Deep Learning Models for the Automatic Segmentation of Lung Nodules

Vinícius Batista Ferreira dos Reis
Master in Computer Science
Department of Computer Science
Faculty of Sciences of the University of Porto
2024

Deep Learning Models for the Automatic Segmentation of Lung Nodules

Vinícius Batista Ferreira dos Reis

Dissertation carried out as part of the Master in Computer Science

Department of Computer Science

2024

Supervisor

Prof. Dr. Helder Filipe Oliveira, Faculty of Sciences

External Supervisor

MSc. Luís Emanuel Pereira Pinto Fernandes, INESC TEC

UNIVERSIDADE DO PORTO

MASTERS THESIS

Deep Learning Models for the Automatic Segmentation of Lung Nodules

Author:
Vinícius BATISTA FERREIRA DOS REIS

Supervisor:
Helder FILIPE OLIVEIRA

A thesis submitted in fulfilment of the requirements
for the degree of MSc. Computer Science

at the

Faculdade de Ciências da Universidade do Porto
Departamento de Ciência de Computadores

December 11, 2024

“ All life is an experiment. The more experiments you make the better. ”

Ralph Waldo *Emerson*

Acknowledgements

First, I thank my supervisors, Hélder Oliveira and Luís Fernandes, for all their support, guidance, and patience. Without them, I wouldn't have known where to begin this work.

I also thank Tânia, Hugo, and everyone else who participated in the update discussions and research direction. Directly and indirectly, they were essential for understanding the research and gaining insights on the path to follow.

I also thank the Universidado do Porto (University of Porto) and INESC TEC for the space and resources provided to complete this work.

Finally, I thank all my friends and family who helped me on this journey, especially my partner in crime, Sonia, who believes in and supports all my ideas.

This work is financed by National Funds through the FCT - Fundação para a Ciência e a Tecnologia, I.P. (Portuguese Foundation for Science and Technology) within the project LUCCA, with reference 2022.03488.PTDC.

Abstract

With millions of cases annually, cancer is one of the most concerning diseases for healthcare systems worldwide. Among the many types of cancer, lung cancer is the deadliest of all. Due to its aggressiveness and late detection, approximately 81% of the nearly 2 million lung cancer cases result in death—an alarming and extremely high figure that demands effort and attention from various scientific fields in the search for better treatments and diagnostic methods.

The focus of this work is precisely to support the diagnostic processes, particularly the segmentation of pulmonary nodules, a task that requires time and skilled professionals and directly impacts the correct treatment of patients.

To achieve this, the study begins with an overview of lung cancer, including introductory concepts about the Deep Learning techniques used in this work. These techniques aim to identify nodules in images from computed tomography systems and create a map with their dimensions and locations within lung tissue. A review of related works is then conducted to understand the current state of research on pulmonary nodule segmentation, the best techniques, and the most significant results. Before initiating the series of experiments, the general methodology, as well as the data and preliminary processes, are presented.

Eight experiments are conducted to determine which components are more or less important, how attention mechanisms improve training performance, and how the correct use of data is essential for generating a robust model. Finally, the proposed model is presented, achieving a final performance of 61.4% DICE Coefficient, along with the conclusions of the study.

The data used for training and validating the models were collected from public datasets (LIDC and Medical Decathlon) and through a partnership with Hospital de São João in Porto, Portugal.

Keywords: Lung Cancer, Pulmonary Nodule Segmentation, Deep Learning, Automatic Nodule Segmentation

Resumo

Com milhões de casos anualmente, câncer é uma das doenças mais preocupantes para os serviços de saúde em todo o mundo e, dentre os mais variados tipos existentes, câncer de púlmão é o mais letal de todos. Devido sua agressividade e detecção tardia, cerca de 81%, dos quase 2 milhões de casos de câncer de pulmão, acabam em óbitos, um número alarmante e extremamente elevado que demanda esforço e atenção de várias áreas científicas na busca de melhores tratamentos e diagnósticos.

O foco deste trabalho se dará justamente em apoiar os processos de diagnóstico, em especial a segmentação dos nódulos pulmonares, algo que demanda tempo e profissionais capacitados e impacta diretamente o correto tratamento dos pacientes.

Para isso, primeiro é feito um levantamento básico do que é o câncer de pulmão, conceitos introdutórios sobre as técnicas de Deep Learning a serem utilizadas durante o trabalho, que irão aprender a identificar os nódulos em imagens vindas de sistemas de tomografia computadorizada, bem como criar um mapa de com as dimensões e localizações no tecido pulmonar. É realizado então um levantamento de trabalhos relacionados para entender o estado corrente dos estudos de segmentação de nódulos pulmonares, melhores técnicas e resultados. Antes de iniciar a série de experimentações é apresentada a metodologia geral e os dados e processos prévios que foram executados.

Oito experimentos são então conduzidos a fim de detectar quais componentes são mais ou menos importantes, como a atenção apoia na melhoria da performance dos treinos e como o correto uso dos dados é essencial para a geração de um modelo coerente. Por fim é apresentado o modelo proposto, com performance final de 61.4% de DICE Coefficient e as conclusões dos estudos.

Os dados utilizados durante o treino e validação dos modelos foram coletados de conjuntos públicos (LIDC e Medical Decathlon) e também através de parceria com o Hospital de São João, da cidade do Porto.

Palavras-chave: Câncer de Pulmão, Segmentação de Nódulos Pulmonares, Deep Learning, Segmentação Automática de Nódulos

Contents

Acknowledgements v

Abstract vii

Resumo ix

Contents xi

List of Figures xiii

List of Tables xv

Glossary xvii

1 Introduction 1
- 1.1 Motivation . . . 2
- 1.2 Objectives and Contributions . . . 3
- 1.3 Structure . . . 4

2 Main Concepts 5
- 2.1 Lungs . . . 5
- 2.2 Lung Cancer . . . 6
  - 2.2.1 Lung Cancer Types . . . 7
- 2.3 Computer-aided Lung Cancer System . . . 7
  - 2.3.1 Computed Tomography . . . 7
  - 2.3.2 Image Segmentation . . . 9
- 2.4 Deep Learning . . . 10
  - 2.4.1 CNN . . . 10
  - 2.4.2 Auto-encoder . . . 12
  - 2.4.3 GAN . . . 13
- 2.5 Datasets . . . 13
- 2.6 Metrics . . . 15

3 Related works 17
- 3.1 Models . . . 17
  - 3.1.1 U-Net Based . . . 18
  - 3.1.2 Non U-Net Based . . . 25

3.2 Summary 32

4 Materials and Methods **35**
4.1 Data 35
4.1.1 Data Analysis 35
4.2 Pre-Processing 40
4.3 Deep Learning and Methodology 42

5 Experiments and Proposed Framework **45**
5.1 Initial Benchmark 45
5.2 Data Augmentation Comparison 47
5.3 Data Selection Test 49
5.4 Attention Mechanisms 50
5.5 Transformer Test 56
5.6 Interactive Segmentation Test 57
5.7 Limitations and failed experiments 60
5.8 Evaluations and Proposed Framework 60
5.8.1 Evaluations 61
5.8.2 Proposed Model 62

6 Conclusions **65**

Bibliography **67**

List of Figures

1.1 Top 10 Causes of Deaths in 2000 and 2019 . . . 2
1.2 Incidence and Mortality of Lung Cancer in Europe Union . . . 3

2.1 Gross Anatomy of the Lungs . . . 5
2.2 Parietal and Visceral Pleurae of the Lungs . . . 6
2.3 Example of a basic CAD System Pipeline for lung cancer treatment . . . 8
2.4 CT Scan Basic Function . . . 8
2.5 Lung nodule Segmentation types . . . 9
2.6 Nodule Segmentation types . . . 10
2.7 An example of CNN architecture for image classification . . . 11
2.8 An example of a 2D convolution operation . . . 11
2.9 U-Net architecture . . . 12
2.10 Example of basic auto-encoder architecture . . . 12
2.11 Denoising autoencoder . . . 13
2.12 Example of a simple GAN architecture . . . 13

3.1 Architecture of Lung_PAYNet . . . 18
3.2 Segmentation example . . . 19
3.3 ViFUnet architecture . . . 20
3.4 Proposed encoder-decoder with ASPP architecture . . . 20
3.5 LL, LH, HL and HH bands example . . . 21
3.6 Wavelet U-Net++ architecture . . . 21
3.7 Two Swin Transformer blocks . . . 22
3.8 CA-UNet Attention Module . . . 22
3.9 CA-UNet architecture . . . 22
3.10 Architecture of DEHA-Net . . . 23
3.11 Full framework of DEHA-Net . . . 24
3.12 MDFN . . . 24
3.13 MSCFS structure . . . 24
3.14 DPBET structure . . . 25
3.15 CSE-GAN block diagram . . . 26
3.16 sScE Module . . . 26
3.17 cSsE Module . . . 27
3.18 CSE Module . . . 27
3.19 Soft Mask Process . . . 28
3.20 Architecture of the SoftGAN . . . 28
3.21 DB-ResNet overview . . . 29
3.22 Semantic generation . . . 30

3.23 Example the synthetic generation . . . 30
3.24 Proposed Network architecture by Qin et al . . . 31
3.25 The 3D deep MTL CNN architecture . . . 31
3.26 End-to-end multi-task framework with MSN and JCN . . . 32

4.1 Sum of Total Annotations by Number of Radiologist . . . 36
4.2 Total number of nodules . . . 37
4.3 Malignancy distribution . . . 37
4.4 Nodule texture distribution . . . 38
4.5 Nodules features distribution . . . 39
4.6 Original CT scan . . . 40
4.7 Example of different crops with: (a) 512x512; (b) 224x224; (c) 96x96; (d) 64x64 41
4.8 Steps of pre-processing the CT Scan . . . 41
4.9 Segmentation Mask Example . . . 42

5.1 Evolution of Dice over Epochs: (a) Training Dice; (b) Validation Dice . . . 46
5.2 Example of the data augmentations transformations . . . 48
5.3 Evolution of Dice over Epochs for Augmentation Comparison: (a) Training Dice; (b) Validation Dice . . . 49
5.4 Basic U-Net with attention block . . . 51
5.5 Attention Blocks Modules. (a) ASPP; (b) PYB; (c) SEB; (d) SAM; (e) ViT . . . 52
5.6 SAMAttentionUNet architecture . . . 54
5.7 MixAttentionUNet architecture . . . 54
5.8 R2AttentionUNetSAM architecture . . . 55
5.9 Point click annotation example. (a) Original ct slice;(b) Annotation . . . 58
5.10 Evaluations Predictions . . . 63

List of Tables

2.1 Dataset Summary 15
2.2 Metric Summary 15

3.1 Frameworks Summary 33

5.1 Initial Benchmarks 46
5.2 Data Augmentation Techniques 47
5.3 Data Augmentation Combinations and Results 48
5.4 Results for Data Selection 50
5.5 Results for Attention Bottleneck for Crop Size 512x512 52
5.6 Results for Attention Bottleneck for Crop Size 96x96 53
5.7 Results for Attention Bottleneck for Crop Size 64x64 53
5.8 Results of Attention Mechanisms on Skip Connections for Crop Size 96x96 . 55
5.9 Results of Attention Mechanisms on Skip Connections for Crop Size 96x96 . 56
5.10 Results of Attention Mechanisms on Skip Connections for Crop Size 64x64 . 56
5.11 Results of Swin-UNet and Swin-UNet V2 57
5.12 Results of Interactive Segmentation Tests for Crop Size 512x512 58
5.13 Results of Interactive Segmentation Tests for Crop Size 96x96 59
5.14 Results of Interactive Segmentation Tests for Crop Size 64x64 59
5.15 Results of Final Evaluation for Crop Size 512x512 on LIDC 61
5.16 Results of Final Evaluation for Crop Size 96x96 on LIDC 61
5.17 Results of Final Evaluation for Crop Size 64x64 on LIDC 62
5.18 Results of Proposed Model Evaluation 63

Glossary

AC	Artificial crawlers
AE	Auto-encoders
ANODE09	Automatic Nodule Detection 2009
BCE	Binary Cross Entropy
BN	Batch Normalization
CAD	Computer Aided Diagnostic
CADe	Computer-Aided Detection
CADx	Computer-Aided Diagnosis
CDC	Centers for Disease Control and Prevention
CF-CNN	Central Focused Convolutional Neural Networks
cGAN	conditional Generative Adversarial Network
CNN	Convolutional Neural Network
CR	Computed Radiography
CT	Computed Tomography
CTIL	CT Image Library
CTV	Clinical Target Volume
DAE	Denoising Auto-encoders
DBN	Deep Belief Network
DL	Deep Learning
DLGC	Deep Learned Prior Based Graph Cut
DSB2017	Data Science Bowl 2017

DSC	Dice Similarity Coefficient
ECIS	European Cancer Information System
ELCAP	The Early Lung Cancer Action Program
FCN	Fully Convolutional Neural Networks
FDA	Food and Drug Administration
FNIH	Foundation for the National Institutes of Health
FPR	False Positive Reduction
G-Convs	Group Convolutions
GAN	Generative Adversarial Networks
GGO	Ground-Glass Opacity
GTV	Gross Tumor Volume
HCRF	Didden Conditional Random Field
HSN	Hierarchical saliency network
HU	Hounsfield Units
ILND	Indian lung nodule dataset
INCA	Instituto Nacional de Câncer
INE	Instituto Nacional de Estatística
JCN	Joint Classification Sub-network
LBP	Local Binary Pattern
LDA	linear discriminate analysis
LIDC/IDRI	Lung Image Database Consortiums Images Collection
LNSM	Lung Nodules with Soft Masks dataset
LUNA16	lung nodule analysis 2016
Lung-RADS	Lung CT Reporting And Data System
MAP	Maximum a posteriori
ML	Machine Learning
MRI	Magnetic Resonance Imaging

MSN	Multi-Channel Segmentation Sub-network
MTL	Multi-task learning
MV-CNN	multi-view convolutional neural networks
NCI	National Cancer Institute
NCI	National Cancer Institute
NELSON	NederlandLeuvens Longkanker Onderzoek
NLST	National Lung Screening Trial
NSCLC	non-small cell lung cancer
OAR	Organ-at-risk
PRN	Progressive resolution network
PSOm	Particle swarm optimizer
RD	Rose diagram
RNN	Recurrent Neural Network
RPN	Region proposal network
RT	Radiation Therapy
SCLC	small cell lung cancer
SDAE	Stacked denoising autoencoder
SFFSA	sequential flood feature selection algorithm
SN	solid nodules
SSN	sub-solid nodules
TianChi	ALIBABA Tianchi challenge
WHO	World Health Organization

Chapter 1

Introduction

Lung cancer is one of the major health problems worldwide, according to the World Health Organization (WHO) [1]. Lung, tracheal, and bronchial cancers rank as the sixth leading cause of death globally, with an increase from 1.2 million deaths in 2000 to 1.6 million deaths in 2019. According to the WHO [2], lung cancer is the second most common cancer globally, trailing only breast cancer. However, it is the most lethal type of cancer, with nearly double the deaths compared to colorectal cancer, which is the second-ranking cancer in terms of mortality.

In the European Union, data from the European Cancer Information System (ECIS) shows that lung cancer represents 11.6% of all cancer cases and 19.5% of cancer-related deaths. Also, it is the fourth most frequently occurring cancer, following prostate, breast, and colorectal cancers, and stands as the leading cause of cancer death. Moreover it is estimated that new cases increased by 2.3% and deaths by 2.4% from 2020 to 2022. [3].

When examining more specific data for Brazil and Portugal, similarities with the global and Europe Union scenario emerge.

In Brazil, according to the Instituto Nacional do Câncer (INCA) [4], lung cancer accounts for 18% of deaths among all types of cancer. In Portugal, according to the Instituto Nacional de Estatística (INE) [5], approximately 15% of cancer deaths are characterized as being related to the trachea, bronchi, and lungs.

When analysing this alarming numbers, it is clear to see that lung cancer is a major issue of general public health. Therefore, it is crucial to develop new solutions for the early detection of cancer so that we can potentially decrease the number of deaths that are associated with this disease.

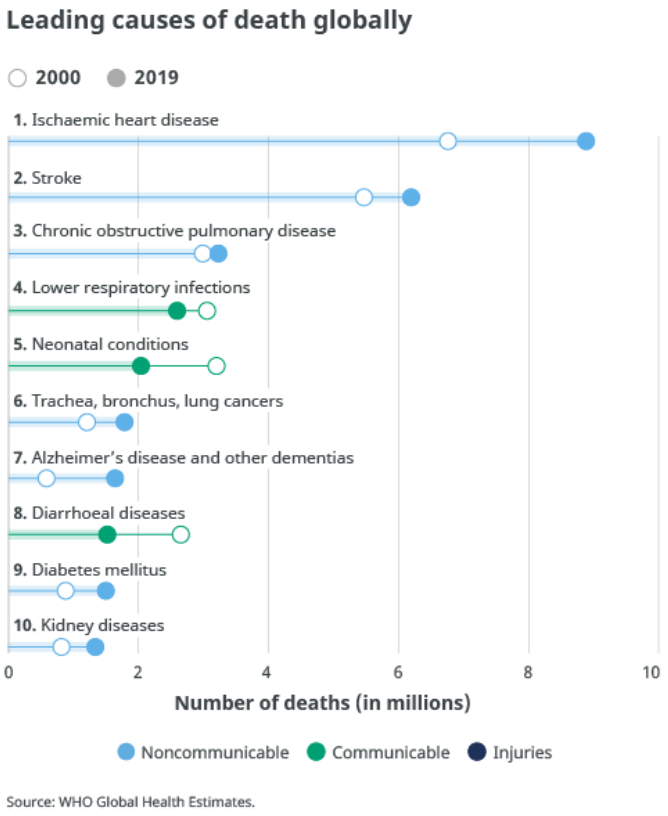

FIGURE 1.1: Top 10 Causes of Deaths in 2000 and 2019. From [1].

1.1 Motivation

When observing the data, it is possible to see that lung cancer has a high mortality rate relative to the total number of cases, with approximately 81% [2] of diagnosed individuals succumbing to this disease. This is largely attributed to late diagnoses and the significant challenge of analyzing medical images. As presented by the American Lung Association [6], in the United Sates, 47% of the lung cancer cases are diagnosed in late stages. In Ciompi et al. [7], a comparative study involving four healthcare professionals with extensive experience demonstrated an average accuracy of 72.9% in the human classification of cancerous nodules. This shows that even for experienced specialists, there is an almost 30% rate of misclassification of lung nodules.

The use of computer systems then emerges as an important tool to help reduce diagnosis time and increase confidence in the correct classification of nodules, in the study by Singh et al. [9], it was found that the use of computational models tends to improve the specificity of mass classification in breast tissue by 20% to 43%. Showing that Computer-Aided System (CAD or CAD System) have significant potential in supporting healthcare

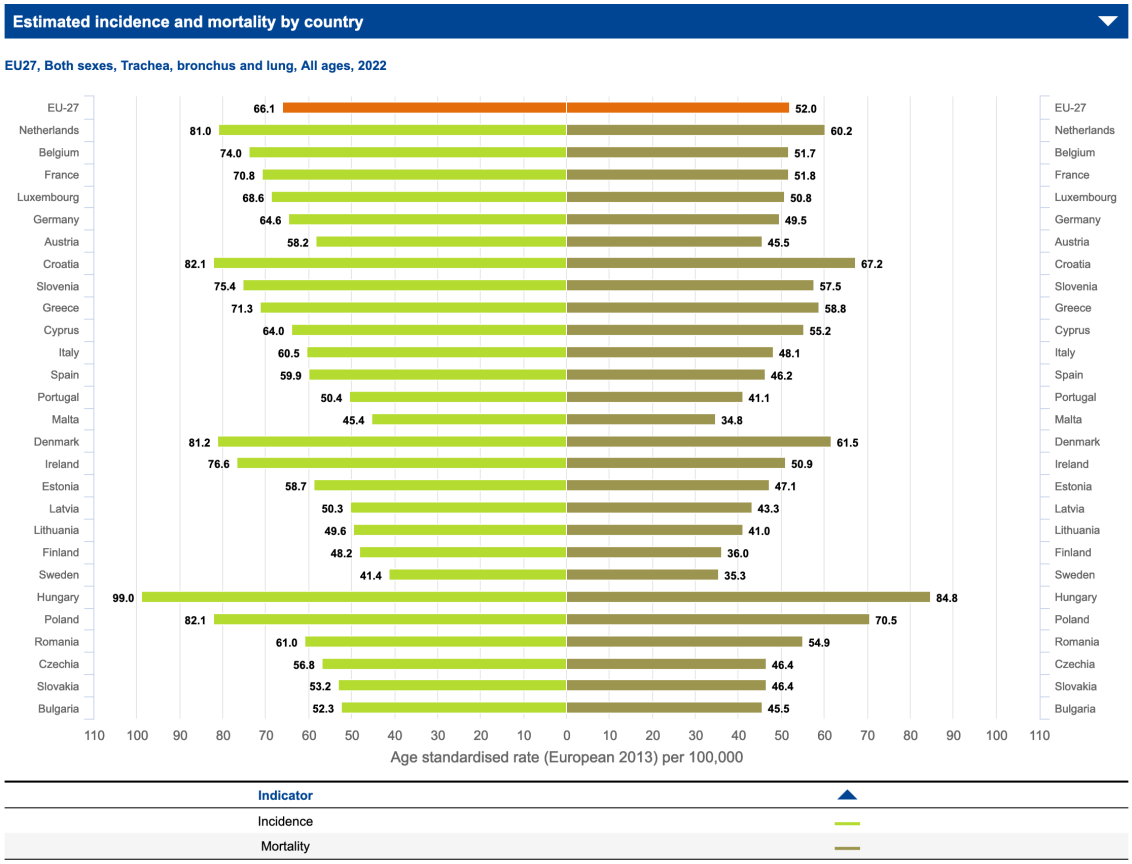

FIGURE 1.2: Incidence and Mortality of Lung Cancer in Europe Union. From [8].

professionals, providing ease and confidence, simplifying and expediting processes.

In an attempt to develope a new CAD System, this work aims to expand the understanding of nodule segmentation in lungs, assess the current state-of-the-art, and investigate methods and solutions to support healthcare in accurately identifying cancerous nodules in the lungs with the use of deep learning, detailed at section 2.4.

1.2 Objectives and Contributions

The main objective of this work is to develop a Deep Learning framework capable of automatically segmenting lung nodules, aiming to assist healthcare professionals in achieving greater speed and efficiency in lung cancer treatment. Initially, a comprehensive study of the current state-of-the-art in nodule segmentation will be conducted to identify the techniques and processes used, as well as potential opportunities for innovation and limitations. In the field of nodule segmentation, there is a wide variety of approaches and

scopes of operation, so it is necessary to consolidate these efforts for an assessment of the current scenario.

Secondly, this work will develop an automatic framework for general use, where limitations related to nodule types, locations, or densities are addressed and do not become obstacles for use in any given scenario. Finally, a range of numerous experiments and comparisons will be conducted to build a centralized and extensive database of nodule segmentation frameworks, aiming to support future work.

1.3 Structure

The chapter sequence will unfold as follows: Chapter 2 will address basic definitions regarding lung anatomy, lung cancer, deep learning, different tasks of deep learning for the analysis of computed tomography scans, and other evaluation methods. Chapter 3 will analyze the state-of-the-art, providing a literature review and presenting data on the main used methods. Chapter 4 will explain the different available datasets and the required preprocessing steps. Chapter 5 is where the experiments, discussions and the proposed method are presented. Finally, Chapter 6 will present the conclusion of the work and additional considerations.

Chapter 2

Main Concepts

In this chapter of the thesis, the main definitions will be presented to establish the object of study. This will include basic concepts of lung anatomy [2.1], an overview of lung cancer and its impact on the organ [2.2], an introduction to the lung CAD System and the various tasks related to lung cancer [2.3], a brief overview of deep learning [2.4], its most commonly used architectures in nodule segmentation [2.4.1], and finally, an exploration of methodologies different than deep learning.

