# LIVER TUMOR SEGMENTATION USING DEEP LEARNING

by

LEMONAKIS GEORGIOS

Electrical and Computer Engineering, Aristotle University of Thessaloniki, 1996

A THESIS

Submitted in partial fulfillment of the requirements for the degree

MASTER OF SCIENCE

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

SCHOOL OF ENGINEERING

HELLENIC MEDITERRANEAN UNIVERSITY

2023

Approved by:
Dr. Karampidis Konstantinos

# Abstract

Liver is a very important organ of the human body, responsible for many crucial functions. Since liver cancer is among the ones that cause many deaths worldwide, and metastases frequently have liver origin, it is essential to have high quality and accurate liver and tumor segmentations for early cancer diagnosis. From the other hand, manual identification and segmentation of lesions in three-dimensional CT scans requires too much time, is difficult to reproduce, and the segmentation results depend on the operator. To overcome these problems, many Deep Learning models have been utilized with promising results.

This work is a Deep Learning approach for liver and tumor segmentation, more specifically two U-net (convolutional neural network (CNN)) variants, namely the Resnet34 and Resnet50 models, both pre-trained and afterwards fine-tuned (transfer learning), and tested on 131 CT scans of the LiTS (Liver Tumor Segmentation) challenge dataset. A 10% of these CT scans were used for evaluation of the models, and the highest DSC scores achieved were 0.926 for liver and 0.619 for tumor segmentation.

# Περίληψη

Το συκώτι είναι ένα πάρα πολύ σημαντικό όργανο του ανθρώπινου σώματος, υπεύθυνο για πολλές ζωτικές λειτουργίες. Επειδή ο καρκίνος του συκωτιού είναι μέσα στους πιο θανατηφόρους σε παγκόσμιο επίπεδο, είναι πολύ σημαντικό να υπάρχουν τεχνικές ανίχνευσής του από ιατρικές εικόνες για την έγκαιρη διάγνωσή του. Από την άλλη πλευρά, η χειροκίνητη διάγνωση του καρκίνου του συκωτιού από ειδικούς απαιτεί πολύ χρόνο, είναι δύσκολη και υποκειμενική διαδικασία. Για αυτούς τους λόγους, πολλές Deep Learning μέθοδοι έχουν προταθεί για την κατάτμηση (segmentation) συκωτιού και όγκων από εικόνες αξονικής τομογραφίας (CT scans) με ελπιδοφόρα αποτελέσματα.

Σε αυτήν την εργασία θα ασχοληθούμε με μια μέθοδο Deep Learning για κατάτμηση (segmentation) συκωτιού και όγκων από εικόνες αξονικής τομογραφίας (CT scans) και πιο συγκεκριμένα με δύο παραλλαγές του U-net (Convolutional neural network (CNN)), τα Resnet34 και Resnet50, τα οποία είναι προ –εκπαιδευμένα (transfer learning) και εκπαιδεύτηκαν περεταίρω (fine-tuning) και δοκιμάστηκαν (testing) σε 131 αξονικές τομογραφίες από το σετ δεδομένων LITS (Liver tumor segmentation). Το 10% από τις αξονικές χρησιμοποιήθηκε για την αξιολόγηση των εξεταζόμενων μοντέλων και οι καλύτερες επιδόσεις όσον αφορά στο DSC score για την κατάτμηση (segmentation) εικόνας ήταν 0.926 για το συκώτι και 0.619 για τους όγκους.

# Acknowledgements

I would like to thank Prof. Georgios Papadourakis and Dr. Konstantinos Karampidis for their support and help, to implement and complete this thesis work.

# Table of contents

Abstract ..... ii
Περίληψη ..... iii
Acknowledgements ..... iv
Table of contents ..... 1
List of figures ..... 4
List of Tables ..... 6
Dedication ..... 7
1 Introduction ..... 8
2 Medical imaging ..... 13
2.1 Imaging Modalities ..... 13
2.1.1 Tomography ..... 13
2.1.2 Magnetic Resonance Imaging (MRI) ..... 14
2.1.3 Ultrasound ..... 14
2.1.4 Nuclear Medicine ..... 15
2.2 Medical image file formats ..... 16
3 Pre-processing techniques for medical imaging ..... 18
3.1 Hounsefield unit ..... 18
3.2 Histogram equalization ..... 19
3.3 Image denoising ..... 20
3.4 Resolution enhancement ..... 20
3.5 Data normalization ..... 21
3.6 Data standardization ..... 21
4 Image segmentation ..... 23
4.1 Segmentation methods ..... 25
4.1.1 Detection of Point, Edge, Line ..... 25
4.1.2 Thresholding ..... 25
4.1.3 Region-Based Segmentation ..... 27
4.2 Segmentation evaluation metrics in medical imaging ..... 28

4.2.1 Spatial overlap-based metrics. ..... 29
4.2.2 Volume based metrics. ..... 33
4.2.3 Semantic Segmentation metrics failures ..... 34
4.3 Liver and liver tumor segmentation ..... 35
4.3.1 Lesion size effect on segmentation performance ..... 35
4.3.2 Image contrast effect on segmentation performance ..... 36
4.3.3 Other challenges ..... 36
5 Deep Learning ..... 38
5.1 Deep learning architectures in liver and liver lesion segmentation ..... 40
5.1.1 Convolutional Neural Networks (CNNs) ..... 41
5.1.2 U-Net Architecture ..... 43
5.1.3 U-Net variants ..... 44
5.1.4 Fully convolution NNs (FCN) ..... 45
5.1.5 Feature pyramid network (FPN) ..... 46
5.1.6 Generative Adversarial Networks (GAN) ..... 47
5.2 Transfer learning ..... 47
5.3 Data augmentation ..... 48
6 Open access liver and lesion segmentation datasets ..... 50
6.1 SLIVER07 dataset ..... 50
6.2 3DIRCADb dataset ..... 50
6.3 CHAOS dataset ..... 50
6.4 PAIP2019 dataset ..... 50
6.5 LiTS dataset (Liver Tumor Segmentation) ..... 51
7 Deep learning-based research proposals for liver and lesion segmentation ..... 54
8 Experimental results ..... 60
8.1 Data ..... 60
8.2 Training ..... 64
8.3 Evaluation Metrics ..... 64
8.4 Experiments ..... 66
8.4.1 Liver and liver tumor segmentation ..... 66
8.4.2 Liver segmentation ..... 73

8.4.3 Tumor Segmentation from all CT scans ..... 78
8.4.4 Tumor Segmentation from Liver masked CT scans ..... 80
8.5 Discussion ..... 82
9 Bibliography ..... 85

# List of figures

Figure 2.1: (a) Head CT scan results (b) MRI machine generating a magnetic field. (c) PET scans (d) Ultrasound technology – Source [5]..... 14
Figure 3.1: An image before and after Pre-processing with Hounsfield Unit – Source [21] ..... 19
Figure 3.2: Histogram equalization – Source [21]..... 19
Figure 3.3: Image denoising with median filter – Source [21]..... 20
Figure 3.4: Resolution enhancement with DWT variant – Source [40]. ..... 21
Figure 4.1: Image segmentation examples – Source [44]..... 23
Figure 4.2: Instance segmentation – Source [47]..... 24
Figure 4.3: Panoptic segmentation – Source [47]..... 24
Figure 4.4: Line detection – Source [42]. ..... 25
Figure 4.5: Basic Global Thresholding – Source [42]. ..... 26
Figure 4.6: Cell image segmentation with Otsu thresholding - Source [42]. ..... 26
Figure 4.7: Region growing: Initial X-Ray image, initial seed, segmented image- Source: [42]. 27
Figure 4.8: Region Splitting and Merging, image from X-ray band telescope, of supernova using various quadregions – Source [42]. ..... 28
Figure 4.9: Common overlap-oriented metrics for segmentation – Source [48] ..... 31
Figure 4.10: Dice coefficient (DSC or DICE) – Source [50] ..... 32
Figure 4.11: Jaccard index (JAC) or Intersection over Union (IoU) ..... 32
Figure 4.12: Effect of object size on DSC. Predictions differ only one pixel. If the object is small, this has extreme impact on DSC (the same happens for IoU) – Source [48]..... 34
Figure 5.1: ML concepts & classes – Source [23]..... 38
Figure 5.2: DL evolution - Source [55]. ..... 39
Figure 5.3: ML vs DL - Source [55]..... 40
Figure 5.4: Typical CNN architecture pipeline - Source [22]. ..... 41
Figure 5.5: Convolutional layer operation - Source [22]..... 42
Figure 5.6: Max pooling layer operation - Source [22]. ..... 42
Figure 5.7: Fully connected layer operation - Source [22]..... 43
Figure 5.8: U-Net architecture for image segmentation - Source [30] ..... 44
Figure 5.9: U-Net++ and UNet3+ architectures - Source [21]. ..... 45

Figure 5.10: FCN architecture - Source [21]. ..... 46
Figure 5.11: a: Feature map extractor FPN architecture, b: with ResNet backbone - Source [21]. ..... 46
Figure 5.12: Architecture of a Generative Adversarial Network (GAN) - Source [21]. ..... 47
Figure 7.1: FCN architecture for liver segmentation, Cohen et al. – Source [77] ..... 55
Figure 7.2: Modified multi-scale CNN for liver segmentation – Source [78] ..... 55
Figure 7.3: a: Resnet50 architecture - Source [68]. ..... 57
Figure 7.4: b, c: Resnet50 architecture, transfer Learning - Source [68] ..... 58
Figure 8.1: Folders containing train and test images. ..... 61
Figure 8.2: Folders containing prediction masks. ..... 63
Figure 8.3: Liver and tumor segmentation, Resnet50 train (a) &validation (b) losses and train accuracy (c). ..... 67
Figure 8.4: Confusion matrix, liver detection. ..... 69
Figure 8.5: Liver and tumor segmentations with good performance. ..... 70
Figure 8.6: Some not close to perfect liver segmentations with lower DSC scores. ..... 71
Figure 8.7: DSC liver: 0.975, tumor: 0.33, small tumor ..... 72
Figure 8.8: DSC liver: 0.44, tumor: 0.074. ..... 72
Figure 8.9: Confusion matrix, tumor detection. ..... 73
Figure 8.10: Liver segmentation, Resnet34, train accuracy (a) and train (b) &validation (c) losses. ..... 74
Figure 8.11: Confusion matrix, liver detection ..... 76
Figure 8.12: Liver segmentations ..... 77
Figure 8.13: Tumor segmentations from abdominal CT scans ..... 79
Figure 8.14: Tumor segmentations from liver masked CT scans ..... 82

# List of Tables

Table 1: Open access datasets – Source [21]. ..... 51
Table 2: Statistics of LITS dataset – Source [29]. ..... 52
Table 3: Segmentation Implementations ..... 58
Table 4: Train and test folders ..... 62
Table 5: Output prediction masks folders ..... 63
Table 6: Model and data parameters ..... 66
Table 7: Liver and tumor segmentation results. ..... 68
Table 8: Liver and tumor detection results. ..... 68
Table 9: Model and data parameters ..... 74
Table 10: Liver segmentation results. ..... 75
Table 11: Liver detection results. ..... 75
Table 12: Model and data parameters ..... 78
Table 13: Tumor segmentation from abdominal CT scans ..... 79
Table 14: Tumor detection from abdominal CT scans ..... 80
Table 15: Tumor segmentation from liver masked CT scans ..... 81
Table 16: Tumor detection from liver masked CT scans ..... 81

# Dedication

To my father.

# 1 Introduction

Liver is a very important organ of the human body. Its shape is wedge, has red and brown color and two lobes with different shape and weight. An average liver is about 1,5 kg and is approximately 15 centimeters in width. Its weight varies, with a range of 970 to 1,860 gr for men and for women being 600 to 1,770 gr. It is the biggest in size and the heaviest organ of human body and it is situated in the right upper size of the abdominal area.

The primary liver's job is to purify the blood that exits the digestive tract, which is afterwards supplied to the body. Besides it processes drugs and cleans up chemicals, it prepares proteins necessary for blood clotting [1]. Moreover, it produces bile, an alkaline fluid containing cholesterol and bile acids that aids in the breakdown of fat, making it an accessory digestive organ. Liver is thought to have around 500 total functions, although estimates on this number vary and liver is responsible for 20% of the body's total oxygen consumption.

Primary liver cancer (PLC) is the fourth cause of deaths related to cancer worldwide. According to Global Cancer Statistics Report (Globocan) [2], there were reported 905,677 cases of liver cancer in 2020 worldwide, which is 4.7% of all cases for that year. Liver cancer deaths were 830.180, making up 8.3 percent of all cancer deaths. WHO (World Health Organization) projects that 1,276,679 people will lose their lives to liver cancer in 2040, based on annual projections. The prevalence and mortality of PLC vary greatly by region [1].

East Asia and sub-Saharan Africa had the highest PLC incidence rates, according to research. With a high incidence rate of 18.3 cases per 100,000 people, China in particular has the highest number of PLC cases [3].

According to NIH (National Cancer Institute, USA) [4], for liver cancer diagnosis, the following tests and procedures may be used:

- Physical examination and medical history
- Alpha-fetoprotein (AFP) tumor marker test
- Liver function tests
- CT scans (computed tomography)
- MRI (Magnetic Resonance Imaging)
- Ultrasound examination

The last three medical imaging methods will be further presented in the following chapter.

Liver cancer concerns many people and that makes its diagnosis and treatment a very important issue. A crucial step in that direction is segmentation of liver and its tumors from medical imaging, which are widely used in clinics, in order to visualize liver and tumors with accuracy and high resolution [5]. However, this method is typically manual and requires time and intensive slice-by-slice delineation.

Manual delineation of these tumors takes a lot of time and depends on the operator [6]. Since liver cancer is among the ones that cause the most deaths, and metastases frequently have liver origin, it is essential to have high quality liver and tumor segmentations [7].

Effective liver tumor segmentation from CT images is still difficult due to the wide variety of their complex shapes, sizes, position, hazy borders, and their intensity diversity [8], [1]. Kang et al. [9] also mentions that in general, tumor boundaries with adjacent non-cancerous liver tissues are not clear, which makes segmentation difficult.

Automatic liver tumor segmentation methods may improve liver tumor surgery success rates and lessen the drawbacks of manual segmentation [1], and therefore there is a great need for research the specific scientific area.

In the direction of automatic methods development with no need of human intervention, Machine Learning (ML) is a very effective approach. Advances in ML have finally ushered in the era of artificial intelligence (AI). ML systems have been used since the late 1990s and they have an important feature-learning phase called feature engineering.

According to Moghbel et al. [6], various computer-assisted methods have been utilized for liver and liver tumor segmentation from computed tomography:

**Statistical shape models (SSM)**: Segmentation based on the shape of the object. A large dataset is required to train an ML model (e.g., k-Nearest Neighbor (kNN) classifier) to segment liver area by its shape [6].

**Probabilistic atlas approaches**: It uses not only shape information like SSM method, but also spatial information (e.g., modeling information of neighboring organs) to perform liver area segmentation by ML models (e.g., kNN) [6].

**Geometric deformable model (or active contours)**: This method is based on borders identification by following intensity variations [6], [10].

**Region growing approaches**: Starting from an initial small area (seed) and expanding it using specific conditions, determines the final segmentation area. This method applies for both liver and liver tumor segmentation [6], [10].

**Graph cut approaches**: The image is represented as a graph whose nodes are the pixels and the edges with their weights correspond to the color or gray level or texture similarity of neighboring pixels [6], [10].

**Texture approaches**: These methods perform segmentation aiming to identify texture properties instead of object contours. After texture feature extraction, ML classifiers like k-means, are utilized to determine the belonging class of the area, to perform segmentation [10].

**Gray level approaches**: By analyzing image statistical data (e.g. histograms), the gray scale intensity distribution of liver is estimated which is used for segmentation (k-means) [6], [10].

**Thresholding approaches**

An intensity threshold can be applied on the image to separate foreground and background. This threshold can be static or based on statistical analysis (Otsu method) of the image depending on its complexity [6], [10].

**Machine learning based methods**

Machine Learning (ML) approaches, like various clustering algorithms and SVMs (Support Vector Machines), have been widely used, in conjunction with other methods. By statistically analyzing data derived from other methods, they increase system performance and automate segmentation process. In composite segmentation tasks, different methods may be used in various steps, where an ML algorithm in the final step is producing the output. In such a case, we are talking about ML based approach [6].

Various ML methods have been utilized such as:

- Neural Networks (NN) [11], [12]
- K-means [13]
- K Nearest Neighbor (KNN) [14]
- Support Vector Machines (SVM) [15]
- Extreme Learning Machine (ELM) [16]
- Fuzzy c-means (FCM) [12], [17]
- AdaBoost classifier [18]
- Bayesian classification [19]

- Bayesian Gaussian mixture models [20]

With the advance of AI algorithms, techniques that are more efficient have been developed, such as deep learning (DL) techniques. These algorithms create systems that automatically learn features from data. Many DL model types have been created for different purposes in medical imaging, like segmentation, classification, and detection of various diseases, including liver tumors. DL models consist of many layers and their task is to extract the image features and perform subsequent classification or regression tasks [21]. All these characteristics makes DL a very useful and effective tool in liver and liver tumor segmentation tasks.

The popularity and effectiveness of DL today is primarily due to three key factors [22]:

- the dramatically increase of processing hardware performance (GPUs)
- the significantly decreased cost of computing hardware
- the significant improvements in ML algorithms

DL already outperforms human capabilities for specific applications in closed environments [23].

Many researchers have worked to create an automatic method for the accurate diagnosis liver tumors, and many automatic and semi-automatic techniques have been proposed recently. In that direction, DL applications for segmentation of medical images have gained a lot of attention due to their excellent performance [21].

Various architectures that are commonly used in medical imaging segmentation tasks including liver and liver tumor [21] are:

- Convolutional Neural Networks (CNNs) [45], [46]
    - U-Net and its variants [24]
- Feature Pyramid Pooling (FPN) [25]
- Fully Convolutional NNs (FCNs) [26], [27]
- Generative Adversarial Networks (GANs) [28]

CNNs have the advantage that they consist of multiple layers, which are robustly trained. They have proved to be very effective and that is why they are very popular in computer vision [22] including liver and liver tumor segmentation ([1], [29]). The U-shaped CNN architecture and its variants are very common in biomedical image segmentation [30] due to their excellent results.

Resnet50 [68] – a pre-trained model- is utilized in this work, and it was trained for liver and liver tumor segmentation from abdominal CT scans.

While liver segmentation results of DL approaches are very close to human detection capabilities, this does not occur for liver tumor segmentation results [21]. In general, segmentation of liver tumor had neither the attention nor the performance if compared to the proposed liver segmentation methods. This is likely due to the difficulties associated with segmentation of liver lesions, and the fact that there are not that many of publicly accessible datasets [6].

Some of the reasons making liver tumor segmentation a difficult task are:

- small size lesions are difficult to detect and segment [29]
- not many large publicly accessible labelled datasets for liver tumor segmentation [31]
- not clear borders between liver and lesions in many [6]
- various tumor types of coexistence [31]
- low contrast between liver and tumors [32]

For all these reasons, it is obvious that liver tumor segmentation is not only crucial, but also a very challenging task and there is a lot of room for improvement to achieve better results. The results of this work confirm all previous conclusions and outcomes of many other works, and are presented in the following chapters.

# 2 Medical imaging

Medical imaging is the technology and the procedure used to represent visually internal body parts for examination and intrusion by a doctor. With the development of quick, more accurate, and minimally invasive instruments, medical imaging has undergone a significant revolution in the past years. In the medical field, accuracy is essential in clinical procedures and equipment development. Greater illumination is required for more effective large data analysis and for igniting the hidden medical knowledge contained in big data. Medical procedures benefit from the relationship between data items and hidden data. As the healthcare sector works to achieve early detection of diseases and minimizing cost, medical imaging plays a determinant role towards that direction [33].

## 2.1 Imaging Modalities

The imaging industry has advanced significantly since the end of the 19th century. For clinical use, numerous imaging methods were developed. Imaging modalities is the term used to describe these imaging techniques. Below some of the imaging modalities used in liver cancer detection are briefly presented.

### *2.1.1 Tomography*

A tomographic optical system is used in tomography to virtually slice an object into a specific cross section without physically breaking it. In contemporary tomography, data from various sides are acquired and passed to a computer software to be processed. Different kinds of tomograms are derived from various physical phenomena. For instance, CT use X-rays, PET use electron positron extermination, SPECT use gamma rays and more [33].

Conventional Tomography uses a moving X-Ray source and a film in unison, to produce a tomogram. The different methods used in traditional tomography include, linear, panoramic and poly tomography and zonography.

**Computer Assisted Tomography (CAT)**, also called Computed Tomography scan (**CT scan)**, data are collected by radiation detectors and a computer rebuilds the image computationally from collected data (Figure 2.1 a) [33].

### *2.1.2 Magnetic Resonance Imaging (MRI)*

MRI (Figure 2.1 b) takes advantage of the relaxation characteristics of hydrogen nuclei of water molecule that have been magnetically stimulated. It is also called NMR tomography. Because its radiations are not ionizing, it is a better choice compared to Computed Tomography. Radio frequency energy is applied to the patient, and with the magnetic field, it causes the nuclei to enter an elevated energy state. The normal, microscopic tumbling of the molecules causes them to release their energy due to relaxation. The images are made using the variations in the rates of tissue relaxation. Strong magnetic fields and radiofrequency, non-ionizing radiation are used in MRI [33].

Magnetic field must be both powerful and consistent for MRI. Based on gadolinium chalets, the contrast materials used in MRI scans. MRI is also regarded as a tomographic imaging technique because it produces a body slice 2D image, similar to how Computed Tomography does. The most recent MRI equipment can generate 3D images [33].

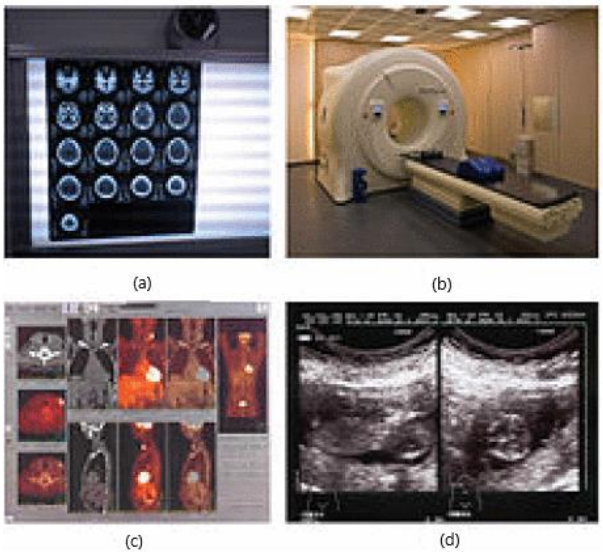



Figure 2.1: (a) Head CT scan results (b) MRI machine generating a magnetic field. (c) PET scans (d) Ultrasound technology – Source [5].

### *2.1.3 Ultrasound*

Medical ultrasound is a diagnostic method to view body organs (Figure 2.1 d). Ultrasound refers to sound waves with frequencies over audio frequencies. By inserting ultrasound pulses into

tissues with the aid of probes, ultrasound images are created. By using various sonographic instruments, different types of imaging can be produced. The most known example is the B-mode image, which shows the 2-D cross sectional acoustic impedance of tissue. The benefits of ultrasound include real-time viewing of the area of interest, portable equipment, a significantly reasonable cost, and avoidance of ionizing radiations. Many people use sonography to carry out both diagnostic and therapeutic procedures.

### *2.1.4 Nuclear Medicine*

Using radioactive substances to diagnose and treat diseases is known as nuclear medicine, a special branch of medicine. Nuclear medicine uses specific isotope characteristics and energetic particles released by radioactive material to diagnose a variety of pathologies. Nuclear medicine makes use of the concept of physiologic assessment as opposed to anatomic radiology. Nuclear medicine exams show the physiological function of the system being studied, unlike the majority of other imaging modalities. In addition to molecular medicine and molecular imaging and therapeutics, nuclear medicine has other names [33].

Unlike conventional radiology imaging where a particular section of the body is focused (for example chest X-Ray, CT scan), nuclear medicine imaging studies are more organ or tissue specific (lung scan, brain scan) in general. Several techniques are available in diagnostic nuclear medicine [33].

Nuclear Medicine techniques include:

- Scintigraphy
- Single Photon Emission Computed Tomography (SPECT)
- Positron Emission Tomography (PET)

Based on specific cellular receptors or functions, some nuclear medicine studies enable imaging of the entire body. Some examples include octreotide scans, PET scans (Figure 2.1 c), and PET/CT scans. Though SPECT and PET can both be used in clinical applications to provide diagnostic information that is similar, these applications differ depending on the radionuclide characteristics that are used. Magnetic resonance imaging (MRI) and positron emission tomography (SPECT) both provide functional information, and fiduciary markers can link that information to anatomical data [33].

## 2.2 Medical image file formats

A standardized method of storing the details describing an image in a computer file is provided by image file formats. Medical image sets usually include [34]:

a) images illustrating the promotion of an organ on an image called projection imaging,
b) an image collection showing slices across a volume called tomographic imaging,
c) a volume data collection called volume or 3D imaging,
d) acquisitions of a volume image over time stamps to create a dynamic acquisition series called 4D imaging.

The file format specifies how the image data should be loaded and displayed correctly by a software program and how software handles pixel data [34].

