis

3.3.1 Dataset Description

This study evaluates the proposed ReMixViT and hybrid architectures using six medical imaging datasets to assess their effectiveness across diverse clinical scenarios. Each dataset is from a different modality and aims to address other separate medical conditions. For example, mammography images detect mass accumulation or calcification in breast tissue; various skin lesions are studied using dermoscopic images, etc. The diversity also lies in the number of classes, i.e., the medical conditions and the imbalance ratio (ratio of the number of samples of the largest class to that of the smallest class). The selection of diverse datasets makes the experiment more robust and generalized. The detailed description of each of the datasets used in this study is presented in Table 3.1. Unless mentioned explicitly, the available samples are divided into 80:20 ratios for training and testing purposes, respectively. The images from all the datasets are resized into the dimension of $224 \times 224 \times 3$.

3.3.2 Experimental Setting

All the codes are written in Python language and are executed using Keras with TensorFlow as the back-end on a computer with an Intel Core i5 processor running at 2.40 GHz using 16 GB of RAM and NVIDIA GeForce RTX 2060 GPU with 6 GB RAM. The code is available at this GitHub repository https://github.com/SusmitaSenGhosh/ReMixViT-Enhancing-Clinical-Image-

Table 3.1: Detailed description of the six medical imaging datasets.

Dataset	Purpose	Modality	#Classes	#Samples	IR
Colorectal Histology (Kather et al., 2016)	Tissue type detection	Histology images	8	5000	01.00
CBIS_DDSM (Sawyer-Lee et al., 2016)	Condition of breast cancer detection	Mammography	5	10,728	32.76
Chestxray (Kermany et al., 2018)	Pneumonia detection	X-ray	2	Train: 4232 Test: 624	02.88
Fundus (DeepDRiD, 2020)	Diabetic retinopathy gradation	Fundus retinal images	5	Train: 1200 Test: 400	07.70
ISIC18 (Codella et al., 2019)	Skin lesion type detection	Dermoscopic images	7	10,015	58.86
PBC (Acevedo et al., 2020)	Blood cell classification	Peripheral blood cells images	8	17092	02.74

#Class: Number of classes, #Samples: Number of samples, IR: Imbalance ratio

Analysis-with-Hybrid-Vision-Transformers-and-Adaptive-Feature-Mix.

I followed the same protocol and hyperparameters for training both models. The mean square error loss function is optimized with Adam optimization algorithms for learning model parameters. The models are trained up to 100 epochs with a batch size of 16. I used an exponential decaying learning rate with the number of epochs with an initial learning rate of 0.0002 and a decaying factor of 0.1. The medical imaging datasets considered in this study have diversity in class imbalance ratios ranging from 1 to 58.86. Thus, to overcome the effect of the existing class imbalance over the classification performance, I incorporate class weights in the loss function. The weight corresponding to i^{th} class is computed by $N/(cN_i)$, where N, N_i and c represent the total number of samples irrespective of classes, number of samples that belong to i^{th} class and the number of classes respectively. The final loss that is backpropagated is the class-wise weighted sum of loss.

I have considered six medical imaging datasets, with the number of classes in each dataset ranging from two to eight and the ratio of class imbalance extending up to 58.86. In this scenario, accuracy is not adequate to quantify the behavior of the model due to its high biasedness toward the majority class. Recall, precision, and F1-score, three widely used metrics in medical applications represent the behavior of the classification model more judiciously as compared to accuracy (Hicks et al., 2022). Recall measures the proportion of samples belonging to a particular class identified correctly, and precision denotes the proportion of samples tagged to a particular class that belongs to that class. F1-score is the harmonic mean between these quantities of inverse relation. As the number of classes extends up to eight, I consider the average class-specific recall, precision, and F1-score (ACSR, ACSP, and

ACSF) as performance metrics. These metrics are not influenced by the performance of the majority class but rather impart equal weight to each of the classes. Let, $\mathbf{C}$ be the confusion matrix for the classification task where element c_{ij} indicates the number of samples of i^{th} class predicted as of j^{th} class. Three metrics, ACSR, ACSP, and ACSF, can be defined in terms of the elements of the confusion matrix as follows.

$$\text{ACSR} = \frac{1}{n}\sum_i \text{Recall}_i = \frac{1}{n}\sum_i \frac{C_{ii}}{\sum_j C_{ij}}, \tag{3.3}$$

$$\text{ACSP} = \frac{1}{n}\sum_i \text{Precision}_i = \frac{1}{n}\sum_i \frac{C_{ii}}{\sum_j C_{ji}}, \tag{3.4}$$

$$\text{ACSF} = \frac{1}{n}\sum_i \text{F1}_i = \frac{1}{n}\sum_i \frac{2 \cdot \text{Recall}_i \cdot \text{Precision}_i}{\text{Recall}_i + \text{Precision}_i}. \tag{3.5}$$

Moreover, I also considered another metric, the mean absolute deviation of F1-scores (MADF), defined by the following equation.

$$\text{MADF} = \frac{1}{n}\sum_i |\, \text{F1}_i - \frac{1}{n}\sum_j \text{F1}_j \,|. \tag{3.6}$$

MADF metric is not capable of quantifying the performance solely. However, it is helpful for the instances where ACSF yields similar results. It measures how class-specific F1-scores are spread out, thus quantifying the degree of handling class imbalance. For a fixed ACSF value, the lower value of MADF indicates better performance.

Along with these metrics, I have also reported the area under the receiver operating characteristics curve, which is independent of the threshold applied to the probability score produced by the models.

3.3.3 Performance of proposed ViT

I start with an experiment for determining the optimum number of encoder blocks of ViT architecture suitable for the context of medical image classification. ViT models with 4, 8,12, and 16 encoder blocks are trained and evaluated on six considered datasets (Figure 3.3). According to the performance metrics (ACSR, ACSP, ACSF, and AUC), the optimum

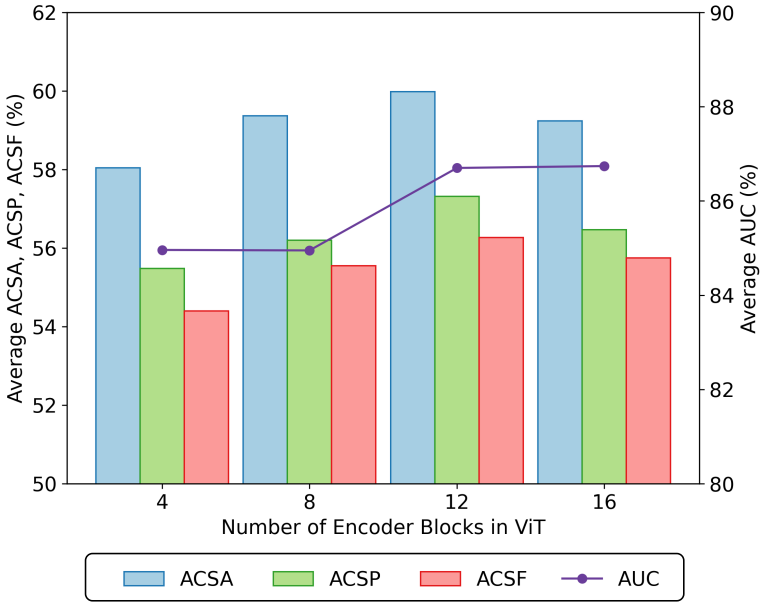

Figure 3.3: Evolution of performance metrics (averaged over six datasets) of ViT architecture with the increasing number of encoder blocks.

number of encoder blocks for ViT architecture is 12 in the experimental setup, as described in the previous section. Hence, throughout this study, I have employed 12 encoder blocks for all four ViT, ReViT, MixViT, and ReMixViT models for impartial comparison.

In this section, I evaluate the performance of the proposed ReMixViT model on six different medical imaging datasets mentioned in section 5.4. I compare the performance of the proposed ReMixViT model with the conventional ViT model (baseline method). The detailed description of the architecture with parameters used in this study is discussed in section 3.2.1 and 3.2.3, respectively.

Fig. 3.4 shows ASCF curves of training and test data with respect to epoch number for both ViT and ReMixViT models. It is explicitly established from Fig. 3.4 that the training performance of ReMixViT surpasses the performance of the ViT model for all the six datasets considered in this study. This indicates that the proposed model can characterize complex patterns more efficiently than ViT. Moreover, the training ACSF curves of ReMixViT saturate with a lesser number of epochs as compared to that of ViT. Thus, I can achieve better performance with fewer epochs using the ReMixViT model. It is also noticed that the training performances and the ACSF graph of the ReMixViT model for test data are elevated compared to the ViT model. This phenomenon also eliminates the possibility of any

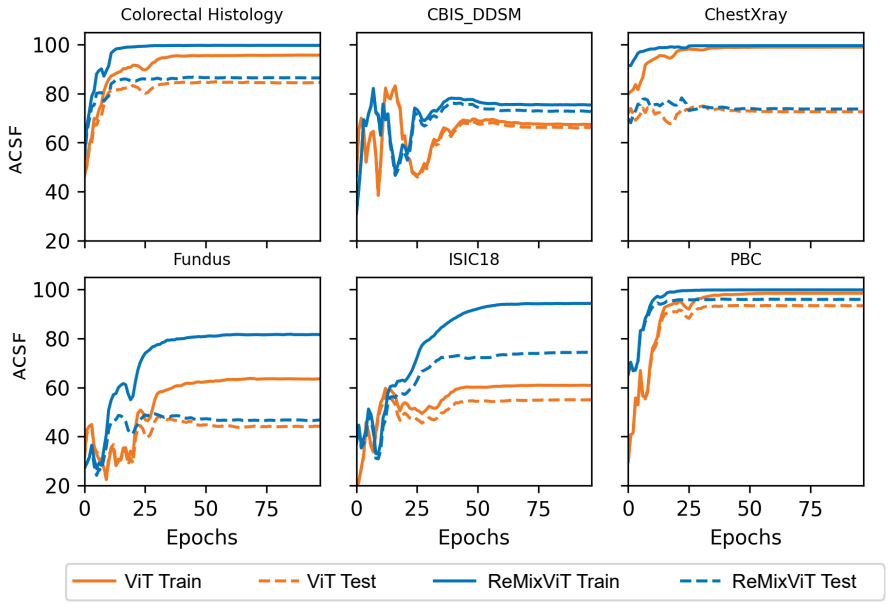

Figure 3.4: Comparison of ASCF performance curve of ViT and ReMixViT model concerning epochs for six medical imaging datasets.

over-fitting. Altogether, the superiority of ReMixViT over conventional ViT is established through this experiment. The performance of these two models at 100^{th} epoch is presented in Table 3.2. For each dataset, I observe the increase in classification performance with the proposed model in terms of ACSR, ACSP, ACSF, and MADF metrics. Using the proposed ReMixViT model, I observed an average increase of 3.72%, 5.01%, and 4.62% in ACSR, ACSP, and ACSF performance, respectively, over ViT with a comparable number of model parameters and floating-point operations (FLOps). Additionally, the proposed modification over ViT has reduced the value MADF metric with improved ACSF for all the datasets, implying the superiority in balancing the disproportion in class-wise samples. Moreover, the performance of the ViT and ReMixViT models are compared by ROC curves and AUROC value (presented in Figure 3.5), which is independent of the threshold applied to the class probabilities yielded by the models, unlike ACSR, ACSP, and ASCF. For the datasets with a number of classes greater than two, I have plotted the macro-average of the one-versus-rest ROC curve. I observe improvements in AUROC values of ReMixViT over the ViT model to be in the range of 0.04 to 0.12 for the six datasets, justifying the proposed modification.

While developing a diagnostic solution, it is important to validate the behavior of the

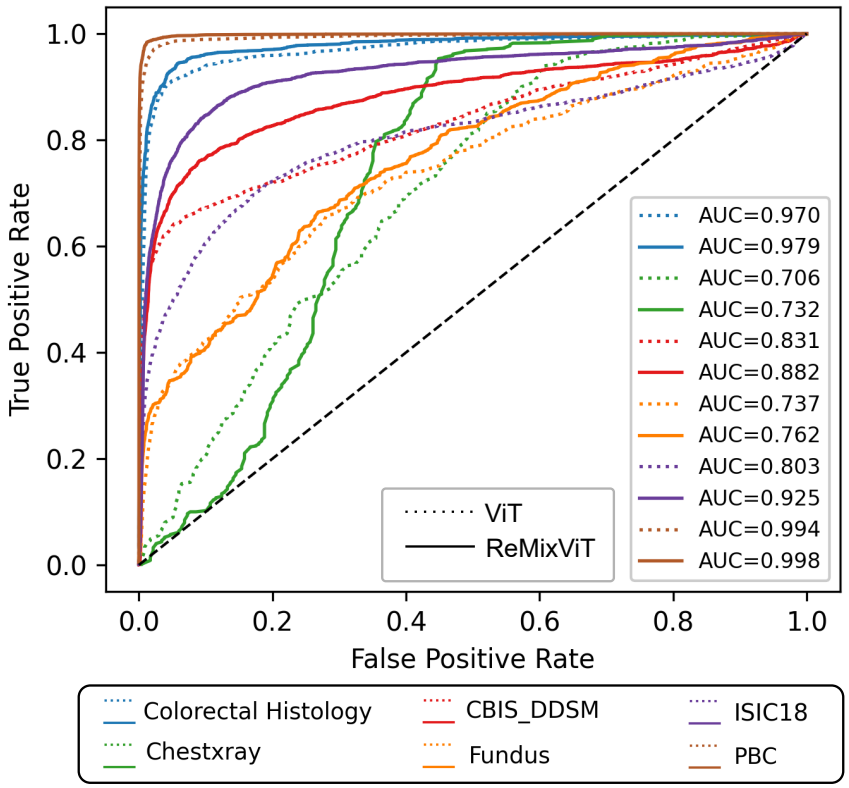

Figure 3.5: Comparison of ROC curve of ViT and ReMixViT model concerning epochs for six medical imaging datasets.

Table 3.2: Comparison of Classification Performance of ViT and ReMixViT (Proposed) Model for Six Medical Imaging Datasets.

Datasets	ViT				ReMixViT			
	ACSR	ACSP	ACSF	MADF	ACSR	ACSP	ACSF	MADF
Colorectal Histology	84.28 ±1.06	84.68 ±1.05	84.39 ±1.06	06.34 ±0.47	**86.32*** ±1.14	**86.42*** ±1.12	**86.34*** ±1.13	**05.96*** ±0.77
CBIS_ DDSM	36.65 ±1.86	27.89 ±2.15	28.30 ±2.01	22.03 ±1.54	**42.5*** ±1.54	**30.46*** ±1.93	**31.45*** ±1.87	**20.84** ±1.23
Chestxray	64.83 ±0.78	76.93 ±2.29	64.71 ±0.89	16.68 ±0.59	**65.38** ±0.58	**82.26*** ±2.13	**65.11** ±0.72	**16.34** ±0.62
Fundus	32.37 ±1.24	32.14 ±1.18	31.00 ±1.06	18.35 ±0.79	**34.95*** ±1.38	**34.87*** ±1.28	**34.76*** ±1.11	**16.96*** ±0.81
ISIC18	50.72 ±2.98	36.14 ±1.87	38.83 ±2.56	12.95 ±0.68	**58.71*** ±3.42	**51.05*** ±2.87	**54.24*** ±3.15	**11.70*** ±0.87
PBC	92.60 ±1.26	92.93 ±1.34	92.75 ±1.30	04.35 ±1.22	**95.90*** ±1.13	**95.68*** ±1.07	**95.78*** ±1.11	**02.60*** ±0.93

* marks the statistically significant improvement.

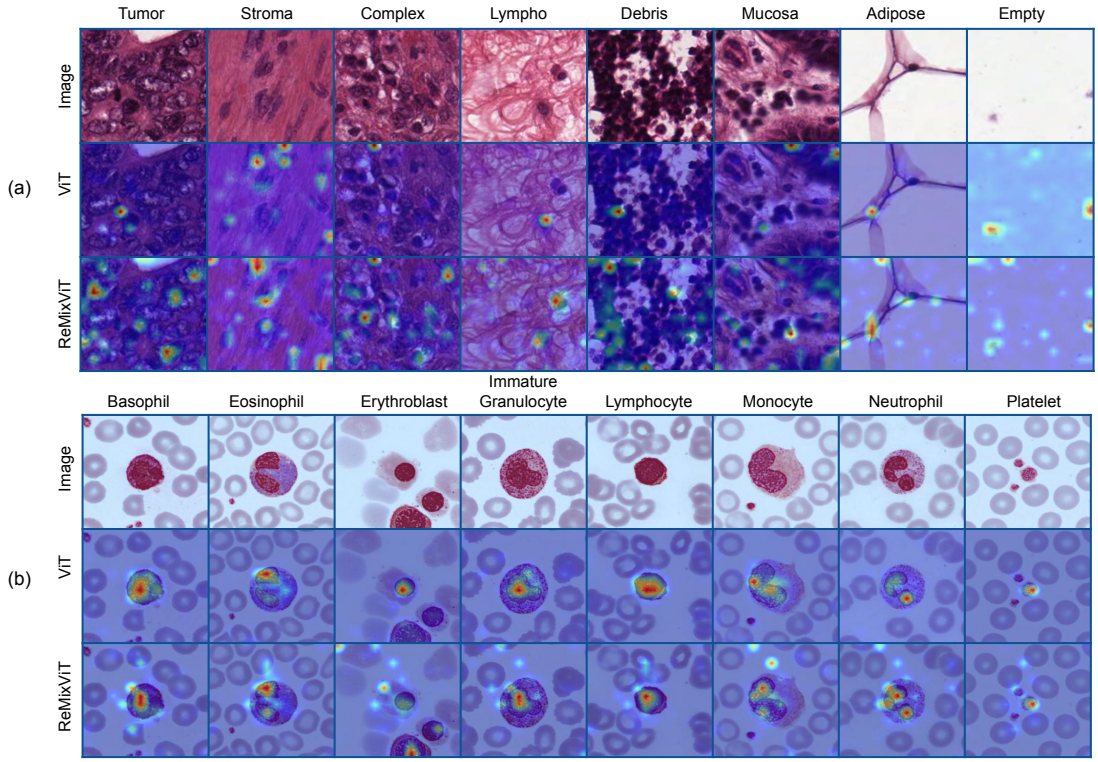

Figure 3.6: Attention maps for Colorectal Histology and PBC dataset are presented in panels (a) and (b), respectively. A single test sample from each class of the datasets is selected, and the corresponding attention maps for both the ViT and ReMixViT models are displayed in this figure.

developed model by inspecting the contributing features. Compared with ViT, the ReMixViT model attains better performances in ACSR, ACSP and ACSF for all six medical imaging datasets. In this section, I analyze attention maps generated from both models to identify the attributes of the ReMixViT model that make it superior to ViT in this study. Fig. 3.6 compares the attention maps of both models. I consider one sample from each class of Colorectal Histology and PBC datasets. The eight different types of tissues are distinguished by their texture. For example, *Tumor* and *Stroma* tissue have different textures as the latter is more loosely structured and arranged in a specific direction Eroschenko (2005), *Adipose* tissues are characterized by their band-like structure, the non-centric nuclei. I observe that the regions highlighted in the attention map of the ReMixViT model include the regions that exhibit such patterns. Histologically, the blood cell types can be characterized by their size, stain of the granules, the morphology of their nucleus, presence of nuclei, number of

Table 3.3: The performances of ViT, ReViT, MixViT, and ReMixViT for six datasets are presented. The relative improvement concerning the performance of the ViT model is indicated in the last row.

Datasets	ViT (FLOPs: 8.6 G, Params.: 19.5 M)			ReViT (FLOPs: 8.6 G, Params.: 19.5 M)			MixViT (FLOPs: 9.57 G, Params.: 20.4 M)			ReMixViT (FLOPs: 9.57 G, Params.: 20.4 M)		
	ACSR	**ACSP**	**ACSF**	**ACSR**	**ACSP**	**ACSF**	**ACSR**	**ACSP**	**ACSF**	**ACSR**	**ACSP**	**ACSF**
Colorectal Histology	84.28	84.68	84.39	89.78	89.66	89.68	82.22	82.15	82.11	86.32	86.42	86.34
CBIS_ DDSM	36.65	27.89	28.30	36.12	27.48	27.80	41.90	30.44	31.67	42.50	30.46	31.45
Chestxray	64.83	76.93	64.71	69.74	81.87	70.70	68.21	82.23	68.81	65.38	82.26	65.11
Fundus	32.37	32.14	31.00	35.60	36.23	33.48	35.70	36.34	35.93	34.95	34.87	34.76
ISIC18	50.72	38.83	36.14	58.70	40.49	43.95	48.70	38.99	42.27	58.71	51.05	54.24
PBC	92.60	92.93	92.75	96.29	95.90	96.08	95.82	95.82	95.81	95.90	95.68	95.78
Average	60.24	58.90	56.22	64.37	61.94	60.28	62.09	61.00	59.43	63.96	63.46	61.28
Improvement over ViT				+4.13	+3.04	+4.07	+1.85	+2.10	+3.22	+3.72	+4.56	+5.07

lobes in nuclei, nucleus to cytoplasm ratio, etc Prinyakupt and Pluempitiwiriyawej (2015). Though both models capture these attributes efficiently, ReMixViT localizes multi-lobe nuclei more accurately as compared to ViT. Thus, class prediction performance and attention map altogether suggest the usefulness of the proposed ReMixViT module.

3.3.4 Ablation Study

I propose the ReMixViT architecture over the conventional ViT architecture by modifying it in two ways. Firstly, I reverse the sequence of the MLP residual block and MSA residual block. Secondly, I incorporate the MLP-Mixer layer by replacing the MLP layers of the MLP residual block. Thus, I consider two more architectures, ReViT and MixViT, to study the effect of each modification separately and conduct a complete ablation study. The architectures of the encoder block of ReViT and MixViT are shown in Fig. 3.2. ReViT is the modified ViT architecture where the only change is the reverse order of occurrence of MLP residual block and MSA residual block. In MixViT architecture, the only difference is the replacement of MLP with the MLP-Mixer layer. The parameters of ReViT and MixViT are similar to those of ViT and ReMixViT, respectively. The protocol and hyperparameters used for training ReViT and MixViT models are as mentioned in section 3.3.2. Table 3.3 presents the performance of ViT, ReViT, and ReMixViT models averaged over all six datasets. It also quantifies the relative improvement of each model with respect to ViT.

It is noted from table 3.3 that each of the modified ViT models with a single change (ReViT, MixViT) as well as ReMixViT shows improvements in classification performance over ViT. The result suggests the significance of each of the modifications that are made individually. Though the performance yielded by ReMixViT does not surpass the performances of ReViT and MixViT in every case individually, it showed the best performance in terms of ACSF when averaged over all the datasets considered here. Thus, the result justifies the considered changes made over conventional ViT to construct ReMixViT for modeling intricate details of clinical images.

3.3.5 Performance of Proposed Hybrid Model

Next, I present the performance of the proposed hybrid architecture by combining the ResNet50 with ReMixViT, referred to as Res-ReMixViT (Fig. 3.2(e)). Since ResNet50 is the backbone of the proposed architecture, along with Res-ReMixViT, the performance of ResNet50 is also reported for comparison purposes. To evaluate the effectiveness of the proposed ReMixViT encoder block compared to the standard ViT encoder block, even in a hybrid architecture, I also include the performance of a ResNet-ViT baseline. ResNet-ViT follows a similar architecture as Res-ReMixViT where ReMixViT encoder block is replaced by the ViT encoder block (Fig. 3.2(e)). The training protocol and all the hyperparameters used for training the models are the same as in the previous sections. All the models have been trained up to 50^{th} epochs for each of the six datasets. For the datasets, that do not explicitly mention the train and test samples (Colorectal Histology, CBIS_DDSM, ISIC18, and PBC), I follow a 5-fold cross-validation strategy, otherwise, I train each model 10 times with different random initialization of the network weights. In table 3.4, I report the mean and standard deviation of performance metrics obtained in 5 folds or 10 random repetitions. I perform unpaired t-tests to assay whether the improvements achieved by the proposed hybrid models are significant or not.

It is evident from the table 3.4 that integrated models (both ResNet-ViT and Res-ReMixViT models) attain better performance for all the datasets as compared to ResNet50, a popular CNN model. ResNet-ViT performs better than the ResNet50 model for all the datasets except the ISIC18 dataset. The proposed hybrid architecture Res-ReMixViT module surpasses both ResNet50 and ResNet-ViT models in terms of all considered metrics. The

Table 3.4: Comparison of classification performances (in terms of mean±standard deviation of ACSF MADF, and AUROC) of baseline models (ResNet50 and ResNet-ViT) and proposed hybrid models (Res-ReMixViT and Res-ReMixViT+) models for six medical imaging datasets.

	ResNet50			**Res-ViT**			**Res-ReMixViT**			**Res-ReMixViT+**		
Datasets	**ACSF**	**MADF**	**AUROC**	**ACSF**	**MADF**	**AUROC**	**ACSF**	**MADF**	**AUROC**	**ACSF**	**MADF**	**AUROC**
Colorectal Histology	94.89 ±0.33	2.97 ±0.33	99.44 ±0.12	95.06 ±0.30	3.15 ±0.32	99.51 ±0.02	95.50* ±0.24	2.79 ±0.22	99.61** ±0.05	95.57* ±0.22	2.84 ±0.24	99.62* ±0.03
CBIS_ DDSM	48.73 ±1.52	18.3 ±0.50	92.87 ±0.50	51.52 ±1.41	17.29 ±0.40	91.34 ±0.99	53.32* ±0.57	17.54 ±0.36	94.87* ±0.23	55.76** ±0.77	17.18 ±0.30	96.49** ±0.23
Chestxray	70.6 ±0.68	13.96 ±0.46	83.71 ±1.05	70.98 ±1.60	13.8 ±1.07	86.39 ±1.29	74.61* ±0.32	11.37* ±0.18	88.39* ±0.95	76.16** ±0.82	10.53** ±0.50	88.54* ±1.45
Fundus	47.11 ±0.21	16.51 ±0.16	79.34 ±0.04	50.57 ±0.55	16.11 ±0.35	81.22 ±0.39	51.42* ±0.74	14.34* ±0.48	81.87* ±0.26	54.09** ±0.58	14.62* ±0.39	81.92* ±0.12
ISIC18	70.25 ±0.29	9.93 ±0.71	93.21 ±0.33	70 ±0.90	10.26 ±0.56	91.44 ±0.47	71.29* ±0.49	9.23 ±0.57	93.76* ±0.16	74.77** ±0.42	09.34* ±0.20	94.81** ±0.20
PBC	98.91 ±0.02	0.83 ±0.09	99.93 ±0.00	98.81 ±0.15	0.8 ±0.11	99.91 ±0.01	99.01* ±0.01	0.66* ±0.11	99.94* ±0.01	99.06* ±0.03	0.62* ±0.06	99.95* ±0.00
Average	71.75	10.42	91.42	72.82	10.24	91.64	74.19	9.32	93.07	75.9	9.19	93.56
Comparison w.r.t to ResNet50				**+1.08**	**−0.18**	**+0.22**	**+2.44**	**−1.10**	**+1.66**	**+4.15**	**−1.23**	**+2.14**
Comparison w.r.t to Res-ViT							**+1.37**	**−0.91**	**+1.44**	**+3.08**	**−1.05**	**+1.92**
Comparison w.r.t to Res-ReMixViT										**+1.71**	**−0.13**	**+0.48**

AUROC values are in the range [0, 1] and are represented here in $\times 10^{-2}$ scale for readability.
Mean value of metrics greater than that for ResNet50 and ResNet-ViT models are marked by bold fonts.
***** indicates the statistically significant improvement in performance metric over both ResNet50 and ResNet-ViT.
****** indicates the statistically significant improvement in performance metric over ResNet50, ResNet-ViT and Res-ReMixViT.

lower value of MADF suggests that the existing class imbalance is handled more successfully by Res-ReMixViT than ResNet-ViT. The outcome of this experiment demonstrates the efficient integration of CNN with ReMixViT encoder blocks for modeling complex patterns in medial images that play a crucial role in detecting underlying medical conditions.

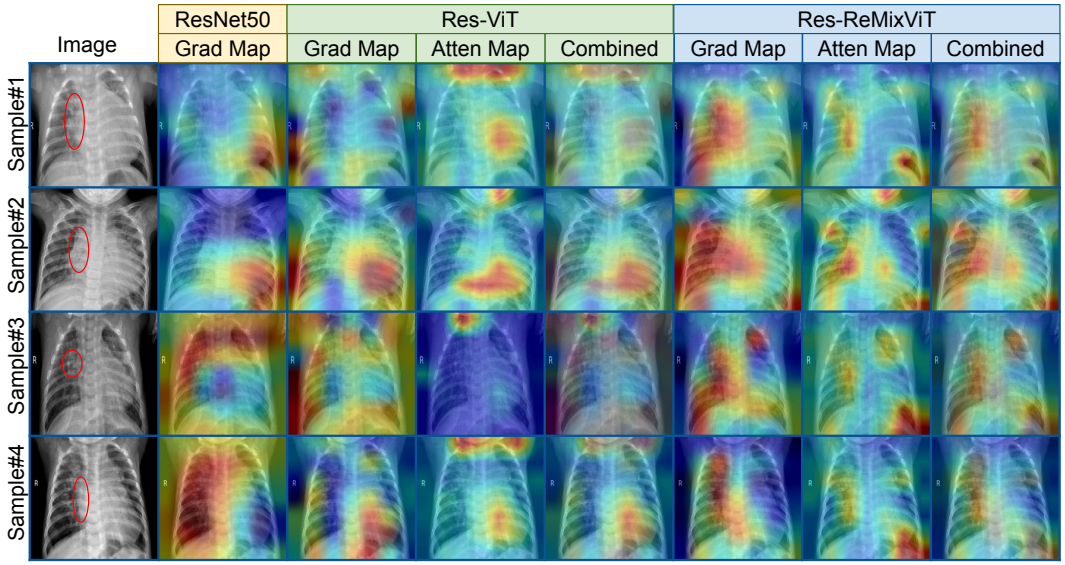

Figure 3.7: Gradient maps for ResNet50, Gradient map, attention map, and combined map for hybrid models (ResNet-ViT and Res-ReMixViT) for samples from the Chestxray dataset are presented. The red line marks consolidations in the right lungs.

Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al., 2017) is an approach to visualize the rough localization of salient regions in the input image that contribute to a model's class prediction. Grad-CAM works by computing the gradients of the class-specific score (i.e., the output corresponding to the predicted class) with respect to the feature maps of the final convolutional layer. These gradients are then globally averaged to obtain importance weights, which are used to compute a weighted combination of the feature maps, producing the localization map. In proposed case, the proposed hybrid networks are a fusion of CNN and ViT or ReMixViT encoder blocks in a cascaded manner. Therefore, I use Grad-CAM to visualize the gradient flow through the final convolutional layer, while attention maps from the encoder blocks are computed as described in Section 3.2.5. Fig. 3.7 shows the gradient maps, attention maps, and combined map for ResNet50, ResNet-ViT, and Res-ReMixViT model for the Chest x-ray dataset, which comprises two classes (Pneumonia

and Normal). I consider samples from the disease class (Pneumonia) and investigate how these three models behave in localizing the part of the lung affected by pneumonia. The consolidations in the right lung are visible in the images (leftmost panel), and the red line highlights the affected areas. It is noted that those consolidations have been localized by Res-ReMixViT more efficiently than the other two models.

The evaluation of the Res-ReMixViT+ model is done using a protocol similar to the one used in the ResNet-ViT and Res-ReMixViT models. Since Res-ReMixViT+ is comprised of two additional two auxiliary classifiers (Fig. 3.2(f)), the final loss is calculated by computing the weighted sum of the losses experienced by the auxiliary classifiers and the loss at the output layer with weights of 0.3, 0.3, and 0.4, respectively. Results reported in Table 3.4 show the performance of Res-ReMixViT+. It is observed that Res-ReMixViT+ outperformed all the other hybrid models used in this study. Overall, it has improved ACSF by 4.73%, 3.89%, and 2.31% concerning ResNet50, ResNet-ViT, and Res-ReMixViT, respectively.