2.1 Lungs

Located at the chest cavity, the lungs are the main organ of the respiratory system, consisting of two parts shaped like pyramids with the function of exchange oxygen and carbon dioxide between the environment and the body. [10]

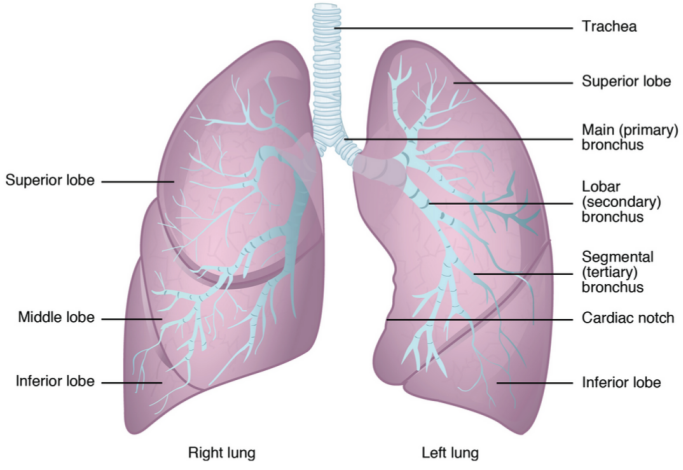

FIGURE 2.1: Gross Anatomy of the Lungs. From [10].

As depicted in Figure 2.1, the two parts are connected to the trachea through the right and left bronchi. The left side is smaller and has a cardiac notch to accommodate the heart, while the right side is slightly wider but shortened due to the higher position of the right diaphragm to make room for the liver.

Both parts are composed of lobes, with 3 on the right and two on the left. On the right, there are the lower, middle, and upper lobes, while on the left, there are the lower and upper lobes. Each lobe has multiple bronchopulmonary segments that receive air, and is supplied with blood by its own artery.

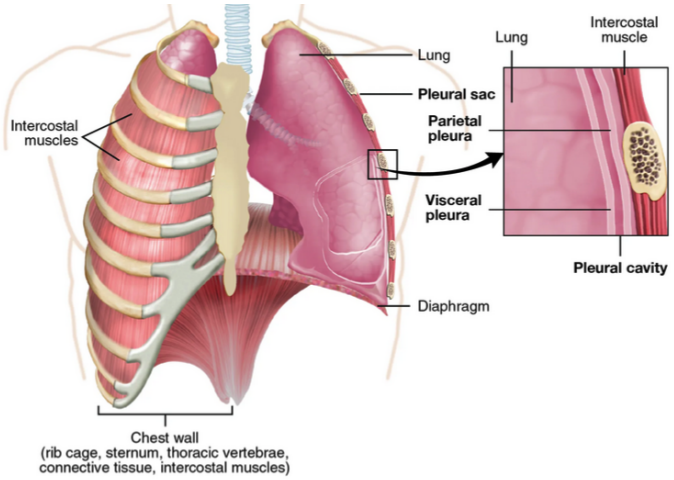

FIGURE 2.2: Parietal and Visceral Pleurae of the Lungs. From [10].

Each lung is enveloped by a membrane called the pleura, which has two layers, the visceral and the parietal. The visceral layer is directly superficial to the lung, while the parietal layer connects with the walls of the thoracic cavity. These layers are connected to each other through the hilum.

2.2 Lung Cancer

The National Cancer Institute defines cancer not as a single disease but as a collection of related diseases characterized, in brief, by uncontrolled cell growth in virtually any part of the body, where the genes of the cells no longer function correctly [11].

Lung cancer specifically occurs when this uncontrolled growth begins in the lung region, although it is not necessarily confined to it, as it can spread throughout the body. Typically, lung cancer is categorized into two types, Small Cell Lung Cancer (SCLC) and Non-Small Cell Lung Cancer (NSCLC) [12], which are detailed further below. Regarding

risk factors for lung cancer smoking is widely recognized as the foremost among them, accounting for approximately 80% of cases. [13].

Other risk factors include, but are not limited to [14]:

- Secondhand smoke;
- Exposure to radon;
- Exposure to asbestos;
- Personal or family history of lung cancer.

2.2.1 Lung Cancer Types

As mentioned earlier, there are two types of lung cancer, SCLC and NSCLC. Both share the same causes, risk factors, and many of the same types of treatment. However, it is worth highlighting some distinctions between them.

NSCLC is the more common type, accounting for about 80% to 85% of cases [12]. It is less aggressive and develops more slowly. Typically, it is classified on a staging scale ranging from 0 (undetected) to 5 (extensively spread to other areas of the body). NSCLC offers a wide range of possible treatments, such as chemotherapy, surgery, radiotherapy, and combinations of them all. [15].

SCLC represents about 10% to 15% of cases [12]. It is highly aggressive and has a more accelerated development. Its classification usually has only two stages: limited stage (confined to the lung) or extensive stage (spread to other areas). Although surgery is an option, the most common treatments are radiotherapy and chemotherapy [16].

2.3 Computer-aided Lung Cancer System

A CAD system for lung cancer treatment can be divided into two main parts, the first involves obtaining lung images through Computed Tomography (CT), and the second involves image segmentation, with the latter having some subdivisions, an example of a basic pipeline is show at Figure 2.3. More details are provided below.

2.3.1 Computed Tomography

Computed Tomography is a technique for producing cross-sectional images of a patient's body through the use of x-rays, developed by Sir Godfrey Hounsfield (Nobel Prize for

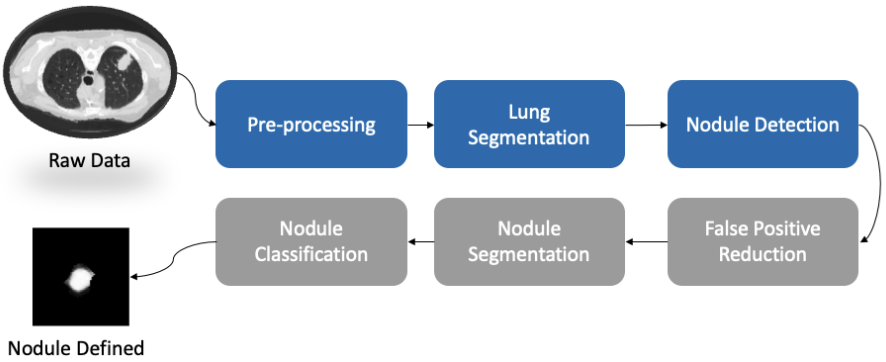

FIGURE 2.3: Example of a basic CAD System Pipeline for lung cancer treatment

Medicine or Physiology in 1979, shared with the physicist Allan Cormack) at EMI research laboratories at the end of the 1960s and early 1970s. The main concept is that by placing the patient in a tubular gantry surrounded by a ring of X-ray sensitive detectors that, when rotated, the X-Ray Beam comes into contact with the tissues and their different densities. They are recorded by the sensitive detector and processed and transformed into 2D or 3D images that can be used to assess patient's problem and to develop algorithms in machine learning (ML) or deep learning (DP). [17]

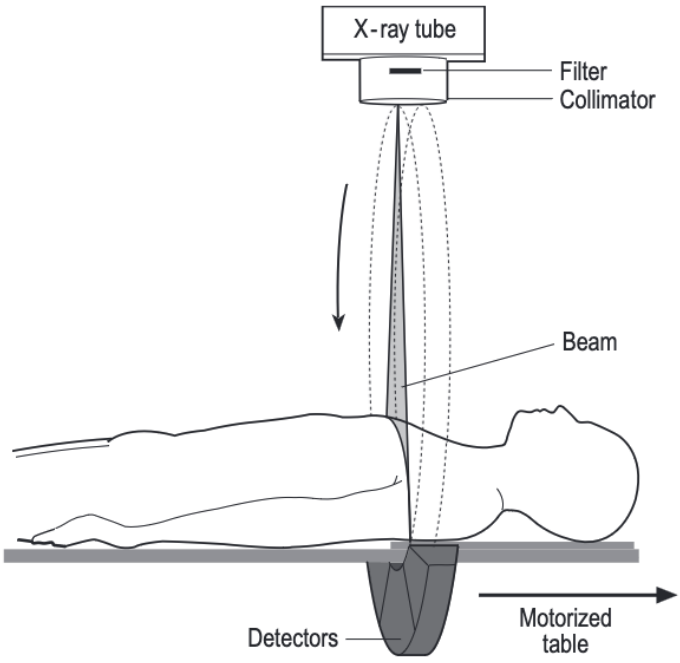

FIGURE 2.4: CT Scan Basic Function. From [18].

2.3.2 Image Segmentation

There are at least five categories of tasks performed using images derived from the CT scan in machine learning or deep learning algorithms. Some works may include more categories, remove, or modify the nomenclature, however for this work they will be defined as show in Figure 2.3.

Lung segmentation is usually, the first task performed in a complete pipeline for evaluating lung images, this step aims to separate the pulmonary parenchyma from other tissues and organs, like demonstrated in Figure 2.5. In other words, it visually isolates the lungs from any other anatomic structures to increase accuracy in subsequent steps.

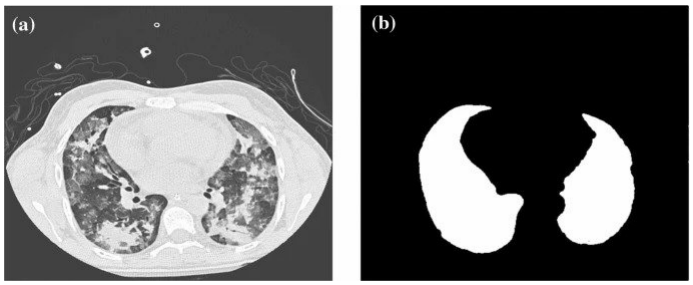

FIGURE 2.5: Lung nodule Segmentation types: **a)** before segmentation, **b)** after segmentation. From [19]

Lung nodule detection is dedicated to identifying all nodules present in lung images with the highest fidelity regarding their characteristics, sizes, and locations to facilitate a complete understanding for subsequent stages. Despite the lung nodule detection step attempting to do this in the best and most accurate way possible, there are still errors and false positives that occur, something expected in the use of DL methods. For this, there is the false positive reduction step that aims to remove incorrectly detected nodules.

An important step and the main subject of this work, nodule segmentation aims to characterize the detected nodule on location, size and density. The classification is divided into three types, that can be seen in Figure 2.6 : Solid nodules, which are the most common and densest type; ground-glass opacity (GGO) or non-solid nodules, with a more opaque and fuzzy appearance; and part-solid nodules, which represent a middle ground between the first two [20].

Finally, lung nodule classification is the classification regarding the malignancy of the nodule. This is crucial for effective and accurate medical treatments in patient diagnosis.

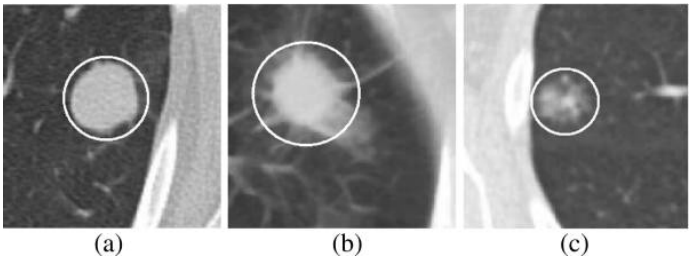

FIGURE 2.6: Nodule Segmentation types: **a)** solid, **b)** part-solid, and **c)** non-solid or GGO. From [21]

2.4 Deep Learning

The main goal of the proposed work is the employment of deep learning techniques in the segmentation of lung nodules in CT images. Descriptions of deep learning, its main architectures, and frameworks that will be used in the experiments will be provided here, as well as a brief summary of the main of those used by other works listed in Chapter 3.

Being classified as a subset of artificial intelligence within machine learning, deep learning is a set of algorithms that uses multiple layers in their neural networks, with origins historically traced to the work of Walter Pitts and Warren McCulloch in 1943 [22], followed by the first known work on deep networks by Ivakhnenko et al. in 1965 [23], until reaching AlexNet, a more contemporary network model that gained fame in the early 2010s by winning the ImageNet competition [24], a object recognition challenge.

There can be numerous definitions of deep learning, as described by Zhang et al. [25]. However, broadly speaking, deep learning can be defined as a class of machine learning algorithms that use multiple hierarchical layers for feature extraction and data representation.

There are various architectural forms of deep learning algorithms, with the following being the main ones used in medical image works, especially those related to lung cancer tasks. Although presented individually, many works use a combination of more than one technique or variations of the same.

2.4.1 CNN

Convolutional Neural Network (CNN) or ConvNet [26] is a recognized and popular architecture for tasks such as image classification and computer vision. It typically consists of a combination of convolutional layers, activation functions, pooling layers, and fully connected layers ([27], [28]), as shown in the example in Figure 2.7.

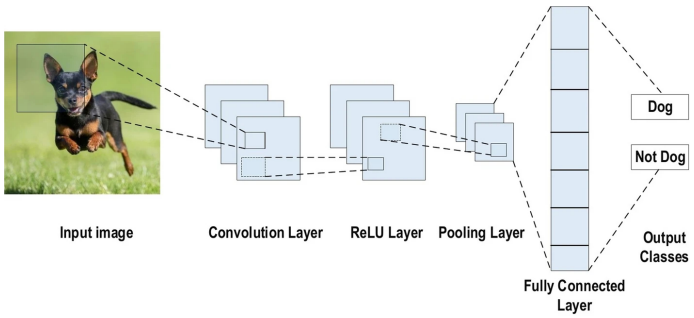

FIGURE 2.7: An example of CNN architecture for image classification. From [28]

A convolution layer is responsible for generating a feature map through the dot product between two matrices: the kernel or filter and the matrix of the input image, as shown in the example in the Figure 2.8.

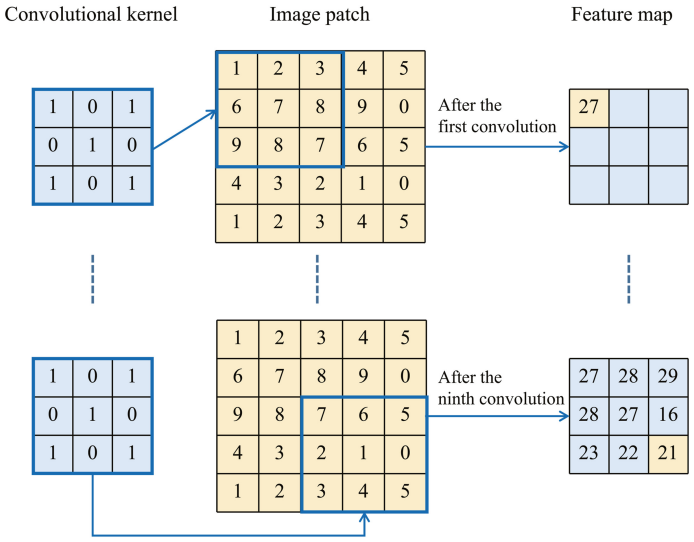

FIGURE 2.8: An example of a 2D convolution operation. From [27]

The pooling layer's main objective is to select and reduce the feature maps generated in the convolutional layers using statistical techniques. This results in a reduction in the quality of the original image but also reduces computational usage.

Finally, the fully connected layer has each neuron connected to all neurons in the previous layers, much like traditional neural networks, serving as the final part of the architecture and the classifier for the input object [28].

There is a wide variety of possible combinations and other techniques for constructing CNNs, but the concepts previously presented are always represented in some form.

In medical image segmentation and its use for the detection of cancerous nodules, the U-Net [29], as shown in Figure 2.9, is one of the most popular CNN architectures. Another example is the ResNet [30], as well as variations of both.

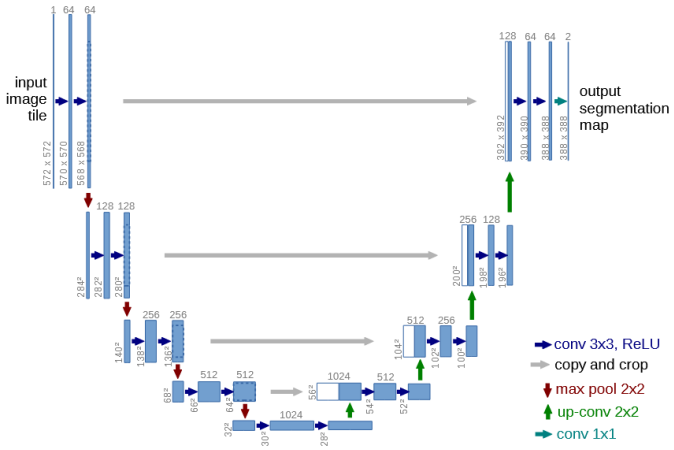

FIGURE 2.9: U-Net architecture where the blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows denote the different operations. From [29]

2.4.2 Auto-encoder

Introduced initially by Rumelhart et al. [31], we can use the definition by Hinton et al.[32] that an autoencoder (AE) has the main purpose of using a set of recognition weights to convert a specific input into an internal code that will be later reconstructed into an output using generative weights. In other words, given an input (encode), an intermediate representation is generated, which will undergo an attempt at reconstruction (decode).

The structure of an AE is divided into an input block, a hidden layer, and an output block, as seen in the Figure. 2.10

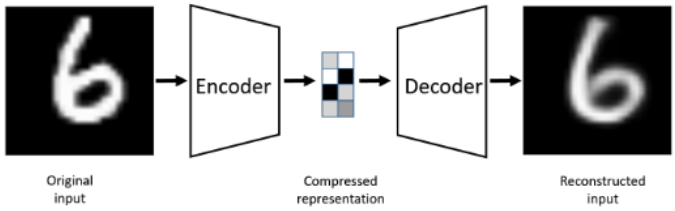

FIGURE 2.10: Example of basic auto-encoder architecture. From [33]

Similar to CNNs, autoencoders have some variations, such as denoising autoencoders (DAE) [34], which can be used for error correction by introducing noise to the input, as illustrated in Figure 2.11. Variational autoencoders (VAE) [35] make use of probability distributions, among other types.

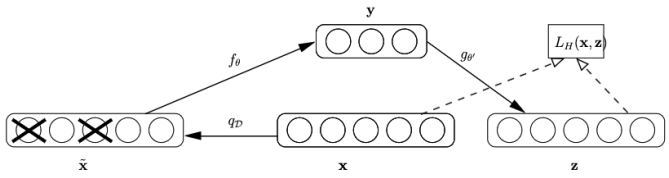

FIGURE 2.11: An example x is corrupted to $\tilde{x}$. The auto-encoder then maps it to y and attempts to reconstruct x. From [34]

2.4.3 GAN

Generative Adversarial Networks (GANs) were initially introduced by Goodfellow et al. [36] in 2014. The framework consists of two neural networks that are trained simultaneously and in competition with each other. One network is responsible for the generative part, while the other handles the discriminative part. In other words, one network tries to produce results, and the other tries to evaluate the truthfulness of such productions working in an adversarial manner.

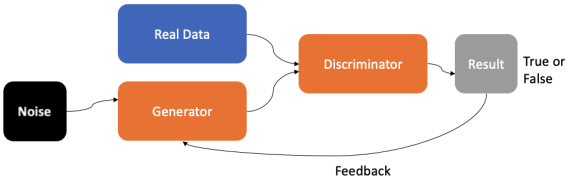

FIGURE 2.12: Example of a simple GAN architecture

2.5 Datasets

For this work, the primary dataset that is going to be used will be the LIDC-IDRI, detailed below along with the most commonly used datasets referenced across all works cited in chapter 3. Details of use, filters and proportion of training, validation and test used on the experimentation phase will be detailed at chapter 4. Below are summaries of the main datasets from lung nodule works.

The Lung Image Database Consortium image collection (LIDC-IDRI) is the largest and most common dataset used in lung cancer studies, it consists of diagnostic and lung cancer CT scans with marked-up annotated lesions. It was first developed by the National Cancer Institute (NCI) and later by the Foundation for the National Institutes of Health (FNIH) and the Food and Drug Administration (FDA) with the collaboration of seven academic centers and eight medical imaging companies. annotated by four experienced professional radiologists, currently contains 244,527 images from 1010 patients[37].

Created as a subset from LIDC-IDRI for the Lung Nodule Analysis 2016 Challenge [38], from where it gets the name LUNA16, by the Grand Challenge Organization. The main difference from LIDC-IDRI is that it had discarded slices with thickness greater than 3mm and scans with inconsistent slice spacing or missing slices, leading to a total of 888 scans [39].

The Automatic Nodule Detection 2009, ANODE09, is a dataset comprised of 55 anonymized CT scans, of wich five scans are examples and the remaining are for testing. It was build by the University Medical Center Utrecht based on the NederlandLeuvens Longkanker Onderzoek (NELSON) study, the largest CT lung cancer screening trial in Europe [40]. Like the LUNA16, it was used on a Challenge by the Grand Challenge Organization in 2010 [41].

The Early Lung Cancer Action Project (ELCAP) was first released in December 2003 and consists of an image set of 50 low-dose documented whole-lung CT scans for detection. The CT scans were obtained in a single breath hold with a 1.25 mm slice thickness with locations of nodules detected provided by radiologists([42] and [43]).

The Lung Nodules with Soft Masks (LNSM) dataset is a dataset developed by Wang et al. [44] where 1500 nodules are annotated with accurate Soft Masks for nodule segmentation. This dataset is generated through an automatic processing pipeline using the DeepLesion datasets as base. [45].

The Indian Lung Nodule Dataset (ILND) dataset was collected on an Indian Hospital by Tyagi et. al [46], it contains around 200 CT scans and each one has between 50 to 700 slices with images of 512x512.

The Lung Nodule database was created for the 2019 LNDb grand challenge [47]. It was a collaboration of the INESC TEC, the Hospital de São João (São João Hospital), the Faculdade de Engenharia da Universidade do Porto (Faculty of Engineering of the University of Porto) and the Faculdade de Medicina da Universidade do Porto (Faculty of Medicine of the University of Porto), all located in Portugal. It is a collection of 294 CT scans from 2016 to 2018. More details can be seen on [48].

The Medical Decathlon data was created for a challenge on Medical Segmentation for multiple organs, like brain, liver, lung and pancreas [49]. For this work, only the lung section was considered, with a total of 96 CT scans from the Stanford University [50].

The table 2.1 has a summary of all the datasets presented.

Dataset	Year	Total Scans
LIDC	2011	1018
LUNA16	2016	888
ANODE09	2009	55
ELCAP	2003	50
LNSM	2022	1500
ILND	2022	200
LNDb	2019	294
Medical Decathlon	2019	96

TABLE 2.1: Dataset Summary

2.6 Metrics

Below there is a short reference for the metrics used or referenced in this work and on the related works as well:

Metric	Description	Formula
Recall (REC)	Measures the proportion of positives that are correctly identified	$\text{REC} = \frac{TP}{TP+FN}$
Precision (PREC)	The proportions of positive results in statistics and diagnostic tests that are truly positive	$\text{PREC} = \frac{TP}{TP+FP}$
Accuracy (ACC)	Classification accuracy of the classifier	$\text{ACC} = \frac{TP+FN}{TP+FN+FP+TN}$
Dice Similarity Coefficient (DSC)	A statistic used to gauge the similarity of two samples	$\text{DSC} = \frac{2 \cdot TP}{2 \cdot TP+FP+FN}$
Intersection over Union (IoU)	Measures the similarity between the predicted area and the real area of the objects in a set of images	$\text{IoU} = \frac{TP}{TP+FP+FN}$
Receiver Operating Characteristic (ROC)	A curve depicting the relationship between sensitivity and specificity (Y-axis is TP rate, X-axis is FP rate)	-
Free Receiver Operating Characteristic (FROC)	Similar to the ROC curve, differing only in the X-axis (FP rate per image or per scan)	-
Area Under Curve (AUC)	Total area under the ROC curve	-

TABLE 2.2: Metric Summary

Chapter 3

Related works

In this section, the main methods found in the literature review will be listed and briefly explained. Initially, more than 35 models were selected and subsequently filtered based on their scores and evaluation metrics, with DSC being the main factor due to being the most common metric across all works and the one that best fits the evaluations used in the experimentation phase in Chapter 5.

For the final comparative list, 15 works dated from 2017 to 2023 were considered. The main reasons for excluding a method were a DSC score below 75%, as well as the lack of a declaration of such a value or any other relevant indicator such as Jaccard, Sensitivity, or PPV. For historical development reasons, other works are mentioned and referenced but not used for the final result comparison.