The main categories of medical image file formats are image standardized formats like DICOM (Digital Imaging and Communications in Medicine) [35] and formats for facilitating and enhancing post-processing analysis. Usually, one of the two following storage configurations is used for medical image files [34]:

- Both metadata and image data are stored in one file, with metadata placed at the file's start. Even though it is permitted by other file formats, DICOM, MINC (Medical Imaging NetCDF) [36], and NIfTI (Neuroimaging Informatics Technology Initiative) [37] file formats use this approach.
- The image data is kept in a separate file from the metadata. In the Analyze file format .img and .hdr filetypes are used.

While many scanners typically produce DICOM format images, many well-known applications used for image manipulation and analysis, handle only images in NIfTI file format, an open format for storing medical imaging data obtained using Magnetic Resonance Imaging techniques. NIfTI, was developed by the National Institutes of Health with the goal of developing a format for neuroimaging that would maintain the benefits of the Analyze format while addressing its drawbacks. Despite the format's ability to store header and data in different files, images are usually stored in one file with the .nii extension that combines both. It has also the capability to store image volume orientation information [34]. NIfTI is a much simpler, more straightforward

format than DICOM. It is used in neuroimaging because it enables researchers to combine and contrast various image processing and analysis software packages [38].

While NIfTI is straightforward and simple to support, DICOM is comprehensive and verbose, making each of these formats well suited for a particular market. DICOM standard is extremely complex, and various scanner manufacturers extend it in a variety of ways, frequently leading to information duplication and incompatibilities between software that is only compatible with a specific DICOM type [38].

# 3 Pre-processing techniques for medical imaging

After acquiring liver CT images, they must be preprocessed. It is necessary to reduce present noise and enhance CTs images borders for efficient liver and tumor segmentation. Various spatial filters are available to improve CT liver images. These are Wiener, bilateral, hybrid filters (the previous two combined), guided, curvature, Kirsch, recursive Gaussian filters, linear interpolation, histogram equalization [1].

Preprocessing is a preparatory process necessary to attain increased performance accuracy in training process of DL models. CT and MRI images can include noise, patient specific or acquiring device related artifacts, and therefore, they must be taken away before segmentation process.

Below some of the most common preprocessing techniques used in medical imaging will be presented [21].

## 3.1 Hounsefield unit

The main formats that apply on medical imaging are the DICOM and NIfTI. In these formats, pixels are represented with their intensity rating, the Hounsfield Unit (HU) that is a number in the range of 1024 to 4000. The eye of a human can only perceive around 100 different variations of gray. Therefore, to illustrate a ROI of a CT scan, pixels must be restricted to a specific region called Hounsfield Unit (HU) windowing. Since CT images contain multiple organs, the HU window areas are different. For liver and its lesions segmentation, HU must be in the range of -200 to 400 [21].

HU windowing affects greatly the segmentation performance, so it can be considered as a hyper-parameter in deep learning models. Patro et al. [39] in order to pre-process CT slices prior to segmentation, experimentally determined that the optimal range of HU for liver segmentation with U-Net, lies in the range of -150 to 250. In Figure 3.1 the effect of processing an image with HU windowing is illustrated [21].

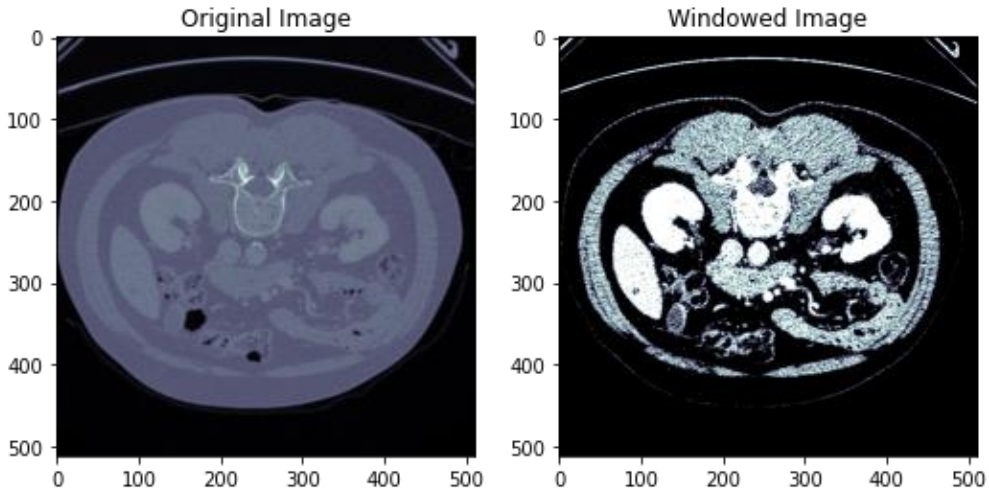



Figure 3.1: An image before and after Pre-processing with Hounsfield Unit – Source [21]

## 3.2 Histogram equalization

Histogram equalization (HE) is a preprocessing technique applied in medical imaging, aiming to spread the pixel intensities of an image in the histogram (compared to initial image), resulting to contrast enhancement. Contrast limited adaptive HE (CLAHE) -a very popular HE method for medical image improvement- is a modified Adaptive HE (AHE) with various improvements. AHE calculates topical histograms with the final image being the bipolar interpolation of them. The drawback of this approach is the amplification of noisy pixels instead of suppressing them. CLAHE on the other side is suitable in enhancing medical images by fastening the histogram to a certain level and avoiding the most common noise in peaks.

The results of HE on CT scans are depicted in Figure 3.2 [21].

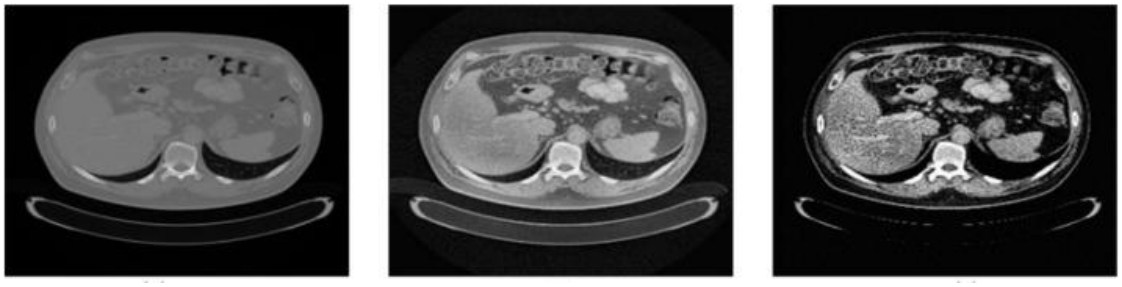


Figure 3.2: Histogram equalization – Source [21]

## 3.3 Image denoising

While acquiring, CT and MR images may capture various undesirable artifacts such as beam hardening, noise, scattering, motion, metallic objects, rings etc. It is important to identify and enhance these findings in CT images so that they will not confuse the detection system. Techniques with multiple filtering like median filtering are very effective in order to reduce noise from images like Gaussian noise, salt & pepper, random noise. Median filtering (Figure 3.3) determines the median value of the data below a certain window size equal to filter's output sample. In case of n-dimension data, windows with length 3, 5, 7 are usually used. A kernel with length 3x3 is proposed for 2D data [21].

Median filters are applied to remove noise, like filling small pits in the liver. This technique outperforms the mean and the Gaussian filters in segmenting liver because it preserves image borders details. Median filter smoothens a CT image and eliminates noise using a 3x3-pixel window size [21].

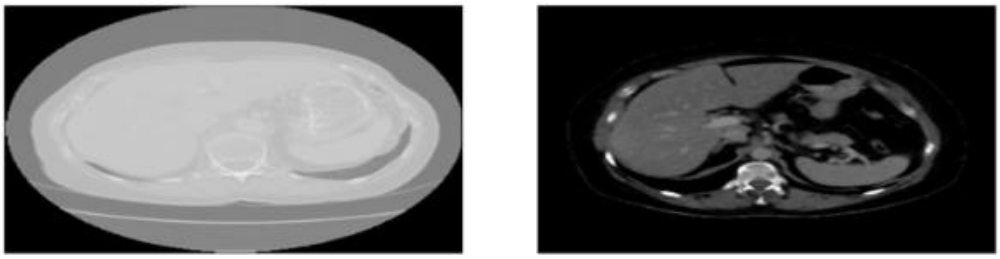


Figure 3.3: Image denoising with median filter – Source [21].

## 3.4 Resolution enhancement

Following noise reduction, edges have lost sharpness. Edge and contour information in the filtered image is preserved through resolution enhancement. Preserving frame and contour information is important for proper image segmentation. Resolution measures image quality after noise removal. Improved Discrete Wavelet Transforms (DWT) can be used to raise resolution, which also preserve edge and contour information [21]. An example of resolution enhancement using DWT is depicted in Figure 3.4.

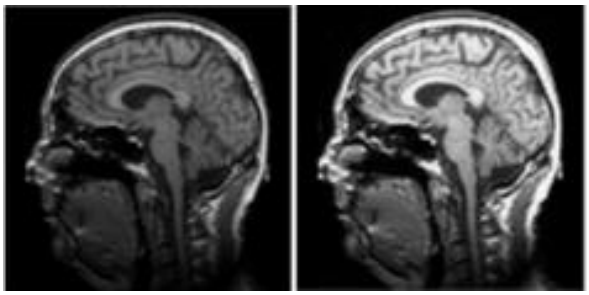


Figure 3.4: Resolution enhancement with DWT variant – Source [40].

## 3.5 Data normalization

A very important task in image preprocessing is normalization. It transforms pixel intensity ranges in a specific scale. The degree of normalization depends on the case and the image type. Prior DL model training process, input data are normalized so that the features have a common scale. This is very significant in speeding up neural network‘s (NN) training process [21].

Normalization is a scaling process, which casts data in a determinate range, usually from 0 to 1, or -1 to +1. It is used in input data with large variances in their values but without many outliers [41].

To calculate the normalized value x’ in case of range [0,1] equation 1 is used.

$$x' = \frac{x-min}{max-min} \quad (1)$$

where x is the initial value, min the minimum value and max the maximum of input data. To normalize data in range [-1, 1] equation 2 is used.

$$x' = 2 * \frac{x-min}{max-min} - 1 \quad (2)$$

Normalization is applied to input data before training process in most medical image segmentation methods.

## 3.6 Data standardization

Another scaling method besides normalization is standardization. It is applied to achieve a uniform pixel intensity value for each specific tissue, slice, patient, MR, body region [21]. While normalization rescales the values into a range of [0,1] or [-1, 1], standardization typically rescales

data to have a mean of 0 and a standard deviation of 1 (unit variance). This method has good results on normal (Gaussian) distribution datasets, but in other cases, it may lead to bad results. Equation 3 presents z-score standardization.

$$x' = \frac{x - mean}{sd} \quad (3)$$

Where *mean* is the mean of values and *sd* represents the standard deviation.

# 4 Image segmentation

Image segmentation is a process, which divides an image into various segments, also referred as image regions, or image objects (pixel sets). Segmentation's target is to make an image's representation more straightforward and/or conducive to analysis [42]. Gonzalez [42] mentions that segmentation's goal is to divide an image into regions. In other words, it aims giving to each image pixel a label so that pixels having the same label value, having similar nature.

Pixels belonging to the same segment share a common characteristic such as intensity, color or texture [43]. Comparable characteristics in adjacent regions have different colors (Figure 4.1).

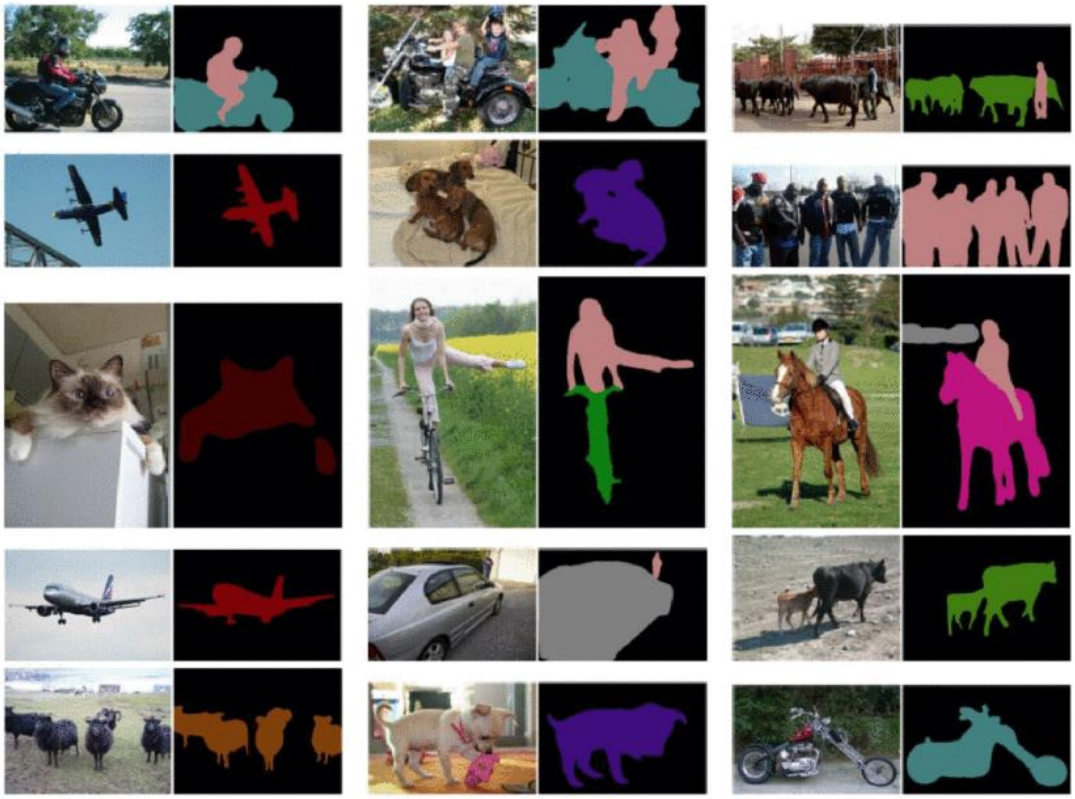


Figure 4.1: Image segmentation examples – Source [44].

Semantic segmentation is a technique that determines the object's class for each individual pixel [45]. For instance, the background is segmented into one object and all the people in a figure into another. Figure 4.1 contains some examples of semantic segmentation.

Instance segmentation (Figure 4.2) identifies an object instance for each pixel. It recognizes every individual object under identified in the image [46] e.g., segmenting each person in a figure into a separate object.



Figure 4.2: Instance segmentation – Source [47].

Panoptic Segmentation (Figure 4.3) combines the two previous approaches. Similar to semantic segmentation, it determines the class to which each pixel belongs. Panoptic segmentation, as opposed to semantic segmentation, is able to separate various instances of the same class [47].

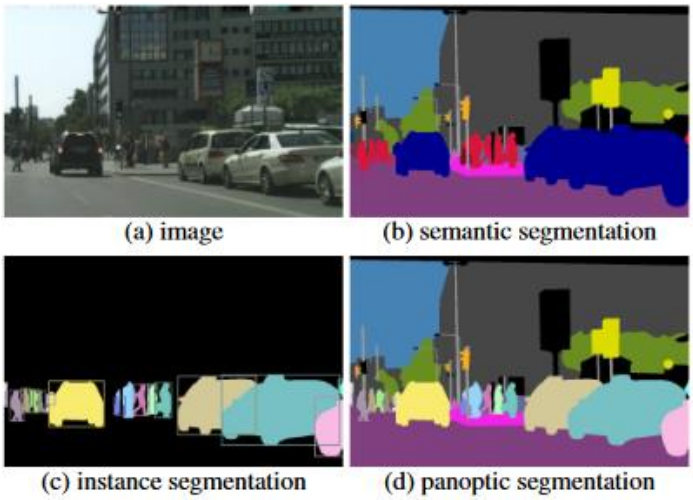



Figure 4.3: Panoptic segmentation – Source [47].

## 4.1 Segmentation methods

Many segmentation algorithms use discontinuity and similarity, which concern intensity. In discontinuity, the strategy is to divide an image on the sharp contrast changes, e.g., edges. In similarity, the main strategy is to divide the image into areas sharing similarities based on predetermined criteria. This category includes techniques like thresholding, region growing, region splitting, and region merging [42]. All following methods in various forms are used to liver and liver tumor segmentation [6], [10].

### *4.1.1 Detection of Point, Edge, Line*

These segmentation methods rely on the detection of sharp, local level intensity variations. The features that are under investigation are isolated lines, edges, points. Edges are found at pixel areas where the intensity changes rapidly. A 'line' could be seen as a line in the image where the pixel intensity of it's both side sub planes differ noticeably from lines' pixels intensities (Figure 4.4) [42].



Figure 4.4: Line detection – Source [42].

### *4.1.2 Thresholding*

**Basic Global Thresholding**

A global intensity threshold is applied on the image if pixel intensities of foreground and background can be sufficiently separated by that threshold value (Figure 4.5) [42].



Figure 4.5: Basic Global Thresholding – Source [42].

**Optimum Global Thresholding** with the use of Otsu's Method

It is possible to think of thresholding as a problem in statistical decision theory.

The goal is to reduce the typical error that occurs when allocating two pixels into two or more categories (or classes). An example of this method is depicted in Figure 4.6.

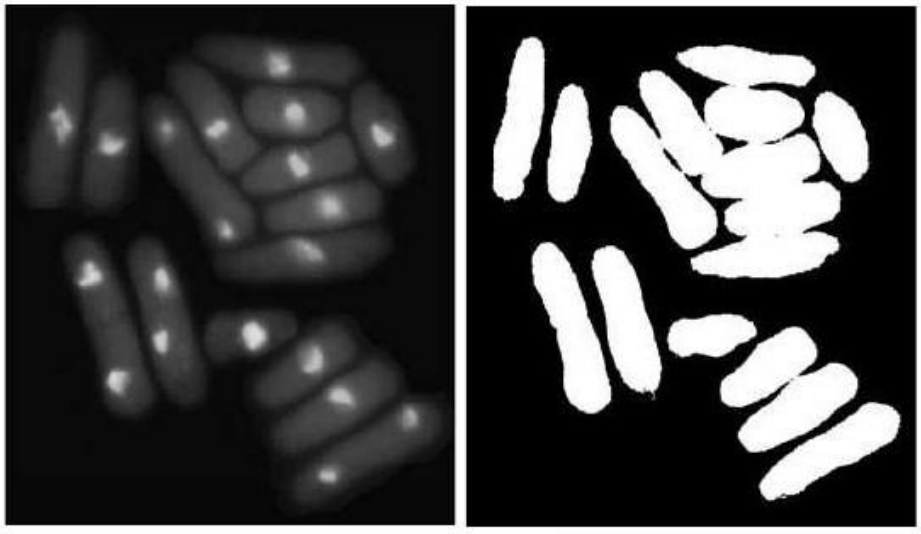


Figure 4.6: Cell image segmentation with Otsu thresholding - Source [42].

### *4.1.3 Region-Based Segmentation*

In this case, the algorithm creates segments by dividing the picture into various components (collection of pixels) with related qualities. This method is based on two fundamental techniques a) Region rowing and b) Region splitting and merging.

- ***Region Growing***

Region growing is a process that assigns pixels or small pixel areas to greater pixel areas using predetermined criteria. The idea initializes the process with some pixels (seeds) and then grow these regions by appending to each initial pixel, neighboring pixels with predefined characteristics similar to initial pixels (like similar intensities, color etc.), as depicted in Figure 4.7 [42].



Figure 4.7: Region growing: Initial X-Ray image, initial seed, segmented image- Source: [42].

- ***Region Splitting and Merging***

To segment an image, these methods use two approaches: region splitting and merging. An image is split into regions with similar properties by repeatedly dividing it into them and merging involves joining neighboring areas sharing similarities to one another (Figure 4.8) [42].

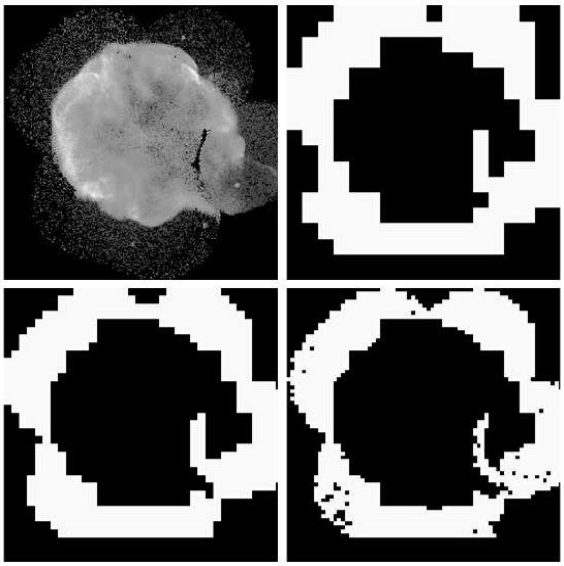


Figure 4.8: Region Splitting and Merging, image from X-ray band telescope, of supernova using various quadregions – Source [42].

- ***Texture***

Texture is a very important feature for in image classification. Statistical features (standard deviation, mean, kurtosis, skewness), secondary statistical features (grayscale co-occurrence matrix (GLCM)), geometric (like area, circularity, diameter, true circularity, perimeter) are very useful for liver segmentation [1].

All these methods, in various forms and variants, including ML implementations, are used in liver and liver tumor segmentation [6]

## 4.2 Segmentation evaluation metrics in medical imaging

In semantic segmentation, one or more labels are applied to each pixel creating this way pixel groups with similar properties, therefore it can also be thought as a pixel-level (or voxel in case of 3D images) classification. In a post-processing step, which is usually performed after classification, relative items are identified as separate objects with clear borders [48].

For assessing the effectiveness of the segmentation applications in medical imaging, various evaluation metrics have been defined. With the aid of these performance evaluation metrics, it is possible to determine the segmentation error.

According to Popovic et al. [49], statistical metrics used for evaluation of medical image segmentation, premise that image elements (voxels) are spatially independent.

Assuming that:

- I (x, y, z) : $R^3 \rightarrow R$ is a three dimensional medical image
- S(I (x, y, z)) : $R^3 \rightarrow \Omega$, $\Omega = \{0, 1\}$, is the binary segmentation mask of the image
- G the ground truth segmentation (mask)
- R the segmentation (mask) result under assessment

It is possible to classify each voxel as:

- true positive: $G(x, y, z) = 1 \wedge R(x, y, z) = 1$
- false positive: $G(x, y, z) = 0 \wedge R(x, y, z) = 1$
- true negative: $G(x, y, z) = 0 \wedge R(x, y, z) = 0$
- false negative: $G(x, y, z) = 1 \wedge R(x, y, z) = 0$

From the above definition, the following statistical parameters concerning the whole image are introduced:

- TP: true positives sum.
- FP: false positives sum.
- TN: true negatives sum.
- FN: false negatives sum.

These statistical parameters are utilized for the calculation of various evaluation metrics for the assessment of 3D medical images segmentation, some of which are commonly used are presented in the following section [50].

### *4.2.1 Spatial overlap-based metrics.*

They are the most frequently used metrics for image segmentation. They are also called overlap metrics because they measure the overlap of ground truth and prediction masks. In the following section, we present some of the most used metrics used for liver and tumor segmentation evaluation.

- **True Positive Rate (TPR)**

TPR (also known as Recall or Sensitivity – equation 4) counts the percentage of positive pixels in target image designated also positive in test image.

$$Recall = Sensitivity = TPR = \frac{TP}{TP+FN} \quad (4)$$

- **True Negative Rate (TNR)**

TNR (also known as Specificity- equation 5) calculates the part of negative pixels, -actually the background of ground truth mask-, which are identified as negative in the prediction mask. Because of their sensitivity to the size of the segment, errors in smaller segments have more impact than errors in larger segments [50].

$$TNR = \frac{TN}{TN+FP} \quad (5)$$

- **False positive rate (FPR)**

FPR (equation 6) counts the percentage of false positive pixels in test image out of the total negative pixels in target image. It represents which portion of the background is wrongly segmented as object:

$$FPR = \frac{FP}{TN+FP} = 1 - TNR \quad (6)$$

- **False negative rate (FNR)**

FNR (equation 7) counts the percentage of false negative pixels in test image out of the total positive pixels in target image. Counter to FPR, FNR represents which portion of the object is not identified and considered as background:

$$FNR = \frac{FN}{TP+FN} = 1 - TPR \quad (7)$$

The similarity of previous metrics leads to the conclusion that only one of each pair metrics should be chosen for validation, e.g., one of FPR, TNR, and one of FNR, TPR.