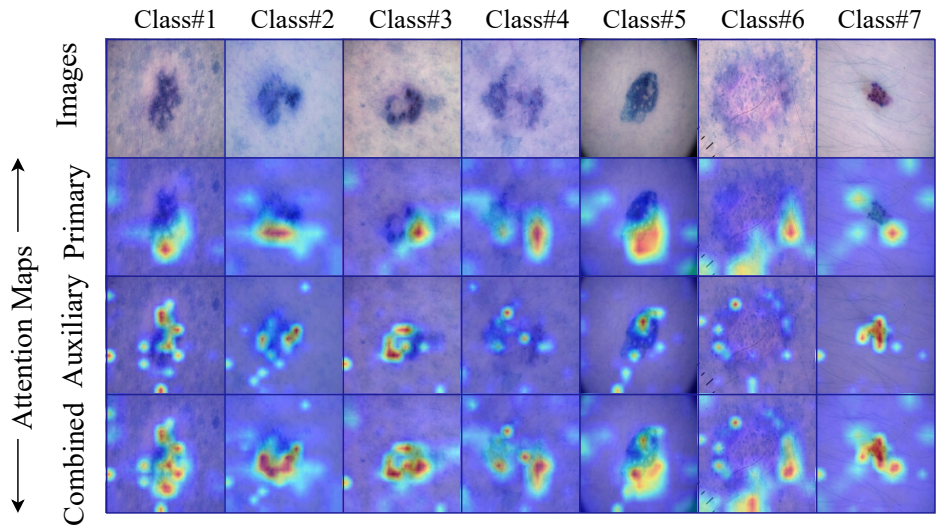

Figure 3.8: Individual attention maps of main encoder blocks and auxiliary encoder blocks, along with the combined attention map, are presented for one of the samples from each of the seven classes of the ISIC18 dataset.

Res-ReMixViT+ comprises four ReMixViT encoder blocks. Two encoder blocks of the central architecture play an important role in predicting and training. The other two encoder blocks that form auxiliary classifiers regulate the loss function in the learning process. I investigate the behavior of these two types of encoder blocks by analyzing the corresponding

attention maps. Fig. 3.8 shows the attention maps of encoder blocks from both the primary classifier and auxiliary classifier separately for one of the samples of each of the seven classes of the ISIC18 dataset. It is observed that the contributing regions highlighted in the attention maps of the encoders from both primary and auxiliary classifiers are mainly confined within the skin lesion area. However, both the highlighted regions play complementary roles in this task. Encoders from main classifiers predominantly enclose the skin lesion part, whereas encoders from auxiliary classifiers pay attention to more minor details.

3.3.6 Comparison with state-of-the-art methods

In this section, I present a comparison between the related state-of-the-art methods and proposed hybrid methods. Along with the CNN and ViT-based hybrid architectures, I have also considered pure CNN and ViT-based methods for comparison. Data augmentation has proven to be an effective process to enhance performance in the context of medical imaging due to the scarcity of available labeled data (Garcea et al., 2022). Like most state-of-the-art methods, I have also introduced variations in the training data by employing augmentation algorithms, namely horizontal and vertical (when applicable) flip, rotation, contrast adjustment, and zoom at random. Consequently, an enhancement in performance is observed. The performances yielded by my proposed method with the state-of-the-art methods are presented in Table 3.5. BreastMNIST and DermaMNIST, two datasets of the same modality and task as compared to CBIS_DDSM and ISIC18, respectively, are used to compare with the state-of-the-art models. Other than the DermaMNIST dataset, my proposed method has outperformed state-of-the-art methods. Additionally, I compare my approach with another study (Chen et al., 2022b) that designed a multi-scale transformer-based classification model for gastric histopathological images and reported an accuracy of 97.97% on HE-GHI-DS dataset (Li et al., 2018). The proposed method (Res-ReMixViT) shows an improvement of 1.03% in accuracy on the same dataset.

3.3.7 Generalization ability of the proposed modifications

I extended the study by investigating the effect of the proposed modifications on other transformer architectures specifically designed for image analysis. For this experiment, I considered

Table 3.5: Comparison with state-of-the-art results including CNN and ViT hybrid architectures.

Dataset	Method, Year	Model	Aug./ Ext. data	Accuracy (%)
Colorectal Histology	Kimianet-IV (Riasatian et al., 2021)	CNN	No	96.80
	CNN Ensemble (Nanni et al., 2021)	CNN	Yes	97.60
	CRCCN-Net (Kumar et al., 2016)	CNN	No	93.50
	CCT (Zeid et al., 2021)	Hybrid	Yes	95.00
	Res-ReMixViT (ours)	**Hybrid**	Yes	97.00
	Res-ReMixViT+ (ours)	**Hybrid**	yes	**98.00**
PBC	EnsDA_C (Nanni et al., 2022)	CNN	Yes	99.12
	BloodCaps (Long et al., 2021)	CNN	Yes	99.30
	MedViT-L (Manzari et al., 2023)	Hybrid	Yes	95.40
	CVM-Cervix (Liu et al., 2022)	Hybrid	Yes	98.48
	Res-ReMixViT (ours)	**Hybrid**	Yes	**99.39**
	Res-ReMixViT+ (ours)	Hybrid	Yes	99.18
Chest-xray	CNN (Hammoudi et al., 2021)	CNN	Yes	92.80
	MedViT-S (Manzari et al., 2023)	Hybrid	Yes	96.10
	Res-ReMixViT (ours)	**Hybrid**	Yes	**97.12**
Fundus	MedViT-S (Manzari et al., 2023)	Hybrid	Yes	56.10
	Res-ReMixViT (ours)	**Hybrid**	Yes	**60.25**
	Res-ReMixViT+ (ours)	**Hybrid**	Yes	**63.25**
Breast-MNIST	CNN (Moon et al., 2020)	CNN	No	88.46
	ViT (Gheflati and Rivaz, 2022)	ViT	Yes	82.00
	MedViT-S (Manzari et al., 2023)	Hybrid	Yes	89.70
	Res-ReMixViT (ours)	**Hybrid**	Yes	86.54
	Res-ReMixViT+ (ours)	**Hybrid**	Yes	**90.38**
Derma-MNIST	MedViT-S (Manzari et al., 2023)	Hybrid	Yes	**78.00**
	Res-ReMixViT (ours)	**Hybrid**	Yes	77.96
	Res-ReMixViT+ (ours)	**Hybrid**	Yes	77.96

the Swin Transformer (SwinT) (Liu et al., 2021), Pyramid Vision Transformer (PVT) (Wang et al., 2021a), and Class-Attention in Image Transformers (CaiT) (Touvron et al., 2021b). Table 3.6 compares the performance of SwinT, PVT, and CaiT with the proposed modifications on the PBC, ChestXray, ISIC18, and Fundus datasets. The proposed modifications, when applied to PVT, Swin Transformer, and CaiT, effectively enhanced classification performance across these architectures. This improvement can be attributed to the replacement of the traditional MLP layer with an MLP-Mixer and the reordering of residual blocks. The MLP-Mixer enhances the model's ability to capture global and local dependencies by enabling more complex interactions between tokens. The reordering of residual blocks also contributes to optimizing the information flow within the network. Consequently, these modifications lead to better feature representation and improved classification accuracy. Importantly, the fact that these enhancements were successfully applied to diverse transformer architectures

Table 3.6: The generalization ability of the proposed modifications on other transformer based architectures, Swin-T, CaiT, and PVT-Tiny.

Method (Params.)	PBC			Chestxray			Fundus		
	ACSR	**ACSP**	**ACSF**	**ACSR**	**ACSP**	**ACSF**	**ACSR**	**ACSP**	**ACSF**
Swin-T	90.48	90.23	90.27	97.91	97.30	97.58	85.03	84.81	84.65
(21.18 M)	± 3.34	± 2.37	± 2.87	± 0.50	± 0.78	± 0.37	± 4.12	± 3.77	± 3.88
Modified Swin-T	**94.69**	**94.76**	**94.65**	**99.02**	**99.17**	**99.08**	**99.55**	**99.78**	**99.66**
(22.29 M)	**± 0.94**	**± 1.18**	**± 1.06**	**± 0.76**	**± 0.51**	**± 0.47**	**± 0.32**	**± 0.12**	**± 0.21**
CaiT	95.44	95.82	95.53	97.59	96.93	97.20	95.62	95.19	95.37
(44.96 M)	± 0.57	± 0.88	± 0.74	± 0.82	± 2.11	± 1.22	± 0.48	± 0.64	± 0.48
Modified CaiT	**95.91**	**96.41**	**96.14**	**99.11**	**99.34**	**99.22**	**96.14**	**96.22**	**96.17**
(46.10 M)	**± 0.51**	**± 0.44**	**± 0.44**	**± 0.54**	**± 0.18**	**± 0.34**	**± 1.21**	**± 1.33**	**± 1.26**
PVT-Tiny	95.20	94.46	94.69	96.72	96.52	96.58	85.31	85.06	83.23
(13.05 M)	± 0.62	± 0.93	± 0.77	± 0.79	± 1.47	± 0.75	± 5.09	± 2.12	± 4.26
Modified PVT-Tiny	**96.07**	**95.97**	**95.98**	**99.39**	**98.75**	**99.05**	**98.42**	**97.04**	**97.53**
(13.39 M)	**± 0.49**	**± 0.80**	**± 0.65**	**± 0.30**	**± 0.99**	**± 0.64**	**± 2.16**	**± 3.38**	**± 3.15**

demonstrates the generalization ability of the proposed modifications, suggesting that they can be effectively integrated into other transformer encoder blocks as well. This further underscores the potential for improving the robustness and effectiveness of vision transformers in complex classification tasks.

3.4 Discussion

I present image-based diagnostic solutions in terms of a deep learning model that not only predicts the underlying medical condition but also gives a visual interpretation of its decision that can be further evaluated by the domain experts. I propose modifications to traditional ViT to attain greater classification performance by modeling complex patterns. In addition to that, I design two hybrid architectures by integrating ResNet50, a popular CNN-based model, and encoder blocks of modified ViT architecture (ReMixViT). The concept of auxiliary classifiers has also been combined with a hybrid model for more substantial convergence. In the context of the number of parameters, these integrated models are comparable with ResNet50 and, thus, suitable for medical image classification, as available samples are scarce. I demonstrate the superiority of the ReMixViT model over the ViT model and hybrid models over the ResNet50 and other hybrid models by validating the result with six medical imaging datasets of diverse nature. Moreover, performance enhancement is also established by adding the ReMixViT encoder-based auxiliary classifiers to the proposed hybrid architecture. My approach has also surpassed the relevant state-of-the-art methods. Furthermore, the portion

of the image that plays a vital role in the process of decision-making is also presented, which can be considered as a visual aid. Altogether, this study offers an image-based diagnostic solution with a visual explanation. Furthermore, an extended study highlights the generalization capability of the proposed modifications, showing that they can be effectively applied to other vision transformer architectures.

This study investigates the potential of the proposed method for diagnostic solutions. However, this study can be extended to explore the prospects of the proposed ViT encoder block in the medical image segmentation domain. Moreover, the usefulness of the proposed method can be investigated in the broad field of computer vision, other than medical imaging.

Chapter 4

Multi-scale Morphology-aided Deep Medical Image Segmentation

Summary

This chapter introduces a novel approach to medical image segmentation designed to address the unique challenges of clinical imaging, such as irregular ROI shapes, variable sizes, and low contrast. Traditional morphological operations have limitations in adapting to these complex features. My solution enhances adaptability by integrating three customized modules based on multi-scale morphological operations: the Multi-scale Morphological Closing, Opening, and Gradient Modules. Unlike conventional techniques, these modules employ trainable structuring elements to capture diverse ROI shapes effectively. To accommodate the full spectrum of ROI sizes, I introduce dilation rates within these morphological operations, allowing for precise, shape-aware segmentation.

The proposed Morph-UNet framework incorporates these modules within the established UNet architecture, creating a lightweight yet powerful segmentation model. Morph-UNet combines deep learning with adaptive morphological operations, demonstrating a unique synergy that results in enhanced segmentation accuracy. Experimental validation across various medical imaging datasets shows Morph-UNet's effectiveness, consistently surpassing thirteen state-of-the-art models and baseline methods across key segmentation metrics. This chapter thus highlights a significant architectural advancement for segmentation tasks in medical imaging, underscoring the framework's adaptability and robustness across multiple imaging modalities and clinical applications.

4.1 Introduction

Medical image segmentation involves identifying and delineating Regions of Interest (ROIs) within clinical images, aiding in the extraction of organs and tissues affected by diseases, etc. While manual segmentation requires expertise and time, computer vision and machine learning offer semi-automatic and fully automatic segmentation methods. Medical image segmentation often involves more complex algorithms compared to general image segmentation due to the intricacies of medical images, including noise, variability in shape and intensity, and the presence of intricate anatomical structures. Despite the widespread use of various imaging modalities, which generate datasets adhering to ethical guidelines, the scarcity of annotated datasets for segmentation tasks remains a challenge. Due to this scarcity, lightweight segmentation architectures emerge as a promising solution, enabling effective segmentation with limited computational resources.

Deep learning algorithms, especially convolutional neural networks (CNNs), have shown remarkable success in medical image segmentation tasks over the years. UNet (Ronneberger et al., 2015) is a convolutional neural network architecture designed for image segmentation, featuring a U-shaped structure that combines downsampling and upsampling pathways to capture context and spatial information effectively. Later, Zhou et al. redesigned the communication between the encoder and decoder block of UNet architecture for improved performance (Zhou et al., 2018). Several studies have adapted UNet architecture as a backbone network upon which integration of different modules has led to segmentation performance enhancement (Valanarasu and Patel, 2022; Wang et al., 2022; Ruan et al., 2022; Chen et al., 2021). Another line of study (Chen et al., 2017a, 2018a, 2017b) focuses on segmenting objects at multiple scales using an Atrous Spatial Pyramid Pooling block based on Atrous Convolution. Few studies have adapted DeeplabV3+ (Chen et al., 2018a) in the domain of clinical image segmentation (da Cruz et al., 2022; Azad et al., 2020).

The task of medical image segmentation entails the delineation of tissues, tumors, glands, lesions, and other regions of interest from their backgrounds. These ROIs exhibit significant variability in terms of shape, size, and texture. While the deep-learning methods discussed above have demonstrated promising performance in segmenting regions of interest in medical images, they have not explored the morphological attributes of these regions, which

are pivotal in medical image segmentation tasks. Thus, I repurpose the traditional mathematical morphological operations, which have previously proven effective in medical image segmentation tasks (Di Rubeto et al., 2000; Mendonca and Campilho, 2006), to the deep learning framework. The traditional mathematical morphological operations consider the predefined structuring elements; thus, previous knowledge of the pattern to be analyzed is required. Shen et al. considered learning the structuring element of morphological operators via training (Shen et al., 2019), thus overcoming its limitation. The trainable mathematical morphological operators are found to be suitable for image classification (Roy et al., 2021; Nogueira et al., 2021), image de-raining (Mondal et al., 2019), image restoration task (Mondal et al., 2020). I present a series of Multi-scale Morphological Modules (MSMMs), designed to extract morphological attributes from feature maps using morphological operations across multiple scales. These modules integrate structuring elements with varying dilation rates to capture attributes at different scales, and the weights of these elements are dynamically estimated during training to accommodate arbitrary shapes in ROI delineation for medical images. I propose a lightweight medical image segmentation network that incorporates the MSMMs with Convolutional Neural Network (CNN) as its backbone. The efficacy of my proposed model is showcased through superior performance across seven clinical datasets for five different medical image segmentation tasks. Against a set of fifteen state-of-the-art segmentation networks designed for diverse medical image segmentation tasks, my proposed method demonstrates superior performance with fewer trainable parameters. Improvements of 0.49%, 4.46%, and 2.88% in intersection-over-union performance metrics for skin lesion, breast tumor, and gland segmentation tasks, respectively are observed. Additionally, it showcases enhancements of 0.99% and 2.92% in the F1-score performance metric for binary cell nuclei and multi-class cell nuclei segmentation tasks, respectively. To the best of my knowledge, no deep learning-based approach has attempted to segment the subtle ROIs of indefinite shapes from clinical images by addressing the morphological attributes.

The contributions of this study are listed as follows.

- I synergize the classical morphological operators, namely *Closing*, *Opening* and *Morphological Gradient*, to the deep learning based medical image segmentation framework. To this end, I introduce three trainable dilation-enabled modules such that the capabil-

ities of morphological operations can be fully explored in the context of medical image segmentation.

- The trainable structuring elements of the proposed morphological modules facilitate the extraction of morphological attributes for irregularly shaped ROIs within medical images. I further introduce the concept of dilation rate within the proposed modules to extract morphological features at multiple scales with a limited number of trainable parameters.

- I propose lightweight medical segmentation architecture ($\approx$ 2.33M parameters) by incorporating the proposed modules within traditional UNet (Ronneberger et al., 2015). Moreover, my framework can be directly coupled with existing CNN encoder architectures such as EfficinetNetB4 ($\approx$ 3.48M parameters) and ResNet34 ($\approx$ 9.28M parameters) for medical image segmentation tasks.

The organization of the rest of the chapter is as follows. The previous works related to this study are discussed in Section 4.2. The preliminary morphological operations are defined in Section 4.3. Methodology and the architectural details of the proposed model are presented in Section 4.4. This section also includes a detailed description of the clinical datasets used to evaluate the proposed method, along with the adapted training protocol and evaluation metrics. Section 4.5 presents the detailed experimental results and discusses their implications as well. Here, I demonstrate the versatile segmentation ability of the proposed model for distinct segmentation tasks — skin lesion segmentation, breast tumor segmentation, and gland and cell nuclei segmentation by using seven medical imaging datasets with diverse modalities, and ROI to image ratios. An elaborate discussion on the appropriateness of a particular proposed module and CNN backbone corresponding to medical images of different characteristics is also comprehended. Finally, Section 4.6 concludes the chapter.

4.2 Related Works

I have examined three distinct medical image segmentation tasks: skin lesions, breast tumors, and cell nuclei or gland segmentation. However, skin lesion and breast tumor images exhibit similar characteristics, including ROI proportions and the number of distinct ROIs in each

image. Consequently, these two tasks often employ similar model approaches. In contrast, segmenting cell nuclei and glands in whole-slide histopathological images differs from dermoscopic or ultrasound images, especially in terms of separate ROI numbers. Moreover, the ratios of each ROI to the image size also. The primary objective for skin lesion and breast tumor segmentation is to accurately delineate and isolate specific regions of interest (lesions or tumors) within medical images. This precision facilitates thorough analysis, diagnosis, and monitoring of these conditions by healthcare experts. In this pursuit, I have explored several deep learning architectures that have emerged in recent years. These architectures can be broadly classified into two main groups: Firstly, there are transformer-based models, which can be relatively heavyweight and may or may not incorporate CNNs. Secondly, there are lightweight models based on the UNet architecture. Following the initial adaptation of the transformer (Vaswani et al., 2017) from the natural language processing domain to image computer vision by Dosovitskiy (Dosovitskiy et al., 2020), several investigations (Zhang et al., 2021c; Chen et al., 2021; Wu et al., 2022) have explored the utilization of transformer encoder blocks, either with or without modifications, for performing segmentation in clinical images. The utilization of Transformers' attention mechanisms in conjunction with CNNs is a notable aspect of TranFuse (Zhang et al., 2021c). The attention mechanism allows the model to focus on relevant features and spatial relationships, potentially enhancing segmentation accuracy. TransUNet (Chen et al., 2021) is introduced as a hybrid architecture that combines the strengths of Convolutional Neural Networks (CNNs) and Transformers. The Transformer serves as a robust encoder, capturing long-range dependencies and contextual information, while the decoder utilizes CNNs for spatial feature extraction. Incorporating an extra transformer branch into the encoder-decoder structure, FAT-Net (Wu et al., 2022) adeptly captures long-range dependencies and global context information. The model introduces a decoder with enhanced memory efficiency, facilitating improved feature fusion between adjacent-level features. This improvement is achieved by selectively activating pertinent channels and suppressing extraneous background noise, ultimately enhancing segmentation accuracy. Authors (Ates et al., 2023) have designed attention modules by combining cross-attentions computed both across the channel and spatial tokens and integrated it in an encoder-decoder segmentation architecture with 30M trainable parameters. The primary objective of these studies (Zhang et al., 2021c; Chen et al., 2021; Wu et al., 2022) was to

incorporate both global and local context in the feature maps extracted by transformer encoder and CNN model, respectively. However, incorporating the transformer multi-head self-attention module into the architectures leads to an increase in the number of trainable parameters. Consequently, this becomes impractical for implementation due to the resulting model complexity and the scarcity of labeled samples required for model training in the medical image segmentation scenario. Given this context, several lightweight image segmentation architectures (Ruan et al., 2022; Valanarasu et al., 2021; Valanarasu and Patel, 2022; Lin et al., 2023) have been designed. MALUNet (Ruan et al., 2022) and Medical Transformer (Valanarasu et al., 2021) are lightweight segmentation models developed by the integration of gated attention mechanisms in UNet shape architecture. These models effectively combine global and local features, utilizing two distinct approaches — one concurrently and the other across multiple stages. UNext, introduced in Valanarasu et al.'s work (Valanarasu and Patel, 2022), presents a streamlined encoder-decoder segmentation model. It integrates an initial convolutional stage and a tokenized MLP block in the latent phase, effectively tokenizing and projecting convolutional features to acquire relevant representation. The authors (Lin et al., 2023) have developed a lightweight hybrid segmentation architecture that combines a CNN and transformer model for medical image segmentation. This architecture incorporates adaptive pruning techniques to remove redundant computations.

Histopathology image segmentation faces several challenges due to the complex nature of tissue structures and the high variability in tissue appearances. Few deep learning-based models are also designed specifically to segment cell nuclei from histopathology images — HoverNet (Graham et al., 2019), HistoSeg (Wazir and Fraz, 2022), SONNET (Doan et al., 2022), TSDF-Net (Ilyas et al., 2022), MCTrans (Ji et al., 2021). DCANet (Ates et al., 2023). These models are also heavy in terms of the number of parameters as compared to other lightweight segmentation models. None of the studies has considered explicitly extracting the morphological features, which play a key role in the medical image segmentation task. However, the irregular shape and size of the ROIs lead to inappropriate extraction of morphological features with predefined structuring elements. This motivates us to design a module for extracting morphological characteristics without having prior knowledge of the shape and size of the region of interest.

4.3 Preliminaries

Mathematical morphological operations are popular tools to extract morphological properties of an image, such as shape, size, texture, etc. The basic mathematical morphological operations, *Erosion* and *Dilation* can be defined as min-sum correlation and max-sum correlation, respectively, presented in equation 5.1 and 5.2.

$$\mathrm{E}_{(i,j,c)}(X) = \Big\|_{c=1}^{C} \min_{m,n=0,\dots K_{SE}} X(i+m, j+n, c) + W_{SE}(m,n,c), \tag{4.1}$$

$$\mathrm{D}_{(i,j,c)}(X) = \Big\|_{c=1}^{C} \max_{m,n=0,\dots K_{SE}} X(i+m, j+n, c) + W_{SE}(m,n,c), \tag{4.2}$$

where $X \in \mathbb{R}^{H\times W\times C}$ is the input feature map and H, W, and C represent respectively the height, width, and number of channels in the feature map. W_{SE} is the structuring element of dimension $K_{SE} \times K_{SE} \times C$ that defines the pattern of interest within the given feature map X. Also, $\|$ denotes the channel-wise concatenation of the feature maps. Unlike 2D *Convolution* operations, morphological operations are independently performed on each input channel. Thus, it results in the same output channels as the input channel, and consequently, the number of parameters to be learned is reduced compared to convolution layers. The *Dilation* and *Erosion* operation expands and shrinks the object's boundary within the image, thus accentuating the shape attribute of the object. The *Opening*, *Closing* and *Morphological Gradients* are three compound operations developed using the primary morphological operations — *Erosion* and *Dilation*. The *Opening* operation is performed by applying an *Erosion* operation followed by a *Dilation* operation, whereas *Erosion* operation can be obtained by the same basic operations in reverse order. The *Morphological Gradient* of an image is defined by the difference between the *Closing* and *Opening* of the same. *Closing* is effective in smoothing object boundaries, filling in small gaps, and connecting nearby structures. *Closing* is often employed to enhance object connectivity, reduce small-scale noise, and ensure more complete object representations. *Opening* is useful for separating objects and removing small, spurious details. It can help improve the discrimination between adjacent structures and reduce noise in the image. *Opening* is commonly used to clean up images by eliminating small structures and enhancing the visibility of distinct objects. Both *Closing* and *Erosion* preserve the shape and size of the object of interest (generally of larger

size) unlike *Erosion* and *Dilation*. *Morphological Gradient*, highlight edges and transitions in intensity within an image. Essentially, when applied to an image, it accentuates the boundary of the objects from its background, making them more distinguishable.

4.4 Proposed Method

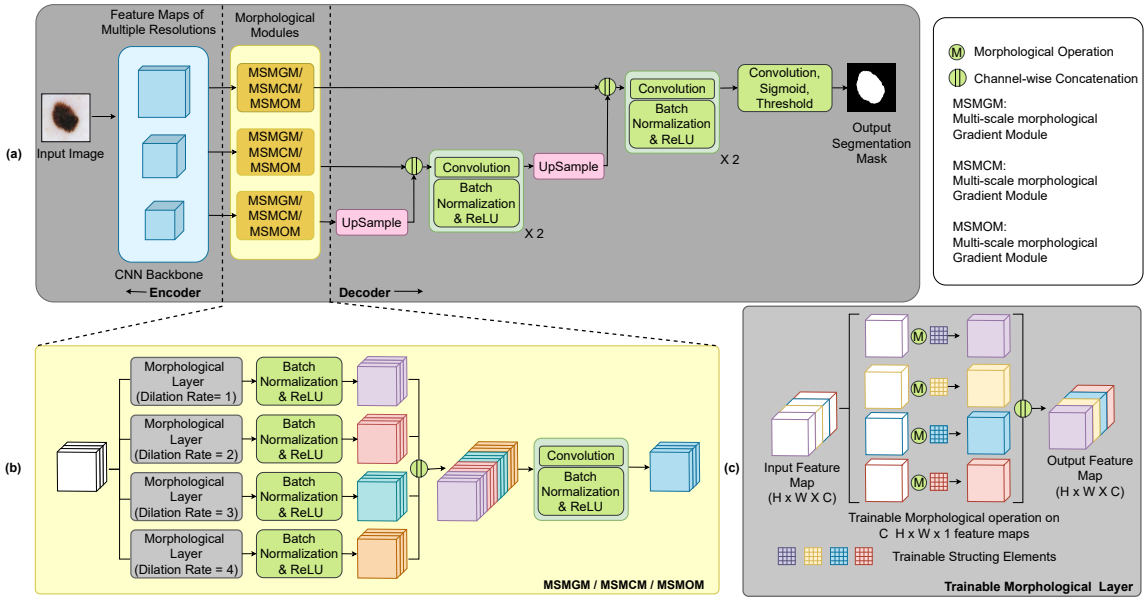

Figure 4.1: (a) The illustration of the proposed architecture designed for medical image segmentation task. (b) The block diagram of the proposed Multi-scale Morphological Modules (MSMM) that are incorporated into the proposed segmentation model. (c) The pictorial depiction of Morphological operations with trainable structuring elements.

4.4.1 Dilated Morphological Operations with Trainable Structuring Elements

The traditional morphological operations have two limitations in using an automated medical image segmentation operation. Firstly, the size of the objects of interest varies from one sample to another; thus, a structuring element of a particular size is insufficient to extract the morphological features. Secondly, the structuring elements are predefined, and the outcome of the morphological operations depends on the structuring elements. However, the irregular shape of the clinical objects limits the traditional morphological operations with

predefined structuring elements to achieve satisfactory results. To deal with the problem of variable object size, the concept of dilating the structuring elements is repurposed within the morphological operations. The successful applications of the dilated kernel in convolution operations have been reported in many studies (Chen et al., 2017a,b, 2018a). This inspired us to incorporate the dilation within the structuring elements of the morphological operation instead of using fixed-size structuring elements. The *Dilated Erosion (DE)* and *Dilated Dilation (DD)* can be defined by the following equations.

$$\begin{aligned} Y_{DD} &= \mathrm{DD}^{r}_{(i,j,c)}(X) \\ &= \Big\Vert_{c=1}^{C} \max_{m,n=0,\dots K} X(i+rm, j+rn, c) + W_{SE}(m,n,c), \end{aligned} \tag{4.3}$$

$$\begin{aligned} Y_{DE} &= \mathrm{DE}^{r}_{(i,j,c)}(X) \\ &= \Big\Vert_{c=1}^{C} \min_{m,n=0,\dots K} X(i+rm, j+rn, c) + W_{SE}(m,n,c), \end{aligned} \tag{4.4}$$

where r is the dilation rate that corresponds to the scale of the structuring element. The pseudo code for Dilated Erosion and Dilated Dilation is presented in Algorithm 1. The dilated morphological operations enable to achieve larger receptive field with no additional computational cost as compared to morphological operations with structuring elements of larger dimensions. To address the second problem, I have considered learning the structuring elements from the training samples themselves via the backpropagation algorithm by adapting the strategy proposed by (Mondal et al., 2019, 2020; Franchi et al., 2020).

Let us assume, I learn structuring elements SE of dilated dilation operation defined by equation 5.4 during training procedure so that the output of this operation Y_{DD} is close to Y'_{DE}, the target output. Let, $\mathcal{L}$ be some loss function and by training, I want to learn the W so that (Y_{DE}, Y'_{DE}) is minimized with respect to W. It is explicit from equation 5.3 that, (Y_{DE}, Y'_{DE}) depends on Y_{DE} and Y_{DE} depends on SE, which is supposed to be learned via training. Thus, following the chain rule for partial derivatives, I obtain,

$$\frac{\partial \mathcal{L}}{\partial W_{\mathrm{SE}}(i,j)} = \sum_i \sum_j \frac{\partial Y_{\mathrm{DD}}(i,j)}{\partial W_{\mathrm{SE}}(i,j)} \frac{\partial \mathcal{L}}{\partial Y_{\mathrm{DD}}(i,j)}, \tag{4.5}$$

where

$$\sum_i \sum_j \frac{\partial Y_{\mathrm{DD}}(i,j)}{\partial W_{\mathrm{SE}}(i,j)}$$
$$= \begin{cases} 1, & if \quad Y_{\mathrm{DD}}(i,j) = X(i + r\partial i, j + r\partial j) + W_{\mathrm{SE}}(\partial i + \partial j) \\ 0, & otherwise. \end{cases}$$

I can update the structuring element by the usual gradient descent iterations as follows,

$$W_{\mathrm{SE}}(i,j) = W_{\mathrm{SE}_{prev}}(i,j) - \eta \frac{\partial \mathcal{L}}{\partial W_{\mathrm{SE}}(i,j)}, \tag{4.6}$$

where $W_{\mathrm{SE}_{prev}}$ is the structuring element in previous iteration, η is the learning rate. I can also update the structuring element of a dilated erosion layer similarly. Thus, the fundamental morphological layers — *Dilated Dilation* and *Dilated Erosion* become trainable.

Dilated Opening (DO), Dilated Closing (DC) and Dilated Morphological Gradients (DG) with trainable structuring elements can be mathematically expressed in terms of *Dilated Erosion (DE), Dilated Dilation (DE)* by the following equations. Mathematically,

$$\mathrm{DO}^r(X) = \mathrm{DD}^r(\mathrm{DE}^r(X)), \tag{4.7}$$

$$\mathrm{DC}^r(X) = \mathrm{DE}^r(\mathrm{DD}^r(X)), \tag{4.8}$$

$$and \quad \mathrm{DG}^r(X) = \mathrm{DD}^r(X) - \mathrm{DE}^r(X). \tag{4.9}$$

Dilated morphological operators with trainable structuring elements in medical image segmentation provide adaptability, robustness, and efficiency. Trainable structuring elements adapt to varying features, making the segmentation method robust across diverse medical images. The approach eliminates manual tuning, saving time in the segmentation process. It efficiently increases the receptive field and enables multi-scale feature extraction. However, potential drawbacks include increased computational complexity, reduced interpretability, and a risk of overfitting to specific features in the training dataset.