3.1 Models

To build the clustering of papers, the main characteristic of each reported model was considered, whether it be its primary architectural foundation, main operational core, or the most relevant evolution it brought to segmentation related works. It is not uncommon for models of one class to have elements of others; again, what defined the classification is the main property presented by the model. There is no single possible definition of how to categorize the numerous frameworks. For this work, a categorization will be adopted based on the main architecture, those that are based on U-Net [29] and its variants, as well as those with other architectures as their main structure. Therefore, the categories are: U-Net Based and Non U-Net Based.

3.1.1 U-Net Based

U-Net [29] can be considered one of the main architectures for medical segmentation methods, especially regarding lung nodule segmentation. Many works use the defined U-Net base with the Encoder and Decoder structure. For instance, in the work by Bruntha et al.[51], U-Net serving as the basis for their architecture, instead of the convolution blocks in the encoders and decoders, the authors use two inverted residual blocks based on the MobileNetV2 solution by Sandler et al.[52] to aid in reducing computational cost. The bridge between the two parts of the network is a pyramidal attention network [53], which extracts pixel-level attention data from the encoder through pyramidal pooling with convolution layers featuring 7×7, 5×5, and 3×3 filters, as shown in Figure 3.1. The solution achieved a performance of 95.7% on the LIDC.

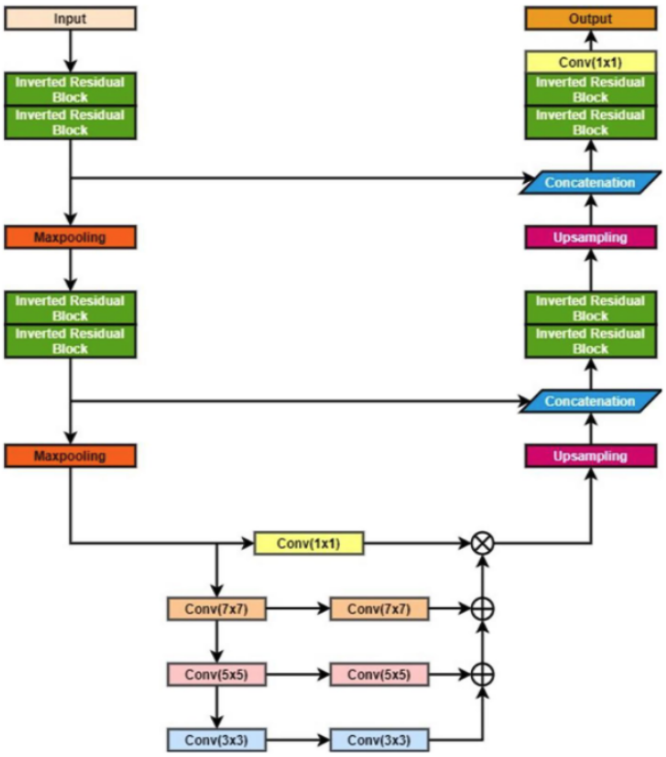

FIGURE 3.1: Architecture of Lung_PAYNet. From [51].

The EANet by Wang et. al [54] is an encoder-decoder network with five components: feature encoder based on VGG-19 [55] for the lightweight nature and feature maps extraction capability; dynamic scale-aware context (on the article refered as DSC, not to get confused with the Dice Score Coefficient) wich is a module to capture the optimal scale context information and identify the target; edge attention preservation (EAP) to handle shape information on semantic boundaries and obtain edge information, feature decoder

for the upsample part, and multi-level pairwise regression (MPR) to make inference of regions and generate improved maps with high resolution.

With the use of the 3 novel blocks proposed, DSC, EAP and MPR, the Dice Score is 87.64% on the TN-SCUI dataset, and the MPR seens to be the part that more contributes to improve the Dice Score, while it is not show the baseline network working with the sole addition of it, it is available that adding MPR to the baseline network with only one of the other blocks increase in 0.15% the score and with both 0.47%, while DSC and EAP adds only 0.04% and 0.02% for the baseline, an example of this comparison can be seen on the Figure 3.2. On the more related LUNA16 dataset, the Dice Score is 98.71%.

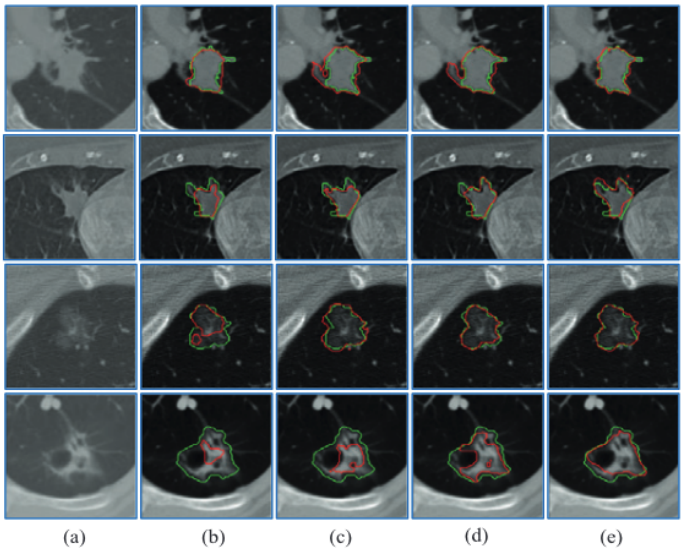

FIGURE 3.2: Segmentation example. (a) CT. (b) Baseline. (c) Baseline+DSC. (d) Baseline+DSC+EAP. (e) Baseline+DSC+EAP+MPR. The red and green lines represent the segmentation results and ground truth, respectively. From [54].

Using a U-Net as base, Chen et al. [56] developed and added to the network a Vision Flash (ViF) attention mechanism module that incorporates the Gated Attention Unit [57] and obtains global and local information to enhance the encoding. One important aspect is that with this the authors were able to reduce the number of total parameters used in the network, from 29 millions on a basic U-Net to 12 millions on ViFUNet, while still achieving a DSC of 86.03% on LUNA16 with lower training speed. The ViF module can be seen on Figure 3.3.

In 2019, Chen et. al [58] proposed a methodology with the following components: a encoder-decoder achitecture (a basic block of 3×3 convolution layer, a BN layer with ReLu, another 3×3 convolution layer and BN layer. A downsampling with 2×2 max-pooling operation on encoder and bilinear interpolation on upsampling) with the additional step of

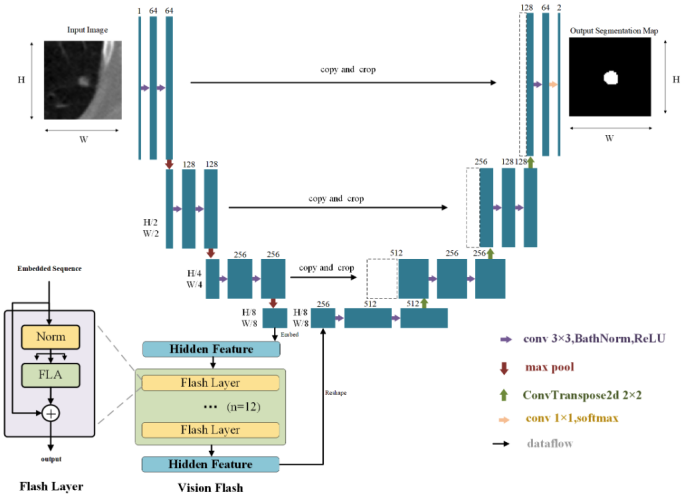

FIGURE 3.3: ViFUnet architecture. From [56].

an Atrous Spatial Pyramid Pooling (ASPP) that is used to obtain multi-scale information using four parallel atrous convolution layers, each with one sample rate so the module can perform extraction at different scales and be combined before the decoder phase as can be seen on Figure 3.4, the framework have a DSC of 90.56% in LIDC.

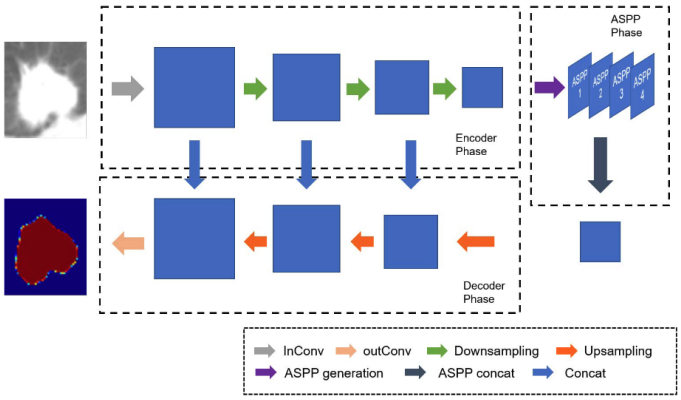

FIGURE 3.4: Proposed encoder-decoder with ASPP architecture. From [58].

Based on the U-Net++ [59], the Wavelet U-Net++ by Agnes et. al [60] brings the Haar wavelet pooling as a way to get frequency information and detail in the ct scan on the downsampling and upsampling of the network. This technique is used on the encoder with the low-frequency component down-sampled using average pooling and the high-frequency component down-sampled using max pooling, and on the decoder block a inverse wavelet transformation followed by a de-pooling operation. By applying this solution, the input image is splited in four segments, LL (Low-Low), LH (Low-High), HL (High-Low) and HH (High-High) bands, as can be seen on Figure 3.5, wich breaks the

image in multiple resolutions and helps on extractions and representations of different features.

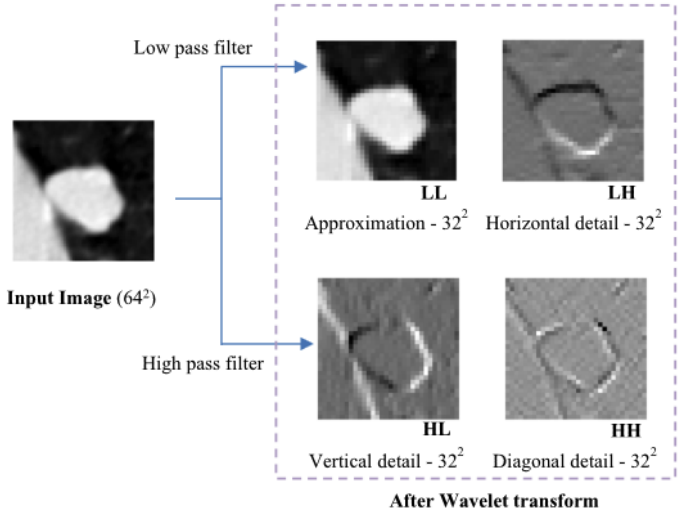

FIGURE 3.5: LL, LH, HL and HH bands example. From [60].

The overall network can be seen on Figure 3.6, with this solution a DSC of 83% was achieved on LIDC.

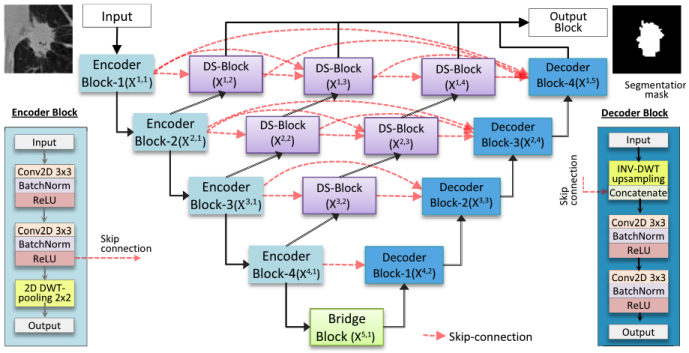

FIGURE 3.6: Wavelet U-Net++ architecture. From [60].

Wang et. al [61] propose two novel concepts on the U-Net based model: the usage of a Swin Transformer block for global feature extraction and a recalculation of weights on the decoder by the channel attention module, calling the framework Convolution and attention fusion for lung nodule segmentation (CA-UNet). The swin transformer module are two successive blocks placed at the skip connection and has Layer-Norm (LN), a multi-headed self-attentive module (first with window-based multi-head self-attention (W-MSA) and later with shifted window-based multi-head self-attention (SW-MSA)), a residual connection, and a 2-layer multilayer perceptron (MLP), as can be seen on 3.7. The W-MSA limits the self-attention calculation to non-overlapping local windows and

SW-MSA bridge the windows of the previous layer, providing a connection between the two.

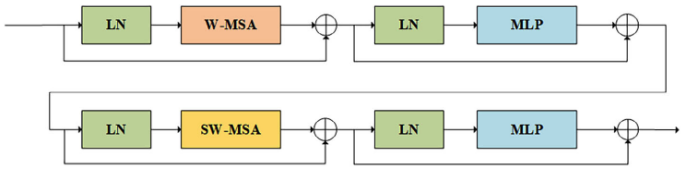

FIGURE 3.7: Two Swin Transformer blocks. From [61].

The Channel Attention aims to expand the feature extraction of the model by compressing the input feature map, passing it to a sequence of average and max pooling operations and obtaining a reweighted feature map. The block can be seen on Figure 3.8, the complete architecture on Figure 3.9 and the DSC on LIDC is 89.86%.

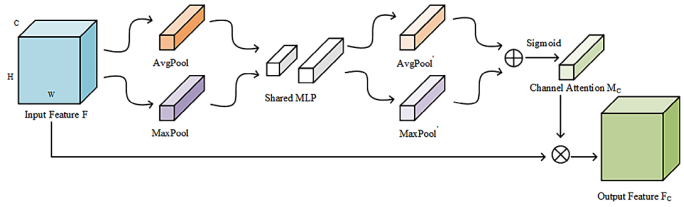

FIGURE 3.8: CA-UNet Attention Module with Weight (W), Height (H) and Channel (C). From [61].

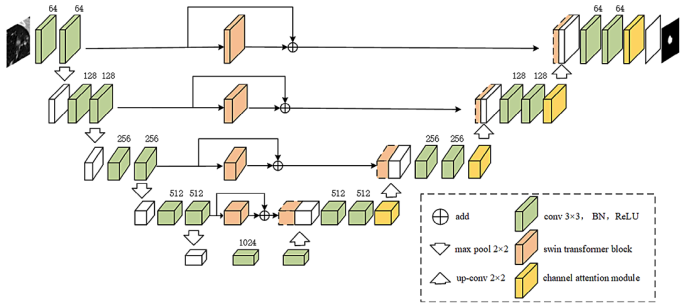

FIGURE 3.9: CA-UNet architecture. From [61].

In Usman et al. [62], a solution is proposed to use the same network in three anatomical planes: axial, sagittal, and coronal. The axial plane is involved in the first stage, while both sagittal and coronal planes are processed simultaneously in the second stage. To achieve this, the authors developed the dual-encoder-based hard attention network (DEHA-Net) with an Adaptive Region of Interest (A-ROI) mechanism, which is an algorithm to estimate the ROI for the next slices. The overall framework aims to adapt to various nodule types, taking the CT Scan and an ROI Mask as inputs for the DEHA-Net. The structure

includes two identical encoders, one for the CT Scans and the other for the ROI Mask, connected through residual connections to a single decoder responsible for producing a segmentation mask for the nth slice. An illustration of the DEHA-Net is presented in Figure 3.10.

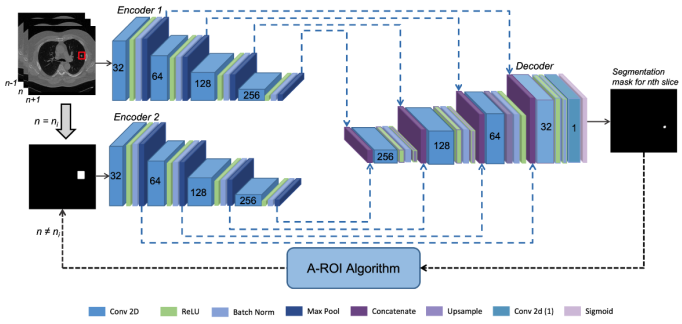

FIGURE 3.10: Architecture of DEHA-Net. From [62]

As previously mentioned, the complete framework has two stages: the first for the axial view and the second for the coronal and sagittal views. To unify the different masks generated by each DEHA-Net, a consensus mechanism is constructed, which is defined by the equations 1 and 2:

$$Ei = thr(\sum_{j=1}^{k} Sij, \tau) \tag{1}$$

$$thr(g, \tau) = \begin{cases} 1, & \text{if } g \geq \tau \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

Where Ei is the consensus value of the ith voxel, Sij represents the prediction of the ith voxel from the jth model, and K denotes the number of models (3 in total for axial, sagittal, and coronal). τ is the threshold that is determined on the validation set. The DSC for this framework is 87.91% on LIDC and the overall architecture with both stages and the consensus are show in Figure 3.11.

The Multi-level Dynamic Fusion Network (MDFN) of Cai et. al [63] is built based on U-Net with 3 novel modules, multi-scale spatial and channel feature selection (MSCFS), self-calibrated encoder and multi-level fusion decoder, as seen in Figure 3.12. First the input image passes through the encoder denoted by E_i for the application of the attention and deep feature mechanism, with each decoder passing by the MSCFS for deep feature filtering, followed by a dual simultaneous decoding of the mask on M_D that are later combined to serve as the input for the final decoder D, with $i \in (1..5)$.

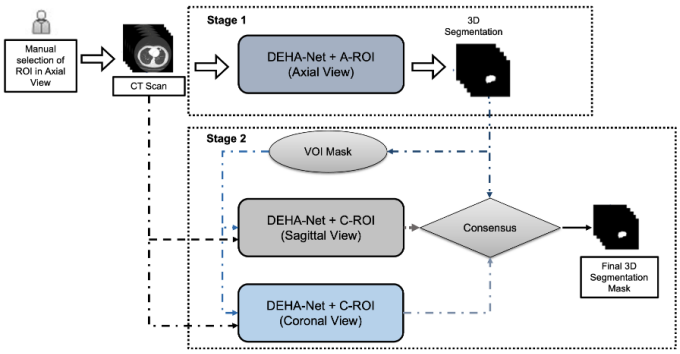

FIGURE 3.11: Full framework of DEHA-Net. From [62].

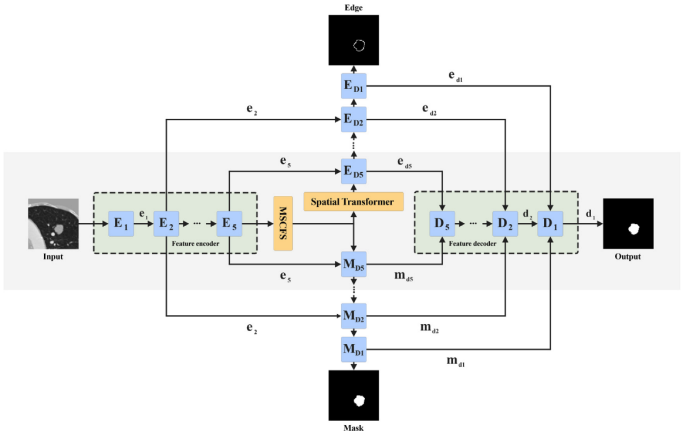

FIGURE 3.12: MDFN view. From [63].

The core MSCFS works by three stages, the first one responsible for multi-scale feature generation, the second for spatial and texture attention maps and the third stage for the integration, see Figure 3.13 for the module structure.The DSC on LUNA16 was 89.19%.

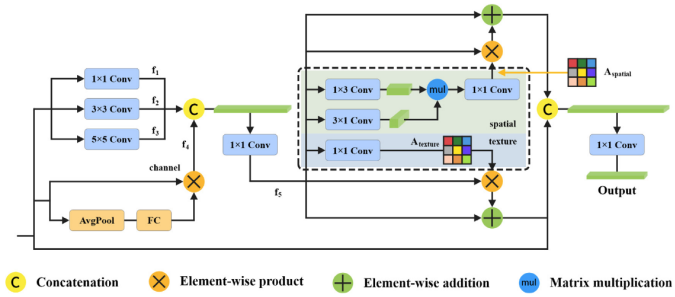

FIGURE 3.13: MSCFS structure. From [63].

The Dual-path lung nodule segmentation model based on boundary enhancement and hybrid transformer (DPBET) by Wang et. al [64] works by aggregating two sub-networks, one called global branch and the other edge branch. The global branch has an encoder-decoder structure with the new Cascade-axial-prune transformer (CAP-Trans)

block working as the bridge between the two phases, wich decomposes the 2D self-attention into two 1D attentions along the height and width axis each passing by a multi-head self-attention (MHSA) and generating the final spatial map. It is worth noting that the two blocks work in cascade mode, with the height output working as input for the width block. The local branch aims to solve blurred edges of lung nodules and generating a more clean edge segmentation mask by using two Down-Attention Sample (DASample) blocks and three decoder blocks, each DASample is composed of three parallel branches that are combined into a CBAM module that captures target spatial information and avoid edge detail. The main architecture can be seen on Figure 3.14 and the DSC on LIDC was 89.86%.

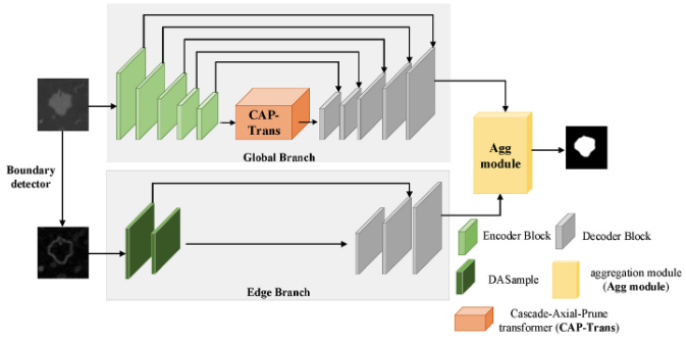

FIGURE 3.14: DPBET structure. From [64].

3.1.2 Non U-Net Based

Despite the popularity of the U-Net architecture, there are several examples of other architectures used in segmentation works. As mentioned in Section 2.4.3, GANs are adversarial networks composed of two parts, the generator and discriminator. Based on this functioning, Tyagi et al. [46] in their CSE-GAN use the generator as a network to segment the nodule and produce the segmentation mask image, while the second network is used to classify these segmentations as true or false as seen in Figure 3.15.

Additionally, an essential part of their strategy lies in the component that gives the architecture its name, concurrent squeeze & excitation (CSE) blocks, used in generation to enhance segmentation. They are composed of two parts, the patial squeeze & channel excitation module (sScE) and the channel squeeze & spatial excitation module (cSsE). The sScE uses height (H), width (W), depth (D) and channel dimensions (C) of the image to perform a squeeze operation along the spatial dimensions to obtain a channel descriptor

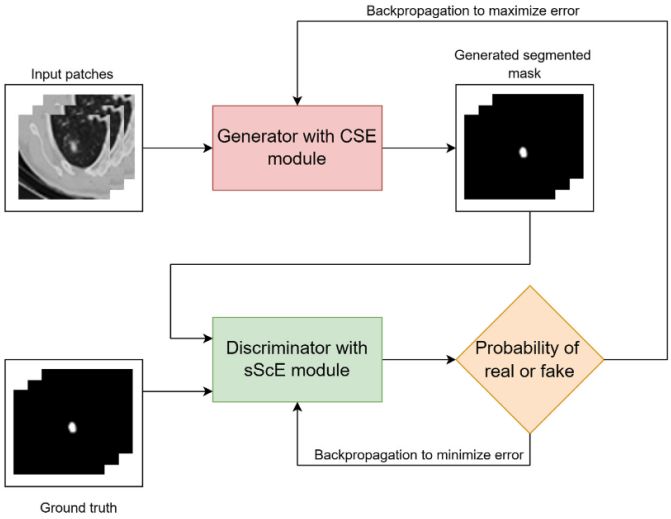

FIGURE 3.15: CSE-GAN block diagram. From [46].

using global average pooling, and then an excitation operation along the channels using a gate mechanism with the sigmoid function as seen in Figure 3.16.