- **Dice similarity coefficient (DICE – DSC)**

The DSC is similar to F1 Score and positively correlated to Intersection over Union (IoU) [48]. It is the primary metric for assessing the accuracy of medical volume segmentations, frequently used to measure reproducibility or repeatability and directly compares predictions with actual segmentation masks. DSC is the intersection between two sets divided by their sum (Equation 8, Figure 4.9, Figure 4.10). As illustrated in Figure 4.9, its value is between 0 when there is no overlap at all, and 1 in the case of full overlap. Maier-Hein et al. in an analysis of biomedical image challenges [51], concluded that the DSC is by far the most common metric in medical image analysis.

$$DICE = \frac{2TP}{2TP+FP+FN} \tag{8}$$

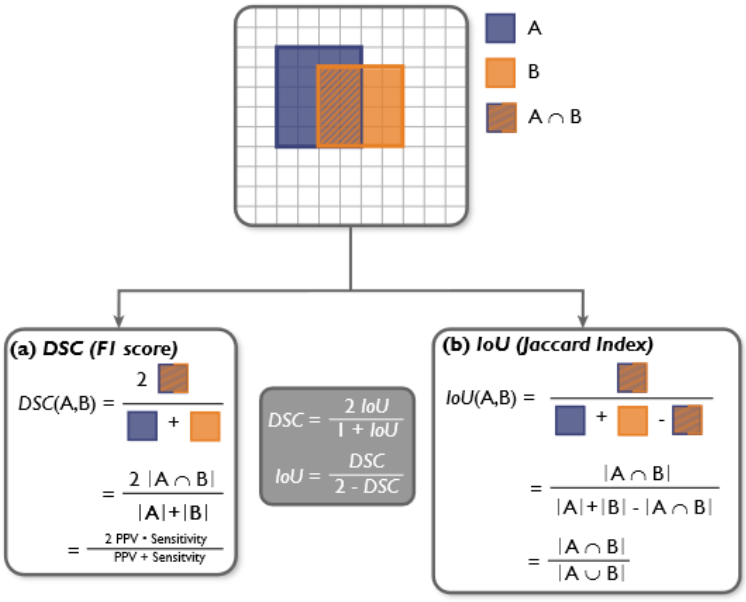



Figure 4.9: Common overlap-oriented metrics for segmentation – Source [48]

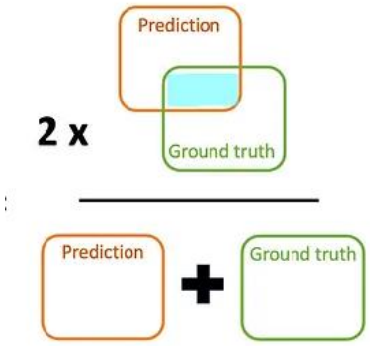



Figure 4.10: Dice coefficient (DSC or DICE) – Source [50]

- **Jaccard index (JAC) or Intersection over Union (IoU)**

Having two sets A, B, their intersection divided by their union gives the JAC metric, also known as Intersection over Union (IoU) (Equation 9, Figure 4.9, Figure 4.11).

$$JAC = \frac{TP}{TP+FP+FN} \tag{9}$$

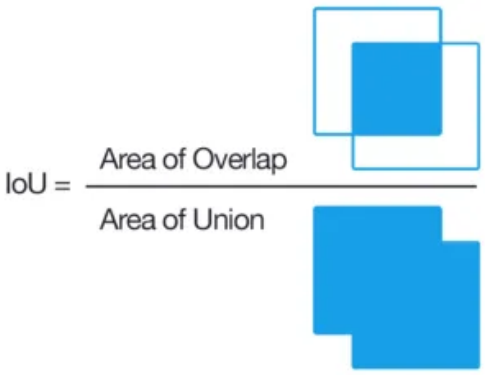



Figure 4.11: Jaccard index (JAC) or Intersection over Union (IoU)

It is obvious that JAC has always a greater value than DSC. They are equal in the extreme case of intersection/union values of 0 or 1. They are also related to each other as in equations 10, 11:

$$JAC = \frac{DICE}{2-DICE} \tag{10}$$

$$DICE = \frac{2JAC}{1+JAC} \quad (11)$$

Both metrics have relevant approach and give similar ranking. Therefore, choosing both of them as validation metrics at the same time, helps for comparison with other experiments using the same metrics and does not add any new information [48].

- **Positive predictive value (PPV)**

Another relative metric is Precision, or positive predictive value (PPV). It is often found in medical images validation or used for F-Measure calculation. It represents the portion of correct positive pixels found out of all positives found (equation 12).

$$PPV = \frac{TP}{TP+FP} \quad (12)$$

### *4.2.2 Volume based metrics.*

Volume based metrics, as their name suggests, are measures that take into account the compared segments volumes to calculate the degree of similarity. Two common metrics in this category are the Volumetric Similarity (VS) and the Volumetric Distance (VD).

Various definitions for VD can be found in the literature, however, Taha et al. [50] suggests the absolute difference of target and prediction volumes divided by their sum (equation 13).

$$VD = \frac{|FN-FP|}{2TP+FP+FN} \quad (13)$$

VS and VD are related as described to equation 14:

$$VS = 1 - VD = 1 - \frac{|FN-FP|}{2TP+FP+FN} \quad (14)$$

### *4.2.3 Semantic Segmentation metrics failures*

Sometimes segmentation metrics do not give an accurate evaluation of results. Reinke et al. [48] exposes some pitfalls of semantic segmentation metrics, with some of them applying also to other areas (Figure 4.12). Segmentation metrics often fail in case where:

- Objects that are small in comparison to pixel size.
- Objects sizes have a high degree of variation.
- Objects come in a variety of shapes.
- Objects volume is particularly important.
- The Objects' center is especially important.

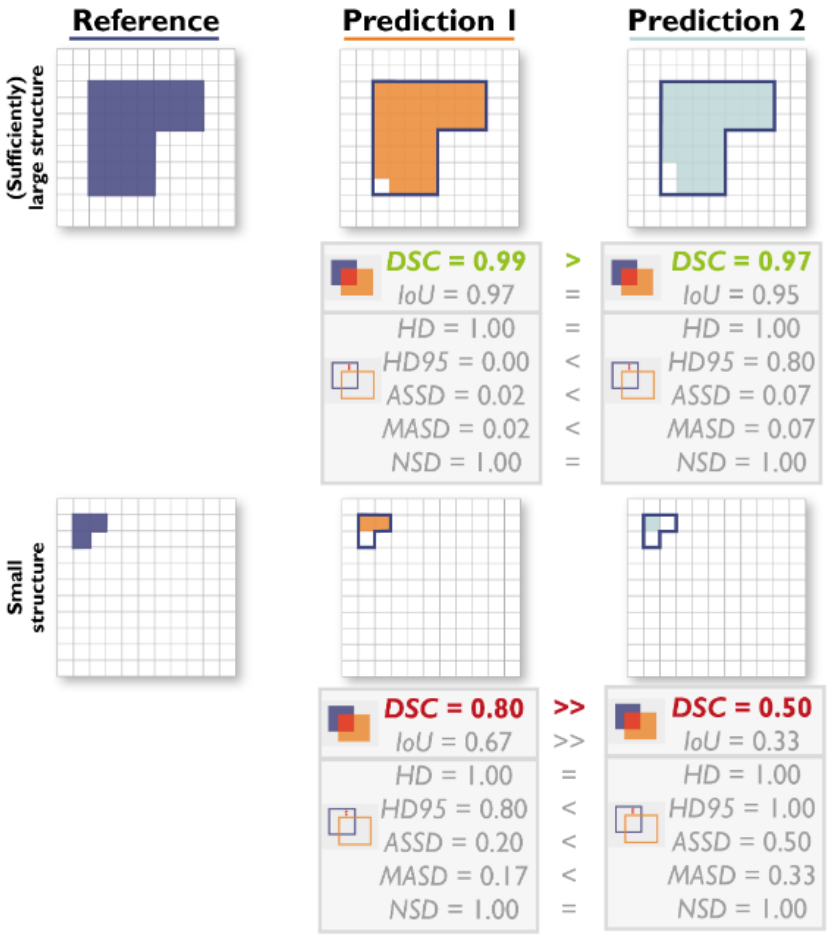



Figure 4.12: Effect of object size on DSC. Predictions differ only one pixel. If the object is small, this has extreme impact on DSC (the same happens for IoU) – Source [48]

- Object boundaries are particularly important.
- Multiple labels per unit might be possible.
- High interrater variability.
- Existence of outliers in the reference annotation is possible.
- Without knowing the target structure, it is possible to reference or predict preferring over vs sub segmentation.

Additional hazards may also be caused by the selection of the threshold the image resolution, the influence, and the confusion matrix, the segmentation versus the accuracy of detection in tasks involving instance segmentation [48].

## 4.3 Liver and liver tumor segmentation

Manual identification of lesions in three-dimensional CTs requires too much time, is difficult to reproduce, and the segmentation results depend on the operator [31]. Unfortunately, segmentation of liver tumor did not have the attention compared to the many proposed liver segmentation methods. This is likely due to the difficulties associated with segmentation of liver tumors and, the fact that there are not that many of publicly accessible datasets [6].

In the following sections, some of the reasons making liver tumor segmentation a difficult task, even for Deep Learning (DL) approaches, are presented.

### *4.3.1 Lesion size effect on segmentation performance*

While many segmentation efforts work fine with sizeable liver lesions, yet they have difficulty in segmenting small size tumors, many of which are only a few voxels in diameter. Furthermore, on large scan resolutions (512 x 512) for each slice, detection of tiny objects is therefore difficult because of the low number of pixels being around that could point out possible lesion borders [29].

This problem is increased by significant noise and artifacts in medical images caused by size similarities. Differences in texture and random shape from the surrounding liver tissue make it hard to differentiate them from real liver tumors. Generally, models with good performance have good operation for large tumor sizes but poorly for small ones. The worst results are obtained with the assessment in a scan with one small tumor (10mm3). Good segmentation results are achieved

with a rather small number of tumors (up to 6) with their total volume being large enough (greater than 40 $mm^3$) [29].

### *4.3.2 Image contrast effect on segmentation performance*

An also important factor which influences segmentation performance is the difference between tumor and liver HU values. Best performing methods work well in volumes with high contrast among liver - lesion. This is more noticeable in tumors that are 40 to 60 HU denser than liver. On the other hand, when the contrast is less than 20 HU, the predictions were poor, which applies also for tumors that are less dense than liver. The mean difference in HU values allows in general to apply simple threshold-derived rules as part of the decision-making process, thus making it easier for the model to distinguish between liver and tumor. Instead, even larger difference values do not give better segmentation [29].

### *4.3.3 Other challenges*

The process creating the labels in datasets by experts increases the possibility of human error and introducing label bias, especially in the case of small lesions whose contours are indeed ambiguous [29].

Data origin, which may come from various sources and locations, information about imaging process, for example scanner type, and demographics, usually are not available in datasets, although they are very important for a deep analysis and extraction of useful conclusions [29].

In general, liver tumor segmentation is a difficult task because of:

- small size tumors, many of which are only a few pixels in diameter and are very difficult to segment. The worst results are obtained with scans with one small tumor (<10mm3) [29]
- not many publicly accessible datasets with large number of scans are available, which are necessary for DL approaches [6]. Since manual identification of lesions in three dimensional CTs (for label creating) requires too much time, is difficult to reproduce and depend on the expert performing it [31]

- edges of many lesions are not clear, which limits the methods which can effectively be used (edge-based) [6]
- a lot of approaches struggle with differences in lesion size, shape and number [31]
- various tumor types coexistence (benign or malignant and various subtypes) [31]
- unconformity in lesion shape, size and location from patient to patient [32]
- low intensity, low contrast, and patient-to-patient anatomic variability of the liver and tumors [32]

Many studies, aimed at determining the optimal method for liver segmentation, are based on datasets either private or custom. Datasets may have biases (for example, people with a specific disease), or contain only a few hundred patients at best. Moreover, it is difficult to compare the various efforts found in the literature because of different metrics. Study variability due to differences in custom datasets and performance measures increases the difficulty of comparing various techniques by the results they declare and reinforces the importance of creating benchmark datasets [32]. Reinke et al. [48] also shows that over segmentation gives better Dice scores compared to under segmentation. Therefore, segmentation methods with focus on achieving maximum DSC of liver tumors have the tension to cause over-segmentation, and the possibility of diagnosing healthy tissue as malicious is increased [32].

# 5 Deep Learning

Machine learning (ML), as opposed to codifying knowledge into computers, aims to automatically discover significant relations and norms from examples [52].

With the recent development of intelligent applications with capacities close to human, which has entered our lives, advances in ML have made it possible for businesses to take advantages of it in many ways like productivity and decision making [53], assistant systems which are able to learn and adapt to users' behaviors, trading stock market applications etc. [23].

The field ML has improved a lot in the field of complex algorithms and effective preprocessing tasks in the past years. A great development was the artificial neural networks (ANNs) into deeper and deeper architectures with higher capabilities to learn and generalize, which gradually led to deep learning (DL) [54].

The relationship of terms is shown in Figure 5.1.

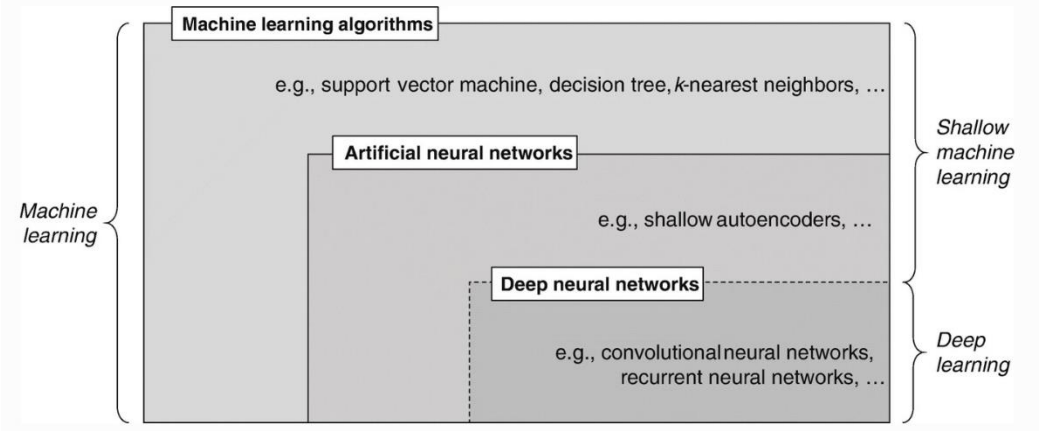



Figure 5.1: ML concepts & classes – Source [23].

The idea behind ML is that a program achieves better results over time in a specific task set by measuring its performance and repeat the process with the target of improving the results. Its target is to automate the development of a model for making cognitive functions e.g., detection of objects in an image or natural language processing (NLP). This is accomplished by using algorithms which repeat in cycles learn directly from the data. This enables the ML applications

to discover complex patterns and hidden insights without the need for being 'programmed' with the classic meaning of the term [52].

Decision trees, Support vector Machine (SVM), Regression models, instance-based algorithms, Naive Bayes, Bayesian methods, ANNs and Support Vector Machine (SVM), are some classes of algorithms that the field of ML offers and each one has advantages in specific tasks. All these models also have multiple specifications and variants [23].

From the other hand, DL models, that are deep ANNs implementations, are becoming increasingly popular because of the hardware and computing increasing capabilities, which allowed the use of DL. Deep neural networks (DNNs) usually have many hidden layers creating complex network architectures. In contrast to simple ANNs, they frequently contain more sophisticated neurons. As a result, they might employ cutting-edge techniques or multiple activations in one neuron. Due to these characteristics, DNNs can be fed with raw data without been preprocessed and automatically find the necessary representations for the specific learning process. Each layer can incrementally perform feature extraction from the data and afterwards it gives them as input to the next layer. The first layer performs the lower-level feature extraction, and the following layers combine them to create a complete representation. Figure 5.2 gives an overview of DL model development.

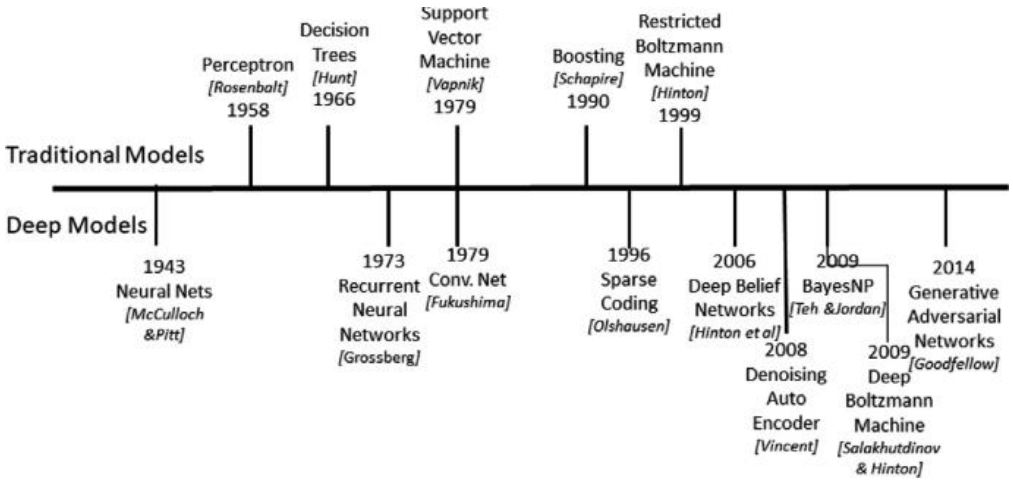



Figure 5.2: DL evolution - Source [55].

While the output of ML models stabilizes once the training data threshold is reached, DL results are improving as the amount of input data increases (Figure 5.3). Nowadays, DL is used in

many fields like Google speech and image recognition, Netflix, Apple's Siri, automated email, text replying engines, chatbots, etc. [55].

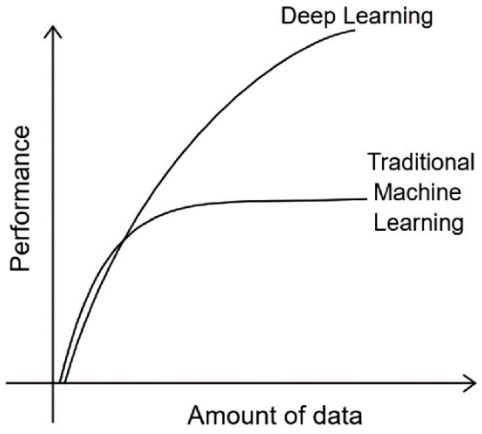



Figure 5.3: ML vs DL - Source [55].

Although DL already outperforms human capabilities for specific applications in closed environments, there are still problems to solve until analytical models are ready for successful application in real business environments, like the appropriate choice of implementation options, bias in the data, hidden properties minimization, usage of preconfigured models (as a service) etc. [23].

The popularity of DL today is primarily due to three key factors: the dramatically improved chip processing capabilities (GPU units), the significantly decreased cost of computing hardware, and the significant improvements in ML algorithms [22].

## 5.1 Deep learning architectures in liver and liver lesion segmentation

In this section, modern DL architectures widely used liver and lesion segmentation liver and liver tumors from three-dimensional volumetric imaging are presented. CNNs, U-Net and its variants, Feature Pyramid Pooling (FPN) architectures, variants of all previous models with pre-trained encoders, Fully Convolutional NNs (FCNs) and Generative Adversarial Networks (GANs) are commonly used for medical imaging segmentation [21].

### *5.1.1 Convolutional Neural Networks (CNNs)*

A very significant DL category is the Convolutional Neural Network (CNN). An advantage of a CNN is that multiple layers are robustly trained. CNNs have proved to be very effective and that is why they are very popular in computer vision [22].

In Figure 5.4 we can see a typical CNN architecture.

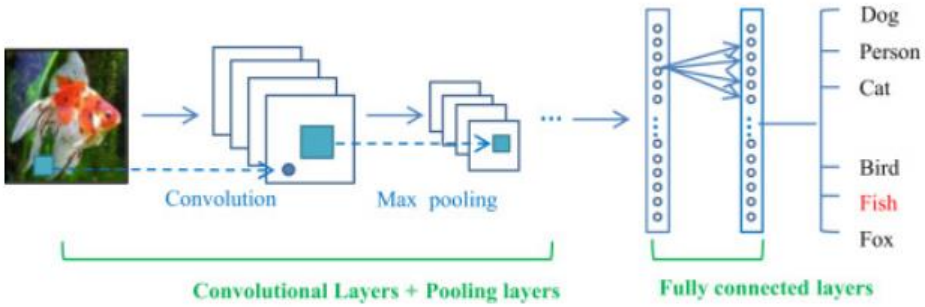



Figure 5.4: Typical CNN architecture pipeline - Source [22].

A CNN consists of convolutional, pooling and fully connected layers and each layer accomplishes a different function. The training process of a CNN includes forward and backward phases. In the forward phase, the input is processed with the initial values of weights and bias of all layers. The loss value is then calculated by comparing the output with target value. In the backward stage, chain rules are applied to calculate the gradient for weights and bias based the loss function results from the previous forward stage. These gradients are used to recalculate the values of weights and bias so they will be used for the next training cycle. The learning process can be stopped when there have been enough forward, and backward iterations and loss is has reached to a minimum value [22].

- ***Convolutional layers***

The role of the convolutional layers in a CNN is to convolve the entire image under process, by utilizing various kernels and the intermediate feature maps and its output is feature maps (Figure 5.5)

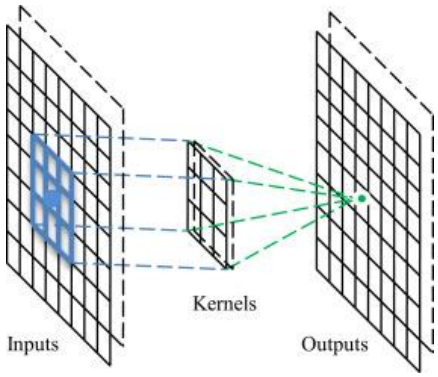



Figure 5.5: Convolutional layer operation - Source [22].

Convolution mechanism has three advantages [56]:

- local connectivity takes advantage of correlations between neighboring pixels.
- object location invariance.
- reduction of the number of parameters caused by the weight sharing operation used in one feature map.

- ***Pooling layers***

The next layer following convolutional is pooling. It can decrease the feature maps dimensionality as also the network parameters. Alike convolutional, the pooling layer is translation invariable since it considers the neighboring pixels in computations. Usual strategies used are average and max pooling. Figure 5.6 illustrates max pooling example. For feature maps with 8x8 dimension, it outputs reduced 4x4 output maps, with a 2x2 size and max-pooling operator [22].

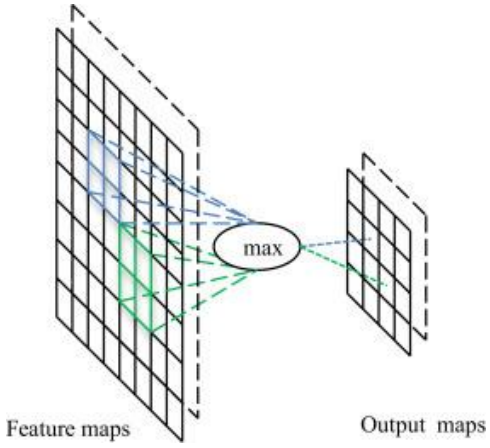



Figure 5.6: Max pooling layer operation - Source [22].

Scherer et al. [57] conducted a comparison for max & average pooling. He concluded that max pooling could speed up convergence, choose better invariable features, and boost generalization. Numerous of quick CNN variants applications using Graphical Processor Units have been presented, with the majority using of the max-pooling strategy [22].

Among the three layers, the pooling layers have received the most research attention. Regarding the pooling layers, there are three popular strategies with distinct goals.

- ***Fully connected layers***

In order to transform feature maps from 2D to 1D vector, various fully connected layers are used as depicted in Figure 5.7

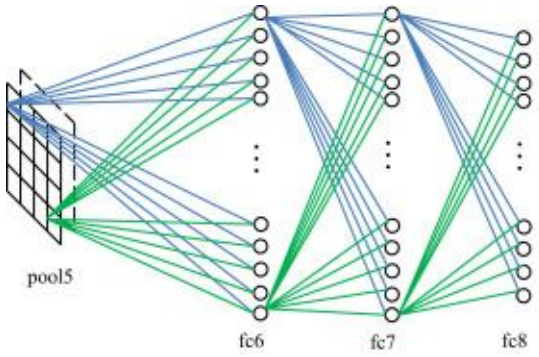



Figure 5.7: Fully connected layer operation - Source [22].

Fully connected layers hold about 90% of CNN properties and function as a classic NN. They enable feed forward of the NN into a vector of predetermined length. One option is to feed forward the vector with a specific number of categories for image categorization or use handle it as feature vector for more processing [22].

The disadvantage of these layers is that they have many parameters, requiring a lot of computational work to train them. For this reason, taking out these layers or connections using a particular strategy is a promising and frequently used. For instance, Google Net [58] switched from fully connected to sparsely connected architectures to design a deep network but maintaining constant computational requirements [22].

### *5.1.2 U-Net Architecture*

U-net [30] was first developed and used in biomedical imaging segmentation. Derived from CNNs, is composed of 2 symmetrical parts (paths), the reduction or left pass or encoder and the

expansion pass (right), also known as decoder. Encoder side encodes the context details while decoder utilizes transposed convolution to encode the regional details. The whole architecture is depicted in Figure 5.8 [21].