4.4.2 Multi-scale Morphological Module (MSMM)

I propose three modules based on dilated morphological operations to extract complex morphological features. The fundamental architecture of these modules is the same and depicted

Algorithm 1: Pseudo code of the *Dilated Erosion* and *Dilated Dilation* operations

Inputs: $X \in \mathbb{R}^{H \times W \times C}$ be the feature map, K_{SE} be the size of the structuring element for morphological operation, r be the dilation rate, *type* represents the morphological operation type, S is the stride the value of which is by default 1

Outputs: $X_{out} \in \mathbb{R}^{H \times W \times C}$ be the output feature map

Initialize: *Initialize structuring elements* $W_{SE} \in \mathbb{R}^{K_{SE} \times K_{SE} \times C}$ *with all zeros*

H, W, $C \leftarrow$ height, width, and channel number of input feature map X
$P = int((S-1)H - S + K_{\text{SE}})/2)$ `// padding for output feature map of same dimension as input feature map, in the proposed case height and width of the feature map is same, thus padding for bpth dimension is same`
$K_e = K_{\text{SE}} + (K_{\text{SE}} - 1)(r-1)$ `// effective kernel size due to dilation rate`
for $i = 0$ **to** H **do**
  for $j = 0$ **to** W **do**
    `// Extract the dilated patches from the feature map`
    $X_{patch} = X[i.S : i.S + K_e : r, j.S : j.S + K_e : r, :]$
    if *type is* $'Dilation'$ **then**
      `// for Dilated Dilation operation`
      $X_{out}[i,j] \leftarrow max(X_{patch} + W_{SE})$;
    else
      if *type is* $'Erosion'$ **then**
        `// for Dilated Erosion operation`
        $X_{out}[i,j] \leftarrow min(X_{patch} - W_{SE})$;
      end
    end
  end
end
return X_{out}

in Figure 4.1(b). Any one of the complex dilated morphological operations (*Dilated Opening*, *Dilated Closing*, and *Dilated Morphological Gradients*) with different dilation rates are applied on the feature map to obtain multi-scale morphological features, which are then channel wise integrated. A convolution block is employed afterward to reduce the number of channels. The number of channels in the input and output feature maps is equal. I designed three such modules based on three types of complex morphological operations included in the module — Multi-scale Morphological Opening Module (*MSMOM*), Multi-scale Morphological Closing Module (*MSMCM*) and Multi-scale Morphological Gradient Module (*MSMGM*). Thus, the

output of an MSMM can be mathematically expressed as

$$Y = \text{MSMM}(X) = \text{ReLU}(\text{BN}(X' * W_{conv})) \tag{4.10}$$

where,

$$X' = \begin{cases} \Big\|_{r=1}^{R} \text{ReLU}(\text{BN}(\text{DC}^r(X))) & \text{for MSMCM,} \\ \Big\|_{r=1}^{R} \text{ReLU}(\text{BN}(\text{DO}^r(X))) & \text{for MSMOM,} \\ \Big\|_{r=1}^{R} \text{ReLU}(\text{BN}(\text{DG}^r(X))) & \text{for MSMGM.} \end{cases} \tag{4.11}$$

The convolution operation between the X' and convolutional kernel W_{conv} is denoted by $*$. ReLU and BN represent the rectifier linear unit and batch normalization operation (Ioffe and Szegedy, 2015) defined by following equations.

$$\text{ReLU}(z) = max(0, z), \tag{4.12}$$

$$\text{BN}(z) = (z - \mu)/\sigma, \tag{4.13}$$

where, μ and σ is the mean and variance of z within a batch. The pseudo code for MSMMs is presented in Algorithm 2. MSMMs excel at extracting complex features by leveraging the non-linearity of morphological operations, the adaptability of structuring elements, and the application of operations at multiple scales. The inherent non-linearity of erosion and dilation, key morphological operations, stems from their reliance on the minimum and maximum functions, making them nonlinear processes that do not adhere to the principles of superposition and homogeneity. The adaptive learning of structuring elements featuring trainable parameters allows the model to automatically adjust to the specific characteristics of input data during training. Additionally, the multi-scale morphological operation utilizes structuring elements of varying scales, enabling the model to analyze image structures at multiple levels of granularity and capture both fine details and broader contextual information.

4.4.3 Morph-UNet

UNet (Ronneberger et al., 2015), the convolutional neural network for medical image segmentation, has been utilized in this study as the baseline model. It follows an encoder-decoder architecture. The encoder gradually decreases the dimensionality of the feature maps, whereas

Algorithm 2: Pseudo code of the Multi-scale Morphological Module (MSMM)

Inputs: $X \in \mathbb{R}^{H \times W \times C}$ be the feature map, K_{SE} be the size of the structuring element for morphological operation, R be the list of dilation rates, S be the stride with default value of 1.

Outputs: $X_{out} \in \mathbb{R}^{H \times W \times C}$ be the output feature map

for $r = 1$ **to** $Length(R)$ **do** `// For loop running over different dilation rates`

  `// DMorphOps is the function to perform dilated morphological operations as defined by algorithm 1`

  if *type is Closing* **then**

    $X_r \leftarrow DMorphOps(DMorphOps(X, \text{'Dilation'}, K_{SE}, r, S), \text{'Erosion'}, K_{SE}, r, S)$;

  else

    if *type is Opening* **then**

      $X_r \leftarrow DMorphOps(DMorphOps(X, \text{'Erosion'}, K_{SE}, r, S), \text{'Dilation'}, K_{SE}, r, S)$;

    end

    if *type is Gradients* **then**

      $X_r \leftarrow DMorphOps(x, \text{'Erosion'}, , K_{SE}, r, S) - DMorphOps(X, \text{'Dilation'}, , K_{SE}, r, S)$;

    end

  end

end

$X \leftarrow \text{concatenate}(X_1, X_2, X_2, ..., X_{length(R)})$; `// channel wise concatenate the morphological feature maps of different dilation rates`

$X \leftarrow \text{Conv2D}(X, \text{kernel} = 1, \text{input\_channel} = Length(R) \cdot C, \text{output\_channel} = C)$; `// Apply pointwise convolution operation to reduce the channels number from` $Length(R) \cdot C$ `to` C

$X_{out} \leftarrow \text{Dropout}(\text{ReLU}(\text{BN}((X))))$; `// apply batch normalization, ReLU activation function and Dropout.`

return X_{out}

the decoder reverses it back to the original dimension. Skip connections from various levels of the encoder stage to the corresponding decoder stage and build a bridge between them. In my proposed segmentation architecture, the encoder yields three feature maps of different dimensions presenting the different levels of complexity and integrity of the spatial position of the pixels. These feature maps are passed through the proposed morphological modules to obtain their morphological characteristics, vital in segmenting the object of interest. Later, the outputs of these morphological modules are merged with the feature maps of decoder blocks

Table 4.1: The number of trainable parameters of the proposed modules and models with different configurations, and comparison with baseline model.

Model	Channels	Modules	Parameters
-	256	MSMCM/MSMOM	0.28 M
-	256	MSMGM	0.27 M
UNet	-	-	1.94 M
Morph-UNet	[64, 128, 256]	MSMCM/MSMOM	2.33 M
		MSMGM	2.31 M
Morph-UNet-EfficientNetB4	[24, 32, 56]	MSMCM/MSMOM	3.48 M
		MSMGM	3.42 M
Morph-UNet-ResNet34	[64, 128, 256]	MSMCM/MSMOM	9.28 M
		MSMGM	9.25 M

of corresponding dimensions via skip connections. Figure 4.1 shows the detailed diagram of my proposed architecture, which is referred to as *Morph-UNet*.

Three variations of *Morph-UNet* are designed in this study with two different backbone networks — ResNet34 and EfficientNetB4. The choice of backbone networks in the Morph-UNet architecture can have significant implications for segmentation performance in various medical imaging tasks. The backbone network serves as the feature extractor, providing a hierarchical representation of image features to the subsequent segmentation layers. Different backbones have distinct architectures, complexities, and pre-trained weights, leading to variations in performance. Transfer learning plays a crucial role in enhancing the performance of any deep learning model. ResNet is known for its deep architecture and ability to capture intricate details, but it comes with a higher number of parameters and computational demands. EfficientNet, on the other hand, emphasizes model efficiency with fewer parameters and computational efficiency. Both architectures support transfer learning. Additionally, in ResNet34, the initial convolutional layers employ a 7×7 kernel, while EfficientNetB4 utilizes a 3×3 kernel for the same purpose. Consequently, when dealing with very small and distinct regions of interest (ROI), ResNet34 as a feature extractor may not exhibit optimal performance in segmentation tasks when compared to EfficientNetB4. This is attributed to the precision deficiency stemming from the larger kernel size in ResNet34. However, for larger and more complex ROIs, ResNet34 outperforms EfficientNetB4 in the segmentation task. The detailed description of the networks is given in Table 4.1.

Algorithm 3: Pseudo code for training proposed Morph-UNet

Inputs: $\mathbb{X} \in \mathbb{R}^{N \times H \times W \times C}$ be the N input images to train the model and $\mathbb{Z} \in \mathbb{R}^{N \times H \times W \times N_s}$ be the corresponding mask of N images, where N_s is the number of classes. For Binary segmentation task, the value of N_s is 2.

Outputs: $\phi_{En}(.)$, $\phi_{De}(.)$ and ϕ_{MSMM_i} be the set of weights for the encoder, decoder part of UNet network and i^{th} MSMM module.

for *iteration* $= 1$ **to** *maximum iteration* **do**
  for *batches* $= 1$ **to** *number of batches* **do**
    $X \leftarrow$ a batch of input images taken from $\mathbb{X}$
    XE_1, XE_2,..., $\text{XE}\,i \leftarrow \phi_{En}(X)$
    for $i = 1$ **to** N_{En} **do** `// For loop running for different stages of encoder`
      $X_{MSMM_i} = \phi_{MSMM_i}(\text{XE}_i)$
    end
    `// decoder is defined by set of parameters` $\phi_{De}(.)$
    $X = X_{MSMM_{N_{En}}}$; `//` $X_{MSMM_{N_{En}}}$ `is the feature map from the bottleneck layer`
    for $j = N_{De}$ **to** 2 **do** `// For loop running for different stages of decoder`
      $X = ConvBlock(concatenate(upsample(x), X_{MSMM_j}))$
    end
    $\hat{Z} \leftarrow Activation(Conv(X))$
    loss $\leftarrow \mathcal{L}(Z, \hat{Z})$; `// Z is the mask corresponding to the batch X`
    update parameters of ϕ_{En}, ϕ_{En} and ϕ_{MSMM_i} using backward propagation of loss
  end
end
return $\phi_{En}(.)$, $\phi_{De}(.)$ and ϕ_{MSMM_i}

4.4.4 General Training Protocol

The weights of the proposed models and also the baseline models are learned by optimizing the loss function. I have used three different loss functions — dice loss (Milletari et al., 2016), a combination of dice loss and binary cross-entropy, and focal loss (Lin et al., 2017) as defined by equation 4.14, 4.16 and 4.17.

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2\sum_{i=1}^{N} z_i \cdot \hat{z}_i}{\sum_{i=1}^{N} z_i + \sum_{i=1}^{N} \hat{z}_i}, \tag{4.14}$$

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N} \left[z_i \log \hat{z}_i + (1 - z_i) \log (1 - \hat{z}_i) \right], \tag{4.15}$$

$$\mathcal{L}_{\text{Comb}} = \mathcal{L}_{\text{Dice}} + 0.5 \cdot \mathcal{L}_{\text{BCE}}, \tag{4.16}$$

$$\mathcal{L}_{\text{Focal}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} \alpha_c (1 - p_{t,c})^{\gamma} z_{i,c} \log \hat{z}_{i,c}, \tag{4.17}$$

where, N is the total number of pixel in an image, C is the total number of classes, z_i and $\hat{z}_i$ represent the ground truth label and predicted probability of i^{th} pixel, α_c is the weighing factor for class c and γ indicates the degree of down-weighting of easy-to-classify pixels. $z_{i,c}$ and $\hat{z}_{i,c}$ represent the ground truth label and predicted probability of i^{th} pixel that belong to c^{th} class. $p_{t,c}$ refers to the predicted probability associated with the ground truth class and is formally defined as follows: $p_{t,c} = \hat{z}_{i,c}$ if the true label $z_{i,c} = 1$, and $p_{t,c} = 1 - \hat{z}_{i,c}$ otherwise. Here, t denotes the ground truth class index for the i^{th} sample.

The ROIs in medical images vary in size and generally take up only a tiny portion of the whole image. This foreground-background imbalance and the size variation of ROIs are both handled by dice loss. I have used dice loss for all experiments that are conducted for the ablation studies. On the other hand, by combining dice loss with binary cross entropy loss, the model benefits from the stable optimization properties of binary cross entropy loss and the ability of dice loss to handle class imbalance, ensuring better performance. This combined loss function is utilized to train the weight of the models while comparing them with state-of-the-art methods. I have also employed focal loss, a variant of cross entropy loss that is designed to address the class imbalance by down-weighting the contribution of easy-to-classify examples and focusing more on hard-to-classify examples. This makes it particularly useful in scenarios where there is a significant imbalance between classes. An example of such a scenario is the PanNuke dataset (for a detailed description refer to section 5.4), where tissues categorized into *Dead* class are very few in number and consequently they become hard to classify. The loss function is optimized using Adam optimizer with an initial learning rate of 0.001 up to 200 epochs. The model with minimum loss is considered the final trained model, which is further used to predict the segmentation mask of the test samples.

4.4.4.1 Metrics

Let, GT and P be the groundtruth mask and predicted mask. True positive, false positive, true negative, and false negative are referred to as tp, fp, tn, and fn, respectively. I considered five pixel-based metrics in this study to quantify segmentation performance, namely — Pixel Accuracy (PA), precision, recall, Dice Score (DS)/ F1-score, and Intersection over Union(IoU). The following equations define the pixel-based metrics — in terms of tp, fp, tn, and fn.

$$\text{Precision}(P, GT) = \frac{tp}{tp + fp}. \tag{4.18}$$

$$\text{Recall}(P, GT) = \frac{tp}{tp + fn}. \tag{4.19}$$

$$\text{DS}(P, GT) = \frac{2(P \cap GT)}{(P \cap GT) + (P \cup GT)} = \frac{2tp}{2tp + fn + fp}. \tag{4.20}$$

$$\text{IoU}(P, GT) = \frac{(P \cap GT)}{(P \cup GT)} = \frac{tp}{tp + fn + fp}. \tag{4.21}$$

The larger value of the pixel-based segmentation metric indicates a better estimation of the segmentation mask.

4.4.4.2 Dataset Description

Seven diverse medical imaging datasets that are popular in the field of medical image segmentation tasks are considered in this study to evaluate the proposed method. These datasets can be divided into three groups according to the type of regions of interest to be segregated from the background, namely — skin lesions, breast tumors, and microscopic nuclei and gland segmentation.

Skin Lesion Segmentation The skin lesion datasets are characterized by their diverse range of lesions, including melanomas, nevi, and benign lesions, reflecting the breadth of skin conditions. The dataset presents challenges in segmentation due to the varied textures, colours, and shapes of skin lesions, making accurate delineation complex. Additionally, skin lesions can appear on different body parts with varying background textures, adding further complexity to segmentation tasks. Additionally, Challenges include class imbalance due to low

ROI to background ratio, thus distribution biases may affect the model's accuracy. Moreover, the difficulty in accurately delineating lesions with ambiguous boundaries or irregular shapes arises due to low contrast foreground and background. Other than the challenges mentioned above, difficulty in the precise segmentation of skin lesions arises from the presence of hairs, pen marks, stickers, bubbles, etc.

For the skin lesion segmentation task, I considered three most popular datasets of dermoscopic images of skin lesions — ISIC 2017 (Codella et al., 2018) and ISIC 2018 (Codella et al., 2019) and HAM10000 (Tschandl et al., 2018). These are three-channel images. There are 2000, 150, and 600 samples for training, validation, and testing purposes in ISIC 2017. However, for ISIC 2018, only 2594 training samples with their ground truth segmentation masks are publicly available. Thus, I randomly selected 20% and 10% of the available samples for testing and validation purposes while the rest are used to train the models. Among the three datasets, HAM10000 is the largest, comprising 100015 dermoscopic images. All the dermoscopic images are resized to 256×256 images.

Breast Tumor Segmentation Breast tumour segmentation plays a vital role in many computer-aided diagnosis systems and is an important task in medical image analysis, particularly in the context of breast ultrasound imaging. It entails defining or highlighting the tumor location in breast pictures obtained using different imaging techniques, including mammography, ultrasound, or MRI. The breast tumor dataset exhibits characteristics such as tumor heterogeneity, with various types like ductal and lobular carcinomas, each possessing distinct features. Tumors can appear at different locations within the breast, and their shapes and sizes vary, adding complexity to segmentation tasks. Delineating precise tumor boundaries is crucial for accurate diagnosis and treatment planning, presenting challenges, especially in cases of irregularly shaped tumors.

I have considered the breast ultrasound images (BUSI) dataset (Al-Dhabyani et al., 2020) and UDIAT (Yap et al., 2017) in this study. The BUSI dataset comprises images of three categories — normal, benign, and malignant along with the masks indicating the tumor and non-tumor pixels. I have only considered the images with tumors, i.e. benign and malignant, following a similar approach as (Ruan et al., 2022). This resulted in 647 ultrasound images with various dimensions, which are resized to the dimension of 256×256. I randomly divided

the dataset into training and test samples with an 80:20 ratio. The UDIAT dataset contains a total of 163 ultrasound images, out of which 110 images are of benign category, and the remaining 53 images represent malignant breast masses. All the images of the UDIAT dataset are resized to the dimension of 256×256. I have followed a four-fold cross-validation approach for splitting the data into training and test data as considered in (Chen et al., 2022a).

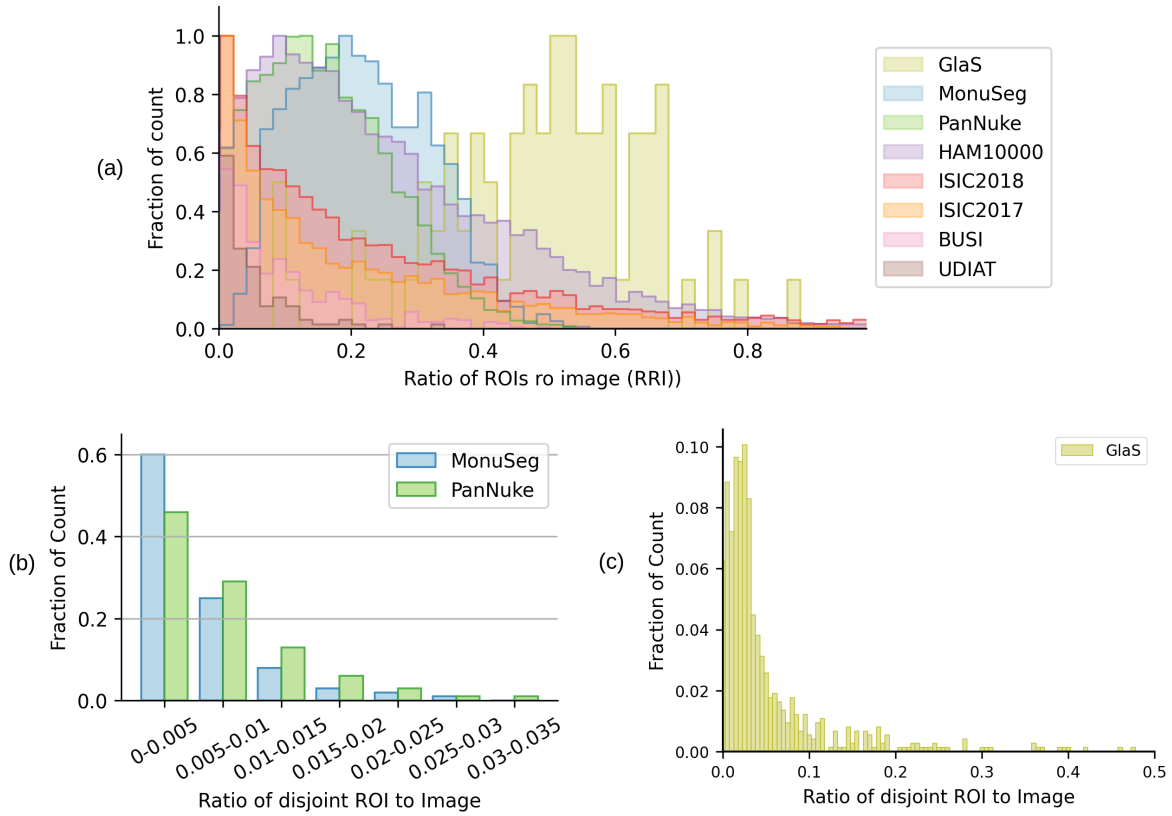

Figure 4.2: Distribution of Ratio of ROI to Image for ISIC2017, ISIC2018, BUSI, MonuSeg and PanNuke dataset (upper panel) and distribution of the ratio of each disjoint ROIs to Image for MonuSeg and PanNuke dataset (middle panel)and GlaS dataset (lower panel) are presented.

Glands and Nuclei Segmentation The third group of datasets used to evaluate the segmentation models is the collection of histopathological images of human body tissues. In histopathology images, gland segmentation tasks involve identifying and delineating various characteristics of glandular structures. These characteristics include diverse shapes, variable density and distribution, staining variability, specific cellular arrangements, accurate depic-

tion of glandular boundaries, handling size variability, leveraging texture and contrast differences, and capturing intra-glandular heterogeneity. Gland segmentation algorithms must be adaptable to these features to provide accurate representations of tissue structures and support comprehensive histopathological analyses. The microscopic nuclei dataset is characterized by high-resolution images capturing detailed cellular structures. Overlapping nuclei pose a challenge in accurately separating individual nuclei during segmentation. The dataset's complexity is further heightened by variations in nuclear size, shape, and staining intensity, contributing to the intricacy of segmentation tasks. Challenges in this dataset include accurately segmenting densely packed nuclei in crowded scenes without merging or undersegmentation. Staining variability across images introduces another challenge, necessitating robust segmentation methods that can account for inconsistencies in staining intensity.

The attributes of this dataset group are relatively different from the other two datasets. In both breast tumor and skin lesion segmentation tasks, the number of disjoint ROIs to be segmented is generally one. However, there are multiple disjoint ROIs in the task of nuclei as well as gland segmentation from microscopic images. Moreover, the areas of each distinct ROI are relatively small in nuclei segmentation tasks compared to breast tumor and skin lesion segmentation tasks (Figure 4.2). I have considered the GlaS dataset for the gland segmentation task and the MonuSeg and PanNuke dataset for the nuclei segmentation task.

GlaS (Sirinukunwattana et al., 2017) is composed of 165 images derived from 16 H&E stained histological sections of stage T3 or T42 colorectal adenocarcinoma and its corresponding ground truth mask indicating the glands. Out of these 165 images, 85 samples are marked as training samples, and the rest are used as test samples. Since the images are of various sizes, I first resize the images to 512×768, thereafter extracting patches of size 256×256 with 50% overlap and without overlap for training and test samples, respectively. This prepossessing step results in 1275 and 480 samples for training and testing purposes.

The MonuSeg dataset (Kumar et al., 2017, 2019) is taken from the MoNuSeg 2018 Challenge. It comprises hematoxylin and eosin (H&E) stained microscopic images of tissues from multiple human body organs captured at 40X resolution. There are 37 (the annotation file of one not readable) and 14 $1000\times$ 1000 dimensional images with ground truth masks for training and testing, respectively. Image patches of size 256×256 are extracted from training and test images with 50% overlap and no overlap, respectively. The 20% of the training

patches are used for validation purposes.

The PanNuke dataset (Gamper et al., 2019) is the collection of whole slide images(WSI) of cancerous tissue slides from which patches 256×256 are extracted for training and testing purposes. The ground truth masks annotate 19 different types of tissues that are broadly categorized into five different classes – *Epithelial*, *Connective/Soft tissue cells*, *Lympho-reticular cells*, *Nervous system cells* and *Dead*. Thus, I pose this segmentation task as a six-class pixel-wise classification task where the sixth class represents the background class. A total of 7901 images are subgrouped into three folds. I follow the original division of folds for training, testing, and validation purposes.

4.4.5 Experimental Protocol

For each dataset used in this study, I have normalized each image with the mean and standard deviation of the pixel intensity. Normalization helps in stabilizing and accelerating the training process by ensuring that the input features (pixel values) have a consistent scale. This aids in faster convergence during optimization. Other than the experiments reported in section 4.5.1, I have applied all or a few of these operations — horizontal flip, vertical flip, shift, scale, rotate, and contrast limited adaptive histogram equalization(CLAHE) on the training dataset using Albumentations (Buslaev et al., 2020) library for data augmentation purpose. Augmentation is used to artificially increase the size of a training dataset by applying various transformations to the original images. These transformations introduce variations in the training data, helping the model generalize better to unseen data.

All the codes are written using PyTorch library and are run on a computer with AMD Ryzen 9 7900X 12-Core Processor running at 3 GHz using 128 GB of RAM and NVIDIA GeForce RTX 4090 GPU with 24 GB RAM. The PyTorch implementation is available at this repository https://github.com/SusmitaSenGhosh/Morph-UNet-An-Approach-Towards-Medical-Image-Segmentation.

4.5 Result and Discussion

4.5.1 Ablation Study

Three multi-scale morphological pyramid pooling modules — MSMGM, MSMCM, and MSMOM are designed to capture multi-scale morphological features. I investigate the efficiency of these modules by incorporating them with the vanilla UNet module. A morphological module is employed on the high-level feature map provided by the encoder block. I refer to these architectures as UNet-MSMGM, UNet-MSMOM and UNet-MSMCM. Along with this, I also consider another module, the Multi-scale depth-wise separable convolutional module, MSDCM, which follows a similar architecture as MSMM, where Morphological operations are replaced by depth-wise separable convolution operations. This architecture is referred to as UNet-MSDCM. This module is incorporated in this ablation study to investigate the effectiveness of the morphological modules over convolutional modules of a similar number of trainable parameters in the context of medical image segmentation. Furthermore, to ensure comprehensive analysis, I have investigated an additional scenario wherein all three morphological modules were concurrently applied to the high-level feature map. This architecture is referred to as UNet-MSMAM. The UNet network is considered here as a baseline model. The performances of these architectures are shown in Figure 4.3 in terms of F1-score, recall, and precision metrics. In medical image segmentation tasks, achieving a balance between precision and recall is essential, directly influencing the segmentation model's performance and clinical applicability. Precision gauges the accuracy of positive predictions, representing the ratio of true positives to the total positives predicted by the model. High precision ensures that identified regions are more likely to be true positives, minimizing the risk of false alarms in clinical settings. Recall assesses the model's ability to capture all relevant instances of a specific class, which is crucial for accurate diagnosis and treatment planning. Striking a balance between precision and recall is challenging, as enhancing one often comes at the cost of the other. This trade-off becomes particularly significant in scenarios where segmentation results guide medical interventions, necessitating careful consideration of precision and recall to ensure clinical efficacy.

The reported results represent the average performance from conducting the same experiments ten times, each with distinct random initialization of network weights. However,

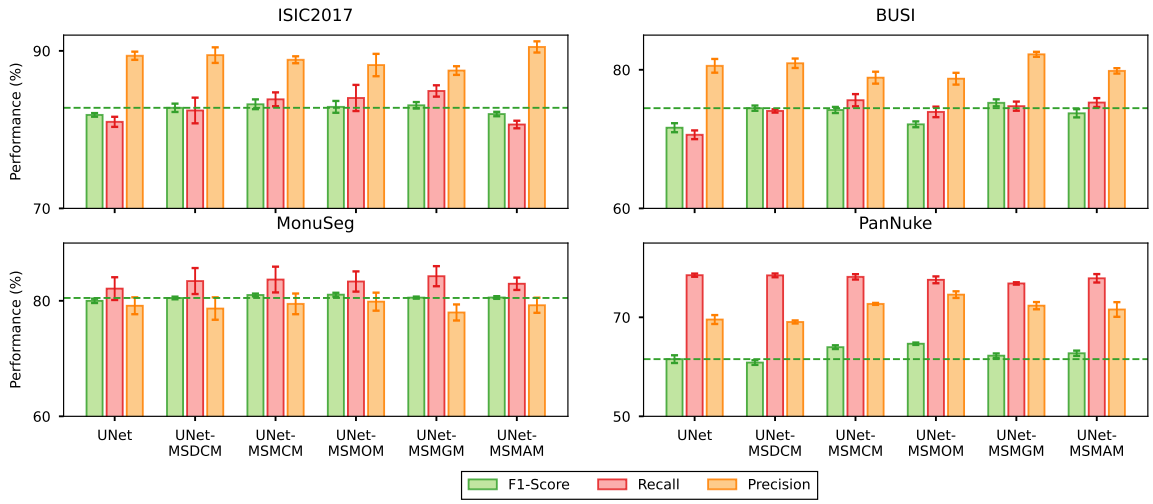

Figure 4.3: The performance comparison of the proposed UNet-MSMCM, UNet-MSMOM, and UNet-MSMGM models against baseline UNet and UNet-MSDCM model in terms of F1-score, recall, and precision. The Green dashed line indicates the best of the F-score achieved by UNet or UNet-MSDCM models.

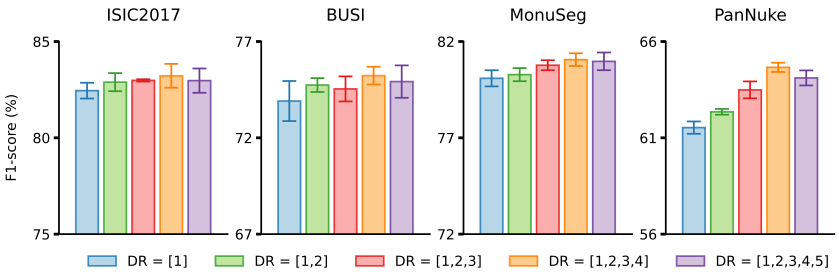

Figure 4.4: The performance comparison of different sets of dilation rates within the proposed modules.

for the PanNuke dataset, the average is calculated across three separate test dataset folds. Among the three datasets for the skin lesion segmentation task, only ISIC2017 is considered in this ablation study as they share similar characteristics. I observe that UNet-MSMCM and UNet-MSMGM models outperformed UNet and UNet-MSDCM models in terms of F1-score for ISIC2017 and BUSI datasets. For both MonuSeg and PanNuke datasets, UNet-MSMOM performed best among all the models considered in this experiment. It should also be noted that the UNet model, when integrated with different types of morphological modules, tends to maintain a balanced relationship between precision and recall values. This is in contrast to both the vanilla UNet and UNet-MSDCM. The proposed MSMMs employ structuring

elements of multiple sizes to perform morphological operations. These structuring elements adapt to the scale of structures in the image. Using multi-scale structuring elements helps in capturing both small and large features, contributing to a more comprehensive segmentation. Moreover, MSMMs with trainable structuring elements offer a mechanism to adapt structuring elements based on the specific requirements of the segmentation task. This adaptability contributes to managing the trade-off between recall and precision by allowing the model to learn and optimize morphological operations for the characteristics of the input data, ultimately enhancing the segmentation performance by balancing between recall and precision. Furthermore, I note that incorporating all three Multi-Scale Morphological Modules (MSMMs) into the UNet architecture does not yield performance improvements compared to the architecture with a single MSMM. However, it introduces an increase in model complexity due to the addition of more trainable parameters.

In the previous experiment, I considered four dilation rates (1,2,3, and 4) within each proposed module and kept them unchanged throughout the experiment. Here, I conduct another ablation study to determine the most effective dilation rates. For each of the four datasets (ISIC2017, BUSI, MonuSeg, and PanNuke), the best-performing modules are chosen to take part in this experiment. I include five sets of dilation rates —- {1},{1,2},{1,2,3},{1,2,3,4} and {1,2,3,4,5} within the proposed modules and investigate their performance using the similar training protocol as used in the previous experiment. From the results depicted in Figure 4.4, I can conclude that the set of dilation rates that performed best for all the datasets is {1,2,3,4}. Thus, this set of dilation rates is followed throughout all the experiments reported in this study.

The proposed Morph-UNet follows an encoder-decoder architecture where features are encoded to three different resolutions corresponding to three different stages. I proposed to incorporate three MSMM blocks in the three skip connection paths between the encoder and decoder. I conducted an ablation study to justify the chosen number of MSMM blocks and their position in the proposed segmentation architecture. I consider all the combinations of the position of one or two three MSMM blocks and reported the performance in terms of IoU and F1-score for ISIC17, BUSI, MonuSeg, and PanNuke dataset in Table 4.2. Comparing segmentation performance, it is explicit that when three MSMM blocks are incorporated after the three feature maps extracted by the convolutional blocks, it yields the best result.

Table 4.2: Performance of the Morph-UNet with varying numbers of MSMMs and their positions.

	ISIC17		BUSI		MonuSeg		PanNuke	
MSMM Position	**IoU**	**F1-Score**	**IoU**	**F1-Score**	**IoU**	**F1-Score**	**IoU**	**F1-Score**
Branch I	73.30 ± 0.58	82.34 ± 0.44	65.23 ± 1.01	74.26 ± 1.31	66.50 ± 0.31	79.59 ± 0.27	55.62 ± 2.69	60.40 ± 2.75
Branch II	73.43 ± 0.86	82.47 ± 0.72	64.48 ± 1.09	73.05 ± 1.27	66.87 ± 0.58	79.81 ± 0.49	57.25 ± 2.85	61.97 ± 2.72
Branch III	73.92 ± 0.91	83.29 ± 0.84	66.91 ± 0.66	75.21 ± 0.89	68.03 ± 0.49	80.69 ± 0.37	60.23 ± 0.48	64.66 ± 0.71
Branch I & II	73.67 ± 0.98	82.73 ± 0.84	64.51 ± 0.99	73.35 ± 0.98	66.59 ± 0.78	79.64 ± 0.59	56.90 ± 0.83	61.66 ± 0.63
Branch I & III	73.77 ± 0.34	82.63 ± 0.31	66.94 ± 0.44	75.06 ± 0.89	67.96 ± 0.78	80.68 ± 0.62	57.83 ± 0.59	62.38 ± 0.61
Branch II & III	73.91 ± 0.66	82.83 ± 0.71	66.25 ± 1.17	74.80 ± 1.29	67.67 ± 0.65	80.45 ± 0.51	58.86 ± 1.34	63.47 ± 1.29
Branch I, II & III	**74.01 ± 0.88**	82.91 ± 0.70	**67.61 ± 0.86**	**75.89 ± 0.90**	**68.14 ± 0.60**	**80.83 ± 0.45**	**60.27 ± 0.25**	**64.73 ± 0.47**

4.5.2 Performance on Skin Lesion and Breast Tumor Segmentation

Considering the shared characteristics between images utilized for skin lesion and breast tumor segmentation, including factors like the count of ROIs per image and the proportional size of each ROI within the image, I address both tasks concurrently to explore the potential of the proposed methodologies. Skin lesions play a vital role in early autoimmune skin disease detection. Therefore, automatically segmenting skin lesions from their surroundings aids medical experts in highlighting the regions of interest for further evaluation. Additionally, precise tumor segmentation can assist radiologists and oncologists in diagnosing breast cancer, planning treatment, or monitoring its progression. The effectiveness of the proposed models is showcased using the ISIC2017, ISIC2018, and HAM10000 datasets, along with the BUSI dataset (refer to Section 5.4 for detailed description), specifically tailored for skin lesion and breast tumor segmentation tasks, respectively.