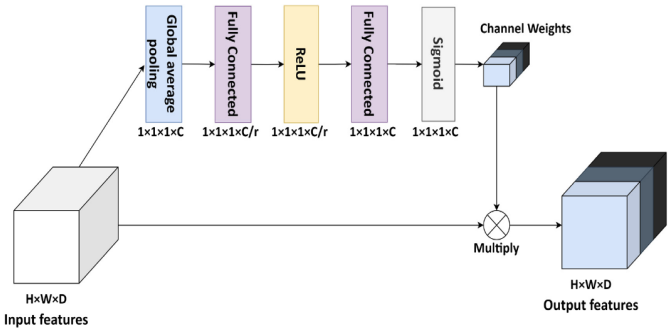

FIGURE 3.16: sScE Module. From [46].

The cSsE is used to recalibrate features along the spatial dimension through a convolution operation that squeezes the channel dimension and then applies a sigmoid activation function as seen in Figure 3.17.

The CSE block, wich adds both sScE and cSsE blocks, see Figure 3.18, is used on the generative part on a U-Net based network after each convolution block on both encoder and decoder. The sScE block is used in the discriminative part after each convolution block in a network consisted of convolutional blocks and dense blocks. The achieved DSC was 80.74% on LUNA16 dataset.

In Wang et. al [44], the authors presented the SoftGAN which is a GAN combined with the Soft Mask technique that is developed as a means to improve the nodule segmentation. The Soft Mask is an alternative to the usual binary mask used generally in the

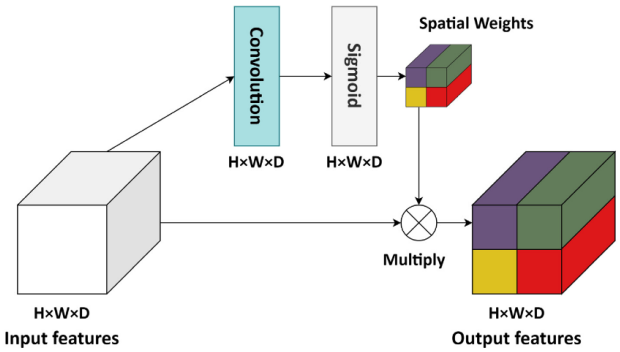

FIGURE 3.17: cSsE Module. From [46].

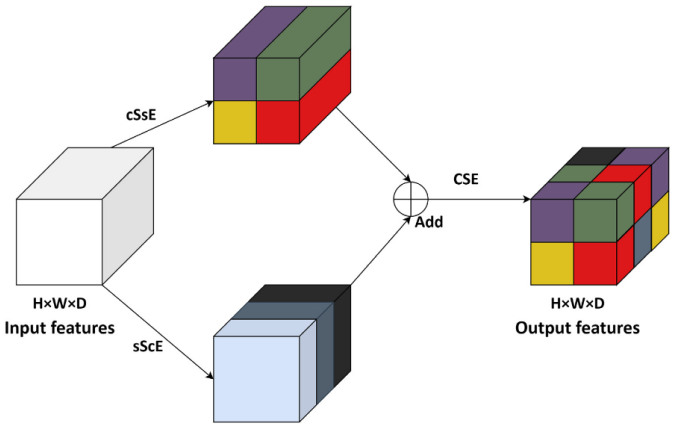

FIGURE 3.18: CSE Module. From [46].

segmentation process. Contrary to binary masks that have discrete values, Soft Masks are able to contain continuous values, meaning that they can hold more and precise information and detail, reducing imprecise boundary annotation.

To produce the Soft Mask, an automatic pipeline was built, where the first 3 strategies are used to initialize trimaps for the lesion region, background region, and the uncertain region. The first strategy is the generation of trimaps obtained through normal binary masks, the second strategy is the generation of trimaps through markings made by experts in datasets, and the third strategy is the generation of trimaps using RECIST, Response Evaluation Criteria In Solid Tumours, that contain the lesion diameter. The second stage of the pipeline consists of using the close-form-matting algorithm to generate the Soft Masks, which is a technique in image processing to perform the matting of images, i.e., estimating the opacity or transparency of each image (more details in [65]). Lastly, the image is binarized to obtain the final mask. The main steps can be seen in the Figure 3.19.

With the Soft Mask the authors then produced the SoftGAN, that is an adversarial

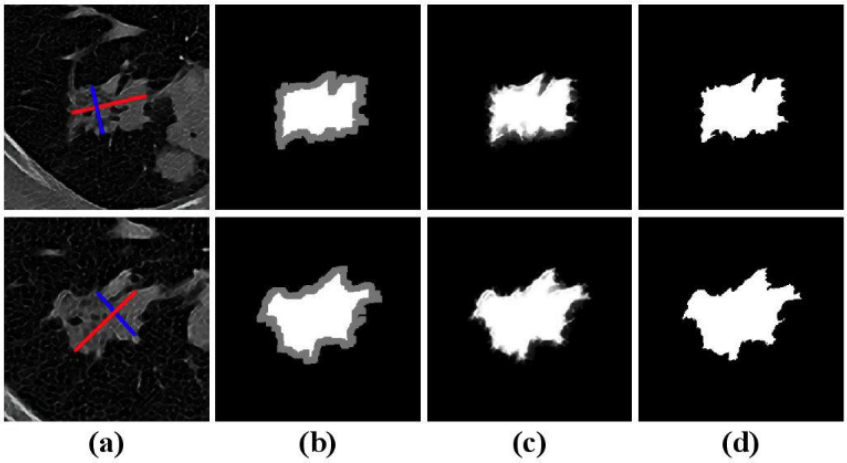

FIGURE 3.19: (a) Lung nodules with RECIST marks (b) Trimaps. Lesion region in white, background region in black and un-certain region in gray. (c) Soft Masks. (d) Binarized Soft Masks. From [44].

training framework, with the segmenter (generator network) consists of a U-Net backbone based altered to use CBAM [66] as the basic block and a Segmentation Head that consists of a 3x3 convolution layer that generates the predicted soft mask and binary mask. The discriminator is composed of 3 losses, the Segmentation Loss in the form of DICE for the supervision of ground-truth binary masks, Soft Mask Loss to make the predicted soft mask has the same distribution as the ground-truth soft mask and GAN Loss in the form of a pixel-wise discriminator that is a network with a sequence of 3x3 convolutional layers that aims to improve the quality of the predicted soft mask. The main architecture can be see in the Figure 3.20. The DSC in the LIDC dataset was 83.21%.

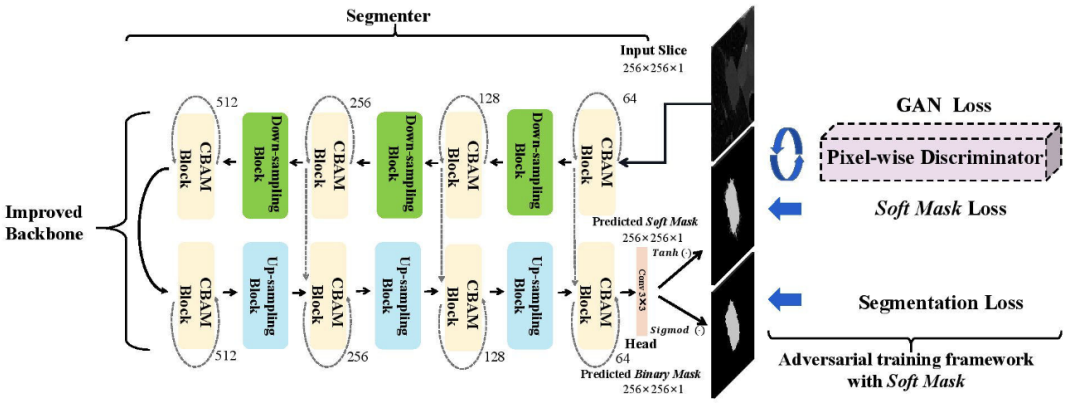

FIGURE 3.20: Architecture of the SoftGAN with adversarial Segmenter and Discriminator. From [44].

In Cao et. al [67] there is an example of multiple networks executing simultaneously with the Dual-Branch Residual Network (DB-ResNet) that aims to capture multi-view

features from current, previous, and subsequent slices centered voxel. The first branch gets multi-views of the current, previous and next slice and the second branch mult-scale for current, both branches have a Central Intensity Pooling that calculates either the center position or the surrounding intensity, that are concatenated at the end to generate the final segmentation. Details can be seen on Figure 3.21 and the DSC on LIDC was 82.74%.

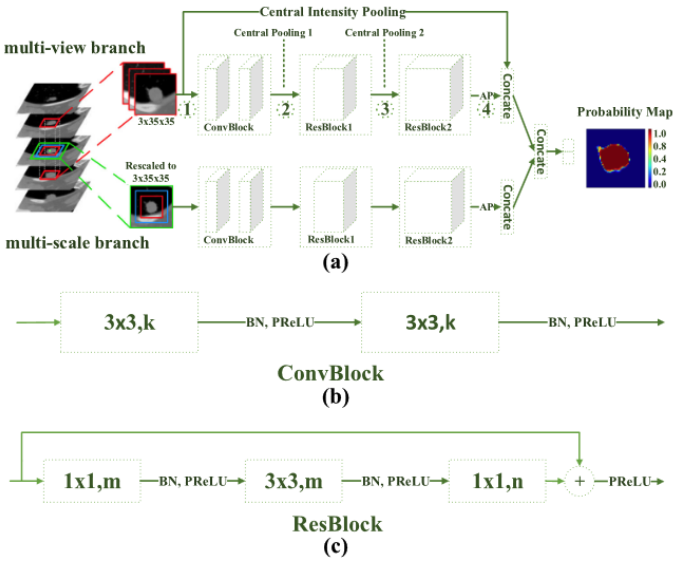

FIGURE 3.21: DB-ResNet overview with: (a) general architecture, (b) the convolution block, and (c) the residual block. K, m and n are the channels. From [67].

Qin et. al [68] divided the proposed framework in two: the increase of the dataset using a conditional generative adversarial network (cGAN) for synthetic image generation, that aims to solve the imbalance on LIDC dataset, and a 3D-CNN model for nodule segmentation.

Since the features on LIDC are not enough for the synthetic generation, first semantic labels are generated from the ground-truth labels, with the extraction and processing of features, ten-channel (9 attributes labels plus the semantic label) images are produced and used as the input of cGAN to generate the synthetic dataset, as seen in Figure 3.22. The cGAN is composed of a U-Net based generator network with strided convolution layer to downsample the image, followed by a batch normalization (BN) layer and a ReLU layer, for the upsample path a transposed convolution a BN layer, ReLU layer, and dropout layer are used. The cGAN discriminator is a FCN designed with, except in the first layer, strided convolution layers followed by a BN and ReLU layer, as seen in Figure 3.23.

With the augmented dataset, the authors developed a 3D-CNN based on 3D U-Net [69] with the introduction of texture maps so the network can differ the type of the current

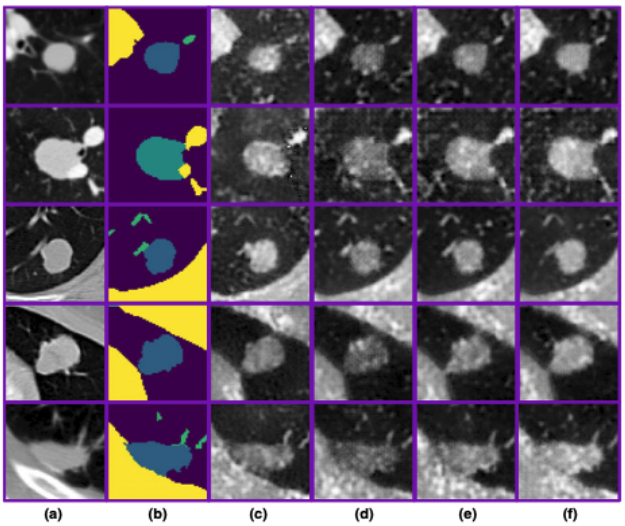

FIGURE 3.22: Semantic generation: (a) Real CT; (b) Semantic labels; (c) Images without nine attributes. (d), (e), and (f) images with nine attributes and texture scores are 1, 3, and 5, respectively. From [68].

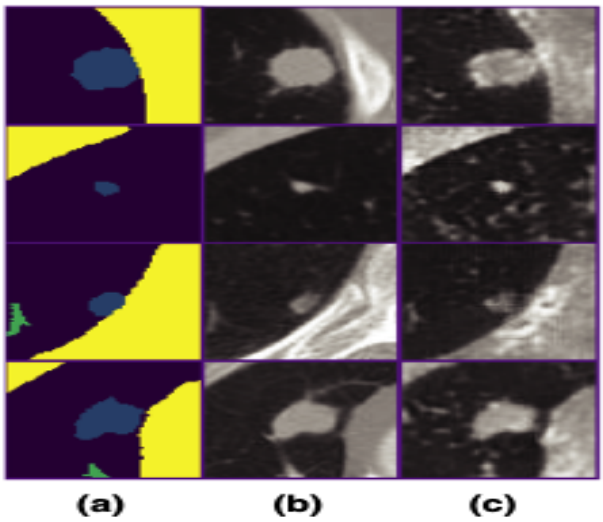

FIGURE 3.23: Example the synthetic generation with (a) Input Labels; (b) CT Images; (c) Images generated. From [68].

nodule. For this end it is used the Local Binary Pattern (LBP) [70] to represent different nodules textures and edge maps that can give information about the limits of a given nodule. For this it is applied two edge detection frameworks from Canny et. al [71] and Sobel et. al [72].

The LBP, both edge maps and the CT Volumes are used as input for the four-channel 3D segmentation network that consists on the contraction part composed of three blocks of residual units (in which are composed by two convolution layers, two BN layers,and one ReLU layer) each followed by a max-pooling layer to reduce the dimension of the

cube. On the the upsample part we can find two convolution layers, two BN layers,and one ReLU layer concatenated with the corresponding shallow features via the skip connection, tha is then used as input for the residual block. Finally, a post block is connected to map the 64-channel feature cube to the size of 1 x 64 x 64 x 64. The DSC of this solution is 84.83% in the LIDC dataset and the Figure 3.24 shows the main architecture.

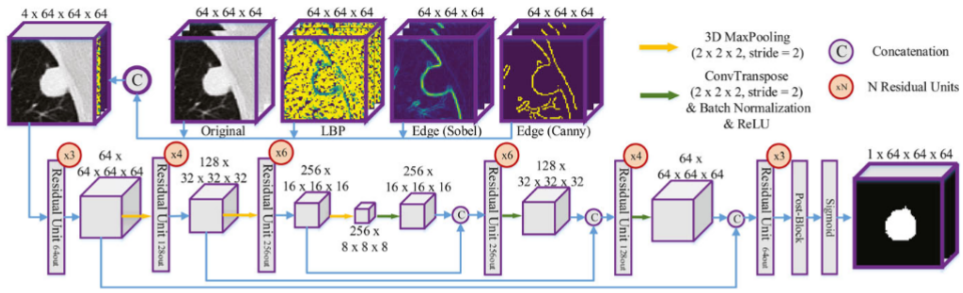

FIGURE 3.24: Proposed Network architecture by Qin et al. From [68].

Khosravan et. al [73] presents a new approach for lung nodule segmentation with the combination of the segmentation phase and false positive reduction (FPR) on the same framework, achieving this by using a multi-task learning (MTL) over a general 3D encoder-decoder network that allows to perform tasks simultaneously. The MTL is a 19 layer network with 14 layers used both by segmentation and FPR tasks, 2 by the end exclusively by segmentation and 3 by FPR. With this solution the authors showed an improvement of 9% from 82% to 91% in DSC with the LUNA16. The Figure 3.25 shows the general architecture.

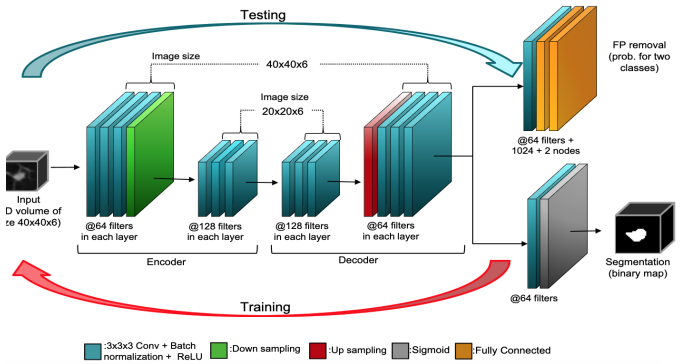

FIGURE 3.25: The 3D deep MTL CNN architecture. From [73].

A end-to-end solution that performs multiple lung cancer image tasks, namely lung nodule segmentation and classification, is also designed by Chen et. al [74] in this case using two sub-networks, Multi-Channel Segmentation Sub-network (MSN) and Joint Classification Sub-network (JCN) as show in Figure 3.26.

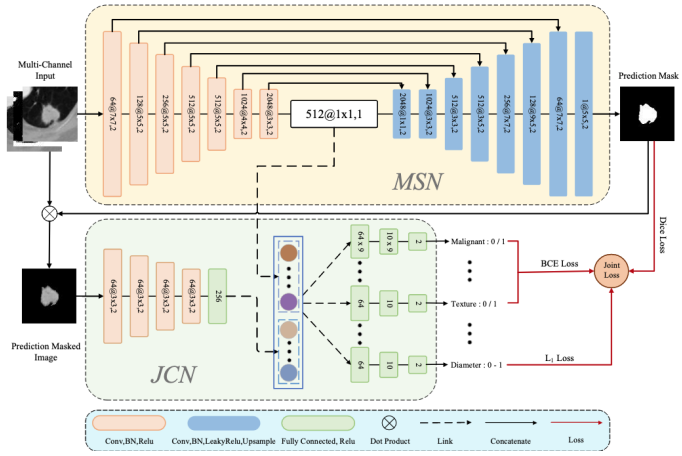

FIGURE 3.26: End-to-end multi-task framework with MSN and JCN. From [74].

The first network is responsible for the lung nodule segmentation phase with U-Net on the downsampling part and a multiple input combination using the original CT image, an OTSU image generated by [75] and SLIC image generated by [76] as show on the equation 3 where X is the final image used as model input and λ are the weights of each channel.

$$X = \lambda_1 X_{ORI} + \lambda_2 X_{OTSU} + \lambda_3 X_{SLIC} \tag{3}$$

The later part, JCN, uses the extracted features of the MSN layer to make the malignancy prediction. The JCN is composed of four convolutional layers, one fully connected layer, a connection to the bottleneck of the MSN and nine branches for different medical features, each with two fully connected layers following Relu that are merged and give the diagnosis prediction. For the loss, they developed a joint loss in the form of equation 4 where N represent malignancy prediction and seven medical features prediction tasks L_{dice}, is the segmentation loss by Dice, L^i_{apl} is the classification loss by binary cross-entropy and L_{dia} is the L_1 loss for the diameter regression. On LIDC the DSC for segmentation was 86.43%. For the malignancy prediction the DSC was 87.07%.

$$L_{joint} = L_{dice} + \sum L^i_{apl} + L_{dia} \tag{4}$$

3.2 Summary

The studies listed in this chapter were primarily selected based on their DSC score, and other metrics used for analysis are listed in the summary below. Some general considerations include:

- Despite the existence of other frameworks, U-Net and its derivatives remain the most widely used models as the main architecture;
- Some solutions are beginning to combine other tasks such as nodule classification or nodule detection to enhance overall performance and unify steps;
- The use of transformers shows promise; however, further study and analysis are still needed;
- With a few exceptions, LIDC is the main dataset for training and testing. While reliable and having a larger amount of data compared to other available datasets, there is a need to expand the training sets to encompass a greater variety of sources;
- Several models specialize in certain types of nodules, while others are more generalist, so caution is required when analyzing the published results.

Below is the table summarizing the frameworks for lung nodule segmentation:

Authors	Year	Dataset	DSC	PREC	REC	IoU	Crop Size
Bruntha et. al [51]	2022	LIDC	95.7%	96.75%	92.57	91.75%	64x64
Tyagi et. al [46]	2022	ILND,LUNA16	80.74%	80.56%	85.46%	72.52%	32x64x64
Wang et. al [44]	2022	LIDC,LNSM	83.21%	83.16%	89.14%	-	256x256
Wang et. al [54]	2022	LIDC	87.64%	-	88.94%	78.7%	96x96
Chen et. al [56]	2022	LUNA16	86.03%	87.65%	88.51%	78.19%	64x64
Chen et. al [58]	2019	LIDC	90.56%	91.35%	90.41%	-	-
Agnes et. al [60]	2023	LIDC	93.7%	-	-	87.8%	64x64
Cao et. al [67]	2019	LIDC	82.74%	79.64%	89.35%	-	3x35x35
Wang et. al [61]	2023	LIDC	89.96%	89.07%	92.44%	82.42%	-
Qin et. al [68]	2018	LIDC	84.83%	88.95%	85.11%	-	10×64×64
Usman et. al [62]	2023	LIDC	87.91%	89.56%	90.84%	-	-
Khosravan et. al [73]	2018	LUNA16	91%	-	98%	-	6x40x40
Chen et. al [74]	2021	LIDC	86.43%	-	-	77.14%	96x96
Cai et. al [63]	2023	LUNA16	89.19%	-	88.89%	80.78%	96x96
Wang et. al [64]	2022	LIDC	89.86%	-	90.5%	-	64x64

TABLE 3.1: Frameworks Summary

Chapter 4

Materials and Methods

In this chapter, a brief description of the materials and methods to be used in the experiments to be conducted later will be provided. Below are details of the data used, as well as a brief analysis of their condition and status. Finally, a summary of the methods employed for the experiments conducted in the chapter 5 is presented.

4.1 Data

As mentioned earlier in Chapter 2, the primary dataset utilized in this work is the LIDC-IDRI dataset. Since it is a publicly available dataset, with a Python library, pylidc [77], that provides support and ease of handling, LIDC-IDRI is the most widely used among all datasets related to lung cancer.

It was built through the collaboration of 7 academic centers and 8 medical imaging companies, comprising a total of 1018 cases for 1010 patients, note that are 8 patients with two scans each. In the section below is a summary of the characteristics of the LIDC-IDRI.

For experimentation and validation, this work also utilized the LNDb and Medical Decathlon datasets. Details of usage are provided in chapter 5.

4.1.1 Data Analysis

The complete dataset contains a total of 7371 nodules marked by at least one specialist. Since this work utilizes the pylidc library as a source of annotations and metadata, a total of 6789 marked nodules are used, as the library fails to handle some markings. The Figure 4.1 illustrates the distribution of markings by radiologists, although it may appear that there are nodules marked by more than 4 radiologists, such behavior is solely due to the

storage format of certain annotations which causes the annotation count of the pylidc API to be incorrect. This should not impact the use of the dataset and this group is considered to be in the 4 annotations group.

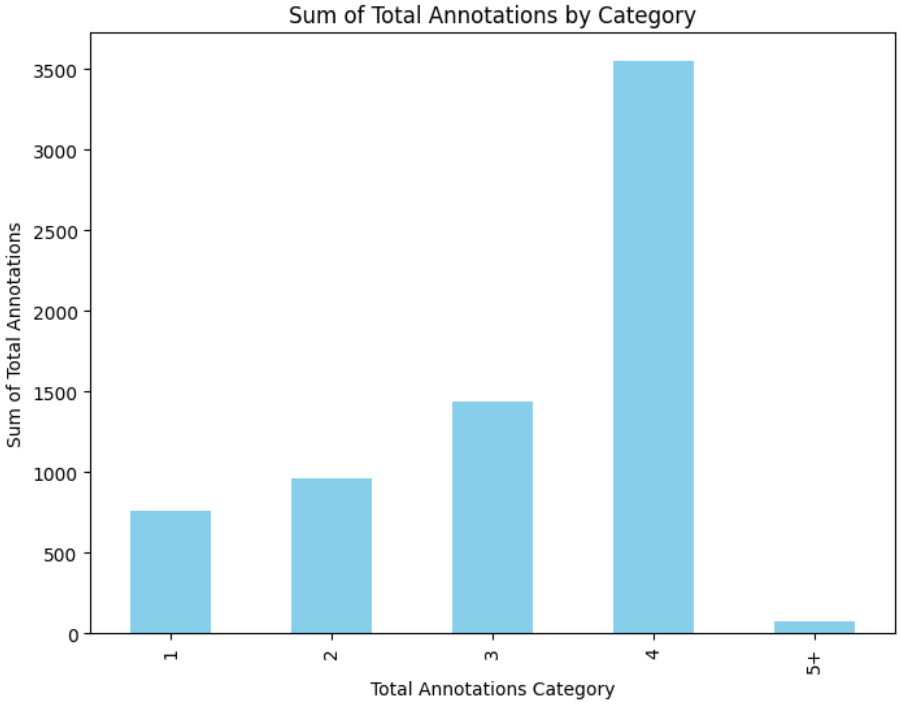

FIGURE 4.1: Sum of Total Annotations by Number of Radiologist

An important aspect when dealing with LIDC-IDRI is understanding its nodule composition. Below, the total distributions of nodules and non-nodules are presented in Figure 4.2, along with their malignancy levels ranging from Highly Unlikely for non-cancerous to Highly Suspicious for cancerous in Figure 4.3.