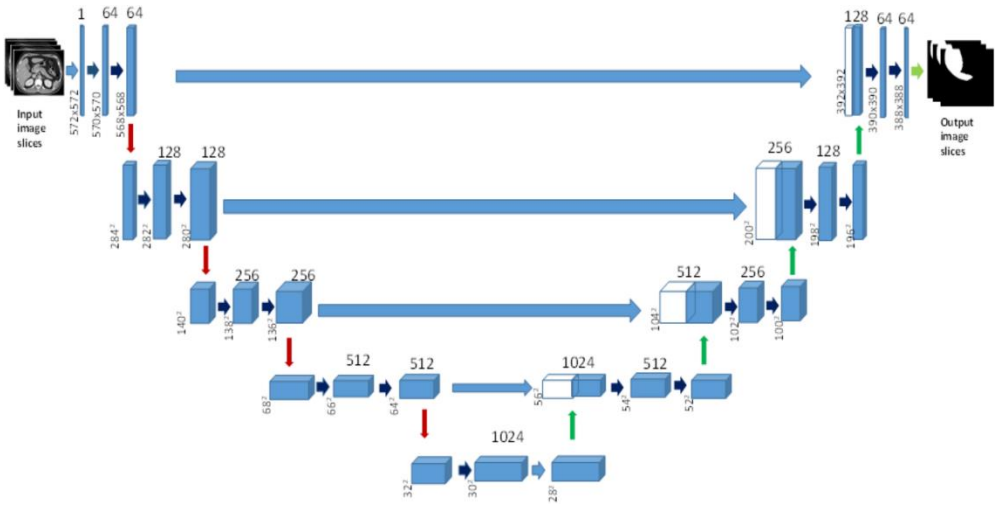



Figure 5.8: U-Net architecture for image segmentation - Source [30]

Initially the image is fed in left pass where it is compressed as it passes. The U-Net architecture has 9 layers as shown in Figure 5.8. Image compression is performed in the first five layers of encoder pass. Encoder's path primary task is to extract as many features as possible from the image. Feature maps are doubled while image passes the compression pass. This compressed image passes from the bottleneck layer where it is up-sampled and given to the right path, the decoder. At the right path, the image is restored in the initial dimensionality as it goes through a number of expansion blocks.

Furthermore, the feature maps are divided as the image is decompressed in decoder pass. Finally, the decoder's final layer output pass is smoothed using a $1 \times 1$ convolution layer and passed to the activation function for the final output, which is afterwards transformed to the initial image dimensionality. U-Net is widely used for segmenting liver and lesions in three-dimensional imaging, because of its excellent performance [21].

### *5.1.3 U-Net variants*

A U-Net variant, U-Net++ [59], contains nested dense-hop connections. U-Net++ concept has to do with the progressive enhancement of the high-resolution feature map from the encoder

pass, with the relative semantically dense feature map by the decoder pass prior synthesis, the network will be more fine-tuned for objects in the foreground, images details are captured more effectively [21].

UNet 3+ was proposed by Huang et al. [60] and considers the effect of full-scale skip links and deep supervisions. Deep supervision learns the structured representations out of the full-scale aggregated feature maps, whilst the full-scale skip links include low-level details and high-level semantics from feature maps with different sizes. Figure 5.9 shows the U-Net++ and U-Net3+ architectures [21].

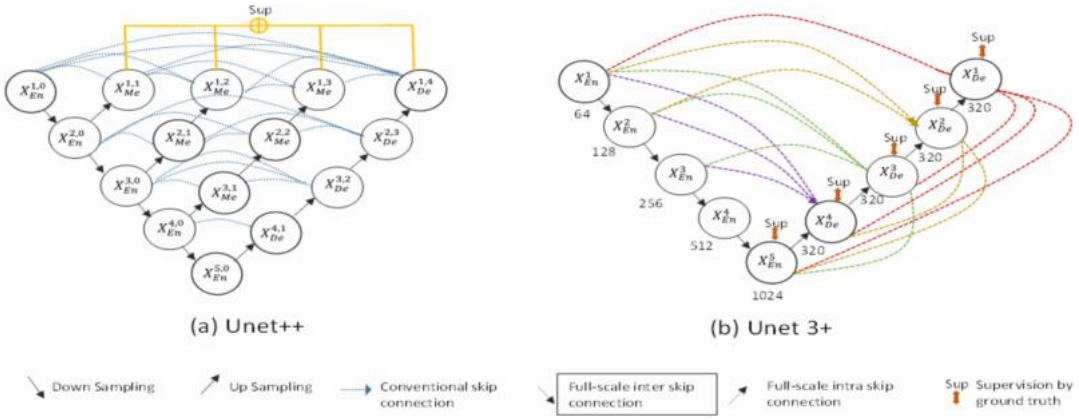



Figure 5.9: U-Net++ and UNet3+ architectures - Source [21].

### *5.1.4 Fully convolution NNs (FCN)*

A full convolutional (FCN) deep learning network was used by Long et al. [61] for semantic segmentation. Unlike traditional convolutional networks (CNNs), FCNs use convolutional layers as an alternative to dense fully connected layers for up-sampling. FCN architecture is shown in Figure 5.10. FCN input may have any dimension while output is resampled and dimensioned spatially [21].

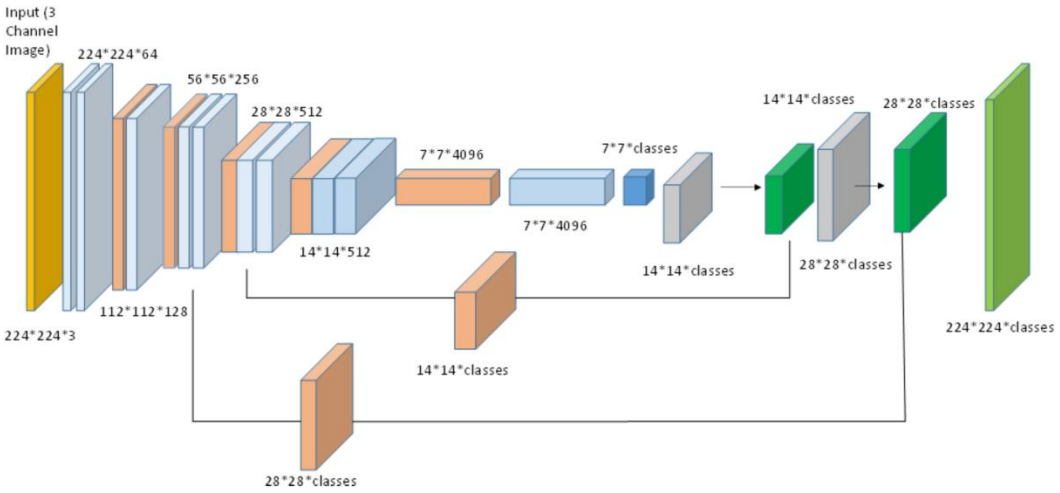



Figure 5.10: FCN architecture - Source [21].

Instead of computing individual patches separately, feedforward computation and backpropagation work faster, allowing the entire image to be computed layer by layer. FCN is a very frequently used DL architecture regarding liver and lesion segmentation from medical imaging [21].

### *5.1.5 Feature pyramid network (FPN)*

FPN is designed for accuracy and speed by building pyramids of images at various scales and detecting objects within them. It replaces feature extractors in detectors such as Faster R-CNN and accord many (multiscale) feature map layers containing elevated quality information than regular feature pyramids. Figure 5.11 shows FPNs architecture [21].

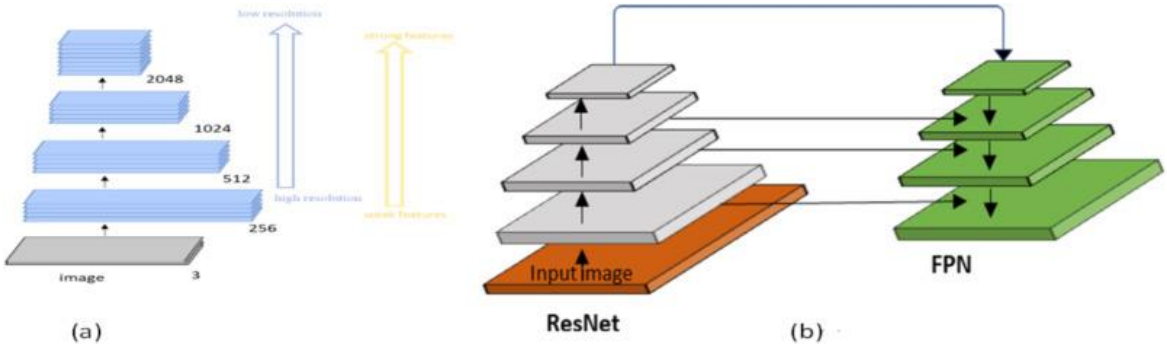



Figure 5.11: a: Feature map extractor FPN architecture, b: with ResNet backbone - Source [21].

FPN contains a bottom-up pass and a top-down pass. For feature extraction, the bottom-up pass has the role of the default convolutional NN. Spatial resolution decreases by going up. The semantic meaning of every layer is higher while additional superstructures are identified [21].

### *5.1.6 Generative Adversarial Networks (GAN)*

Goodfellow et al. [62] launched Generative Adversarial Networks (GANs) in 2014. A GAN contains two NNs, the generator and discriminator. The first is used to create new data, and the second assess these new data based on its similarity to the training data.

Generator learns from the data in training set and creates fake data, which are apparently real. Next, discriminator's job is to distinguish fake data from real dataset data. Afterwards, discriminator gives its conclusion to generator, to improve generated data to be more real like. During training process, both performances of generator and discriminator are improving until they reach a balance [63].

For example, if there is an image-related problem, the generator initially outputs an image using random noise. Discriminator compares the generator image with the real dataset image and output the probability for each one to be real or generated, feeding it back to generator. Figure 5.12 shows a GAN layout [21].

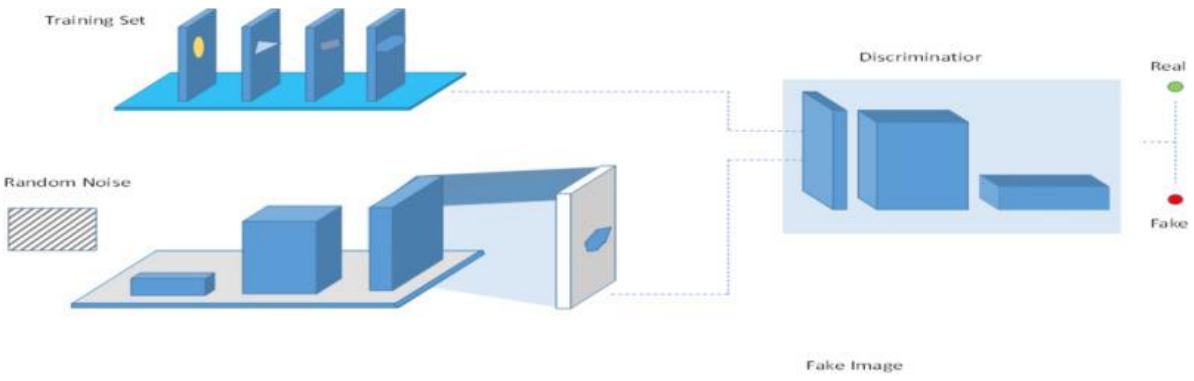



Figure 5.12: Architecture of a Generative Adversarial Network (GAN) - Source [21].

## 5.2 Transfer learning

Transfer learning is utilized to address the problem of insufficient training data. With transfer learning, a model is already trained in a grade in a relative domain and the training process on the target domain may be performed with much less data and much faster, as it does not start from zero [64].

Transfer learning is very popular in DL for cases with insufficient data. The primary purpose of transfer learning is to utilize the experience of an already trained model by another to save time and input data. Transfer learning consists of two primary categories, one that uses the same pre-trained NN for feature extraction and second utilizing a pre-trained model to fine-tune the target model [21].

The fine-tuning phase is essential for improving models so they can be tailored to particular tasks and datasets. These datasets that are used to calculate the losses, fine-tuning typically requires class labels. In this scenario, the new model takes the initial parameter values (weights, bias etc.) from a pre-trained model, except for the final output layer -whose outputs have the same number as the number of class labels in the new dataset- that will take random parameter values. Fine tuning increases the ability of the network to generalize [22], and has the advantages of no need to train the model from scratch, and passing current features to the other model [21].

Medical image segmentation is a great challenge in imaging. To train a DL model a task can last weeks. Therefore, the use of pre-trained models, which is used as a DL network initializer, would solve many problems. Commonly used models for transfer learning are constant or repeated architectures (VGG variant) [65], initial modules (GoogleNet, Inception, etc.) [66], residual modules (ResNet34, Resnet50, resnet101) [67], [68], and dense blocks (DenseNet variants) [69], with novelty structure. Such already trained CNN structures are often utilized to perform as encoders in feature extraction blocks in imaging segmentation architectures, leveraging novelty CNN models, instead of training the model from the beginning [21].

## 5.3 Data augmentation

If a visual object is processed by a CNN for recognition, often data augmentation is used to create extra data with no need for additional labelling work.

Data augmentation may involve creating translations, reflections, or changing the RGB intensities of inputs. By adding pixel regions and manipulating color, Howard [70] used AlexNet as a starting point and after adding changes that enhanced the position and color invariances. Another method is an unsupervised feature learning using data augmentation: based on random selection of small image areas and declaring them as a substitute class, then extending these areas by translating, scaling, adding color, and contrast, and training the CNN to distinguish these

substitute classes. Network's learned features performed well on a number of classification tasks. In addition to the traditional techniques like rotating, scaling and cropping, the use of lens distortion, color casting, and vignetting techniques led to the production of more training examples with comprehensive coverage [22].

# 6 Open access liver and lesion segmentation datasets

Challenges on liver and lesion segmentation gave access to new data so they were available when challenges ended and can be downloaded. A quick review of these datasets and their properties is provided below, while Table 1 summarizes the properties of the datasets.

## 6.1 SLIVER07 dataset

Sliver07 [71] is a MICCAI 2007 challenge dataset for liver segmentation. It contains 40 CT scans out of which 20 are intended for training purposes, 10 for first testing and another 10 for online competition. It contains target segmentation masks for the training data, which where manually annotated by experts and every CT scan has 64-502 slices, with 512×512 resolution and distance among them varying from 0.7 to 5 mm [21].

## 6.2 3DIRCADb dataset

In 2016 the 3DIRCADb2 [72] dataset also became open to access. It was further subdivided into 3DIRCADb-01 and 3DIRCADb-02. 3DIRCABb-01 contains of 20 DICOM images while 3DIRCADb-02 contains two 3D CT scans. The images in both datasets are anonymized. All data, images, labels, masks are available and downloadable from 3DIRCADb website [21].

## 6.3 CHAOS dataset

CHAOS [73] (Combined (CT-MRI) Healthy abdominal organs segmentation) was another challenge which offers a dataset for liver segmentation. It contains 3D abdominal area images, specifically, 40 MRI (20 training and 20 testing) and 40 CT (20 training and 20 testing) scans. CHAOS includes abdominal scans from sane people and the primary target of CHAOS challenge was liver segmenting [21].

## 6.4 PAIP2019 dataset

For the PAIP2019 challenge [74], the dataset was prepared for a period of 3 years (2018 to 2021) and contained more than 3000 images with 6 types of tumors. The data came from Seoul University Bundang and Seoul National University Hospitals, and SMG-SNU Boramae Medical center.

Table 1: Open access datasets – Source [21].

| Dataset | Year | Modality | Image format | Type of tumor | Total Images | No. of slices | Image labels |
|---|---|---|---|---|---|---|---|
| SLIVER07 | 2007 | CT | Header file (.mhd) and image file (.raw) | All cases (HCC, secondary tumor) | 40 (20 training, 10 first test, 10 for online contest | 64-502 per CT | V |
| 3DIRCADb-01 | 2016 | CT | DICOM | 75% cases of hepatic tumor | 20 (15 train, 5 test) | 128 per CT | V |
| LITS | 2017 | Contrast enhanced CT | DICOM | Primary, secondary, metastases derived from colorectal, lung | 201 (131 training, 70 for challenge testing) | 42-1026 per CT | V |
| CHAOS | 2019 | CT, MRI | DICOM | Healthy organs (no tumors) | 80 (20 train and 20 test MRI, 20 train and 20 test CT) | 77-105 per CT, 1367 train, 1408 test | V |
| PAIP2019 | 2019 | DP (digital pathology) slides | SVS format | Hepatocellular carcinoma | 50 WSI (whole slide image) for train, 10 for validation, 40 for test | 3000 | V |

## 6.5 LiTS dataset (Liver Tumor Segmentation)

LiTS [29], which is the dataset used in this work, is also an available public dataset for liver and lesion segmentation. The LITS Benchmark took place during the 2017 IEEE International Symposium on Biomedical Imaging (ISBI) and the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) in 2017 and 2018 [29].

LITS benchmark dataset consists of 201 contrast enhanced abdomen computed tomography volumes following DICOM standard, of which 131 scans are labeled (have liver and liver tumor masks) and proposed for training, and 70 CT scans without labels proposed for testing in order to rank each participation effort for the benchmark. From them, 194 scans contain 0 until 75 liver tumors, which can be primary, secondary or metastases. The process for creating liver and lesion masks (labels) was manual by experienced radiologists and oncologists with value 1 indicating liver and two lesions [21].

LITS image sizes are 512x512, resolution between 0.98x0.98 mm2 to 0.45x0.45 $mm^2$. Each CT scan has 75 until 987 slices, each having depth 5mm to 0.45mm [24].

All data are anonymized, and images are visually checked to rule out the presence of personally identifiable information. The only processing applied to the image is a consistent conversion to his NIfTI format using Python's NiBabel [29].

LITS dataset CT scans contain images depicting both primary liver cancer types (cholangiocarcinoma, hepatocellular carcinoma) and secondary cancer types (colorectal origin metastases, lung and breast primary tumors).

Tumors compared to background, presented either higher or lower intensities. The dataset contains abdominal CT scans of patients before and after therapy, which were acquired with various models of CT scanners, acquisition protocols, and included various imaging artifacts (for example metal object artifacts) often found in medical images [29].

For this reason, resolution and image quality was very diverse. The in-plane resolution of the images had a range of 0.56mm until 1.0mm, and 0.45mm until 6.0mm concerning slice thickness and the axial slices number was in the range of 42 to 1026. Tumors number varies per image, from 0 to 12. Tumors sizes diversion is from $38mm^3$ to $1231mm^3$. Testing images contained more tumors than train images [29].

Table 2: Statistics of LITS dataset – Source [29].

| | **Train** | **Test** | **p-value** |
|---|---|---|---|
| **In-plane resolution (mm)** | 0.76 (0.7, 0.85) | 0.74 (0.69, 0.8) | 0.042 |
| **Slice thickness (mm)** | 1.0 (0.8, 1.5) | 1.5 (0.8, 4.0) | 0.004 |
| **Volume size** | 512 × 512 × 432 (512 × 512 × 190, 512 × 512 × 684) | 512 × 512 × 270 (512 × 512 × 125, 512 × 512 × 622) | 0.058 |
| **Number of tumors** | 3 (1, 9) | 5 (2, 12) | 0.016 |
| **Tumor volume ($mm^3$)** | $16.11 \times 10^3$ ($3.40 \times 10^3$, $107.77 \times 10^3$) | $34.78 \times 10^3$ $10.41 \times 10^3$, $98.90 \times 10^3$) | 0.039 |
| **Liver volume ($mm^3$)** | $1586.48 \times 10^3 \pm 447.14 \times 10^3$ ($1337.75 \times 10^3$, $1832.46 \times 10^3$) | $1622.64 \times 10^3 \pm 546.48 \times 10^3$ ($1315.94 \times 10^3$, $1977.27 \times 10^3$) | 0.60 |

The p-value of statistical test is 0.6, which means training, and test volume images do not have great differences. Concerning tumor HU value, the average is 65 for the train and 59 for the test sets. The statistics of LiTS dataset can be found in Table 2. The split ratio of training and test data sets is 2:1 and both sets are similar, with a center distribution [29].

Any image part, which was not assigned to a label, was assumed background. Another 3 experts further verified these masks who had not seen the primary segmentation, and the more experienced expert, in case of labeling conflicts had the final decision. Scans with tiny and uncertain structures that may looked like lesions were omitted and not annotated [29].

# 7 Deep learning-based research proposals for liver and lesion segmentation

Many studies concerned liver and lesion segmentation, but some focused on liver segmentation only. DL liver segmentation results are very close to human detection capabilities as we see from the results of various models. In general, different datasets, metrics and metrics calculation make it rather difficult to compare the results of various models. For this reason, the safer way to compare is from the results of benchmarks and challenges where the dataset, metrics and metrics calculation are common for all approaches. Publications concerning liver segmentation techniques may be divided into the flowing major classes [29]:

- Earlier knowledge of shape and geometry
- Intensity distribution & spatial context
- DL techniques

The U-shaped CNN architecture [30] and its variants are very common in biomedical image segmentation [1], [30], [44], [59], [75], [76] and were proved very efficient and robust in various segmentation tasks. The methods with best performance share a common multi-step process, including initially a three-dimensional CNN for image segmentation and post processing the derived probability maps using Markov random fields. Many older deep learning liver segmentation algorithms combined NNs with specialized post processing procedures. In the approach of Christ et al. [26], a fully connected 3D NN integrated with conditional random fields were used for liver segmentation with the 3DIRCAD dataset and achieved 0.94 Dice score [29].

Moghbel et al. [6] mentions that liver segmentation using DL has reached very high precision, comparable to human, but from the other hand, robust liver tumor segmentation remains a significant challenge. Some reasons for this are the unavailability of public liver and liver tumor labelled datasets, privacy concerns and the lengthy process of annotating images, which limits access to private datasets [21].

Cohen et al. [77] with an FCN (based on VGG 16, Figure 7.1) trained for 50 epochs, conducted 3D liver segmentation on images from twenty CT taken from SLIVER07 challenge dataset and the Dice scores for his experiments were from 0.87 to 0.89. He also participated in the LITS 2017 benchmark achieving Dice score 0.962 (ranked 2nd) for liver and 0,62 (ranked 11th) for tumor segmentation [29].

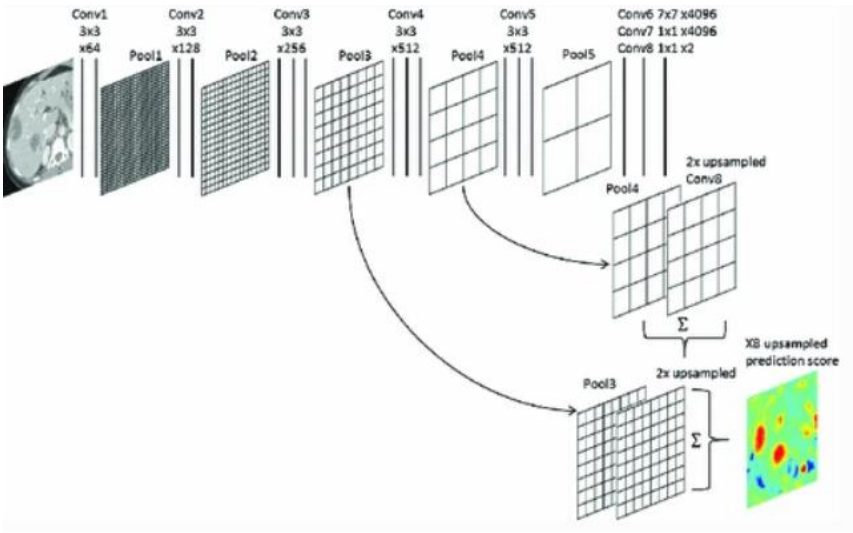



Figure 7.1: FCN architecture for liver segmentation, Cohen et al. – Source [77]

Tang et al. [78], utilized a modified CNN (Figure 7.2), trained on 151 and validated on 25 CT scans from SLIVER07 [79], LiTS17 [29] and a private dataset, tested on 40 CT scans of a private dataset and achieved a Dice score of 0.94 for liver segmentation, compared to experts' annotations.

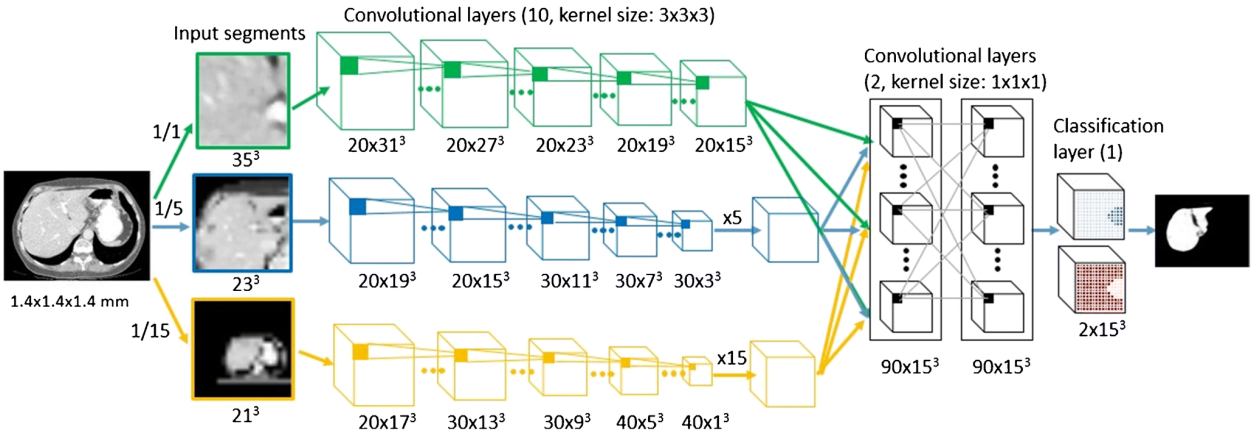



Figure 7.2: Modified multi-scale CNN for liver segmentation – Source [78]

Yuan [76] created a fully convolutional-deconvolutional hierarchal NN (CDNN), which was trained and tested with the MICCAI LiTS 2017 dataset. It was assessed with 131 CTs for training and 70 CTs for testing and achieved a dice score coefficient (DSC) 0.657 for lesion segmentation [21].