Table 4.3 and Table 4.5 compare the proposed models' performance with the state-of-the-art segmentation models in the context of skin lesion segmentation and breast tumor segmentation tasks, respectively. To report a fair comparison, I consider the similar training protocol and evaluation metrics as followed in (Valanarasu and Patel, 2022). For all the datasets except UDIAT, I split the training and test data in an 80:20 ratio three times, and the mean and standard deviation of the performance metrics, namely IoU and F1-score are reported. However, in the case UDIAT dataset I have used four-fold cross-validation as used in (Chen et al., 2022a). I have considered three variations of the Morph-UNet architectures of different CNN backbones, resulting in networks with different numbers of trainable parameters. Along with segmentation performances, I also compare the space, time, and computational complexity of the proposed methods with the state-of-the-art methods. The space complexity of a model is quantified by the number of trainable parameters of the

Table 4.3: Performance comparison of the proposed methods with the state-of-the-art skin lesion segmentation methods in terms of IoU and F1-score for ISIC2017, ISIC2018, and HAM10000 datasets.

Method	Module	ISIC2017		ISIC2018		HAM10000	
		IoU	F1-score	IoU	F1-score	IoU	F1-score
MALUNet (Ruan et al., 2022)	-	80.42 ± 0.18*	88.91 ± 0.11*	78.74 ± 0.33*	87.84 ± 0.21*	87.16 ± 0.62*	93.03 ± 0.37*
UNext (Valanarasu and Patel, 2022)	-	82.42 ± 0.20*	90.18 ± 0.16*	81.95 ± 0.41*	89.88 ± 0.28*	89.04 ± 0.31*	94.12 ± 0.18*
APFormer (Lin et al., 2023)	-	83.59 ± 0.25*	90.94 ± 0.14*	82.20 ± 0.30*	90.02 ± 0.17*	89.37 ± 0.08*	94.32 ± 0.03*
TransFuse (Zhang et al., 2021c)	-	83.95 ± 0.44*	91.13 ± 0.26*	82.38 ± 1.14*	90.12 ± 0.71*	89.55 ± 0.41*	94.41 ± 0.24*
FAT-Net (Wu et al., 2022)	-	85.08 ± 0.18*	91.81 ± 0.12*	83.48 ± 0.61	90.82 ± 0.36	89.93 ± 0.21*	94.64 ± 0.11*
UNet (Baseline)	-	82.87 ± 0.04*	90.53 ± 0.02*	80.90 ± 1.01*	89.24 ± 0.62*	88.83 ± 0.32*	94.00 ± 0.18*
Morph-UNet (Ours)	MSMCM	83.35 ± 0.28	90.78 ± 0.16	82.67 ± 0.53	90.33 ± 0.33	89.60 ± 0.14	94.45 ± 0.08
	MSMOM	82.03 ± 1.00	89.93 ± 0.64	81.70 ± 0.76	89.74 ± 0.52	89.45 ± 0.17	94.37 ± 0.09
	MSMGM	83.36 ± 0.24	90.78 ± 0.14	82.22 ± 1.09	90.07 ± 0.68	89.58 ± 0.25	94.44 ± 0.15
Morph-UNet-EfficinetNetB4 (Ours)	MSMCM	81.29 ± 0.65	89.56 ± 0.45	80.50 ± 0.23	88.97 ± 0.15	89.13 ± 0.24	94.18 ± 0.14
	MSMOM	80.74 ± 1.25	89.14 ± 0.78	80.27 ± 0.19	88.79 ± 0.15	89.00 ± 0.25	94.10 ± 0.15
	MSMGM	81.37 ± 0.76	89.58 ± 0.48	81.31 ± 0.88	89.47 ± 0.57	89.42 ± 0.29	94.35 ± 0.17
Morph-UNet-ResNet34 (Ours)	MSMCM	**85.46 ± 0.03**	**92.08 ± 0.02**	**84.13 ± 1.02**	**91.23 ± 0.64**	**90.37 ± 0.17**	**94.89 ± 0.10**
	MSMOM	**85.12 ± 0.23**	**91.84 ± 0.16**	**84.03 ± 0.94**	**91.16 ± 0.56**	**90.15 ± 0.14**	**94.76 ± 0.08**
	MSMGM	**85.21 ± 0.24**	**91.92 ± 0.15**	**84.10 ± 0.74**	**91.22 ± 0.43**	**90.32 ± 0.18**	**94.85 ± 0.10**

All the models are trained and evaluated following the recommended setting of (Valanarasu and Patel, 2022).
Bold metrics signify improvement over the state-of-the-art, with red and green denoting the top two metrics, respectively.
* indicates the methods that are surpassed by the proposed method with a statistically significant margin (Wilcoxon signed-rank test with a significance level of 0.05)
In a comparison of my proposed Morph-UNet with MSMCM with the lightweight models MALUNet (Ruan et al., 2022) and UNext (Valanarasu and Patel, 2022), the former model achieved significant improvement with few more parameters and comparable inference time for all three datasets. APFormer (Lin et al., 2023) takes a larger inference time as compared to Morph-UNet. Proposed Morph-UNet-ResNet34 with MSMCM shows further improvement over TransFuse (Zhang et al., 2021c) and FAT-Net (Wu et al., 2022) with 3-4 times fewer parameters and much lesser inference time.

model, computation complexity is measured by a number of Multiply-Accumulate operations (MACs) that involve a multiplication followed by an addition. On the other hand, the time complexities of the models are measured by inference time i.e. the time taken by the model time taken by a model to process a given input image. Reported inference times are specific to the ISIC2017 dataset.

Table 4.3 provides a comprehensive comparison of the proposed methods with state-of-the-art approaches, focusing on segmentation performance measured by IoU and F1-score. A comparison of time efficiency and memory efficiency is also presented in terms of inference time and number of trainable parameters and Multiply-Accumulate Operations (MACs) respectively in Table 4.4. The inference time is the time taken to predict the segmentation mask of one image. Reported inference times are specific to the ISIC2017 dataset. I have the following observations from Table 4.3 and 4.4. While compared among different MSMMs, the integration of MSMCM and MSMGM module within the Morph-UNet led to superior performance as compared to MSMOM. Given the irregular shapes and textures often associated with skin lesions, *Closing* proves effective in ensuring the lesion is perceived as a connected region, even

Table 4.4: Complexity comparison of the proposed methods with the state-of-the-art skin lesion segmentation methods in terms of number of trainable parameters, Multiply-Accumulate Operations (MACs), and inference time for ISIC2017.

Method	**Module**	**Parameters (M)**	**MACs (G)**	**Inference Time (ms)**
MALUNet (Ruan et al., 2022)	-	00.18	00.08	3.92
UNext (Valanarasu and Patel, 2022)	-	01.47	00.50	2.33
APFormer (Lin et al., 2023)	-	02.60	01.40	6.48
TransFuse (Zhang et al., 2021c)	-	26.27	05.93	7.12
FAT-Net (Wu et al., 2022)	-	34.00	29.43	5.54
UNet (Baseline)	-	01.94	02.54	1.50
Morph-UNet (Ours)	MSMCM	02.33	02.74	3.68
	MSMOM	02.33	02.74	3.66
	MSMGM	02.29	02.74	3.01
Morph-UNet-EfficinetNetB4 (Ours)	MSMCM	00.35	00.66	5.54
	MSMOM	00.35	00.66	5.51
	MSMGM	00.34	00.66	4.73
Morph-UNet-ResNet34 (Ours)	MSMCM	09.28	02.81	4.67
	MSMOM	09.28	02.81	4.78
	MSMGM	09.24	02.81	3.92

in the presence of minor interruptions. *Closing* operation is particularly advantageous for handling texture irregularities, fostering a more uniform representation in the segmentation results. In the context of diverse textures exhibited by skin lesions, *Closing* contributes to a consistent portrayal across the entire lesion area. *Morpand hological Gradient* is sensitive to variations in texture. Skin lesions often exhibit distinct texture characteristics, and the gradient is effective in capturing these textural differences for segmentation. Leveraging the characteristics of skin lesions and the capabilities inherent in *Morphological Closing* and *Gradient* operations, my proposed model incorporating Multi-Scale Morphological Closing and Gradient Modules (MSMCM, MSMGM) has demonstrated superior performance compared to Multi-Scale Morphological opening Modules (MSMOMs). Secondly, the proposed lightweight Morph-UNet model with MSMCM demonstrated superior performance compared to both the baseline UNet model and other lightweight models, including MALUNet (Ruan et al., 2022), UNext (Valanarasu and Patel, 2022) and APFormer (Lin et al., 2023), across the ISIC2018 and HAM10000 datasets. Nevertheless, on the ISIC2017 dataset, the proposed Morph-UNet model with MSMCM surpassed all state-of-the-art lightweight segmentation models except for APFormer (Lin et al., 2023). Though the inference time associated with my proposed method is longer than the baseline method (UNet) due to the inclusion of MSMMs within the architecture, it is worth noting that it either matches or outperforms the existing state-of-the-art methods in this context except (inference time of UNext). It is worth mentioning

that, my proposed approach (Morph-UNet-ResNet3 with MSMCM) has fewer trainable parameters, MACs and has lesser inference time as compared to the TransFuse (Zhang et al., 2021c) and FAT-Net (Wu et al., 2022), the two best-performing state-of-the-art-methods, yet demonstrated superior performance when compared to other state-of-the-art methods. Thus altogether, it surpasses the state-of-the-art methods in terms of performance as well as time, space, and computational complexity, making it well-suited for practical applications in the field of skin lesion segmentation.

Table 4.5: Performance comparison of the proposed methods with the state-of-the-art medical image segmentation methods in terms of IoU and F1-score for BUSI and UDIAT datasets.

Method	Module	Params (M)/ MACs (G)	Inference Time (ms)	Data Aug.	BUSI		UDIAT	
					IoU	F1-score	IoU	F1-score
MedT, (Valanarasu and Patel, 2022)	-	00.18/00.97	52.36	No	63.89 ± 0.55*	76.93 ± 0.11*	-	-
UNext, (Valanarasu and Patel, 2022)	-	01.47/00.50	02.33	Yes	64.69 ± 1.79*	77.60 ± 1.64*	67.89 ± 1.31*	80.48 ± 1.02*
APFormer, (Lin et al., 2023)	-	02.60/01.40	06.48	Yes	67.82 ± 0.17*	80.25 ± 0.15*	64.52 ± 1.72*	77.74 ± 1.21*
TransUNet, (Valanarasu and Patel, 2022)	-	105.0/01.44	-	Yes	66.92 ± 0.75*	79.30 ± 0.37*	-	-
DPCTN, (Song et al., 2023)	-	- / -	-	Yes	57.97	70.22	-	-
AAUNet, (Chen et al., 2022a)	-	50.00/-	-	No	68.82 ± 0.44*	77.51 ± 0.68*	69.10 ± 2.98*	78.14 ± 2.41*
UNet (Baseline)	-	01.94/01.50	02.54	Yes	65.46 ± 1.12*	78.34 ± 0.66*	66.67 ± 1.23*	79.42 ± 1.02*
Morph-UNet (Ours)	MSMCM	02.33/02.74	03.68	Yes	65.84 ± 0.23	78.52 ± 0.36	64.46 ± 1.26	78.00 ± 1.01
	MSMOM	02.33/02.74	03.66	Yes	63.35 ± 0.66	76.77 ± 0.68	64.88 ± 0.80	78.06 ± 0.54
	MSMGM	02.31/02.74	03.01	Yes	68.49 ± 0.46	**80.37 ± 0.10**	**69.83 ± 1.98**	**81.67 ± 1.42**
Morph-UNet-EfficinetNetB4 (Ours)	MSMCM	00.35/00.66	05.54	Yes	56.89 ± 0.52	71.56 ± 0.59	35.60 ± 1.01	51.60 ± 1.31
	MSMOM	00.35/00.66	05.51	Yes	55.91 ± 0.94	70.77 ± 1.06	44.43 ± 0.42	60.33 ± 0.53
	MSMGM	00.34/00.66	04.73	Yes	58.28 ± 0.33	72.71 ± 0.04	42.34 ± 0.86	58.80 ± 0.80
Morph-UNet-ResNet34 (Ours)	MSMCM	09.28/02.81	04.67	Yes	**70.46 ± 0.30**	**81.72 ± 0.14**	76.01 ± 0.91	85.85 ± 0.65
	MSMOM	09.28/02.81	04.78	Yes	**70.63 ± 0.12**	**81.75 ± 0.36**	**76.04 ± 0.65**	**85.99 ± 0.46**
	MSMGM	09.25/02.81	03.92	Yes	**72.28 ± 0.27**	**83.17 ± 0.09**	**78.01 ± 1.20**	**87.24 ± 0.87**

Along with UNext (Valanarasu and Patel, 2022) and APFormer (Lin et al., 2023), all versions of the proposed models are trained and evaluated following the recommended setting of (Valanarasu and Patel, 2022). Results reported for MedT (Valanarasu et al., 2021) and TransUNet (Chen et al., 2021) are taken from (Valanarasu and Patel, 2022), whereas, performance measures of DPCTN (Song et al., 2023) and AAUNet (Chen et al., 2022a) are obtained from respective original articles.
Metrics highlighted in bold type indicate improvements over the state-of-the-art, with red and green indicating the top two metrics, respectively.
* indicates the methods that are surpassed by the proposed method with a statistically significant margin (Wilcoxon signed-rank test with a significance level of 0.05)
The proposed Morph-UNet with MSMGM has surpassed all the state-of-the-art models by a significant margin. MedT (Valanarasu et al., 2021) and APFormer (Lin et al., 2023) have significantly high inference time, whereas TransUNet (Chen et al., 2021) and AAUNet (Chen et al., 2022a) have a huge number of parameters as compared to Morph-UNet.

I draw the following observations from Table 4.5 that compares the performance of the proposed method with state-of-the-art and baseline methods in the context of breast tumor segmentation tasks. I observe that all three versions of proposed models with the MSMGM module consistently perform better than other morphological modules. This aligns with the results obtained in the ablation study (Section 4.5.1). The morphological gradient boosts the contrast between distinct regions in an image. It operates as an edge detection operator, emphasizing the transitions between tumor and non-tumor regions within the image. For breast tumor segmentation, this is valuable for improving the visibility of subtle features within

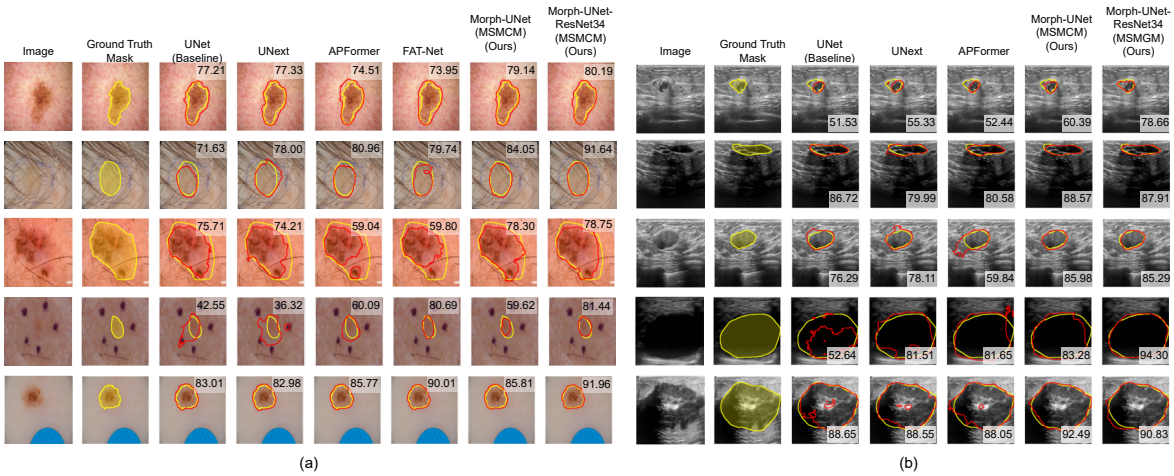

Figure 4.5: The qualitative comparison of proposed segmentation models with baseline and state-of-the-art model for (a) skin lesion segmentation and (b) breast tumor segmentation task. Yellow and red contours indicate the boundary of the ground truth mask and predicted mask respectively.

and around tumor areas, aiding in the accurate identification and delineation of tumors. Consequently, the Morph-UNet incorporating Multi-Scale Morphological Gradient Modules (MSMGM) has demonstrated commendable performance in tasks related to breast tumor segmentation task. The proposed methods, both Morph-UNet and Morph-UNet-ResNet34 with MSMGM, show significant improvement over baseline and related state-of-the-art models. The best result was obtained through the Morph-UNet-ResNet34 model with MSMGM modules, which beats the AAUNet (Chen et al., 2022a) model by 3.46% and 5.66% in terms of IoU and F1-score respectively for BUSI dataset. Similarly, I also observe an enhancement of 8.91% and 9.1% in the IOU and F1-score measure for the UDIAT dataset.

In these two tasks, both lesions and tumors are generally represented by one single continuous ROI, which has a comparatively higher ratio of ROI to the image. The main challenge here is to precisely define the segmentation boundary rather than the small areas with similar characteristics as the ROI since such small areas are usually absent in medical images of these categories. This might have led to the inferior performance of the MSMOM module as compared to other morphological modules, skin lesions, and breast tumor segmentation as morphological *Opening* operation primarily targets the removal of small objects. Moreover, the ResNet34 CNN backbone performs better than the EfficientNetB4 CNN backbone in both tasks. This demonstrates that, for skin lesions and breast tumors, the ROIs are better

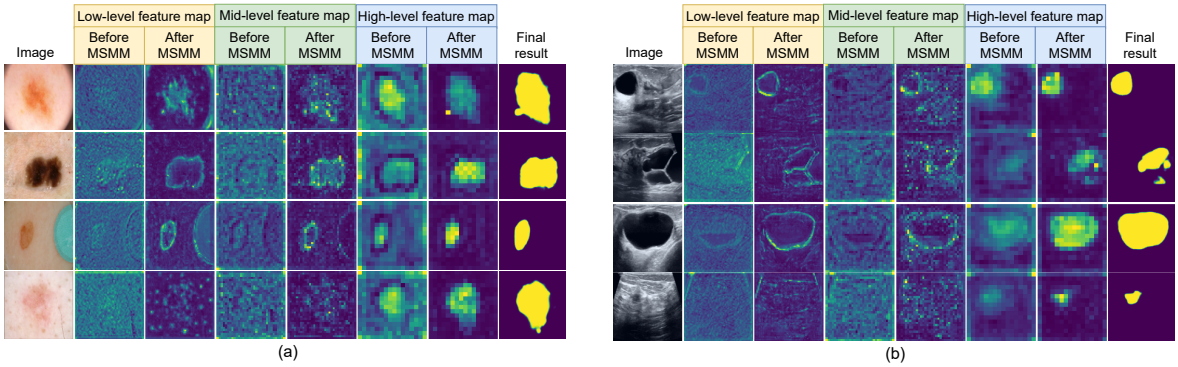

Figure 4.6: The channel-wise average of the low-level, mid-level, and high-level feature map extracted by ResNet34 backbone and their corresponding feature maps after the application of Multi-scale Morphological Modules (MSMM) shown are shown for (a) skin lesion segmentation task and (b) breast tumor segmentation task.

characterized by the comparatively larger initial kernel size of ResNet34 than by the smaller 3×3 initial kernel of EfficientNetB4.

Along with the quantitative results, I also present the qualitative results for comparison among the baseline (UNet), state-of-the-art models (UNext and APFormer and FAT-Net), and proposed models (Morph-UNet and Morph-UNet-ResNet34 with MSMCM and MSMGM modules respectively for skin lesion segmentation and breast tumor segmentation task respectively). Figure 4.5 (a) and (b) show a few samples from the datasets used for skin lesion and breast tumor segmentation tasks, respectively, along with their ground truth masks and predicted masks obtained by various methods. It includes images with different ratios of ROIs to images and noises, namely reference stickers, hair, and pen markings. The proposed methods delineate the boundary of skin lesions and breast tumors with more precision as compared to baseline and state-of-the-art models.

In Figure 4.6, I present the feature maps obtained in three stages of the CNN backbone along with the corresponding feature maps after applying morphological modules. I consider Morph-UNet-ResNet34 with MSMCM and MSMGM modules for this analysis in skin lesion and breast tumor segmentation tasks, respectively, which yield the best performance in the corresponding domain. Incorporating skip connections from all stages is pivotal in intricate image segmentation tasks (Ronneberger et al., 2015). As I progress from low-level to high-level feature maps, there is a trade-off between spatial resolution and enriched abstract and semantic information. From Figure 4.6, it is apparent that the application of morphological

Table 4.6: Performance comparison of the proposed methods with the state-of-the-art medical image segmentation methods in terms of IoU and F1-score for GlaS and MonuSeg dataset.

Method	Module	Param.	Inference Time (ms)	Data Aug.	GlaS		MonuSeg	
					IoU	F1-score	IoU	F1-score
MedT (Valanarasu et al., 2021)	-	01.40 M	52.36	No	69.61	81.02	66.17	79.55
HistoSeg (Wazir and Fraz, 2022)	-	29.00 M	-	Yes	76.73	98.07	71.06	75.08
DCA (Ates et al., 2023)	-	30.68 M	-	Yes	81.68	89.90	65.97	79.50
UCtransNet (Wang et al., 2022)	-	65.60 M	05.94	Yes	82.24	89.84	66.68	79.87
UNet(Baseline)	-	01.94 M	01.50	Yes	81.86 ± 0.27	89.54 ± 0.23	65.08 ± 0.39	78.60 ± 0.29
Morph-UNet (Ours)	MSMCM	02.33 M	03.68	Yes	**83.25 ± 0.11**	90.50 ± 0.10	65.22 ± 0.15	78.68 ± 0.12
	MSMOM	02.33 M	03.66	Yes	81.64 ± 0.65	89.41 ± 0.47	67.11 ± 0.53	**80.05 ± 0.41**
	MSMGM	02.31 M	03.01	Yes	**82.68 ± 0.24**	90.08 ± 0.12	65.61 ± 0.42	79.00 ± 0.32
Morph-UNet-EfficinetNetB4 (Ours)	MSMCM	03.48 M	05.54	Yes	**83.27 ± 0.58**	90.49 ± 0.37	67.52 ± 0.27	**80.37 ± 0.22**
	MSMOM	03.48 M	05.51	Yes	**83.16 ± 0.47**	90.43 ± 0.29	68.17 ± 0.36	**80.86 ± 0.28**
	MSMGM	03.42 M	04.73	Yes	**83.47 ± 0.72**	90.50 ± 0.52	67.64 ± 0.25	**80.48 ± 0.18**
Morph-UNet-ResNet34 (Ours)	MSMCM	09.28 M	04.67	Yes	**84.64 ± 0.10**	91.31 ± 0.05	66.04 ± 0.36	79.30 ± 0.27
	MSMOM	09.28 M	04.78	Yes	84.12 ± 0.31	91.00 ± 0.22	66.31 ± 0.16	79.50 ± 0.12
	MSMGM	09.25 M	03.92	Yes	**85.12 ± 0.21**	91.63 ± 0.14	66.16 ± 0.15	79.39 ± 0.12

Bold metrics signify improvement over the state-of-the-art, with red and green denoting the top two metrics, respectively. All three versions of the proposed models (Morph-UNet, Morph-UNet-EfficinetNetB4, and Morph-UNet-ResNet34) are much lighter than HistoSeg, DCA, and UCTransNet and take less time to infer as compared to MedT.

modules at each stage has enhanced the morphological information of the ROI, which plays a crucial role in predicting the mask indicating the lesion or tumor in the clinical images.

4.5.3 Performance on Glands and Nuclei Segmentation

In medical pathology, gland segmentation is crucial for the analysis of tissue samples. It aids pathologists in studying the morphology and distribution of glands within tissues, contributing to diagnosing diseases such as cancer. Additionally, cell nuclei, serving as indicators of diverse cellular functions and diseases, hold significant importance. The segmentation of cell nuclei extracted from microscopy images is crucial for quantitatively analyzing cellular processes and related medical conditions. I have evaluated the performance of the proposed model in the context of segmenting glands and nuclei from microscopic histopathology images. The task of gland and nuclei segmentation differs from the segmentation tasks explained in the previous section in terms of the number of disjoint ROIs. The number of disjoint ROIs in skin lesion or breast tumor segmentation tasks is generally one, whereas, gland and nuclei segmentation tasks involve segmenting a large number of disjoint ROIs. Moreover, the nuclei segmentation task is more intricate due to the comparatively smaller areas of each disjoint ROI. These characteristics of microscopic cellular images challenge the model's capability to segment the region of interest.

In Table 4.6, I report the quantitative analysis of the performance of the proposed method

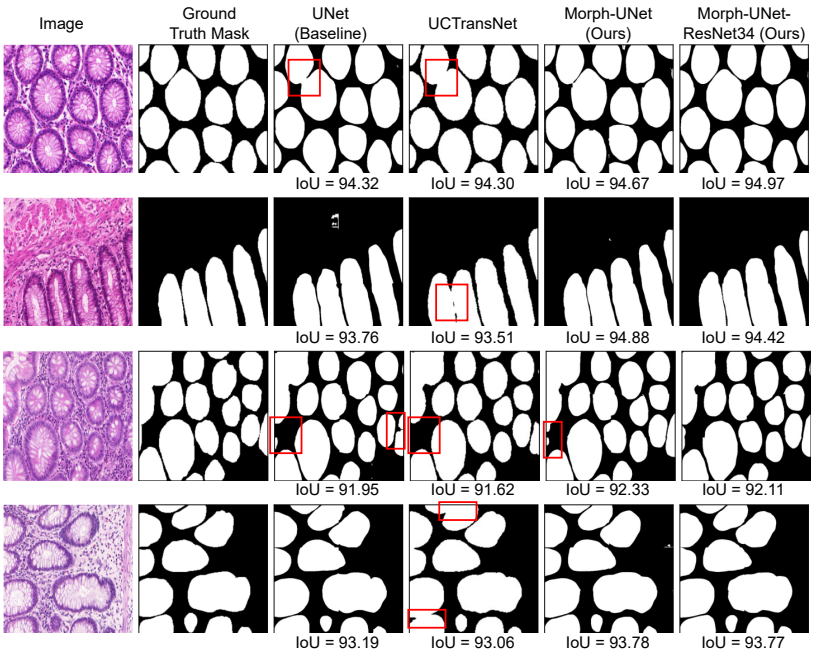

Figure 4.7: The qualitative comparison of proposed segmentation models with baseline and state-of-the-art model for GlaS dataset.

along with the baseline and other medical image segmentation models in extracting nuclei and glands from microscopic images using the MonuSeg and GlaS datasets, respectively. The proposed Morph-UNet has surpassed UNet, the baseline model, and state-of-the-art methods in gland segmentation tasks in terms of IoU metric. Among the MSMMs, MSMGM and MSMCM have performed better than MSMOM in all three versions of the proposed Morph-UNet. The effectiveness of *Morphological Closing* and *Gradient* operations in gland segmentation tasks can be attributed to their ability to capture specific morphological features and enhance the visibility of glandular structures. *Morphological Closing*, which involves the sequential application of dilation followed by erosion, is effective in smoothing irregularities, *Closing* small gaps in the boundaries of glandular structures and minimizing noise and artifacts. Thus, MSMCM ensures that the segmentation algorithm focuses on the actual glandular structures, leading to more reliable results. The morphological gradient operation involves computing the intensity differences between pixels, emphasizing regions of rapid intensity change. In the context of gland segmentation, it highlights boundaries and fine details within the glandular structures. Thus, MSMGM enhances the contrast between glandular and non-glandular regions, making it easier for segmentation algorithms to identify

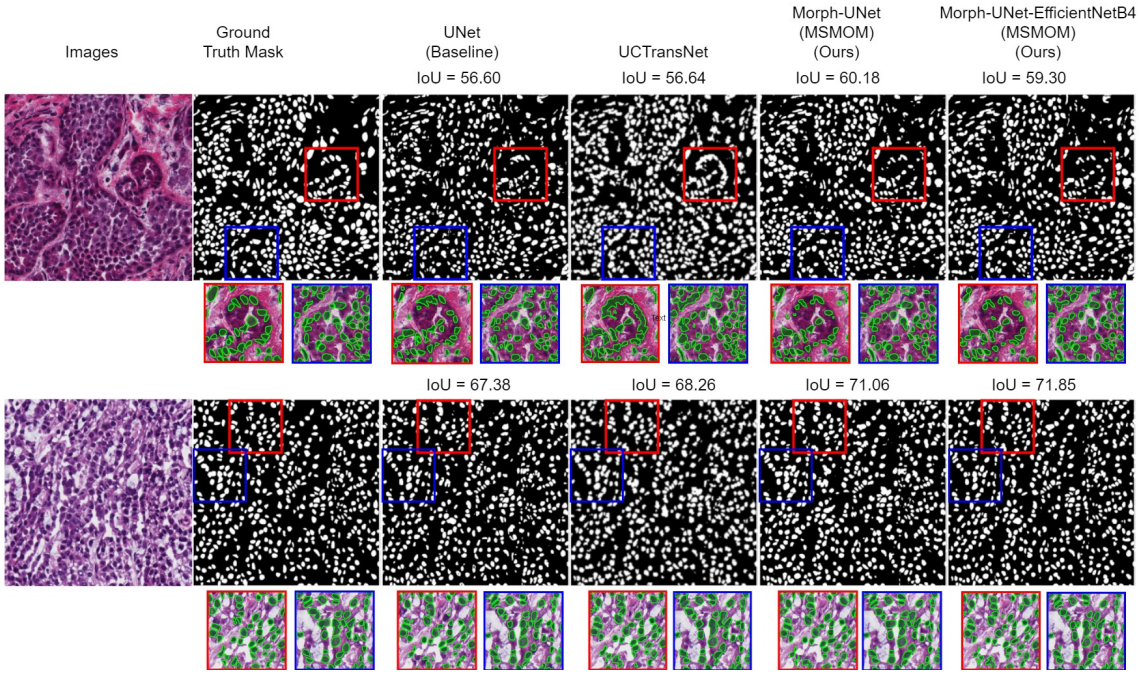

Figure 4.8: The qualitative comparison of proposed segmentation models with baseline and state-of-the-art model for the MonuSeg dataset.

and delineate gland boundaries accurately. Furthermore, among the two CNN backbones that have been fused with the proposed architecture, ResNet34 performed better in the context of gland segmentation tasks due to the relatively large RRI, which is more accurately accommodated by the larger receptive field of the initial convolutional layer of ResNet34. Morph-Unet-ResNet34 architecture with MSMGM has shown an improvement of 2.88% in IoU with respect to the state-of-the-art method, even with a significantly low number of parameters. Figure 4.7 illustrates the qualitative comparison of the proposed models with both baseline (UNet) and state-of-the-art methods (UCTransNet). While UNet and UCTransNet fail to separate closely bound glands, the Morph-UNet and Morph-UNet-ResNet34 with MSMGM precisely delineate the separating boundary.

Among the state-of-the-art methods, UCTransNet (Wang et al., 2022) achieved the best F1-score in nuclei segmentation task using the MonuSeg dataset. Thus, I have followed the same train, validation, and test split as mentioned in (Wang et al., 2022). While comparing the different morphological modules, I note that the MSMOM module performs better than the other morphological modules in this context. In terms of F1-score, the proposed lightweight model, Morph-UNet has surpassed all the state-of-the-art methods, whereas it

Table 4.7: Performance comparison of the proposed methods with the state-of-the-art medical image segmentation methods in terms of F1-score for the PanNuke dataset.