Another attribute that is crucial for evaluating nodule segmentation's is their texture. The distribution can be observed in Figure 4.4.

A general distribution of nodule features is shown at Figure 4.5. Additional information and characteristics can be analyzed through pylidc, and further details are available in [78].

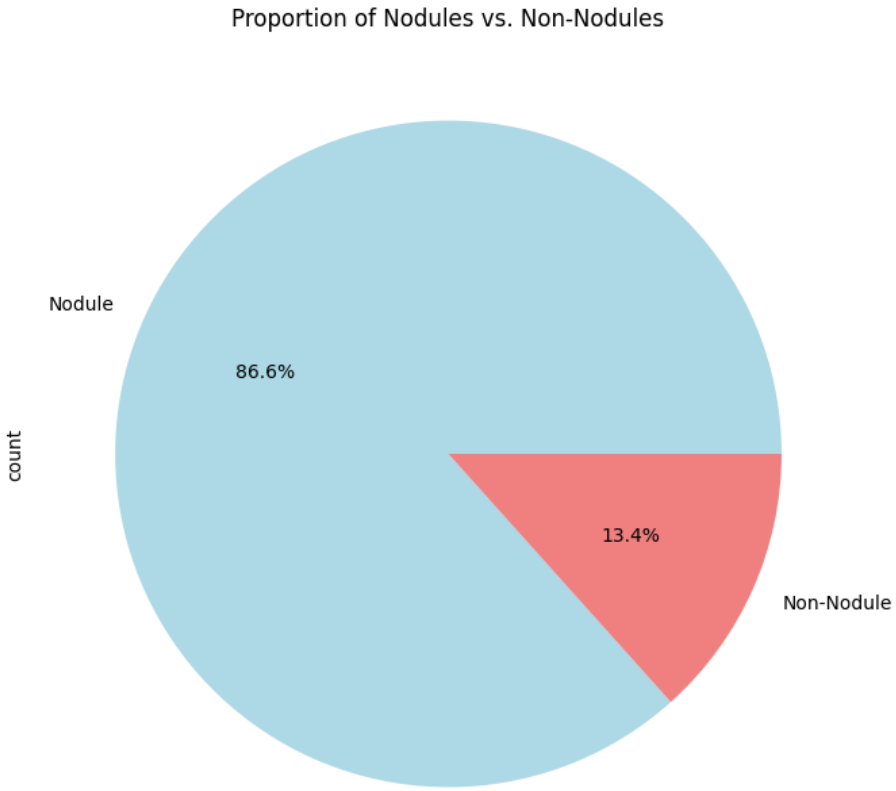

FIGURE 4.2: Total number of nodules

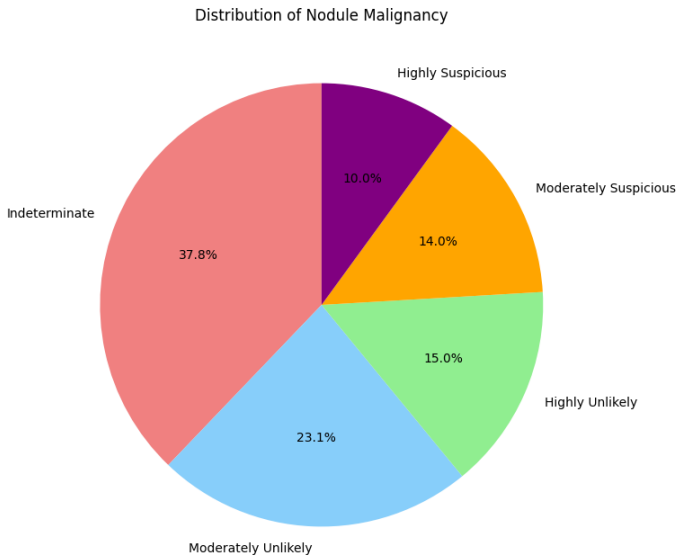

FIGURE 4.3: Malignancy distribution

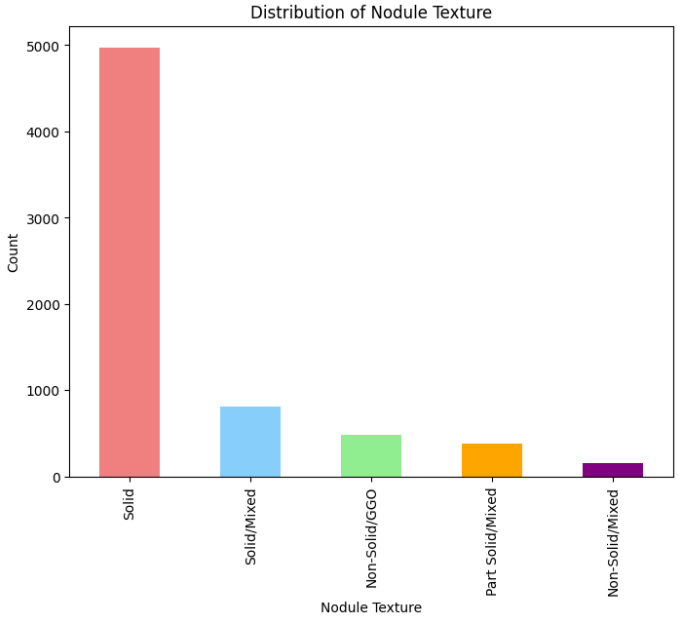

FIGURE 4.4: Nodule texture distribution

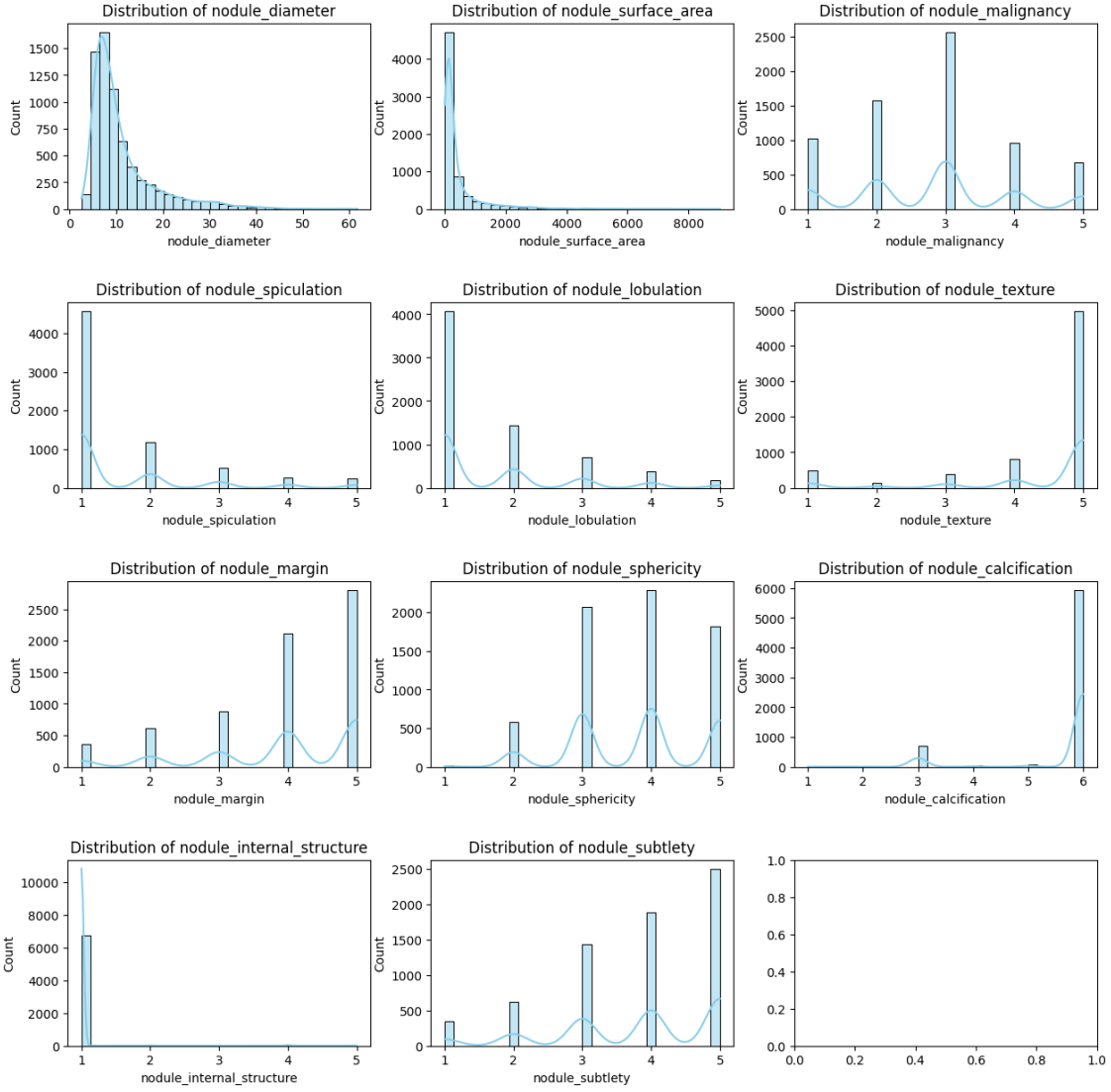

FIGURE 4.5: Nodules features distribution

4.2 Pre-Processing

The main objective of the pre-processing step is to segment the lung from the other tissues and organs present in each CT scan, as mentioned in section 2.3.2. This allows the networks to be better trained and isolate the area of interest, removing unnecessary anatomical structures and image artifacts that could otherwise lead to incorrect training. In this work, both 2D and 3D data were used. Although they share the same core process, a few key differences exist, which are detailed in the sections below.

The first step is to input the original image into the segmentation process. In this work, four different crops were used: 64x64, 96x96, 224x224, and 512x512. The use of different crops serves the following purposes: 1) to compare with current works that typically use 64x64 and 96x96 crops; 2) to properly process models using the Swin Transformer, which requires 224x224 crops; 3) to use the full image from the CT scan at 512x512 as this is the main goal. The Figure 4.6 shows the original CT slice scan that is used as input on the lung segmentation. The Figure 4.7 shows examples of the different crop sizes.

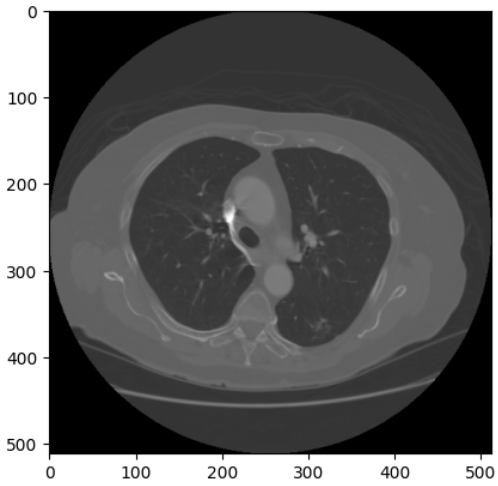

FIGURE 4.6: Original CT scan

The subsequent steps are as follows: Normalization, for standardizing the pixels in the image; Filtering, for removing noise using median and anisotropic filters; Clustering with K-Means to separate the Region of Interest (ROI) from the background; Mask generation, using morphological operations to remove distortions from the image and produce the final mask; and finally, Lung segmentation, which applies the generated mask to the original image to extract the lung. The code for the lung extraction is based mainly on the code from [79] and [80]. More detailed explanation of each step can be seen on [81].

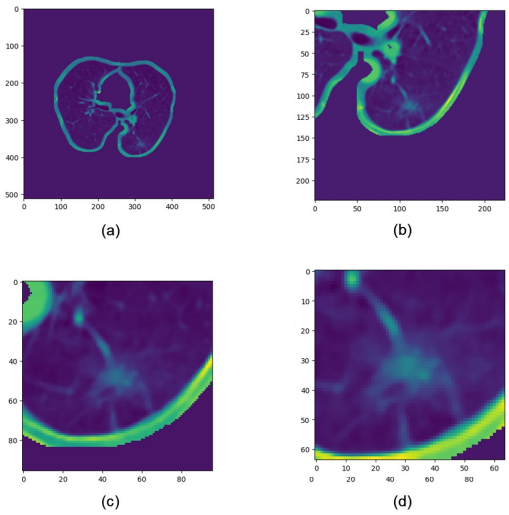

FIGURE 4.7: Example of different crops with: (a) 512x512; (b) 224x224; (c) 96x96; (d) 64x64

The Figure 4.8 shows the step-by-step of the lung segmentation process and the Figure 4.9 shows an example of a segmentation mask for a nodule.

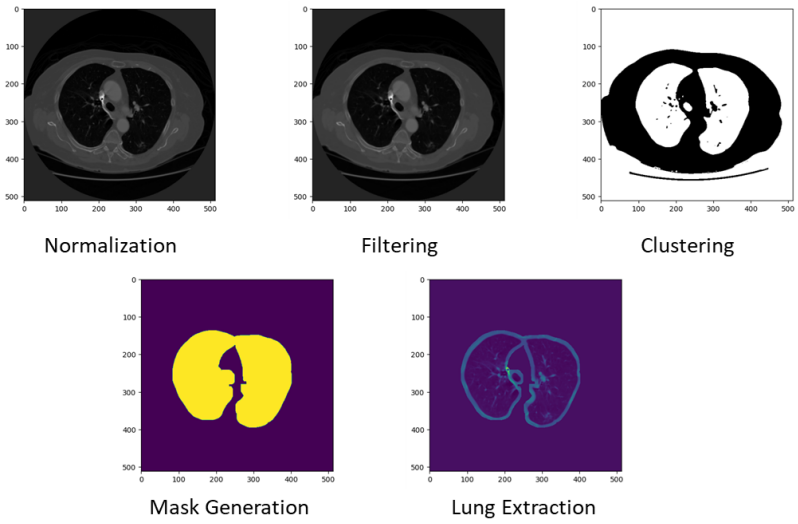

FIGURE 4.8: Steps of pre-processing the CT Scan

With the core lung segmentation function, 2D lung segmentation of the LIDC dataset can be achieved. Using the pylidc library, the CT scans of each patient are loaded and undergo a process that separates scans with nodule annotations from scans without nodule annotations. For the nodules with annotations, a consensus mechanism on the annotations is applied, aiming to select only slices with at least 50% consensus among specialists' annotations. Afterward, the selected slices undergo the lung segmentation process, and a set of images with masks and scans is saved for future use.

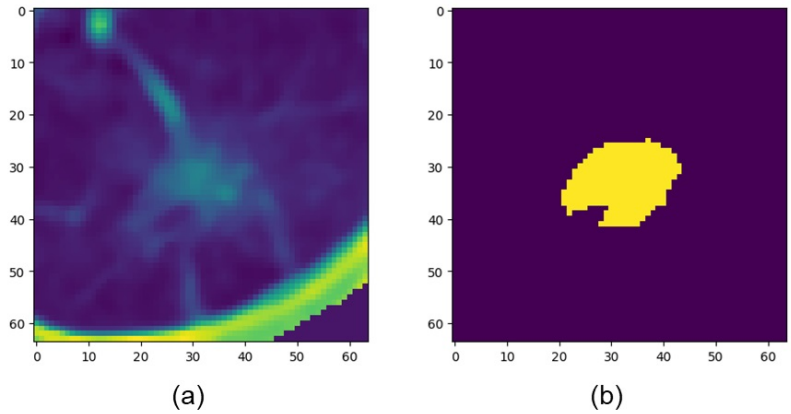

FIGURE 4.9: Segmentation Mask Example: (a) Lung Scan; (b) Nodule Mask

Throughout the entire process, a file containing metadata of both nodules and non-nodules is constructed and later used in creating training, validation, and test datasets. Initially, in this study, the LIDC dataset is split into 6160 images for training, 1735 for validation, and 2110 for testing.

In addition to the LIDC dataset, this work utilized the LNDb dataset and the Medical Decathlon dataset. For each dataset, a simpler process was employed due to the lack of consensus mechanisms. This process involves only the conversion of the original data to numpy arrays, as well as lung extraction and mask generation. For the LNdb dataset 668 images were used for training, 166 for validation and 199 for testing. For the Medical Decathlon, 967 images were used for training, 344 for validation and 346 for testing.

4.3 Deep Learning and Methodology

As seen in the review of related works in Chapter 3, the use of Deep Learning techniques is the most common and promising approach for lung nodule segmentation. Within this field, there are numerous possibilities and varieties of frameworks. For this work, simplified architectures based on U-Net were chosen.

The motivation for this choice lies in the fact that U-Net is the most commonly used architecture and serves as the core for many current models, demonstrating its effectiveness. Another reason is that it is more basic and simplified, which facilitates conducting experiments and tests without major difficulties or necessary modifications. A few different U-Net base models were used, more details on chapter 5.

In addition to the U-Net, a test was conducted using the Swin Transformer to understand the impact and feasibility of using transformer-based models in segmentation, a more recent yet promising technology in various fields.

Unless otherwise specified in the testing sections, the models were standardized with a learning rate of $1e-4$, which offers a good balance between performance and precision, enabling good results without significantly impacting computational resource consumption. Further details are described in the sections of Chapter 5.

The general approach for conducting the tests was first to establish a baseline understanding of how basic models perform in the segmentation task, in order to gain knowledge and assess the possibilities. Subsequently, attempts and experiments were carried out to determine which components or processes could improve the results of the standard models and which variant would provide the best combination to achieve a general model.

At the end of the tests, a model is selected and evaluated using the previously presented datasets to assess its final performance, define the proposal of this work, and highlight the challenges and limitations encountered.

Chapter 5

Experiments and Proposed Framework

To construct a robust and dependable framework, a series of experiments and tests were conducted. The following sections detail key findings, including insights into data augmentation, transformer utilization, and attention mechanisms. All code is accessible in the GitHub repository: https://github.com/vinibfr/lung_nodule_segmentation.

5.1 Initial Benchmark

In order to establish a baseline for future comparisons, four of the most common CNN models were selected to work with the LIDC dataset. These models included U-Net [29], U-Net++ [59], UT-Net [82] and Attention U-Net [83]. The implementations for U-Net and U-Net++ were based on the code available in the GitHub repository by Jaeho Kim [84], while the UT-Net implementation was sourced from the official repository by Gao et. al [85], and the Attention U-Net implementation was obtained from the GitHub repository of Lee Jun Hun [86].

All models underwent training for 80 epochs with a learning rate of 10^{-3}, a batch size of 20, and a basic data augmentation technique. The loss function used was a combination of Binary Cross Entropy (BCE)([87] and [88]) and Dice coefficient (DSC) [89]. The BCE equation is defined below at 1, where N is the total number of samples, y_ijk is the ground truth for the i-th sample at jk pixel position and $\hat{y}_ijk$ is the predicted probability for the i-th sample at same position.

$$BCE = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j,k}\Big[y_{ijk}\log(\hat{y}_{ijk}) + (1-y_{ijk})\log(1-\hat{y}_{ijk})\Big] \quad [88] \tag{1}$$

The DSC is defined at 2, where $intersection_i$ is the y and $\hat{y}$ intersection and α is a operator to avoid division by zero with value 10^{-5}.

$$DiceLoss = 1 - \frac{1}{N}\sum_{i=1}^{N}\frac{2\cdot(\sum_{j,k} y_{ijk}\hat{y}_{ijk} + \alpha)}{\sum_{j,k} y_{ijk} i + \sum_{j,k}\hat{y}_{ijk} + \alpha} \tag{2}$$

The final loss is defined at 3, where β is a weighting factor for the BCE with value 0.5.

$$BCEDiceLoss = \beta * BCE + DSC \tag{3}$$

The metric used for evaluation is also DSC, as defined at 4, the only difference is the $\hat{y}^{>}_{ijk}$ that is the pixel-wise prediction after applying a 0.5 threshold. In this initial set of results, the best performing model was U-Net++ with DSC of 53.4%. In table 5.1 is possible to see the best validation dice of each model . The training and validation dice is shown at Figure 5.1.

$$DiceMetric = \frac{1}{N}\sum_{i=1}^{N}\frac{2\cdot(\sum_{j,k} y_{ijk}\hat{y}^{>}_{ijk} + \alpha)}{\sum_{j,k} y_{ijk} i + \sum_{j,k}\hat{y}^{>}_{ijk} + \alpha} \tag{4}$$

Model	**DSC**
UT-Net	49.8%
U-Net	51.9%
Attention U-Net	52.3%
U-Net++	**53.4%**

TABLE 5.1: Initial Benchmarks

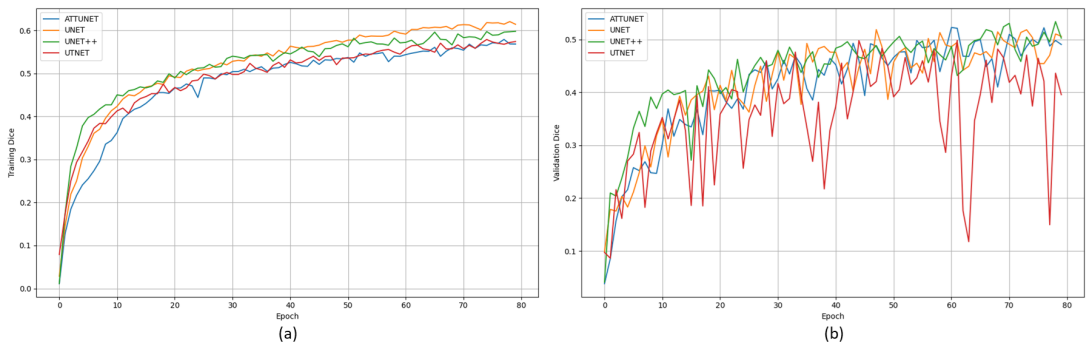

FIGURE 5.1: Evolution of Dice over Epochs: (a) Training Dice; (b) Validation Dice

This initial setting provided the basic parameters and baselines on which this work will investigate and experiment. For an out-of-the-box solution, U-Net++ seems to be the better model, but with the correct adjustments the results may vary.

5.2 Data Augmentation Comparison

When developing deep learning models, it is usually necessary to have a large volume of data that not only provides a vast quantity but also a wide variety of data. This allows the models to achieve good performance on specific tasks and also avoid over fitting problems. Unfortunately, due to the sensitive nature of personal medical data, such characteristics of quantity and variety are hardly achieved in available datasets. To overcome this problem, the technique of data augmentation is applied, which consists of a series of mechanisms that enable enlargements and alterations in the dataset to be employed [90].

In this work we conducted a comparison with 5 different augmentation techniques and a total of 13 combinations, based on the work by Goceri [91]. The table 5.2 list the 5 techniques used, their description and usage detail. Figure 5.2 show an example of each transformation techinique.

Technique	Description	Usage Detail
Shearing	Slides along the vertical or horizontal axis of the image	Random angle from -15 to 15
Translation	Translates (or shift) an image along a particular direction	Random value from -15 to 15
Rotation	Apply a rotation degree over the image	Random value from 0º to 35º
Flipping	Vertical or horizontal mirroring of the image	-
Gaussian Noise	Add intensity noise in the pixels of the image with a Gaussian distribution	Intensity from 0.3 to 2

TABLE 5.2: Data Augmentation Techniques

All combinations were evaluated with U-Net++, besides the epoch number of 40 instead 80, all parameters were the same as the mentioned in section 5.1. The list of combinations and final DSC is detailed at table 5.3.