Hu et al. [80], utilized an eleven-layer, three-dimensional Convolutional Neural Network for liver segmentation to the Sliver07-I [79] dataset, which contains 20 CT scans. The results of the experiments had an average DSC of 97.24 with a deviation of 0.71, while for Sliver07-II [79] dataset containing 10 CT scans, the average of DSC values were 97.25 with a deviation of 0.65. The same approach, applied on a 12 CTs nonpublic dataset, achieved an average DSC 97.28 and 0.39 deviation.

Vorontsov et al. [27] also created a cascade FCN variant, which has the U-Net structure utilizing short connections between layers, and long (or skip) connections across the network. The study was designed to classify tumors either as primary as hepatocellular carcinoma or as secondary like colon metastases, breast, lung etc. LiTS dataset was also used in this case with 115 CTs for training the model, 15 CTs to validate the model, and 26 CTs to test it. The DSC scores were 0.14 for sizes smaller than 10mm, 0.53 for size between 10mm to 2 mm & 0.68 for larger than 20m [21]. The achieved DSC score of this method in LITS benchmark was 0.645 [29].

An encoder-decoder was proposed by Zhang et al. [75] that used 2D convolutions in encoder bottom to reduce maziness, and three-dimensional convolutions in other layers to investigate spatial and temporal information. To decrease the parameters, DSTS (depth wise and spatiotemporal separate) factorization of 3D convolutions was utilized. For model evaluation, they used LiTS [29] and 3Dircadb [72] and the scores achieved were 73.0% and 94.1% Dice per case respectively for tumor segmentation [21].

Peng et al. [68] utilized the ResNet50 model, a ResNet variant, to detect the presence of HCC (hepatocellular carcinoma) in scans of 562 patients under treatment and achieved 84.3 accuracy value.

Chen et al. [81] also utilized a pre-trained ResNet50 as encoder of the model, which was trained on the LIITS2017 dataset scans for 2D liver tumor segmentation. The model achieved a Dice score of 0.65 and was submitted in the LITS 2018 benchmark where its score was 0.611.

ResNet50 Architecture is illustrated in Figure 7.3, Figure 7.4 a, b.

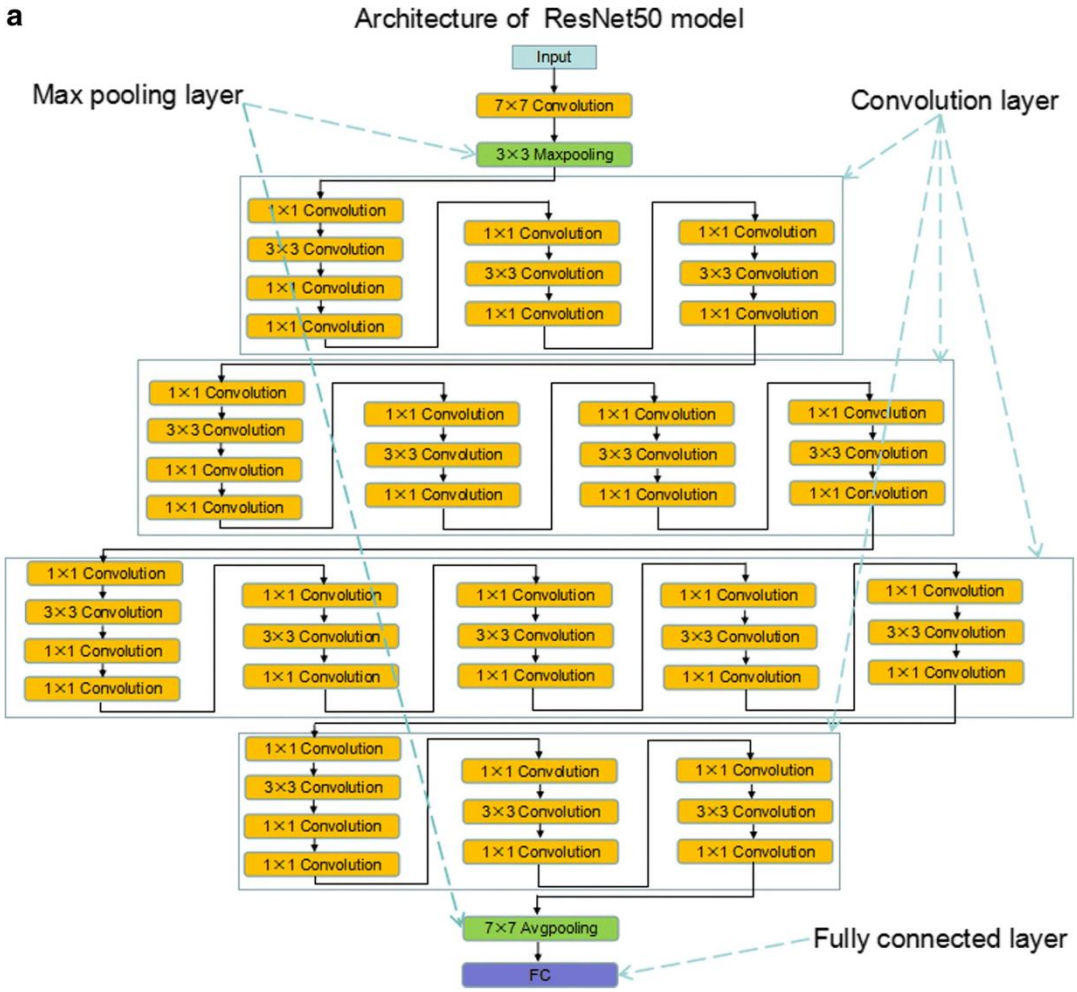



Figure 7.3: a: Resnet50 architecture - Source [68].

In Figure 7.4 c, we see ResNet50 with transfer learning. ResNet50 model has been pre-trained with 1M images of the ImageNet dataset, which may enlist images to 1,000 object classes. A transfer-learning model was trained on CT scans by Peng et al. [68], significantly reducing training time and increasing accuracy.

Chlebus et al. [82], proposed a two-dimension FCN model for liver tumor segmentation and identification, from the LITS dataset CTs. The model consisted of two cascading which increased tumor identification. For tumor segmentation the mean DSC score was 0.72, a little higher than human performance (0.69), but concerning tumor detection, it performed much better, achieving a recall value 92 compared to human 63. The same model participated in the LITS 2017 benchmark achieving a DSC score 0.652 [29].

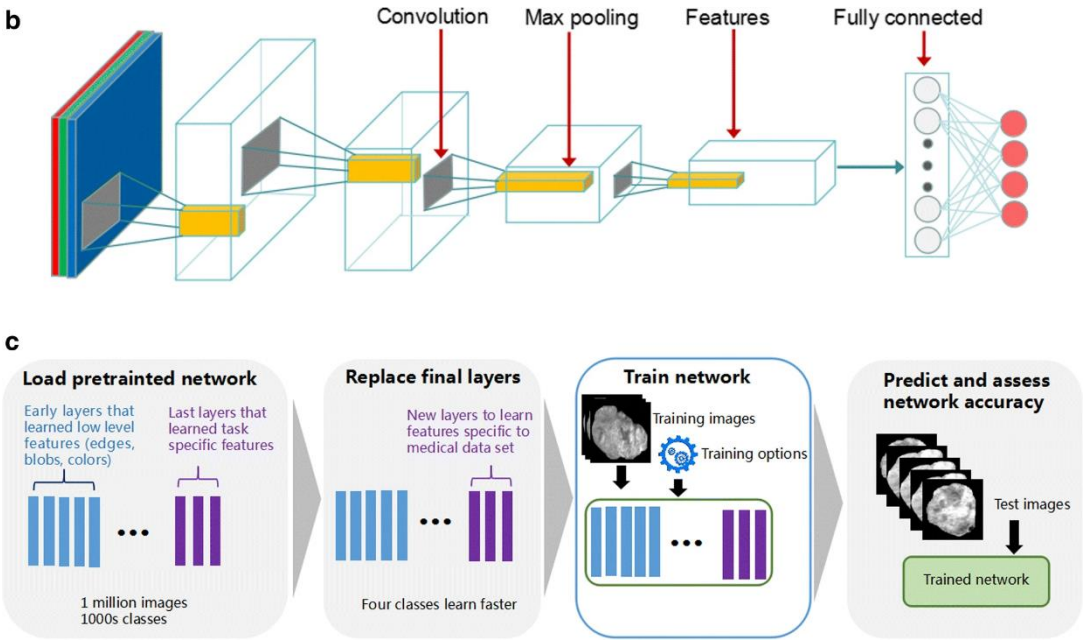



Figure 7.4: b, c: Resnet50 architecture, transfer Learning - Source [68].

Isensee et al. [83] proposed a U-net variant, nnU-Net, a self-configurable DL segmentation approach. The DSC score of this method in LITS dataset was 0.739 for liver tumor segmentation, and 0.962 for liver segmentation, which were the highest achieved scores for LITS 2018 benchmark. Han et al. [84], with the same dataset and a 2.5 dimension Deep CNN model (U-Net, ResNet variant) achieved the highest DSC score 0.674 for LITS ISBI 2017 tumor segmentation [29].

Table 3 shows a comparative view of the previous implementations.

Table 3: Segmentation Implementations

| Implementation | Model | Dataset | Liver/Tumor | Dice Score |
|---|---|---|---|---|
| Christ et al. [26] | 3D FCN integrated with conditional random fields | 3DIRCAD | Liver | 0.94 |

| | | | | |
|---|---|---|---|---|
| Vorontsov et al. [27] | Cascade FCN variant | LiTS17 | Liver/Tumor | 0.951 / 0.645 |
| Zhou et al. [59] | Unet++ | LiTS18 | Liver | 82.90 (IoU) |
| Zhang et al. [75] | CNN | LiTS17/ 3Dircadb | Tumor | 0.73 |
| Yuan [76] | CNN | LiTS17 | Liver/Tumor | 0.963 / 0.657 |
| Cohen et al. [77] | FCN (VGG 16) | SLIVER07, 20 CT scans | Liver/Tumor | 0.88 (LITS 0.962 / 0.62) |
| Hu et al.[80] | CNN | Sliver07/nonpublic | Liver | 0.972 |
| Tang et al. [78] | CNN | SLIVER07, LiTS17 | Liver | 0.94 |
| Chen et al. [81] | CNN Resnet50 | LiTS18 | Tumor | 0.65 (0.611 rated in LITS, and 0.954 in liver) |
| Chlebus et al.[82] | Cascade 2D FCN | LiTS17 | Tumor | 0.72 (0.676 in LITS, for liver 0.96) |
| Isensee et al. [83] | nnU-Net | LiTS2018 | Liver/Tumor | 0.962 / 0.739 |
| Han et al. [84] | 2.5D CNN | LiTS17 | Tumor | 0.674 (also rated in LITS 0.943 for liver) |

# 8 Experimental results

For liver and tumor segmentation, four different experiments were implemented:

- Liver and tumor segmentation
- Liver segmentation
- Tumor Segmentation from all CT scans
- Tumor Segmentation from Liver masked CT scans

For each experiment, transfer learning Resnet34 and Resnet50 pre-trained models were used, with various data, test combinations and parameters (code and output files with results can be found in git [86]).

The following resources and tools were used:

- Hardware: NVIDIA DGX Station™ A100 (HMU Intelligent Systems and Computer Architecture lab, ISCA)
- Programming environment: Python
- Deep learning package: FASTAI Library
- Image processing: OpenCV – Open Computer Vision Library
- NIfTI images processing: NiBabel Library

Below, the steps which were followed and the whole process, are explained in more detail.

## 8.1 Data

For training and testing the models, the LITS benchmark dataset was used. It consists of 131 patients CT scans (in NIfTI format, .nii file type) for training and another 70 unlabeled (without ground truth masks of liver and tumors) CT scans for testing purposes (organizer used them to evaluate submissions). In our implementation, only the 131-labeled scans were used, for training and testing.

Test images of LITS benchmark dataset (70 CT scans) were not used in this work, because no labels were available for them to test results.

After downloading the dataset, the images were preprocessed and split into a training and test set. More specifically the preprocessing techniques applied to the initial dataset images were:

- Hounsfield Unit (HU) windowing

- Histogram equalization (HE)

Afterwards, images were saved in .jpg format and masks in .png format in the corresponding folders (test/train, images/masks).

From the LITS dataset, 14 volumes out of 131 were used for testing (volumes 87-99 and 9, about 11% of volumes), corresponding to 10481 images (about 18% of images, see Table 4, test_images, test_masks). The rest 48.157 (from the remaining 117 volumes, see Table 4, train_images, train_masks) were used for training purposes. The total number of dataset images is 58638.

As we see in Table 2, liver and tumor volumes in LITS test images are larger than the ones used in this work (train/test images are all from LITS train images with known masks). This means that if this implementation was performed in LITS train/test images, it would possibly achieve even higher DSC scores, since segmenting larger structures is easier.

**New masks creation**

Dataset mask files contain 0 for background, 1 for liver and 2 for liver tumor in the corresponding pixels of the corresponding CT scan image. However, to train and test the models for all approaches for liver and tumor segmentation, various folders (Figure 8.1) containing test and train files were created after processing initial dataset's .nii images with a python application [86], and keep mask values and/or areas, depending their purpose, for various segmentation approaches:

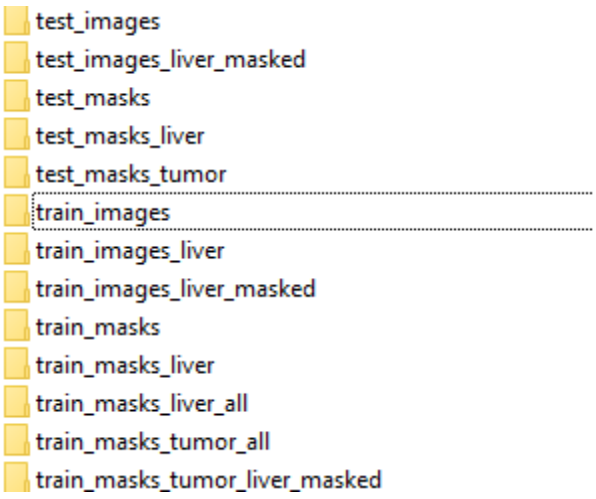



Figure 8.1: Folders containing train and test images.

In Table 4 train and test data folders are explained and demonstrated.

Table 4: Train and test folders

| **Folders** | **Images number** | **Images** | **Mask (Background/ liver/ tumor)** | **Approach** |
|---|---|---|---|---|
| Train_images, train_masks | 48157 | 90% for training | 0/1/2 | Liver and tumor segmentation (initial masks and images) |
| Test_images, test_masks | 10481 | 10% for testing | 0/1/2 | Liver and tumor segmentation (initial masks and images) |
| Train_images_liver_masked Train_masks_tumor_liver_ masked | 16110 | Images containing liver, liver area only | 0/0/1 | Tumor segmentation |
| Train_masks_liver_all (with train_images for liver segmentation, train from all CTs) | 48157 | train_images | 0/1/1 | Liver segmentation |
| Train_masks_tumor_all (tumor segmentation from all CTs) | 48157 | train_images | 0/0/1 | Tumor segmentation |
| Test_masks_liver | 3053 | train_images* | 0/1/1 | Liver segmentation |
| Test_images_liver_masked (with test_masks_tumor, for tumor segmentation liver masked) | 3053 | Images containing liver, liver area only* | 0/0/1 | Tumor segmentation |
| Test_masks_tumor | 1356 | Images containing tumor* | 0/0/1 | Tumor segmentation |

* For testing**,** if no mask file is found for an image file, it is considered that it does not contain liver and/or tumor and mask image is assumed full of zeros.

Prediction folders (Figure 8.2) contain the masks with the output prediction masks of the various models.

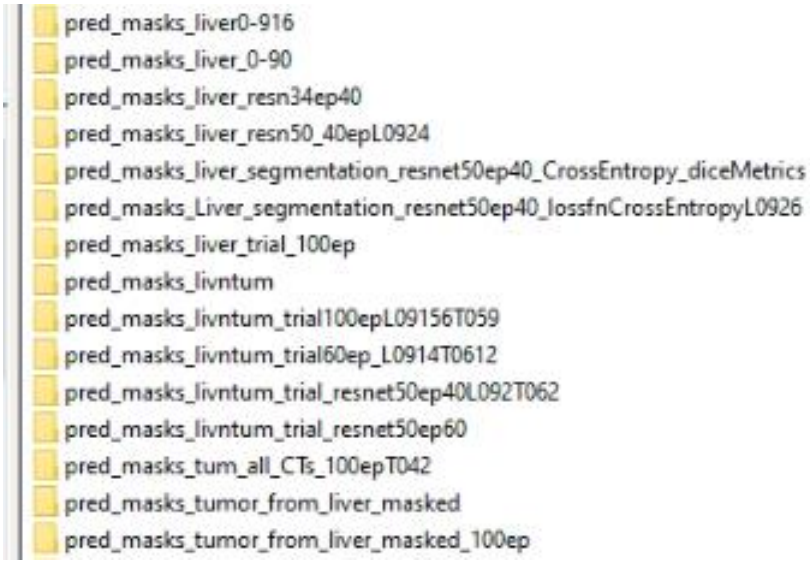



**Figure 8.2:** Folders containing prediction masks.

In Table 5, prediction masks folders are explained and demonstrated (**xxx** denotes the experiment performed for specific model and train parameters).

Table 5: Output prediction masks folders

| **Folders** | **Approach** | **Mask (Background/ liver/ tumor)** |
|---|---|---|
| Pred_masks_livntum_xxx | Liver and tumor segmentation | 0/1/2 |
| Pred_masks_liver_xxx | Liver segmentation | 0/1/1 |
| Pred_masks_tumor_from_liver_masked_xxx | Tumor segmentation | 0/0/1 |

| | | |
|---|---|---|
| ( models trained with train_images_liver_masked and train_masks_tumor) | | |
| Pred_masks_tum_all_CTs_xxx (models trained with train_images and train_masks_tumor_all) | Tumor segmentation | 0/0/1 |

## 8.2 Training

After creating train and test data folders, the next step was to create and train the model.

After many efforts, the optimal model parameters resulting to higher DSC scores were the following.

**Model training data parameters**:

- **Validation** data set to 15% of training data.
- **Seed** set to 42 for reproducibility (use the same input images for validation and train every time we train the models).
- **Resize** input images from 512x512 to 256x256 for faster processing and system stability. Scaling to 128x128 reduced training time, but also model's performance. Higher resolutions increased performance and training time. If no transformation used and image sizes left to 512x512 model training was very slow and was collapsing after a few epochs.
- Input images were normalized (ImageNet normalization: mean = [0.485, 0.456, 0.406] and std = [0.229, 0.224, 0.225] since Resnet34 and Resnet50 are pre-trained on ImageNet data).

## 8.3 Evaluation Metrics

**Segmentation metrics**

For each image, each pixel of the prediction mask was compared to the corresponding test mask (ground truth) pixel, and based on this comparison, for the whole prediction mask image, the following metrics were calculated:

- TP, TN, FP, FN: pixels of the prediction mask.
- Accuracy
- Precision
- Recall
- F1_score
- IoU
- DSC
- FNR (False negative rate)
- FPR (False positive rate)
- Sensitivity (Sens)
- Specificity (Spec)

To validate the model, the averages of each one of the above metrics from all test images were calculated. It is worth mentioning that images with only background (only 0 pixel values) which are correctly predicted and all pixels are TN, do not contribute to the average of DSC and IoU, since they have NaN value and there is no positive impact in the results.

All metrics containing TN (like Accuracy, Specificity) return high scores, very close to 1, especially for tumors, because background (0) pixels are many more than liver (1) and tumor (2) and they are easily predicted.

The most suitable metrics are DSC and IoU which are segmentation oriented, and they are correlated and close to each other. DSC was used in all liver and tumor segmentation implementations, so it was selected as the basic metric to compare their results.

In the following sections, all results and metrics for each segmentation effort with various parameters and models are presented (average scores of all test images). From there, it is obvious that DSC and IoU give the most relative and representative results of segmentation performance, as recorded in the .csv report file with the per image results for each segmentation effort.

**Liver and tumor detection per image**

To measure the ability of each model to correctly predict the existence of liver and/or tumor in images the following metrics were calculated:

- TP, TN, FP, FN: for liver or tumor detection in each image. Two methods were used to find liver or tumor existence in image: Existence of pixels with liver or tumor value in mask, or detection of areas with values 1 or 2 in masks with python OpenCV library function. These two methods gave the same results.
- Accuracy
- Precision
- Recall
- F1_score

Because of the presence of many TN images, these metrics give in some cases high results, which are not representative of the performance.

All above metrics for each experiment can be found in csv files in git repository [86].

## 8.4 Experiments

On experiments implementation, for Resnet50 model, training time was about 30 minutes per epoch, while for Resnet34 it was about 08:30 minutes (for 256x256 image resolution).

In the following sections, more details about the performed experiments are presented.

### *8.4.1 Liver and liver tumor segmentation*

The first conducted experiment was for segmenting liver and tumor simultaneously from the initial CT scans images. The resulting masks contain 1 in the areas with liver, 2 in tumor areas and 0 elsewhere. The dataset in its initial form was used for these experiments, without any transformation.

For performing the experiments, the following and folders were utilized:

- Train files folders: train_images, train_masks
- Test files folders: test_images, test_masks
- Output prediction masks folder: pred_masks_livntum_trial_xxx

Model creation and parameters:

Pre-trained transfer learning resnet34 and resnet50 models were utilized. Parameters resulting to highest DSC scores for liver and tumor segmentation in this work are in Table 6.

Table 6: Model and data parameters

| Parameter | Value |
|---|---|
| **Input data** | |
| Transformations | Resize 256x256, normalization |
| Seed | 42 |
| Validation data (train) | 15% |
| **Model** | |
| Model | Resnet50 |
| Learning rate | 0.001 |
| Activation function | Mish |
| Optimizer | Ranger |
| Epochs | 40 |
| Weight decay | 0.1 |

Train and validations losses and training accuracy diagrams for Resnet50, 40 epochs for liver and tumor segmentation are depicted in Figure 8.3**.**

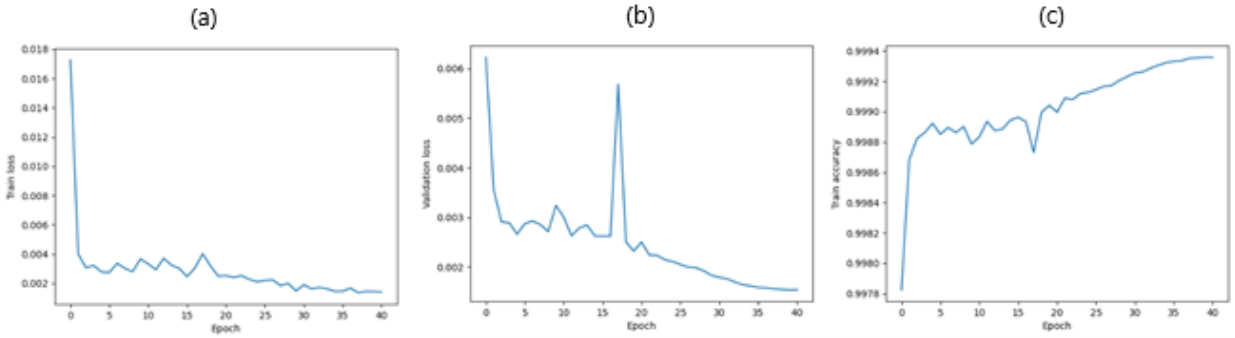



Figure 8.3: Liver and tumor segmentation, Resnet50 train (a) &validation (b) losses and train accuracy (c).

As we can see, the models results (Table 7, Table 8) are very close with Resnet50 giving slightly better results. From there we can see that top DSC score achieved for **liver segmentation** was **0.9226** (Table 7**,** experiment 7) by a Resnet50 model trained for 40 epochs. For **tumor segmentation**, the best performance was obtained by a Resnet50 for liver and tumor segmentation, trained for 60 epochs giving a DSC score **0.619** (Table 7**,** experiment 8)**,** very close to the same model trained for 40 epochs, which achieved a score of 0.6189 (Table 7**,** experiment 7). These

models were very close in tumor segmentation, but as we see in Table 7, the model trained for 60 epochs achieved lower DSC (0.9096) for liver segmentation, compared to the one trained for 40 epochs (0.9226). From these we conclude that concerning the overall performance in liver and tumor segmentation, the Resnet50 model trained for 40 epochs was the most efficient.

Resnet34 is very close in performance with DSC highest scores 0.9157 (experiment 6) for liver and 0.6116 (experiment 4) for tumor segmentation, having the advantages of much faster training process and much smaller model file sizes.