Model	Module	Param.	Neoplastic	Epithelial	Inflammatory	Connective	Dead	Average
Hover-Net (Graham et al., 2019)	-	45.00 M	0.55	0.49	0.42	0.39	0.14	0.40
TSFD-Net (Ilyas et al., 2022)	-	21.90 M	0.65	0.57	0.57	0.53	0.43	0.55
MCTrans (Ji et al., 2021)	-	07.64 M	0.84	0.78	0.68	0.65	0.46	0.68
SONNET (Doan et al., 2022)	-	-	0.64	0.60	0.52	0.47	0.37	0.52
UNet	-	01.94 M	0.60±0.012	0.66±0.112	0.46±0.009	0.54±0.007	0.97±0.006	0.65±0.023
Morph-UNet	MSMCM	02.33 M	0.60±0.031	0.77±0.041	0.46±0.007	0.54±0.025	0.96±0.011	0.67±0.016
	MSMOM	02.33 M	0.60±0.031	**0.79±0.008**	0.45±0.011	0.55±0.011	0.96±0.013	0.67±0.011
	MSMGM	02.31 M	0.64±0.012	0.72±0.026	0.46±0.008	0.54±0.035	0.97±0.006	0.66±0.007
Morph-UNet-EfficientNetB4	MSMCM	03.48 M	0.68±0.039	**0.87±0.017**	0.48±0.024	0.55±0.030	0.90±0.034	**0.70±0.018**
	MSMOM	03.48 M	0.72±0.004	**0.85±0.009**	0.49±0.010	0.58±0.003	0.96±0.007	**0.72±0.003**
	MSMGM	03.42 M	0.71±0.019	**0.85±0.014**	0.48±0.009	0.56±0.008	0.95±0.009	**0.71±0.003**
Morph-UNet-ResNet34	MSMCM	09.28 M	0.66±0.033	**0.81±0.014**	0.47±0.006	0.55±0.005	0.91±0.049	0.68±0.021
	MSMOM	09.28 M	0.65±0.034	**0.81±0.033**	0.47±0.017	0.56±0.012	0.94±0.028	**0.69±0.009**
	MSMGM	09.25 M	0.65±0.021	0.76±0.046	0.47±0.005	0.56±0.008	0.95±0.010	0.68±0.014

Bold metrics signify improvement over the state-of-the-art, with red and green denoting the top two metrics, respectively. Morph-UNet-EfficinetNetB4 is much lighter than Hover-Net, TSFD-Net, and MCTrans and shows significant improvement over the state-of-the-art model with MSMOM.

yields the second-best performance when evaluated using the IoU metric. Furthermore, the Morph-UNet-EfficientNetB4 has outperformed the UCtransNet (Wang et al., 2022) model by 0.99% in the F1-score. The EfficientNetB4 backbone has shown better performance than the ResNet34 backbone. Due to the larger receptive field in the initial convolution layer of ResNet34, it lacks precision in segmenting comparatively small cell nuclei. The qualitative comparison of the proposed models with baseline as well as state-of-the-art methods is also presented in Figure 4.8. It is observed that the proposed method performed better in preserving the shapes of ROIs as well as detecting disjoint ROIs as compared to UNet and UCTransNet, the baseline and state-of-the-art methods respectively.

All the segmentation tasks used to evaluate the proposed method so far are binary, i.e., there are only two discrete categories (ROI and background) in which the pixels are grouped. However, the ROIs of the PanNuke dataset are divided into five categories according to the cell types. Thus, the task here is to segment the nuclei in the microscopic images and identify the type of its cell, which further challenges the medical segmentation models. In Table 4.7, I compare the quantitative results obtained by the proposed method with the other segmentation method. Focal loss (Lin et al., 2017) with $\gamma = 0.75$ is used as the loss function, which is optimized in the training process. Table 4.7 presents the quantitative analysis of the proposed methods, baseline, and state-of-the-art models in terms of individual classwise F1-score and average F1-score. In average F1-score, Morph-UNet-EfficientNetB4 outperforms both baseline and state-of-the-art methods by 7% and 4%, respectively, while utilizing a

notably lower number of trainable parameters. However, while considering the F1-score for individual classes, the proposed method has surpassed all the state-of-the-art models except MCTrans (Ji et al., 2021). The EfficientNet backbone yields improved performance as compared to the ResNet34 backbone. The features of small-sized cell nuclei are more precisely captured by the smaller receptive field of the initial convolution operation of EfficientNetB4 compared to the ResNet34 model. Among different morphological modules, MSMOM has consistently exhibited the best performance in nuclei segmentation tasks. The *Morphological Opening* operation entails the sequential application of erosion followed by dilation, a process designed to uphold object boundaries while eliminating smaller, surrounding structures. Its efficacy extends to the separation of touching objects by eroding the regions where they converge and subsequently dilating them back to their original size, facilitating more precise segmentation of individual nuclei. Given the typically well-defined boundaries of nuclei, the use of *Opening* proves advantageous in preserving their integrity while simultaneously removing background noise or artifacts. Thus, this is particularly beneficial in addressing common challenges such as touching nuclei in histological images, by adeptly filtering out small-scale noise and extraneous details in the background, *Opening* ensures the maintenance of the overall structural fidelity of the nuclei. Following an elaborate experimentation process, my findings demonstrate that the proposed architecture with Multi-Scale Morphological Opening Modules (MSMOM) outperformed MSMGM or MSMCM in the nuclei segmentation task, yielding superior results for both the MonuSeg and PanNuke datasets.

4.5.4 MSMM in Decoder

All three versions of the proposed architecture incorporate MSMM blocks on the feature maps extracted by the encoder. I conduct another experiment to investigate the performance of those architectures while the MSMM modules are included in the decoder path after each upsampling operation. In Table 4.8, I showcase the evaluation performance of the proposed model with the above-said modification on all seven medical imaging datasets. Among all the variations of the proposed method, I specifically focus on the variant that demonstrated the most favorable outcomes, both with and without any CNN backbone. Additionally, I integrate the aforementioned modification into that architecture. It is found that, though the above modification yielded an improved performance as compared to the corresponding

Table 4.8: Performance of proposed method with MSMM blocks incorporated in the decoder path. E-D skip path indicates Encoder-Decoder skip path.

Dataset	Backbone	Module, position	IoU	F1-score
ISIC2017	No Backbone	MSMCM, E-D skip path	83.35 ± 0.28	90.78 ± 0.16
	No Backbone	MSMCM, decoder	**83.40 ± 0.44**	**90.82 ± 0.26**
	ResNet34	MSMCM, E-D skip path	**85.46 ± 0.03**	**92.08 ± 0.02**
	ResNet34	MSMCM, decoder	85.22 ± 0.29	91.92 ± 0.19
ISIC2018	No Backbone	MSMCM, E-D skip path	**82.67 ± 0.53**	**90.33 ± 0.33**
	No Backbone	MSMCM, decoder	82.05 ± 0.68	89.95 ± 0.43
	ResNet34	MSMCM, E-D skip path	**84.13 ± 1.02**	**91.23 ± 0.64**
	ResNet34	MSMCM, decoder	84.03 ± 0.76	91.18 ± 0.45
HAM10000	No Backbone	MSMCM, E-D skip path	**89.60 ± 0.14**	**94.45 ± 0.08**
	No Backbone	MSMCM, decoder	89.41 ± 0.32	94.34 ± 0.18
	ResNet34	MSMCM, E-D skip path	**90.37 ± 0.17**	**94.89 ± 0.10**
	ResNet34	MSMCM, decoder	90.21 ± 0.28	94.79 ± 0.16
BUSI	No Backbone	MSMGM, E-D skip path	**68.49 ± 0.46**	**80.37 ± 0.10**
	No Backbone	MSMGM, decoder	66.55 ± 2.18	79.18 ± 1.40
	ResNet34	MSMGM, E-D skip path	**72.28 ± 0.27**	**83.17 ± 0.09**
	ResNet34	MSMGM, decoder	71.51 ± 1.09	82.49 ± 0.32
MonuSeg	No Backbone	MSMOM, E-D skip path	**67.11 ± 0.53**	**80.05 ± 0.41**
	No Backbone	MSMOM, decoder	66.73 ± 0.00	79.73 ± 0.00
	EfficientNetB4	MSMOM, E-D skip path	**68.17 ± 0.36**	**80.86 ± 0.28**
	EfficientNetB4	MSMOM, decoder	66.07 ± 0.26	79.30 ± 0.20
GlaS	No Backbone	MSMCM, E-D skip path	**83.25 ± 0.11**	**90.50 ± 0.10**
	No Backbone	MSMCM, decoder	83.06 ± 0.35	90.32 ± 0.23
	ResNet34	MSMGM, E-D skip path	**85.12 ± 0.21**	**91.63 ± 0.14**
	ResNet34	MSMGM, decoder	84.20 ± 0.07	91.01 ± 0.05
PanNuke	No Backbone	MSMCM, E-D skip path	**62.66 ± 1.07**	**67.09 ± 1.09**
	No Backbone	MSMCM, decoder	61.43 ± 3.57	65.98 ± 3.71
	EfficientNetB4	MSMOM, E-D skip path	**67.20 ± 0.29**	**71.82 ± 2.64**
	EfficientNetB4	MSMOM, decoder	64.58 ± 1.47	69.09 ± 1.50

state-of-the-art method for a few datasets, the incorporation of MSMM modules within the decoder did not show any improvement over the proposed design. I can infer from this experiment that the proposed position of MSMM blocks within the encoder path of Morph-UNet is the ideal position in the context of medical image segmentation.

4.6 Discussion

I introduced a lightweight network designed for the specific task of medical image segmentation, particularly suited for scenarios where the subtle regions of interest (ROIs) exhibit irregular shapes and sizes. Three separate multi-scale morphological modules are developed based on three morphological operations — *Opening*, *Closing*, and *Morphological Gradient*. These

modules were designed to capture morphological details from regions of interest (ROIs), which are crucial in segmenting these ROIs. The suggested modules can be seamlessly integrated into various deep learning-based segmentation models to improve performance. Investigating this potential enhancement is a topic for future research. To validate the effectiveness of my approach, I applied it to three distinct medical image segmentation tasks – skin lesion segmentation, breast tumor segmentation, and cell nuclei segmentation. I conducted experiments using eight publicly available datasets to demonstrate the versatility and robustness of my proposed method. The performance and quality of my proposed models were evaluated using both quantitative metrics and qualitative assessments. Through this comprehensive analysis, I effectively highlighted the superiority of my suggested models in addressing these complex medical image segmentation challenges with a limited number of training samples. The effectiveness of the proposed models can be further investigated in other medical image segmentation tasks as a potential avenue for future research.

The proposed multi-scale morphological modules with trainable structuring elements can be plugged into any other segmentation model. To this extent, the proposed module could be used as the post-processing algorithm to enhance the precise delineation of the boundary after the cluster-based segmentation algorithms (Kumar et al., 2022, 2023).

The performance of Morph-UNet with specific MSMM can vary based on the specific segmentation task and the characteristics of the images. It may not offer universal optimization across all image types or objects, necessitating thoughtful consideration of the particular task. Proposed approach involved individually incorporating three morphological operations (*Opening*, *Closing*, and *Morphological Gradient*) into MSMM during its design. However, there is an opportunity to explore the inclusion of other morphological operations, such as top-hat, bottom-hat, etc, in MSMM. The exploration must aim to assess their potential in diverse medical image segmentation tasks by analyzing the intrinsic characteristics of the respective datasets.

Recently, the Segment Anything Model (SAM) (Kirillov et al., 2023)) has been developed to perform segmentation tasks across a wide variety of images using prompt-based inputs, such as points, boxes, or masks. While SAM has shown promise in medical image segmentation (Ma et al., 2024; Mazurowski et al., 2023; Huang et al., 2024), I have chosen not to consider this approach in this thesis for several reasons. First, MedSAM, which builds on

SAM's framework, relies on large-scale pretraining and prompt-based segmentation. However, its high computational complexity, large parameter count (over 200 million parameters), and dependence on user inputs make it less practical for resource-constrained clinical settings.

In contrast, my proposed segmentation model, Morph-UNet, is explicitly designed to address the unique challenges of medical image segmentation, such as irregular ROI shapes, limited dataset sizes, and the need for fully automated processing. Unlike MedSAM, which requires prompt-based user inputs, Morph-UNet operates autonomously, aligning with the central goal of this research. Furthermore, the small medical datasets used in this thesis are unlikely to benefit from MedSAM's large-scale pretraining requirements, which demand extensive labeled data, and fine-tuning such models with a large number of parameters on these limited datasets increases the risk of overfitting. My models, on the other hand, are specifically designed and optimized to excel in a data-scarce environment, addressing one of the most pressing challenges in medical image analysis.

Chapter 5

MA-DTNet: Multi-task Learning with Morphological Attention for Medical Image Analysis

Summary

In medical image analysis, combining classification and segmentation tasks into a unified framework enhances both performance and efficiency. Multi-task learning (MTL) approaches enable shared representations and contextual information across tasks, leading to mutual benefits in medical imaging where both pixel-level segmentation and image-level classification are essential. To this end, I propose a lightweight MTL framework, MA-DTNet, designed to simultaneously handle segmentation and classification tasks using a shared encoder architecture. MA-DTNet integrates a Spatial Morphological Attention (SMA) module and a Channel Attention (CA) mechanism to refine key morphological features and emphasize informative channels within the shared feature representations, improving task performance with fewer parameters. Specifically, in the context of breast tumor segmentation and malignancy detection, MA-DTNet demonstrates superior performance on two public breast ultrasound datasets. It achieves a 3.28% improvement in dice score for segmentation and a 1.05% increase in F1-score for classification on the UDIAT dataset, and a 1.62% and 4.96% improvement in dice score and accuracy, respectively, on the BUSI dataset. MA-DTNet's efficiency, with significantly fewer trainable parameters, highlights its potential for real-world clinical applications. The model's generalization ability is further demonstrated on additional tasks, including gland segmentation and classification in histology images, and skin lesion analysis in dermoscopic

images.

5.1 Introduction

5.1.1 Overview

Medical image analysis underpins various healthcare applications, including diagnosis, treatment planning, and prognosis. Accurate segmentation and classification of anatomical structures and abnormalities within these images are essential for practical analysis. These tasks have been traditionally addressed using separate models trained for each specific task. However, this approach can be inefficient and resource-intensive, mainly when dealing with limited datasets or complex tasks requiring extensive training data. The emergence of multi-task learning (MTL) has revolutionized the field of medical image analysis, offering a compelling alternative to single-task learning approaches. MTL leverages the inherent correlations and shared information between related tasks within a single model, leading to several potential benefits like improved data efficiency (Zhang and Yang, 2021), reduced overfitting (Ruder, 2017), enhanced feature learning (Ruder, 2017), among others.

Successful applications of MTL in medical image analysis include glioma segmentation and isocitrate dehydrogenase genotyping from brain MRI (Cheng et al., 2022), skin lesion segmentation and classification from dermoscopic images (Xie et al., 2020b), and kidney segmentation and domain translation from urographic images (Zeng et al., 2021). Building upon this growing body of research, I focus on the tasks of breast tumor segmentation and malignancy detection, as documented in recent studies highlighting the effectiveness of MTL for these specific tasks (Zhang et al., 2021a; Chowdary et al., 2022; Xu et al., 2022, 2023). These MTL networks comprise trainable parameters in the range of 90-110 M. However, MTL models with fewer parameters are highly desirable for practical implementation in real-world settings. I design a lighter MTL architecture of 22.95 M parameters to address this challenge.

I focus on breast tumor segmentation and malignancy detection in ultrasound images as my primary task in this study. Ultrasound imaging typically identifies cancerous tumors as hypoechoic regions with poorly defined borders (Gokhale, 2009). Consequently, more accurate tumor segmentation improves tumor-type diagnosis. Through experimentation, I found that the encoder-shared multi-task learning (MTL) model outperforms individual models de-

signed separately for segmentation and classification. However, there is potential to further enhance segmentation performance and, by extension, malignancy detection. To address this, I propose a Channel and Spatial Morphological Attention Module (C-SMA), which integrates channel attention (CA) to prioritize crucial feature maps and spatial morphological attention (SMA) to focus on morphological attributes within the feature maps. While the successful integration of morphological operations in deep learning has been reported for tasks like image de-raining and image restoration (Mondal et al., 2019), this work is the first, to my knowledge, to incorporate a trainable, multi-scale morphological operation in the form of morphological attention within an MTL network. My MTL architecture integrates the C-SMA block into the skip connections between the encoder and decoder at various stages. Although I do not directly use the segmentation outcome to detect tumor malignancy, the encoder shared by both segmentation and classification branches is optimized to improve segmentation results, thereby enhancing malignancy detection. The contributions of this study are as follows:

- I propose a novel C-SMA module that integrates channel attention with morphological operations. The spatial morphological operations with trainable, multi-scale structuring elements effectively highlight the morphological attributes of feature maps, allowing for the identification of regions of interest (ROIs) with variable shapes and sizes.

- I propose a lightweight multi-task learning network for breast tumor segmentation and malignancy detection from ultrasound images by incorporating the C-SMA module within the multi-task framework.

- To evaluate efficiency, I employed the proposed method on public datasets (UDIAT and BUSI), outperforming both single-task and multi-task baselines and SOTA methods. I achieve 3.28% and 1.62% improvement in dice score in tumor segmentation tasks for UDIAT and BUSI datasets, respectively. In the classification task, an enhancement of 1.05% in the F1-score and 4.96% in accuracy is observed for the UDIAT and BUSI datasets, respectively.

- The Generalization ability of my method is extended to two other multi-tasking scenarios: segmenting and classifying skin lesions in dermoscopic images and segmenting and predicting malignancy of glands in histology images.

The rest of the chapter is organized as follows. Section 5.2 describes the existing literature concerning relevant MTL and their limitations. Section 5.3 elaborates on the proposed MA-DTNet and its components. Section 5.5 presents the experimental result, description of the datasets, and training protocol followed in this study. Lastly, section 5.6 concludes the study.

5.2 Related Studies

In this section, I delve into the existing literature concerning pertinent segmentation, classification, and multi-task learning for breast tumor segmentation and malignancy detection from ultrasound images, aiming to identify the gaps and limitations in current methodologies that my research seeks to address.

In 2021, Zhang *et al.* (Zhang et al., 2021a) devised an integrated segmentation and classification network, incorporating attention gates to utilize information from lesion regions effectively. However, this approach fails to address the computational cost of the model, which is a critical factor for real-world applications, especially in clinical settings where resources may be limited. This highlights the need to develop more computationally efficient models.

In 2022, Xu *et al.* (Xu et al., 2022) introduced an MTL framework for segmenting breast ultrasound tumors and predicting their malignancy. This approach leverages segmentation outcomes as prior knowledge to enhance contextual relationships. Later, in 2023, Xu *et al.* (Xu et al., 2023) introduced a regional-attentive multi-task learning framework by integrating a regional attention (RA) module. This incorporation enhances representation, improving performance in segmentation and classification tasks for each breast ultrasound image. Despite these advancements, the models used by Xu *et al.* employed a self-attention mechanism with a large number of trainable parameters (109 M and 93 M). This makes them less suitable for real-time applications due to their computational demands. Existing research shows a key challenge: developing lightweight and efficient MTL models that maintain performance for real-world clinical use. To address this, I propose a more computationally efficient MTL network with fewer parameters, making it suitable for real-time applications.

Moreover, while existing methods have integrated various attention mechanisms within segmentation and MTL networks, none have explored the use of trainable morphological operations with adaptive structuring elements. To my knowledge, this is the first attempt

to integrate such operations within MTL networks. The novel C-SMA mechanism aims to enhance both segmentation and classification tasks by focusing on important channels and spatial locations in the feature maps based on their morphological attributes while maintaining a lightweight model architecture.

By addressing the computational inefficiencies and introducing a novel attention mechanism, my work fills a critical gap in the existing literature, contributing to the development of more practical and effective MTL models for breast tumor segmentation and malignancy detection from ultrasound images.

5.3 Proposed method

Spatial Morphological Attention (SMA):

I introduce a Spatial Morphological Attention (SMA) module designed to emphasize the spatial pixel position in the feature maps according to the morphological characteristics. This module performs two types of morphological operations — dilation (D) and erosion(E) on the inputs feature map X, described by the equation 5.2 and 5.1 respectively.

$$\mathrm{E}_{(i,j,k)}(X) = \bigparallel_{c=1}^{C} \min_{m,n=0,1,\ldots K_{SE}} X(i-m, j-n, c) + W_{SE}(m,n,c), \tag{5.1}$$

$$\mathrm{D}_{(i,j,k)}(X) = \bigparallel_{c=1}^{C} \max_{m,n=0,1,\ldots K_{\mathrm{SE}}} X(i-m, j-n, c) + W_{\mathrm{SE}}(m,n,c), \tag{5.2}$$

where $i = 1, 2, 3, \ldots, H$ and $j = 1, 2, 3, \ldots, W$. Here, X represents the input feature map, and W_{SE} denotes the structuring element of size $K_{\mathrm{SE}} \times K_{\mathrm{SE}}$ that characterizes the pattern of interest within the provided feature map X. The traditional morphological operations have two limitations. Firstly, structuring elements of a particular size is insufficient to capture the morphological characteristics of the ROIs of different sizes. I have considered structuring elements of three different sizes to overcome this drawback. This is achieved by incorporating dilation rates within the structuring elements to achieve a larger receptive field with the same computational memory. The dilated erosion and dilation operation can be defined by the

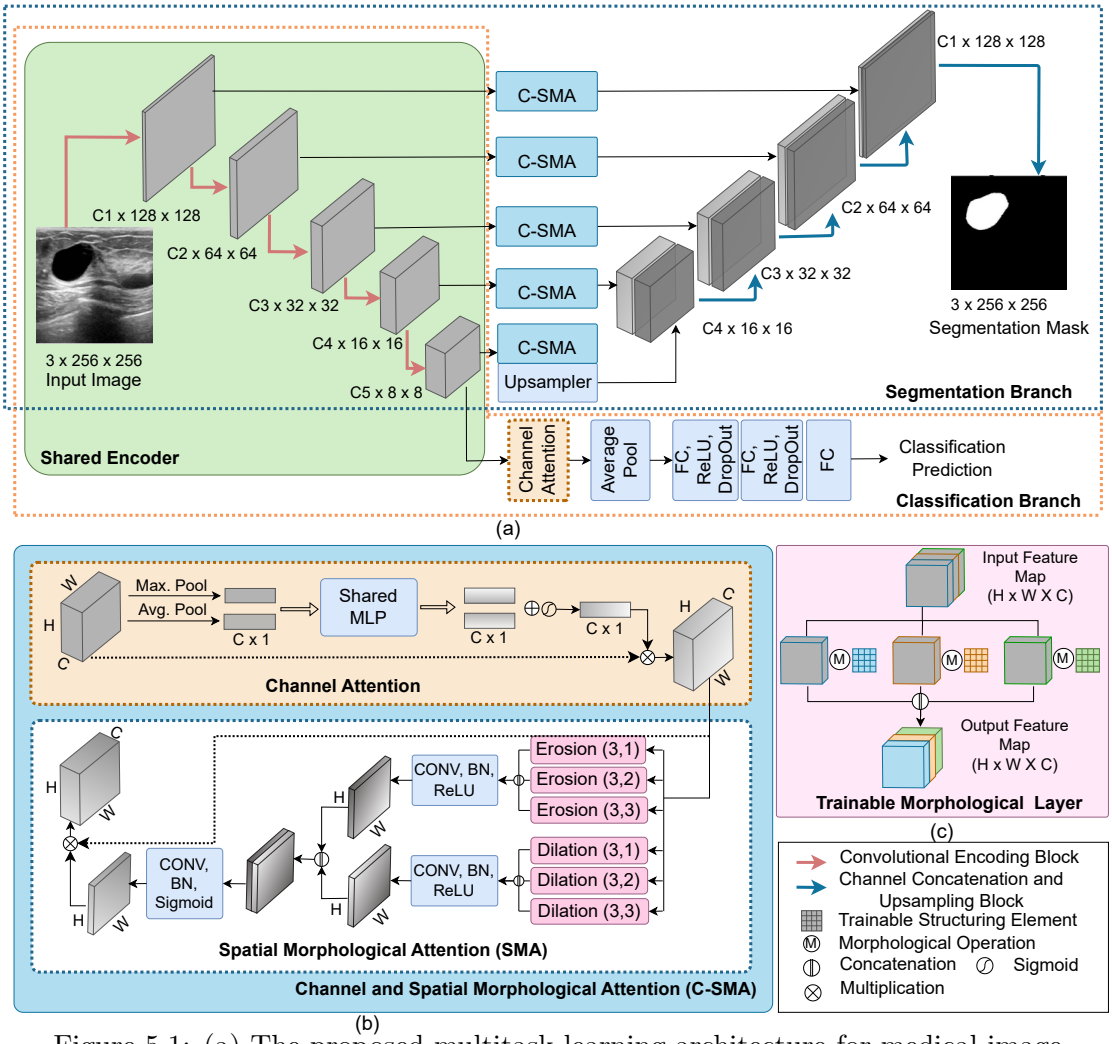

Figure 5.1: (a) The proposed multitask learning architecture for medical image segmentation and classification, (b) The channel and spatial morphological attention module (C-SMA).

equation 5.4 and 5.3, respectively.

$$\mathrm{DE}^{r}_{(i,j,k)}(X) = \Big\Vert_{c=1}^{C} \min_{m,n=0,1,\ldots K_{\mathrm{SE}}} X(i-rm, j-rn, c) + W_{\mathrm{SE}}(m,n,c), \tag{5.3}$$

$$\mathrm{DD}^{r}_{(i,j,k)}(X) = \Big\Vert_{c=1}^{C} \max_{m,n=0,1,\ldots K_{\mathrm{SE}}} X(i-rm, j-rn, c) + W_{\mathrm{SE}}(m,n,c), \tag{5.4}$$

where r is the dilation rate. Secondly, a predefined structuring element is necessary to perform traditional morphological operations, which are unsuitable for precisely segmenting

tumours of various shapes. Unlike traditional morphological operations where the structuring elements are predefined, I learn the structuring elements via backpropagation while training (Roy et al., 2021).

For a particular feature map, each of two morphological operations, i.e., dilated erosion and dilated dilation operations, are carried out with three sizes of structuring elements implemented using dilated structuring elements, thus consequently capturing morphological features at three different scales. The three resulting feature maps from morphological operations are merged using a convolutional block that includes convolution, batch normalization, and a ReLU activation function. These two resulting feature maps, each corresponding to one of the two morphological operations, are combined and fed into another convolutional block, generating a weight map (W_{SMA}) that estimates the weights for every pixel position.

$$W_{\text{SMA}} = \sigma(\text{BN}(\text{Conv}(\|(X_{\text{DE}}, X_{\text{DD}})))), \tag{5.5}$$

where,

$$X_{\text{DE}} = \text{ReLU}(\text{BN}(\text{Conv}(\|_{r=1,2,3} \text{DE}^r(X)))), \tag{5.6}$$

$$\textit{and} \quad X_{\text{DD}} = \text{ReLU}(\text{BN}(\text{Conv}(\|_{r=1,2,3} \text{DD}^r(X)))). \tag{5.7}$$

The original feature map multiplied by the weight maps results in the output of the proposed SMAM module (Equation 5.8).

$$X' = W_{\text{SMA}} \odot X, \tag{5.8}$$

where $\odot$ denotes element-wise multiplication. Thus, spatial attention is computed by the morphological characteristics of the feature map, which is learned via training.

Channel Attention(CA):

Given an input feature map X with dimensions $H \times W \times C$, where C is the number of channels and H and W are the spatial dimensions, the channel attention mechanism computes attention weights W_{CA} as follows (Woo et al., 2018).

$$W_{CA} = \sigma(\text{MLP}(X_{avg}) + \text{MLP}(X_{max})), \tag{5.9}$$

where X_{avg} and X_{max} are feature descriptors obtained by performing average-pooling and max-pooling operations on input X. MLP comprises two fully connected layers with ReLU activation function and are shared by both parallel paths. MLP squeezes pooled feature descriptor of dimension $C \times 1 \times 1$ to $C/\rho \times 1 \times 1$, with a reduction ratio of ρ and again excites it back to $C \times 1 \times 1$. The sigmoid function (σ) normalizes the attention scores across channels, allowing the model to selectively amplify or suppress specific channels based on their importance for the task. Finally, these attention weights are applied element-wise to the input feature map X to obtain the attended feature map X'.

$$X' = W_{\text{CA}} \odot X, \tag{5.10}$$

Channel and Spatial Morphological Attention (C-SMA):

In line with the inspiration drawn from the Convolutional Block Attention Module (CBAM) (Woo et al., 2018), I systematically integrate a channel attention module with the Spatial Morphological Attention Module to enhance overall performance as channel attention leverages the association among the channels of the feature map. I denote this combined attention as channel and spatial morphological attention (C-SMA).

$$X_{C-SMA} = W_{\text{CA}} \odot (W_{\text{SMA}} \odot X). \tag{5.11}$$

5.3.1 MA-DTNet

The proposed network, MA-DTNet architecture for multi-task learning, is rooted in the encoder-decoder structure of the UNet (Ronneberger et al., 2015) architecture. There are two paths, one for segmentation and the other for classification. These two paths share the encoder part of the UNet. This encoder generates feature maps at different scales, which are later combined with the decoder's output at their respective levels. Instead of passing the outputs from the encoder directly to the decoder, I route the feature maps through my proposed C-SMA to enrich the morphological features. This concludes the segmentation network pathway. The prediction for the classification outcome is generated by the classification head, which is attached to the encoder's final output via the channel attention module. The detailed diagram of the proposed architecture is shown in Figure 5.1. I used a popular CNN network,

ResNet34, as the backbone of MA-DTNet.

5.4 Dataset Description and Training Protocol

5.4.1 Datasets

In this study, I utilize two publicly accessible datasets, namely UDIAT (Yap et al., 2017) and BUSI (Al-Dhabyani et al., 2020), to assess the efficacy of the proposed multitask learning approach. The UDIAT dataset comprises 163 ultrasound images, of which 110 pertain to benign tumors and the remaining depict malignant tumors, while the BUSI dataset encompasses 647 ultrasound images of breast tumors, with 437 depicting benign tumors and the remainder of malignancies. The primary objective for both datasets involves tumor segmentation, followed by the secondary task of malignancy detection. To standardize the datasets for analysis, I resize both the images and corresponding masks to dimensions of 256×256. Employing a 5-fold cross-validation methodology, I compute the mean and standard deviation of the performance metrics. To enhance the diversity of the training dataset and improve model robustness, I incorporate data augmentation techniques such as horizontal and vertical flips and rotations.

Other than these, I have also employed two publicly available datasets (HAM10000 (Tschandl et al., 2018) and GlaS (Sirinukunwattana et al., 2017)) to investigate the generalization ability of my proposed method. The HAM10000 dataset is a collection of dermatoscopic images encompassing a diverse range of skin lesions. It comprises 10,015 images of skin lesions, along with corresponding segmentation masks and diagnostic classifications. The diagnostic categories include actinic keratoses and intraepithelial carcinoma, basal cell carcinoma, benign keratosis-like lesions, dermatofibroma, melanoma, melanocytic nevi, and vascular lesions. I have used this dataset to evaluate my proposed multi-task algorithm, where the tasks are to segment the lesion and predict its diagnostic category. GlaS dataset comprised of colon histology images that are used to segment the glands and predict the malignancy of the gland. There are a total 165 images, among them 85 images (37 benign, 48 malignant) are used for training and 80 images (37 benign, 43 malignant) are used for testing following the original division. In preparation for deep learning model training, all images and their corresponding segmentation masks underwent preprocessing to establish a

uniform size of 256×256 pixels, ensuring compatibility and consistency within the dataset.

5.4.2 Training Protocol

The overall loss of the proposed network is computed by a weighted combination of the segmentation loss and classification loss given by the following equation.

$$Loss = \lambda \cdot \mathcal{L}_{segmentation} + (1 - \lambda) \cdot \mathcal{L}_{classification}. \tag{5.12}$$

The value of λ lies between 0 and 1 and is determined to be 0.8 for optimal performance (please refer 5.5.4 for detailed experiment). $\mathcal{L}segmentation$ consists of binary cross-entropy loss and Dice loss between the original segmentation mask and the predicted segmentation mask. For the classification task, $\mathcal{L}classification$ is generally computed using binary cross-entropy loss, as most datasets used in this work involve binary classification. However, for the HAM10000 dataset, which involves multi-class classification of skin lesions, categorical cross-entropy loss is used instead. This adjustment ensures a more suitable and effective loss function for handling multiple classes. The proposed multitask network is trained up to 1000 epochs by optimizing the total loss using the Adam optimizer with an initial learning rate of 0.0001. The PyTorch implementation is available at the GitHub repository https://github.com/SusmitaSenGhosh/MA-DTNET.