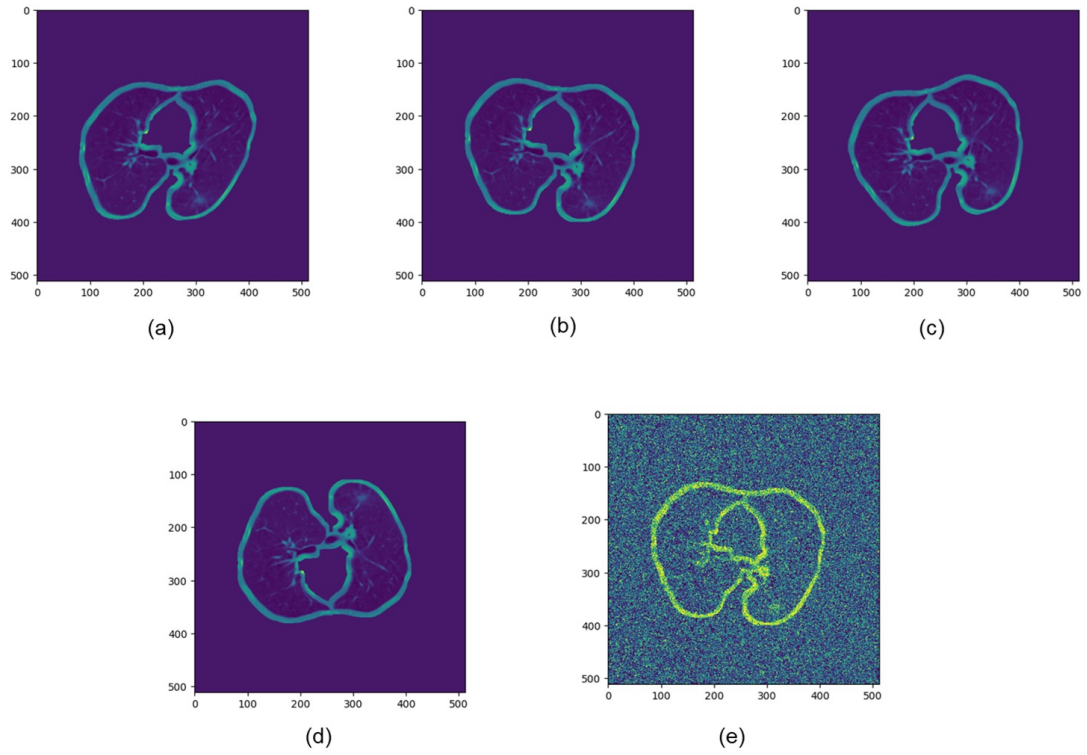

FIGURE 5.2: Example of the data augmentations transformations: (a) Shearing; (b) Translation; (c) Rotation; (d) Vertical Flipping; (e) Gaussian Noise

Combination	Technique(s)	DSC
1	**Shearing**	**56%**
2	Translation	55%
3	Horizontal Flipping	52%
4	Rotation	53%
5	Vertical Flipping	49%
6	**Shearing and Translation**	**56%**
7	Horizontal Flipping and Rotation	53%
8	Horizontal Flipping, Rotation and Vertical Flipping	45%
9	Shearing, Translation, Horizontal Flipping and Rotation	49%
10	Shearing, Translation, Vertical Flipping and Rotation	49%
11	Gaussian Noise	35%
12	Gaussian Noise and Rotation	38%
13	Gaussian Noise, Rotation and Horizontal Flipping	33%

TABLE 5.3: Data Augmentation Combinations and Results

Consonant with the work of Goceri, shearing and translation are the techniques with best the results. When used alone they present DSC of 56% and 55% respectively, in combination the DSC is 56%. The Figure 5.3 shows the training and validation dice evolution.

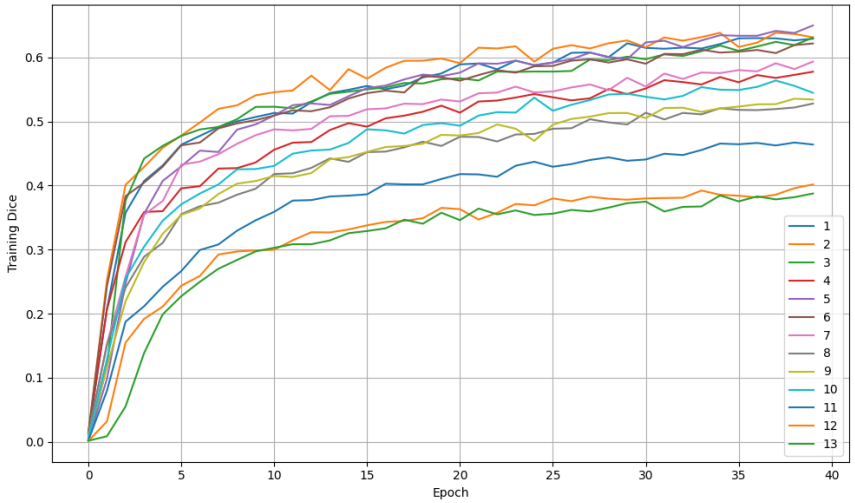

FIGURE 5.3: Evolution of Dice over Epochs for Augmentation Comparison: (a) Training Dice; (b) Validation Dice

For this work, the augmentation technique used from now on is number 6 on table 5.3.This technique, compared to the basic U-Net++, demonstrated a 2.6% improvement in DSC. It is worth noting that several of the techniques used had a negative impact on the models' performance, achieving DSCs below 40% in the techniques that use Gaussian Noise, thus demonstrating the importance of choosing the correct technique to maximize results rather than the opposite.

5.3 Data Selection Test

One of the most important aspects regarding the training of a neural network is the data used and its proper selection. So far, a broad set of nodules from the LIDC has been used, which contains all nodules with at least 50% consensus among the specialists who performed the annotations. Tests were then conducted with U-Net and U-Net++ by filtering only the nodules classified with malignancy ratings of 4 and 5 in the LIDC annotations, namely, 4 being 'Moderately Suspicious' and 5 'Highly Suspicious'. In contrast, ratings of 1 to 3 are considered 'Highly Unlikely', 'Moderately Unlikely', and 'Indeterminate'.

It can be observed that by applying a new data filter, the models show a considerable improvement, with the U-Net having an improvement of almost 7% and the U-Net++

Model	DSC	Malignancy Scope	Crop Size
U-Net	61.2%	1,2,3,4 and 5	512x512
U-Net++	56%	1,2,3,4 and 5	512x512
U-Net	**68.1%**	4 and 5	512x512
U-Net++	67.7%	4 and 5	512x512

TABLE 5.4: Results for Data Selection

of almost 12%. Nevertheless, it will be seen in the evaluation section 5.8 that the use of filtered data negatively impacts the model's ability to generalize its use. From now on, when the data selection with all nodules appears, it will be referred to as "complete," and the one with the filtered nodules will be referred to as "filtered."

5.4 Attention Mechanisms

As seen in Chapter 3, multiple solutions employ attention mechanisms to enhance the performance of the network. The attention mechanism can be implemented at various layers, such as the encoder path, bottleneck, skip connection, or decoder path. Next, a series of studies with different mechanisms will be conducted to comprehend and provide insight on performance gains, downsides, and upsides in the utilization of attention features. Two classes of tests were made, Bottleneck Attention and Skip Connection Attention.

This first test will investigate how attention mechanisms applied to the bottleneck layer can impact the overall network. To this end, the U-Net will be used as it is simpler than other networks and has fewer components, thereby minimizing potential experimental variables. The network will undergo training for 80 epochs with a learning rate of 10^{-4}, a batch size of 20, and employing data augmentation technique 6 from Section 5.2, specifically the one involving shearing and translation. Additionally, tests will also be conducted with different crop sizes on the images, specifically 64x64, 96x96, and 224x224. This will allow for a broader basis of comparison with the works listed in the state-of-the-art. For the cropped images, the only difference is the batch size used, which is 64 instead of 20. With lower resolution, the batch size can be increased without compromising computational resources.

In this work, 5 different attention blocks were tested, which are: the Atrous Spatial Pyramid Pooling (ASPP) based on the work by Chen et. al [92] which is a block consisted of multiple parallel filters with different rates (or strides) on standard convolution that

aims to classify the center pixel of the input; Pyramidal Attention Block (PYB) based on the work of Bruntha et. al [51] that uses a pyramidal pooling block with three convolutional layers (7×7, 5×5 and 3×3) for spatial attention on the input; Squeeze-and-Excitation Block (SEB) inspired on the work by Hu et. al [93] using a block that uses a squeeze operation and a excitation operation that uses a self-gating mechanism to produce a feature map with modulation weights; the Spatial Attention Module (SAM) from the work by Guo et. al [94] which is a block that uses the input to generate a spatial attention map from the input using a concatenation of average and max-pooling operations followed by a convolutional layer, sigmoid activation and element-wise multiplication; finally, the last block is the Vision Transformer (ViT) by Dosovitskiy et. al [95] that slices the original image into multiple patches and passes each one of those patchs to a structure composed of multiheaded self-attention (MSA) and multi-layer perceptron (MLP) that classifies the patches.

The Figure 5.4 shows the basic U-Net layout with the addition of the attention block. And the Figure 5.5 shows the diagram of all attention blocks used.

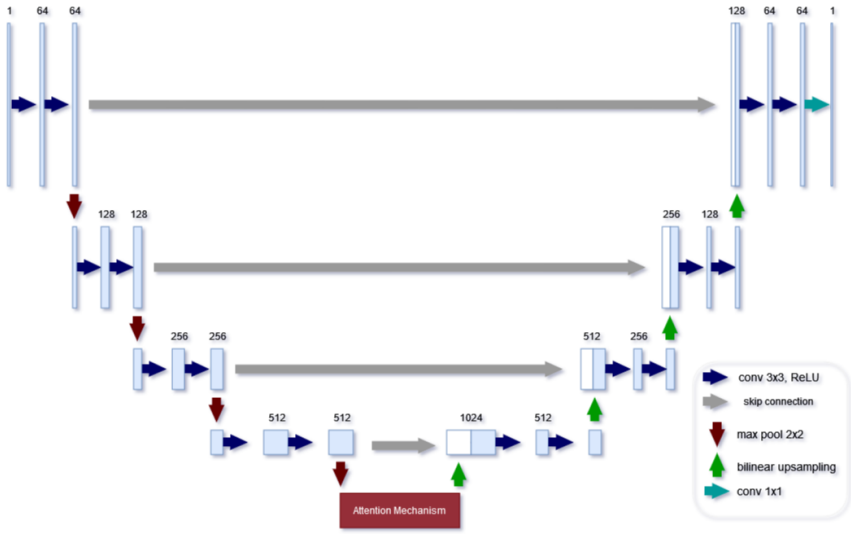

FIGURE 5.4: Basic U-Net with attention block

Tables 5.5, 5.6, 5.7 contains the results from the test of attention mechanism on bottleneck layer. In all tables, U-Net and U-Net++ are presented for comparison.

In the original image size, even though the results were close to each other, the use of Spatial Attention techniques showed a clear advantage over the others, with improvements ranging from 0.4% to 1.5% in DSC compared to other attention blocks and 10.3% compared to the base U-Net. It is also worth highlighting SAM's ease and transparency

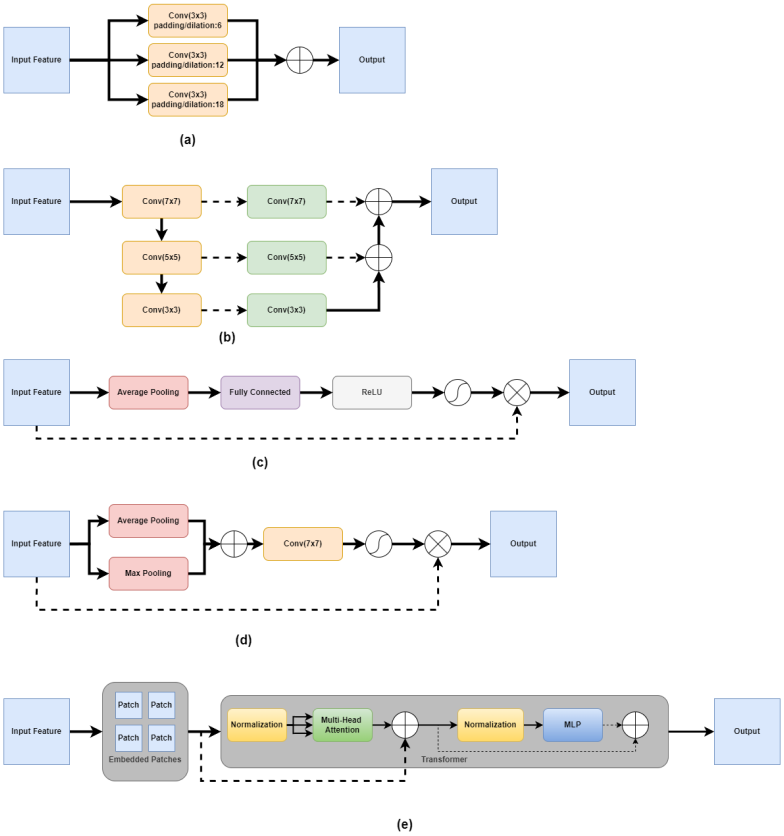

FIGURE 5.5: Attention Blocks Modules. (a) ASPP; (b) PYB; (c) SEB; (d) SAM; (e) ViT

Model	DSC	Batch Size	Crop	Data Split
ASPP U-Net	**64%**	20	512x512	All
PYB U-Net	62.6%	20	512x512	All
SEB U-Net	59.9%	20	512x512	All
SAM U-Net	62.3%	20	512x512	All
ViT U-Net	57.9%	20	512x512	All
U-Net	61.3%	20	512x512	All
U-Net++	60.7%	20	512x512	All
ASPP U-Net	**68.8%**	20	512x512	Filtered
PYB U-Net	68.5%	20	512x512	Filtered
SEB U-Net	67.4%	20	512x512	Filtered
SAM U-Net	**68.8%**	20	512x512	Filtered
ViT U-Net	63.6%	20	512x512	Filtered
U-Net	68.1%	20	512x512	Filtered
U-Net++	67.7%	20	512x512	Filtered

TABLE 5.5: Results for Attention Bottleneck for Crop Size 512x512

Model	DSC	Batch Size	Crop	Data Split
ASPP U-Net	82.2%	64	96x96	All
PYB U-Net	82.8%	64	96x96	All
SEB U-Net	**82.9%**	64	96x96	All
SAM U-Net	82.8%	64	96x96	All
ViT U-Net	82.4%	64	96x96	All
U-Net	82.8%	64	96x96	All
U-Net++	82.1%	64	96x96	All
ASPP U-Net	83.5%	64	96x96	Filtered
PYB U-Net	84.1%	64	96x96	Filtered
SEB U-Net	83.9%	64	96x96	Filtered
SAM U-Net	**84.4%**	64	96x96	Filtered
ViT U-Net	82.6%	64	96x96	Filtered
U-Net	83.9%	64	96x96	Filtered
U-Net++	79.7%	64	96x96	Filtered

TABLE 5.6: Results for Attention Bottleneck for Crop Size 96x96

Model	DSC	Batch Size	Crop	Data Split
ASPP U-Net	83.2%	64	64x64	All
PYB U-Net	**83.4%**	64	64x64	All
SEB U-Net	**83.4%**	64	64x64	All
SAM U-Net	83.1%	64	64x64	All
ViT U-Net	82.8%	64	64x64	All
U-Net	83.7%	64	64x64	All
U-Net++	82.7%	64	64x64	All
ASPP U-Net	83.1%	64	64x64	Filtered
PYB U-Net	83.7%	64	64x64	Filtered
SEB U-Net	83.7%	64	64x64	Filtered
SAM U-Net	83.5%	64	64x64	Filtered
ViT U-Net	82.9%	64	64x64	Filtered
U-Net	**83.9%**	64	64x64	Filtered
U-Net++	82.5%	64	64x64	Filtered

TABLE 5.7: Results for Attention Bottleneck for Crop Size 64x64

of integration into architectures. Regarding the use of the two types of data selection, it is noted that there is a general decrease in performance when all nodules are used. This reduction is more easily observed in the full image (512x512) than in the cropped images (64x64 and 96x96), this is likely due to the greater impact that the size and variety of the nodules have on the training process compared to the size of the image.

Another common use of attention mechanisms occurs in skip connections, which are used to facilitate the decoding process and reduce potential information loss. Employing attention mechanisms helps enhance the overall performance and adaptability of the network. In this work, AttentionUNet [83] is used as the basis for the tests, along with

R2AttentionUNet [96] and some variations of both utilizing the SAM introduced earlier in this section.

Three variations were generated: SAMAttentionUNet, which replaces the attention gate of AttentionUNet with SAM, as seen in Figure 5.6; MixAttentionUNet, which applies SAM immediately after the standard attention gate, as shown in Figure 5.7; and finally, R2AttentionUNetSAM, which also replaces the attention gate with SAM, as depicted in Figure 5.8.

U-Net and U-Net++ are in the results table for comparison.

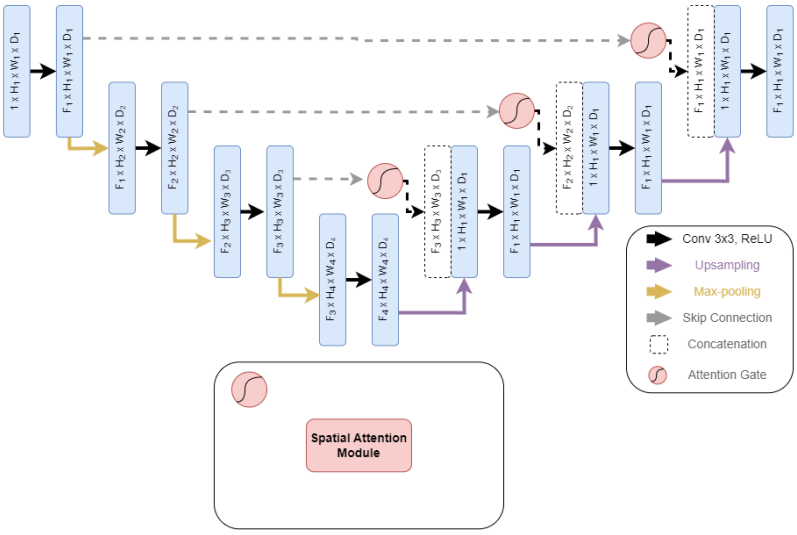

FIGURE 5.6: SAMAttentionUNet architecture

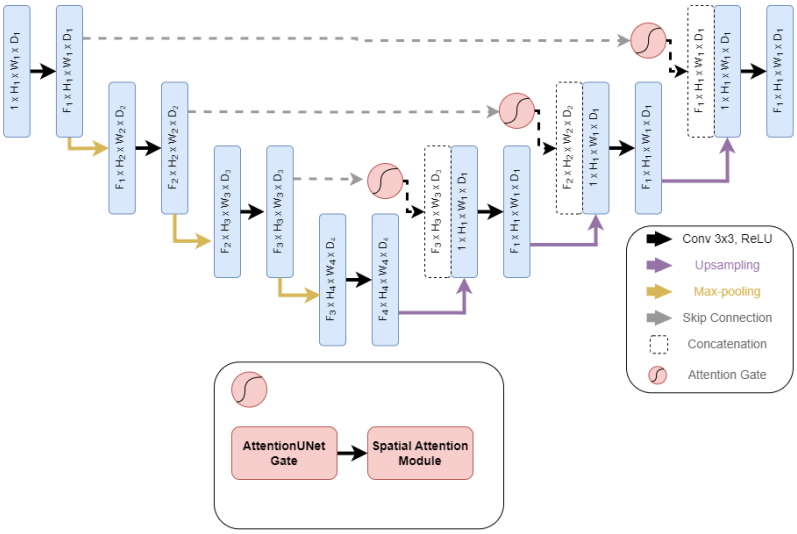

FIGURE 5.7: MixAttentionUNet architecture

The results of those experiments can be seen on tables 5.8, 5.9 and 5.10.

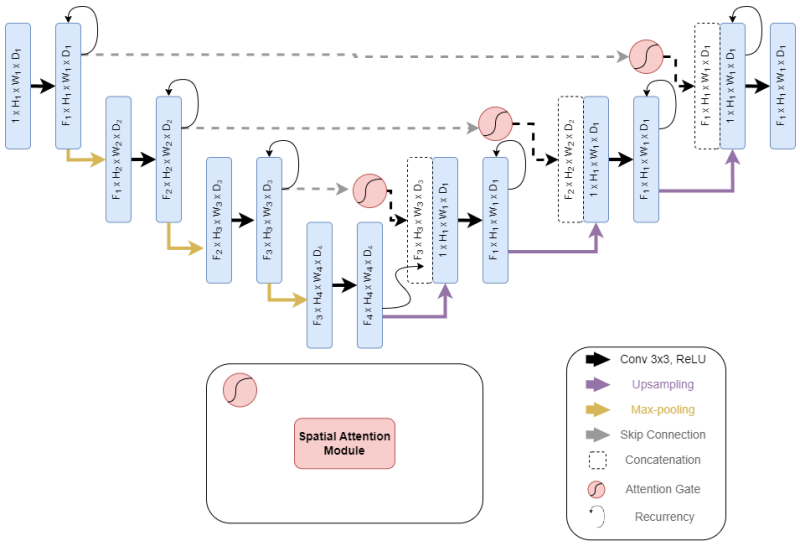

FIGURE 5.8: R2AttentionUNetSAM architecture

Model	DSC	Crop	Data Split
Mixed Attention U-Net	60.6%	512x512	All
Attention U-Net	60.3%	512x512	All
SAM Attention U-Net	**61.4%**	512x512	All
R2 Attention U-Net	44.1%	512x512	All
R2 Attention U-Net SAM	42.1%	512x512	All
U-Net	61.3%	512x512	All
U-Net++	60.7%	512x512	All
Mixed Attention U-Net	67.3%	512x512	Filtered
Attention U-Net	66.7%	512x512	Filtered
SAM Attention U-Net	64.2%	512x512	Filtered
R2 Attention U-Net	52.8%	512x512	Filtered
R2 Attention U-Net SAM	58.4%	512x512	Filtered
U-Net	**68.1%**	512x512	Filtered
U-Net++	67.7%	512x512	Filtered

TABLE 5.8: Results of Attention Mechanisms on Skip Connections for Crop Size 96x96

In the tests conducted using attention gates, no significant advantage was detected compared to other models. Overall, the performance was similar to or inferior to that of the basic U-Net. For models based on R2Attention U-Net, the difference is even more pronounced, as they exhibited the lowest performance and have a high computational cost, making them unsuitable for use in the context of this study.

Model	DSC	Crop	Data Split
Mixed Attention U-Net	82.3%	96x96	All
Attention U-Net	82.6%	96x96	All
SAM Attention U-Net	82.5%	96x96	All
R2 Attention U-Net	77.6%	96x96	All
R2 Attention U-Net SAM	76.0%	96x96	All
U-Net	**82.8%**	96x96	All
U-Net++	82.1%	96x96	All
Mixed Attention U-Net	83.4%	96x96	Filtered
Attention U-Net	83.6%	96x96	Filtered
SAM Attention U-Net	82.4%	96x96	Filtered
R2 Attention U-Net	82.1%	96x96	Filtered
R2 Attention U-Net SAM	80.3%	96x96	Filtered
U-Net	**83.9%**	96x96	Filtered
U-Net++	79.7%	96x96	Filtered

TABLE 5.9: Results of Attention Mechanisms on Skip Connections for Crop Size 96x96

Model	DSC	Crop	Data Split
Mixed Attention U-Net	83.1%	64x64	All
Attention U-Net	83.0%	64x64	All
SAM Attention U-Net	82.7%	64x64	All
R2 Attention U-Net	77.8%	64x64	All
R2 Attention U-Net SAM	74.9%	64x64	All
U-Net	**83.7%**	64x64	All
U-Net++	82.7%	64x64	All
Mixed Attention U-Net	**84.2%**	64x64	Filtered
Attention U-Net	83.6%	64x64	Filtered
SAM Attention U-Net	84%	64x64	Filtered
R2 Attention U-Net	81.7%	64x64	Filtered
R2 Attention U-Net SAM	80.7%	64x64	Filtered
U-Net	83.9%	64x64	Filtered
U-Net++	82.5%	64x64	Filtered

TABLE 5.10: Results of Attention Mechanisms on Skip Connections for Crop Size 64x64

5.5 Transformer Test

Despite having been used before, with the ViT-Unet presented in Section 5.1, this work also dedicated time to conduct studies using another vision transformer, the Swin Transformer [97], a component that has shown success in multiple applications of image segmentation. For the tests, two versions of networks were used: Swin-UNet [98] and Swin-UNetV2 [99], which incorporates the latest version of the vision transformer block. For both, two tests were performed, one without a pre-trained base model and another with a pre-trained model. The base code and pre-trained model were obtained from the official

repository [100]. Table 5.11 details the results of each test. For comparison, U-Net, SAM U-Net and U-Net++ were used to assess the performance.