Table 7: Liver and tumor segmentation results.

| Exp/nt | Task | Resnet | Epochs | Seed | Reso | pct | Accuracy | Precision | Recall | F1_score | IOU | DSC | FNR | FPR | Spec |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Liver | 34 | | no | 128 | 0.20 | 0.9982 | 0.9120 | 0.8984 | 0.9303 | 0.8235 | 0.8693 | 0.1016 | 0.0880 | 0.9991 |
| | Tumor | 34 | | no | 128 | 0.20 | 0.9997 | 0.5470 | 0.7493 | 0.7950 | 0.4430 | 0.5116 | 0.2507 | 0.4530 | 0.9997 |
| 2 | Liver | 50 | | no | 128 | 0.20 | 0.9982 | 0.9241 | 0.8843 | 0.9244 | 0.8223 | 0.8706 | 0.1157 | 0.0759 | 0.9992 |
| | Tumor | 50 | | no | 128 | 0.20 | 0.9997 | 0.5855 | 0.7512 | 0.7858 | 0.4695 | 0.5410 | 0.2488 | 0.4145 | 0.9998 |
| 3 | Liver | 34 | | no | 256 | 0.20 | 0.9983 | 0.9005 | 0.9104 | 0.9334 | 0.8228 | 0.8641 | 0.0896 | 0.0995 | 0.9991 |
| | Tumor | 34 | | no | 256 | 0.20 | 0.9996 | 0.5290 | 0.7830 | 0.8052 | 0.4441 | 0.5008 | 0.2170 | 0.4710 | 0.9997 |
| 4 | Liver | 34 | 60 | 42 | 256 | 0.15 | 0.9987 | 0.9458 | 0.9223 | 0.9472 | 0.8772 | 0.9139 | 0.0777 | 0.0542 | 0.9995 |
| | Tumor | 34 | 60 | 42 | 256 | 0.15 | 0.9997 | 0.6721 | 0.7763 | 0.8326 | 0.5453 | **0.6116** | 0.2237 | 0.3279 | 0.9998 |
| 5 | Liver | 34 | 60 | 42 | 256 | 0.10 | 0.9986 | 0.9271 | 0.9232 | 0.9467 | 0.8607 | 0.8968 | 0.0768 | 0.0729 | 0.9993 |
| | Tumor | 34 | 60 | 42 | 256 | 0.10 | 0.9997 | 0.6588 | 0.7933 | 0.8302 | 0.5454 | 0.6086 | 0.2067 | 0.3412 | 0.9998 |
| 6 | Liver | 34 | 100 | 42 | 256 | 0.15 | 0.9987 | 0.9484 | 0.9212 | 0.9451 | 0.8782 | 0.9157 | 0.0788 | 0.0516 | 0.9994 |
| | Tumor | 34 | 100 | 42 | 256 | 0.15 | 0.9997 | 0.6474 | 0.7735 | 0.8249 | 0.5212 | 0.5903 | 0.2265 | 0.3526 | 0.9998 |
| 7 | Liver | 50 | 40 | 42 | 256 | 0.15 | 0.9987 | 0.9545 | 0.9217 | 0.9448 | 0.8855 | **0.9226** | 0.0783 | 0.0455 | 0.9995 |
| | tumor | 50 | 40 | 42 | 256 | 0.15 | 0.9997 | 0.6640 | 0.7990 | 0.8178 | 0.5488 | **0.6189** | 0.2010 | 0.3360 | 0.9998 |
| 8 | Liver | 50 | 60 | 42 | 256 | 0.15 | 0.9987 | 0.9417 | 0.9206 | 0.9470 | 0.8727 | 0.9096 | 0.0794 | 0.0583 | 0.9994 |
| | Tumor | 50 | 60 | 42 | 256 | 0.15 | 0.9997 | 0.6694 | 0.7917 | 0.8276 | 0.5524 | **0.6190** | 0.2083 | 0.3306 | 0.9998 |

Concerning liver and tumor detection, by considering existence of pixels of tumor or liver in predict mask as detection criterion, the results are very good compared to LITS benchmark results (Table 8).

From there we can see that for **liver detection** most models gave very good results. Best results came from Resnet50 liver and tumor segmentation model, trained for 40 epochs (Table 8, experiment 7).

Table 8: Liver and tumor detection results.

| Exp/nt | Task | Resnet | Epochs | Seed | Resol | pct | IOU | DSC | TP | FP | FN | TN | Accurac | Precisic | Recall | F1_score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 4 | Liver | 34 | 60 | 42 | 256 | 0.15 | 0.8772 | 0.9139 | 2996 | 46 | 57 | 7382 | 0.9902 | 0.9849 | 0.9813 | 0.9831 |
| | Tumor | 34 | 60 | 42 | 256 | 0.15 | 0.5453 | **0.6116** | 1066 | 195 | 290 | 8930 | 0.9537 | 0.8454 | 0.7861 | 0.8147 |
| 5 | Liver | 34 | 60 | 42 | 256 | 0.10 | 0.8607 | 0.8968 | 2989 | 97 | 64 | 7331 | 0.9846 | 0.9686 | 0.9790 | 0.9738 |
| | Tumor | 34 | 60 | 42 | 256 | 0.10 | 0.5454 | 0.6086 | 1087 | 214 | 269 | 8911 | 0.9539 | 0.8355 | 0.8016 | 0.8182 |
| 6 | Liver | 34 | 100 | 42 | 256 | 0.15 | 0.8782 | 0.9157 | 2996 | 32 | 57 | 7396 | 0.9915 | 0.9894 | 0.9813 | 0.9854 |
| | Tumor | 34 | 100 | 42 | 256 | 0.15 | 0.5212 | 0.5903 | 1030 | 216 | 326 | 8909 | 0.9483 | 0.8266 | 0.7596 | 0.7917 |
| 7 | Liver | 50 | 40 | 42 | 256 | 0.15 | 0.8855 | **0.9226** | 2997 | **13** | 56 | **7415** | **0.9934** | **0.9957** | 0.9817 | **0.9886** |
| | tumor | 50 | 40 | 42 | 256 | 0.15 | 0.5488 | **0.6189** | 1089 | 178 | 267 | 8947 | 0.9575 | 0.8595 | 0.8031 | 0.8303 |
| 8 | Liver | 50 | 60 | 42 | 256 | 0.15 | 0.8727 | 0.9096 | **3000** | 63 | **53** | 7365 | 0.9889 | 0.9794 | **0.9826** | 0.9810 |
| | Tumor | 50 | 60 | 42 | 256 | 0.15 | 0.5524 | **0.6190** | **1090** | **175** | **266** | **8950** | **0.9579** | **0.8617** | **0.8038** | **0.8317** |

Best Recall value was 0.986 (same model trained for 60 epochs, experiment 8)

Indicatively, detection metrics for liver detection of best performing Resnet50 model (experiment 7), trained for 40 epochs were (confusion matrix in Figure 8.4):

- TP: 2997
- FP: 13
- FN: 56
- TN: 7415

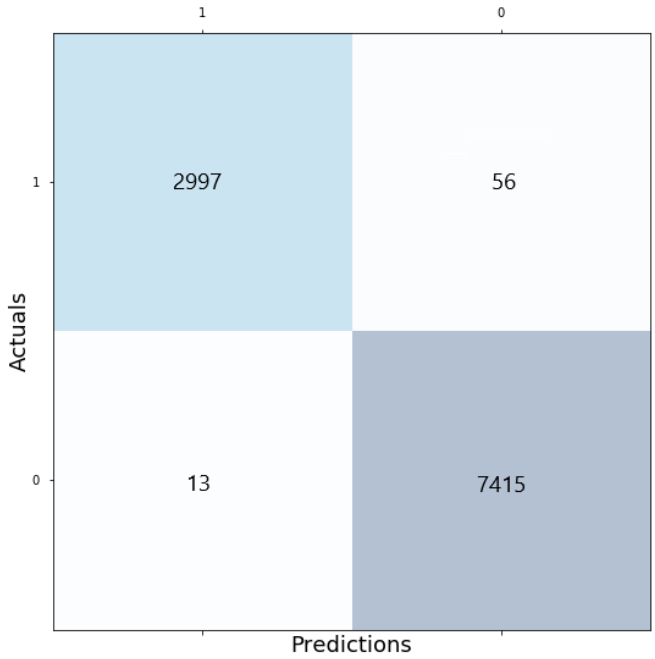



Figure 8.4: Confusion matrix, liver detection.

For tumor detection, from the same table, we can see that best results were all from Resnet50 trained for 60 epochs (experiment 8), slightly better from the one trained for 40 epochs (it worth mentioning from the other hand that the model trained for 40 epochs had better performance in liver segmentation than the one trained for 60 epochs).

In Figure 8.5, some high DSC score segmentations are depicted with their corresponding metrics from Resnet50 model, trained for 40 epochs (average DSC liver 0.9226, tumor 0.6189).

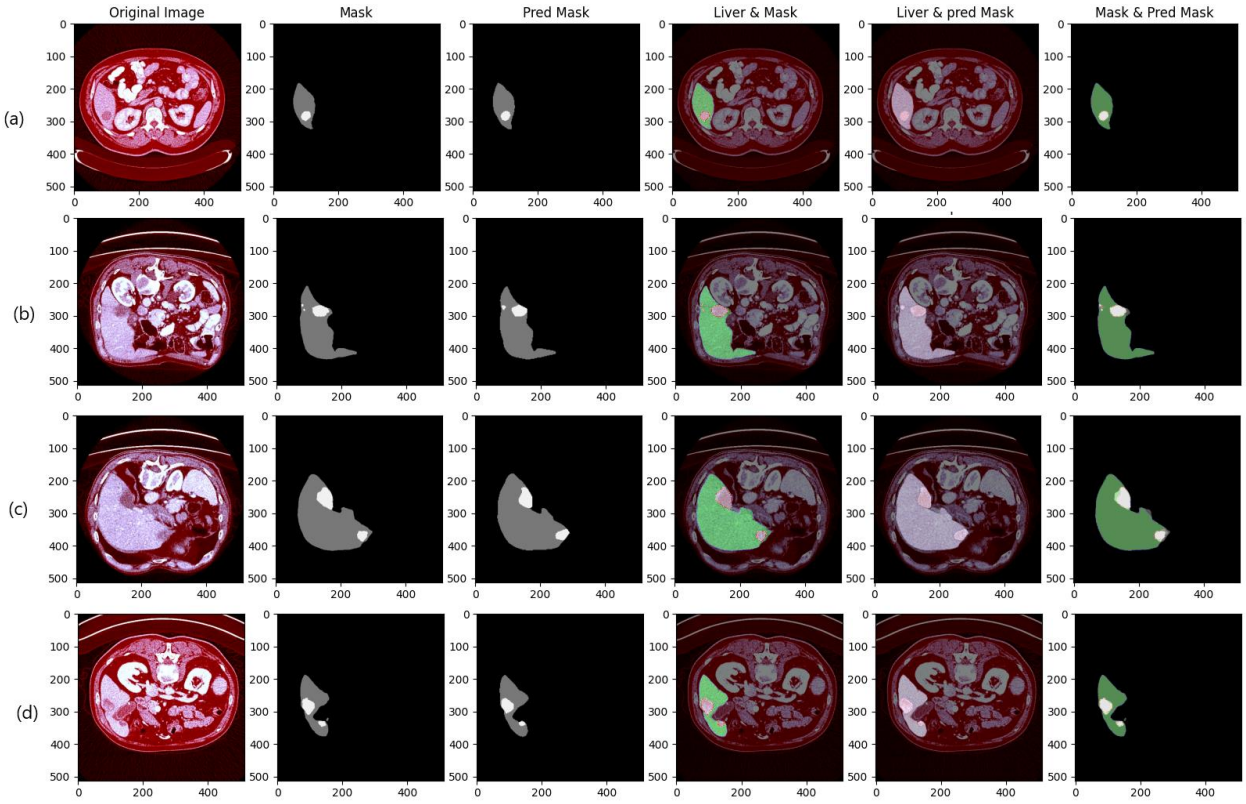



Figure 8.5: Liver and tumor segmentations with good performance.

More specifically, the DSC scores were:

a) liver: 0.97, tumor 0.97.

b) liver: 0.977, tumor: 0.934.

c) liver: 0.967, tumor: 0.936.

d) liver: 0.955, tumor: 0.974.

As we notice, in these images liver and tumors sizes have sufficient number of pixels and tumor numbers per image are not more than three. In some other cases, DSC score and segmentation results are not very high due to the already mentioned difficulties of liver and especially tumor segmentation. Some of these cases are shown in Figure 8.6. We also notice that all cases concern small and multiple tumors, which are more difficult to segment.

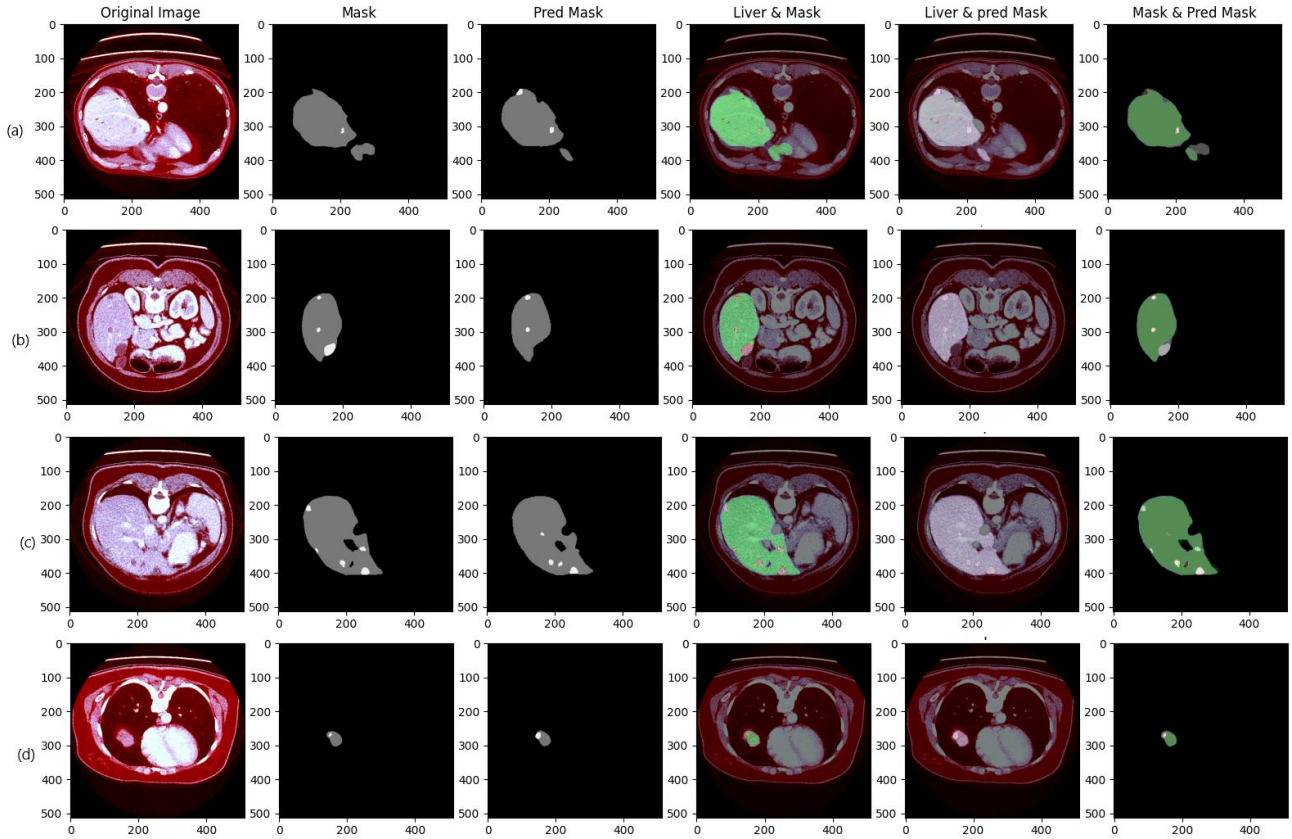



Figure 8.6: Some not close to perfect liver segmentations with lower DSC scores**.**

More specifically the DSC scores were:

a) liver: 0939, Tumor: 0.385, found 2 instead 1 tumor.

b) liver: 0.976, tumor: 0.3, found 2 instead of three tumors.

c) liver: 0.967, tumor: 0.762, many small tumors difficult to segment.

d) liver: 0.945, tumor: 0.46, tumor a few pixels size, low DSC. Although it is a rather good segmentation, for a small size tumor, which is over-segmented, DSC score drops.

In Figure 8.7, image from volume 96, slice 630 with DSC liver: 0.975, tumor: 0.33, we have a tumor of few pixels size. The model although detects tumor existence, it does not manage to segment it very correctly and detects wrongly another small tumor.

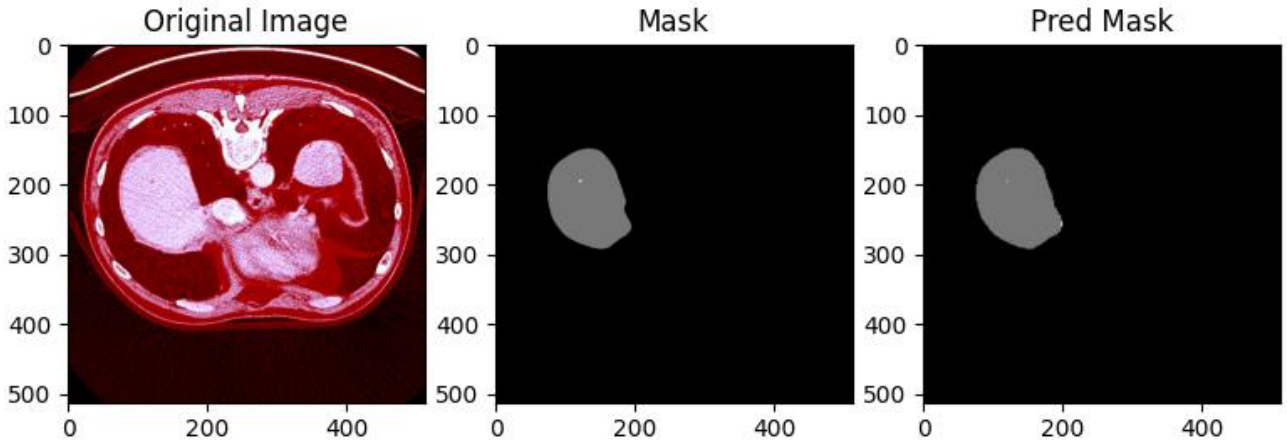



Figure 8.7: DSC liver: 0.975, tumor: 0.33, small tumor

Another case is in Figure 8.8. Here the model did not manage to segment tumor which is larger than liver and segments only a part of liver and a very small part of tumor.

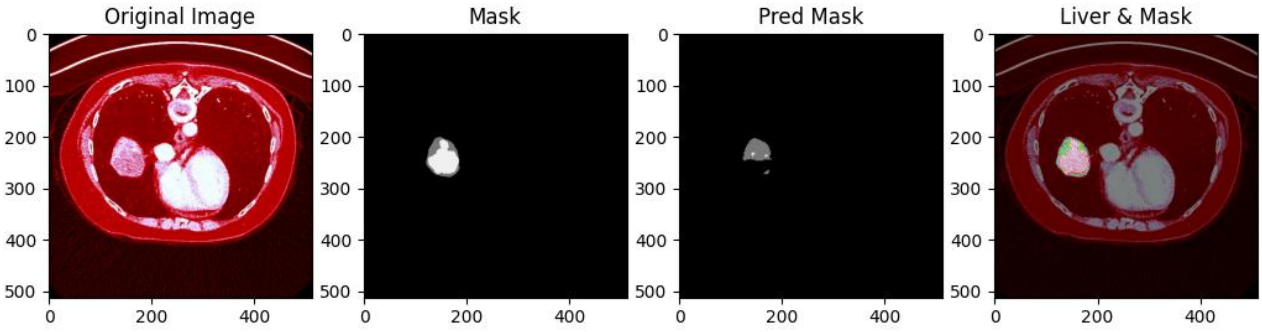



Figure 8.8: DSC liver: 0.44, tumor: 0.074.

To the above cases, FN and FP cases should be added, which in the case of tumors are not negligible.

Out of 1356 images with tumors, 1090 were detected correctly (TP), **266 were not detected** (FN), which are rather many, and another **175 which were identified as tumors but**

**were not** (FP), which are also many, considering how crucial the process is (see confusion matrix Figure 8.9).

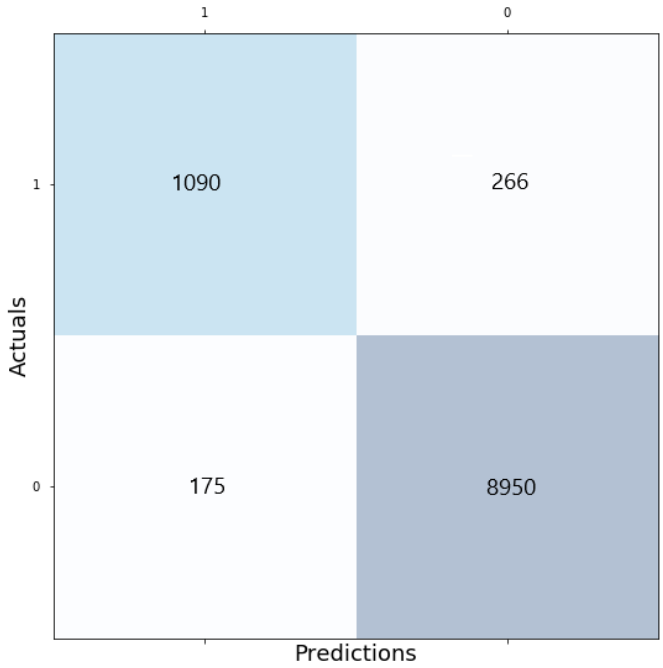



Figure 8.9: Confusion matrix, tumor detection.

### *8.4.2 Liver segmentation*

The second conducted experiment was for segmenting liver only from the initial CT scans. The resulting masks contain 1 in the areas with liver otherwise 0. Initial dataset masks were transformed to contain only liver values as described in section 8.1.

For performing the experiments, the following folders were used:

- Train files folders: train_images, train_masks_liver_all
- Test files folders: test_images, test_masks_liver
- Output prediction masks folder: pred_masks_liver_xxx

Model creation and parameters:

Pre-trained transfer learning resnet34 and resnet50 models were utilized. Parameters resulting to highest DSC scores for liver and tumor segmentation in this work are in Table 9.

Table 9: Model and data parameters

| Parameter | Value |
|---|---|
| **Input data** | |
| Transformations | Resize 256x256, normalization |
| Seed | 42 |
| Validation data (train) | 15% |
| **Model** | |
| Model | Resnet50 |
| Learning rate | 0.001 |
| Activation function | Mish |
| Optimizer | Ranger |
| Loss function | CrossEntropyLossFlat |
| epochs | 40 |
| Weight decay | 0.1 |

Train and validations losses and training accuracy diagrams for Resnet34, 40 epochs for liver and tumor segmentation are depicted in Figure 8.10.

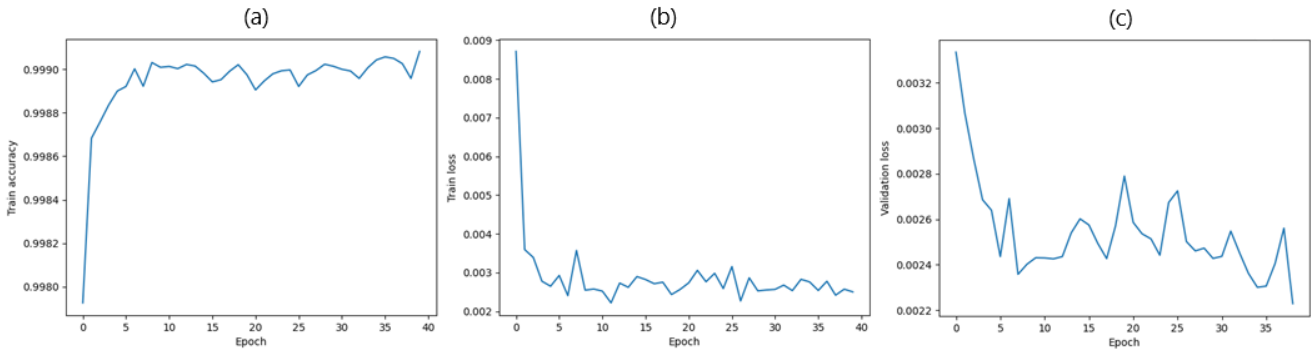



Figure 8.10: Liver segmentation, Resnet34, train accuracy (a) and train (b) &validation (c) losses.

Table 10 contains liver segmentation results. Like liver and tumor segmentation, a Resnet50 model trained for 40 epochs achieved top DSC score also which was **0.9262** (experiment 9), higher than the one achieved in liver and tumor segmentation experiments.