5.4.3 Classification Performance Metrics

The classification (malignancy detection) performance, is evaluated using three metrics —- accuracy, F1-score, and area under the receiver operating characteristic (AUROC). Accuracy measures the ratio of correctly predicted instances to the total number of instances. On the other hand, the F1-score combines both precision and recall into a single value. Precision measures the proportion of true positive predictions among all positive predictions, while recall measures the proportion of true positive predictions among all actual positives. In the context of malignancy detection, precision measures the accuracy of the model in correctly identifying malignant cases among all the cases it predicts as malignant. Recall measures the ability of the model to correctly identify all malignant cases, including those it may have missed. Thus, the F1-score combines both precision and recall into a single metric, providing

a balanced evaluation. Mathematically,

$$\mathrm{F1} = \frac{2 \cdot \mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}. \tag{5.13}$$

Both accuracy and F1-score are threshold-dependent measures, while AUROC is a threshold-independent measure. The ROC curve plots the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. AUROC represents the area under the ROC curve and provides a single scalar value to evaluate the performance of a binary classifier across all possible classification thresholds.

5.4.4 Segmentation Performance Metrics

I have used pixel similarity measuring metrics — dice score (DS), intersection over union (IoU), sensitivity, specificity, and precision metric to quantify the segmentation performance. The definitions of these metrics are presented in the following paragraph.

Let, Z and $\hat{Z}$ be the groundtruth mask and predicted mask respectively. Pixel-wise true positive, false positive, true negative, and false negative are denoted by tp, fp, tn, and fn, respectively. The pixel similarity segmentation metrics — Dice Score (DS), Intersection over Union (IoU), sensitivity, specificity, and precision used in this study are defined as follows.

$$\mathrm{DS}(Z, \hat{Z}) = \frac{2(Z \cap \hat{Z})}{(Z \cap \hat{Z}) + (Z \cup \hat{Z})} = \frac{2\,\mathrm{tp}}{2\,\mathrm{tp} + \mathrm{fn} + \mathrm{fp}}. \tag{5.14}$$

$$\mathrm{IoU}(Z, \hat{Z}) = \frac{(Z \cap \hat{Z})}{(Z \cup \hat{Z})} = \frac{\mathrm{tp}}{\mathrm{tp} + \mathrm{fn} + \mathrm{fp}}. \tag{5.15}$$

$$\mathrm{Sensitivity}(Z, \hat{Z}) = \frac{\mathrm{tp}}{\mathrm{tp} + \mathrm{fn}}. \tag{5.16}$$

$$\mathrm{Specificity}(Z, \hat{Z}) = \frac{\mathrm{tn}}{\mathrm{tn} + \mathrm{fp}}. \tag{5.17}$$

$$\mathrm{Precision}(Z, \hat{Z}) = \frac{\mathrm{tp}}{\mathrm{tp} + \mathrm{fp}}. \tag{5.18}$$

Recall, precision, and specificity focus on the algorithm's ability to correctly identify tumor and non-tumor regions, without considering the extent of the overlap between the predicted and ground truth masks. On the other hand, IoU and Dice score quantify the extent of

overlap between the predicted and ground truth masks, providing a measure of segmentation accuracy and similarity. The Dice score focuses on the overlap between the predicted and ground truth masks, giving equal importance to false positives and false negatives. IoU considers the size of the regions, as it measures the ratio of the intersection to the union.

Besides the pixel similarity based measures, I also considered shape similarity based measures such as Hausdorff distance which can be defined by the following equation.

$$\mathrm{HD}(Z, \hat{Z}) = \max(\mathrm{hd}(Z, \hat{Z}), \mathrm{hd}(\hat{Z}, Z) \tag{5.19}$$

where,

$$\mathrm{hd}(X, Y) = \max_{x \epsilon X} \min_{y \epsilon Y} ||x - y||. \tag{5.20}$$

The Hausdorff distance measures how much the predicted segmentation deviates from the ground truth segmentation in terms of spatial extent. It provides a measure of how well the predicted tumor boundaries align with the ground truth tumor boundaries. A lower Hausdorff distance indicates better agreement between the predicted and ground truth masks, meaning the predicted boundaries closely match the ground truth boundaries. Specifically, I utilized the 95^{th} percentile of Hausdorff distance (HD95), computed by considering the distances at the 95^{th} percentile rather than the maximum distance, as outlined by equation 5.19.

Moreover, histological images typically contain numerous glands with varying shapes and sizes. To assess the performance of my proposed method on GlaS dataset, I adopted two established metrics: object-level Dice score and Hausdorff distance. These metrics align with the evaluation criteria employed in (Dabass et al., 2022). The object-level Dice score specifically considers each gland as an individual entity during the evaluation process. Let, $Z_i \in Z$ and $\hat{Z}_j \in \hat{Z}$ be the i^{th} groundtruth and and j^{th} segmented glands, where $i = \{1, 2, \ldots N_Z\}$ and $j = \{1, 2, \ldots \hat{N_Z}\}$. Z'_i represents the glandular object groundtruth that is comprehensively overlaying with segmented gland $\hat{Z}_i$, whereas, $\hat{Z}'_j$ indicates the segmented object that is comprehensively overlaying with groundtruth object Z_i. Mathematically, object-level dice score can be computed by the following equation.

$$\mathrm{DS}_{obj}(Z, \hat{Z}) = \frac{1}{2}\Big[\sum_{i=1}^{N_Z} w_i \,\mathrm{DS}(Z_i, \hat{Z}_i) + \sum_{j=1}^{N_{\hat{Z}}} w'_j \,\mathrm{DS}(Z'_j, \hat{Z}'_j)\Big] \tag{5.21}$$

where, $w_i = \frac{|Z_i|}{\sum_{n=1}^{N_Z} |Z_n|}$, $w'_j = \frac{|\hat{Z}_j|}{\sum_{n=1}^{N_{\hat{Z}}} |\hat{Z}_n|}$ and DS is simple dice score computed using equation 2.

5.5 Experimental Results

5.5.1 Comparison with SOTA

In this section, I compare the performance of the proposed model with state-of-the-art methods for classification, segmentation, and MTL. My evaluation includes state-of-the-art generalized segmentation models like UNet (Ronneberger et al., 2015), UNet++ (Zhou et al., 2018), DeepLabV3+ (Chen et al., 2018a), utilizing ResNet34, a popular CNN backbone, as well as classification models like ResNet34. Furthermore, I take into account the contemporary segmentation, classification, and multitask models (Luo et al., 2022; Chen et al., 2022a; Wu et al., 2023a; Cui et al., 2022; Dabass et al., 2022; Zhang et al., 2021a) that are tailored to the breast tumor segmentation as well as malignancy detection task. I have considered Dice score, IOU, sensitivity, and specificity as segmentation performance accessing metrics and accuracy, F1-score, and AUROC as classification metrics following (Xu et al., 2023).

Result of the above experiments for UDIAT and BUSI dataset are reported in Table 5.1 and 5.2 respectively, where the top three performances are highlighted in red, blue, and green, corresponding to the first, second, and third best results. From both the tables, it is evident that my proposed method has MA-DTNet outperformed others in both segmentation and classification tasks, achieving the highest scores in terms of almost all the metrics considered here. Compared to the state-of-the-art MTL network, my proposed approach demonstrates a 3.28% increase in dice score and a 3.14% improvement in IoU metrics. Additionally, I observe a 1.05% enhancement in the F1-score for the UDIAT dataset. Furthermore, my proposed method exhibits a better balance between sensitivity and specificity compared to other MTL methods. A similar trend is also observed in the case of the BUSI dataset (Table 5.2). The segmentation performance experienced an improvement of 1.62% and 1.52% in terms of dice-score and IoU metrics, respectively, while 4.96% improvement is observed in accuracy metrics for the malignancy detection task. While the segmentation specificity of the proposed model may not surpass that of the other MTL models, it does exhibit a superior balance between sensitivity and specificity. High sensitivity ensures that the model does not miss any tumor

Table 5.1: A comparison of the performance of the proposed method with a related state-of-the-art method for the UDIAT dataset. The first, second, and third-best performances are highlighted in red, blue, and green, respectively.

Task	Method	Params.	Segmentation Performance				Classification Performance		
			DS↑	IoU↑	Sen.↑	Spec.↑	Acc.↑	F1-score↑	AUROC↑
Seg.	UNet (Ronneberger et al., 2015)	22.72 M	87.79 ± 2.55	79.86 ± 3.01	91.10 ± 3.01	99.44 ± 0.11	-	-	-
	UNet++ (Zhou et al., 2018)	24.42 M	87.36 ± 2.58	79.50 ± 3.20	89.33 ± 2.33	99.52 ± 0.10	-	-	-
	DeepLabV3+ (Chen et al., 2018a)	22.44 M	87.45 ± 2.89	79.34 ± 3.20	90.32 ± 1.97	99.48 ± 0.17	-	-	-
	AAU-net (Chen et al., 2022a)†	-	78.14 ± 2.41	69.10 ± 2.98	82.22 ± 3.84	98.82 ± 0.35	-	-	-
	ESKNet (Chen et al., 2024)†	44.57 M	78.71 ±2.37	70.20 ±2.28	82.41 ±2.84	97.47 ±0.35	-	-	-
	NU-net (Chen et al., 2023)†	77.05 M	80.80 ±0.57	72.03 ±0.82	84.13 ±1.73	98.96 ±0.17	-	-	-
	SMU-Net (Ning et al., 2021)†	-	87.03 ±1.25	78.49 ±1.49	88.85 ±1.72	-	-	-	-
Clas.	ResNet34 (He et al., 2016)	21.35 M	-	-	-	-	94.53 ± 3.94	91.21 ± 6.57	0.937 ± 0.066
	HoVer-Trans (Mo et al., 2023)†	79.38 M	-	-	-	-	77.40 ±6.10	61.90 ±9.90	0.781 ±0.118
	EPTM (Singh et al., 2024)†	192 M	-	-	-	-	90.90	93.20	0.932
MT	MTL-Net (Xu et al., 2023)†	93.50 M	80.95 ± 5.00	72.73 ± 5.49	84.28 ± 5.05	99.25 ± 0.14	87.08 ± 2.79	90.51 ± 2.29	0.936 ± 0.046
	MTL-COSA (Xu et al., 2022)†	109.24 M	84.07 ± 3.25	76.05 ± 3.71	86.97 ± 2.76	99.27 ± 0.25	91.44 ± 3.90	93.85 ± 2.58	0.946 ± 0.034
	RMTL-Net (Xu et al., 2023)†	93.51 M	85.69 ± 2.00	77.84 ± 2.45	89.51 ± 0.91	99.25 ± 0.19	95.74 ± 3.45	92.84 ± 5.98	0.935 ± .081
	MA-DTNet (ours)	22.95 M	88.97 ± 2.50	80.98 ± 3.17	91.55 ± 1.98	99.45 ± 0.23	96.35 ± 3.95	93.89 ± 6.81	0.954 ± 0.062

Seg. : Segmentation, Clas. : Classification, MT : Multitask, DS : Dice score, IoU : Intersection over union, Sen. : Sensitivity, Spec. : Specificity, Acc. : Accuracy
Results reported that the methods marked by † are taken from respective studies. The proposed method outperformed all the multitask learning models for segmentation and classification tasks with significantly fewer trainable parameters.

regions, which is critical for accurate diagnosis and treatment planning. High specificity helps minimize false alarms or misclassifications of non-tumor regions as tumors, which can reduce unnecessary medical interventions. In tumor segmentation, striking the right balance between sensitivity and specificity is essential to ensure the model's effectiveness in accurately identifying tumor regions while minimizing false positives and negatives. Moreover, MA-DTNet has at least four times fewer parameters than other MTL models, indicating an efficient model design that outperforms the other MTL methods in both tumor segmentation and malignancy detection tasks.

5.5.2 Comparison among various attention mechanisms

The effectiveness of the proposed C-SMA is compared with the performance of the other attention mechanisms for segmentation tasks. For this experiment, I have considered CBAM (Woo et al., 2018) and multi-head self-attention (MHSA) (Vaswani et al., 2017). CBAM

Table 5.2: A comparison of the performance of the proposed method with a related state-of-the-art method for the BUSI dataset. The first, second, and third-best performances are highlighted in red, blue, and green, respectively.

Task	Method	Params.	Segmentation Performance				Classification Performance		
			DS↑	IoU↑	Sen.↑	Spec.↑	Acc.↑	F1-score↑	AUROC↑
Seg.	UNet (Ronneberger et al., 2015)	22.72 M	81.10 ± 1.60	73.32 ± 1.64	83.33 ± 3.20	97.83 ± 0.55	-	-	-
	UNet++ (Zhou et al., 2018)	24.42 M	81.09 ± 1.36	72.92 ± 1.47	83.51 ± 2.47	97.91 ± 0.34	-	-	-
	DeepLabV3+ (Chen et al., 2018a)	22.44 M	80.88 ± 1.56	72.69 ± 1.67	82.68 ± 2.23	97.80 ± 0.26	-	-	-
	AAU-net (Chen et al., 2022a)†	-	77.51 ± 0.68	68.82 ± 0.44	80.10 ± 0.52	97.57 ± 0.24	-	-	-
	ESKNet (Chen et al., 2024)†	44.57 M	79.92 ± 2.21	71.65 ± 2.39	82.66 ± 1.40	99.01 ± 0.35	-	-	-
	AMS-PAN (Lyu et al., 2023)†	-	80.71	68.53	79.30	98.54	-	-	-
	NU-net (Chen et al., 2023)†	77.05 M	78.62 ± 1.38	70.35 ± 1.54	82.46 ± 1.02	97.48 ± 0.49	-	-	-
Clas.	ResNet34	21.35 M	-	-	-	-	- 95.21 ± 1.74	92.49 ± 2.71	0.968 ± 0.019
	HoVer-Trans (Mo et al., 2023)†	79.38 M	-	-	-	-	85.50 ± 5.00	87.20 ± 8.00	0.865 ±0.066
	MS-GOF (Zhong et al., 2023)†	105.69M	-	-	-	-	76.48 ± 5.93	74.90 ± 7.53	0.790 ± 0.064
MT	MTL-Net (Xu et al., 2023)†	93.50 M	77.76 ± 3.11	69.33 ± 2.89	78.91 ± 2.22	98.30 ± 0.25	90.18 ± 3.25	93.07 ± 2.41	0.962 ± 0.021
	MTL-COSA (Xu et al., 2022)†	109.24 M	78.90 ± 2.03	70.65 ± 2.01	79.31 ± 2.48	98.31 ± 0.11	91.49 ± 3.02	93.66 ± 2.36	0.968 ± 0.016
	RMTL-Net (Xu et al., 2023)†	93.51 M	80.04 ± 2.47	71.93 ± 2.15	82.54 ± 2.31	98.00 ± 0.3	91.02 ± 3.42	93.32 ± 3.35	0.0967 ± 0.015
	MA-DTNet (ours)	22.95 M	81.66 ± 1.56	73.45 ± 1.55	83.61 ± 2.51	97.85 ± 0.11	95.98 ± 1.48	93.71 ± 2.29	0.978 ± 0.010

Seg. : Segmentation, Clas. : Classification, MT : Multitask, DS : Dice score, IoU : Intersection over union, Sen. : Sensitivity, Spec. : Specificity, Acc. : Accuracy
Results reported the methods that are marked by † are taken from respective studies. The proposed method outperformed all the multi-task learning models for both segmentation and classification tasks with a significantly lower number of trainable parameters.

is a spatial and channel attention mechanism designed to enhance CNN's representational power, while self-attention is a mechanism commonly used in transformer architectures to capture dependencies within input sequences. Along with vanilla UNet, I have considered three types of attention modules — CBAM, MHSA, and C-SMA that are fused with UNet architecture. In UNet with CBAM and C-SMA, attention computation is conducted within each skip connection connecting the encoder to the decoder. Conversely, in UNet with MHSA, MHSA is exclusively applied to three skip connections containing smaller feature maps. This selective application is attributed to the heightened memory usage and computational cost associated with MHSA, rendering it impractical to employ across all skip connections.

Table 5.3 summarizes the results of the experiment. All three attention mechanisms (MHSA, CBAM, and C-SMA) yielded improvements in segmentation metrics compared to

Table 5.3: Performance comparison of different attention mechanisms integrated within UNet architecture for segmentation task.

Dataset	Method	Params.	Segmentation Performance				
			Dice Score↑	IoU↑	Precision↑	Recall↑	HD95↓
UDIAT	UNet	1.81 M	79.98 ± 2.82	70.48 ± 3.78	80.20 ± 2.97	83.14 ± 2.81	24.72 ± 7.35
	UNet-CBAM	1.82 M	82.55 ± 1.35	73.10 ± 1.93	82.26 ± 1.35	86.13 ± 1.21	19.68 ± 4.87
	UNet-MHSA	3.31 M	84.97 ± 3.53	75.87 ± 4.12	83.13 ± 3.01	**88.80 ± 3.38**	16.91 ± 6.81
	UNet-CA	1.95 M	83.48 ± 1.79	74.31 ± 2.30	82.49 ± 2.51	86.43 ± 2.37	18.41 ± 3.51
	UNet-SMA	1.98 M	83.50 ± 2.30	74.49 ± 2.74	83.98 ± 2.55	87.74 ± 2.63	19.43 ± 5.96
	UNet-C-SMA (Ours)	1.99 M	**85.14 ± 0.94***	**75.95 ± 1.27***	**85.35 ± 2.07***	87.34 ± 1.26	**15.38 ± 1.40**
BUSI	UNet	1.81 M	74.79 ± 1.53	65.92 ± 1.57	78.86 ± 1.28	76.65 ± 2.19	36.00 ± 4.77
	UNet-CBAM	1.82 M	78.63 ± 1.52	69.88 ± 1.48	81.42 ± 1.30	80.32 ± 1.45	27.17 ± 2.83
	UNet-MHSA	3.31 M	79.61 ± 1.70	70.72 ± 1.83	80.58 ± 2.08	**83.01 ± 2.21**	27.90 ± 2.95
	UNet-CA	1.95 M	78.88 ± 1.27	70.02 ± 1.06	80.19 ± 1.24	82.02 ± 1.55	29.45 ± 1.37
	UNet-SMA	1.98 M	76.87 ± 1.79	67.97 ± 1.65	79.54 ± 2.14	79.52 ± 1.24	31.96 ± 3.98
	UNet-C-SMA (Ours)	1.99 M	**79.73 ± 1.61***	**71.00 ± 1.37***	**81.92 ± 1.75***	81.49 ± 1.20*	**26.29 ± 1.98**

UNet-C-SMA achieved statistically significant performance improvements over UNet-CBAM, as evidenced by Wilcoxon signed-rank test results (indicated by *). UNet-C-SMA exhibited greater efficiency compared to UNet-MHSA in terms of dice score, intersection over union (IoU), and Hausdorff Distance (HD95) metrics, while also requiring fewer parameters. Notably, UNet-C-SMA maintained a superior balance between precision and recall.

the baseline model without attention. Notably, MHSA led to significant improvement, but at the cost of nearly doubling the trainable parameters. Conversely, CBAM and C-SMA increased the parameter count by a smaller margin. Focusing on the comparison between UNet-CBAM and UNet-C-SMA, the latter achieved statistically significant (Wilcoxon signed-rank test) improvements of 2.59% and 1.15% on the UDIAT and BUSI datasets, respectively. This improvement came at the expense of only 0.17 million additional trainable parameters.

To assess the individual contributions of CA and SMA components within the proposed C-SMA module, I evaluate the segmentation performance of the UNet network integrated with each component separately. These results, presented alongside the performance with the full C-SMA module, are crucial for understanding the efficacy of each component. As expected, the inclusion of either CA or SMA independently improves segmentation performance compared to the baseline UNet. Notably, integrating the combined C-SMA module (CA+SMA) further enhances segmentation accuracy, demonstrating the synergistic effect of these attention mechanisms.

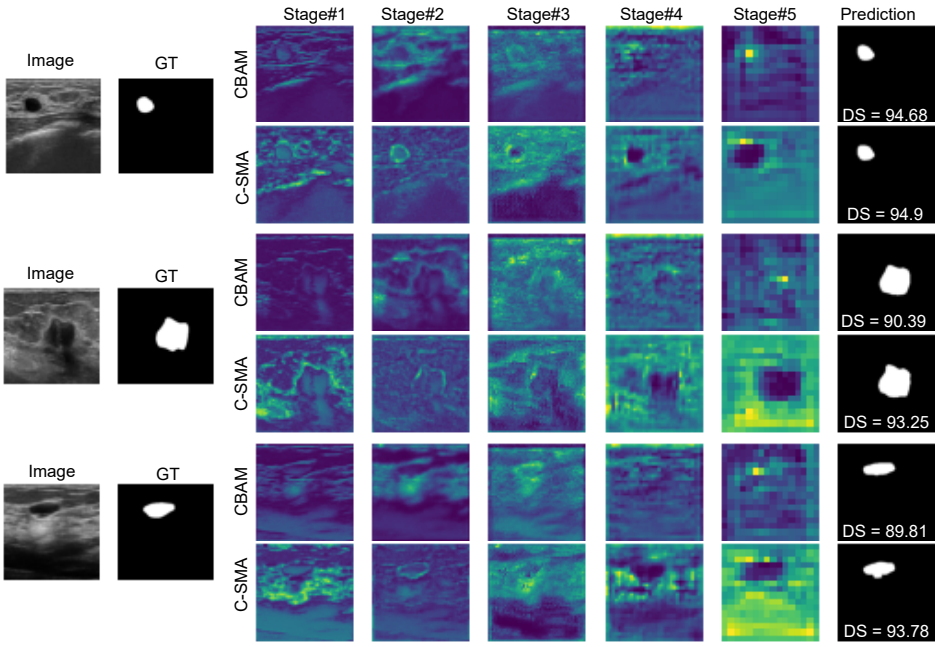

Figure 5.2: The qualitative comparison of the feature maps obtained at different stages of the encoder after the application of CBAM and proposed C-SMA.

5.5.3 Ablation study

The qualitative analysis of the CBAM and C-SMA module is also conducted by analyzing the feature maps obtained at different stages of the encoder post-application of respective attention mechanisms. It is evident from Figure 5.2 that C-SMA is capable of capturing morphological attributes such as the shape and size of the tumor in each of the feature maps more accurately by distinctly enhancing the boundary pixels.

MA-DTNet tackles multi-task learning by sharing an encoder for improved efficiency. It seamlessly integrates C-SMA for enhanced segmentation while incorporating CA within the classification branch to boost performance. I performed an ablation study to validate each component of the proposed network. This involved training a series of models with each component progressively removed. The performance of these ablated models was then compared to the full model's performance on the combined task. Additionally, to evaluate the effectiveness of the multi-task learning approach, I compared the performance of individual tasks (segmentation and classification) trained in isolation to their performance within the MA-DTNet framework. The segmentation task leverages a U-shaped network architecture, serving as the foundation for the proposed multi-task learning (MTL) network. For the

Table 5.4: Performance comparison of the ablation study on each component of the proposed MA-DTNet for UDIAT and BUSI dataset.

Dataset	Task	Method	Segmentation Performance					Classification Performance		
			DS↑	IoU↑	Prec.↑	Rec.↑	HD95↓	Acc.↑	F1-score↑	AUROC↑
UDIAT	Seg.	UNet	79.98 ± 2.82	70.48 ± 3.78	80.20 ± 2.97	83.14 ± 2.81	24.72 ± 7.35	-	-	-
	Clas.	UNet Encoder	-	-	-	-	-	90.21 ± 5.79	84.62 ± 8.88	0.8886 ± 0.0811
	MT	ES-MTL	80.88 ± 4.31	71.34 ± 4.87	81.56 ± 5.38	84.37 ± 3.37	24.21 ± 5.75	90.81 ± 3.67	84.23 ± 6.94	0.8831 ± 0.0756
		ES-MTL +C-SMA	84.99 ± 1.79	75.79 ± 2.06	**86.98 ± 2.44**	86.41 ± 2.56	16.98 ± 2.85	93.26 ± 3.30	89.27 ± 4.94	0.9175 ± 0.0416
		ES-MTL +C-SMA+CA	**85.33 ± 2.41**	**76.13 ± 2.91**	84.54 ± 3.00	**88.66 ± 1.09**	**16.39 ± 5.53**	**93.26 ± 2.57**	**89.63 ± 4.66**	**0.9445 ± 0.0457**
BUSI	Seg.	UNet	74.79 ± 1.53	65.92 ± 1.57	78.86 ± 1.28	76.65 ± 2.19	36.00 ± 4.77	-	-	-
	Clas.	UNet Encoder	-	-	-	-	-	89.18 ± 1.79	81.86 ± 3.92	0.9205 ± 0.0266
	MT	ES-MTL	75.65 ± 1.44	66.67 ± 1.50	78.39 ± 2.00	79.04 ± 1.84	36.20 ± 3.00	91.97 ± 1.28	87.46 ± 1.92	0.9429 ± 0.0155
		ES-MTL +C-SMA	**79.56 ± 1.42**	70.73 ± 1.31	81.72 ± 1.20	81.42 ± 1.24	27.15 ± 0.74	92.12 ± 1.83	87.54 ± 2.91	0.9441 ± 0.0254
		ES-MTL +C-SMA+CA	79.42 ± 1.69	**70.78 ± 1.44**	**82.35 ± 1.38**	**80.67 ± 2.63**	**25.51 ± 2.16**	**93.20 ± 1.58**	**89.14 ± 3.11**	**0.9521 ± 0.0116**

Seg. : Segmentation, Clas. : Classification, MT : Multitask
ES-MTL: Encoder-shared multitask learning model, ES-MTL+C-SMA: Encoder-shared multitask learning model with channel and spatial morphological attention (C-SMA) in encoder-decoder skip connection path, ES-MTL+MA+CA: Encoder-shared multitask learning model with channel and spatial morphological attention (C-SMA) in encoder-decoder skip connection path and channel attention (CA) in encoder-classifier path.

classification task, the encoder portion of the U-Net is directly integrated with a classification head. The result of the ablation study is reported in Table 5.4 for both datasets.

It can be observed that the encoder-shared vanilla MTL network surpassed both individual segmentation and classification performance for both UDIAT and BUSI datasets, suggesting the efficiency of an encoder-shared MTL (referred to as ES-MTL). The Encoder-shared MTL with C-SMAs integrated within the skip connections of the encoder and decoder (referred to as ES-MTL+C-SMA) further improves the segmentation performance in terms of all the segmentation metrics by a significant margin. Notably, this enhancement in segmentation performance was accompanied by a corresponding improvement in classification performance, particularly for the UDIAT dataset and marginally for BUSI. This suggests that accurate tumor segmentation plays a crucial role in achieving better malignancy detection, thus supporting the effectiveness of the proposed C-SMA for the MTL framework. Finally, the inclusion of CA within the classification path yielded a further improvement in classification performance, underlining its importance in this context.

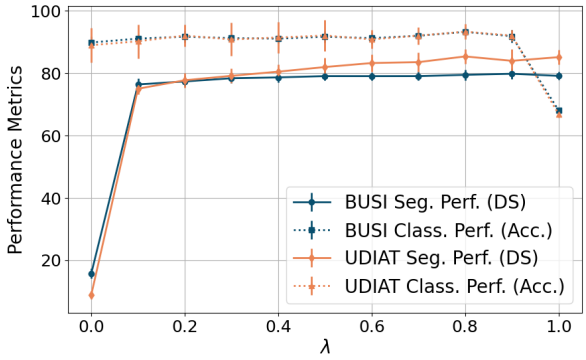

Figure 5.3: The segmentation (dice score) and classification performances (accuracy) of ES-MTL+C-SMA+CA for different λ

Table 5.5: Performance comparison for different sizes of the structuring element in C-SMA

Dataset	K_{SE}	**Param.**	**DS↑**	**HD95↓**
BUSI	3	1.95 M	79.73 ± 1.44	26.29 ± 1.77
	5	2.04 M	79.37 ± 1.33	25.94 ± 1.47
UDIAT	3	1.95 M	85.14 ± 0.84	15.38 ± 1.25
	5	2.04 M	83.66 ± 0.75	15.70 ± 3.28

5.5.4 Hyperparameter Optimization

The optimal value of the hyperparameter λ in equation 5.12 was determined by conducting a grid search across λ values ranging from 0 to 1 in increments of 0.1. As shown in figure 5.3, a λ value of 0.8 yields the best balance between segmentation performance and malignancy detection performance.

The size of the structuring element (K_{SE}) of morphological operation within the C-SMA module is another hyperparameter that requires optimization. It is also determined empirically. As shown in table 5.5, increasing the value of K_{SE} from 3 to 5 leads to a growth in the number of trainable parameters. However, this increase did not translate to significant performance improvements. Therefore, I opted for 3×3 as a value of K_{SE} for the optimal configuration for the C-SMA module, balancing model complexity with performance.

5.5.5 Generalization Ability

The efficacy of the proposed methodology is demonstrated in the context of segmenting breast tumors and detecting malignancies in ultrasound images. This section delves into examining the generalization capacity of the proposed method across two additional multitask

learning scenarios pertinent to medical imaging. The first scenario involves segmenting skin lesions and classifying diseases using dermoscopic images, while the second entails segmenting glands and predicting malignancies from histological images. To evaluate these tasks, I utilized the HAM10000 (Tschandl et al., 2018) and GlaS (Sirinukunwattana et al., 2017) datasets (for detailed description refer to the supplementary material). Comparative analyses of the proposed method's performance against state-of-the-art multitask learning approaches relevant to these tasks are presented in Table 5.6.

Table 5.6: Comparison of the performance of the proposed method on HAM10000 and GlaS dataset with related state-of-the-art methods.

Skin lesion segmentation and classification (HAM10000)			**Gland segmentation and malignancy detection (GlaS)**				
	Seg. perf.	**Clas. Perf.**		**Seg. perf. (GlaS A/ GlaS B)**		**Clas. perf.**	
Method	**DS↑**	**Acc.↑**	**Method**	**Object DS↑**	**HD↓**	**F1-score↑**	**Acc.↑**
CTAN (Kim et al., 2023)	92.91	57.85	MTUNet (Dabass et al., 2022)	**95.60**/90.90	23.17/71.53	97.77	97.50
MA-DTNet (Ours)	**94.75 ± 0.17**	**85.83 ± 0.47**	MA-DTNet (Ours)	92.99/**91.51**	**19.89/21.08**	**100.00**	**100.00**

Seg. perf. : Segmentation performance, Clas. perf. : Classification performance
For GlaS, the original train test split was used, whereas, for HAM10000, 5-fold cross-validation is adapted for the proposed method.

Regarding the HAM10000 dataset, the proposed method has exhibited superior performance compared to CTAN (Kim et al., 2023), surpassing it by 1.84% in dice score and 27.98% in accuracy metrics for skin lesion segmentation and classification tasks, respectively. The assessment of segmentation performance on the GlaS dataset entails an evaluation of two distinct subsets of the test data, GlaS A and GlaS B. Both the object dice score (Dabass et al., 2022) and Hausdorff Distance (HD) metrics exhibit enhancements on the GlaS B dataset when compared to MTUNet (Dabass et al., 2022). Conversely, for GlaS A, an improvement over MTUNet is observed solely in the Hausdorff distance metric. Despite these segmentation disparities, my proposed method achieves perfect accuracy in predicting gland malignancy. Overall, beyond its success in segmenting and classifying breast tumors in ultrasound images, the method performs well in the other two multitask settings, suggesting its generalizability for medical image analysis.

5.6 Discussion

This study introduces a novel Multi-Task Learning (MTL) model specifically designed for breast tumor segmentation and malignancy detection from ultrasound images. The model consists of a shared UNet encoder and UNet decoder, along with novel Channel-Spatial Morphological Attention Modules (C-SMAs) integrated at multiple resolution stages for semantic segmentation. Additionally, it includes a classification head with a channel attention module for disease grade prediction. The channel attention module highlights significant channels within a multitude, while the C-SMA simultaneously focuses on important channels and spatial locations in the feature maps based on their morphological attributes.

The proposed approach demonstrates effectiveness through experiments on ultrasound datasets for breast tumor segmentation and malignancy detection. To further test the versatility of the proposed method, it has been generalized to other multitask scenarios such as microscopic gland segmentation and detection, as well as skin lesion segmentation and disease detection. Further research must explore its adaptability to a wider range of medical imaging modalities beyond those investigated here.

Furthermore, the C-SMA module currently utilizes fundamental morphological operations. Future research could potentially investigate the application of more complex morphological operations like opening, closing, and gradients to improve model performance.

Chapter 6

Conclusion

Summary

This chapter evaluates the contributions presented in the earlier chapters of this thesis. It provides a concise summary of the key attributes of my work, emphasizing their significance in advancing research efforts in medical image analysis, particularly in the areas of classification, segmentation, and multitask learning. Additionally, I explore potential future directions inspired by my contributions, aiming to foster new research opportunities in this field.