Model	DSC	Pre-trained	Crop	Data Split
Swin-U-Net	**84.9%**	Yes	224x224	All
Swin-U-Net	83.9%	No	224x224	All
Swin-U-Net V2	84.6%	Yes	224x224	All
Swin-U-Net V2	83.4%	No	224x224	All
U-Net	82.1%	No	224x224	All
SAM U-Net	82.7%	No	224x224	All
U-Net++	81.4%	No	224x224	All
Swin-U-Net	81.3%	Yes	224x224	Filtered
Swin-U-Net	77.8%	No	224x224	Filtered
Swin-U-Net V2	79.2%	Yes	224x224	Filtered
Swin-U-Net V2	76.2%	No	224x224	Filtered
U-Net	**83.3%**	No	224x224	Filtered
SAM U-Net	82.9%	No	224x224	Filtered
U-Net++	76.7%	No	224x224	Filtered

TABLE 5.11: Results of Swin-UNet and Swin-UNet V2

As seen in Table 5.11, it is noticeable that the original version of the network achieves better results than the second version, which was not initially expected. It is also noteworthy that in both networks, the use of pre-trained models improves the results by almost 3% in the filtered data split and around 1% on all data, demonstrating that the use of pre-trained models in segmentation of different natures positively impacts lung nodule segmentation tasks, at least for the Swin Transformer. A curious result was that on Filtered data split, none of the models with Swin were able to achieve better performance than the base U-Net or the SAM U-Net, being only better than the U-Net++. This highlights the importance of having a greater variance of training images for models using the Swin Transformer, with a lesser impact on more traditional models.

5.6 Interactive Segmentation Test

Interactive Segmentation is an approach that involves using, in addition to the usual image and mask, annotations indicating the location of the object of interest, such as the annotator's clicks on the mask, a small bounding box at the object's central point, or even indicative texts [101]. With the help of such annotations, models can learn where to focus attention and thus achieve better performance and more precise results. Inspired by this

capability and by some works from other areas ([102] and [103]), this study conducted experiments to validate the potential use in lung nodule segmentation.

For this purpose, artificial points simulating the annotator's click in the areas of nodule location were generated and added as a new channel to the numpy array. These annotations were added to both the images and masks. Figure 5.9 exemplifies the addition of five points in the central region.

An additional model was tested based on the Mix Attention U-Net, where the click layer is used only in the attention gate and not throughout the entire training process, to evaluate whether there would be any distinction between using the consolidated image or separately. Tables 5.12, 5.13 and 5.14 contains the results of all the tests.

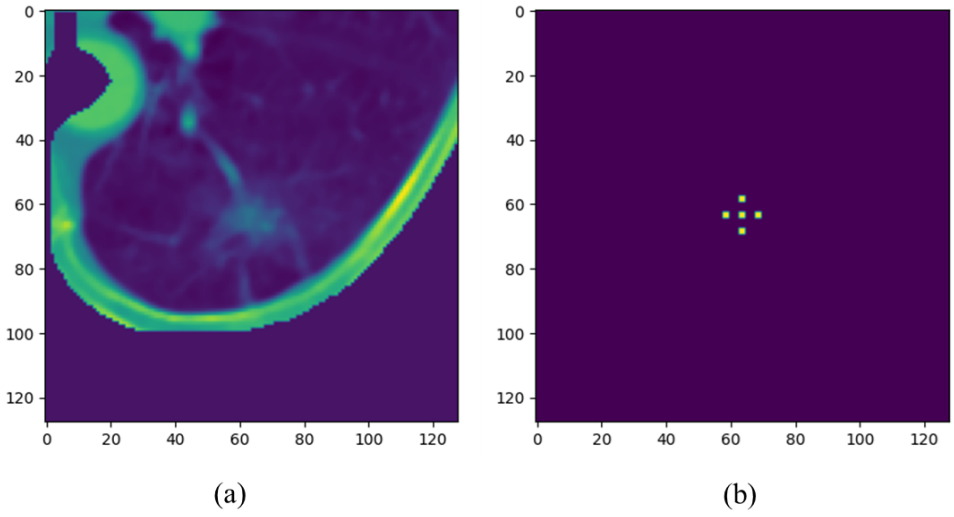

FIGURE 5.9: Point click annotation example. (a) Original ct slice;(b) Annotation

Model	DSC	Crop	Data Split
U-Net	58.6%	512x512	All
U-Net++	59.1%	512x512	All
SAM U-Net	**59.3%**	512x512	All
Attention U-Net	59.1%	512x512	All
Mix Attention U-Net	56.5%	512x512	All
Mix Attention U-Net Separated	55.6%	512x512	All
U-Net	62.6%	512x512	Filtered
U-Net++	61.6%	512x512	Filtered
SAM U-Net	**63.5%**	512x512	Filtered
Attention U-Net	62.1%	512x512	Filtered
Mix Attention U-Net	61.5%	512x512	Filtered
Mix Attention U-Net Separated	62.1%	512x512	Filtered

TABLE 5.12: Results of Interactive Segmentation Tests for Crop Size 512x512

Model	DSC	Crop	Data Split
U-Net	85.2%	96x96	All
U-Net++	84.8%	96x96	All
SAM U-Net	85.1%	96x96	All
Attention U-Net	85.1%	96x96	All
Mix Attention U-Net	85.2%	96x96	All
Mix Attention U-Net Separated	**85.9%**	96x96	All
U-Net	85.6%	96x96	Filtered
U-Net++	83.8%	96x96	Filtered
SAM U-Net	85.5%	96x96	Filtered
Attention U-Net	85.3%	96x96	Filtered
Mix Attention U-Net	**85.7%**	96x96	Filtered
Mix Attention U-Net Separated	85.6%	96x96	Filtered

TABLE 5.13: Results of Interactive Segmentation Tests for Crop Size 96x96

Model	DSC	Crop	Data Split
U-Net	85%	64x64	All
U-Net++	84.2%	64x64	All
SAM U-Net	85.2%	64x64	All
Attention U-Net	84.6%	64x64	All
Mix Attention U-Net	84.7%	64x64	All
Mix Attention U-Net Separated	**85.7%**	64x64	All
U-Net	84.9%	64x64	Filtered
U-Net++	83.9%	64x64	Filtered
SAM U-Net	85.2%	64x64	Filtered
Attention U-Net	85.1%	64x64	Filtered
Mix Attention U-Net	**85.9%**	64x64	Filtered
Mix Attention U-Net Separated	85.4%	64x64	Filtered

TABLE 5.14: Results of Interactive Segmentation Tests for Crop Size 64x64

In the experiments, none data augmentation were used, more details on section 5.7.

5.7 Limitations and failed experiments

Unfortunately, not all ideas and experiments yield satisfactory results. Some fail, and others are not feasible to execute. Here, a brief summary of some tests that were not well-executed will be presented.

Firstly, there was an attempt to expand the scope of the work to include the use of 3D datasets, as this format is widely used in the production of CT scans and can significantly support medical work by providing more detailed insights into lung nodules. However, when producing the 3D dataset, a unification of slices occurs, reducing the total number of data points to be learned while considerably increasing the size of each object. In the tests, a dataset of 4x512x512 was produced, resulting in a DSC of less than 30%. Due to the time required, further tests were postponed for future works.

Another test that did not yield better results involved Interactive Segmentation. Despite achieving good results, it was not possible to apply an effective data augmentation technique. When annotations were used together with data augmentation, the annotations were distorted, worsening the model's overall performance. Given the time constraints, there was no further investigation into a technique that could be effectively utilized.

Regarding Interactive Segmentation, since the aim of the work is to provide a general-use model that can be applied to various scenarios, using a model trained with Interactive Segmentation in real-world applications becomes impractical because clicks would not exist in the original scan. Thus, an attempt was made to evaluate the use of knowledge transfer from a model trained with annotations to one that would not use them. Unfortunately, the results were inferior to those of models trained directly on the original images.

Lastly, it was not possible to obtain good results with Swin Transformers when using full 512x512 images. As these components are designed for 224x224, the adaptation proved inadequate. Despite this, it remains a solution that needs to be further studied to harness its potential for real-world applications.

5.8 Evaluations and Proposed Framework

In this section, the final evaluation results on the datasets used will be presented. The evaluations are conducted on the LIDC datasets (with both data splits). Subsequently, the

best model was selected based on the overall best performance.

5.8.1 Evaluations

In this section, the models previously presented are evaluated on the testing data as described in the chapter 4. For this evaluations, the models selected were: ASPP U-Net, PYB U-Net, SAM U-Net, Attention U-Net, SAM Attention U-Net, U-Net and U-Net++. They were selected based on the overall performance so far on all crop sizes. While each model was trained in both data split, for the evaluation only the "All" split is used.

Model	Train DSC	Eval DSC	Data Split in Training
U-Net	61.2%	58.7%	All
U-Net	68.2%	47.9%	Filtered
U-Net++	60.7%	59.9%	All
U-Net++	67.7%	46.8%	Filtered
SAM U-Net	62.3%	**61.3%**	All
SAM U-Net	68.8%	50.6%	Filtered
ASPP U-Net	64%	60.6%	All
ASPP U-Net	68.8%	46.9%	Filtered
PYB U-Net	62.6%	60.7%	All
PYB U-Net	68.5%	47.4%	Filtered
Attention U-Net	60.3%	57.4%	All
Attention U-Net	66.7%	45.6%	Filtered
SAM Attention U-Net	61.4%	58.3%	All
SAM Attention U-Net	64.2%	54.2%	Filtered

TABLE 5.15: Results of Final Evaluation for Crop Size 512x512 on LIDC

Model	Train DSC	Eval DSC	Data Split in Training
U-Net	82.8%	**83.8%**	All
U-Net	83.9%	79.5%	Filtered
U-Net++	82.1%	81.6%	All
U-Net++	79.7%	78.7%	Filtered
SAM U-Net	82.8%	82.6%	All
SAM U-Net	84.4%	78.1%	Filtered
ASPP U-Net	82.2%	82%	All
ASPP U-Net	83.5%	77%	Filtered
PYB U-Net	82.8%	83%	All
PYB U-Net	84.1%	78.8%	Filtered
Attention U-Net	82.6%	82.7%	All
Attention U-Net	83.6%	77.1%	Filtered
SAM Attention U-Net	82.8%	81.7%	All
SAM Attention U-Net	83.3%	77.7%	Filtered

TABLE 5.16: Results of Final Evaluation for Crop Size 96x96 on LIDC

Model	Train DSC	Eval DSC	Data Split in Training
U-Net	83.7%	**83.8%**	All
U-Net	83.9%	79.5%	Filtered
U-Net++	82.7%	82%	All
U-Net++	82.5%	78.8%	Filtered
SAM U-Net	83.1%	82.9%	All
SAM U-Net	83.5%	78.9%	Filtered
ASPP U-Net	83.2%	83.4%	All
ASPP U-Net	83.1%	77.1%	Filtered
PYB U-Net	83.4%	83%	All
PYB U-Net	83.7%	79.5%	Filtered
Attention U-Net	83%	82.9%	All
Attention U-Net	83.6%	78.2%	Filtered
SAM Attention U-Net	83.1%	82%	All
SAM Attention U-Net	83.5%	80%	Filtered

TABLE 5.17: Results of Final Evaluation for Crop Size 64x64 on LIDC

5.8.2 Proposed Model

After all the evaluations, it can be noted the similarity and proximity among all the evaluated models. The difference between the models trained on the data split that only includes cancerous nodules and the models trained with all types of nodules is also clear. For the final model, this work chooses the SAM U-Net, given its overall stability during the various tests and its ability to achieve the best performance in the target study object, namely images with a total size of 512x512. As described in section 5.4, the model consists of a basic U-Net with an added spatial attention block in the bottleneck.

Finally, the chosen model is evaluated against the Medical Decathlon and LNDb datasets, presented in Section 4. Four training versions were created: one previously used with LIDC, one exclusively using Medical Decathlon, one exclusively using LNDb, and one combining all three datasets. This approach allows for evaluating the differences and capabilities of each dataset. All versions were tested on all datasets. In Figure 5.10, a set of evaluations and predictions is depicted.

As can be seen in Table 5.18, the best-performing model is the one trained with the combination of the three datasets. Despite having lower performance during training, the combination of datasets provided greater generalization capability, making this model superior to all others in the four evaluations. Is evident how a smaller dataset, such as LNDb and Medical Decathlon, negatively impacts the model's training and usability. The limited variety of the LNDb dataset prevent the proper training of models when trained solely on this data, resulting in low performance.

Model	Train DSC	Eval DSC	Dataset in training	Dataset in Evaluation
SAM U-Net	62.3%	61.3%	LIDC	LIDC
SAM U-Net	62.3%	48.6%	LIDC	LNDb
SAM U-Net	62.3%	0%	LIDC	Medical Decathlon
SAM U-Net	62.3%	53.5%	LIDC	Combined
SAM U-Net	0.06%	0.6%	LNDb	LIDC
SAM U-Net	0.06%	22.1%	LNDb	LNDb
SAM U-Net	0.06%	0.1%	LNDb	Medical Decathlon
SAM U-Net	0.06%	0.7%	LNDb	Combined
SAM U-Net	43%	0%	Medical Decathlon	LIDC
SAM U-Net	43%	0%	Medical Decathlon	LNDb
SAM U-Net	43%	50%	Medical Decathlon	Medical Decathlon
SAM U-Net	43%	0.4%	Medical Decathlon	Combined
SAM U-Net	60.8%	**61.4%**	Combined	LIDC
SAM U-Net	60.8%	**55.1%**	Combined	LNDb
SAM U-Net	60.8%	**56.3%**	Combined	Medical Decathlon
SAM U-Net	60.8%	**60.6%**	Combined	Combined

TABLE 5.18: Results of Proposed Model Evaluation

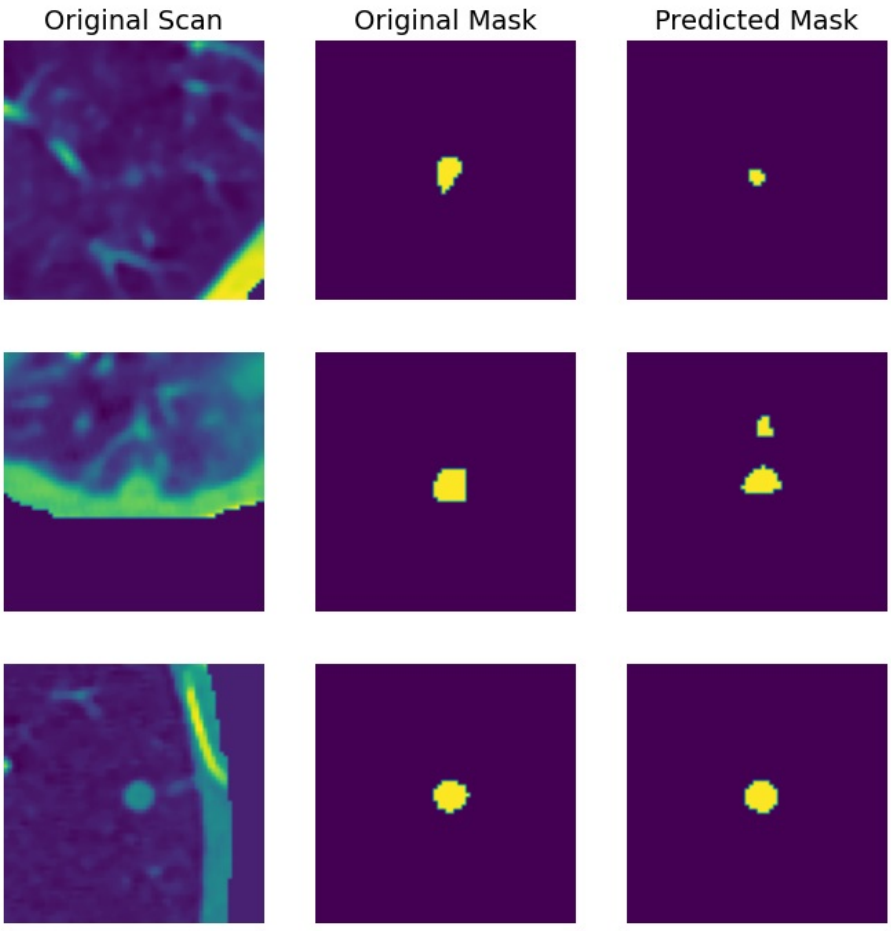

FIGURE 5.10: Evaluations Predictions Example

Chapter 6

Conclusions

Lung cancer is a major health issue due to its difficult and often late diagnosis, making the disease highly lethal and challenging to treat. Therefore, it is essential that efforts are directed towards improving all stages of treatment, facilitating the detection and correct segmentation of nodules by healthcare professionals, and consequently increasing the chances that the disease will not be fatal.

To contribute to this effort against lung cancer, this work focused its research on segmentation methods using neural networks. An extensive review of the state-of-the-art was conducted, which enabled the evaluation of numerous techniques and different ideas on how to approach the problem, as well as understanding the current limitations and the overall state of the technology.

After the initial data collection, a series of experiments and ideas were tested. The best data augmentation technique was identified, using translation and shearing on the training images, which resulted in a 7% increase in DSC compared to a U-Net model without any technique. Techniques that negatively impact the model's performance were also identified, with Gaussian noise reducing the DSC by up to 15%, thus demonstrating the importance of choosing appropriate techniques to expand the dataset to be used.

In addition to effective data augmentation techniques, two different methods of data split were tested: one containing only cancerous nodules and the other including all types of nodules. Although the first method performed better in numerous tests, it was found that the model's ability to generalize was lower than those models trained with the entire range of nodules. The latter models were much more stable and consistent in all results.

The experiments then focused on attention mechanisms, first in the bottleneck of neural networks and later with attention gates in the skip connections. Despite being close

in efficiency to standard models, it was possible to identify models with potential for use, especially methods utilizing spatial attention. Another test involved transformers, a more recent technique for nodule segmentation but with proven high capability in other areas. The results for 224x224 images were satisfactory, with DSC reaching up to 84%, but it was not possible to use higher resolution images.

A final test was conducted to examine the behavior and impact of Interactive Segmentation. The main idea was to train the models with an extra channel containing artificial clicks on the CT Scan, simulating the behavior of a specialist, and use those clicks as guides for the model's learning. Unfortunately, this approach did not prove to be effective, yielding worse results than the models without Interactive Segmentation A transfer knowledge from the interactive model to standard models were also unsuccessful.

The final proposed model, SAM U-Net, showed stable results for all crop sizes across different tests. For the main focus of the study, 512x512 images, the final performance was 61.4% on the LIDC. For the 64x64 and 96x96 crops, the performances, respectively, were 82.9% and 82.6%, values that are comparable to those obtained by other state-of-the-art works. Additionally, the final model was trained with a combination of the LIDC, LNDb, and Medical Decathlon datasets, thus achieving greater generalization and usability.

Finally, despite the results not reaching a high level of performance in the target image size, valuable insights and a knowledge base have been generated, allowing future studies to benefit from the tests and results provided here.

Bibliography

[1] World Health Organization, "The top 10 causes of death," Dec. 2020, accessed on 17.11.2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death [Cited on pages 1 and 2.]

[2] ——, "Cancer," Feb. 2022, accessed on 17.11.2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cancer [Cited on pages 1 and 2.]

[3] European Cancer Information System, "2022 new cancer cases and cancer deaths on the rise in the eu," Oct. 2023, accessed on 20.11.2023. [Online]. Available: https://ecis.jrc.ec.europa.eu/pdf/factsheets/CancerEstimates2022_factsheet.pdf [Cited on page 1.]

[4] Instituto Nacional de Câncer, "Estatísticas de câncer," Jun. 2022, accessed on 17.11.2023. [Online]. Available: https://www.gov.br/inca/pt-br/assuntos/cancer/numeros [Cited on page 1.]

[5] Instituto Nacional de Estatística, "Causas de morte," May. 2023, accessed on 17.11.2023. [Online]. Available: https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_destaques&DESTAQUESdest_boui=594417880&DESTAQUESmodo=2 [Cited on page 1.]

[6] American Lung Association, "State of lung cancer," accessed on 10.01.2023. [Online]. Available: https://www.lung.org/getmedia/381ca407-a4e9-4069-b24b-195811f29a00/solc-2020-report-final.pdf [Cited on page 2.]

[7] F. Ciompi, K. Chung, S. J. V. Riel, A. A. A. Setio, P. K. Gerke, C. Jacobs, E. T. Scholten, C. Schaefer-Prokop, M. M. Wille, A. Marchianò, U. Pastorino, M. Prokop, and B. V. Ginneken, "Towards automatic pulmonary nodule management in lung cancer screening with deep learning," *Scientific Reports*, vol. 7, 4 2017. [Online].

Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5395959/ [Cited on page 2.]

[8] European Cancer Information System, "Estimates of cancer incidence and mortality in 2022, for all countries," Oct. 2023, accessed on 20.11.2023. [Online]. Available: https://ecis.jrc.ec.europa.eu/index.php [Cited on page 3.]

[9] S. Singh, J. Maxwell, J. A. Baker, J. L. Nicholas, and J. Y. Lo, "Computer-aided classification of breast masses: Performance and interobserver variability of expert radiologists versus residents," *Radiology*, vol. 258, pp. 73–80, 2011, pMID: 20971779. [Online]. Available: https://doi.org/10.1148/radiol.10081308 [Cited on page 2.]

[10] J. G. Betts, K. A. Young, J. A. Wise, E. Johnson, B. Poe, D. H. Kruse, O. Korol, J. E. Johnson, M. Womble, and P. DeSaix, *Anatomy and Physiology*, 2nd ed. OpenStax, Apr 2022. [Online]. Available: https://openstax.org/books/anatomy-and-physiology-2e/pages/1-introduction [Cited on pages 5 and 6.]

[11] National Cancer Institute, "Understanding cancer - nci," accessed on 19.11.2023. [Online]. Available: https://www.cancer.gov/about-cancer/understanding [Cited on page 6.]

[12] Centers for Disease Control and Prevention, "What is lung cancer? — cdc," accessed on 19.11.2023. [Online]. Available: https://www.cdc.gov/lung-cancer/about/index.html [Cited on pages 6 and 7.]

[13] J. Malhotra, M. Malvezzi, E. Negri, C. L. Vecchia, and P. Boffetta, "Risk factors for lung cancer worldwide series thoracic oncology," *Eur Respir J*, vol. 48, pp. 889–902, 2016, accessed on 17.11.2023. [Online]. Available: https://erj.ersjournals.com/content/48/3/889 [Cited on page 7.]

[14] American Cancer Society, "Lung cancer risk factors," accessed on 19.11.2023. [Online]. Available: https://www.cancer.org/cancer/types/lung-cancer/causes-risks-prevention/risk-factors.html [Cited on page 7.]

[15] National Cancer Institute, "Non-small cell lung cancer treatment (pdq®)–health professional version," accessed on 19.11.2023. [Online]. Available: https://www.cancer.gov/types/lung/hp/non-small-cell-lung-treatment-pdq [Cited on page 7.]

[16] ——, "Small cell lung cancer treatment (pdq®)–health professional version," accessed on 19.11.2023. [Online]. Available: https://www.cancer.gov/types/lung/hp/small-cell-lung-treatment-pdq [Cited on page 7.]