Concerning liver detection, by considering existence of pixels of liver in predict mask as detection criterion, the results were very good if compared against LITS benchmark results (Table 11). The same top performing model (experiment 9) for liver segmentation achieved the higher liver detection scores:

Table 10: Liver segmentation results.

| Exp/nt | Resnet | Epochs | Seed | Reso | pct | Accuracy | Precision | Recall | F1_score | IOU | DSC | FNR | FPR | Spec |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 34 | | no | 128 | 0.20 | 0.9987 | 0.9419 | 0.9232 | 0.9417 | 0.8708 | 0.9061 | 0.0768 | 0.0581 | 0.9994 |
| 2 | 50 | | no | 128 | 0.20 | 0.9980 | 0.9292 | 0.8700 | 0.9202 | 0.8137 | 0.8589 | 0.1300 | 0.0708 | 0.9993 |
| 3 | 34 | 30 | no | 256 | 0.15 | 0.9988 | 0.9427 | 0.9312 | 0.9480 | 0.8809 | 0.9161 | **0.0688** | 0.0573 | 0.9995 |
| 4 | 34 | 100 | no | 256 | 0.15 | 0.9986 | 0.9193 | 0.9240 | 0.9434 | 0.8519 | 0.8864 | 0.0760 | 0.0807 | 0.9994 |
| 5 | 34 | 100 | 42 | 256 | 0.15 | 0.9986 | 0.9436 | 0.9164 | 0.9422 | 0.8678 | 0.9053 | 0.0836 | 0.0564 | 0.9994 |
| 6 | 50 | 40 | 42 | 256 | 0.15 | 0.9987 | 0.9665 | 0.9172 | 0.9488 | 0.8889 | 0.9240 | 0.0828 | **0.0335** | 0.9995 |
| 7 | 50 | 60 | 42 | 256 | 0.15 | 0.9987 | 0.9501 | 0.9253 | 0.9465 | 0.8830 | 0.9188 | 0.0747 | 0.0499 | 0.9995 |
| 8 | 34 | 40 | 42 | 256 | 0.15 | 0.9987 | 0.9472 | 0.9252 | 0.9452 | 0.8799 | 0.9152 | 0.0748 | 0.0528 | 0.9995 |
| 9 (Cross entr) | 50 | 40 | 42 | 256 | 0.15 | 0.9988 | 0.9595 | 0.9256 | 0.9463 | **0.8909** | **0.9262** | 0.0744 | 0.0405 | 0.9995 |

- Accuracy 0.9941
- Precision 0.9970
- F1-score 0.9898
- Recall 0.9826

Table 11: Liver detection results.

| Exp/nt | Resnet | Epochs | Seed | Resol | pct | IOU | DSC | TP | FP | FN | TN | Accuracy | Precision | Recall | F1_score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3 | 34 | 30 | no | 256 | 0.15 | 0.8809 | 0.9161 | 3020 | 69 | 33 | 7359 | 0.9903 | 0.9777 | **0.9892** | 0.9834 |
| 4 | 34 | 100 | no | 256 | 0.15 | 0.8519 | 0.8864 | 2985 | 123 | 68 | 7305 | 0.9818 | 0.9604 | 0.9777 | 0.9690 |
| 5 | 34 | 100 | 42 | 256 | 0.15 | 0.8678 | 0.9053 | 2977 | 42 | 76 | 7386 | 0.9887 | 0.9861 | 0.9751 | 0.9806 |
| 6 | 50 | 40 | 42 | 256 | 0.15 | 0.8889 | 0.9240 | 2987 | 10 | 66 | 7418 | 0.9927 | 0.9967 | 0.9784 | 0.9874 |
| 7 | 50 | 60 | 42 | 256 | 0.15 | 0.8830 | 0.9188 | 2995 | 25 | 58 | 7403 | 0.9921 | 0.9917 | 0.9810 | 0.9863 |
| 8 | 34 | 40 | 42 | 256 | 0.15 | 0.8799 | 0.9152 | 3002 | 40 | 51 | 7388 | 0.9913 | 0.9869 | 0.9833 | 0.9851 |
| 9 (Cross entr) | 50 | 40 | 42 | 256 | 0.15 | 0.8909 | **0.9262** | 3000 | 9 | 53 | 7419 | **0.9941** | **0.9970** | 0.9826 | **0.9898** |

Indicatively, detection metrics for liver detection of best performing Resnet50 model, trained for 40 epochs were (confusion matrix in Figure 8.11):

- TP: 3000
- FP: 9

- FN: 53
- TN: 7419

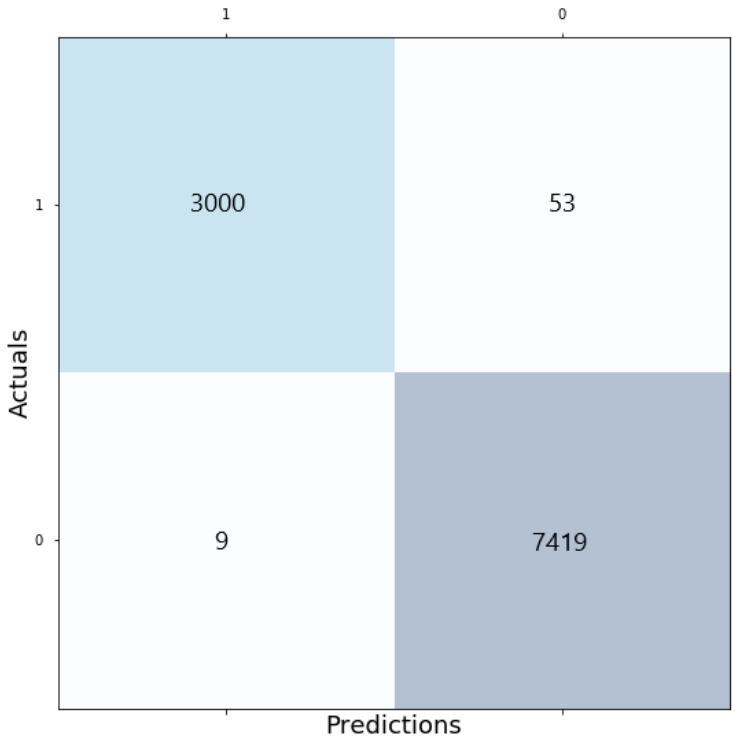



Figure 8.11: Confusion matrix, liver detection

The results of this model are better than the results of liver and tumor segmentation models. In Figure 8.12, various liver segmentation cases from the top performing model (experiment 9) are depicted, with high and low DSC scores.

More specifically, DSC values for liver segmentation of images were:

a) 0.98
b) 0.644 (small liver area)
c) 0.151
d) 0
e) 0 (liver detected in non-liver area, only a few cases like this)
f) 0.38
g) 0.76

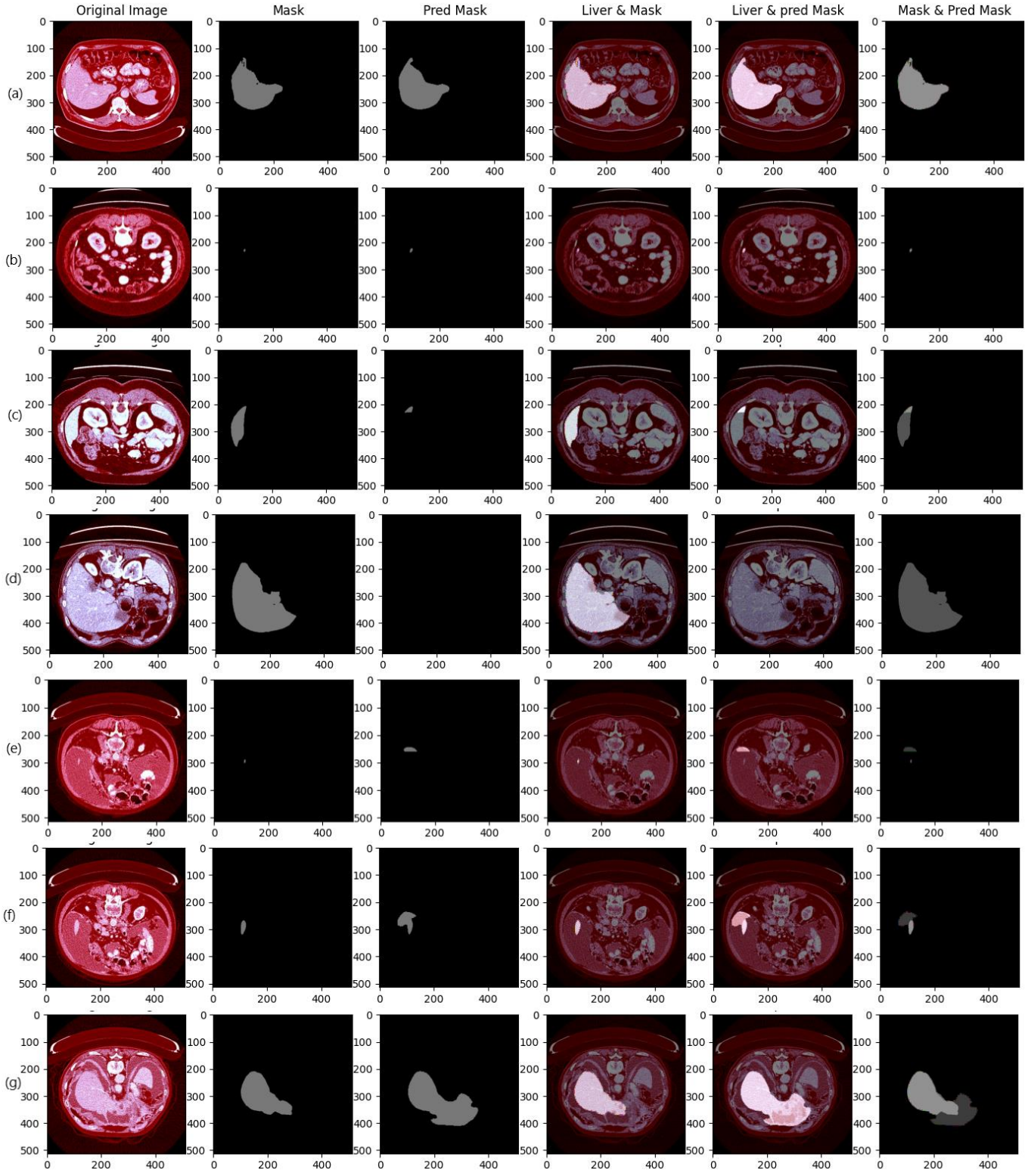



Figure 8.12: Liver segmentations

### *8.4.3 Tumor Segmentation from all CT scans*

The third conducted experiment was for segmenting tumor only, directly from the abdominal CT scans. The resulting masks contain 1 in the areas with tumor, otherwise 0. Initial dataset masks were transformed to contain only tumor values as described in section 8.1.

For performing the experiments, the following folders were used:

- Train files folders: train_images, train_masks_tumor_all
- Test files folders: test_images, test_masks_tumor
- Output prediction masks folder: pred_masks_tum_all_CTs_xxx

Model creation and parameters:

Pre-trained transfer learning resnet34 and resnet50 models were utilized. Parameters resulting to highest DSC scores for liver and tumor segmentation in this work are in Table 12.

Table 12: Model and data parameters

| Parameter | Value |
|---|---|
| **Input data** | |
| Transformations | Resize 256x256, normalization |
| Seed | 42 |
| Validation data (train) | 15% |
| **Model** | |
| Model | Resnet50 |
| Learning rate | 0.001 |
| Activation function | Mish |
| Optimizer | Ranger |
| epochs | 40 |
| Weight decay | 0.1 |

The results of these efforts can be found in

Table 13, from where it is obvious that the DSC scores achieved are much lower compared to liver and tumor segmentation scores. Here, the highest DSC score achieved was **0.465**.

Table 13: Tumor segmentation from abdominal CT scans

| Exper/nt | Resnet | Epochs | Seed | Reso | pct | Accuracy | Precision | Recall | F1_score | IOU | DSC | FNR | FPR | Spec |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 34 | | no | 128 | 0.20 | 0.9996 | 0.7334 | **0.4873** | 0.7045 | **0.3837** | **0.4652** | **0.5127** | 0.2666 | 0.9998 |
| 2 | 34 | | no | 128 | 0.20 | 0.9996 | 0.7086 | 0.3805 | 0.6552 | 0.3126 | 0.3894 | 0.6195 | 0.2914 | 0.9999 |
| 3 | 50 | 5 | no | 128 | 0.20 | 0.9996 | 0.7327 | 0.4184 | **0.7046** | 0.3380 | 0.4126 | 0.5816 | 0.2673 | 0.9999 |
| 4 | 50 | | no | 128 | 0.20 | 0.9996 | 0.7271 | 0.3487 | 0.6465 | 0.2909 | 0.3653 | 0.6513 | 0.2729 | 0.9999 |
| 5 | 34 | 100 | no | 256 | 0.15 | 0.9996 | **0.7610** | 0.4237 | 0.6888 | 0.3503 | 0.4272 | 0.5763 | **0.2390** | 0.9999 |

Concerning tumor detection, the results (Table 14) are also lower than liver and tumor segmentation experiments.

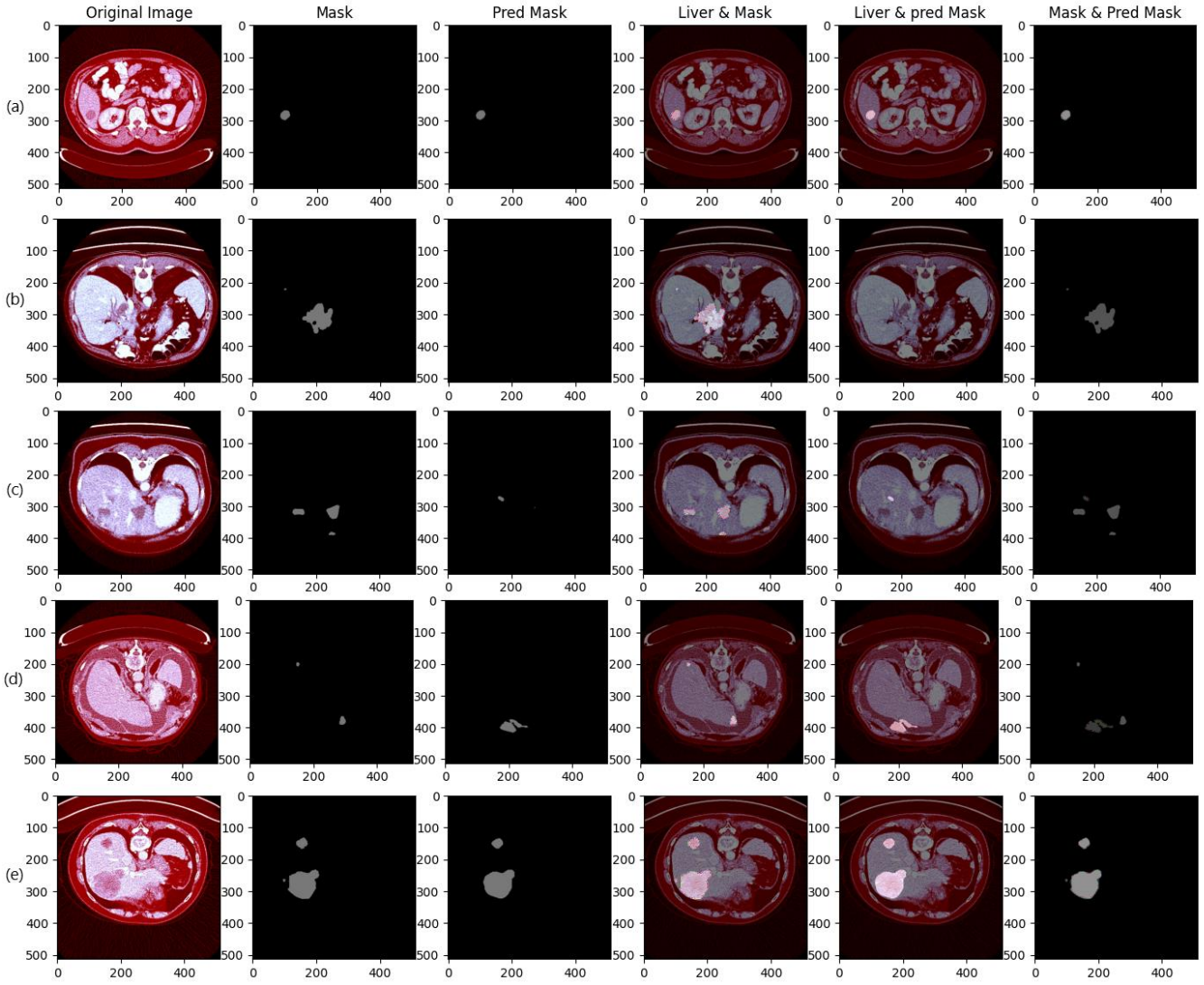



Figure 8.13: Tumor segmentations from abdominal CT scans

Table 14: Tumor detection from abdominal CT scans

| Exper/nt | Resnet | Epochs | Seed | Resol | pct | IOU | DSC | TP | FP | FN | TN | Accuracy | Precision | Recall | F1_score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | 34 | 100 | no | 256 | 0.15 | 0.3503 | 0.4272 | 900 | 63 | 456 | 9062 | 0.950 | 0.935 | 0.664 | 0.776 |

In Figure 8.13, various tumor segmentations are depicted, with high and low DSC scores.
More specifically, DSC values for tumor segmentation:

a) 0.95
b) 0 (FN)
c) 0.00532
d) 0 (tumor detection in non-tumorous area, only few cases like this)
e) 0.956.

### *8.4.4 Tumor Segmentation from Liver masked CT scans*

The fourth conducted experiment was also for tumor segmentation, with the difference that the images contained liver areas only (liver masked CT scans). In non-liver areas, pixel values were zero. These images, for training and testing, were prepared with python code from the initial CT scan images (as described in section 8.1). The resulting prediction masks contain one in the areas with tumor, otherwise zero.

For the implementation, the following folders were used:

- Train files folders: train_images_liver_masked, train_masks_tumor_liver_masked
- Test files folders: test_images_liver_masked, test_masks_tumor
- Output prediction masks folder: pred_masks_tumor_from_liver_masked_.xxx

Table 15 contains the results of various experiments for tumor segmentation from liver masked images. The highest DSC score achieved was **0.5258**. Although the results are better than tumor segmentation from all CT scans (Table 13, DSC score 0.465), still they are not as good as liver and tumor segmentation results (Table 7, DSC score 0.619).

Table 15: Tumor segmentation from liver masked CT scans

| Exper/nt | Resnet | Epochs | Seed | Reso | pct | Accuracy | Precision | Recall | F1_score | IOU | DSC | FNR | FPR | Spec |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 34 | | no | 128 | 0.20 | 0.9990 | 0.7226 | 0.5646 | 0.7407 | 0.4271 | 0.5098 | 0.4354 | 0.2774 | 0.9995 |
| 2 | 50 | 10 | no | 128 | 0.20 | 0.9990 | 0.6875 | 0.5904 | **0.7522** | 0.4317 | 0.5122 | **0.4096** | 0.3125 | 0.9994 |
| 3 | 50 | | no | 128 | 0.20 | 0.9988 | 0.6672 | 0.4811 | 0.7088 | 0.3528 | 0.4309 | 0.5189 | 0.3328 | 0.9995 |
| 4 | 50 | 20 | no | 256 | 0.15 | 0.9988 | 0.6858 | **0.5849** | 0.7425 | 0.4255 | 0.5068 | 0.4151 | 0.3142 | 0.9995 |
| 5 | 34 | 20 | no | 256 | 0.15 | 0.9991 | 0.7545 | 0.5576 | 0.7425 | 0.4382 | 0.5213 | 0.4424 | 0.2455 | 0.9996 |
| 6 | 34 | 20 | no | 128 | 0.15 | 0.9988 | 0.6943 | 0.5535 | 0.7299 | 0.4121 | 0.4989 | 0.4465 | 0.3057 | 0.9994 |
| 7 | 34 | 40 | no | 256 | 0.15 | 0.9990 | 0.7410 | 0.5612 | 0.7470 | 0.4375 | 0.5217 | 0.4388 | 0.2590 | 0.9995 |
| 8 | 34 | 100 | no | 256 | 0.15 | 0.9990 | **0.7823** | 0.5369 | 0.7368 | **0.4395** | **0.5258** | 0.4631 | **0.2177** | 0.9996 |

Table 16 contains the metrics for tumor detection from liver masked CT scans. The same model with top tumor segmentation performance achieved higher tumor detection metrics. Comparing the results with the previous approaches, we conclude that, like tumor segmentation, they are better than the tumor segmentation from all CT scans (see section 8.4.3), still they are not as good as liver and tumor segmentation results (see section 8.4.1). Besides, this approach presupposes liver segmentation (to create liver masked images), and any loses in this process, further decrease the performance of tumor segmentation.

Table 16: Tumor detection from liver masked CT scans

| Exper/nt | Resnet | Epochs | Seed | Resol | pct | IOU | DSC | TP | FP | FN | TN | Accurac | Precisic | Recall | F1_score |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 4 | 50 | 20 | no | 256 | 0.15 | 0.4255 | 0.5068 | **1066** | 175 | **290** | 1522 | 0.848 | 0.859 | **0.786** | 0.821 |
| 5 | 34 | 20 | no | 256 | 0.15 | 0.4382 | 0.5213 | 1039 | 104 | 317 | 1593 | 0.862 | 0.909 | 0.766 | 0.832 |
| 6 | 34 | 20 | no | 128 | 0.15 | 0.4121 | 0.4989 | 1040 | 129 | 316 | 1568 | 0.854 | 0.890 | 0.767 | 0.824 |
| 7 | 34 | 40 | no | 256 | 0.15 | 0.4375 | 0.5217 | 1054 | 126 | 302 | 1571 | 0.860 | 0.893 | 0.777 | 0.831 |
| 8 | 34 | 100 | no | 256 | 0.15 | **0.4395** | **0.5258** | 1021 | **55** | 335 | **1642** | **0.872** | **0.949** | 0.753 | **0.840** |

In Figure 8.14, various tumor segmentations from liver masked images are depicted, with high and low DSC scores.

More specifically tumor segmentation DSC values for each image are:

a) 0 (tumor detection in non-tumorous area, only few cases like this)

b) 0.00966, 4 pixels out of 824 found

c) 0, false positive

d) 0.964

e) 0.02895

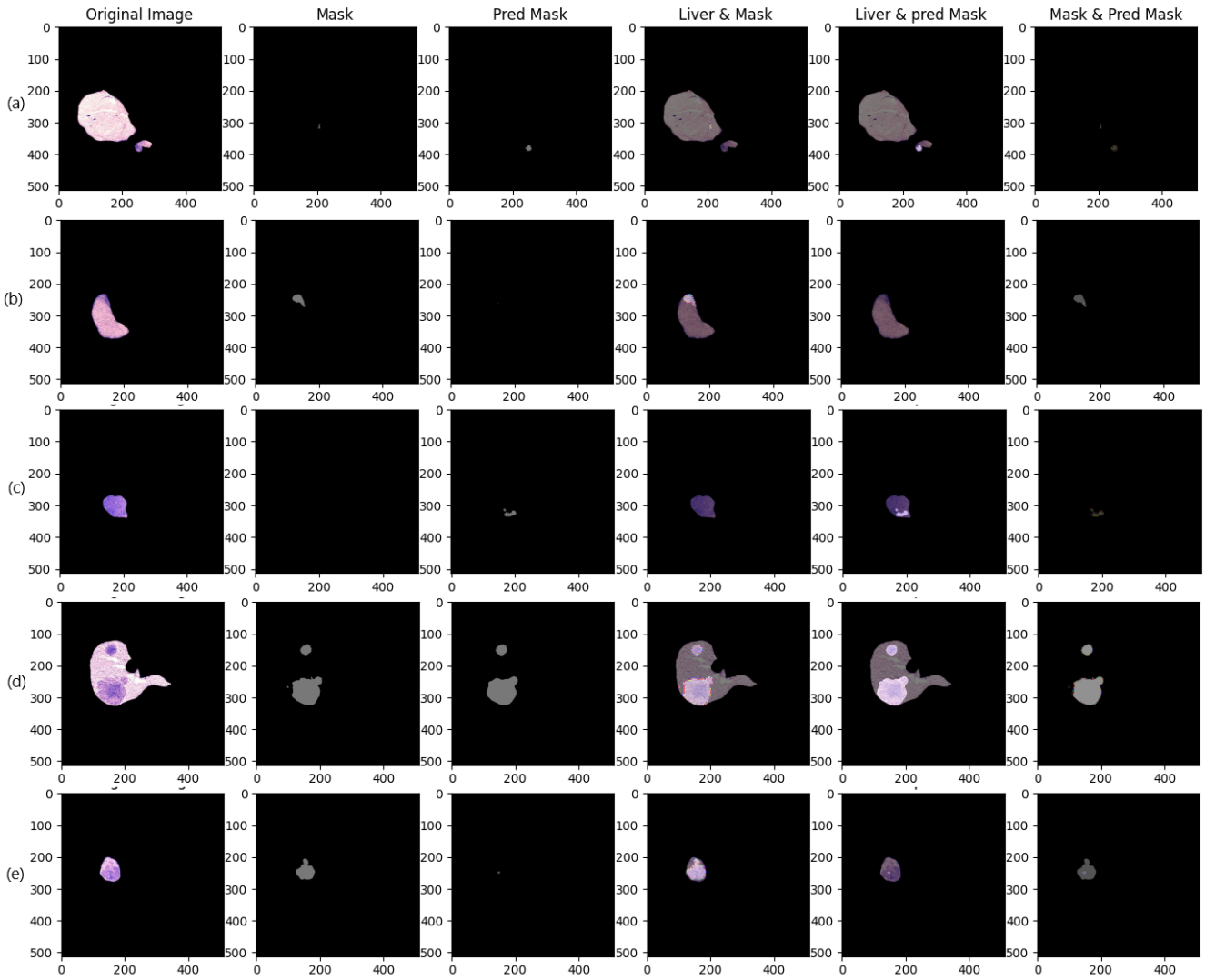



Figure 8.14: Tumor segmentations from liver masked CT scans

## 8.5 Discussion

The experimental results confirmed the challenges mentioned by other implementations and efforts using the LITS benchmark.