6.1 Evaluation of contributions

This thesis has made significant contributions to the field of deep learning for medical image analysis by addressing key challenges in classification, segmentation, and multi-task learning. Throughout this work, novel architectures have been developed to tackle the complexities of medical data, such as irregular shapes, low contrast, and the limited availability of annotated datasets. These include advancements in hybrid architectures that combine the strengths of convolutional neural networks (CNNs) and transformer models, as well as lightweight segmentation networks that effectively balance accuracy and computational efficiency. Additionally, the proposed multitask learning frameworks highlight the potential of shared representations to enhance performance across multiple tasks, such as classification and segmentation, while reducing computational overhead. In this section, I briefly evaluate the different contributions made in this thesis.

- **Spectral-Spatial Domain Integration for Classification:** The second chapter

introduced a novel classification framework that integrates spatial and spectral domain features, using Discrete Cosine Transform (DCT) and Discrete Wavelet Transform (DWT) to enhance feature extraction. The proposed architecture demonstrated significant improvements in classification accuracy by leveraging complementary spatial and spectral features. Through experiments on COVID-19 detection tasks and other medical imaging datasets, the model achieved a performance boost of over 4% in the F1-score, demonstrating the value of incorporating spectral domain information alongside spatial content for disease classification tasks.

- **ReMixViT – A Novel Vision Transformer for Classification:** The third chapter introduced the ReMixViT, a novel transformer-based architecture designed for medical image classification. By incorporating elements of ResNet and an MLP-Mixer, the architecture overcame the challenges faced by traditional CNNs, such as limited generalization due to restricted receptive fields. This hybrid design achieved a 4.62% improvement in F1-score over vanilla ViT models and other hybrid approaches. The ReMixViT was further evaluated across six diverse medical imaging datasets, showing consistent performance improvements. This contribution highlights the potential of transformer-based architectures in improving classification accuracy, particularly in the domain of medical imaging, where detailed feature extraction is critical.

- **Morph-UNet – Morphological Operations in Segmentation:** In the fourth chapter, the focus shifted to segmentation tasks with the introduction of Morph-UNet, a novel architecture combining traditional UNet with multi-scale morphological operations. The segmentation framework, featuring trainable morphological modules, was particularly effective in handling irregularly shaped regions of interest (ROIs) in medical images. Morph-UNet was evaluated on several medical imaging datasets and achieved superior segmentation performance, with an average Dice Similarity Coefficient (DSC) improvement of 2.4% compared to traditional models. The approach effectively balanced precision and computational efficiency, demonstrating that morphological operations can significantly enhance the ability of deep learning models to capture complex medical features.

- **Multitask Learning Framework for Classification and Segmentation:** The fifth

chapter proposed a multitask learning framework, MA-DTNet, designed to simultaneously handle segmentation and classification tasks. By integrating spatial morphological attention and channel attention mechanisms, the model enhanced the feature extraction process for both tasks, achieving better performance with fewer parameters. MA-DTNet was tested on breast ultrasound datasets and demonstrated improvements in both classification and segmentation metrics, with an average 3% boost in accuracy for classification tasks. The multitask approach proved to be efficient in leveraging shared representations, reducing computational complexity, and offering a holistic solution for medical imaging tasks.

Each contribution has been evaluated using diverse and widely recognized medical imaging datasets, validating the effectiveness of the proposed models in both classification and segmentation tasks. The consistent performance improvements across different datasets underscore the robustness and scalability of the architectures. The integration of transformers, novel morphological operations, and multitask learning frameworks has pushed the boundaries of what can be achieved with deep learning in medical image analysis.

6.2 Future Possibilities

The findings and contributions of this thesis offer a foundation for future research. Several promising avenues can be explored to further advance the field of medical image analysis:

- **3D Medical Image Processing:** Three-dimensional image analysis has become increasingly important as modern scanners produce volumetric data (e.g., CT, MRI) critical for diagnosis and treatment planning. Conventional convolutional neural networks (CNNs) have been extended to three dimensions to leverage spatial context in these volumes (Litjens et al., 2017). However, 3D CNNs face challenges of limited receptive field and high computational cost. Recent work shows that attention-based transformer architectures can significantly enhance 3D processing by capturing global context. For example, SegFormer3D — a hierarchical 3D transformer model — applies multiscale attention across volumetric features and even uses a lightweight all-MLP decoder to achieve accuracy comparable to much larger networks (Xie et al., 2023). Such models

demonstrate that transformers can *surpass traditional 3D CNNs* by effectively modeling long-range dependencies in volume data (Isensee et al., 2021). Hybrid approaches that combine CNNs (to capture fine local detail) with transformer modules (for global context) are also promising (Hatamizadeh et al., 2022).

Despite these advances, key research questions remain. For example, Leader *et al.* found that naively using 3D CNNs did not automatically outperform 2D models for lung nodule detection, and emphasized the need for further research to develop novel 3D-CNN architectures that can fully leverage the rich information in 3D volumetric images (Leader et al., 2021). Future work should explore new network designs and training strategies to fully exploit volumetric data. Possible directions include multi-scale 3D network layouts, cascaded patch-based pipelines, self-supervised pretraining on unlabeled 3D datasets, and hierarchical transformers. Effective 3D position encodings and fusion of 3D imaging with temporal sequences or different modalities (e.g., PET-CT) are other active areas.

Improved 3D processing has clear clinical applications. Volumetric segmentation of tumors and organs is a fundamental task in medical imaging, as accurate 3D delineations inform diagnosis, surgical planning, and treatment (Xie et al., 2023). Higher-fidelity 3D analysis enables precise quantification of lesion volumes, supports navigation in image-guided interventions, and improves dynamic imaging applications (e.g., cardiac MRI).

- **Exploring Advanced Morphological Operations:** I have utilized several core morphological operations such as erosion, dilation, opening, closing, and gradient for improving the segmentation of medical images. These operations were chosen for their ability to handle irregularities in region shapes and to enhance features in medical images for tasks like tumor delineation or organ boundary extraction.

  As a potential future extension of this work, other morphological operations like the Top-Hat Transform, Black-Hat Transform, and Hit-or-Miss Transform could be incorporated to further enhance segmentation accuracy. For instance, the Top-Hat Transform could be applied to highlight small bright regions within medical scans, potentially improving the detection of microcalcifications or other subtle abnormalities. The Black-

Hat Transform, on the other hand, could help in accentuating darker regions that may represent hidden cavities or shadowed areas in the scans. Furthermore, the Hit-or-Miss Transform could be utilized for shape detection, particularly for identifying specific anatomical structures or pathologies based on predefined templates.

Integrating these additional morphological operations would provide a more comprehensive set of tools to tackle the intricate variations in medical images, enabling better segmentation and detection outcomes.

- **Integrating Multimodal Data for Comprehensive Diagnosis:** I have demonstrated the effectiveness of combining different feature extraction methods to enhance classification tasks in medical imaging. Future work can build on this by incorporating multimodal data—integrating medical images with complementary clinical information such as textual reports, genomic profiles, and patient histories. Modern diagnostic workflows increasingly rely on multiple data modalities, including imaging, electronic health records, genomics, and physician notes. Combining these diverse sources leads to more robust and comprehensive computational systems. As highlighted in recent surveys, models capable of processing multimodal data (e.g., image and text) are expected to play a pivotal role in next-generation biomedical computing (Bracci et al., 2023). Vision–language models for healthcare, which fuse visual and textual inputs using transformer architectures, have shown promise in tasks like radiology report generation and visual question answering (Zhang et al., 2023).

  In oncology, fusing histopathology images with gene expression data using integrated strategies has improved survival prediction and provided insights into disease-related biological mechanisms (Mobadersany et al., 2018). Looking ahead, critical research challenges include designing effective fusion techniques for modalities with heterogeneous formats and incomplete data. Transformer-based models and contrastive pretraining using large-scale multimodal datasets (e.g., chest X-rays paired with clinical notes) offer potential solutions. Additional promising directions include modular fusion networks, self-supervised learning objectives, and graph-based contextual modeling.

  Multimodal computational frameworks have broad clinical utility, including cancer prognosis, disease staging, personalized therapy planning, and automated medical re-

porting. Systems that can seamlessly integrate imaging with both structured and unstructured clinical data will be foundational to reliable and interpretable medical decision support.

Appendix A

Supplementary for Chapter 2

Table A.1: Detailed description of medical imaging datasets used in this study.

Dataset	Type of Data	Disease Type	Number of Classes	Number of Training Samples	Imbalance ratio
CBIS-DDSM (Sawyer-Lee et al., 2016)	Mammography images	Abnormality in breast tissue	2	8941	6.55
Diabetic Retinopathy (DR) [1]	Eye images	Diabetic Retinopathy	2	13052	1.33
Colorectal Histology (Kather et al., 2016)	Histology images	Colorectal cancer tissue type	8	4000	1
ISIC18 (Codella et al., 2019)	Dermoscopic images of skin	Skin disease	7	8166	58.86
Chestxray1 (Wang et al., 2020)	Chest x-ray images	COVID-19	3	13898	16.84
Chestxray2 (Kermany et al., 2018)	Chest x-ray images	Pneumonia	2	5232	2.88
BHI (Janowczyk and Madabhushi, 2016)	Breast Histopathology images	Invasive Ductal Carcinoma	2	12952	2.79

Table A.2: The performances of pixel, DCT, DWT, and pixel+DCT+DWT in identifying different medical condition for different datasets as listed in Table 2.3

	Pixel				DCT				DWT					Combined			
Dataset	ACSA	ASCP	ASCF	MADF	ACSA	ASCP	ASCF	MADF	ACSA	ASCP	ASCF	MADF	Combination Type	ACSA	ASCP	ASCF	MADF
BHI	89.32 (±0.53)	90.44 (±0.44)	89.85 (±0.28)	5.09 (±0.18)	87.15 (±0.51)	87.00 (±0.34)	87.07 (±0.21)	6.33 (±0.18)	85.52 (±0.45)	86.47 (±0.57)	85.97 (±0.33)	7.02 (±0.18)	pixel+DCT+DWT, SF	**89.86** (±0.3)	**90.98** (±0.23)	**90.40** (±0.17)	**4.81** (±0.1)
CBIS_DDSM	91.19 (±0.52)	**94.59** (±0.3)	92.79 (±0.26)	5.36 (±0.21)	77.58 (±1.55)	80.46 (±1.42)	78.82 (±0.43)	15.78 (±0.53)	86.60 (±0.89)	87.58 (±0.82)	87.05 (±0.26)	9.51 (±0.23)	pixel+DCT+DWT, FF	**91.98** (±1.24)	94.17 (±1.14)	**93.00** (±0.48)	**5.17** (±0.37)
CH	95.28 (±0.34)	95.42 (±0.25)	95.32 (±0.32)	3.34 (±0.33)	92.36 (±0.42)	92.38 (±0.29)	92.33 (±0.36)	5.07 (±0.42)	94.05 (±0.2)	94.03 (±0.19)	94.02 (±0.2)	4.04 (±0.28)	pixel+DWT, FF	**95.63** (±0.24)	**95.67** (±0.21)	**95.62** (±0.23)	**2.96** (±0.24)
CHESTXRAY1	89.97 (±0.46)	90.31 (±0.42)	90.00 (±0.45)	1.04 (±0.21)	87.33 (±0.47)	88.30 (±0.38)	87.34 (±0.49)	**0.82** (±0.24)	91.23 (±0.45)	91.37 (±0.46)	90.02 (±0.45)	91.95 (±0.22)	pixel+DCT+DWT, FF	**93.63** (±0.43)	**93.69** (±0.42)	**93.64** (±0.43)	1.44 (±0.34)
CHESTXRAY2	71.62 (±1.86)	86.31 (±0.58)	72.87 (±2.17)	12.44 (±1.37)	68.90 (±1.28)	85.41 (±0.37)	69.58 (±1.44)	14.57 (±0.89)	74.46 (±1.35)	88.00 (±0.4)	76.18 (±1.51)	10.49 (±0.93)	pixel+DWT, SF	**74.68** (±1.28)	**88.08** (±0.37)	**76.42** (±1.44)	**10.35** (±0.89)
DR	69.62 (±0.41)	69.42 (±0.35)	69.49 (±0.36)	4.24 (±0.3)	60.17 (±0.48)	60.28 (±0.35)	60.14 (±0.45)	6.67 (±1.51)	65.02 (±0.77)	64.79 (±0.58)	64.73 (±0.67)	4.19 (±1.3)	pixel+DCT, FF	**70.08** (±0.68)	**69.76** (±0.63)	**69.84** (±0.64)	**3.80** (±0.38)
ISIC18	75.89 (±0.63)	76.40 (±1.53)	75.86 (±0.87)	**9.49** (±0.44)	62.70 (±0.97)	66.97 (±0.89)	64.51 (±0.76)	13.93 (±0.82)	63.44 (±1.99)	65.81 (±1.36)	64.03 (±1.21)	15.17 (±0.74)	pixel+DCT+DWT, FF	**78.53** (±1.38)	**77.55** (±1.35)	**77.69** (±0.79)	9.69 (±0.75)

[1] https://www.kaggle.com/c/diabetic-retinopathy-detection

Appendix B

Supplementary for Chapter 3

Table B.1: Comparison of classification performances (in terms of mean±standard deviation of ACSR, ACSP, ACSF MADF, and AUROC) of baseline models (ResNet50 and ResNet-ViT) and proposed hybrid models (Res-ReMixViT and Res-ReMixViT+) models for six medical imaging datasets.

	ResNet50					**ResNet-ViT**					**Res-ReMixViT**					**Res-ReMixViT+**				
Datasets	**ACSR**	**ACSP**	**ACSF**	**MADF**	**AUROC** (10^{-2})	**ACSR**	**ACSP**	**ACSF**	**MADF**	**AUROC** (10^{-2})	**ACSR**	**ACSP**	**ACSF**	**MADF**	**AUROC** (10^{-2})	**ACSR**	**ACSP**	**ACSF**	**MADF**	**AUROC** (10^{-2})
Colorectal Histology	94.88 ±0.32	95.02 ±0.30	94.89 ±0.33	02.97 ±0.33	99.44 ±0.12	95.06 ±0.29	95.14 ±0.30	95.06 ±0.30	03.15 ±0.32	99.51 ±0.02	**95.50*** ±0.24	**95.54*** ±0.23	**95.50*** ±0.24	**02.79** ±0.22	**99.61**** ±0.05	**95.56*** ±0.22	**95.64*** ±0.21	**95.57*** ±0.22	**02.84** ±0.24	**99.62*** ±0.03
CBIS_ DDSM	52.60 ±1.29	47.90 ±2.09	48.73 ±1.52	18.30 ±0.50	92.87 ±0.50	55.59 ±1.01	49.23 ±1.86	51.52 ±1.41	17.29 ±0.40	91.34 ±0.99	52.69 ±0.47	**54.69*** ±1.60	**53.32*** ±0.57	17.54 ±0.36	**94.87*** ±0.23	54.28 ±0.67	**57.98**** ±1.14	**55.76**** ±0.77	**17.18** ±0.30	**96.49**** ±0.23
Chestxray	69.74 ±0.56	86.20 ±0.16	70.60 ±0.68	13.96 ±0.46	83.71 ±1.05	70.15 ±1.29	86.56 ±0.30	70.98 ±1.60	13.80 ±1.07	86.39 ±1.29	**73.08*** ±0.27	**87.07*** ±0.24	**74.61*** ±0.32	**11.37*** ±0.18	**88.39*** ±0.95	**74.46**** ±0.73	**88.16**** ±0.20	**76.16**** ±0.82	**10.53**** ±0.50	**88.54*** ±1.45
Fundus	46.92 ±0.11	48.98 ±0.17	47.11 ±0.21	16.51 ±0.16	79.34 ±0.04	49.28 ±0.44	55.20 ±0.59	50.57 ±0.55	16.11 ±0.35	81.22 ±0.39	**52.02*** ±0.83	51.52 ±0.66	**51.42*** ±0.74	**14.34*** ±0.48	**81.87*** ±0.26	**53.16**** ±0.33	**57.25**** ±1.16	**54.09**** ±0.58	**14.62*** ±0.39	**81.92*** ±0.12
ISIC18	70.31 ±0.38	70.53 ±0.37	70.25 ±0.29	09.93 ±0.71	93.21 ±0.33	70.00 ±0.70	70.59 ±1.02	70.00 ±0.90	10.26 ±0.56	91.44 ±0.47	**72.64*** ±0.83	**70.84** ±1.08	**71.29*** ±0.49	**09.23** ±0.57	**93.76*** ±0.16	**73.59**** ±0.48	**76.32**** ±0.44	**74.77**** ±0.42	**09.34*** ±0.20	**94.81**** ±0.20
PBC	98.91 ±0.02	98.92 ±0.06	98.91 ±0.02	0.83 ±0.09	99.93 ±0.00	98.83 ±0.08	98.79 ±0.22	98.81 ±0.15	0.80 ±0.11	99.91 ±0.01	**98.99*** ±0.03	**99.03*** ±0.02	**99.01*** ±0.01	**0.66*** ±0.11	**99.94*** ±0.01	**99.06*** ±0.04	**99.06*** ±0.10	**99.06*** ±0.03	**0.62*** ±0.06	**99.95*** ±0.00
Average	72.23	74.59	71.75	10.42	91.42	73.15	75.92	72.82	10.24	91.64	**74.15**	**76.45**	**74.19**	**09.32**	**93.07**	**75.02**	**79.07**	**75.90**	**09.19**	**93.56**
Comparison w.r.t to ResNet50						+0.92	+1.33	+1.08	−0.18	+ 0.22	**+1.93**	**+1.86**	**+2.44**	**−1.10**	**+1.66**	**+2.79**	**+4.48**	**+4.15**	**−1.23**	**+2.14**
Comparison w.r.t to ResNet-ViT											**+1.00**	**+0.53**	**+1.37**	**−0.91**	+1.44	**+1.87**	**+3.15**	**+3.08**	**−1.05**	**+1.92**
Comparison w.r.t to Res-ReMixViT																**+0.87**	**+2.62**	**+1.71**	**−0.13**	**+0.48**

Mean value of metrics greater than that for ResNet50 and ResNet-ViT models are marked by bold fonts.
***** indicates the statistically significant improvement in performance metric over both ResNet50 and ResNet-ViT.
****** indicates the statistically significant improvement in performance metric over ResNet50, ResNet-ViT and Res-ReMixViT.

Appendix C

List of Datasets

Table C.1: List of medical imaging datasets used in this thesis.

Name of Dataset	**Type of Images**	**Purpose of Use**
BHI (Janowczyk and Madabhushi, 2016)	Breast histopathology images	Classification
BUSI (Al-Dhabyani et al., 2020)	Breast ultrasound images	Segmentation, Classification
CBIS DDSM (Sawyer-Lee et al., 2016)	Mammography Images	Classification
Chestxray (Kermany et al., 2018)	Chest X-ray for Pneumonia detection	Classification
Chestxray1 (Wang et al., 2020)	Chest X-ray for COVID detection	Classification
Colorectal Histology (Kather et al., 2016)	Histological images of human colorectal cancer	Classification
Diabetic Retinopathy	Retina scan images	Classification
Fundus Images (DeepDRiD, 2020)	Fundus retinal images	Classification
GlaS (Sirinukunwattana et al., 2017)	Histological images of colorectal adenocarcinoma	Segmentation, Classification

HAM10000 (Tschandl et al., 2018)	Dermoscopic Images	Segmentation, Classification
ISIC2017 (Codella et al., 2018)	Dermoscopic images	Segmentation
ISIC2018 Codella et al. (2019)	Dermoscopic images	Segmentation, Classification
MonuSeg Kumar et al. (2017)	Histological microscopic images of tissues from multiple human body organs	Segmentation
PanNuke (Gamper et al., 2019)	Histological image patches of cancerous tissue	Segmentation
PBC Acevedo et al. (2020)	Peripheral blood cells images	Classification
UDIAT (Yap et al., 2017)	Breast ultrasound images	Segmentation, Classification

Appendix D

GitHub Repositories for Thesis Chapters

Below are the GitHub repositories corresponding to each chapter of this thesis.

- Chapter 2: `https://github.com/SusmitaSenGhosh/COVID-detection-from-X-ray-using-DWT-DCT-`
- Chapter 3: `https://github.com/SusmitaSenGhosh/ReMixViT-Enhancing-Clinical-Image-Analysis-with-Hybrid-Vision-Transformers-and-Adaptive-Feature-Mix`
- Chapter 4: `https://github.com/SusmitaSenGhosh/Morph-UNet-An-Approach-Towards-Medical-Image-Segmentation`
- Chapter 5: `https://github.com/SusmitaSenGhosh/MA-DTNET`

List of Publications

- Susmita Ghosh, Swagatam Das and Mallipeddi, Rammohan, "*A Deep Learning Framework Integrating the Spectral and Spatial Features for Image-Assisted Medical Diagnostics*," in IEEE Access, Volume 9, Pages 163686-163696, 2021, doi: `https://doi.org/10.1109/ACCESS.2021.3133338`.

- Susmita Ghosh and Swagatam Das, "*Multi-scale morphology-aided deep medical image segmentation*," in Engineering Applications of Artificial Intelligence, Volume 137, Part A, 2024, Pages 109047, doi: `https://doi.org/10.1016/j.engappai.2024.109047`.

- Susmita Ghosh and Swagatam Das, "*Enhancing Medical Image Analysis with MA-DTNet: A Dual Task Network Guided by Morphological Attention*," in Proceedings of 27th International Conference on Pattern Recognition 2024, Lecture Notes in Computer Science, Volume 15312, Springer, Cham, `https://doi.org/10.1007/978-3-031-78198-8_19`.

- Susmita Ghosh and Swagatam Das, "*ReMixViT: Enhancing Clinical Image Analysis with Hybrid Vision Transformers and Adaptive Feature Mixing*", Accepted for publication in Applied Soft Computing, (Manuscript Number: ASOC-D-24-07069), May 2025.

References

A. Abbas, M. M. Abdelsamea, and M. M. Gaber. DeTrac: Transfer learning of class decomposed medical images in convolutional neural networks. *IEEE Access*, 8:74901–74913, 2020. 16

A. Acevedo, A. Merino, S. Alférez, Á. Molina, L. Boldú, and J. Rodellar. A dataset of microscopic peripheral blood cell images for development of automatic recognition systems. *Data in Brief, ISSN: 23523409, Vol. 30,(2020)*, 2020. 65, 149

U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, M. Adam, A. Gertych, and R. San Tan. A deep convolutional neural network model to classify heartbeats. *Computers in biology and medicine*, 89:389–396, 2017. 34

P. Afshar, S. Heidarian, F. Naderkhani, A. Oikonomou, K. N. Plataniotis, and A. Mohammadi. COVID-CAPS: A capsule network-based framework for identification of COVID-19 cases from X-ray images. *Pattern Recognition Letters*, 138:638–643, 2020. 36

N. Ahmed, T. Natarajan, and K. R. Rao. Discrete cosine transform. *IEEE transactions on Computers*, 100(1):90–93, 1974. 38

W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy. Dataset of breast ultrasound images. *Data in brief*, 28:104863, 2020. 97, 126, 148

F. Almalik, M. Yaqub, and K. Nandakumar. Self-Ensembling Vision Transformer (SEViT) for Robust Medical Image Classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part III*, pages 376–386. Springer, 2022. 17

L. Alzubaidi, M. A. Fadhel, O. Al Shamma, J. Zhang, J. Santamaría, and Y. Duan. Robust application of new deep learning tools: An experimental study in medical imaging. *Multimedia Tools and Applications*, pages 1–29, 2022. 54

I. D. Apostolopoulos and T. A. Mpesiana. Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks. *Physical and Engineering Sciences in Medicine*, page 1, 2020. 35

G. C. Ates, P. Mohan, and E. Celik. Dual cross-attention for medical image segmentation. *arXiv preprint arXiv:2303.17696*, 2023. 22, 84, 85, 110

M. Ayoub, Z. Liao, S. Hussain, L. Li, C. W. Zhang, and K. K. Wong. End to end vision transformer architecture for brain stroke assessment based on multi-slice classification and

localization using computed tomography. *Computerized Medical Imaging and Graphics*, 109:102294, 2023. 57

R. Azad, M. Asadi-Aghbolaghi, M. Fathy, and S. Escalera. Attention Deeplabv3+: Multi-level context attention mechanism for skin lesion segmentation. In *European conference on computer vision*, pages 251–266. Springer, 2020. 81

A. T. Azar and S. M. El-Metwally. Decision tree classifiers for automated medical diagnosis. *Neural Computing and Applications*, 23:2387–2403, 2013. 11

V. Badrinarayanan, A. Kendall, and R. Cipolla. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017. 19

Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166, 1994. 13

B. Bracci, E. Alfonsi, F. Coppola, L. Grassi, F. Gallivanone, M. Sollini, and M. Kirienko. Multimodal learning in medical ai: a systematic review of models integrating imaging and text data. *European Journal of Nuclear Medicine and Molecular Imaging*, 50(1):1–17, 2023. 143

A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin. Albumentations: fast and flexible image augmentations. *Information*, 11(2):125, 2020. 100

Y. Cai and Y. Wang. MA-Unet: An improved version of unet based on multi-scale and attention mechanism for medical image segmentation. In *Third international conference on electronics and communication; network and computer technology (ECNCT 2021)*, volume 12167, pages 205–211. SPIE, 2022. 21

Z. Camlica, H. R. Tizhoosh, and F. Khalvati. Medical image classification via svm using lbp features from saliency-based folded data. In *2015 IEEE 14th international conference on machine learning and applications (ICMLA)*, pages 128–132. IEEE, 2015. 11

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*, pages 205–218. Springer, 2022. 22

Y. Celik, M. Talo, O. Yildirim, M. Karabatak, and U. R. Acharya. Automated invasive ductal carcinoma detection based using deep transfer learning with whole-slide images. *Pattern Recognition Letters*, 2020. 34

G. Chen, L. Li, Y. Dai, J. Zhang, and M. H. Yap. AAU-net: an adaptive attention U-net for breast lesions segmentation in ultrasound images. *IEEE Transactions on Medical Imaging*, 2022a. 21, 98, 104, 107, 108, 130, 131, 132

G. Chen, L. Li, J. Zhang, and Y. Dai. Rethinking the unpretentious u-net for medical ultrasound image segmentation. *Pattern Recognition*, 142:109728, 2023. 21, 131, 132

G. Chen, L. Zhou, J. Zhang, X. Yin, L. Cui, and Y. Dai. ESKNet: An enhanced adaptive selection kernel convolution for ultrasound breast tumors segmentation. *Expert Systems with Applications*, 246:123265, 2024. 131, 132

H. Chen, C. Li, G. Wang, X. Li, M. M. Rahaman, H. Sun, W. Hu, Y. Li, W. Liu, C. Sun, et al. GasHis-Transformer: A multi-scale visual transformer approach for gastric histopathological image detection. *Pattern Recognition*, 130:108827, 2022b. 57, 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. 21, 81, 84, 107

L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017a. 20, 81, 88

L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017b. 20, 81, 88

L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018a. 20, 81, 88, 130, 131, 132

R. J. Chen, C. Chen, Y. Li, T. Y. Chen, A. D. Trister, R. G. Krishnan, and F. Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16144–16155, 2022c. 57

S. Chen, Z. Wang, J. Shi, B. Liu, and N. Yu. A multi-task framework with feature passing module for skin lesion classification and segmentation. In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 1126–1129. IEEE, 2018b. 23

J. Cheng, J. Liu, H. Kuang, and J. Wang. A fully automated multimodal MRI-based multi-task learning for glioma segmentation and IDH genotyping. *IEEE Transactions on Medical Imaging*, 41(6):1520–1532, 2022. 24, 119

J. Cho, K. Lee, E. Shin, G. Choy, and S. Do. Medical image deep learning with hospital PACS dataset. *arXiv preprint arXiv:1511.06348*, 2015. 14

J. Chowdary, P. Yogarajah, P. Chaurasia, and V. Guruviah. A multi-task learning framework for automated segmentation and classification of breast tumors from ultrasound images. *Ultrasonic imaging*, 44(1):3–12, 2022. 119

A. Chung. Figure1-covid-chestxray-dataset. `https://github.com/agchung/Actualmed-COVID-chestxray-dataset`, 2020a. (Accessed on 07/14/2020). 41

A. Chung. Actualmed-covid-chestxray-dataset. `https://github.com/agchung/Figure1-COVID-chestxray-dataset`, 2020b. (Accessed on 07/14/2020). 41

F. Ciompi, B. de Hoop, S. J. van Riel, K. Chung, E. T. Scholten, M. Oudkerk, P. A. de Jong, M. Prokop, and B. van Ginneken. Automatic classification of pulmonary peri-fissural nodules in computed tomography using an ensemble of 2d views and a convolutional neural network out-of-the-box. *Medical image analysis*, 26(1):195–202, 2015. 14

N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC). *arXiv preprint arXiv:1902.03368*, 2019. 43, 65, 97, 145, 149

N. C. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 168–172. IEEE, 2018. 97, 149

J. P. Cohen, P. Morrison, L. Dao, K. Roth, T. Q. Duong, and M. Ghassemi. COVID-19 image data collection: Prospective predictions are the future. *arXiv preprint arXiv:2006.11988*, 2020. 41

D. Cozzi, M. Albanesi, E. Cavigli, C. Moroni, A. Bindi, S. Luvarà, S. Lucarini, S. Busoni, L. N. Mazzoni, and V. Miele. Chest X-ray in new Coronavirus Disease 2019 (COVID-19) infection: findings and correlation with clinical outcome. *La Radiologia Medica*, page 1, 2020. 34

W. Cui, Y. Peng, G. Yuan, et al. FMRNet: A fused network of multiple tumoral regions for breast tumor classification with ultrasound images. *Medical Physics*, 49(1):144–157, 2022. 130

L. B. da Cruz, D. A. D. Júnior, J. O. B. Diniz, A. C. Silva, J. D. S. de Almeida, A. C. de Paiva, and M. Gattass. Kidney tumor segmentation from computed tomography images using deeplabv3+ 2.5 d model. *Expert Systems with Applications*, 192:116270, 2022. 81

M. Dabass, . Vashisth, and R. Vig. MTU: A multi-tasking U-net with hybrid convolutional learning and attention modules for cancer classification and gland Segmentation in Colon Histopathological Images. *Comput. Biol. Med.*, 150:106095, 2022. 23, 129, 130, 137

Y. Dai, Y. Gao, and F. Liu. TransMed: Transformers Advance Multi-modal Medical Image Classification. *Diagnostics*, 11(8):1384, 2021. 54

O. Dalmaz, M. Yurt, and T. Çukur. ResViT: residual vision transformers for multimodal medical image synthesis. *IEEE Transactions on Medical Imaging*, 41(10):2598–2614, 2022. 17

DeepDRiD. The 2nd diabetic retinopathy – grading and image quality estimation challenge. `https://isbi.deepdr.org/data.html`, 2020. 65, 148

C. Di Rubeto, A. Dempster, S. Khan, and B. Jarra. Segmentation of blood images using morphological operators. In *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, volume 3, pages 397–400. IEEE, 2000. 82

T. N. Doan, B. Song, T. T. Vuong, K. Kim, and J. T. Kwak. Sonnet: A self-guided ordinal regression neural network for segmentation and classification of nuclei in large-scale multi-tissue histology images. *IEEE Journal of Biomedical and Health Informatics*, 26(7):3218–3228, 2022. 85, 113

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 16, 54, 58, 84

J. S. Duncan and N. Ayache. Medical image analysis: Progress over two decades and the challenges ahead. *IEEE transactions on pattern analysis and machine intelligence*, 22(1): 85–106, 2000. 11

V. Eroschenko. diFiore's atlas of histology. *Lippincott Williams & Wilkins, Philadelphia-USA, p162-4*, 314:322, 2005. 70

A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542 (7639):115–118, 2017. 34

A. Foo, W. Hsu, M. L. Lee, G. Lim, and T. Y. Wong. Multi-task learning for diabetic retinopathy grading and lesion segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13267–13272, 2020. 24

G. Franchi, A. Fehri, and A. Yao. Deep morphological networks. *Pattern Recognition*, 102: 107246, 2020. 88

H. Fu, J. Cheng, Y. Xu, C. Zhang, D. W. K. Wong, J. Liu, and X. Cao. Disc-aware ensemble network for glaucoma screening from fundus image. *IEEE Transactions on Medical Imaging*, 37(11):2493–2501, 2018. 16

J. Gamper, N. A. Koohbanani, K. Benes, A. Khuram, and N. Rajpoot. Pannuke: an open pan-cancer histology dataset for nuclei instance segmentation and classification. In *European Congress on Digital Pathology*, pages 11–19. Springer, 2019. 100, 149

F. Gao, H. Yoon, T. Wu, and X. Chu. A feature transfer enabled multi-task deep learning model on medical imaging. *Expert Systems with Applications*, 143:112957, 2020. 23

F. Garcea, A. Serra, F. Lamberti, and L. Morra. Data augmentation for medical imaging: A systematic literature review. *Computers in Biology and Medicine*, page 106391, 2022. 76