[17] D. Panetta, "2.03 - computed tomography," in *Comprehensive Biomedical Physics*, A. Brahme, Ed. Oxford: Elsevier, 2014, pp. 65–88. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780444536327002033 [Cited on page 8.]

[18] P. Allisy-Roberts and J. Williams, "Chapter 7 - computed tomography," in *Farr's Physics for Medical Imaging (Second Edition)*, 2nd ed., P. Allisy-Roberts and J. Williams, Eds. W.B. Saunders, 2008, pp. 103–119. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780702028441500119 [Cited on page 8.]

[19] C. Liu, R. Zhao, W. Xie, and M. Pang, "Pathological lung segmentation based on random forest combined with deep model and multi-scale superpixels," *Neural Processing Letters*, vol. 52, 10 2020, accessed on 20.11.2023. [Online]. Available: https://link.springer.com/article/10.1007/s11063-020-10330-8 [Cited on page 9.]

[20] K. Loverdos, A. Fotiadis, C. Kontogianni, M. Iliopoulou, and M. Gaga, "Lung nodules: A comprehensive review on current approach and management," *Annals of Thoracic Medicine*, vol. 14, p. 226, 10 2019, accessed on 20.11.2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6784443/ [Cited on page 9.]

[21] S. Mukhopadhyay, "A segmentation framework of pulmonary nodules in lung ct images," *Journal of digital imaging*, vol. 29, 06 2015, accessed on 20.11.2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4722030/ [Cited on page 10.]

[22] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics*, vol. 5, no. 4, pp. 115–133, Dec 1943, accessed on 10.11.2023. [Online]. Available: https://doi.org/10.1007/BF02478259 [Cited on page 10.]

[23] A. Ivakhnenko, V. Lapa, and P. U. L. I. S. O. E. ENGINEERING., *Cybernetic Predicting Devices*, ser. JPRS 37, 803. Joint Publications Research Service

[available from the Clearinghouse for Federal Scientific and Technical Information], 1965. [Online]. Available: https://books.google.co.in/books?id=l38DHQAACAAJ [Cited on page 10.]

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, p. 84–90, may 2017, accessed on 10.11.2023. [Online]. Available: https://doi.org/10.1145/3065386 [Cited on page 10.]

[25] W. Zhang, G. Yang, Y. Lin, C. Ji, and M. M. Gupta, "On definition of deep learning," in *2018 World Automation Congress (WAC)*, 2018, pp. 1–5, accessed on 27.11.2023. [Cited on page 10.]

[26] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998, accessed on 27.11.2023. [Cited on page 10.]

[27] Y. Gu, J. Chi, J. Liu, L. Yang, B. Zhang, D. Yu, Y. Zhao, and X. Lu, "A survey of computer-aided diagnosis of lung nodules from ct scans using deep learning," *Computers in Biology and Medicine*, vol. 137, p. 104806, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482521006004 [Cited on pages 10 and 11.]

[28] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, "Review of deep learning: concepts, cnn architectures, challenges, applications, future directions," *Journal of Big Data*, vol. 8, p. 53, 12 2021, accessed on 27.11.2023. [Online]. Available: /pmc/articles/PMC8010506//pmc/articles/PMC8010506/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC8010506/ [Cited on pages 10 and 11.]

[29] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015, accessed on 27.11.2023. [Online]. Available: https://doi.org/10.48550/arXiv.1505.04597 [Cited on pages 11, 12, 17, 18, and 45.]

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015, accessed on 27.11.2023. [Cited on page 11.]

[31] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, Oct 1986. [Online]. Available: https://doi.org/10.1038/323533a0 [Cited on page 12.]

[32] G. E. Hinton and R. Zemel, *Autoencoders, Minimum Description Length and Helmholtz Free Energy*, J. Cowan, G. Tesauro, and J. Alspector, Eds. Morgan-Kaufmann, 1993, vol. 6, accessed on 10.11.2023. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/1993/file/9e3cfc48eccf81a0d57663e129aef3cb-Paper.pdf [Cited on page 12.]

[33] D. Bank, N. Koenigstein, and R. Giryes, "Autoencoders," 2021, accessed on 10.11.2023. [Online]. Available: https://doi.org/10.48550/arXiv.2003.05991 [Cited on page 12.]

[34] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," *Proceedings of the 25th International Conference on Machine Learning*, pp. 1096–1103, 01 2008, accessed on 10.11.2023. [Online]. Available: https://www.cs.toronto.edu/~larocheh/publications/icml-2008-denoising-autoencoders.pdf [Cited on pages 12 and 13.]

[35] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2022, accessed on 10.11.2023. [Online]. Available: https://doi.org/10.48550/arXiv.1312.6114 [Cited on page 12.]

[36] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020, accessed on 10.11.2023. [Online]. Available: https://doi.org/10.48550/arXiv.1406.2661 [Cited on page 13.]

[37] S. G. Armato, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman, E. A. Kazerooni, H. MacMahon, E. J. R. van Beek, D. Yankelevitz, A. M. Biancardi, P. H. Bland, M. S. Brown, R. M. Engelmann, G. E. Laderach, D. Max, R. C. Pais, D. P. Qing, R. Y. Roberts, A. R. Smith, A. Starkey, P. Batra, P. Caligiuri, A. Farooqi, G. W. Gladish, C. M. Jude, R. F. Munden, I. Petkovska, L. E. Quint, L. H. Schwartz, B. Sundaram, L. E. Dodd, C. Fenimore, D. Gur, N. Petrick, J. Freymann, J. Kirby, B. Hughes, A. V. Casteele, S. Gupte, M. Sallam, M. D. Heath, M. H. Kuhn,

E. Dharaiya, R. Burns, D. S. Fryd, M. Salganicoff, V. Anand, U. Shreter, S. Vastagh, B. Y. Croft, and L. P. Clarke, "The lung image database consortium (lidc) and image database resource initiative (idri): A completed reference database of lung nodules on ct scans," *Medical Physics*, vol. 38, pp. 915–931, 2 2011, accessed on 20.11.2023. [Online]. Available: https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=1966254 [Cited on page 13.]

[38] Grand Challenge, "Lung nodule analysis 2016," 2017, accessed on 17.11.2023. [Online]. Available: https://luna16.grand-challenge.org/ [Cited on page 14.]

[39] A. A. A. Setio, A. Traverso, T. de Bel, M. S. Berens, C. van den Bogaard, P. Cerello, H. Chen, Q. Dou, M. E. Fantacci, B. Geurts, R. van der Gugten, P. A. Heng, B. Jansen, M. M. de Kaste, V. Kotov, J. Y.-H. Lin, J. T. Manders, A. Sóñora-Mengana, J. C. García-Naranjo, E. Papavasileiou, M. Prokop, M. Saletta, C. M. Schaefer-Prokop, E. T. Scholten, L. Scholten, M. M. Snoeren, E. L. Torres, J. Vandemeulebroucke, N. Walasek, G. C. Zuidhof, B. van Ginneken, and C. Jacobs, "Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The luna16 challenge," *Medical Image Analysis*, vol. 42, pp. 1–13, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841517301020 [Cited on page 14.]

[40] B. van Ginneken, S. G. Armato, B. de Hoop, S. van Amelsvoort-van de Vorst, T. Duindam, M. Niemeijer, K. Murphy, A. Schilham, A. Retico, M. E. Fantacci, N. Camarlinghi, F. Bagagli, I. Gori, T. Hara, H. Fujita, G. Gargano, R. Bellotti, S. Tangaro, L. Bolaños, F. D. Carlo, P. Cerello, S. Cristian Cheran, E. Lopez Torres, and M. Prokop, "Comparing and combining algorithms for computer-aided detection of pulmonary nodules in computed tomography scans: The anode09 study," *Medical Image Analysis*, vol. 14, no. 6, pp. 707–722, 2010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841510000587 [Cited on page 14.]

[41] Grand Challenge, "Automatic nodule detection 2009," 2010, accessed on 17.11.2023. [Online]. Available: https://anode09.grand-challenge.org/Home/ [Cited on page 14.]

[42] Cornell University, "Elcap - simba chest health analysis system public lung image database," 2010, accessed on 17.11.2023. [Online]. Available: https://www.via.cornell.edu/databases/simbadb.html [Cited on page 14.]

[43] A. P. Reeves, Y. Xie, and S. Liu, "Large-scale image region documentation for fully automated image biomarker algorithm development and evaluation," *Journal of Medical Imaging*, vol. 4, no. 2, p. 024505, 2017. [Online]. Available: https://doi.org/10.1117/1.JMI.4.2.024505 [Cited on page 14.]

[44] C. Wang, R. Xu, S. Xu, W. Meng, J. Xiao, Q. Peng, and X. Zhang, "Softgan: Towards accurate lung nodule segmentation via soft mask supervision," in *2022 IEEE International Conference on Multimedia and Expo (ICME)*, 2022, pp. 1–6, accessed on 10.11.2023. [Online]. Available: https://doi.org/10.1109/ICME52920.2022.9859677 [Cited on pages 14, 26, 28, and 33.]

[45] K. Yan, X. Wang, L. Lu, and R. M. Summers, "Deeplesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning," *Journal of medical imaging*, vol. 5, no. 3, pp. 036 501–036 501, 2018, accessed on 10.11.2023. [Online]. Available: https://doi.org/10.1117/1.jmi.5.3.036501 [Cited on page 14.]

[46] S. Tyagi and S. N. Talbar, "Cse-gan: A 3d conditional generative adversarial network with concurrent squeeze-and-excitation blocks for lung nodule segmentation," *Computers in Biology and Medicine*, vol. 147, p. 105781, 2022, accessed on 10.11.2023. [Online]. Available: https://doi.org/10.1016/j.compbiomed.2022.105781 [Cited on pages 14, 25, 26, 27, and 33.]

[47] Grand Challenge, "Lndb challenge," 2019. [Online]. Available: https://lndb.grand-challenge.org/ [Cited on page 14.]

[48] J. Pedrosa, G. Aresta, C. Ferreira, M. Rodrigues, P. Leitão, A. S. Carvalho, J. Rebelo, E. Negrão, I. Ramos, A. Cunha *et al.*, "Lndb: a lung nodule database on computed tomography," *arXiv preprint arXiv:1911.08434*, 2019. [Online]. Available: https://doi.org/10.48550/arXiv.1911.08434 [Cited on page 14.]

[49] M. Antonelli, A. Reinke, S. Bakas, K. Farahani, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze, O. Ronneberger, R. M. Summers *et al.*, "The

medical segmentation decathlon," *Nature communications*, vol. 13, no. 1, p. 4128, 2022. [Online]. Available: https://doi.org/10.1038/s41467-022-30695-9 [Cited on page 14.]

[50] A. L. Simpson, M. Antonelli, S. Bakas, M. Bilello, K. Farahani, B. Van Ginneken, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze *et al.*, "A large annotated medical image dataset for the development and evaluation of segmentation algorithms," *arXiv preprint arXiv:1902.09063*, 2019. [Online]. Available: https://doi.org/10.48550/arXiv.1902.09063 [Cited on page 14.]

[51] P. M. Bruntha, S. I. A. Pandian, K. M. Sagayam, S. Bandopadhyay, M. Pomplun, and H. Dang, "Lung_paynet: a pyramidal attention based deep learning network for lung nodule segmentation," *Scientific Reports*, vol. 12, no. 1, p. 20330, 2022. [Online]. Available: https://doi.org/10.1038/s41598-022-24900-4 [Cited on pages 18, 33, and 51.]

[52] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520. [Online]. Available: https://doi.org/10.48550/arXiv.1801.04381 [Cited on page 18.]

[53] H. Li, P. Xiong, J. An, and L. Wang, "Pyramid attention network for semantic segmentation," *arXiv preprint arXiv:1805.10180*, 2018. [Online]. Available: https://doi.org/10.48550/arXiv.1805.10180 [Cited on page 18.]

[54] K. Wang, X. Zhang, X. Zhang, Y. Lu, S. Huang, and D. Yang, "Eanet: Iterative edge attention network for medical image segmentation," *Pattern Recognition*, vol. 127, p. 108636, 2022. [Online]. Available: https://doi.org/10.1016/j.patcog.2022.108636 [Cited on pages 18, 19, and 33.]

[55] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014. [Online]. Available: https://doi.org/10.48550/arXiv.1409.1556 [Cited on page 18.]

[56] T. Chen, F. Luo, F. Zhou, F. Liu, and Y. Zha, "Vifunet: a vision flash based unet for lung nodules segmentation task," in *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2022, pp. 1481–1486. [Online].

Available: https://doi.org/10.1109/BIBM55620.2022.9995350 [Cited on pages 19, 20, and 33.]

[57] W. Hua, Z. Dai, H. Liu, and Q. Le, "Transformer quality in linear time," in *International Conference on Machine Learning*. PMLR, 2022, pp. 9099–9117. [Online]. Available: https://doi.org/10.48550/arXiv.2202.10447 [Cited on page 19.]

[58] S. Chen and Y. Wang, "Pulmonary nodule segmentation in computed tomography with an encoder-decoder architecture," in *2019 10th International Conference on Information Technology in Medicine and Education (ITME)*. IEEE, 2019, pp. 157–162. [Online]. Available: https://doi.org/10.1109/ITME.2019.00045 [Cited on pages 19, 20, and 33.]

[59] 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: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*. Springer, 2018, pp. 3–11. [Online]. Available: https://doi.org/10.48550/arXiv.1807.10165 [Cited on pages 20 and 45.]

[60] S. A. Agnes, A. A. Solomon, and K. Karthick, "Wavelet u-net++ for accurate lung nodule segmentation in ct scans: Improving early detection and diagnosis of lung cancer," *Biomedical Signal Processing and Control*, vol. 87, p. 105509, 2024. [Online]. Available: https://doi.org/10.1016/j.bspc.2023.105509 [Cited on pages 20, 21, and 33.]

[61] T. Wang, F. Wu, H. Lu, and S. Xu, "Ca-unet: Convolution and attention fusion for lung nodule segmentation," *International Journal of Imaging Systems and Technology*, 2023. [Online]. Available: https://doi.org/10.1002/ima.22878 [Cited on pages 21, 22, and 33.]

[62] M. Usman and Y.-G. Shin, "Deha-net: A dual-encoder-based hard attention network with an adaptive roi mechanism for lung nodule segmentation," *Sensors*, vol. 23, no. 4, p. 1989, 2023. [Online]. Available: https://doi.org/10.3390/s23041989 [Cited on pages 22, 23, 24, and 33.]

[63] Y. Cai, Z. Liu, Y. Zhang, and Z. Yang, "Mdfn: A multi-level dynamic fusion network with self-calibrated edge enhancement for lung nodule segmentation," *Biomedical Signal Processing and Control*, vol. 87, p. 105507, 2024. [Online]. Available: https://doi.org/10.1016/j.bspc.2023.105507 [Cited on pages 23, 24, and 33.]

[64] S. Wang, A. Jiang, X. Li, Y. Qiu, M. Li, and F. Li, "Dpbet: A dual-path lung nodules segmentation model based on boundary enhancement and hybrid transformer," *Computers in Biology and Medicine*, vol. 151, p. 106330, 2022. [Online]. Available: https://doi.org/10.1016/j.compbiomed.2022.106330 [Cited on pages 24, 25, and 33.]

[65] A. Levin, D. Lischinski, and Y. Weiss, "A closed-form solution to natural image matting," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 2, pp. 228–242, 2007. [Online]. Available: https://doi.org/10.1109/TPAMI.2007.1177 [Cited on page 27.]

[66] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. [Cited on page 28.]

[67] H. Cao, H. Liu, E. Song, C.-C. Hung, G. Ma, X. Xu, R. Jin, and J. Lu, "Dual-branch residual network for lung nodule segmentation," *Applied Soft Computing*, vol. 86, p. 105934, 2020. [Online]. Available: https://doi.org/10.1016/j.asoc.2019.105934 [Cited on pages 28, 29, and 33.]

[68] Y. Qin, H. Zheng, X. Huang, J. Yang, and Y.-M. Zhu, "Pulmonary nodule segmentation with ct sample synthesis using adversarial networks," *Medical physics*, vol. 46, no. 3, pp. 1218–1229, 2019. [Online]. Available: https://doi.org/10.1002/mp.13349 [Cited on pages 29, 30, 31, and 33.]

[69] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*. Springer, 2016, pp. 424–432. [Online]. Available: https://doi.org/10.48550/arXiv.1606.06650 [Cited on page 29.]

[70] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 7, pp. 971–987, 2002. [Online]. Available: https://doi.org/10.1109/TPAMI.2002.1017623 [Cited on page 30.]

[71] J. Canny, "A computational approach to edge detection," *IEEE Transactions on pattern analysis and machine intelligence*, no. 6, pp. 679–698, 1986. [Online]. Available: https://doi.org/10.1109/TPAMI.1986.4767851 [Cited on page 30.]

[72] I. Sobel, R. Duda, P. Hart, and J. Wiley, "Sobel-feldman operator," *Preprint at https://www. researchgate. net/profile/Irwin-Sobel/publication/285159837. Accessed*, vol. 20, 2022. [Cited on page 30.]

[73] N. Khosravan and U. Bagci, "Semi-supervised multi-task learning for lung cancer diagnosis," in *2018 40th Annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. IEEE, 2018, pp. 710–713. [Online]. Available: https://doi.org/10.1109/EMBC.2018.8512294 [Cited on pages 31 and 33.]

[74] W. Chen, Q. Wang, D. Yang, X. Zhang, C. Liu, and Y. Li, "End-to-end multi-task learning for lung nodule segmentation and diagnosis," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 6710–6717. [Online]. Available: https://doi.org/10.1109/ICPR48806.2021.9412218 [Cited on pages 31, 32, and 33.]

[75] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE transactions on systems, man, and cybernetics*, vol. 9, no. 1, pp. 62–66, 1979. [Online]. Available: https://cw.fel.cvut.cz/b201/_media/courses/a6m33bio/otsu.pdf [Cited on page 32.]

[76] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012. [Online]. Available: https://doi.org/10.1109/TPAMI.2012.120 [Cited on page 32.]

[77] M. C. Hancock and J. F. Magnan, "Lung nodule malignancy classification using only radiologist-quantified image features as inputs to statistical learning algorithms: probing the lung image database consortium dataset with two

statistical learning methods," *Journal of Medical Imaging*, vol. 3, no. 4, pp. 044 504–044 504, 2016. [Online]. Available: https://doi.org/10.1117/1.JMI.3.4.044504 [Cited on page 35.]

[78] M. Hancock, "Pylidc," 2016. [Online]. Available: https://pylidc.github.io/index.html [Cited on page 36.]

[79] J. Kim, "Lidc-idri-preprocessing," 2020. [Online]. Available: https://github.com/jaeho3690/LIDC-IDRI-Preprocessing [Cited on page 40.]

[80] B. A. Hamilton, "Dsb3tutorial," 2016. [Online]. Available: https://github.com/booz-allen-hamilton/DSB3Tutorial/tree/master [Cited on page 40.]

[81] Z. Ali, A. Irtaza, and M. Maqsood, "An efficient u-net framework for lung nodule detection using densely connected dilated convolutions," *The Journal of Supercomputing*, vol. 78, no. 2, pp. 1602–1623, 2022. [Online]. Available: https://doi.org/10.1007/s11227-021-03845-x [Cited on page 40.]

[82] R. Hussain and H. Basak, "Ut-net: Combining u-net and transformer for joint optic disc and cup segmentation and glaucoma detection," *arXiv preprint arXiv:2303.04939*, 2023. [Online]. Available: https://arxiv.org/abs/2303.04939 [Cited on page 45.]

[83] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz *et al.*, "Attention u-net: Learning where to look for the pancreas. arxiv 2018," *arXiv preprint arXiv:1804.03999*, 2018. [Online]. Available: https://arxiv.org/abs/1804.03999 [Cited on pages 45 and 53.]

[84] J. Kim, "Lidc segmentation of lung nodule with u-net, u-net++," 2020. [Online]. Available: https://github.com/jaeho3690/LIDC-IDRI-Segmentation [Cited on page 45.]

[85] Y. Gao, "Utnet," 2021. [Online]. Available: https://github.com/yhygao/UTNet [Cited on page 45.]

[86] L. J. Hun, "Pytorch implementation of u-net, r2u-net, attention u-net, attention r2u-net," 2018. [Online]. Available: https://github.com/LeeJunHyun/Image_Segmentation [Cited on page 45.]

[87] U. Ruby and V. Yendapalli, "Binary cross entropy with deep learning technique for image classification," *Int. J. Adv. Trends Comput. Sci. Eng*, vol. 9, no. 10, 2020. [Online]. Available: https://www.warse.org/IJATCSE/static/pdf/file/ijatcse175942020.pdf [Cited on page 45.]

[88] A. Semenov, V. Boginski, and E. L. Pasiliao, "Neural networks with multidimensional cross-entropy loss functions," in *Computational Data and Social Networks: 8th International Conference, CSoNet 2019, Ho Chi Minh City, Vietnam, November 18–20, 2019, Proceedings 8*. Springer, 2019, pp. 57–62. [Online]. Available: https://doi.org/10.1007/978-3-030-34980-6_5 [Cited on pages 45 and 46.]

[89] K. H. Zou, S. K. Warfield, A. Bharatha, C. M. Tempany, M. R. Kaus, S. J. Haker, W. M. Wells III, F. A. Jolesz, and R. Kikinis, "Statistical validation of image segmentation quality based on a spatial overlap index1: scientific reports," *Academic radiology*, vol. 11, no. 2, pp. 178–189, 2004. [Online]. Available: https://doi.org/10.1016/S1076-6332(03)00671-8 [Cited on page 45.]

[90] S. Connor and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019. [Online]. Available: https://doi.org/10.1186/s40537-019-0197-0 [Cited on page 47.]

[91] E. Goceri, "Medical image data augmentation: techniques, comparisons and interpretations," *Artificial Intelligence Review*, vol. 56, no. 11, pp. 12 561–12 605, 2023. [Online]. Available: https://doi.org/10.1007/s10462-023-10453-z [Cited on page 47.]

[92] 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*, vol. 40, no. 4, pp. 834–848, 2017. [Online]. Available: https://doi.org/10.48550/arXiv.1606.00915 [Cited on page 50.]

[93] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141. [Online]. Available: https://doi.org/10.48550/arXiv.1709.01507 [Cited on page 51.]

[94] 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). IEEE, 2021, pp. 1236–1242. [Online]. Available: https://doi.org/10.48550/arXiv.2004.03696 [Cited on page 51.]

[95] 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. [Online]. Available: https://doi.org/10.48550/arXiv.2010.11929 [Cited on page 51.]

[96] Q. Zuo, S. Chen, and Z. Wang, "R2au-net: attention recurrent residual convolutional neural network for multimodal medical image segmentation," *Security and Communication Networks*, vol. 2021, pp. 1–10, 2021. [Online]. Available: https://doi.org/10.1155/2021/6625688 [Cited on page 54.]

[97] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022. [Online]. Available: https://doi.org/10.48550/arXiv.2103.14030 [Cited on page 56.]

[98] 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," 2021. [Online]. Available: https://doi.org/10.48550/arXiv.2105.05537 [Cited on page 56.]

[99] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong *et al.*, "Swin transformer v2: Scaling up capacity and resolution," *arXiv preprint arXiv:2111.09883*, 2021. [Cited on page 56.]

[100] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet," 2021. [Online]. Available: https://github.com/HuCaoFighting/Swin-Unet [Cited on page 57.]

[101] K. Sofiiuk, I. A. Petrov, and A. Konushin, "Reviving iterative training with mask guidance for interactive segmentation," in *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2022, pp. 3141–3145. [Online]. Available: https://doi.org/10.48550/arXiv.2102.06583 [Cited on page 57.]

[102] K. Li, G. Vosselman, and M. Y. Yang, "Interactive image segmentation with cross-modality vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 762–772. [Online]. Available: https://doi.org/10.48550/arXiv.2307.02280 [Cited on page 58.]

[103] Y. Huang, H. Yang, K. Sun, S. Zhang, L. Cao, G. Jiang, and R. Ji, "Interformer: Real-time interactive image segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 22 301–22 311. [Online]. Available: https://doi.org/10.48550/arXiv.2304.02942 [Cited on page 58.]