Liver segmentation results are better because liver size is larger compared to tumors, and due to the fact that liver is present in much more images (for testing 3053 images contain liver, while only 1356 contain tumor - the same is observed in training images. Small size tumors, especially if they are many, often result to poor performance (see section 4.3.3). If the tumor size

is a few pixels, it is difficult to segment and detect it and besides even a small deviation in segmentation results, result to a large decrease of DSC score. Furthermore, the fact that many tumor edges are not clear in liver area and the low contrast among liver and tumors, makes segmentation process more difficult.

Concerning comparison with results of other efforts, many implementations were on datasets with not many images, private, so in many cases the results cannot always be compared reliably. Even in the LITS benchmark dataset where the data are same for all, there was a possibility of submission of manually segmented images, as the organizers mentioned, and the train and test data used were not exactly the same with the submissions of LITS benchmark (see 8.1). Something worth mentioning is that liver and tumor volumes in LITS benchmark test images (70 unlabeled CTs volumes), are larger than the ones used in this work (10% of 110 train CTs volumes, see section 8.1). This fact means that if our models were trained in all LITS benchmark dataset train images (110 volumes), and tested in test images (70 unlabeled volumes), just as LITS benchmark submissions, they would possibly achieve even higher DSC scores.

The higher DSC scores achieved in this implementation for segmentation were **0.926** for liver (liver segmentation, Table 10) and **0.619** for tumors (liver and tumor segmentation, Table 7), which are good compared to other efforts, including LITS benchmark.

Tumor only segmentation efforts -either from all abdominal CTs or from liver masked CTs-, did not give as good results as the models with liver and tumor segmentation together or liver only segmentation. This occurred because liver segmentation, when performed simultaneously, is helpful in locating and segmenting tumors. Liver only segmentation gave slightly better results from liver and tumor segmentation together (higher DSC score 0.923), concerning liver segmentation.

Although LITS is one of the largest labelled public dataset concerning liver and tumor segmentation, still CT scans of 120 patients are not enough for training a DL model and to cover large number of cases. That is explained if we consider that manual liver and tumor segmentation for labeled dataset creation is a difficult and a time-consuming task, and the need for larger open access datasets is crucial and highlighted.

As future work, with the experience achieved, we plan to experiment with new DL architectures, train from scratch and utilize additional preprocessing techniques, to improve segmentation performance. Moreover, if possible, submitting to LITS benchmark with LITS

unlabeled test CT scans for more objective segmentation assessment and comparison with other efforts would be challenging.

# 9 Bibliography


[1] M. Rela, N. Suryakari, P. R.-2020 IEEE-HYDCON, and undefined 2020, "Liver tumor segmentation and classification: A systematic review," *ieeexplore.ieee.org*, Accessed: Jun. 08, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9242757/

[2] International Agency for Research on Cancer, "Global Cancer Statistics Report 2020," *https://gco.iarc.fr/today/data/factsheets/populations/900-world-fact-sheets.pdf*, Mar. 21, 2020.

[3] M. Feng, Y. Pan, R. Kong, S. S.-T. Innovation, and undefined 2020, "Therapy of primary liver cancer," *Elsevier*, Accessed: Jun. 08, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666675820300321

[4] National Cancer Institute NIH, "Liver Cancer Diagnosis," *https://www.cancer.gov/types/liver/what-is-liver-cancer/diagnosis*, Jul. 01, 2022.

[5] Y. Li, B. Zou, Q. L.-B. and B. Engineering, and undefined 2021, "A deep attention network via high-resolution representation for liver and liver tumor segmentation," *Elsevier*, Accessed: Jun. 08, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0208521621001030

[6] M. Moghbel, S. Mashohor, R. Mahmud, and M. I. Bin Saripan, "Review of liver segmentation and computer assisted detection/diagnosis methods in computed tomography," *Artif Intell Rev*, vol. 50, no. 4, pp. 497–537, Dec. 2018, doi: 10.1007/S10462-017-9550-X.

[7] J. H. Grendell, K. R. McQuaid, and S. L. Friedman, "Current diagnosis & treatment in gastroenterology," p. 867, 2003.

[8] X. Deng, G. D.-M. workshop, and undefined 2008, "3D segmentation in the clinic: a grand challenge II-liver tumor segmentation," *researchgate.net*, Accessed: Jun. 08, 2023. [Online]. Available: https://www.researchgate.net/profile/Guangwei-Du-2/publication/267699971_Editorial_3D_Segmentation_in_the_Clinic_A_Grand_Challenge_II_-Liver_Tumor_Segmentation/links/549a28f40cf2b8037135913b/Editorial-3D-Segmentation-in-the-Clinic-A-Grand-Challenge-II-Liver-Tumor-Segmentation.pdf

[9] J. H. Kang, S. H. Kim, and J. K. Han, "Poorly-differentiated colorectal neuroendocrine tumour: CT differentiation from well-differentiated neuroendocrine tumour and poorly-

differentiated adenocarcinomas," *Eur Radiol*, vol. 27, no. 9, pp. 3867–3876, Sep. 2017, doi: 10.1007/S00330-017-4764-5.

[10] S. L. luo@newcastle. edu. au$The U. of Newcastle, A. L. U. of Newcastle, A. L. C. C. Informatics, and Australia, "Review on the Methods of Automatic Liver Segmentation from Abdominal Images," *Journal of Computer and Communications*, vol. 02, no. 02, pp. 1–7, Jan. 2014, doi: 10.4236/JCC.2014.22001.

[11] D. Tsai, N. T.-P. of 1994 I. International, and undefined 1994, "Neural-network-based boundary detection of liver structure in ct images for 3-d visualization," *ieeexplore.ieee.org*, Accessed: Jul. 11, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/374895/

[12] S. S. Kumar, R. S. Moni, and J. Rajeesh, "Liver tumor diagnosis by gray level and contourlet coefficients texture analysis," *2012 International Conference on Computing, Electronics and Electrical Technologies, ICCEET 2012*, pp. 557–562, 2012, doi: 10.1109/ICCEET.2012.6203881.

[13] M. Goryawala, S. Gulec, R. Bhatt, … A. M.-B. research, and undefined 2014, "A low-interaction automatic 3D liver segmentation method using computed tomography for selective internal radiation therapy," *hindawi.com*, Accessed: Jul. 11, 2023. [Online]. Available: https://www.hindawi.com/journals/bmri/2014/198015/

[14] E. Marjolein Van Rikxoort, Y. Arzhaeva, E. Van Rikxoort, and B. Van Ginneken, "Automatic segmentation of the liver in computed tomography scans with voxel classification and atlas matching," *researchgate.net*, 2007, Accessed: Jul. 11, 2023. [Online]. Available: https://www.researchgate.net/profile/Yulia-Arzhaeva/publication/46686749_Automatic_segmentation_of_the_liver_in_computed_tomograpy_scans_with_voxel_classification_and_atlas_matching/links/0046352325d1f05d33000000/Automatic-segmentation-of-the-liver-in-computed-tomograpy-scans-with-voxel-classification-and-atlas-matching.pdf

[15] J. Zhou *et al.*, "Semi-automatic segmentation of 3D liver tumors from CT scans using voxel classification and propagational learning," *researchgate.net*, 2008, doi: 10.54294/rfkjix.

[16] W. Huang, N. Li, Z. Lin, … G. H.-2013 35th A., and undefined 2013, "Liver tumor detection and segmentation using kernel-based extreme learning machine,"

*ieeexplore.ieee.org*, Accessed: Jul. 11, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/6610337/

[17] A. M. Anter, A. Ella Hassanien, M. Abu Elsoud, M. Tolba, M. A. Abu ElSoud, and M. F. Tolba, "Neutrosophic sets and fuzzy c-means clustering for improving ct liver image segmentation," *Springer*, vol. 303, pp. 193–203, 2014, doi: 10.1007/978-3-319-08156-4_20.

[18] C. Huang, X. Li, F. J.-P. of the V. challenge at ISBI, and undefined 2014, "Automatic liver segmentation using multiple prior knowledge models and free-form deformation," *ceur-ws.org*, Accessed: Jul. 11, 2023. [Online]. Available: https://ceur-ws.org/Vol-1194/visceralISBI14-2.pdf

[19] M. Freiman, O. Eliassaf, Y. Taieb, L. Joskowicz, and J. Sosna, "A Bayesian approach for liver analysis: Algorithm and validation study," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 5241 LNCS, no. PART 1, pp. 85–92, 2008, doi: 10.1007/978-3-540-85988-8_11.

[20] A. S. Maklad *et al.*, "Blood vessel-based liver segmentation using the portal phase of an abdominal CT dataset," *Med Phys*, vol. 40, no. 11, 2013, doi: 10.1118/1.4823765.

[21] S. Gul, M. S. Khan, A. Bibi, A. Khandakar, M. A. Ayari, and M. E. H. Chowdhury, "Deep learning techniques for liver and liver tumor segmentation: A review," *Comput Biol Med*, vol. 147, p. 105620, Aug. 2022, doi: 10.1016/J.COMPBIOMED.2022.105620.

[22] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, "Deep learning for visual understanding: A review," *Neurocomputing*, vol. 187, pp. 27–48, Apr. 2016, doi: 10.1016/J.NEUCOM.2015.09.116.

[23] C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," *Electronic Markets*, vol. 31, no. 3, pp. 685–695, Sep. 2021, doi: 10.1007/S12525-021-00475-2.

[24] H. Seo, C. Huang, M. Bassenne, … R. X.-I. transactions on, and undefined 2019, "Modified U-Net (mU-Net) with incorporation of object-dependent high level features for improved liver and liver-tumor segmentation in CT images," *ieeexplore.ieee.org*, 2019, doi: 10.1109/TMI.2019.2948320.

[25] H. Li *et al.*, "3D IFPN: Improved Feature Pyramid Network for Automatic Segmentation of Gastric Tumor," *Front Oncol*, vol. 11, p. 618496, May 2021, doi: 10.3389/FONC.2021.618496/BIBTEX.

[26] P. F. Christ *et al.*, "Automatic liver and lesion segmentation in CT using cascaded fully convolutional neural networks and 3D conditional random fields," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9901 LNCS, pp. 415–423, 2016, doi: 10.1007/978-3-319-46723-8_48.

[27] E. Vorontsov, M. Cerny, P. Régnier, … L. D. J.-R. A., and undefined 2019, "Deep learning for automated segmentation of liver lesions at CT in patients with colorectal cancer liver metastases," *pubs.rsna.org*, vol. 1, no. 2, Mar. 2019, doi: 10.1148/ryai.2019180014.

[28] X. Xiao *et al.*, "Radiomics-guided GAN for Segmentation of Liver Tumor Without Contrast Agents," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11765 LNCS, pp. 237–245, 2019, doi: 10.1007/978-3-030-32245-8_27/TABLES/1.

[29] P. Bilic *et al.*, "The Liver Tumor Segmentation Benchmark (LiTS)," *Med Image Anal*, vol. 84, p. 102680, Feb. 2023, doi: 10.1016/J.MEDIA.2022.102680.

[30] O. Ronneberger *et al.*, "U-net: Convolutional networks for biomedical image segmentation," *Springer*, Accessed: Jun. 08, 2023. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-24574-4_28

[31] M. Todorov, J. Paetzold, O. Schoppe, G. Tetteh, … S. S.-N., and undefined 2020, "Machine learning analysis of whole mouse brain vasculature," *nature.com*, Accessed: Jun. 08, 2023. [Online]. Available: https://www.nature.com/articles/s41592-020-0792-1

[32] M. Y. Ansari *et al.*, "Practical utility of liver segmentation methods in clinical surgeries and interventions," *BMC Med Imaging*, vol. 22, no. 1, Dec. 2022, doi: 10.1186/S12880-022-00825-2.

[33] A. Elangovan, T. J.-2016 I. C. on, and undefined 2016, "Medical imaging modalities: a survey," *ieeexplore.ieee.org*, Accessed: Jun. 08, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7603066/

[34] M. Larobina and L. Murino, "Medical image file formats," *J Digit Imaging*, vol. 27, no. 2, pp. 200–206, 2014, doi: 10.1007/S10278-013-9657-9.

[35] "(PDF) Overview of the DICOM standard." https://www.researchgate.net/publication/224370654_Overview_of_the_DICOM_standard (accessed Jul. 23, 2023).

[36] R. D. Vincent *et al.*, "MINC 2.0: A flexible format for multi-modal images," *Front Neuroinform*, vol. 10, no. AUG, p. 206796, Aug. 2016, doi: 10.3389/FNINF.2016.00035/BIBTEX.

[37] National Institute of Health (NIH), "NIfTI: — Neuroimaging Informatics Technology Initiative," *https://nifti.nimh.nih.gov/*, Dec. 18, 2013.

[38] X. Li, P. Morgan, J. Ashburner, … J. S.-J. of neuroscience, and undefined 2016, "The first step for neuroimaging data analysis: DICOM to NIfTI conversion," *Elsevier*, Accessed: Jun. 08, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165027016300073

[39] S. Patro, K. S. preprint arXiv:1503.06462, and undefined 2015, "Normalization: A preprocessing stage," *arxiv.org*, Accessed: Jun. 08, 2023. [Online]. Available: https://arxiv.org/abs/1503.06462

[40] W. Azani Mustafa *et al.*, "Image Enhancement Based on Discrete Cosine Transforms (DCT) and Discrete Wavelet Transform (DWT): A Review," *IOP Conf Ser Mater Sci Eng*, vol. 557, no. 1, p. 012027, Jun. 2019, doi: 10.1088/1757-899X/557/1/012027.

[41] P. Jamal, M. Ali, R. H. Faraj, P. J. M. Ali, and R. H. Faraj, "Data normalization and standardization: a technical report," *researchgate.net*, vol. 1, no. 1, pp. 1–6, 2014, Accessed: Jun. 29, 2023. [Online]. Available: https://www.researchgate.net/profile/Peshawa-Muhammad-Ali/publication/340579135_Data_Normalization_and_Standardization_A_Technical_Report/links/5e91b65d299bf130798fc1bd/Data-Normalization-and-Standardization-A-Technical-Report.pdf

[42] R. C. Gonzalez and R. E. Woods, "Digital image processing," 2008, Accessed: Jun. 08, 2023. [Online]. Available: http://103.62.146.201:8081/jspui/bitstream/1/2827/1/dip-rafael.pdf

[43] F. Nielsen, R. N.-2003 I. C. S. Conference, and undefined 2003, "On region merging: The statistical soundness of fast sorting, with applications," *ieeexplore.ieee.org*, 2003, doi: 10.1109/CVPR.2003.1211447.

[44] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, “Image Segmentation Using Deep Learning: A Survey,” *IEEE Trans Pattern Anal Mach Intell*, vol. 44, no. 7, pp. 3523–3542, Jul. 2022, doi: 10.1109/TPAMI.2021.3059968.

[45] D. Guo, Y. Pei, K. Zheng, H. Yu, … Y. L.-I. T. on, and undefined 2019, “Degraded image semantic segmentation with dense-gram networks,” *ieeexplore.ieee.org*, doi: 10.1109/TIP.2019.2936111.

[46] J. Yi *et al.*, “Multi-scale cell instance segmentation with keypoint graph based bounding boxes,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11764 LNCS, pp. 369–377, 2019, doi: 10.1007/978-3-030-32239-7_41.

[47] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár, “Panoptic segmentation,” *openaccess.thecvf.com*, Accessed: Jun. 08, 2023. [Online]. Available: http://openaccess.thecvf.com/content_CVPR_2019/html/Kirillov_Panoptic_Segmentation_CVPR_2019_paper.html

[48] A. Reinke, M. Tizabi, C. Sudre, … M. E. preprint arXiv, and undefined 2021, “Common limitations of image processing metrics: A picture story,” *arxiv.org*, Accessed: Jun. 08, 2023. [Online]. Available: https://arxiv.org/abs/2104.05642

[49] A. Popovic, M. de la Fuente, M. Engelhardt, and K. Radermacher, “Statistical validation metric for accuracy assessment in medical image segmentation,” *Int J Comput Assist Radiol Surg*, vol. 2, no. 3–4, pp. 169–181, Jul. 2007, doi: 10.1007/S11548-007-0125-1/METRICS.

[50] A. A. Taha and A. Hanbury, “Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool,” *BMC Med Imaging*, vol. 15, no. 1, Aug. 2015, doi: 10.1186/S12880-015-0068-X.

[51] L. Maier-Hein *et al.*, “Erratum to: Why rankings of biomedical image analysis competitions should be interpreted with care (Nature Communications, (2018), 9, 1, (5217), 10.1038/s41467-018-07619-7),” *Nat Commun*, vol. 10, no. 1, Dec. 2019, doi: 10.1038/S41467-019-08563-W.

[52] C. Bishop and N. Nasrabadi, “Pattern recognition and machine learning,” 2006, Accessed: Jun. 08, 2023. [Online]. Available: https://link.springer.com/book/9780387310732

[53] Y. Shrestha, V. Krishna, G. von K.-J. of B. Research, and undefined 2021, “Augmenting organizational decision-making with deep learning algorithms: Principles, promises, and

challenges," *Elsevier*, Accessed: Jun. 08, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0148296320306512

[54] I. Goodfellow, Y. Bengio, and A. Courville, "Deep learning," 2016, doi: 10.4258/hir.2016.22.4.351.

[55] A. Mathew, P. Amudha, and S. Sivakumari, "Deep learning techniques: an overview," *Advances in Intelligent Systems and Computing*, vol. 1141, pp. 599–608, 2021, doi: 10.1007/978-981-15-3383-9_54/TABLES/2.

[56] M. Zeiler, "Hierarchical convolutional deep learning in computer vision," 2013, Accessed: Jun. 08, 2023. [Online]. Available: https://search.proquest.com/openview/62c046242f67ce115a76b9224e66a69c/1?pq-origsite=gscholar&cbl=18750

[57] D. Scherer, A. Müller, and S. Behnke, "Evaluation of pooling operations in convolutional architectures for object recognition," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6354 LNCS, no. PART 3, pp. 92–101, 2010, doi: 10.1007/978-3-642-15825-4_10.

[58] C. Szegedy *et al.*, "Going deeper with convolutions," *cv-foundation.org*, Accessed: Jun. 08, 2023. [Online]. Available: https://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Szegedy_Going_Deeper_With_2015_CVPR_paper.html

[59] Z. Zhou, M. R. Siddiquee, … N. T.-D. L. in, and undefined 2018, "Unet++: A nested u-net architecture for medical image segmentation," *Springer*, 2018, Accessed: Jun. 08, 2023. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-00889-5_1

[60] H. Huang *et al.*, "Unet 3+: A full-scale connected unet for medical image segmentation," *ieeexplore.ieee.org*, Accessed: Jun. 08, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9053405/

[61] J. Long, E. Shelhamer, … T. D. on computer, and undefined 2015, "Fully convolutional networks for semantic segmentation," *openaccess.thecvf.com*, Accessed: Jun. 08, 2023. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2015/html/Long_Fully_Convolutional_Networks_2015_CVPR_paper.html

[62] I. Goodfellow *et al.*, “Generative adversarial networks,” *dl.acm.org*, vol. 63, no. 11, pp. 139–144, Oct. 2020, doi: 10.1145/3422622.

[63] S. Xun *et al.*, “Generative adversarial networks in medical image segmentation: A review,” *Comput Biol Med*, vol. 140, p. 105063, Jan. 2022, doi: 10.1016/J.COMPBIOMED.2021.105063.

[64] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, “A survey on deep transfer learning,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11141 LNCS, pp. 270–279, 2018, doi: 10.1007/978-3-030-01424-7_27.

[65] B. M. Anderson *et al.*, “Automated Contouring of Contrast and Noncontrast Computed Tomography Liver Images With Fully Convolutional Networks,” *Adv Radiat Oncol*, vol. 6, no. 1, p. 100464, Jan. 2021, doi: 10.1016/J.ADRO.2020.04.023.

[66] L. Balagourouchetty, J. K. Pragatheeswaran, B. Pottakkat, and G. Ramkumar, “GoogLeNet-Based Ensemble FCNet Classifier for Focal Liver Lesion Diagnosis,” *IEEE J Biomed Health Inform*, vol. 24, no. 6, pp. 1686–1694, Jun. 2020, doi: 10.1109/JBHI.2019.2942774.

[67] M. N. Ul Haq, A. Irtaza, N. Nida, M. A. Shah, and L. Zubair, “Liver Tumor Segmentation using Resnet based Mask-R-CNN,” *Proceedings of 18th International Bhurban Conference on Applied Sciences and Technologies, IBCAST 2021*, pp. 276–281, Jan. 2021, doi: 10.1109/IBCAST51254.2021.9393194.

[68] J. Peng *et al.*, “Residual convolutional neural network for predicting response of transarterial chemoembolization in hepatocellular carcinoma from CT imaging,” *Eur Radiol*, vol. 30, no. 1, pp. 413–424, Jan. 2020, doi: 10.1007/S00330-019-06318-1.

[69] Q. Peng *et al.*, “Liver tumor segmentation and classification using FLAS-UNet++ and an improved DenseNet,” *Technology and Health Care*, vol. 30, no. 6, pp. 1475–1487, Jan. 2022, doi: 10.3233/THC-213655.

[70] A. G. Howard, “Some improvements on deep convolutional neural network based image classification,” *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*, 2014, Accessed: Jun. 08, 2023. [Online]. Available: http://code.google.com/p/cuda-convnet

[71] T. Heimann *et al.*, “Comparison and evaluation of methods for liver segmentation from CT datasets,” *IEEE Trans Med Imaging*, vol. 28, no. 8, pp. 1251–1265, Aug. 2009, doi: 10.1109/TMI.2009.2013851.

[72] L. Soler, “Liver segmentation 3D-IRCADb-01,” *https://www.ircad.fr/research/data-sets/liver-segmentation-3d-ircadb-01/*, 2010.

[73] V. V. Valindria *et al.*, “Multi-modal Learning from Unpaired Images: Application to Multi-organ Segmentation in CT and MRI,” *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, vol. 2018-January, pp. 547–556, Mar. 2018, doi: 10.1109/WACV.2018.00066.

[74] Y. J. Kim *et al.*, “PAIP 2019: Liver cancer segmentation challenge,” *Med Image Anal*, vol. 67, Jan. 2021, doi: 10.1016/J.MEDIA.2020.101854.

[75] J. Zhang *et al.*, “Light-Weight Hybrid Convolutional Network for Liver Tumor Segmentation.,” *ijcai.org*, 2019, Accessed: Jun. 08, 2023. [Online]. Available: https://www.ijcai.org/Proceedings/2019/0593.pdf

[76] Y. Yuan, “Hierarchical Convolutional-Deconvolutional Neural Networks for Automatic Liver and Tumor Segmentation,” Oct. 2017, Accessed: Jun. 08, 2023. [Online]. Available: http://arxiv.org/abs/1710.04540

[77] A. Ben-Cohen, I. Diamant, E. Klang, M. Amitai, and H. Greenspan, “Fully convolutional network for liver segmentation and lesions detection,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10008 LNCS, pp. 77–85, 2016, doi: 10.1007/978-3-319-46976-8_9/FIGURES/3.

[78] X. Tang *et al.*, “Whole liver segmentation based on deep learning and manual adjustment for clinical use in SIRT,” *Eur J Nucl Med Mol Imaging*, vol. 47, no. 12, pp. 2742–2752, Nov. 2020, doi: 10.1007/S00259-020-04800-3/FIGURES/7.

[79] T. Heimann, “SLIVER07 Challenge,” *https://sliver07.grand-challenge.org/*, Jan. 01, 2009.

[80] P. Hu, F. Wu, J. Peng, P. Liang, and D. Kong, “Automatic 3D liver segmentation based on deep learning and globally optimized surface evolution,” *Phys Med Biol*, vol. 61, no. 24, p. 8676, Nov. 2016, doi: 10.1088/1361-6560/61/24/8676.

[81] X. Chen, R. Zhang, and P. Yan, “Feature fusion encoder decoder network for automatic liver lesion segmentation,” *Proceedings - International Symposium on Biomedical Imaging*, vol. 2019-April, pp. 430–433, Apr. 2019, doi: 10.1109/ISBI.2019.8759555.

[82] G. Chlebus, A. Schenk, J. H. Moltz, B. van Ginneken, H. K. Hahn, and H. Meine, “Automatic liver tumor segmentation in CT with fully convolutional neural networks and object-based postprocessing,” *Scientific Reports 2018 8:1*, vol. 8, no. 1, pp. 1–7, Oct. 2018, doi: 10.1038/s41598-018-33860-7.

[83] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, “nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation,” *Nat Methods*, vol. 18, no. 2, pp. 203–211, Feb. 2021, doi: 10.1038/S41592-020-01008-Z.

[84] X. Han, “Automatic Liver Lesion Segmentation Using A Deep Convolutional Neural Network Method,” Apr. 2017, doi: 10.1002/mp.12155.

[85] P. Christ, “LiTS - Liver Tumor Segmentation Challenge,” *https://competitions.codalab.org/competitions/17094*, 2017.

[86] Lemonakis Georgios, “Liver tumor segmentation, implementation code and results,” *https://github.com/glemon1/LiverTumorSegmentation*, Aug. 06, 2023.