L. Gaur, U. Bhatia, N. Jhanjhi, G. Muhammad, and M. Masud. Medical image-based detection of COVID-19 using deep convolution neural networks. *Multimedia systems*, 29(3): 1729–1738, 2023. 15

B. Gheflati and H. Rivaz. Vision transformers for classification of breast ultrasound images. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 480–483. IEEE, 2022. 57, 77

S. Ghosh, S. Das, and R. Mallipeddi. A deep learning framework integrating the spectral and spatial features for image-assisted medical diagnostics. *IEEE Access*, 9:163686–163696, 2021. 54

S. Gokhale. Ultrasound characterization of breast masses. *Indian Journal of Radiology and Imaging*, 19(03):242–247, 2009. 119

B. Graham, A. El-Nouby, H. Touvron, P. Stock, A. Joulin, H. Jégou, and M. Douze. LeViT: a vision transformer in ConvNet's clothing for faster inference. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12259–12269, 2021. 22

S. Graham, Q. D. Vu, S. E. A. Raza, A. Azam, Y. W. Tsang, J. T. Kwak, and N. Rajpoot. Hover-Net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical image analysis*, 58:101563, 2019. 85, 113

S. Graham, Q. D. Vu, M. Jahanifar, S. E. A. Raza, F. Minhas, D. Snead, and N. Rajpoot. One model is all you need: multi-task learning enables simultaneous histology image segmentation and classification. *Medical Image Analysis*, 83:102685, 2023. 24

C. Guo, M. Szemenyei, Y. Yi, W. Wang, B. Chen, and C. Fan. SA-UNet: Spatial attention u-net for retinal vessel segmentation. In *2020 25th international conference on pattern recognition (ICPR)*, pages 1236–1242. IEEE, 2021. 21

A. Gupta, S. Gupta, R. Katarya, et al. InstaCovNet-19: A deep learning classification model for the detection of COVID-19 patients using Chest X-ray. *Applied Soft Computing*, 99: 106859, 2021. 36

K. Hammoudi, H. Benhabiles, M. Melkemi, F. Dornaika, I. Arganda-Carreras, D. Collard, and A. Scherpereel. Deep learning on chest X-ray images to detect and evaluate pneumonia cases at the era of COVID-19. *Journal of medical systems*, 45(7):75, 2021. 77

A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, A. Myronenko, B. Landman, H. Roth, and D. Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 574–584, 2022. 142

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 131

S. He, P. E. Grant, and Y. Ou. Global-local transformer for brain age estimation. *IEEE Transactions on Medical Imaging*, 41(1):213–224, 2021. 57

T. Heimann and H.-P. Meinzer. Statistical shape models for 3d medical image segmentation: a review. *Medical image analysis*, 13(4):543–563, 2009. 11

S. A. Hicks, I. Strümke, V. Thambawita, M. Hammou, M. A. Riegler, P. Halvorsen, and S. Parasa. On evaluation metrics for medical applications of artificial intelligence. *Sci. Rep.*, 12(1):5979, 2022. 65

G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 20

X. Huang, Z. Deng, D. Li, and X. Yuan. MISSFormer: An effective medical image segmentation transformer. *arXiv preprint arXiv:2109.07162*, 2021. 22

Y. Huang, X. Yang, L. Liu, H. Zhou, A. Chang, X. Zhou, R. Chen, J. Yu, J. Chen, C. Chen, et al. Segment anything model for medical images? *Medical Image Analysis*, 92:103061, 2024. 116

Z. Huang, X. Zhu, M. Ding, and X. Zhang. Medical image classification using a light-weighted hybrid neural network based on pcanet and densenet. *IEEE Access*, 8:24697–24712, 2020. 15

N. Ibtehaz and M. S. Rahman. MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation. *Neural networks*, 121:74–87, 2020. 20, 22

T. Ilyas, Z. I. Mannan, A. Khan, S. Azam, H. Kim, and F. De Boer. TSFD-Net: Tissue specific feature distillation network for nuclei segmentation and classification. *Neural Networks*, 151:1–15, 2022. 85, 113

S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015. 91

A. Iqbal and M. Sharif. UNet: A semi-supervised method for segmentation of breast tumor images using a U-shaped pyramid-dilated network. *Expert Systems with Applications*, 221: 119718, 2023. 21

F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein. nnu-net: Self-adapting framework for u-net-based medical image segmentation. *Nature methods*, 18(2):203–211, 2021. 142

A. M. Ismael and A. Şengür. Deep learning approaches for COVID-19 detection based on chest X-ray images. *Expert Systems with Applications*, 164:114054, 2021. 36

D. K. Jain et al. An evaluation of deep learning based object detection strategies for threat object detection in baggage security imagery. *pattern recognition letters*, 120:112–119, 2019. 54

A. Janowczyk and A. Madabhushi. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of pathology informatics*, 7, 2016. 43, 145, 148

Y. Ji, R. Zhang, H. Wang, Z. Li, L. Wu, S. Zhang, and P. Luo. Multi-compound transformer for accurate biomedical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*, pages 326–336. Springer, 2021. 85, 113, 114

Y. Jiang, L. Chen, H. Zhang, and X. Xiao. Breast cancer histopathological image classification using convolutional neural networks with small se-resnet module. *PloS one*, 14(3):e0214587, 2019. 15

J. N. Kather, C.-A. Weis, F. Bianconi, S. M. Melchers, L. R. Schad, T. Gaiser, A. Marx, and F. G. Zöllner. Multi-class texture analysis in colorectal cancer histology. *Scientific reports*, 6(1):1–11, 2016. 43, 65, 145, 148

T. Kaur and T. K. Gandhi. Automated brain image classification based on VGG-16 and transfer learning. In *2019 international conference on information technology (ICIT)*, pages 94–98. IEEE, 2019. 15

D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131, 2018. 43, 65, 145, 148

D. Khaledyan, T. J. Marini, T. M. Baran, A. O'Connell, and K. Parker. Enhancing breast ultrasound segmentation through fine-tuning and optimization techniques: sharp attention unet. *Plos one*, 18(12):e0289195, 2023. 21

H. E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros, and T. Ganslandt. Transfer learning for medical image classification: a literature review. *BMC medical imaging*, 22(1):69, 2022. 16

S. Kim, T. G. Purdie, and C. McIntosh. Cross-Task Attention Network: Improving Multi-task Learning for Medical Imaging Applications. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 119–128. Springer, 2023. 137

A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 116

P. Kora, C. P. Ooi, O. Faust, U. Raghavendra, A. Gudigar, W. Y. Chan, K. Meenakshi, K. Swaraja, P. Plawiak, and U. R. Acharya. Transfer learning techniques for medical image analysis: A review. *Biocybernetics and Biomedical Engineering*, 42(1):79–107, 2022. 16

A. Kumar, J. Kim, D. Lyndon, M. Fulham, and D. Feng. An ensemble of fine-tuned convolutional neural networks for medical image classification. *IEEE journal of biomedical and health informatics*, 21(1):31–40, 2016. 16, 77

A. Kumar, O. S. Ajani, S. Das, and R. Mallipeddi. GridShift: A Faster Mode-seeking Algorithm for Image Segmentation and Object Tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8131–8139, 2022. 116

A. Kumar, S. Das, and R. Mallipeddi. UEQMS: UMAP embedded quick mean shift algorithm for high dimensional clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 8386–8395, 2023. 116

N. Kumar, R. Verma, S. Sharma, S. Bhargava, A. Vahadane, and A. Sethi. A dataset and a technique for generalized nuclear segmentation for computational pathology. *IEEE Transactions on Medical Imaging*, 36(7):1550–1560, 2017. 99, 149

N. Kumar, R. Verma, D. Anand, Y. Zhou, O. F. Onder, E. Tsougenis, H. Chen, P.-A. Heng, J. Li, Z. Hu, et al. A multi-organ nucleus segmentation challenge. *IEEE Transactions on Medical Imaging*, 39(5):1380–1391, 2019. 99

G. Latif, D. A. Iskandar, J. M. Alghazo, and N. Mohammad. Enhanced MR image classification using hybrid statistical and wavelets features. *IEEE Access*, 7:9634–9644, 2018. 11

J. K. Leader, R. Wang, C. Fuhrman, T. E. Hartman, S. Bidic, X. Wang, D. O. Wilson, D. Gur, Z. Lu, et al. Fully 3d convolutional networks for sub-centimeter pulmonary nodule detection in ct imaging. *Academic Radiology*, 28(2):207–214, 2021. 142

C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu. Deeply-supervised nets. In *Artificial intelligence and statistics*, pages 562–570. Pmlr, 2015. 63

H. Li, J. Fang, S. Liu, X. Liang, X. Yang, Z. Mai, M. T. Van, T. Wang, Z. Chen, and D. Ni. CR-Unet: A composite network for ovary and follicle segmentation in ultrasound images. *IEEE journal of biomedical and health informatics*, 24(4):974–983, 2019. 21

Q. Li, W. Cai, X. Wang, Y. Zhou, D. D. Feng, and M. Chen. Medical image classification with convolutional neural network. In *2014 13th international conference on control automation robotics & vision (ICARCV)*, pages 844–848. IEEE, 2014. 14

X. Li, T. Pang, B. Xiong, W. Liu, P. Liang, and T. Wang. Convolutional neural networks based transfer learning for diabetic retinopathy fundus image classification. In *2017 10th international congress on image and signal processing, biomedical engineering and informatics (CISP-BMEI)*, pages 1–11. IEEE, 2017. 16

Y. Li, X. Li, X. Xie, and L. Shen. Deep learning based gastric cancer identification. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 182–185. IEEE, 2018. 76

S. Liang, R. Nie, J. Cao, X. Wang, and G. Zhang. FCF: Feature complement fusion network for detecting COVID-19 through CT scan images. *Applied Soft Computing*, 125:109111, 2022. 57

T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 94, 113

X. Lin, L. Yu, K.-T. Cheng, and Z. Yan. The lighter the better: Rethinking transformers in medical image segmentation through adaptive pruning. *IEEE Transactions on Medical Imaging*, 2023. 85, 105, 106, 107

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*, 42:60–88, 2017. 141

W. Liu, C. Li, N. Xu, T. Jiang, M. M. Rahaman, H. Sun, X. Wu, W. Hu, H. Chen, C. Sun, et al. CVM-Cervix: A hybrid cervical Pap-smear image classification framework using CNN, visual transformer and multilayer perceptron. *Pattern Recognition*, 130:108829, 2022. 77

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*, pages 10012–10022, 2021. 55, 77

F. Long, J. J. Peng, W. Song, X. Xia, and J. Sang. Bloodcaps: A capsule network based model for the multiclassification of human peripheral blood cells. *Computer methods and programs in biomedicine*, 202:105972, 2021. 77

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*, pages 3431–3440, 2015. 18

H. Luo, Y. Changdong, and R. Selvan. Hybrid ladder transformers with efficient parallel-cross attention for medical image segmentation. In *International Conference on Medical Imaging with Deep Learning*, pages 808–819. PMLR, 2022. 130

Y. Lyu, Y. Xu, X. Jiang, J. Liu, X. Zhao, and X. Zhu. AMS-PAN: Breast ultrasound image segmentation model combining attention mechanism and multi-scale features. *Biomedical Signal Processing and Control*, 81:104425, 2023. 132

D. Ma, M. R. Hosseinzadeh Taher, J. Pang, N. U. Islam, F. Haghighi, M. B. Gotway, and J. Liang. Benchmarking and boosting transformers for medical image classification. In *MICCAI Workshop on Domain Adaptation and Representation Transfer*, pages 12–22. Springer, 2022. 17

J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024. 116

M. S. Majib, M. M. Rahman, T. S. Sazzad, N. I. Khan, and S. K. Dey. VGG-SCNet: A VGG net-based deep learning framework for brain tumor detection on MRI images. *IEEE Access*, 9:116942–116952, 2021. 15

H. Malik, T. Anees, M. Din, and A. Naeem. CDC_Net: Multi-classification convolutional neural network model for detection of COVID-19, pneumothorax, pneumonia, lung Cancer, and tuberculosis using chest X-rays. *Multimedia Tools and Applications*, 82(9):13855–13880, 2023. 54

O. N. Manzari, H. Ahmadabadi, H. Kashiani, S. B. Shokouhi, and A. Ayatollahi. MedViT: A robust vision transformer for generalized medical image classification. *Computers in Biology and Medicine*, 157:106791, 2023. 17, 57, 58, 77

M. A. Mazurowski, H. Dong, H. Gu, J. Yang, N. Konz, and Y. Zhang. Segment anything model for medical image analysis: an experimental study. *Medical Image Analysis*, 89: 102918, 2023. 116

A. M. Mendonca and A. Campilho. Segmentation of retinal blood vessels by combining the detection of centerlines and morphological reconstruction. *IEEE Transactions on Medical Imaging*, 25(9):1200–1213, 2006. 82

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)*, pages 565–571. Ieee, 2016. 19, 94

E. Miranda, M. Aryuni, and E. Irwansyah. A survey of medical image classification techniques. In *2016 international conference on information management and technology (ICIMTech)*, pages 56–61. IEEE, 2016. 11

Y. Mo, C. Han, Y. Liu, M. Liu, Z. Shi, J. Lin, B. Zhao, C. Huang, B. Qiu, Y. Cui, et al. HoVer-Trans: Anatomy-aware HoVer-Transformer for ROI-free breast cancer diagnosis in ultrasound images. *IEEE Transactions on Medical Imaging*, 2023. 131, 132

P. Mobadersany, S. Yousefi, M. Amgad, D. A. Gutman, J. S. Barnholtz-Sloan, J. E. Vega, D. J. Brat, and L. A. D. Cooper. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proceedings of the National Academy of Sciences*, 115(13): E2970–E2979, 2018. 143

A. K. Mondal, A. Bhattacharjee, P. Singla, and A. Prathosh. xViTCOS: Explainable vision transformer based COVID-19 screening using radiography. *IEEE Journal of Translational Engineering in Health and Medicine*, 10:1–10, 2021. 57

R. Mondal, P. Purkait, S. Santra, and B. Chanda. Morphological networks for image de-raining. In *International Conference on Discrete Geometry for Computer Imagery*, pages 262–275. Springer, 2019. 82, 88, 120

R. Mondal, M. S. Dey, and B. Chanda. Image restoration by learning morphological opening-closing network. *Mathematical Morphology-Theory and Applications*, 4(1):87–107, 2020. 82, 88

W. K. Moon, Y. W. Lee, H. H. Ke, S. H. Lee, C. S. Huang, and R. F. Chang. Computer-aided diagnosis of breast ultrasound images using ensemble learning from convolutional neural networks. *Computer methods and programs in biomedicine*, 190:105361, 2020. 77

W. H. Nailon. Texture analysis methods for medical image characterisation. *Biomedical imaging*, 75:100, 2010. 11

L. Nanni, S. Ghidoni, and S. Brahnam. Ensemble of convolutional neural networks for bioimage classification. *Applied Computing and Informatics*, 17(1):19–35, 2021. 77

L. Nanni, M. Paci, S. Brahnam, and A. Lumini. Feature transforms for image data augmentation. *Neural Computing and Applications*, 34(24):22345–22356, 2022. 77

P. Nardelli, D. Jimenez-Carretero, D. Bermejo-Pelaez, G. R. Washko, F. N. Rahaghi, M. J. Ledesma-Carbayo, and R. S. J. Estépar. Pulmonary artery–vein classification in CT images using deep learning. *IEEE Transactions on Medical Imaging*, 37(11):2428–2440, 2018. 54

Z. Ning, S. Zhong, Q. Feng, W. Chen, and Y. Zhang. SMU-Net: Saliency-guided morphology-aware U-Net for breast lesion segmentation in ultrasound image. *IEEE Transactions on Medical Imaging*, 41(2):476–490, 2021. 21, 131

K. Nogueira, J. Chanussot, M. Dalla Mura, and J. A. Dos Santos. An introduction to deep morphological networks. *IEEE Access*, 9:114308–114324, 2021. 82

Y. Oh, S. Park, and J. C. Ye. Deep learning COVID-19 features on CXR using limited training data sets. *IEEE transactions on medical imaging*, 39(8):2688–2700, 2020. 35

O. Oktay. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018. 20, 22

T. Ozturk, M. Talo, E. A. Yildirim, U. B. Baloglu, O. Yildirim, and U. R. Acharya. Automated detection of covid-19 cases using deep neural networks with x-ray images. *Computers in Biology and Medicine*, page 103792, 2020. 35, 48

Y. Pan, W. Huang, Z. Lin, W. Zhu, J. Zhou, J. Wong, and Z. Ding. Brain tumor grading based on neural networks and convolutional neural networks. In *2015 37th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 699–702. IEEE, 2015. 14

S. Park and J. C. Ye. Multi-task distributed learning using vision transformer with random patch permutation. *IEEE Transactions on Medical Imaging*, 2022. 57

J. Prinyakupt and C. Pluempitiwiriyawej. Segmentation of white blood cells and comparison of cell morphology by linear and naïve bayes classifiers. *Biomedical engineering online*, 14: 1–19, 2015. 71

W. Qi, H.-C. Wu, and S.-C. Chan. Mdf-net: A multi-scale dynamic fusion network for breast tumor segmentation of ultrasound images. *IEEE Transactions on Image Processing*, 2023. 21

R. Rajasree, C. C. Columbus, and C. Shilaja. Multiscale-based multimodal image classification of brain tumor using deep learning method. *Neural Computing and Applications*, 33 (11):5543–5553, 2021. 15

P. Rajpurkar, J. Irvin, R. L. Ball, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. P. Langlotz, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the chexnext algorithm to practicing radiologists. *PLoS medicine*, 15(11):e1002686, 2018. 35

R. Ranftl, A. Bochkovskiy, and V. Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021. 55

R. Rasti, H. Rabbani, A. Mehridehnavi, and F. Hajizadeh. Macular OCT classification using a multi-scale convolutional neural network ensemble. *IEEE Transactions on Medical Imaging*, 37(4):1024–1034, 2017. 15

A. Riasatian, M. Babaie, D. Maleki, S. Kalra, M. Valipour, S. Hemati, M. Zaveri, A. Safarpoor, S. Shafiei, M. Afshari, et al. Fine-tuning and training of densenet for histopathology image representation using tcga diagnostic slides. *Medical Image Analysis*, 70:102032, 2021. 77

M. M. Rocha, G. Landini, and J. B. Florindo. Medical image classification using a combination of features from convolutional neural networks. *Multimedia Tools and Applications*, 82(13):19299–19322, 2023. 54

O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 18, 81, 83, 91, 109, 125, 130, 131, 132

H. R. Roth, L. Lu, J. Liu, J. Yao, A. Seff, K. Cherry, L. Kim, and R. M. Summers. Improving computer-aided detection using convolutional neural networks and random view aggregation. *IEEE Transactions on Medical Imaging*, 35(5):1170–1181, 2015. 14

R. Rouhi, M. Jafari, S. Kasaei, and P. Keshavarzian. Benign and malignant breast tumors classification based on region growing and cnn segmentation. *Expert Systems with Applications*, 42(3):990–1002, 2015. 14

S. K. Roy, R. Mondal, M. E. Paoletti, J. M. Haut, and A. Plaza. Morphological convolutional neural networks for hyperspectral image classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:8689–8702, 2021. 82, 124

RSNA. Rsna pneumonia detection challenge. `https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/data`, 2018. (Accessed on 07/14/2020). 41

RSNA. Covid-19 radiography database. `https://www.kaggle.com/tawsifurrahman/covid19-radiography-database`, 2020. (Accessed on 07/14/2020). 41

J. Ruan, S. Xiang, M. Xie, T. Liu, and Y. Fu. Malunet: A multi-attention and light-weight unet for skin lesion segmentation. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1150–1156. IEEE, 2022. 20, 81, 85, 97, 105, 106

S. Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017. 119

D. Sarwinda, R. H. Paradisa, A. Bustamam, and P. Anggia. Deep learning in image classification using residual network (resnet) variants for detection of colorectal cancer. *Procedia Computer Science*, 179:423–431, 2021. 15

R. Sawyer-Lee, F. Gimenez, A. Hoogi, and D. Rubin. Curated breast imaging subset of ddsm, 2016. URL `https://wiki.cancerimagingarchive.net/x/lZNXAQ`. 43, 65, 145, 148

R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 74

W. Shen, M. Zhou, F. Yang, C. Yang, and J. Tian. Multi-scale convolutional neural networks for lung nodule classification. In *Information Processing in Medical Imaging: 24th International Conference, IPMI 2015, Sabhal Mor Ostaig, Isle of Skye, UK, June 28-July 3, 2015, Proceedings 24*, pages 588–599. Springer, 2015. 15

Y. Shen, X. Zhong, and F. Y. Shih. Deep morphological neural networks. *arXiv preprint arXiv:1909.01532*, 2019. 82

M. J. Shensa. The discrete wavelet transform: wedding the a trous and mallat algorithms. *IEEE Transactions on signal processing*, 40(10):2464–2482, 1992. 39

H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging*, 35(5):1285–1298, 2016. 54

V. K. Singh, E. M. Mohamed, and M. Abdel-Nasser. Aggregating efficient transformer and CNN networks using learnable fuzzy measure for breast tumor malignancy prediction in ultrasound images. *Neural Computing and Applications*, pages 1–17, 2024. 131

K. Sirinukunwattana, J. P. Pluim, H. Chen, X. Qi, P.-A. Heng, Y. B. Guo, L. Y. Wang, B. J. Matuszewski, E. Bruni, U. Sanchez, et al. Gland segmentation in colon histology images: The glas challenge contest. *Medical image analysis*, 35:489–502, 2017. 99, 126, 137, 148

D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg. SmoothGrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017. 47

L. Song, J. Lin, Z. J. Wang, and H. Wang. An end-to-end multi-task deep learning framework for skin lesion analysis. *IEEE journal of biomedical and health informatics*, 24(10):2912–2921, 2020. 23

P. Song, Z. Yang, J. Li, and H. Fan. DPCTN: Dual path context-aware transformer network for medical image segmentation. *Engineering Applications of Artificial Intelligence*, 124: 106634, 2023. 107

J. C. Souza, J. O. B. Diniz, J. L. Ferreira, G. L. F. da Silva, A. C. Silva, and A. C. de Paiva. An automatic method for lung segmentation and reconstruction in chest X-ray using deep neural networks. *Computer methods and programs in biomedicine*, 177:285–296, 2019. 35

R. Sun, Y. Li, T. Zhang, Z. Mao, F. Wu, and Y. Zhang. Lesion-aware transformers for diabetic retinopathy grading. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10938–10947, 2021. 57

K. Suzuki. Overview of deep learning in medical imaging. *Radiological physics and technology*, 10(3):257–273, 2017. 34

C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 54

C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 55

M. Talo, O. Yildirim, U. B. Baloglu, G. Aydin, and U. R. Acharya. Convolutional neural networks for multi-class brain disease detection using MRI images. *Computerized Medical Imaging and Graphics*, 78:101673, 2019. 35

M. Toğaçar, B. Ergen, and Z. Cömert. COVID-19 detection using deep learning models to exploit Social Mimic Optimization and structured chest X-ray images using fuzzy color and stacking approaches. *Computers in biology and medicine*, 121:103805, 2020. 36

I. O. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, et al. MLP-Mixer: An all-MLP architecture for vision. *Advances in Neural Information Processing Systems*, 34, 2021. 55, 60

H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021a. 54

H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, and H. Jégou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 32–42, 2021b. 77

P. Tschandl, C. Rosendahl, and H. Kittler. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5 (1):1–9, 2018. 97, 126, 137, 149

J. M. J. Valanarasu and V. M. Patel. UNext: MLP-based rapid medical image segmentation network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 23–33. Springer, 2022. 81, 85, 104, 105, 106, 107

J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel. Medical transformer: Gated axial-attention for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*, pages 36–46. Springer, 2021. 22, 58, 85, 107, 110

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 16, 54, 84, 131

H. Wang, P. Cao, J. Wang, and O. R. Zaiane. UCTransNet: rethinking the skip connections in u-net from a channel-wise perspective with transformer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 2441–2449, 2022. 22, 81, 110, 112, 113

L. Wang, Z. Q. Lin, and A. Wong. COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. *Scientific Reports*, 10(1): 1–12, 2020. xiv, 35, 43, 47, 50, 51, 145, 148

W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 568–578, 2021a. 55, 77

X. Wang, M. Xu, J. Zhang, L. Jiang, and L. Li. Deep multi-task learning for diabetic retinopathy grading in fundus images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2826–2834, 2021b. 24

S. Wazir and M. M. Fraz. HistoSeg: Quick attention with multi-loss function for multi-structure segmentation in digital histology images. In *2022 12th International Conference on Pattern Recognition Systems (ICPRS)*, pages 1–7. IEEE, 2022. 85, 110

S. Woo, J. Park, J. Y. Lee, and I. S. Kweon. CBAM: Convolutional block attention module. In *Proceedings of the ECCV*, pages 3–19, 2018. 124, 125, 131

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*, 76:102327, 2022. 84, 105, 106, 107

H. Wu, X. Huang, X. Guo, Z. Wen, and J. Qin. Cross-image dependency modelling for breast ultrasound segmentation. *IEEE Transactions on Medical Imaging*, 2023a. 130

R. Wu, P. Liang, X. Huang, L. Shi, Y. Gu, H. Zhu, and Q. Chang. MHorUNet: High-order spatial interaction UNet for skin lesion segmentation. *Biomedical Signal Processing and Control*, 88:105517, 2024. 20

X. Wu, Y. Feng, H. Xu, Z. Lin, T. Chen, S. Li, S. Qiu, Q. Liu, Y. Ma, and S. Zhang. CTransCNN: Combining transformer and CNN in multilabel medical image classification. *Knowledge-Based Systems*, 281:111030, 2023b. 17

L. Xia, J. An, C. Ma, H. Hou, Y. Hou, L. Cui, X. Jiang, W. Li, and Z. Gao. Neural network model based on global and local features for multi-view mammogram classification. *Neurocomputing*, 536:21–29, 2023. 57

L. Xie, L. Zhang, T. Hu, H. Huang, and Z. Yi. Neural networks model based on an automated multi-scale method for mammogram classification. *Knowledge-Based Systems*, 208:106465, 2020a. 15

Y. Xie, J. Zhang, Y. Xia, and C. Shen. A mutual bootstrapping model for automated skin lesion segmentation and classification. *IEEE Transactions on Medical Imaging*, 39(7): 2482–2493, 2020b. 119

Y. Xie, J. Zhang, Y. Xia, Y. Zhang, and D. Shen. Segformer3d: 3d medical image segmentation with transformers. *Medical Image Analysis*, 87:102859, 2023. 141, 142

C. Xin, Z. Liu, K. Zhao, L. Miao, Y. Ma, X. Zhu, Q. Zhou, S. Wang, L. Li, F. Yang, et al. An improved transformer network for skin cancer classification. *Computers in Biology and Medicine*, 149:105939, 2022. 57

M. Xu, K. Huang, and X. Qi. Multi-task learning with context-oriented self-attention for breast ultrasound image classification and segmentation. In *2022 IEEE 19th ISBI*, pages 1–5. IEEE, 2022. 23, 119, 121, 131, 132

M. Xu, K. Huang, and X. Qi. A regional-attentive multi-task learning framework for breast ultrasound image segmentation and classification. *IEEE Access*, 11:5377–5392, 2023. 119, 121, 130, 131, 132

Z. Xue, B. Xin, D. Wang, and X. Wang. Radiomics-enhanced multi-task neural network for non-invasive glioma subtyping and segmentation. In *International Workshop on Radiomics and Radiogenomics in Neuro-oncology*, pages 81–90. Springer, 2019. 24

Z. Yang, L. Ran, S. Zhang, Y. Xia, and Y. Zhang. EMS-Net: Ensemble of multiscale convolutional neural networks for classification of breast cancer histology images. *Neurocomputing*, 366:46–53, 2019. 16

M. H. Yap, G. Pons, J. Martí, S. Ganau, M. Sentis, R. Zwiggelaar, A. K. Davison, and R. Marti. Automated breast ultrasound lesions detection using convolutional neural networks. *IEEE journal of biomedical and health informatics*, 22(4):1218–1226, 2017. 97, 126, 149

S. Yu, K. Ma, Q. Bi, C. Bian, M. Ning, N. He, Y. Li, H. Liu, and Y. Zheng. MIl-VT: Multiple instance learning enhanced vision transformer for fundus image classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24*, pages 45–54. Springer, 2021. 57

F. Yuan, Z. Zhang, and Z. Fang. An effective CNN and Transformer complementary network for medical image segmentation. *Pattern Recognition*, 136:109228, 2023. 17

M. A. E. Zeid, K. El Bahnasy, and S. Abo Youssef. Multiclass colorectal cancer histology images classification using vision transformers. In *2021 Tenth International Conference on Intelligent Computing and Information Systems (ICICIS)*, pages 224–230. IEEE, 2021. 77

W. Zeng, W. Fan, R. Chen, et al. Accurate 3d kidney segmentation using unsupervised domain translation and adversarial networks. In *2021 IEEE 18th ISBI*, pages 598–602. IEEE, 2021. 119

Z. Zeng, W. Xie, Y. Zhang, and Y. Lu. RIC-Unet: An improved neural network based on Unet for nuclei segmentation in histology images. *IEEE Access*, 7:21420–21428, 2019. 21

B. Zhang, J. Gong, D. Jin, L. Zhang, and X. Zhang. Large multimodal models for medical applications: A survey. *arXiv preprint arXiv:2303.14794*, 2023. 143

G. Zhang, K. Zhao, Y. Hong, et al. SHA-MTL: soft and hard attention multi-task learning for automated breast cancer ultrasound image segmentation and classification. *Int. j. comput. assist. radiol. surg.*, 16:1719–1725, 2021a. 23, 119, 121, 130

H. Zhang, J. Yang, K. Zhou, Z. Chai, J. Cheng, S. Gao, and J. Liu. BioNet: Infusing Biomarker Prior into Global-to-Local Network for Choroid Segmentation in Optical Coherence Tomography Images. *arXiv preprint arXiv:1912.05090*, 2019a. 21

H. Zhang, J. Lian, Z. Yi, R. Wu, X. Lu, P. Ma, and Y. Ma. HAU-Net: Hybrid CNN-transformer for breast ultrasound image segmentation. *Biomedical Signal Processing and Control*, 87:105427, 2024. 21

J. Zhang, Y. Xie, Q. Wu, and Y. Xia. Medical image classification using synergic deep learning. *Medical image analysis*, 54:10–19, 2019b. 16

Y. Zhang and Q. Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(12):5586–5609, 2021. 119

Y. Zhang, H. Li, J. Du, J. Qin, T. Wang, Y. Chen, B. Liu, W. Gao, G. Ma, and B. Lei. 3d multi-attention guided multi-task learning network for automatic gastric tumor segmentation and lymph node classification. *IEEE Transactions on Medical Imaging*, 40(6): 1618–1631, 2021b. 24

Y. Zhang, H. Liu, and Q. Hu. TransFuse: Fusing transformers and CNNs for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*, pages 14–24. Springer, 2021c. 84, 105, 106, 107

C. Zhao, R. Shuai, L. Ma, W. Liu, and M. Wu. Improving cervical cancer classification with imbalanced datasets combining taming transformers with T2T-ViT. *Multimedia tools and applications*, 81(17):24265–24300, 2022. 57

S. Zhong, C. Tu, X. Dong, Q. Feng, W. Chen, and Y. Zhang. MsGoF: breast lesion classification on ultrasound images by multi-scale gradational-order fusion framework. *Computer Methods and Programs in Biomedicine*, 230:107346, 2023. 132

Y. Zhou, H. Chen, Y. Li, Q. Liu, X. Xu, S. Wang, P.-T. Yap, and D. Shen. Multi-task learning for segmentation and classification of tumors in 3d automated breast ultrasound images. *Medical Image Analysis*, 70:101918, 2021. 23

Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 3–11. Springer, 2018. 81, 130, 131, 132

M. Zhu, Z. Chen, and Y. Yuan. DSI-Net: Deep synergistic interaction network for joint classification and segmentation with endoscope images. *IEEE Transactions on Medical Imaging*, 40(12):3315–3325, 2021. 24

W. Zhu, J. Tian, M. Chen, L. Chen, and J. Chen. Mss-unet: A multi-spatial-shift mlp-based unet for skin lesion segmentation. *Computers in Biology and Medicine*, 168:107719, 2024. 20

X. Zhu, H. Hu, H. Wang, J. Yao, W. Li, D. Ou, and D. Xu. Region aware transformer for automatic breast ultrasound tumor segmentation. In *International Conference on Medical Imaging with Deep Learning*, pages 1523–1537. PMLR, 2022. 21

J. Zhuang, J. Cai, R. Wang, J. Zhang, and W.-S. Zheng. Deep knn for medical image classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*, pages 127–136. Springer, 2020. 11